ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

Carla
Díaz Louzao

Tese de doutoramento

Applying Joint Modelling Regression Approaches in Biomedical Data Science

Santiago de Compostela, 2022

Programa de Doutoramento en Epidemioloxía e Saúde Pública

DOCTORAL THESIS

APPLYING JOINT MODELLING REGRESSION APPROACHES IN BIOMEDICAL DATA SCIENCE

Carla Díaz Louzao

INTERNATIONAL PHD SCHOOL OF THE UNIVERSITY OF SANTIAGO DE COMPOSTELA

PHD PROGRAMME IN EPIDEMIOLOGY AND PUBLIC HEALTH

SANTIAGO DE COMPOSTELA

2022

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DO AUTOR/A DA TESE

D./Dna. **Carla Díaz Louzao**

Título da tese: **Applying Joint Modelling Regression Approaches in Biomedical Data Science**

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.
4) A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En **Santiago de Compostela**, **14 de Setembro de 2022**.

Sinatura:

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DO DIRECTOR DA TESE

Applying Joint Modelling Regression Approaches in Biomedical Data Science

D. Francisco Gude Sampedro

INFORMA:

Que a presente tese, correspóndese co traballo realizado por Dna. Carla Díaz Louzao, baixo a miña dirección, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reprodución de publicacións, nas que a participación da doutoranda foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En **Santiago de Compostela, 05 de Outubro de 2022**

Francisco Gude Sampedro

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DO TITOR DA TESE

Applying Joint Modelling Regression Approaches in Biomedical Data Science

D. Adolfo Figueiras Guzmán

INFORMA:

Que a presente tese, correspóndese co traballo realizado por Dna. Carla Díaz Louzao, baixo a miña titorización, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como titor desta non incorre nas causas de abstención establecidas na Lei 40/2015.

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reprodución de publicacións, nas que a participación da doutoranda foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En **Santiago de Compostela, 05 de Outubro de 2022**

Adolfo Figueiras Guzmán

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecementos-Acknowledgments

A realización desta tese foi de todo menos doada, chegando a haber momentos nos que o abandono da mesma era unha posibilidade moi real, case un feito. Pero este proxecto que comezou xa hai case seis anos chega agora ao seu desexado desenlace. Isto non sería posible sen todas as persoas que me axudaron, apoiaron e acompañaron durante esta etapa crucial da miña vida. Este traballo tamén é voso.

A Francisco Gude, Pachicho, teño que agradecerlle que finalmente me convecese de emprender este camiño baixo a súa dirección e a de Carmen Cadarso. Moitas gracias por toda a axuda, todo o apoio e toda a confianza depositada en min.

Carmen é, sen dúbida, a persoa á que máis lle teño que agradecer. Dende un inicio me apoiou e creu en min, incluso cando nin eu mesma o facía. Sempre me animou e non me deixou abandonar cando as cousas se puxeeron de costas, apostando por min e polo meu traballo. Ela entendía que a estatística debía traspasar o plano teórico para poñerse ao servicio da comunidade, visión que eu tamén comparto, e desenvolveu toda a súa carreira traballando cóbado con cóbado cos clínicos para logralo. Ela tiña moita fe neste proxecto de tese tan interdisciplinario (feito por unha estatística nun programa da rama médica) e prodúceme un fondo pesar que non o puidese chegar a ver rematado. Gústame pensar que estaría moi orgullosa.

Gracias ao programa de doutoramento en Epidemioloxía e Saúde Pública e, en especial, ao meu titor Adolfo Figueiras e ao director da CAPD Agustín Montes, por permitirme desenvolver este proxecto de tese tan estatístico e tan pouco habitual dentro da rama médica. Gracias tamén por toda a axuda proporcionada, sobre todo na etapa final.

Gustaríame agradecer tamén a Vicente e María, de Biostatech. Dende o primeiro momento me deron a mellor acollida e o seu apoio e axuda incondicionais. Tanto foi así que a idea inicial foi que realizase o doutoramento dentro da empresa, pero iso non foi posible. Quero destacar que, de non ser por eles, non tería sido capaz de resolver o papeleo de todo un ano en tan só dúas semanas, podendo así solicitar a bolsa predoutoral da Xunta de Galicia, cuxa concesión supoñía a miña marcha da empresa. E unha vez me fun (non sen pena), seguíronme apoiando e axudando a pesar da consecuente disminución do contacto.

USC
DE SANTIAGO
DE COMPOSTELA

A concesión da bolsa predoutoral foi, por outra banda, como un regalo do ceo, xa que supuxo a miña entrada no GRID-BDS e na Unidade de Bioestatística da Facultade de Medicina. A xente coa que tiven a sorte de toparme aquí non só foron os meus compañeiros ao longo desta aventura, senón que se convirtiron en grandes amigos aos que levarei sempre comigo. Ana Bouzas, sen a cal estariamos todos perdidos na inmensidade da burocracia, sempre cunha palabra de apoio e disposta a axudar no que sexa. Berta, capaz de sacarme un sorriso ata nos momentos máis difíciles. Jeni, sen cuxa orientación e axuda, incluso despois de ela marchar do grupo para seguir o seu propio camiño na docencia, non tería saído adiante. Óscar, un dos meus grandes apoios e compañeiro de viaxes, que pouco a pouco me ensinou a tomarme as cousas con filosofía e máis tranquilidade, evitando así tolear completamente. Ipek, que me axudou moito na miña incursión no difícil mundo dos JMLS. Juan, o meu compañeiro na última etapa. E Ana Cao, que non é tecnicamente do grupo, pero cuxa presencia e apoio me permitiu sobrelevar os nervios propios deste final de camiño. A todos eles, moitísimas gracias de todo corazón.

To Nadja Klein, thanks a lot for hosting me in Berlin and for her help and guidance in the boosting for JMLS. It was a pity that the change of PhD programme meant that I had to give up this very interesting topic and, above all, the opportunity to continue working with her and learning from her.

Aos investigadores do CHUS cos que tiven a sorte de colaborar, especialmente a Santi Tomé, sen cuxa axuda, esforzo e paciencia sería imposible este traballo.

I have to thank the members of Riseup-PPD, and especially Raquel Costa, Ana Mesquita and Emma Motrico, for the opportunity to participate in a project of such magnitude and for the affection and trust placed in me and my work.

Recordo tamén con especial cariño a Amanda, quen fixo da miña estancia en Berlín unha experiencia inolvidable. Cada congreso no que me atopo con ela mellora enormemente de forma inmediata.

Xa para rematar, téñolles moito que agradecer ás miñas persoas máis queridas. Ás miñas amigas Noe, Míriam e Nuria, por sempre estar aí animándome e aturándome (o cal non debe ser fácil). A Antonio, cuxa chegada trouxo luz á miña vida, e que dende o primeiro momento estivo ao meu lado incansablemente. A miña nai, meu pai e meu irmán, por apoiarme e animarme de forma sincera e incondicional. Non sería quen son hoxe sen todos eles. Espero poder chegar a recompensalos algún día por todo o que me aportan.

A todos eles, sempre estarei infinitamente agradecida.

Carla Díaz Louzao
Santiago de Compostela, Setembro 2022

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Fundings: This thesis forms part of the *"Novo modelo de vida saudable"* (New model of healthy living) challenge, focusing on the area of *"prevención, diagnose e tratamento de enfermidades"* (disease prevention, diagnosis and treatment). This challenge is financed by the programme *"Axudas de apoio á etapa predoutoral"* (Aid to the Predoctoral Stage) of the Galician Regional Authority (*Consellería de Cultura, Educación e Ordenación Universitaria, Xunta de Galicia*) and European Social Fund 2014/2020, which provided a grant to the present author in 2017 (*ED481A-2017/130*).

This research was also supported by the projects *COV20/00404: "Pronóstico de los pacientes infectados por el nuevo coronavirus SARS-CoV-2"*, and *MTM2017-83513-R: "Flexible Joint Modelling in diabetes, cardiovascular disease and cancer"*, and co-financed by the Carlos III Health Institute, Ministry of Economy and Competitiveness (Spain) and the European Regional Development Fund (ERDF).

This work was also supported by grants from the Galician Government: *Rede INBIOEST: "Novas Ferramentas Estatisticas e Computacionais Aplicadas á Investigación en Saúde, Deporte e Medio Ambiente" (ED341D-R2016/032)*, *Grupo de Referencia Competitiva (ED431C 2016-025)*, *Grupo de Potencial Crecimiento (IN607B 2018/1)*, and *"Grupo de Bioestadística e Ciencia de Datos Biomédicos (GRID-BDS)- GI2127" (ED431C 2020/20)*.

This PhD thesis was also supported by the Spanish Ministry of Health, by grants from the Carlos III Institute of Health (*Instituto de Salud Carlos III-ISCIII/ PI20/01069/* Co-funded by European Union), the *"Red de Investigación en Cronicidad, Atención Primaria y promoción de la salud"* (*Instituto de Salud Carlos III-ISCIII/ RD21/0016/0022/* Co-funded by European Union), and the *"Axencia Galega de Innovación - Grupos de referencia competitiva" (GAIN-GRC/IN607A/2021/02/Xunta de Galicia)*.

Finally, a part of this project falls within the framework of the COST Action *CA18138 - Research Innovation and Sustainable Pan-European Network in Peripartum Depression Disorder (Riseup-PPD)* and was supported by COST under COST Action Riseup-PPD CA18138.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abstract

The research work carried out in this thesis is based on the collaboration of two research groups: a biostatistics group (*Group in Biostatistics and Biomedical Data Science, GI-2127: GRID-BDS*) belonging to the University of Santiago de Compostela (USC), and a clinical medicine group (*Research Methods Group, C017*) belonging to the *Instituto de Investigación Sanitaria de Santiago de Compostela* (IDIS). These groups have been coordinating for years to carry out interdisciplinary research in the field of biostatistics. This collaboration is particularly noteworthy in the current era, in which technology makes it possible to collect huge amounts of all kinds of variables, thus creating enormously complex databases that pose a challenge when it comes to processing, analysing and interpreting them. This complexity of data requires the collaboration of statisticians and clinicians to be able to carry out a biomedical research study with guarantees, applying data science tools.

In this thesis, biostatistical models for multivariate response are studied. In general, in this type of modelling, the different response variables are related to each other, so that the individual analysis of each one can lead to biased results. Moreover, sometimes the importance of the study lies not only in the modelling of the responses, but also in their correlation. It is in this context that joint modelling is born and, despite their great usefulness, the use of these techniques is not yet widespread in biomedicine. Consequently, the main objective of this thesis is to bring this type of models closer to clinical practice, and to future epidemiological studies. To this end, we have real databases on liver damage in patients with COVID-19, perinatal mental health during the COVID-19 pandemic, and thyroid-related hormones in a healthy adult population.

Specifically, we start with a review of joint distributional modelling for bivariate response, following the methodology of Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS), introduced by Marra and Radice (2017). These models allow the distribution of two response variables to be fully determined by flexibly modelling each parameter of the response variable (not just its mean) as a function of other covariates of interest. They also allow modelling of the correlation between the two responses. This methodology is applied in the study of

USC
DE SANTIAGO
DE COMPOSTELA

the correlation between transaminases and inflammation markers in patients with COVID-19, as well as in the study of the influence of the measures taken by governments during the pandemic on perinatal mental health.

More recently, Multivariate Conditional Transformation Models (MCTM; Klein et al., 2019) have emerged as a more flexible and faster alternative to CGAMLSS models. However, the software for its implementation is still under development, and its implementation is limited to a single explanatory variable. We yapply this methodology to data from healthy patients to determine the joint distribution of the three main thyroid-related hormones (TSH, free T3 and free T4), as well as their variation with patient age. The aim is that it can be used as an aid in the diagnosis of thyroid diseases.

We should also highlight a specific type of joint modelling whose response variables are, on the one hand, longitudinal markers (repeated measures) and, on the other hand, one or several events of interest (e.g. death and the development of a certain disease) and the time to these events. In longitudinal studies, in which the values of some markers are collected repeatedly over time, it is very common to find variables of this kind. To study the association between longitudinal and survival processes, it is necessary to analyse them jointly by treating the longitudinal markers as response variables, and then incorporating their corresponding estimates into the survival model as covariates. Joint regression models for longitudinal and survival data (JMLS; Tsiatis and Davidian, 2004) have emerged precisely for this type of analysis. In this thesis we introduce the most recent and complete extension of these models (Mauff et al., 2020), illustrating it in practice by analysing the influence of three transaminases (AST, ALT and GGT), collected over time, on survival in patients with COVID-19.

Finally, in order to bring these models closer to clinical practice, we provide a tutorial with the necessary code for the analyses carried out during the course of this doctoral thesis.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumo

O traballo de investigación realizado na presente tese parte da colaboración de dous grupos de investigación: un de bioestatística (*Grupo de Investigación e Ciencia de Datos Biomédicos, GI-2127: GRID-BDS*) pertencente á Universidade de Santiago de Compostela (USC) e coordinado pola profesora Carmen Cadarso, e outro de medicina clínica (*Metodoloxía da Investigación, C017*) pertencente ao Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS) e coordinado polo doutor Francisco Gude. Estes grupos coordínanse dende hai anos para levar a cabo investigación interdisciplinaria no ámbito da bioestatística. Esta colaboración é especialmente destacable na era actual, na que a tecnoloxía permite recompilar cantidades inxentes de todo tipo de variables, creándose así bases de datos enormemente complexas (incluíndo *big data*) que supoñen todo un reto á hora de procesalas, analizalas e interpretalas de forma correcta. Esta complexidade nos datos fai necesaria a colaboración de estatísticos e clínicos para poder levar a cabo con garantías un estudo de investigación biomédico, aplicando ferramentas de ciencia de datos.

O termo "Ciencia de Datos" ou *Data Science* foi acuñado hai xa seis décadas polo estatístico John Tuckey (Tukey, 1962) para describir un campo de estudo emerxente cuxo principal interese era aprender directamente dos datos. Nos últimos anos, esta área foi gañando presenza debido á xa mencionada explosión tecnolóxica, que permite a xeración e colección de enormes cantidades de datos de diversa natureza en diferentes ámbitos. Aínda que non hai unha definición universalmente aceptada para a Ciencia de Datos Biomédicos (BDS polas súas siglas en inglés), pódese aplicar ao caso das bases de datos biomédicas a definición dada por *USA National Institutes of Health*, quen definen a ciencia de datos moderna como "o campo interdisciplinario de investigación no que se desenvolven e utilizan enfoques, procesos e sistemas cuantitativos e analíticos para extraer coñecementos e ideas de conxuntos de datos cada vez máis grandes e/ou complexos." (National Institutes of Health (NIH), 2018).

O carácter interdisciplinario da BDS queda claramente establecido ao ter en conta que esta disciplina científica xurdiu como conxunción das áreas de Estatística, Informática e Saúde, permitindo aos analistas de datos obter coñecementos a partires

USC DE SANTIAGO DE COMPOSTELA

de datos biomédicos complexos. Resulta obvio que as competencias requiridas para ser un científico de datos non son fáciles de atopar nunha soa persoa. Isto leva inevitablemente á necesidade de crear grupos de traballo multidisciplinares formados por diferentes especialistas das distintas áreas implicadas, que combinen as súas habilidades e coñecementos para levar a cabo o proceso da BDS. Este proceso consta das seguintes seis fases:

(i) Establecemento as preguntas de investigación ás que se lles quere dar resposta da forma máis clara posible.

(ii) Identificación de todas as bases de datos dispoñibles.

(iii) Realización dunha limpeza previa dos datos.

(iv) Análise exploratoria dos datos.

(v) Análise estatística en profundidade dos datos nos casos nos que sexa necesario.

(vi) Comunicación dos resultados obtidos á comunidade científica.

Aínda que poida parecer que cada unha das fases do proceso está orientada cara a un perfil específico de todos os que compoñen a BDS, requírese unha combinación dos diferentes perfís en cada unha delas. De novo, evidénciase a necesidade dun grupo interdisciplinario cuxos membros interactúen entre eles durante todo o proceso.

Para poder levar a cabo este proceso con total garantía, precísanse utilizar as técnicas estatísticas máis adecuadas. A análise estatística está en constante evolución e cada vez xorden novas metodoloxías que permiten analizar mellor os complexos e diversos datos dispoñibles, de forma que se pode extraer información máis precisa e facer predicións más exactas.

Nesta tese estúdanse modelos bioestatísticos de resposta multivariante, nos cales a importancia do estudo non só reside na modelización das respostas, senón tamén da súa correlación. Este tipo de estudos nos que as diferentes variables resposta están relacionadas entre si son moi comúns en investigación biomédica, xa que habitualmente se obteñen varias medicións de distintas variables para o mesmo paciente. As estratexias habituais de análise consisten en tratar estas variables resposta como independentes entre si, ou aplicar unha transformación preestablecida que dea lugar a un problema univariante. Estes enfoques poden depreciar a información útil que xorde da distribución conxunta das medicións continuas e levar a resultados nesgados, de xeito que se obteñen predicións menos precisas (tanto individuais como doutro tipo). É neste contexto onde nace a modelización conxunta (JM polas

USC
DE SANTIAGO
DE COMPOSTELA

súas siglas en inglés), cuxo uso aínda non está moi estendido en biomedicina a pesar da súa grande utilidade, xa que permite mellorar a aplicación da regresión multivariante para enfermidades cuxo diagnóstico se basea en varios biomarcadores continuos, ou avaliacións clínicas que se basean en medicións correlacionadas. Debido a isto, o obxectivo principal desta tese de doutoramento é achegar este tipo de modelos á práctica clínica, e a futuros estudos epidemiolóxicos. Para iso, contamos con bases de datos reais relativas a problemas epidemiolóxicos de grande importancia e actualidade.

Durante a elaboración desta tese, apareceu a pandemia de COVID-19, declarada emerxencia de saúde pública pola Organización Mundial da Saúde (WHO polas súas siglas en inglés). Segundo o informe do 21 de setembro de 2022, o número acumulado de casos confirmados reportado a nivel mundial supera xa os 609 millóns, e o número acumulado de decesos está por riba dos 6,5 millóns de persoas (World Health Organization (WHO), 2022). Debido a isto, a maioría das contribucións incluídas neste documento están altamente relacionadas con esta enfermidade. En concreto, estudamos por unha banda o dano hepático en pacientes con COVID-19 ingresados no Complexo Hospitalario Universitario de Santiago de Compostela durante a primeira vaga da enfermidade, ocorrida entre marzo e xuño de 2020. Por outra banda, realizamos un estudo para ver a como a pandemia e as diferentes medidas implementadas polos países para intentar paliala afectaron a saúde mental perinatal (ansiedade e depresión en mulleres embarazadas ou con fillos menores de 6 meses) durante a pandemia de COVID-19 en 12 países, 9 deles europeos, entre a primeira e a segunda vaga da enfermidade. Este estudo enmárcase dentro do proxecto europeo "*Research Innovation and Sustainable Pan-European Network in Peripartum Depression Disorder - Riseup-PPD*" (Cost Action 18138). Finalmente, aplicáronse tamén as técnicas JM para un estudo non relacionado coa COVID-19, pero de grande interese epidemiolóxico: a determinación da distribución das hormonas tiroideas en poboación adulta sa en relación coa idade dos pacientes.

Especificamente, empezamos cunha revisión da modelización conxunta distribucional para respostas bivariadas, seguindo a metodoloxía dos modelos cópula aditivos xeralizados para localización, escala e forma (CGAMLSS), introducida por Marra and Radice (2017). Estes constitúen a extensión ao campo das respostas bivariadas dos modelos aditivos xeralizados para localización, escala e forma (GAMLSS; Rigby and Stasinopoulos, 2005), os cales se centran en regresión distribucional, de xeito que se pode modelizar de forma flexible o efecto das covariables en todos os parámetros da distribución da resposta (por exemplo, a varianza), en lugar de limitarnos soamente ao estudo da media. Ademais, os CGAMLSS permiten á súa vez modelizar, tamén de forma flexible, a correlación entre as dúas respostas a través das funcións cópula.

Na literatura existen diversas cópulas que permiten diferentes estruturas de de-

USC DE SANTIAGO DE COMPOSTELA

pendencia non Gaussianas (non normais), engadindo máis flexibilidade e precisión ao modelo. Por exemplo, existen funcións cópula que permiten considerar estruturas de dependencia asimétricas para o caso no que as dúas variables aleatorias mostren unha asociación positiva máis forte para valores pequenos que grandes (cópula Clayton) ou viceversa (cópulas Joe ou Gumbel), sendo posible tamén considerar versións rotadas das mesmas para modelizar estruturas de correlación negativa. Por outra banda, existen tamén funcións cópula para o caso non que non se observe ningunha dependencia nos extremos (por exemplo, as cópulas Gaussiana ou a Frank). Nos modelos cópula, a asociación entre as variables resposta preséntase habitualmente en termos do coeficiente de asociación τ de Kendall. Este estimador toma valores entre -1 e 1, podéndose facer unha interpretación similar do mesmo que cando se usa o coeficiente de correlación linear de Pearson: canto máis preto de 0, menor é a correlación.

A metodoloxía CGAMLSS aplícase en dous estudos derivados da pandemia de COVID-19. En primeiro lugar, estúdase a correlación ao longo do tempo entre transaminasas (AST, ALT, GGT e ALP) e marcadores de inflamación (PCR, Dímero-D, IL-6 e linfocitos) en pacientes con esta enfermidade ingresados no Complexo Hospitalario Universitario de Santiago de Compostela durante a primeira vaga, entre marzo e xuño de 2020. A motivación deste estudo radica en que, a pesares de que o principal órgano afectado son os pulmóns, viuse que o dano hepático tamén é frecuente durante a infección (Richardson et al., 2020; Phipps et al., 2020; Popkin et al., 2020), sendo o mecanismo exacto causante deste dano hepático descoñecido. Existen varias hipóteses (Marjot et al., 2021), como son o efecto da resposta inflamatoria, os fármacos administrados, anoxia tisular ou unha acción directa do virus. A implementación dos modelos CGAMLSS neste estudo permitiu ver unha dinámica diferente nos pacientes que sobreviviron e nos que faleceron, sobre todo tras a segunda semana despois do ingreso, momento a partir do cal xa non hai (apenas) acción directa do virus. Tendo en conta que a inmensa maioría dos pacientes foron tratados con antivirais (lopinavir-ritonavir) e antimaláricos (hidroxicloroquina), pero non con fármacos inmunosupresores como corticoesteroides, os resultados obtidos dan pé a pensar que a explicación para o potencial mecanismo de dano hepático pode ser a propia resposta inflamatoria inducida.

Por outra banda, aplícanse as técnicas CGAMLSS para analizar o papel das medidas gobernamentais e factores individuais en síntomas de ansiedade e depresión en mulleres en período perinatal (embarazadas ou con fillos de menos de 6 meses), considerando o Índice de Desenvolvemento Humano axustado por desigualdade (IHDI polas súas siglas en inglés) e outros factores individuais como a idade ou a exposición a COVID-19. Para isto cóntase con datos procedentes de enquisas *online* a mulleres maiores de idade en período perinatal residentes en 12 países entre xuño e outubro de 2020. Para medir os niveis de ansiedade e depresión úsanse os

USC
DE SANTIAGO
DE COMPOSTELA

tests *Generalised Anxiety Disorder Screener* (GAD-7; Spitzer et al., 2006) e *Edinburgh Postnatal Depression Scale* (EPDS; Cox et al., 1987) respectivamente. Como indicador da resposta gobernamental á pandemia emprégase a media dos 30 días previos á enquisa do Índice de Contención e Saúde (CHI polas súas siglas en inglés; Hale et al., 2020). Ademais, inclúese tamén o número de mortes por millón no momento de realización da enquisa como indicador do impacto da pandemia na poboación de cada país. Deste estudo destácase que o número relativo de mortes debido á pandemia en cada país actúa como factor de risco para unha mala saúde mental perinatal, mentres que un maior CHI actúa como factor protector, indicando que as mulleres se senten más protexidas canto máis duras son as medidas implementadas polos gobernos para intentar frear a pandemia e paliar os seus efectos.

Dous anos despois da aparición dos modelos CGAMLSS, no 2019, xurdiron os modelos de transformación condicional multivariantes (MCTM; Klein et al., 2019), unha alternativa máis flexible e rápida aos modelos anteriores, que ademais se adapta de forma inmediata a situacións con máis de dúas variables resposta. Porén, o *software* para a súa implementación aínda está en desenvolvemento, e o seu uso vese limitado a unha soa variable explicativa.

Aplicamos esta metodoloxía a datos de pacientes sans para determinar a distribución conxunta das tres principais hormonas tiroideas (TSH, T3 libre e T4 libre), así como a súa variación coa idade dos pacientes. O obxectivo é que poida servir de axuda no diagnóstico de enfermidades tiroideas. Isto é de vital importancia debido a que a disfunción tiroidea é unha enfermidade común que trae asociados un gran número de diferentes síntomas, como a fatiga, unha maior sensibilidade ao frío, unha redución da memoria e da concentración, dispnea, ganancia de peso ou incluso depresión. A glándula tiroides, cuxo funcionamento está controlado pola glándula pituitaria a través da hormona TSH, é a encargada da regularización do metabolismo, polo que o seu funcionamento incorrecto é daniño para todo o corpo, de aí a grande importancia dun correcto diagnóstico. Como resultado, podemos ver un diferente comportamento das tres hormonas en función da idade. A T3 libre diminúe coa idade, sendo esta diminución máis acusada a partires dos 70 anos; a T4 libre mostra unha forma de "U", de forma que os valores máis altos se presentan especialmente en individuos ata 30 anos, e tamén a partir dos 70 anos; en cambio, a TSH non parece verse influída pola idade do suxeito. En canto á correlación entre as respostas (que, a diferencia do CGAMLSS, vén dada polo coeficiente de correlación de rangos de Spearman, ρ), podemos ver que a relación entre a TSH e a T3 libre é negativa e estatisticamente significativa para individuos entre os 30 e os 50. Ademais, a correlación entre a TSH e a T4 libre é negativa e a correlación entre a T3 e a T4 libres é positiva e decrece coa idade.

Como última metodoloxía debemos destacar un tipo específico de modelización conxunta cuxas variables resposta son, por un lado, marcadores lonxitudinais (me-

USC DE SANTIAGO DE COMPOSTELA

didas repetidas) e, por outro, un ou varios eventos de interese (por exemplo, a morte ou o desenvolvemento dunha certa enfermidade) e o tempo a ditos eventos. Nos estudos lonxitudinais, nos que se adoitan recoller os valores dalgúns marcadores de forma repetida ao longo do tempo no mesmo paciente, é moi común atoparnos con variables desta clase. Para estudar a asociación entre os procesos lonxitudinais e de supervivencia, é preciso analizalos de forma conxunta tratando os marcadores lonxitudinais como variables resposta, para despois incorporar as súas correspondentes estimacións no modelo de supervivencia como covariables. Os modelos de regresión conxunta para datos lonxitudinais e de supervivencia (JMLS; Tsiatis and Davidian, 2004) xurdiron precisamente para este tipo de análises. Este tipo de JM é o máis estendido tanto na práctica como na literatura, onde continuamente xorden novas extensións e avances dos mesmos. Nesta tese facemos especial fincapé na máis recente e completa extensión deste modelos introducida por Mauff et al. (2020).

A primeira forma de modelizar este tipo de datos é cos coñecidos como modelos en dúas fases ou *two-stage* (Self and Pawitan, 1992), nos que os dous procesos son analizados de forma separada, primeiro o lonxitudinal (xeralmente implementando un modelo linear mixto, LMM) e logo o de supervivencia (a través dun modelo de Cox de riscos proporcionais; Cox, 1972). Esta aproximación ten vantaxes computacionais, porén leva a resultados nesgados debido á dependencia dos datos lonxitudinais do evento en cuestión, así como do tempo ao mesmo. Os JMLS toman esta idea, pero introducindo unha distribución conxunta na que a estrutura de dependencia entre os dous procesos se define a través de efectos aleatorios latentes comúns, téndose así os chamados modelos de efectos aleatorios compartidos.

Estes modelos así descritos poden ser restritivos, no sentido de que só permiten unha variable lonxitudinal continua e un só evento terminal, é dicir, sen posibilidade de repetición. Con respecto ás respostas lonxitudinais, xurdiron novos avances que permiten considerar outro tipo de respostas, tales como binarias ou outras respostas continuas non Gaussianas, axustando un modelo linear xeralizado mixto (GLMM) no canto dun LMM. Ademais, estas extensións permiten incluír no modelo varios marcadores lonxitudinais a un mesmo tempo, simplemente asumindo unha distribución normal multivariante para os efectos aleatorios e incorporando no preditor linear do modelo de supervivencia os valores esperados de cada lonxitudinal. A última extensión presentada por Mauff et al. (2020) permite, ademais de todo o anterior, considerar calquera tipo de censura nos datos (censura pola dereita (a máis habitual), pola esquerda e incluso censura por intervalos), así como a inclusión de riscos competitivos ou modelos multiestado, onde se conta con máis dun evento de interese.

Con esta nova extensión, gáñase unha enorme flexibilidade, comparándoa cos JMLS básicos. Esta nova extensión é matematicamente directa, pero extremada-

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

mente custosa computacionalmente. Isto levou a abordar o problema dende unha perspectiva Bayesiana utilizando cadeas de Markov, o cal mellora o problema pero non o atalla completamente. Por iso, a solución práctica máis comunmente implementada é o uso dunha aproximación en dúas fases, primeiro axustando os modelos mixtos multivariantes e, a continuación, utilizando os resultados obtidos no modelo de supervivencia. Como isto pode levar a estimacións nesgadas, aplícase un factor de corrección baseado en teoría de mostraxe de importancia (*importance sampling*) para rectificar ese posible erro de estimación.

A implementación destes modelos na práctica clínica ten como un dos seus obxectivos principais a realización de predicións precisas das probabilidades de supervivencia para poder axustar o tratamento médico. A este respecto, para un paciente en concreto e para un tempo t específico durante o seguimento do mesmo, é factible usar toda a información dispoñible ata ese tempo t determinado, de forma que se poida ir actualizando a predición a medida que se vai obtendo nova información lonxitudinal. Estas predicións individualizadas, coñecidas como predicións dinámicas, poden xogar un papel destacado na medicina personalizada, constituíndo un dos grandes atractivos destes modelos.

Así como tamén ocorre nos modelos habitualmente usados na investigación biomédica, existen medidas de discriminación baseadas nas curvas ROC (*Receiver Operating Characteristic*; Heagerty and Zheng, 2005), as cales poden ser moi útiles para avaliar a capacidade preditiva do modelo de supervivencia. Neste caso, ao tratarse de variables dependentes do tempo, a sensibilidade e a especificidade deben calcularse para cada momento. Ademais, como as predicións dos JMLS teñen unha natureza dinámica, Rizopoulos (2011) introduciu novas definicións destes parámetros, de forma que se ten en conta non só o momento no que se está facendo a predición, se non tamén o tempo horizonte para o que se quere predicir, existindo unha curva ROC para cada variación destes dous tempos, así como a súa área baixo a curva ou AUC asociada (canto máis preto de 1, maior capacidade discriminativa ten o modelo).

Implementamos a última extensión de JMLS para a análise da influencia de tres transaminasas (AST, ALT e GGT), recollidas ao longo do tempo, na supervivencia de pacientes ingresados con COVID-19 durante a primeira vaga da pandemia en Santiago de Compostela. Este estudo vén motivado polas diferencias na evolución das transaminasas atopadas entre os pacientes que faleceron e os que non no primeiro estudo descrito nesta tese. Grazas a estes modelos, podemos ver que tanto AST como ALT parecen influír de forma significativa na supervivencia dos pacientes, pero de forma oposta (a maior AST maior risco de falecemento, mentres que o contrario ocorre para ALT). GGT non mostrou efecto significativo no risco de falecemento dos suxeitos de estudo. A presencia de diabetes, incluída como marcador non lonxitudinal no modelo de supervivencia, tamén constitúe un factor de risco de

USC DE SANTIAGO DE COMPOSTELA

exitus en pacientes con COVID-19.

Finalmente, e co obxectivo de achegar estes modelos á práctica clínica, proporcionamos un titorial co código necesario para as análises realizadas durante o transcurso da presente tese de doutoramento.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

List of Figures **xix**

List of Tables **xxiii**

1 Introduction **1**
1.1 Biomedical Data Science (BDS) 1
1.2 Joint Modelling (JM) Regression 4
1.2.1 General Joint Modelling Framework 6
1.2.2 Copula Distributional Regression Models 9
1.2.3 Multivariate Conditional Transformation Models (MCTM) . . 9
1.2.4 Joint Modelling for Longitudinal and Survival Data (JMLS) . 10
1.3 General Objectives of the Thesis 12
1.4 Structure of the Thesis 13

2 Motivating Research Questions **15**
2.1 Liver Injury in COVID-19 (Liver-19) Data 15
2.1.1 Ethical Issues 17
2.1.2 Descriptive Analysis of Liver-19 Database 17
2.2 Peripartum Anxiety and Depression (PPAD) Data 23
2.2.1 Ethical Issues 24
2.2.2 Descriptive Analysis of PPAD Database 24
2.3 Thyroid-Related Hormones (TRH) Data 27
2.3.1 Ethical Issues 28
2.3.2 Descriptive Analysis of TRH Database 29

3 Copula Regression Models for Bivariate Responses **31**
3.1 Introduction 31
3.2 Bivariate Copula Models 32
3.2.1 Model Formulation 33
3.2.2 Estimation and Inference 35

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.3 Bivariate Joint Modelling of Liver-19 Data . . . 36
3.3.1 Motivation of the Study . . . 36
3.3.2 Statistical Analysis of Liver-19 Data . . . 38
3.3.3 Discussion . . . 47
3.4 Bivariate Joint Modelling of PPAD Data . . . 51
3.4.1 Motivation of the Study . . . 51
3.4.2 Statistical Analysis of PPAD Data . . . 52
3.4.3 Discussion . . . 61

4 An Alternative Approach to Copula JM: Multivariate Conditional Transformation Models (MCTM) 63
4.1 Introduction . . . 63
4.2 Multivariate Conditional Transformation Models . . . 64
4.2.1 Model Formulation . . . 64
4.2.2 Estimation and Inference . . . 67
4.3 Simulation Study for the Bivariate Case . . . 67
4.3.1 Simulation Design . . . 68
4.3.2 Evaluation of the Performance . . . 68
4.3.3 Results of the Simulation Study . . . 69
4.4 Trivariate Joint Modelling of TRH Data . . . 70
4.4.1 Motivation of the Study . . . 70
4.4.2 Statistical Analysis of TRH Data . . . 71

5 Joint Modelling for Longitudinal and Survival Data (JMLS) 75
5.1 Longitudinal Analysis: Linear Mixed Model (LMM) . . . 76
5.2 Survival Analysis: Cox Proportional Hazards Model . . . 77
5.3 Joint Modelling for Longitudinal and Survival Data . . . 78
5.3.1 The Basic JMLS . . . 78
5.3.2 Model Extensions . . . 79
5.3.3 Dynamic Predictions . . . 81
5.3.4 Time-Dependent ROC Curve . . . 81
5.4 JMLS Regression of Liver-19 Data . . . 82
5.4.1 Motivation of the Study . . . 82
5.4.2 Statistical Analysis of Liver-19 Data . . . 82

6 Software and Programming Code 91
6.1 Code for Section 3.3: "Bivariate Joint Modelling of Liver-19 Data" . . . 91
6.1.1 GAMM regression model for ALT . . . 93
6.1.2 CGAMLSS regression model for ALT and CRP . . . 97
6.2 Code for Section 3.4: "Bivariate Joint Modelling of PPAD Data" . . . 106

USC DE SANTIAGO DE COMPOSTELA

6.3 Code for Section 4.4: "Trivariate Joint Modelling of TRH Data" 116
6.4 Code for Section 5.4: "JMLS Regression of Liver-19 Data" 129

7 General Conclusions and Future Research **141**
7.1 Chapter 3: "Copula Regression Models for Bivariate Responses" . . . 141
7.1.1 Future work concerning CGAMLSS 143
7.2 Chapter 4: "An Alternative Approach to Copula JM: Multivariate Conditional Transformation Models (MCTM)" 143
7.3 Chapter 5: "Joint Modelling for Longitudinal and Survival Data (JMLS)" 144
7.4 Chapter 6: "Software and Programming Code" 145

References **147**

Further Information **165**
Conflict of Interests Statement 165
Publications Supporting the Contributions of this PhD Thesis Work 165
Published Papers 165
Unpublished Papers 168

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Figures

1.1 Venn diagram of biomedical data science (BDS; Hoyt and Muenchen, 2019). 2
1.2 Phases of BDS in a scientific study. 3
1.3 Smoothed trajectories and 95% confident bands using lasso (Tibshirani, 1996) of AST levels (blue) and CRP levels (pink) of patients diagnosed with COVID-19 and admitted to the *Complejo Hospitalario Universitario de Santiago de Compostela* in Spain during the first wave, separated by final outcome. Left: Discharged patients. Right: Deceased patients. AST and CRP raw values were divided by their maximum normal values as a form of weighting, because not all samples were analysed with the same apparatus (hence shown without units). . . . 5

2.1 Exploratory curves for weighted values (when needed) of transaminases (left) and inflammation markers (right) over time for deceased (pink) and living (blue) patients, along with 95% confidence intervals. Estimation made using a lasso regression analysis. 20
2.2 Kaplan-Meier estimated curve for overall survival, including 95% confidence intervals. 21
2.3 Kaplan-Meier estimated curve for survival, including 95% confidence intervals, for female (pink) and male (blue) subjects. P-value from log-rank test for difference between curves is included. 22
2.4 Kaplan-Meier estimated curve for survival, including 95% confidence intervals, for patients with (pink) and without (blue) diabetes. P-value from log-rank test for difference between curves is included. . 22
2.5 Smooth-effect of age in the survival probability, from a Cox additive model including only the baseline covariates. 95% confidence interval is depicted, as well as a pink line indicating 0, i.e. no significance. 23
2.6 Boxplot for GAD-7 (left) and EPDS (right) raw scores for the whole sample. Grey line indicates the cutoff for diagnosis (10 for GAD-7 and 13 for EPDS). 26

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2.7 Two-by-two scatter plot of the thyroid-related hormones. Left: log-TSH vs. Free T3. Centre: log-TSH vs. Free T4. Right: Free T4 vs. Free T3. 29

3.1 Contour plots of different copula functions with standard normal margins for simulated data using a Kendall's τ coefficient = 0.5. 34

3.2 Estimated centred smooth effect (with 95% CI) of time (left) and age (right) over the weighted and Box–Cox transformed transaminases extracted from the generalised additive mixed models in Table 3.1. Pink stands for deceased patients, blue for the living, and black for the total number of patients. The horizontal line indicates zero, and when it lies inside the 95% CI, no statistical significance was found. . 41

3.3 Histogram and estimated density curve (left panel) and Q-Q plot (right panel) for CRP (top) and lymphocytes (bottom) considering a gamma marginal distribution. 43

3.4 Time variation of the correlation between the weighted and Box–Cox transformed transaminases and the weighted and/or Box–Cox transformed (if required) inflammation markers for deceased (pink) and living (blue) patients, expressed by means of Kendall's τ coefficient (with 95% CI). The horizontal line indicates zero correlation, and when it lies inside the 95% CI, no correlation is found. 46

3.5 Variation of the correlation between the weighted and Box–Cox transformed transaminases and the weighted and/or Box–Cox transformed (if required) inflammation markers across the age of the patients, expressed by means of Kendall's τ coefficient (with 95% CI). The horizontal red line indicates zero correlation, and when it lies inside the 95% CI, no correlation is found. 47

3.6 Histogram and estimated density curve (left panel) and Q-Q plot (right panel) for GAD-7 (left) and EPDS (right) in the pregnant dataset, considering a Discrete Generalised Pareto (DGP, on top) and a Negative Binomial - type I (NBI, at the bottom) as marginal distributions. 53

3.7 Histogram and estimated density curve (left panel) and Q-Q plot (right panel) for GAD-7 (left) and EPDS (right) in the mother dataset, considering a Discrete Generalised Pareto (DGP, on top) and a Negative Binomial type I (NBI, at the bottom) as marginal distributions. 54

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.8 Estimated effect of the Age of the mother on GAD-7 (upper left panel), EPDS (upper right panel) raw scores, and their correlation (bottom panel) for pregnant women. Red lines are the mean values of the respective raw scores. 58
3.9 Estimated random effects of the country of origin on GAD-7 (top) and EPDS (middle) raw scores, and on the correlation between the outcomes (bottom), for pregnant women (left) and postpartum women (right). Red line is the 0 and indicates no effect. 60

4.1 Representation of Bernstein basis of order 2, 3, 5, and 7. 66
4.2 Function estimates $\hat{\lambda}(x)$ for the effect $\lambda(x) = x^2$ on the correlation parameter for CGAMLSS (left) and MCTM (right). The black line represents the true function and the grey lines are the estimations of the $R = 100$ simulation samples. 69
4.3 Left: $\mathrm{RMSE}\left(\lambda, \hat{\lambda}^{(r)}\right)$ for CGAMLSS and MCTM. Right: $\mathrm{SD}\left(\lambda, \hat{\lambda}^{(r)}\right)$ for CGAMLSS and MCTM. 69
4.4 Diagram of the functioning of thyroid hormones (Source: UCLA Endocrine Center (UCLA Health) (nd)). 71
4.5 Estimated marginal conditional CDFs $F_j(y_j|\text{age})$, and conditional density functions $f_j(y_j|\text{age})$, for j = log-TSH (left), j = Free T3 (centre), and j = Free T4 (right), from age = 20 (brown) to age = 90 (purple). . 72
4.6 Scatter plot of thyroid-related hormones vs. age. In red, percentile curves are shown. Left: log-TSH levels. Centre: Free T3 levels. Right: Free T4 levels. 73
4.7 Two-by-two conditional Spearman's $\rho^S(\text{Age})$ for the thyroid-related hormones (left: log-TSH and Free T3; centre: log-TSH and Free T4; right: Free T3 and Free T4). Grey area represents the 95% bootstrap confidence interval, while black line represents the bootstrapped point estimate. Red line indicates the 0. 73

5.1 Dynamic predictions for patient id=180, with 95% confidence interval. Left: estimation of transaminases, with AST on the top, ALT on the middle, and GGT on the bottom. Right: estimation of the survival probability. 87
5.2 Dynamic predictions for patient id=183, with 95% confidence interval. Left: estimation of transaminases, with AST on the top, ALT on the middle, and GGT on the bottom. Right: estimation of the survival probability. 88

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.3 AUC for different the survival time points for the JMLS regression model fitted. We assume different time intervals to update the survival predictions (dt = 1, 3, and 7). Horizontal pink line represents AUC = 0.5, indicating that the discrimination ability of the model is zero. 89

6.1 Graphical output of the command `gam.check`. If the model is adequate: theoretical quantiles will be very close to the straight red line (upper left panel), residuals will not follow a clearly distinguishable pattern (top right panel), the histogram will resemble that of a normal variable (Gaussian bell shape, bottom left panel), and response vs. fitted values will follow an increasing linear pattern (bottom right panel). 94
6.2 Plot for all the smooth effects in model `mod1`, obtained with the `R` command `plot(mod1, pages = 1, shade = T)`. 96
6.3 Plot of the smooth effect of `Age` in model `mod1`. 97
6.4 Plot of the smooth effect of `Time` in the correlation between `CRP` and `ALT_trans` according to model `fit1`. 104
6.5 Correlation between `ALT_trans` and `CRP`, expressed as Kendall's τ, as a function of `Time` and `Exitus` according to model `fit1`. In pink, the effect for the deceased is shown, while the effect for the people who survived is depicted in blue. Black line indicates no correlation. 106
6.6 Smooth effect of `AgeMother` on `GAD7` according to model `fit_p`. Left: Estimated smooth effect on the linear predictor. Right: Predicted smooth effect on the response. 114
6.7 Random effects of `Country` on GAD7, with 95% Confidence Intervals. 116
6.8 Traceplot for the 4 chains generated for the estimation of the β parameter for the first part of the effect of the longitudinal time on `AST_trans`. 136
6.9 ROC curve for day 15 with information up to day 7. 140

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Tables

2.1 Baseline characteristics of the study population, stratified by outcome. Age, AST, ALT, GGT, ALP, Bilirubin, CRP, D-Dimer, IL-6, Lymphocytes, BMI, and oxygen saturation are expressed in median and interquartile range (Me [IQR]). 18

2.2 Time course of transaminase levels according to final outcome. Data are expressed in medians and interquartile ranges (Me [IQR]). 19

2.3 Demographic characteristics of participants in the "Perinatal Mental Health and COVID-19 Pandemic" study. 25

2.4 Country-related variables in the PPAD database. CHI stands for Containment and Health Index mean of the previous 30 days, and IHDI is the Inequality adjusted Human Development Index. 26

2.5 Raw scores obtained from EPDS (Edinburgh Postnatal Depression Scale) and GAD-7 (the Generalised Anxiety Disorder Screener). Data are expressed in median and interquartile range (Me (IQR)). 27

2.6 Descriptive measurements of the age and the three outcomes: log-TSH, free T4 and free T3. 29

3.1 Results for the generalised additive mixed models for the weighted and Box–Cox transformed values of transaminases. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. edf $= 1$ indicates linear dependence. 40

3.2 Comparison of AIC and BIC values for the candidate marginal distributions for CRP and lymphocytes, both in the bigger set (with AST, ALT and GGT) and the smaller set (for ALP). 42

3.3 Best copula function for the regression models for each transaminase (columns) and IM (rows) pair. 44

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.4 Regression models for correlation between weighted and Box–Cox transformed values of transaminases and weighted and/or Box–Cox transformed (if required) IM. From the general model type (3.5), every regression model in this table only includes the covariates that were found to be statistically significant (p<0.05). $s(\cdot)$ indicates the smooth effect of the covariate. For Sex, the reference level is "Male"; for $Exitus$, is "Living"; and for AT (antiviral treatment), the reference level is "No". The smooth effects of $s_1(Age_l)$ and $s_2(Time_l, Exitus_l)$ in the correlation are depicted in Figures 3.5 and 3.4, respectively. . . 45
3.5 Comparison of AIC and BIC values for the candidate marginal distributions for GAD-7 and EPDS raw scores, both for pregnant and postpartum women. 52
3.6 Numerical results for the marginal models (using two negative binomial - type I distributions) for pregnant women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. Ratio (95%CI) is the exponential of the β estimates and their 95% Confidence Intervals. 57
3.7 Numerical results for the correlation model between GAD-7 and EPDS for pregnant women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. 57
3.8 Numerical results for the marginal models (using two negative binomial - type I distributions) for postpartum women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. Ratio (95%CI) is the exponential of the β estimates and their 95% Confidence Intervals. 59
3.9 Numerical results for the correlation model between GAD-7 and EPDS for postpartum women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. 59

5.1 Summary for the longitudinal sub-model for AST. SD is standard deviation of the estimated coefficients β, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(t, 15)$ and $B_n(t, 30.61)$ denote the B-spline basis matrix for a natural spline of order 1 for t, and σ is the estimation of the standard deviation of the error terms ε_{ij}. 84

USC
DE SANTIAGO
DE COMPOSTELA

5.2 Summary for the longitudinal sub-model for ALT. SD is standard deviation of the estimated coefficients β, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(t, 15)$ and $B_n(t, 30.61)$ denote the B-spline basis matrix for a natural spline of order 1 for t, and σ is the estimation of the standard deviation of the error terms ε_{ij}. 84
5.3 Summary for the longitudinal sub-model for GGT. SD is standard deviation of the estimated coefficients β, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(t, 15)$ and $B_n(t, 3.61)$ denote the B-spline basis matrix for a natural spline of order 1 for t, and σ is the estimation of the standard deviation of the error terms ε_{ij}. 85
5.4 Summary for the survival sub-model. SD is standard deviation of the estimated coefficients γ and α, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R-hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(Age, d_1)$ and $B_n(Age, d_2)$ denote the B-spline basis matrix for a natural cubic splines with 2 degrees of freedom for Age. 85

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Introduction

1.1 Biomedical Data Science (BDS)

The technological revolution in biomedicine today is creating new challenges in the processing, management, analysis and interpretation of increasingly complex biomedical databases (including big data). This problem covers a wide range of data structures, that can be classified in three data varieties:

- **Structured data:** Data that can be stored processed, and manipulated in a traditional relational database.
- **Unstructured data:** Data that does not fit into a structured database format.
- **Semi-structured data:** Data that does not fit into a structured database format, but which can nevertheless be classified by means of labels that bring some order to the data.

Examples of these can be multivariate observations, clustered data, repeated measurements, functional data, correlated spatio-temporal data, images, etc. In addition to this complexity, in many cases the amount of data available is enormous, which makes the analysis of these data a very difficult task. In this context, the field of Biomedical Data Science (BDS) can be of great help.

Data science was first acknowledged as a science six decades ago by John Tuckey in his article *The Future of Data Analysis*. According to the author, "Data analysis" includes:

> "(...) among other things: procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and

USC DE SANTIAGO DE COMPOSTELA

> all the machinery and results of (mathematical) statistics which apply to analysing data." (Tukey, 1962).

This definition reflects the existence of a field of study whose interest is learning from data, laying the foundations for what is known today as the *Data Science* area. This area has been gaining presence in recent years due to the technological explosion, which allows the generation and collection of huge amounts of data in different fields. So much so that Davenport and Patil (2012) proclaimed the figure of data scientist to be "the Sexiest Job of the 21st Century".

USA National Institutes of Health defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data." (National Institutes of Health (NIH), 2018). Although there is no universally accepted definition of BDS, the above definition for biomedical databases can be applied.

The interdisciplinary character of BDS is clearly established if one takes into account that this scientific discipline emerged in the areas of Statistics, Computer Science and Health, responsibly enabling data analysts to gain knowledge from complex biomedical data (see Figure 1.1).

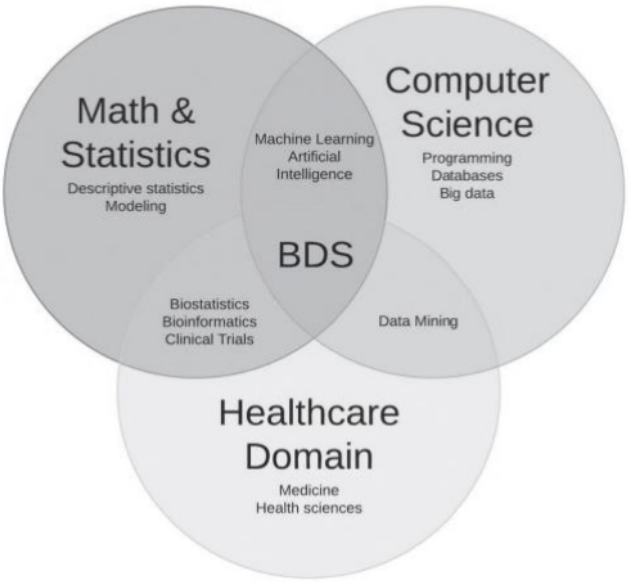

Figure 1.1: Venn diagram of biomedical data science (BDS; Hoyt and Muenchen, 2019).

USC
DE SANTIAGO
DE COMPOSTELA

It is obvious that the skills required to be a data scientist are not easy to develop

in a single person. This inevitably leads to the need for the creation of multidisciplinary working groups made up of different specialists from the different areas involved, who will combine their skills in order to carry out the BDS process, represented in Figure 1.2.

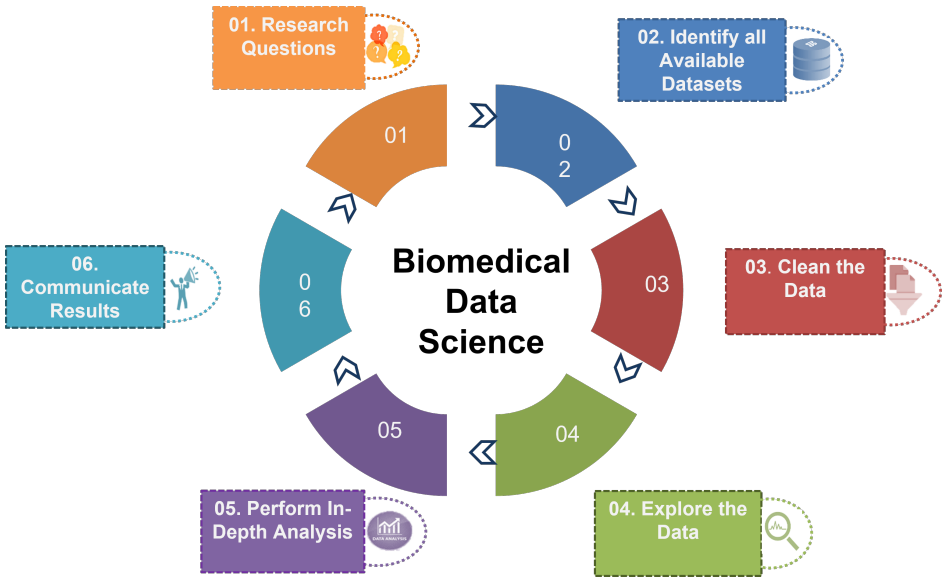

Figure 1.2: Phases of BDS in a scientific study.

The phases of BDS process are:

(i) **Clearly state the research questions you want to answer.** Being clear about the questions to be answered and the initial hypotheses is fundamental to focus the study and the whole process of data collection and analysis.

(ii) **Identify all available databases.** It is important to consider, from the outset of the study, all possible databases and variables that can (and need to) be collected in order to answer the questions posed.

(iii) **Clean the data.** Before in-depth statistical analysis, the databases should be checked to put them in the right format and to correct any errors that may have arisen during data collection. Errors not detected prior to an in-depth analysis may lead to incorrect results, as well as to a large loss of time and work due to the possible need for re-analysis. In this respect, Broman and Woo (2018) provided a very nice guide for organising spreadsheets.

(iv) **Explore the data.** Data exploration is the first step in statistical analysis. It

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

will not only provide a better overview of the collected data, guiding future analyses, but also helps to detect possible errors in the database.

(v) **Perform in-depth analysis.** This analysis (not always necessary, sometimes the exploration of the data is sufficient) is intended to answer the research questions.

(vi) **Communicate results**. The final step of every research project. Once the results were obtained and interpreted, the results should be communicated to the scientific community.

Although it seems that each phase of the process is oriented towards a specific profile of those who make up the BDS, the truth is that different profiles are required in each of them. For example, both in the exploration and in the in-depth analysis of the data, in addition to the statistician or computer scientist who carries it out, the intervention of an expert in the medical field in question is necessary to interpret the results obtained and guide the successive steps. Again, the need for an interdisciplinary group whose members interact with each other throughout the process is evident.

1.2 Joint Modelling (JM) Regression

Statistical analysis is constantly evolving and new methodologies are emerging all the time, allowing for a better analysis of the complex and diverse data available, so that more accurate information can be extracted and more precise predictions can be made.

This thesis deals with biomedical multivariate response studies, in which not only the modelling of the response variables is important, but also their correlation. These studies are pretty common in biomedical research, for instance, to study i) the relationship between different biomarkers in the same patient (as glucose and glycated haemoglobin levels, both used to diagnose diabetes), ii) the relationship between different conditions in the same patient (as diabetes and hypertension), iii) the same pathology in duplicated organs (e.g. to detect diabetic retinopathy in both eyes), iv) of studies made in twins or family units. In all this kind of situations, it is also interesting to asses, together with the effect of the covariates, the structure of dependence between the different outcomes, being Joint Modelling (JM) regression models the most adequate models to deal with this.

As a motivating example, which will be discussed in greater depth throughout this thesis, we can consider data from COVID-19 patients at the *Complejo Hospitalario Universitario de Santiago de Compostela* during the first wave of the disease, from

USC DE SANTIAGO DE COMPOSTELA

March to June 2020. In this first study, clinicians were interested in studying the relationship between the transaminase (indicators of liver damage) and inflammation markers levels, and also patient survival. Biomarker levels were taken longitudinally at different time points of the patients' stay in the hospital. Figure 1.3 depicts the smoothed trajectories over time (in days, up to 30 days) of AST transaminase (in blue) and CRP inflammation marker (in pink) for people who died (right) and people who survived (left). One can intuit from the figure an opposite behaviour of the two variables over time (when one goes up, the other goes down and vice versa), implying a strong correlation between them over time. This simple descriptive analysis shows the great importance of modelling the association, and JM are the most appropriate technique for doing so.

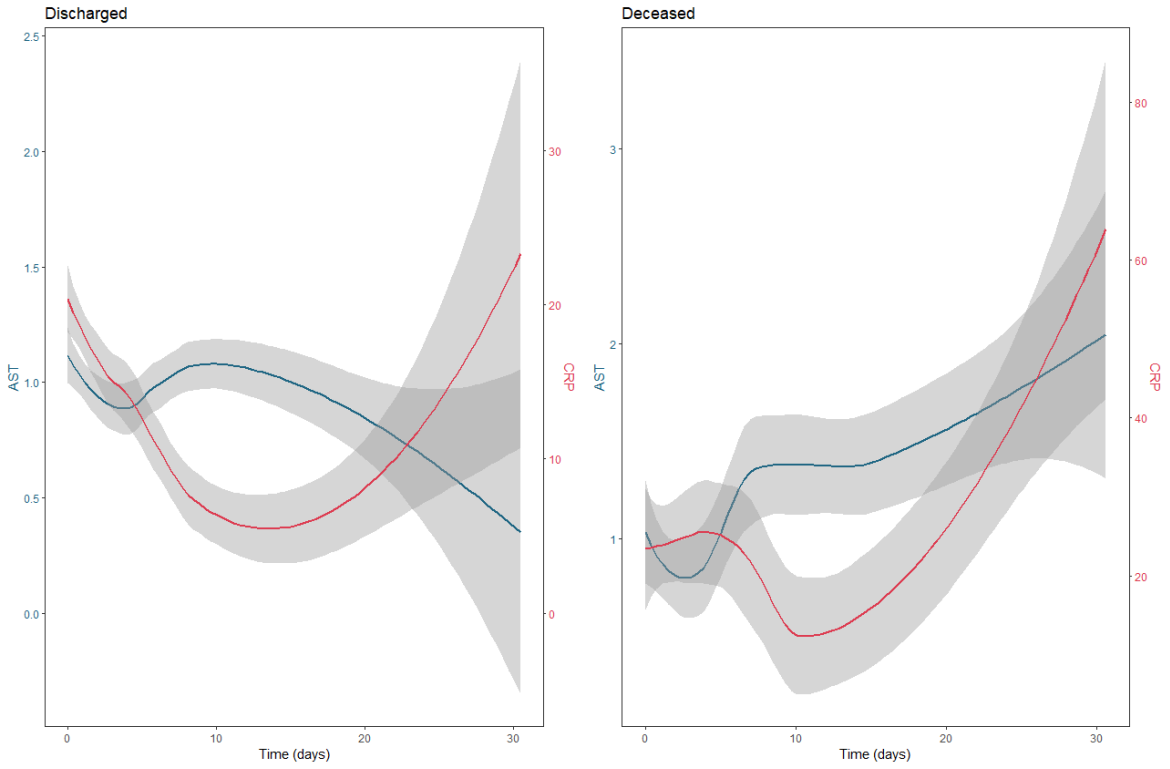

Figure 1.3: Smoothed trajectories and 95% confident bands using lasso (Tibshirani, 1996) of AST levels (blue) and CRP levels (pink) of patients diagnosed with COVID-19 and admitted to the *Complejo Hospitalario Universitario de Santiago de Compostela* in Spain during the first wave, separated by final outcome. Left: Discharged patients. Right: Deceased patients. AST and CRP raw values were divided by their maximum normal values as a form of weighting, because not all samples were analysed with the same apparatus (hence shown without units).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

JM approaches presented in this PhD project represent a leap forward in the multivariate regression analysis application to real clinical problems. To date, multivariate regression usage in the clinical setting is anecdotal (Duarte et al., 2017; Espasandín-Domínguez et al., 2018, 2019). Indeed, currently, when several correlated measurements are available for the same patient, they are treated as mutually independent variables, or a pre-specified transformation is applied to them resulting in a univariate problem. This common strategy may depreciate useful information arising from the continuous measurements joint distribution, information that allows us, among other things, to make much more accurate predictions (individual or otherwise). Hence, JM may enhance the application of multivariate regression for diseases whose diagnosis is based on several continuous biomarkers, or clinical assessments that are based on correlated measurements.

1.2.1 General Joint Modelling Framework

Classically, regression models provide a function describing the relationship between a set of independent variables (also known as predictors or covariates) and a response variable, also known as dependent variable or outcome. Regression analysis allows for the construction of mathematical models to explain possible relationships between a set of variables (for instance, transaminases and inflammation markers, or levels of perinatal anxiety and depression in relation to COVID-19), enabling us to understand and describe some processes, which is why it has become one of the most widespread statistical tools.

In this thesis, the focus is on cases where we have multiple responses that are supposed to be related to each other, and the study of the relationship between them is also of great importance. As a simplification, let us consider two response variables, Y_1 and Y_2, and let $\boldsymbol{X}$ be a set of covariates observed in the same individuals. Let also $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3 \subseteq \boldsymbol{X}$ be subsets of potentially different predictors in $\boldsymbol{X}$. In this context, the general framework of JM can be expressed as in (1.1):

$$\begin{cases} Y_1 \sim F_1(\boldsymbol{X}_1) \\ Y_2 \sim F_2(\boldsymbol{X}_2) \\ \mathrm{Assoc}(Y_1, Y_2) \sim F_3(\boldsymbol{X}_3) \end{cases} \tag{1.1}$$

where $\mathrm{Assoc}(Y_1, Y_2)$ represents the possible association between the two outcomes, and F_1, F_2 and F_3 are the different regression functions for each term. It should be noted that for the fitting of each of the sub-models, it is not necessary (although it is possible) to include the same covariates in all of them. In this type of regression, the separate response variables are referred to as marginals, and most of times, a univariate regression model can be considered for their modelling.

USC DE SANTIAGO DE COMPOSTELA

Univariate Regression Models

The nature of the response variables can be very diverse (continuous variables, categorical variables, etc.), and the regression functions used can also be of different types. Let y_i, be the observations of the outcomes for each subject $i \in \{1, \ldots, n\}$, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ be the set of p covariates observed in the same individuals. With this in mind, different models can be considered for the marginals:

- **Linear regression Models (LM).** This was the earliest form of regression, and dates back to the end of the 19th century. This is the simplest regression model and it assumes the relationship between the response and the covariates - also including a random error term ε_i - to be linear:

  $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

  where β_j, $j \in \{0, \ldots, p\}$ are the unknown regression coefficients to be estimated. Linear regression models assume that the response variable is continuous and normally distributed, which is very restrictive.

- **Generalised Linear regression Models (GLM).** To consider other kinds of response variables such as binary, or other non-normal continuous outcomes, Nelder and Wedderburn (1972) proposed the GLM models. In this type of models, instead of directly modelling the (conditional) expectation $\mathbb{E}(Y_i|x_i) = \mu_i$, this is linked to a linear predictor η through a known link function g as follows:

  $$\eta_i = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i. \tag{1.2}$$

  By making this transformation, the normality assumption is relaxed and the response variable is now assumed to belong to the exponential family distribution (for instance, Poisson, binomial, or gamma), of which the normal distribution is a part. This leads to the fact that linear regression models are a particular case of GLM, with the identity as link function.

  Despite the flexibility gained in terms of response variable distribution, the assumption of linearity for the covariate effects remains as the main limitation of this methodology.

- **Generalised Additive regression Models (GAM).** Introduced in early 1990s by Hastie and Tibshirani (1990b), GAM regression models replace the linear predictor in (1.2) by an additive one:

  $$\eta_i = g(\mu_i) = \beta_0 + s_1(x_{i1}) + \cdots + s_p(x_{ip}) + \varepsilon_i, \tag{1.3}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

where β_0 is a global intercept and s_j, $j \in \{1, \ldots, p\}$ are unknown smooth functions to be be estimated. In this way, GAM models allow for the consideration of different types of covariate effects, including, but not restricted to, smooth estimation of the effect of continuous variables, categorical covariates, and even random effects. It should be noted that when all s_j are centred linear functions (with no intercept), we are dealing with a GLM. The flexibility gained with the GAM is mainly due to the introduction of smooth functions. In practice, multiple smoothers can be considered, such as penalised splines (P-splines), or thin plate regression splines, among others. The reader is advised to consult Wood (2017) for more technical and detailed information on GAM models and the different types of smoothers.

GAM models are also limited, in the sense that they are restricted to the estimation of the mean of the outcomes. In many occasions, the interest lays in the global estimation of the distribution of the outcomes.

- **Generalised Additive Models for Location, Scale and Shape (GAMLSS).** Introduced by Rigby and Stasinopoulos (2005), these models focus on distributional regression. The goal is to characterise the effects of covariates on all the parameters of the response's distribution, such as variance, or to know the complete distribution of the response, instead of limiting ourselves to the study of the mean. In GAMLSS framework, Y_i is assumed to follow a complex distribution, which has a fully parametric specification, $Y_i|\sim\ D(\mu_i, \sigma_i, \nu_i, \tau_i)$. All the parameters of the response distribution can be related to an additive predictor, similarly as in (1.3). Moreover, the distribution is not restricted to the simple exponential family. GAMLSS regression models are defined as follows:

$$\begin{cases} \eta_i^{\mu} = g_1(\mu_i) = \beta_0^{\mu} + s_1^{\mu}(x_{i1}) + \cdots + s_p^{\mu}(x_{ip}) + \varepsilon_i^{\mu} \\ \eta_i^{\sigma} = g_2(\sigma_i) = \beta_0^{\sigma} + s_1^{\sigma}(x_{i1}) + \cdots + s_p^{\sigma}(x_{ip}) + \varepsilon_i^{\sigma} \\ \eta_i^{\nu} = g_3(\nu_i) = \beta_0^{\nu} + s_1^{\nu}(x_{i1}) + \cdots + s_p^{\nu}(x_{ip}) + \varepsilon_i^{\nu} \\ \eta_i^{\tau} = g_4(\tau_i) = \beta_0^{\tau} + s_1^{\tau}(x_{i1}) + \cdots + s_p^{\tau}(x_{ip}) + \varepsilon_i^{\tau} \end{cases}, \tag{1.4}$$

where the first equation in (1.4) refers to the location parameter μ of y_i (similar as in equation (1.3) for the GAM models), the second refers to the scale parameter σ and the last two are related to the shape parameters, ν and τ.

There are, in addition to those mentioned above, other alternatives, such as the least absolute shrinkage and selection operator regression (lasso; Tibshirani, 1996), which also performs variable selection, or non-parametric alternatives to GAMLSS regression working from a Bayesian framework, known as distribution regression

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(DR) models (Klein et al., 2015a,b). In this thesis, our work will be focused on GAM and GAMLSS regression models. Compared to DR models, the former offer similar results, but are faster, have lower computational cost and the available software is more user-friendly (Espasandín Domínguez, 2019).

1.2.2 Copula Distributional Regression Models

In recent years, several regressions models for multivariate responses have been proposed. However, in most of the existing multivariate regression models, the marginal distributions as well as the joint distribution are constrained to be of the same type (in general, Gaussian); which might be restrictive in some studies (Klein and Kneib, 2016). Copula distributional regression models with additive predictors have overcome these problems. A major advantage of the copula approach is that the marginal distributions may come from different families, and the dependence between the responses does not need to be linear. In this PhD thesis, we have adopted Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS; Marra and Radice, 2017). In a nutshell, CGAMLSS extends the class of GAMLSS (Rigby and Stasinopoulos, 2005) to the case where all the parameters of a bivariate response distribution can be modelled simultaneously as functions of additive predictors by allowing for several types of covariate effects (such as non-linear, random and spatial effects). In the statistical literature, other copula-based regression approaches have been developed but, as compared to Marra and Radice (2017), they either only cover some of the flexibility of the CGAMLSS or are based on less efficient estimation approaches (for instance, Gijbels et al., 2011; Acar et al., 2013; Sabeti et al., 2014; Vatter and Chavez-Demoulin, 2015; Yee, 2015; Klein and Kneib, 2016).

It should be mentioned that the availability of software for fitting copula regression models with additive predictors is scarce, so the fact that the CGAMLSS methodology is implemented in `GJRM R` package is another point in favour of its use.

1.2.3 Multivariate Conditional Transformation Models (MCTM)

Multivariate Conditional Transformation Models (MCTM; Klein et al., 2019) represent a very interesting alternative for CGAMLSS regression models when the outcomes are continuous. It is a non-parametric approach (i.e. without any prior assumptions about the data) and computationally faster than CGAMLSS. These models characterize the covariates effects not only on the correlations (relationship) between the outcomes, but also on the Cumulative Distribution Functions (CDF) of

USC DE SANTIAGO DE COMPOSTELA

each response. It is also possible to obtain percentile curves from these models for each outcome, facilitating interpretability. Because of all these advantages, its implementation also has a place in this thesis, for the study of thyroid-related hormones in relation to the age of the patients.

Despite outperforming the CGAMLSS methodology in several aspects, its use is currently restricted to very specific cases due to its software development, as it can so far only be implemented for continuous responses and with only one explanatory variable. But this is a technique worth considering, both now (when the previous requirements are fulfilled) and for possible future advancement and extensions.

1.2.4 Joint Modelling for Longitudinal and Survival Data (JMLS)

In many biomedical studies, patients are followed during the research study, and information is collected repeatedly within a time period. In these follow-up studies, in addition to repeated measurements, it is very common to consider one or several events and the time to the event for each patient. Usually, the main purpose is the simultaneous analysis of these two kind of outcomes: longitudinal biomarkers and time-to-event. In most of the cases, clinicians are usually interested in time to recovery, recurrence of a disease or mortality in terms of baseline covariates and repeated measurements that are taken during the research study. The main goal is often to identify prognostic factors that can be used to guide clinical management of patients. Examples are: i) the study of CD4 cell counts over time and survival in HIV patients, ii) or the study of bilirubin levels in relation to the development of cirrhosis. The previously introduced COVID-19 related example (see Figure 1.3) can be (and, in fact, will be later on this thesis) also analysed from this perspective since, in this case, the final event is known (deceased or not), as well as the follow-up time, so one may be interested in looking at the extent to which the available longitudinal markers can help predict patient mortality.

Joint models have received considerable attention in the past decades, focusing on this objective. In fact, in literature, when talking about JM they usually refer to a very specific class of models, involving longitudinal and survival outcomes, referred as Joint Models for Longitudinal and Survival data (JMLS; Tsiatis and Davidian, 2004).

JMLS does not exactly adapt to the modelling structure expressed in (1.1). Let Y be the longitudinal process, S is the survival process and U is a latent random effect. The main idea of a joint model consists of a joint likelihood calculation of the two processes (S and Y) using the latent random effect U. Various approaches of joint models have been proposed in the literature, depending on the process of interest:

- **Selection Models.** In this type of model the focus is only on the time-to-event

USC DE SANTIAGO DE COMPOSTELA

process, thus can be used for survival analysis with endogenous variables, which are explanatory variables depending on a certain way on the outcome. In these models, a latent random effect, U, underlines only the longitudinal process Y, and the calculation of joint likelihood consists of a factorization into the conditional distribution of the longitudinal process given the random effect on the one hand, and the conditional distribution of the survival process given the longitudinal outcome on the other hand:

$$f(Y, S, U) = f(U)f(Y|U)f(S|Y)$$

- **Pattern-Mixture Models.** This type of models can be used for the longitudinal studies with a drop-out process generated by a non-ignorable mechanism. They are similar to the selection models, but factorization is reversed, i.e. a latent random effect, U, underlines only the survival process S, so that the calculation of joint likelihood consists of a factorization into the conditional distribution of the survival process given the random effect on the one hand, and the conditional distribution of the longitudinal process given the survival outcome on the other hand:

$$f(Y, S, U) = f(U)f(S|U)f(Y|S)$$

- **Shared Random Effects Models.** In these models the latent random effect U underlines both longitudinal and survival process, thus being the ones indicated when the interest is in both processes equally:

$$f(Y, S, U) = f(U)f(Y|U)f(S|U)$$

The joint modelling approaches benefit both the analysis of the biomarker and time-to-event of interest: i) by giving the appropriate representation of the quantitative influence of the time-varying factors on the time-to-event, and ii) by taking into account the informative missing-data mechanism on the longitudinal biomarker due to event of interest.

There are also earlier regression modelling techniques in the literature addressing these questions. Focusing on a case with a single longitudinal biomarker and a time-to-event data, one of the earliest methods is the extended Cox model (Cox and Oakes, 1984), which assumes that the covariates are external and not related to the failure mechanism and does not take into account the measurement error of the longitudinal process. Thus, Self and Pawitan (1992) introduced two-stage models as an alternative approach where the both processes are modelled separately. To study the association between the longitudinal and survival data, joint models are required to feature this correlation (Tsiatis and Davidian, 2004).

USC DE SANTIAGO DE COMPOSTELA

There are multiple software implementations of the JMLS approaches that allow different extensions depending on the research study. Guo and Carlin (2004) have described the use of markov chain Monte Carlo approach, which is implemented in `WinBUGS` (Lunn et al., 2009) in a joint model within a Bayesian framework. In the context of the free software `R` (R Core Team, 2020), the `joineR` package developed by Philipson et al. (2018) is based on an extended Wulfsohn and Tsiatis (1997) model, allowing for a single time-to-event and competing risks. Rizopoulos (2010) developed the `JM` package, which fits joint models for normal longitudinal responses and time-to-event data using maximum likelihood. From a Bayesian perspective, Köhler et al. (2017) developed flexible Bayesian additive joint models, which have a high degree of flexibility as they are based on Bayesian GAM regression models, and are implemented in `bamlss` package. In the context of multiple longitudinal outcomes, Rizopoulos (2016) developed `JMbayes` package, which also adds flexibility with respect to his previous package and works within the Bayesian approach. `JMbayes2` (Rizopoulos et al., 2022) allows, in addition, multiple event times (competing risks and multi-state processes), also from a Bayesian perspective. Finally, `lcmm` package (Proust-Lima et al., 2022) can be used for fitting latent class joint models, which involves a different approach to the problem from those explained above.

1.3 General Objectives of the Thesis

Nowadays, the different JM methodologies are very advanced in terms of software (Marra and Radice, 2017; Rizopoulos et al., 2022), but less so in terms of implementation. The main objective of this PhD thesis is to bring the novel JM regression approaches closer to clinical practice, and to future epidemiological studies. To this end, this thesis focuses on the application of these type of models in various disciplines of medicine and psychiatry, in particular in studies concerning liver damage, thyroid-related hormones and peripartum anxiety and depression. In these fields, JM prove to be appropriate and very useful, but they are hardly present. During the course of this thesis, the COVID-19 pandemic emerged, declared a public health emergency by the World Health Organization (WHO). According to the latest report of 21 September 2022, the cumulative number of confirmed cases reported globally is now over 609 million and the cumulative number of deaths is over than 6.5 million (World Health Organization (WHO), 2022). This lead us to the fact that most of the contributions included here are highly related to that disease.

In this PhD project we place special emphasis on the BDS process as a guide for the optimal conduct of a biomedical study. Therefore, a great effort was made to offer as clearly as possible the software necessary for the implementation of these methodologies, using the free statistical software `R` (R Core Team, 2020).

USC DE SANTIAGO DE COMPOSTELA

1.4 Structure of the Thesis

This chapter is devoted to preliminaries, introduction to Biomedical Data Science and the Joint Modelling approaches, as well as the objectives and the structure of the theses. This thesis is structured in 7 chapters with the objective of implementing Joint Modelling approaches in biomedical databases. The remaining chapters of this work are structured as follows.

Chapter 2 includes the presentation of the databases used in the analyses.

Chapter 3 focuses on the distributional JM methodology using copula models to bond the two response variables. On the one hand, this chapter gives the statistical background of the methodology (model formulation, estimation and inference). On the other hand, it presents the application of this methodology for both continuous and both counting data outcomes. JM methods were applied to the study of the relationship between transaminases levels and inflammation markers in patiens affected with COVID-19, and also to the study of anxiety and depression levels of peripartum women during the COVID-19 pandemic. Some of the contents of this chapter were published in 2022 by *Scientific Reports* (Díaz-Louzao et al., 2022). There is another part of the contents of this chapter which, although not yet published (the corresponding article is currently submitted also to *Scientific Reports*), has an online preprint (Mesquita et al., 2022).

Chapter 4 introduces MCTM models as an alternative of distributional JM approaches, with an application to thyroid-related hormones. Some of the contents of this chapter were published in Díaz-Louzao et al. (2021).

Chapter 5 is dedicated to Bayesian JMLS for multiple longitudinal outcomes. This chapter includes the statistical background of these models, including the formulation of the models and inference, but also dynamic predictions and time dependent ROC curves. Dynamic predictions are specially interesting because of its proximity to so-called precision medicine, as it allows the prediction for each individual to be updated as new information is collected. This methodology will be applied to the study of the survival of COVID-19 patients regarding information on transaminase levels.

Chapter 6 is entirely dedicated to the software implemented in the application of all these JM methodologies to real databases. Programming code is provided to the user by way of a tutorial for the application of this methods on practice in the open source statistical software R.

Finally, Chapter 7 discusses the main results of the dissertation and future work.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 2

Motivating Research Questions

The research objectives of this thesis are questioned by real data coming from various clinical research studies.

The first research questions come from a follow-up study with several longitudinal biomarkers and the survival of patients with COVID-19. The objective of the clinicians is double: on the one hand, to explore the relationship between a transaminase and a inflammation biomarker in SARS-CoV-2 infected persons; on the other hand, to study the effect of longitudinally collected transaminase levels on the survival of these patients. In order to address these medical problems, the distributional JM methodology for the former and JMLS for the latter will be applied.

Another research question of interest addressed in this PhD thesis focuses on how the pandemic context can affect the mental health of women in perinatal situation. To this end, distributional JM models will be applied to study the levels of anxiety and depression (and the relationship between them) in perinatal women during the COVID-19 pandemic.

The clinical problem is the only one not related to COVID-19, but is nonetheless of vital importance. The aim is to determine the distribution of TSH, free T3 and free T4 values in a healthy population and in relation to the age of the patients, which will help to give a more accurate diagnosis of possible thyroid dysfunctions. For this purpose, the MCTM methodology will be use.

2.1 Liver Injury in COVID-19 (Liver-19) Data

For this motivating database, data were collected from the medical reports of patients diagnosed with COVID-19 and admitted to the *Complejo Hospitalario Universitario de Santiago de Compostela* in Spain, from March 12, 2020 (date of first COVID-19 diagnosis) to June 11, 2020, i.e. during the first wave of the pandemic. A con-

USC DE SANTIAGO DE COMPOSTELA

firmed case of COVID-19 was defined as a positive result in the reverse transcription polymerase chain reaction (RT-PCR) test on samples obtained from nasal and throat swabs performed in accordance with WHO protocol (World Health Organization (WHO), 2020b). The study time for each included patient was from admission until the patient died or was discharged from hospital.

Using `Python 3.8`, clinicians designed a specific programme with a graphical user interface through which each patient's data were imported and automatically recorded from the electronic medical record. In all cases, the data were encrypted prior to filing. Clinicians collected all demographic data, admission history, and medical comments on the patient's progression, clinical constants (heart rate, blood pressure, and body temperature), pulse oximetry, radiological studies, and pharmacological prescriptions, including the date and time of their prescription. Among the laboratory determinations, all parameters of the complete blood count, and blood biochemistry (glucose, urea, creatinine, uric acid, sodium, potassium, calcium, magnesium, albumin, cholesterol, triglycerides, aspartate transaminase (AST), alanine aminotransferase (ALT), gamma glutamyl aminotransferase (GGT), alkaline phosphatase (ALP), bilirubin, and coagulation), which included the international normalised ratio, prothrombin, blood gases (pO_2, pCO_2, O_2 saturation, bicarbonate, lactate), procalcitonin, lactate dehydrogenase, vitamin D, and haemoglobin A1c were recorded. C-reactive protein (CRP), D-Dimer, interleukin-6 (IL-6), and lymphocyte counts were considered as a specific group of inflammation markers (IM), and reference limits of normality for each parameter and laboratory measurement were recorded. An abnormal increase in transaminase levels has been considered any value that exceeds the normal range. The transaminase elevation profile has been grouped according to R-factor and similar how drug-induced liver damage (DILI) is classified depending on whether the predominant elevation is due to AST and ALT (hepatocellular pattern), GGT and ALP (cholestatic pattern) or elevation similar to one another (mixed). The R-value is defined as serum alanine aminotransferase/upper limit of normal (ULN) divided by serum ALP/ULN. By common convention, $R \geq 5$ is labelled as hepatocellular, $R < 2$ is labelled as cholestatic, and $2 < R < 5$ is labelled as "mixed" (Chalasani et al., 2014).

For those cases in which specific treatment had been started before hospital admission, the start date of each drug was manually recorded. The programme classified each drug into its corresponding pharmacological group, adding this as an additional variable in each registry. The records were processed and archived in a relational `sqlite3` database, with independent tables for physical constants, pharmacological prescriptions, symptoms, exploratory signs, radiological signs, laboratory parameters, complications, and status (emergency room, hospital ward, critical unit, discharge or death), indexed by case number and start date and linked to a general table with demographic characteristics, antecedents, and form of infection.

USC DE SANTIAGO DE COMPOSTELA

Prior to any statistical analysis, all variables of interest were checked and included in a new reduced database.

2.1.1 Ethical Issues

The study was conducted in accordance with the guidelines of the Declaration of Helsinki and the principles of good practice and was approved by the Institutional Review Board of the Galician health service on April 3, 2020 (# 2020/194). Informed consent documents were waived by the Institutional Review Board.

2.1.2 Descriptive Analysis of Liver-19 Database

During the study period, 212 patients were admitted to the hospital with COVID-19. We restricted the maximum follow-up time to 30.61 days, which was the longest follow-up time for the deceased, in order to make the two groups comparable over the entire time period considered. Table 2.1 shows the sample's demographic and clinical characteristics. The descriptive analysis of the sample was performed according to the demographic variables, treatments administered, and absolute values of transaminases and IMs for the total sample and separately for the patients who were eventually discharged (Alive) or died (Exitus). The continuous variables are expressed as median and interquartile range, and the categorical variables are expressed as their absolute frequency and percentage. Differences between the surviving and deceased patients were assessed by means of Fisher's exact test (Fisher, 1935) for binary variables, while the Mann-Whitney U test (Mann and Whitney, 1947) was employed for the continuous variables.

The median age was 68 years, and 56% were men. In our series, 37 (17.4%) patients died. The patients who died were significantly older, and there was a significantly higher percentage of patients with a history of ischemic heart disease, diabetes, and a lower oxygen saturation rate during the hospitalization. There were no differences between the deceased and those who survived with respect to the treatment undergone, past HBV or HCV infection, and statin treatment. 53%, 56%, 42% and 19% of the patients presented elevated AST, ALT, GGT and ALP respectively. Of those patients: 12%, 50% and 38% presented elevated transaminase levels with a hepatocellular, cholestatic, and mixed pattern, respectively. The surviving patients who presented a hepatocyte pattern had a significant increase in ALT. The median AST and ALP values were significantly higher in the patients who died. There was a significant difference in IL-6, D-Dimer, CRP, and lymphocyte values between the patients who died and those who survived (Table 2.1).

The progression of transaminase levels at fixed intervals every 7 days during the first 3 weeks of hospitalization (Table 2.2) shows a characteristic elevation, from the

USC DE SANTIAGO DE COMPOSTELA

first week, with respect to AST, ALT, GGT, and ALP, especially in those patients who died. The same progression for IMs also demonstrated a characteristic elevation for CRP and D-Dimer, but from day zero, with a second elevation at the end of the observation period. This movement was less evident in the case of IL-6. Conversely, the lymphocytes count fell from the first week for the patients who died.

Table 2.1: Baseline characteristics of the study population, stratified by outcome. Age, AST, ALT, GGT, ALP, Bilirubin, CRP, D-Dimer, IL-6, Lymphocytes, BMI, and oxygen saturation are expressed in median and interquartile range (Me [IQR]).

		TOTAL	**EXITUS**	**ALIVE**	**P-VALUE**
Period under study		01/03/2020-19/06/2020			
Number of patients		n = 212	n = 37	n = 175	
Age (years)		68.40 [56.58, 77.18]	82.52 [73.35, 86.65]	65.48 [55.10, 74.57]	<0.001
Male		120 (56.60%)	27 (72.97%)	93 (53.14%)	0.030
Lopinavir/Ritonavir and /or Hydroxychloroquine		190 (89.62%)	31 (83.78%)	159 (90.86%)	0.233
Statins		78 (36.79%)	15 (45.86%)	63 (36.00%)	0.708
Ischemic heart Disease		19 (7.55%)	9 (24.32%)	10 (5.71%)	0.001
Diabetes		48 (22.64%)	17 (45.95%)	31 (17.71%)	<0.001
HCV*		1 (0.56%)	0 (0.00%)	1 (0.64%)	1.000
HBV**		12 (6.63%)	1 (4.35%)	11 (6.96%)	1.000
Chronic Liver Disease		13 (6.13%)	0 (0.00%)	13 (7.43%)	0.131
High liver enzymes	AST	112/212 (52.83%)	19/37 (51.35%)	93/175 (53.14%)	0.858
	ALT	119/212 (56.13%)	10/37 (27.03%)	109/175 (62.29%)	<0.001
	GGT	90/212 (42.45%)	13/37 (35.14%)	77/175 (44.00%)	0.364
	ALP***	23/121 (19.01%)	7/24 (29.17%)	16/97 (16.49%)	0.160
	Hepatocellular pattern	15/121 (12.40%)	2/24 (8.33%)	13/97 (13.40%)	0.078
	Cholestatic pattern	60/121 (49.59%)	17/24 (70.83%)	43/97 (44.33%)	
	Mixed pattern	46/121 (38.02%)	5/24 (20.83%)	41/97 (42.27%)	
High bilirubin		48 (22.64%)	10 (27.03%)	38 (21.81%)	0.518
AST		30 [22, 45]	37 [24, 59]	29 [22, 42]	<0.001
ALT		35 [22, 61]	31 [18, 64]	36 [23, 60]	0.007
GGT		45 [26, 86]	59 [29, 102]	44 [26, 80]	0.007
ALP		111 [79, 166]	152 [98, 219]	102 [70, 146]	<0.001
Bilirubin		0.5 [0.4, 0.8]	0.5 [0.4, 0.8]	0.5 [0.4, 0.8]	0.615
CRP		4.16 [1.27, 9.24]	7.82 [3.30, 15.23]	3.64 [1.08, 8.54]	<0.001
D-Dimer		764 [470, 1411]	1262 [690, 4616]	717 [465, 4335]	<0.001
IL-6		7.47 [7.40, 19.23]	7.40 [7.34, 17.68]	7.49 [7.43, 19.40]	<0.001
Lymphocytes		16.30 [8.50, 24.80]	6.90 [4.90, 10.95]	18.85 [11.20, 27.02]	<0.001
BMI (kg/m^2)****		30.12 [27.22, 34.05]	32.71 [30.03, 34.18]	29.73 [26.91, 33.34]	0.089
Oxygen saturation (%)		96 [96, 97]	94 [94, 97]	96 [95, 97]	<0.001

(*) The effective "n" was a total of 180: 23 exitus and 157 living; (**) The effective "n" was a total of 181: 23 exitus and 158 living. (***) The effective "n" was a total of 121: 24 exitus and 97 living. (****) The effective "n" was a total of 93 patients.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 2.2: Time course of transaminase levels according to final outcome. Data are expressed in medians and interquartile ranges (Me [IQR]).

		Day 0	Day 7	Day 14	Day 21
AST (UI/L)	Deceased	34 [20, 47]	69 [22, 89]	42 [39, 55]	40 [30, 50]
	Living	32 [25, 44]	28 [24, 44]	28 [24, 52]	24 [20, 29]
ALT (UI/L)	Deceased	20 [17, 33]	98 [28, 118]	71 [31, 150]	106 [90, 121]
	Living	33 [25, 52]	39 [27, 72]	66 [43, 137]	57 [41, 74]
GGT (UI/L)	Deceased	39 [23, 70]	53 [46, 87]	110 [91, 125]	176 [138, 214]
	Living	42 [27, 64]	42 [27, 60]	74 [32, 102]	113 [83, 153]
ALP (UI/L)	Deceased	130 [91, 265]	210 [181, 235]	111 [93, 197]	223 [173, 272]
	Living	85 [63, 138]	124 [79, 157]	115 [80, 122]	88 [67, 94]
Bilirubin (mg/dL)	Deceased	0.6 [0.4, 0.8]	0.7 [0.5, 1.3]	0.4 [0.4, 0.8]	0.4 [0.4, 0.5]
	Living	0.5 [0.4, 0.7]	0.6 [0.4, 0.7]	0.4 [0.3, 0.5]	0.3 [0.3, 0.7]
CRP (mg/dL)	Deceased	7.8 [4.8, 12.1]	4.3 [1.3, 9.2]	9.2 [1.4, 18.2]	11.9 [6.0, 17.8]
	Living	6.0 [2.4, 10.2]	1.9 [0.8, 4.7]	1.3 [0.5, 3.5]	0.2 [0.1, 1.0]
D-Dimer (ng/mL)	Deceased	949 [637, 1452]	835 [777, 1010]	897 [611, 1843]	3999 [3999, 3999]
	Living	644 [417, 1112]	709 [490, 1024]	1050 [582, 1596]	600 [361, 1096]
IL-6 (pg/mL)	Deceased	7.4 [7.4, 7.5]	7.5 [7.4, 76.8]	7.4 [7.3, 7.4]	7.3 [7.3, 7.4]
	Living	7.5 [7.4, 8.9]	7.5 [5.1, 10.1]	7.5 [7.5, 22.3]	7.4 [7.4, 44.5]
Lymphocytes ($x10^9$)	Deceased	11 [6, 18]	5 [2, 8]	5 [4, 5]	6 [5, 7]
	Living	18 [12, 27]	17 [8, 25]	18 [11, 21]	19 [11, 24]
O_2 Sat (%)	Deceased	95 [93, 97]	94 [90, 97]	93 [91, 95]	-
	Living	96 [95, 98]	96 [95, 97]	96 [94, 97]	96 [96, 98]

Given that not all blood samples were analysed with the same apparatus, the raw values of certain variables were not comparable. Therefore, at this point, we divided the raw values for transaminases, IL-6, and CRP by their maximum normal values as a form of weighting. For these variables, we thereby obtain the proportion with respect to the acceptable maximum (<1 implies normal levels, and >1 implies high levels). As transaminases and IMs do not follow a normal distribution, we estimated the exploratory curves for these variables over time for the deceased and living patients by lasso regression analysis (Tibshirani, 1996), since this technique does not assume any parametric distribution for data. Figure 2.1 represents the mean estimate with a 95% confidence interval of these values over time and as a function of the outcome. The left part of the figure shows the transaminase behaviours in these patient groups. The patients who died showed a greater increase in transaminase levels than those who survived, although this increase was only notable for AST and ALP. Similarly, IMs showed a greater increase in the patients who died (right side of Figure 2.1), although only CRP and D-Dimer were significantly elevated compared with the levels in the patients who survived. Lymphocyte counts showed an inverse behaviour, decreasing in the patients who died and rising in the patients who survived.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

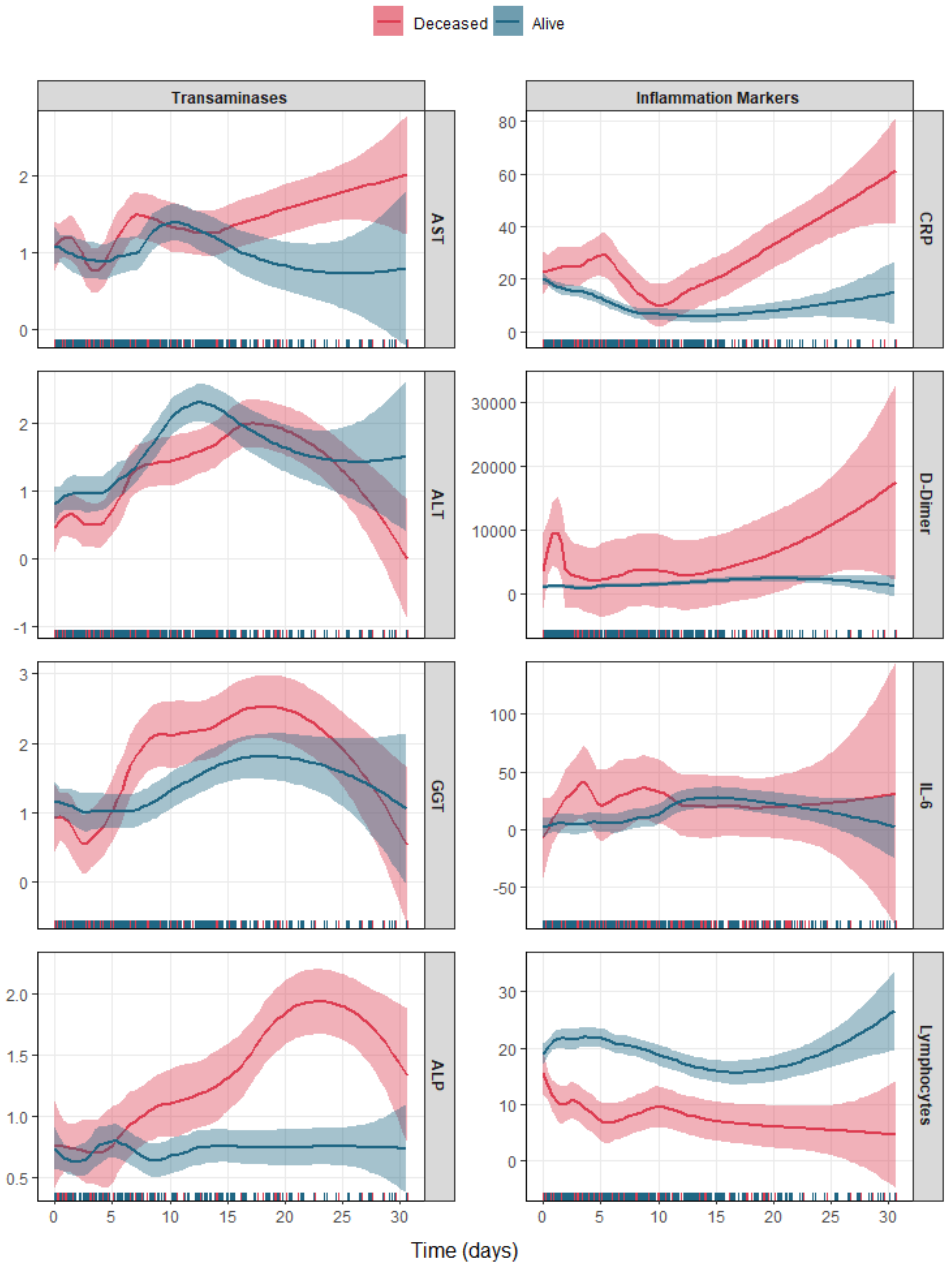

Figure 2.1: Exploratory curves for weighted values (when needed) of transaminases (left) and inflammation markers (right) over time for deceased (pink) and living (blue) patients, along with 95% confidence intervals. Estimation made using a lasso regression analysis.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

We then normalised the transaminases, D-Dimer, and also IL-6 values using a Box–Cox transformation (Box and Cox, 1964) to facilitate statistical modelling calculations. We considered that transforming the CRP and lymphocytes values would not produce any improvement, so we decided to use the weighted values (when necessary) for these covariates.

Survival Exploratory Analysis

As we stated at the beginning of this section, this dataset will also be used in the study of the effect of longitudinal transaminase levels on the survival of the patients. Due to missingness, of the initial 212 individuals, 205 are included in this study, of whom 36 have died (17.56%). Of the total number of patients, 118 (57.56%) where males, and 47 (22.93%) suffered from diabetes.

The graphic representation of the Kaplan-Meier estimators for survival over time (Kaplan and Meier, 1958) is depicted, both globally (Figure 2.2), by gender (Figure 2.3), and diabetes (Figure 2.4). For gender and diabetes, log-rank test (Mantel, 1966) for differences between curves was performed, and the p-value was included in the figure. This test showed significant differences according to diabetes status (p-value = 0.002), with a higher probability of survival among non-diabetics. In contrast, no significant differences were found according to gender (p-value = 0.083).

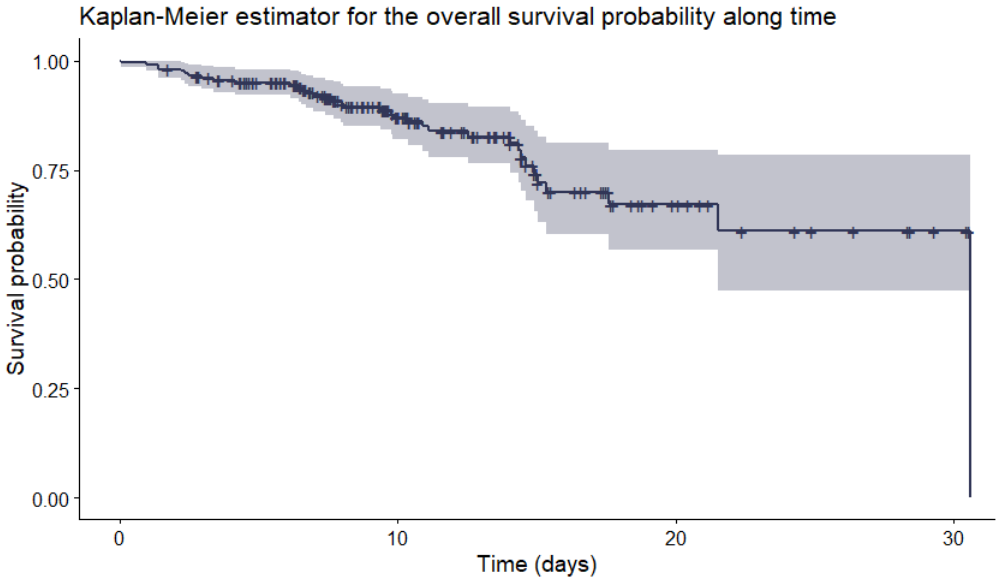

Figure 2.2: Kaplan-Meier estimated curve for overall survival, including 95% confidence intervals.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

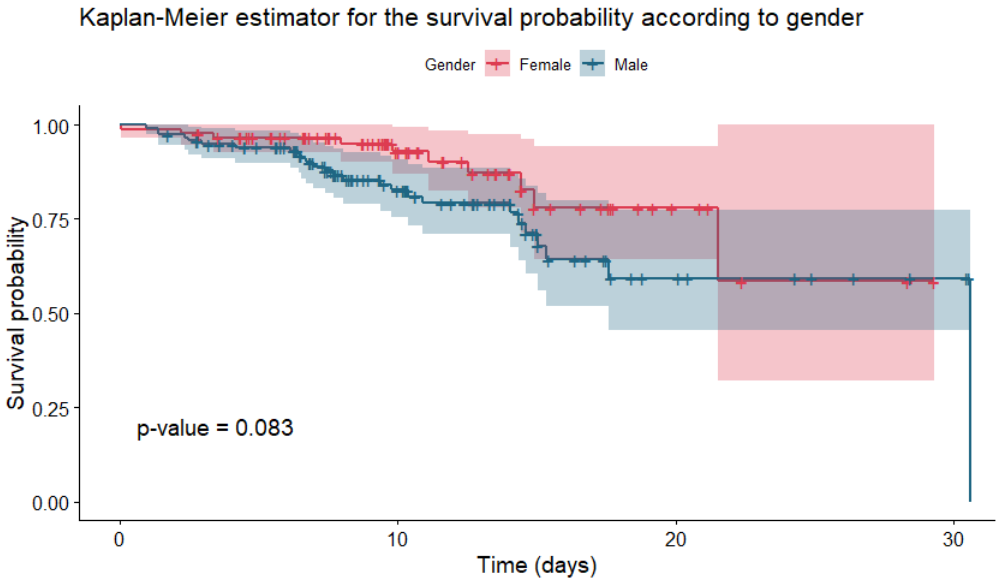

Figure 2.3: Kaplan-Meier estimated curve for survival, including 95% confidence intervals, for female (pink) and male (blue) subjects. P-value from log-rank test for difference between curves is included.

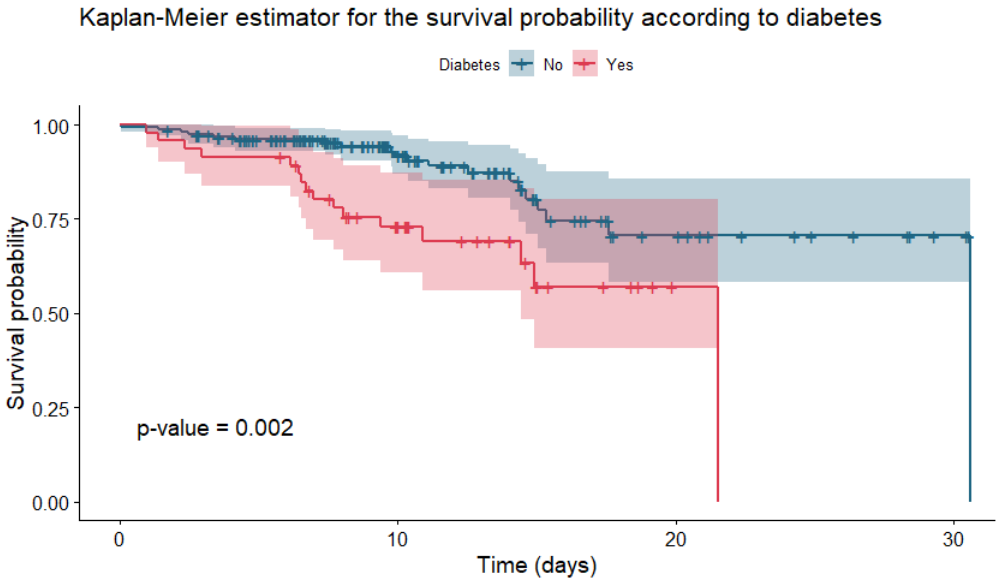

Figure 2.4: Kaplan-Meier estimated curve for survival, including 95% confidence intervals, for patients with (pink) and without (blue) diabetes. P-value from log-rank test for difference between curves is included.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

For exploring the effect of age in survival, we fitted a flexible Cox regression model (Hastie and Tibshirani, 1990a) including gender and diabetes, as well as a smooth effect of age, simply to see if it could be considered in a linear way. In Figure 2.5 this effect is depicted, alongside its 95% confidence interval, and including a horizontal pink line indicating 0. The effect of age is statistically significant from about 45 years of age, and it is not completely linear.

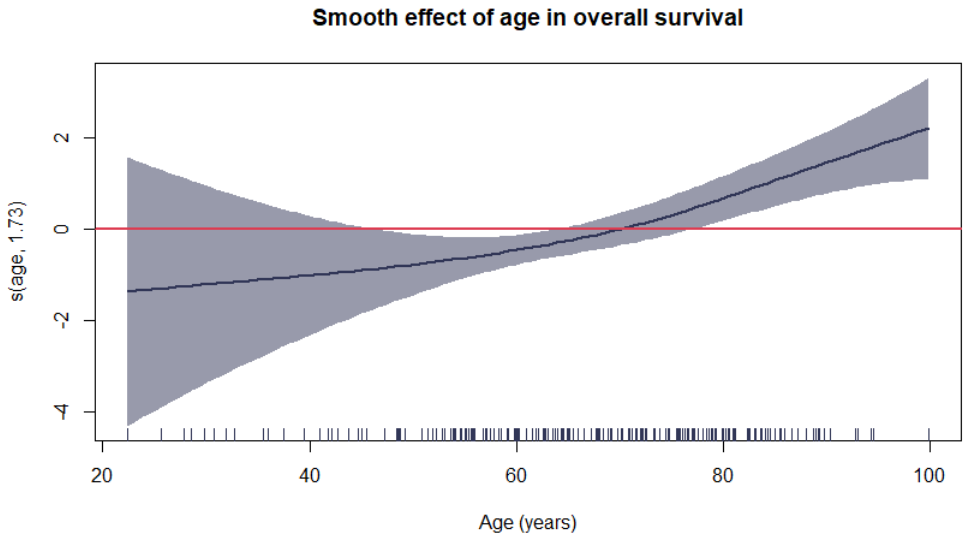

Figure 2.5: Smooth-effect of age in the survival probability, from a Cox additive model including only the baseline covariates. 95% confidence interval is depicted, as well as a pink line indicating 0, i.e. no significance.

2.2 Peripartum Anxiety and Depression (PPAD) Data

Within the "Research Innovation and Sustainable Pan-European Network in Peripartum Depression Disorder - Riseup-PPD" (Cost Action 18138), an international prospective cohort study is underway to assess the impact of COVID-19 in perinatal mental health, carried out by the "Perinatal Mental Health and COVID-19 Pandemic" task force (Motrico et al., 2021). The data available of our second database come from the aforementioned task force. An outline of the study Riseup-PPD is available at www.clinicaltrials.gov, code NCT04595123.

The data come from online questionnaires to women over the age of 18 in the perinatal period (pregnant or with a child of 6 months or less) with residence in

USC DE SANTIAGO DE COMPOSTELA

9 European countries (Albania, Bulgaria, Cyprus, Greece, Malta, Portugal, Spain, Turkey, and the United Kingdom), Israel, Brazil and Chile, from June 2020 to October 2020. These data collected socio-demographic information (age, country of birth (whether born in the country of residence or not), country of residence, educational level (secondary education or lower vs. higher education), cohabitation with partner (yes vs. no), professional status (employed vs. unemployed), first pregnancy (yes vs. no), exposure to COVID-19 (no vs. yes if exposed to at least one of the following: history of COVID-19 diagnosis, COVID-19 related symptoms, direct physical contact with and infected person, or death of a relative with COVID-19), among others), symptoms of depression, assessed by the Edinburgh Postnatal Depression Scale (EPDS; Cox et al., 1987), and symptoms of anxiety, assessed by the Generalised Anxiety Disorder Screener (GAD-7; Spitzer et al., 2006). Moreover, 2019 estimates of Inequality-adjusted Human Development Index (IHDI) were included, in addition to macro variables related to the epidemiological situation and public health measures regarding COVID-19, that are available for each participant's country in the Oxford COVID-19 Government Response Tracker (OxCGRT). In particular, the Containment and Health index (CHI; Hale et al., 2020) was used to assess the level of governmental response across the COVID-19 pandemic period. We used the mean of the 30 days before the completion of the questionnaires for each woman. The number of deaths per million at the time of the questionnaire completion was also included in the study, as an indicator of the impact of COVID-19 on each country's population. Macro variables at country level were collected online from Global Change Data Lab (nd).

2.2.1 Ethical Issues

The study is conducted in accordance with the guidelines of the Declaration of Helsinki and the principles of good practice. Ethical approval was obtained from all the countries before starting the project. Electronic informed consent was obtained from all the participants, and the confidentiality of all the information provided will be ensured.

2.2.2 Descriptive Analysis of PPAD Database

A descriptive analysis of the variables is presented below. Median (Me) and interquartile range (IQR) were computed for continuous variables, and absolute frequencies and percentage were computed for categorical variables. Differences between the pregnant and postpartum women were assessed by means of Fisher's exact test for categorical variables, and the Mann-Whitney U test was employed for the continuous variables.

USC DE SANTIAGO DE COMPOSTELA

A total of 7,642 women, of whom 3,502 were pregnant and 4,140 had already given birth responded to the survey. In Table 2.3, the participants' characteristics are shown, separating according to pregnant and postpartum women, and also primiparous and multiparous women. Multiparous women, either pregnant or in the postpartum period, were older (p-value < 0.001/p-value < 0.001), had lower education (p-value = 0.001/p = 0.019), and were less exposed to COVID-19 (p-value = 0.003/p-value < 0.001) compared to women with a first pregnancy.

Table 2.3: Demographic characteristics of participants in the "Perinatal Mental Health and COVID-19 Pandemic" study.

		Pregnant women				Postpartum women			
		Primiparous n = 2036 n (%)	**Multiparous n = 1466 n (%)**	**Total n = 3502 n (%)**	**p-value**	**Primiparous n = 2366 n (%)**	**Multiparous n = 1774 n (%)**	**Total n = 4140 n (%)**	**p-value**
Maternal age	≤ 24	228 (11.8)	64 (4.6)	292 (8.8)	<0.001	212 (9.4)	62 (3.6)	274 (6.9)	<0.001
	25-35	1420 (73.6)	898 (64.6)	2318 (69.8)		1667 (73.7)	1056 (61.8)	2723 (68.6)	
	≥ 36	281 (14.6)	428 (30.8)	709 (21.4)		382 (16.9)	973(24.5)	937 (24.5)	
	NA	107 (5.2)	76 (5.2)	183 (5.2)		105 (4.4)	170 (4.1)	170 (4.1)	
Maternal country of birth	Country born	1790 (94.1)	1253 (94.5)	3043 (94.2)	0.604	2031 (94.8)	1485 (93.7)	3516 (94.3)	0.141
	Foreign	113 (5.9)	73 (5.5)	186 (5.8)		111 (5.2)	100 (6.3)	211 (5.7)	
	NA	133 (6.5)	140 (9.6)	273 (7.8)		224 (9.5)	159 (9.0)	413 (10.0)	
Live with partner	Yes	1855 (93.0)	1351 (94.5)	3206 (93.7)	0.073	2153 (93.1)	1635 (93.9)	3788 (93.4)	0.291
	No	139 (7.0)	78 (5.5)	217 (6.3)		160 (6.9)	106 (6.1)	266 (6.6)	
	NA	42 (2.1)	37 (2.5)	79 (2.3)		53 (2.2)	33 (1.9)	86 (2.1)	
Exposure to COVID-19	Yes	371 (18.2)	327 (22.3)	698 (19.9)	0.003	463 (19.6)	446 (25.1)	909 (22.0)	<0.001
	No	1665 (81.8)	1139 (77.7)	2804 (80.1)		1903 (80.4)	1328 (74.9)	3231 (78.0)	
	NA	0 (0.0)	0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	0 (0.0)	
History of Mental Problems	Yes	532 (26.1)	382 (26.1)	914 (26.1)	0.962	627 (26.5)	478 (26.9)	1105 (26.7)	0.749
	No	1504 (73.9)	1084 (73.9)	2588 (73.9)		1739 (73.5)	1296 (73.1)	3035 (73.3)	
	NA	0 (0.0)	0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	0 (0.0)	
Mother's education	Secondary education or lower	480 (24.0)	417 (29.2)	897 (26.2)	0.001	549 (23.6)	465 (26.8)	1014 (25.0)	0.019
	Higher education	1521 (76.0)	1009 (70.8)	2530 (73.8)		1779 (76.4)	1270 (73.2)	3049 (75.0)	
	NA	35 (1.7)	40 (2.7)	75 (2.1)		38 (1.6)	39 (2.2)	77 (1.9)	
Unemployment	Yes	362 (17.8)	240 (16.4)	602 (17.2)	0.280	322 (13.6)	229 (12.9)	551 (13.3)	0.520
	No	1674 (82.2)	1225 (83.6)	2899 (82.8)		2044 (86.4)	1543 (87.1)	3587 (86.7)	
	NA	0 (0.0)	1 (0.07)	1 (0.03)		0 (0.0)	2 (0.001)	2 (0.0005)	
Country	Albania	14 (0.7)	6 (0.4)	20 (0.6)	<0.001	14 (0.6)	3 (0.2)	17 (0.4)	<0.001
	Brazil	124 (6.1)	144 (9.8)	268 (7.7)		328 (13.9)	269 (15.2)	597 (14.4)	
	Bulgaria	31 (1.5)	13 (0.9)	44 (1.3)		26 (1.1)	14 (0.8)	40 (1.0)	
	Chile	102 (5.0)	70 (4.8)	172 (4.9)		140 (5.9)	132 (7.4)	272 (6.6)	
	Cyprus	138 (6.8)	82 (5.6)	220 (6.3)		152 (6.4)	97 (5.5)	249 (6.0)	
	Greece	236 (11.6)	101 (6.9)	337 (9.6)		260 (11.0)	116 (6.5)	376 (9.1)	
	Israel	74 (3.6)	142 (9.7)	216 (6.2)		110 (4.6)	222 (12.5)	332 (8.0)	
	Malta	73 (3.6)	41 (2.8)	114 (3.3)		96 (4.1)	53 (3.0)	149 (3.6)	
	Portugal	441 (21.7)	312 (21.3)	753 (21.5)		385 (16.3)	284 (16.0)	669 (16.2)	
	Spain	182 (8.9)	171 (11.7)	353 (10.1)		258 (10.9)	217 (12.2)	475 (11.5)	
	Turkey	524 (25.7)	282 (19.2)	806 (23.0)		407 (17.2)	213 (12.0)	620 (15.0)	
	UK	97 (4.8)	102 (7.0)	199 (5.7)		190 (8.0)	154 (8.7)	344 (8.3)	
	NA	0 (0.0)	0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	0 (0.0)	

Country-related variables are displayed in Table 2.4. Brazil and Chile are the countries with the highest mean CHI, whereas the UK and Malta are the ones with highest IHDI. The total deaths per million was higher in Spain, Brazil and Chile.

USC DE SANTIAGO DE COMPOSTELA

Table 2.4: Country-related variables in the PPAD database. CHI stands for Containment and Health Index mean of the previous 30 days, and IHDI is the Inequality adjusted Human Development Index.

Country	Period of data collection	CHI				IHDI	Total Deaths per million
		Median	P25	P75	Min-Max		Min-Max
Albania	15/07 - 23/08	56.70	56.50	56.77	55.52-56.82	0.708	35.10 - 86.87
Brazil	01/07 - 31/10	67.35	67.35	68.04	61.66-72.83	0.570	285.25 - 752.19
Bulgaria	12/06 - 30/10	39.63	39.23	39.76	37.21-46.69	0.721	24.75 - 180.47
Chile	02/10 - 31/10	70.95	70.95	70.95	70.95-70.95	0.709	673.09 - 743.19
Cyprus	23/06 - 29/10	53.99	53.80	54.72	50.83-62.87	0.805	21.69 - 28.54
Greece	23/06 - 31/10	51.37	49.25	54.31	47.52-59.31	0.791	18.23 - 60.06
Israel	01/07 - 29/10	64.87	52.68	65.54	39.66-72.92	0.814	38.13 - 291.72
Malta	03/07 - 31/10	46.53	45.80	49.02	39.91-53.02	0.823	20.38 - 140.42
Portugal	29/06 - 29/10	62.41	62.20	62.79	61.32-66.81	0.761	153.78 - 238.12
Spain	16/06 - 31/10	57.24	52.97	57.53	47.70-58.99	0.783	580.39 - 767.37
Turkey	22/06 - 31/10	60.65	53.26	66.97	49.39-67.57	0.683	58.98 - 121.56
UK	09/06 - 30/10	60.03	59.39	60.19	55.43-62.10	0.856	573.70 - 682.31

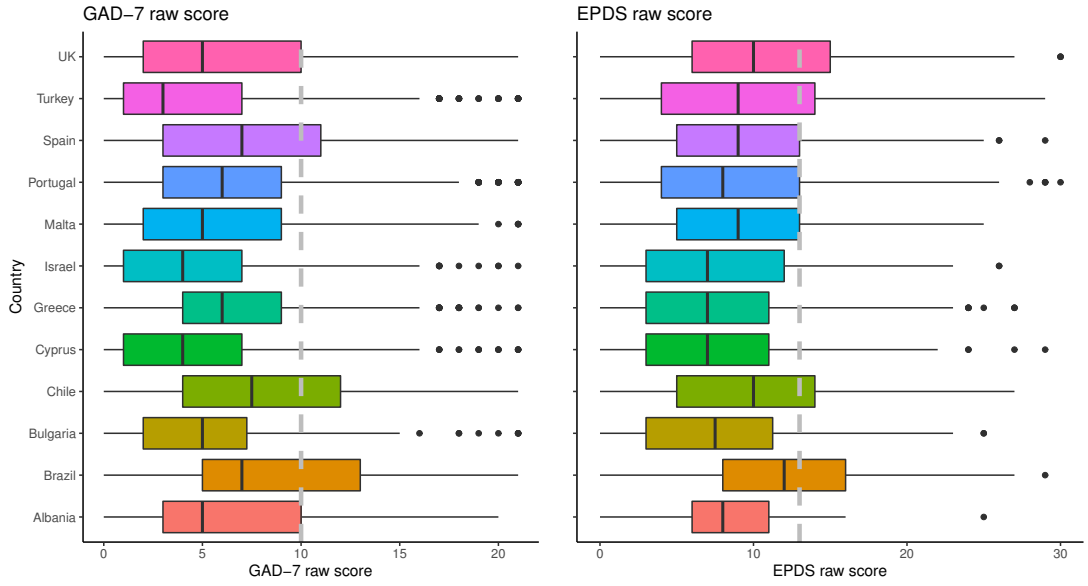

Figure 2.6: Boxplot for GAD-7 (left) and EPDS (right) raw scores for the whole sample. Grey line indicates the cutoff for diagnosis (10 for GAD-7 and 13 for EPDS).

Figure 2.6 shows the boxplot for GAD-7 (left) and EPDS (right) raw scores for the whole sample. Vertical dashed grey lines indicates the cutoff for diagnosis (10 for GAD-7 and 13 for EPDS). In Table 2.5, the results for the GAD-7 and EPDS raw scores for anxiety and depression are depicted. Overall, Brazil stands out as the

USC DE SANTIAGO DE COMPOSTELA

country with the highest depression symptoms. In contrast, Turkey has lower levels of anxiety than the rest of the countries.

Table 2.5: Raw scores obtained from EPDS (Edinburgh Postnatal Depression Scale) and GAD-7 (the Generalised Anxiety Disorder Screener). Data are expressed in median and interquartile range (Me (IQR)).

Country	EPDS			GAD-7		
	Pregnant women n = 3492	Postpartum women n = 4131	Total n=7623	Pregnant women n = 3496	Postpartum women n = 4121	Total n = 7617
Albania	8.0 (6.0-13.5)	8.0 (6.0-10.5)	8.0 (6.0-11.5)	5.0 (1.5-10.8)	5.0 (3.0-7.0)	5.0 (3.0-10.0)
Brazil	11.0 (7.0-16.0)	12.0 (8.0-17.0)	12.0 (8.0-16.0)	7.0 (4.0-12.8)	7.0 (5.0-13.0)	7.0 (4.5-13.0)
Bulgaria	8.0 (3.0-12.8)	7.0 (3.3-11.0)	7.5 (3.0-11.8)	5.0 (1.3-8.8)	4.5 (2.0-7.0)	5.0 (2.0-7.8)
Chile	9.0 (5.0-14.0)	11.0 (5.3-14.0)	10.0 (5.0-14.0)	7.0 (4.0-12.0)	8.0 (4.0-13.0)	7.5 (4.0-12.0)
Cyprus	6.0 (2.0-11.0)	7.0 (3.0-12.0)	7.0 (3.0-11.0)	3.5 (1.0-7.0)	4.0 (1.0-7.0)	4.0 (1.0-7.0)
Greece	6.0 (3.0-11.0)	7.0 (3.0-12.0)	7.0 (3.0-11.5)	6.0 (4.0-8.0)	6.0 (4.0-10.0)	6.0 (4.0-9.0)
Israel	7.0 (4.0-13.0)	7.0 (3.0-12.0)	7.0 (3.0-12.0)	4.0 (1.8-7.0)	3.0 (1.0-6.0)	4.0 (1.0-7.0)
Malta	9.0 (4.0-13.0)	9.0 (5.0-13.5)	9.0 (5.0-13.0)	5.0 (2.0-7.3)	5.0 (2.0-9.5)	5.0 (2.0-9.0)
Portugal	7.0 (4.0-12.0)	9.0 (5.0-13.0)	8.0 (4.0-13.0)	5.0 (3.0-8.0)	6.0 (3.0-9.0)	6.0 (3.0-9.0)
Spain	8.0 (4.0-13.0)	10.0 (5.0-14.0)	9.0 (5.0-13.0)	6.0 (3.0-11.0)	7.0 (3.0-12.0)	7.0 (3.0-11.0)
Turkey	8.0 (4.0-13.0)	9.0 (5.0-14.0)	9.0 (4.0-14.0)	3.0 (0.0-6.0)	4.0 (1.0-7.0)	3.0 (1.0-7.0)
UK	9.0 (4.0-14.0)	11.0 (6.8-15.0)	10.0 (6.0-15.0)	4.0 (2.0-8.0)	6.0 (2.0-10.3)	5.0 (2.0-10.0)
Total	8.0 (4.0-13.0)	9.0 (5.0-14.0)	9.0 (5.0-13.0)	5.0 (2.0-8.0)	6.0 (3.0-10.0)	5.0 (2.0-9.0)

2.3 Thyroid-Related Hormones (TRH) Data

The data we have available comes from A-Estrada Glycation and Inflammation Study (AEGIS), which is a cross-sectional population-based study being performed in the municipality of A Estrada (NW, Spain). Its aim is to investigate the association between glycation (the attachment of a carbohydrate to another biomolecule, such as a protein or lipid), inflammation, lifestyles, and their association with common diseases and to study discordances between markers for glycaemia (Gude et al., 2017). An outline of the AEGIS study is available at www.clinicaltrials.gov, code NCT01796184.

USC DE SANTIAGO DE COMPOSTELA

An age-stratified random sample of the population aged 18 years and older was drawn from Spain's National Health System (NHS) Registry, which covers more than 95% of the population and contains the name, birth date and address of every person entitled to NHS primary care. The sample was stratified into the age groups by decades, with the youngest group from 18 to 29 years, and the oldest group from 80 years of age onwards. First, a computer programme generated a random sample of equal numbers of subjects in each age group (n = 500). Of this initial sample of 3,500 individuals, were excluded:

- 428 due to completion of the recruitment period,
- 84 due to death,
- 211 due to non-response,
- 134 due to change of address,
- 19 due to inability to provide written informed consent (patients with dementia, mental retardation, cerebrovascular disease, terminal cancer, or inability to communicate).

Of the remaining 2,624 persons, a total of 1,516 subjects agreed to participate in the study (overall participation rate, 68%). Participation was lower among men (65%), when compared to women (71%), and also in the youngest and oldest age groups. There were no significant differences in terms of age or residence (rural/urban) between subjects who did and did not agree to participate in the study. From November 2012 through March 2015, all subjects were successively summoned for a one-day evaluation at the Primary Care Centre, which included an interviewer-administered structured questionnaire and fasting venous blood sampling (Gude et al., 2017).

As our objective was to estimate the distribution of thyroid hormone values in a healthy population, we excluded extreme values of TSH, free T4 and free T3 (since they clearly denoted a non diagnosed thyroid-related problem), and also the treated individuals due to high variability of the values, leaving a total of 1,282 healthy (with no thyroid-related problems) subjects.

2.3.1 Ethical Issues

The study was conducted in accordance with the guidelines of the Declaration of Helsinki and the principles of good practice and was approved by Clinical Research Ethics Committee from Galicia, Spain (CEIC2012-025). Written informed consent documents were waived by the Institutional Review Board.

USC DE SANTIAGO DE COMPOSTELA

2.3.2 Descriptive Analysis of TRH Database

We studied the relationship between TSH (mIU/L), free T3 (pmol/L), and free T4 (ng/dL). The relationship between TSH and T4 was broadly studied since it seems that this tandem has the greatest diagnostic value for primary thyroid failure (see, for instance, Kumar et al., 1977) and in most of these studies the logarithm of the TSH (log-TSH) was considered (Spencer et al., 1990; Hoermann et al., 2010; Hadlow et al., 2013), so we are considering it too.

The dataset consists of 1,282 healthy individuals, of which 606 (47.2%) are males. Table 2.6 show the descriptive analysis (median, interquartile range and range) for the age, and the three thyroid-related hormones considered.

Table 2.6: Descriptive measurements of the age and the three outcomes: log-TSH, free T4 and free T3.

	Median	**P25**	**P75**	**Min-Max**
Age (years)	52	39	67	19 - 91
log-TSH	0.69	0.34	1.07	-4.61 - 2.76
Free T4 (ng/dL)	1.17	1.06	1.77	0.68 - 1.77
Free T3 (pmol/L)	3.38	3.14	3.65	2.08 - 6.3

In Figure 2.7, the two-by-two scatter plots of the outcomes are shown. As expected, a strong direct relationship between free T4 and free T3 can be seen. Although less obvious, there also appears to be an inverse relationship between log-TSH and free T4.

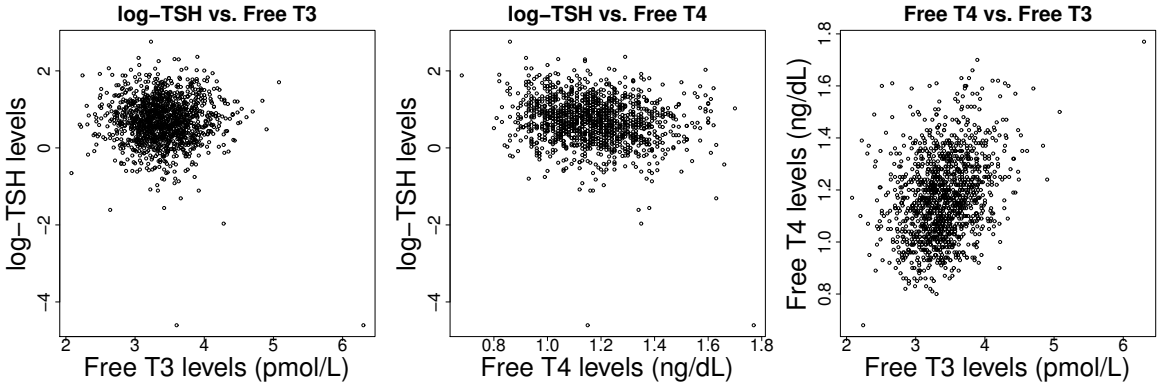

Figure 2.7: Two-by-two scatter plot of the thyroid-related hormones. Left: log-TSH vs. Free T3. Centre: log-TSH vs. Free T4. Right: Free T4 vs. Free T3.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Copula Regression Models for Bivariate Responses

Some of the contents of this chapter were published in Díaz-Louzao et al. (2022). There is also a part of this chapter which has not yet been published (it is currently under submission to *Scientific Reports*), but is available online as a non-peer reviewed preprint (Mesquita et al., 2022).

3.1 Introduction

In December 2019, China reported to the World Health Organisation (WHO) several cases of pneumonia of unknown origin in Wuhan (Huang et al., 2020). Later, it was confirmed that these cases were due to a new coronavirus (SARS-CoV-2), and the disease it caused was named COVID-19 (Lu et al., 2020). This disease spread rapidly to most countries in the world. By 10 July 2020, more than 12.1 million people had been infected, of whom 551,046 had died (World Health Organization (WHO), 2020a) worldwide. At present, the cumulative number of confirmed cases is over 609 million, with around 6.5 million deaths, according to the latest report of 21 September (World Health Organization (WHO), 2022).

Clinical manifestations of COVID-19 range from asymptomatic infection to pneumonia, which can progress to acute respiratory distress syndrome (ARDS), multi-organ failure and ultimately death (Chen et al., 2020; Guan et al., 2020). About 80% of reported cases have mild symptoms, but 15-20% progress to pneumonia, which causes death in 1-5% of patients.

In this context, it is essential to correctly identify the factors that influence the development of a serious illness that leads to admission to the ICU or death, as well as understanding the impact this pandemic situation can have on people's lives.

USC DE SANTIAGO DE COMPOSTELA

This chapter has a twofold objective: on the one hand, the study of liver damage in patients with COVID-19; on the other hand, the study of the impact of the pandemic situation on perinatal mental health. To meet these objectives, we will use JM regression models based on copula models applied, respectively, to the Liver-19 and PPAD databases, described in Chapter 2. In particular, bivariate Copula Generalised Additive Models for Location, Scale and Shape (CGAMLSS; Marra and Radice, 2017) are used.

CGAMLSS regression models have been chosen because of their great flexibility, as they are a extension of GAMLSS (Rigby and Stasinopoulos, 2005) to the bivariate case. There are other conditional copula regression techniques proposed in the literature, but lack the flexibility of the latter, either because they do not allow for additive predictors (Acar et al., 2013), or because they operate under the assumption of Gaussian marginals (Sabeti et al., 2014). There are also proposals where the estimation is done in two steps. Gijbels et al. (2011) estimates covariates effects on two responses correlation, after adjusting for covariates effects on their marginal distributions. Vatter and Chavez-Demoulin (2015) estimation also requires two steps, firstly marginal models were estimated by applying an univariate location, scale, and shape estimation (i.e. GAMLSS), and then both models were joined by a conditional parametric copula. Klein and Kneib (2016) proposed a simultaneous estimation procedure from a Bayesian estimation perspective. CGAMLSS regression models is an adaptation of the aforementioned model under the frequentist perspective. A very interesting alternative is the use of copulas as a special class of bivariate distribution in the framework of Vector Generalised Additive Models (VGAM; Yee, 2015). Like CGAMLSS, VGAMs allow the use of additive predictors for each parameter of a bivariate non-standard response. However, smoothing parameter selection is more difficult and thus leading to a potentially greater bias than in CGAMLSS estimations.

Despite the great flexibility of CGAMLSS models and the enormous potential for their use in biomedical fields, these models are scarcely used in the clinical setting due to the difficult interpretation of the distributional parameters. Therefore, this chapter aims to bring the use of these techniques closer to clinical practice.

3.2 Bivariate Copula Models

Copula functions are multivariate distribution functions whose one-dimensional margins are uniform on the interval $[0, 1]$ (Nelsen, 2007). In the bivariate setting, let us assume a function, $C : [0, 1]^2 \rightarrow [0, 1]$ whose domain is $[0, 1]^2$ and its image is $[0, 1]$. We can say that C is a copula function if it verifies the following properties (Palaro and Hotta, 2006):

USC
DE SANTIAGO
DE COMPOSTELA

(i) $C(y_1, y_2) = 0,\ \forall (y_1, y_2) \in [0,1]^2 \Leftrightarrow y_1 = 0$ and/or $y_2 = 0$.

(ii) $C(y_1, 1) = C(1, y_2) = 1,\ \forall (y_1, y_2) \in [0,1]^2$.

(iii) $\forall (a_1, a_2),\ (b_1, b_2) \in [0,1]^2$, $C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) + C(a_1, b_1) \geq 0$.

More specifically, if y_1 and y_2 are two response variables, by Sklar's theorem (Sklar, 1959) we can say that their joint cumulative distribution function, given the covariates information $\boldsymbol{x}$, can be expressed in terms of the marginal cumulative distribution functions (CDF) and a copula function $C(\cdot, \cdot|\boldsymbol{x})$ that binds them together:

$$F(y_1, y_2|\boldsymbol{x}) = C(F_1(y_1|\boldsymbol{x}), F_2(y_2|\boldsymbol{x})),$$

where $F_1(y_1|\boldsymbol{x})$ and $F_2(y_2|\boldsymbol{x})$ are the marginal CDFs of $y_1|\boldsymbol{x}$ and $y_2|\boldsymbol{x}$, which take values of (0,1). In addition, if y_1 and y_2 are continuous variables, this copula function C is also unique.

Copula functions allows a great flexibility on the responses because their joint distribution might be constructed by combining any parametric distribution using a parametric copula. In addition, literature counts with several copulas allowing different non-Gaussian structures of dependence (see Figure 3.1 for a graphical illustration). For example, the Clayton copula allows for the consideration of asymmetric structures of dependence when two random variables show a stronger positive association at smaller values than at larger values. The Joe or Gumbel copula, in contrast, addresses the opposite situation, and it is also possible to consider rotated versions of the these copulas for modelling negative structures of dependence. Gaussian, FGM and Frank copulas, for instance, do not have any tail dependence.

In copula models, association between the response variables is usually presented in terms of Kendall's association coefficient τ. This estimator takes values between -1 and 1, and a similar interpretation can be made as when using Pearson's linear correlation coefficient r: the closer to 0, the lower the correlation.

3.2.1 Model Formulation

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ be a bivariate continuous variable, observed for each subject $i \in \{1, \ldots, n\}$, where $\boldsymbol{y}_1$, and $\boldsymbol{y}_2$ are the marginal response variables. Also, let $\boldsymbol{X}$ be a set of covariates observed in the same individuals. The conditional copula model can be written as:

$$\begin{aligned} F\left(y_1, y_2, \mid \boldsymbol{x}\right) = C\Big(& F_1\left(y_1 \mid \mu_1(\boldsymbol{x}_{\mu_1}), \sigma_1(\boldsymbol{x}_{\sigma_1}), \nu_1(\boldsymbol{x}_{\nu_1})\right), \\ & F_2\left(y_2 \mid \mu_2(\boldsymbol{x}_{\mu_2}), \sigma_2(\boldsymbol{x}_{\sigma_2}), \nu_2(\boldsymbol{x}_{\nu_2})\right); \\ & \theta(\boldsymbol{x}_\theta)\Big), \end{aligned} \tag{3.1}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

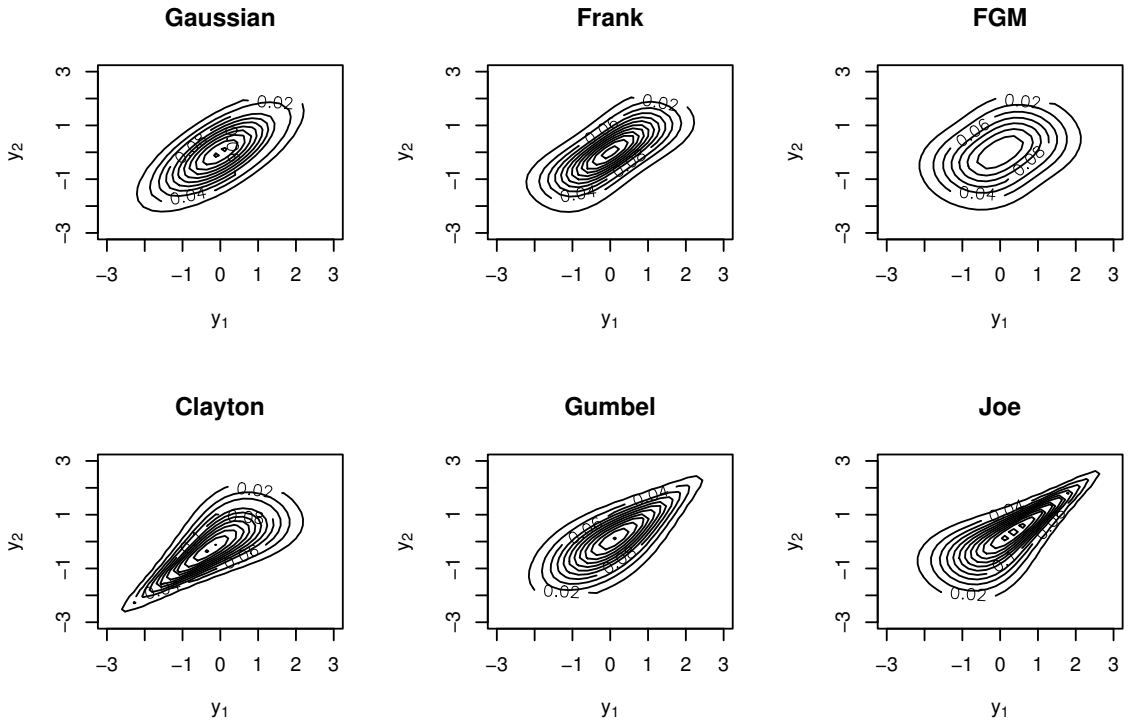

Figure 3.1: Contour plots of different copula functions with standard normal margins for simulated data using a Kendall's τ coefficient = 0.5.

where C represents a parametric copula function defined by the association parameter θ, and F_r are univariate distributions defined by the parameters μ_r, σ_r, and ν_r (which sometimes represents location, scale and shape) for $r = 1, 2$. Here, $\boldsymbol{x}_\zeta$, is a p_ζ-dimensional sub-vector of the complete variable vector $\boldsymbol{x}$, for $\zeta \in Z = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \nu_1, \nu_2, \theta\}$. The model (3.1) is completed by defining the following additive predictors for the marginal and copula parameters, as in (1.3) or (1.4):

$$g_\zeta(\zeta_i) = \eta_{\zeta i} = \beta_0^\zeta + s_1^\zeta(\boldsymbol{x}_{\zeta i1}) + \cdots + s_{p_\zeta}^\zeta(\boldsymbol{x}_{\zeta i p_\zeta}), \tag{3.2}$$

where $\zeta \in Z = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \nu_1, \nu_2, \theta\}$ are the parameters for the univariate marginal distributions and the copula, and g_ζ is the transformation of the corresponding parameter, given by the marginal distributions and copula function selected. Here, for each parameter, β_0^ζ represents a general intercept; $s_j^\zeta(\boldsymbol{x}_{\zeta ij})$ is the generic effect of covariates for the i^{th} individual (being $x_{\zeta j}$ for $j \in \{1, \ldots, p_\zeta\}$ sub-vectors of the complete variable vector $\boldsymbol{x}$). It is noteworthy that different covariates from the entire available set can be used in each additive predictor, i.e. it is not necessary to always

USC
DE SANTIAGO
DE COMPOSTELA

include the same covariates in all predictors.

The choice of the link function is determined by the distribution of the marginals and the copula, in a way that the restrictions that apply to the parameter space of the corresponding parameter are fulfilled. For instance, to model the standard deviation as a function of covariates $g_\sigma(\sigma) = \eta_\sigma$, $\log(\cdot)$ function can be used to ensure positive values (see Marra and Radice (2017) for a more detailed information).

3.2.2 Estimation and Inference

As a general form, each generic effect $s_j^\zeta(\boldsymbol{x}_{\zeta ij})$ can be approximated by a linear combination of basis functions and regression coefficients, as:

$$s_j^\zeta(\boldsymbol{x}_{\zeta ij}) = \sum_{k_j=1}^{K_j} b_{jk_j}^\zeta(\boldsymbol{x}_{\zeta ij})\beta_{jk_j}^\zeta,$$

where $b_{jk_j}^\zeta$ is the $k_j{}^{\text{th}}$ basis function, and $\beta_{jk_j}^\zeta$ the corresponding parameter. Thus, the vector of evaluations $\{s_j^\zeta(\boldsymbol{x}_{\zeta 1j}), \ldots, s_j^\zeta(\boldsymbol{x}_{\zeta nj})\}^\top$ can be written as $\boldsymbol{X}_j^\zeta\boldsymbol{\beta}_j^\zeta$, with $\boldsymbol{\beta}_j^\zeta = (\beta_{j1}^\zeta, \ldots, \beta_{jk_j}^\zeta)^\top$ the vector of the coefficients. This means that the predictor in (3.2) can be formulated as:

$$g_\zeta(\zeta) = \eta_\zeta = \beta_0^\zeta \mathbf{1}_n + \sum_{j=1}^{p_\zeta} \boldsymbol{X}_j^\zeta\boldsymbol{\beta}_j^\zeta,$$

where $\mathbf{1}_n$ is an n-dimensional vector made up all of ones. The effect of covariates can take many forms, including non linear ones, for which smooth regression is advisable. In the case where penalised smooth regression is applied, a penalty term must be taken into account in the estimation of the parameters. For each penalised smooth function s_j^ζ, with vector of coefficients $\boldsymbol{\beta}_j^\zeta$, the penalty term is:

$$\lambda_j^\zeta \left(\boldsymbol{\beta}_j^\zeta\right)^\top \boldsymbol{S}_j^\zeta\boldsymbol{\beta}_j^\zeta,$$

where the penalty matrix $\boldsymbol{S}_j^\zeta$ penalises the squared differences of order d $(1, 2, 3, \ldots)$ of the adjacent parameters $\boldsymbol{\beta}_j^\zeta$, and the parameter $\lambda_j^\zeta \in [0, \ldots, \infty)$ balances bias and variance and it is decisive for the form taken by f_j^ζ. The overall penalty for each predictor ζ can be written as $\left(\boldsymbol{\beta}^\zeta\right)^\top \boldsymbol{S}^\zeta\boldsymbol{\beta}^\zeta$, where $\boldsymbol{S}^\zeta = \operatorname{diag}\left(0, \lambda_1^\zeta\boldsymbol{S}_1^\zeta, \ldots, \lambda_{p_\zeta}^\zeta\boldsymbol{S}_{p_\zeta}^\zeta\right)$, and $\boldsymbol{\beta}^\zeta = \left(\left(\beta_0^\zeta\right)^\top, \left(\boldsymbol{\beta}_1^\zeta\right)^\top, \ldots, \left(\boldsymbol{\beta}_{p_\zeta}^\zeta\right)^\top\right)^\top$. Having this, the overall penalty for the copula model is expressed as:

USC DE SANTIAGO DE COMPOSTELA

$$\boldsymbol{\beta}^\top \boldsymbol{S} \boldsymbol{\beta}, \tag{3.3}$$

where $\boldsymbol{\beta} = \left(\left(\boldsymbol{\beta}^{\mu_1}\right)^\top, \left(\boldsymbol{\beta}^{\sigma_1}\right)^\top, \left(\boldsymbol{\beta}^{\nu_1}\right)^\top, \left(\boldsymbol{\beta}^{\mu_2}\right)^\top, \left(\boldsymbol{\beta}^{\sigma_2}\right)^\top, \left(\boldsymbol{\beta}^{\nu_2}\right)^\top, \left(\boldsymbol{\beta}^{\theta}\right)^\top \right)^\top$ is the vector of all the parameters of the model and $\boldsymbol{S} = \operatorname{diag}\left(\boldsymbol{S}^{\mu_1}, \boldsymbol{S}^{\sigma_1}, \boldsymbol{S}^{\nu_1}, \boldsymbol{S}^{\mu_2}, \boldsymbol{S}^{\sigma_2}, \boldsymbol{S}^{\nu_2}, \boldsymbol{S}^{\theta}\right)$.

The inference of the additive models parameters is based on a maximum likelihood estimator, that can be written as:

$$\begin{aligned} \ell(\boldsymbol{\beta}) = & \sum_{i=1}^{n} \log \left\{ c\left(F_1\left(y_{1i} \mid \mu_{1i}, \sigma_{1i}, \nu_{1i}\right), F_2\left(y_{2i} \mid \mu_{2i}, \sigma_{2i}, \nu_{2i}\right); \theta_i\right) \right\} + \\ & \sum_{i=1}^{n} \sum_{r=1}^{2} \log \left\{ f_r\left(y_{ri} \mid \mu_{ri}, \sigma_{ri}, \nu_{ri}\right) \right\}, \end{aligned}$$

where c is the copula density function, f_r are de density functions for the marginals, and the distributional parameters $\zeta_i = g_\zeta^{-1}(\eta_{\zeta i})$. In the case of penalised regression, the penalised likelihood estimator is the one to be used, constructed by combining the previous formula with (3.3) . Thus regression parameters are given by the maximisation of:

$$\ell_{pen}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S} \boldsymbol{\beta}$$

This is based on a trust region algorithm, whose details can be found in Marra and Radice (2017).

Selection of the marginal distributions, and also selection of the suitable copula can be done according to Akaike's Information Criterion (AIC; Sakamoto et al., 1986) and the Bayesian Information Criterion (BIC; Schwarz et al., 1978). AIC and BIC are defined as AIC$= -2\ell(\hat{\boldsymbol{\beta}}) + 2edf$ and BIC$= -2\ell(\hat{\boldsymbol{\beta}}) + \log(n)$, where the log-likelihood is evaluated at the penalised parameter estimates and edf are the effective degrees of freedom, indicating the number of parameters needed in the model. According to these criteria, lower the AIC or BIC, the better the model.

3.3 Bivariate Joint Modelling of Liver-19 Data

3.3.1 Motivation of the Study

Since the notification in December 2019 of several cases of pneumonia caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the infection has

USC DE SANTIAGO DE COMPOSTELA

spread rapidly throughout the world, reaching pandemic levels. As of the beginning of June 2021, the pandemic has caused over 173 million infections and more than 3.5 million deaths. The disease caused by this agent has been coined coronavirus disease 2019 (COVID-19) by the World Health Organization (World Health Organization (WHO), 2020c). The main organ affected is the lungs, and most patients present respiratory symptoms; however, gastrointestinal symptoms and especially liver damage are also frequent during the infection (Richardson et al., 2020; Phipps et al., 2020; Popkin et al., 2020).

The presence of liver damage during the course of SARS-CoV-2 infection has been reported in patients with and without previous liver disease (Jothimani et al., 2020; Cai et al., 2020; Xu et al., 2020; Paštrovic et al., 2021). This involvement is characterized by a predominant elevation of alanine transaminase (ALT) and aspartate transaminase (AST) levels, which affects 15%-62% of patients and can be accompanied by significant increases in bilirubin levels. Most of these alterations are characterized by being mild-moderate transient increases and are related to the severity of the viral infection. Autopsies of some patients who died during infection have shown liver abnormalities compatible with hepatic steatosis and portal inflammation (Sun et al., 2020).

The exact mechanism of liver damage is currently unknown, but various hypotheses have been proposed (Marjot et al., 2021):

- **The direct action of the virus.** There are at least 3 types of receptors for the virus in the liver (angiotensin-converting enzyme 2, transmembrane serine protease 2, and furin), distributed in cholangiocytes and hepatocytes, which suggests that the direct action of the virus on liver cells induces liver damage, in a similar manner as at the pulmonary level, by means of a virus-receptor interaction (Sun et al., 2020; Pirola and Sookoian, 2020).

- **Induced by the systemic inflammatory response.** There is also the possibility that the liver damage is related to the degree of systemic inflammatory response, in the scenario of the so-called "cytokine storm" and that the liver behaves as an innocent bystander.

- **Anoxia.** This is secondary to the respiratory involvement that occurs mainly in the most severe cases of infection, and could lead to sustained hypoxia of the liver tissue, with liver cell damage consequently.

- **The drug toxicity.** Possibly derived from the multiple treatments administered, especially during the first wave, could explain the subsequent liver damage due to hepatotoxicity.

USC DE SANTIAGO DE COMPOSTELA

An exhaustive review of articles published to date generally shows a "still photo" of liver disorders in COVID-19, such as altered transaminase and bilirubin levels at a certain point in the disease progression. Few studies have dynamically analysed these alterations over time, and even fewer have studied their potential relationships with biological markers such as inflammation markers (IMs) (Phipps et al., 2020; Bloom et al., 2021; Lei et al., 2020; Oussalah et al., 2020; Li et al., 2020a; Xu et al., 2021). The moment a liver disorder appears can indicate whether there is any relationship with the inflammatory response, drug administration, or tissue anoxia and can infer whether the damage is due to the direct action of the virus. Specifically directed studies, with repeatedly and exhaustively collected variables, might therefore shed light on whether the temporary establishment of liver disorders is due to the immediate direct effect of the virus or to other mechanisms. The intention of this study is to describe the liver disorders in a cohort of patients from the *Complejo Hospitalario Universitario de Santiago de Compostela* (Spain) corresponding to the "first wave" of COVID-19. By employing an innovative statistical joint modelling methodology, we seek to investigate the dynamics of the modifications in transaminase levels analysed longitudinally over time and to correlate them with IMs, C-reactive protein (CRP), interleukin-6 (IL-6), D-Dimer, and lymphocytes as a function of survival.

3.3.2 Statistical Analysis of Liver-19 Data

Statistical modelling of Liver-19 data was performed in two phases: first, a univariate response analysis for each transaminase (AST, ALT, GGT, ALP), to assess the factors that could influence them; second, JM to study the correlation between each transaminase and each inflammation factor.

It is worth noting that, due to missingness in ALP (see Table 2.1 from Chapter 2), and in the interest of losing as little data as possible, different databases were used depending on the variables involved. In this case, a bigger dataset was used when the transaminases were AST, ALT or GGT (n = 212 patients), and a smaller one for when ALP was the transaminase of interest (n = 121).

Univariate Analysis of Transaminase Values

In a preliminary analysis, each transaminase figure was evaluated individually. To do so, we fitted a Generalised Additive Mixed Model (GAMM; Wood, 2017) for each Box-Cox transformed transaminase. This transformation was performed in order to normalised the response variables and, for now on, we are going to refer to the transformed variables with their original names, for simplicity. We adjust for sex, exitus (binary for alive or deceased), antiviral treatment (binary for lopinavir/ritonavir

USC DE SANTIAGO DE COMPOSTELA

and/or hydroxychloroquine administration), age, IM-exitus association, and time-exitus association. We also included random intercepts by patient to consider the variability within each individual, since the values recorded for the same individual are expected to be more closely related to each other than to the values of another individual. In a GAMM regression model, the GAM framework is used, and the random effects are included in the model using P-splines. This means that the models adjusted were of the form (1.3). In our case, the smooth effects of continuous covariates were considered by means of thin-plate spline smoothers.

The GAMM regression model considered here was as follows (3.4):

$$\begin{aligned} Transaminase_{ij} = \beta_0^{t,im} + Sex_i\beta_1^{t,im} + s_1^{t,im}(Age_i) + AT_i\beta_2^{t,im} + Exitus_i\beta_3^{t,im} + \\ s_2^{t,im}(IM_{ij}, Exitus_i) + s_3^{t,im}(Time_{ij}, Exitus_i) + u_i, \end{aligned} \tag{3.4}$$

where $Transaminase$ and t refer to AST, ALT, GGT, or ALP, AT stands for antiviral treatment, and IM and im are CRP, D-Dimer, IL-6, or lymphocytes. In the above equation, $i \in \{1, \ldots, n\}$ indicates de individual, and $j \in \{1, \ldots, n_j\}$ indicates the longitudinal observation of individual i. $\beta_0^{t,im}$ indicates the intercept of the model, $s_2^{t,im}(IM_{ij}, Exitus_i)$ represents the IM curve according to $Exitus$, i.e. a possible different non-linear effect of IM for deceased and living patients, and the same for $s_3^{t,im}(Time_{ij}, Exitus_i)$. u_i indicates the random intercepts by patient. All the smoothed effects are centred to ensure uniqueness.

A total of 16 regression models was adjusted, 4 for each transaminase. After these first models, a second GAMM for each transaminase was fitted, including for each transaminase only the statistically significant covariates from all four corresponding previous models (3.4). Statistical significance was established for p-values less than 0.05 and, in the case of smoothed continuous covariates, for those ranges of the variable in which the estimated 95% confidence intervals do not include zero. For the categorical covariates, as well as for the continuous variables whose smoothed effect was linear, we reported the corresponding estimated β coefficient. The effective degrees of freedom (edf) were also reported for all covariates. The edf number represents the parameters needed to estimate the effect of a covariate, i.e. edf indicates the complexity of the functional form of the effect. When the effect is linear, the edf is approximately 1. The results of this first analysis are shown in Table 3.1.

In Table 3.1 we can see that ALT is associated with age, increasing in middle age and decreasing significantly after the eighth decade. Despite its lack of statistical significance, the trend was similar in the analysis of age and GGT and ALP. The following variables were statistically significant in this analysis: AST increased with IL-6 (p-value < 0.001), ALT decreased with CRP, and GGT decreased with lymphocytes. These 3 variables (AST, ALT and GGT) increased in turn with D-Dimer.

USC
DE SANTIAGO
DE COMPOSTELA

Figure 3.2 shows the smooth effect of time, according to exitus (left part of the figure), and age on the transaminase levels. With regard to the time variable, ALT and ALP increased significantly, reaching a maximum from the second week in the patients who died. In addition, the AST levels decreased significantly and the GGT increased in those who survived. The latter drops significantly at the end of the observation period in both these patients who lived and in those who died.

Table 3.1: Results for the generalised additive mixed models for the weighted and Box–Cox transformed values of transaminases. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. edf $= 1$ indicates linear dependence.

Outcome	Variables	Estimate (β)	SE	edf	P-value
AST	Intercept	-2.697	1.090	1.000	0.014
	Exitus (Ref: Alive)	0.060	0.108	1.000	0.579
	D-Dimer	0.952	0.452	1.000	0.036
	IL-6	0.102	0.029	1.000	<0.001
	s_1^{AST}(Age)	See Figure 3.2		1.000	0.715
	s_2^{AST}(Time, Exitus = Alive)	See Figure 3.2		4.892	<0.001
	s_2^{AST}(Time, Exitus = Deceased)	See Figure 3.2		1.001	0.270
ALT	Intercept	-1.657	0.713	1.000	0.020
	Exitus (Ref: Alive)	-0.251	0.126	1.000	0.046
	CRP	-0.008	0.001	1.000	<0.001
	D-Dimer	0.579	0.265	1.000	0.029
	s_1^{ALT}(Age)	See Figure 3.2		2.988	0.003
	s_2^{ALT}(Time, Exitus = Alive)	See Figure 3.2		4.689	<0.001
	s_2^{AST}(Time, Exitus = Deceased)	See Figure 3.2		3.965	<0.001
GGT	Intercept	-4.028	1.135	1.000	<0.001
	Gender (Ref: Male)	0.339	0.126	1.000	0.008
	Exitus(Ref: Alive)	-0.131	0.190	1.000	0.490
	CRP	-0.002	0.002	1.000	0.373
	D-Dimer	1.549	0.465	1.000	<0.001
	IL-6	-0.017	0.030	1.000	0.566
	Lymphocytes	-0.007	0.003	1.000	0.021
	s_1^{GGT}(Age)	See Figure 3.2		2.777	0.085
	s_2^{GGT}(Time, Exitus = Alive)	See Figure 3.2		5.103	<0.001
	s_2^{GGT}(Time, Exitus = Deceased)	See Figure 3.2		4.194	<0.001
ALP (UI/L)	Intercept	-0.584	0.055	1.000	<0.001
	Exitus (Ref: Alive)	0.242	0.131	1.000	0.065
	s_1^{ALP}(Age)	See Figure 3.2		2.054	0.569
	s_2^{ALP}(Time, Exitus = Alive)	See Figure 3.2		4.397	0.004
	s_2^{ALP}(Time, Exitus = Deceased)	See Figure 3.2		3.871	<0.001

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

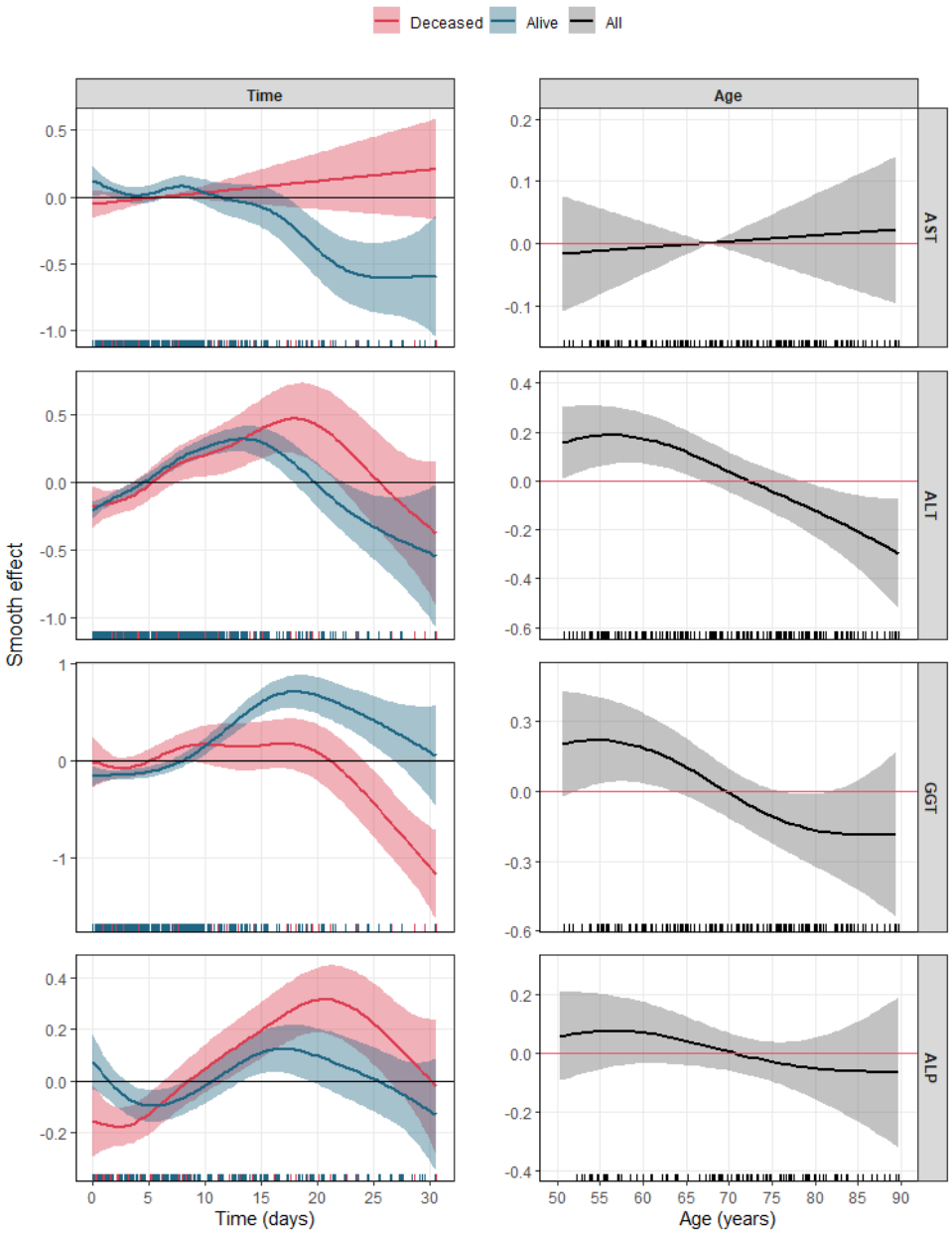

Figure 3.2: Estimated centred smooth effect (with 95% CI) of time (left) and age (right) over the weighted and Box–Cox transformed transaminases extracted from the generalised additive mixed models in Table 3.1. Pink stands for deceased patients, blue for the living, and black for the total number of patients. The horizontal line indicates zero, and when it lies inside the 95% CI, no statistical significance was found.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Longitudinal Correlations between Transaminases and Inflammation Markers

The temporal correlation between transaminase levels and IM was estimated using the bivariate CGAMLSS. Previous to the adjustment of the model, the best distribution for each marginal had to be chosen. We did not consider three parameter distributions as these typically require larger sample sizes to produce reliable results. We selected the best marginal distribution according to AIC and BIC criteria, and also graphically with the quantile-quantile plots (Q-Q plots) of the normalised quantile residuals. All the Box–Cox transformed covariates followed a Gaussian distribution (as this was the aim of the transformation, no selection was done here), and the CRP and lymphocytes each followed a gamma distribution. In Table 3.2, the AIC and BIC values for the possible marginal distributions for CRP and lymphocytes are shown. The selection of the best marginal distribution for CRP and lymphocytes was performed in both the dataset including AST, ALT and GGT, and the dataset including ALP, leading in both cases to the same marginal distribution.

Table 3.2: Comparison of AIC and BIC values for the candidate marginal distributions for CRP and lymphocytes, both in the bigger set (with AST, ALT and GGT) and the smaller set (for ALP).

	CRP				**Lymphocytes**			
	AST, ALT and GGT set		**ALP set**		**AST, ALT and GGT set**		**ALP set**	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Gamma	6739.027	6748.683	1595.464	1602.260	7839.351	7849.248	2478.703	2486.313
Weibull	6752.364	6762.019	1597.650	1604.447	7841.349	7851.247	2483.622	2491.232
Log-normal	6988.187	6997.842	1625.543	1632.339	7921.924	7931.822	2488.691	2496.301
Reverse Gumbel	7279.160	7288.815	1809.658	1816.454	7905.502	7915.400	2528.398	2536.009
Inverse Gamma	8128.935	8138.590	1739.518	1746.314	7981.711	7991.609	2499.839	2507.449
Logistic	7582.111	7591.766	1893.867	1900.663	8090.987	8100.885	2611.806	2619.417
Normal	7689.161	7698.816	1919.079	1925.876	8100.041	8109.939	2623.605	2631.215
Gumbel	259.930	8269.585	2042.212	2049.008	8561.432	8571.330	2786.468	2794.078

In Figure 3.3, the histogram and the estimated density of CRP and lymphocytes, along with the Q-Q plots, for the two data sets and the gamma marginal distribution are depicted. We can see that the dots in the Q-Q plot fit the red line very well, indicating that the chosen distribution may be adequate. However, there are some departures for higher and lower values of the variables.

To assess the potential effect of time, drugs, sex, and age on the correlation between transaminases and IMs, the regression model used for each transaminase-IM pair is determined by the model (3.5). As we are not interested here in modelling transaminases and inflammation markers individually, but rather in their correlation, θ is the only covariate-adjusted parameter.

USC DE SANTIAGO DE COMPOSTELA

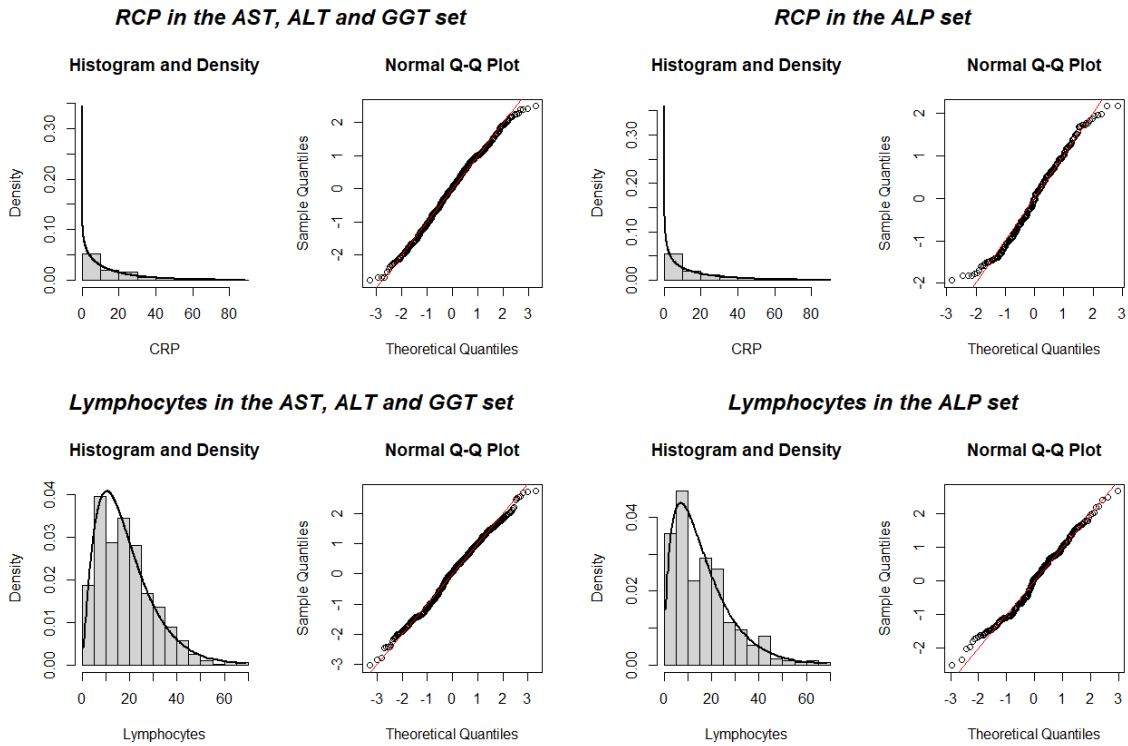

Figure 3.3: Histogram and estimated density curve (left panel) and Q-Q plot (right panel) for CRP (top) and lymphocytes (bottom) considering a gamma marginal distribution.

$$\begin{cases} \eta_l^{\mu_1} = \beta_0^{\mu_1} \\ \eta_l^{\mu_2} = \beta_0^{\mu_2} \\ \eta_l^{\sigma_1} = \beta_0^{\sigma_1} \\ \eta_l^{\sigma_2} = \beta_0^{\sigma_2} \\ \eta_l^{\theta} = \beta_0^{\theta} + Sex_l\beta_1 + s_1(Age_l) + AT_l\beta_2 + Exitus_l\beta_3 + s_2(Time_l, Exitus_l) \end{cases}, \quad (3.5)$$

where $l \in \{1, \ldots, N\}$ denotes the number of observation, not the individual (note that the total number of observations, $N = \sum_{i=1}^{n} n_i$, is different from the number of individuals n).

The best copula function for each of the 16 regression models was chosen following AIC and BIC criteria, and making sure the model converged (results not shown). Table 3.3 shows the chosen copula function for each transaminase-IM pair.

USC DE SANTIAGO DE COMPOSTELA

Table 3.3: Best copula function for the regression models for each transaminase (columns) and IM (rows) pair.

	AST	**ALT**	**GGT**	**ALP**
CRP	Normal	Normal	Normal	Plackett
D-Dimer	Frank	Normal	Plackett	Gumbel 180º
IL-6	Normal	Normal	Plackett	Plackett
Lymphocytes	Plackett	Frank	Plackett	Frank

Gumbel 180º refers to Gumbel copula rotated by 180º, addressing a situation with stronger positive association at smaller values instead that at larger values, as happens with the original Gumbel copula function.

Table 3.4 shows the regression models, indicating only the statistically significant terms. Sex affected the correlation between D-Dimer and AST and between D-Dimer and ALT, increasing in women (β = 1.231 [p-value = 0.007] and β = 0.272 [p-value $<$ 0.001], respectively). Sex also had an influence (but inverse), making the correlation lower in women between AST and lymphocytes, between ALT and CRP, between GGT and CRP, between GGT and lymphocytes, and between ALP and lymphocytes (β = -0.450 [p-value = 0.020]; β = -0.187 [p-value = 0.007]; β = -0.234 [p-value = 0.023]; β = -0454, [p-value = 0.028]; and β = -2.118 [p-value = 0.010], respectively). Lastly, the patients who underwent antiviral therapy showed an increased correlation between D-Dimer and GGT (β = 0.410; p-value = 0.046), and this same variable showed a decreased correlation between lymphocytes and ALP (β = -2.118; p-value = 0.010). The association of these variables did not affect the final model. For the sake of a better understanding, although not entirely correct, we have denoted as $\eta_l^\theta(Trans, IM)$ the estimator of the parameter for the transaminase $Trans$ and the inflammation marker IM.

Effect of time on the correlation between transaminases and IMs. The temporal correlation between transaminases and IMs (Figure 3.4) showed a different dynamic between the survivors and the deceased. In the former (Figure 3.4, in blue), CRP and AST showed a positive correlation that weakened from the second week of hospitalization. In this patient group and from the second week on, CRP and ALT showed a negative relationship, as did CRP and GGT. In contrast, the correlation between CRP and ALP in the surviving patients was positive. D-Dimer showed an initially positive correlation to AST and ALT that became negative from the second week in the patients who survived. In contrast, the relationship between D-Dimer, GGT, and ALP in the surviving patients was positive. IL-6 and AST in the surviving patients showed an initially positive correlation that weakened after 15 days. The correlation between IL-6 and GGT was also positive in these patients.

USC DE SANTIAGO DE COMPOSTELA

Table 3.4: Regression models for correlation between weighted and Box–Cox transformed values of transaminases and weighted and/or Box–Cox transformed (if required) IM. From the general model type (3.5), every regression model in this table only includes the covariates that were found to be statistically significant ($p<0.05$). $s(\cdot)$ indicates the smooth effect of the covariate. For Sex, the reference level is "Male"; for $Exitus$, is "Living"; and for AT (antiviral treatment), the reference level is "No". The smooth effects of $s_1(Age_l)$ and $s_2(Time_l, Exitus_l)$ in the correlation are depicted in Figures 3.5 and 3.4, respectively.

$\eta_l^\theta(AST, CRP) = s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(AST, D-Dimer) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(AST, IL-6) = s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(AST, Lymphocytes) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALT, CRP) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALT, D-Dimer) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALT, IL-6) = s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALT, Lymphocytes) = Exitus_l\beta_3 + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(GGT, CRP) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(GGT, D-Dimer) = s_1(Age_l) + AT_i\beta_2 + Exitus_l\beta_3 + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(GGT, IL-6) = s_2(Time_l, Exitus_l)$
$\eta_l^\theta(GGT, Lymphocytes) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALP, CRP) = s_1(Age_l) + AT_i\beta_2 + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALP, D-Dimer) = s_1(Age_l) + s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALP, IL-6) = s_2(Time_l, Exitus_l)$
$\eta_l^\theta(ALP, Lymphocytes) = Sex_l\beta_1 + s_1(Age_l) + s_2(Time_l, Exitus_l)$

In the patients who died (Figure 3.4, in pink), the correlation was positive between CRP and AST and between GGT and ALP, again after the second week. Although the correlation between D-Dimer and ALT was not statistically significant, it was positive between AST, GGT, and ALP in the patients who died. In this group of patients, the correlation was significant and positive for IL-6 and AST, GGT, ALP, and ALT, although at the end of the observation period the correlation became negative for ALT.

Lastly, the lymphocyte counts in the survivors showed a negative correlation for AST and a positive one for GGT and ALP from the second week. In the patients who died, the correlation was statistically significant and negative for AST, GGT, and ALP and especially for ALT, although it became positive at the end of the observation period.

USC DE SANTIAGO DE COMPOSTELA

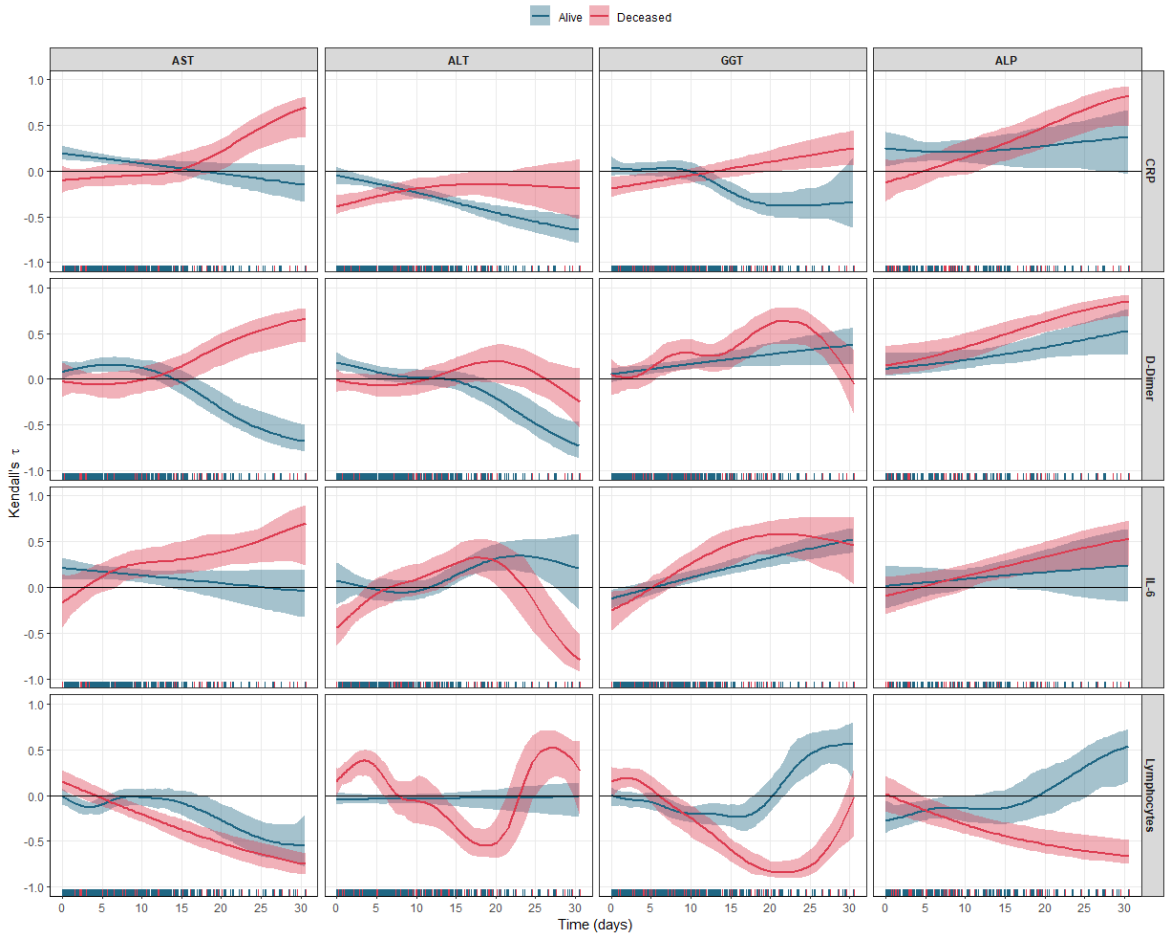

Figure 3.4: Time variation of the correlation between the weighted and Box–Cox transformed transaminases and the weighted and/or Box–Cox transformed (if required) inflammation markers for deceased (pink) and living (blue) patients, expressed by means of Kendall's τ coefficient (with 95% CI). The horizontal line indicates zero correlation, and when it lies inside the 95% CI, no correlation is found.

Effect of age on the correlation between transaminases and IMs. Figure 3.5 shows the correlation between transaminases and IMs as a function of age; this correlation was lost in those older than 80 years. The AST and CRP correlation weakened in these older patients and became negative with ALT and GGT. With D-Dimer and AST, ALT had a negative correlation in this age subgroup and again weakened in its relationship with GGT and ALP, although the correlation with the latter continued to be positive. The correlation between IL-6, ALT, and GGT again became negative. In contrast to these markers in lymphocytes, the correlation expressed a positive

USC
DE SANTIAGO
DE COMPOSTELA

trend as age advanced.

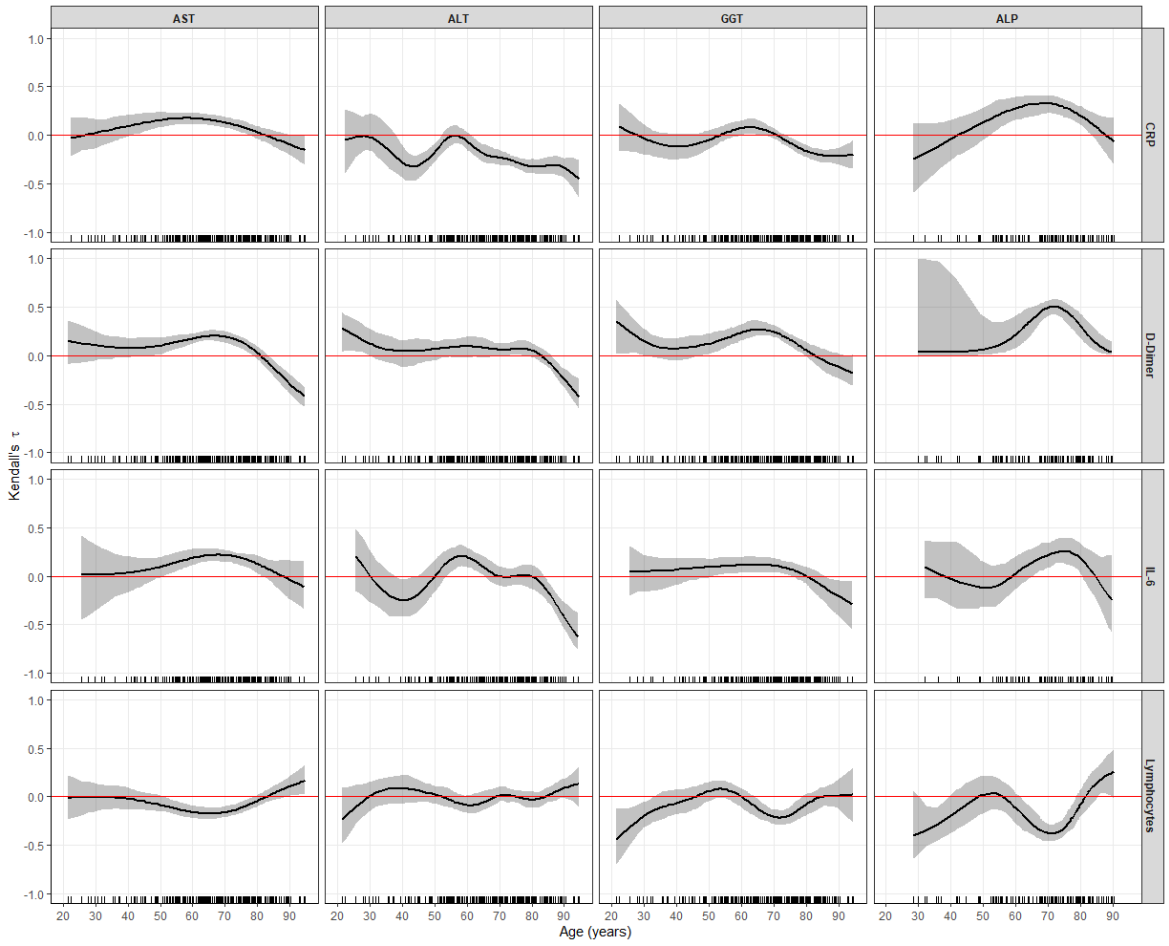

Figure 3.5: Variation of the correlation between the weighted and Box–Cox transformed transaminases and the weighted and/or Box–Cox transformed (if required) inflammation markers across the age of the patients, expressed by means of Kendall's τ coefficient (with 95% CI). The horizontal red line indicates zero correlation, and when it lies inside the 95% CI, no correlation is found.

3.3.3 Discussion

There is currently an abundance of articles indicating the high frequency of liver disorders in COVID-19. Nevertheless, it is still unclear why the liver is so frequently affected and what mechanism is ultimately involved (Cai et al., 2020; Hundt et al., 2020; Phipps et al., 2020; Richardson et al., 2020; Xu et al., 2020; Kunutsor and

USC DE SANTIAGO DE COMPOSTELA

Laukkanen, 2021; Paštrovic et al., 2021; Sharma et al., 2021). Understanding why the liver experiences these changes is key to understanding the pathophysiology of COVID-19.

As with other studies, our study found that a high proportion (60%) of patients with COVID-19 had elevated transaminase levels. Unlike other studies, the percentage of patients with a hepatocellular and those with a cholestatic pattern was similar (Hundt et al., 2020; Phipps et al., 2020). It is well documented that patients with previous liver damage appear more likely to present alterations with infection (Guan et al., 2020; Jothimani et al., 2020; Ridruejo and Soza, 2020; Tian and Ye, 2020; Wang et al., 2020; Marjot et al., 2021; Sharma et al., 2021). Our cohort was characterized by a low incidence of underlying liver disease, as well as of HCV and HBV infection. In addition, they were "first wave" patients treated with antivirals (lopinavir-ritonavir) and antimalarials (hydroxychloroquine) but not with immunosuppressive drugs such as corticosteroids. We believe that both characteristics facilitate the explanation of the potential mechanisms of damage, given that they are not masked by certain therapeutic interventions or previous pathology.

Another common finding in other studies is that older patients with ischaemic heart disease and diabetes died more frequently from COVID-19. There is also a trend toward higher mortality in patients with higher body mass index. The missing data in this study have probably prevented us from observing a significant effect for obesity on prognosis, as reported in other publications (Arentz et al., 2020; Grasselli et al., 2020; Jothimani et al., 2020; Popkin et al., 2020; Richardson et al., 2020). We also observed that mortality was higher in those patients with greater higher increase in IMs, as has been reported in other studies (Guan et al., 2020; Herold et al., 2020).

We found that the patients who died showed a significant increase in AST and ALP levels, whereas the patients who survived showed a significant decrease in AST levels. The increase in ALP levels and their temporal dynamics is especially striking. The increases in this enzyme in COVID-19 have been described in the medical literature as moderate (Phipps et al., 2020; Marjot et al., 2021). The peculiar dynamics of ALP in this study have been revealed by the statistical methodology employed, given that we performed a continuous temporal analysis, the effects of which can be obviated in an analysis at a specific moment.

The mechanisms of liver damage have been associated with a direct action of the virus, the inflammatory response, hepatotoxicity, and anoxia (Pirola and Sookoian, 2020; Sun et al., 2020; Marjot et al., 2021). Although the present study was not designed to determine the direct effects of the virus on its target organ, the statistical methodology applied allows us to infer the indirect mechanisms of damage. A correlation analysis found an association between IMs and transaminases. From the second week of hospitalization, the patients who died showed positive and significant 2-to-2 correlations between CRP and AST and between GGT and ALP. The D-

USC DE SANTIAGO DE COMPOSTELA

Dimer correlations are equally significant for this with AST, GGT, and ALP, which also occurs with IL-6. In contrast, the surviving patients showed negative correlations between CRP, ALT, and GGT, an effect that was equally significant in those patients who survived, from the second week of hospitalization, between D-Dimer, AST, and ALT. The relationship between Il-6 and transaminases in these surviving patients is irrelevant.

In the patients who died, an inverse correlation between lymphocytes and AST, GGT, ALP, and even ALT is especially striking, although this effect was lost with this enzyme at the end of the observation period. Also of particular interest is the markedly negative correlation between ALP and lymphocytes for the patients who died and the positive correlation for those who survived. In adjusting the correlations with covariates, antiviral drugs exert a modifying effect, increasing the correlation between GGT and D-Dimer and decreasing the correlation between ALT and CRP; however, these effects did not affect the form of the relationship. On the other hand, these drugs were administered in a similar proportion to the patients who lived and to those who died, which makes it impossible to attribute a significant effect to their use, both in terms of non-mortality and hepatotoxicity. We therefore believe that the treatment administered did not play a role in the liver damage. In this work, a sequential alteration of inflammatory markers followed by enzyme elevation has been detected, which could suggest that the latter is a consequence of inflammation, although a direct cytopathic effect cannot be ruled out. This would justify the performance of specially directed studies to validate this hypothesis. The observed cholestatic pattern of elevated ALP levels in the patients who died in our study is more biologically plausible given that the different distribution of specific receptors for the virus in hepatocytes and cholangiocytes is well documented and is predominant in the latter (Pirola and Sookoian, 2020; Marjot et al., 2021; Sharma et al., 2021).

The effect of age is particularly interesting. In a regression analysis, age behaved as a variable associated with ALT; ALT levels increase early in life and decrease from the eighth decade. When we established the correlations between age and the correlation itself with the markers of inflammation and transaminases, we observed that they became negative after the eighth decade. There is no clear biological explanation for this phenomenon, although the immunological senescence observed in vaccine trials could point to the effect (Herold et al., 2020; Liu et al., 2020). Patients with higher increases in inflammation parameters showed higher mortality (Anderson et al., 2020; Ramasamy et al., 2020), which contrasts with the fact that older patients despite showing less inflammation die more frequently; however, it is also true that older patients are affected by higher rates of diabetes, obesity, and heart disease, which influence mortality.

Various studies with longitudinal measurements of transaminases have been

USC DE SANTIAGO DE COMPOSTELA

published, and a number of them have included inflammation parameters (Lei et al., 2020; Li et al., 2020a; Oussalah et al., 2020; Bloom et al., 2021; Xu et al., 2021). In a cross-sectional study, Phipps et al. (2020) found a relationship between ferritin and IL-6 with severe liver damage in a large cohort of patients with COVID-19. Another longitudinal study established a relationship between ferritin and AST elevation but not with other IMs (Bloom et al., 2021). Another study observed a relationship between leukocyte counts, decreased lymphocyte counts, and increased transaminase levels (Anderson et al., 2020). In the remaining publications, specific associations were found between inflammation and liver disorders on one hand and mortality on the other but not specifically between IMs and liver disorders (Guan et al., 2020; Li et al., 2020a; Oussalah et al., 2020; Xu et al., 2021).

In the present study, we used innovative flexible regression joint modelling techniques that, applied in this area, allow us to assess the dynamics of the temporal correlation between IMs and elevated transaminase levels. Furthermore, the advantage of this statistical methodology is that it allows us to adjust the temporal correlation for covariates such as age, sex, and treatment administration. For the successful application of this methodology, an exhaustive data capture was crucial, in which the measurements have been taken continuously over time, employing a specifically designed database. The use of these techniques has allowed us to establish a dynamic temporal representation of transaminases different from that of other studies. As far as we know, ours is the first study to use this type of tool to analyse liver disorders in SARS-CoV-2.

The retrospective nature of this study limits our conclusions. Despite exhaustive data capture, certain data was lacking to establish more robust conclusions. We also lacked histological studies. A histology compatible with immune damage would reinforce our conclusions. In addition, we only analysed a certain IM battery, which limits the conclusions to the analysed markers, although the statistical techniques employed can be extended to a larger panel of markers.

In conclusion, the present study once again confirms the frequent alteration of transaminase levels in patients with COVID-19 infection. A joint-modelling statistical methodology established that the patients who died from COVID-19 had characteristically elevated AST and ALP levels and showed a significant and sequential correlation over time with IMs. The patients who survived were better protected from inflammation-mediated liver damage. This relationship between transaminases and IMs weakened with age.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3.4 Bivariate Joint Modelling of PPAD Data

3.4.1 Motivation of the Study

During the perinatal period, the monitoring of anxiety and depression is of vital importance because they are related to poorer child development, both physically and cognitively, fostering the onset of behavioural problems and even mental disorders (Beardselee et al., 1998; Weissman et al., 2006; Goodman et al., 2011). Among the factors contributing to perinatal mental health disorders, we should highlight exposure to natural disasters or stressful life events (Harville et al., 2010). This fact, coupled with the global COVID-19 pandemic situation declared by the WHO on 11th March 2020, stresses the importance of surveillance during the current situation, as aggravation of negative emotional outcomes in these women is to be expected (Chmielewska et al., 2021; Iyengar et al., 2021).

After the global pandemic situation was declared, governments in many countries implemented a wide range of early non-pharmaceutical interventions in an attempt to reduce the burden of disease and to enable health systems to prepare themselves (Hale et al., 2020; Kucharski et al., 2020; Ngonghala et al., 2020; Teslya et al., 2020; Regmi and Lwin, 2021), showing disparate effectiveness in containing the disease (Liu et al., 2021). These interventions had a large economic and social impacts, namely in terms of mental health (Bäuerle et al., 2020; Farkhad and Albarracín, 2021) and health-related quality of life (Hay et al., 2021). Consequently, consideration of the contribution of government measures implemented to contain the disease is crucial to fully understand the impact of the COVID-19 pandemic on the mental health of the global population and, in particular, of women in the prenatal and postnatal period.

According to different studies conducted since the beginning of the pandemic, the prevalence of anxiety almost doubled compared to pre-pandemic rates (Fawcett et al., 2019), ranging from 34% to 42% in pregnant (Yan et al., 2020; Sun et al., 2021; Fan et al., 2021). Furthermore, rates of prenatal depression were reported varying from 25% to 31% (Yan et al., 2020; Fan et al., 2021; Sun et al., 2021), and around 22% for the postpartum period (Yan et al., 2020). It is noteworthy that not all women seem to be equally affected, with higher risk for perinatal mental disorders found among those living in socio-economic disadvantaged contexts (Effati-Daryani et al., 2020; Wu et al., 2020). This circumstance raises awareness to socio-economic inequalities when considering the COVID-19 impact on mental health during the perinatal period.

The main goal of this study is to analyse the role of governmental measurements and individual factors on symptoms of depression and anxiety in pregnant women and postpartum women, while considering country Inequality-adjusted Human

USC DE SANTIAGO DE COMPOSTELA

Development Index (IHDI) and individual factors such as age, parity, and exposure to COVID-19.

3.4.2 Statistical Analysis of PPAD Data

Multivariate analyses were performed separately for pregnant and for postpartum women, with GAD-7 and EPDS raw scores as dependent variables. Test and questionnaires scores, although often treated as continuous variables, are not. In this case, they are discrete variables with a large number of possible values (GAD-7 takes integer values from 0 to 21, while EPDS takes integer values from 0 to 30). We are dealing with what is called "counting data". It follows, then, that the distribution of the marginal variables is not Gaussian, so we have to select the appropriate distribution. As we did in Section 3.3, the selection fo the marginals was performed using AIC and BIC criteria, as well as graphical tools (histogram and estimated density plot and Q-Q plot). In Table 3.5, the AIC and BIC values for the possible marginal distributions for GAD-7 and EPDS raw scores are shown, for pregnant and postpartum women separately.

Table 3.5: Comparison of AIC and BIC values for the candidate marginal distributions for GAD-7 and EPDS raw scores, both for pregnant and postpartum women.

	Pregnant womem				**Postpartum women**			
	GAD-7		**EPDS**		**GAD-7**		**EPDS**	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Discrete Generelised Pareto	19674.57	19686.89	21469.51	21481.82	24071.77	24084.42	4.00	16.653
Negative Binomial - type I	19830.07	19842.39	21833.84	21846.15	24326.61	24339.25	26374.21	26386.861
Negative Binomial - type II	19830.07	19842.39	21833.84	21846.15	24326.61	24339.25	26374.21	26386.861
Poisson Inverse Gaussian	20171.22	20183.54	22163.98	22176.30	24721.08	24733.73	26773.96	26786.615
Poisson	26093.15	26099.31	27898.61	27904.76	32013.21	32019.54	32640.74	32647.065

In Figures 3.6 and 3.7, the histogram and the estimated density of GAD-7 and EPDS, along with the Q-Q plots, for the two best distributions according to AIC and BIC are depicted, for pregnant and postpartum women respectively. We can see that, even though Discrete Generalised Pareto (DGP) distribution was the best distribution, according to AIC and BIC criteria, the fit is not good. In fact, this could already be sensed at least for EPDS in postpartum women, as the difference between AIC and BIC for this distribution compared to the others is too large (AIC and BIC of 4.00 and 16.653 respectively, compared to amounts of more than 26000). In contrast, although in the Q-Q plot, the points do not fit perfectly to the red line for high values, the Negative Binomial - type I (NBI) distribution seems to be more adequate. This is an example of the importance of not being guided by analytical criteria alone, and that a graphical analysis can be of great help on many occasions.

USC DE SANTIAGO DE COMPOSTELA

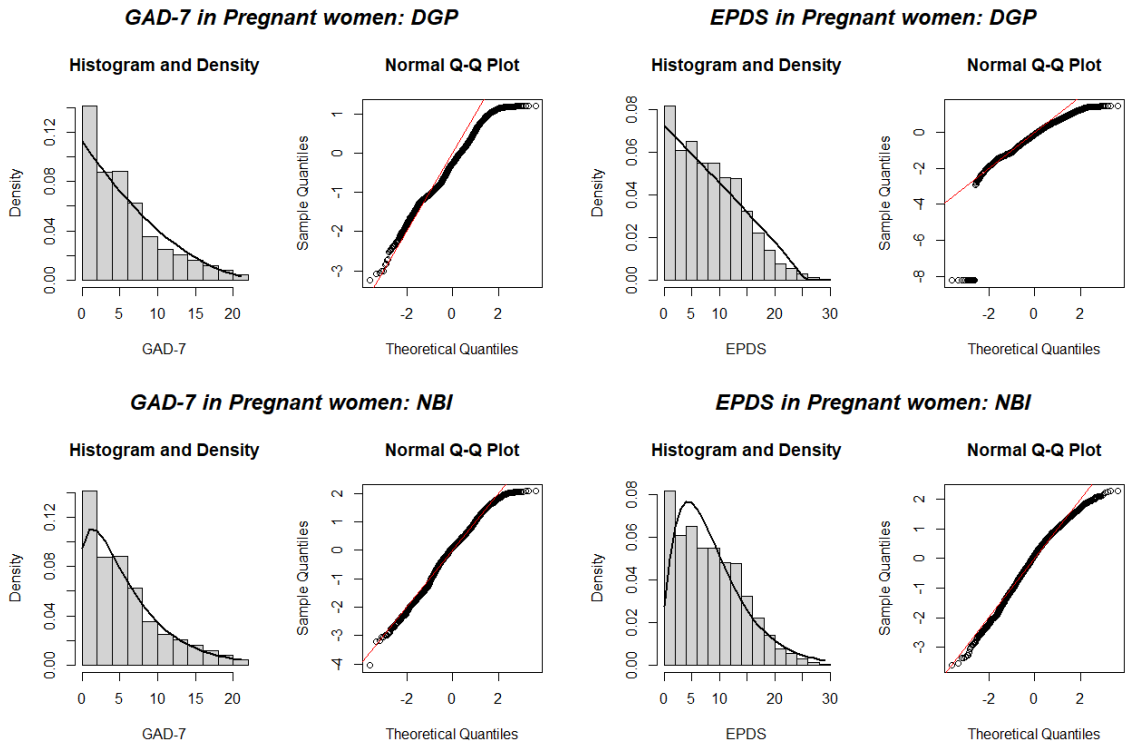

Figure 3.6: Histogram and estimated density curve (left panel) and Q-Q plot (right panel) for GAD-7 (left) and EPDS (right) in the pregnant dataset, considering a Discrete Generalised Pareto (DGP, on top) and a Negative Binomial - type I (NBI, at the bottom) as marginal distributions.

Once that the marginal distributions were properly selected, the CGAMLSS for GAD-7 and EPDS raw scores were fitted, according to parity (multiparous vs. primiparous), exposure to COVID-19 (no vs. yes if exposed to at least one of the following: history of COVID-19 diagnosis, COVID-19 related symptoms, direct physical contact with an infected person, or death of a relative with COVID-19), age (in years), and country-related covariates (Containment and Health Index (CHI), Deaths per million and IHDI). For interpretability, the total number of deaths per million will be counted in hundreds. As we stated when summarising the variables in Chapter 2, CHI is computed as the mean value for the 30 previous day to the completion of the questionnaire. It should be noted that other indexes related to government measurements (as the Stringency Index) were also considered at first in previous univariate regression analysis, but was the CHI the best performer, even thought results were very similar (not shown). After many trials, we have come to the model (3.6) for pregnant women.

USC DE SANTIAGO DE COMPOSTELA

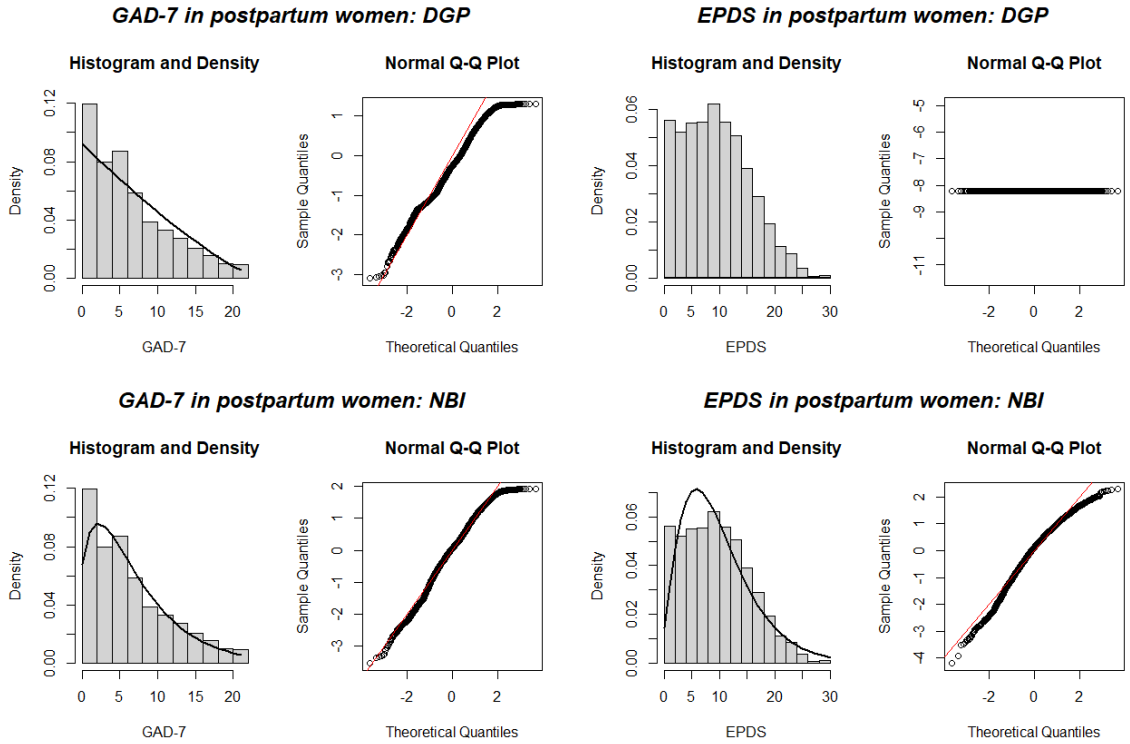

Figure 3.7: Histogram and estimated density curve (left panel) and Q-Q plot (right panel) for GAD-7 (left) and EPDS (right) in the mother dataset, considering a Discrete Generalised Pareto (DGP, on top) and a Negative Binomial type I (NBI, at the bottom) as marginal distributions.

$$
\begin{cases}
\eta_{ij}^{\mu_1} = \beta_0^{\mu_1} + fp_{ij}\beta_1^{\mu_1} + ce_{ij}\beta_2^{\mu_1} + tdh_{ij}\beta_3^{\mu_1} + CHI_{ij}d_{m,ij}\beta_4^{\mu_1} + CHI_{ij}d_{p,ij}\beta_5^{\mu_1} + \\
\qquad IHDI_i d_{m,ij}\beta_6^{\mu_1} + IHDI_i d_{p,ij}\beta_7^{\mu_1} + s_1^{\mu_1}(Age_{ij}) + u_i^{\mu_1} \\
\eta_{ij}^{\mu_2} = \beta_0^{\mu_2} + fp_{ij}\beta_1^{\mu_2} + ce_{ij}\beta_2^{\mu_2} + tdh_{ij}\beta_3^{\mu_2} + CHI_{ij}d_{m,ij}\beta_4^{\mu_2} + CHI_{ij}d_{p,ij}\beta_5^{\mu_2} + \\
\qquad IHDI_i d_{m,ij}\beta_6^{\mu_2} + IHDI_i d_{p,ij}\beta_7^{\mu_2} + s_1^{\mu_2}(Age_{ij}) + u_i^{\mu_2} \\
\eta_{ij}^{\sigma_1} = \beta_0^{\sigma_1} \\
\eta_{ij}^{\sigma_2} = \beta_0^{\sigma_2} \\
\eta_{ij}^{\theta} = \beta_0^{\theta} + fp_{ij}\beta_1^{\theta} + ce_{ij}\beta_2^{\theta} + tdh_{ij}\beta_3^{\theta} + CHI_{ij}d_{m,ij}\beta_4^{\theta} + CHI_{ij}d_{p,ij}\beta_5^{\theta} + \\
\qquad IHDI_i d_{m,ij}\beta_6^{\theta} + IHDI_i d_{p,ij}\beta_7^{\theta} + s_1^{\theta}(Age_{ij}) + u_i^{\theta}
\end{cases}, \quad (3.6)
$$

where $i \in \{1, \ldots, 12\}$ denotes the number of countries of origin for the participants, $j \in \{1, \ldots, n_i\}$ are the participants from each country, fp indicates first pregnancy, ce indicates COVID-19 exposure, tdh is total of deaths due to COVID-19 in hun-

USC DE SANTIAGO DE COMPOSTELA

dreds, and $dm = \mathbb{I}_{\{fp=\text{“}Multiparous''\}}$ is the indicator operators for a woman being multiparous ($dm = 1$) or not ($dm = 0$), and the same for $dp = \mathbb{I}_{\{fp=\text{“}Primiparous''\}}$. This indicators are just an artificial tool needed for introducing linear interactions. For instance, $CHI_{ij}d_{m,ij}\beta_4 + CHI_{ij}d_{p,ij}\beta_5$ just indicates the interaction between CHI and fp, which means that the effect of CHI is allowed to be different depending on whether this is the woman's first pregnancy.

Finally, u is the random intercept indicating the effect of the country. By introducing the random effect, intra-country correlation is taking into account, since it is expected that women from the same country are more closely related to each other than to women from other countries.

In (3.6), we used the same formulas for modelling the mean values of the two marginal distributions, and the correlation between the outcomes. The variance was assumed constant. The model for the postpartum women, displayed in (3.7) is very similar, with the only difference that the age of the mother is considered linear.

$$\begin{cases} \eta_{ij}^{\mu_1} = \beta_0^{\mu_1} + fp_{ij}\beta_1^{\mu_1} + ce_{ij}\beta_2^{\mu_1} + tdh_{ij}\beta_3^{\mu_1} + CHI_{ij}d_{m,ij}\beta_4^{\mu_1} + CHI_{ij}d_{p,ij}\beta_5^{\mu_1} + \\ \qquad IHDI_i d_{m,ij}\beta_6^{\mu_1} + IHDI_i d_{p,ij}\beta_7^{\mu_1} + Age_{ij}\beta_8^{\mu_1} + u_i^{\mu_1} \\ \eta_{ij}^{\mu_2} = \beta_0^{\mu_2} + fp_{ij}\beta_1^{\mu_2} + ce_{ij}\beta_2^{\mu_2} + tdh_{ij}\beta_3^{\mu_2} + CHI_{ij}d_{m,ij}\beta_4^{\mu_2} + CHI_{ij}d_{p,ij}\beta_5^{\mu_2} + \\ \qquad IHDI_i d_{m,ij}\beta_6^{\mu_2} + IHDI_i d_{p,ij}\beta_7^{\mu_2} + Age_{ij}\beta_8^{\mu_2} + u_i^{\mu_2} \\ \eta_{ij}^{\sigma_1} = \beta_0^{\sigma_1} \\ \eta_{ij}^{\sigma_2} = \beta_0^{\sigma_2} \\ \eta_{ij}^{\theta} = \beta_0^{\theta} + fp_{ij}\beta_1^{\theta} + ce_{ij}\beta_2^{\theta} + tdh_{ij}\beta_3^{\theta} + CHI_{ij}d_{m,ij}\beta_4^{\theta} + CHI_{ij}d_{p,ij}\beta_5^{\theta} + \\ \qquad IHDI_i d_{m,ij}\beta_6^{\theta} + IHDI_i d_{p,ij}\beta_7^{\theta} + Age_{ij}\beta_8^{\theta} + u_i^{\theta} \end{cases}, \quad (3.7)$$

For both model for pregnant and postpartum women, the best copula was the Gumbel copula function (AIC = 36194.75 and BIC = 36545.28 for pregnant women, and AIC = 44617.28 and BIC = 44952.77 for postpartum women). This makes sense, since the Gumbel copula allows for the consideration of asymmetric structures of dependence when two random variables show a stronger positive association at larger values than at smaller ones (see Figure 3.1).

In Tables 3.6 and 3.8, numerical results for the marginal models for pregnant and postpartum women, respectively, are shown. Exposure to COVID-19 (ce) was associated with increased anxiety symptoms both in pregnant and postpartum women (more specifically, women exposed to COVID-19 had a GAD-7 score 12.1% and 17.1% higher than those not exposed). Women's age was associated with anxiety symptoms; this was a linear association for postpartum women (symptoms of anxiety decreased with age, 1% every year) and a non-linear association for pregnant women, in whom anxiety is higher for younger women, and decreases linearly until the age of 30, from which it remains approximately constant (see upper left panel of

USC DE SANTIAGO DE COMPOSTELA

Figure 3.8). The total number of deaths per million (in hundreds, tdh) was associated with increased anxiety symptoms, but only in postpartum women (5.7% increase in GAD-7 raw score for each one-hundredth increase in the number of deaths). IHDI and CHI had a differential association with anxiety among pregnant and postpartum women. Multiparous women tended to have higher anxiety in lower IHDI contexts, both in pregnant and postpartum women. Symptoms of anxiety were also higher for multiparous women lower CHI contexts, but only for postpartum women. In Figure 3.9, the estimated centred random effects of the countries are depicted. We can see that, both pregnant and postpartum women from Chile, Greece and Portugal presented higher levels anxiety than the other countries, while, on the other hand, women from Cyprus and Turkey presented lower levels. In addition, postpartum women from Malta presented also higher anxiety levels, while postpartum women from Israel and Spain presented lower levels than postpartum women from the other countries.

With regard to symptoms of depression, these were lower for primiparous compared to multiparous (70.1% and 52.7% for pregnant and postpartum women respectively). COVID-19 exposure and the total number of national deaths per million (in hundreds) were associated with increased depression both in pregnant and postpartum women. Women's age was associated with depression symptoms, too; as well as for anxiety, this was a linear association for postpartum women (symptoms of depression decreased about 0.7% per year of age) and a non-linear association for pregnant women, in whom anxiety is higher for younger women, and decreases linearly until the age of 30 (see upper right panel of Figure 3.8). CHI was associated with symptoms of depression in multiparous women, with lower levels of depression with higher CHI. In addition, contexts of higher IHDI were associated with lower symptoms of depression in multiparous pregnant women. According to Figure 3.9, Maltese pregnant women presented higher depression symptoms than the average, while for Cypriot and Spanish pregnant women was the opposite. In Bulgaria and Spain, postpartum women presented lower levels of depression symptoms, while the opposite situation was the case in Malta, Portugal, Turkey and UK.

Tables 3.7 and 3.9 present the numerical results for the correlation models between GAD-7 and EPDS. All significant fixed effects for correlation in both models are related to COVID-19. For pregnant women, the total number of deaths per million (in hundreds) is associated with higher correlation between depression and anxiety symptoms. For postpartum women, exposure to COVID-19 increases correlation, while higher levels of CHI implies lower correlation, but only in multiparous women. Finally, Figure 3.9 shows that the correlation between the two outcomes in both pregnant and postpartum women from Chile was higher, and it was lower for women from Turkey. The correlation was also lower in Cyprus only for pregnant women. In Greece, correlation has the opposite effect in pregnant women (where it

USC
DE SANTIAGO
DE COMPOSTELA

was higher than average) than in postpartum women (where it was lower).

Table 3.6: Numerical results for the marginal models (using two negative binomial - type I distributions) for pregnant women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. Ratio (95%CI) is the exponential of the β estimates and their 95% Confidence Intervals.

		Estimate (β)	**SE**	**Ratio (95%CI)**	**edf**	**P-value**
GAD-7 raw score	Intercept	3.36	0.71	28.789 (7.159, 115.766)	1	<0.001
	fp (ref: Multiparous)	-0.983	0.561	0.374 (0.125, 1.124)	1	0.080
	ce (ref: No)	0.114	0.04	1.121 (1.036, 1.212)	1	0.004
	tdh	0.029	0.016	1.029 (0.998, 1.062)	1	0.061
	CHI for Multiparous	-0.008	0.004	0.992 (0.984, 1.000)	1	0.082
	CHI for Primiparous	-0.004	0.004	0.996 (0.988, 1.004)	1	0.344
	IHDI for Multiparous	-1.606	0.798	0.201 (0.042, 0.959)	1	0.044
	IHDI for Primiparous	-0.633	0.793	0.531 (0.112, 2.512)	1	0.424
	s_1(Age)	See Figure 3.8			2.851	<0.001
	Country (random effects)	See Figure			8.448	<0.001
EPDS raw score	Intercept	3.511	0.404	33.482 (15.168, 73.908)	1	<0.001
	fp (ref: Multiparous)	-1.208	0.454	0.299 (0.123, 0.727)	1	0.008
	ce (ref: No)	0.108	0.032	1.114 (1.046, 1.186)	1	0.001
	tdh	0.028	0.008	1.028 (1.012, 1.045)	1	0.001
	CHI for Multiparous	-0.008	0.003	0.992 (0.986, 0.998)	1	0.022
	CHI for Primiparous	0.001	0.003	1.001 (0.995, 1.007)	1	0.651
	IHDI for Multiparous	-1.284	0.376	0.277 (0.133, 0.579)	1	0.001
	IHDI for Primiparous	-0.443	0.371	0.642 (0.310, 1.329)	1	0.233
	s_1(Age)	See Figure 3.8			4.670	<0.001
	Country (random effects)	See Figure 3.9			4.901	0.027

Table 3.7: Numerical results for the correlation model between GAD-7 and EPDS for pregnant women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect.

	Estimate (β)	**SE**	**edf**	**P-value**
Intercept	-0.219	1.102	1	0.843
fp (ref: Multiparous)	0.975	1.063	1	0.359
ce (ref: No)	-0.025	0.074	1	0.733
tdh	0.059	0.026	1	0.023
CHI for Multiparous	0.006	0.008	1	0.508
CHI for Primiparous	-0.002	0.007	1	0.789
IHDI for Multiparous	-0.047	1.112	1	0.966
IHDI for Primiparous	-0.682	1.101	1	0.536
s_1(Age)	See Figure 3.8		3.991	0.333
Country (random effects)	See Figure 3.9		6.572	<0.001

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

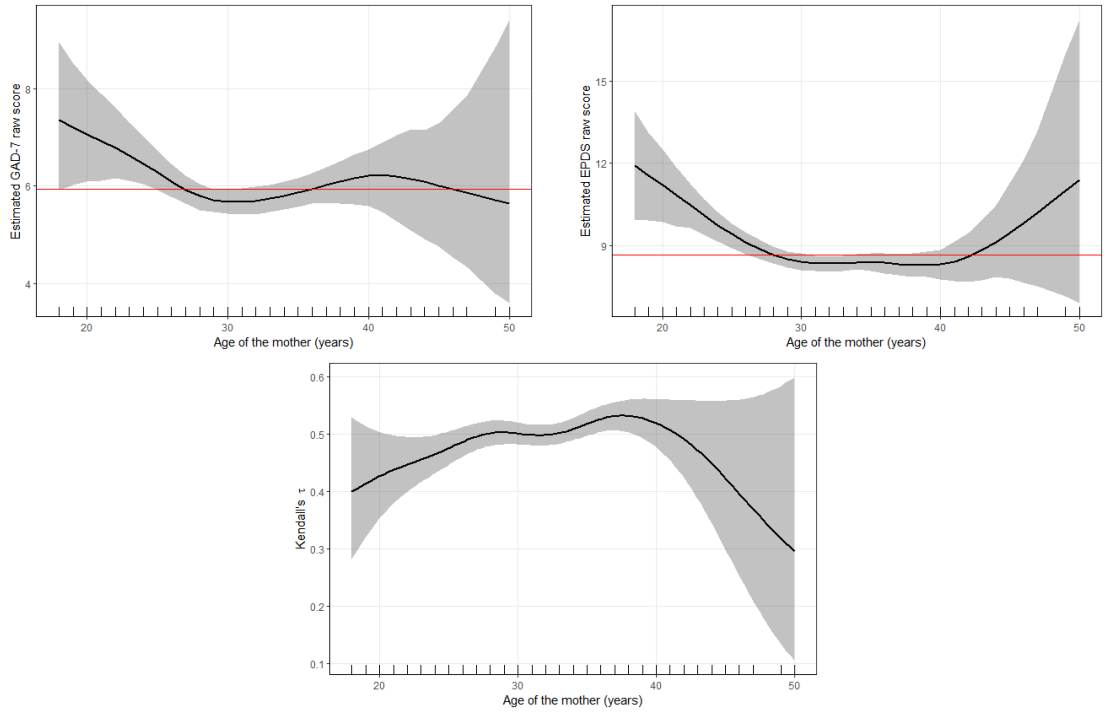

Figure 3.8: Estimated effect of the Age of the mother on GAD-7 (upper left panel), EPDS (upper right panel) raw scores, and their correlation (bottom panel) for pregnant women. Red lines are the mean values of the respective raw scores.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Table 3.8: Numerical results for the marginal models (using two negative binomial - type I distributions) for postpartum women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect. Ratio (95%CI) is the exponential of the β estimates and their 95% Confidence Intervals.

		Estimate (β)	SE	Ratio (95% CI)	edf	P-value
GAD-7 raw score	Intercept	3.733	0.533	41.804 (14.707, 118.826)	1	<0.001
	fp (ref: Multiparous)	-0.796	0.459	0.451 (0.183, 1.109)	1	0.083
	ce (ref: No)	0.158	0.033	1.171 (1.098, 1.249)	1	<0.001
	tdh	0.055	0.012	1.057 (1.032, 1.082)	1	<0.001
	Age	-0.01	0.003	0.990 (0.984, 0.996)	1	0.001
	CHI for Multiparous	-0.013	0.004	0.987 (0.979, 0.995)	1	<0.001
	CHI for Primiparous	-0.007	0.004	0.993 (0.985, 1.001)	1	0.063
	IHDI for Multiparous	-1.259	0.555	0.284 (0.096, 0.843)	1	0.023
	IHDI for Primiparous	-0.693	0.543	0.500 (0.173, 1.450)	1	0.202
	Country (random effects)	See Figure 3.9			7.180	<0.001
EPDS raw score	Intercept	3.614	0.481	37.114 (14.458, 95.273)	1	<0.001
	fp (ref: Multiparous)	-0.748	0.358	0.473 (0.235, 0.955)	1	0.036
	ce (ref: No)	0.087	0.026	1.091 (1.037, 1.148)	1	0.001
	tdh	0.042	0.011	1.043 (1.021, 1.066)	1	<0.001
	Age	-0.007	0.002	0.993 (0.989, 0.997)	1	0.001
	CHI for Multiparous	-0.009	0.003	0.991 (0.985, 0.997)	1	0.002
	CHI for Primiparous	-0.004	0.003	0.996 (0.990 , 1.002)	1	0.172
	IHDI for Multiparous	-0.969	0.537	0.379 (0.132, 1.087)	1	0.071
	IHDI for Primiparous	-0.379	0.529	0.685 (0.243, 1.931)	1	0.473
	Country (random effects)	See Figure 3.9			8.096	<0.001

Table 3.9: Numerical results for the correlation model between GAD-7 and EPDS for postpartum women. SE means standard error. The edf number represents the parameters needed to estimate the effect of a covariate, indicating the complexity of the functional form of the smooth effect.

	Estimate (β)	SE	edf	P-value
Intercept	2.263	2.125	1	0.287
fp (ref: Multiparous)	-0.727	0.895	1	0.417
ce (ref: No)	0.172	0.064	1	0.007
tdh	-0.021	0.041	1	0.615
Age	0.000	0.006	1	0.934
CHI for Multiparous	-0.016	0.007	1	0.027
CHI for Primiparous	-0.009	0.008	1	0.250
IHDI for Multiparous	-1.511	2.659	1	0.570
IHDI for Primiparous	-1.047	2.656	1	0.693
Country (random effects)	See Figure 3.9		9.150	<0.001

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

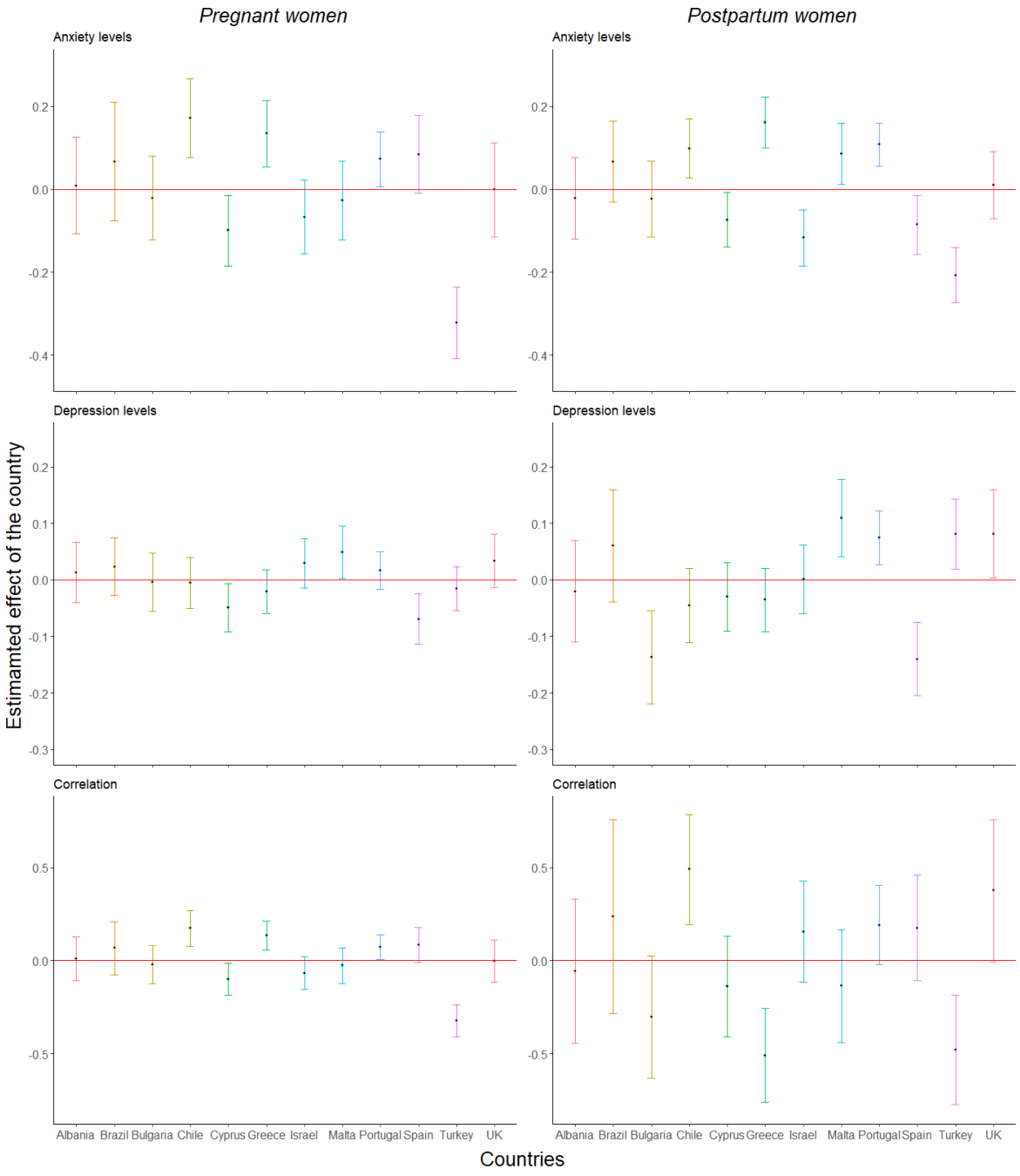

Figure 3.9: Estimated random effects of the country of origin on GAD-7 (top) and EPDS (middle) raw scores, and on the correlation between the outcomes (bottom), for pregnant women (left) and postpartum women (right). Red line is the 0 and indicates no effect.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3.4.3 Discussion

In our study, a major focus of attention were governmental responses to control the spread of COVID-19, which differed between countries. Increased CHI were associated with decreased symptoms of anxiety and depression, as well as a lower association between them in postpartum women. In other words, the number and intensity of closure and containment policies and those aimed at disease surveillance have a positive effect on women's perinatal mental health. These results contradict the view of several authors that such restrictive measures, such as social isolation, can negatively affect mental health (Leigh-Hunt et al., 2017; Brooks et al., 2020; Rogers et al., 2020). One explanation for this could be that the COVID-19 pandemic represents a lot of uncertainty at different levels, not only at the health level, and that these governmental interventions may be perceived as protective, providing individuals with a sense of security and control (Wu et al., 2020). In fact, lockdown measures during the pandemic were associated with some positive outcomes during pregnancy, for instance, low birth-weight births (Philip et al., 2020) or a decrease in preterm (De Curtis et al., 2021), which may be due to pregnant women resting more at home and and having better nutritional intake (Ranjbar et al., 2021).

Women living in countries with high IHDI displayed lower symptoms of anxiety and depression, underlining that exposure to multiple risk factors (a phenomenon characteristic of countries with higher inequalities) is associated with increased risk of mental health problems. This finding contrasts with the higher depression symptoms in countries with higher HDI reported in a systematic review by Lee et al. (2021), although it should be borne in mind that in those studies, the HDI was used, whereas in our study the IHDI was used.

Symptoms of anxiety and depression increased with the dimension of the pandemic's negative effects, such as the number of deaths and exposure to COVID-19. For every 100 deaths per million, the depressive symptoms increased between 2.8 and 4.3% in both pregnant and postpartum women, and the anxiety symptoms increased 9.4% in postpartum women. Correlation between anxiety and depression symptoms in postpartum women was also higher as this number increases . These data supports the assumption that COVID-19 may lead to national mental health crises, particularly in countries with greater infection levels (Dong and Bouey, 2020).

Most of the participants had high levels of education in our study. This can be limitation, since it may lead to underestimating mental health problems, because a higher educational level is a protective factor for mental health problems (Nisar et al., 2020). As a result, the generalisability of the results may be limited. Similarly, women under 18 years old, who may be especially vulnerable to the adverse effects of the COVID-19 pandemic, were not included in the study due to ethical issues.

It is also important to note that this study reports data collected between June

USC DE SANTIAGO DE COMPOSTELA

and October 2020, between the first and the second wave of COVID-19 in Europe, and the long-term impact of prolonged exposure to such restrictive governmental measures as seen in this period remains to be investigated.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

An Alternative Approach to Copula JM: Multivariate Conditional Transformation Models (MCTM)

The contents of this chapter were published in Díaz-Louzao et al. (2021).

4.1 Introduction

Thyroid dysfunction is a common disease. For instance, the estimated prevalence of hyperthyroidism ranges from 0.2% to 1.3% in parts of the world with enough iodine (Taylor et al., 2018), being larger in those places with a lack of iodine.

Primary thyroid failure brings with it a number of different symptoms, raging from fatigue, sensitivity to cold or poor memory and concentration, to dyspnoea, weight gain or even depression. As this gland is in charge of the regularization of the metabolism, its malfunction is harmful to the whole body.

It seems obvious that the simultaneous study of the thyroid-related hormones may improve the diagnosis of primary thyroid failure. To this aim, the characterisation of the distribution function of thyroid hormones in healthy subjects is a good tool, as it allows the detection of individuals with abnormal values. In order to achieve this objective, Multivariate Conditional Transformation Models (MCTM; Klein et al., 2019) were considered, applied to TRH database, presented in Chapter 2.

MCTM were recently proposed as a new multivariate regression technique. They characterise jointly the covariates effects on the marginal distributions of the responses and their correlations without requiring parametric assumptions. Flexibility, in both the responses and covariates effects are achieved using Bernstein basis

USC DE SANTIAGO DE COMPOSTELA

polynomials (Bernstein, 1912). Unfortunately, the application of these models is limited to the inclusion of a single covariate, due to the current state of development of the software required for their implementation.

In this chapter, we compare MCTM with the well established CGAMLSS, introduced in Chapter 3, for the bivariate response case. In addition, we will go a step beyond since, in this application to real data, the response is trivariate.

4.2 Multivariate Conditional Transformation Models

Multivariate Conditional Transformation Models (MCTM; Klein et al., 2019) were developed in the context of multivariate quantile regression analysis as a non parametric way of characterising the joint distribution function. These models are the natural extension of univariate conditional transformation models (CTM; Hothorn et al., 2018) to the multivariate case. As a result, this framework makes six substantial contributions, which cannot so far be found simultaneously in any other multivariate regression methodology:

- MCTM allows for direct estimation and inference of the multivariate conditional cumulative distribution function (CDF).
- Model estimation is done for all model components at once, in a single step.
- Theoretical results on optimality properties, such as asymptotic normality, are available.
- Unlike CGAMLSS, strong parametric assumptions are not required.
- MCTM allow the association between responses to be non-linear depending on covariates.
- The methodology adapts very well to situations with more than two response variables.

As in the case of copulas, joint distributions are constructed by their decomposition into marginals and the dependence structure.

4.2.1 Model Formulation

Multivariate Transformation Models were developed for a J-dimensional, absolutely continuous response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_J)^\top \in \mathbb{R}^J$. In these models, an unknown, bijective, strictly monotonically increasing transformation function h :

USC DE SANTIAGO DE COMPOSTELA

$\mathbb{R}^J \to \mathbb{R}^J$ maps the outcome $\boldsymbol{Y}$ to a set of J independent and identically distributed, absolutely continuous random variables $Z_j \sim F_Z, j \in \{1, \ldots, J\}$, where F_Z is a predefined distribution (usually, $Z_j \sim \mathrm{N}(0, 1)$). That is:

$$h(\boldsymbol{Y}) = (h_1(\boldsymbol{Y}), \ldots, h_J(\boldsymbol{Y}))^\top = (Z_1, \ldots, Z_J)^\top = \boldsymbol{Z} \in \mathbb{R}^J. \tag{4.1}$$

MCTM are obtained by extending the transformation function (4.1) to include a vector of potential covariates $\boldsymbol{X}$, resulting in:

$$h(\boldsymbol{y}|\boldsymbol{x}) = (h_1(\boldsymbol{y}|\boldsymbol{x}), \ldots, h_J(\boldsymbol{y}|\boldsymbol{x}))^\top. \tag{4.2}$$

A triangular structure is imposed on the transformation function h in order to simplify calculations, without imposing any limitation on the estimation of the responses. This means that the j^{th} component of the transformation function depends only on the first j elements of $\boldsymbol{y}$. Therefore, for each $j \in \{1, \ldots, J\}$, the components of (4.2) can be written as:

$$h_j(\boldsymbol{y}|\boldsymbol{x}) = h_j((y_1, \ldots, y_J)|\boldsymbol{x}) = h_j((y_1, \ldots, y_j)|\boldsymbol{x}). \tag{4.3}$$

Finally, the h_j transformation functions in (4.3) are set to be linear combinations of marginal transformation functions $\tilde{h}_j : \mathbb{R} \to \mathbb{R}$:

$$h_j(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\jmath=1}^{j-1} \lambda_{j\jmath}(\boldsymbol{x}) \tilde{h}_\jmath(y_\jmath|\boldsymbol{x}) + \tilde{h}_j(y_j|\boldsymbol{x}),$$

where $\lambda_{j\jmath}(\boldsymbol{x})$ and $\tilde{h}_j(y_j|\boldsymbol{x})$ are expressed in terms of basis function expansions. Thus:

$$\tilde{h}_j(y_j|\boldsymbol{x}) = \boldsymbol{a}_j(y_j)^\top \boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_{j,2}, \tag{4.4}$$

and:

$$\lambda_{j\jmath}(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\gamma}_{j\jmath}, \quad 1 \leq \jmath < j \leq J. \tag{4.5}$$

Here, $\boldsymbol{a}_j$ and $\boldsymbol{b}$ are polynomial Bernstein basis functions for the response and the covariates respectively, and $\boldsymbol{\beta}_{j\jmath} = \left(\boldsymbol{\vartheta}_{j,1}^\top, \boldsymbol{\vartheta}_{j,2}^\top, \boldsymbol{\gamma}_{j\jmath}^\top\right)^\top$ are the parametric coefficients of the model. In Figure 4.1, a representation of the Bernstein basis of four different orders is shown.

Summing up the specifications of the model, MCTM are characterised by a set of marginal conditional transformations $\tilde{h}_j(y_j|\boldsymbol{x})$, $j \in \{1, \ldots, J\}$ (computed as in (4.4)), each of which applying to only one component of $\boldsymbol{Y}$ and the covariates $\boldsymbol{X}$, and by

USC DE SANTIAGO DE COMPOSTELA

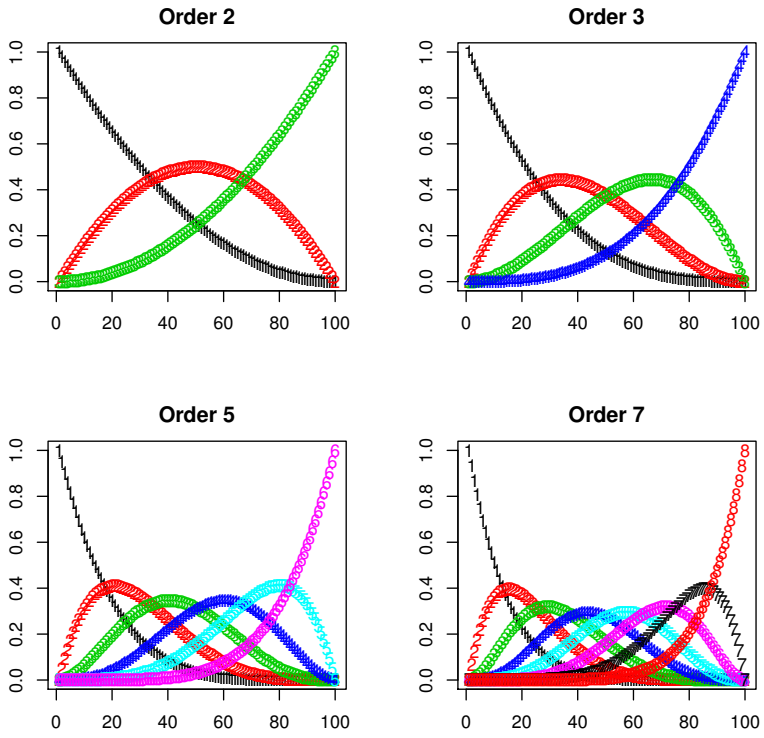

Figure 4.1: Representation of Bernstein basis of order 2, 3, 5, and 7.

a lower triangular $(J \times J)$ matrix of transformation coefficients $\boldsymbol{\Lambda}(\boldsymbol{x})$:

$$\boldsymbol{\Lambda}(\boldsymbol{x}) = \begin{pmatrix} 1 & & & & 0 \\ \lambda_{21}(\boldsymbol{x}) & 1 & & & \\ \lambda_{31}(\boldsymbol{x}) & \lambda_{32}(\boldsymbol{x}) & 1 & & \\ \vdots & \vdots & & \ddots & \\ \lambda_{J1}(\boldsymbol{x}) & \lambda_{J2}(\boldsymbol{x}) & \dots & \lambda_{J,J-1}(\boldsymbol{x}) & 1 \end{pmatrix},$$

with $\lambda_{jj}(\boldsymbol{x})$, $1 \leq \jmath < j \leq J$, defined as in (4.5). Under the standard normal reference distribution $F_Z = \mathrm{N}(0,1)$, the coefficients of $\boldsymbol{\Lambda}(\boldsymbol{x})$ characterise the dependence structure of the responses. This is so because $\tilde{\boldsymbol{Z}}(\boldsymbol{x}) = (\tilde{Z}_1(\boldsymbol{x}), \dots, \tilde{Z}_J(\boldsymbol{x}))^\top$, defined by the random variables $\tilde{Z}_j(\boldsymbol{x}) = \tilde{h}_j(Y_j|\boldsymbol{x})$, follows a multivariate Gaussian distribution $\tilde{\boldsymbol{Z}}(\boldsymbol{x}) \sim \mathrm{N}(\mathbf{0}_J, \boldsymbol{\Sigma}(\boldsymbol{x}))$, with $\boldsymbol{\Sigma}(\boldsymbol{x}) = \boldsymbol{\Lambda}(\boldsymbol{x})^{-1}(\boldsymbol{\Lambda}(\boldsymbol{x})^{-1})^\top$.

In the MCTM framework, association between the response variables is usually presented in terms of Spearman's rank correlation coefficient ρ^S.

USC DE SANTIAGO DE COMPOSTELA

4.2.2 Estimation and Inference

In MCTM, the estimation of the parameters is based on the model log-likelihood. Let $\boldsymbol{\vartheta}_1 = \left(\boldsymbol{\vartheta}_{1,1}^\top, \ldots, \boldsymbol{\vartheta}_{J,1}^\top\right)^\top$ and $\boldsymbol{\vartheta}_2 = \left(\boldsymbol{\vartheta}_{1,2}^\top, \ldots, \boldsymbol{\vartheta}_{J,2}^\top\right)^\top$be the parameters for all the marginal transformations functions $\tilde{h}_j(\boldsymbol{y}|\boldsymbol{x})$, $j \in \{1, \ldots, J\}$, let $\boldsymbol{\gamma}$ be the vector containing the parameters for all the unknown elements of $\boldsymbol{\Lambda}$, and let $\boldsymbol{\beta} = \left(\boldsymbol{\vartheta}_1^\top, \boldsymbol{\vartheta}_2^\top, \boldsymbol{\gamma}^\top\right)^\top$ the set of all the parameters involved in the model. The maximum likelihood estimator can be written as:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Bigg(-\frac{1}{2}\sum_{j=1}^{J}\Bigg(\sum_{\jmath=1}^{j-1}\left(\boldsymbol{b}(\boldsymbol{x}_i)^\top\boldsymbol{\gamma}_{j\jmath}\right)\left(\boldsymbol{a}_\jmath(y_{ij})^\top\boldsymbol{\vartheta}_{\jmath,1} - \boldsymbol{b}_\jmath(\boldsymbol{x}_i)^\top\boldsymbol{\vartheta}_{\jmath,2}\right) + \boldsymbol{a}_j(y_{ij})^\top\boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}_j(\boldsymbol{x}_i)^\top\boldsymbol{\vartheta}_{j,2}\Bigg)^2 + \log\left(\boldsymbol{a}_j'(y_{ij})^\top\boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}_j'(\boldsymbol{x}_i)^\top\boldsymbol{\vartheta}_{j,2}\right)\Bigg) \tag{4.6}$$

Maximising (4.6), we find that the expression for the estimated conditional cumulative distribution function (CDF) for each outcome, $\mathrm{F}_{\boldsymbol{Y}}(Y_j|\boldsymbol{x})$, is given by:

$$\mathbb{P}(Y_j \leq y_j|\boldsymbol{X} = \boldsymbol{x}) = \Phi_{0,\hat{\sigma}_j}\left(\tilde{h}_j(y_j|\boldsymbol{x})\right) = \mathrm{F}_{\boldsymbol{Z}}\left(\boldsymbol{a}_j(y_j)^\top\hat{\boldsymbol{\vartheta}}_{j,1} - \boldsymbol{b}_j(\boldsymbol{x})^\top\hat{\boldsymbol{\vartheta}}_{j,2}\right),$$

and the joint CDF can be expressed as:

$$\hat{\mathrm{F}}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x}) = \Phi_{\mathbf{0},\hat{\boldsymbol{\Sigma}}}\left(\boldsymbol{a}_1(y_1)^\top\hat{\boldsymbol{\vartheta}}_{1,1} - \boldsymbol{b}_1(\boldsymbol{x})^\top\hat{\boldsymbol{\vartheta}}_{1,2}, \ldots, \boldsymbol{a}_J(y_J)^\top\hat{\boldsymbol{\vartheta}}_{J,2} - \boldsymbol{b}_J(\boldsymbol{x})^\top\hat{\boldsymbol{\vartheta}}_{J,2}\right),$$

where $\Phi_{0,\hat{\sigma}_j^2}$ is the univariate Gaussian distribution function with mean = 0 and standard deviation = $\hat{\sigma}_j$, and $\Phi_{\mathbf{0},\hat{\boldsymbol{\Sigma}}}$ is the multivariate Gaussian distribution function with mean = $\mathbf{0}$ and $\hat{\boldsymbol{\Sigma}}$ as variance-covariance matrix.

Inference on the responses correlations is done by using parametric bootstrap. See Klein et al. (2019) for details on the algorithm.

4.3 Simulation Study for the Bivariate Case

In Section 5 of their paper, Klein et al. (2019) evaluated the performance of MCTM with a simulation of a bivariate response depending on one covariate, comparing the results of MCTM with two other approaches: Bayesian structured additive distributional regression models (Klein et al., 2015b), and vector generalised additive

USC DE SANTIAGO DE COMPOSTELA

models (VGAM; Yee, 2015). In this Section, we are going to replicate that simulation study, considering now as a competitor the conditional CGAMLSS.

As we have seen in Chapter 3, CGAMLSS are bivariate copula models with parametric marginal distributions, one-parameter copulas, and semi-parametric specification for the predictors of all parameters of both the marginal and the copula jointly, using a penalised likelihood framework. Unlike MCTM, this approach is only limited to the bivariate case for continuous responses. Even so, we do not see that this is a good enough reason not to compare them when we have two possibly related responses.

4.3.1 Simulation Design

In the simulation set-up, $R = 100$ data sets of size $n = 1000$ were simulated and one single covariate $x \sim \mathrm{U}[-0.9, 0.9]$ was considered. The two continuous outcomes were considered following a Dagum distribution. The bivariate response dependence structure depends on x, being $\lambda(x) = x^2$ (to see all of the details, go to Klein et al. (2019), Subsection 5.1). With this construction, the first margin is independent of the covariate x, but the scale parameter of the second margin is not (not so the shape parameters a and p, which remain constant).

For the MCTM, we applied Bernstein polynomials of order 8 and 3 in the responses and the covariate respectively.

For the CGAMLSS, we used the true specification of the model, i.e. a Gaussian Copula and Dagum marginals, with the correlation parameter and the b_2 parameter of the Dagum distribution for y_2 dependent on x, while all of the parameters of the first marginal, as well as the shape parameters a_2 and p_2 of the second one did not depend on any covariate. For this purpose, both the predictor for b_2 and the correlation parameter ρ^S of the copula were specified using cubic B-splines with 20 inner knots, as it was done for the other approaches compared in the original paper. Likewise, the other parameters of the model were estimated as constants.

4.3.2 Evaluation of the Performance

The performance of the methods was evaluated by estimating the lower element of $\mathbf{\Lambda}^{(r)}$, $\hat{\lambda}^{(r)}(x)$, from every data replicate $r \in \{1, \ldots, R\}$, on a grid of length $G = 100$ within the range of x, and comparing it to the real one, $\lambda(x)$. On the one hand, we considered the estimated curves as they were; on the other hand, we computed the root mean squared error $\mathrm{RMSE}\left(\lambda, \hat{\lambda}^{(r)}\right) = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\lambda(x_g) - \hat{\lambda}(x_g)\right)^2}$ and the standard deviation $\mathrm{SD}\left(\hat{\lambda}^{(r)}\right)$ of each sample as a numeric measure of discordance.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4.3.3 Results of the Simulation Study

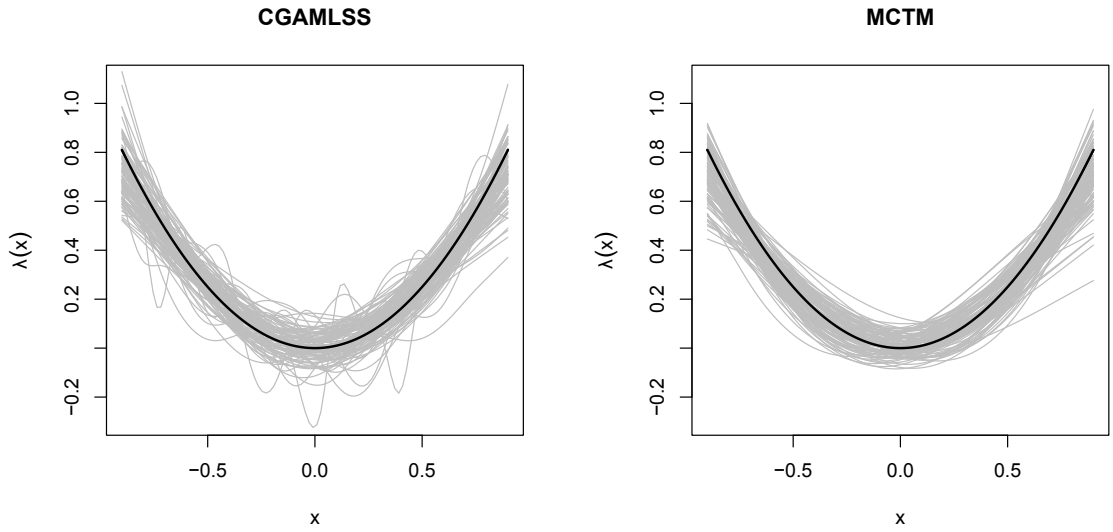

Figure 4.2: Function estimates $\hat{\lambda}(x)$ for the effect $\lambda(x) = x^2$ on the correlation parameter for CGAMLSS (left) and MCTM (right). The black line represents the true function and the grey lines are the estimations of the $R = 100$ simulation samples.

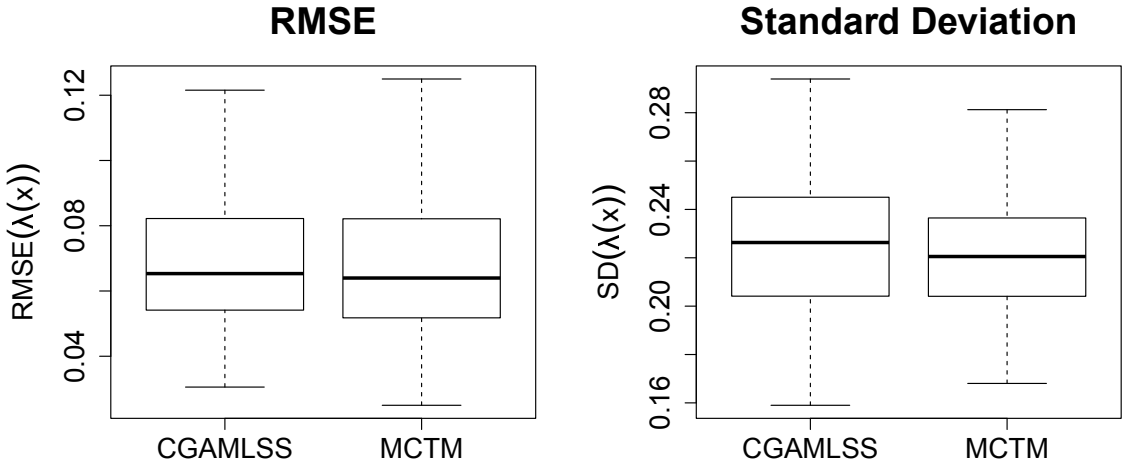

Figure 4.3: Left: $\mathrm{RMSE}\left(\lambda, \hat{\lambda}^{(r)}\right)$ for CGAMLSS and MCTM. Right: $\mathrm{SD}\left(\lambda, \hat{\lambda}^{(r)}\right)$ for CGAMLSS and MCTM.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Figure 4.2 shows the estimations for $\lambda(x) = x^2$ of the 100 simulated data sets for CGAMLSS (left panel) and MCTM (right panel). Even though both models seemed to capture the general form properly, MCTM led to better smoothed curves. Although CGAMLSS was fitted using the actual model specification, in terms of RMSE and SD both approaches were very similar (Figure 4.3). This is highly remarkable because MCTM was fitted without specifying parametric marginal distributions, a requirement that is restrictive, and an obstacle in practice. Besides this, we have found a big difference in the execution times, since, on average, the adjustment of MCTM was less than a second (0.480 seconds), while a CGAMLSS model took almost 5 seconds (4.807 seconds), run on a single core of a 3.60 GHz Intel (R) Core (TM) i7-7700. Therefore, a MCTM approach seemed to be more advisable than a CGAMLSS approach.

4.4 Trivariate Joint Modelling of TRH Data

4.4.1 Motivation of the Study

The thyroid gland is a small gland located in front of the neck, that performs a vitally important function: the control of the metabolism. This is fulfilled by taking the iodine found in many foods and converting it into two thyroid hormones: triiodothyronine (T3), and thyroxine (T4). The normal thyroid gland produces about 20% T3, and about 80% T4 and every cell in the body depends upon thyroid hormones for regulation of their metabolism. Only a small fraction of the thyroid hormones is free to enter tissues and has a biologic effect.

The thyroid gland is under the control of the pituitary gland. This small gland located at the base of the brain, produces Thyroid Stimulating Hormone (TSH) when the T3 and T4 levels drop too low, in order to stimulate the thyroid to produce more hormones (see Figure 4.4).

But this mechanism does not always work well, and when that happens, it can lead to profound adverse effects if undiagnosed or untreated. This gland is in charge of the regularization of the metabolism, so the whole body would be affected in the case of malfunction.

The relationship between TSH and free T4 was broadly studied since it seems that this tandem has the greatest diagnostic value for primary thyroid failure (Kumar et al., 1977). Many studies suggested an inverse, linear relationship between log-TSH and T4, being this taken as a generally accepted principle of thyroid physiology thanks to Spencer et al. (1990). It was not until 2010 that studies began to emerge refuting the linearity of the relationship (Hoermann et al., 2010; Hadlow et al., 2013).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

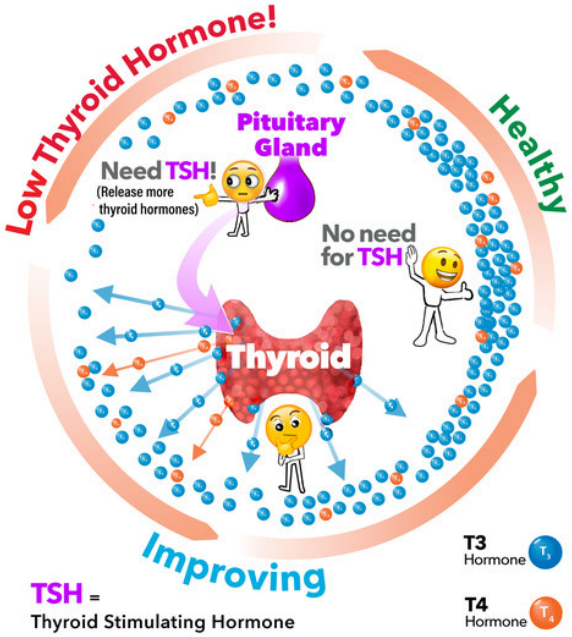

Figure 4.4: Diagram of the functioning of thyroid hormones (Source: UCLA Endocrine Center (UCLA Health) (nd)).

The diagnosis of primary thyroid failure may be improved by studying simultaneously the three thyroid-related hormones. In addition, it is also possible that the relationship between these hormones varies with age and should be included as a covariate. To this aim, MCTM were applied to study the relationship between log-TSH, free T3, and free T4, adjusting for the age of the subjects.

4.4.2 Statistical Analysis of TRH Data

Based on a $F_Z = \mathrm{N}(0,1)$ distribution, our marginal conditional distribution is parametrised as:

$$\tilde{h}_j(y_j|\text{Age}) = \boldsymbol{a}_j(y_j)^\top \boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}(\text{Age})\boldsymbol{\vartheta}_{j,2},\ j \in \{\text{log-TSH, Free T3, Free T4}\},$$

where basis functions $\boldsymbol{a}$ are Bernstein polynomials of order eight and $\boldsymbol{b}$ is a Bernstein polynomial of order three. The coefficients of $\boldsymbol{\Lambda}$ were parametrised as:

$$\lambda_{jj}(\text{Age}) = \boldsymbol{b}(\text{Age})\boldsymbol{\gamma}_{jj}, \quad \jmath < j \in \{\text{log-TSH, Free T3, Free T4}\},$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Results for Marginal Distributions

The estimated marginal conditional CDFs $F_j(y_j|\text{age})$, and densities $f_j(y_j|\text{age})$, are shown in Figure 4.5. The different colours represent the ages of the individuals. We can see that the shapes differ for the three hormones, and that hormone levels decrease as age increases for log-TSH and Free T3, specially for the latter. For Free T4, we can see that higher values are obtained in middle-aged individuals. Another way of seeing this is as depicted in Figure 4.6, with the scatter plot of the hormones as a function of age, and the percentile lines shown in red. Here it can be clearly seen that, although log-TSH levels appear to decrease with age, this change is very small, so the log-TSH levels can be considered invariant with respect to age.

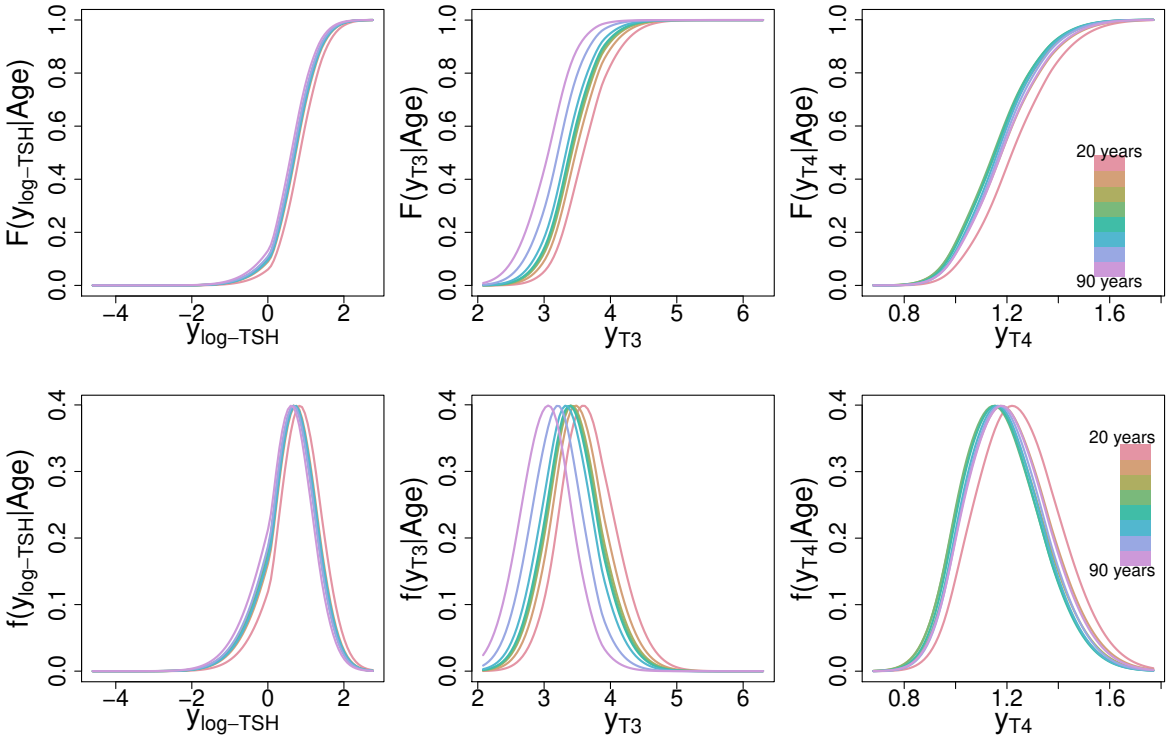

Figure 4.5: Estimated marginal conditional CDFs $F_j(y_j|\text{age})$, and conditional density functions $f_j(y_j|\text{age})$, for j = log-TSH (left), j = Free T3 (centre), and j = Free T4 (right), from age = 20 (brown) to age = 90 (purple).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

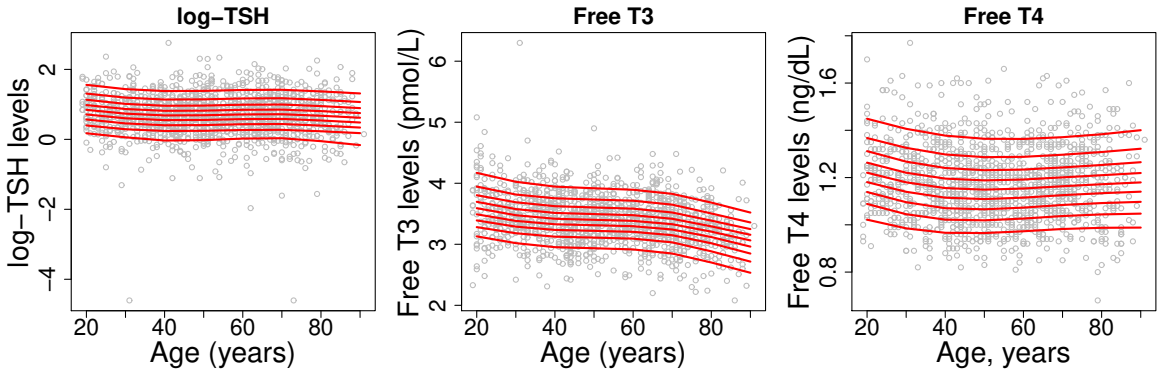

Figure 4.6: Scatter plot of thyroid-related hormones vs. age. In red, percentile curves are shown. Left: log-TSH levels. Centre: Free T3 levels. Right: Free T4 levels.

Results for Dependence Structure

In Figure 4.7 it is shown that the two-by-two conditional Spearman's $\rho^S(\text{Age})$ for the three hormones, is statistically significant in all three cases. The grey area represents the confidence interval estimates at 95% from 1000 parametrically drawn bootstrap as explained in Subsection 3.4 in Klein et al. (2019). We can see that the relationship between log-TSH and Free T3 is negative, and statistically significant for individuals between 30 and 50 years old. Correlation between log-TSH and Free T4 is negative (as expected). Finally, correlation between Free T3 and Free T4 is positive and decreases with age.

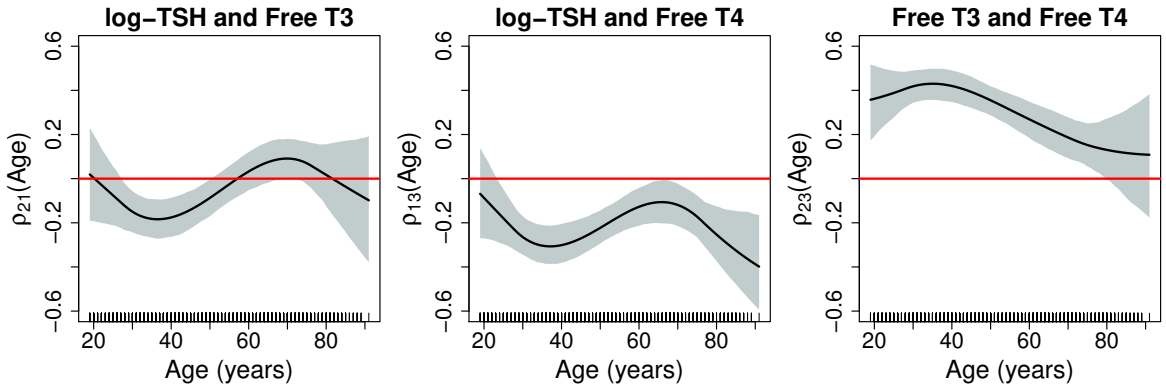

Figure 4.7: Two-by-two conditional Spearman's $\rho^S(\text{Age})$ for the thyroid-related hormones (left: log-TSH and Free T3; centre: log-TSH and Free T4; right: Free T3 and Free T4). Grey area represents the 95% bootstrap confidence interval, while black line represents the bootstrapped point estimate. Red line indicates the 0.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

Joint Modelling for Longitudinal and Survival Data (JMLS)

In recent years there has been great progress in the biomedical field thanks to advances in diagnostic techniques, as well as the evolution of omics (proteomics, genomics, metabolomics) that have facilitated the discovery of new biomarkers. This improvement requires the development of new statistical tools with great flexibility. On the other hand, these measurements of clinical parameters and markers are usually carried out repeatedly in the same individual, which is common when studying the evolution of a disease. These measurements, known as longitudinal data, require a specific statistical methodology.

In the last years there has been a great advance in the development of longitudinal data analysis (Verbeke and Molenberghs, 2000; Diggle et al., 2002; Pinheiro and Bates, 2006; Demidenko, 2013). However, many biomedical researchers still do not apply it on a regular basis, using other statistical techniques that are not appropriate, although its use is spreading.

Frequently, in studies with long-term data, multiple response variables of different nature (continuous, categorical, discrete) are evaluated over time for each subject. For example, the appearance of exacerbations in people with chronic obstructive pulmonary disease, associated with the evolution of the forced expiratory volume.

On the other hand, it may also happen that a response is linked to censored results for each subject and the objective is to study these results together with longitudinal data. A very clear example is described in clinical trials conducted in patients with HIV, whose covariates include the assigned treatment, clinical characteristics, socio-demographic information and a series of repeated measurements related to viral load and CD4 count, which are performed at each visit. The survival time of each patient is also recorded, being one of the usual objectives the evaluation of the

USC
DE SANTIAGO
DE COMPOSTELA

viral load in relation to the treatment and the progression to death.

There are various possible approaches to joint modelling longitudinal and survival processes, including the use of copula models (Zhang et al., 2022), but random effects models have prevailed. These models are based on the idea that unobserved random effects u explain the association between the outcomes, and standard methods make use of two sub-models (for the longitudinal and survival processes), in order to specify the full model.

In recent years, JM for longitudinal and survival data (JMLS) has acquired great importance in the field of biostatistics, where some nice reviews can be found (Tsiatis and Davidian, 2004; Sousa, 2011; Alsefri et al., 2020). But it has also become increasingly important in biomedicine, with a wide number of clinical applications (Touloumi et al., 2004; Deslandes and Chevret, 2007; Li et al., 2009; Luo and Wang, 2014; Zhang et al., 2017; Kiplimo et al., 2021).

The aim of this chapter is to introduce the reader to the latest developments in JMLS, showing their potential in clinical practice.

5.1 Longitudinal Analysis: Linear Mixed Model (LMM)

Let Y be the longitudinal biomarker, collected at a set of t_{ij} time points, and let $\boldsymbol{X} = (\boldsymbol{x}_1|\ldots|\boldsymbol{x}_p)$ a set of p baseline covariates, for each subject $i \in \{1,\ldots,n\}$ (n is the number of subjects), and each observation within each subject $j \in \{1,\ldots,n_i\}$ (n_i the number of observations per individual, and $N = \sum_{i=1}^{n} n_i$ the total number of observations). We are assuming that this biomarker is collected with some error, so we need to estimate its "true" value to be used in the survival analysis. For this purpose, a linear mixed-effects model is typically used (5.1).

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + \beta_{p+1} t_{ij} + u_{0i} + u_{1i} t_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathrm{N}(0, \sigma), \tag{5.1}$$

where the βs are the "fixed effects" (i.e., the coefficients for the global covariates of the model) and u_{0i} and u_{1i} are the random intercept and random slope effects respectively, i.e., they represent the variation of each subject from the global intercept the global time effect. It is assumed for the random effects that:

$$\boldsymbol{u}_i = \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} \right),$$

with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{u_0} & \sigma_{u_0}\sigma_{u_1}\rho_{01} \\ \sigma_{u_0}\sigma_{u_1}\rho_{01} & \sigma^2_{u_1} \end{pmatrix},$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

where σ_{u_0} and σ_{u_1} are the standard deviations of the random effects and ρ_{01} indicates the correlation between them.

In a general form, (5.1) can be written as:

$$\begin{cases} y_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \\ \boldsymbol{u}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}),\ \varepsilon_i(t) \sim \mathrm{N}\,(0, \sigma) \end{cases}$$

where $\boldsymbol{x}_i^\top(t)$ are the design vectors for the fixed effects $\boldsymbol{\beta}$, $\boldsymbol{z}_i^\top$ are the design vectors for the random effects $\boldsymbol{u}_i$ and $\varepsilon_i(t)$ are the error terms.

5.2 Survival Analysis: Cox Proportional Hazards Model

The most striking feature of survival processes is the presence of censoring. As a general rule, in a survival study it is assumed that all individuals suffer at some point the event under study (e.g. death or development of a certain disease), even if it is not observed for some subjects in the sample. The most common case of censoring, and the one we are going to focus on, is right censoring. In this case the follow-up and observation time of the subject is less than the "true" time to event. In this scenario, let T_i^* be the "true" time-to-event for individual $i \in \{1, \ldots, n\}$, and be C_i their censoring time (for instance, the end of the study). So, the information available for each subject is:

- The observed time-to-event: $T_i = \min(T_i^*, C_i)$.
- The event indicator:

$$\delta_i = \begin{cases} 1, & \text{in case the event is observed} \\ 0, & \text{in case of censoring} \end{cases}$$

The main objective of survival analysis is to make valid inference for T_i^* using $\{T_i, \delta_i\}$. In this context, there are two basic functions to be estimated. Let $f(t)$ the density function of the "true" time-to-event T^*, with corresponding distribution function $F(t) = \mathbb{P}(T^* \leq t)$. Then, the survival function of T^* is the probability of being alive up to time t:

$$S(t) = \mathbb{P}(T^* > t) = \int_t^\infty f(x)dx = 1 - F(t). \tag{5.2}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The hazard function is the instantaneous risk of an event at time t, given that the event has not occurred until t:

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T^* < t + dt | T^* \geq t)}{dt} = \frac{f(t)}{S(t)}, \; t > 0. \tag{5.3}$$

The hazard function can be interpreted as the expected number of events per individual per unit of time. From (5.2) and (5.3), we can establish the following relationship:

$$S(t) = \exp\left\{ - \int_0^t h(s) ds \right\}$$

In survival analysis, the most extended approach is through modelling the hazard function (5.3) using a Cox proportional hazards model (Cox, 1972):

$$h_i(t) = h_0(t) \exp\left\{ \gamma_1 w_{i1} + \ldots \gamma_Q w_{iQ} \right\},$$

where $h_0(t)$ denotes the baseline hazard, representing the hazard for an event when all the covariates w_q or all the coefficients γ_q, $q \in \{1, \ldots, Q\}$, are 0. In this model, if a covariate has a beneficial effect, $h(t)$ increases and vice versa. Cox proportional hazards models does not so much seek to estimate the function $h_0(t)$, which is identical for all subjects, as the relationship between the risks of death between two individuals exposed to different factors. To do this, the model starts from a fundamental assumption: that the risks are proportional or, in other words, that the effect of the covariates does not depend on time.

There are extensions of this model to the case of time-dependent covariates (Cox and Oakes, 1984), and also allowing for a non-linear effect of the covariates (Hastie and Tibshirani, 1990a; Fahrmeir et al., 2011), among others.

5.3 Joint Modelling for Longitudinal and Survival Data

5.3.1 The Basic JMLS

In this section we are focusing on the case of just one continuous longitudinal outcome. The intuitive idea behind these models is:

(i) Describe the evolution of the biomarker over time for each individual using an appropriate model. Let $y_i(t)$ be the observed longitudinal response. Then, taking model (5.1) as a reference, we have the following longitudinal sub-model (Rizopoulos, 2012):

$$y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \; \varepsilon_i(t) \sim \mathrm{N}(0, \sigma), \tag{5.4}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

where $\boldsymbol{x}_i(t)$ and $\boldsymbol{\beta}$ represent the fixed part, and $\boldsymbol{z}_i(t)$ and $\boldsymbol{u}_i$ indicates the random effects part, with $\boldsymbol{u}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

(ii) Use the estimates $m_i(t)$ obtained with the above model (5.4) in a Cox model to predict survival:

$$h_i(t|\mathcal{M}_i(t)) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha m_i(t)\right\},$$

where $\mathcal{M}_i(t) = \{m_i(s)|0 \leq s < t\}$ indicates the longitudinal history, $\boldsymbol{\gamma}$ are the effect of the baseline covariates $\boldsymbol{w}_i$, and α quantifies the strength of the association between the longitudinal biomarker and the risk for an event.

By looking at (5.4), we see that we have considered the true value of the response variable to be included in the survival process. Another popular strategy is to simply consider the random effects, i.e., after fitting the longitudinal sub-model, considering only $m_i(t) = \boldsymbol{z}_i^\top(t)\boldsymbol{u}_i$ (Wulfsohn and Tsiatis, 1997). This strategy has computational advantages, but makes it difficult to interpret α.

The simplest approximations to this problem follow a two-stage approach, considering the two previous processes as independent (Self and Pawitan, 1992). This has the advantage of computational facilities. However, a separate likelihood calculation may lead to biased estimates (Tsiatis and Davidian, 2004; Rizopoulos, 2012). Then, the JMLS itself involves a third step: defining a model for joint distribution as follows (Tsiatis and Davidian, 2004):

$$f(y_i, T_i, \delta_i) = \int f(y_i|\boldsymbol{u}_i)\{h(T_i|\boldsymbol{u}_i)^{\delta_i} S(T_i|\boldsymbol{u}_i)\} f(\boldsymbol{u}_i) du_i,$$

where $\boldsymbol{u}_i$ is the vector of random effects, $f(\cdot)$ indicates the density functions, $h(\cdot)$ denotes the hazard function, and $S(\cdot)$ is the survival function. Rizopoulos (2012) is a very good guide to the world of JMLS.

5.3.2 Model Extensions

So far we have introduced the basic JMLS. But these models can be restrictive, in the sense that only allows for one longitudinal continuous variable and for a single time-to-event. With regard to longitudinal outcomes, it is possible to consider other kind of longitudinal outcomes, such as binary or other non-Gaussian continuous responses, fitting a Generalised Linear Mixed model (GLMM) instead of the LMM. Also, multiple longitudinal markers can be included in the model at the same time, just by assuming a multivariate normal distribution for all the random effects and

USC DE SANTIAGO DE COMPOSTELA

incorporating in the linear predictor of the survival model the expected value of each longitudinal. As for the survival process, also left and interval censoring can be allowed (not only right censoring), as well as competing risks or multi state models. In this section we will introduce the latest extension, presented by Mauff et al. (2020).

Let $\mathcal{D}_n = \{T_i, T_i^U, \delta_i, \boldsymbol{y}_i\}$ denote a sample, with $i \in \{1, \ldots, n\}$ refers to every subject on it. Here, T_i^* indicates de true time-to-event, and T_i and T_i^U indicate the observed event times; $\delta_i \in \{0, 1, 2, 3\}$ is the event indicator (0 for right censoring, 1 for a true event, 2 for left censoring and 3 for interval censoring); and $\boldsymbol{y}_i$ represent the K longitudinal markers, so that $\boldsymbol{y}_{ki}$ denotes the response vector for the k^{th} outcome and the i^{th} individual. The general form of the extended JMLS can be expressed as follows (5.5):

$$\begin{cases} g_k\left(\mathbb{E}\left(y_{ki}(t)|\boldsymbol{u}_{ki}\right)\right) = \eta_{ki}(t) = \boldsymbol{x}_{ki}^\top(t)\boldsymbol{\beta}_k + \boldsymbol{z}_{ki}^\top(t)\boldsymbol{u}_{ki}, \\ h_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i(t)) = h_0(t)\exp\left\{\boldsymbol{\gamma}^\top\boldsymbol{w}_i(t) + \sum_{k=1}^{K}\sum_{l=1}^{L_k} p_{kl}\left(\mathcal{M}_{ki}(t), \boldsymbol{w}_i(t), \boldsymbol{u}_{ki}, \boldsymbol{\alpha}_{kl}\right)\right\}. \end{cases} \tag{5.5}$$

In the longitudinal sub-models of (5.5), $g_k(\cdot)$ is the link function for the k^{th} longitudinal sub-model, $y_{ki}(t)$ is the value of the k^{th} longitudinal outcome for the i^{th} subject at time t, and $\boldsymbol{x}_{ki}$ and $\boldsymbol{z}_{ki}$ indicate the design matrix for the fixed effects $\boldsymbol{\beta}_k$ and the random effects $\boldsymbol{u}_{ki}$. The association between the longitudinal markers is built by assuming that the random effects follow a multivariate Gaussian distribution:

$$\boldsymbol{u}_i = \left(\boldsymbol{u}_{1i}^\top, \ldots, \boldsymbol{u}_{Ki}^\top\right)^\top \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right)$$

In the survival sub-model of (5.5), $\mathcal{M}_{ki}(t) = \{\eta_{ki}(s)|0 \leq s < t\}$ indicates the longitudinal history up to time t for the k^{th} longitudinal process and subject i; $h_0(\cdot)$ denotes the baseline hazard function, which is estimated in a flexible way through a B-splines approach; and $\boldsymbol{\gamma}$ are the regression coefficients for the vector of possibly time-varying covariates $\boldsymbol{w}_i(t)$. Functions $p_{kl}(\cdot)$, with their corresponding vector of parameters $\boldsymbol{\alpha}_{kl}$, allow the different parametrisations of the contribution of each longitudinal sub-model to the survival process to be specified. The most common one is to use $p_{kl}\left(\mathcal{M}_{ki}(t), \boldsymbol{w}_i(t), \boldsymbol{u}_{ki}, \boldsymbol{\alpha}_{kl}\right) = \alpha_{kl}\eta_{ki}(t)$ (and $L_k = 1$), but others can be assumed (see Mauff et al. (2020) for a more complete information).

With this new extension, a lot of flexibility was gained, compared to the basic JMLS. This extension is mathematically straightforward, but extremely computationally expensive. This has led to the adoption of a Bayesian perspective, but this is not enough. The most commonly used practical solution is the use of a two-stage approach, first fitting the multivariate mixed model and, subsequently, the results obtained are used in the survival sub-model. As mentioned above, this can lead to

USC DE SANTIAGO DE COMPOSTELA

biased estimates, so the use of a correction factor based on importance sampling theory is proposed to overcome this obstacle. Please, see Mauff et al. (2020) for further details.

5.3.3 Dynamic Predictions

Accurate predictions of survival probabilities is of great interest to physicians in order to adjust medical care. In this regard, for a specific patient at a specific time t during the follow-up study one would like to use all available information until time t, updating the prediction as additional longitudinal information is obtained. Such individualised predictions can play a highlighted role in personalized medicine, and have therefore attracted a great deal of interest within the joint modelling framework on this front (Proust-Lima and Taylor, 2009; Rizopoulos, 2011).

Based on the fitted joint model, we are interested in the survival predictions of a new subject i which has a set of longitudinal biomarkers $\mathcal{Y}_i(t) = \{\boldsymbol{y}_i(s); 0 \geq s \geq t\}$. Rizopoulos (2011) focused on the conditional probability of survival time $v > t$ given survival up to t:

$$\pi_i(v|t) = \mathbb{P}\{T_i^* \geq v | T^* > t, \mathcal{Y}_i(t), \mathcal{D}_n\},$$

where v is the follow-up time ($v > t$), $\mathcal{D}_n$ denotes the sample on which joint model was fitted, and $\mathcal{Y}_i(t)$ are the longitudinal markers. Given this definition, it is easy to see that when new information is recorded at a new time $t' > t$, the predictions can be updated just by computing $\pi_i(v|t')$, with $v < t'$.

5.3.4 Time-Dependent ROC Curve

Discrimination measures based on Receiver Operating Characteristic (ROC; Heagerty and Zheng, 2005) methodology are very useful to assess the predictive capacity of the survival model. In the base of a binary outcome Y, the accuracy of a prediction is summarised through sensitivity $\mathbb{P}(p_i > c|Y_i = 1)$, indicating the probability of obtaining a true positive, and specificity $\mathbb{P}(p_i < c|Y_1 = 0)$, which is the probability of obtaining a true negative prediction. Here, p_i is the prediction and c is a criterion for classification. But when dealing with time-dependent variables, this is not enough, as sensitivity and specificity need to be computed for every time moment. Also, we have seen that the predictions of a JM have a dynamic nature, as they can be updated as time passes and new information is obtained.

Rizopoulos (2011) presented new prediction rules which can be used for JMLS, leading to new definitions of sensitivity and specificity, restricted to a certain rele-

USC DE SANTIAGO DE COMPOSTELA

vant time interval $(t, t + dt)$:

$$\begin{aligned}\text{sensitivity} &= \text{TP}_t^{dt} = \mathbb{P}\{\{p_i(s) \geq c_s; k \geq s \geq t\} | t < T_i^* \leq t + dt\} \\ \text{specificity} &= 1 - \text{FP}_t^{dt} = \mathbb{P}\{\{p_i(s) < c_s; k \geq s \geq t\} | t < t + dt < T_i^*\},\end{aligned}$$

where TP indicates the true positives, FP are the false positives (individuals wrongly classified as positive), p_i is the prediction at each time point t, c is the criterion of classification and k specifies which past values are used. Using this definition, it follows that at any fixed time t, the entire sample is classified as a case or control based on the vital state at that time t. Therefore, each individual plays the role of control when they have not yet experienced the event and of case in the opposite situation. This is called the cumulative/dynamic approach. Now, the time-dependent ROC curve por each interval $(t, t + dt)$ is:

$$\text{ROC}_t^{dt}(p) = \text{TP}_t^{dt}\{\left(\text{FP}_t^{dt}\right)^{-1}(p)\},$$

and the Area Under the Curve (AUC) is:

$$\text{AUC}_t^{dt} = \int_0^1 \text{ROC}_t^{dt}(p)dp.$$

$\text{AUC}_t^{dt} \in (0, 1)$, and the closer to 1, the greater the discriminative capacity of the model.

5.4 JMLS Regression of Liver-19 Data

5.4.1 Motivation of the Study

In Section 3.3 we have seen that liver function is impaired in patients with COVID-19, resulting in increased transaminase values. In addition, we have seen clear differences in the progression over time of transaminases between people who survived the disease and people who unfortunately died. Knowing this, we wanted to analyse these data from another perspective and study whether transaminase values can help predict mortality in patients with COVID-19, using JMLS regression models.

5.4.2 Statistical Analysis of Liver-19 Data

In total there are 205 patients available for analysis, of whom 36 have died (17.56%). Of the total number of patients 118 (57.56%) where males, and 47 (22.93%) suffered

USC DE SANTIAGO DE COMPOSTELA

from diabetes. The transaminases included in the model are AST, ALT and GGT. ALP is excluded from this study due to missingness in that variable. Missing values are also the reason that we picked only 205 patients, instead of the 212 patients in total in the database, since 7 patients did not have values for at least one of the three transaminases considered.

In this section, a JMLS regression model is fitted including the three transaminase covariates, AST, ALT and GGT, adjusting for age of the patient and the presence of diabetes mellitus. In view of Figure 2.1 from Chapter 2, and taking into account our previous knowledge that transaminases seem to behave in a certain way during the first 15 days of infection and differently from the second week onwards, in the longitudinal sub-models, time was considered as a two-part linear function, with a change occurring after 15 days, using B-splines for polynomial splines of order 1 (De Boor, 1978).

So, for each longitudinal covariate, the same formulation for the model was considered (5.6):

$$trans_{kij} = \beta_0 + \beta_{k1} B_n(t_{ij}, 15) + \beta_{k2} B_n(t_{ij}, 30.61) + u_{k0i} + u_{k1i} t_{ij}, \tag{5.6}$$

where $trans_k$ is one of the three transaminases considered (AST, ALT, and GGT); $B_n(t_{ij}, 15)$ and $B_n(t_{ij}, 30.61)$ denote the B-spline basis matrix for a natural spline of order 1 (De Boor, 1978), representing each one of the two-part linear function of the longitudinal time (t_{ij}) (the first part up to day 15 and the second one up to day 30.61, which is the maximum observation time); and u_{0i} and u_{1i} are the random intercept and slope.

For survival sub-model, a Cox model was implemented, the formulation of which can be seen below in (5.7):

$$h_i(t|\mathcal{M}_i(t)) = h_0(t) \exp\Bigg\{ \gamma_1 Sex_i + \gamma_2 B_n(Age_i, d_1) + \gamma_3 B_n(Age_i, d_2) + \gamma_3 dm_i + \sum_{k=1}^{3} \alpha_k trans_{ki} \Bigg\} \tag{5.7}$$

where $\mathcal{M}_i(t) = \{\boldsymbol{trans}_i(s) | 0 \leq s < t\}$ indicates history of the longitudinal markers previous to time t, dm represents diabetes mellitus (whether the patient suffers from it or not), and $trans_k$ indicates the estimated values for each transaminase obtained from the corresponding model (5.6). Here, $B_n(Age_i, d_l)$, for $l = \{1, 2\}$ indicates the B-spline basis matrix for a natural cubic spline with two degrees of freedom (df) (De Boor, 1978). We decided to model Age with a quadratic non-linear smooth function based on Figure 2.5 from Chapter 2, where we have seen that the effect of Age on survival follows this form.

USC DE SANTIAGO DE COMPOSTELA

JMLS is built under a Bayesian approach, so some initial parameters need to be established. The model was fitted for 4 chains, with 55,000 iterations per chain, 5,000 burn-in per chain and a thinning of 50 (these are the default values of `BayesX` programme (Belitz et al., 2015), which were used as reference to ensure convergence). The model took 11.8min to run on a single core of a 3.60 GHz Intel (R) Core (TM) i7-7700. After ensuring that the effective size was always > 50 and all potential scale reduction factors ($R - hat$) were between 0.90 and 1.05 (necessary requirements for convergence of the models, without which the results are not reliable; see Morita et al., 2008; and Vehtari et al., 2021), the output of the models is shown below. Table 5.1 shows the results for the longitudinal sub-model for AST; in Table 5.2, the results for the longitudinal sub-model for ALT are depicted; Table 5.3 shows the results for the longitudinal sub-model for GGT; finally, in Table 5.4 the results for the survival sub-model can be seen.

Table 5.1: Summary for the longitudinal sub-model for AST. SD is standard deviation of the estimated coefficients β, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(t, 15)$ and $B_n(t, 30.61)$ denote the B-spline basis matrix for a natural spline of order 1 for t, and σ is the estimation of the standard deviation of the error terms ε_{ij}.

Covariates	Mean estimate β	SD	2.5%CI	97.5%CI	P-value	$R - hat$
(Intercept)	-0.2139	0.0496	-0.3102	-0.1148	< 0.0001	0.9997
$B_n(t, 15)$	-0.1641	0.0805	-0.3229	-0.0069	0.0400	0.9998
$B_n(t, 30.61)$	-0.6855	0.2524	-1.1903	-0.1930	0.0050	1.0000
σ	0.3206	0.0100	0.3013	0.3407	< 0.0001	1.0002

Table 5.2: Summary for the longitudinal sub-model for ALT. SD is standard deviation of the estimated coefficients β, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(t, 15)$ and $B_n(t, 30.61)$ denote the B-spline basis matrix for a natural spline of order 1 for t, and σ is the estimation of the standard deviation of the error terms ε_{ij}.

Covariates	Mean estimate β	SD	2.5%CI	97.5%CI	P-value	$R - hat$
(Intercept)	-0.4776	0.0603	-0.5953	-0.3593	< 0.0001	0.9999
$B_n(t, 15)$	0.8784	0.1077	0.6700	1.0944	< 0.0001	1.0015
$B_n(t, 30.61)$	-0.0906	0.3287	-0.7310	0.5435	0.7975	1.0059
σ	0.3163	0.0106	0.2968	0.3381	< 0.0001	1.0002

It can be seen that in the three longitudinal sub-models, t is statistically significant, as expected.

USC DE SANTIAGO DE COMPOSTELA

Table 5.3: Summary for the longitudinal sub-model for GGT. SD is standard deviation of the estimated coefficients β, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(t, 15)$ and $B_n(t, 3.61)$ denote the B-spline basis matrix for a natural spline of order 1 for t, and σ is the estimation of the standard deviation of the error terms ε_{ij}.

Covariates	Mean estimate β	SD	2.5%CI	97.5%CI	P-value	$R - hat$
(Intercept)	-0.6825	0.1528	-0.9933	-0.3845	< 0.0001	1.0000
$B_n(t, 15)$	0.4289	0.1151	0.1895	0.6469	0.0010	0.9999
$B_n(t, 30.61)$	-1.4885	0.3679	-2.2154	-0.7714	< 0.0001	1.0002
σ	0.2714	0.0156	0.2461	0.3064	< 0.0001	1.0006

Table 5.4: Summary for the survival sub-model. SD is standard deviation of the estimated coefficients γ and α, 2.5%CI and 97.5%CI are the limits of the corresponding 95% confidence intervals for the estimates, and $R - hat$ are the potential scale reduction factors, that should be between 0.90 and 1.05. $B_n(Age, d_1)$ and $B_n(Age, d_2)$ denote the B-spline basis matrix for a natural cubic splines with 2 degrees of freedom for Age.

Covariates	Mean estimate (γ and α)	SD	2.5%CI	97.5%CI	P-value	$R - hat$
Sex (ref: Female)	0.3625	0.5550	-0.6807	1.4751	0.5185	0.9996
$B_n(Age, d_1)$	1.5299	3.5726	-4.4840	9.3874	0.7285	1.0015
$B_n(Age, d_2)$	3.6210	0.9188	1.8813	5.5295	< 0.0001	1.0033
dm (ref: No)	1.3052	0.4025	0.5359	2.1078	0.0010	0.9996
AST	1.9705	0.6807	0.6810	3.3285	0.0020	1.0037
ALT	-1.2997	0.4296	-2.1599	-0.4610	0.0015	1.0012
GGT	-0.1590	0.3073	-0.7761	0.4406	0.6045	1.0015

Table 5.4 shows that Age and dm have a statistically significant effect on the risk of event (in this case, death), but not Sex, as expected after the exploratory analysis of the survival (Figures 2.3 and 2.4 from Chapter 2). Regarding the longitudinal covariates, the effect of AST and ALT on survival is statistically significant, but this is not the case with GGT. The coefficients here should be interpreted as in a logistic regression: the sign indicates whether the risk increases or decreases, but the true value of the effect is seen after applying the exponential function. In this case, instead of the OR typical for binary regression models, we have the HR (hazard ratio). Thus, risk for death increases with Age and it is also higher for the people suffering from diabetes (HR for dm: 3.69, which means that these patients have a risk for death more than 3 times higher than people without diabetes). Risk for death also increases with AST (the higher the AST, the higher the risk for death), whereas decreases with ALT (the higher the ALT, the lower the risk for death). These results

USC DE SANTIAGO DE COMPOSTELA

are consistent with the initial plots depicted in Figure 2.1 from Chapter 2.

Dynamic Predictions

As time passes and new longitudinal data is being collected, survival probability for each patient is expected to change. This can be seen by performing dynamic predictions, one of the most interesting features of JMLS regression models, which can be of great use for so-called personalised medicine. For illustrative purposes, the dynamic prediction for two patients is shown below. The individuals chosen are:

- id=180: Male, 73 years, non-diabetic, deceased. Follow-up time: 30.61 days.
- id=183: Male, 71 years, non-diabetic, discharged. Follow-up time: 30.48 days.

Both individuals are of similar characteristics (excluding the longitudinal ones), but one of them died, while the other was discharged from hospital. The dynamic predictions are graphically depicted for id=180 in Figure 5.1, and for id=183 in Figure 5.2. In both cases, estimations were updated on day 7, 15, 21 and 29, and the 30 days survival probability was considered. In the figures, both survival estimation up to day 30 and transaminase estimation up to the day of the estimation can be seen.

For each observation time, the graphs from Figures 5.1 and 5.2 show, on the left hand side, the estimated transformed transaminase values over time up to the time of prediction (7, 15, 21 and 29 days), with their corresponding confidence bands (AST in grey, ALT in green and GGT in blue). The dots indicate the true recorded observations. The right-hand side shows the estimated probability of survival, in our case always at 30 days from admission to the ward. It can be seen that as the observation time increases and new transaminase measurements are obtained, the estimates are updated as a consequence of the incorporation of this new information.

In these two subjects (id=180 and id=183), of similar age and both non-diabetic, the difference in their measurements is found in the transaminase values. In the survival predictions with information up to days 7 and 15 there are hardly any differences between them. On the other hand, for the predictions with information up to days 21 and 29 (already in the second fortnight, where we had seen that it was when there were more differences between the transaminase values between the people who survived and those who finally died) the patient with id=183 always shows a higher survival probability than id=180.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

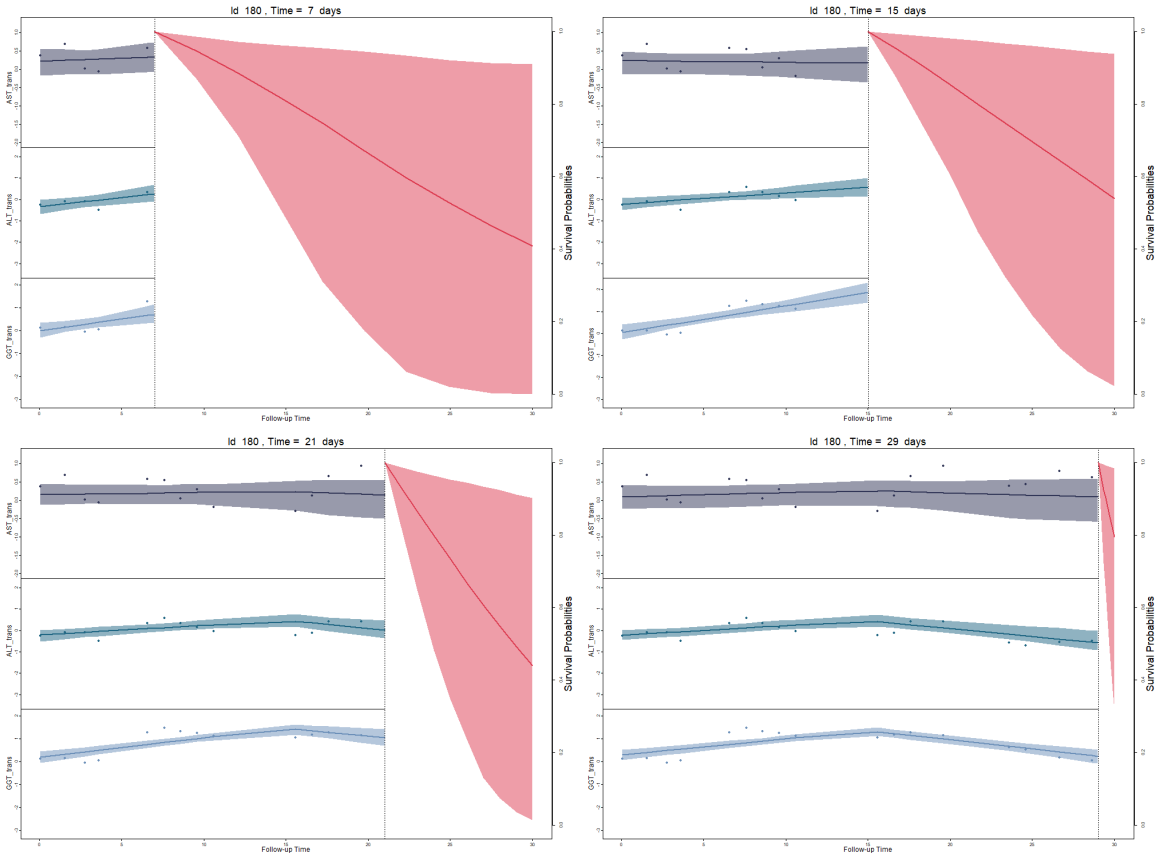

Figure 5.1: Dynamic predictions for patient id=180, with 95% confidence interval. Left: estimation of transaminases, with AST on the top, ALT on the middle, and GGT on the bottom. Right: estimation of the survival probability.

Time-Dependent ROC Curve

As we already stated, ROC curves and AUCs are computed for certain time moment t and in a certain time interval $(t + dt)$. This brings up the problem of dt selection. Depending on the biomedical problem, a small dt might be preferable in order to detect differences within days or make predictions for near future time, as is the case in this application. But for the computation of the AUC, an event is required to occur in $(t, t + dt)$. If the time interval is very short, it is likely that no event will occur within it, so the AUC cannot be calculated. If it is, the number of events in the interval is likely to be very small, leading to a very volatile curve. As an illustration, Figure 5.3 represents the AUC values for the JMLS regression model fitted for specific time intervals: 1, 3, and 7 days. In other words, they are the AUC

USC DE SANTIAGO DE COMPOSTELA

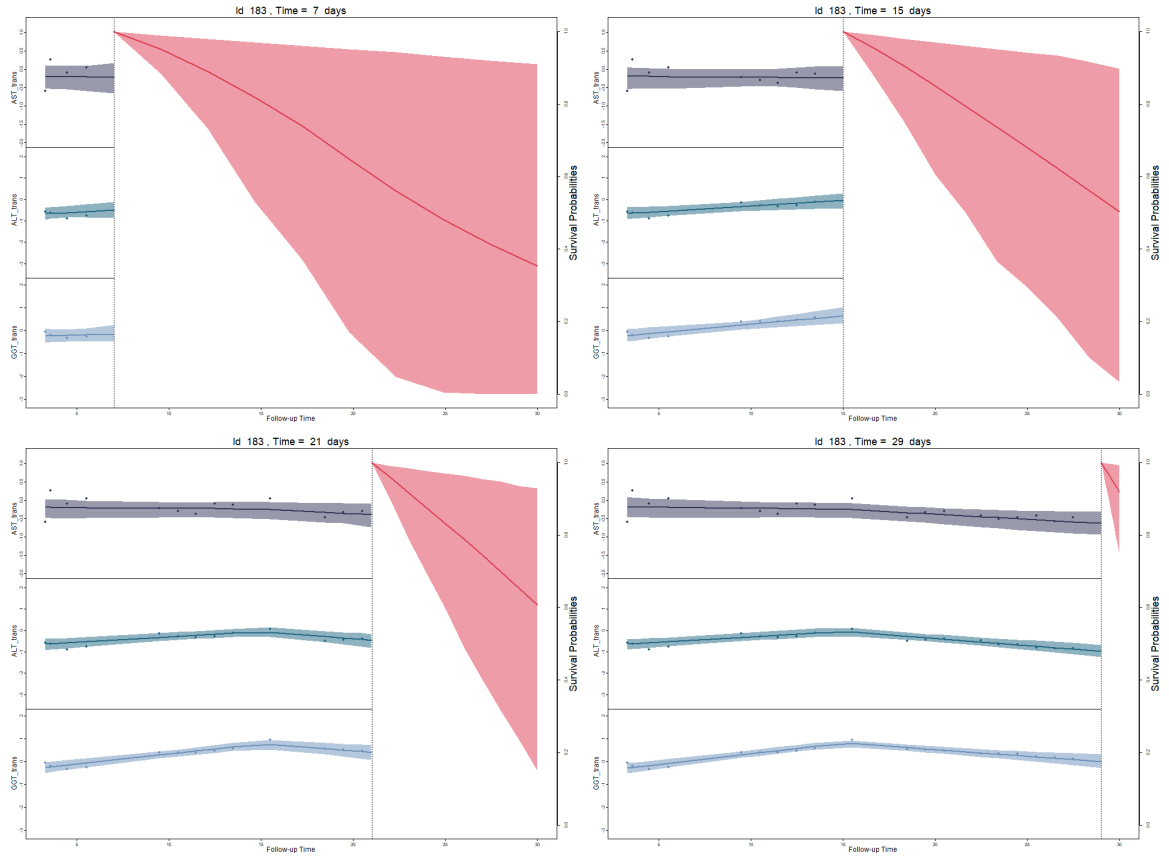

Figure 5.2: Dynamic predictions for patient id=183, with 95% confidence interval. Left: estimation of transaminases, with AST on the top, ALT on the middle, and GGT on the bottom. Right: estimation of the survival probability.

computed for every survival time we have in our sample using information from 1, 3 or 7 days before.

In Figure 5.3 we can see that if we take 1 day as the length of the time interval from observation to prediction (top graph), the discrimination ability is very variable and is around 0.5 (i.e. no discrimination). In addition, we observe periods where the AUC could not be calculated because no events occurred in those 1-day time intervals considered. This indicates that the time period considered is very short. However, as we increase the length of the prediction interval, the AUC becomes less variable and there are fewer periods of time not calculated. For the case of the 3 and 7 day intervals, although there is more variability in the AUC values in the smaller one, the trend seems to be the same, around 0.65 approximately. This ex-

USC
DE SANTIAGO
DE COMPOSTELA

ample highlights the need for further study of time-dependent AUC and the choice of the length of the prediction interval in the context of JMLS.

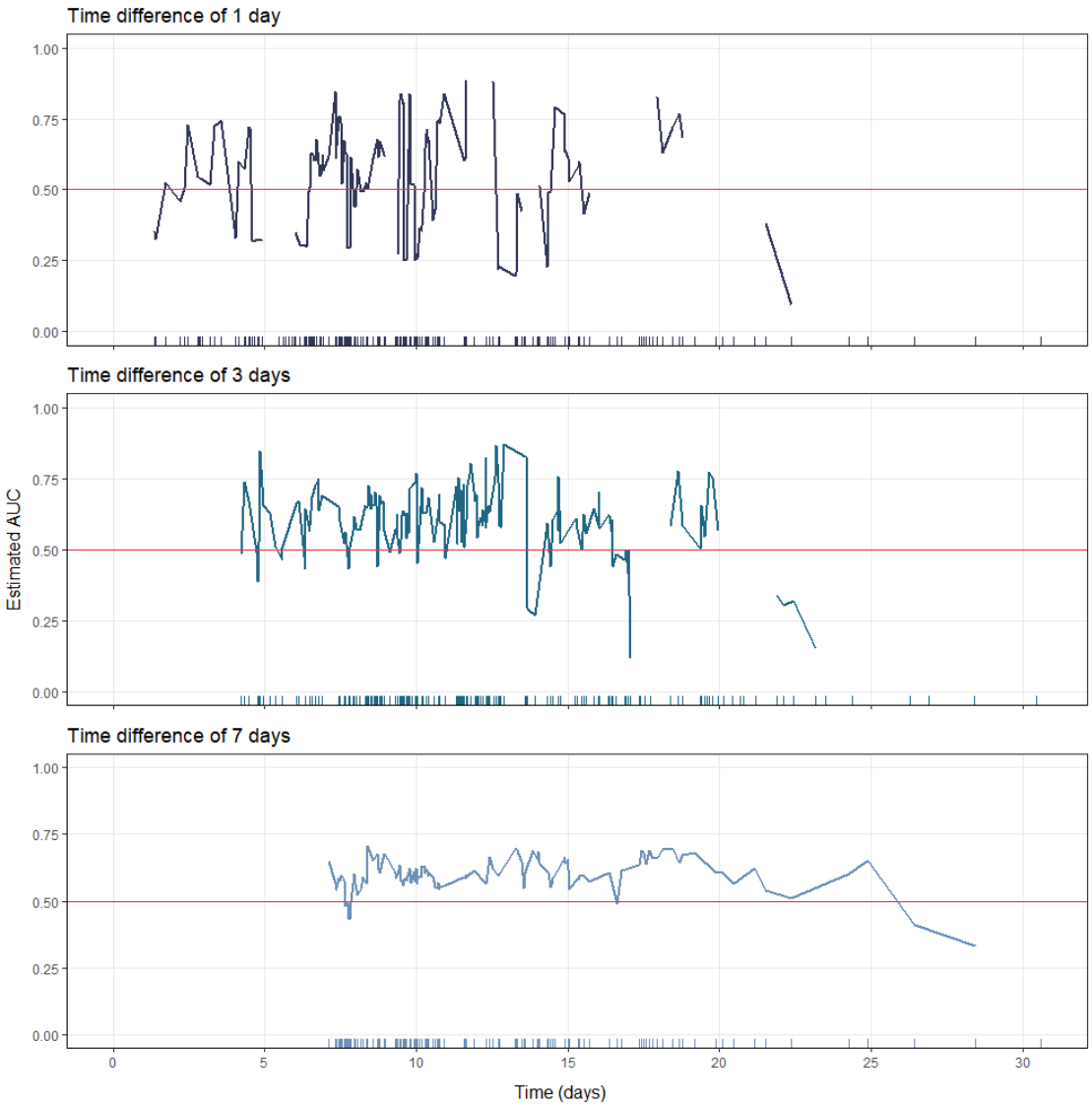

Figure 5.3: AUC for different the survival time points for the JMLS regression model fitted. We assume different time intervals to update the survival predictions (dt = 1, 3, and 7). Horizontal pink line represents AUC = 0.5, indicating that the discrimination ability of the model is zero.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 6

Software and Programming Code

This chapter contains the most important parts of the code necessary for the data analysis carried out throughout this thesis. All computations were performed in `R` (R Core Team, 2020), version 4.1.1. Although throughout the thesis, most of the graphs have been made using the `ggplot2` package (Wickham, 2016), in this section we will show how to do it in R-basic to simplify the code for the user.

6.1 Code for Section 3.3: "Bivariate Joint Modelling of Liver-19 Data"

In this section, we present the basic code we used in the analysis of the Liver-19 database discussed in Section 3.3 from Chapter 3. As several GAMM and CGAMLSS models have been fitted, we will reproduce the code for only one of these models. Specifically, as an example of GAMM we will take the model for `ALT` shown in Table 3.1 from Chapter 3, and we will use the copula model for outcomes `ALT` and `CRP` as an example of CGAMLSS. The analyses with the other response variables would be performed in a similar way.

Before starting, it is necessary to indicate that each row in the database has to be the entry of a given patient at a given time. For instance, let `dat` be the database in `R data.frame` format:

```
> head(dat)
```

```
  id  Sex      Age   Exitus  AT       Time       AST       ALT
1  1 Male 77.57377 Deceased  No 0.00000000       NA        NA
2  1 Male 77.57377 Deceased  No 0.01944444 1.054054 0.4444444
3  1 Male 77.57377 Deceased  No 0.45833333       NA        NA
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
4  1 Male 77.57377 Deceased  No 0.85763889       NA        NA
5  1 Male 77.57377 Deceased Yes 1.20277778       NA        NA

        GGT ALP      CRP D_Dimer      IL6 Lymphocytes
1        NA  NA       NA      NA 0.186801          NA
2 0.6823529  NA 49.39394    1428       NA        22.6
3        NA  NA       NA      NA       NA          NA
4        NA  NA       NA      NA 0.182963          NA
5        NA  NA       NA      NA       NA          NA
```

In this database, `id` is the indicator of the subject, `Sex` is their gender (`Male` or `Female`) `Age` is their age in years on admission, `Exitus` is their final vital status (`Deceased` of `Alive`), `AT` indicates whether they were under antiviral treatment in that precise moment and `Time` is the time at which the longitudinal measurements were taken. `NA` indicates missing values.

At this point, we then normalised the transaminases, D-dimer, and also IL-6 values using a Box–Cox transformation. For this to work, it is necessary that the database contains no missing data for these variables. Also, GAMM and CGAMLSS only work with complete cases, so the first thing to do is to take a sub-sample with only the variables of interest and eliminate non-complete cases.

```
dat_1 <- na.omit(dat[, c(1:6, 8, 11:12)])
```

Here, `c(1:6, 8, 11:12)` is how we indicate the number of the columns we are interested in keeping. To create the Box-Cox transformed variables, `ALT_trans` and `D_Dimer_trans`, we can use the following function:

```
boxcox_trans <- function(x, lambda = seq(-2, 2, 0.01)){
  if(min(x) <= 0){
    stop("All values of the variable must be positive")
  }
  require(MASS)
  result = boxcox(x ~ 1, lambda = lambda, plotit = F)
  mylambda = result$x[which.max(result$y)]
  if(mylambda == 0){
    output <- log(x)
  }else{
    output <- (x^mylambda - 1)/mylambda
  }
  rm(result)
  return(list(output, mylambda))
}
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
dat_1$ALT_trans <- boxcox_trans(dat_1$ALT)[[1]]
dat_1$D_Dimer_trans <- boxcox_trans(dat_1$D_Dimer)[[1]]
```

6.1.1 GAMM regression model for ALT

As stated at the beginning of this section, we are providing the code to reproduce the model for `ALT` shown in Table 3.1 from Chapter 3. In order to do so, we are going to use `mgcv` package (Wood, 2017).

```
library(mgcv)

mod1 <- gam(ALT_trans ~  Sex + s(Age) + AT + CRP + D_Dimer +
              Exitus + s(Time, by = Exitus) + s(id, bs = "re"),
            data = dat_1, method = "REML")

par(mfrow = c(2,2))
gam.check(mod1)
```

Here, $\sim$ is the symbol used to separate the outcome from the formula with the covariates; `s(Age)` indicates a smooth function (non linear) for the effect of variable `Age`; `s(Time, by = Exitus)` is the interaction between the smooth function of `Time` and `Exitus`, which allows for a different smooth effect of `Time` for `Alive` and `Deceased` patients; and, finally, `s(id, bs = "re")` is how random intercepts for each subject are introduced in the model.

The function `gam.check(mod1)` is meant for a quick check of the model fitting, and returns the following output:

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-0.0003600757,4.235505e-06]
(score 659.7408 & scale 0.1516056).
Hessian positive definite, eigenvalue range [0.2207489,440.0226].
Model rank =  238 / 238

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                                    k'    edf k-index p-value
s(Age)                            9.00   2.94    1.03   0.875
s(Time):as.factor(Exitus)Alive    9.00   4.99    0.95   0.095 .
s(Time):as.factor(Exitus)Deceased 9.00   5.50    0.95   0.095 .
```

USC DE SANTIAGO DE COMPOSTELA

```
s(id)                              205.00 170.86      NA      NA

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As all p-values are higher than 0.05, in principle, there is no problem with the fit of the model. This command also provides the graphical output that can be seen in the Figure 6.1.

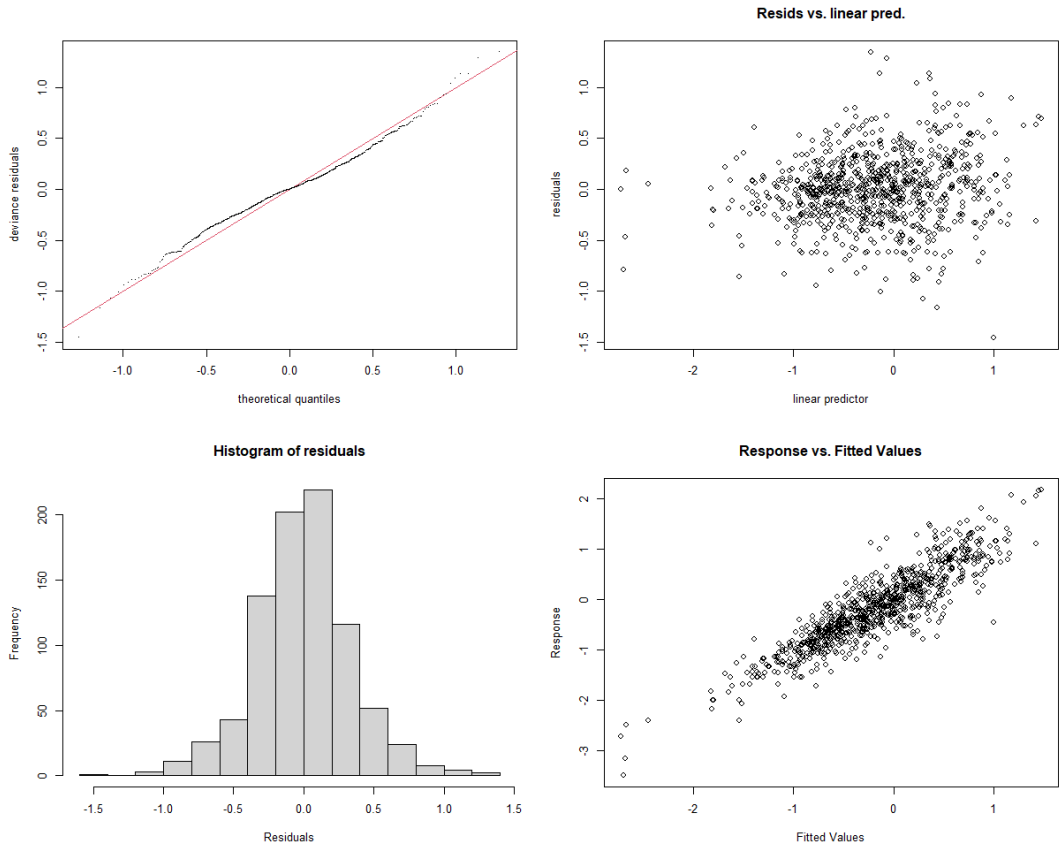

Figure 6.1: Graphical output of the command `gam.check`. If the model is adequate: theoretical quantiles will be very close to the straight red line (upper left panel), residuals will not follow a clearly distinguishable pattern (top right panel), the histogram will resemble that of a normal variable (Gaussian bell shape, bottom left panel), and response vs. fitted values will follow an increasing linear pattern (bottom right panel).

From Figure 6.1, we can say that the model fits well since theoretical quantiles are very close to the straight red line (upper left panel), residuals do not follow a clearly

USC DE SANTIAGO DE COMPOSTELA

distinguishable pattern (top right panel), the histogram resembles that of a normal variable, showing a Gaussian bell shape (bottom left panel), and the representation of the response vs. fitted values seems to follow an increasing linear pattern (bottom right panel). `par(mfrow = c(2,2))` just indicates that we want to divide the plot window into four panels, distributed into two rows and two columns.

Once the model was checked, the numerical results for the model can be obtained with `summary(mod1)`:

```
> summary(mod1)

Family: gaussian
Link function: identity

Formula:
ALT_trans ~ +Sex + s(Age) + AT + CRP + D_Dimer + Exitus + s(Time,
    by = as.factor(Exitus)) + s(id, bs = "re")

Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.706e-01  7.165e-02  -2.381   0.0175 *
SexMale         8.355e-02  8.465e-02   0.987   0.3240
ATYes           4.859e-02  4.192e-02   1.159   0.2468
CRP            -7.927e-03  1.388e-03  -5.711  1.7e-08 ***
D_Dimer        -3.981e-06  3.751e-06  -1.061   0.2889
ExitusDeceased -2.235e-01  1.273e-01  -1.756   0.0796 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                      edf  Ref.df      F
s(Age)                              2.940   3.062  4.249
s(Time):as.factor(Exitus)Alive      4.987   6.114 18.658
s(Time):as.factor(Exitus)Deceased   5.505   6.575  4.662
s(id)                             170.859 201.000  7.789
                                   p-value
s(Age)                             0.00504 **
s(Time):as.factor(Exitus)Alive     < 2e-16 ***
s(Time):as.factor(Exitus)Deceased 8.87e-05 ***
s(id)                              < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
R-sq.(adj) =  0.738   Deviance explained = 79.6%
-REML = 659.74  Scale est. = 0.15161    n = 849
```

In `Parametric coefficients:` chart, `Estimate` is the estimated value of the coefficients for the lineal part of the model, `Std. Error` is their standard error and `Pr(>|t|)` is their p-value. For the smooth part, instead of the coefficients of the basis, we obtain the effective degrees of freedom `edf` and their `p-value`. From the summary we also extract some parameters relating the performance of the model: the adjusted R^2 (`R-sq.(adj)`), and the percentage of deviance explained (`Deviance explained`).

We can obtain a graphical representation of all centred smooth effects in one plot, displayed in Figure 6.2:

```
plot(mod1, pages = 1, shade = T)
```

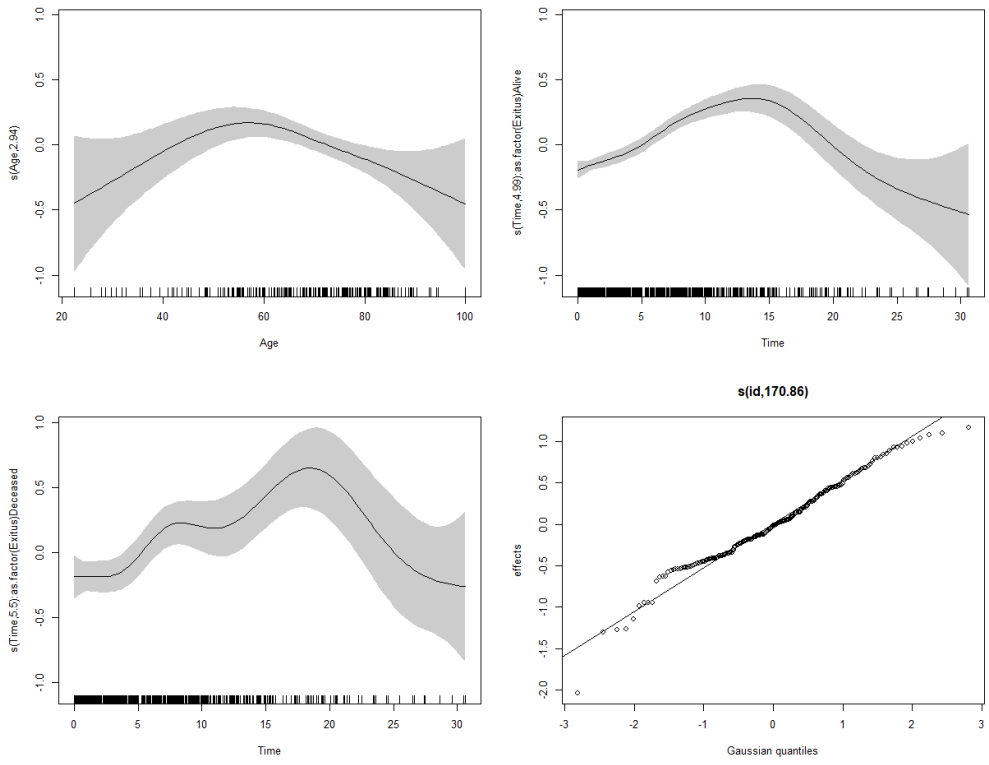

Figure 6.2: Plot for all the smooth effects in model `mod1`, obtained with the `R` command `plot(mod1, pages = 1, shade = T)`.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Smooth effects can be represented separately, by choosing the number of the covariate in the parameter `select`. For instance, the plot of the smooth effect of `Age`, which is the first smooth variable, is depicted in Figure 6.3.

```
par(mfrow = c(1, 1))
plot(mod1, shade = T, se = T, select = 1, xlab = "Age (years)",
     main = "Age (years)", scale = 0)
abline(h = 0, col = "red")
```

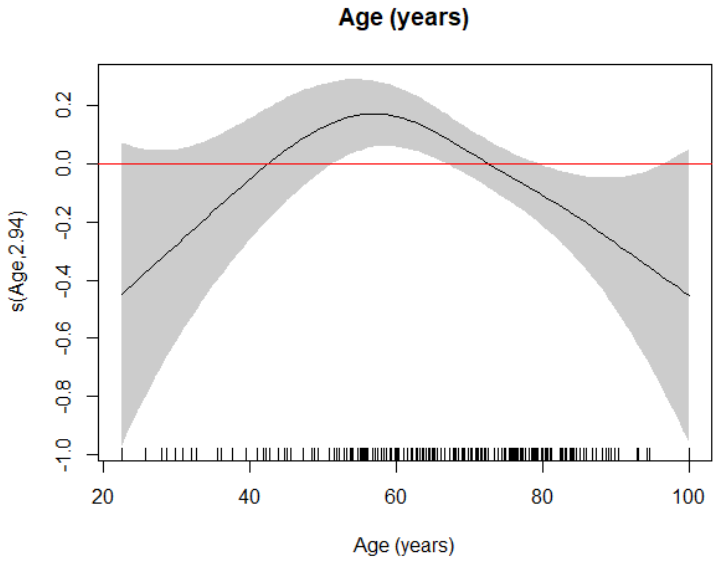

Figure 6.3: Plot of the smooth effect of `Age` in model `mod1`.

`abline(h = 0, col = "red")` just draws an horizontal red line indicating the 0. Where this line falls within the confidence bands, the effect of the variable is not statistically significant.

6.1.2 CGAMLSS regression model for ALT and CRP

In order to fig a CGAMLSS, we need to use `gjrm` package (Marra and Radice, 2022). The first step is to choose an appropriate distribution for the outcomes. As we are working with `ALT_trans`, which was transformed to be normal, we only need to look for a suitable distribution for `CRP`. In order to do that, the function `best_dist` is of great help:

USC DE SANTIAGO DE COMPOSTELA

```
library(GJRM)

best_dist = function(y, discrete = F){
  if(discrete == F){
    dist = c("N","LN","GP","GU","rGU","LO","WEI","iG","GA","DAGUM",
             "SM","FISK")
  }else{
    dist = c("DGP","PO","ZTP","NBI","NBII","PIG")
  }
  AICs = BICs = ind = NULL
  if(!require("GJRM")){
    install.packages("GJRM", dependencies = T)
  }
  require(GJRM)
  for(k in 1:length(dist)){
    skip_to_next <- FALSE
    tryCatch({
      f <- resp.check(y, margin = dist[k], loglik = T, plots = F)
    }, error = function(e){skip_to_next <<- TRUE})
    if(skip_to_next){
      ind <- c(ind, k)
      next
    }
    AICs = c(AICs, AIC(f))
    BICs = c(BICs, BIC(f))
  }
  if(is.null(ind) == F){
    dist <- dist[-ind]
  }
  df = data.frame(Distributions = dist, AIC = round(AICs, 3),
                  BIC = round(BICs, 3))
  return(df[order(df[, 2]), ])
}

bd1 <- best_dist(dat_1$CRP)
bd1
```

`best_dist` gives us the possible distributions of the variable ordered from lower to higher AIC. We can see that, in this case, the best distribution (according to AIC criterion) was the `DAGUM`, which has 3 parameters. For simplicity, and since there seem to be no differences between this distribution and a gamma (`GA`), we prefer this second one because it has only 2 parameters.

USC DE SANTIAGO DE COMPOSTELA

```
   Distributions      AIC      BIC
10         DAGUM 6207.003 6221.236
9             GA 6208.933 6218.421
7            WEI 6220.655 6230.143
11            SM 6222.655 6236.887
3             GP 6256.812 6266.300
12          FISK 6399.460 6408.948
2             LN 6446.573 6456.062
5            rGU 6682.120 6691.608
6             LO 6955.652 6965.140
1              N 7050.623 7060.112
8             iG 7555.086 7564.574
4             GU 7580.075 7589.563
```

With function `resp.check` we can check graphically whether our choice is appropriate (see the top left panel of Figure 3.3 from Chapter 3):

```
resp.check(dat_1$CRP, margin = "GA",
           main = "Histogram and Density", xlab = "CRP")
```

Now that we have chosen the best distribution for the marginals, we need to choose the best copula. To do so, we need to establish first all the sub-models involved. In our case, we are just modelling the correlation as a function of covariates, so the other sub-models (the ones corresponding to the parameters of the marginal distributions) need to be set as constants.

```
mu.1 = ALT_trans ~ 1

mu.2 = CRP ~ 1

sigma.1 = ~ 1
sigma.2 = ~ 1

theta = ~ Sex + s(Age) + Exitus + AT + s(Time, by = Exitus)

f <- list(mu.1, mu.2, sigma.1, sigma.2, theta)
```

Once the sub-models have been established, we use the function `best_cop` to help us choose the best copula.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
best_cop = function(f = f, marginals = c("N", "N"), data = NULL,
                    plus = F, gamlssfit = FALSE){
```

```
  if(plus == F){
    copulas = c("N", "C0", "C90", "C180", "C270", "J0", "J90",
                "J180", "J270", "G0", "G90", "G180", "G270", "F",
                "PL")
  }else{
    copulas = c("N", "C0", "C90", "C180", "C270", "J0", "J90",
                "J180", "J270", "G0", "G90", "G180", "G270", "F",
                "PL", "AMH", "FGM", "T", "HO")
  }
  modelos = list()
  ind = NULL
  AICs = BICs = conv = NULL
  for(k in 1:length(copulas)){
    skip_to_next <- FALSE
    tryCatch({
      if(is.null(data) == T){
        fit = gjrm(f, margins = marginals, BivD = copulas[k],
                   Model = "B", gamlssfit = gamlssfit)
      }else{
        fit = gjrm(f, margins = marginals, BivD = copulas[k],
                   Model = "B", data = data, gamlssfit = gamlssfit)
      }
    }, error = function(e){skip_to_next <<- TRUE})
    if(skip_to_next){
      ind <- c(ind, k)
      next
    }
    modelos[[k]] = fit
    AICs = c(AICs, AIC(fit))
    BICs = c(BICs, BIC(fit))
    conv = c(conv, fit$fit$converged)
    print(k)
  }
  if(is.null(ind) == F){
    copulas <- copulas[-ind]
  }
  df = data.frame(Copulas = copulas, AIC = AICs, BIC = BICs,
                  Convergence = conv)
  return(list(df = df, fits = modelos))
}

cop_best <- best_cop(f, marginals = c("N", "GA"), data = dat_1,
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
                    gamlssfit = T, plus = F)
head(cop_best$df[order(cop_best$df$AIC),])
```

In this case, `marginals = c("N", "GA")` indicates that the marginal for the first outcome `ALT_trans` is the normal distribution, and the marginal for the second outcome `CRP` is the gamma distribution. The first six best copulas, according to AIC are:

```
   Copulas      AIC      BIC Convergence
5     C270 7959.266 8081.725        TRUE
3      C90 7973.201 7982.317        TRUE
13    G270 7997.122 8043.647        TRUE
1        N 8022.544 8113.147        TRUE
14       F 8022.672 8114.903        TRUE
7      J90 8023.156 8053.432        TRUE
```

The convergence indicator can be a bit treacherous, since even if it is established that the model has converged, it does not always converge to what it should. For instance, if we want to fit a model using the best copula `C270`, we have the following warning message:

```
> fit <- gjrm(f, margins = c("N", "GA"), BivD = "C270",
+             Model = "B", data = dat_1, gamlssfit = T)

Warning message:
Largest absolute gradient value is not close to 0.
 Read the WARNINGS section in ?gjrm.
```

This message indicates that the model is inadequate, and the same is true for `C90` and `G270`. Therefore, we have chosen normal copula as the best one. With the function `summary(fit)` we obtain the numerical results of the model, similar to those obtained in a GAMM regression model.

```
> fit <- gjrm(f, margins = c("N", "GA"), BivD = "N", Model = "B",
+             data = dat_1, gamlssfit = T)

> summary(fit)

COPULA:   Gaussian
MARGIN 1: Gaussian
MARGIN 2: gamma

EQUATION 1
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
Link function for mu.1: identity
Formula: ALT_trans ~ 1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.23397    0.02634  -8.884   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2
Link function for mu.2: log
Formula: CRP ~ 1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.65297    0.03821   69.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 3
Link function for sigma.1: log
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.27833    0.02379   -11.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 4
Link function for sigma.2: log
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14872    0.02082   7.142 9.21e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
EQUATION 5
Link function for theta: atanh
Formula: ~Sex + s(Age) + Exitus + AT + s(Time, by = Exitus)

Parametric coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.44909    0.06897  -6.511 7.46e-11 ***
SexMale         0.17335    0.07255   2.389   0.0169 *
ExitusDeceased -0.02198    0.09953  -0.221   0.8252
ATYes           0.07900    0.07182   1.100   0.2714
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
                         edf Ref.df Chi.sq  p-value
s(Age)                 8.362  8.880 48.256 5.05e-09 ***
s(Time):ExitusAlive    1.001  1.001 31.061  < 2e-16 ***
s(Time):ExitusDeceased 1.735  2.165  2.665    0.322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

sigma.1 = 0.757(0.727,0.789)  sigma.2 = 1.16(1.11,1.21)
theta = -0.26(-0.442,-0.0616)  tau = -0.175(-0.304,-0.0411)
n = 849  total edf = 19.1
```

`EQUATION 5` shows the correlation results. We can plot the centred smooth effect of the covariates for the correlation parameter θ just by using the function `plot()`, as in a GAMM regression model, but here we have a new parameter `eq` to indicate the number of equation we are refering to. For instance, if we would like to plot the centred smooth effect of `Time` (see Figure 6.4):

```
par(mfrow = c(1, 2))
plot(fit, eq = 5, shade = T, select = 2, xlab = "Time (days)",
     main = "Alive")
abline(h = 0, col = 2)
plot(fit, eq = 5, shade = T, select = 3, xlab = "Time (days)",
     main = "Exitus")
abline(h = 0, col = 2)
```

USC DE SANTIAGO DE COMPOSTELA

The smooth effect shown in Figure 6.4 is not really interpretable. In this case, it is much more interesting to represent this effect on the scale of Kendall's τ. To

do so, we need to use the `pred.mvt`, which make predictions, so it is necessary to create a new dataset in which the variables not involved in the plot (all but `Time`, `Exitus`, `ALT_trans` and `CRP`) are constants. In this case, we have set them all to their reference values except for `Age`, which we have used their mean value.

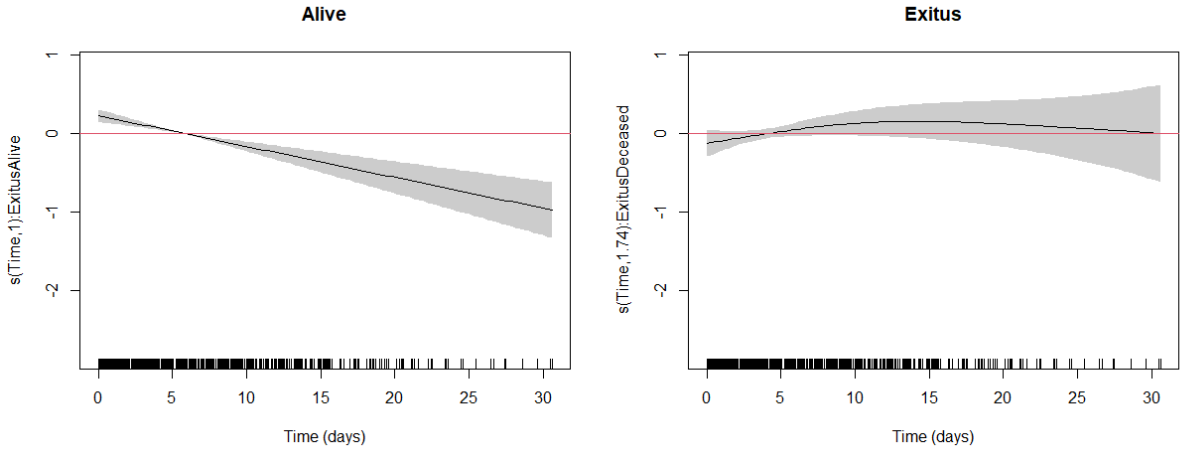

Figure 6.4: Plot of the smooth effect of `Time` in the correlation between `CRP` and `ALT_trans` according to model `fit1`.

```
dat_id <- dat[!duplicated(dat$id), ]
Age_m <- mean(dat_id$Age)
newdata <- data.frame(Time = dat_1$Time, Exitus = dat_1$Exitus,
                      Age = Age_m, Sex = "Female", AT = "No",
                      ALT_trans = dat_1$ALT_trans, CRP = dat_1$CRP)
```

```
pred <- pred.mvt(fit, eq = 5, fun = "tau", newdata = newdata)
dataset <- data.frame(tau = pred$pred, lci = pred$CIpred$CItauLB,
                      uci = pred$CIpred$CItauUB,
                      Exitus = dat_1$Exitus, Time = dat_1$Time)
```

In this new dataset, `lci` and `uci` are the lower and upper bounds for the 95% confidence bands for Kendall's τ. In order to plot this, we need to split the dataset into two: one for the people who survived (`Alive`) and one for the people who died (`Deceased`).

```
alive <- dataset[dataset$Exitus == "Alive", ]
deceased <- dataset[dataset$Exitus == "Deceased", ]
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

As we would like the legend to be placed on top outside the plot frame, we need to change the top margin of the plot window. We do this with the parameter `mar` of the function `par`.

```
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 5.1, 2.1), xpd = TRUE)
```

Now, we plot the first line (for example, the corresponding to the people who survived). As Kendall's τ goes from -1 to 1, we force the limits of the plot frame to those values using `ylim = c(-1, 1)`.

```
plot(alive$tau ~ alive$Time, col = "#1F6683", type = "l",
     ylim = c(-1, 1), lwd = 2, xlab = "Time (days)",
     ylab = expression("Kendall's"*phantom(x)*tau))
```

We add the confidence interval using the function `polygon`. As the pairs of points are to be indicated in an ordered way, we indicate the points corresponding to the upper margin in reverse order using the function `rev`.

```
polygon(x = c(alive$Time, rev(alive$Time)),
        y = c(alive$lci, rev(alive$uci)),
        col = "#1F668359")
```

Then, we do the same for the deceased, but changing the colour. Note that the function `lines` needs to be used instead of `plot` because we do not want to start a new plot.

```
lines(deceased$tau ~ deceased$Time, col = "#DD3C51", type = "l",
      lwd = 2)
polygon(c(deceased$Time, rev(deceased$Time)),
        c(deceased$lci, rev(deceased$uci)), col = "#DD3C5159")
```

Finally, we add an horizontal line indicating the 0 (no correlation) and we place the legend on top of the graphic. The final plot for the correlation can be seen in Figure 6.5.

```
segments(-1, 0, 31.7, 0, lwd = 2)
legend ("top", c("Alive", "Deceased"),
        fill = c("#1F6683","#DD3C51"), inset = c(0, -0.3),
        bty = "n", horiz = T)
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

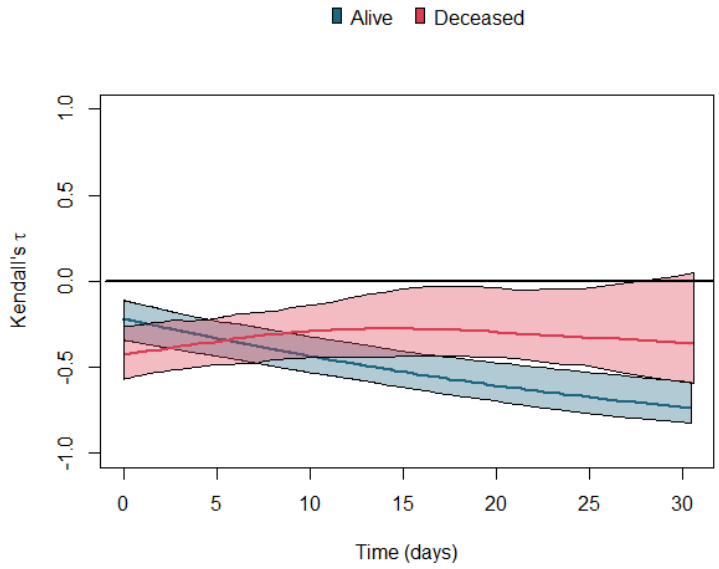

Figure 6.5: Correlation between `ALT_trans` and `CRP`, expressed as Kendall's τ, as a function of `Time` and `Exitus` according to model `fit1`. In pink, the effect for the deceased is shown, while the effect for the people who survived is depicted in blue. Black line indicates no correlation.

6.2 Code for Section 3.4: "Bivariate Joint Modelling of PPAD Data"

In this section, we present the basic code we used in the analysis of the PPAD database discussed in Section 3.4 from Chapter 3. Two CGAMLSS models were fitted in this analysis, one for pregnant women and the other for postpartum women. Here we will reproduce the code for the model for pregnant women. We start our analysis with a brief look at the database, called `pregnant`.

```
> head(pregnant)

  GAD7 EPDS FirstPregnancy CovidExposure total_deaths CHI_mean
2   11    7    Primiparous           Yes      0.35096 56.81609
3   12   16    Primiparous            No      0.35096 56.81609
4    0    8    Primiparous            No      0.35096 56.81609
5   15    6    Multiparous           Yes      0.35096 56.81609
6    5    6    Primiparous            No      0.35096 56.81609
7    4    1    Primiparous            No      0.35096 56.81609
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
   IHDI AgeMother Country
2 0.708        29 Albania
3 0.708        28 Albania
4 0.708        31 Albania
5 0.708        25 Albania
6 0.708        27 Albania
7 0.708        31 Albania
```

As we did in the previous section, we need to load `GJRM` package and select the best marginal distribution for the outcomes. For that purpose, we used the `best_dist` and `resp.check` functions, whose results were shown in Table 3.5 and Figure 3.6 from Chapter 3. In this case, we have to indicate `discrete = T` as our outcomes are not continuous.

```
library(GJRM)

best_dist(pregnant$EPDS, discrete = T)
best_dist(pregnant$GAD7, discrete = T)

par(oma = c(0, 0, 2, 0))
resp.check(pregnant$EPDS, margin = "DGP",
           main = "Histogram and Density", "EPDS")
mtext("EPDS in Pregnant women: DGP", outer = T, cex = 1.5,
      font = 4)
resp.check(pregnant$EPDS, margin = "NBI",
           main = "Histogram and Density", "EPDS")
mtext("EPDS in Pregnant women: NBI", outer = T, cex = 1.5,
      font = 4)
resp.check(pregnant$GAD7, margin = "DGP",
           main = "Histogram and Density", "GAD-7")
mtext("GAD-7 in Pregnant women: DGP", outer = T, cex = 1.5,
      font = 4)
resp.check(pregnant$GAD7, margin = "NBI",
           main = "Histogram and Density", "GAD-7")
mtext("GAD-7 in Pregnant women: NBI", outer = T, cex = 1.5,
      font = 4)
```

`par(oma = c(0, 0, 2, 0))` was used to enlarge the top margin of the graphics window, and with the function `mtext` we were able to place a title on top of the plot.

Now that we have chosen a negative binomial - type I distribution (`NBI`) for the outcomes, we need to define the models for all the distributional parameters.

USC DE SANTIAGO DE COMPOSTELA

```
mu.1 = GAD7 ~ FirstPregnancy + CovidExposure + total_deaths +
  CHI_mean:FirstPregnancy + IHDI:FirstPregnancy + s(AgeMother) +
  s(Country, bs = "re")

mu.2 = EPDS ~ FirstPregnancy + CovidExposure + total_deaths +
  CHI_mean:FirstPregnancy + IHDI:FirstPregnancy + s(AgeMother) +
  s(Country, bs = "re")

sigma.1 = ~ 1
sigma.2 = ~ 1

theta = ~ FirstPregnancy + CovidExposure + total_deaths +
  CHI_mean:FirstPregnancy + IHDI:FirstPregnancy + s(AgeMother) +
  s(Country, bs = "re")
```

`CHI_mean:FirstPregnancy` indicates the interaction between `CHI_mean`, considered linear, and `FirstPregnancy`, allowing for the linear effect of `CHI_mean` to be different in `Primiparous` and `Multiparous` (similarly, the same applies to `IHDI:FirstPregnancy`).

To choose the best copula, again we can use the function `cop_best` to identify the best copula.

```
> cop_best <- best_cop(f, marginals = c("NBI", "NBI"),
+                      data = pregnant, gamlssfit = T)
> cop_best$df[order(cop_best$df$AIC),]

   Copulas      AIC      BIC Convergence
10      G0 36219.01 36516.45        TRUE
4     C180 36345.59 36642.39        TRUE
6       J0 36385.98 36681.93        TRUE
14       F 36402.19 36700.87        TRUE
1        N 36417.58 36724.52        TRUE
15      PL 36425.89 36720.93        TRUE
12    G180 36804.89 37105.95        TRUE
2       C0 37301.61 37607.89        TRUE
8     J180 37403.03 37708.51        TRUE
5     C270 39174.21 39418.61        TRUE
3      C90 39174.21 39418.62        TRUE
7      J90 39174.21 39418.62        TRUE
9     J270 39174.21 39418.62        TRUE
11     G90 39174.21 39418.62        TRUE
13    G270 39174.21 39418.62        TRUE
```

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

As the best copula seems to be a Gumbel (`G0`), we fit the model using that copula function, and see the `summary` for the numerical results of the fitted model.

```
> fit_p <- gjrm(f, margins = c("NBI", "NBI"), BivD = "G0",
+               Model = "B", data = pregnant, gamlssfit = T)
> summary(fit_p)

COPULA:   Gumbel
MARGIN 1: Negative Binomial - Type I
MARGIN 2: Negative Binomial - Type I

EQUATION 1
Formula: ~FirstPregnancy + CovidExposure + total_deaths +
    CHI_mean:FirstPregnancy + IHDI:FirstPregnancy +
    s(AgeMother) + s(Country, bs = "re")

Parametric coefficients:
                                  Estimate Std. Error z value
(Intercept)                        3.359623   0.709659   4.734
FirstPregnancyPrimiparous         -0.983086   0.560664  -1.753
CovidExposureYes                   0.113972   0.039900   2.856
total_deaths                       0.029039   0.015512   1.872
FirstPregnancyMultiparous:CHI_mean -0.007609   0.004381  -1.737
FirstPregnancyPrimiparous:CHI_mean -0.003677   0.003882  -0.947
FirstPregnancyMultiparous:IHDI     -1.605755   0.798459  -2.011
FirstPregnancyPrimiparous:IHDI     -0.633355   0.792812  -0.799
                                  Pr(>|z|)
(Intercept)                        2.2e-06 ***
FirstPregnancyPrimiparous           0.07953 .
CovidExposureYes                    0.00428 **
total_deaths                        0.06121 .
FirstPregnancyMultiparous:CHI_mean  0.08239 .
FirstPregnancyPrimiparous:CHI_mean  0.34350
FirstPregnancyMultiparous:IHDI      0.04432 *
FirstPregnancyPrimiparous:IHDI      0.42436
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
              edf Ref.df Chi.sq  p-value
s(AgeMother) 2.851  3.586  22.92 0.000111 ***
s(Country)   8.448 10.000  95.98  < 2e-16 ***
```

USC DE SANTIAGO DE COMPOSTELA

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2
Link function for mu.2: log
Formula: ~FirstPregnancy + CovidExposure + total_deaths +
    CHI_mean:FirstPregnancy + IHDI:FirstPregnancy +
    s(AgeMother) + s(Country, bs = "re")

Parametric coefficients:
                                  Estimate Std. Error z value
(Intercept)                        3.510824   0.403920   8.692
FirstPregnancyPrimiparous         -1.208458   0.454277  -2.660
CovidExposureYes                   0.107841   0.032413   3.327
total_deaths                       0.027597   0.008474   3.257
FirstPregnancyMultiparous:CHI_mean -0.007782   0.003395  -2.292
FirstPregnancyPrimiparous:CHI_mean  0.001345   0.002971   0.453
FirstPregnancyMultiparous:IHDI     -1.284374   0.376416  -3.412
FirstPregnancyPrimiparous:IHDI     -0.442766   0.371132  -1.193
                                  Pr(>|z|)
(Intercept)                        < 2e-16 ***
FirstPregnancyPrimiparous         0.007810 **
CovidExposureYes                  0.000877 ***
total_deaths                      0.001127 **
FirstPregnancyMultiparous:CHI_mean 0.021899 *
FirstPregnancyPrimiparous:CHI_mean 0.650655
FirstPregnancyMultiparous:IHDI     0.000645 ***
FirstPregnancyPrimiparous:IHDI     0.232863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
               edf Ref.df Chi.sq  p-value
s(AgeMother) 4.670  5.668 24.286 0.000295 ***
s(Country)   4.901 11.000  9.597 0.027031 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 3
Link function for sigma.1: log
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.39409    0.03233  -12.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 4
Link function for sigma.2: log
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.84912    0.03234  -26.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 5
Link function for theta: log(· - 1)
Formula: ~FirstPregnancy + CovidExposure + total_deaths +
    CHI_mean:FirstPregnancy + IHDI:FirstPregnancy +
    s(AgeMother) + s(Country, bs = "re")

Parametric coefficients:
                                  Estimate Std. Error z value
(Intercept)                     -0.218762   1.102108  -0.198
FirstPregnancyPrimiparous        0.975222   1.063317   0.917
CovidExposureYes                -0.025110   0.073501  -0.342
total_deaths                     0.058809   0.025847   2.275
FirstPregnancyMultiparous:CHI_mean  0.005582   0.008430   0.662
FirstPregnancyPrimiparous:CHI_mean -0.001990   0.007438  -0.268
FirstPregnancyMultiparous:IHDI     -0.047487   1.112231  -0.043
FirstPregnancyPrimiparous:IHDI     -0.681849   1.101041  -0.619
                                  Pr(>|z|)
(Intercept)                         0.8427
FirstPregnancyPrimiparous           0.3591
CovidExposureYes                    0.7326
total_deaths                        0.0229 *
FirstPregnancyMultiparous:CHI_mean   0.5078
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
FirstPregnancyPrimiparous:CHI_mean   0.7891
FirstPregnancyMultiparous:IHDI       0.9659
FirstPregnancyPrimiparous:IHDI       0.5357
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
               edf Ref.df Chi.sq  p-value
s(AgeMother) 3.991  4.959  5.293    0.333
s(Country)   6.572 10.000 42.271 3.16e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

sigma.1 = 0.674(0.629,0.72)  sigma.2 = 0.428(0.398,0.456)
theta = 2.29(2.03,2.59)  tau = 0.553(0.498,0.603)
n = 3305  total edf = 57.4
```

`Pr(>|z|)` indicates de p-value for the linear covariates. The results displayed in `EQUATION 1` and `EQUATION 2` correspond with the modelisation of the mean of the marginals, i.e. `GAD7` and `EPDS`. As the marginals follow a negative binomial distribution, by doing the exponential of the estimated coefficients we can compute their effect in terms of a ratio between the two categories (for the binary covariates) or the increment in just one unit in the case of the continuous ones. This can be done very easily with the following code lines:

```
table <- round(as.data.frame((summary(fit_p)$tableP1)), 3)
r <- NULL
for(i in 1:dim(summary(fit_p)$tableP1)[1]){
  est <- round(exp(table$Estimate[i]), 3)
  low <- round(exp(table$Estimate[i] -
                     qnorm(1 - 0.05/2)*table$‘Std. Error‘[i]), 3)
  up <- round(exp(table$Estimate[i] +
                    qnorm(1 - 0.05/2)*table$‘Std. Error‘[i]), 3)
  r <- rbind(r, paste0(est, " (", inf, ", ", sup, ")"))
}
```

Here, `(summary(fit_p)$tableP1)` is the table with the numerical results for the linear covariates in `EQUATION 1`, `est` is the exponential of the estimates, and `low` and `up` are the lower and upper bounds of the confidence interval for the ratio. By doing this, we obtain a table with the ratio and the corresponding 95% CI for each variable:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
     [,1]
[1,] "28.789 (7.159, 115.766)"
[2,] "0.374 (0.125, 1.124)"
[3,] "1.121 (1.036, 1.212)"
[4,] "1.029 (0.998, 1.062)"
[5,] "0.992 (0.984, 1)"
[6,] "0.996 (0.988, 1.004)"
[7,] "0.201 (0.042, 0.959)"
[8,] "0.531 (0.112, 2.512)"
```

If we go to the third row, which corresponds to `CovidExposure`, we can see that the ratio between the category `Yes` and the category `No` is 1.121, meaning that having been exposed to COVID-19 increases `GAD7` in a 12.1% ($(1.121-1)*100$). The analogous can be done for `EPDS` if we replace `tableP1` for `tableP2`. Note that since the transformation in `EQUATION 5` is not exactly `log`, it makes no sense to apply the exponential function there.

With the function `plot`, we can see the effect of the smooth effect of Age on the scale of the linear predictor (see the left panel of Figure 6.6 for an example for the marginal of `GAD7`). If we want it in the scale of the response (for the mean of the marginals) or as Kendall's τ (for the correlation), we need to use the function `pred.mvt` (see right panel of Figure 6.6 for an example for the marginal of `GAD7`).

Using the `plot` function:

```
op <- par(mfrow = c(1, 2))
plot(fit_p, eq = 1, scale = 0, shade = T, select = 1, se = T,
     xlab = "Age of the mother (years)", main = "Linear predictor")
abline(h = 0, col = 2)
```

Using the `pred.mvt` function, note that we need to create a new data frame with all covariates as constants except for `AgeMother`.

```
newdata <- data.frame(GAD7 = pregnant$GAD7, EPDS = pregnant$EPDS,
                      FirstPregnancy = "Multiparous",
                      CovidExposure = "No",
                      total_deaths = mean(pregnant$total_deaths),
                      CHI_mean = mean(pregnant$CHI_mean),
                      IHDI = unique(pregnant$IHDI[pregnant$Country
                                    == "Albania"]),
                      AgeMother = pregnant$AgeMother,
                      Country = "Albania")

pred <- pred.mvt(fit_p, eq = 1, fun = "mean", newdata = newdata)
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
dataset <- data.frame(mean = pred$pred, lci = pred$CIpred[, 1],
                      uci = pred$CIpred[, 2],
                      Age = pregnant$AgeMother)

dataset <- dataset[with(dataset, order(Age)), ]
plot(dataset$mean ~ dataset$Age, col = "black", type = "l",
     ylim = c(4, 11.5), xlab = "Age of the mother (years)",
     ylab = "GAD-7 raw score", main = "Response")
polygon(x = c(dataset$Age, rev(dataset$Age)),
        y = c(dataset$lci, rev(dataset$uci)),
        col = "lightgrey", border = NA)
lines(dataset$mean ~ dataset$Age, col = "black", type = "l")
```

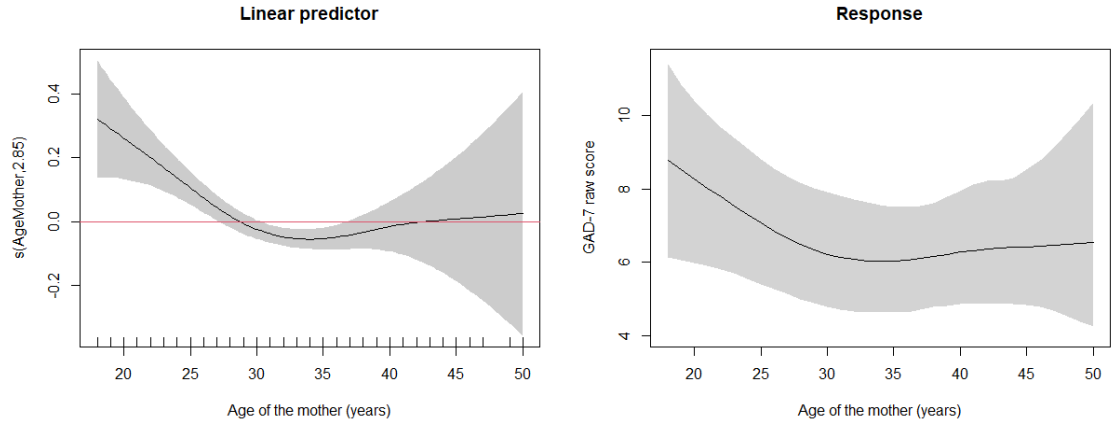

Figure 6.6: Smooth effect of `AgeMother` on `GAD7` according to model `fit_p`. Left: Estimated smooth effect on the linear predictor. Right: Predicted smooth effect on the response.

It should be noted that confidence bands are wider in the scale os the response, since it is a prediction. For `EPDS` we need to change to `eq = 2`, and for the correlation, in addition to `eq = 5`, the argument `fun = "mean"` has to be changed for `fun = "tau"`.

Finally, if we want to graphically represent the random intercepts, we have to use the `predict` function with the argument `type = "terms"`, and then choose those corresponding to our random variable, in our case `'s(Country)'` (note that we should write it as in the output of the summary of the model). If we state `se = T`, we can also extract their standard errors.

USC DE SANTIAGO DE COMPOSTELA

```
ri <- as.data.frame(predict(fit_p, eq = 1, type = "terms"))$
  's(Country)'
se <-  as.data.frame(predict(fit_p, eq = 1, type = "terms",
                             se = T)$se.fit)$'s(Country)'
```

Then, we just match the random effects with their corresponding countries and eliminate duplicities.

```
nd <- data.frame(Country = as.factor(pregnant$Country),
                 intercept = ri, se = se)
nd <- nd[!duplicated(nd$Country), ]
```

To plot the random effects, we recommend use $1:12$ (we have 12 countries) as x variable instead of `nd$Country` because, otherwise, it plots horizontal segments instead of points. Since we do not use `nd$Country` in the plot, we just add `xaxt = "n"` to remove the marks on the x axis and add the ones we want with the function `axis`. To do the bars for the confident intervals, we use the function `arrows`. The resulting plot is shown in Figure 6.7.

```
par(mfrow = c(1, 1))
plot(1:12, nd$intercept, ylim = c(-0.45, 0.3), xlab = "Country",
     ylab = "Estimated effect", pch =  16, xaxt = "n")
arrows(x0 = 1:12, y0 = nd$intercept - nd$se, x1 = 1:12,
       y1 = nd$intercept + nd$se, code = 3, angle = 90,
       length = 0.1, col = rainbow(12), lwd = 2)
axis(1, at = 1:12, labels = nd$Country)
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

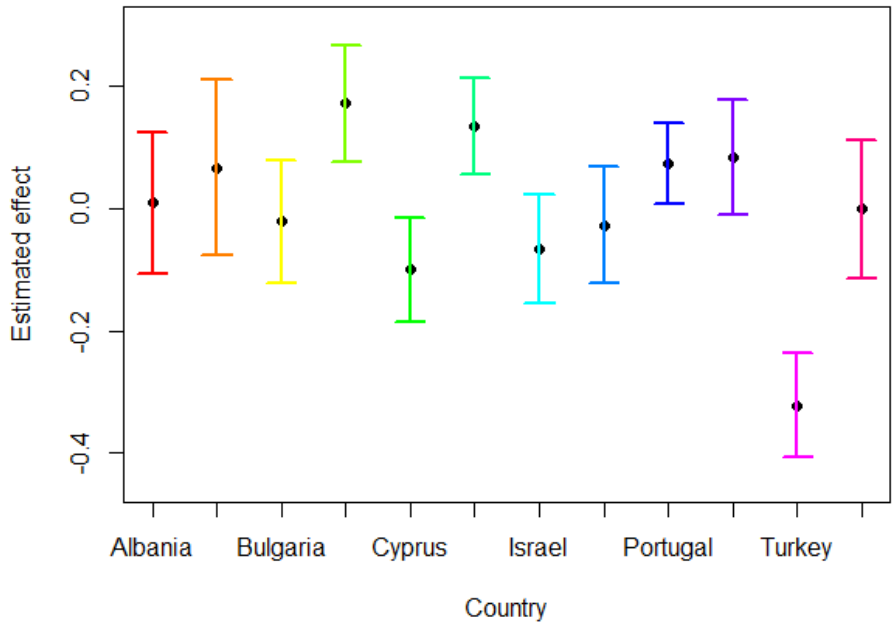

Figure 6.7: Random effects of `Country` on GAD7, with 95% Confidence Intervals.

6.3 Code for Section 4.4: "Trivariate Joint Modelling of TRH Data"

In this section, we present the basic code we used in the analysis of the TRH database discussed in Section 4.4 from Chapter 4. Before starting, we show a small sample of our database:

```
      TSH   T3   T4 Age
1  0.5306283 3.19 1.22  47
2  1.5496879 4.19 1.62  43
3  1.0367369 3.30 1.15  65
4  0.2070142 4.42 1.32  56
5 -0.4942963 3.35 1.33  72
6  0.7323679 3.48 1.34  67
```

At the time of this study, there was not yet a package in R to carry it out. The developers of the methodology themselves have provided us with the necessary

USC DE SANTIAGO DE COMPOSTELA

functions for the implementation of these methods. Before we begin, we will run these functions, which we will use later. We highly recommend to save this first block of code in a separate `R` script named, for instance `mmlt.R`.

```
library("mlt")
library("alabama")
library("Matrix")
library("sandwich")
library("numDeriv")

mmlt <- function(by, bx = NULL, todistr = "Normal", data,
                 coef = NULL, ...) {
  todistr <- match.arg(todistr)
  stopifnot(todistr == "Normal")
  J <- length(by)

  dofit <- is.null(coef)

  mctm <- vector(mode = "list", length = J)
  mmlt <- vector(mode = "list", length = J)
  mctm[[1]] <- ctm(by[[1]], shift = bx, todistr = todistr)
  mmlt[[1]] <- mlt(mctm[[1]], data = data, dofit = dofit, ...)
  if (!is.null(coef))
    coef(mmlt[[1]]) <- coef[[1]]
  pdat <- data
  htotal <- "~ 1"

  for (j in 2:J) {
    hhat <- paste("hhat", j - 1, sep = "_")
    htotal <- c(htotal, hhat)
    if (todistr == "Normal") {
      data[[hhat]] <- predict(mmlt[[j - 1]], newdata = pdat,
                              type = "trafo")
    } else {
      ### FIXME: we need to multiply with sigma_j NOT known here!
      ### </FIXME>
      data[[hhat]] <- qnorm(predict(mmlt[[j - 1]], newdata = pdat,
                                    type = "distribution"))
    }
    pdat[[hhat]] <- 0
    bhi <- as.basis(as.formula(paste(htotal, collapse = "+")),
                    data = data, remove_intercept = TRUE)
    if (!is.null(bx)) {
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
      bh <- as.basis(as.formula(paste(htotal, collapse = "+")),
                     data = data)
      mctm[[j]] <- ctm(by[[j]],
                       shift = c(bhi = bhi,
                                 bb = b(bh = bh, bx = bx)),
                       todistr = todistr)
    } else {
      mctm[[j]] <- ctm(by[[j]], shift = bhi, todistr = todistr)
    }
    mmlt[[j]] <- mlt(mctm[[j]], data = data, dofit = dofit, ...)
    if (!is.null(coef))
      coef(mmlt[[j]]) <- coef[[j]]
  }
  ret <- list(mctm = mctm, mmlt = mmlt, by = by, bx = bx,
              data = data, todistr = todistr)
  class(ret) <- "mmlt"
  ret
}


refit <- function(object, quiet = TRUE, Hessian = TRUE, ...) {
  cf <- coef(object)
  i <- cumsum(sapply(cf, length))
  idx <- cbind(c(1, i[-length(i)] + 1), i)
  ui <- do.call("bdiag", lapply(object$mmlt, function(m)
    attr(model.matrix(m$model,m$data), "constraint")$ui))
  ci <- do.call("c", lapply(object$mmlt, function(m)
    attr(model.matrix(m$model,m$data), "constraint")$ci))
  ui <- as(ui[is.finite(ci),,drop = FALSE], "matrix")
  ci <- ci[is.finite(ci)]
  r0 <- rowSums(abs(ui)) == 0
  ui <- ui[!r0,,drop = FALSE]
  ci <- ci[!r0]

  umod <- function(beta) {
    tmpcf <- lapply(1:nrow(idx), function(i)
      beta[idx[i,1]:idx[i,2]])
    mmlt(object$by, object$bx, todistr = object$todistr,
         data = object$data, coef = tmpcf)
  }

  f <- function(beta)
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

```
    -logLik(umod(beta))
  g <- function(beta)
    do.call("c", lapply(estfun(umod(beta)), colSums))

  o <- alabama::auglag(par = unlist(cf), fn = f, gr = g,
                       hin = function(par) ui %*% par - ci,
                       hin.jac = function(par) ui,
                       control.outer = list(...))

  ret <- umod(o$par)
  if (Hessian)
      ret$numHessian <- hessian(f, x = unlist(cf))
  ret
}

coef.mmlt <- function(object)
  lapply(object$mmlt, coef)


logLik.mmlt <- function(object) {
  ll <- lapply(object$mmlt, logLik)
  ret <- do.call("sum", ll)
  return(ret)
  df <- sum(sapply(ll, function(l) attr(l, "df")))
  attr(ret, "df") <- df
  class(ret) <- "logLik"
  ret
}

### add derivatives of conditioning terms
estfun.mmlt <- function(object) {
  ef <- lapply(object$mmlt, function(m) estfun(m))
  xm <- lapply(object$mmlt, function(m)
    c(m$score(coef(m, fixed = TRUE), weights(m), Xmult = FALSE)))
  cf <- coef(object)
  for (j in 1:(length(object$mmlt) - 1)) {
    X <- model.matrix(object$mmlt[[j]]$model, data = object$data)
    hcf <- grep("hhat", colnames(X))
    wh <- 1:ncol(X)
    if (length(hcf) > 0)
      wh <- wh[-hcf]
    for (i in (j + 1):length(object$mmlt))
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

```
      ef[[j]][, wh] <- ef[[j]][, wh] -
      X[, wh] * xm[[i]] * cf[[i]][paste("hhat", j, sep = "_")]
  }
  ef
}

### gives the conditional covariance matrix from the model
vcov.mmlt <- function(object, newdata = data.frame(1),
                      precchol = TRUE) {
  J <- length(object$mmlt)
  savedata <- newdata
  ret <- matrix(0, nrow = nrow(newdata), ncol = J * (J - 1) / 2)
  jtmp <- 1
  for (j in 2:J) {
    hhat <- paste("hhat", 1:(j - 1), sep = "_")
    tmp <- diag(length(hhat))
    colnames(tmp) <- hhat
    retj <- vector(mode = "list", length = NROW(tmp))
    newdata[, hhat] <- 0
    yval <- mkgrid(object$mmlt[[j]]$model, n = 1)[[1]][1]
    newdata[[object$mmlt[[j]]$model$response]] <- yval
    sh <- predict(object$mmlt[[j]], newdata = newdata,
                  terms = "bshift", type = "trafo")
    for (i in 1:NROW(tmp)) {
      for (h in hhat) newdata[[h]] <- tmp[i,h]
      ret[, jtmp] <- predict(object$mmlt[[j]], newdata = newdata,
                             terms = "bshift", type = "trafo") - sh
      jtmp <- jtmp + 1
    }
  }
  if (precchol)
    return(cbind(savedata, ret))
  lapply(1:nrow(ret), function(i) {
    A <- diag(J)
    A[lower.tri(A)] <- ret[i,]
    Ainv <- solve(A)
    return(tcrossprod(Ainv))
  })
}
```

Now we can begin with the analysis. Flexible estimation is based on Bernstein polynomial basis. Before going any further, we need to define the response variables in terms of Bernstein basis.

USC DE SANTIAGO DE COMPOSTELA

```
By <- lapply(c("TSH","T3","T4"), function(y) {
  v <- numeric_var(y, support = quantile(dat[[y]],
                                         prob = c(.1, .9)),
                   bounds = c(-Inf, Inf))
  Bernstein_basis(var = v, order = 8, ui = "increasing")
})
```

According to Klein et al. (2019), developers of the methodology, the order of the basis should be between 5 and 10. We chose `order = 8` for that reason. The above function uses `lapply`, allowing to define the Bernstein basis of the outcomes all at once. We then define the predictor variable `Age` in terms of Bernstein basis.

```
Bx <- Bernstein_basis(
  numeric_var("Age",
              support = quantile(dat$Age, prob = c(.1, .9)),
              bounds = c(0, 100)), order = 3, ui = "zero")
```

Note that for the covariate, there is no need for such an order of the bases (we established `order = 3`). We would also like to point out that for the outcomes it is mandatory that the coefficients of the bases go in increasing order (indicated in the code as `ui = "increasing"`), while this restriction is not necessary for the explicative variable (`ui = "zero"`).

To fit the model, we use the function `mmlt`.

```
m = mmlt(By,bx = Bx, data = dat)
m = refit(m, trace = T, checkGrad = F, triter = 1, maxit = 100)
```

Refitting the model using `refit` is recommended to obtain a better estimation, but is not compulsory, and it takes quite a while.

Here is the code to obtain the marginal results for each outcome. First, we need to establish for which values of the covariate `Age` we want to calculate the marginal functions. Then, we estimate the variance-covariance matrix.

```
ageseq <- seq(20, 90, by = 10)
aest2 <- vcov(m, newdata = data.frame(Age = ageseq), prec = F)
```

We then calculate the marginal conditional distribution functions for our outcomes and store their values in `p1s`, `p2s` and `p3s`.

```
p1s = matrix(NA, nrow = 100, ncol = length(aest2))
p2s = p1s
p3s = p1s
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
q1 <- seq(from = min(dat[[1]]), to = max(dat[[1]]),
length.out = 100)
q2 <- seq(from = min(dat[[2]]), to = max(dat[[2]]),
length.out = 100)
q3 <- seq(from = min(dat[[3]]), to = max(dat[[3]]),
length.out = 100)

for(i in 1:length(aest2)){
  sd <- sqrt(diag(aest2[[i]]))
  p1s[,i] <- pnorm(predict(m$mmlt[[1]],
                           newdata = data.frame(1,
                                                Age = ageseq[i]),
                           q = q1, type = "trafo") / sd[1])
  p2s[,i] <- pnorm(predict(m$mmlt[[2]],
                           newdata = data.frame(1, hhat_1 = 0,
                                                Age = ageseq[i]),
                           q = q2, type = "trafo") / sd[2])
  p3s[,i] <- pnorm(predict(m$mmlt[[3]],
                           newdata = data.frame(1, hhat_1 = 0,
                                                hhat_2 = 0,
                                                Age = ageseq[i]),
                           q = q3, type = "trafo") / sd[3])
}
```

Now, the results for the marginal conditional density functions, stored in `d1s`, `d2s` and `d3s`.

```
d1s = d2s = d3s = p1s
for(i in 1:length(aest2)){
  sd <- sqrt(diag(aest2[[i]]))
  d1s[,i] <- dnorm(predict(m$mmlt[[1]],
                           newdata = data.frame(1,
                                                Age = ageseq[i]),
                           q = q1, type = "trafo") / sd[1])
  d2s[,i] <- dnorm(predict(m$mmlt[[2]],
                           newdata = data.frame(1, hhat_1 = 0,
                                                Age = ageseq[i]),
                           q = q2, type = "trafo") / sd[2])
  d3s[,i] <- dnorm(predict(m$mmlt[[3]],
                           newdata = data.frame(1, hhat_1 = 0,
                                                hhat_2 = 0,
                                                Age = ageseq[i]),
                           q = q3, type = "trafo") / sd[3])}
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

In order to plot the estimates of the marginals, we first need to establish a colour scale, as we are going to do this for different age ranges, stored in `cols`:

```
library("colorspace")
library("fields")
cols = (rainbow_hcl(length(aest2), start = 0, end = 300))
```

With the following code, we can graphically represent the cumulative distribution function (CDF) for the outcomes for different ages of the patients, as shown on top in Figure 4.5 from Chapter 4.

```
par(mfrow = c(1, 3), mar = c(5.5, 6.5, 3.5, 1.9) - 1)
plot(q1, p1s[, 1], type = "l",
     xlab = expression(paste(y[log-TSH], sep = "")),
     ylab = expression(paste("F(", y[log-TSH],"|Age)", sep = "")),
     col = cols[1], cex.lab = 2.5, cex.axis = 2, lwd = 2)
for(i in 2:length(aest2)){
  lines(q1, p1s[, i], col = cols[i], lwd = 2)
}

plot(q2, p2s[, 1], type = "l",
     xlab = expression(paste(y[T3], sep = "")),
     ylab = expression(paste("F(", y[T3], "|Age)", sep = "")),
     col = cols[1], cex.lab = 2.5, cex.axis = 2, lwd = 2)
for(i in 2:length(aest2)){
  lines(q2, p2s[, i], col = cols[i], lwd = 2)
}

plot(q3, p3s[, 1], type = "l",
     xlab = expression(paste(y[T4], sep = "")),
     ylab = expression(paste("F(", y[T4], "|Age)", sep = "")),
     col = cols[1], cex.lab = 2.5, cex.axis = 2, lwd = 2)
for(i in 2:length(aest2)){
  lines(q3, p3s[, i], col = cols[i], lwd = 2)
}

colorbar.plot(1.6, 0.25, 1:length(aest2), strip.width = 0.1,
              zrange = NULL, adj.x = 0.5, adj.y = 0.5,
              col = cols[length(aest2):1], horizontal = FALSE)
text(1.6, 0.4, "20 years")
text(1.6, 0.1, "90 years")
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

`colorbar.plot` function is used to create the age colour bar. In a similar way, the conditional density functions of the outcomes for different ages can be graphically represented, as shown in the lower part of Figure 4.5 from Chapter 4.

```
par(mfrow = c(1, 3), mar = c(5.5, 6.5, 3.5, 1.9) - 1)
plot(q1,d1s[,1],type="l",lwd = 2,
     xlab=expression(paste(y[log-TSH],sep="")),
     ylab=expression(paste("f(",y[log-TSH],"|Age)",sep="")),
     col=cols[1],cex.lab=2.5,cex.axis=2)
for(i in 2:length(aest2)){
  lines(q1,d1s[,i],col=cols[i], lwd = 2)
}


plot(q2,d2s[,1],type="l",lwd = 2,
     xlab=expression(paste(y[T3],sep="")),
     ylab=expression(paste("f(",y[T3],"|Age)",sep="")),
     col=cols[1],cex.lab=2.5,cex.axis=2)
for(i in 2:length(aest2)){
  lines(q2,d2s[,i],col=cols[i], lwd = 2)
}


plot(q3,d3s[,1],type="l",lwd = 2,
     xlab=expression(paste(y[T4],sep="")),
     ylab=expression(paste("f(",y[T4],"|Age)",sep="")),
     col=cols[1],cex.lab=2.5,cex.axis=2)
for(i in 2:length(aest2)){
  lines(q3,d3s[,i],col=cols[i], lwd = 2)
}

colorbar.plot(1.7,0.25,1:length(aest2), strip.width = 0.1,
              zrange = NULL, adj.x = 0.5, adj.y = 0.5,
              col = cols[length(aest2):1], horizontal = FALSE)
text(1.65,0.19,"80 years",cex=1.5)
text(1.65,0.31,"20 years",cex=1.5)
```

Next, the scatter plot of thyroid-related hormones vs. `Age`, with the percentile curves, as in Figure 4.6 from Chapter 4. We need the function `aspline` from package for the computation of the percentile curves.

```
library(akima)
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
par(mfrow = c(1, 3), mar = c(5.5, 6.5, 3.5, 1.9) - 1)
cuants = numeric()
for(i in 1:length(aest2)){
  cuants = rbind(cuants,
                 aspline(x = p1s[, i], y = q1,
                         xout = seq(0.10, 0.90, 0.10))$y)
}
cuants
plot(dat$TSH ~ dat$Age, col = "grey", cex.lab = 2.5, cex.axis=2,
     cex.main = 2,main = "log-TSH", xlab = "Age (years)",
     ylab = "log-TSH levels")
for(i in 1:dim(cuants)[2]){
  lines(cuants[, i] ~ ageseq, type = "l", col = 2, lwd = 2)
}

# Second plot
cuants = numeric()
for(i in 1:length(aest2))
  cuants = rbind(cuants,
                 aspline(x=p2s[,i],y=q2,
                         xout=seq(0.10,0.90,0.10))$y)
cuants
plot(dat$T3~dat$Age,col="grey",cex.lab=2.5,cex.axis=2,cex.main=2,
     main="Free T3", xlab="Age (years)",
     ylab="Free T3 levels (pmol/L)")
for(i in 1:dim(cuants)[2])
  lines(cuants[,i]~ageseq,type="l",col=2,lwd=2)

# Third plot
cuants = numeric()
for(i in 1:length(aest2))
  cuants = rbind(cuants,aspline(x=p3s[,i],y=q3,
                                xout=c(seq(0.10,0.90,0.10)))$y)
cuants
plot(dat$T4~dat$Age,col="grey",cex.lab=2.5,cex.axis=2,cex.main=2,
     main="Free T4", xlab="Age, years",
     ylab="Free T4 levels (ng/dL)")
for(i in 1:dim(cuants)[2]) lines(cuants[,i]~ageseq,type="l",col=2,
                                 ylim=c(70,200),lwd=2)
```

Finally, we present the code for the graphical representation of the correlation, as in Figure 4.7 from Chapter 4. As we already explained in that chapter, the computation of the correlation is immediate, but if we want to obtain confidence bands, we

USC DE SANTIAGO DE COMPOSTELA

need to apply bootstrap resampling. In this part we show the code we used for that. To save time, we used packages `doParallel` and `foreach`, and thus be able to parallelise the process. In `makeCluster(12)`, the number 12 indicates the number of cores we want to use.

```
library(doParallel)
library(foreach)
cl = makeCluster(12)
registerDoParallel(cl)

r = foreach(i=1:1000,.errorhandling = "pass")%dopar%{
  source("mmlt.R")
  library(mvtnorm)
  norig=length(dat$TSH)
  ageseq = seq(min(dat$Age),max(dat$Age),by=1)

  nsamp = length(ageseq)
  r12s=matrix(NA,nsamp,length(ageseq))
  r13s=r12s
  r23s=r12s
  aest2 <- (vcov(m, newdata = data.frame(Age = ageseq),prec=F))

  yi <- matrix(0, nrow = norig, ncol = 3)
  xcage <- dat$Age

    for(i in 1:norig)
  {
    il=which(xcage[i]==ageseq)
    xi=data.frame(Age=ageseq[il])
    resid <- (rmvnorm(1,mean=c(0,0,0),sigma=aest2[[il]]))
    q1 <- seq(from = min(dat[[1]]), to = max(dat[[1]]),
              length.out = 1000)
    q2 <- seq(from = min(dat[[2]]), to = max(dat[[2]]),
              length.out = 1000)
    q3 <- seq(from = min(dat[[3]]), to = max(dat[[3]]),
              length.out = 1000)
    t1 <- predict(m$mmlt[[1]],
                  newdata = data.frame(1,Age=ageseq[il]),
                  q = q1, type = "trafo")
    t2 <- predict(m$mmlt[[2]],
                  newdata = data.frame(1, hhat_1 = 0,
                                       Age=ageseq[il]),
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
                  q = q2, type = "trafo")
    t3 <- predict(m$mmlt[[3]],
                  newdata = data.frame(1, hhat_1 = 0,
                                       hhat_2 = 0, Age=ageseq[il]),
                  q = q3, çtype = "trafo")

yi[i, ] <- c(q1[which.min((resid[1] - t1)^2)],
                 q2[which.min((resid[2] - t2)^2)],
                 q3[which.min((resid[3] - t3)^2)])
  }

  dati = dat
  dati[[1]]=yi[,1]
  dati[[2]]=yi[,2]
  dati[[3]]=yi[,3]

  mmi <- mmlt(By, bx = Bx, data = dati)


  a3 <- (vcov(mmi, newdata = data.frame(Age = ageseq),prec=F))
  for(i in 1:length(ageseq))
  {
    a3[[i]] = cov2cor(a3[[i]])
  }
  r12=a3[[1]][1,2]
  r13=a3[[1]][1,3]
  r23=a3[[1]][2,3]
  for(k in 2:length(ageseq))
  {
    r12=c(r12,a3[[k]][1,2])
    r13=c(r13,a3[[k]][1,3])
    r23=c(r23,a3[[k]][2,3])
  }

  cbind(r12,r13,r23)
}
```

Now that we have already done the bootstrap resampling, we can plot the two-by-two correlations as a function of `Age`, as shown in Figure 4.7 from Chapter 4.

```
r12s=do.call(cbind,lapply(r,function(x) x[,1]))
r13s=do.call(cbind,lapply(r,function(x) x[,2]))
r23s=do.call(cbind,lapply(r,function(x) x[,3]))
```

USC DE SANTIAGO DE COMPOSTELA

```
ageseq = seq(min(dat$Age),max(dat$Age),by=1)
par(mfrow=c(1,3),mar=c(5.5,6.5,3.5,1.5)-1)
plot(ageseq, apply(r12s,MARGIN=1,FUN="mean"), type="l",lwd=2,
     ylab=expression(paste(rho[21],"(Age)",sep="")),cex.axis=1.75,
     cex.lab=2,cex.main=2,ylim=c(-0.6,0.6),col=NA,
     xlab="Age (years)",main="log-TSH and Free T3")
polygon(c(ageseq, rev(ageseq)),
        c(apply(r12s,MARGIN=1,FUN=function(x){
          quantile(x,prob=0.025)}),
          rev(apply(r12s,MARGIN=1,FUN=function(x){
            quantile(x,prob=0.975)}))),
        col = "azure3", border = NA)
lines(ageseq, apply(r12s,MARGIN=1,FUN="mean"), type="l",lwd=2,
      xlab="age",ylab=expression(paste(rho[21],"(age)",sep="")),
      cex.axis=1.75,cex.lab=2,cex.main=2,ylim=c(-0.6,0.6))
rug(dat$Age)
abline(h=0,col=2,lwd=2)

# Plot 2
plot(ageseq, apply(r13s,MARGIN=1,FUN="mean"), type="l",lwd=2,
     ylab=expression(paste(rho[13],"(Age)",sep="")),
     cex.axis=1.75,cex.lab=2,cex.main=2,ylim=c(-0.6,0.6),col=NA,
     xlab="Age (years)",main="log-TSH and Free T4")
polygon(c(ageseq, rev(ageseq)),
        c(apply(r13s,MARGIN=1,FUN=function(x){
          quantile(x,prob=0.025)}),
          rev(apply(r13s,MARGIN=1,FUN=function(x){
            quantile(x,prob=0.975)}))),
        col = "azure3", border = NA)
lines(ageseq, apply(r13s,MARGIN=1,FUN="mean"), type="l",lwd=2,
      xlab="age",ylab=expression(paste(rho[13],"(age)",sep="")),
      cex.axis=1.75,cex.lab=2,cex.main=2,ylim=c(-0.2,0.6))
rug(dat$Age)
abline(h=0,col=2,lwd=2)

# Plot 3
plot(ageseq, apply(r23s,MARGIN=1,FUN="mean"), type="l",lwd=2,
     ylab=expression(paste(rho[23],"(Age)",sep="")),
     cex.axis=1.75,cex.lab=2,cex.main=2,ylim=c(-0.6,0.6),col=NA,
     xlab="Age (years)",main="Free T3 and Free T4")
polygon(c(ageseq, rev(ageseq)),
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

```
        c(apply(r23s,MARGIN=1,FUN=function(x){
          quantile(x,prob=0.025)}),
          rev(apply(r23s,MARGIN=1,FUN=function(x){
            quantile(x,prob=0.975)}))),
        col = "azure3", border = NA)
lines(ageseq, apply(r23s,MARGIN=1,FUN="mean"),lwd=2)
rug(dat$Age)
abline(h=0,col=2,lwd=2)
```

6.4 Code for Section 5.4: "JMLS Regression of Liver-19 Data"

In the last section of this chapter, we present the basic code we used in the analysis of the Liver-19 database discussed in Section 5.4 from Chapter 5. As we did with the other three applications, we start by taking a quick look at the data `dat`.

```
    id      age gender exitus   Exitus     time_l   time_s
2    1 77.57377   Male      1 Deceased 0.01944444 6.437500
6    1 77.57377   Male      1 Deceased 2.80763889 6.437500
736  2 71.71311   Male      1 Deceased 0.09375000 6.528472
737  2 71.71311   Male      1 Deceased 0.97986111 6.528472
740  2 71.71311   Male      1 Deceased 2.96666667 6.528472
741  2 71.71311   Male      1 Deceased 3.94166667 6.528472

    diabetes   AST_trans  ALT_trans  GGT_trans
2        Yes  0.05212067 -0.8952217 -0.4010644
6        Yes  0.29993959 -0.5714548 -0.9550756
736      Yes -0.75847563 -2.4801284 -2.1715167
737      Yes -1.01116470 -2.7193496 -2.7504089
740      Yes -0.79300488 -3.1474640 -3.1881345
741      Yes -0.47049957 -3.4871435 -2.0727406
```

We can see that here we have already the Box-Cox transformed transaminases. If this is not the case, simply apply the Box-Cox transformation as explained in Section 6.1 from this chapter. In this database `Exitus` is the event indicator (`Deceased` or `Discharged`), and `exitus` just indicates the same, but in a numerical way (1 for `Deceased` and 0 for `Discharged`). Also, we can see two different time variables: `time_l` indicates the time moment of the longitudinal observations, whereas `time_s` indicates the time of the survival process, the time-to-event.

For this kind of analysis, we need a second database, containing only 1 observation per individual. One way of constructing it is the following:

USC DE SANTIAGO DE COMPOSTELA

```
dat_id <- dat[!duplicated(dat$id), ]
```

This way, we are removing all the observations per patient but the first one. This new database looks as follows:

```
    id      age gender exitus     Exitus     time_l    time_s
2     1 77.57377   Male      1   Deceased 0.01944444  6.437500
736   2 71.71311   Male      1   Deceased 0.09375000  6.528472
1536  4 64.33060   Male      0 Discharged 0.23750000 10.174306
1648  6 81.01644 Female      0 Discharged 0.00000000 11.643750
1737  7 84.57650 Female      1   Deceased 0.07083333  2.209028
1753  8 64.59290 Female      0 Discharged 0.05902778 10.761111

     diabetes   AST_trans  ALT_trans  GGT_trans
2         Yes  0.05212067 -0.8952217 -0.4010644
736       Yes -0.75847563 -2.4801284 -2.1715167
1536       No  0.21394470  0.1546846 -0.9550756
1648       No -0.14926147 -1.0039524 -0.4167237
1737       No -0.11693384 -1.3021376 -1.1510192
1753       No -0.57478222 -1.2352271 -0.2057895
```

To fit a JMLS with the methodology explained in Chapter 5, we need to install and load `JMbayes2` package (Rizopoulos et al., 2022). In turn, to install this package, we need the package `remotes`:

```
remotes::install_github("drizopoulos/jmbayes2")
library(JMbayes2)
```

In order to fit a JMLS, we need, on the one hand, the survival sub-model. This model is to be fitted using the reduced database. Please note that no longitudinal variables are included in this sub-model.

```
CoxFit <- coxph(Surv(time_s, exitus == 1) ~ gender + ns(age, 2) +
                  diabetes, data = dat_id, model = T)
```

With `ns(age, 2)` we are indicating that variable `age` is considered non-linear, using natural cubic splines with 2 degrees of freedom.

Now is the turn of the longitudinal sub-models. For each transformed transaminase, we fit a sub-model with the same structure (it is not compulsory). `bs(time_l, knots = 15, degree = 1)` indicates that we are using linear B-splines to create a piece-wise function for the longitudinal time, so that it behaves in a certain way for the first 15 days, and in a different way thereafter.

USC DE SANTIAGO DE COMPOSTELA

```
fm1 <- lme(AST_trans ~ bs(time_l, knots = 15, degree = 1),
           data = dat, random = ~ time_l | id)

fm2 <- lme(ALT_trans ~ bs(time_l, knots = 15, degree = 1),
           data = dat, random = ~ time_l | id)

fm3 <- lme(GGT_trans ~ bs(time_l, knots = 15, degree = 1),
           data = dat, random = ~ time_l | id)
```

Once the sub-models have been established, the JMLS regression model can be adjusted using the function `jm`. In this function, `time_var` indicates the longitudinal times. We have set very high parameters for the Bayesian algorithm (`n_iter`, `n_burnin` and `n_thin`) to ensure the convergence of the model, but this inevitably increases the fitting time of the model.

```
joint_trans <- jm(CoxFit, list(fm1, fm2, fm3), time_var = "time_l",
                  n_iter = 55000L, n_burnin = 5000L, n_thin = 50L,
                  n_chains = 4)
```

There are several ways to assess whether the model has converged. First, we can check if the so-called effective size is greater than 50 in any case.

```
> joint_trans$statistics$Effective_Size

$bs_gammas
 bs_gammas_1  bs_gammas_2  bs_gammas_3  bs_gammas_4  bs_gammas_5
   3546.6145    2564.9700    1777.6730    1417.5403    1048.4797
 bs_gammas_6  bs_gammas_7  bs_gammas_8  bs_gammas_9 bs_gammas_10
    451.6182     238.5197     148.3984     126.7650     143.6313
bs_gammas_11
    163.2591

$tau_bs_gammas
tau_bs_gammas
     1164.111

$gammas
 genderMale ns(age, 2)1 ns(age, 2)2 diabetesYes
   3677.836    2490.557    2454.270    2583.043

$alphas
value(AST_trans) value(ALT_trans) value(GGT_trans)
```

USC DE SANTIAGO DE COMPOSTELA

```
        2922.380         2744.139         3251.299

$W_std_gammas
[1] 2425.229

$Wlong_std_alphas
[1] 2573.397

$D
 D[1, 1]  D[2, 1]  D[3, 1]  D[4, 1]  D[5, 1]  D[6, 1]  D[2, 2]
2944.943 1523.484 2659.683 1585.791 2962.540 2634.556 1597.632
 D[3, 2]  D[4, 2]  D[5, 2]  D[6, 2]  D[3, 3]  D[4, 3]  D[5, 3]
1781.711 1688.063 2199.664 2347.125 3306.882 1940.126 3096.627
 D[6, 3]  D[4, 4]  D[5, 4]  D[6, 4]  D[5, 5]  D[6, 5]  D[6, 6]
2958.279 2031.758 2488.674 2610.811 2635.870 2902.877 1929.959

$betas1
                          (Intercept)
                             3650.108
bs(time_l, knots = 15, degree = 1)1
                             1585.240
bs(time_l, knots = 15, degree = 1)2
                             1466.781

$betas2
                          (Intercept)
                              4000.0000
bs(time_l, knots = 15, degree = 1)1
                              1296.2216
bs(time_l, knots = 15, degree = 1)2
                              833.8217

$betas3
                          (Intercept)
                             1780.598
bs(time_l, knots = 15, degree = 1)1
                             1389.922
bs(time_l, knots = 15, degree = 1)2
                              639.433

$sigmas
sigmas_1 sigmas_2 sigmas_3
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
3535.195 3817.657 4000.000
```

We also need to check that all potential scale reduction factors ($R - hat$) were between 0.90 and 1.05.

```
> gelman_diag(joint_trans)

$betas1
Potential scale reduction factors:

                                    Point est. Upper C.I.
(Intercept)                              0.999          1
bs(time_l, knots = 15, degree = 1)1      1.000          1
bs(time_l, knots = 15, degree = 1)2      1.000          1

Multivariate psrf

1

$betas2
Potential scale reduction factors:

                                    Point est. Upper C.I.
(Intercept)                                  1       1.00
bs(time_l, knots = 15, degree = 1)1          1       1.00
bs(time_l, knots = 15, degree = 1)2          1       1.01

Multivariate psrf

1

$betas3
Potential scale reduction factors:

                                    Point est. Upper C.I.
(Intercept)                                  1          1
bs(time_l, knots = 15, degree = 1)1          1          1
bs(time_l, knots = 15, degree = 1)2          1          1

Multivariate psrf

1
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
$sigmas
Potential scale reduction factors:

         Point est. Upper C.I.
sigmas_1          1       1.00
sigmas_2          1       1.00
sigmas_3          1       1.01

Multivariate psrf

1

$D
Potential scale reduction factors:

        Point est. Upper C.I.
D[1, 1]      1.001      1.003
D[2, 1]      1.006      1.021
D[3, 1]      1.002      1.007
D[4, 1]      1.010      1.032
D[5, 1]      1.003      1.010
D[6, 1]      1.002      1.007
D[2, 2]      1.007      1.022
D[3, 2]      1.008      1.024
D[4, 2]      1.008      1.024
D[5, 2]      1.005      1.016
D[6, 2]      1.003      1.010
D[3, 3]      1.003      1.008
D[4, 3]      1.008      1.027
D[5, 3]      1.001      1.005
D[6, 3]      1.001      1.004
D[4, 4]      1.007      1.019
D[5, 4]      1.008      1.027
D[6, 4]      1.003      1.011
D[5, 5]      1.005      1.016
D[6, 5]      1.002      1.008
D[6, 6]      0.999      0.999

Multivariate psrf

1.02
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
$bs_gammas
Potential scale reduction factors:

             Point est. Upper C.I.
bs_gammas_1        1.00       1.00
bs_gammas_2        1.00       1.00
bs_gammas_3        1.00       1.00
bs_gammas_4        1.00       1.00
bs_gammas_5        1.00       1.00
bs_gammas_6        1.00       1.01
bs_gammas_7        1.01       1.03
bs_gammas_8        1.03       1.05
bs_gammas_9        1.03       1.06
bs_gammas_10       1.03       1.05
bs_gammas_11       1.03       1.05

Multivariate psrf

1.01

$tau_bs_gammas
Potential scale reduction factors:

              Point est. Upper C.I.
tau_bs_gammas       1.01       1.02


$gammas
Potential scale reduction factors:

            Point est. Upper C.I.
genderMale           1       1.01
ns(age, 2)1          1       1.00
ns(age, 2)2          1       1.00
diabetesYes          1       1.00

Multivariate psrf

1

$alphas
Potential scale reduction factors:
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
                  Point est. Upper C.I.
value(AST_trans)          1          1
value(ALT_trans)          1          1
value(GGT_trans)          1          1

Multivariate psrf

1
```

We can also asses convergence graphically, through `ggtraceplot` function. These graphs represent the generated MCMC chains. In Figure 6.8 we can see, as an example, the traceplot for the 4 chains generated for the estimation of the β parameter for the first part of the effect of the longitudinal time on `AST_trans`. There are no differences between curves, and the curves seem to oscillate around a fixed value (-0.2) at all times. This is an indicator of convergence.

```
ggtraceplot(joint_trans, parm = "betas",
            custom_theme = c("1" = "#DD3C51", "2" = "#313657",
                             "3" = "#1F6683", "4" = "#6C90B9"))
```

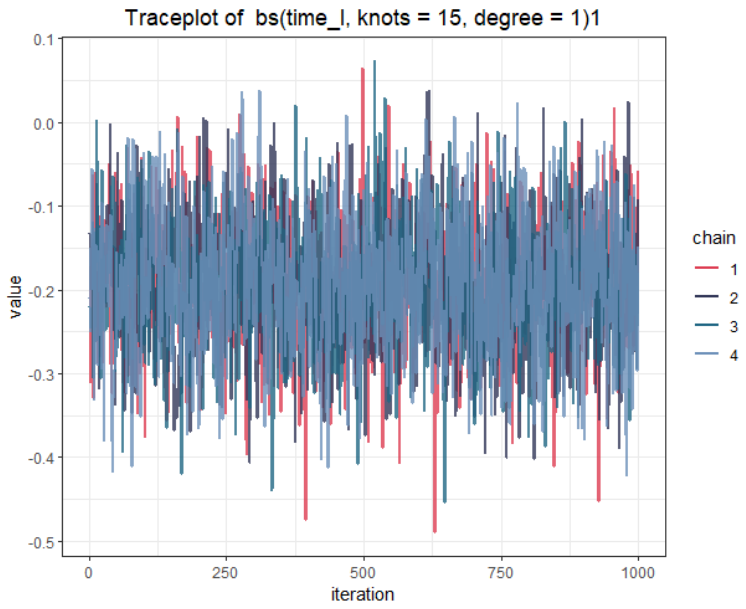

Figure 6.8: Traceplot for the 4 chains generated for the estimation of the β parameter for the first part of the effect of the longitudinal time on `AST_trans`.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Now that we are sure that the model has converge, so it is trustable, we can take a look at its summary to see the numerical results for the model:

```
> summary(joint_trans)

Call:
jm(Surv_object = CoxFit, Mixed_objects = list(fm1, fm2, fm3),
    time_var = "time_l", n_iter = 55000L, n_burnin = 5000L,
    n_thin = 50L,  n_chains = 4)

Data Descriptives:
Number of Groups: 205 Number of events: 36 (17.6%)
Number of Observations:
  AST_trans: 858
  ALT_trans: 858
  GGT_trans: 858

                 DIC     WAIC      LPML
marginal    4315.762 8866.655 -6511.786
conditional 1982.090 2516.807 -4484.722

Random-effects covariance matrix:

       StdDev   Corr
(Intr) 0.4784 (Intr)  time_l  (Intr)  time_l  (Intr)
time_l 0.0557 -0.4405
(Intr) 0.6327 0.7652  -0.4959
time_l 0.0653 -0.2711 0.8505  -0.3755
(Intr) 0.8509 0.4200  -0.3206 0.5344  -0.3395
time_l 0.0536 0.0939  0.3775  -0.0724 0.4563  -0.3070

Survival Outcome:
                     Mean  StDev    2.5%   97.5%      P   Rhat
genderMale         0.4645 0.5662 -0.5902  1.5790 0.4310 1.0023
ns(age, 2)1        1.3366 3.4809 -4.5155  8.7524 0.7515 1.0006
ns(age, 2)2        3.6453 0.9252  1.9183  5.5252 0.0000 1.0005
diabetesYes        1.2267 0.4094  0.4379  2.0507 0.0025 1.0007
value(AST_trans)   2.0299 0.6204  0.8428  3.2888 0.0015 1.0002
value(ALT_trans) -1.3680 0.3850 -2.1299 -0.6227 0.0005 1.0001
value(GGT_trans) -0.0401 0.2927 -0.6310  0.5303 0.9010 1.0007

Longitudinal Outcome: AST_trans (family = gaussian,
                                 link = identity)
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
                   Mean  StDev    2.5%   97.5%      P   Rhat
(Intercept)    -0.2125 0.0498 -0.3106 -0.1154 0.0000 0.9994
b(_,k=15,d=1)1 -0.1935 0.0778 -0.3448 -0.0413 0.0135 0.9998
b(_,k=15,d=1)2 -0.4985 0.3663 -1.2320  0.1963 0.1630 1.0004
sigma           0.3219 0.0099  0.3034  0.3421 0.0000 0.9998

Longitudinal Outcome: ALT_trans (family = gaussian,
                                 link = identity)
                   Mean  StDev    2.5%   97.5%      P   Rhat
(Intercept)    -0.4812 0.0599 -0.6001 -0.3645 0.0000 1.0003
b(_,k=15,d=1)1  0.7937 0.1067  0.5815  0.9993 0.0000 1.0004
b(_,k=15,d=1)2 -0.1227 0.4924 -1.0816  0.8326 0.8095 1.0010
sigma           0.3187 0.0108  0.2990  0.3408 0.0000 1.0006

Longitudinal Outcome: GGT_trans (family = gaussian,
                                 link = identity)
                   Mean  StDev    2.5%   97.5%      P   Rhat
(Intercept)    -0.5768 0.2032 -0.9956 -0.1922 0.0045 1.0009
b(_,k=15,d=1)1  0.4381 0.1241  0.1874  0.6808 0.0010 0.9996
b(_,k=15,d=1)2 -2.5888 0.6127 -3.7888 -1.3901 0.0000 1.0012
sigma           0.2861 0.0182  0.2573  0.3301 0.0000 1.0021

MCMC summary:
chains: 4
iterations per chain: 55000
burn-in per chain: 5000
thinning: 50
time: 16.7 min
```

From this output, `Random-effects covariance matrix:` is the matrix Σ linking all the random effects. We can see that only the lower triangular part of the matrix is shown, but this is because the matrix is symmetrical. After that, we have the summary for the `Survival Outcome`. `value(AST_trans)` indicates that the whole estimation of the linear predictor `AST_trans` is being plugged into the survival model. Then, we find the summary for each `Longitudinal Outcome`. Finally, we have the specifications for the MCMC algorithm and the fitting time.

The code for estimating and plotting the dynamic predictions is shown below. First, we have to establish the time horizon to which we want to predict `tmax`, nd the time up to which we have information `t0`.

```
tmax <- 30
t0 <- 15
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Then, we make a subset with the data of the individual were interested in and we stick to the information up to time `t0`. We also create a new event variable `status <- 0`, since we know that this patient was alive at `t0`. Finally, we establish `t0` as the new survival time, since we are supposing that this is the maximum follow-up time that we have on the patient.

```
ND <- dat[dat$id == 180, ]
ND <- ND[ND$time_l< t0, ]
ND$status2 <- 0
ND$time_s <- t0
```

Now, we have to make predictions for the longitudinal and the survival part, using function `predict`.

```
predLong <- predict(joint_trans, newdata = ND,
                    times = seq(t0, tmax, length.out = 10),
                    return_newdata = TRUE)
predSurv <- predict(joint_trans, newdata = ND, process = "event",
                    times = seq(t0, tmax, length.out = 10),
                    return_newdata = TRUE)
```

Once the predictions are made, it only remains to plot them. With the code below we obtain, for instance, the upper right graph of Figure 5.1 from Chapter 5.

```
plot(predLong, predSurv, outcomes = 1:3, subject = id,
     fun_event = function (x) 1 - x,
     ylab_event = "Survival Probabilities",
     ylab_long = c("AST_trans", "ALT_trans", "GGT_trans"),
     col_points = c("#313657", "#1F6683", "#6C90B9"),
     col_line_long = c("#313657", "#1F6683", "#6C90B9"),
     col_line_event = "#DD3C51",
     fill_CI_long = c("#31365780", "#1F668380", "#6C90B980"),
     fill_CI_event = "#DD3C5180",
     pos_ylab_long = c(-0.5, -0.5, -0.5),
     main = paste("Id ", id, ", Time = ", t0, " days"),
     cex_main = 1.5,
     cex_ylab_long = 1.5,
     cex_ylab_event = 1.5)
```

ROC curve with information up to day `tstart` and predictions for `dt` days after can be computed very easily with the function `tvROC`. As an illustration, lets say that we have to assess the accuracy of the fits of day 15 with information up to 7 days:

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
tstart <- 7
dt <- 8
roc <- tvROC(joint_trans, newdata = dat, Tstart = tstart, Dt = dt,
             add_density = T)
plot(roc, main = paste0("Day ",  tstart + dt,
                        " with information from day ", tstart))
```

Figure 6.9 shows the ROC curve computed with the code above:

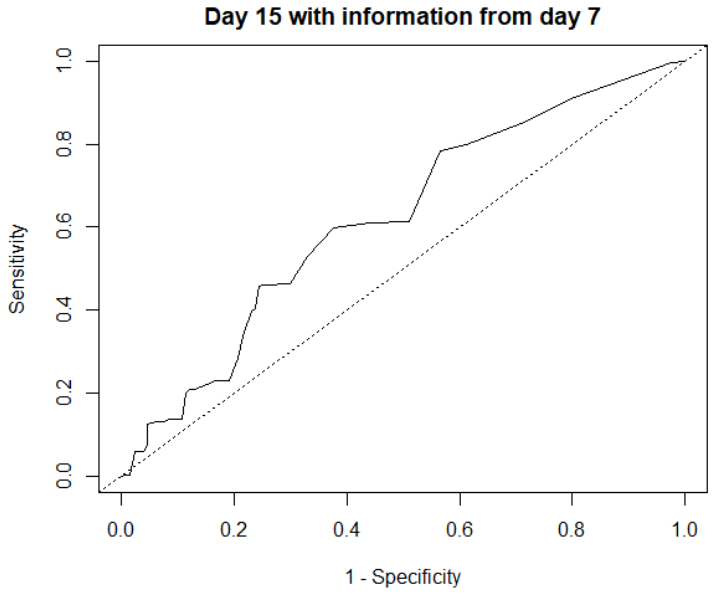

Figure 6.9: ROC curve for day 15 with information up to day 7.

Finally, AUC can be computed from a JMLS model or from a ROC curve, using function `tvAUC`. The AUC for the ROC curve in Figure 6.9 is:

```
> tvAUC(roc)

Time-dependent AUC for the Joint Model joint_trans

Estimated AUC: 0.6217
At time: 15
Using information up to time: 7 (143 subjects still at risk)
```

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Chapter 7

General Conclusions and Future Research

The main objective of this thesis is the use of modern and novel joint modelling (JM) regression techniques for multiple response variables in fields of medicine and, in particular, epidemiology where, until now, they had not been applied and their use could make a big difference. We hope that both the applications of these techniques to real data and the provision of the necessary code for their programming will help, at least a little, to increase the use of these techniques in biomedicine. In this work, the application of JM techniques addresses epidemiological important situations: the COVID-19 pandemic on the one hand, and thyroid-related hormones on the other. In Chapter 2, the databases used were introduced, as well as the research questions that each of them was intended to answer. Different JM techniques were applied, depending on the question to be answered and the nature of the data.

In this chapter, we present a summary of the main results obtained in this thesis (Chapters 3-6) together with some conclusions and interesting lines for future research.

7.1 Chapter 3: "Copula Regression Models for Bivariate Responses"

In univariate (one response) regression analysis, the emergence of Generalised Additive Models (GAM; Hastie and Tibshirani, 1990b), which replace the linear predictor in GLM models by an additive one, was a breakthrough due to the great flexibility of the models compared to the alternatives that had existed up to that time. 25 years later, Rigby and Stasinopoulos (2005) introduced the Generalised Additive Models for Location, Scale and Shape (GAMLSS), focusing on distributional

USC DE SANTIAGO DE COMPOSTELA

regression, since it is not only a question of modelling the mean but also each of the parameters of the distribution, such as variance. In addition, these models increased the variety of possible outcome distributions. In Chapter 3, Bivariate Copula Generalised Additive Models for Location Scale and Shape (CGAMLSS; Marra and Radice, 2017) are explained. These models extend the GAMLSS models in a frequentist framework to the case of a bivariate response. Two important biomedical contributions are made in this chapter through the application of these techniques to our data.

Liver damage in COVID-19 patients admitted to hospital. We were able to establish a pattern of evolution of transaminases over time, with an increase in transaminase values during the first few days, that began to reverse after the second week, with a greater recovery in the patients who finally survived. The determination of this pattern, hitherto unestablished in the clinical literature, was made possible by fitting GAMM models. These models allow for the inclusion of random effect variables in the GAM framework. But the truly remarkable results in this section are those obtained in the study of the correlation between transaminases and inflammation markers (IM). Applying CGAMLSS regression models, we were able to find significant correlation between transaminases over time, especially after 15 days. This is particularly striking considering that, with the COVID-19 strain existing in the first wave (when the data were collected), it has been established that, in most cases, the direct action of the virus only took place in the first 15 days. These results suggest that the liver damage in COVID-19 patients is therefore due to the inflammatory response of the organism, although due to the limitations of the study this cannot be categorically stated.

Future work in this field needs to be done. As this was a retrospective study, with data taken at a time of hospital saturation, some variables of interest were not collected correctly and could therefore not be included in the study (e.g. body mass index). In addition, the small number of patients limits the analyses. Data has continued to be collected, so that the database has grown beyond the first wave of the disease. This, together with the fact that other variables of interest in the study of liver damage (such as ferritin) were also collected, leads us to, in the near future, the extension of the study, both in number of patients and in possible response variables.

Perinatal mental health during COVID-19 pandemic. We were able to study the importance of governmental responses to the COVID-19 pandemic on perinatal mental health. The study found that a higher hardness of these measures plays a protective role for both anxiety and depressive symptoms. Since anxiety and

USC DE SANTIAGO DE COMPOSTELA

depression are two closely related conditions, it is not unreasonable to think that analysing them together could lead to more accurate results. This led us to apply CGAMLSS regression models, which also allowed us to study the correlation between the two responses as a function of some covariates of interest. In doing so, we were able to determine that not only the levels of anxiety and depression are affected by the pandemic situation, but also the relationship between them.

In this study we have limited ourselves to a handful of variables chosen at the discretion of the psychologists and psychiatrists involved. But the database we have is huge, so there are many other variables that may be of interest in the study of perinatal mental health that have not yet been studied. Related to this, future work will focus on determining risk factors for both anxiety and perinatal depression in times of pandemic. There are many known risk factors, such as age (Curry et al., 2019) or the mother's previous history of mental health problems (Kinser et al., 2018), but there are new factors that need to be assessed. In this line, we are already carrying out the first works in which we apply the latest boosting techniques for the selection of variables, both univariate (i.e. a single response), with the methodology proposed by Strömer et al. (2022), and bivariate with Boosting Distributional Copula Regression models (Hans et al., 2022).

7.1.1 Future work concerning CGAMLSS

It is noteworthy that, although we have longitudinal data in the COVID-19 study, we have not treated them that way in the bivariate regression (despite the fact that this methodology allows for it, as we have seen in the study of perinatal mental health). This is due to the large number of random effects that would have to be introduced in each sub-model, one per patient (in our case 212), which greatly impairs fit and performance. Exploring how random effects of this type (with more groups than observations per group) could be incorporated is of great interest to us. Given that most follow-up studies in biomedicine are of this type, this may be a new step towards a more accurate analysis of these data.

7.2 Chapter 4: "An Alternative Approach to Copula JM: Multivariate Conditional Transformation Models (MCTM)"

USC DE SANTIAGO DE COMPOSTELA

In Chapter 4 we modelled jointly, for the first time, the age effect on the concentrations of the three thyroid-related hormones, offering a better understanding of these

endocrinological measurements. This has been made possible by the implementation of MCTM regression models, introduced in this chapter. MCTM allow for a joint estimation of the covariates effects on the distribution of each response, and on their correlations. MCTM performance was evaluated in comparison with the frequentist conditional copula regression models.

Thyroid dysfunction diagnosis have been the object of research in several works. However, given the state of art of regression analysis, these studies have been conducted considering thyroid hormone concentrations as mutually independent variables. For instance, Jonklaas and Razvi (2019) focused on the clinical variables effect over the thyroid hormones distribution, with independence of any pathological status. Hoermann et al. (2016) discussed the convenience of estimating a trivariate reference region for TSH, T3, and T4. Indeed, they proved that univariate reference curves for each hormone cause thyroid dysfunction overdiagnosis. Therefore, trivariate regression modelling might be advantageous for thyroid dysfunction diagnosis. In this context, the proposal of trivariate reference regions - i.e. regions characterizing the 95% of healthy patients' results - are desirable in the clinical setting. In future work, we will investigate the MCTM extension for conditional trivariate reference region estimation.

7.3 Chapter 5: "Joint Modelling for Longitudinal and Survival Data (JMLS)"

In Chapter 5 we studied the effect of three longitudinal transaminase variables (AST, ALT and GGT) on the survival of patients with COVID-19, adjusting also for age, gender and diabetes mellitus. We have seen that risk for death increases significantly as AST increases, and ALT decreases. GGT showed no significant effect on survival. We have also seen that diabetes mellitus is positively linked to the risk for death.

Chapter 5 is devoted to joint modelling for longitudinal and survival data, or JMLS, which are the regression models that have allowed us to carry out the analysis described above.. We introduced these models to the reader, both the basic model, for just one continuous longitudinal covariate and a single time-to-event, and the latest extension to it, which allows for several possibly non-normal longitudinal biomarkers and also competing risks or multi state models in the survival part. We also introduced the dynamic predictions, one of the major strengths of these models, since the predictions of the survival probabilities can be updated with any new information collected. We also discussed the time-dependent ROC curve as a measure for assessing model discrimination.

USC DE SANTIAGO DE COMPOSTELA

As expected in this pandemic situation, the survival of patients with COVID-19 has been the subject of many studies (see, for instance, Li et al., 2020b; Shu et al., 2021; Hueso et al., 2022), but to the best of our knowledge, Tong-Minh et al. (2022) were the only ones to apply JMLS regression models, using them to assess the impact on survival of 4 different inflammation markers. Some of these are included in our database Liver-19: CRP and IL-6. As future work we would also like to apply JMLS to include markers of inflammation as longitudinal covariates, but so far we have had many convergence problems, we believe due to the high correlation between them. That is why we must continue to work on this to try to overcome these problems.

In general, the discriminative ability of the models, as measured by ROC curves and AUC, are good tools for assessing the performance of the model, and it is also a very useful technique for comparing different models. But we have seen that, depending on the width of the prediction interval used, AUC can have an undesirable behaviour due to its volatility and lack of continuity in some cases. This problem leads us to want to study AUC in depth in the context of JMLS. In particular, we believe that a smoothed version of the AUC curve, which would give an idea of its general trend, could be very useful, and we have therefore already started to investigate it.

7.4 Chapter 6: "Software and Programming Code"

Chapter 6 consists of a tutorial with a shortened version of the code used in our analyses, in order to facilitate the implementation of JM methodologies in clinical practice, which is the ultimate goal of this thesis.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

References

Acar, E., Craiu, R., and Yao, F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7:2822–2850.

Alsefri, M., Sudell, M., García-Fiñana, M., and Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Medical Research Methodology*, 20(1):1–17.

Anderson, E. J., Rouphael, N. G., Widge, A. T., Jackson, L. A., Roberts, P. C., Makhene, M., Chappell, J. D., Denison, M. R., Stevens, L. J., Pruijssers, A. J., et al. (2020). Safety and immunogenicity of SARS-CoV-2 mRNA-1273 vaccine in older adults. *New England Journal of Medicine*, 383(25):2427–2438.

Arentz, M., Yim, E., Klaff, L., Lokhandwala, S., Riedo, F. X., Chong, M., and Lee, M. (2020). Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State. *JAMA: The Journal of the American Medical Association*, 323(16):1612–1614.

Bäuerle, A., Steinbach, J., Schweda, A., Beckord, J., Hetkamp, M., Weismüller, B., Kohler, H., Musche, V., Dörrie, N., Teufel, M., et al. (2020). Mental health burden of the COVID-19 outbreak in Germany: predictors of mental health impairment. *Journal of Primary Care & Community Health*, 11:2150132720953682.

Beardselee, W. R., Versage, E. M., and Giadstone, T. R. (1998). Children of affectively ill parents: a review of the past 10 years. *Journal of the American Academy of Child & Adolescent Psychiatry*, 37(11):1134–1141.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015). BayesX: Software for Bayesian inference in structured additive regression models. Version 3.0.2. Available online on http://www.BayesX.org/.

Bernstein, S. (1912). Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Communications of the Kharkov Mathematical Society*, 13(1):1–2.

USC DE SANTIAGO DE COMPOSTELA

Bloom, P. P., Meyerowitz, E. A., Reinus, Z., Daidone, M., Gustafson, J., Kim, A. Y., Schaefer, E., and Chung, R. T. (2021). Liver biochemistries in hospitalized patients with COVID-19. *Hepatology*, 73(3):890–900.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Broman, K. W. and Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1):2–10.

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., and Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The lancet*, 395(10227):912–920.

Cai, Q., Huang, D., Yu, H., Zhu, Z., Xia, Z., Su, Y., Li, Z., Zhou, G., Gou, J., Qu, J., Sun, Y., Liu, Y., He, Q., Chen, J., Liu, L., and Xu, L. (2020). COVID-19: Abnormal liver function tests. *Journal of Hepatology*, 73(3):566–574.

Chalasani, N. P., Hayashi, P. H., Bonkovsky, H. L., Navarro, V. J., Lee, W. M., Fontana, R. J., and the Practice Parameters Committee of the American College of Gastroenterology (2014). ACG Clinical Guideline: the diagnosis and management of idiosyncratic drug-induced liver injury. *Official Journal of the American College of Gastroenterology—ACG*, 109(7):950–966.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, 395(10223):507–513.

Chmielewska, B., Barratt, I., Townsend, R., Kalafat, E., van der Meulen, J., Gurol-Urganci, I., O'Brien, P., Morris, E., Draycott, T., Thangaratinam, S., et al. (2021). Effects of the COVID-19 pandemic on maternal and perinatal outcomes: a systematic review and meta-analysis. *The Lancet Global Health*, 9(6):e759–e772.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*. Chapman and Hall/CRC.

Cox, J. L., Holden, J. M., and Sagovsky, R. (1987). Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale. *The British Journal of Psychiatry*, 150(6):782–786.

USC DE SANTIAGO DE COMPOSTELA

Curry, S. J., Krist, A. H., Owens, D. K., Barry, M. J., Caughey, A. B., Davidson, K. W., Doubeni, C. A., Epling, J. W., Grossman, D. C., Kemper, A. R., and US Preventive Services Task Force (2019). Interventions to prevent perinatal depression: US Preventive Services Task Force recommendation statement. *JAMA: The Journal of the American Medical Association*, 321(6):580–587.

Davenport, T. H. and Patil, D. (2012). Data scientist. *Harvard Business Review*, 90(5):70–76.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag.

De Curtis, M., Villani, L., and Polo, A. (2021). Increase of stillbirth and decrease of late preterm infants during the COVID-19 pandemic lockdown. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 106(4):456–456.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Deslandes, E. and Chevret, S. (2007). Assessing surrogacy from the joint modelling of multivariate longitudinal data and survival: application to clinical trial data on chronic lymphocytic leukaemia. *Statistics in Medicine*, 26(30):5411–5421.

Díaz-Louzao, C., Barrera-Lopez, L., Lopez-Rodriguez, M., Casar, C., Vazquez-Agra, N., Pernas-Pardavila, H., Marques-Afonso, A., Vidal-Vazquez, M., Montoya, J. G., Andrade, A. H., Fernandez-Castro, I., Varela, P., Gonzalez-Quintela, A., Otero, E., Gude, F., Cadarso-Suarez, C., and Tome, S. (2022). Longitudinal relationship of liver injury with inflammation biomarkers in COVID-19 hospitalized patients using a joint modeling approach. *Scientific Reports*, 12(1):1–13.

Díaz-Louzao, C., Lado-Baleato, Ó., Gude, F., and Cadarso-Suárez, C. (2021). Multivariate conditional transformation models. application to thyroid-related hormones. In *International Conference on Computational Science and Its Applications*, pages 341–351. Springer.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Dong, L. and Bouey, J. (2020). Public mental health crisis during COVID-19 pandemic, China. *Emerging Infectious Diseases*, 26(7):1616.

Duarte, E., de Sousa, B., Cadarso-Suárez, C., Espasandín-Domínguez, J., Lado-Baleato, O., Marra, G., Radice, R., and Rodrigues, V. (2017). Applying spatial copula additive regression to breast cancer screening data. In *International Conference on Computational Science and Its Applications*, pages 586–599. Springer.

USC DE SANTIAGO DE COMPOSTELA

Effati-Daryani, F., Zarei, S., Mohammadi, A., Hemmati, E., Ghasemi Yngyknd, S., and Mirghafourvand, M. (2020). Depression, stress, anxiety and their predictors in Iranian pregnant women during the outbreak of COVID-19. *BMC Psychology*, 8(1):1–10.

Espasandín Domínguez, J. (2019). *Contributions to distributional regression models. Applications in biomedicine*. PhD thesis, Universidade de Santiago de Compostela.

Espasandín-Domínguez, J., Cadarso-Suárez, C., Kneib, T., Marra, G., Klein, N., Radice, R., Lado-Baleato, O., González-Quintela, A., and Gude, F. (2019). Assessing the relationship between markers of glycemic control through flexible copula regression models. *Statistics in Medicine*, 38(27):5161–5181.

Espasandín-Domínguez, J., Carollo-Limeres, C., Coladas-Uría, L., Cadarso-Suárez, C., Lado-Baleato, O., and Gude, F. (2018). Bivariate copula additive models for location, scale and shape with applications in biomedicine. In *The Mathematics of the Uncertain*, pages 135–146. Springer.

Fahrmeir, L., Kneib, T., et al. (2011). Bayesian smoothing and regression for longitudinal, spatial and event history data. *Oxford University Press Catalogue*.

Fan, S., Guan, J., Cao, L., Wang, M., Zhao, H., Chen, L., and Yan, L. (2021). Psychological effects caused by COVID-19 pandemic on pregnant women: a systematic review with meta-analysis. *Asian Journal of Psychiatry*, 56:102533.

Farkhad, B. F. and Albarracín, D. (2021). Insights on the implications of COVID-19 mitigation measures for mental health. *Economics & Human Biology*, 40:100963.

Fawcett, E. J., Fairbrother, N., Cox, M. L., White, I. R., and Fawcett, J. M. (2019). The prevalence of anxiety disorders during pregnancy and the postpartum period: a multivariate Bayesian meta-analysis. *The Journal of clinical psychiatry*, 80(4):1181.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society Series*, 98:39–54.

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, 55(5):1919–1932.

Global Change Data Lab (n.d.). https://ourworldindata.org/. Accessed: 01-06-2021.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Goodman, S. H., Rouse, M. H., Connell, A. M., Broth, M. R., Hall, C. M., and Heyward, D. (2011). Maternal depression and child psychopathology: a meta-analytic review. *Clinical Child and Family Psychology Review*, 14(1):1–27.

Grasselli, G., Zangrillo, A., Zanella, A., Antonelli, M., Cabrini, L., Castelli, A., Cereda, D., Coluccello, A., Foti, G., Fumagalli, R., Iotti, G., Latronico, N., Lorini, L., Merler, S., Natalini, G., Piatti, A., Ranieri, M. V., Scandroglio, A. M., Storti, E., Cecconi, M., Pesenti, A., and COVID-19 Lombardy ICU Network (2020). Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy. *JAMA: The Journal of the American Medical Association*, 323(16):1574–1581.

Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., Liu, L., Shan, H., Lei, C. L., Hui, D. S., Du, B., Li, L. J., Zeng, G., Yuen, K. Y., Chen, R. C., Tang, C. L., Wang, T., Chen, P. Y., Xiang, J., Li, S. Y., Wang, J. L., Liang, Z. J., Peng, Y. X., Wei, L., Liu, Y., Hu, Y. H., Peng, P., Wang, J. M., Liu, J. Y., Chen, Z., Li, G., Zheng, Z. J., Qiu, S. Q., Luo, J., Ye, C. J., Zhu, S. Y., and Zhong, N. S. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*, 382(18):1708–1720.

Gude, F., Díaz-Vidal, P., Rúa-Pérez, C., Alonso-Sampedro, M., Fernández-Merino, C., Rey-García, J., Cadarso-Suárez, C., Pazos-Couselo, M., García-López, J. M., and Gonzalez-Quintela, A. (2017). Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of Diabetes Science and Technology*, 11(4):780–790.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16–24.

Hadlow, N. C., Rothacker, K. M., Wardrop, R., Brown, S. J., Lim, E. M., and Walsh, J. P. (2013). The relationship between TSH and free T4 in a large population is complex and nonlinear and differs by age and sex. *The Journal of Clinical Endocrinology & Metabolism*, 98(7):2936–2943.

Hale, T., Angrist, N., Kira, B., Petherick, A., Phillips, T., and Webster, S. (2020). Variation in government responses to COVID-19.

Hans, N., Klein, N., Faschingbauer, F., Schneider, M., and Mayr, A. (2022). Boosting distributional copula regression. *arXiv preprint arXiv:2202.12851*.

Harville, E., Xiong, X., and Buekens, P. (2010). Disasters and perinatal health: a systematic review. *Obstetrical & Gynecological Survey*, 65(11):713.

USC DE SANTIAGO DE COMPOSTELA

Hastie, T. and Tibshirani, R. (1990a). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46:1005–1016.

Hastie, T. and Tibshirani, R. (1990b). *Generalized additive models*. Chapman and Hall/CRC.

Hay, J. W., Gong, C. L., Jiao, X., Zawadzki, N. K., Zawadzki, R. S., Pickard, A. S., Xie, F., Crawford, S. A., and Gu, N. Y. (2021). A US population health survey on the impact of COVID-19 using the EQ-5D-5L. *Journal of General Internal Medicine*, 36(5):1292–1301.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Herold, T., Jurinovic, V., Arnreich, C., Lipworth, B. J., Hellmuth, J. C., von Bergwelt-Baildon, M., Klein, M., and Weinberger, T. (2020). Elevated levels of IL-6 and CRP predict the need for mechanical ventilation in COVID-19. *Journal of Allergy and Clinical Immunology*, 146(1):128–136.

Hoermann, R., Eckl, W., Hoermann, C., and Larisch, R. (2010). Complex relationship between free thyroxine and TSH in the regulation of thyroid function. *European Journal of Endocrinology*, 162(6):1123–1129.

Hoermann, R., Larisch, R., Dietrich, J. W., and Midgley, J. E. (2016). Derivation of a multivariate reference range for pituitary thyrotropin and thyroid hormones: diagnostic efficiency compared with conventional single-reference method. *European journal of endocrinology*, 174(6):735–743.

Hothorn, T., Möst, L., and Bühlmann, P. (2018). Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134.

Hoyt, R. and Muenchen, R. (2019). *Introduction to Biomedical Data Science*. Lulu.com.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, china. *The Lancet*, 395(10223):497–506.

Hueso, T., Godron, A. S., Lanoy, E., Pacanowski, J., Levi, L. I., Gras, E., Surgers, L., Guemriche, A., Meynard, J. L., Pirenne, F., Idri, S., Tiberghien, P., Morel, P., Besson, C., Duléry, R., Lamure, S., Hermine, O., Gagneux-Brunon, A., Freymond, N., Grabar, S., and Lacombe, K. (2022). Convalescent plasma improves overall survival in patients with B-cell lymphoid malignancy and COVID-19: a longitudinal cohort and propensity score analysis. *Leukemia*, 36(4):1025–1034.

USC DE SANTIAGO DE COMPOSTELA

Hundt, M. A., Deng, Y., Ciarleglio, M. M., Nathanson, M. H., and Lim, J. K. (2020). Abnormal liver tests in COVID-19: a retrospective observational cohort study of 1,827 patients in a major US hospital network. *Hepatology*, 72(4):1169–1176.

Iyengar, U., Jaiprakash, B., Haitsuka, H., and Kim, S. (2021). One year into the pandemic: a systematic review of perinatal mental health outcomes during COVID-19. *Frontiers in Psychiatry*, page 845.

Jonklaas, J. and Razvi, S. (2019). Reference intervals in the diagnosis of thyroid dysfunction: treating patients not numbers. *The Lancet Diabetes & Endocrinology*, 7(6):473–483.

Jothimani, D., Venugopal, R., Abedin, M. F., Kaliamoorthy, I., and Rela, M. (2020). COVID-19 and the liver. *Journal of Hepatology*, 73(5):1231–1240.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kinser, P. A., Thacker, L. R., Lapato, D., Wagner, S., Roberson-Nay, R., Jobe-Shields, L., Amstadter, A., and York, T. P. (2018). Depressive symptom prevalence and predictors in the first half of pregnancy. *Journal of Women's Health*, 27(3):369–376.

Kiplimo, R., Kosgei, M., Mwangi, A., Onyango, E., Ogero, M., and Koske, J. (2021). Longitudinal-survival models for case-based tuberculosis progression. *Frontiers in Public Health*, 9:543750.

Klein, N., Hothorn, T., Barbanti, L., and Kneib, T. (2019). Multivariate conditional transformation models. *Scandinavian Journal of Statistics*.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, 26(4):841–860.

Klein, N., Kneib, T., and Lang, S. (2015a). Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110(509):405–419.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015b). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, 9(2):1024–1052.

Köhler, M., Umlauf, N., Beyerlein, A., Winkler, C., Ziegler, A. G., and Greven, S. (2017). Flexible Bayesian additive joint models with an application to type 1 diabetes research. *Biometrical Journal*, 59(6):1114–1165.

USC DE SANTIAGO DE COMPOSTELA

Kucharski, A. J., Klepac, P., Conlan, A. J. K., Kissler, S. M., Tang, M. L., Fry, H., Gog, J. R., Edmunds, W. J., Emery, J. C., Medley, G., Munday, J. D., Russell, T. W., Leclerc, Q. J., Diamond, C., Procter, S. R., Gimma, A., Sun, F. Y., Gibbs, H. P., Rosello, A., van Zandvoort, K., Hué, S., Meakin, S. R., Deol, A. K., Knight, G., Jombart, T., Foss, A. M., Bosse, N. I., Atkins, K. E., Quilty, B. J., Lowe, R., Prem, K., Flasche, S., Pearson, C. A. B., Houben, R. M. G. J., Nightingale, E. S., Endo, A., Tully, D. C., Liu, Y., Villabona-Arenas, J., O'Reilly, K., Funk, S., Eggo, R. M., Jit, M., Rees, E. M., Hellewell, J., Clifford, S., Jarvis, C. I., Abbott, S., Auzenbergs, M., Davies, N. G., and Simons, D. (2020). Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(10):1151–1160.

Kumar, M. S., Safa, A. M., Deodhar, S. D., and Schumacher, O. P. (1977). The relationship of thyroid-stimulating hormone (TSH), thyroxine (T4), and triiodothyroxine (T3) in primary thyroid failure. *American Journal of Clinical Pathology*, 68(6):747–751.

Kunutsor, S. K. and Laukkanen, J. A. (2021). Markers of liver injury and clinical outcomes in COVID-19 patients: a systematic review and meta-analysis. *Journal of Infection*, 82(1):159–198.

Lee, Y., Lui, L. M., Chen-Li, D., Liao, Y., Mansur, R. B., Brietzke, E., Rosenblat, J. D., Ho, R., Rodrigues, N. B., Lipsitz, O., Nasri, F., Cao, B., Subramaniapillai, M., Gill, H., Lu, C., and McIntyre, R. S. (2021). Government response moderates the mental health impact of COVID-19: a systematic review and meta-analysis of depression outcomes across countries. *Journal of Affective Disorders*, 290:364–377.

Lei, F., Liu, Y. M., Zhou, F., Qin, J. J., Zhang, P., Zhu, L., Zhang, X. J., Cai, J., Lin, L., Ouyang, S., Wang, X., Yang, C., Cheng, X., Liu, W., Li, H., Xie, J., Wu, B., Luo, H., Xiao, F., Chen, J., Tao, L., Cheng, G., She, Z. G., Zhou, J., Wang, H., Lin, J., Luo, P., Fu, S., Zhou, J., Ye, P., Xiao, B., Mao, W., Liu, L., Yan, Y., Liu, L., Chen, G., Li, H., Huang, X., Zhang, B. H., and Yuan, Y. (2020). Longitudinal association between markers of liver injury and mortality in COVID-19 in China. *Hepatology*, 72(2):389–398.

Leigh-Hunt, N., Bagguley, D., Bash, K., Turner, V., Turnbull, S., Valtorta, N., and Caan, W. (2017). An overview of systematic reviews on the public health consequences of social isolation and loneliness. *Public Health*, 152:157–171.

Li, L., Hu, B., and Greene, T. (2009). A semiparametric joint model for longitudinal

USC DE SANTIAGO DE COMPOSTELA

and survival data with application to hemodialysis study. *Biometrics*, 65(3):737–745.

Li, Q., Xu, W., Li, W., Huang, C., and Chen, L. (2020a). Dynamics of cytokines and lymphocyte subsets associated with the poor prognosis of severe COVID-19. *European Review for Medical and Pharmacological Sciences*, 24(23):12536–12544.

Li, Y., Li, H., Zhu, S., Xie, Y., Wang, B., He, L., Zhang, D., Zhang, Y., Yuan, H., Wu, C., Sun, W., Zhang, Y., Li, M., Cui, L., Cai, Y., Wang, J., Yang, Y., Lv, Q., Zhang, L., and Xie, M. (2020b). Prognostic value of right ventricular longitudinal strain in patients with COVID-19. *Cardiovascular Imaging*, 13(11):2287–2299.

Liu, F., Li, L., Xu, M., Wu, J., Luo, D., Zhu, Y., Li, B., Song, X., and Zhou, X. (2020). Prognostic value of interleukin-6, C-reactive protein, and procalcitonin in patients with COVID-19. *Journal of Clinical Virology*, 127:104370.

Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., and Jit, M. (2021). The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Medicine*, 19(1):1–12.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W. J., Wang, D., Xu, W., Holmes, E. C., Gao, G. F., Wu, G., Chen, W., Shi, W., and Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224):565–574.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Luo, S. and Wang, J. (2014). Bayesian hierarchical model for multiple repeated measures and survival data: an application to Parkinson's disease. *Statistics in Medicine*, 33(24):4279–4291.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170.

USC DE SANTIAGO DE COMPOSTELA

Marjot, T., Webb, G. J., Barritt, A. S., Moon, A. M., Stamataki, Z., Wong, V. W., and Barnes, E. (2021). COVID-19 and liver disease: mechanistic and clinical perspectives. *Nature Reviews Gastroenterology & Hepatology*, 18(5):348–364.

Marra, G. and Radice, R. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112:99–113.

Marra, G. and Radice, R. (2022). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-6.

Mauff, K., Steyerberg, E., Kardys, I., Boersma, E., and Rizopoulos, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, 30(4):999–1014.

Mesquita, A. R., Costa, R., Bina, R., Cadarso-Suárez, C., Gude, F., Díaz-Louzao, C., Dikmen-Yildiz, P., Osório, A., Mateus, V., Domínguez-Salas, S., Vousoura, E., Levy, D., Alfayumi-Zeadna, S., Wilson, C. A., Contreras-García, Y., Saldivia, S., Christoforou, A., Hadjigeorgiou, E., Felice, E., Buhagiar, R., Hancheva, C., Ajaz, E., Uka, A., Motrico, E., and Riseup-PPD-COVID-19 Group (2022). The impact of governmental responses to the COVID-19 pandemic on perinatal mental health: evidence from 12 countries. *A preprint, available online at SSRN: https://ssrn.com/abstract=4006137*.

Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.

Motrico, E., Bina, R., Domínguez-Salas, S., Mateus, V., Contreras-García, Y., Carrasco-Portiño, M., Ajaz, E., Apter, G., Christoforou, A., Dikmen-Yildiz, P., Felice, E., Hancheva, C., Vousoura, E., Wilson, C. A., Buhagiar, R., Cadarso-Suárez, C., Costa, R., Devouche, E., Ganho-Ávila, A., Gómez-Baya, D., Gude, F., Hadjigeorgiou, E., Levy, D., Osorio, A., Rodriguez, M. F., Saldivia, S., González, M. F., Mattioli, M., Mesquita, A., and Riseup-PPD-COVID-19 Group (2021). Impact of the Covid-19 pandemic on perinatal mental health (Riseup-PPD-COVID-19): protocol for an international prospective cohort study. *BMC Public Health*, 21(1):1–11.

National Institutes of Health (NIH) (2018). NIH Strategic Plan for Data Science. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf. Accessed: 16-08-2022.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer.

USC DE SANTIAGO DE COMPOSTELA

Ngonghala, C. N., Iboi, E., Eikenberry, S., Scotch, M., MacIntyre, C. R., Bonds, M. H., and Gumel, A. B. (2020). Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus. *Mathematical Biosciences*, 325:108364.

Nisar, A., Yin, J., Waqas, A., Bai, X., Wang, D., Rahman, A., and Li, X. (2020). Prevalence of perinatal depression and its determinants in Mainland China: a systematic review and meta-analysis. *Journal of Affective Disorders*, 277:1022–1037.

Oussalah, A., Gleye, S., Urmes, I. C., Laugel, E., Barbé, F., Orlowski, S., Malaplate, C., Aimone-Gastin, I., Caillierez, B. M., Merten, M., Jeannesson, E., Kormann, R., Olivier, J. L., Rodriguez-Guéant, R. M., Namour, F., Bevilacqua, S., Thilly, N., Losser, M.-R., Kimmoun, A., Frimat, L., Levy, B., Gibot, S., Schvoerer, E., and Guéant, J. L. (2020). The spectrum of biochemical alterations associated with organ dysfunction and inflammatory status and their association with disease outcomes in severe COVID-19: a longitudinal cohort and time-series design study. *EClinicalMedicine*, 27:100554.

Palaro, H. P. and Hotta, L. K. (2006). Using conditional copula to estimate value at risk. *Journal of Data Science*, 4:93–115.

Paštrovic, F., Lucijanic, M., Atic, A., Stojic, J., Barisic Jaman, M., Tjesic Drinkovic, I., Zelenika, M., Milosevic, M., Medic, B., Loncar, J., Mijic, M., Filipec Kanizaj, T., Kralj, D., Lerotic, I., Virovic Jukic, L., Ljubicic, N., Luetic, K., Grgic, D., Majerovic, M., Ostojic, R., Krznaric, Z., Luksic, I., Piskac Zivkovic, N., Keres, T., Grabovac, V., Persec, J., Barsic, B., and Grgurevic, I. (2021). Prevalence and prognostic impact of deranged liver blood tests in COVID-19: experience from the regional COVID-19 center over the cohort of 3812 hospitalized patients. *Journal of Clinical Medicine*, 10(18):4222.

Philip, R. K., Purtill, H., Reidy, E., Daly, M., Imcha, M., McGrath, D., O'Connell, N. H., and Dunne, C. P. (2020). Unprecedented reduction in births of very low birthweight (VLBW) and extremely low birthweight (ELBW) infants during the COVID-19 lockdown in Ireland: a "natural experiment" allowing analysis of data from the prior two decades. *BMJ Global Health*, 5(9):e003075.

Philipson, P., Sousa, I., Diggle, P. J., Williamson, P., Kolamunnage-Dona, R., Henderson, R., and Hickey, G. L. (2018). *joineR: Joint Modelling of Repeated Measurements and Time-to-Event Data*. R package version 1.2.6.

Phipps, M. M., Barraza, L. H., LaSota, E. D., Sobieszczyk, M. E., Pereira, M. R., Zheng, E. X., Fox, A. N., Zucker, J., and Verna, E. C. (2020). Acute liver injury

USC DE SANTIAGO DE COMPOSTELA

in COVID-19: prevalence and association with clinical outcomes in a large US cohort. *Hepatology*, 72(3):807–817.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer.

Pirola, C. J. and Sookoian, S. C. (2020). SARS-CoV-2 virus and liver expression of host receptors: Putative mechanisms of liver involvement in COVID-19. *Liver International*, 40(8):2038–2040.

Popkin, B. M., Du, S., Green, W. D., Beck, M. A., Algaith, T., Herbst, C. H., Alsukait, R. F., Alluhidan, M., Alazemi, N., and Shekar, M. (2020). Individuals with obesity and COVID-19: a global perspective on the epidemiology and biological relationships. *Obesity Reviews*, 21(11):e13128.

Proust-Lima, C., Philipps, V., Diakite, A., and Liquet, B. (2022). *lcmm: Extended Mixed Models Using Latent Classes and Latent Processes*. R package version: 2.0.0.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment PSA: a joint modeling approach. *Biostatistics*, 10(3):535–549.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramasamy, M. N., Minassian, A. M., Ewer, K. J., Flaxman, A. L., Folegatti, P. M., Owens, D. R., Voysey, M., Aley, P. K., Angus, B., Babbage, G., Belij-Rammerstorfer, S., Berry, L., Bibi, S., Bittaye, M., Cathie, K., Chappell, H., Charlton, S., Cicconi, P., Clutterbuck, E. A., Colin-Jones, R., Dold, C., Emary, K. R. W., Fedosyuk, S., Fuskova, M., Gbesemete, D., Green, C., Hallis, B., Hou, M. M., Jenkin, D., Joe, C. C. D., Kelly, E. J., Kerridge, S., Lawrie, A. M., Lelliott, A., Lwin, M. N., Makinson, R., Marchevsky, N. G., Mujadidi, Y., Munro, A. P. S., Pacurar, M., Plested, E., Rand, J., Rawlinson, T., Rhead, S., Robinson, H., Ritchie, A. J., Ross-Russell, A. L., Saich, S., Singh, N., Smith, C. C., Snape, M. D., Song, R., Tarrant, R., Themistocleous, Y., Thomas, K. M., Villafana, T. L., Warren, S. C., Watson, M. E. E., Douglas, A. D., Hill, A. V. S., Lambe, T., Gilbert, S. C., Faust, S. N., Pollard, A. J., and the Oxford COVID Vaccine Trial Group (2020). Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3 trial. *The Lancet*, 396(10267):1979–1993.

Ranjbar, F., Allahqoli, L., Ahmadi, S., Mousavi, R., Gharacheh, M., Eshraghi, N., and Alkatout, I. (2021). Changes in pregnancy outcomes during the COVID-19 lockdown in Iran. *BMC Pregnancy and Childbirth*, 21(1):1–6.

USC DE SANTIAGO DE COMPOSTELA

Regmi, K. and Lwin, C. M. (2021). Factors associated with the implementation of non-pharmaceutical interventions for reducing coronavirus disease 2019 (COVID-19): a systematic review. *International Journal of Environmental Research and Public Health*, 18(8):4274.

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., Barnaby, and the Northwell COVID-19 Research Consortium (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA: The Journal of the American Medical Association*, 323(20):2052–2059.

Ridruejo, E. and Soza, A. (2020). The liver in times of COVID-19: what hepatologists should know. *Annals of Hepatology*, 19(4):353–358.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*, 35(9):1–33.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. Chapman and Hall/CRC.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72(7):1–45.

Rizopoulos, D., Papageorgiou, G., and Miranda Afonso, P. (2022). *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data*. R package version 0.2-4.

Rogers, J. P., Chesney, E., Oliver, D., Pollak, T. A., McGuire, P., Fusar-Poli, P., Zandi, M. S., Lewis, G., and David, A. S. (2020). Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic. *The Lancet Psychiatry*, 7(7):611–627.

Sabeti, A., Wei, M., and Craiu, R. (2014). Additive models for conditional copulas. *Statistics*, 3:300–312.

USC DE SANTIAGO DE COMPOSTELA

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology*, pages 231–255. Springer.

Sharma, A., Jaiswal, P., Kerakhan, Y., Saravanan, L., Murtaza, Z., Zergham, A., Honganur, N.-S., Akbar, A., Deol, A., Francis, B., Patel, S., Mehta, D., Jaiswal, R., Singh, J., Patel, U., and Malik, P. (2021). Liver disease and outcomes among COVID-19 hospitalized patients–a systematic review and meta-analysis. *Annals of Hepatology*, 21:100273.

Shu, M., Bowen, R. S., Herrmann, C., Qi, G., Santacatterina, M., and Zabih, R. (2021). Deep survival analysis with longitudinal X-rays for COVID-19. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4046–4055.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Sousa, I. (2011). A review on joint modeling of longitudinal measurements and time-to-event. *REVSTAT - Statistical Journal*, 9(1):57–81.

Spencer, C., LoPresti, J., Patel, A., Guttler, R., Eigen, A., Shen, D., Gray, D., and Nicoloff, J. (1990). Applications of a new chemiluminometric thyrotropin assay to subnormal measurement. *The Journal of Clinical Endocrinology & Metabolism*, 70(2):453–460.

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Strömer, A., Staerk, C., Klein, N., Weinhold, L., Titze, S., and Mayr, A. (2022). Deselection of base-learners for statistical boosting—with an application to distributional regression. *Statistical Methods in Medical Research*, 31(2):207–224.

Sun, F., Zhu, J., Tao, H., Ma, Y., and Jin, W. (2021). A systematic review involving 11,187 participants evaluating the impact of COVID-19 on anxiety and depression in pregnant women. *Journal of Psychosomatic Obstetrics & Gynecology*, 42(2):91–99.

Sun, J., Aghemo, A., Forner, A., and Valenti, L. (2020). COVID-19 and liver disease. *Liver International*, 40(6):1278–1281.

USC DE SANTIAGO DE COMPOSTELA

Taylor, P. N., Albrecht, D., Scholz, A., Gutierrez-Buey, G., Lazarus, J. H., Dayan, C. M., and Okosieme, O. E. (2018). Global epidemiology of hyperthyroidism and hypothyroidism. *Nature Reviews Endocrinology*, 14(5):301.

Teslya, A., Pham, T. M., Godijk, N. G., Kretzschmar, M. E., Bootsma, M. C., and Rozhnova, G. (2020). Impact of self-imposed prevention measures and short-term government-imposed social distancing on mitigating and delaying a COVID-19 epidemic: a modelling study. *PLoS Medicine*, 17(7):e1003166.

Tian, D. and Ye, Q. (2020). Hepatic complications of COVID-19 and its treatment. *Journal of Medical Virology*, 92(10):1818–1824.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tong-Minh, K., van der Does, Y., van Rosmalen, J., Ramakers, C., Gommers, D., van Gorp, E., Rizopoulos, D., and Endeman, H. (2022). Joint modeling of repeated measurements of different biomarkers predicts mortality in COVID-19 patients in the Intensive Care Unit. *Biomarker Insights*, 17:11772719221112370.

Touloumi, G., Pantazis, N., Babiker, A. G., Walker, S. A., Katsarou, O., Karafoulidou, A., Hatzakis, A., Porter, K., Collaboration, C., et al. (2004). Differences in HIV RNA levels before the initiation of antiretroviral therapy among 1864 individuals with known HIV-1 seroconversion dates. *AIDS*, 18(12):1697–1705.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–834.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.

UCLA Endocrine Center (UCLA Health) (n.d.). What are normal thyroid hormone levels? https://www.uclahealth.org/medical-services/surgery/endocrine-surgery/conditions-treated/thyroid/normal-thyroid-hormone-levels#:~:text=TSH\%20normal\%20values\%20are\%200.5\%20to\%205.0\%20mIU%2FL.. Accessed: 25-04-2021.

Vatter, T. and Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P. C. (2021). Rank-normalization, folding, and localization: an improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718.

USC DE SANTIAGO DE COMPOSTELA

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longtitudinal Data*. Springer.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., and Peng, Z. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA: The Journal of the American Medical Association*, 323(11):1061–1069.

Weissman, M. M., Wickramaratne, P., Nomura, Y., Warner, V., Pilowsky, D., and Verdeli, H. (2006). Offspring of depressed parents: 20 years later. *American Journal of Psychiatry*, 163(6):1001–1008.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

World Health Organization (WHO) (2020a). Coronavirus disease 2019 (COVID-19) situation report-172. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200710-covid-19-sitrep-172.pdf?sfvrsn=70724b90_2. Accessed: 11-07-2020.

World Health Organization (WHO) (2020b). Coronavirus disease (COVID-19) technical guidance: Laboratory testing for 2019-nCoV in humans. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance/. Accessed: 08-06-2021.

World Health Organization (WHO) (2020c). WHO coronavirus (COVID 19) dashboard. https://covid19.who.int/. Accessed: 08-06-2021.

World Health Organization (WHO) (2022). Weekly epidemiological update on COVID-19 - 21 September 2022. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—21-september-2022. Accessed: 21-09-2022.

Wu, Y., Zhang, C., Liu, H., Duan, C., Li, C., Fan, J., Li, H., Chen, L., Xu, H., Li, X., Guo, Y., Wang, Y., Li, X., Li, J., Zhang, T., You, Y., Li, H., Yang, S., Tao, X., Xu, Y., Lao, H., Wen, M., Zhou, Y., Wang, J., Chen, Y., Meng, D., Zhai, J., Ye, Y., Zhong, Q., Yang, X., Zhang, D., Zhang, J., Wu, X., Chen, W., Dennis, C.-L., and feng Huang, H. (2020). Perinatal depressive and anxiety symptoms of pregnant women during the coronavirus disease 2019 outbreak in China. *American Journal of Obstetrics and Gynecology*, 223(2):240.e1–240.e9.

USC
DE SANTIAGO
DE COMPOSTELA

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Xu, L., Liu, J., Lu, M., Yang, D., and Zheng, X. (2020). Liver injury during highly pathogenic human coronavirus infections. *Liver International*, 40(5):998–1004.

Xu, W., Huang, C., Fei, L., Li, Q., and Chen, L. (2021). Dynamic changes in liver function tests and their correlation with illness severity and mortality in patients with COVID-19: a retrospective cohort study. *Clinical Interventions in Aging*, 16:675–685.

Yan, H., Ding, Y., and Guo, W. (2020). Mental health of pregnant and postpartum women during the coronavirus disease 2019 pandemic: a systematic review and meta-analysis. *Frontiers in Psychology*, 11:617001.

Yee, T. (2015). *Vector generalized linear and additive models: with an implementation in R*. Springer.

Zhang, D., Chen, M. H., Ibrahim, J. G., Boye, M. E., and Shen, W. (2017). Bayesian model assessment in joint modeling of longitudinal and survival data with applications to cancer clinical trials. *Journal of Computational and Graphical Statistics*, 26(1):121–133.

Zhang, Z., Charalambous, C., and Foster, P. (2022). Joint modelling of longitudinal measurements and survival times via a multivariate copula approach. *Journal of Applied Statistics*, 0(0):1–21.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Further Information

Some additional information regarding this PhD thesis work is provided in this final part.

Conflict of Interests Statement

The PhD student declares that she has no conflict of interest in relation to the PhD thesis.

Publications Supporting the Contributions of this PhD Thesis Work

Some information on the papers that support the contributions of this PhD thesis work is provided, in compliance with the rules of doctoral studies at *Universidade de Santiago de Compostela* collected in *Regulamento dos estudos de doutoramento na USC, DOG de 16 de setembro de 2020*.

Published Papers

The following works, corresponding to parts of Chapters 2, 3 and 4, have been published. Information about the articles, journals and copyright permissions is provided hereafter.

Chapters 2 and 3: Díaz-Louzao et al. (2022)

Title: Longitudinal relationship of liver injury with inflammation biomarkers in COVID-19 hospitalized patients using a joint modeling approach.
Authors: C. Díaz-Louzao[1,2,3], L. Barrera-Lopez[5], M. Lopez-Rodriguez[5], C. Casar[5], N. Vazquez-Agra[5], H. Pernas-Pardavila[5], A. Marques-Afonso[5], M. Vidal-Vazquez[5],

USC DE SANTIAGO DE COMPOSTELA

J.G. Montoya[5], A.H. Andrade[5], I. Fernandez-Castro[5], P. Varela[5], A. Gonzalez-Quintela[5], E. Otero[5], F. Gude[6], C. Cadarso-Suarez[2,3,4], and S. Tome[1,5].
Affiliations:

[1] Department of Psychiatry, Radiology, Public Health, Nursing and Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain.

[2] CITMAga, Santiago de Compostela, Spain.

[3] Group of Biostatistics and Biomedical Data Science, University of Santiago de Compostela, Santiago de Compostela, Spain.

[4] Department of Statistics, Mathematical Analysis and Optimisation, University of Santiago de Compostela, Santiago de Compostela, Spain.

[5] Hepatology Unit, Department of Internal Medicine, Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Spain.

[6] Epidemiology and Research Unit, Complejo Hospitalario Universitario de Santiago de Compostela, Instituto de Investigaciones Sanitarias (IDIS) Santiago de Compostela, Santiago de Compostela, Spain.

Year: 2022.
Journal: Scientific Reports.
Volume: 12.
Number: 1.
Pages: 1–13.
ISSN: 20452322.
Publisher: Springer Nature.
DOI: https://doi.org/10.1038/s41598-022-09290-x.
SJR Impact Factor and Quartile: 1.005 [Q1 in Multidisciplinary (11/120)].
Cite Score: 6.9.
Information Regarding Copyright and Use: This is an open access article distributed under the terms of the Creative Commons CC BY license, with permission to reproduce in whole or in part. See https://s100.copyright.com/AppDispatchServlet?title=Longitudinal%20relationship%20of%20liver%20injury%20with%20inflammation%20biomarkers%20in%20COVID-19%20hospitalized%20patients%20using%20a%20joint%20modeling%20approach&author=Carla%20Diaz-Louzao%20et%20al&contentID=10.1038%2Fs41598-022-09290-x&copyright=The%20Author%28s%29&publication=2045-2322&publicationDate=2022-04-01&publisherName=SpringerNature&orderBeanReset=true&oa=CC%20BY
PhD Candidate's Contributions: The candidate contributed to the conceptualization of the works, data cleaning and analysis, and to the manuscripts preparation.

USC DE SANTIAGO DE COMPOSTELA

Chapters 2 and 4: Díaz-Louzao et al. (2021)

Title: Multivariate Conditional Transformation Models. Application to Thyroid-Related Hormones.
Authors: C. Díaz-Louzao[1,2], O. Lado-Baleato[2,3], F. Gude[4], and C. Cadarso-Suárez[2,3].
Affiliations:

[1] Department of Psychiatry, Radiology, Public Health, Nursing, and Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain.

[2] Group of Biostatistics and Biomedical Data Science, University of Santiago de Compostela, Santiago de Compostela, Spain.

[3] Department of Statistics, Mathematical Analysis, and Optimisation, University of Santiago de Compostela, Santiago de Compostela, Spain.

[4] Epidemiology and Research Unit, Complejo Hospitalario Universitario de Santiago de Compostela, Instituto de Investigaciones Sanitarias (IDIS) Santiago de Compostela, Santiago de Compostela, Spain.

Year: 2021.
Conference Proceedings: Computational Science and Its Applications – ICCSA 2021.
Book Series: Lecture Notes in Computer Science.
Volume: 12949.
Pages: 341–351.
ISSN: 0302-9743.
Publisher: Springer Nature.
DOI: https://doi.org/10.1007/978-3-030-86653-2_25.
SJR Impact Factor and Quartile: 0.407 [Q3 in Mathematics: Theoretical Computer Sciences (78/123)].
Cite Score: 2.1.
Information Regarding Copyright and Use: For a thesis/dissertation reuse request: *Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.*
Complete information about copyright and reuses can be found at:
https://www.springer.com/gp/rights-permissions/obtaining-permissions/882.
PhD Candidate's Contributions: The candidate contributed to the conceptualization of the works, simulation study, data analysis and to the manuscripts preparation.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Unpublished Papers

The following work, corresponding to parts of Chapters 2 and 3 has not been peer-rewiewed, but has been submitted to *Scientific Reports*. Currently, it can be found online as a preprint available at SSRN.

Chapters 2 and 3: Mesquita et al. (2022)

Title: The Impact of Governmental Responses to the COVID-19 Pandemic on Perinatal Mental Health: Evidence From 12 Countries.
Authors: A. Mesquita[1], R. Costa[2,3,4], R. Bina[5], C. Cadarso-Suárez[6,7], F. Gude[8], C. Díaz-Louzao[7,9], P. Dikmen-Yildiz[10], A. Osorio[11], V. Mateus[11], S. Domínguez-Salas[12], E. Vousoura[13], D. Levy[5], S. Alfayumi-Zeadna[14,15], C.A. Wilson[16], Y. Contreras-García[17], S. Saldivia[18], A. Christoforou[19], E. Hadjigeorgiou[20], E. Felice[21], R. Buhagiar[21], C. Hancheva[22], E. Ajaz[23], A. Uka[23], and Emma Motrico[12], and Riseup-PPD-COVID-19 Group.
Affiliations:

[1] School of Psychology, University of Minho, Braga, Portugal.

[2] EPIUnit, Institute of Public Health, University of Porto, Porto, Portugal.

[3] Laboratory for Integrative and Translational Research in Population Health (ITR), Porto, Portugal.

[4] Hei-Lab: Digital Human-Environment Interaction Lab, Faculty of Psychology, Education and Sports, Lusófona University, Porto, Portugal.

[5] School of Social Work, Bar Ilan University, Ramat Gan, Israel.

[6] Department of Statistics, Mathematical Analysis, and Optimisation, University of Santiago de Compostela, Santiago de Compostela, Spain.

[7] Group of Biostatistics and Biomedical Data Science, University of Santiago de Compostela, Santiago de Compostela, Spain.

[8] Epidemiology and Research Unit, Complejo Hospitalario Universitario de Santiago de Compostela, Instituto de Investigaciones Sanitarias (IDIS) Santiago de Compostela, Santiago de Compostela, Spain.

[9] Department of Psychiatry, Radiology, Public Health, Nursing, and Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain.

[10] Department of Psychology, Kirklareli University, Kirklareli, Turkey.

USC DE SANTIAGO DE COMPOSTELA

[11] Graduate Program on Developmental Disorders, Center for Biological and Health Sciences, Mackenzie Presbyterian University, São Paulo, Brazil.

[12] Psychology Department, Universidad Loyola Andalucia, Sevilla, Spain

[13] First Department of Psychiatry, Medical School, National and Kapodestrian University of Athens, Eginition Hospital, Athens, Greece.

[14] The Center for Women's Health Studies and Promotion, Ben-Gurion University of the Negev, Israel.

[15] MAP Centre for Urban Health Solutions, Li Ka Shing Knowledge Institute, Michael's Hospital, Unity Health Toronto, Toronto, Canada.

[16] Section of Women's Mental Health, King's College London and South London and Maudsley NHS Foundation Trust, London, UK.

[17] Department of Obstetrics and Childcare, School of Medicine, Universidad de Concepción, Concepción, Chile.

[18] Department of Psychiatry and Mental Health, Universidad de Concepción, Concepción, Chile.

[19] Department of Social and Behavioural Sciences, European University Cyprus, Nicosia, Cyprus.

[20] Department of Nursing, School of Health Science, Cyprus University of Technology, Limassol, Cyprus.

[21] Department of Psychiatry, University of Malta, Malta.

[22] Sofia University "St. Kliment Ochridski", Sofia, Bulgaria.

[23] Department of Education and English Language, Beder University College, Tirana, Albania.

Year: 2022.
DOI: http://dx.doi.org/10.2139/ssrn.4006137.
PhD Candidate's Contributions: The candidate contributed to the data cleaning and analysis, and to the manuscripts preparation.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

In recent years, the technological revolution is allowing the collection of an enormous amount of data of different types, creating enormously complex databases that require the collaboration of statisticians and clinicians to carry out a biomedical study with guarantees, applying the tools of data science. This requires the development of new statistical techniques. This thesis focuses on joint modelling regression models for multivariate responses. Specifically, we study the cases of two and three continuous outcomes, as well as models for longitudinal and survival data. These techniques are applied in three studies of major epidemiological importance: liver damage and survival in COVID-19 patients, perinatal mental health during the COVID-19 pandemic, and the study of thyroid-related hormones.