USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Selección de variables en modelos de regresión

Jacobo Casares Lorenzo

2023-2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Traballo Fin de Grado

Selección de variables en modelos de regresión

Jacobo Casares Lorenzo

Junio, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estatística e Investigación de Operacións
Título: Selección de variables en modelos de regresión
Breve descripción del contenido:
A selección de variables é un tema de moita actualidade especialmente nun contexto de abundancia de covariables posibles para seren integradas nun modelo. Os métodos para esta selección divídense en aqueles que supoñen un modelo coñecido (usualmente lineal) ou aqueles que só dependen da natureza das variables implicadas (resposta e covariables). Entre as primeiras se inclúen métodos clásicos como a regresión stepwise ou máis modernos como a regresión LASSO que se aplican a modelos lineais. Entre os segundos cabe destacar o método mRMR (minimum Redundancy Maximum Relevance) ou algoritmos baseados na correlación de distancias. O obxectivo do traballo é presentar estas técnicas e analizar o seu funcionamento tanto en exemplos de simulación como en datos reais.
Recomendaciones: Bo nivel de programación en R
Otras observaciones

Índice

Resumen

La selección de variables es un tema de gran relevancia, especialmente en contextos donde la abundancia de posibles covariables dificulta la integración en modelos. Los métodos para esta selección se dividen en aquellos que asumen una estructura de modelo conocida, típicamente lineal, y aquellos que dependen únicamente de la naturaleza de las variables involucradas (respuesta y covariables). En la primera categoría se encuentran enfoques clásicos como la regresión Stepwise y técnicas más modernas como la regresión LASSO, aplicables a modelos lineales. La segunda categoría incluye métodos como mRMR (minimum Redundancy Maximum Relevance) y algoritmos basados en correlación de distancias. Este trabajo tiene como objetivo presentar estas técnicas y analizar su eficacia mediante ejemplos de simulación y datos reales.

Abstract

Variable selection is a highly relevant topic, particularly in contexts where an abundance of potential covariates complicates model integration. Methods for selection can be categorized into those assuming a known model structure, typically linear, and those relying solely on the nature of involved variables (response and covariates). Classic approaches like Stepwise regression and modern techniques such as LASSO fall into the former category, applicable to linear models. The latter category includes methods like mRMR (minimum Redundancy Maximum Relevance) and distance correlation-based algorithms. This paper aims to introduce these techniques, analyzing their efficacy through simulations and real-world data examples.

Resumo

A selección de variables é un tema de gran relevancia, especialmente en contextos nos que a abundancia de posíbeis covariables complica a integración en modelos. Os métodos para esta selección divídense en aqueles que asumen unha estrutura de modelo coñecida, tipicamente lineal, e aqueles que dependen exclusivamente da natureza das variables implicadas (resposta e covariables). Na primeira categoría atopamos enfoques clásicos como a regresión Stepwise e técnicas máis modernas como a regresión LASSO, aplicables a modelos lineais. A segunda categoría inclúe métodos como mRMR (minimum Redundancy Maximum Relevance) e algoritmos baseados na correlación de distancias. Este traballo ten como obxectivo presentar estas técnicas e analizar a súa eficacia mediante exemplos de simulación e datos reais, explorando a súa aplicabilidade na

práctica científica e analítica contemporánea.

Introducción

La selección de variables es un paso fundamental en los modelos de regresión. En muchos conjuntos de datos, existe una abundancia de variables, algunas de las cuales pueden ser irrelevantes o redundantes, lo que puede afectar a la precisión y eficiencia del modelo. La correcta identificación de las variables más relevantes puede mejorar significativamente la interpretabilidad y el rendimiento del modelo.

Este informe se centra en la evaluación de diversos métodos de selección de variables, abarcando tanto enfoques clásicos como modernos, como son:

- **Selección por Pasos (Stepwise Selection)**
- **Selección Hacia Adelante (Forward Selection)**
- **Eliminación Hacia Atrás (Backward Elimination)**
- **Best-subset**
- **LASSO (Least Absolute Shrinkage and Selection Operator)**
- **LASSO agrupado**
- **mRMR (Minimum Redundancy Maximum Relevance)**

El objetivo de este TFG es comparar la eficacia de estos métodos en la selección de variables relevantes, considerando su precisión, interpretabilidad y eficiencia computacional. Para ello, se llevará a cabo un estudio de simulación exhaustivo bajo diferentes escenarios de datos sintéticos. Además, se incluirá un análisis práctico utilizando un conjunto de datos reales, proporcionando una aplicación concreta de estos métodos en situaciones del mundo cotidiano.

Con esta evaluación integral, se busca destacar las fortalezas y debilidades de cada enfoque y ofrecer recomendaciones sobre su uso en distintos contextos de análisis de datos y modelado predictivo.

Se puede presentar el problema de selección de variables de la siguiente forma:

Dado un conjunto de variables predictoras $M = \{X_1, X_2, \ldots, X_p\}$ y una variable respuesta Y, el objetivo es encontrar un subconjunto $N \subseteq M$ que maximice alguna medida de rendimiento del modelo, sujeto a ciertas restricciones. Las restricciones pueden incluir:

- Limitaciones en el tamaño del subconjunto N.
- Restricciones de interpretabilidad del modelo.
- Consideraciones computacionales (como las limitaciones de tiempo de cálculo).

La solución a este problema puede ser computacionalmente costosa debido a la combinatoria de posibles subconjuntos. Por ello, se estudiarán los diferentes métodos para poder diferenciar los escenarios en los que aplicar cada uno.

Capítulo 1

Preliminares

Para afrontar el problema de selección de variables expuesto en la introducción, se necesita entender el concepto de regresión. Se trata de una técnica estadística utilizada para examinar la relación entre una variable dependiente (también llamada variable respuesta) y una o más variables independientes (también llamadas variables predictoras o explicativas). Dependiendo de las características de la función con la que se modela la relación entre los dos tipos de variables, se pueden tener regresiones lineales y no lineales. En este caso, se abordará el problema de la selección de variables en modelos de regresión lineal general.

La formulación del modelo lineal dependerá en primera instancia de la cantidad de variables explicativas (denotadas normalmente por X), ya que el modelo que relaciona la variable respuesta (denotada por Y) con una única variable explicativa, se conoce como modelo de regresión lineal simple, y el que relaciona la variable respuesta con varias variables predictoras son los modelos de regresión lineal múltiple (en este caso, las variables explicativas serán X_i con $i \in \{1, \ldots, p\}$ con p el número de variables). Ante esta situación y con el objetivo de afrontar un problema de selección de variables, se presentará únicamente el modelo lineal general con más de una variable explicativa.

Sean ahora una variable respuesta Y, p variables explicativas y n observaciones, se formula el modelo de regresión lineal general, de la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

donde $\beta_1, \ldots, \beta_p$ son los parámetros asociados a las p variables explicativas y β_0 será el intercepto. Además, como se dispone de n observaciones se tiene que $Y = (Y_1, \ldots, Y_n)^T$ y $X_j = (x_{1j}, \ldots, x_{nj})^T$ para todo $j \in \{1, \ldots, p\}$, y $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. En este caso, el modelo de regresión

lineal múltiple se expresaría con la siguiente forma matricial:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i2} & \cdots & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

siendo X una matriz no aleatoria $n \times (p+1)$, donde cada fila representa un individuo distinto y cada columna una característica concreta (x_{ij}). Teniendo en cuenta que $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, se obtiene otra expresión matricial más abreviada equivalente a la anterior:

$$Y = X\beta + \epsilon$$

Observación 1.1. Este es el caso del modelo de regresión lineal múltiple, la primera columna estará formada por todo 1, que representará la participación del intercepto β_0 en el modelo.

Ahora bien, todo modelo lineal general debe cumplir las siguientes hipótesis:

- **Linealidad:** Por la propia definición de modelo lineal general, se tiene la siguiente expresión lineal: $Y = X\beta + \epsilon$.
- **Independencia:** Los errores aleatorios son independientes entre sí, es decir, $E(\epsilon_i, \epsilon_j) = 0 \forall i \neq j$.
- **Homocedasticidad:** Todos los errores tienen la misma varianza: $\text{Var}(\epsilon_i) = \sigma^2$.
- **Normalidad:** La distribución de los errores aleatorios sigue una distribución normal: $\epsilon \sim N(0, \sigma^2)$.

1.1. Obtención de los parámetros

Para la obtención de los parámetros del modelos, se tienen: un modelo que relaciona la variable Y con las diferentes variables X_i, unas hipótesis que ha de cumplir dicho modelo, y se necesita estimar los valores de las distintas componentes del vector β, el vector de parámetros de la regresión. Para ello, se usará el método de los mínimos cuadrados (que bajo la hipótesis es equivalente a la máxima verosimilitud, ideal en la mayoría de los métodos estadísticos), buscando encontrar el vector $\hat{\beta}$ que minimice la función $RSS(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ donde y_i es el valor observado de la variable dependiente para la observación i, mientras que, $\hat{y}_i$ es el valor predicho por

el modelo en esa misma observación. Además, RSS es la suma residual de cuadrados, denotada así por sus siglas en inglés. Se puede definir el vector buscado como:

$$\hat{\beta} = \text{argmin}_\beta RSS(\beta) = \text{argmin}_\beta (Y - X\beta)^T (Y - X\beta) = \text{argmin}_\beta ||(Y - X\beta)||^2$$

Derivando la función RSS con respecto a β e igualando a 0 se consiguen las ecuaciones normales de la regresión:

$$X^T X\beta = X^T Y$$

de las cuales se obtendrán los $p+1$ parámetros del modelo. Despejando β se obtiene el estimador por mínimos cuadrados:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.1}$$

Observación 1.2. Como anotación se debe tener en cuenta que para definir este estimador es necesario que la matriz $X^T X$ sea invertible. Esta es cuadrada de orden $p + 1$, simétrica y semi-definida positiva, de manera que, su rango coincide con la dimensión del espacio lineal en el que se encuentran las columnas $1, x_i 1, \ldots, x_i p$. De esta forma, se necesitan al menos $p + 1$ individuos $(n \geq (p + 1))$ para poder definir $X^T X$ y por tanto, poder definir $(X^T X)^{-1}$.

Con la estimación del vector β obtenida, se pueden calcular las predicciones para cualquier individuo de la muestra usando la siguiente expresión:

$$\hat{Y} = X\hat{\beta}$$

Sustituyendo 1.1 en la expresión anterior se obtiene que:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

donde la matriz $H = X(X^T X)^{-1} X^T$ se denomina matriz hat. Los elementos de la diagonal de esta matriz se conocen como leverages y muestran la influencia de cada observación en la predicción de sí misma. Gracias al concepto de leverage se pueden detectar las observaciones que más influyen en el modelo ajustado, es decir, aquellas que en el caso de omitirse, harían que el modelo fuese distinto.

Ahora bien, la definición de las predicciones y de la matriz hat conduce a la definición del vector de residuos, el cual es clave para poder estimar la varianza del modelo. Se define el vector de residuos como la diferencia entre las predicciones y los valores reales de la variable respuesta, es decir:

$$\hat{\epsilon} = Y - \hat{Y} = (I_n - H)Y = MY$$

siendo M una matriz de orden $n - (p + 1)$ denominada matriz generadora de residuos.

La importancia de este vector de residuos surge por el desconocimiento del valor de los errores, y la imposibilidad de estimar la varianza del modelo. Una vez definido el vector $\hat{\epsilon}$, se define el estimador de la varianza del error de la siguiente forma:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \vec{x}_i\hat{\beta})^2 = \frac{RSS(\hat{\beta})}{n-p-1}$$

Además, usando que $RSS(\hat{\beta}) = \hat{\epsilon}^T\hat{\epsilon}$ y $\hat{\epsilon} = MY$, entonces se tiene que:

$$\hat{\sigma}^2 = \frac{(MY)^T MY}{n-p-1} = \frac{Y^T M^T MY}{n-p-1} = \frac{Y^T MY}{n-p-1}$$

donde en la última igualdad se aplica la propiedad de idempotencia de la matriz M, es decir, $M \cdot M = M$, o equivalentemente, $M^2 = M$.

Observación 1.3. Análogamente a la estimación de $\hat{\beta}$, solo se obtendrá un estimador para la varianza cuando $X^T X$ sea invertible. Además, el denominador tomado es $n-p-1$ para asegurar que dicho estimador sea insesgado.

1.2. Problemas en los métodos de regresión

En el marco del modelo de regresión lineal general, y específicamente en los modelos de regresión lineal múltiple pueden surgir diversos problemas. Entre los destacados aparece el problema de la multicolinealidad y los modelos heterocedásticos.

Problema de multicolinealidad

Para entender el problema de la multicolinealidad, basta suponer la situación en que alguna variable explicativa es linealmente dependiente con otras. Esto provoca que la matriz $(X^T X)^{-1}$ obtenida al estimar tanto $\hat{\beta}$ como $\hat{\sigma}^2$ sea no invertible. Análogamente, cuando los coeficientes de correlación simples o múltiples sean 1, se encuentra con la misma problemática.

En la práctica, se miden los diferentes grados de colinealidad entre las variables. En estos casos no se incumple ninguna de las hipótesis de partida y se puede estimar el modelo como se explicó con anterioridad. Lo interesante reside en medir esos grados de colinealidad, para ello se suelen usar la matriz de correlación y el factor de inflación de la varianza que se denominará (VIF) por sus siglas en inglés. Por un lado, la matriz de correlación recoge las relaciones lineales entre cada par de variables utilizando el coeficiente de Pearson. Por otro lado, para definir `VIF`, es necesario definir la varianza de $\hat{\beta}_j$, que viene dada por la siguiente expresión:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{nS_{xj}^2} \quad \text{para } j \in \{1,\ldots,p-1\} \tag{1.2}$$

donde el segundo término de 1.2 es la varianza del coeficiente β_j si se realiza una regresión con esa única variable, siendo $S_{xj}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ la varianza muestral de la j-ésima variable explicativa. Mientras que, el primer término de 1.2 será el `VIF` donde R_j^2 es el coeficiente de determinación de la j-ésima variable explicativa al hacer regresión de ella sobre las demás variables explicativas. Teniendo en cuenta esto, el $\texttt{VIF} = \frac{1}{1-R_j^2}$, y dependerá del R_j^2.

Observación 1.4. En el caso de que $R_j^2 = 0$, entonces $\texttt{VIF} = 1$. En este caso, la varianza del estimador se corresponderá con la obtenida en una regresión simple con la misma σ^2.

Observación 1.5. Además, cuanto mayor sea el grado de multicolinealidad, más próximo a 0 estará el determinante de X^TX, por lo que los estimadores serán indeterminados y las varianzas tenderán a infinito. También se debe tener en cuenta que una alta multicolinealidad perjudica la capacidad predictora del modelo de varias maneras. Entre ellas destacan: la dificultad en la interpretación de los coeficientes debido a que es difícil discernir los efectos de cada variable explicativa, la inflación de las varianzas de los coeficientes, el Overfitting o sobreajuste, o la insensibilidad de los coeficientes que pueden sufrir ante pequeñas perturbaciones de los datos.

Dentro de las causas, aparecen la introducción de dos o más variables directamente relacionadas (causa-efecto), así como, tener muestras pequeñas en relación al número de variables, o que los valores de cierta variable sean parecidos y, consecuentemente, estén todos muy próximos a la media. Con respecto a las formas de localizar los grados elevados de multicolinealidad, obtener estimadores con signos contrarios a los que deberían tener es un signo claro de estar ante la problemática. Además, una manera muy práctica a la hora de la localización es el uso del `VIF` para el cual se plantea un umbral de $\texttt{VIF} > 5$ que indicará cuando se considera un grado de colinealidad alto. (Sin embargo, este forma de actuar se debe combinar con otras, ya que puede existir colinealidad entre variables con un `VIF` bajo, en el caso de que el determinante de la matriz de correlación es mayor que 0.5)

Finalmente, para corregir dicho problema, se puede aumentar el tamaño muestral o encontrar las variables causantes de la multicolinealidad y, en caso de detectar relaciones lineales directas entre ellas, y seleccionar solo una de ellas.

Modelos heterocedásticos

Con respecto al problema de los modelos heterocedásticos, aparece una situación en la que los valores de las varianzas de los errores condicionados a X no se mantienen constantes. Ante estos modelos, se puede proceder realizando la estimación de los parámetros usando el método de mínimos cuadrados, que bajo la hipótesis de normalidad es equivalente al de máxima verosimilitud, ideal para la mayoría de los métodos estadísticos. Sin embargo, existe la posibilidad de que el estimador deja de ser óptimo y algunos de los contrastes no son válidos.

Este problema de heterocedasticidad puede ser causado por la existencia de factores que influ-

yan en la varianza de los errores producida por una variable explicativa, o cuando la distribución de la variable respuesta es una Binomial, una Gamma o una Poisson, entre otras. Ahora bien, para detectar esta problemática se puede proceder de varias maneras entre las que se encuentran la aplicación de diferentes test y contrastes que serán mencionados y cuya explicación se encuentra en [10]. Las posibles vías a seguir son:

- Representar los residuos en función de cada variable explicativa X_j, de modo que el modelo puede ser heterocedástico si no se observa la misma dispersión para cualquier valor de X_j.
- Cuando se dispone de muestras pequeñas, se puede aplicar el test de Goldfeld-Quandt.
- Aplicando el test de Breusch-Pagan, que consiste en ajustar un modelo de regresión lineal con variable respuesta los residuos al cuadrado del modelo original.
- En caso de tener muestras grandes, también podemos usar el test de White.

Capítulo 2

Métodos clásicos de regresión lineal

Para abordar el problema de selección de variables en un modelo lineal general existen diversos métodos y estrategias. Partiendo del modelo de regresión lineal general con p variables explicativas, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$ se tienen diferentes maneras de buscar un subconjunto $N \subseteq M$ de variables explicativas. Entre ellas, se podría pensar en buscar a través de todos los subconjuntos posibles (por ejemplo, para $p > 40$ es inabordable puesto que las combinaciones posibles serían 2^{40}). Otra forma sería buscar un buen camino a través de ellos. Aparecen así los conocidos como métodos clásicos de regresión lineal: métodos Forward (o selección hacia delante), Backward (o selección hacia atras) o la combinación de ambas direcciones (Stepwise). La cuestión se encuentra en saber hasta cuando se incluyen o se excluyen las variables explicativas en función al tipo de método que se esté usando. Para ello, se expondrán unos criterios que ayudan a evaluar los diversos modelos y escoger el subconjunto N óptimo.

2.1. Criterios de evaluación del orden de métodos estadísticos

Los criterios estadísticos que se presentan están basados en la idea del principio de parsimonia, priorizando así la selección de modelos con el mínimo número de variables posible. Esto, siempre y cuando, RSS sea pequeño, obteniendo así un modelo óptimo, además de parsimonioso. De todas formas, a la hora de seleccionar se deben tener en cuenta los aspectos como las diferentes opciones de tamaño del subconjunto de variables y los consecuentes puntos de vista de la magnitud de estas diferencias, que son relevantes a la hora de la comparación de modelos.

Por todo esto, se presentan los siguientes criterios:

2.1.1. El coeficiente de determinación (R^2)

El coeficiente de determinación (R^2) es un criterio global que muestra la variabilidad de la variable respuesta explicada por la regresión, y que se formula de la manera siguiente:

$$R^2 = 1 - \frac{RSS}{TSS}$$

donde $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ es la suma total de los cuadrados, denominada así por sus siglas.

Dentro de las principales propiedades de este coeficiente destaca que el valor varía entre 0 y 1, lo que supone que, cuanto más cercano es el valor a 1 mejor será el ajuste del modelo a la variable dependiente. El problema reside en que aumenta su valor al aumentar el número de variables explicativas del modelo.

2.1.2. El coeficiente de determinación ajustado ($R^2_{\mathbf{aj}}$)

El coeficiente de determinación ajustado (R^2_{aj}) es un criterio global que, a diferencia del R^2, mide la proporción de la varianza total de la variable dependiente que es explicada por el modelo de regresión, ajustada al número de variables explicativas en el modelo y al tamaño de la muestra. De este modo, se corrige el problema anterior con dicho ajuste. Este coeficiente se define de la siguiente manera:

$$R^2_{\text{aj}} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

siendo p el número de grados de libertad de RSS y n el tamaño muestral.

De manera análoga al coeficiente de determinación R^2, los valores del R^2_{ajust} están entre 0 y 1, siendo 0 la peor explicación posible y 1 la mejor. Cabe destacar que cuando se incorpora la cantidad de parámetros en el denominador de la ecuación, sucede que cuantas más variables se incorporen, menor será el valor de la expresión (son inversamente proporcionales), esto supone que se penaliza la incorporación de nuevas variables, corrigiendo así el problema que presentaba R^2.

2.1.3. Criterio de Información de AIC

El Criterio de Información de Akaike (AIC) es un criterio global que está definido por Akaike en [1] con la siguiente formulación:

$$AIC = -2\ln(L(\hat{\theta})) + 2p$$

donde $L(\hat{\theta})$ es la función de máxima verosimilitud y p es el número de variables explicativas del modelo. Este criterio maximiza el logaritmo de verosimilitud al aplicar $-2\ln(\theta)$, justificando

su uso bajo el supuesto de la teoría estándar de muestras grandes de la estimación de máxima verosimilitud.

En la práctica, cuando se use el AIC para modelos de regresión lineal, y en específico, para el modelo de regresión lineal múltiple, se utilizará la siguiente formulación:

$$AIC = n\ln(2\pi) + n\ln(RSS/n) + n + 2p$$

donde, igual que anteriormente, RSS es la suma residual de cuadrados, p es el número de variables predictoras en el modelo de regresión y n es el número de observaciones de la muestra.

Finalmente, cabe destacar que a la hora de seleccionar un modelo bajo este criterio, se escogerá el modelo con menor valor AIC. De este modo el criterio es combinable con cualquiera de los métodos clásicos de regresión lineal. Sin embargo, dicho criterio padece de una cierta inconsistencia, provocada por que $2p$ no crece suficientemente rápido en comparación con el aumento de verosimilitud con más parámetros. Esto lleva a pensar en la búsqueda de un criterio que la mejorase.

2.1.4. Criterio de Información de Bayes (BIC)

El Criterio de Información de Bayes o Criterio de Schwarz (BIC) es un criterio global de selección de modelos desde la perspectiva bayesiana, que aparece con el objetivo de mejorar la inconsistencia que presentaba el AIC. Dicho criterio fue presentado por Schwarz [12] y su formulación general es la siguiente:

$$BIC = -2\ln(L(\hat{\theta}) + p\ln(n)$$

Análogamente al criterio AIC, se maximiza el logaritmo de verosimilitud al aplicar $-2\ln(L(\hat{\theta}))$, justificando su uso bajo el supuesto de la teoría estándar de muestras grandes de la estimación de máxima verosimilitud. En este caso, para los modelos de regresión lineal múltiple se tiene la siguiente formulación:

$$BIC = n\ln(2\pi) + n\ln(RSS/n) + n + p\ln(n)$$

donde RSS es la suma residual de cuadrados, p el número de variables explicativas en el modelo de regresión y n el número de observaciones de la muestra.

Finalmente, destacar que el criterio de BIC comparte muchas propiedades con el criterio de AIC. Además, cuánto mayor sea el número de observaciones más fiabilidad tendrá la estimación realizada. De manera análoga, al criterio de AIC, presentado anteriormente, se selecciona el modelo con menor valor BIC, y podrá ser combinado con cualquiera de los métodos que se presentarán en la siguiente sección.

2.1.5. Criterio de AIC corregido ($AICc$)

Ahora bien, pese al problema de inconsistencia mejorado con este criterio, sigue existiendo un problema con el sesgo de AIC. Para solucionar dicho problema se plantea un criterio global que está aproximado por p, los cuales son constantes y no tienen variabilidad. Para un modelo lineal múltiple se obtiene de [19] que una corrección del sesgo de logaritmo de la función de máxima verosimilitud será:

$$Sesgo = \frac{n(p+1)}{n-p-2}$$

donde n es el número de observaciones y p el número de variables predictoras del modelo.

Surge así, el criterio de AIC corregido ($AICc$), que al aplicar en un modelo de regresión lineal múltiple, se formula de la siguiente manera:

$$AICc = n\ln(2\pi) + n\ln(RSS/n) + n + 2\frac{n(p+1)}{n-p-2}$$

De nuevo, para terminar, tener en cuenta que análogamente al BIC y al AIC, se seleccionará el modelo con menor $AICc$, y dicho criterio es combinable con todos los métodos clásicos.

2.2. Métodos clásicos de regresión lineal

Tras definir los criterios estadísticos que se emplearán para la selección de variables, se presentan los métodos clásicos de regresión lineal, con los que se combinan los criterios presentados, anteriormente, para obtener un modelo simplificado con las variables explicativas suficientes. Estos métodos también son conocidos como parcialmente heurísticos y consisten en la comparación de diferentes modelos formados por distintas agrupaciones de las variables y la selección del mejor de ellos. Los métodos que se presentan son:

2.2.1. Backward

El método Backward, o de selección hacia atrás, consiste en comenzar con un modelo complejo (generalmente con el modelo con todas las variables explicativas), que incorpora todos los efectos que razonablemente pueden influir en la variable respuesta, y se van eliminando variables usando un criterio de evaluación. En cada paso se parte del modelo que ha sido seleccionado en el paso anterior. El método continúa hasta llegar a la mejor combinación de variables según el criterio de evaluación preestablecido. Una estructuración del procedimiento en pasos es la siguiente:

1. Se inicializa el método con el modelo formado por todas las variables explicativas:
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$

2. Se calcula el modelo con cada variable removida X_j, una a una. A continuación, se evalúan usando un criterio de la sección anterior y se selecciona la variable cuya eliminación mejora más el modelo según el criterio elegido.

3. Remover la variable seleccionada del modelo, en este caso la X_j:
   $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_p X_p + \epsilon$.

4. Se continúan eliminando variables (repitiendo los pasos 2 y 3).

5. El procedimiento termina cuando no se puede eliminar ninguna variable adicional que mejore el modelo significativamente.

Con respecto a la implementación en R, de igual manera que se usará en el estudio de simulación, es posible utilizar el comando `step` con `direction='backward'` para realizar un método Backward.

2.2.2. Forward

El método Forward, o de selección hacia adelante, es un algoritmo que consiste en comenzar con un modelo lo más simple posible, generalmente con el modelo nulo $Y = \beta_0 + \epsilon$, y ir añadiendo términos, guiándose por el criterio de evaluación elegido. Una estructuración del procedimiento en pasos es la siguiente:

1. Se inicializa el método con el modelo nulo: $Y = \beta_0 + \epsilon$.

2. Se calcula el modelo con cada variable adicional X_j que no esté actualmente en el modelo. A continuación, se evalúan usando un criterio de la sección anterior y se selecciona la variable cuya adición mejora más el modelo según la regla elegida.

3. Se agrega la variable seleccionada al modelo: $Y = \beta_0 + \beta_j X_j + \epsilon$.

4. Se continúan agregando variables (repitiendo los pasos 2 y 3).

5. El procedimiento termina cuando no se puede agregar ninguna variable adicional que mejore el modelo significativamente.

Este método puede parecer menos óptimo que otros (como el Best-subset, que se presentará más adelante), pero existen dos razones por la que elegirlo frente a otros. En primer lugar, computacionalmente hablando, siempre se puede calcular la secuencia hacia adelante (incluso cuando $p > N$). Por otro lado, estadísticamente hablando, la selección hacia adelante implica

una búsqueda más restringida y como resultado se obtendrá una varianza más reducida, aunque posiblemente, un mayor más sesgo.

Con respecto a la implementación en R, de igual manera que se usará en el estudio de simulación, es posible utilizar el comando `step` con `direction='forward'` para realizar un método Forward.

2.2.3. Stepwise

El método Stepwise es una combinación de los dos métodos, el Backward y el Forward, de modo que se combinan pasos de ambas técnicas, por lo que se añaden, se eliminan o se intercambian las variables. Una estructuración del procedimiento en pasos es la siguiente:

1. Se inicializa el método con un modelo con unas variables determinadas:
   $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ con $k < p$.
2. Se calcula el modelo con cada variable adicional X_j que no esté actualmente en el modelo. A continuación, se evalúan usando un criterio de los explicados anteriormente y se selecciona la variable cuya adición mejora más el modelo según el criterio elegido.
3. Se agrega la variable seleccionada al modelo: $Y = \beta_0 + \ldots + \beta_j X_j + \epsilon$ (suponiendo que están ordenadas y que $j > k$, en otro caso la última variable sería X_k).
4. Se calcula el modelo con cada variable removida X_i, una a una. A continuación, se evalúan usando un criterio de la sección 2.1 y se selecciona la variable cuya eliminación mejora más el modelo según la regla elegida.
5. Se repiten los pasos del 2 hasta el 4.
6. El procedimiento termina cuando no se puede agregar o eliminar ninguna variable adicional que mejore el modelo significativamente.

Con respecto a la implementación en R, de igual manera que se usará en el estudio de simulación, es posible utilizar el comando `step` con `direction='both'` para realizar un método stepwise.

2.2.4. Best-subset

El método Best-subset, o del mejor subconjunto, ajusta para cada $k \in \{0, 1, 2, \ldots, p\}$, el subconjunto de tamaño k que proporciona el mínimo RSS, entre cada una de las posibles combinaciones de las p variables. Aparecen diferentes procedimientos entre los que se destaca el de

saltos y límites, explicado en [17], que es eficaz a la hora de realizar la tarea del método para valores de p tan grandes como 30 o 40. Además, tomando como ejemplo, el mejor subconjunto de tamaño 3, el método asegura que no se tienen porque incluir necesariamente las variables que estaban en el mejor subconjunto de tamaño 2. En este caso, la estructura del método en pasos es la siguiente:

1. Se inicializa el proceso generando el modelo inicial M_0 donde $Y = \beta_0 + \epsilon$.
2. Se generan p modelos introduciendo una variable en cada uno. A continuación, se elije el mejor modelo utilizando un criterio de evaluación del orden de métodos estadísticos y se denota como M_1 al modelo seleccionado.
3. Se repite el paso anterior para modelos de dos variables. Se selecciona así el correspondiente M_2, y así sucesivamente hasta tener seleccionados los modelos $M_1, M_2, \ldots, M_p$.
4. Se selecciona el mejor modelo entre $M_0, M_1, \ldots, M_p$ según determine el criterio de evaluación usado.

Observación 2.1. Como apunte, tener en cuenta que este método no es factible para $p = 40$, ya que de lo contrario, se tendría un coste computacional altísimo, consecuencia del elevado número de variables a comparar. En particular, para $p = 40$ se tendrían $2^{40} = 1099511627776$ modelos diferentes. Además, estadísticamente hablando, se incurre en un costo en términos de varianza al elegir el mejor subconjunto de cada tamaño, en comparación a los métodos Forward.

Capítulo 3

Métodos de reducción: LASSO

Los métodos de selección de subconjuntos de variables vistos hasta ahora mantienen solo un subconjunto de los predictores y elimina los demás creando un modelo que es fácil de interpretar. Según [16], al ser un proceso discreto se tiende a no reducir significativamente el error de predicción y a tener una alta variabilidad. Una forma de solucionar esto son los métodos de reducción, que son más continuos y, por lo tanto, no sufren tanto de alta variabilidad. Entre estos aparecen, la regresión Ridge, la regresión LASSO y la Elastic Net, entre otras, acompañadas de sus variantes y extensiones. Sin embargo, para el problema de selección de variables, se presentará la regresión LASSO, que penaliza ciertas variables llegando a anularlas del modelo y así realizar una selección de variables.

3.1. Least Absolute Shrinkage and Selection Operator(LASSO)

El método LASSO (Least Absolute Shrinkage and Selection Operator) es un método de reducción desarrollado en 1996 por Robert Tibshirani en [15]. Su objetivo es mejorar las predicciones y la interpretabilidad de los modelos de regresión lineal generalizado, minimizando la suma de los cuadrados de los residuos aplicando una penalización.

Considerando $Y \in R^n$ una variable respuesta continua, X la matriz del diseño $n \times (p+1)$, y β el vector de coeficientes de las p variables del modelo de regresión lineal generalizado. En este caso, se desarrollará para una regresión lineal múltiple. Se define LASSO mediante la formulación siguiente:

$$\hat{\beta}_{\text{LA}} = \text{argmin}_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.1}$$

donde $\hat{\beta}_{\text{LA}}$ es el estimador y $\lambda > 0$ es el parámetro que controla la penalización.

Observación 3.1. Se destaca que $\|\cdot\|_2^2$ es la norma euclidiana al cuadrado y $|\cdot|$ el valor absoluto, siendo la norma euclidiana al cuadrado:

$$\|u\|_2^2 = \sum_{i=1}^{n} u_i^2 \tag{3.2}$$

De esta manera, aplicando 3.2 a 3.1, se obtiene:

$$\hat{\beta}_{\text{LA}} = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.3}$$

Cabe destacar que el término de penalización de LASSO, $\lambda \sum_{j=1}^{p} |\beta_j|$, está basado en la penalización L_1, lo que hace que no haya una expresión en forma cerrada ya que las soluciones son no lineales en y_i. Por otro lado, a medida que λ aumenta, la reducción correspondiente se intensifica, lo que conlleva a que para valores grandes de λ, la penalización resulte en $\beta_j = 0$ para algún $j \in \{1, 2, \ldots, p\}$, es decir, varios coeficientes se vuelven cero y se anulen sus respectivas variables. Consecuentemente, el método busca identificar aquellos estimadores que mejoren la precisión de la predicción y que al mismo tiempo reduzcan la varianza de los valores estimados.

Ahora bien, a la hora de la práctica para aplicar el método, se toma un conjunto de datos $\{(x_i, y_i)\}_{i=1,\ldots,n}$, que, sin pérdida de generalidad, se centrarán y estandarizarán las variables explicativas, es decir, $\forall j = 1, \ldots, p$, se tendrá que:

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{y} \quad \sum_{i=1}^{n} \frac{x_{ij}^2}{n} = 1$$

Ademas, en el caso la penalización sobre β_0 no se impone ya que si no el proceso dependería del origen elegido para Y. Por otro lado, se destaca que la solución de la regresión LASSO no es invariante ante cambios de escala, si se agregara una constante a cada variable y_i, esta situación no se corregiría aplicando el mismo cambio en la predicción. Con todo esto, se estimará β_0 por $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, mientras que el resto de las variables utilizando el método de mínimos cuadrados, sin tener en cuenta al intercepto. Se llega así, a que el método LASSO, 3.3, se puede formular como un problema de minimización:

$$\hat{\beta}_{\text{LA}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ sujeto a } \sum_{j=1}^{p} |\beta_j| \leq t. \tag{3.4}$$

donde $\sum_{j=1}^{p} |\beta_j| \leq t$ es la restricción impuesta, y en consecuencia, al hacer t lo suficientemente pequeño, algunos coeficientes se convierten en exactamente cero. Ahora bien, de [16] se obtiene que para la elección de t, que es clave a la hora de minimizar una estimación del error de predicción, se debe tener en cuenta que:

$$\begin{cases} t > t_0 = \sqrt{\sum_{j=1}^{p} |\beta_j|^2}, & \text{donde} \quad \hat{\beta}_j = \hat{\beta}_{ls,j} \\ t \leq t_0 = \sqrt{\sum_{j=1}^{p} |\beta_j|^2} & \text{entonces} \quad \hat{\beta}_{ls,j} \quad \text{se reducen en promedio} \end{cases}$$

siendo $\hat{\beta}_{ls,j}$ los coeficientes del modelo de regresión lineal generalizado estimados mediante mínimos cuadrados.

Finalmente, hay que tener en cuenta ciertas consideraciones a la hora de utilizar el método LASSO. Entre otras muchas, destacan:

- En el caso $p > n$, LASSO selecciona a lo sumo n variables antes de saturarse.
- Si existe un grupo de variables en las que la correlación es muy alta, tiende a elegir una variable del grupo sin importar cuál sea.
- En el caso $n > p$ o $n < p$, si existe una alta correlación entre las variables predictoras, en general, la predicción no es óptima ya que es un modelo sesgado.

3.2. Elección del parámetro λ

La selección de variables explicativas está influenciada por el valor de λ. Como se explicó anteriormente, a medida que este parámetro aumenta, la penalización de los coeficientes de regresión se intensifica y estos tienden a aproximarse a cero. Por esto último, es importante realizar una buena elección del parámetro. Para ello existen diversos métodos, pero en este trabajo se realizará un método de validación cruzada. Dentro de los métodos de validación cruzada (Cross-Validation) existen diferentes tipos en función a la elección de los subconjuntos de aplicación:

- **La validación cruzada dejando uno fuera (Leave-one-out cross-validation)**: es una técnica para evaluar el rendimiento de un modelo predictivo. Se entrena el modelo

utilizando todos los datos excepto uno, que se reserva como conjunto de prueba. Luego, se repite este proceso para cada punto de datos, calculando la métrica de rendimiento deseada. Al promediar los resultados de todas las iteraciones, se obtiene una estimación robusta del rendimiento del modelo en datos no vistos. Este método es particularmente útil cuando se dispone de una cantidad limitada de datos y se busca una evaluación exhaustiva del modelo. El principal inconveniente de este enfoque es su alta carga computacional.

- **La validación cruzada de K pliegues (K-fold cross-validation)**: es una técnica para evaluar el rendimiento de un modelo predictivo cuando los datos son limitados. La idea principal es dividir los datos disponibles en K partes, aproximadamente iguales, llamadas pliegues. Luego, se entrena y evalúa el modelo K veces, cada vez utilizando un pliegue diferente como conjunto de prueba y los K-1 pliegues restantes como conjunto de entrenamiento. Este proceso proporciona una estimación más robusta del rendimiento del modelo al promediar los resultados de las K iteraciones. Aunque esta técnica es precisa, su coste computacional es significativo, por lo que se suele utilizar 10 iteraciones para equilibrar precisión y eficiencia.

- **La validación cruzada aleatoria**: es una técnica para evaluar el rendimiento de un modelo predictivo utilizando una selección aleatoria de conjuntos de entrenamiento y prueba. En este enfoque, se dividen aleatoriamente los datos en un conjunto de entrenamiento y un conjunto de prueba en cada iteración. Esto se repite múltiples veces para obtener una estimación más robusta del rendimiento del modelo. La validación cruzada aleatoria es especialmente útil cuando se trabaja con conjuntos de datos grandes y se desea una evaluación rápida y eficiente del modelo. El principal inconveniente de este enfoque es que no ofrece garantías de que pueda calcular la predicción de manera uniforme para todos los datos.

En caso de disponer de grandes conjuntos de datos, existen otros métodos como el de retención o holdout method. Este enfoque implica dividir el conjunto de datos en dos partes: la primera se denomina conjunto de entrenamiento (`training set`) y su complemento se conoce como conjunto de validación (`test set`). Esta división se realiza con el propósito de entrenar el modelo utilizando el conjunto de entrenamiento y luego evaluar su rendimiento con el conjunto de validación. Aunque no requiere un alto costo computacional, los resultados están influenciados en gran medida por la forma en que se realizan las particiones.

3.3. LASSO agrupado

El método LASSO agrupado es una variante del método LASSO que aparece para solucionar aquellas situaciones en las que, además de disponer de variables explicativas continuas se tienen variables categóricas (o factores), y donde se seleccionan variables ficticias individuales en vez de variables categóricas completas, por lo que esta solución dependería de la codificación de estas variables ficticias. Para solucionar este problema lo idóneo sería agrupar y seleccionar conjuntamente los miembros de un grupo, aparece así el LASSO agrupado propuesto por [20] y ,posteriormente, generalizado por [18]. Con este método se supera el problema incorporando una extensión de la penalización LASSO. Las formulación del método ha sido obtenida de [21].
Para introducir el método, se supone que hay p variables explicativas distribuidas en G grupos, por lo tanto, el estimador de LASSO agrupado se define como:

$$\hat{\beta}_{\text{LA}} = \text{argmin}_{\beta} \left\{ \|Y - X\beta\|_2^2 + \sum_{g=1}^{G} \lambda_g \|\beta_{I_g}\|_1 \right\} \tag{3.5}$$

siendo I_g el conjunto de índices que pertenecen al g-ésimo grupo de variables, con $g \in \{1, \ldots, G\}$. Ahora bien, de igual manera que en la regresión LASSO, usando 3.2, se puede reescribir la formulación 3.5 de la siguiente manera:

$$\hat{\beta}_{\text{LA}} = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 \mathbf{1} - \sum_{g=1}^{G} x_{ig} \beta_g \right)^2 + \sum_{g=1}^{G} \lambda_g \sqrt{p_g} \|\beta_g\|_1 \right\} \tag{3.6}$$

siendo p_g el número de variables explicativas en el grupo g con $g \in \{1, \ldots, G\}$. Además, se utiliza la notación x_{ig}^{gr} para representar los elementos de la matriz X, la cual contiene las variables explicativas correspondientes al grupo $g \in \{1, \ldots, G\}$, junto con β_g su coeficiente contenido en el vector de coeficientes correspondiente β. Y de este modo, el método del LASSO agrupado será también presentado como problema de minimización sujeto a restricciones:

$$\underset{\beta}{\text{argmin}} ||y - \beta_0 \mathbf{1} - \sum_{g=1}^{G} X_g \beta_g||_2^2 + \sum_{g=1}^{G} \lambda_g \sqrt{p_g} ||\beta_g||_1 \tag{3.7}$$

donde los λ_g es el lambda del grupo g-ésimo. A diferencia de la solución LASSO anterior, para esta variante se escoge un λ para cada grupo de forma que se tengan en cuenta las características de cada grupo a la hora de penalizar al modelo.

Observación 3.2. En la formulación del método, los tamaños de cada grupo se tienen en cuenta en los términos $\sqrt{p_g}$. Además, se fomenta la dispersión (a nivel grupal e individual), ya que, para algunos λ, todo un grupo de variables explicativas puede quedar anulado. Esto último se debe a que $\|\beta_g\|_2 = 0 \iff \beta_g = \mathbf{0}$, es decir, todas las componentes son nulas.

Capítulo 4

Minimum Redundancy Maximum Relevance (mRMR)

El método de Mínima Redundancia Máxima Relevancia (mRMR) para la selección de variables tiene como propósito identificar las variables más relevantes para predecir la variable respuesta, eliminando aquellas cuya contribución puede ser duplicada por otras. Este método fue presentado en [3] y ofrece una implementación formal de un procedimiento de selección de variables que considera, explícitamente, el equilibrio entre relevancia y redundancia. Los conceptos de este capítulo están basados en [2], [14] y [3].

Antes de la explicación de los conceptos de relevancia y redundancia, se fija un número k de variables seleccionadas del conjunto de datos inicial ($k \leq p$), que se agregarán a un subconjunto N, $|N| = k$, donde $|N|$ denota el tamaño del subconjunto N, es decir, el número de variables incluidas en él.

4.1. Relevancia y Redundancia

Los conceptos de Relevancia y Redundancia se presentarán a continuación:

Definición 4.1. La relevancia VC y la redundancia WC se definen como:

$$VC(X_i) = \frac{1}{|N|} \sum_{X_i \in N} I(X_i, Y) \tag{4.1}$$

$$WC(X_i, X_j) = \frac{1}{|N|^2} \sum_{X_i, X_j \in N} |I(X_i, X_j)| \tag{4.2}$$

siendo N el conjunto de variables seleccionadas, $|N|$ su cardinal e I una función que mide el grado de relación entre dos variables. Por lo tanto, es natural pensar que la relevancia de X_i se

mide por su asociación con la variable de respuesta Y, representada por $I(X_i, Y)$, mientras que, la redundancia entre X_i y X_j está dada por $I(X_i, X_j)$.

Análogamente a métodos como la regresión LASSO o LASSO grupal, se puede expresar esta problemática de selección de covariables como un problema de optimización, buscando maximizar la relevancia y minimizar la redundancia simultáneamente. Matemáticamente hablando, se puede plantear el problema de la siguiente forma:

$$\max_{N \subseteq M} (VC(N) - WC(N))$$

donde:

- M es el conjunto de todas las características.
- N es un subconjunto de características seleccionadas.
- $VC(N)$ es una medida de la relevancia de las características en el conjunto N con respecto al objetivo del modelo.
- $WC(N)$ es una medida de la redundancia entre las características en el conjunto N.

Para explicar el funcionamiento del método mRMR, se tiene en cuenta que el problema de selección de variables se puede expresar como uno de optimización, que se resuelve mediante los siguientes pasos:

1. El procedimiento se inicia seleccionando la variable más relevante, representada por el valor t_i, de tal manera que, el conjunto $N_i = \{t_i\}$ maximice VC(N) entre todos los conjuntos unitarios de tipo $N_j = \{t_j\}$.
2. Posteriormente, las variables se incorporan secuencialmente al conjunto N de variables previamente seleccionadas, siguiendo el criterio de maximizar la diferencia:

   $$F_{CD} = \max(VC - WC) = \max \left(\frac{1}{|N|} \sum_{X_i \in N} I(X_i, Y) - \frac{1}{|N|^2} \sum_{X_i, X_j \in N} |I(X_i, X_j)| \right)$$

   o, de manera alternativa, el cociente:

   $$F_{CQ} = \max \left(\frac{VC}{WC} \right) = \max \left(\frac{1}{|N|} \sum_{X_i \in N} I(X_i, Y) \div \frac{1}{|N|^2} \sum_{X_i, X_j \in N} |I(X_i, X_j)| \right)$$

3. Por último, se pueden considerar distintas reglas de detención. Como, por ejemplo, en el estudio de simulación del capítulo 5 se usará el criterio de AIC para seleccionar el óptimo de los modelos obtenidos tras realizar el mRMR. Entre otros criterios se encuentran los expuestos en el capítulo 2.

4.1.1. Ventajas/Desventajas de implementar mRMR

Algunas de las ventajas a destacar son:

- Al emplear mRMR se descartan variables irrelevantes y redundantes, lo que resulta en un modelo más conciso.
- Al eliminar las variables redundantes y poco significativas, el modelo resultante será más sencillo, lo que facilitará su interpretación.

Por otro lado, un inconveniente que se puede deducir del mRMR es que al fijar inicialmente el número de variables explicativas que vamos a utilizar, la relevancia y redundancia de las variables pueden diferir. Además, cuantas más variables predigamos, más complejo se vuelve computacionalmente.

4.2. Eleccción de I

La elección de la medida de asociación I entre variables (definida en [2]) es un aspecto crítico. Dado que existen diversas propuestas disponibles [2], se requiere una cuidadosa consideración para seleccionar la medida más adecuada para un conjunto de datos específico. En este trabajo, se presentan y utilizan varias medidas de asociación $I(X, Y)$ para examinar la relación entre variables, cada una con sus propias características y aplicaciones.

4.2.1. Coeficiente de correlación de Pearson

Definición 4.2. Dadas un par de variables aleatorias X e Y, se define el coeficiente de correlación de Pearson como:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}$$

donde $\mathrm{Cov}(X, Y)$ es la covarianza entre las variables X e Y, y $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$ son sus respectivas varianzas.

Esta elección presenta algunas desventajas como la no caracterización de la independencia y la no adecuación a la hora de capturar asociaciones no lineales. Además, dado que el valor del coeficiente de Pearson varía entre $[-1, 1]$, donde los valores extremos indican una mayor correlación entre las variables, se toma el valor absoluto de este coeficiente.

4.2.2. Información Mutua

La medida de Información Mutua (denotada por MI por sus siglas en inglés) es la de la versión original del método mRMR. Se ha extraído de [14] la definición siguiente del concepto:

Definición 4.3. Dadas dos variables aleatorias continuas X e Y, la Información Mutua viene dada por la expresión:

$$MI(X,Y) = \int_x \int_y f_{X,Y}(x,y) \cdot \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, dx \, dy$$

siendo $f_X(x)$ y $f_Y(y)$ las funciones de densidad de X e Y respectivamente, y $f_{X,Y}(x,y)$ la distribución de densidad conjunta.

Además, como se explica en [7], el $MI(X,Y)$ medirá la información que proporciona X acerca de Y.

Observación 4.4. Entre las propiedades más destacadas de esta medida se tienen: (obtenidas de [14] y [9])

1. $MI(X,Y)$ cuantifica la distancia entre la distribución conjunta $f_{X,Y}(x,y)$ y la situación de independencia $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
2. $MI(X,Y) = MI(Y,X)$, es decir, es simétrica con respecto a las variables X e Y.
3. $M(X,Y) \geq 0$
4. $MI(X,Y) = 0$ si $\iff$ X e Y son independientes.

Cuando se dispone de variables aleatorias continuas, se puede usar el método de máxima verosimilitud en el caso de conocer la distribución de los datos. Cuando esta es desconocida, se usan métodos no paramétricos como las Ventanas de Parzen explicadas profundamente en [11]. Sin embargo, a la hora de la práctica, según [2] se aproxima la medida $MI(X,Y)$ agrupando, si es necesario, los valores de X e Y en intervalos representados por a_i, b_j, de manera que se discretizan ambos valores. Se obtiene así la siguiente aproximación:

$$\widehat{MI(X,Y)} = \sum_{i,j} \log \left(\frac{P(X = a_i, Y = b_j)}{P(X = a_i) \cdot P(Y = b_j)} \right) \cdot P(X = a_i, Y = b_j)$$

donde las probabilidades pueden ser estimadas de forma empírica por las frecuencias relativas.

4.2.3. Covarianza de distancias

Se considera una muestra aleatoria simple de n observaciones $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ de dos vectores aleatorios de distinta dimensión $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$. Se tiene en cuenta que X e Y tienen

dimensión arbitraria y medias finitas. (los momentos de primer orden de X e Y serán finitos). En este contexto de suposiciones, se define el concepto de covarianza de distancias de la siguiente forma:

Definición 2 (Covarianza de distancias). La covarianza de distancias entre vectores aleatorios X e Y con momentos de primer orden finitos es el número no negativo $V(X,Y)$ definido por:

$$V^2(X,Y) = \|\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)\|^2$$
$$= \frac{1}{c_p c_q} \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} \frac{|\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2}{|t|^{1+p}|s|^{1+q}} dtds$$

donde:

- $\phi_{X,Y}$ es la función característica conjunta de X e Y.
- ϕ_X y ϕ_Y son las funciones características de X e Y respectivamente.
- $w(u,v)$ es una función de peso definida como $w(u,v) = (c_p c_q |t|^{1+p}|s|^{1+q})^{-1}$, con $c_d = \frac{\pi(1+d)}{2\Gamma((1+d)/2)}$ siendo la mitad del área superficial de la esfera unitaria en $\mathbb{R}^{d+1}$, y $|\cdot|_1$ denota la norma L^1 en $\mathbb{R}^d$.

De manera análoga, la varianza de distancia se define como la raíz cuadrada de $V^2(X) = V^2(X,X) = \|\phi_{X,X}(t,s) - \phi_X(t)\phi_X(s)\|^2$, es decir:

$$V(X) = \sqrt{pV^2(X,X)} = \sqrt{q||\phi_{X,X}(t,s) - \phi_X(t)\phi_X(s)||^2}$$

La covarianza de distancia explica la discrepancia entre las funciones características conjuntas y el producto de las funciones características marginales $\phi_X(u)\phi_Y(v)$ ponderada por la función de peso $w(u,v)$

.

Observación 4.5. Propiedades de esta definición

1. X e Y pueden tener dimensiones diferentes.
2. $V^2(X,Y) = 0 \iff X$ e Y son independientes.
3. La elección indicada para los pesos $w(u,v)$ proporciona valiosas propiedades de equivarianza para $V^2(X,Y)$.
4. La cantidad puede ser estimada, consistentemente, a partir de las distancias mutuas $|X_i - X_j|_p$ y $|Y_i - Y_j|_q$ entre los valores muestrales X_i e Y_j (no se necesita discretización).

4.2.4. Correlación de distancia

Definición 4.6. La correlación de distancia se define como:

$$R^2(X,Y) = \begin{cases} \frac{V^2(X,Y)}{\sqrt{V^2(X)V^2(Y)}}, & \text{si } V^2(X)V^2(Y) > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

En el artículo [14] se recogen las siguientes propiedades de la correlación de distancias.

Observación 4.7. Se presentan las propiedades principales de la correlación de distancia:

- $R(X,Y)$ se define para X e Y de dimensión arbitraria
- $R(X,Y) = 0$ caracteriza la independencia de X e Y
- La correlación de distancia satisface $0 \leq R \leq 1$ y $R = 0 \iff X$ e Y son independientes
- En el caso normal bivariable, R es una función de ρ y $R(X,Y) \leq |\rho(X,Y)|$ con igualdad cuando $\rho = \pm 1$

Esta última propiedad, viene justificada por el siguiente teorema extraído de [14], donde viene desarrollada su demostración.

Teorema 4.8. *Si X e Y son normales estándar con correlación $\rho = \rho(X,Y)$, entonces:*

1. $R(X,Y) \leq |\rho|$,
2. $R^2(X,Y) = \rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin \frac{\rho}{2} - \sqrt{4-\rho^2+1}$,
3. $\text{ínf}_\rho \frac{R(X,Y)}{|\rho|} = \lim_{\rho \to 0} \frac{R(X,Y)}{|\rho|} = \frac{1}{2(1+\frac{\pi}{3}-\sqrt{3})^{1/2}} \approx 0{,}89066.$

Capítulo 5

Estudio piloto

Con el objetivo de comparar los modelos presentados a lo largo del informe, se expondrán diversos escenarios con diferentes particularidades, para estudiar así el comportamiento de cada método ante las adversidades. Los modelos escogidos para este estudio serán los métodos clásicos de regresión lineal (Stepwise, Forward, Backward y Best-subset) junto con el LASSO y el mRMR. El caso del Lasso agrupado no se aplica al estudio debido a la necesidad de unos datos con variables explicativas agrupadas. Para la implementación de los distintos modelos y gráficas en R, se han utilizado los paquetes `MASS, car, glmnet, leaps, mRMRe, ggplot2, xtable`. El indicador de calidad a tener en cuenta en este estudio será la selección por parte del método de variables relevantes frente a la de variables irrelevantes, donde la idoneidad del método estaría en seleccionar todas las variables explicativas relevantes y ninguna irrelevante. Además, se mostrarán gráficas comparativas de las veces que los diferentes métodos han seleccionado más o menos variables tanto relevantes como irrelevantes. Para simplificar los nombres de las gráficas se denominarán a los métodos Backward, Forward, Stepwise, Best subset, LASSO y mRMR como `Ba, Fo, Ste, Besub, Las, mRMR`, respectivamente.

5.1. Escenario A

En este Escenario A, se ha fijado un β_1 y un modelo de regresión lineal múltiple con $p = 20$ variables y $n = 100$ observaciones. Se parte entonces de la consideración de que las 10 primeras variables son relevantes y las siguientes 10 irrelevantes.

$$\beta_1 = \texttt{(0.5,0.2,1,1.7,-0.2,-4,2,2.4,-1.8,-0.25,0,0,0,0,0,0,0,0,0,0)}$$

Tras la realización de 100 simulaciones, se ha obtenido como resultado los expuestos en la siguiente gráfica:

Figura 5.1: Promedios de selección de variables relevantes e irrelevantes en el escenario A

En 5.1 se observa que los métodos Forward, Backward y Stepwise se han comportado de una manera análoga a la hora de incluir las variables relevantes siendo el Forward el que menos irrelevantes incluye. Con respecto a los otros tres métodos, el LASSO, el Best-subset y el mRMR han seguido la línea de los tres anteriores, a la hora de incluir variables relevantes, con la diferencia de que en promedio han seleccionado mayor número de irrelevantes, destacando el mRMR y el LASSO, por lo que según estos resultados obtenidos, en base al indicador de calidad propuesto, estos dos serían los que peor funcionan ante este escenario acompañado del Best subset que se queda ligeramente atrás.

A continuación se exponen las gráficas con los resultados de las simulaciones en cada método:

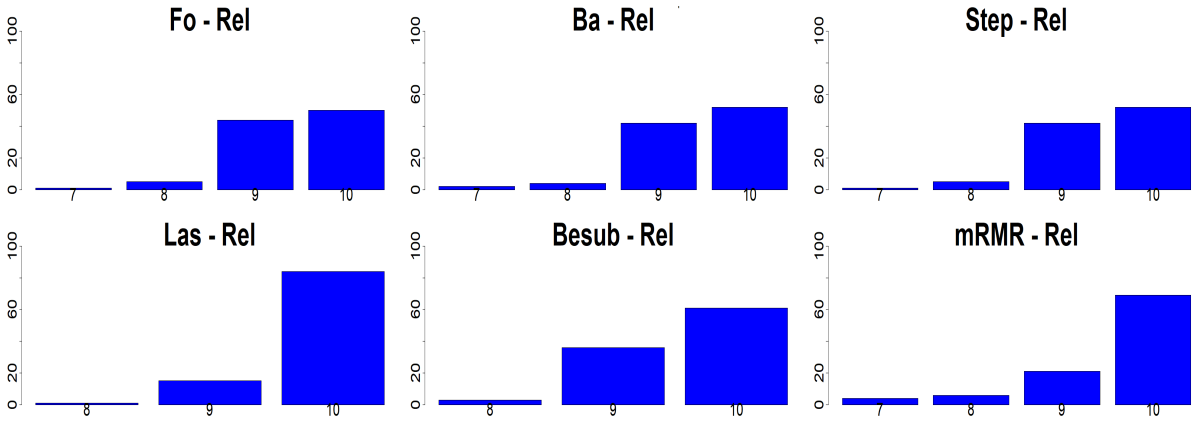

Figura 5.2: Número de veces que se seleccionan variables relevantes en la simulación A

En 5.2, con respecto a las variables relevantes, los métodos Forward, Backward, Stepwise y mRMR han seleccionado unas pocas veces modelos con solo 7 relevantes, mientras que el LASSO y el Best-subset como mínimo 8, siendo destacable que la LASSO selecciona aproximadamente 80 modelos con las 10 relevantes. En el caso del mRMR, añadir que hasta en 60 ocasiones el método selecciona 10 variables relevantes.

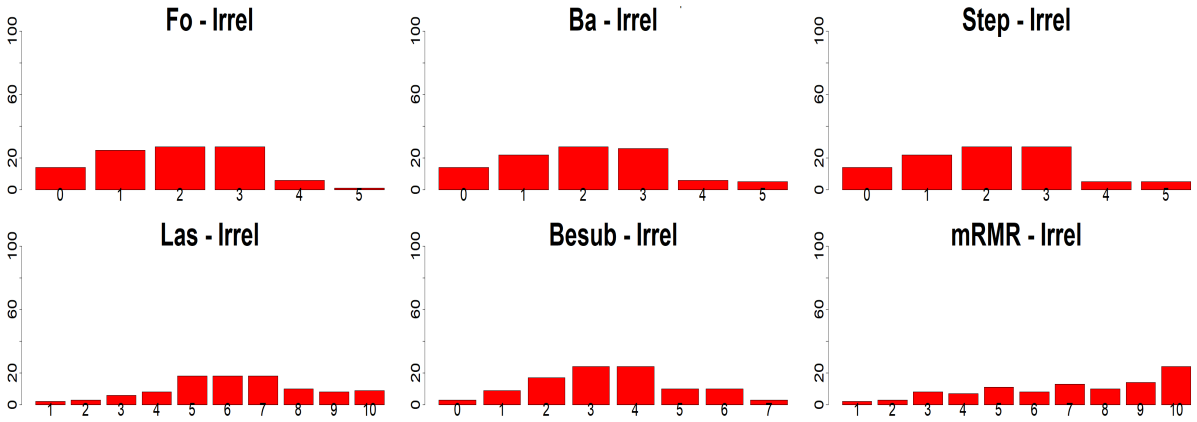

Figura 5.3: Número de veces que se seleccionan variables irrelevantes en la simulación A

En 5.3, con respecto a las variables irrelevantes, los métodos Forward, Backward y Stepwise son los que menos seleccionan, tomando modelos con como mucho 5 variables irrelevantes. Por otro lado, el Best-subset llega a seleccionar modelos con 6 y 7, mientras que, el mRMR y el LASSO se disparan con incluso modelos con todas las irrelevantes.

5.2. Escenario B

En este Escenario B, se ha fijado un β_2 y un modelo de regresión lineal múltiple con $p = 20$ y $n = 100$ observaciones. Además, se ha incluido un problema de heterocedasticidad en los errores del modelo, para estudiar las respuestas de los diferentes modelos ante el mismo escenario pero con la aparición de esta problemática.

$$\beta_2 = \texttt{(0.5,0.2,1,1.7,-0.2,-4,2,2.4,-1.8,-0.25,0,0,0,0,0,0,0,0,0,0)}$$

Tras la realización de 100 simulaciones, se ha obtenido como resultado los expuestos en la siguiente gráfica:

Figura 5.4: Promedios de selección de variables relevantes e irrelevantes en el escenario B

En 5.4, el mRMR se sitúa junto al LASSO como los peores métodos con respecto al indicador de calidad, sin embargo, son los métodos que incluyen más relevantes en sus modelos, destacando que el mRMR casi alcanza en promedio el máximo de variables de este tipo por modelo. Por otro lado, son los métodos clásicos de regresión lineal los que mayor equilibrio muestran y, en especial, el Forward, mientras que el Best susbet se mantiene con un promedio similar al primer escenario, incluyendo menos irrelevantes.

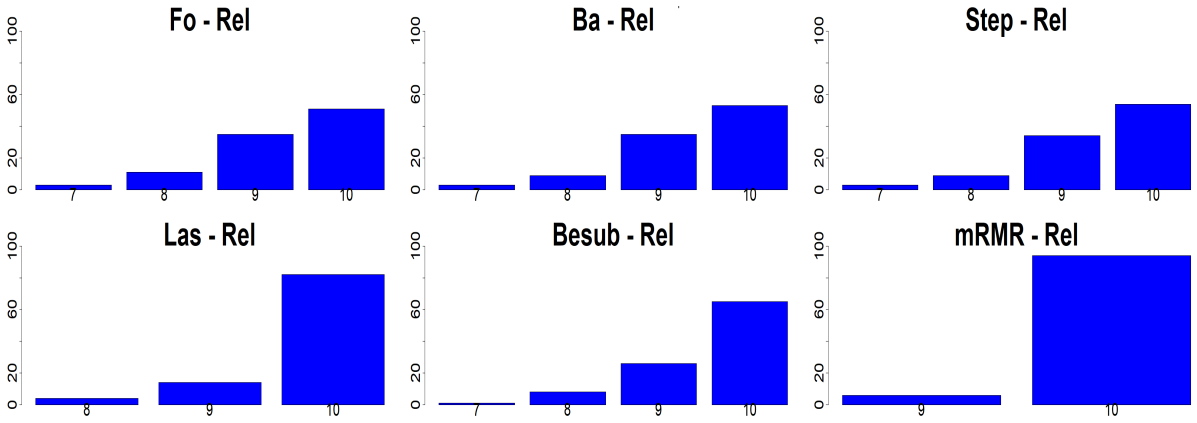

Figura 5.5: Número de veces que se seleccionan variables relevantes en la simulación B

En 5.5, se han extraído unos resultados muy similares a los del escenario A. En lo referido a las variables relevantes, los métodos Forward, Backward, Stepwise y mRMR han seleccionado unas pocas veces modelos con solo 7 relevantes, mientras que el LASSO y el Best-subset como mínimo 8, siendo destacable que la LASSO selecciona aproximadamente 80 modelos con las 10 relevantes y el mRMR con, de nuevo, aproximadamente 60 modelos con las mismas variables de este tipo.

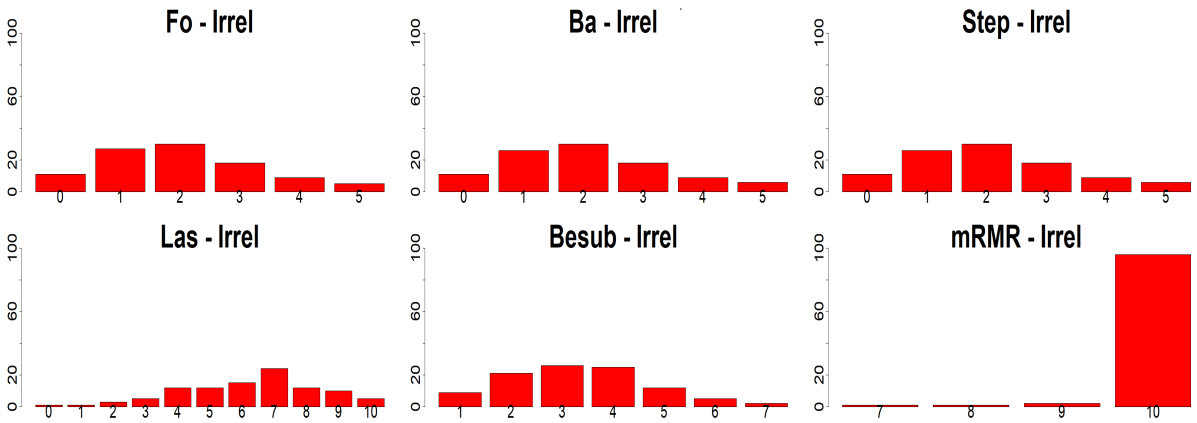

Figura 5.6: Número de veces que se seleccionan variables irrelevantes en la simulación B

En 5.6, con respecto a las variables irrelevantes, los métodos Forward, Backward y Stepwise son los que menos seleccionan, tomando modelos con como mucho 5 variables irrelevantes, mientras que, el Best-subset llega a seleccionar modelos con 6 y 7. De nuevo, el LASSO se disparan con incluso modelos con todas las irrelevantes. Finalmente, hay que destacar que el mRMR ha

seleccionado casi en todas las ocasiones todas las variables de este tipo.

5.3. Escenario C

En este Escenario C, se ha fijado un β_3 y un modelo de regresión lineal múltiple con $p = 20$ y $n = 100$ observaciones. Además, se ha provocado que las variables sean dependientes, tomando, las 15 primeras como relevantes.

$$\beta_3 = \texttt{(0.5,0.2,1,1.7,-0.2,-4,2,2.4,-1.8,-0.25,1,-1.3,0.2,-0.67,0.4,0,0,0,0,0)}$$

Tras la realización de 100 simulaciones, se ha obtenido como resultado los expuestos en la siguiente gráfica:

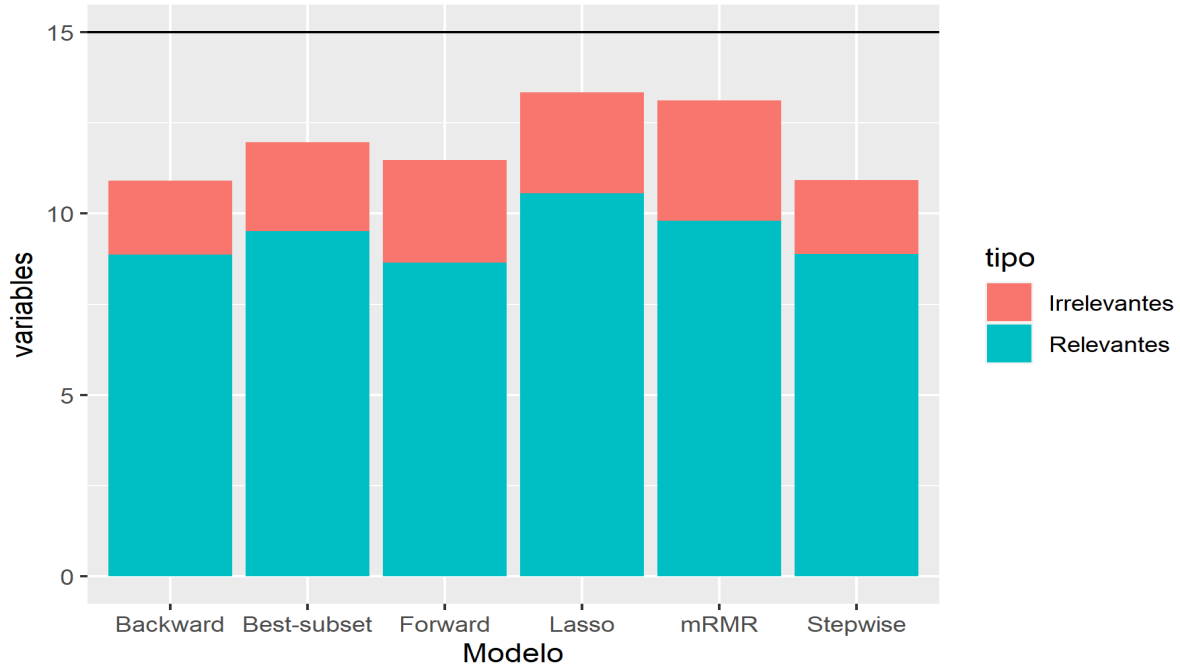

Figura 5.7: Promedios de selección de variables relevantes e irrelevantes en el escenario C

En 5.7, los métodos Forward, Backward y Stepwise son los que menos variables relevantes en promedio incluyen en sus modelos, mientras que, el LASSO, el Best-subset y el mRMR son los mejores en cuánto al indicador de calidad que se está buscando, destacando que es el LASSO el que, en promedio, más relevantes incluye en sus modelos.

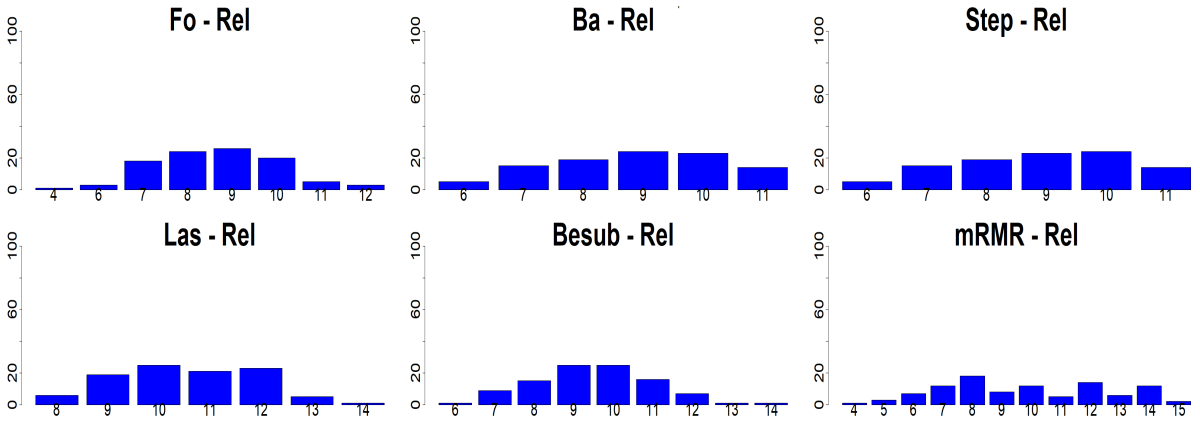

Figura 5.8: Número de veces que se seleccionan variables relevantes en la simulación D

En 5.8, con respecto a las relevantes, todos los modelos han seleccionado modelos con bastante variabilidad en el número de variables de este tipo. Se puede destacar, que el método mRMR ha tomado todas las relevantes en alguna ocasión, seguido de los métodos LASSO y Best subset que han llegado a incluir 14. En lo referido a los otros tres, solo el Forward, en ocasiones contadas, ha llegado a las 12 variables de este tipo.

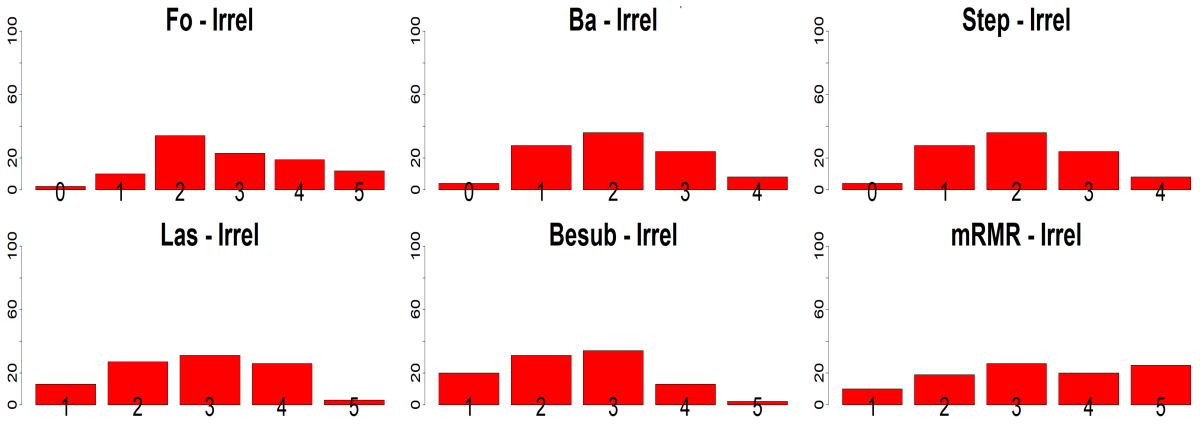

Figura 5.9: Número de veces que se seleccionan variables irrelevantes en la simulación C

En 5.9, con respecto a las variables irrelevantes, de nuevo, el mRMR es el método que más cantidad de estas variables selecciona, escogiendo en múltiples ocasiones todas las irrelevantes. Por otro lado, los demás métodos reparten más los casos, siendo el Forward y el Backward los únicos que no seleccionaron todas las irrelevantes en alguna simulación.

5.4. Escenario D

En este Escenario D, se ha fijado un β_4 y un modelo de regresión lineal múltiple con $p = 30$ variables y $n = 200$ observaciones. Claramente, se ha presentado un modelo con 15 variables relevantes y otras 15 irrelevantes.

$$\beta_4 = (0.5,\ 0.2,\ 1,\ 1.7,\ -0.2,\ -4,\ 2,\ 2.4,\ -1.8,\ -0.25,$$
$$1,\ -1.3,\ 0.2,\ -0.67,\ 0.4,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)$$

Tras la realización de 100 simulaciones, se ha obtenido como resultado los expuestos en la siguiente gráfica:

Figura 5.10: Promedios de selección de variables relevantes e irrelevantes en el escenario D

En 5.10 se observa un comportamiento muy similar al del escenario A, donde los métodos clásicos de regresión lineal reinan en lo que respecta al indicador de calidad, mientras que, el LASSO parte con una clara desventaja, ya que en promedio incluye casi todas las variables relevantes y peca de la inclusión de muchas irrelevantes. Cabe destacar, la pobreza del mRMR a la hora de incluir variables relevantes en su modelo, que incluso, en promedio, son las mismas que las irrelevantes.

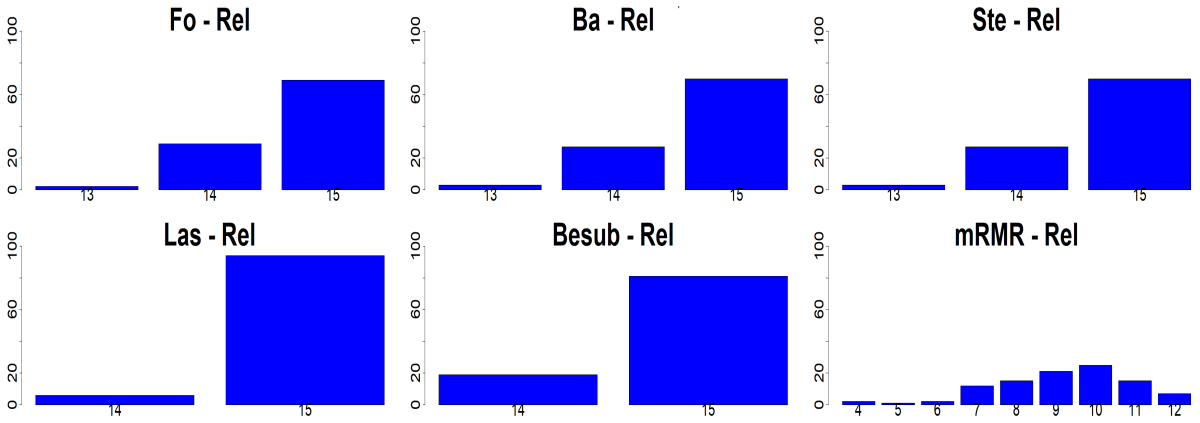

Figura 5.11: Número de veces que se seleccionan variables relevantes en la simulación D

En 5.11, en lo referido a las relevantes, el mRMR reparte más los casos llegando solo a incluir 13 de ellas, en pocas ocasiones. Sin embargo, los demás métodos demuestran una gran eficacia a la hora de incluir este tipo de variable, siendo el LASSO el mejor en este escenario a la hora de incluir relevantes, en casi todas las simulaciones selecciona las 15.

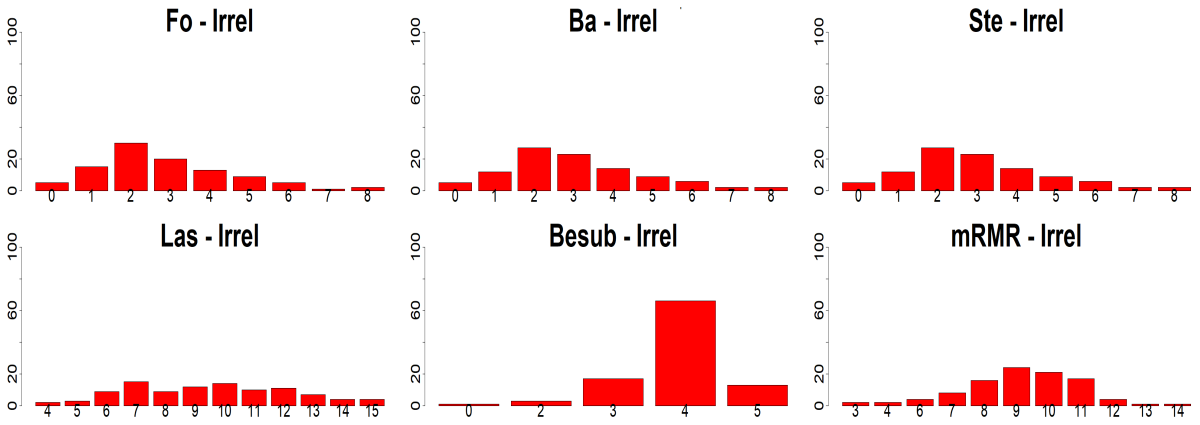

Figura 5.12: Número de veces que se seleccionan variables irrelevantes en la simulación D

En 5.12, con respecto a las irrelevantes, destacar que el método Best subset es, con diferencia, el que menos variables de este tipo incluye, llegando como máximo a la cantidad de 5 y seguido por el Backward, Forward y Stepwise que han obtenido unos números más consistentes que el mRMR y el LASSO, los cuales han llegado a incluir hasta 14 y 15, respectivamente, en alguna ocasión.

5.5. Escenario E

En este Escenario E se ha fijado un β_5 y un modelo de regresión lineal múltiple con $p = 30$ variables y $n = 200$ observaciones. Además, se ha incluido un problema de heterocedasticidad en los errores del modelo, para estudiar las respuestas de los diferentes modelos ante el mismo escenario pero con la aparición de esta problemática. Claramente, se ha presentado un modelo con 15 variables relevantes y otras 15 irrelevantes.

$$\beta_5 = \texttt{(0.5, 0.2, 1, 1.7, -0.2, -4, 2, 2.4, -1.8, -0.25,}$$
$$\texttt{1, -1.3, 0.2, -0.67, 0.4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)}$$

Tras la realización de 100 simulaciones, se ha obtenido como resultado los expuestos en la siguiente gráfica:

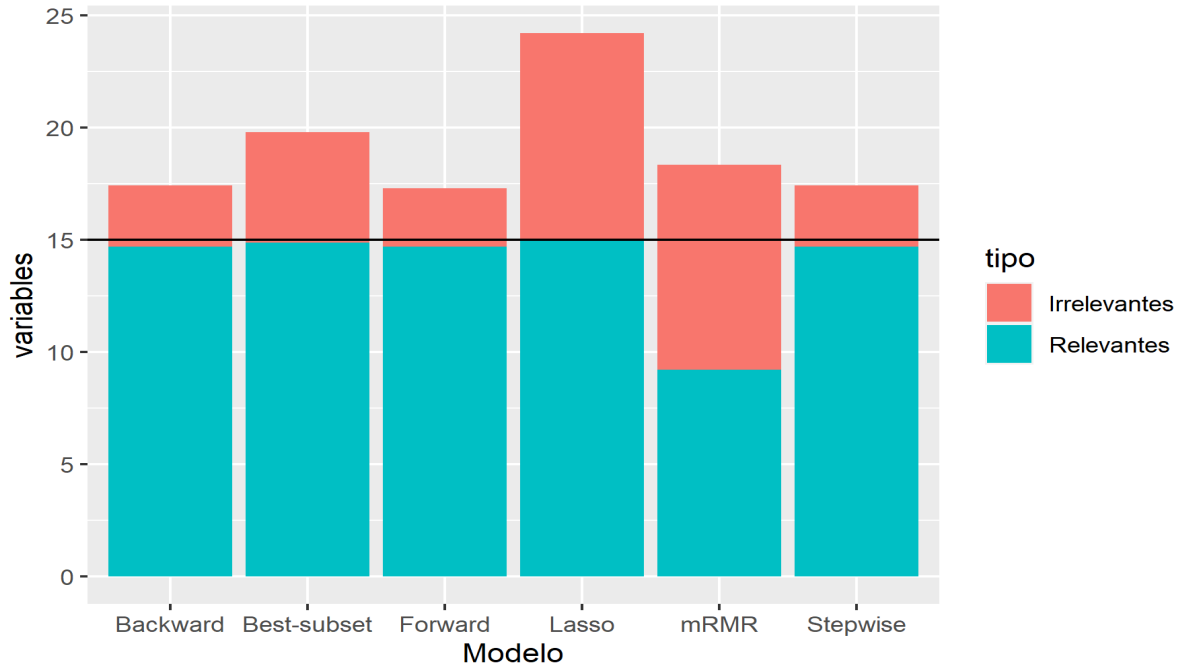

Figura 5.13: Promedios de selección de variables relevantes e irrelevantes en el escenario E

En 5.13 se obtienen unos resultados muy similares al escenario anterior, donde, de nuevo, los métodos clásicos de regresión lineal lideran en lo referido al indicar propuesto, sin embargo, el LASSO, en promedio, incluye casi todas las variables relevantes y peca de la inclusión de muchas irrelevantes. Cabe destacar, la pobreza del mRMR a la hora de incluir variables relevantes en su modelo, que incluso, en promedio, son las mismas que las irrelevantes.

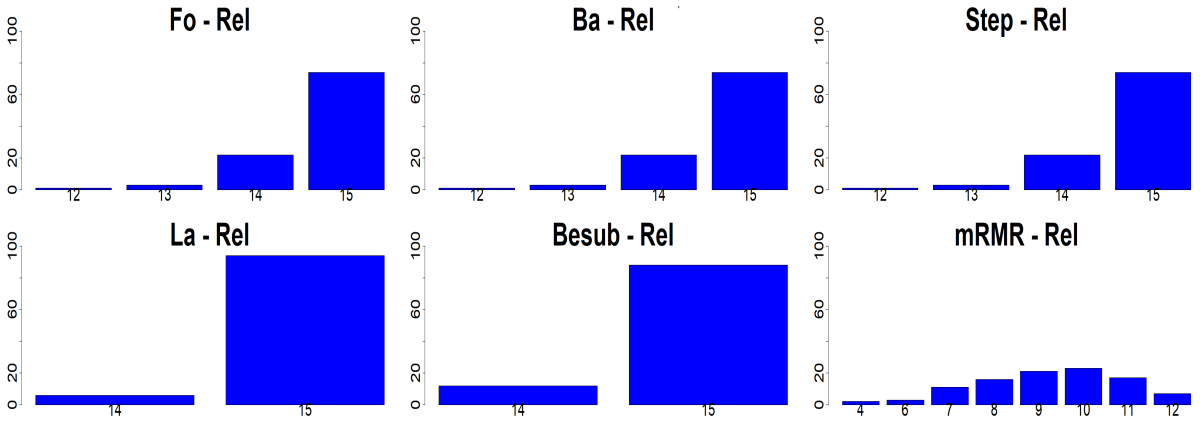

Figura 5.14: Número de veces que se seleccionan variables relevantes en la simulación E

En 5.14, con respecto a las variables relevantes, de nuevo, se asemeja mucho al escenario anterior, donde el mRMR reparte más los casos llegando solo a incluir 13 de ellas, en alguna ocasiones. Mientras, los demás métodos demuestran una gran eficacia a la hora de incluir este tipo de variable, volviendo, el LASSO, a repetir liderazgo a la hora de incluir relevantes, en casi todas las simulaciones selecciona todas.

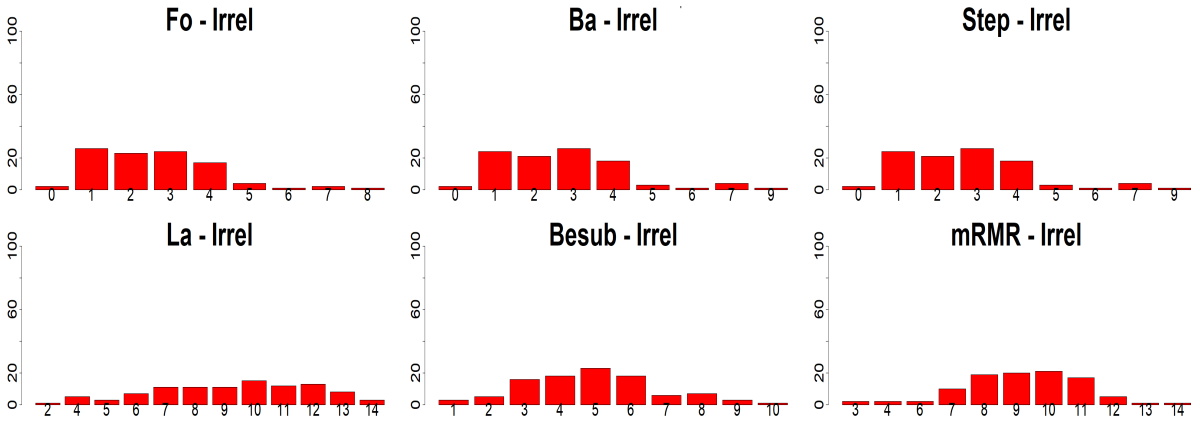

Figura 5.15: Número de veces que se seleccionan variables irrelevantes en la simulación E

En 5.15, con respecto a las variables irrelevantes, si se compara con el escenario anterior, los métodos Forward, Backward y Stepwise, aumentan los modelos de entre 2 y 4 irrelevantes reduciendo los de entre 5 y 8 a escasos casos. El LASSO Y el mRMR, prácticamente, mantienen los resultados, mientras que, el Best subset pasa de ser super eficaz en este tipo de variables, a repartir más las casuísticas asemejándose ligeramente a los dos anteriores.

5.6. Escenario F

En este Escenario F, se ha fijado un β_6 y un modelo de regresión lineal múltiple con $p = 30$ variables y $n = 200$ observaciones. Además, se ha provocado que todas las variables sean dependientes y tomando las 15 primeras como relevantes.

$$\beta_6 = (0.5,\ 0.2,\ 1,\ 1.7,\ -0.2,\ -4,\ 2,\ 2.4,\ -1.8,\ -0.25,$$
$$1,\ -1.3,\ 0.2,\ -0.67,\ 0.4,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)$$

Tras la realización de 100 simulaciones, se ha obtenido como resultado los expuestos en la siguiente gráfica:

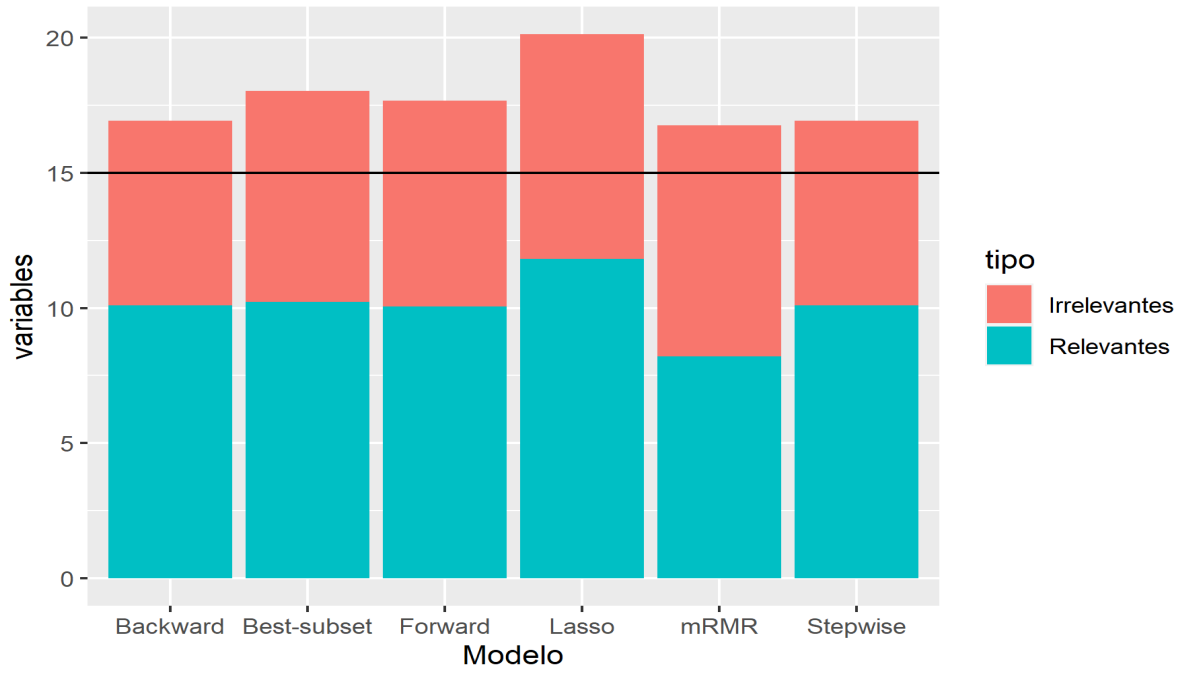

Figura 5.16: Promedios de selección de variables relevantes e irrelevantes en el escenario F

En 5.16, todos los métodos pierden protagonismo en comparación a los escenarios D y E, ya que disminuyen el promedio de variables relevantes que incluyen, quitando el mRMR que mantiene, aroximadamente, el mismo. Destacar que el LASSO es el que más promedio de relevantes incluye en sus modelos en comparación a los demás, pero quedándose más lejos que en otros escenarios, del promedio de relevantes idóneo. En este escenario, todos los métodos promedian un alto número de irrelevantes.

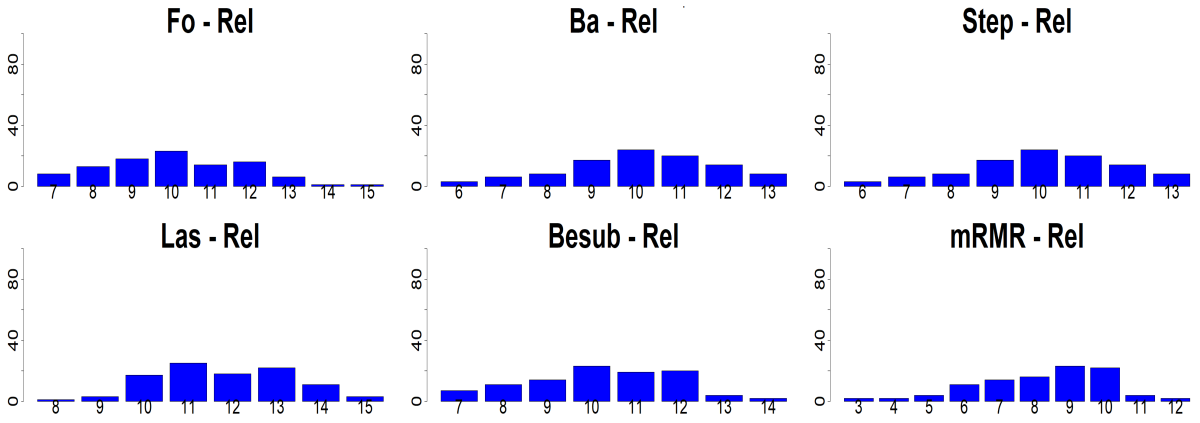

Figura 5.17: Número de veces que se seleccionan variables relevantes en la simulación F

En 5.17, con respecto a las variables relevantes, es de todos los escenarios en el que más se sigue un comportamiento similar entre todos los métodos. Destacar que el método LASSO seguido del Forward y del Best-subset son los que mas relevantes incluyen en sus modelos.

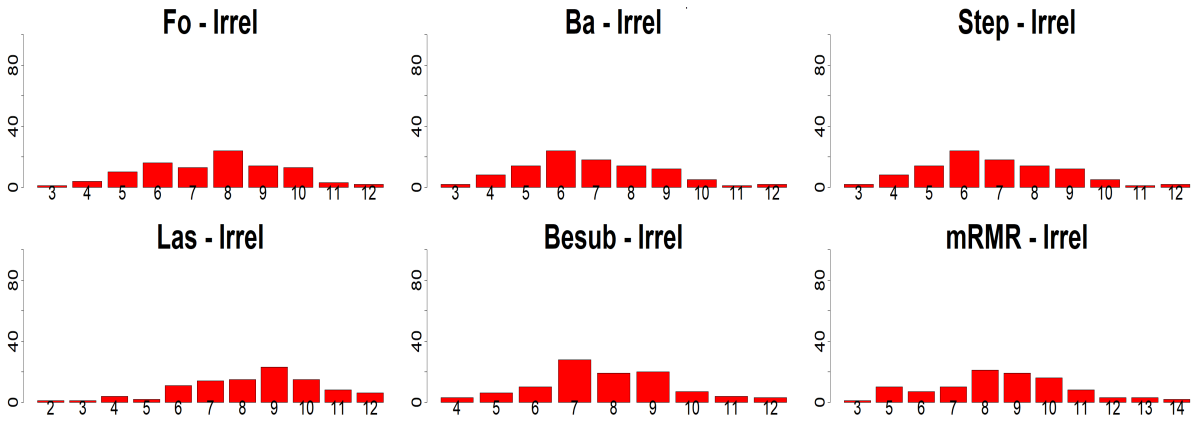

Figura 5.18: Número de veces que se seleccionan variables irrelevantes en la simulación F

En 5.18, con respecto a las variables irrelevantes, también, es de todos los escenarios en el que más se sigue un comportamiento similar entre todos los métodos. Además, la dependencia ha provocado que se hayan incluido, por lo general, muchas más variables irrelevantes en los modelos generados.

5.7. Conclusiones

Finalmente, para resumir el estudio, se pueden destacar una serie de conclusiones obtenidas tras la realización de las simulaciones en los 6 escenarios.

5.7.1. Conclusiones por métodos de selección de variables

Métodos Clásicos (Forward, Backward, Stepwise, Best-subset): Los métodos clásicos demostraron consistencia en varios escenarios al seleccionar variables relevantes y excluir las irrelevantes de manera efectiva. Forward y Stepwise destacaron por su capacidad para mantener un equilibrio entre inclusión de variables relevantes y exclusión de irrelevantes, es decir, para ser eficaz frente al indicador de calidad, siendo especialmente adecuados cuando se busca interpretabilidad en el modelo.

LASSO: LASSO mostró eficacia en la inclusión de variables relevantes, especialmente en escenarios complejos donde otras técnicas podrían enfrentar dificultades. Sin embargo, tendió a incluir más variables irrelevantes en comparación con los métodos clásicos, lo que puede comprometer la interpretabilidad del modelo.

mRMR: mRMR exhibió una capacidad variable en la selección de variables relevantes, mostrando resultados competitivos en algunos escenarios pero también incluyendo frecuentemente variables irrelevantes. Es particularmente útil cuando se considera la relevancia de las variables dentro de un contexto de dependencia y redundancia mínima.

5.7.2. Comparativa por escenario

Escenarios A y B: Los métodos clásicos como Forward y Stepwise fueron superiores al seleccionar un alto número de variables relevantes mientras minimizaban la inclusión de irrelevantes. Estos métodos mostraron una capacidad robusta para manejar modelos con datos heterogéneos y mantener la calidad del modelo.

Escenario C: El Forward y Stepwise continuaron destacando al mantener una selección consistente de variables relevantes. LASSO y mRMR mostraron una inclusión similar de variables relevantes pero con una tendencia mayor a incluir irrelevantes, lo que sugiere una adaptabilidad variable según la complejidad del modelo.

Escenarios D y E y F: Los métodos clásicos mantuvieron su eficacia al seleccionar variables relevantes y manejar la complejidad del modelo de manera efectiva. LASSO y mRMR enfrentaron desafíos adicionales debido a la inclusión inconsistente de variables relevantes e irrelevantes, lo

que sugiere que su uso debe ser cuidadosamente considerado en contextos complejos.

5.7.3. Conclusión general

Finalmente, se extrae una conclusión general basada en los escenarios simulados y los métodos evaluados:

- **Preferencia por Métodos clásicos de regresión lineal:** Para la mayoría de los escenarios, los métodos Forward y Stepwise son recomendados debido a su capacidad para seleccionar variables relevantes de manera efectiva y mantener la interpretabilidad del modelo.
- **Consideraciones sobre regularización:** LASSO y mRMR pueden ser útiles en situaciones donde se necesite manejar grandes conjuntos de datos o se priorice la inclusión de un alto número de variables relevantes. Sin embargo, se debe tener cuidado con la inclusión de variables irrelevantes, lo que puede afectar la interpretabilidad y la precisión del modelo.

Capítulo 6

Ejemplo ilustrativo

En este capítulo, se presenta un análisis de selección de variables utilizando datos reales. Se aplicarán los métodos expuestos anteriormente con el fin de reforzar los resultados obtenidos de los estudios de simulación previos.

6.1. Descripción de los Datos

Los datos utilizados en este análisis provienen de:
`https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html`
En este conjunto de datos se han tomado a 442 pacientes de diabetes y se obtuvieron 10 medidas basales: `AGE` (edad),`SEX` (sexo), `BMI` (índice de masa corporal), `BP` (promedio de la presión arterial), y 6 medidas de suero sanguíneo: `TC` (hormona triacontanol), `LDL` (colesterol-LDL), `HDL` (colesterol-HDL), `TCH` (hormona tiroidea), `LTG` (molécula lamotrigina) y `GLU` (glucosa). Todas ellas con el objetivo de medir el progreso de la enfermedad después de un año, por lo que la variable respuesta será un indice que mide dicho progreso. Cabe destacar que la variable `SEX` se ha tratado como continua ya que ha sido estandarizada para tener media 0 y longitud al cuadrado igual a 1 ($\sum x^2 = 1$), de igual modo que las otras 9 variables explicativas resultantes. Tras la realización de un ajuste de regresión lineal, se encuentran dos problemáticas. Por un lado, hay 5 variables dependientes, ya que tienen un VIF >5. Por otro lado, tras realizar el bp test en R, se encuentran evidencias de heterocedasticidad en el modelo, tras realizar el test de Brendon-Praguen. En esta 6.1 quedan resumidos los coeficientes del ajuste así como los valores del VIF, de manera que, además de mostrar el problema de dependencia de algunas variables, se conforme una idea, a priori, de cuales pueden ser las más importantes en este modelo de regresión lineal.

Variable	`VIF`	**Coeficiente**
(Intercept)		152.13348
AGE	1.217307	-10.01220
SEX	1.278073	-239.81909
BMI	1.509446	519.83979
BP	1.459429	324.39043
TC	59.203786	-792.18416
LDL	39.194379	476.74584
HDL	15.402352	101.04457
TCH	8.890986	177.06418
LTG	10.076222	751.27932
GLU	1.484623	67.62539

Cuadro 6.1: Coeficientes de la regresión lineal y `VIF`

6.2. Aplicación de los métodos

Se ha decidido realizar una selección de variables usando los métodos expuestos en el estudio piloto de la sección anterior, con el fin de obtener un modelo simplificado para explicar la variable respuesta. En la 6.2 se recogen los resultados de manera que 1 significa que el método ha incluido dicha variable y 0 cuando ha sido excluida.

Variable	Forward	Backward	Stepwise	LASSO	BestSubset	mRMR
AGE	0	0	0	0	0	0
SEX	1	1	1	1	1	1
BMI	1	1	1	1	1	1
BP	1	1	1	1	1	1
TC	1	1	1	1	1	1
LDL	1	1	1	0	1	1
HDL	0	0	0	1	0	1
TCH	0	0	0	0	1	0
LTG	1	1	1	1	1	1
GLU	0	0	0	1	1	1

Cuadro 6.2: Resultados de la selección de variables

6.3. Conclusiones

Se concluye que para medir el progreso de la diabetes en pacientes durante un año, las variables más importantes según los resultados obtenidos en 6.2 son el sexo `SEX`, el índice de masa corporal (`IMC`), la presión arterial media `BP`, la hormona triacontanol `TC` y la molécula lamotrigina (`LTG`). Esto se puede ver apoyado en que el sexo puede afectar a la susceptibilidad y al manejo de la diabetes debido a diferencias hormonales y metabólicas entre hombres y mujeres, el `IMC` refleja la cantidad de grasa corporal y está relacionado con la resistencia a la insulina, la presión arterial alta puede ser tanto causa como consecuencia de la diabetes, el `TC` puede estar implicada en la regulación del metabolismo y la respuesta a la insulina, y el `LTG` que aunque no es comúnmente asociada con la diabetes, su inclusión en los modelos sugiere que podría haber una interacción entre este compuesto y el metabolismo de la glucosa, lo cual podría influir en el progreso de la enfermedad en ciertos contextos. Por otro lado, variables como la edad `AGE` y los niveles de glucosa `GLU` no mostraron ser determinantes en este contexto específico y con los métodos utilizados. Este enfoque puede guiar estrategias clínicas para monitorear y gestionar la diabetes, destacando la importancia de factores específicos en el manejo de esta enfermedad crónica.

Bibliografía

[1] Akaike, H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 19(6), 716–723.

[2] Berrendero, J. R., Cuevas, A., & Torrecilla, J. L. (2016). *The mRMR variable selection method: A comparative study for functional data.* Journal of Statistical Computation and Simulation, 86(5), 891–907.

[3] Ding, C., & Peng, H. (2005). *Minimum redundancy feature selection from microarray gene expression data.* Journal of Bioinformatics and Computational Biology, 3(02), 185–205.

[4] Vilar Fernández, J. M. (2006). *Modelos estadísticos aplicados.* Universidade da Coruña, Servicio de Publicacións.

[5] Freijeiro González, L. *Modelos de predicción y clasificación con alta dimensión en el número de covariables.*

[6] Hall, M. A. (1999). *Correlation-based feature selection for machine learning.*

[7] Peng, H., Long, F., & Ding, C. (2005). *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226–1238.

[8] Hastie, T., Tibshirani, R., Friedman, J., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction, volume 2.* Springer.

[9] Mac Kay, D. J. C. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

[10] Kaufman, R. L. (2013). *Heteroskedasticity in regression: Detection and correction.* Sage Publications.

[11] García Laencina, P. J., & Sancho Gómez, J. L. (2010). *Estimación de densidad de probabilidad mediante ventanas de parzen.* In Jornadas de introducción a la investigación de la UPCT, (3) (pp. 68–70).

[12] Schwarz, G. (1978). *Estimating the dimension of a model.* The annals of statistics, 461–464.

[13] Székely, G. J., & Rizzo, M. L. (2017). *The energy of data.* Annual Review of Statistics and Its Application, 4, 447–479.

[14] Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). *Measuring and testing dependence by correlation of distances.* The Annals of Statistics, 35(6), 2769–2794.

[15] Tibshirani, R. (1996). *Regression shrinkage and selection via the LASSO.* Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[16] Friedman, J., Hastie, T., & Tibshirani, R. (2017). *The Elements of Statistical Learning.* Springer.

[17] Furnival, G. M., & Wilson, R. W. (1974). *Regressions by leaps and bounds.* Technometrics, 16(4), 499–511.

[18] Yuan, M., & Lin, Y. (2006). *Model selection and estimation in regression with grouped variables.* J. R. Stat. Soc. Ser. B, 68(1), 49–67.

[19] Bozdogan, H. (1987). *Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions.* Psychometrika, 52(3), 345–379.

[20] Bakin, S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems.*

[21] Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). *Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates.* International Statistical Review, 90, 118–145.