USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

Aspectos computacionales de los contrastes de bondad de ajuste para modelos de regresión lineales

Guillermo Fernández Fernández

2024/2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO EN MATEMÁTICAS

Trabajo Fin de Grado

Aspectos computacionales de los contrastes de bondad de ajuste para modelos de regresión lineales

Guillermo Fernández Fernández

Julio, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa

Título: Aspectos computacionales de los contrastes de bondad de ajuste para modelos de regresión lineales

Breve descripción del contenido

Los modelos de regresión en media son una herramienta de gran interés en el ámbito de la Estadística, ya que permiten establecer la relación de dependencia entre una variable de interés (que habitualmente se conoce como variable respuesta) y una o varias variables explicativas.

Los modelos clásicos de regresión asumen que la relación entre las variables explicativas y la variable respuesta se puede establecer a través de efectos lineales, surgiendo así los modelos de regresión lineales múltiples.

A lo largo de este TFG, trataremos de presentar contrastes de hipótesis que nos permitan testear si realmente el efecto de las variables explicativas sobre la variable respuesta es o no lineal. Para llevar a cabo dichos contrastes emplearemos la función de regresión integrada introducida por Stute (1997).

A modo de orientación, el trabajo podría organizarse en las siguientes secciones:

- Los modelos de regresión lineales múltiples.
- Presentación del contraste de bondad de ajuste para modelos de regresión lineales.
- El calibrado del contraste de bondad de ajuste.
- Implementación del contraste de bondad de ajuste, junto con el plan de remuestreo.

Para ello utilizaremos el software estadístico libre R (`https://www.r-project.org/`). Además, ilustramos el buen comportamiento en la práctica del contraste de bondad de ajuste introducido, utilizando tanto conjuntos de datos reales como datos simulados.

Recomendaciones
Otras observaciones

Índice

Resumen

Resumen

En este trabajo se realizará una introducción al modelo de regresión lineal múltiple y a los modelos de regresión paramétricos, así como a las hipótesis clásicas que se suponen sobre los mismos. Una de ellas es la hipótesis de linealidad, respectivamente la forma paramétrica considerada, del modelo; que derivará en que nos centremos en el contraste de bondad de ajuste presentado por Stute (1997), el cuál resulta de gran utilidad a la hora de verificar si un modelo especificado verifica una cierta forma paramétrica. Debido a la complejidad a la hora de estimar la distribución del estadístico propuesto para dicho contraste, se presentará una aproximación *bootstrap*, más concretamente un *wild bootstrap* sobre los residuos del modelo considerado para llevar a cabo el calibrado del test en la práctica.

Más adelante, programaremos dicho contraste en  y realizaremos un estudio de simulación con el objetivo de comprobar el buen comportamiento del contraste de forma empírica; verificando que respeta el nivel de significación bajo la hipótesis nula y que muestra una buena potencia, es decir, que es capaz de rechazar la hipótesis nula cuando consideramos modelos bajo la hipótesis alternativa. Por último, presentaremos una aplicación a datos reales que nos permitirá ilustrar la utilidad del procedimiento presentado en la práctica.

Resumo

Neste traballo realizarase unha introdución ao modelo de regresión lineal múltiple e aos modelos de regresión paramétricos, así como ás hipóteses clásicas que se supoñen sobre os mesmos. Unha delas é a hipótese de linealidade, respectivamente a forma paramétrica considerada, do modelo; que derivará en que nos centremos no contraste de bondade de axuste presentado por

Stute (1997), o cal resulta de grande utilidade á hora de verificar se un modelo especificado verifica unha certa forma paramétrica. Debido á complexidade á hora de estimar a distribución do estatístico proposto para dito contraste, presentarase unha aproximación *bootstrap*, máis concretamente un *wild bootstrap* sobre os residuos do modelo considerado para levar a cabo o calibrado do test na práctica.

Máis adiante, programaremos dito contraste en R e realizaremos un estudo de simulación co obxectivo de comprobar o bo comportamento do contraste de xeito empírico; verificando que respeta o nivel de significación baixo a hipótese nula e que amosa unha boa potencia, é dicir, que é capaz de rexeitar a hipótese nula cando consideramos modelos baixo a hipótese alternativa. Por último, presentaremos unha aplicación a datos reais que nos permitirá ilustrar a utilidade do procedemento presentado na práctica.

Abstract

This work will present an introduction to the multiple linear regression model and to parametric regression models, as well as to the classical assumptions typically imposed on them. One of these assumptions is the linearity, respectively the parametric form, of the model; which will lead us to focus on the goodness-of-fit test introduced by Stute (1997), which is highly useful in order to check if a certain model follows some specified parametric form. Due to the complexity of estimating the distribution of the proposed test statistic, we will present a bootstrap approximation, more specifically the wild bootstrap about the residuals of the considered model in order to carry out the calibration of the test in practice.

Later, we will implement this test in R and we will carry out a simulation study with the aim of assessing the good performance of the test empirically; verifying that it respects the nominal level under the null hypothesis and it shows a good power, that is, the test will reject the null hypothesis when we consider models under the alternative hypothesis. Finally, we will present a real data application which will allow us to illustrate the utility of the presented procedure in practice.

Capítulo 1

Introducción

En la actualidad hay muchas situaciones en las que resulta razonable explicar una cierta característica de una población a partir de diferentes factores que creemos que pueden estar afectando en mayor o menor medida. Es aquí cuando surge la **regresión lineal múltiple**, que busca generar un modelo que permita determinar el valor de cierta variable respuesta a partir de la combinación lineal de un conjunto de variables conocidas como variables explicativas o covariables.

Con el objetivo de entender estos conceptos, realizaremos una breve y clara introducción a los modelos de regresión, centrándonos en cuatro aspectos clave que desarrollaremos a lo largo de este capítulo: la formulación del modelo; las hipótesis que se imponen sobre este; la estimación de los parámetros; y la distribución en el muestreo de los estimadores, que nos permitirá aplicar herramientas de Inferencia Estadística sobre los parámetros.

1.1. El modelo de regresión lineal múltiple

A la hora de formular un modelo de regresión, partiremos de una cierta variable respuesta Y y un conjunto de variables explicativas X_1, X_2, ..., X_p dando lugar a la formulación extendida del modelo:

$$Y = m(X) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

donde ε representa el error del modelo y β_0, β_1, ..., β_p representan respectivamente el intercepto y los coeficientes asociados a cada variable explicativa del modelo; siendo $m(x) = \mathbb{E}(Y \mid X = x)$, lo que se conoce como **función de regresión**. Podemos reescribir el modelo en formulación matricial de la siguiente manera:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{2}$$

donde $\mathbf{Y} \in \mathbb{R}^n$ es el vector de valores de la variable respuesta; $\mathbf{X}$ es una matriz de dimensión $n \times (p+1)$ llamada matriz de diseño, donde cada fila representa a un individuo y cada columna una variable explicativa; $\beta \in \mathbb{R}^{p+1}$ es el vector de parámetros desconocidos que hay que estimar; y $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ es el vector que contiene las observaciones del error.

Presentamos a continuación, de manera extendida, la formulación matricial presentada en (2):

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Nótese que la primera columna de la matriz de diseño formada por unos permite la aparición del intercepto β_0 en (1). En la expresión anterior, asumimos un diseño fijo, lo cual significa que los valores de las variables explicativas $X_1, X_2, ..., X_p$ no son aleatorios, y por eso los valores son denotados por x_{ij}. Es decir, tratamos el diseño experimental como fijo y conocido, y la única fuente de aleatoriedad proviene de los errores aleatorios.

1.2. Hipótesis del modelo

Al adoptar un modelo de regresión lineal para el estudio de la dependencia de una variable Y en función de un conjunto de variables $X_1, ..., X_p$ se dan por válidas ciertas hipótesis básicas asociadas a este tipo de modelos, las cuales son: linealidad, homocedasticidad, normalidad e independencia de los errores. El incumplimiento de cualquiera de estas hipótesis supone la extracción de conclusiones erróneas sobre los datos, así como predicciones equivocadas e incluso contrarias a la realidad; de ahí la especial importancia de comprobar siempre que en efecto sí estamos bajo dichas hipótesis. Es decir, resulta crucial llevar a cabo lo que se conoce como **validación del modelo**. A continuación, describiremos en detalle las hipótesis asociadas a un modelo de regresión lineal múltiple.

1.2.1. Linealidad

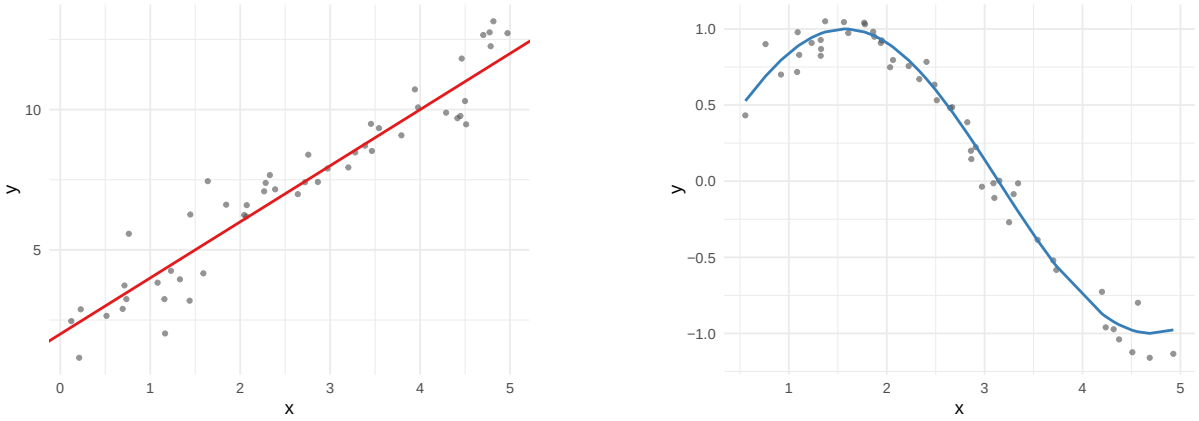
Al considerar un modelo lineal múltiple, estamos suponiendo que nuestra variable respuesta Y guarda una relación lineal con el resto de variables explicativas, es decir, se asume que

$$m(x) = \mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

Es importante destacar que por relación lineal no entendemos que forzosamente los datos siguen una combinación lineal perfecta, sino que dicha relación efectivamente puede ser aproximada a

ello.

Ejemplo 1.1. Presentamos en la Figura 1.1, a modo de ejemplo, dos conjuntos de datos simulados de tamaño de muestra 100. En ambas muestras, la variable X sigue una distribución uniforme en $[0, 5]$. A la izquierda presentamos el modelo lineal $Y = 2 + 2X + \varepsilon$, donde el error ε sigue una distribución normal de media 0 y desviación típica 1 ($\sigma = 1$); mientras que a la derecha, hemos simulado un modelo senoidal, $Y = sin(X) + \varepsilon$, donde el error vuelve a seguir una distribución normal de media cero, pero esta vez $\sigma = 0.1$.



(a) Modelo lineal.

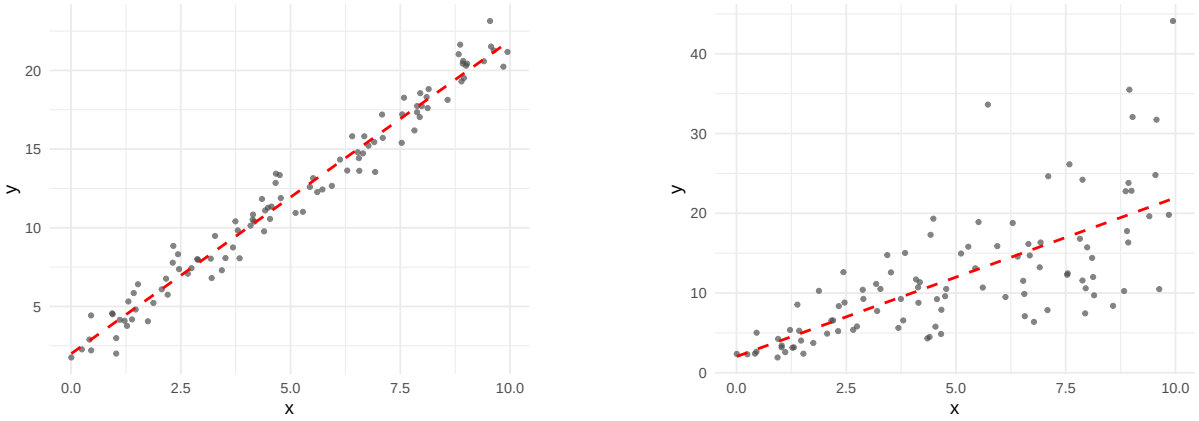
(b) Modelo no lineal.

Figura 1.1: Muestras de tamaño 100 de ejemplos de relaciones de dependencia lineal (Figura (a)) y no lineal (Figura (b)) entre dos variables simuladas.

1.2.2. Homocedasticidad

Por homocedasticidad entendemos que la varianza de ε no depende de las variables explicativas, es decir, $\text{Var}(\varepsilon \mid X = x) = \sigma^2$ es constante para todo posible valor x de las variables explicativas.

Ejemplo 1.2. En la Figura 1.2 hemos representado dos modelos de datos simulados con un tamaño de muestra 100, para los cuales la variable X sigue una distribución uniforme en $[0, 10]$ y ε sigue una distribución normal de media cero y desviación típica 1. A la izquierda presentamos el modelo lineal homocedástico $Y = 2+2X+\sigma(X)\varepsilon$, donde $\sigma(X) = 1$; mientras que a la derecha, tenemos un modelo lineal heterocedástico, $Y = 2 + 2X + \sigma(X)\varepsilon$, donde $\sigma(X) = X+ 0.5$.



(a) Modelo homocedástico.

(b) Modelo heterocedástico.

Figura 1.2: Muestras de tamaño 100 de ejemplos de modelos lineales con errores homocedásticos (Figura (a)) y heterocedásticos (Figura (b)) entre dos variables simuladas.

1.2.3. Normalidad

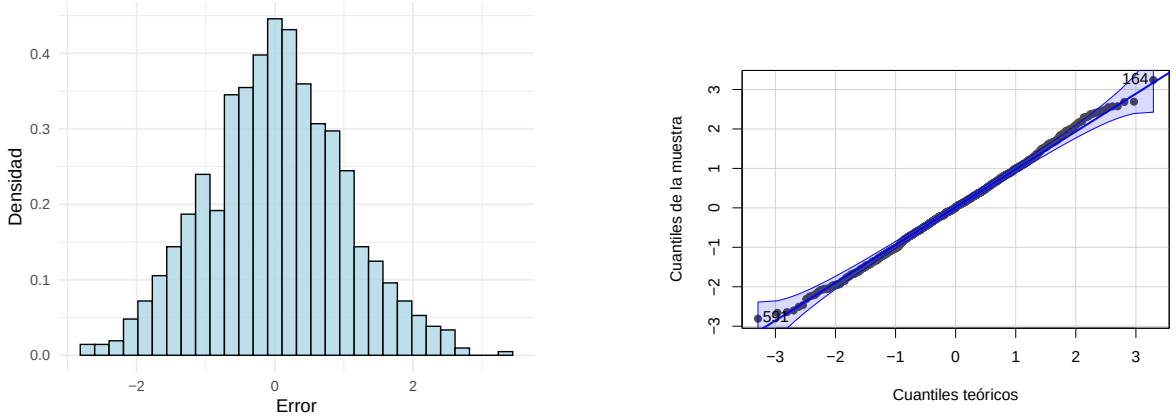
En cuanto a la normalidad, se trata simplemente de suponer que los errores del modelo siguen una distribución normal, es decir, $\varepsilon \in N(0, \sigma^2)$. Para testear esta hipótesis, será necesario realizar un test de bondad de ajuste sobre la distribución normal. En la práctica, los contrastes más utilizados serán el de Shapiro-Wilk, Kolmogorov-Smirnov o el ji-cuadrado.

Ejemplo 1.3. En la Figura 1.3 podemos observar a la izquierda el histograma[1] y a la derecha el QQ-plot[2] asociados a los errores de una muestra de datos de tamaño 1000. Dichos errores proceden de una distribución normal de media cero y $\sigma = 1$.

Por otro lado, en la Figura 1.4, podemos observar a la izquierda el histograma y a la derecha el QQ-plot asociados a los errores de una muestra de datos de tamaño 1000. En este caso, los errores siguen una distribución ji-cuadrado con 3 grados de libertad.

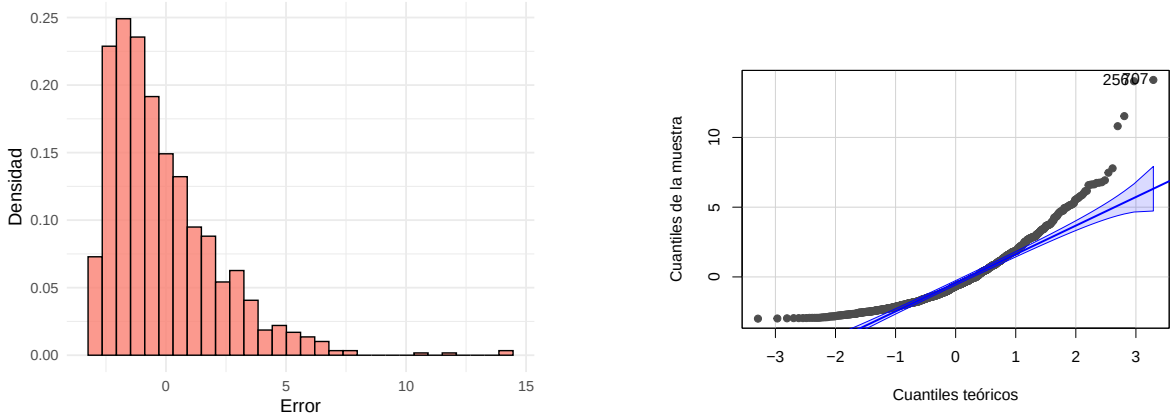
[1]Un histograma es una representación gráfica que "divideüna muestra de datos en intervalos y muestra la densidad con la que ocurren los valores dentro de cada intervalo. Es útil para visualizar la distribución de un conjunto de datos.

[2]Un QQ plot (Quantile-Quantile plot) es una herramienta gráfica que se usa para comparar la distribución de una muestra de datos con una distribución teórica, en nuestro caso la distribución normal.



(a) Histograma. (b) QQ-plot.

Figura 1.3: Histograma (Figura (a)) y QQ-plot (Figura (b)) asociados a muestra de tamaño 1000 de errores que siguen una distribución normal de media cero y $\sigma = 1$.



(a) Histograma. (b) QQ-plot.

Figura 1.4: Histograma (Figura (a)) y QQ-plot (Figura (b)) asociados a muestra de tamaño 1000 de errores que siguen una distribución ji-cuadrado con 3 grados de libertad.

1.2.4. Independencia

Por último, en lo que se refiere a la hipótesis de independencia de los errores, es habitual que podamos suponerla cierta directamente, simplemente porque las observaciones son tomadas directamente de individuos distintos y sin relaciones entre ellos. No obstante, en otras circunstancias surgen diferentes tests capaces de contrastar el cumplimiento de dicha hipótesis.

Esta hipótesis implica que no hay correlación entre los errores del modelo (recordemos que bajo normalidad, incorrelación e independencia son equivalentes), es decir, el valor del error asociado a una observación no debe influir ni estar relacionado con el de otra observación.

Las tres últimas hipótesis se suponen ciertas al asumir lo siguiente:

$$\boldsymbol{\varepsilon} \in N_n\left(0, \sigma^2 I_n\right) \tag{3}$$

siendo σ^2 la varianza del error, que también habrá que estimar, e I_n la matriz identidad de dimensión $n \times n$.

1.3. Estimación de los parámetros

En esta sección nos centraremos en el problema de la estimación de los parámetros del modelo, tanto σ^2 como β. En primer lugar, estimaremos el vector de coeficientes β utilizando el **método de mínimos cuadrados** y luego nos centraremos en la estimación de σ^2.

1.3.1. Estimación del vector de coeficientes

La estimación por **mínimos cuadrados** está basada en minimizar la **suma de los residuos al cuadrado** que cometemos a la hora de realizar cada predicción, esto es,

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - x_i\beta)^2 \tag{4}$$

donde x_i representa la fila i-ésima de la matriz de diseño $\mathbf{X}$, que se corresponde con los valores de las variables explicativas del i-ésimo individuo en nuestra muestra.

El problema de minimización (4) en notación matricial se reescribe como sigue:

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) = \varphi(\beta)$$

siendo φ la función objetivo. Derivando la función φ respecto de β e igualando a cero obtenemos lo que se conoce como **ecuaciones normales de la regresión**:

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$$

que tiene por solución el estimador de β por mínimos cuadrados

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Nótese que para que el estimador esté bien definido, $\mathbf{X}^T\mathbf{X}$ ha de ser una matriz invertible y una vez lo hayamos calculado podremos calcular los ajustes del modelo como sigue:

$$\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\beta}$$

1.3.2. Estimación de la varianza del error

Para estimar la varianza del error vamos a recordar que los residuos se definen como la diferencia entre observaciones y predicciones, es decir,

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - x_i\widehat{\beta}, \quad i \in \{1, \dots, n\}$$

Puesto que los errores ε no se observan, vamos a hacer uso de los residuos previamente definidos para realizar una estimación de su varianza. El **estimador de la varianza del error** es el que sigue

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(Y_i - x_i\widehat{\beta})^2 = \frac{\text{RSS}}{n-(p+1)} \tag{5}$$

donde hemos empleado la notación RSS para la suma residual de cuadrados. El denominador $(n-(p+1))$, aunque no vamos a profundizar en ello, cobra sentido si se hiciese un estudio del sesgo del estimador, pues con este cociente conseguimos que sea insesgado, es decir, que la esperanza del estimador sea el propio parámetro σ^2.

1.4. Inferencia sobre los parámetros

En esta sección trataremos de presentar procedimientos de Inferencia Estadística sobre los parámetros en nuestro modelo lineal, tanto β como σ^2. Para ello, vamos a presentar previamente un resultado teórico que nos será de gran utilidad para nuestro propósito.

Teorema 1.4 (Teorema de Fisher)**.** *Supongamos que los errores $\varepsilon_1, \dots, \varepsilon_n$ son independientes y tienen distribución común $N(0, \sigma^2)$, y que $\boldsymbol{X}$ es una matriz de orden $n \times (p+1)$ de rango $p+1$. Entonces se cumplen las siguientes propiedades:*

(i) El estimador de mínimos cuadrados $\widehat{\beta}$ sigue una distribución normal multivariante:

$$\widehat{\beta} \in N_{p+1}\left(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right).$$

(ii) El cociente entre RSS y σ^2 sigue una distribución ji-cuadrado con $n-(p+1)$ grados de libertad:

$$\frac{RSS}{\sigma^2} = \frac{(n-(p+1))\widehat{\sigma}^2}{\sigma^2} \in \chi^2_{n-(p+1)}.$$

(iii) El estimador $\widehat{\beta}$ y la suma residual de cuadrados (RSS), o equivalentemente $\widehat{\sigma}^2$, son independientes.

Demostración. La demostración de este teorema podemos encontrarla en el Capítulo 3 de Montgomery et al. (2021). □

A la vista del Teorema 1.4, nos encontramos ya en condiciones de realizar tareas de Inferencia Estadística sobre los parámetros del modelo lineal múltiple. Nótese que para poder demostrar el Teorema 1.4 es necesario suponer las cuatro hipótesis clásicas del modelo lineal múltiple: linealidad, homocedasticidad, normalidad e independencia.

1.4.1. Inferencia sobre el vector de coeficientes

Del apartado (i) del Teorema 1.4 y aplicando propiedades de la distribución normal multivariante obtenemos que

$$\frac{(\widehat{\beta}-\beta)^T(\mathbf{X}^T\mathbf{X})(\widehat{\beta}-\beta)}{\sigma^2} \in \chi^2_{p+1}$$

De esta forma, si σ^2 fuese conocida ya tendríamos un pivote para generar regiones de confianza y contrastes de hipótesis sobre β. Ahora bien, en la práctica lo habitual es que σ^2 sea desconocida y por tanto la sustituiremos por $\widehat{\sigma}^2$, viéndose alterada la distribución en el muestreo. Apoyándonos en el Teorema 1.4 seremos capaces de llegar a lo siguiente

$$\frac{(\widehat{\beta}-\beta)^T(X^TX)(\widehat{\beta}-\beta)}{(p+1)\widehat{\sigma}^2} \in F_{p+1,n-(p+1)}$$

puesto que tenemos una distribución ji-cuadrado tanto en el numerador como en el denominador y son independientes. El resultado de esta expresión sigue una distribución F de Snedecor con $p+1$ y $n-(p+1)$ grados de libertad. Esta distribución aparece en muchas pruebas estadísticas, especialmente cuando se evalúan diferencias cuadráticas en modelos de regresión.

Si nuestro objetivo fuese un único coeficiente, β_i, bastaría con extraer el elemento correspondiente de la diagonal de $(\mathbf{X}^T\mathbf{X})^{-1}$ que denotaremos por $(\mathbf{X}^T\mathbf{X})^{-1}_{ii}$. Obteniendo finalmente el pivote

$$\frac{\widehat{\beta}_i-\beta_i}{\widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \in T_{n-(p+1)} \tag{6}$$

que sigue una distribución T de Student con $n-(p+1)$ grados de libertad.

Apoyándonos en el pivote (6), podemos obtener intervalos de confianza[3] de nivel $(1-\alpha)$ para cada coeficiente del modelo de la siguiente forma:

$$\left(\widehat{\beta}_i - t_{n-(p+1),1-\alpha/2}\ \widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}},\quad \widehat{\beta}_i + t_{n-(p+1),1-\alpha/2}\ \widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}\right)$$

donde $t_{n-(p+1),1-\alpha/2}$ es el cuantil de orden $1-\frac{\alpha}{2}$ de la distribución T de Student con $n-(p+1)$ grados de libertad.

Cabe destacar que también existe la posibilidad de considerar un grupo de coeficientes, procediéndose de manera similar a cuando se consideran todos ellos. De hecho, cuando nos centramos en dos coeficientes podemos representar fácilmente la región de confianza en la que se encuentran. Veamos a continuación un ejemplo con datos simulados.

Ejemplo 1.5. Simularemos en R una muestra de tamaño $n = 100$ de un modelo lineal múltiple. El modelo propuesto es el siguiente:

$$Y = 2 + 3X_1 - 2X_2 + \varepsilon,$$

donde

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right)$$

y ε sigue una distribución normal de media cero y $\sigma = 1$.

En la Figura 1.5 podemos observar representada la región de confianza de nivel 99 % de los coeficientes de este modelo, donde además aparece marcado en el centro el estimador de mínimos cuadrados.

Al mismo tiempo, el pivote (6) nos permitirá realizar contrastes de hipótesis sobre los coeficientes. De especial interés resultarán los **contrastes de significación** que permitirán evaluar el efecto de cada variable explicativa sobre la variable respuesta. Recordemos que un contraste de significación es un contraste de la forma:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{cases}$$

en el cual vamos a rechazar la hipótesis nula si $|T| = \left|\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}}\right| > t_{n-(p+1),1-\alpha/2}$, siendo $t_{n-(p+1),1-\alpha/2}$ el cuantil de orden $1-\frac{\alpha}{2}$ de la distribución T de Student con $n-(p+1)$ grados de libertad. Nótese que rechazar la hipótesis nula implicaría que la variable X_i no tiene un efecto sobre la variable respuesta.

[3]Un intervalo de confianza para un parámetro poblacional es un intervalo aleatorio, calculado a partir de los datos muestrales, que contiene al verdadero valor del parámetro con una determinada probabilidad (habitualmente alta).

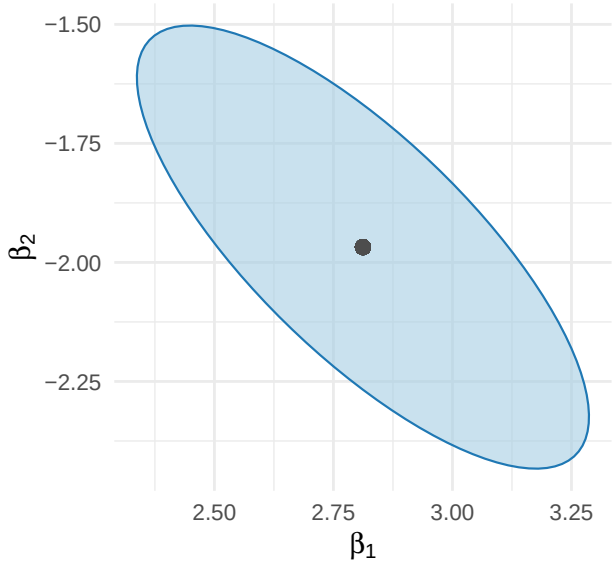


Figura 1.5: Región de confianza y estimador de mínimos cuadrados (punto central) para los coeficientes del modelo $Y = 2 + 3X_1 - 2X_2 + \varepsilon$, asociado a una muestra simulada de tamaño 100.

1.4.2. Inferencia sobre la varianza del error

Basándonos en el apartado (ii) del Teorema 1.4 seremos capaces de realizar tareas de Inferencia Estadística sobre la varianza del error. Recordemos que el pivote que utilizaremos será:

$$\frac{(n-(p+1))\widehat{\sigma}^2}{\sigma^2} \in \chi^2_{n-(p+1)} \tag{7}$$

donde $\widehat{\sigma}^2$ es el estimador de la varianza basado en los residuos del modelo definido en (5).

Este resultado nos permite construir un intervalo de confianza para σ^2. Dado un nivel de confianza $1-\alpha$, los límites del intervalo estarán dados por:

$$\left(\frac{(n-(p+1))\widehat{\sigma}^2}{\chi^2_{n-(p+1),1-\alpha/2}}, \quad \frac{(n-(p+1))\widehat{\sigma}^2}{\chi^2_{n-(p+1),\alpha/2}} \right)$$

donde $\chi^2_{n-(p+1),\alpha/2}$ y $\chi^2_{n-(p+1),1-\alpha/2}$ son los cuantiles de orden $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$, respectivamente, de una distribución ji-cuadrado con $n-(p+1)$ grados de libertad.

Al igual que comentamos en la subsección anterior, el pivote (7) también nos permitirá llevar a cabo contrastes de hipótesis sobre el parámetro σ^2, aunque en este caso resultan de menor utilidad.

1.5. Modelos de regresión paramétricos

Hasta el momento hemos trabajado bajo la suposición de que el comportamiento de nuestra variable respuesta puede ser explicado mediante una combinación lineal de nuestras variables explicativas; no obstante, esta restricción puede no ser adecuada pues podríamos encontrarnos con situaciones reales en las cuales la relación no sea lineal pero sí parametrizable.

Ejemplo 1.6. Simulamos en R un modelo paramétrico no lineal para un tamaño de muestra $n = 100$. El modelo es el siguiente:

$$Y = 2 + 1.5X - 0.2X^2 + \varepsilon,$$

donde X sigue una distribución uniforme en el intervalo $[0, 10]$ y ε sigue una distribución normal de media cero y $\sigma = 1$.

En la Figura 1.6 hemos representado dicho modelo: la línea roja discontinua se corresponde con el ajuste lineal del modelo, mientras que la línea azul se corresponde con el ajuste cuadrático del mismo. Como bien podemos apreciar, suponer que estos datos proceden de un modelo lineal es erróneo y nos llevaría a extraer conclusiones equivocadas sobre los datos analizados.

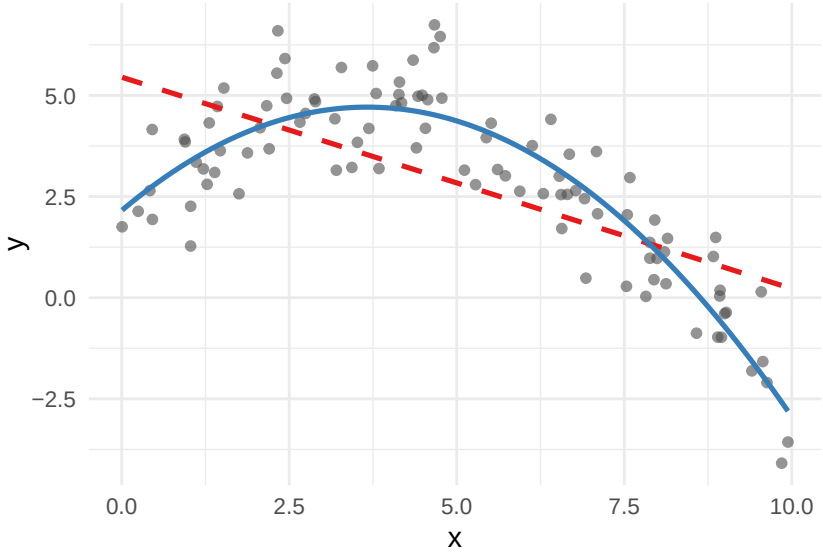


Figura 1.6: Representación de una muestra de datos simulados del modelo $Y = 2 + 1.5X - 0.2X^2 + \varepsilon$ para un tamaño de muestra $n = 100$. La línea roja discontinua representa el ajuste lineal del modelo y la línea azul continua el ajuste cuadrático del mismo.

De esta necesidad surgen los **modelos de regresión paramétricos**, donde planteamos la existencia de una función $m(x, \theta)$ que denominaremos función de regresión paramétrica y será

conocida salvo por el parámetro θ. Dicha función describirá el valor esperado de la variable respuesta en función de las variables explicativas, es decir,

$$\mathbb{E}[Y \mid X = x] = m(x, \theta), \quad x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^d,$$

donde θ es un vector de parámetros desconocidos que caracteriza la familia de funciones $m(x, \theta)$, y el conjunto $\mathcal{X}$ representa el dominio de las variables explicativas. La formulación del modelo paramétrico será la que sigue:

$$Y = m(x, \theta) + \varepsilon.$$

De manera análoga a lo presentado en (4), empleando el método de mínimos cuadrados obtenemos el estimador de θ:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \left(Y_i - m(x_i, \theta)\right)^2.$$

Suponiendo de nuevo cierto (3), es decir, las hipótesis de homocedasticidad, normalidad e independencia de los errores, llegamos a la distribución del estimador:

$$\widehat{\theta} \in N_d\left(\theta, \frac{\sigma^2}{n} J^{-1}\right)$$

donde $J = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta m(x_i, \theta) \nabla_\theta m(x_i, \theta)^\top$, siendo $\nabla_\theta m(x_i, \theta)$ el gradiente de la función $m(x_i, \theta)$ con respecto al parámetro θ, es decir, un vector columna de dimensión d donde cada componente es la derivada parcial de $m(x_i, \theta)$ respecto a cada una de las componentes de θ:

$$\nabla_\theta m(x_i, \theta) = \begin{pmatrix} \frac{\partial m(x_i, \theta)}{\partial \theta_1} \\ \frac{\partial m(x_i, \theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial m(x_i, \theta)}{\partial \theta_d} \end{pmatrix} \in \mathbb{R}^d.$$

Para consultar esto último en más detalle, podemos visitar Amemiya (1985). No profundizaremos más en el modelo paramétrico, pero es de vital importancia presentarlo pues el objetivo de este trabajo se centrará más adelante en contrastar la hipótesis de linealidad o, de forma más general, la forma paramétrica de un modelo de regresión.

1.6. Estructura y objetivo de este trabajo

Como bien hemos presentado previamente, a la hora de estimar un modelo lineal múltiple damos por válidas ciertas hipótesis, pues bien, el **objetivo principal** de este trabajo será **contrastar la hipótesis de linealidad**, o en general, la forma paramétrica del modelo. Para ello, dividiremos el trabajo en varios capítulos que se presentan a continuación.

En el Capítulo 2 se introduce el contraste propuesto por Winfried Stute en 1997, definiendo la función de regresión integrada y profundizando en el estadístico del contraste así como en la distribución que sigue. Nos centraremos también en los aspectos computacionales de dicho contraste, definiendo un procedimiento *bootstrap* para aproximar la distribución del estadístico en la práctica.

En el Capítulo 3 realizaremos lo que se conoce como un estudio de simulación por Monte Carlo, que consistirá en la implementación en R del contraste anteriormente presentado para su posterior puesta a prueba mediante la simulación de diferentes modelos de regresión. Este estudio de simulación tendrá por objetivo comprobar si dicho contraste ajusta bien el nivel de significación, es decir, si es acorde a los posibles niveles de significación establecidos a la hora de aceptar o rechazar la hipótesis nula, cuando estamos bajo la misma. Así mismo, también se considerarán modelos bajo la hipótesis alternativa para testear la potencia del contraste planteado.

En el Capítulo 4 comprobaremos de nuevo cómo se comporta nuestro contraste pero esta vez aplicándolo a datos reales. Para ello, recopilaremos una serie de datos del fútbol de élite con el objetivo de ajustar un modelo de regresión lineal múltiple que sea capaz de explicar los puntos por partido que un equipo logra a lo largo de una temporada. Dicho modelo será validado con nuestro contraste previamente programado.

Finalmente, las principales conclusiones de este trabajo serán presentadas en el Capítulo 5. Así mismo, en el Anexo I se reproduce todo el código de R que ha sido desarrollado para la obtención de las gráficas presentadas en el Capítulo 1, los resultados del estudio de simulación y los de la aplicación a datos reales. En el Anexo II se presenta la base de datos completa que ha sido empleada en el Capítulo 4.

Capítulo 2

Contraste de bondad de ajuste

En este capítulo presentaremos el contraste introducido por Stute (1997) así como el estadístico propuesto, el cual sigue una distribución compleja de usar en la práctica al depender de varias cantidades desconocidas. Motivados por esto mismo, desarrollaremos una sección en la que nos centraremos en el denominado método *bootstrap*, que nos permitará aproximar la distribución del estadístico de contraste en la práctica. Nos centraremos concretamente en el *wild bootstrap* y probaremos que en efecto supone una aproximación válida en el contexto de la bondad de ajuste para modelos de regresión paramétricos.

2.1. Presentación del contraste

El objetivo principal de Stute (1997) es presentar una nueva propuesta metodológica para testear la bondad de ajuste de un modelo de regresión paramétrico. Para ello introduciremos la notación empleada en dicho artículo y trataremos de explicar y entender los aspectos detrás del contraste que se propone.

Sea pues (X, Y) un vector aleatorio en $\mathbb{R}^d \times \mathbb{R}$, y asumiremos que Y es integrable[4], de tal manera que la función de regresión asociada es la siguiente:

$$m(x) = \mathbb{E}[Y \mid X = x], \quad x \in \mathbb{R}^d. \tag{8}$$

En el contexto de la regresión paramétrica se supone que la función de regresión m pertenece a una familia paramétrica de funciones conocida, salvo por un cierto parámetro $\theta \in \mathbb{R}^p$, que denotaremos como sigue:

$$\mathcal{M}_\theta = \{m(\cdot, \theta) : \theta \in \Theta \subset \mathbb{R}^p\},$$

[4]Al asumir integrabilidad, nos aseguramos de que la función de regresión asociada está bien definida salvo en un cojunto de medida cero.

que se traduce en suponer la existencia de un parámetro verdadero $\theta \in \Theta$ tal que:

$$m(x) = m(x, \theta).$$

Un caso particular es el modelo de regresión lineal, presentado en el Capítulo 1, en el que $m(x, \theta) = x\theta$.

El trabajo desarrollado por Stute (1997) tiene como objetivo resolver el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : m \in \mathcal{M}_\theta \\ H_a : m \notin \mathcal{M}_\theta \end{cases}$$

es decir, testear si en efecto nuestra función de regresión pertenece a una cierta familia paramétrica de funciones frente a alternativas totalmente no paramétricas.

Para ello, se introduce la **función de regresión integrada**, que viene dada por:

$$I(x) = \int_{-\infty}^{x} m(z)\, dF(z), \quad x \in \mathbb{R},$$

donde F denota la función de distribución del vector aleatorio $X \in \mathbb{R}^d$. Ahora bien, por la propia definición de la función de regresión m se obtiene que:

$$I(x) = \mathbb{E}\left[\mathbf{1}_{\{X \leq x\}} Y\right],$$

donde $\mathbf{1}$ denota la **función indicadora**[5].

Si suponemos que (X_i, Y_i), con $i = 1, \ldots, n$ es una muestra independiente e idénticamente distribuida con la misma distribución que el par (X, Y), entonces el **estimador empírico**[6] de la función de regresión integrada viene dado por:

$$I_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}} Y_i$$

el cual es insesgado dado que:

$$\mathbb{E}[I_n(x)] = I(x).$$

Presentamos ahora la conocida Ley Fuerte de los Grandes Números con el objetivo de entender más adelante el comportamiento asintótico de nuestro estimador I_n.

[5]La función indicadora $\mathbf{1}_{\{X \leq x\}}$ toma el valor 1 si se cumple que cada componente de $X \in \mathbb{R}^d$ es menor o igual que el correspondiente componente de $x \in \mathbb{R}^d$, es decir,

$$\mathbf{1}_{\{X \leq x\}} = \begin{cases} 1, & \text{si } X_j \leq x_j \text{ para todo } j = 1, \ldots, d \\ 0, & \text{en caso contrario.} \end{cases}$$

[6]Estimación construida a partir de una muestra de datos. Sustituye cantidades teóricas por sus análogos muestrales. En este caso, la función de regresión integrada empírica se obtiene reemplazando la esperanza condicional por un promedio de los datos observados.

Teorema 2.1 (Ley Fuerte de los Grandes Números)**.** *Supongamos que $Z_1, Z_2, \ldots, Z_n$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas con esperanza finita $\mathbb{E}[Z_i] = \mu < \infty$. Entonces, se cumple que:*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{c.s.} \mu \quad \textit{cuando} \quad n \to \infty,$$

es decir, la media muestral ***converge de forma casi segura***[7] *a la esperanza μ.*

Demostración. La demostración de este teorema la podemos encontrar en el Capítulo 8 de Ross (2014). □

Como consecuencia directa del Teorema 2.1, obtenemos lo siguiente:

$$\lim_{n\to\infty} I_n(x) = I(x) \quad \text{con probabilidad 1.}$$

Con el objetivo de obtener límites no degenerados, Stute (1997) propone un proceso estandarizado alternativo, que denominaremos **proceso de residuos integrados** que se define de la siguiente forma:

$$R_n(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}}[Y_i - m(X_i, \theta)], \tag{9}$$

tratándose de una suma de variables aleatorias independientes e idénticamente distribuidas, centradas condicionalmente en X_i, con $1 \le i \le n$, y cuya varianza está dada por

$$T(x) = \int_{-\infty}^{x} \mathrm{Var}(Y \mid X = u)\, dF(u).$$

Si nuestro objetivo fuese testear la hipótesis nula simple de que $m(x) = m(x, \theta_0)$ con θ_0 conocido, es decir, contrastar si se trata de cierta función completamente determinada, entonces nos sería suficiente con (9) pues basta con reemplazar $m(X_i, \theta)$ por $m(X_i, \theta_0)$ y utilizar un estadístico del tipo Kolmogorov-Smirnov basado en R_n para resolver el contraste. Para mayor profundidad en esto mismo, se puede consultar Stute (1997), aunque nosotros nos vamos a centrar en el caso de mayor complejidad que presentamos en la siguiente sección junto con el correspondiente estadístico de contraste.

[7] Una sucesión de variables aleatorias $\{X_n\}$ converge *casi seguramente* a una variable aleatoria X, y se denota $X_n \xrightarrow{c.s.} X$, si

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Esto significa que, salvo en un conjunto de probabilidad nula, los valores de X_n se aproximan a X cuando n tiende a infinito. Es una forma fuerte de convergencia en probabilidad.

2.2. El estadístico de contraste

Para una hipótesis nula compuesta, es decir, si nuestro objetivo fuese contrastar si (8) pertenece a cierta familia de funciones $\mathcal{M}_\theta$, entonces $m(x,\theta)$ debe ser reemplazada por $m(x,\widehat{\theta})$, donde $\widehat{\theta}$ es un estimador adecuado de θ. Esto se traduce en que el estadístico de contraste ha de estar basado en el siguiente proceso:

$$R_n^1(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} \left[Y_i - m(X_i, \widehat{\theta}) \right].$$

Presentamos ahora dos suposiciones así como un teorema con el objetivo de entender el comportamiento asintótico del proceso R_n^1.

SUPUESTO 1. Bajo la hipótesis nula H_0, es decir, $m = m(\cdot,\theta)$ para algún valor desconocido $\theta \in \Theta$, el estimador $\widehat{\theta}$ admite una representación del tipo:

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l(X_i, Y_i, \theta) + o_P(1)^8$$

para alguna función l tal que:

(i) $\mathbb{E}\left[l(X,Y,\theta)\right] = 0$;

(ii) $L(\theta) := \mathbb{E}\left[l(X,Y,\theta) l^\top(X,Y,\theta)\right]$ existe.

SUPUESTO 2.

(i) La función $m(x,\theta)$ es continuamente diferenciable respecto de θ en el interior de Θ. Además, denotaremos su vector gradiente de la siguiente forma:

$$m^{(1)}(x,\theta) = \frac{\partial m(x,\theta)}{\partial \theta} \equiv (m_1^{(1)}(x,\theta), \ldots, m_p^{(1)}(x,\theta))^T.$$

(ii) Existe una función integrable $M(x)$ respecto de F, tal que:

$$|m_i^{(1)}(x,\theta)| \leq M(x) \quad \text{para todo } \theta \in \Theta \text{ y } 1 \leq i \leq p.$$

[8] El término $o_P(1)$ denota una cantidad aleatoria que converge en probabilidad a cero cuando $n \to \infty$. Es decir, para todo $\varepsilon > 0$, se cumple que $\mathbb{P}(|o_P(1)| > \varepsilon) \to 0$. Este término representa una parte residual del desarrollo asintótico del estimador, cuya contribución se vuelve despreciable en muestras de tamaño grande.

Observación 2.2. El Supuesto 2 implica que la función

$$M^{(1)}(x,\theta) = \left(M_1^{(1)}(x,\theta), \ldots, M_p^{(1)}(x,\theta)\right)^T,$$

donde

$$M_i^{(1)}(x,\theta) = \int_{-\infty}^{x} g_i(u,\theta)\, dF(u),$$

con $1 \leq i \leq p$. La función $M^{(1)}(x,\theta)$ está bien definida y es continua en θ para cada θ en el interior de Θ. Además, la continuidad de $M^{(1)}$ asegura la continuidad del proceso límite de R_n^1.

Teorema 2.3. *Supóngase que $\mathbb{E}[Y^2] < \infty$ y que se cumplen los Supuestos 1 y 2. Entonces, bajo la hipótesis $m = m(\cdot,\theta)$, se tiene que, uniformemente en x,*

$$R_n^1(x) = R_n(x) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M^{(1)}(x,\theta)^T l(X_i, Y_i, \theta) + o_P(1).$$

Demostración. La demostración de este resultado podemos encontrarla en Stute (1997). □

Este resultado nos muestra el efecto que tiene en nuestro proceso R_n^1 la estimación del parámetro θ. Estamos ya en condiciones de presentar el siguiente corolario, que establece el comportamiento asintótico del proceso empírico R_n^1.

Corolario 2.4. *Bajo las hipótesis del Teorema 2.3, se tiene que $R_n^1 \to R_\infty^1$ en distribución en el espacio de Skorokhod* [9] *$D[-\infty,\infty]$, donde R_∞^1 es un proceso gaussiano de media cero con función de covarianza*

$$\begin{aligned} K^1(s,t) = {} & T(s \wedge t) + M^{(1)}(s,\theta)^T L(\theta) M^{(1)}(t,\theta) \\ & - M^{(1)}(s,\theta)^T \mathbb{E}\left[\mathbf{1}_{\{X \leq t\}} (Y - m(X,\theta))\, l(X,Y,\theta)\right] \\ & - M^{(1)}(t,\theta)^T \mathbb{E}\left[\mathbf{1}_{\{X \leq s\}} (Y - m(X,\theta))\, l(X,Y,\theta)\right]. \end{aligned}$$

Donde T representa la varianza del proceso de residuos integrados presentado en (9).

A partir de este proceso, se podría considerar como estadístico de contraste cualquier norma del mismo. En particular Stute (1997) introduce el estadístico de tipo *Cramér–von Mises*[10]:

$$W_n^2 = \int_{\mathbb{R}^d} [R_n^1(x)]^2\, dF_n(x), \tag{10}$$

[9] El espacio de Skorokhod $D[a,b]$ es el conjunto de funciones $f : [a,b] \to \mathbb{R}$ que son continuas por la derecha y poseen límite por la izquierda en cada punto.

[10] Un estadístico de tipo Cramér–von Mises consiste en evaluar la discrepancia entre un objeto empírico y su correspondiente versión poblacional mediante la integración del cuadrado de sus diferencias. Este tipo de estadístico es particularmente sensible a desviaciones globales, ya que acumula las discrepancias a lo largo de todo el dominio de estudio.

donde $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$ es la función de distribución empírica asociada al vector aleatorio compuesto por las variables explicativas.

El estadístico (10) mide la discrepancia acumulada entre el modelo estimado bajo la hipótesis nula, $m(x, \widehat{\theta})$, y los datos observados. Valores grandes de W_n^2 sugieren que el modelo está mal especificado, es decir, que el modelo paramétrico no es correcto para explicar el comportamiento de la variable respuesta y que por lo tanto debemos recurrir a modelos no paramétricos más flexibles.

Otra alternativa en lugar de utilizar el estadístico (10), podría ser un estadístico de tipo Kolmogorov-Smirnov, que sería un estadístico de la forma:

$$D_n = \sup_{x \in \mathbb{R}^d} \left| R_n^1(x) \right|.$$

La complicada estructura de $K^1(s,t)$ no permite una representación sencilla del proceso límite R_∞^1, lo que en la práctica dificulta la estimación de la distribución límite del proceso para decidir cuando rechazar o aceptar la hipótesis nula. Además, la convergencia del proceso límite es lenta y su implementación tiene un alto coste computacional. Es por ello que en Stute et al. (1998) se propone una aproximación *bootstrap*, que desarrollaremos en la siguiente sección, para llevar a cabo el calibrado del contraste en la práctica.

2.3. Calibrado del estadístico de contraste

Como bien hemos indicado previamente, en la práctica, no va a ser habitual poder estimar la distribución límite del proceso R_n^1. Ante esta problemática, surge la aproximación *bootstrap*, la cual, a partir de una muestra de datos, será capaz de proporcionar una aproximación de dicha distribución límite. La aproximación *bootstrap* consistirá en el cálculo del p-valor asociado a cierto funcional[11] escogido, para así decidir cuándo aceptar o rechazar la hipótesis nula del contraste.

Observación 2.5. Este funcional al que nos referimos formalmente, puede ser entendido como el estadístico que queramos utilizar, sea por ejemplo el presentado en (10) o el de tipo Kolmogorov-Smirnov.

Sea pues $(X_i^\star, Y_i^\star)$, con $1 \leq i \leq n$, una muestra artificial que se concretará más adelante, y sea $\widehat{\theta}^\star$ el estimador de $\theta^\star$ obtenido a partir de dicha muestra. Entonces, se define la *versión*

[11] Aplicación que asigna un número real a cada función de un cierto espacio funcional. Es decir, si $\mathcal{F}$ es un conjunto de funciones, un funcional es una aplicación $\Psi : \mathcal{F} \to \mathbb{R}$. Un ejemplo común es la integral definida de una función, que asocia a cada función su área bajo la curva.

bootstrap del proceso R_n^1 como:

$$R_n^{1\star}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{1}_{\{X_i^\star \leq x\}} \left(Y_i^\star - m(X_i^\star, \widehat{\theta}^\star) \right).$$

Sea $\Psi(R_n^1)$ un funcional continuo, diseñado para contrastar la hipótesis nula H_0. Llamemos además $c_{1-\alpha}$ al punto crítico correspondiente a un test de nivel α, es decir,

$$\mathbb{P}\left(\Psi(R_n^1) > c_{1-\alpha}\right) = \alpha,$$

entonces $c_{1-\alpha}$ se aproxima por $c_{1-\alpha}^\star$, el cual satisface que:

$$\mathbb{P}^\star\left(\Psi(R_n^{1\star}) > c_{1-\alpha}^\star\right) = \alpha,$$

donde $\mathbb{P}^\star$ denota la probabilidad inducida por el procedimiento *bootstrap*.

En la práctica, el valor $c_{1-\alpha}^\star$ se estima mediante el método *Monte Carlo*[12]:

1. Se generan B réplicas *bootstrap* independientes a partir de la muestra original.
2. En cada réplica j, se calcula el estadístico $D_j^* = \Psi(R_n^{1\star j})$.
3. Por último, se ordenan los valores obtenidos y se estima el punto crítico $c_{1-\alpha}^\star$ como aquel que se corresponde con la posición $D_{[B(1-\alpha)]}^\star$.

Podemos encontrar en Glasserman (2004) una revisión más en profundidad sobre el método *Monte Carlo* así como múltiples aplicaciones del mismo.

Existen diversas aproximaciones *bootstrap* para generar las réplicas *bootstrap*, como podría ser: el *bootstrap* por bloques, que se desarrolla en Politis y Romano (1994); el *bootstrap* circular que se aborda en Liu y Singh (1997); el *bootstrap* clásico que podemos visitar Efron (1979) para profundizar en él; el *bootstrap* suavizado que ha sido objeto de estudio en Silverman y Young (1987) o el *wild bootstrap* que será en el que nos centraremos y utilizaremos en nuestro posterior estudio de simulación así como en la aplicación a datos reales para llevar a cabo el calibrado del test presentado por Stute (1997).

2.3.1. Procedimiento *bootstrap* propuesto por Stute et al. (1998)

En Stute et al. (1998) se presenta un completo estudio de simulación así como una aplicación utilizando datos reales para ilustrar el comportamiento de los diferentes métodos *bootstrap*

[12]Técnica estadística que utiliza simulaciones aleatorias repetidas para estimar valores numéricos complejos. Se basa en generar múltiples escenarios posibles mediante muestreo aleatorio, lo que permite analizar sistemas donde interviene la incertidumbre o el azar.

expuestos. De dicho estudio se extrae como conclusión que el procedimiento más adecuado es el *wild bootstrap*. Además, de dicho análisis se deriva que los estadísticos de tipo Cramér-Von Mises y Kolmogorov-Smirnov tienen una potencia similar a la hora de realizar el contraste estudiado. Es por esto, que emplearemos la aproximación *wild bootstrap* y el estadístico de tipo Cramér-Von Mises presentado en (10) en nuestro posterior estudio de simulación y aplicación a datos reales.

El *wild bootstrap* consiste en tomar como muestra artificial lo siguiente:

$$X_i^\star = X_i, \quad Y_i^\star = m(X_i, \widehat{\theta}) + \varepsilon_i^\star, \quad i = 1, \ldots, n,$$

es decir, dejamos fijo los valores de las variables explicativas X_i y generamos una muestra *bootstrap* de la variable respuesta $Y_i^\star$ haciendo uso de la función de regresión predicha $m(X_i, \widehat{\theta})$ y añadiéndole un error *bootstrap* $\varepsilon_i^\star$, donde:

- $\varepsilon_i^\star = \widehat{\varepsilon}_i V_i^\star$
- $\widehat{\varepsilon}_i = Y_i - m(X_i, \widehat{\theta})$ son los residuos del modelo ajustado,
- $V_i^\star$ es una muestra aleatoria generada de tal forma que sus observaciones son independientes e idénticamente distribuidas tales que:

$$\mathbb{E}^\star[V_i^\star] = 0, \qquad \mathrm{Var}^\star[V_i^\star] = 1; \tag{11}$$

$$|V_i^\star| \leq c < \infty \tag{12}$$

  para cierto c finito, pues de esta forma conseguimos que la esperanza de los residuos siga siendo 0 y su varianza se mantenga constante.

Este esquema genera por tanto errores $\varepsilon_i^\star$ que conservan la esperanza y la varianza del modelo original:

$$\mathbb{E}^\star[\varepsilon_i^\star] = 0, \quad \mathrm{Var}^\star[\varepsilon_i^\star] = \mathrm{Var}[\widehat{\varepsilon}_i].$$

Nuestro objetivo ahora será probar que en efecto el *wild bootstrap* supone una aproximación válida de R_n^1 bajo la hipótesis nula $H_0 : m \in \mathcal{M}_\theta$. Es decir, H_0 afirma que los datos satisfacen el modelo

$$Y_i = m(X_i, \theta) + \varepsilon_i,$$

donde ε_i representa un error no observable tal que

$$\mathbb{E}[\varepsilon_i \mid X_i] = 0, \quad 1 \leq i \leq n.$$

Presentamos ahora una suposición que será necesaria para el teorema que nos proporcionará el comportamiento asintótico de $R_n^{1\star}$ con la aproximación *bootstrap* propuesta.

SUPUESTO 3. La matriz $p \times p$ $A = \mathbb{E}[m(X)m^T(X)]$ existe y es definida positiva. Además, $M_{jk} = \mathbb{E}[m_j(X)m_k(X)\varepsilon^2]$, existe para $1 \leq j, k \leq p$.

Se define a partir de ello:

$$\mathbf{M} = (M_{jk})_{1 \leq j,k \leq p}.$$

Teorema 2.6. *Sea cierto el Supuesto 3. Entonces, bajo H_0, se tiene con probabilidad 1 que,*

$$R_n^{1\star} \to R_\infty^{1\star}$$

en distribución en el espacio de Skorokhod $D[-\infty, \infty]$. Donde $R_\infty^{1\star}$ y R_∞^1 tienen la misma distribución y la muestra bootstrap se genera vía wild bootstrap bajo las condiciones (11) *y* (12).

Demostración. La demostración se puede encontrar en el apéndice de Stute et al. (1998). □

Este resultado quiere decir que, bajo las hipótesis adecuadas, que en la práctica cumpliremos, el proceso *bootstrap* $R_n^{1\star}$ y el proceso R_n^1 tienen la misma distribución asintótica. Hemos probado pues que R_n^1 puede ser aproximado en distribución mediante el *wild bootstrap*. Este proceso sirve como base para múltiples estadísticos que tienen por objetivo testear la validez de un modelo de regresión paramétrico.

2.4. Implementación del contraste propuesto

Vamos a realizar una breve explicación de cómo habría que proceder en la práctica a la hora de implementar el contraste propuesto por Stute (1997), incluyendo la aproximación *wild bootstrap* expuesta en la sección anterior:

1. Partimos de nuestra muestra (X_i, Y_i), asociada a un modelo de regresión paramétrico de la forma:

   $$Y_i = m(X_i, \theta) + \varepsilon$$

   y calculamos $\widehat{\theta}$ que será el estimador de mínimos cuadrados de θ.

2. Con esto podemos calcular los residuos del modelo, $\widehat{\varepsilon_i}$, que nos permitirán calcular R_n^1 y el posterior estadístico, que en nuestro caso será:

   $$W_n^2 = \int_{\mathbb{R}^d} [R_n^1(x)]^2 \, dF_n(x) = \frac{1}{n} \sum_{i=1}^n R_n^1(x_i)^2.$$

3. Generamos B remuestras *bootstrap* $(X_i, Y_i^\star)$ como hemos explicado en la Subsección 2.3.1, para luego calcular el respectivo estimador $\widehat{\theta^\star}$, es decir:

3.1. Obtenemos nuestra muestra *bootstrap* $Y_i^\star = m(X_i, \widehat{\theta}) + \varepsilon_i^\star$, donde $\varepsilon_i^\star = \widehat{\varepsilon}_i V_i^\star$, siendo $\widehat{\varepsilon}_i$ los residuos del modelo ajustado y $V_i^\star$ la muestra aleatoria de observaciones independientes e idénticamente distribuidas satisfaciendo (11) y (12).

3.2. Para cada muestra $(X_i, Y_i^\star)$ calculamos el estimador de mínimos cuadrados $\widehat{\theta}^\star$.

4. Para cada una de estas muestras obtenemos los respectivos residuos del modelo *bootstrap* y calculamos el valor de $R_n^{1\star}$ que nos permite obtener el valor del estadístico *bootstrap*:

$$W_n^{2,\star} = \int_{\mathbb{R}^d} [R_n^{1\star}(x)]^2 \, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} R_n^{1\star}(x_i)^2.$$

5. Por último, calculamos el p-valor asociado al contraste para nuestra muestra inicial (X_i, Y_i) como sigue:

$$\text{p-valor} = \frac{1}{B} \sum_{k=1}^{B} \mathbf{1}_{\{W_n^2 < W_n^{2,\star,(k)}\}},$$

donde $W_n^{2,\star,(k)}$ representa el valor del estadístico de la *k-ésima* muestra *bootstrap* generada.

En el Capítulo 3 nuestro objetivo será implementar en R el contraste expuesto en este Capítulo 2 para poder llevar a cabo un estudio de simulación y comprobar su buen comportamiento en la práctica mostrando evidencias empíricas.

Capítulo 3

Estudio de simulación

El objetivo de este capítulo será ilustrar el comportamiento en la práctica del test propuesto por Stute (1997) gracias a un estudio de simulación utilizando el método de Monte Carlo. Es decir, la idea será utilizar datos generados de manera artificial para comprobar si el contraste presentado en el Capítulo 2 respeta el nivel de significación bajo la hipótesis nula, así como evaluar su potencia bajo la hipótesis alternativa. Para ello, detallaremos la implementación en  del contraste estudiado en Stute (1997) y luego procederemos al estudio y extracción de conclusiones sobre el mismo. En el Anexo I se encuentra el código completo empleado para la obtención de los resultados que se presentarán a lo largo de este capítulo.

Partimos pues de una muestra de tamaño n de las variables $X, Y \in \mathbb{R}$, que denotaremos por `x` e `y`, respectivamente. Nuestro objetivo será contrastar la hipótesis nula

$$H_0 : Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

frente a alternativas no paramétricas. El primer paso será estimar el modelo bajo la hipótesis nula así como los residuos y las predicciones que se derivan del mismo. El código utilizado es el siguiente:

```
modelo <- lm(y ~ x) # estima el modelo
y.hat <- fitted(modelo) # predicciones del modelo
residuos <- resid(modelo) # residuos del modelo
```

A continuación, calculamos el estadístico de contraste observado. Para ello ordenamos los valores de la muestra `x` de manera creciente para calcular correctamente el estadístico de contraste W_n^2 definido en (10).

```
orden <- order(x) # calcula el indice que ordena los valores de la variable
    explicativa de manera creciente
x.ord <- x[orden] # ordena el vector x en orden creciente
```

```
res.ord <- residuos[orden] # reordena los residuos
n=length(y)
Rn1 <- cumsum(res.ord) / sqrt(n) # calculamos Rn^1(Xi)
Wn2 <- sum(Rn1^2) / n # calculamos Wn^2
```

Ahora, procedemos a implementar el procedimiento *bootstrap* para aproximar la distribución del estadístico y poder decidir si aceptamos o no la hipótesis nula. Para ello, generamos la muestra V_i^*, siguiendo las ideas propuestas por Stute et al. (1998); que proceden de una distribución discreta que toma dos posibles valores, de manera que se satisfacen las condiciones (11) y (12). En concreto, consideramos una distribución que toma los valores $-\frac{\sqrt{5}-1}{2}$ y $\frac{\sqrt{5}+1}{2}$ con probabilidad $\frac{\sqrt{5}+1}{2\sqrt{5}}$ y $\frac{\sqrt{5}-1}{2\sqrt{5}}$, respectivamente. Tendremos que calcular diferentes remuestras *bootstrap* $\{Y_1^\star, \ldots, Y_n^\star\}$ para así calcular el valor del estadístico *bootstrap* $W_n^{2,\star}$ para cada una de ellas. El código utilizado para esta parte es el siguiente:

```
B=499
Wn2.bootstrap <- numeric(B)
for(j in 1:B){
  P=c((sqrt(5)+1)/(2*sqrt(5)),(sqrt(5)-1)/(2*sqrt(5)))
  z<-c(-(sqrt(5)-1)/(2),(sqrt(5)+1)/(2))
  V<-runif(n,min=0,max=1)
  for (k in 1:n) { if (V[k]<P[1]) {V[k]=z[1]}
    else {V[k]=z[2]}
    } # genera la muestra Vi*
  y.boot <- y.hat + residuos * V # muestra bootstrap
  modelo.boot <- lm(y.boot ~ x) # Estima modelo bootstrap
  residuos.boot <- resid(modelo.boot)
  # Calculamos el estadistico boostrap:
  res.ord.boot <- residuos.boot[order(x)] #ordenamos residuos
  Rn1.boot <- cumsum(res.ord.boot) / sqrt(n)
  Wn2.bootstrap[j] <- sum(Rn1.boot^2) / n
  }
```

Por último, procedemos a realizar el calibrado del test, es decir, vamos a comparar el estadístico observado con las remuestras *bootstrap* del estadístico para así calcular el p-valor o nivel crítico de la siguiente forma:

```
p.valor <- sum(Wn2.bootstrap > Wn2)/n
```

Ahora que ya hemos presentado la base del código que vamos a utilizar, podemos comenzar con el estudio de simulación. Presentaremos diversos casos prácticos con el objetivo de ver cómo se comporta el procedimiento introducido por Stute (1997) ante diferentes propiedades del modelo de regresión que estemos utilizando.

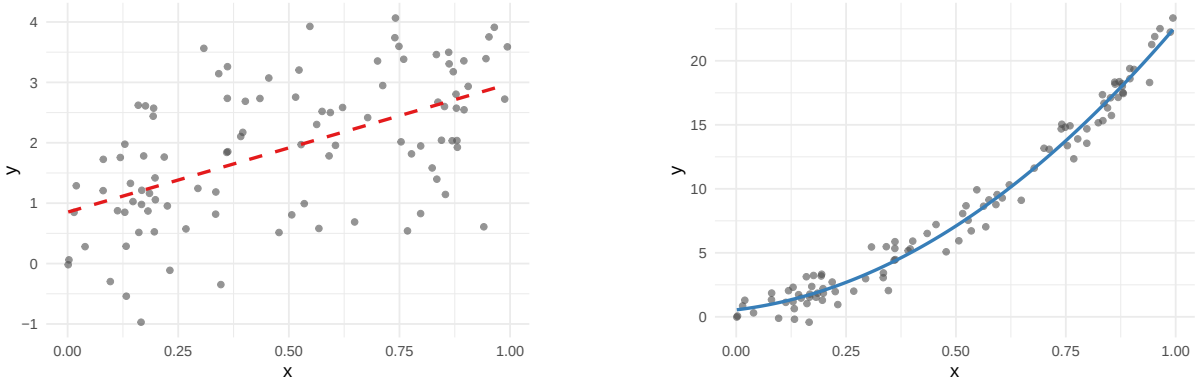
3.1. Escenario 1

En primer lugar, consideraremos el siguiente modelo de regresión:

$$\text{Modelo 1:}\quad Y = 1 + 2X + aX^2 + \varepsilon,$$

donde la variable explicativa X sigue una distribución uniforme en el intervalo $[0, 1]$ y el error ε sigue una distribución normal de media 0 y desviación típica σ. Nótese que nuestro objetivo será contrastar la hipótesis nula de que el modelo lineal simple es correcto, por lo tanto, el parámetro a representa la desviación del modelo de la H_0.

En la Figura 3.1 hemos representado dos muestras del Modelo 1 obtenidas para un tamaño de muestra $n = 100$, $\sigma = 1$ y coeficiente $a = 0$ y $a = 20$ respectivamente, para así ilustrar el efecto que tiene el parámetro a en las muestras que estamos simulando.



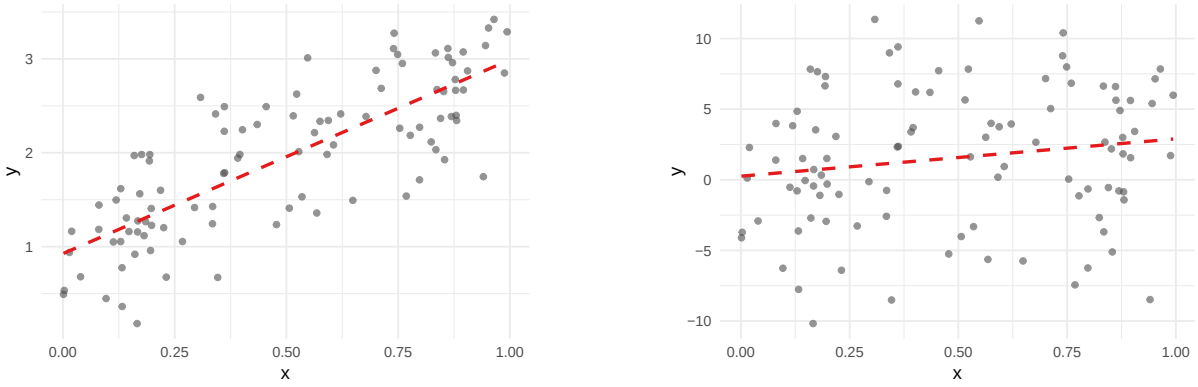
(a) $a = 0$.

(b) $a = 20$.

Figura 3.1: Muestras de datos simulados del Modelo 1 para un tamaño de muestra $n = 100$, $\sigma = 1$ y $a = 0$ (Figura (a)) y $a = 20$ (Figura (b)). La muestra de la Figura (a) se halla bajo H_0 mientras que el de la Figura (b) se halla bajo H_a.

De manera análoga, en la Figura 3.2 hemos representado dos muestras del Modelo 1 obtenidas para un tamaño de muestra $n = 100$, coeficiente $a = 0$ y desviaciones típicas $\sigma = 0.5$ y $\sigma = 5$ respectivamente. Podemos observar con facilidad el efecto que tiene sobre una muestra de datos una alta variabilidad en el error. Es decir, a medida que aumenta la varianza del error, los puntos de la muestra se alejarán más del modelo poblacional.

Para el Modelo 1, la hipótesis nula vendrá representada por el modelo lineal simple y, por la contra, como hipótesis alternativa tendremos cualquier otra alternativa no paramétrica. Hemos establecido como niveles de significación los tres más habituales, es decir, $\alpha = (0.1, 0.05, 0.01)$;



(a) $\sigma = 0.5$.

(b) $\sigma = 5$.

Figura 3.2: Muestras de datos simulados del Modelo 1 para un tamaño de muestra $n = 100$, coeficiente $a = 0$ y desviación típica del error $\sigma = 0.5$ (Figura (a)) y $\sigma = 5$ (Figura (b)).

cuatro tamaños de muestra n para cada valor de a; y tres valores de σ. Los resultados de nuestro estudio de simulación están recogidos en la Tabla 3.1, en la que se presenta la proporción de rechazo asociada al test propuesto por Stute (1997) asociada al Modelo 1. Esperamos encontrarnos que, bajo la hipótesis nula, es decir, cuando $a = 0$, el test respete los niveles de significación y a mayor tamaño muestral, se aproxime aún más a estos valores. A medida que aumentamos el parámetro a, es decir, según nos desviamos de la hipótesis alternativa, el test debería rechazar la hipótesis nula con probabilidades cada vez más cercanas a 1 con el aumento del tamaño de muestra. El aumento de σ, es decir, de la variabilidad del error, esperamos que suponga cierta pérdida de potencia del test a la hora de rechazar la hipótesis nula cuando estamos bajo la hipótesis alternativa.

A la vista de la Tabla 3.1, centrémonos primero en qué ocurre cuando tomamos $\sigma = 1$. Cuando $a = 0$, nos hallamos bajo la hipótesis nula de linealidad, es por ello que la probabilidad de rechazo es cercana a los niveles de significación y a medida que aumentamos el tamaño de la muestra es aún más precisa. Esto resulta lógico, pues a mayor tamaño muestral nuestra aproximación *bootstrap* estima mejor la distribución del estadístico de contraste bajo la hipótesis nula.

Para $a = 1$ nos encontramos bajo la hipótesis alternativa y, a medida que aumenta el tamaño de muestra, también lo hace la probabilidad de rechazo. Tomando $a = 2$, para tamaños muestrales elevados se observa una gran mejoría a la hora de detectar que estamos bajo la hipótesis alternativa y, por tanto, rechazar la hipótesis nula. Para cuando imponemos $a = 5$, obtenemos que la probabilidad de rechazo converge a 1.

A modo de resumen, el contraste respeta los niveles de significación establecidos bajo la hipótesis nula y es capaz de distinguir cuándo nos encontramos bajo la hipótesis alternativa y

a	n	$\sigma = 1$			$\sigma = 2$			$\sigma = 5$		
		α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01
0	50	0.106	0.038	0.008	0.106	0.038	0.008	0.106	0.038	0.008
	100	0.098	0.053	0.015	0.098	0.053	0.015	0.098	0.053	0.015
	250	0.113	0.063	0.020	0.113	0.063	0.020	0.113	0.063	0.020
	500	0.101	0.041	0.007	0.101	0.041	0.007	0.101	0.041	0.007
1	50	0.125	0.068	0.007	0.104	0.054	0.005	0.104	0.044	0.005
	100	0.177	0.101	0.020	0.121	0.056	0.015	0.096	0.050	0.014
	250	0.294	0.207	0.066	0.170	0.092	0.033	0.118	0.065	0.019
	500	0.436	0.311	0.134	0.180	0.114	0.041	0.112	0.060	0.011
2	50	0.216	0.128	0.027	0.125	0.068	0.007	0.106	0.048	0.004
	100	0.387	0.261	0.092	0.177	0.101	0.020	0.112	0.054	0.016
	250	0.707	0.591	0.354	0.294	0.207	0.066	0.143	0.075	0.025
	500	0.928	0.880	0.687	0.436	0.311	0.134	0.152	0.089	0.029
5	50	0.734	0.592	0.296	0.294	0.184	0.049	0.125	0.068	0.007
	100	0.952	0.913	0.769	0.528	0.392	0.169	0.177	0.101	0.020
	250	1.000	1.000	0.993	0.855	0.770	0.576	0.294	0.207	0.066
	500	1.000	1.000	1.000	0.986	0.972	0.910	0.436	0.311	0.134

Tabla 3.1: Proporciones de rechazo asociadas al contraste propuesto por Stute (1997), para diferentes valores de los parámetros n (tamaño de muestra), σ (desviación típica del error) y a (parámetro de desviación de la hipótesis nula) para el Modelo 1.

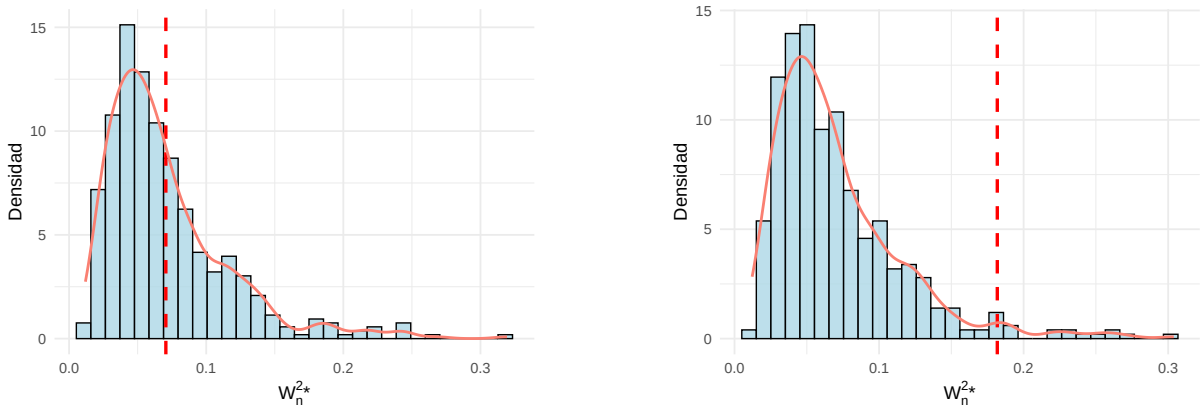
cuándo nos desviamos considerablemente de la hipótesis nula. Al verse aumentado a, aumenta la desviación del modelo de la hipótesis nula, aumentando consecuentemente la potencia del test.

Hemos obtenido también que, bajo la hipótesis nula, a pesar de aumentar la varianza del error, las probabilidades de rechazo se mantienen prácticamente invariantes. Sin embargo, bajo la hipótesis alternativa, al aumentar σ, es más complicado detectar si las posibles desviaciones de los residuos se deben a una mala especificación del modelo o simplemente al azar, y consecuentemente el test se vuelve menos sensible.

Esta pérdida de potencia se hace mucho más notable cuando establecemos $\sigma = 5$, de hecho, para $a = 1$, actúa como si estuviésemos bajo la hipótesis nula, esto es debido a que la componente cuadrática no tiene suficiente peso en los datos en comparación con la variabilidad del error. Para $a = 2$ y $a = 5$ sí que rechaza con mayor frecuencia pero como es lógico la pérdida de potencia provoca que la probabilidad de rechazo no sea tan elevada como en situaciones de menor variabilidad.

En la Figura 3.3 hemos representado dos histogramas asociados a remuestras *bootstrap* del

estadístico de contraste dado en (10) junto con una estimación no paramétrica de la función de densidad correspondiente. Lo hemos hecho para $B = 499$ muestras *bootstrap* para el Modelo 1 con tamaño de muestra 500, siendo además $\sigma = 1$ y $a = 0$ y $a = 1$, respectivamente. La línea roja discontinua se corresponde con el valor observado del estadístico de contraste.



(a) $a = 0$. (b) $a = 1$.

Figura 3.3: Histogramas asociados a remuestras *bootstrap* del estadístico de contraste junto con una estimación no paramétrica de la función de densidad correspondiente. Se ha tomado $B = 499$ muestras *bootstrap* para modelos de tamaño 500, con $\sigma = 1$ y con coeficiente $a = 0$ Figura (a) y $a = 1$ Figura (b).

A la vista de la Figura 3.3, observamos que, a medida que nos alejamos de la hipótesis nula, aumentando el valor de a, el estadístico observado toma un mayor valor, la línea roja discontinua se desplaza a la derecha. Esto supone que cada vez un número mayor de réplicas *bootstrap* del estadístico caen a la izquierda del estadístico observado y, por tanto, el p-valor es cada vez más pequeño, provocando que se rechace con más frecuencia la hipótesis nula.

3.2. Escenario 2

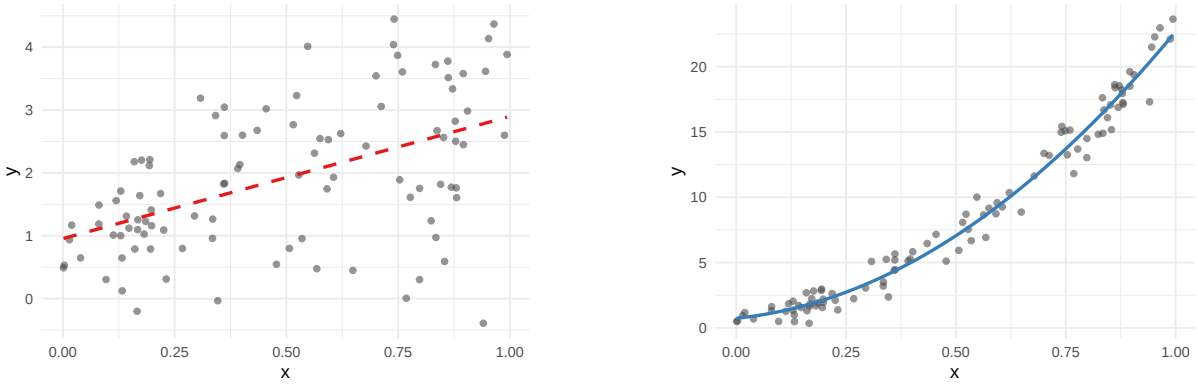
En segundo lugar, vamos a considerar el siguiente modelo de regresión:

$$\text{Modelo 2:}\quad Y = 1 + 2X + aX^2 + (X + 0.5)\varepsilon,$$

donde la variable explicativa X, de nuevo, sigue una distribución uniforme en el intervalo $[0, 1]$ y el error ε sigue una distribución normal de media 0 y desviación típica σ. La diferencia con respecto al Modelo 1, es que ahora los errores son heterocedásticos. Nuestro objetivo será, por tanto, comprobar si el contraste respeta el nivel de significación a pesar de la pérdida de homocedasticidad de los errores. Asimismo, también estamos interesados en conocer cómo la falta

de homocedasticidad afecta a la potencia del test. Nótese que, nuevamente, la hipótesis nula a contrastar es el modelo lineal simple frente a alternativas no paramétricas.

En la Figura 3.4 hemos representado dos muestras de datos simulados del Modelo 2 obtenidas para un tamaño de muestra $n = 100$, $\sigma = 1$ y coeficiente $a = 0$ y $a = 20$, respectivamente. Comparando con la Figura 3.1, podemos observar el efecto que tiene sobre las observaciones la presencia de estos errores heterocedásticos: los valores de la muestra estarán más próximos al modelo poblacional para valores bajos de la variable explicativa mientras que estarán más desviados a medida que aumente el valor de la covariable.



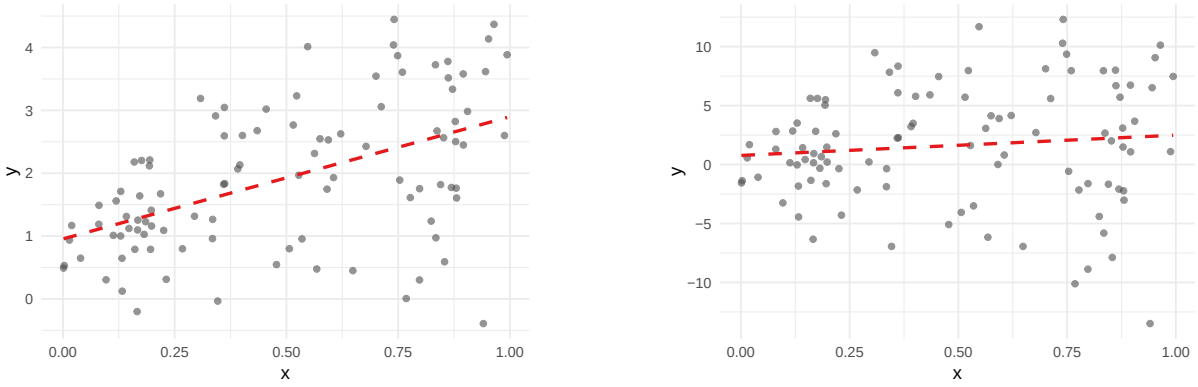
(a) $a = 0$. (b) $a = 20$.

Figura 3.4: Muestras de datos simulados del Modelo 2 para tamaño de muestra $n = 100$, $\sigma = 1$ y $a = 0$ (Figura (a)) y $a = 20$ (Figura (b)). La Figura (a) se halla bajo la hipótesis nula mientras que la Figura (b) se halla bajo la hipótesis alternativa.

Análogamente, en la Figura 3.5 hemos representado dos muestras de datos simulados del Modelo 2 obtenidas para un tamaño de muestra $n = 100$, coeficiente $a = 20$ y desviaciones típicas $\sigma = 1$ y $\sigma = 5$, respectivamente. Observamos que, para valores cercanos a 1 de la variable explicativa X, los valores de la variable respuesta Y se dispersan más al aumentar el valor de σ.

Procedemos pues a la realización del estudio de simulación sobre el Modelo 2. Las probabilidades de rechazo asociadas al test propuesto por Stute (1997) han sido recogidas en la Tabla 3.2. y, de nuevo, establecemos los mismos valores de n, a, σ y α. La hipótesis nula a contrastar vuelve a ser la de linealidad, que se cumple si $a = 0$.

A la vista de la Tabla 3.2, al igual que con el Modelo 1, cuando $a = 0$, es decir, cuando nos hallamos bajo la hipótesis nula, el contraste respeta los niveles de significación. De nuevo, bajo la hipótesis nula, los resultados se mantienen prácticamente invariantes ante el aumento de la variabilidad del error.



(a) $\sigma = 1$.

(b) $\sigma = 5$.

Figura 3.5: Muestras de datos simulados del Modelo 2 para tamaño de muestra $n = 100$, coeficiente $a = 0$ y $\sigma = 1$ (Figura (a)) y $\sigma = 5$ (Figura (b)).

a	n	$\sigma = 1$			$\sigma = 2$			$\sigma = 5$		
		α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01
0	50	0.090	0.039	0.011	0.090	0.039	0.011	0.090	0.039	0.011
	100	0.098	0.043	0.005	0.098	0.043	0.005	0.098	0.043	0.005
	250	0.096	0.048	0.008	0.096	0.048	0.008	0.096	0.048	0.008
	500	0.107	0.065	0.015	0.107	0.065	0.015	0.107	0.065	0.015
1	50	0.134	0.057	0.015	0.103	0.051	0.010	0.098	0.046	0.010
	100	0.163	0.087	0.021	0.111	0.055	0.008	0.091	0.043	0.006
	250	0.246	0.158	0.046	0.130	0.069	0.019	0.106	0.047	0.011
	500	0.410	0.293	0.123	0.187	0.115	0.030	0.118	0.068	0.018
2	50	0.216	0.129	0.030	0.134	0.057	0.015	0.103	0.047	0.010
	100	0.342	0.249	0.081	0.163	0.087	0.021	0.101	0.050	0.007
	250	0.649	0.524	0.264	0.246	0.158	0.046	0.113	0.062	0.019
	500	0.901	0.831	0.612	0.410	0.293	0.123	0.162	0.089	0.027
5	50	0.661	0.533	0.260	0.274	0.175	0.043	0.134	0.057	0.015
	100	0.935	0.881	0.681	0.456	0.349	0.136	0.163	0.087	0.021
	250	1.000	1.000	0.997	0.811	0.722	0.473	0.246	0.158	0.046
	500	1.000	1.000	1.000	0.978	0.960	0.849	0.410	0.293	0.123

Tabla 3.2: Proporciones de rechazo asociadas al contraste propuesto por Stute (1997), para diferentes valores de los parámetros n (tamaño de muestra), σ (desviación típica del error) y a (parámetro de desviación de la hipótesis nula) para el Modelo 2.

Con $a = 1$, el contraste es capaz de detectar que nos hallamos bajo la hipótesis alternativa,

presentando para $\sigma = 1$ y $n = 500$ una probabilidad de rechazo similar a la que obteníamos en la Tabla 3.1. Al aumentar σ, el test pierde potencia, de hecho, para $\sigma = 5$, se comporta como si estuviésemos bajo la hipótesis nula como ya habíamos observado en el Escenario 1.

Con el aumento del parámetro a, las probabilidades de rechazo son cada vez mayores y para cuando establecemos $a = 5$, la probabilidad de rechazo para $\sigma = 1$ converge a 1 a medida que aumentamos el tamaño de muestra. Asímismo, si establecemos $\sigma = 2$ el test rechaza con probabilidad muy próxima a 1 y con $\sigma = 5$ encontramos probabilidades de rechazo elevadas a pesar de la alta varianza del error.

En la Figura 3.6 hemos representado la estimación de la función de potencia del contraste para el Modelo 2. Hemos tomado como parámetros del modelo: $n = 100$, $\sigma = 1$ y $\alpha = 0.1$. Los valores de a fueron tomados con saltos de 0.2 en el intervalo $[-5, 5]$ y hemos realizado los cálculos basándonos en $M = 100$ muestras. La línea discontinua se corresponde con el nivel de significación establecido. Como podemos observar, para valores alejados de a del 0, la probabilidad de rechazo es muy elevada y, a medida que nos aproximamos al valor $a = 0$, el test es capaz de respetar el nivel de significación establecido.

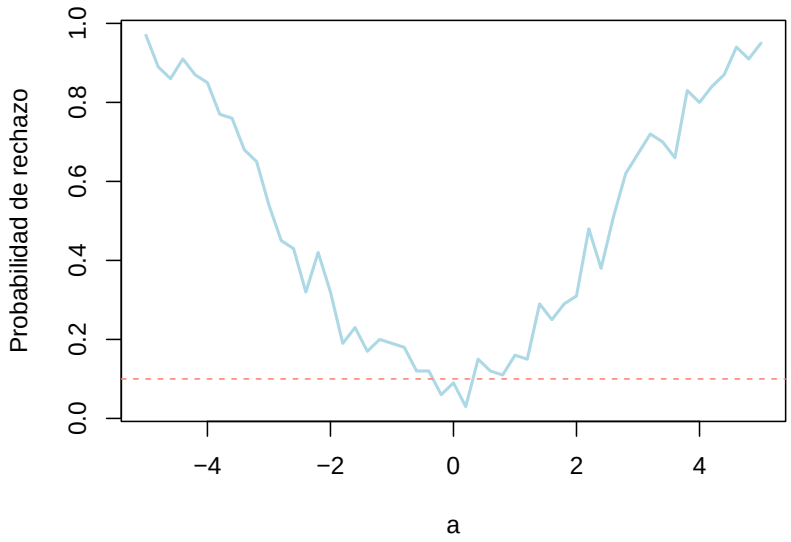


Figura 3.6: Función de potencia del test propuesto por Stute (1997) estimada para el Modelo 2, considerando un tamaño de muestra $n = 100$, $\sigma = 1$ para diferentes valores del parámetro a. La línea discontinua se corresponde con el nivel de significación considerado que es $\alpha = 0.1$.

La conclusión que extraemos del estudio realizado sobre este escenario es clara; aunque se

pierda la hipótesis clásica de homocedasticidad de los errores, el contraste sigue respetando los niveles de significación bajo la hipótesis nula y no sufre pérdida de potencia bajo la hipótesis alternativa; es capaz de diferenciar cuándo estamos bajo la hipótesis nula y cuándo estamos bajo la alternativa. Esto mismo concuerda con el estudio de simulación realizado en Stute et al. (1998), el cual concluía que la aproximación *wild bootstrap* seguía siendo válida para un modelo con errores heterocedásticos.

3.3. Escenario 3

Por último, consideraremos el siguiente modelo de regresión con dos variables explicativas:

$$\text{Modelo 3:}\quad Y = 2 + 5X_1 - X_2 + aX_1X_2 + \varepsilon,$$

que ha sido previamente considerado en Stute et al. (1998). El vector de variables (X_1, X_2) sigue una distribución uniforme en el cuadrado unidad $[0,1] \times [0,1]$ y las variables X_1 y X_2 son independientes. El error ε sigue una distribución normal de media 0 y desviación típica σ, y además, nótese que habrá interacción entre las variables explicativas cuando $a \neq 0$. Nuestro objetivo será contrastar la hipótesis nula de que el modelo lineal bivariante es correcto, por lo que a representa, una vez más, la desviación del modelo de la hipótesis nula.

Como estamos ante un modelo con más de una variable explicativa, será necesario implementar una pequeña modificación sobre el código que presentamos al comienzo de este capítulo. Hemos implementado el contraste de tal manera que el código siga funcionando independientemente del número de variables explicativas consideradas. El primer cambio realizado será en el cálculo del estadístico de contraste observado que sería el siguiente:

```
for (w in 1:n) {
   indicador=numeric(n)
   for(h in 1:p){ # p representa el numero de covariables consideradas
     indicador=indicador+as.numeric(X[,h]<=X[w,h])
   }
   Rn1[w] <- sum( (indicador==p) * residuos) / sqrt(n)
 }
```

donde la matriz X que aparece se define previamente como aquella que contiene los valores de las variables explicativas consideradas por columnas. El segundo y último cambio a realizar será a la hora de calcular el estadístico de contraste de las remuestras *bootstrap*, la modificación es la que sigue:

```
for (d in 1:n) {
      indicadora=numeric(n)
      for(l in 1:p){
```

```
        indicadora=indicadora+as.numeric(X[,1]<=X[d,1])
      }
      Rn1.boot[d] <- sum( (indicadora==p) * residuos.boot) / sqrt(n)
    }
```

En el Anexo I de este documento se puede consultar el código completo correspondiente a este Escenario 3.

Procedemos entonces a realizar el estudio de simulación sobre el Modelo 3. Es de esperar que el test siga respetando los niveles de significación bajo la hipótesis nula, es decir, cuando $a = 0$; que sea capaz de detectar cuándo nos desviamos de la hipótesis nula con el aumento del parámetro a y que con el aumento de la variabilidad del error, es decir, el aumento de σ, se pierda cierta potencia en el test.

En la Tabla 3.3 presentamos la proporción de rechazo asociada al test propuesto por Stute (1997) para el Modelo 3 considerando distintos valores de los parámetros n, σ y a. A la vista de la Tabla 3.3, como era de esperar, bajo la hipótesis nula, es decir, cuando $a = 0$, el test respeta el nivel de significación establecido. El aumento en la variabilidad del error bajo la hipótesis nula, de nuevo, prácticamente no supone ningún cambio en los resultados obtenidos.

Adicionalmente, observamos que, como era de esperar, el aumento de la variabilidad del error cuando consideramos modelos bajo la hipótesis alternativa provoca cierta pérdida de potencia en el test. Esta pérdida de potencia se hace muy notable cuando imponemos $\sigma = 5$, presentando probabilidades de rechazo de H_0 mucho más bajas que los respectivos modelos con menor varianza del error.

Por otra parte, a medida que aumenta la desviación del modelo de la hipótesis nula, es decir, a medida que aumenta el valor del parámetro a, observamos como las probabilidades de rechazo de la hipótesis nula son cada vez mayores. De hecho, para cuando establecemos $a = 25$, la probabilidad de rechazo converge a 1 con el aumento del tamaño de muestra, independientemente de la varianza del error considerada.

Hemos considerado también modelos con errores heterocedásticos. A la vista de la Tabla 3.3, es fácil comprobar que el contraste propuesto por Stute (1997) sigue funcionando correctamente incluso cuando se pierde la hipótesis clásica de homocedasticidad de los errores, es decir, respeta el nivel de significación bajo la hipótesis nula y rechaza con alta probabilidad cuando nos hallamos bajo la hipótesis alternativa.

De hecho, podemos observar que las probabilidades de rechazo bajo la hipótesis alternativa son mucho mayores para los modelos con errores heterocedásticos que para los modelos con errores homocedásticos pero de alta variabilidad ($\sigma = 5$). La potencia del test en este caso es muy similar a cuando consideramos $\sigma = 2$; sin embargo, no es tan elevada como para $\sigma = 1$.

a	n	$\sigma = 1$			$\sigma = 2$			$\sigma = 5$			$\sigma = 2(0.5 + X_1)$		
		α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01	α=0.1	α=0.05	α=0.01
0	50	0.089	0.046	0.006	0.089	0.046	0.006	0.089	0.046	0.006	0.109	0.050	0.006
	100	0.115	0.061	0.010	0.115	0.061	0.010	0.115	0.061	0.010	0.094	0.033	0.008
	250	0.100	0.049	0.006	0.100	0.049	0.006	0.100	0.049	0.006	0.086	0.039	0.008
	500	0.103	0.053	0.008	0.103	0.053	0.008	0.103	0.053	0.008	0.115	0.045	0.013
5	50	0.568	0.381	0.109	0.204	0.109	0.020	0.117	0.060	0.009	0.234	0.131	0.026
	100	0.894	0.827	0.538	0.376	0.261	0.082	0.148	0.080	0.016	0.405	0.269	0.090
	250	1.000	1.000	0.998	0.816	0.692	0.452	0.239	0.156	0.047	0.799	0.691	0.419
	500	1.000	1.000	1.000	0.986	0.957	0.862	0.406	0.274	0.117	0.978	0.953	0.849
15	50	0.999	0.988	0.888	0.846	0.719	0.334	0.260	0.142	0.035	0.853	0.735	0.388
	100	1.000	1.000	1.000	0.995	0.990	0.932	0.506	0.354	0.130	0.999	0.994	0.937
	250	1.000	1.000	1.000	1.000	1.000	1.000	0.932	0.868	0.650	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000	0.148	0.080	0.016	1.000	1.000	1.000
25	50	1.000	0.999	0.981	0.994	0.971	0.784	0.568	0.381	0.109	0.992	0.974	0.795
	100	1.000	1.000	1.000	1.000	1.000	1.000	0.894	0.827	0.538	1.000	1.000	1.000
	250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.3: Proporciones de rechazo asociadas al contraste propuesto por Stute (1997), para diferentes valores de los parámetros n (tamaño de muestra), σ (desviación típica del error) y a (parámetro de desviación de la hipótesis nula) para el Modelo 3.

En definitiva, la metodología introducida por Stute (1997), cuando consideramos modelos de regresión lineales múltiples, sigue respetando el nivel de significación, es decir, las probabilidades de rechazo se ajustan al nivel de significación establecido. Asimismo, el contraste es capaz de rechazar la hipótesis nula cuando consideramos modelos bajo la hipótesis alternativa. Hemos comprobado además que el test se sigue comportando bien ante la pérdida de la hipótesis clásica de homocedasticidad de los errores, concordando con las conclusiones extraídas del estudio de simulación realizado en Stute et al. (1998).

Cabe destacar que la implementación del contraste considerado se hace computacionalmente más costosa a medida que aumenta el número de variables explicativas consideradas, es lo que se conoce como el desastre de la dimensión. Este hecho se ha apreciado al considerar el Escenario 3, donde los tiempos de ejecución del test son mucho mayores a los de los Escenarios 1 y 2.

Capítulo 4

Aplicación a datos reales

El Real Club Deportivo de La Coruña (https://www.rcdeportivo.es/) es uno de los clubes de fútbol más grandes de España y es el club gallego más laureado de la historia. Forma parte de un selecto grupo de equipos dado que es uno de los nueve clubes españoles que han logrado conquistar el Campeonato Liguero Español. Recientemente, el 19 de mayo de 2025, se cumplieron 25 años de la famosa liga del *SuperDépor* que entrenaba Javier Irureta, es por ello que, a modo de homenaje, centraremos nuestra aplicación a datos reales en el equipo que maravilló a Europa durante una década.

Nuestro objetivo en este capítulo será ajustar un modelo de regresión lineal múltiple para intentar explicar los puntos por partido que un equipo puede lograr a lo largo de una temporada, para luego testear con el contraste presentado en el Capítulo 2 que en efecto la hipótesis de linealidad es correcta.

Para la obtención de la base de datos hemos consultado la página web **https://fbref.com/**, creada por *Sports-Reference* y lanzada en Junio de 2018, que recoge una amplia gama de datos con información sobre los equipos de la élite del fútbol, con cobertura de la liga para seis naciones: Inglaterra, Francia, España, Italia, Alemania y los Estados Unidos. Esta extensa serie de datos abarca múltiples aspectos que se ven reflejados en un terreno de juego: desde facetas básicas, como los goles que un equipo anota o el número de pases que un equipo realiza; hasta matices más complejos como el porcentaje de posesión en el último tercio del campo o la altura media con la que se realizan los robos de balón. Hemos tomado específicamente las temporadas 1999/00, 2000/01 y 2001/02, que se corresponden con las tres mejores posiciones en liga del *SuperDépor* de Javier Irureta. Al tratarse de campañas en las que no se contaba con las tecnologías de las que hoy en día sí se dispone, no tendremos datos tan específicos como en la actualidad.

En el Anexo I hemos incluido el código de  empleado para la obtención de los resultados que se presentan en este capítulo. Asimismo, en el Anexo II de este documento, se puede consultar

la versión completa de la base de datos que vamos a utilizar, que recoge las siguientes variables:

- Los goles a favor que denotaremos por (`GF`), es decir, los tantos que un equipo marca a lo largo de una temporada.
- Los goles en contra que denotaremos por (`GC`), es decir, los tantos que un equipo recibe en toda una campaña.
- Las tarjetas amarillas que denotaremos por (`TA`), es decir, el número de tarjetas amarillas que un equipo recibe a lo largo de una temporada.
- Las tarjetas rojas que denotaremos por (`TR`), es decir, el total de tarjetas rojas que un equipo recibe a lo largo de una campaña.
- Las porterías a cero que denotaremos por (`PaC`), es decir, el número de encuentros en los que el equipo no concedió ningún gol.
- Los tiros a puerta a favor que denotaremos por (`TaPF`), es decir, los lanzamientos a puerta que realiza un equipo en toda una temporada.
- Los tiros a puerta en contra que denotaremos por (`TaPC`), es decir, los disparos a puerta que concede un equipo en toda una campaña.
- Los puntos por partido que denotaremos por (`PTP`), es decir, los puntos por partido que un equipo logra a lo largo de una temporada.

Nótese que, para un mismo club, puede haber hasta tres observaciones distintas (en función de posibles ascensos o descensos), cada una correspondiente a cada una de las tres temporadas consideradas, siendo distintas entre ellas.

A la vista de la base de datos considerada, nuestra variable respuesta vendrá dada por los puntos por partido (`PTP`) logrados en una temporada. Las variables explicativas consideradas serán: goles a favor (`GF`), goles en contra (`GC`), tarjetas amarillas recibidas (`TA`), tarjetas rojas recibidas (`TR`), el número de porterías a cero (`PaC`), los tiros a puerta realizados (`TaPF`) y los tiros a puerta en contra (`TaPC`).

Partiremos, pues, del siguiente modelo:

$$\texttt{PTP} = \beta_0 + \beta_1\texttt{GF} + \beta_2\texttt{GC} + \beta_3\texttt{TA} + \beta_4\texttt{TR} + \beta_5\texttt{PaC} + \beta_6\texttt{TaPF} + \beta_7\texttt{TaPC} + \varepsilon. \tag{13}$$

Podemos preguntarnos por qué considerar estas variables explicativas y no otras, la razón detrás de ello es la siguiente: a la hora de sumar puntos en un partido de fútbol, lo único que importa es el marcador final, que refleja el desempeño tanto ofensivo como defensivo del equipo; pues bien, estas variables explicativas son factores clave en ambos aspectos.

Presentamos a continuación el ajuste del modelo (13):

```
> summary(modelo)
Call:
lm(formula = PTP ~ GF + GC + TA + TR + PaC + TaPF + TaPC)

Residuals:
      Min         1Q     Median         3Q        Max
-0.230206 -0.075224  0.002942  0.083173  0.263145

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.542e-01  3.579e-01   2.666   0.0102 *
GF           1.628e-02  2.315e-03   7.031 4.41e-09 ***
GC          -1.504e-02  3.202e-03  -4.697 1.97e-05 ***
TA           1.072e-03  1.345e-03   0.797   0.4289
TR          -9.888e-04  6.412e-03  -0.154   0.8780
PaC          1.460e-02  8.206e-03   1.780   0.0810 .
TaPF         3.862e-05  8.654e-04   0.045   0.9646
TaPC         5.560e-04  1.065e-03   0.522   0.6037
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1183 on 52 degrees of freedom
Multiple R-squared:  0.8511,  Adjusted R-squared:  0.831
F-statistic: 42.45 on 7 and 52 DF,  p-value: < 2.2e-16
```

A la vista de la anterior salida de R, en la columna `Estimate` podemos emcontrar las estimaciones de los parámetros β_i. Para este primer modelo, observamos ciertos aspectos a destacar acerca de los coeficientes estimados: los goles a favor (`GF`) tienen un coeficiente positivo, así como las porterías a cero (`PaC`) y los tiros a puerta a favor (`TaPF`); mientras que, los goles en contra (`GC`) y las tarjetas rojas recibidas (`TR`) tienen peso negativo en el modelo. Esto resulta lógico, pues los primeros son aspectos positivos sobre el juego de un equipo y los últimos son todos aspectos negativos que repercuten negativamente a la hora de ganar un partido.

Un coeficiente positivo significa que, al aumentar en una unidad la variable explicativa a la que está asociado, manteniendo constantes el resto de variables explicativas, la variable respuesta aumentará el valor de ese coeficiente. Esto se traduce en que, por ejemplo, cuando aumentamos la cantidad de goles a favor, los puntos por partido aumentan. En contrapartida, un coeficiente negativo implica que el aumento de la variable explicativa asociada, manteniendo el resto constante, provocará una disminución en la variable respuesta. Se puede entender como aspectos que repercuten positiva y negativamente en el desempeño de un equipo en los encuentros a lo largo de una temporada.

El valor de estos coeficientes tiene también gran relevancia: aquellos coeficientes más alejados

del cero, tienen una mayor influencia en el modelo. Una interpretación sencilla de esta salida es la siguiente: en un partido, es más importante anotar un gol más que encajar un tanto menos.

En la Figura 4.1 hemos representado las predicciones del modelo (13) frente a los valores observados para hacernos una idea sobre la bondad del modelo presentado. La línea roja representa la diagonal del primer cuadrante y en torno a la misma se deben distribuir los puntos si el modelo de regresión considerado es correcto. Por lo tanto, la Figura 4.1 nos permite intuir que el modelo (13) parece cumplir las hipótesis de linealidad y homocedasticidad de los errores, dado que los puntos se distribuyen en torno a la diagonal y la variabilidad de los mismos en torno a la recta roja parece uniforme. Hemos marcado en diferente color las tres puntuaciones que supusieron la consecución del campeonato en cada una de las tres temporadas consideradas: en dorado, el RC Deportivo de La Coruña en la temporada 1999/00; en rosa, el Real Madrid en la 2000/01; y, en azul, el Valencia CF en la 2001/02.

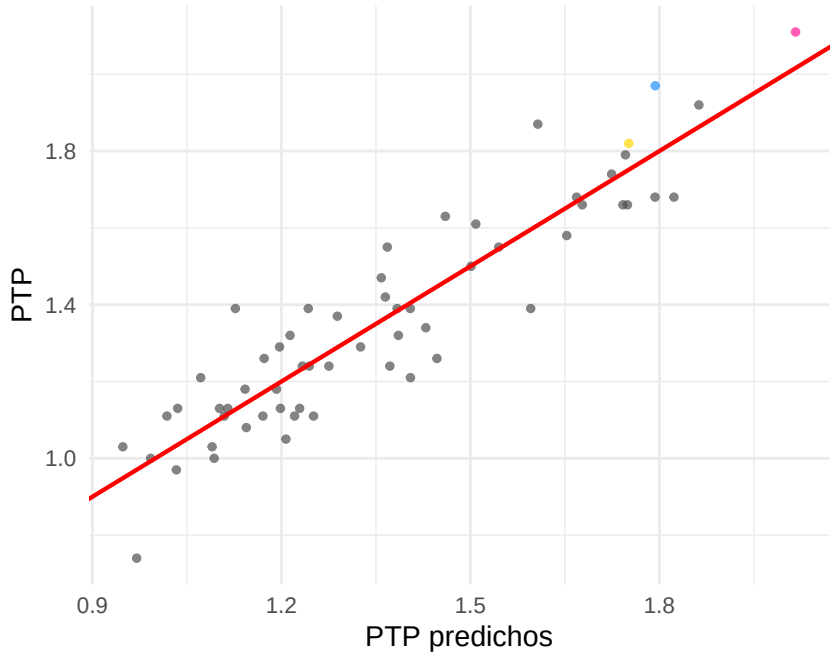


Figura 4.1: Predicciones obtenidas con el modelo (13) frente a los valores observados. Hemos marcado las tres puntuaciones que supusieron la consecución del campeonato: en dorado, el RC Deportivo de La Coruña en la temporada 1999/00; en rosa, el Real Madrid en la 2000/01; y, en azul, el Valencia CF en la 01/02.

Observamos, adicionalmente, en la salida la existencia de variables explicativas que no son significativas en nuestro modelo (13), que son aquellas con p-valor asociado al contraste de significación (columna `Pr(>|t|)`) menor que el nivel de significación fijado que en este caso es $\alpha = 0.05$. Surge, por tanto, la necesidad de aplicar un proceso de selección de variables.

Emplearemos un procedimiento de selección de variables mediante el criterio de información de Akaike[13] (AIC), la salida obtenida es la siguiente:

```
> step(modelo)
Start:  AIC=-246.68
PTP ~ GF + GC + TA + TR + PaC + TaPF + TaPC

       Df Sum of Sq     RSS     AIC
- TR    1   0.00059 0.26204 -248.56
- TaPC  1   0.00109 0.26254 -248.47
- TaPF  1   0.00155 0.26301 -248.38
- TA    1   0.00760 0.26905 -247.24
<none>              0.26145 -246.68
- GC    1   0.11080 0.37226 -231.01
- PaC   1   0.13842 0.39987 -227.43
- GF    1   0.63707 0.89852 -186.95

Step:  AIC=-248.56
PTP ~ GF + GC + TA + PaC + TaPF + TaPC

       Df Sum of Sq     RSS     AIC
- TaPC  1   0.00122 0.26327 -250.33
- TaPF  1   0.00172 0.26376 -250.24
- TA    1   0.01021 0.27226 -248.65
<none>              0.26204 -248.56
- GC    1   0.11520 0.37725 -232.34
- PaC   1   0.13798 0.40002 -229.41
- GF    1   0.63939 0.90144 -188.79

Step:  AIC=-250.33
PTP ~ GF + GC + TA + PaC + TaPF

       Df Sum of Sq     RSS     AIC
- TaPF  1   0.00219 0.26546 -251.91
- TA    1   0.00987 0.27314 -250.49
<none>              0.26327 -250.33
- GC    1   0.13273 0.39600 -231.92
- PaC   1   0.13677 0.40003 -231.41
- GF    1   0.63819 0.90146 -190.79

Step:  AIC=-251.92
PTP ~ GF + GC + TA + PaC

```

[13]El AIC es una medida global utilizada para comparar modelos de regresión. Tiene en cuenta tanto el ajuste del modelo como su complejidad, penalizando aquellos modelos con un mayor número de parámetros. Dicho criterio se define como: $\text{AIC} = 2k - 2\ln(L)$, donde k es el número de parámetros estimados y L es la verosimilitud del modelo. Consideraremos aquellos modelos que minimizan el valor del AIC.

```
       Df Sum of Sq     RSS     AIC
- TA    1   0.00959 0.27505 -252.14
<none>              0.26546 -251.91
- GC    1   0.13172 0.39718 -233.77
- PaC   1   0.13589 0.40136 -233.25
- GF    1   1.28973 1.55519 -165.52

Step:  AIC=-252.14
PTP ~ GF + GC + PaC

       Df Sum of Sq     RSS     AIC
<none>              0.27505 -252.14
- GC    1   0.13326 0.40831 -234.39
- PaC   1   0.13452 0.40957 -234.23
- GF    1   1.42965 1.70470 -162.93

Call:
lm(formula = PTP ~ GF + GC + PaC)

Coefficients:
(Intercept)           GF           GC          PaC
   0.788375     0.015533    -0.009481     0.027318
```

El modelo (13) presentó un AIC=-246.68. A través del proceso iterativo que se presenta en la salida de R, se fueron eliminando variables, lo que permitió reducir progresivamente dicho valor. Las variables eliminadas fueron `TR`, `TaPC`, `TaPF` y `TA`.

El modelo final seleccionado es el siguiente, con un total de 3 variables explicativas:

$$\texttt{PTP} = \beta_0 + \beta_1 \texttt{GF} + \beta_2 \texttt{GC} + \beta_3 \texttt{PaC} + \varepsilon. \tag{14}$$

El modelo (14) tiene asociado un AIC=-252.14, indicando una mejora con respecto al modelo (13). A continuación presentamos su ajuste:

```
> summary(modeloF)

Call:
lm(formula = PTP ~ GF + GC + PaC)

Residuals:
      Min        1Q    Median        3Q       Max
-0.142034 -0.063655  0.006142  0.063080  0.123419

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.788375   0.165679   4.758 1.97e-05 ***
```

```
GF           0.015533   0.001005  15.463  < 2e-16 ***
GC          -0.009481   0.002008  -4.721 2.23e-05 ***
PaC          0.027318   0.005759   4.743 2.07e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07733 on 46 degrees of freedom
Multiple R-squared:  0.9261,  Adjusted R-squared:  0.9212
F-statistic:   192 on 3 and 46 DF,  p-value: < 2.2e-16
```

Observamos que todas las variables explicativas consideradas son significativas, por lo que eliminar más variables explicativas del modelo carece de sentido. Tanto `GF` como `PaC` tienen asociados coeficientes positivos; resulta intuitivo, pues son aspectos positivos del juego de un equipo: cuantos más goles anoten y más porterías a cero consigan, más puntos lograrán en sus partidos. Por otro lado, `GC` tiene un coeficiente negativo, que también es lógico pues si encajas más goles, será más difícil ganar los partidos, obteniendo menores puntuaciones.

Adicionalmente, hemos obtenido un R^2 superior a 0.92, es decir, nuestro modelo es capaz de explicar más del 92 % de la variabilidad de la variable respuesta, por lo que parece que estamos ante un muy buen modelo. En este punto, resulta crucial testear la hipótesis de linealidad puesto que en caso de que nos falle dicha hipótesis, el modelo dejará de ser correcto. Vamos a utilizar el test propuesto por Stute (1997) para contrastarla.

En la Figura 4.2 hemos representado un histograma de las $B = 499$ réplicas *bootstrap* del estadístico de contraste para este caso concreto, junto con una estimación no paramétrica de la función de densidad correspondiente. La línea roja discontinua representa el valor observado del estadístico de contraste. Como bien podemos observar, el estadístico de contraste observado se posiciona bastante a la izquierda, por lo que un gran número de réplicas *bootstrap* del estadístico de contraste caen a su derecha, provocando que obtengamos un elevado p-valor.

Más concretamente, hemos obtenido un p-valor igual a 0.6192, por lo que podemos aceptar la hipótesis nula de linealidad para los niveles de significación habituales. Concluimos entonces que el modelo (14) está bien especificado. En la Figura 4.3 hemos representado las predicciones del modelo (14) frente a los valores observados.

A la vista de la Figura 4.3, observamos que concuerda con lo obtenido con el test propuesto por Stute (1997), parece cumplirse la hipótesis de linealidad, el modelo está bien especificado.

Hemos sido capaces, por tanto, de plantear un modelo de regresión lineal múltiple para explicar los puntos por partido que un equipo podía lograr a lo largo de una temporada y de testear con nuestro contraste que en efecto la relación entre la variable de interés y las variables explicativas es lineal.

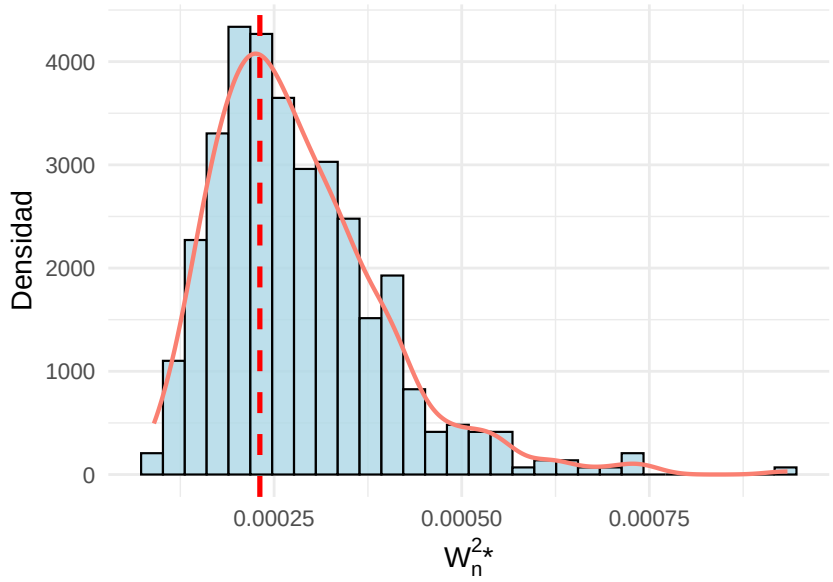


Figura 4.2: Histograma de las $B = 499$ réplicas *bootstrap* del estadístico de contraste, junto con una estimación no paramétrica de la función de densidad correspondiente. La línea roja discontinua representa el valor observado del estadístico de contraste asociado al modelo (14).

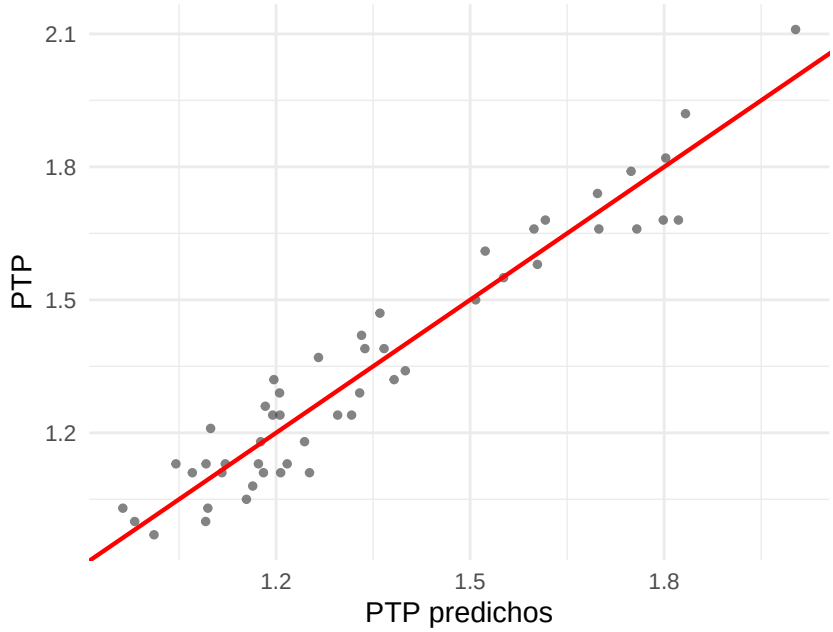


Figura 4.3: Predicciones obtenidas con el modelo (14) frente a los valores observados.

A la vista del modelo (14), podemos extraer que, de cara a ser un equipo de fútbol exitoso, hay que trabajar tanto en aspectos defensivos como ofensivos, es tan importante el anotar goles, como

no concederlos y mantener la portería imbatida; al final, los puntos se reparten en función del marcador, es decir, en función de los goles que marca cada uno de los equipos. Aquel *SuperDepor*, que logró la hazaña de quitarle una liga a los gigantes del fútbol español, combinó ambos aspectos: fue el segundo equipo más goleador de aquella campaña y el quinto que menos concedió, además de ser el segundo equipo con mayor número de porterías a cero. Se cumplen 25 años de aquella liga, pero perdurará para siempre en la memoria histórica del fútbol.

Capítulo 5

Conclusión

A lo largo de este trabajo se ha presentado la necesidad de considerar **contrastes de bondad de ajuste** en el contexto de los **modelos de regresión paramétricos**, centrándonos en el contraste propuesto por Stute (1997). El objetivo de este trabajo ha sido presentar dicho contraste para más adelante implementarlo en R y ponerlo a prueba mediante un estudio de simulación y una aplicación a datos reales. Es decir, comprobar que respeta el nivel de significación cuando consideramos modelos bajo la hipótesis nula y evaluar su potencia cuando consideramos modelos bajo la hipótesis alternativa.

Primeramente, en el Capítulo 1, se ha realizado una **introducción a los modelos de regresión lineales múltiples** ajustados por el método de mínimos cuadrados, así como a las hipótesis clásicas que se asumen sobre ellos: linealidad, homocedasticidad, normalidad e independencia de los errores. A continuación, se presentó un resultado teórico, el teorema de Fisher, que establece la distribución en el muestreo de los estimadores asociados a un modelo lineal, con el objetivo de realizar tareas de Inferencia Estadística sobre los parámetros. Se utilizó el software estadístico R para ilustrar la utilidad de dicho resultado que nos permite, por ejemplo, calcular la región de confianza para los coeficientes de un modelo de regresión lineal con dos variables explicativas. Para concluir este primer capítulo, se realizó una breve introducción a los modelos de regresión paramétricos, concepto clave a la hora de presentar el contraste planteado en Stute (1997).

En el Capítulo 2, se presentó el contraste que se desarrolla en Stute (1997), exponiendo la denominada **función de regresión integrada** que dará lugar al proceso de residuos integrados que será la base del estadístisco de contraste. Más adelante, se proporcionaron varios resultados teóricos que establecieron el comportamiento asintótico de dicho proceso. A partir de esto, se introduce un estadístico de tipo Cramér-Von Mises que será el de nuestro contraste. En la práctica, la dificultad de estimar la distribución límite del estadístico de contraste, nos conduce a introducir una **aproximación *bootstrap***, más concretamente un procedimiento *wild boots-*

trap propuesto por Stute et al. (1998), que explicamos en detalle, pues será el método que más adelante empleamos tanto en el estudio de simulación como en la aplicación a datos reales para llevar a cabo el calibrado del test. Concluimos el capítulo mostrando un esquema sobre cómo se procede a la hora de implementar el contraste en R.

Teniendo en cuenta el enfoque práctico de este trabajo, se ha desarrollado íntegramente el código de R que permite ejecutar el test para modelos de regresión lineales considerando tantas variables explicativas como sea preciso. Esta implementación se ha validado en el Capítulo 3 a través de un **estudio de simulación utilizando el método Monte Carlo**, en el que se han generado datos simulados bajo distintos escenarios:

- modelos bajo la hipótesis nula para evaluar el ajuste del nivel de significación,
- modelos bajo la hipótesis alternativa para evaluar la potencia del test,
- diferentes distribuciones del error del modelo, y
- diferentes tamaños de muestra para estudiar su efecto en el comportamiento del contraste.

En primer lugar, se consideró un modelo de regresión lineal simple, es decir, con una sola variable explicativa. Dicho modelo cumplía con las hipótesis de homocedasticidad, normalidad e independencia de los errores y contaba con un posible efecto cuadrático de la variable explicativa escalado por cierto parámetro a (constante que controla si el modelo está bajo la hipótesis nula o alternativa). Se consideraron también modificaciones en la variabilidad del error así como en el tamaño de muestra. La conclusión de dicho escenario fue clara, el contraste propuesto por Stute (1997) respeta los niveles de significación establecidos bajo la hipótesis nula y es capaz de rechazar con alta probabilidad cuando nos desviamos hacia la hipótesis alternativa (aumentando el parámetro a). Asimismo, el aumento de la variabilidad del error, como era de esperar, supuso cierta pérdida de potencia.

En el segundo escenario se introdujo un pequeño cambio, esta vez se consideraron modelos con errores heterocedásticos con el objetivo de comprobar que el *wild bootstrap* seguía siendo una aproximación válida pese a la pérdida de la hipótesis clásica de homocedasticidad de los errores. Tras analizar este escenario de simulación se concluyó que, en efecto, la presencia de errores heterocedásticos no supuso pérdida de potencia, por lo que dicha aproximación *bootstrap* seguía siendo perfectamente válida. De dicho escenario concluimos que el procedimiento *wild bootstrap* se podría utilizar en contextos más generales en los cuales no se considere la hipótesis de homocedasticidad de los errores.

Finalizando con el Capítulo 3, se presentó un tercer y último escenario, un modelo de regresión lineal con dos variables explicativas y con cierto parámetro a que controlaba la interacción entre

ellas. Este escenario se propuso con el objetivo de implementar un código más general que no dependiese del número de variables explicativas consideradas en el modelo, que resultaría necesario más adelante en la aplicación a datos reales. De nuevo, mediante el estudio de simulación realizado, se comprobó el buen funcionamiento del contraste incluso cuando aumenta la dimensión de las variables explicativas aunque esto tenga un claro efecto en el tiempo de computación del contraste.

Para concluir con este trabajo, en el Capítulo 4 hemos empleado el contraste en un **caso real relacionado con el fútbol de élite** para ilustrar la utilidad del contraste propuesto por Stute (1997) en la práctica. Hemos tomado 3 de las temporadas más exitosas del Real Club Deportivo de La Coruña, con el objetivo de plantear un modelo de regresión lineal múltiple que explique los puntos por partido que un equipo logra a lo largo de una campaña. Dicho modelo fue sometido a un procedimiento de selección automática de variables, utilizando un procedimiento backward con criterio global, el criterio de información de Akaike (conocido habitualmente como AIC). Como consecuencia de dicho procedimiento, nos quedaremos con sólo las variables explicativas que más ayudan a explicar el comportamiento de la variable respuesta y a llevar a cabo predicciones.

El modelo final contaba con tan sólo 3 variables explicativas: los goles a favor, los goles en contra y las porterías a cero. Dicho modelo presentó un coeficiente de determinación superior a 0.92, por lo que obtuvimos un muy buen ajuste. Por último, validamos el modelo utilizando nuestro contraste implementado en  , obteniendo un p-valor $= 0.6192$, por lo que aceptamos la hipótesis nula de linealidad para los niveles de significación habituales. Por lo tanto, hemos sido capaces de **aplicar el contraste a una situación real**, además de validar el modelo propuesto.

En resumen, podemos afirmar que el contraste propuesto por Stute (1997), complementado con el método de aproximación *wild bootstrap* propuesto por Stute et al. (1998), constituye una eficaz y robusta herramienta a la hora de validar la forma paramétrica de un modelo de regresión en media ajustado por mínimos cuadrados. Su implementación en la práctica, exigente a nivel técnico, es factible y proporciona resultados fácilmente interpretables tanto para situaciones con datos simulados como con datos reales. En contra posición a métodos más clásicos, como viene siendo el análisis gráfico de los residuos del modelo o los contrastes de especificación como puede ser el de Ramsey (1969), el contraste que hemos estudiado está fundamentado en una base teórica sólida y no precisa de especificaciones de un modelo concreto bajo la hipótesis alternativa, convirtiéndolo en una opción muy atractiva en diversas situaciones.

No obstante, conviene señalar algunas de las limitaciones del test propuesto por Stute (1997). En primer lugar, la aproximación *bootstrap* empleada, aunque eficaz, provoca que nuestro test sea más lento y además, a medida que aumenta el número de variables explicativas consideradas, el test pierde potencia (debido a la consideración de la función indicadora en la definición del estadístico de contraste), es lo que se conoce como el **desastre de la dimensión**. El hecho de

aumentar la dimensión del modelo se hacía notar en el estudio de simulación, que con el aumento del número de variables explicativas consideradas, provocaba un importante aumento en el tiempo de computación del contraste. Asimismo, aunque el contraste se comporta bien bajo condiciones ideales, acusa su sensibilidad ante modelos que presentan errores con alta variabilidad, provocando pérdida de potencia cuando consideramos modelos bajo la hipótesis alternativa.

A modo resumen, este trabajo ha contribuido no sólo a la comprensión teórica del contraste estudiado por Stute (1997), sino también a su puesta en práctica, mostrando su utilidad como herramienta de validación para modelos de regresión paramétricos. Hemos probado su buen comportamiento ante modelos bajo la hipótesis de linealidad, siendo el test capaz de respetar el nivel de significación, así como su potencia a la hora de rechazar la hipótesis nula cuando se consideraron modelos bajo la hipótesis alternativa. Asimismo, hemos sido capaces de aplicarlo a un caso real, exhibiendo su utilidad más allá de los marcos teóricos.

Anexo I

Código de 

En este anexo presentamos el código de  que ha sido desarrollado para la obtención de todas las gráficas que se presentan a lo largo de este trabajo, así como el código desarrollado para la obtención de los resultados del estudio de simulación y de la aplicación a datos reales.

I.1. Gráficas del Capítulo 1

```r
#Linealidad vs no linealidad
set.seed(123)
# Generar 50 datos con relacion lineal creciente
n<-50
x <- runif(n, min = 0, max = 5)
y <- 2 + 2* x + rnorm(n, mean = 0, sd = 1)
df_lineal <- data.frame(x = x, y = y)
ggplot(df_lineal, aes(x = x, y = y)) +
  geom_point(color = "grey30", alpha = 0.6) +
  geom_abline(intercept = 2, slope = 2, color = "#E41A1C", size = 1) +
  labs(
    x = "x",
    y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

# Generar 50 datos sin relacion lineal (senoidal)
x <- runif(n, min = 0, max = 5)
y <- sin(x) + rnorm(n, mean = 0, sd = 0.1)
df_seno <- data.frame(x = x, y = y, y_true = sin(x))
ggplot(df_seno, aes(x = x, y = y)) +
  geom_point(color = "grey30", alpha = 0.6) +
```

```r
  geom_line(aes(y = y_true), color = "#377EB8", size = 1) +
  labs(
    x = "x",
    y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

# Homocedasticidad vs heterocedasticidad
set.seed(123)
n <- 100
x <- runif(n, 0, 10)
beta_0 <- 2
beta_1 <- 2
# -------- Modelo HOMOCEDASTICO --------
epsilon_homo <- rnorm(n, mean = 0, sd = 1)
y_homo <- beta_0 + beta_1 * x + epsilon_homo
df_homo <- data.frame(x = x, y = y_homo)

ggplot(df_homo, aes(x = x, y = y)) +
  geom_point(color = "grey30", alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
# -------- Modelo HETEROCEDASTICO --------
sd_hetero <- 1
epsilon_hetero <- rnorm(n, mean = 0, sd = sd_hetero)
y_hetero <- beta_0 + beta_1 * x + (x+0.5)*epsilon_hetero
df_hetero <- data.frame(x = x, y = y_hetero)

ggplot(df_hetero, aes(x = x, y = y)) +
  geom_point(color = "grey30", alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))


# Normalidad vs no normalidad
install.packages("car")
library(car)
set.seed(123)
```

```r
n <- 1000
# -------- Errores NORMALES --------
eps_normal <- rnorm(n)
df_normal <- data.frame(error = eps_normal, tipo= "Normal")
# -------- Errores NO NORMALES (asimetricos) --------
eps_nonnormal <- rchisq(n, df = 3) - 3  # centrado para media 0
df_nonnormal <- data.frame(error = eps_nonnormal, tipo= "No normal")
# -------- Graficas separadas con facet_wrap --------
ggplot(df_normal, aes(x = error, y = ..density..)) +
  geom_histogram(bins = 30, fill = "light blue", color = "black", alpha = 0.8) +
  facet_wrap(~ tipo, scales = "free") +
  labs(
       x = "Error", y = "Densidad") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))+
  theme(strip.text = element_blank())

ggplot(df_nonnormal, aes(x = error, y = ..density..)) +
  geom_histogram(bins = 30, fill = "salmon", color = "black", alpha = 0.8) +
  facet_wrap(~ tipo, scales = "free") +
  labs(
    x = "Error", y = "Densidad") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5)) +
  theme(strip.text = element_blank())
qqPlot(eps_normal,
       xlab = "Cuantiles teoricos",
       ylab = "Cuantiles de la muestra",
       col = "grey30", pch = 19) # qqplot
qqPlot(eps_nonnormal,
       xlab = "Cuantiles teoricos",
       ylab = "Cuantiles de la muestra",
       col = "grey30", pch = 19) # qqplot

#### GRAFICA REGION DE CONFIANZA PARA DOS COEFICIENTES DE UN MODELO DE REGRESION
    LINEAL MULTIPLE
set.seed(42)
n <- 100
# Definir media y matriz de covarianza para X1 y X2
mu <- c(0, 0)
Sigma_x <- matrix(c(1, 0.7, 0.7, 1), nrow = 2)  # correlacion 0.7 entre X1 y X2
# Simular variables explicativas correlacionadas
X <- mvrnorm(n, mu, Sigma_x)
x1 <- X[,1]
x2 <- X[,2]
# Generar variable respuesta y
y <- 2 + 3 * x1 - 2 * x2 + rnorm(n)
```

```r
# Ajustar modelo
modelo <- lm(y ~ x1 + x2)
beta_hat <- coef(modelo)[2:3]
Sigma_hat <- vcov(modelo)[2:3, 2:3]
nivel <- 0.99
# Crear elipse de confianza
elipse <- as.data.frame(ellipse(Sigma_hat, centre = beta_hat, level = nivel))
colnames(elipse) <- c("beta1", "beta2")

 ggplot(elipse, aes(x = beta1, y = beta2)) +
  geom_polygon(fill = "#A6CEE3", color = "#1F78B4", alpha = 0.6) +
  geom_point(aes(x = beta_hat[1], y = beta_hat[2]), color = "grey30", size = 3)
    +
  labs(
    x = expression(beta[1]),
    y = expression(beta[2])
  ) +
  coord_fixed() +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold")
  )

# Modelo parametrico no lineal
set.seed(123)
n <- 100
x <- runif(n, 0, 10)
y <- 2 + 1.5 * x - 0.2 * x^2 + rnorm(n, 0, 1)

# Crear data frame
datos <- data.frame(x = x, y = y)

# Ajustar dos modelos
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
modelo_B <- lm(y ~ x + I(x^2), data = datos)     # Modelo cuadratico

# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
df_pred$cuadratico <- predict(modelo_B, newdata = df_pred)

# Graficar
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
```

```
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  geom_line(data = df_pred, aes(y = cuadratico), color = "#377EB8", size = 1.2)
    +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
```

I.2. Estudio de simulación

Presentamos en esta sección el código de  desarrollado a lo largo del Capítulo 3.

I.2.1. Gráficas de los modelos

En primer lugar, se introduce el código asociado a las diferentes representaciones gráficas presentadas.

```
##### MODELO 1 CON a=0
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 0 * x^2 + rnorm(n, 0, 1)
datos <- data.frame(x = x, y = y)
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

##### MODELO 1 CON a=20
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
```

```r
y <- 1 + 2 * x + 20 * x^2 + rnorm(n, 0, 1)
datos <- data.frame(x = x, y = y)
modelo_B <- lm(y ~ x + I(x^2), data = datos)     # Modelo cuadratico
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$cuadratico <- predict(modelo_B, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = cuadratico), color = "#377EB8", size = 1.2)
    +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

##### MODELO 1 CON SIGMA=1/2
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 0 * x^2 + rnorm(n, 0, 1/2)
datos <- data.frame(x = x, y = y)
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

###### MODELO 1 SIGMA=5
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 0 * x^2 + rnorm(n, 0, 5)
datos <- data.frame(x = x, y = y)
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
```

```r
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

# MODELO 2 CON a=0
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 0 * x^2 + (x+0.5)*rnorm(n, 0, 1)
datos <- data.frame(x = x, y = y)
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

###### MODELO 2 CON a=20
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 20 * x^2 + (x+0.5)*rnorm(n, 0, 1)
datos <- data.frame(x = x, y = y)
modelo_B <- lm(y ~ x + I(x^2), data = datos)     # Modelo cuadratico
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$cuadratico <- predict(modelo_B, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
```

```
  geom_line(data = df_pred, aes(y = cuadratico), color = "#377EB8", size = 1.2) +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

#### MODELO 2 CON SIGMA=1
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 0 * x^2 + (x+0.5)*rnorm(n, 0, 1)
datos <- data.frame(x = x, y = y)
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

###### MODELO 2 CON SIGMA=5
set.seed(123456)
n <- 100
x <- runif(n, 0, 1)
y <- 1 + 2 * x + 0 * x^2 + (x+0.5)*rnorm(n, 0, 5)
datos <- data.frame(x = x, y = y)
modelo_A <- lm(y ~ x, data = datos)              # Modelo lineal
# Generar predicciones para graficar suavemente
x_seq <- seq(min(x), max(x), length.out = 200)
df_pred <- data.frame(x = x_seq)
df_pred$lineal <- predict(modelo_A, newdata = df_pred)
ggplot(datos, aes(x, y)) +
  geom_point(alpha = 0.6, size = 2, color = "grey30") +
  geom_line(data = df_pred, aes(y = lineal), color = "#E41A1C", size = 1.2,
    linetype = "dashed") +
  labs(
    x = "x", y = "y"
  ) +
```

```
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

####### GRAFICA DE LA FUNCION DE POTENCIA DEL MODELO 2
set.seed(123456789)
n=100
sigma=1
a=0
M=100
B=499
alpha=c(0.1)
kind=1
a_vals <- seq(-5, 5, by = 0.2) # valores de a
power <- numeric(length(a_vals))
pb <- txtProgressBar(min = 0, max = length(a_vals), style = 3)
for (h in seq_along(a_vals)){
  setTxtProgressBar(pb, h)
  a <- a_vals[h]
  rechazar=matrix(NA,ncol=length(alpha),nrow=M)
for(i in 1:M){
  erro=rnorm(n,sd=sigma) # error de los datos
  x=runif(n,min=0,max=1) # muestra x
  y=1+2*x+a*x^2+(x+0.5)*erro # muestra y

  #Estimar el modelo bajo H0:
  modelo <- lm(y ~ x) # estima el modelo
  y.hat <- fitted(modelo) # valores predecidos
  residuos <- resid(modelo) # residuos del modelo

  # Calcular el estadistico de contraste:
  orden <- order(x)
  x.ord <- x[orden]
  res.ord <- residuos[orden]
  Rn1 <- cumsum(res.ord) / sqrt(n)
  Wn2 <- sum(Rn1^2) / n

  Wn2.bootstrap <- numeric(B)
  for(j in 1:B){
    # Muestra bootstrap
    P=c((sqrt(5)+1)/(2*sqrt(5)),(sqrt(5)-1)/(2*sqrt(5)))
    z<-c(-(sqrt(5)-1)/(2),(sqrt(5)+1)/(2))
    V<-runif(n,min=0,max=1)
    for (k in 1:n) {
      if (V[k]<P[1]) {
        V[k]=z[1]
      }
      else {
```

```
        V[k]=z[2]
      }
    } # genera la muestra i.i.d. con media 0, varianza 1 y finita.
    y.boot <- y.hat + residuos * V
    modelo.boot <- lm(y.boot ~ x)
    residuos.boot <- resid(modelo.boot)
    res.ord.boot <- residuos.boot[order(x)]
    Rn1.boot <- cumsum(res.ord.boot) / sqrt(n)
    Wn2.bootstrap[j] <- sum(Rn1.boot^2) / n
  }
  p.valor <- mean(Wn2.bootstrap > Wn2)
  rechazar[i,] <- p.valor < alpha
}
  power[h]=apply(rechazar, 2, mean)
  kind=kind+1 # barra de progreso
}
plot(a_vals, power, type = "l", lwd = 2, col = "light blue",
     xlab = expression(a),
     ylab = "Probabilidad de rechazo")
abline(h = alpha, col = "salmon", lty = 2)
```

I.2.2. Código del escenario 1

A continuación, presentamos a modo de ejemplo, uno de los modelos de regresión considerados en el contexto del Escenario 1 del Capítulo 3.

```
set.seed(1234) # semilla para reproducibilidad
n=50 # elementos en cada muestra
sigma=1 # desviacion tipica
a=0 # parametro de desviacion de H0
M=1000 # muestras aleatorias generadas
B=499 #muestras bootstrap
alpha=c(0.1,0.05,0.01) # nivel de significacion
rechazar=matrix(NA,ncol=length(alpha),nrow=M) # matriz de rechazo
pb=txtProgressBar(style=3) # barra de progreso
kind=1
tempo0=proc.time() # contador de tiempo
# pesos de Mammen
P=c((sqrt(5)+1)/(2*sqrt(5)),(sqrt(5)-1)/(2*sqrt(5)))
z<-c(-(sqrt(5)-1)/(2),(sqrt(5)+1)/(2))

# bucle del proceso entero:
for(i in 1:M){
  setTxtProgressBar(pb,kind/M)
  erro=rnorm(n,sd=sigma) # error de los datos
  x=runif(n,min=0,max=1) # muestra x
```

```
y=1+2*x+a*x^2+erro # muestra y

#Estimar el modelo bajo H0:
modelo <- lm(y ~ x) # estima el modelo
y.hat <- fitted(modelo) # valores predecidos
residuos <- resid(modelo) # residuos del modelo

# Calcular el estatistico de contraste:
orden <- order(x) # calcula el indice que ordena los valores de manera creciente
x.ord <- x[orden] # ordena el vector x en orden creciente usando los indices calculados
res.ord <- residuos[orden] # reordena los residuos
Rn1 <- cumsum(res.ord) / sqrt(n) # calculamos Rn^1
Wn2 <- sum(Rn1^2) / n # calculo del estadistico Wn^2 (aproximamos la integral mediante la suma de cuadrados entre n)

# Procedimiento bootstrap:
Wn2.bootstrap <- numeric(B)
for(j in 1:B){
  # Muestra bootstrap
  V<-runif(n,min=0,max=1)
  for (k in 1:n) {
    if (V[k]<P[1]) {
      V[k]=z[1]
    }
    else {
      V[k]=z[2]
    }
  } # genera la muestra Vi*
  y.boot <- y.hat + residuos * V

  # Estimar modelo bootstrap
  modelo.boot <- lm(y.boot ~ x)
  residuos.boot <- resid(modelo.boot)

  # Calcular el estadistico boostrap
  res.ord.boot <- residuos.boot[order(x)] #ordenamos residuos de manera creciente
  Rn1.boot <- cumsum(res.ord.boot) / sqrt(n)
  Wn2.bootstrap[j] <- sum(Rn1.boot^2) / n
}
# Calibrar el test: comparar estadistico observado con bootstrap
p.valor <- mean(Wn2.bootstrap > Wn2) # crea un vector logico que contiene 1 y 0 segun se cumpla o no
rechazar[i,] <- p.valor < alpha # para cada muestra guarda en el vector si rechaza (1) o si acepta (0) H0.
```

```
  kind=kind+1 # barra de progreso
}
apply(rechazar, 2, mean) # vector probabilidad de rechazo de cada nivel de
    significacion
proc.time()-tempo0
```

I.2.3. Código del escenario 2

Seguidamente, presentamos a modo de ejemplo, uno de los modelos de regresión considerados en el contexto del Escenario 2 del Capítulo 3.

```
set.seed(123456789)
n=50 # elementos en cada muestra
sigma=1 # desviacion tipica
a=0 # parametro de desviacion de H0
M=1000 # muestras aleatorias generadas
B=499 #muestras bootstrap
alpha=c(0.1,0.05,0.01) # nivel de significacion
rechazar=matrix(NA,ncol=length(alpha),nrow=M) # matriz de rechazo
pb=txtProgressBar(style=3) # barra de progreso
kind=1
tempo0=proc.time() # contador de tiempo
# pesos de Mammen
P=c((sqrt(5)+1)/(2*sqrt(5)),(sqrt(5)-1)/(2*sqrt(5)))
z<-c(-(sqrt(5)-1)/(2),(sqrt(5)+1)/(2))

# bucle del proceso entero:
for(i in 1:M){
  setTxtProgressBar(pb,kind/M)
  erro=rnorm(n,sd=sigma) # error de los datos
  x=runif(n,min=0,max=1) # muestra x
  y=1+2*x+a*x^2+(x+0.5)*erro # muestra y

  #Estimar el modelo bajo H0:
  modelo <- lm(y ~ x) # estima el modelo
  y.hat <- fitted(modelo) # valores predecidos
  residuos <- resid(modelo) # residuos del modelo
  # Calcular el estatistico de contraste:
  orden <- order(x) # calcula el indice que ordena los valores de manera
    creciente
  x.ord <- x[orden] # ordena el vector x en orden creciente usando los indices
    calculados
  res.ord <- residuos[orden] # reordena los residuos
  Rn1 <- cumsum(res.ord) / sqrt(n) # calculamos Rn^1
  Wn2 <- sum(Rn1^2) / n
  # Procedimiento bootstrap:
```

```
  Wn2.bootstrap <- numeric(B)
  for(j in 1:B){
    # Muestra bootstrap
    V<-runif(n,min=0,max=1)
    for (k in 1:n) {
      if (V[k]<P[1]) {
        V[k]=z[1]
      }
      else {
        V[k]=z[2]
      }
    } # genera la muestra Vi*
    y.boot <- y.hat + residuos * V

    # Estimar modelo bootstrap
    modelo.boot <- lm(y.boot ~ x)
    residuos.boot <- resid(modelo.boot)
    # Calcular el estadistico boostrap
    res.ord.boot <- residuos.boot[order(x)] #ordenamos residuos de manera
    creciente
    Rn1.boot <- cumsum(res.ord.boot) / sqrt(n)
    Wn2.bootstrap[j] <- sum(Rn1.boot^2) / n
  }
  p.valor <- mean(Wn2.bootstrap > Wn2)
  rechazar[i,] <- p.valor < alpha
  kind=kind+1 # barra de progreso
}
apply(rechazar, 2, mean) # vector probabilidad de rechazo de cada nivel de
    significacion
proc.time()-tempo0
```

I.2.4. Código del escenario 3

Presentamos ahora, a modo de ejemplo, uno de los modelos de regresión considerados en el contexto del Escenario 3 del Capítulo 3.

```
set.seed(1234)
p=2 # numero de variables explicativas
n=50
sigma=1
a=0
M=1000
B=499
alpha=c(0.1,0.05,0.01)
rechazar=matrix(NA,ncol=length(alpha),nrow=M)
pb=txtProgressBar(style=3)
```

```
kind=1
tempo0=proc.time()
# pesos de Mammen
P=c((sqrt(5)+1)/(2*sqrt(5)),(sqrt(5)-1)/(2*sqrt(5)))
z<-c(-(sqrt(5)-1)/(2),(sqrt(5)+1)/(2))

# bucle del proceso entero:
for(i in 1:M){
  setTxtProgressBar(pb,kind/M)
  erro=rnorm(n,sd=sigma)
  x1=runif(n,min=0,max=1) # muestra x1
  x2=runif(n,min=0,max=1) # muestra x2
  y=2+5*x1-x2+a*x1*x2+erro # muestra y
  X=cbind(x1,x2) # matriz con las variables explicativas

  #Estimar el modelo bajo H0:
  modelo <- lm(y ~ x1+x2) # estimar el modelo
  y.hat <- fitted(modelo)
  residuos <- resid(modelo)

  # Calcular el estadistico de contraste (para cualquier dimension de las
    covariables):
  Rn1 <- numeric(n)
  for (w in 1:n) {
    indicador=numeric(n)
    for(h in 1:p){
      indicador=indicador+as.numeric(X[,h]<=X[w,h])
    }
    Rn1[w] <- sum( (indicador==p) * residuos) / sqrt(n)
  }
  Wn2 <- mean(Rn1^2)

  # Procedimiento bootstrap:
  Wn2.bootstrap <- numeric(B)
  for(j in 1:B){
    # Muestra bootstrap
    V<-runif(n,min=0,max=1)
    for (k in 1:n) {
      if (V[k]<P[1]) {
        V[k]=z[1]
      }
      else {
        V[k]=z[2]
      }
    }
    y.boot <- y.hat + residuos * V

```

```
    # Estimar modelo bootstrap
    modelo.boot <- lm(y.boot ~ x1+x2)
    residuos.boot <- resid(modelo.boot)

    # Calcular el estadistico boostrap
    Rn1.boot <- numeric(n)
    for (d in 1:n) {
      indicadora=numeric(n)
      for(l in 1:p){
        indicadora=indicadora+as.numeric(X[,l]<=X[d,l])
      }
      Rn1.boot[d] <- sum( (indicadora==p) * residuos.boot) / sqrt(n)
    }
    Wn2.bootstrap[j] <- mean(Rn1.boot^2)
  }
  p.valor <- mean(Wn2.bootstrap > Wn2)
  rechazar[i,] <- p.valor < alpha
  kind=kind+1 # barra de progreso
}
apply(rechazar, 2, mean) # vector probabilidad de rechazo de cada nivel de
    significacion
proc.time()-tempo0
```

I.3. Aplicación a datos reales

A continuación se presenta el código de  que ha sido desarrollado para llevar a cabo la aplicación del test propuesto por Stute (1997) a un conjunto de datos reales.

```
datos=read.table("base de datos TFG.txt",header=TRUE)
View(datos)
n=nrow(datos) #numero de equipos analizados
attach(datos)
GF<-datos$GF  #goles a favor
GC<-datos$GC  #goles en contra
TA<-datos$TA  #tarjetas amarillas
TR<-datos$TR  #tarjetas rojas
PaC<-datos$PaC  #porterias a cero
TaPF<-datos$TaPF  #tiros a puerta a favor
TaPC<-datos$TaPC  #tiros a puerta en contra
PTP<-datos$PTP  #puntos por partido
summary(PTP)
p=8 #numero de variables
#estimacion del modelo
modelo=lm(PTP~GF+GC+TA+TR+PaC+TaPF+TaPC)
```

```r
summary(modelo) #Con este primer modelo ya observamos ciertos aspectos con sentido:
# los GF suman positivamente, asi como las porterias a cero y los tiros a puerta a favor; mientras que los GC y las tarjetas rojas recibidas tienen peso negativo en el modelo.
PTPhat<-modelo$fitted.values #predicciones
modelo$residuals #residuos
deviance(modelo) #estimacion de la varianza residual
#diagrama de dispersion
df <- data.frame(
  ajustado = modelo$fitted.values,
  observado = PTP
)
df$color <- ifelse(1:nrow(df) == 8, "gold",
                   ifelse(1:nrow(df) == 35, "deeppink",
                          ifelse(1:nrow(df) == 57, "dodgerblue", "grey30")))
ggplot(df, aes(x = ajustado, y = observado)) +
  geom_point(aes(color = color), alpha=0.7) +  # puntos (pch=19 en base R)
  scale_color_identity() +
  geom_abline(slope = 1, intercept = 0, color = "red", size = 1) +
  labs(x = "PTP predichos", y = "PTP") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

summary(modelo)
step(modelo)
# seleccion de variables mediante criterio de informacion Akaike (AIC), el
# modelo final es aquel que minimiza AIC, que en este caso contiene 3 variables explicativas,
# los goles a favor, los goles en contra y las porterias a cero.
modeloF=lm(formula = PTP ~ GF + GC + PaC)
summary(modeloF)

####### CONTRASTAR LINEALIDAD
X=cbind(GF,GC,PaC)
p=3
set.seed(1234) # semilla para reproducibilidad
B=499 #muestras bootstrap
# bucle del proceso entero:
#Estimar el modelo bajo H0:
y.hat <- fitted(modeloF) # valores predecidos
residuos <- resid(modeloF) # residuos del modelo

# Calcular el estadistico de contraste:
Rn1 <- numeric(n)
for (w in 1:n) {
  indicador=numeric(n)
```

```
  for(h in 1:p){
    indicador=indicador+as.numeric(X[,h]<=X[w,h])
  }
  Rn1[w] <- sum( (indicador==p) * residuos) / sqrt(n)
}
Wn2 <- mean(Rn1^2)

# Procedimiento bootstrap:
Wn2.bootstrap <- numeric(B)
for(j in 1:B){
  # Muestra bootstrap
  P=c((sqrt(5)+1)/(2*sqrt(5)),(sqrt(5)-1)/(2*sqrt(5)))
  z<-c(-(sqrt(5)-1)/(2),(sqrt(5)+1)/(2))
  V<-runif(n,min=0,max=1)
  for (k in 1:n) {
    if (V[k]<P[1]) {
      V[k]=z[1]
    }
    else {
      V[k]=z[2]
    }
  } # genera la muestra i.i.d. con media 0, varianza 1 y finita.
  y.boot <- y.hat + residuos * V

  # Estimar modelo bootstrap
  modelo.boot <- lm(y.boot ~ GF+GC+PaC)
  residuos.boot <- resid(modelo.boot)

  # Calcular el estadistico bootstrap
  Rn1.boot <- numeric(n)
  for (d in 1:n) {
    indicadora=numeric(n)
    for(l in 1:p){
      indicadora=indicadora+as.numeric(X[,l]<=X[d,l])
    }
    Rn1.boot[d] <- sum( (indicadora==p) * residuos.boot) / sqrt(n)
  }
  Wn2.bootstrap[j] <- mean(Rn1.boot^2)
}

# Calibrar el test: comparar estadistico observado con bootstrap
p.valor <- mean(Wn2.bootstrap > Wn2)
p.valor

# histograma bootstrap
df_bootstrap <- data.frame(
  valor = Wn2.bootstrap,
```

```
  tipo = "Distribucion Bootstrap"
)
ggplot(df_bootstrap, aes(x = valor, y = ..density..)) +
  geom_histogram(bins = 30, fill = "light blue", color = "black", alpha = 0.8) +
  geom_vline(xintercept = Wn2, color = "red", linetype = "dashed", linewidth =
    1.2) +
  geom_density(color = "salmon", linewidth = 1) +
  facet_wrap(~ tipo, scales = "free") +
  labs(
    x = expression(W[n]^2*"*"),
    y = "Densidad"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    strip.text = element_blank()
  )
```

Anexo II

Base de datos reales

A lo largo de este anexo, se presenta la base de datos completa que ha sido utilizada en el Capítulo 4 para ilustrar el comportamiento del contraste propuesto por Stute (1997) en la práctica.

EQUIPO	GF	GC	TA	TR	PaC	TaPF	TaPC	PTP
Alaves99	40	37	108	7	17	196	171	1.61
Athletic99	47	57	125	3	8	187	174	1.32
Atletico99	48	64	121	6	6	215	181	1.00
Barça99	70	46	85	5	14	246	164	1.68
Betis99	33	56	100	9	11	178	186	1.11
Celta99	44	43	123	11	10	221	152	1.39
Espanyol99	51	48	103	4	7	184	169	1.24
Depor99	64	44	94	9	16	198	151	1.82
Malaga99	54	50	98	8	11	203	180	1.26
Mallorca99	51	45	85	6	9	172	177	1.34
Numancia99	46	59	108	6	11	170	179	1.18
Oviedo99	44	60	104	2	10	169	202	1.18
Santander99	52	50	98	7	10	195	188	1.21
Rayo99	49	53	100	5	8	230	191	1.37
Madrid99	56	48	95	7	7	251	197	1.63
RealSociedad99	42	49	105	11	8	185	192	1.24
Sevilla99	42	65	111	6	6	183	184	0.74
Valencia99	56	39	103	11	12	187	193	1.68
Valladolid99	35	44	105	4	11	175	188	1.39
Zaragoza99	59	40	116	4	10	193	163	1.66

Alaves00	55	59	113	5	9	181	196	1.29
Athletic00	44	60	111	11	8	154	185	1.13
Barça00	79	56	91	9	10	259	192	1.66
Celta00	49	49	125	11	9	182	165	1.55
Espanyol00	44	44	89	5	12	190	187	1.32
Depor00	73	44	103	7	12	226	170	1.92
LasPalmas00	42	62	117	9	9	148	203	1.21
Malaga00	60	61	95	6	8	176	226	1.47
Mallorca00	59	43	93	5	10	200	163	1.87
Numancia00	38	64	103	5	7	182	222	1.03
Osasuna00	40	54	91	3	8	149	178	1.11
Oviedo00	51	67	109	6	8	196	238	1.08
Santander00	47	62	98	6	6	168	198	1.03
Rayo00	54	68	109	9	4	165	237	1.13
Madrid00	78	40	78	2	14	258	176	2.11
RealSociedad00	51	68	95	2	4	169	196	1.13
Valencia00	53	34	117	7	15	202	164	1.66
Valladolid00	41	50	109	6	11	181	185	1.11
Villarreal00	57	52	116	0	12	211	168	1.50
Zaragoza00	53	57	86	6	4	212	195	1.11
Alaves01	39	44	135	8	13	169	188	1.42
Athletic01	54	66	108	6	3	198	225	1.39
Barça01	63	37	88	3	14	260	169	1.68
Betis01	40	34	110	7	17	201	151	1.55
Celta01	63	46	123	10	10	225	159	1.58
Espanyol01	47	56	122	8	8	191	216	1.24
Depor01	64	41	80	4	13	217	156	1.79
LasPalmas01	40	50	118	7	8	149	201	1.05
Malaga01	43	44	129	8	12	182	179	1.39
Mallorca01	40	52	95	5	11	154	201	1.13
Osasuna01	34	49	98	4	13	145	186	1.11
Rayo01	41	52	103	10	10	168	188	1.29
Madrid01	66	44	86	2	11	239	175	1.74
RealSociedad01	48	54	89	4	10	189	193	1.24
Sevilla01	51	40	126	8	12	185	186	1.39
Tenerife01	32	58	126	8	9	142	226	1.00
Valencia01	50	27	112	5	16	210	134	1.97
Valladolid01	45	58	97	3	9	178	214	1.26

Villarreal01	46	55	118	4	7	187	218	1.13
Zaragoza01	35	54	98	6	7	147	206	0.97

Bibliografía

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Efron, B. (1979). *Bootstrap methods: Another look at the jackknife*. The Annals of Statistics, **7**(1), 1–26.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer.

Liu, J. y Singh, K. (1997). *A Circular Bootstrap Method for Dependent Data*. Journal of the American Statistical Association, **92**(440), 1354-1367.

Montgomery, D. C., Peck, E. A. y Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. Wiley.

Politis, D. N. y Romano, J. P. (1994). *The Stationary Bootstrap*. Journal of the American Statistical Association, **89**(428), 1303-1313.

Ramsey, J. B. (1969). *Tests for Specification Errors in Classical Linear Least Squares Regression Analysis*. Journal of the Royal Statistical Society: Series B (Methodological), **31**(2), 350–371.

Ross, S. (2014). *A First Course in Probability*. Pearson Education.

Silverman, B. W. y Young, G. A. (1987). *The bootstrap: To smooth or not to smooth?* Biometrika, **74**(3), 469–479.

Stute, W. (1997). *Nonparametric model checks for regression*. The Annals of Statistics, **25**(2), 613–641.

Stute, W., González-Manteiga, W. y Presedo Quindimil, M. (1998). *Bootstrap approximations in model checks for regression*. Journal of the American Statistical Association, **93**(441), 141–149.