Estimación e inferencia de parámetros en un modelo de regresión normal múltiple multivariado mediante el *Bootstrap* y el *Jackknife*

Karen Manuela Torres García

Universidad El Bosque
Facultad de Ciencias
Departamento de Matemáticas
Programa de Estadística
Bogotá D.C, Colombia
2023

Estimación e inferencia de parámetros en un modelo de regresión normal múltiple multivariado mediante el *Bootstrap* y el *Jackknife*

Karen Manuela Torres García

Tesis como requisito parcial para optar al título de:
Estadístico

Director:
PhD. Mario José Pacheco López

Universidad El Bosque
Facultad de Ciencias
Departamento de Matemáticas
Programa de Estadística
Bogotá D.C, Colombia
2023

Agradecimientos

A todos los profesores que acompañaron mi proceso de aprendizaje durante estos años

A mi director Mario José Pacheco por su dedicación, acompañamiento y recomendaciones durante el desarrollo de este proyecto

A mi padre por su sacrifico y amor diario

A mi madre y hermana por su amor y apoyo incondicional, su paciencia y su comprensión durante todo el proceso

A mi novio, por su constante apoyo y motivación

Resumen

En este proyecto se describe el procedimiento *Bootstrap* y *Jackknife* para los modelos lineales múltiples multivariados y se crea una función que estima los parámetros tanto por *Bootstrap* como por *Jackknife*. Además, se construyen escenarios de simulación para evaluar el algortimo cuando los datos siguen una distribución normal multivariada. Y por último, se realiza una aplicación de la función donde se comparan las estimaciones obtenidas por mínimos cuadrados ordinales y las dos técnicas de remuestreo.

Palabras clave: Modelos lineales, remuestreo, *Bootstrap*, *Jackknife*, parámetros, estimación.

Abstract

This Project describes the process of Bootstrap and Jackknife for Multivariate Multiple Regression and creates a function that estimates the parameters of Bootstrap and Jackknife techniques. Further, simulate scenarios are built to evaluate the algorithm when the data are distributed normal multivariate. Finally, an application of the function is performed where the estimates obtained by ordinary least squares and both resampling techniques are compared.

Keywords: Linear Models, resampling, Bootstrap, Jackknife, parameters, estimation.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

Los modelos de regresión lineales múltiples multivariados (MLMM) son una extensión de los modelos de regresión lineales múltiples (MLM), su mayor diferencia se encuentra en la respuesta que corresponde a más de una variable. Los MLM se expresan como $y = X\beta + \epsilon$, donde y es un vector que contiene la variable de respuesta, y en el caso de los MLMM Y es una matriz que contiene desde y_1, y_2, hasta y_p variables respuestas.

Estos modelos pueden ser aplicados en diferentes áreas tales como: negocios, economía e investigación médica. Por ejemplo, Monge (1977) examina el modelo de regresión múltiple multivariado y explora su aplicabilidad en otros ámbitos tales como la comunicación, Eyvazian et al. (2011) utiliza la regresión múltiple multivariada para caracterizar la calidad de un proceso mediante dos o más perfiles, Clack (2017) realiza estimaciones de irradiación mediante un esquema de regresión lineal múltiple multivariado en 10 ubicaciones. Así mismo, autores como Quick y James (2013), aplican un modelo de regresión múltiple multivariado para identificar cuales atletas pueden competir en un campeonato nacional de levantamiento de pesas.

Tanto los MLM como los MLMM poseen supuestos sobre la distribución de los datos, estos se deben cumplir para obtener estimaciones confiables, sin embargo, a veces es difícil cumplir estos supuestos distribucionales y es necesario recurrir a métodos o modelos diferentes. Por lo anterior, resulta pertinente desarrollar una alternativa eficiente y eficaz para la estimación de los parámetros de los MLMM. En este caso, se hará uso de técnicas de remuestreo como el *Bootstrap* y el *Jackknife*. Estas metodologías son de fácil aplicabilidad debido al avance tecnológico que ha habido en las últimas décadas. Cabe resaltar que desde el año 1982 Efron introduce en su libro " *The Jackknife, the Bootstrap and other resampling plan*"la aplicabilidad de los métodos en estadística y Davison y Hinkley (1997) presentan un acercamiento más detallado al *Bootstrap* con el cual es posible estimar los parámetros de un MLM, intervalos de confianza y realizar pruebas de hipótesis.

El *Jackknife*, propuesto por Quenouille en el año 1949 es una técnica de remuestreo que permite la reducción de sesgo en la estimación de parámetros (Bradley, 1982). Esta técnica tiene como idea básica estimar el parámetro de interés n veces, en la cual en cada iteración se elimina la *i-ésima* observación para realizar la estimación, donde n es el tamaño de muestra. En el año 1958 Tukey propuso el uso general de esta técnica para reducir sesgos y la estimación de intervalos de confianza cuando los procedimientos estadísticos estándar podían no existir o son difíciles de aplicar (Miller, 1964).

Por otro lado, el *Bootstrap* es un acercamiento no paramétrico para la inferencia estadística donde el objetivo es obtener r muestras *Bootstrap*, usualmente $r = 500$, partiendo de la muestra original. Dado que el *Bootstrap* no requiere suposiciones distribucionales puede proporcionar inferencias más precisas cuando los datos no se comportan bien o cuando el tamaño de la muestra es pequeño (Fox, 2016).

Esa así, como en este trabajo se describe la metodología *Bootstrap* y *Jackkni-*

fe aplicada a los modelos lineales múltiples multivariados. Además, se construyó una función para estimar los parámetros por medio de estás técnicas, con la finalidad de poder realizar estimaciones de los parámetros del modelo cuando los supuestos distribucionales no se cumplen. Se realizó una aplicación de la función y se compararon las estimaciones obtenidas tanto por mínimos cuadrados ordinales como por *Jakknife* y *Bootstrap*.

El documento se organiza en 9 capítulos que se describen brevemente a continuación. En este primer capítulo se realiza la introducción del documento dando una corta explicación de la temática a tratar. En el segundo capítulo se encuentra el planteamiento del problema. En el tercer capítulo la justificación. En el cuarto capítulo los objetivos generales y específicos del proyecto. En el quinto capítulo se encontrará el marco teórico, con los fundamentos de los modelos MLMM, el *Bootstrap* para estimación e inferencia y el *Jackknife* para estimación y reducción de sesgo. En el sexto capítulo la metodología del trabajo. En el séptimo los resultados obtenidos en la investigación. Por último, en los capítulos ocho y nueve se encontrarán la discusión y las conclusiones.

2. Planteamiento del problema

En estadística para lograr estimar y realizar inferencia de parámetros se requieren supuestos sobre la muestra. Como en los modelos MLMM, donde las variables de respuesta en la matriz ($\mathbf{Y}$) deben seguir una distirbución normal multivariada y en consecuencia los errores del modelo siguen una distribución normal multivariada con vector de medias de 0 y matriz de varianzas y covarianzas Σ. Sin embargo, estos supuestos distribucionales en ocasiones no se cumplen y no es posible aplicar los métodos ya propuestos para obtener estimaciones confiables. Además, cuando se tienen muestras pequeñas también es difícil cumplir los supuestos distribucionales y es necesario recurrir a métodos no parámetricos para lograr estimar los parámetros de forma adecuada.

3. Justificación

Los MLMM necesitan ciertos supuestos para poder ser usados, se debe cumplir normalidad multivariada en los errores y en consecuencia, las columnas de la matriz de parámetros $\boldsymbol{\beta}$ también sigue una distribución normal. Además, cuando se ajusta un modelo lineal múltiple multivariado, se tienen otros dos supuestos necesarios para un buen ajuste: independencia entre las observaciones y la consistencia en la matriz de covarianzas. Aunque, la realidad presenta muchos escenarios donde es muy difícil o casi imposible, por la misma naturaleza de los datos, obtener la distribución requerida para los estimadores, tanto así que su distribución puede llegar a ser desconocida. Los anteriores supuestos son requeridos para poder obtener estimaciones fiables de los parámetros y, cuando estos supuestos no se cumplen lo primero que se piensa es en buscar otro modelo que no requiera estos supuestos distirbucionales. Los avances en al ámbito computacional han permitido que la aplicabilidad de técnicas de remuestreo sean mucho más sencillas, como el *Bootstrap* o el *Jackknife*, que brindan una solución alterna cuando hay carencia en los supuestos distribucionales de los datos o incluso cuando la muestra es considerada pequeña. De esta forma se pueden aplicar métodos clásicos de la estadística con algunas modificaciones que permiten brindar confiabilidad en las estimaciones y en la toma de decisiones.

4. Objetivos

4.1. Objetivo general

Estimar y realizar las inferencias sobre los parámetros de un modelo de regresión lineal múltiple multivariado por medio de *Bootstrap* y *Jackknife*.

4.2. Objetivos específicos

1. Describir el procedimiento *Bootstrap* y *Jackknife* para los modelos lineales múltiples multivariados.
2. Crear un algoritmo que permita estimar los parámetros del modelo lineal múltiple multivariado por medio de *Bootstrap* y *Jackknife*.
3. Estimar la matriz de covarianzas por *Bootstrap* y *Jackknife*, estimar intervalos de confianza para los parámetros del modelo y realizar prueba de hipótesis para la significancia de los coeficientes del modelo.
4. Realizar un pequeño estudio de simulación para comparar las estimaciones que se obtienen por MCO y las dos técnicas de remuestreo.
5. Realizar una aplicación del algoritmo y comparar las estimaciones obtenidas por *Bootstrap* y *Jackknife*.

5. Marco Teórico

5.1. Antecedentes

Los modelos de regresión lineales múltiples multivariados se suelen confundir con los modelos de regresión lineales múltiples, debido a que en esencia son similares y la estimación de sus parámetros es paralela. Sin embargo, en la literatura se logran encontrar algunas aplicaciones de estos modelos, como se muestra a continuación

Monge (1977) en su trabajo "*Multivariate Multiple Regression in Communication Research*" examina el modelo de regresión múltiple multivariado y explora su aplicabilidad en la investigación en comunicación, donde se analiza la partición de una matriz de datos, como un dispositivo heurístico para distinguir de modelos de regresión alternativos. Además, se evalúan los supuestos del modelo y la interpretación de los coeficientes para dar respuesta a las leyes de la comunicación humana.

Eyvazian et al. (2011) en el estudio "*Phase II Monitoring of Multivariate Multiple Linear Regression Profiles*" utiliza la regresión lineal múltiple multivariada para caracterizar la calidad de un proceso mediante dos o más perfiles, investigan los problemas relacionados con una estructura lineal múltiple multivariada y se proponen 4 métodos para monitorear perfiles de regresión lineal múltiple multivariada. Los problemas que se suelen presentar en el ajuste de los modelos están relacionados con la independencia entre las observaciones y la falta de normalidad en los errores. Para el monitoreo de los ajustes que se realizan mediante estos modelos de regresión se encontró, por medio de simulación, que el método de esquema combinado de gráficos de control MEWMA y chi-cuadrado es el que mayor rendimiento presenta.

Nkurunziza y Ahmed (2011) en su artículo "*Estimation strategies for the regression coefficient parameter matrix in multivariate multiple regression*" examinaron el rendimiento relativo de dos modelos de regresión múltiple multivariados: el primero incluyendo todos los predictores y el segundo restringiendo los coeficientes a un subespacio lineal candidato basado en información previa, combinando dos técnicas de estimación. Se desarrolló una teoría para muestras grandes para los estimadores que incluye la derivación del sesgo y el riesgo de distribución asintótico de los estimadores. Por último, se llevaron a cabo simulaciones de Monte Carlo para evaluar el desmpeño de los estimadores sugeridos con los estimadores clásicos.

Jeong et al. (2012) en el estudio "*Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator*" proporciona un procedimiento de reducción de escala estadístico híbrido multisitio, para esto se emplea la regresión lineal múltiple multivariada, el modelo de cadenas de Márkov y la técnica de mapeo de distribución para reproducir la variabilidad temporal y la dependencia espacial multisitio de las series de precipitación. Un problema de estas series de precipitación generadas por el modelo, es que generalmente tienen diferentes propiedades estadísticas que las series de precipitación observadas, esto da como resultado que la matriz residual no se distribuya normalmente y produzca

sesgo. Para superar este problema se adaptó la técnica de mapeo de distribución de probabilidad.

Quick y James (2013) en su documento "*Multivariate Multiple Regression with Applications to Powerlifting Data*" desarrolla la teoría del modelo lineal múltiple multivariado y sus propiedades. Además, se realiza una aplicación relacionada con la selección de atletas que pueden competir en un campeonato nacional de levantamiento de pesas, sin embargo, se tuvieron dos problemas con los datos: 1) con la independencia entre observaciones y 2) con la linealidad de las respuestas respecto a la variable edad.

Amiri et al. (2014) en el estudio "*Diagnosis Aids in Multivariate Multiple Linear regression Profiles Monitoring*" identifican los perfiles y parámetros que han cambiado durante el proceso en la estructura de perfiles de regresión lineales multivariantes los cuales permiten caraterizar procesos y observar la calidad de un producto. Para esto se buscaron aquellos perfiles o parámetros que estuvieran fuera de control por medio de simulaciones Monte carlo y un estudio de casos reales en términos de porcentaje de precisión.

Clack (2017) en el estudio "*Modeling Solar Irradiance and Solar PV Power Output to Create a Resource Assessment Using Linear Multiple Multivariate Regression*" realizan estimaciones de irradiación mediante un esquema de regresión lineal múltiple multivariado en 10 ubicaciones, para después ingresarlas en un algoritmo de modelado de energía solar fotovoltaica para calcular las estimaciones de estas energías. En el estudio se asume que los errores son independientes entre las especies de irradiancia, cuando estas realmente son dependientes, sin embargo, asumir que no los son no cambiaron los resultados significativamente.

Por otra parte, el *Bootstrap* y el *Jackknife* son técnicas de remuestreo para la estimación de parámetros y de la varianza de los estimadores de dichos parámetros. A continuación se presenta literatura relacionada con el uso y aplicación de estas técnicas de remuestreo en modelos de regresión lineal múltiple y en otros aspectos de la estadística como en las pruebas de hipótesis.

Bradley (1982) en la sección 5.7 del libro "*The Jackknife, the Bootstrap and Other Resampling Plans*" habla sobre la utilidad de los métodos de remuestreo para situaciones donde se tienen muchas muestras y una variedad de estructuras de datos más complicadas, como en los modelos de regresión lineales múltiples, donde es posible realizar la estimación de los parámetros por medio de *Bootstrap*.

Fox (2016) en el capítulo 21 del libro "*Applied Regressions Analysis Generalized Linear Models*" da una introducción al *Bootstrap* y menciona como esta técnica puede ser útil en estadística. Explica como se usa para la estimación de los parámetros de un modelo de regresión lineal y para prueba de hipótesis.

Fox y Weisberg (2012) en su apéndice titulado "*Bootstrapping Regression Models in R*" desarrollan con mayor detalle el tema de los modelos de regresión Bootstrap, explicando el funcionamiento del Bootstrap y la función Boot() del paquete car en R, la cual permite estimar los parámetros de un modelo de regresión, no solamente lineal múltiple, sino que gracias a modificaciones recientes también puede ser utilizada en modelos lineales generalizados.

Se puede evidenciar que los modelos lineales multivariados pueden ser útiles en varios campos. Si bien, hay técnicas que permiten superar ciertas complicaciones en el proceso de ajuste, aún no hay una aproximación que permita realizar las estimaciones por técnicas de remuestreo. Existe un acercamiento a estas estimaciones para modelos de regresión lineal y modelos lineales generalizados y sería ideal extenderla a los modelos lineales múltiples multivariados.

5.2. Modelo lineal múltiple multivariado

Se describe a continuación las características principales del modelo lineal múltiple multivariado.

5.2.1. El modelo

De acuerdo con Rencher (1998) el modelo lineal múltiple multivariado es una extensión del modelo de regresión lineal múltiple, en el caso multivariado se miden varías variables dependientes $y_1, y_2, \ldots, y_p$ para un conjunto de variables independientes $x_1, x_2, \ldots, x_q$. Cada una de las $y_1, y_{2\ldots}, y_p$ deben ser predichas por todas las variables explicativas $x_1, x_2, \ldots, x_q$. Si se tienen n observaciones, la matriz que representa los valores de cada una de las y's está dada por

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{bmatrix} \tag{5.1}$$

Las filas de la matriz $\mathbf{Y}$ contienen las observaciones de las p variables en cada uno de los sujetos y cada columna representa las p variables medidas en n sujetos. Análogamente al modelo lineal, la matriz correspondiente a las variables predictoras se expresa como

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nq} \end{bmatrix} = \begin{bmatrix} 1' \\ x_1' \\ \vdots \\ x_n' \end{bmatrix} \tag{5.2}$$

Se asume que la matriz $\mathbf{X}$ es fija. Debido a que las p variables de la matriz $\mathbf{Y}$ dependerán de las x's de forma distinta, será necesario estimar distintos coeficientes β para cada una de las columnas de la matriz. De esta forma el modelo múltiple multivariado se expresa como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.3}$$

donde $\mathbf{Y}$ tiene dimensiones de $n \times p$, $\mathbf{X}$ de $n \times (q+1)$, $\boldsymbol{\beta}$ de $(q+1) \times p$ y $\boldsymbol{\varepsilon}$ es el término que hace referencia al error del modelo, en este caso una matriz con dimesiones $n \times p$.

5.2.2. Supuestos del modelo

Los supuestos necesarios para obtener estimaciones confiables en el modelo son:

1. $E(\boldsymbol{\varepsilon}) = \mathbf{O}$ o $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$
2. $cov(\mathbf{y}_i) = \Sigma$ para toda $i = 1, 2, \ldots, n$, donde $\mathbf{y}_i'$ es la *i-ésima* fila de $\mathbf{Y}$
3. $cov(\mathbf{y}_i \mathbf{y}_j) = \mathbf{O}$ para toda $i \neq j$

5.2.3. Estimación de los parámetros del modelo múltiple multivariado por mínimos cuadrados

Para los parámetros en la matriz $\boldsymbol{\beta}$ se buscan estimadores que minimicen la suma de cuadrados de los valores observados de $\mathbf{Y}$ y de sus valores predichos

$$\begin{aligned}\mathbf{SCE} &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}} - (\mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{Y} + (\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{X}\widehat{\boldsymbol{\beta}})\end{aligned}$$

Ahora derivando con respecto a $\widehat{\boldsymbol{\beta}}$

$$\frac{\partial \hat{\varepsilon}^2}{\partial \widehat{\boldsymbol{\beta}}} = -\mathbf{Y}'\mathbf{X} - \mathbf{Y}'\mathbf{X} + \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{X}$$

E igualando a 0 para obtener la expresión para $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{5.4}$$

Como se observa la estimación de los parámetros del modelo (5.4) es análoga a la estimación de un modelo lineal múltiple. Cada columna de $\widehat{\boldsymbol{\beta}}$ corresponde a cada una de las estimaciones de las columnas de $\mathbf{Y}$, es decir, cada una de las *y's* es predicha de forma diferente por $\mathbf{X}$. Además, mediante **SCE** es posible obtener una estimación de $cov(y_i) = \Sigma$ dada por

$$\mathbf{S}_e = \frac{\mathbf{E}}{n-q-1} = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n-q-1} \tag{5.5}$$

Con el denominador $n - q - 1$, $\mathbf{S}_e$ es un estimador insesgado de Σ, debido a que $E(\mathbf{S}_e) = \Sigma$.

5.2.4. Propiedades de $\widehat{\boldsymbol{\beta}}$ y $\hat{\Sigma}$

Cada fila en el modelo (5.3) se puede expresar como

$$y_i' = x_i'\boldsymbol{\beta} + \varepsilon_i', \;\; i = 1, 2, \ldots, n \tag{5.6}$$

o en su forma transpuesta

$$y_i = \boldsymbol{\beta}' x_i + \varepsilon_i \tag{5.7}$$

A los tres supuestos en 5.2.2 se añade normalidad multivariada de y_i, es decir, y_i se distribuye $N_p(\boldsymbol{\beta}' x_i, \Sigma)$, lo que es equivalente a decir que los errores ε_i se distribuyen $N_p(\mathbf{0}, \Sigma)$. Si este supuesto distirbucional se cumple $\widehat{\boldsymbol{\beta}}$ y $\hat{\Sigma}$ tienen las siguientes propiedades

1. Cada columna de $\widehat{\boldsymbol{\beta}}$ es normal: $\widehat{\boldsymbol{\beta}}_{(j)}$ es $N_{q+1}[\boldsymbol{\beta}_{(j)}, \sigma_{jj}(\mathbf{X}'\mathbf{X})], j = 1, 2, \ldots, p$.
2. $n\hat{\Sigma} = E$ se distribuye $W_p(n - q - 1, \Sigma)$, donde $\mathbf{E} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$.
3. cada $\hat{\beta}_{jk}$ y $[n/(n-q-1)]\hat{\sigma}_{jk}$ es el estimador insesgado de varianza mínima de β_{jk} y σ_{jk}.

5.2.5. Diagnóstico sobre los Residuales

El análisis de residuales permite evaluar la calidad del ajuste del modelo, con el objetivo de buscar patrones o estructuras en los residuales que indiquen que las suposiciones del modelo como lo son la normalidad, la homogeniedad de varianzas y la linealidad se cumplen. Caroni (1987) extiende la idea del análisis de residuales de un modelo lineal múltiple a un modelo lineal múltiple multivariado para considerar simultáneamente las diferentes variables de respuesta.

Se pueden desarrollar estadísticas equivalentes para el caso de respuesta multivariante. Como se observó anteriormente el modelo lineal múltiple multivariado es $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, el estimador por mínimos cuadrados ordinarios es $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ y sus respectivos residuales son $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbf{Y}\widehat{\boldsymbol{\beta}}$. La fila i de $\hat{\boldsymbol{\varepsilon}}$ es $\hat{\varepsilon}_i'$ con varianza $(1 - v_{ii})\Sigma$, donde v_{ii} es el *i-ésimo* elemento de la diagonal de $V = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ conocida como la matriz *hat*. Una forma de examinar estos vectores de residuos es reduciendolos a escalares como distancias con respecto a alguna norma, con versiones estudentizadas interna y externamente.

- **Residuales estudentizados internamente**

$$R_i^2 = \frac{\hat{\varepsilon}_i'\hat{\Sigma}\hat{\varepsilon}_i}{(1 - v_{ii})}$$

donde $\hat{\Sigma}$ es $\hat{\varepsilon}\hat{\varepsilon}/(n - q)$

- **Residuales estudentizados externamente**

$$T_i^2 = \frac{\hat{\varepsilon}_i' \hat{\Sigma}_{(i)} \hat{\varepsilon}_i}{(1 - v_{ii})}$$

donde $\hat{\Sigma}_{(i)}$ es igual que en los residuales estudentizados internamente con la diferencias de que se estima sin la observación i.

Con respecto a medidas de influencia se tiene las distancia de Cook, una opción propia del problema multivariado, es cuando se toma una fila de la matriz $\widehat{\boldsymbol{\beta}}$. Esto corresponde a los coeficientes de regresión para todas las variables en un predictor particular y ha sido una elección utilizada en las aplicaciones probadas hasta ahora. Sea $\hat{\beta}_j$ la fila j de $\widehat{\boldsymbol{\beta}}$, entonces bajo supuestos de normalidad

$$g_j^{-1/2}(\beta_j - \hat{\beta}_j) \sim N_p(0, \Sigma) \tag{5.8}$$

donde g_{jj} es el *j-ésimo* elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$. Esto indica como construir una medida de Cook similar basada en elipsoides de confianza análogos. Específicamente

$$D_i = \frac{(\hat{\beta}_{j(i)} - \hat{\beta}_j)'(g_{jj}\hat{\Sigma})^{-1}(\hat{\beta}_{j(i)} - \hat{\beta}_j)(n - p - q + 1)}{p(n - q)} \tag{5.9}$$

donde $\hat{\beta}_{j(i)}$ es las estimación de $\hat{\beta}_j$ después de haber eliminado la obervación i, y está medida puede ser convertida a un percentil de la distribución F con p y $n-q-p+1$ grados de libertad.

5.3. Pruebas de hipótesis para los parámetros del modelo

Se mostrarán tres pruebas de hipótesis para los $\beta' s$. Se asumirá que y_i es $N_p(\boldsymbol{\beta}' x_i, \Sigma)$ para obtener pruebas F.

5.3.1. Prueba de regresión general

Según Rencher (1998) se considera la hipótesis de que ninguna de las variables regresoras $x_1, x_2, \ldots, x_q$ predice alguna de las variables respuesta $y_1, y_2, \ldots, y_p$, la cual puede ser expresada como $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{O}$, donde $\boldsymbol{\beta}_1$ incluye todas las filas de $\boldsymbol{\beta}$ exceptuando la primera

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}'_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} = \begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ - & - & - & - \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{q1} & \beta_{q2} & \dots & \beta_{qp} \end{pmatrix}$$

No se incluye $\boldsymbol{\beta}'_0 = \mathbf{0}'$ en la hipótesis, porque esto restringiría a todas las $y's$ a tener interceptos de cero. La hipótesis alternativa es $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{O}$, la cual implica que se quiere saber si al menos uno de los $\beta_{jk} \neq 0, j = 1, 2, \dots, q;\ k = 1, 2 \dots, p$

Se puede escribir la $SCE = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'$ entonces

$$\mathbf{Y}'\mathbf{Y} = \left(\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'y\right) + \widehat{\boldsymbol{\beta}}'\mathbf{X}'Y \tag{5.10}$$

lo cual particiona a $\mathbf{Y}'\mathbf{Y}$ en una parte debida a $\boldsymbol{\beta}$ y a una parte debida a las desviaciones del modelo ajustado. Además, para corregir a Y por su media y así evitar la inclusión de $\boldsymbol{\beta}'_0 = \mathbf{0}'$, se resta $n\bar{y}\bar{y}'$ en ambos lados para obtener

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} - n\bar{y}\bar{y} &= \left(\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'Y\right) + \left(\widehat{\boldsymbol{\beta}}'\mathbf{X}'Y - n\bar{y}\bar{y}'\right) \\ &= SCE + SCR \\ &= \mathbf{E} + \mathbf{H} \end{aligned} \tag{5.11}$$

donde $Y'Y - n\bar{y}\bar{y}' = \sum_i^n (y_i - \bar{y})^2$ lo cual corresponde a la suma total de cuadrados ajustados por la media y $SCR = \widehat{\boldsymbol{\beta}}'\mathbf{X}'Y - n\bar{y}\bar{y}'$ es la suma de cuadrados de la regresión general ajustada por el intercepto.

Se puede probar $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{O}$ por medio del estadístico

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} = \frac{|\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'Y|}{|\mathbf{Y}'\mathbf{Y} - n\bar{y}\bar{y}'|} \tag{5.12}$$

el cual se distribuye como $\Lambda_{p,q,n-q-1}$, cuando H_0 es verdadera y donde Λ corresponde a la distribución Lambda de Wilks. Se rechaza H_0 si $\Lambda \leq \Lambda_{\alpha,p,q,n-q-1}$. Si $\mathbf{H}$ es grande debido a los valores grandes de $\hat{\beta}_{jk}$, entonces se esperaría que $|\mathbf{E} + \mathbf{H}|$ fuera lo suficientemente más grande que $|\mathbf{E}|$, por tanto, Λ llevaría al rechazo de H_0. Además, podemos encontar los valores propios $\lambda_1, \lambda_{2,}, \dots, \lambda_s$ de $\mathbf{E}^{-1}\mathbf{H}$, los cuales nos indicarán el rango de $\boldsymbol{\beta}_1$, en el caso de un solo valor propio diferente de cero, el rango de $\boldsymbol{\beta}_1$ es 1. Hay varias maneras de que pueda ocurrir, por ejemplo, $\boldsymbol{\beta}_1$ podría tener una fila diferente de cero, lo cual indicaría que solo una de las x's predice las y's.

5.3.2. Prueba para un subconjunto de $\boldsymbol{\beta}$

Se considera la hipótesis de que las y's no dependen de la última h de las x's, x_{q-h-1}, $x_{q-h+2}, \ldots, x_q$. Con esto se quiere decir que ninguna de las y's es predicha por ninguna de estás h x's. Para expresar esta hipótesis, se escribe la matriz $\boldsymbol{\beta}$ en forma dividida

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_d \end{pmatrix} \tag{5.13}$$

donde el subíndice r denota el subconjunto de β_{jk} que se mantendrán en el modelo reducido y d representa el subconjunto de β_{jk} que se eliminarán si no son predictores significativos de las y's. Así, $\boldsymbol{\beta}_d$ tiene h filas. La hipótesis se puede expresar como

$$H_0 : \boldsymbol{\beta}_d = \boldsymbol{O} \tag{5.14}$$

Si $\mathbf{X}_r$ contiene las columnas de $\mathbf{X}$ correspondientes a $\boldsymbol{\beta}_r$, entonces el modelo reducido es

$$\mathbf{Y} = \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon} \tag{5.15}$$

Para comparar el ajuste del modelo completo y el modelo reducido, usamos la diferencia entre la matriz de suma de cuadrados y productos de la regresión para el modelo completo $\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$, y las matriz de suma de cuadrados y productos de la regresión para el modelo reducido $\widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y}$. Esta diferencia se convierte en la matriz $\mathbf{H}$

$$\mathbf{H} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y} \tag{5.16}$$

Por tanto, la prueba de $H_0 : \boldsymbol{\beta}_d = \boldsymbol{O}$ es una prueba de la significancia completa y reducida de $x_{q-h+1}, x_{q-h+2}, \ldots, x_q$ por encima y más allá de $x_1, x_2, \ldots, x_{q-h}$.

Para hacer la prueba, se usa la matriz $\mathbf{E}$ basada en el modelo completo, $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$. Entonces

$$\begin{aligned} \mathbf{E} + \mathbf{H} &= (\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}) + (\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y}) \\ &= \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y} \end{aligned} \tag{5.17}$$

y el estadístico Λ de Wilks esta dado por

$$\begin{aligned} \Lambda(x_{q-h+1}, \ldots, x_q | x_1, \ldots, x_{q-h}) &= \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} \\ &= \frac{|\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y}|} \end{aligned} \tag{5.18}$$

el cual se distribuye $\Lambda_{p,h,n-q-1}$ cuando $H_0 : \boldsymbol{\beta}_d = \boldsymbol{O}$ es verdadera. El Λ de Wilks en (5.18) proporciona una prueba del modelo completo y reducido, se puede expresar en terminos de Λ para el modelo completo y un Λ similar para el modelo reducido. En el denominador de (5.18) se tiene $\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y}$, la cual es la matriz de error para el modelo reducido $\mathbf{Y} = \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$ en (5.15). Esta matriz de error puede ser usada en una prueba para la significancia de la regresión general en el modelo reducido, como en (5.12).

$$\Lambda_r = \frac{|\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - n\bar{y}\bar{y}'|} \tag{5.19}$$

Como Λ_r en (5.19) tiene le mismo denominador que Λ en (5.12), se reconoce a (5.18) como el ratio entre el Λ de Wilks para la prueba general de la regresión en el modelo completo para el Λ de Wilks para la prueba de regresión general en el modelo reducido

$$\begin{aligned}\Lambda(x_{q-h+1},\ldots,x_q|x_1,\ldots,x_{q-h}) &= \frac{|\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{Y}|} \\ &= \frac{\frac{|\mathbf{Y}'\mathbf{Y}-\widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y}-n\bar{y}\bar{y}'|}}{\frac{|\mathbf{Y}'\mathbf{Y}-\widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y}-n\bar{y}\bar{y}'|}} \\ &= \frac{\Lambda_f}{\Lambda_r}\end{aligned} \tag{5.20}$$

donde Λ_f esta dado por (5.12). Los valores críticos o las pruebas aproximadas para estos estadísticos de prueba están basados en $v_H = h$, $v_E = n - q - 1$.

5.3.3. Prueba e intervalos de confianza individuales β_{jk}

Una prueba o intervalo de confianza para un único β_{jk} es fácil de obtener. Dado que $\hat{\beta}_{(k)}$ se distribuye $N_{q+1}\left(\beta_{(k)}, \sigma_{kk}(X'X)^{-1}\right)$, donde $\beta_{(k)}$ es la k-*ésima* columna de $\boldsymbol{\beta}$. Por tanto, el estimador $\hat{\beta}_{jk}$ se distirbuye como una normal univariada $N\left(\beta_{jk}, \sigma_{kk}g_{jj}\right)$, donde g_{jj} es el *j-ésimo* elemento de la diagonal de $(X'X)^{-1}$. Se estima σ_{kk} por s_k^2, el elemento *k-ésimo* de la diagonal de $S_e = (Y'Y - \hat{\boldsymbol{\beta}}'X'Y)/(n-q-1)$. Por tanto, se puede probar $H_0 : \beta_{jk} = 0$ con

$$t = \frac{\hat{\beta}_{jk}}{s_k\sqrt{g_{jj}}} \tag{5.21}$$

Se rechaza H_0 si $|t| \geq t_{\alpha/2,n-q-1}$. El intervalo correspondiente para β_{jk} es

$$\hat{\beta}_{jk} \pm t_{\alpha/2,n-q-1}s_k\sqrt{g_{jj}} \tag{5.22}$$

5.4. El *Bootstrap*

La idea básica del *Bootstrap* es crear muestras del conjunto original de datos a partir de las cuales se puede evaluar la variabilidad de los estimadores sin necesidad de recurrir a un análisis complejo. Fue introducido por Bradley Efron en el año 1979 y desde esa época diferentes autores como Robert Tibshirani y David Hinkley han realizado aportes en el tema.

Suponga que se tiene una muestra aleatoria $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ de una variable X cuya función de distirbución se denota como F, muestra la cual se utilizará para realizar inferencia sobre un parámetro (θ) de la población. La función de distribución empírica, la cual se denotará como $\hat{F}$, es una estimación de toda la distribución F, $\hat{F}$ se obtiene a través de muestras *Bootstrap* $\mathbf{x}^* = (\mathbf{x}_1^*, \mathbf{x}_3^*, \ldots, \mathbf{x}_r^*)$ las cuales se calculan muestreando aleatoriamente r veces los datos originales.(Efron y Tibshirani, 1993)

5.4.1. Intervalos por percentiles

Un enfoque simple para construir intervalos de confianza de los parámetros es el mencionado por Fox (2016), donde se usan los cuantiles de la distribución de muestreo *Bootstrap* del estimador con el objetivo de establecer los puntos finales del intervalo. Sea $\hat{\theta}^*_{(r)}$ las estimaciones *Bootstrap* ordenadas, y suponga que se quiere construir un intervalo al $(100 - \alpha)\,\%$. Si el número de réplicas *Bootstrap* r es grande, entonces

$$\hat{\theta}^*_{(I)} < \theta < \hat{\theta}^*_{(S)} \tag{5.23}$$

donde $I = r\alpha/2$ y $S = r(1 - \alpha/2)$.

5.4.2. Intervalos *Bootstrap* mejorados (Bc_a)

De acuerdo con Efron y Tibshirani (1993) y Fox (2016) para mejorar la precisión de los intervalos de confianza por percentiles se requiere el valor unitario de la normal con probabilidad $\alpha/2$ a la derecha, y dos factores de correción, Z y A, definidos como

$$Z \equiv \Phi^{-1}\left[\frac{\#_{b=1}^{r}(\hat{\theta}^*_{(b)} < \hat{\theta})}{r}\right] \tag{5.24}$$

donde Φ^{-1} es la inversa de la distribución normal estándar y $\#(\hat{\theta}^*_{(b)} < \hat{\theta})/r$ es la proporción de las réplicas Bootstrap por debajo la estimación $\hat{\theta}$. Ahora, sea $\hat{\theta}_{(-i)}$ el valor producido cuando se elimina la *í-esima* observación de la muestra y $\bar{\theta}$ representa el promedio de los $\hat{\theta}_{(-i)}$, es decir, $\theta \equiv \sum_{i=1}^{n} \hat{\theta}_{(-i)}$ (Proceso *Jackknife*). Luego se calcula

$$A \equiv \frac{\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(i)})^3}{6\left[\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(i)})^2\right]^{2/3}} \tag{5.25}$$

con los factores de correción A y Z se obtiene

$$A_1 \equiv \Phi\left[Z + \frac{Z - z_{\alpha/2}}{1 - A(Z - z_{\alpha/2})}\right], \tag{5.26}$$

$$A_2 \equiv \Phi\left[Z + \frac{Z + z_{\alpha/2}}{1 - A(Z + z_{\alpha/2})}\right] \tag{5.27}$$

donde Φ es la función de distribución acumulada de la normal estándar. Los valores de A_1 y A_2 son usados para localizar los puntos finales del intervalo de confianza de percentiles corregido

$$\hat{\theta}^*_{(I^*)} < \theta < \hat{\theta}^*_{(S^*)} \tag{5.28}$$

donde $I^* = rA_1$ y $S^* = rA_2$.

5.4.3. Intervalos *Bootstrap*-t

A través del uso de *Bootstrap* es posible obtener intervalos precisos sin tener que hacer suposiciones teóricas de la distribución normal. Este procedimiento estima la distribución del estadístico Z directamente de los datos, en esencia se calculan los cuantiles que son apropiados para el conjunto de datos en cuestión. Estos cuantiles se usan para construir los intervalos de confianza exactamente de la misma forma que los cuatiles que se usarían de una normal y una distribución t. Los cuantiles *Bootstrap* se construyen generando r muestras *Bootstrap*, y con estas se calcula la versión *Bootstrap* de Z para cada una.

Se generan r muestras *Bootstrap* $\mathbf{x}^*_1, \mathbf{x}^*_3, \ldots, \mathbf{x}^*_r$ y para cada una se calcula

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{se}^*(b)} \tag{5.29}$$

donde $\hat{\theta}^*(b)$ es el valor de $\hat{\theta}$ para la submuestra *Bootstrap* $\mathbf{x}^*_r$ y $\widehat{se}^*(b)$ es el error estándar estimado de $\hat{\theta}^*$ para la submuestra $\mathbf{x}*_r$. El percentil α de $Z^*(b)$ es estimado por el valor de $\hat{t}_{(\alpha)}$ tal que

$$\alpha = \frac{\#\{Z^*(b) \leq \hat{t}_{(\alpha)}\}}{r} \tag{5.30}$$

Finalmente, el intervalo *Bootstrap*-t es

$$(\hat{\theta} - \hat{t}_{(1-\alpha)}\widehat{se}, \hat{\theta} - \hat{t}_{(\alpha)}\widehat{se}). \tag{5.31}$$

5.4.4. Pruebas de hipótesis *Bootstrap*

De acuerdo con Godfrey (2009), se supone que se dispone de una muestra aleatoria simple de una población normal y se desea probar que la media de la población es igual a cero, sin ninguna restricción sobre la varianza σ^2, excepto que sea finita y positiva.

Ahora, suponga que las variables aleatorias $y_1, y_2, \ldots, y_n$ se distirbuyen normalmente, identica e independientemente. La hipótesis nula a probar es $H_0 : E(y) = 0$. Como punto de partida se supone que la hipótesis alternativa es $H_1 : E(y) > 0$. Claramente H_0 solo determina uno de los dos parámetros que juntos definen un solo miembro de la familia normal de distribuciones. El parámetro σ^2 permanece desconocido si H_0 es verdadera o no.

Parámetros desconocidos que no son determinados por la hipótesis nula a veces son llamados parámetros *nuisance*. Una estrategia común para lidiar con estos parámetros es remplazar estos términos desconocidos por estimadores consistentes. Para el ejemplo, es conveniente estimar σ^2 por

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5.32}$$

en el cual

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.33}$$

denota la media muestral la cual es insesgada para la $E(y)$, sea H_0 verdadera o no.

Es conocido que, cualquiera que sea el valor verdadero de $E(Y)$

$$\frac{\bar{y} - E(y)}{\sqrt{s_y^2/n}} \sim t_{n-1} \tag{5.34}$$

Por tanto, cuando $H_0 : E(y) = 0$ es verdadera, el estadístico de prueba esta dado por

$$\hat{\tau} = \frac{\bar{y}}{\sqrt{s_y^2/n}} \tag{5.35}$$

el cual también sigue una distribución t_{n-1}. Un estadístico de prueba, como $\hat{\tau}$, que bajo su hipótesis nula asociada, tiene una distribución que no depende de ningún parámetro desconocido es llamado estadístico pivote. Cuando el estadístico de prueba es pivote, es posible obtener un test exacto usando métodos de simulación.

Ahora, suponga que no dispone de información precisa que permita espicificar las forma de su distribución. Sin embargo, una prueba asintótica se obtiene fácilmente. Bajo restricciones leves, se puede apelar a un teorema del límite central y así

$$\hat{\tau} = \frac{\bar{y}}{\sqrt{s_y^2/n}} \sim N(0,1) \tag{5.36}$$

cuando la hipótesis nula es veradera. En consecuencia, es asintóticamente válido utilizar valores críticos de la distribución normal estándar. Se debe tener en cuenta que solo la distribución asintótica del estadístico de prueba $\hat{\tau}$ es independiente de los parámetros desconocidos y por tanto, es asintóticamente pivote, pero no es un pivote exacto. La distribución muestral finita de $\hat{\tau}$ bajo la hipótesis nula depende, en parte, de la distribución desconocida de un término típico y_i. Si se utilizan métodos de simulación en un intento de obtener mejor control de los niveles de singificancia de muestras finitas que el proporcionado por la teória asintótica, la distirbución desconocida de cada y_i tendrá que imitarse en el esquema de simulación.

Es conveniente usar F, la distribución de probabilidad acumulada (CDF) de los y_i independientes e idénticamente distribuidos, para representar la distribución desconocida del parámetro. La CDF evaluada en algún número c se define como

$$F(c) = Pr(y \leq c) \tag{5.37}$$

la cual es estimada consistentemente por la proporción muestral correspondiente

$$\frac{1}{n}\sum_{i=1}^{n}(y_i \leq c) = \frac{\#(y_i \leq c)}{n} \tag{5.38}$$

donde $\#(y_i \leq c)$ denota el número de veces que y_i es menor o igual a c. Esta proporción muestral se puede reinterpretar como la CDF para una variable artificial y°, definida condicionalmente sobre los datos observados, con

$$Pr(y^\circ = y_i) = \frac{1}{n}, i = 1, \ldots, n \tag{5.39}$$

ya que, con esta disitribución de probabilidad

$$Pr(y^\circ \leq c) = F^\circ(c) = \frac{\#(y_i \leq c)}{n} \tag{5.40}$$

la cual corresponde a la distribución de probabilidad empírica de los datos reales $\hat{F}$.

En la proporción muestral en (5.38), es tentador que se piense usar (5.39) para derivar los datos artificiales. En el esquema de simulación basado en (5.39), a cada valor de la muestra real se le asigna la misma probabilidad. Así, se obtienen las muestras *Bootstrap* de tamaño n, mediante un muestreo aleatorio simple con remplazo de los

datos originales. Sin embargo, hay un problema con este proceso de generación de muestras *Bootstrap*.

El valor esperado de y° en el mundo simulado, condicionado a los datos observados, es

$$E^\circ(y^\circ) = \sum_{i=1}^{n} y_i Pr(y^\circ = y_i) = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{5.41}$$

La idea es aproximar el comportamiento de $\hat{\tau}$ bajo la hipótesis nula, que específica una media poblacional de cero. Es decir, que el esquema de simulación utilizado en (5.39) no pertenece a la familia de distribuciones en donde la hipótesis nula es verdadera.

Así que lo ideal es centrar los datos restando $\bar{y}$ a cada valor. Dados estos datos centrados, se obtienen r muestras aleatorias *Bootstrap* con remplazo de tamaño n, denotadas por $y^*_{r1}, y^*_{r2}, \ldots, y^*_{rn}$, del modelo de probabilidad *Bootstrap* definido por

$$Pr(y^* = y_i - \bar{y}) = \frac{1}{n}, i = 1, \ldots, n \tag{5.42}$$

para el cual $E^*(y^*) = 0$. Las contrapartes del *Bootstrap* de la media muestral y el estimador de la varianza de los datos reales están dados por

$$\bar{y}^*_r = \frac{1}{n}\sum_{i=1}^{n} y^*_{ri} \tag{5.43}$$

$$s^{*2}_{yr} = \frac{1}{n-1}\sum_{i=1}^{n}(y^*_{ri} - \bar{y}^*_r)^2 \tag{5.44}$$

respectivamente. De igual forma, $\hat{\tau}$ por las muestras *Bootstrap* es

$$\tau^*_r = \frac{\bar{y}^*_r}{\sqrt{s^{*2}_{yr}/n}}. \tag{5.45}$$

5.4.5. El *Bootstrap* para regresión lineal múltiple

Se han desarrollado dos metodologías para realizar las estimaciones de los parámetros del modelo lineal múltiple por *Bootstrap*. La primera consiste en tomar los regresores como aleatorios y seleccionar muestras directamente de las observaciones. La segunda trata a los regresores como fijos y toma muestras de los residuos del modelo de regresión ajustado. Fox (2016) sugiere que se deben seguir los siguientes pasos para la estimación de los parámetros de una regresión lineal

1. Se estiman los coeficientes del modelo $\beta_0, \beta_1, \ldots, \beta_p$ para la muestra original, y se calculan los valores predichos y su respectivos residuos para cada observación.

2. Se selecciona una muestra *Bootstrap* para los residuales e_b^* y con estos, se calculan los valores y *Bootstrap* mediante $y_b^* = \hat{y}_i + e_{bi}^*$.

3. Se realizan las estimaciones *Bootstrap* de los coeficientes con los valores fijos de las $x's$, si se utiliza el método de mínimos cuadrados entonces $\beta_b^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y_b^*$.

5.5. El *Jackknife*

El *Jackknife* es una técnica para estimar el sesgo y el error estándar de una estimación. El *Jackknife* es anterior al *Bootstrap* y es, al igual que el *Bootstrap*, una técnica de remuestreo. Suponga que tiene una muestra $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ y un estimador $\hat{\theta} = s(\mathbf{x})$. Se desea estimar el sesgo y el error estándar de $\hat{\theta}$. El *Jackknife* se centra en las muestras que dejan fuera una observación a la vez

$$\mathbf{X}_{(i)} = (x_1, x_2, \ldots, x_{l-i}, x_{l+i}, \ldots, x_n) \tag{5.46}$$

para $i = 1, 2, \ldots, n$, llamada muestras *Jackknife*. La *i-ésima* muestra *Jackknife* consiste en el conjunto de datos con la *i-ésima* observación eliminada. Sea $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$ la *i-ésima* réplica *Jackknife* de $\hat{\theta}$.

La estimación del sesgo *Jackknife* está definida por

$$\widehat{bias}_{jack} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta}) \tag{5.47}$$

donde

$$\hat{\theta}_{(.)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n. \tag{5.48}$$

La estimación del error estándar *Jackknife* se define por

$$\widehat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2}. \tag{5.49}$$

6. Metodología

Inicialmente se describió la teoría del *Bootstrap* y el *Jackknife* para los modelos lineales múltiples multivariados, para luego construir la función *Boot_rmm* para la estimación e inferencia de los parámetros del modelo. Específicamente, con la metodología *Bootstrap* se estimaron los coeficientes y la varianza del modelo al igual que con el *Jackknife*, sin embargo, las pruebas de hipótesis solo se realizaron mediante el *Bootstrap* debido a que el objetivo principal del *Jackknife* es estimar el sesgo y la varianza de los parámetros.

Como primer paso, el algoritmo estima los coeficientes del modelo, la matriz $\hat{\boldsymbol{\beta}}$, con la muestra original y obtiene los valores predichos para cada variable respuesta y su respectiva matriz de residuales. Posteriormente, se obtiene una muestra *Bootstrap* para construir la matriz ε_b^*, con la cual se calculan los valores "nuevos" de cada y, por medio de $Y_r^* = \hat{Y}_i + \varepsilon_{ir}^*$. Por último, se vuelve a estimar la matriz de los coefiecientes y se repiten los pasos anteriores tantas veces como se le indique a la función. Para finalmente poder promediar las estimaciones obtenidas en cada muestra. Adicionalmente, la función permite seleccionar el tipo de intervalo que se desea construir con su respectivo nivel de confianza.

Posterior a la creación de la función se construyeron escenarios de simulación con normalidad multivariadda y con diferentes tamaños de muestra, donde se buscó evaluar como trabaja dicha función. Por último, se realizó una aplicación del algoritmo donde se obtuvieron la estimaciones de los parámetros tanto por mínimos cuadrados como por los métodos de remuestreo *Bootstrap* y *Jackknife*, se realizó la inferencia de los coeficientes y se comparó lo obtenido por MCO y las técnicas de remuestreo.

Los datos provienen de estudiantes matriculados entre los cursos de tercero y once durante el año 2019, cuenta con variables asociadas a aspectos familiares, sociodemográficos y comportamiento del estudiante, se tomó como variables respuesta los puntajes de las variables *Debilidades* y *Motivación*. Estos datos se obtuvieron mediante encuestas que los estudiantes respondieron.

Los procedimientos anteriormente descritos se realizaron mediante el lenguaje de programación R en su versión 4.2.2. Además, se utilizaron las librerías *fastDummies* de Kaplan (2020), *dplyr* de Wickham et al. (2022), *ggplot2* de Wickham (2016), *car* de Fox y Weisberg (2019) y *qqplotr* de Almeida et al. (2018), para la creación de variables *dummies*, la manupilación de *data frames* y la creación de gráficos.

7. Resultados

7.1. Estimación por *Bootstrap*

En está sección se explicará como se realizarón las estimaciones y la inferencia de los parámetros por medio del *Bootstrap*.

7.1.1. Estimación de los parámetros de un Modelo Lineal Múltiple Multivariado por *Bootstrap*

Para estimar la matriz $\widehat{\boldsymbol{\beta}}$ es necesario obtener la muestra a partir de los errores los cuales se obtienen realizando la estimación por MCO, es decir, estimando $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ con la muestra original. Con los coeficientes estimados se calculan las estimaciones de $\hat{\mathbf{Y}}$ y se obtiene la matriz de errores mediante $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$. Con base a esta matriz, se obtiene un vector con las posiciones para la nueva muestra con remplazo de tamaño n.

Ahora, se encuentra la nueva matriz $\mathbf{Y}_r^* = \hat{\mathbf{Y}}_i + \hat{\boldsymbol{\varepsilon}}_{ir}$, la cual contiene los valores *Bootstrap* de $\mathbf{Y}$ y se realiza una regresión sobre estos valores para obtener los coeficientes *Bootstrap*, entonces $\widehat{\boldsymbol{\beta}}_r^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_r^*$. Lo anterior se repite r veces, para promediar los valores obtenidos en cada iteración para cada uno de los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}^*$, la cual se define como

$$\widehat{\boldsymbol{\beta}}^* = \begin{bmatrix} \frac{\sum_{b=1}^{r}\hat{\beta}_{11(b)}}{r} & \frac{\sum_{b=1}^{r}\hat{\beta}_{12(b)}}{r} & \cdots & \frac{\sum_{b=1}^{r}\hat{\beta}_{1p(b)}}{r} \\ \frac{\sum_{b=1}^{r}\hat{\beta}_{21(b)}}{r} & \frac{\sum_{b=1}^{r}\hat{\beta}_{22(b)}}{r} & \cdots & \frac{\sum_{b=1}^{r}\hat{\beta}_{2p(b)}}{r} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\sum_{b=1}^{r}\hat{\beta}_{(q+1)1(b)}}{r} & \frac{\sum_{b=1}^{r}\hat{\beta}_{(q+1)2(b)}}{r} & \cdots & \frac{\sum_{b=1}^{r}\hat{\beta}_{(q+1)p(b)}}{r} \end{bmatrix}. \tag{7.1}$$

Por último, se estima la matriz $\boldsymbol{\Sigma}$ correspondiente a las varianzas y covarianzas de las $\mathbf{y}$, mediante (5.5), con $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^*$.

7.1.2. Prueba de hipótesis para los β_{jk} del Modelo Lineal Múltiple Multivariado

Para saber si cada uno de los coeficientes es significativamente diferente de 0, se quiere probar $H_0 : \beta_{jk} = 0$ y $H_1 : \beta_{jk} \neq 0$, mediante el estadístico de prueba en (5.21), el cual se podría comparar con el estadístico calculado $|t| \geq t_{\alpha/2,n-q-1}$, sin embargo, no se estarían obteniendo inferencias confiables ya que cada uno de los $\hat{\beta}_{jk}^*$ se estimaron basandose en la función empírica de los datos por *Bootstrap*. Por lo cual, fue adecuado construir la distribución empírica del estadístico de prueba t.

Para calcular cada t_r^*, es necesario calcular la correspondiente matriz g^* y $\mathbf{S}_e^*$ para cada una de las r iteraciones *Boostrap*, es decir, si se tienen 500 muestras *Bootstrap* se calculan 500 veces tanto la matriz g y como la matriz $\mathbf{S}_e$, con el objetivo de calcular el error estándar de cada uno de los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}^*$.

Otro aspecto que se debe tener en cuenta es que se debe generar el esquema de simulación bajo la hipótesis nula, centrando cada uno de los $\hat{\beta}_{jk}^*$ por medio de la esperanza de la matriz $\boldsymbol{\beta}$. Por las propiedades de mínimos cuadrados el estimador $\widehat{\boldsymbol{\beta}}$ es insesgado, es decir, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, por tanto, para construir los estadísticos t_r^*, cada uno de los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}_r^*$ de cada iteración se centra mediante $\widehat{\boldsymbol{\beta}}$, es decir

$$t_r^* = \frac{\hat{\beta}_{jk(r)}^* - \hat{\beta}_{jk}}{s_{k_{(r)}}^* \sqrt{g_{jj_{(r)}}^*}} \tag{7.2}$$

donde $s_{k_{(r)}}^*$ es el *k-ésimo* elemento de la diagonal de $\mathbf{S}_e^*$ y $g_{jj_{(r)}}^*$ es el *j-ésimo* elemento de la diagonal de $g^* = (\mathbf{X}'\mathbf{X})^{-1}$ en la iteración r. Al haber obtenido las r réplicas *Bootstrap* del estadístico de prueba, la estimación *Boostrap* del valor-p para H_0 es

$$\hat{p}^* = \frac{\#t_{jk_{(r)}}^* \geq |t_{jk}|}{r} \tag{7.3}$$

donde t_{jk} es el estadístico de prueba de $\hat{\beta}_{jk}$ de la muestra original.

7.1.3. Estimación de la varianza para los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}$

Con las estimaciones de los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}^*$ y cada una de las estimaciones de dicha matriz obtenidas en cada una de las réplicas *Bootstrap* es posible estimar la varianza *Boot* de los coeficientes del modelo, la cual se define como

$$\widehat{\boldsymbol{\beta}}_{var}^* = \begin{bmatrix} \frac{\sum_{b=1}^r \hat{\beta}_{11(r)} - \hat{\beta}_{11}^*}{r} & \frac{\sum_{b=1}^r \hat{\beta}_{12(r)} - \hat{\beta}_{12}^*}{r} & \cdots & \frac{\sum_{b=1}^r \hat{\beta}_{1p(r)} - \hat{\beta}_{1p}^*}{r} \\ \frac{\sum_{b=1}^r \hat{\beta}_{21(r)} - \hat{\beta}_{21}^*}{r} & \frac{\sum_{b=1}^r \hat{\beta}_{22(r)} - \hat{\beta}_{22}^*}{r} & \cdots & \frac{\sum_{b=1}^r \hat{\beta}_{2p(r)} - \hat{\beta}_{2p}^*}{r} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\sum_{b=1}^r \hat{\beta}_{(q+1)1(r)} - \hat{\beta}_{(q+1)1}^*}{r} & \frac{\sum_{b=1}^r \hat{\beta}_{(q+1)2(r)} - \hat{\beta}_{(q+1)2}^*}{r} & \cdots & \frac{\sum_{b=1}^r \hat{\beta}_{(q+1)p(r)} - \hat{\beta}_{(q+1)p}^*}{r} \end{bmatrix}. \tag{7.4}$$

Por ejemplo, $\hat{\beta}_{12}^*$ corresponde al elemento de la fila uno y columna dos de la matriz $\widehat{\boldsymbol{\beta}}^*$ y $\hat{\beta}_{12(r)}^*$ al elemento de la fila uno y columna dos de la matriz $\widehat{\boldsymbol{\beta}}_{(r)}^*$. Con la varianza estimada es posible calcular el error estándar de los coeficientes como $\widehat{\boldsymbol{\beta}}_{se}^* = \sqrt{\widehat{\boldsymbol{\beta}}_{var}^*}$.

7.1.4. Intervalos *Bootstrap* para los coeficientes del modelo

Intervalos por percentiles

La estimación de los intervalos por percentiles se realiza como se muestra en la sección (5.4.1), donde $\hat{\theta}^*$ es cada uno de los coeficientes en la matriz $\widehat{\boldsymbol{\beta}}^*$, por tanto, se debe ordenar cada una de las estimaciones *Bootstrap* obtenidas para el respectivo $\hat{\beta}^*_{jk}$ y encontrar los cuantiles de la distribución empírica de los parámetros. La expresión análoga a (5.23) para cada $\hat{\beta}^*_{jk}$ es

$$\hat{\beta}^*_{jk_{(I)}} < \hat{\beta}_{jk} < \hat{\beta}^*_{jk_{(S)}} \tag{7.5}$$

donde $I = r\alpha/2$ y $S = r(1 - \alpha/2)$.

Intervalos *Bootstrap* mejorados (BC_a)

De la misma forma, los intervalos de confianza mejorados para los coeficientes se calculan encontrando los dos factores de correción, como se muestra en (5.25) y (5.24), para cada $\hat{\beta}^*_{jk}$. Luego, se calculan los valores de A_1 y A_2, los cuales permiten encontrar los cuantiles de la distribución empírica de cada coeficiente y hallar los límites de los intervalos. Por lo anterior, la expresión para cada intervalo se resume en

$$\hat{\beta}^*_{jk_{(I^*)}} < \hat{\beta}_{jk} < \hat{\beta}^*_{jk_{(S^*)}} \tag{7.6}$$

donde $I^* = rA_1$ $S^* = rA_2$.

Intervalos *Bootstrap-t*

Por último, para estimar los intervalos de confianza *Bootstrap-t* se aprovecharon los valores del estadístico t calculados para la prueba de hipótesis *Bootstrap*. Se calculan los respectivos valores del percentil de α con (5.30) para cada coeficiente $\hat{\beta}^*_{jk}$ y se calcula el cuantil de la distribución t con dicho valor de α. Al igual que en (5.31) el intervalo de confianza se expresa en este caso como

$$(\hat{\beta}^*_{jk} - \hat{t}_{(1-\alpha)}\widehat{\boldsymbol{\beta}}^*_{se_{(jk)}}, \hat{\beta}^*_{jk} - \hat{t}_{(\alpha)}\widehat{\boldsymbol{\beta}}^*_{se_{(jk)}}) \tag{7.7}$$

donde $\hat{t}$ es la distribución empírica de los valores t y $\widehat{\boldsymbol{\beta}}^*_{se_{(jk)}}$ es el error estándar *Bootstrap* del coeficiente jk.

7.2. Estimación por *Jackknife*

A continuación se explica como se realizaron las estimaciones de los coeficientes mediante el *Jackknife*.

7.2.1. Estimación de los coeficientes de un Modelo Lineal Múltiple Multivariado por *Jackknife*

Para estimar los coeficientes del modelo se elimina el *m-ésimo* elemento en la iteración m, por tal razón, la matriz con las variables de respuesta $\mathbf{Y}$ y la matriz con las covariables $\mathbf{X}$ tendrán $n-1$ filas en cada iteración. Cabe resaltar que el número de iteraciones es igual al número de observaciones de la muestra, así, al final se tendrían n estimaciones de la matriz $\widehat{\boldsymbol{\beta}}$, donde cada una se estimaria como

$$\widehat{\boldsymbol{\beta}}_m^J = (\mathbf{X}'_{(-m)}\mathbf{X}_{(-m)})^{-1}\mathbf{X}'_{(-m)}\mathbf{Y}_{(-m)} \tag{7.8}$$

Con las n matrices es posible calcular cada coeficente realizando un promedio entre los valores obtenidos en cada iteración,

$$\widehat{\boldsymbol{\beta}}^J = \begin{bmatrix} \frac{\sum_{m=1}^n \hat{\beta}_{11m}}{n} & \frac{\sum_{m=1}^n \hat{\beta}_{12m}}{n} & \cdots & \frac{\sum_{m=1}^n \hat{\beta}_{1pm}}{n} \\ \frac{\sum_{m=1}^n \hat{\beta}_{21m}}{n} & \frac{\sum_{m=1}^n \hat{\beta}_{22m}}{n} & \cdots & \frac{\sum_{m=1}^n \hat{\beta}_{2pm}}{n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\sum_{m=1}^n \hat{\beta}_{(q+1)1m}}{n} & \frac{\sum_{m=1}^n \hat{\beta}_{(q+1)2m}}{n} & \cdots & \frac{\sum_{m=1}^n \hat{\beta}_{(q+1)pm}}{n} \end{bmatrix}. \tag{7.9}$$

7.2.2. Estimación de la varianza de los coeficientes mediante *Jackknife*

El procedimiento de la varianza de las estimaciones mediante *Jackknife* es similar al *Bootstrap*. Se calcula la diferencia entre el promedio de las estimaciones y la estimación en cada iteración. Para cada uno de los β_{jk}^J, se calcula de la siguiente forma

$$\widehat{\beta}_{var_{jk}}^J = \frac{n-1}{n}\sum_{m=1}^{n}(\hat{\beta}_{jk_m} - \beta_{jk}^J)^2 \tag{7.10}$$

7.2.3. Intervalos de confianza para los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}$

Finalmente, se construyeron intervalos para las estimaciones *Jackknife*. Tomando las estimaciones en $\widehat{\boldsymbol{\beta}}^J$ y la varianza estimada en (7.2.2) es posible calcular los límites de confianza del intervalo como

$$\widehat{\boldsymbol{\beta}}_{jk}^J \pm t_{\alpha/2,n-q-1}\widehat{\beta}_{se_{jk}}^J. \tag{7.11}$$

7.3. Simulaciones

Para construir los escenarios de simulación, primero se generaron los errores $\hat{\boldsymbol{\varepsilon}}^s \sim N_p(\mathbf{0}, \Sigma)$, los cuales siguen una distribución normal multivariada y tienen vector de media $\mathbf{0}$. Luego, se generó la matriz $\mathbf{X}^s$ con la principal condición de evitar la multicolinealidad y se fijaron los valores de la matriz $\widehat{\boldsymbol{\beta}}^s$ como 1 para cada coeficiente. Finalmente, se generaron los valores de la matriz $\mathbf{Y}^s$ como $\mathbf{Y}^s = \mathbf{X}^s\widehat{\boldsymbol{\beta}}^s + \hat{\boldsymbol{\varepsilon}}^s$. Lo anterior, se realizó para cada método y para 5 tamaños de muestra diferentes.

Los resultados obtenidos se muestran en la tabla 7.1, como se generó una matriz $\widehat{\boldsymbol{\beta}}^s$ de tamaño 3×2, se tienen ocho coeficientes en el modelo. Se observa que para cada uno de los métodos el RMSE es similar. Además, el valor del RMSE disminuye a medida de que el tamaño de muestra aumenta, es decir, que el valor estimado es más cercano al valor real cuando la muestra es más grande.

Método	n	Σ	Beta1	Beta2	Beta3	Beta4	Beta5	Beta6	Beta7	Beta8
MCO	25	0.25	1.01	0.23	0.22	0.21	1.03	0.22	0.22	0.22
MCO	50	0.25	1.00	0.15	0.16	0.15	1.01	0.14	0.14	0.15
MCO	100	0.25	1.01	0.10	0.10	0.10	1.01	0.11	0.10	0.10
MCO	500	0.25	1.00	0.04	0.05	0.04	1.00	0.05	0.04	0.04
MCO	1000	0.25	1.00	0.03	0.03	0.03	1.00	0.03	0.03	0.03
BOOT	25	0.25	1.02	0.22	0.22	0.22	1.03	0.23	0.23	0.22
BOOT	50	0.25	1.00	0.15	0.15	0.15	1.01	0.15	0.14	0.15
BOOT	100	0.25	1.01	0.10	0.10	0.10	1.01	0.10	0.10	0.10
BOOT	500	0.25	1.00	0.05	0.04	0.05	1.00	0.04	0.05	0.04
BOOT	1000	0.25	1.00	0.03	0.03	0.03	1.00	0.03	0.03	0.03
JACKK	25	0.25	1.01	0.23	0.22	0.21	1.03	0.22	0.22	0.22
JACKK	50	0.25	1.00	0.15	0.16	0.15	1.01	0.14	0.14	0.15
JACKK	100	0.25	1.01	0.10	0.10	0.10	1.01	0.11	0.10	0.10
JACKK	500	0.25	1.00	0.04	0.05	0.04	1.00	0.05	0.04	0.04
JACKK	1000	0.25	1.00	0.03	0.03	0.03	1.00	0.03	0.03	0.03

Tabla 7.1: *Resultados de las simulaciones con datos normales multivariados*

7.4. Aplicación

El conjunto de datos utilizado contaba con registros de 940 estudiantes entre los cursos de tercero hasta once de seis colegios diferentes, donde se registraron variables sociodemográficas, de comportamientos de los estudiantes y de aspectos familiares. La variables independientes utilizadas consistian en el colegio al que pertence, el rango de edad del estudiante, el sexo, si los recursos son o no compartidos, si lleva o no onces al colegio y el tiempo que comparten con su familia.

Se seleccionaron dos variables dependientes: puntaje de motivación y puntaje de debilidades en los estudiantes. Si el puntaje en debilidades es alto significa que los estudiantes reconocen las actitudes negativas que no los permiten estar bien con ellos

mismos y sienten que el colegio los acompaña para mejorar las acitudes que no les permite estar bien con ellos mismos. Por otro lado, si el puntaje de motivación es alto significa que el estudiante se siente llamado a ayudar a los demás, su relación con Dios es el principal motor de su vida, se interesa por compartir lo que es auténticamente con los demás y se siente llamado a vivir sus cualidades auténticas en situaciones conretas.

Como se observa en la Tabla 7.2 y 7.3 se obtuvieron las estimaciones de los coeficientes del modelo por mínimos cuadrados ordinales, *Bootstrap* y *Jackknife*. Tanto para la variable *Debilidades* y la variable *Motivación* los valores estimados son similares, sobretodo entre el *Jackknife* y el MCO.

Al analizar los coeficientes obtenidos por el modelo, se observa que el puntaje de motivación y debilidades es menor en aquellos que comparten parcialmente y no comparten con sus familiares que aquellos que sí comparten, al igual que en aquellos estudiantes que son hombres y tienen entre 14 y 16 años. Por otro lado, el puntaje de debilidades es menor en aquellos que no llevan onces porque no quieren que en aquellos que si lo llevan. Por último, se observa que el puntaje de motivación es mayor en los colegios cuatro y cinco en comparación con el uno.

7.4.1. Estimaciones de los coeficientes de la matriz $\widehat{\beta}$

	Debilidades		
Coeficientes	MCO	BOOT	JACKK
Intercept	3.10	3.10	3.10
Colegio_2	-0.20	-0.20	-0.20
Colegio_3	-0.09	-0.09	-0.09
Colegio_4	0.13	0.13	0.13
Colegio_5	0.13	0.13	0.13
Colegio_6	0.05	0.04	0.05
Edad_14.16	-0.10	-0.11	-0.10
Edad_17.20	0.01	0.01	0.01
Edad_8.9	-0.12	-0.12	-0.12
Sexo_Masculino	-0.16	-0.16	-0.16
Recursos_No.Compartidos	0.02	0.02	0.02
Onces_No.traigo.porque.no.puedo	-0.16	-0.17	-0.16
Onces_No.traigo.porque.no.quiero	-0.29	-0.29	-0.29
Compartir_Familiar_No.Comparten	-0.17	-0.17	-0.17
Compartir_Familiar_Parcial	-0.21	-0.21	-0.21

Tabla 7.2: *Estimaciones de los coeficientes del modelo para la variable Debilidades por los tres métodos.*

	Motivación		
Coeficientes	MCO	BOOT	JACKK
Intercept	3.13	3.13	3.13
Colegio_2	-0.06	-0.06	-0.06
Colegio_3	0.08	0.08	0.08
Colegio_4	0.21	0.21	0.21
Colegio_5	0.18	0.18	0.18
Colegio_6	0.05	0.05	0.05
Edad_14.16	-0.16	-0.16	-0.16
Edad_17.20	0.08	0.07	0.08
Edad_8.9	0.01	0.01	0.01
Sexo_Masculino	-0.20	-0.20	-0.20
Recursos_No.Compartidos	-0.07	-0.06	-0.07
Onces_No.traigo.porque.no.puedo	-0.22	-0.23	-0.22
Onces_No.traigo.porque.no.quiero	-0.21	-0.20	-0.21
Compartir_Familiar_No.Comparten	-0.49	-0.49	-0.49
Compartir_Familiar_Parcial	-0.40	-0.40	-0.40

Tabla 7.3: *Estimaciones de los coeficientes del modelo para la variable Motivación por los tres métodos.*

7.4.2. Significancia de los $\hat{\beta}_{jk}$

Se realizaron las pruebas correspondientes para los coeficientes de cada variable. Para los métodos MCO y *Bootstrap* se uso el estadístico de prueba en (5.21) y en (7.2), respectivamente. Para el *Jackknife* se usaron los intervalos de confianza estimados para verificar la significancia en este método. Como se observa en la tabla 7.4 en algunos de los coeficientes la significancia cambia según el método para la variable *Debilidades*, para el *Bootstrap* el colegio 4 y 5 si son significativamente diferentes al colegio 1 pero para el MCO y el *Jackknife* no lo son. Por otro lado, aquellos que no llevan refrigerio porque no quieren no son significativamente diferentes de aquellos que comen refrigerio en *Jackknife*. De igual forma aquellos que no comparten con su familia no son significativamente diferentes de aquellos que si comparten con su familia. Las pruebas y los intervalos se construyeron con un nivel de significancia del 5 %.

	Debilidades		
Coeficientes	MCO	BOOT	JACKK
Intercept	Sí	Sí	Sí
Colegio_2	Sí	Sí	Sí
Colegio_3	No	No	No
Colegio_4	No	Sí	No
Colegio_5	No	Sí	No
Colegio_6	No	No	No
Edad_14.16	Sí	Sí	Sí
Edad_17.20	No	No	No
Edad_8.9	No	No	No
Sexo_Masculino	Sí	Sí	Sí
Recursos_No.Compartidos	No	No	No
Onces_No.traigo.porque.no.puedo	No	No	No
Onces_No.traigo.porque.no.quiero	Sí	Sí	No
Compartir_Familiar_No.Comparten	Sí	Sí	No
Compartir_Familiar_Parcial	Sí	Sí	Sí

Tabla 7.4: *Significancia de los coeficientes para la variable Debilidades*

7.4.3. Prueba de Normalidad multivariada en los residuales

Se realizaron pruebas de normalidad multivariada en los residuales para saber si se cumple el supuesto distribucional en cada uno de los métodos. Se usaron dos pruebas: Mardia y Royston, en donde el *valor-p* en MCO, *Boostrap* y *Jackknife* fue menor al nivel de significancia 5 %, por tanto, se rechazó la hipótesis nula de normalidad multivariada para cada una de las pruebas. Es decir, que los residuales no se distribuyen normales multivariados en ningun caso.

7.4.4. Gráficos Residuales

Distancia de Mahalanobis

Se calculó la distancia de Mahalanobis de la matriz de residuales $\hat{\boldsymbol{\varepsilon}}$, dicha distancia sigue aproximadamente una distribución χ^2 si $Y \sim N_p(\mu, \sigma)$. Por lo anterior, permite visualizar si los residuos cumplen el supuesto distribucional. Además, se construyeron bandas de confianza al 95 % por medio de la función *qqPlot*. Se observa en las figuras 7.1, 7.2 y 7.3 que hay valores que se encuentra fuera de las bandas de confianza e incluso los valores 36 y 62 de la distancia de Mahalanobis parecen estar influyendo en la falta de normalidad multivariada de los datos. Dicho gráfico se construyó para cada método, sin embargo, los resultados son similares.

	Motivación		
Coeficientes	MCO	BOOT	JACKK
Intercept	Sí	Sí	Sí
Colegio_2	No	No	No
Colegio_3	No	No	No
Colegio_4	Sí	Sí	Sí
Colegio_5	Sí	Sí	Sí
Colegio_6	No	No	No
Edad_14.16	Sí	Sí	Sí
Edad_17.20	No	No	No
Edad_8.9	No	No	No
Sexo_Masculino	Sí	Sí	Sí
Recursos_No.Compartidos	No	No	No
Onces_No.traigo.porque.no.puedo	No	No	No
Onces_No.traigo.porque.no.quiero	No	No	No
Compartir_Familiar_No.Comparten	Sí	Sí	Sí
Compartir_Familiar_Parcial	Sí	Sí	Sí

Tabla 7.5: *Significancia de los coeficientes para la variable Motivación*

Residuales estudentizados

En la Figura 7.4 se observan los residuales estudentizados para cada método, los resultados son muy similares para cada uno de ellos. Hay algunos valores que se alejan de la nube de puntos, posiblemente sean valores atípicos en el modelo e influyen en la falta de normalidad multivariada en los errores.

7.4.5. Distancia de Cook

En el gráfico 7.5 se observa la distancia de Cook para 4 de las filas de la matriz $\widehat{\boldsymbol{\beta}}$, hay varias observaciones que se alejan de la nube de puntos, sin embargo, ninguna es mayor a 1.

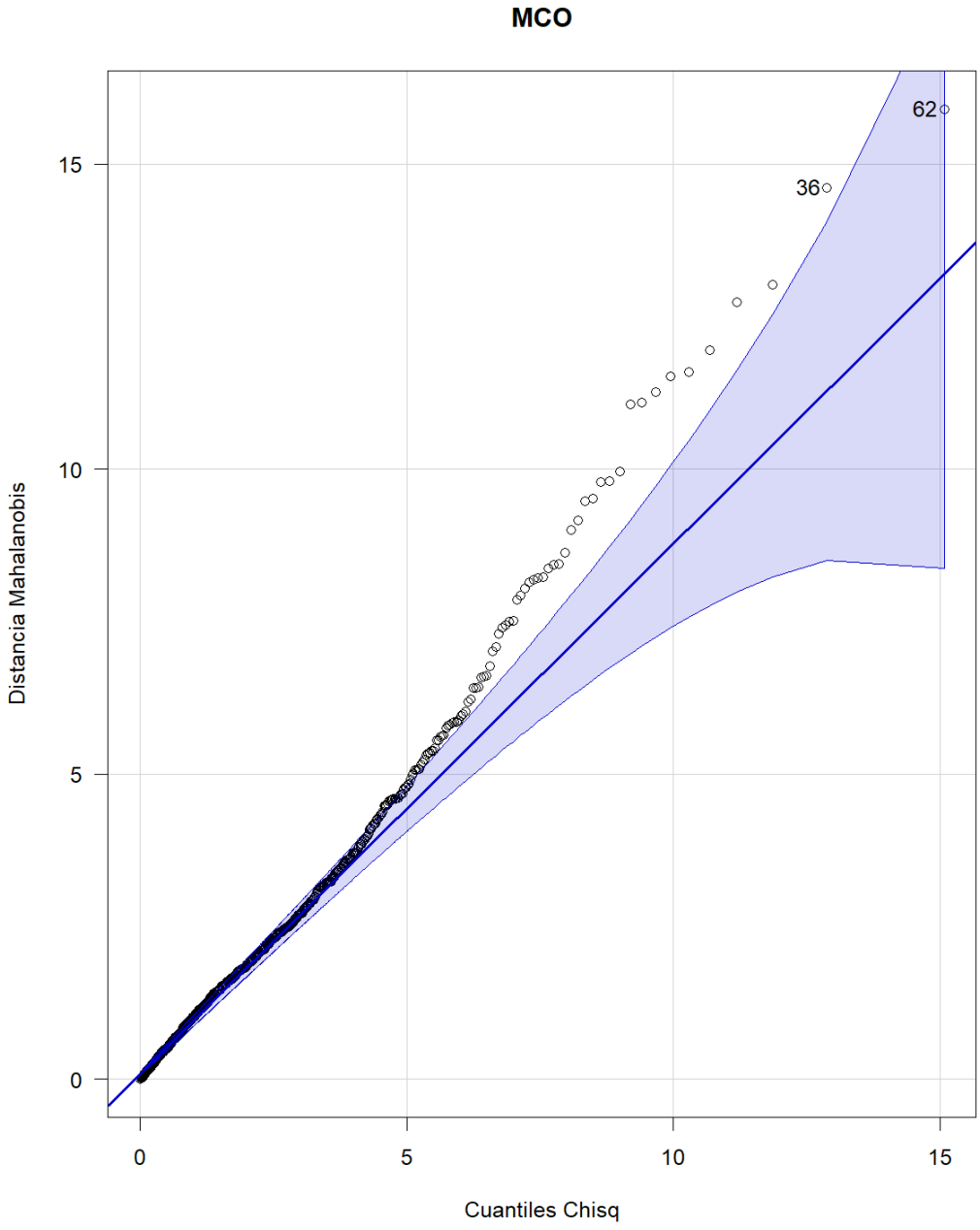

Figura 7.1: Distancia de Mahalanobis y cuantiles de una distribución χ^2 para el MCO

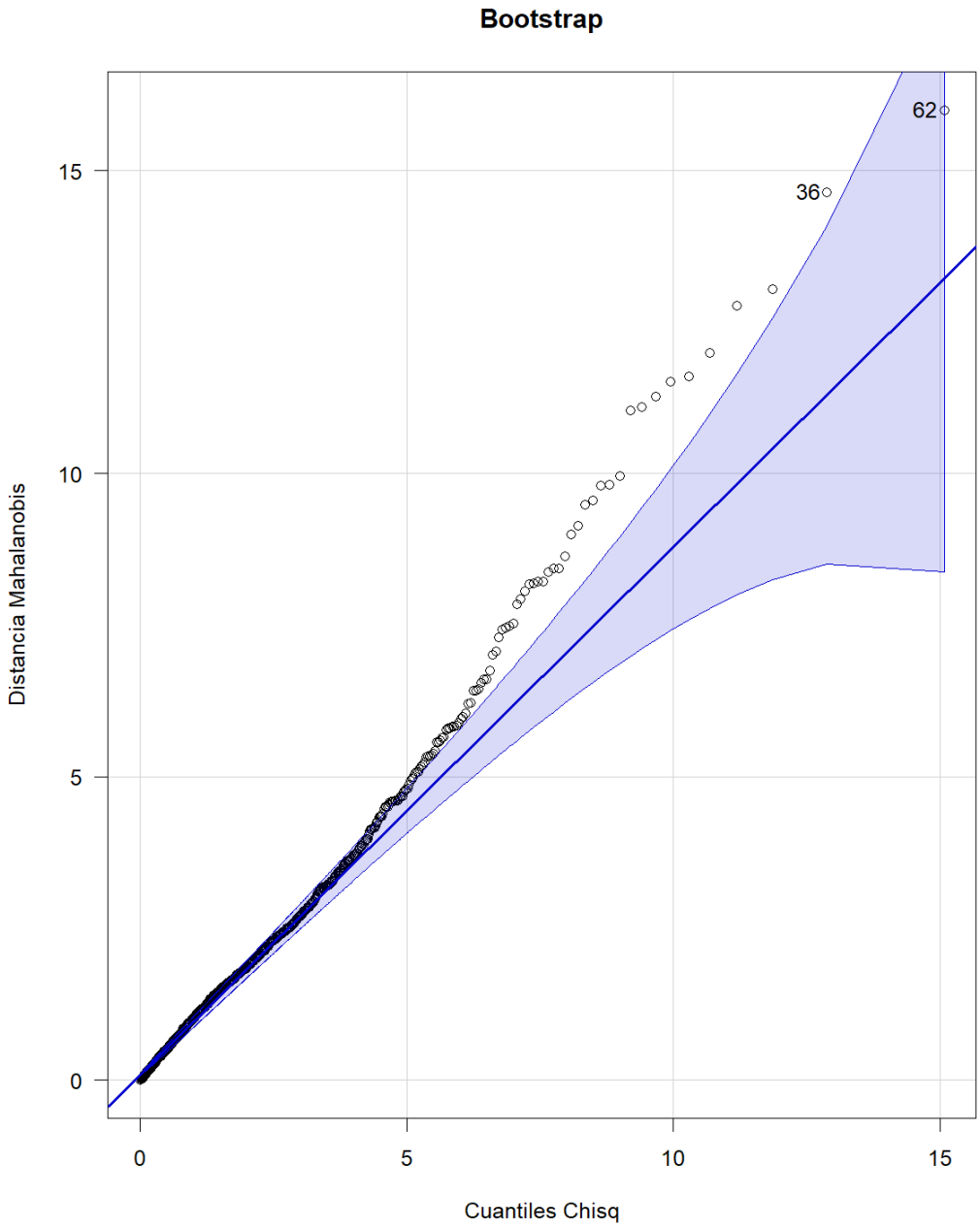

Figura 7.2: Distancia de Mahalanobis y cuantiles de una distribución χ^2 para el *Bootstrap*

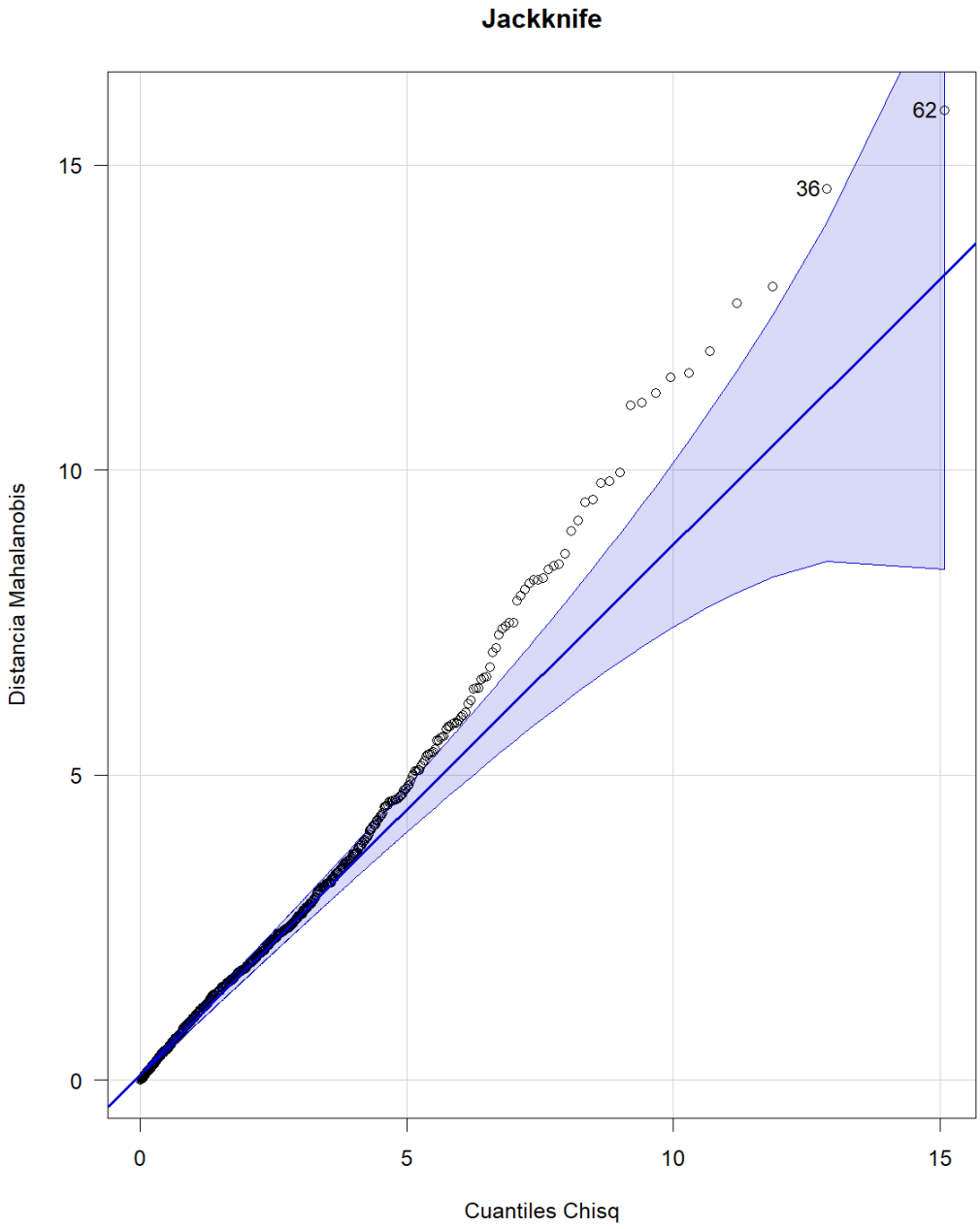

Figura 7.3: Distancia de Mahalanobis y cuantiles de una distirbución χ^2 para el *Jackknife*

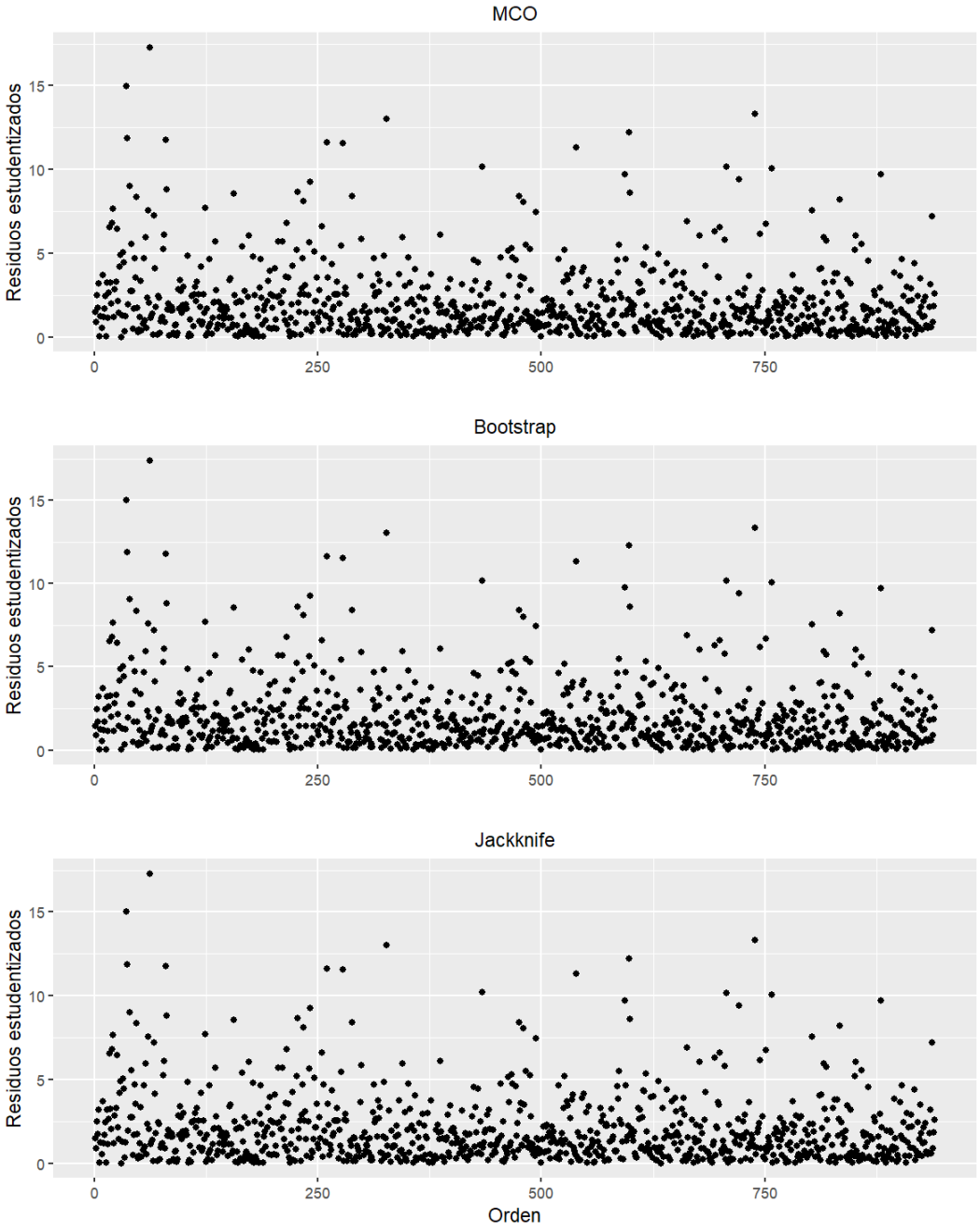

Figura 7.4: Residuales estudentizados del modelo para cada método

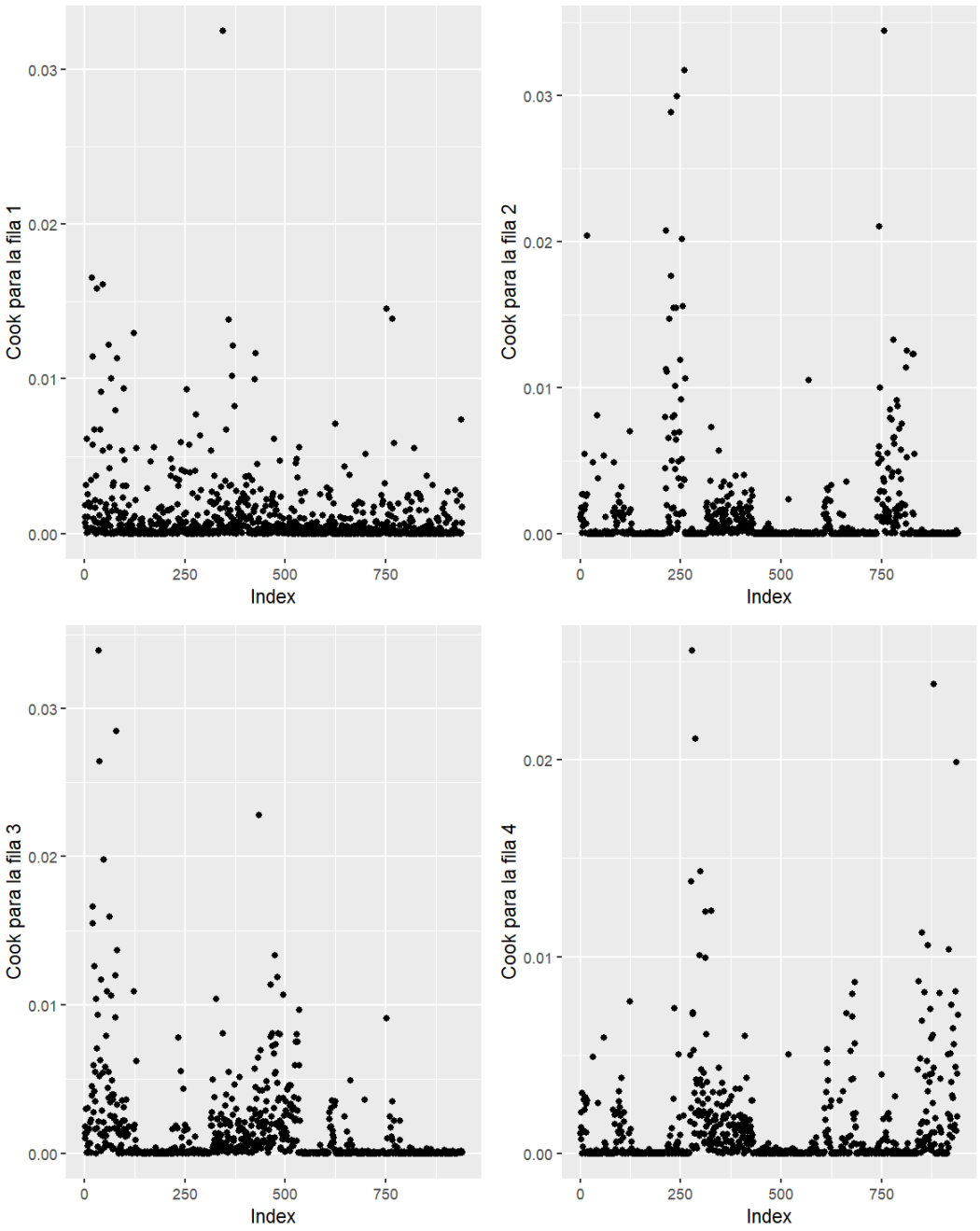

Figura 7.5: Distancia de Cook para 4 filas de la matriz $\widehat{\boldsymbol{\beta}}$

8. Discusión

Al obtener los resultados de la aplicación se observa que las estimaciones son similares para cada uno de los métodos, posiblemente esto se deba a que la esencia de las estimaciones sigue siendo la ecuación presentada en (5.4). Además, si las estimaciones fueran completamente diferentes se estaría afirmando que hay un problema en el proceso de estimación mediante las técnicas de remuestreo. No obstante, la mayor diferencia se observa en la significancia de los coeficientes, por lo cual es importante tener dos aspectos en cuenta. Primero, los datos utilizados para el ejemplo no cumplen el supuesto de normalidad multivariada, por tanto, cualquier inferencia de los parámetros pierde su validez y confianza. Segundo, la significancia de los coeficientes, principalmente por las pruebas de hipótesis *Bootstrap*, no utilizan ningún supuesto distribucional en los datos, en cambio se utiliza la distribución empírica $\hat{F}$, en consecuencia las inferencias realizadas mediante este método son confiables.

Por otra parte, en las simulaciones realizadas se observa un comportamiento similar en la estimación de los coeficientes cuando se tiene normalidad multivariada, por lo que las estimaciones que está realizando el algoritmo no son muy lejanas a las estimaciones que se realizan cuando se cumple el supuesto de normalidad, esto indica que las estimaciones obtenidas por las técnicas de remuestreo están siendo adecuadas y son cosistentes bajo el supuesto de normalidad.

8.1. Trabajo Futuro

Se pone en consideración realizar un estudio completo de simulación donde se desvíe adecuadamente de la normalidad multivariada, cumpliendo que la matriz de residuales tenga vector de medias de $\mathbf{0}$. Por otro lado, ampliar el desarrollo de la teoría del análisis de residuales para la adecuación del modelo y crear una librería en R que incluya una función para las predicciones del modelo.

9. Conclusiones

Se logró describir los procedimientos teóricos y prácticos de estimación de los parámetros de un modelo lineal múltiple multivariado de las técnicas de remuestreo *Bootstrap* y *Jackknife*, así como los procedimientos de inferencia sobre los coeficientes, específicamente pruebas de hipótesis e intervalos de confianza. Se construyeron, además, dos funciones en R que realizan todo el procedimiento de estimación e inferencia, la primera, denominada *rmm*, para la estimación por mínimos cuadrados ordinales, ya que no se contaba con una librería que realizara el procedimiento. La segunda, denominada *Boot_rmm*, utiliza los métodos de remuestreo para la estimación e inferencia, para el *Bootstrap* es posible realizar pruebas de hipótesis y construir tres tipos de intervalos diferentes, por otro lado, para el *Jackknife* se puede realizar la estimación de los parámetros y construir intervalos de confianza.

Se realizó una aplicación sobre un conjunto de datos reales en el que se modeló de manera conjunta un par de variables (puntaje) relacionadas con las debilidades y motivación de los estudiantes de seis colegios en la ciudad de Bogotá, empleando como variables explicativas características sociodemográficas e información de los padres. Como resultado se observa que dependiendo del proceso de inferencia sobre los parámetros (MCO o *Bootstrap/ Jackknife*) la significancia de los coeficientes puede variar debido a la violación de los supuestos sobre el modelo.

10. Bibliografía

Almeida, A., Loy, A., y Hofmann, H. (2018). *ggplot2 Compatible Quantile-Quantile Plots in R*.

Amiri, A., Saghaei, A., Mohseni, M., y Zerehsaz, Y. (2014). Diagnosis aids in multivariate multiple linear regression profiles monitoring. *Communications in Statistics - Theory and Methods*, 43:3057–3079.

Bradley, E. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Capital City Press.

Caroni, C. (1987). Residuals and influence in the multivariate linear model. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 36(4):365–370.

Clack, C. T. (2017). Modeling solar irradiance and solar pv power output to create a resource assessment using linear multiple multivariate regression. *Journal of Applied Meteorology and Climatology*, 56:109–125.

Davison, A. y Hinkley, D. (1997). Bootstrap methods and their application.

Efron, B. y Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, volume 1. Chapman Hall.

Eyvazian, M., Noorossana, R., Saghaei, A., y Amiri, A. (2011). Phase ii monitoring of multivariate multiple linear regression profiles. *Quality and Reliability Engineering International*, 27:281–296.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*.

Fox, J. y Weisberg, S. (2012). Bootstrapping regression models in r. an appendix to an r companion to applied regression. *Unpublished Manuscript. Accessed at http://socserv. mcmaster. ca/jfox/Books/Companion/appendix. html*.

Fox, J. y Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition.

Godfrey, L. (2009). *Bootstrap Tests for Regression Models*. Palgrave Macmillan.

Jeong, D. I., St-Hilaire, A., Ouarda, T. B., y Gachon, P. (2012). Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator. *Climatic Change*, 114:567–591.

Kaplan, J. (2020). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. R package version 1.6.3.

Miller, R. G. (1964). A trustworthy jackknife. *The Annals of Mathematical Statistics*, 35(4):1594–1605.

Monge, P. R. (1977). Multivariate multiple regression in communication research.

Nkurunziza, S. y Ahmed, S. E. (2011). Estimation strategies for the regression coefficient parameter matrix in multivariate multiple regression. *Statistica Neerlandica*, 65:387–406.

Quick, C. y James, K. (2013). Multivariate multiple regression with applications to powerlifting data.

Rencher, A. (1998). *Multivariate statistical inference and applications*.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., y Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.

Anexos

A. Funciones creadas

A.1. *rmm*

Función que estima los parámetros de un modelo lineal múltiple multivariado por mínimos cuadrados ordinales. Realiza pruebas de hipótesis individuales para los coeficientes y estima intervalos de confianza estándar e intervalos de confianza usando la varianza estimada por el *Jackknife*.

A.1.1. Argumentos

Y	matriz con las variables respuesta
X	matriz con las variables explicativas
alpha	nivel de significancia, por defecto 0,05
na.action	omitir valores faltantes, por defecto FALSE
VarJackk	TRUE para estimar los intervalos de confianza con varianza *Jackknife*

A.1.2. Salida

betas	p matrices de los coeficientes con sus respectivos IC
Sigma	matriz de varianzas y covarianzas
Residuals	residuales del modelo
pred	estimaciones de las variables de respuesta

A.2. *Boot_rmm*

Función que esitma los parámetros de un modelo lineal múltiple multivariado por medio del *Boostrap* o *Jackknife*. Con el método *Bootstrap* se estiman parámetros, se hacen pruebas de hipótesis y se construyen intervalos de confianza. Por *Jackknife* se estiman los coeficientes y se construyen intervalos de cofnianza.

A.2.1. Argumentos

Y	matriz con las variables respuesta
X	matriz con las variables explicativas
alpha	nivel de significancia, por defecto 0,05
Boot	número de muestras *Bootstrap*, por defecto 500
Jackk	TRUE para realizar las estimaciones por *Jackknife*
type	tipo de intervalo de confianza, por defecto "*per*"
na.action	omitir valores faltantes, por defecto FALSE

A.2.2. Salida

betas	p matrices de los coeficientes con sus respectivos $p - valor$ e IC
Sigma	matriz de varianzas y covarianzas
sdJackk	error estándar *Jackknife*
sdBoot	error estándar *Bootstrap*
Residuals	residuales del modelo

B. Pseudoalgoritmos

B.1. Estimación de los coeficientes por *Bootstrap*

Para estimar los valores de la matriz $\widehat{\boldsymbol{\beta}}^*$ se deben seguir los siguientes pasos

1. Estimar la matriz $\widehat{\boldsymbol{\beta}}$ por MCO con la muestra original.
2. Calcular las estimaciones de $\hat{\mathbf{Y}}$.
3. Estimar la matriz de errores como $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$.
4. Calcular una muestra de tamaño n con remplazo a partir de los errores.
5. Encontrar la nueva matriz $\mathbf{Y}_r^*$ con $\hat{\mathbf{Y}}_i + \hat{\boldsymbol{\varepsilon}}_{ir}$, denominados valores **Bootstrap** de $\mathbf{Y}$.
6. Calcular $\widehat{\boldsymbol{\beta}}_r^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_r^*$.

Repetir los pasos del 4 al 6 r veces y promediar cada uno de los valores de las matrices $\widehat{\boldsymbol{\beta}}_r^*$.

B.2. Estimación de los coeficientes por *Jackknife*

Para la estimación con *Jackknife* se deben seguir los siguientes pasos

1. Eliminar la *í-esima* observación de los datos.
2. Estimar los coeficientes de la matriz $\widehat{\boldsymbol{\beta}}$.
3. Repetir paso 1 y 2 n veces.

Promediar cada uno de los valores estimados para obtener la matriz $\widehat{\boldsymbol{\beta}}^J$.

C. Definición de las variables

Variable	Definición
Debilidad	Variable sobre la percepción de debilidad en los estudiantes
Motivación	Variable sobre la percepción de motivación en los estudiantes
Colegio	Estudiante al que pertenece el estudiante
Edad	Edad del estudiante
Sexo	Sexo del estudiante
Recursos	Si comparte o no los recursos del hogar
Onces	Si lleva o no onces al colegio
Compartir.Familia	El tiempo que suelen compartir con su familia

Tabla C.1: *Definición variables del modelo*