Propuesta metodológica para imputar valores no influyentes en modelos de regresión lineal múltiple con información incompleta

José Alfredo Jiménez Moscoso

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Matemáticas y Estadística
Santafé de Bogotá, Colombia

1999

Propuesta metodológica para imputar valores no influyentes en modelos de regresión lineal múltiple con información incompleta

José Alfredo Jiménez Moscoso

Tesis presentada como requisito parcial para optar al título de:

Maestría en Estadística

Director: MSc. Luis Francisco Rincón

Profesor Asociado

Línea de Investigación:

Modelos lineales

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Matemáticas y Estadística
Santafé de Bogotá, Colombia
1999

HOJA DE ACEPTACIÓN

LUIS FRANCISCO RINCÓN
Director de Tesis

ADRIANA PÉREZ
Jurado

LUIS ALBERTO LÓPEZ
Jurado

DAVID OSPINA
Director del Postgrado

Santafé de Bogotá, D.C., diciembre de 1999

DEDICATORIAS

A *Dios* todopoderoso de quien he recibido fortaleza a lo largo de mi vida.

Con amor, orgullo y cariño a mi hija *Camila Andrea*, quien fue mi inspiración en la culminación de este trabajo.

Con amor y cariño a *Nathaly C.*, quien estuvo a mi lado durante estos años de estudio.

A mis padres *María D.* y *Antonio*.

Con respeto y admiración a mi hermano *Sidney G.*

AGRADECIMIENTOS

Deseo manifestar mi gratitud al profesor Luis F. Rincón, por sus conocimientos y valiosa colaboración en el desarrollo de este trabajo.

A mis amigos quienes me apoyaron en los momentos más difíciles, en especial a *Mauricio*, *Ricardo*, *Fernando*, *Angélica* y *Luz Denny*.

Tabla de contenido

Resumen

Propuesta metodológica para imputar valores no influyentes en modelos de regresión lineal múltiple con información incompleta. Tesis 1999. José Alfredo Jiménez M. Facultad de Ciencias. Departamentos de Matemáticas y Estadística. Universidad Nacional de Colombia.

En esta tesis se presenta un método para imputar datos faltantes en un modelo de regresión lineal múltiple; estos datos se establecerán de tal manera que no sean influyentes, es decir, que el impacto (cambio) ejercido sobre la suma de los cuadrados residuales del modelo sea pequeño. La estimación de los parámetros del modelo de regresión lineal múltiple se calcula mediante el método de mínimos cuadrados ordinarios.

Una generalización de la estadística DFBeta se establece en el Teorema 4.1 para cuantificar el impacto de las observaciones imputadas en la estimación de mínimos cuadrados del modelo de regresión lineal múltiple. Además, el Teorema 4.5 presenta una nueva expresión para la estadística Q_k, con la cual se cuantifica el impacto que ejercen las observaciones imputadas en el modelo sobre la suma de cuadrados de los residuos.

Palabras clave: modelos lineales, mínimos cuadrados, formas cuadráticas, observaciones influyentes, estadística DFBeta, estadística Q_k.

Abstract

Methodological proposal to impute non-influential values in multiple linear regression models with incomplete information. Thesis 1999. José Alfredo Jiménez M. Faculty of Sciences. Departments of Mathematics and Statistics. Universidad Nacional de Colombia.

This thesis presents a method to impute missing data in a multiple linear regression model; these data are established in such a way that they are non-influential, that is, that the impact (change) exerted on the sum of the residual squares of the model is small. The estimation of the parameters of the multiple linear regression model is compute by the ordinary least squares method.

A generalization of the DFBeta statistic is establishes in Theorem 4.1 to quantify the impact of the imputed observations on the least squares estimation of the multiple linear regression model. In addition, Theorem 4.5 presents a new expression for the Q_k statistic, which quantifies the impact of the imputed observations in the model exert on the sum of squares of the residuals.

Key words: Linear models, Least squares, Quadratic forms, influential observations, DFBeta statistics, Q_k Statistics.

Introducción

En este trabajo se plantea una alternativa para imputar valores no influyentes en modelos de regresión lineal múltiple $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ con pérdida de información en la variable respuesta, es decir, se propone una metodología que permita completar la información faltante y luego estimar el modelo, usando algún método de estimación, por ejemplo Mínimos Cuadrados Ordinarios (MCO), Mínimos Cuadrados Ponderados (MCP), Estimación Máxima Verosimilitud (EMV), etc. La propuesta se basa en el análisis de la influencia e impacto de la información imputada en el ajuste del modelo, utilizando estadísticas diseñadas para detectar observaciones atípicas o influyentes y el método de estimación vía mínimos cuadrados ordinarios (MCO).

Entre los métodos descritos en la literatura estadística para estimar modelos de regresión lineal con información incompleta, asignando valores en la información faltante, no se conoce uno en el cual se investigue de manera directa la influencia e impacto que pueda generar la información imputada en la estimación de los parámetros y en la suma de cuadrados de residuales. Resulta así una propuesta novedosa, en la cual, el criterio para rellenar la información faltante sea el análisis de la influencia de los valores imputados. Los resultados teóricos obtenidos en el desarrollo de este trabajo aparecen resumidos en los **Teoremas 4.1, 4.3** y **4.5**.

1 Objetivos

1.1. Objetivo General

Diseñar una propuesta metodológica para imputar valores no influyentes en modelos de regresión lineal múltiple $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ con información incompleta, cuando los datos faltantes se presentan en la variable respuesta, usando como método de estimación los mínimos cuadrados ordinarios, y bajo el supuesto de que se conozca la distribución de los residuales. La propuesta se desarrolla completamente para el supuesto de que los residuales siguen una distribución normal estándar con media cero y varianza constante, es decir, $\epsilon_i \sim N(0, \sigma^2)$.

1.2. Objetivos Específicos

Considerando el modelo de regresión múltiple particionado

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} + \begin{bmatrix} \vec{\gamma}_1 \\ \vec{\gamma}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}$$

donde el bloque $\vec{Y}_1$ representa la información incompleta. Se pretende

1.2.1. Establecer las relaciones g_1, g_2 que satisfagan

$$DFBeta(\vec{Y}_1) = g_1(X, \vec{\gamma}) \qquad \text{y} \qquad Q_k = g_2(\vec{\epsilon}_1, \vec{\gamma}_1, X_1, X)$$

donde $\vec{\gamma} = \begin{bmatrix} \vec{\gamma}_1 \\ \vec{\gamma}_2 \end{bmatrix}$ es un vector de constantes, $DFBeta(\vec{Y}_1)$ es la generalización de la estadística $DFBeta(i)$ presentada en Belsley et al. (1980) y Q_k es la estadística presentada en Draper and John (1981).

1.2.2. Bajo el supuesto $\epsilon_i \sim N(0, \sigma^2)$ determinar la distribución de probabilidad de las estadísticas $DFBeta(\vec{Y}_1)$ y Q_k.

1.2.3. Utilizar la distribución de probabilidad de la estadística Q_k, para establecer el criterio de influencia de las k-observaciones imputadas.

2 Planteamiento

En el desarrollo del trabajo se considera el bloque $\vec{Y_1}$ del modelo

$$\begin{bmatrix} \vec{Y_1} \\ \vec{Y_2} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon_1} \\ \vec{\epsilon_2} \end{bmatrix}$$

como información incompleta en la variable Y. Conociendo la distribución de los residuales interesa imputar valores en este bloque y que cumplan la relación $\hat{Y_1} = g(X_1, \hat{\beta}, \vec{\gamma})$ para un vector $\vec{\gamma}$ seleccionado en una región tal que dado un nivel de significancia α, para las observaciones en $\vec{Y_1}$ se satisfaga

$$P(|Q_k| \leq m) = 1 - \alpha \tag{2.1}$$

y además, se busca con esta metodología que la información imputada no sea influyente en la estimación del modelo, lo cual se verifica utilizando la estadística Q_k, que mide el cambio en la suma de cuadrados de los residuales.

Los resultados en la estimación del modelo después de usar la metodología de imputación propuesta, se comparan con los obtenidos con el análisis de caso completo (CC), con los obtenidos al imputar la media de los $\vec{Y_2}$. Con el método propuesto se llega a los resultados obtenidos en Bartlett (1937), si el vector $\vec{\gamma}$ es el vector de ajustes definido en Draper and John (1981), es decir, cuando los nuevos residuales en la información imputada son cero, o a los resultados obtenidos mediante el algoritmo de estimación E.M. desarrollado por Dempster et al. (1977), si se establece como criterio de convergencia la condición dada en (2.1).

3 Marco Teórico

3.1. Estadísticas para detectar puntos influyentes y outliers

Dado el modelo de regresión lineal múltiple

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} \tag{3.1}$$

donde,

$\vec{Y}$ es el vector de respuestas $n \times 1$.

X es la matriz de constantes de dimensión $n \times r$, con $r = p + 1$, el rango de la matriz X.

$\vec{\beta} = (\beta_0, \beta_1, ...\beta_p)'$ es el vector de parámetros de dimensión $r \times 1$.

$\vec{\epsilon}$ vector $n \times 1$ de variables aleatorias normales independientes con esperanzas $E(\vec{\epsilon}) = \vec{0}$ y matriz de varianza-covarianza $\text{Var}(\vec{\epsilon}) = \sigma^2 I_n$.

La literatura estadística cuando emplea la técnica de modelos de regresión para un análisis estadístico, realiza un estudio de los residuales para detectar las observaciones que más influyen en la suma de cuadrados de los residuales. La base teórica de la propuesta se centra en el estudio de las estadísticas diseñadas para analizar y detectar observaciones atípicas o influyentes, particularmente la aplicación de las estadísticas Q_k y $DFBeta(i)$ presentadas a continuación.

3.1.1. Estadística Q_k

Para el modelo de Regresión (3.1), Draper and John (1981) desarrollaron una metodología para detectar un grupo de k observaciones influyentes o atípicas, equivalente a la propuesta por Bartlett (1937), citada en Little and Rubin (1986), para estimar los parámetros del modelo de regresión lineal cuando existen observaciones faltantes en la variable respuesta.

En la propuesta de Draper and John (1981) se analiza el modelo (3.1) particionado

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 & I \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} \vec{\beta} \\ \vec{\gamma} \end{bmatrix} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix} \tag{3.2}$$

donde $\vec{Y}_1$ es el bloque conformado por las observaciones consideradas atípicas. Las estimaciones $\vec{\beta}$ y $\vec{\gamma}$ del modelo (3.2) están definidas por:

$$\hat{\beta} = (X_2' X_2)^{-1} X_2' \vec{Y}_2 \qquad \text{y} \qquad \hat{\gamma} = (I - H_{11})^{-1} \hat{\epsilon}_1 \tag{3.3}$$

donde $H_{ij} = X_i (X'X)^{-1} X_j'$, es submatriz de la matriz $H = X(X'X)^{-1}X'$ la cual se conoce como "Matriz Hat", Tukey (1977); $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ y el cambio en la suma de cuadrados de residuales está dado por la estadística

$$Q_k = \hat{\epsilon}_1' (I - H_{11})^{-1} \hat{\epsilon}_1. \tag{3.4}$$

En resumen, el método descrito permite detectar el grupo de observaciones atípicas en base al cambio en la suma de cuadrados de residuales, lo cual se cuantifica con la estadística Q_k. Sin embargo, el método no mide la influencia de estas observaciones en la estimación de los parámetros.

3.1.2. Estadística DFBeta(i)

Entre las metodologías aplicadas en la estimación mínimos cuadrados ordinarios para detectar específicamente la influencia que un grupo de observaciones seleccionadas ejer-

ce en la estimación de los parámetros, la estadística más reconocida al respecto es la estadística $DFBeta(i)$ la cual se presenta en Belsley et al. (1980) como una medida de diagnóstico en regresión. Esta estadística mide la influencia que ejerce la i-ésima observación sobre el estimador de mínimos cuadrados del vector $\vec{\beta}$ asociado al modelo de regresión lineal (3.1). Para la i-ésima observación esta estadística se obtiene a partir de la expresión:

$$DFBeta(i) = \frac{\hat{\epsilon}_i}{1 - h_{ii}} c_i, \qquad 1 \leq i \leq n \tag{3.5}$$

con c_i la i-ésima línea de la matriz $C = (X'X)^{-1}X'$, $\hat{\epsilon}_i = \vec{Y}_i - \hat{Y}_i$ y h_{ii} el i-ésimo elemento en la diagonal de la matriz $H = X(X'X)^{-1}X'$.

Esta estadística es generalizable y permite detectar la influencia que un grupo de observaciones ejerce en la estimación de los parámetros en un modelo de regresión lineal simple Rincón and López (1997).

3.1.3. Distribución de probabilidad de la estadística $\hat{\gamma}_i$

En Espinosa (1998) para modelos de regresión lineal simple, se expresa la estadística $\hat{\gamma}_i$ como

$$\hat{\gamma}_i = \frac{\hat{\epsilon}_i}{1 - h_{ii}}, \qquad 1 \leq i \leq n$$

y bajo el supuesto de que los residuales siguen una distribución normal estándar con media cero y varianza constante, es decir, $\epsilon_i \sim N(0, \sigma^2)$, se obtiene que

$$\hat{\epsilon}_i \sim N(0, \sigma^2(1 - h_{ii}))$$

concluyendo finalmente que

$$\hat{\gamma}_i \sim N\left(0, \frac{\sigma^2}{1 - h_{ii}}\right) \tag{3.6}$$

y puesto que $S^2 = \frac{SSE}{n-2}$ es el estimador insesgado de σ^2 se obtiene que

$$\frac{\hat{\gamma}_i\sqrt{1-h_{ii}}}{S} = T_1 \tag{3.7}$$

la cual se distribuye t-student con $(n-2)$ grados de libertad.

3.1.4. Distribución de probabilidad de las componentes de la estadística DFBeta(i)

En Espinosa (1998) para modelos de regresión lineal simple se expresan las componentes de la estadística $DFBeta(i)$ como

$$\hat{\beta}_0(i) - \hat{\beta}_0^*(i) = \hat{\gamma}_i\left(\frac{1}{n} - \bar{x}\nu_i\right) \qquad \text{y} \qquad \hat{\beta}_1(i) - \hat{\beta}_1^*(i) = \hat{\gamma}_i\nu_i$$

donde $\nu_i = \frac{x_i - \bar{x}}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$ y utilizando (3.6), la distribución de $\hat{\gamma}_i$, concluye que

$$\hat{\beta}_0(i) - \hat{\beta}_0^*(i) \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\left[\frac{1}{n} - \bar{x}\nu_i\right]^2\right)$$
$$\hat{\beta}_1(i) - \hat{\beta}_1^*(i) \sim N\left(0, \frac{\sigma^2\,\nu_i^2}{1-h_{ii}}\right)$$

puesto que S^2 es el estimador insesgado de σ^2 se obtiene que

$$\frac{\hat{\beta}_0(i) - \hat{\beta}_0^*(i)}{\sqrt{\frac{S^2}{1-h_{ii}}\left[\frac{1}{n} - \nu_i\bar{x}\right]^2}} = \frac{\hat{\beta}_0(i) - \hat{\beta}_0^*(i)}{\frac{S}{\sqrt{1-h_{ii}}}\left[\frac{1}{n} - \nu_i\bar{x}\right]} = T_2 \tag{3.8}$$

$$\frac{\hat{\beta}_1(i) - \hat{\beta}_1^*(i)}{\sqrt{\frac{S^2\,\nu_i^2}{1-h_{ii}}}} = \frac{\hat{\beta}_1(i) - \hat{\beta}_1^*(i)}{\frac{S\,\nu_i}{\sqrt{1-h_{ii}}}} = T_3 \tag{3.9}$$

En dicho trabajo se llega a la conclusión final de que las estadísticas T_1, T_2, T_3 son estadísticas equivalentes; es decir, $T_1 = T_2 = T_3$ con la misma distribución t-students con $n-2$ grados de libertad.

4 Desarrollo de la propuesta

4.1. Generalización de la Estadística DFBeta

Para el modelo de regresión lineal múltiple

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}$$

definido en (3.1), usando el método de estimación mínimos cuadrados se obtiene el estimador $\hat{\beta}$ de los parámetros, los valores estimados de $\hat{Y}$, los errores estimados de $\hat{\epsilon}$ y la suma de cuadrados de los residuales (SSE) según las siguientes expresiones

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'\vec{Y} \\
\hat{Y} &= X\hat{\beta} = X(X'X)^{-1}X'\vec{Y} = H\vec{Y} \quad \text{con} \quad H = X(X'X)^{-1}X' \\
\hat{\epsilon} &= \vec{Y} - \hat{Y} = \vec{Y} - H\vec{Y} = (I-H)\vec{Y} \\
SSE &= \hat{\epsilon}'\hat{\epsilon} = [(I-H)\vec{Y}]'(I-H)\vec{Y} = \vec{Y}'(I-H)\vec{Y}
\end{aligned}$$

Cuando se plantea el modelo

$$\vec{Y}^* = X\vec{\beta}^* + \vec{\epsilon}^* \tag{4.1}$$

siendo $\vec{Y}^* = \vec{Y} + \vec{\gamma}$, para $\vec{\gamma}$ un vector de constantes conocido, interesa establecer las expresiones de los nuevos estimadores, en función de $\vec{\gamma}$, $\vec{Y}$ y de los estimadores descritos para el modelo (3.1).

El estimador del vector $\vec{\beta}^*$ por el método de mínimos cuadrados ordinarios corresponde a la expresión

$$\hat{\beta}^* = (X'X)^{-1}X'\vec{Y}^*$$

reemplazando $\vec{Y}^*$ se tiene

$$\begin{aligned}\hat{\beta}^* =&(X'X)^{-1}X'(\vec{Y}+\vec{\gamma})\\ =&(X'X)^{-1}X'\vec{Y}+(X'X)^{-1}X'\vec{\gamma}\\ =&\hat{\beta}+(X'X)^{-1}X'\vec{\gamma},\end{aligned}$$

de donde se concluye que

$$\hat{\beta}-\hat{\beta}^* = -(X'X)^{-1}X'\vec{\gamma}. \tag{4.2}$$

De la misma forma, el nuevo vector de predicciones $\hat{Y}^*$ se obtiene según

$$\begin{aligned}\hat{Y}^* =&X\hat{\beta}^*\\ =&X(\hat{\beta}+(X'X)^{-1}X'\vec{\gamma})\\ =&\hat{Y}+X(X'X)^{-1}X'\vec{\gamma}\\ =&\hat{Y}+H\vec{\gamma}\end{aligned} \tag{4.3}$$

Con la misma metodología se obtiene el vector de errores estimado para el modelo (4.1) según

$$\begin{aligned}\hat{\epsilon}^* =&\vec{Y}^*-\hat{Y}^*\\ =&(\vec{Y}+\vec{\gamma})-(\hat{Y}+H\vec{\gamma})\\ =&(\vec{Y}-\hat{Y})+(\vec{\gamma}-H\vec{\gamma})\\ =&\hat{\epsilon}+(I-H)\vec{\gamma}\end{aligned} \tag{4.4}$$

Conociendo los nuevos errores es posible calcular la nueva suma de cuadrados de los residuales (SSE^*), según

$$
\begin{aligned}
SSE^* &= (\vec{Y}^*)'(I-H)\vec{Y}^* \\
&= (\vec{Y}+\vec{\gamma})'(I-H)(\vec{Y}+\vec{\gamma}) \\
&= \vec{Y}'(I-H)\vec{Y} + \vec{Y}'(I-H)\vec{\gamma} + \vec{\gamma}'(I-H)\vec{Y} + \vec{\gamma}'(I-H)\vec{\gamma}
\end{aligned}
$$

como

$$\vec{Y}'(I-H)\vec{\gamma} = \vec{\gamma}'(I-H)\vec{Y}.$$

Se tiene que

$$
\begin{aligned}
SSE^* &= SSE + 2\vec{\gamma}'(I-H)\vec{Y} + \vec{\gamma}'(I-H)\vec{\gamma} \\
&= SSE + \vec{\gamma}'(I-H)[2\vec{Y}+\vec{\gamma}] \qquad (4.5)
\end{aligned}
$$

Sin pérdida de generalidad, los resultados anteriores se pueden particularizar para el modelo (3.1) particionado

$$\begin{bmatrix}\vec{Y}_1\\ \vec{Y}_2\end{bmatrix} + \begin{bmatrix}\vec{\gamma}_1\\ \vec{0}\end{bmatrix} = \begin{bmatrix}X_1\\ X_2\end{bmatrix}\vec{\beta}^* + \begin{bmatrix}\vec{\epsilon}_1^*\\ \vec{\epsilon}_2^*\end{bmatrix} \qquad (4.6)$$

con $\vec{Y}_1$ de dimensión $k \times 1, k < n$ y en tal caso

$$
\begin{aligned}
\hat{\beta}^* =& \hat{\beta} + (X'X)^{-1}X'\vec{\gamma} \\
=& \hat{\beta} + (X'X)^{-1}\begin{bmatrix}X_1' & X_2'\end{bmatrix}\begin{bmatrix}\vec{\gamma}_1\\ \vec{0}\end{bmatrix} \\
\hat{\beta} - \hat{\beta}^* =& -(X'X)^{-1}X_1'\vec{\gamma}_1 \qquad (4.7)
\end{aligned}
$$

Bajo esta partición la ecuación (4.4) se expresa como

$$\begin{bmatrix}\hat{\epsilon}_1^*\\ \hat{\epsilon}_2^*\end{bmatrix} = \begin{bmatrix}\hat{\epsilon}_1\\ \hat{\epsilon}_2\end{bmatrix} + \begin{bmatrix}I-H_{11} & -H_{12}\\ -H_{21} & I-H_{22}\end{bmatrix}\begin{bmatrix}\vec{\gamma}_1\\ \vec{0}\end{bmatrix}$$

con $H_{ij} = X_i(X'X)^{-1}X_j'$.

De tal manera que el vector $\vec{\gamma}_1$ que hace $\hat{\epsilon}_1^* = 0$ está dado por

$$\hat{\gamma}_1 = -(I - H_{11})^{-1}\hat{\epsilon}_1 \tag{4.8}$$

expresión que reemplazada en la ecuación (4.7) proporciona la generalización de la estadística $DFBeta(\vec{Y}_1)$ según

$$DFBeta(\vec{Y}_1) = (X'X)^{-1}X_1'(I - H_{11})^{-1}\hat{\epsilon}_1 \tag{4.9}$$

Nótese que $DFBeta(\vec{Y}_1)$ es un vector de dimensión $r \times 1$ y como expresión mide el efecto que tienen los k registros del bloque $\vec{Y}_1$, en la estimación vía mínimos cuadrados ordinarios en cada una de las componentes del vector de parámetros $\vec{\beta}$.

El anterior resultado se puede resumir en el siguiente teorema.

Teorema 4.1. *En un modelo de regresión lineal $Y = X\vec{\beta} + \vec{\epsilon}$, particionado como:*

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}$$

donde $\vec{Y}_1$ es de dimensión $k \times 1$, si los registros que conforman dicho bloque se eliminan, entonces el cambio en los parámetros estimados se calcula por:

$$DFBeta(\vec{Y}_1) = (X'X)^{-1}X_1'(I - H_{11})^{-1}\hat{\epsilon}_1$$

donde $H_{11} = X_1(X'X)^{-1}X_1'$. El vector $DFBeta(\vec{Y}_1)$ es de dimensión $r \times 1$.

4.1.1. Distribución de probabilidad de la estadística $\mathbf{DFBeta(\tilde{Y}_1)}$

Para establecer la distribución de la estadística $DFBeta$ generalizada se debe conocer la distribución de la estadística $\vec{\gamma}_1$.

En su construcción Rincón (1999) utiliza el siguiente teorema que se enuncia sin demostración

Teorema 4.2. *Si* $E(X) = \mu$, $Var(X) = \Sigma$ *entonces para* $Y = TX$ *se cumple que* $E(Y) = T\mu$ *y* $Var(Y) = T\Sigma T'$.

Para el modelo particionado (4.6) y bajo el supuesto de normalidad de los residuales

$$\hat{\epsilon}_1 \sim N(0, \sigma^2(I - H_{11}))$$

se muestra que $\hat{\gamma}_1$ definido en (4.8) satisface

$$\hat{\gamma}_1 \sim N(0, \sigma^2(I - H_{11})^{-1}) \tag{4.10}$$

es decir que cada una de las componentes $\gamma_i,\ i = 1, ..., k$ de $\hat{\gamma}_1$ se distribuye según

$$\hat{\gamma}_i \sim N(0, \sigma^2 H_i) \tag{4.11}$$

donde H_i es el i-ésimo elemento de la diagonal de la matriz $(I - H_{11})^{-1}$. Para obtener finalmente con $\hat{\sigma}^2 = \frac{SSE}{n-r}$ que las estadísticas T_i definidas en (4.12) se distribuyen t-Student con $(n - r)$ grados de libertad.

$$\frac{\hat{\gamma}_i}{S\sqrt{H_i}} = T_i \sim T_{(n-r)} \tag{4.12}$$

Conocida la distribución de $\hat{\gamma}_1$ obtenida en (4.10) y reescribiendo la estadística $DFBeta(\vec{Y}_1)$ como

$$DFBeta(\vec{Y}_1) = -(X'X)^{-1}X_1'\hat{\gamma}_1$$

la aplicación del Teorema 4.2 proporciona

$$E(DFBeta(\vec{Y}_1)) = -(X'X)^{-1}X_1'E(\hat{\gamma}_1) = \vec{0}$$

y

$$\begin{aligned} Var(DFBeta(\vec{Y}_1)) &= (X'X)^{-1}X_1'Var(\hat{\gamma}_1)[(X'X)^{-1}X_1']' \\ &= \sigma^2 C(I - H_{11})^{-1}C' \end{aligned}$$

con $C = (X'X)^{-1}X_1'$ para establecer que

$$DFBeta(\vec{Y_1}) \sim N(\vec{0}, \sigma^2 C(I - H_{11})^{-1}C'). \tag{4.13}$$

En particular denotando por M_j el j-ésimo elemento de la diagonal de $C(I - H_{11})^{-1}C'$ para cada $j = 1, ..., r$ la dimensión de la $DFBeta(\vec{Y_1})$, resulta que

$$DFBeta_j(\vec{Y_1}) \sim N(\vec{0}, \sigma^2 M_j). \tag{4.14}$$

Y usando el estimador insesgado de σ^2, $\hat{\sigma}^2 = \frac{SSE}{n-r}$ obtener que para cada $j = 1, 2, ..., r$

$$\frac{DFBeta_j(\vec{Y_1})}{S\sqrt{M_j}} = T_j \sim T_{(n-r)}. \tag{4.15}$$

Para el caso particular del modelo de regresión lineal simple, cuando $\vec{Y_1}$ consta de una única observación, las estadísticas definidas en (4.15) junto con sus distribuciones, coinciden con los resultados de Espinosa (1998), expuestos en las secciones 3.2.3 y 3.2.4.
Es conveniente mencionar que los valores de las estadísticas T_j definidos en (4.15) difieren para cada j en el modelo de regresión lineal múltiple, cuando $\vec{Y_1}$ tiene más de una observación, lo cual indica que el bloque $\vec{Y_1}$ puede ser influyente sobre alguno o algunos de los r parámetros y no influyente para los demás, concluyendo que la influencia de una observación sobre los parámetros estimados es diferente cuando la observación se considera aislada a cuando se considera como componente del bloque $\vec{Y_1}$ y la influencia se asigna al bloque.
El anterior resultado se puede resumir en el siguiente teorema.

Teorema 4.3. *La estadística $DFBeta(\vec{Y_1})$ tiene distribución Normal*

$$DFBeta(\vec{Y_1}) \sim N(\vec{0}, \sigma^2 C(I - H_{11})^{-1}C')$$

donde $H_{11} = X_1(X'X)^{-1}X_1'$ y $C = (X'X)^{-1}X_1'$.

4.2. Estadística Q_k

La expresión (4.5) que mide la variación en la suma de cuadrados de los residuales, en el modelo particionado (4.6) corresponde a

$$SSE^* = SSE + \vec{\gamma}_1'(I - H_{11})[2\vec{Y}_1 + \vec{\gamma}_1] \tag{4.16}$$

dado que $\hat{\epsilon}_1 = (I - H_{11})\vec{Y}_1$ y $\hat{\gamma}_1 = -(I - H_{11})^{-1}\hat{\epsilon}_1$ de la expresión anterior se obtiene que

$$Q_k = SSE - SSE^* = \hat{\epsilon}_1{}'(I - H_{11})^{-1}\hat{\epsilon}_1$$

expresión (3.4) presentada en Draper and John (1981), es decir, Q_k expresada en función de los residuales estimados en el bloque $\vec{Y}_1$, o se obtiene la expresión equivalente

$$Q_k = SSE - SSE^* = \hat{\gamma}_1{}'(I - H_{11})\hat{\gamma}_1, \tag{4.17}$$

nueva expresión de Q_k en función de $\hat{\gamma}_1$ que para los objetivos de este trabajo es más atractiva, ya que su distribución será utilizada para establecer el criterio que evalúe la influencia del bloque imputado $\vec{Y}_1$.

4.2.1. Distribución de probabilidad de la estadística Q_k

Para establecer la distribución de la estadística Q_k se enuncian sin demostración el teorema 4.4. y su corolario citados en Searle (1971).

Teorema 4.4. *Si X es un vector aleatorio de dimensión $n \times 1$ distribuido $N(\mu, V)$, entonces $X'AX \sim \chi^{2'}_{(\nu,\lambda)}$, con $\nu = rango(A)$ y parámetro de no centralidad $\lambda = \frac{1}{2}\mu' A\mu$, si y solo si AV es idempotente.*

Corolario 4.4.1. *Si $X \sim N(0, V)$ entonces $X'AX \sim \chi^2_\nu$ con $\nu = rango(A)$ si y solo si AV es idempotente.*

Dado que Q_k se puede expresar como la siguiente forma cuadrática

$$\hat{Q}_k = \hat{\gamma}_1{}'(I - H_{11})\hat{\gamma}_1$$

con

$$\hat{\gamma}_1 \sim N(\vec{0}, \sigma^2(I - H_{11})^{-1})$$

por la aplicación del corolario 4.4.1. se concluye que

$$\frac{\hat{Q}_k}{\sigma^2} \sim \chi^2_\nu \qquad \text{con} \qquad \nu = \text{rango}(I - H_{11}) = k$$

ya que $\frac{(I-H_{11})}{\sigma^2}\sigma^2(I - H_{11})^{-1}$ es idempotente.

El anterior resultado se puede resumir en el teorema que se anuncia a continuación

Teorema 4.5. *En un modelo de regresión lineal $Y = X\vec{\beta} + \vec{\epsilon}$, particionado como:*

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}$$

donde $\vec{Y}_1$ es de dimensión $k \times 1$, si los registros que conforman dicho bloque se eliminan, entonces el cambio en la suma de los residuales se distribuye Chi-cuadrado con k grados de libertad, es decir

$$\frac{\hat{Q}_k}{\sigma^2} = \frac{\hat{\gamma}_1'(I - H_{11})\hat{\gamma}_1}{\sigma^2} \sim \chi^2_{(k)}$$

donde $\hat{\gamma}_1 = -(I - H_{11})^{-1}\hat{\epsilon}_1$.

Por otra parte, si se reescribe la expresión (4.17), se tiene que

$$\begin{aligned} SSE &= SSE^* + Q_k \\ (n-r)\frac{SSE}{n-r} &= (n-k-r)\frac{SSE^*}{n-k-r} + Q_k \end{aligned} \tag{4.18}$$

Puesto que el estimador insesgado de σ^2 es $\hat{\sigma}^2 = S^2 = \frac{SSE}{n-r}$ y el estimador insesgado de $(\sigma^*)^2$ es $(\hat{\sigma}^*)^2 = (S^*_{(\vec{Y}_1)})^2 = \frac{SSE^*}{n-k-r}$ el cual denota la estimación usual de σ^2 sin el bloque

$\vec{Y_1}$ de observaciones; si se reemplaza en (4.18) se obtiene que

$$(n-r)S^2 = (n-k-r)(S^*_{(\vec{Y}_1)})^2 + Q_k \tag{4.19}$$
$$(n-r)\frac{S^2}{\sigma^2} = (n-k-r)\frac{(S^*_{(\vec{Y}_1)})^2}{\sigma^2} + \frac{Q_k}{\sigma^2}.$$

En esta última expresión cada uno de los términos se distribuyen Chi-cuadrado ya que

$$(n-r)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-r)} \qquad \frac{Q_k}{\sigma^2} \sim \chi^2_{(k)}$$
$$(n-k-r)\frac{(S^*_{(\vec{Y}_1)})^2}{\sigma^2} \sim \chi^2_{(n-k-r)}. \tag{4.20}$$

Luego, si se dividen las dos primeras expresiones y cada una por sus correspondientes grados de libertad se tiene que

$$\frac{\frac{Q_k}{k\sigma^2}}{\frac{(n-r)S^2}{(n-r)\sigma^2}} = \frac{Q_k}{k \cdot S^2} \sim F_{(k,n-r)} \tag{4.21}$$

Esta expresión no sirve para medir la influencia del bloque $\vec{Y_1}$ imputado, debido a que como se divide por el número de registros imputados hace que esta expresión sea muy pequeña y por lo tanto, enmascara algunos registros influyentes de dicho bloque.
Si se reescribe la ecuación (4.19) se tiene que

$$\hat{Q}_k = (n-r)S^2 - (n-r-k)(S^*_{(\vec{Y}_1)})^2$$
$$\frac{\hat{\gamma}_1'(I-H_{11})\hat{\gamma}_1}{S^2} = \frac{(n-r)S^2 - (n-r-k)(S^*_{(\vec{Y}_1)})^2}{S^2}. \tag{4.22}$$

Si en esta ecuación se toma $k = 1$, es decir, se considera que se quita una única observación, se obtiene que

$$\frac{\hat{\gamma}_i^2(1-h_{ii})}{S^2} = \frac{(n-r)S^2 - (n-r-1)(S^*_{(i)})^2}{S^2} \tag{4.23}$$

donde $(1-h_{ii})$ es el i-ésimo elemento de la diagonal de la matriz $(I-H_{11})$ y $(S^*_{(i)})^2$ denota la estimación usual de σ^2 sin la i-ésima observación.

Por la ecuación (4.18) se tiene que

$$\frac{\hat{\gamma}_i}{S\sqrt{H_i}} = T_i \sim T_{(n-r)}$$

donde H_i es el i-ésimo elemento de la diagonal de la matriz $(I - H_{11})^{-1}$, elevando esta expresión al cuadrado se obtiene

$$\frac{\hat{\gamma}_i^2}{S^2 H_i} = T_i^2 \sim T_{(n-r)}^2.$$

Esta última expresión es equivalente a la ecuación (4.23), por lo tanto, se tiene que

$$T_i^2 = \frac{(n-r)S^2 - (n-r-1)(S_{(i)}^*)^2}{S^2} \sim T_{(n-r)}^2 \tag{4.24}$$

Dado que la estadística Q_k se puede calcular como

$$Q_k = \sum_{i=1}^{k} \hat{\gamma}_i^2 (1 - h_{ii}) - 2 \sum_{i<j}^{k} \hat{\gamma}_i \hat{\gamma}_j h_{ij}$$

donde h_{ij} es el elemento de la i-ésima fila y j-ésima columna de la matriz $H_{11} = X_1(X'X)^{-1}X_1'$, si se divide esta expresión por S^2 y, además, se utilizan las relaciones (4.23) y (4.24), se obtiene

$$\frac{Q_k}{S^2} = \sum_{i=1}^{k} \frac{\hat{\gamma}_i^2 (1 - h_{ii})}{S^2} - 2 \sum_{i<j}^{k} \frac{\hat{\gamma}_i \hat{\gamma}_j}{S^2} h_{ij} = \sum_{i=1}^{k} T_i^2 - 2 \sum_{i<j}^{k} \frac{\hat{\gamma}_i \hat{\gamma}_j}{S^2} h_{ij},$$

es decir, que la estadística Q_k para un bloque se puede obtener como la suma de las estadísticas $Q_k(i)$ de las observaciones que conforman el bloque. Por consiguiente, se tiene que

$$\frac{Q_k}{S^2} = \frac{(n-r)S^2 - (n-r-k)(S_{(\vec{Y}_1)}^*)^2}{S^2} \sim T_{(n-r)}^2.$$

Luego, si se toma raíz cuadrada a esta expresión se obtiene

$$\frac{\sqrt{Q_k}}{S} \sim T_{(n-r)}. \tag{4.25}$$

Esta expresión (4.25) será útil para establecer el criterio que permita evaluar la influencia del bloque imputado $\vec{Y}_1$, en la variación porcentual de la suma de cuadrados residual, criterio que se presenta en la siguiente sección.

4.3. Definición de influencia y criterio para evaluarla

En este trabajo, la influencia de una o varias observaciones en la estimación del modelo (3.1) solo se considera, analizando el impacto que ella o ellas generan en la suma de cuadrados de los residuales (SSE). Por consiguiente, para alcanzar los objetivos de este trabajo, en la sección 4.3.1. se establece la definición de influencia, en la sección 4.3.2. se presenta el criterio para establecer cuando un bloque es influyente y en la sección 4.3.3. se presentan algunos ejemplos con diferente número de variables regresoras para ilustrar la aplicación del criterio expuesto en 4.3.2.

4.3.1. Definición de influencia

En la estimación del modelo (3.1) particionado, un bloque $\vec{Y}_1$ es influyente, si genera un cambio estadísticamente significativo en la SSE, cuando es eliminado.

La definición anterior requiere de un criterio que permita clasificar un bloque como influyente, analizando su aporte en la suma de cuadrados de los residuales. Una estadística que sirve a este propósito es la obtenida en (4.25) según la cual

$$\frac{\sqrt{Q_k}}{S} \sim T_{(n-r)}.$$

4.3.2. Criterio para evaluar la influencia

Para el modelo (3.1) particionado

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}$$

Interesa probar la hipótesis

H_0 : El bloque $\vec{Y}_1$ no es influyente

H_a : Algún registro del bloque $\vec{Y}_1$ es influyente

El estadístico de prueba propuesto en este trabajo para la prueba de la hipótesis a un nivel de significancia de α, está dado por la expresión

$$\mathcal{P} = \frac{\sqrt{Q_k}}{S}.$$

Donde $Q_k = SSE - SSE^*$ es el cambio en la suma de cuadrados de los residuales en el modelo (3.1) y $S = \sqrt{\frac{SSE}{n-r}}$ es la raíz cuadrada del cuadrado medio de los residuales del modelo (3.1) en el cual se quiere observar si el bloque $\vec{Y}_1$ es influyente. Esta estadística $\mathcal{P}$ se distribuye $t_{(n-r,\alpha/2)}$ y con este estadístico de prueba, la hipótesis nula H_0 se rechaza a un nivel de significancia $\alpha\,\%$ si

$$\mathcal{P} > t_{(n-r,\alpha/2)} \tag{4.26}$$

para $t_{(n-r,\alpha/2)}$ tal que $P\left(\mathcal{P} < t_{(n-r,\alpha/2)}\right) = 1 - \alpha$. Este criterio es equivalente a, calcular la probabilidad mínima para la cual el valor observado del estadístico de prueba es significativo, es decir, calcular el p valor de la prueba, a través de la expresión

$$p = P(t > \mathcal{P}) \tag{4.27}$$

y considerar que el bloque $\vec{Y}_1$ es influyente si $p < \alpha$. Para el caso particular del modelo de regresión lineal múltiple, cuando $\vec{Y}_1$ consta de una única observación el criterio de influencia definido en esta sección coincide con el criterio de influencia propuesto en Rincón (1999).

4.3.3. Ejemplos

En las siguientes secciones se relacionan tres ejemplos con diferente número de variables regresoras para ilustrar la aplicación del criterio expuesto anteriormente.

4.3.3.1. Modelo de regresión simple

Para la ilustración y aplicación del criterio de influencia definido en 4.3.2. se utilizan los datos dados en Mickey et al. (1967) citados en Draper and John (1981). Los datos se presentan en la tabla 4.1, junto con las estadísticas Q_k, la variación porcentual que cada observación genera en la suma de cuadrados residual (SSE), cuando es eliminada, la estadística $\mathcal{P}$ de cada observación y el p-valor del criterio definido 4.3.2. De este análisis individual sobre la influencia de cada observación, se concluye que es la 8a. observación la única clasificada como influyente según el criterio definido anteriormente.

Tabla 4.1: Datos de Mickey et al. (1967) con el análisis individual de influencia

OBS	X	Y	Q_k	**Variación**	$\mathcal{P}$	**valor de** p
1	42	57	88.10526	3.81642	0.85154	0.40508
2	10	83	259.80303	11.25377	1.46226	0.16001
3	10	83	259.80303	11.25377	1.46226	0.16001
4	11	84	192.53973	8.34016	1.25882	0.22334
5	11	86	139.63370	6.04845	1.07201	0.29714
6	7	113	133.44259	5.78027	1.04798	0.30780
7	26	71	108.37029	4.69423	0.94441	0.35682
8	17	121	968.56197	41.95477	2.82337	0.01086
9	15	102	85.66407	3.71067	0.83966	0.41154
10	9	91	82.01509	3.55261	0.82158	0.42151
11	12	105	78.93614	3.41924	0.80601	0.43021
12	20	94	47.91421	2.07548	0.62797	0.53750
13	11	102	21.68686	0.93940	0.42248	0.67742
14	9	96	14.97643	0.64873	0.35108	0.72939
15	18	93	12.35811	0.53531	0.31892	0.75327
16	8	104	10.72944	0.46476	0.29716	0.76957
17	11	100	6.74813	0.29231	0.23567	0.81621
18	15	95	4.33256	0.18767	0.18883	0.85223
19	10	100	2.07958	0.09008	0.13083	0.89729
20	10	100	2.07958	0.09008	0.13083	0.89729
21	20	87	0.12034	0.00521	0.03147	0.97522

En la tabla 4.2 se presenta la estadística $\mathcal{P}$ calculada para $\vec{Y}_1, \vec{Y}_2, \vec{Y}_3, \vec{Y}_4, \vec{Y}_5$ diferentes bloques que incluyen o excluyen el registro 8, único clasificado como influyente, en el análisis individual de influencia y que genera una variación porcentual de la suma de cuadrados de los residuales del 41.95 %. Para conformar los citados bloques se incluyó de manera forzosa el registro 8 en los bloques $\vec{Y}_1, \vec{Y}_2$ y $\vec{Y}_4$ seleccionando de manera aleatoria las 3 restantes componentes del bloque. En los bloques $\vec{Y}_3$ y $\vec{Y}_5$ se seleccionaron aleatoriamente las cuatro componentes que los conforman, resultando según lo anterior:

- El bloque $\vec{Y}_1$ conformado por los registros 8, 11, 21, 3.
- El bloque $\vec{Y}_2$ conformado por los registros 8, 7, 2, 19.
- El bloque $\vec{Y}_3$ conformado por los registros 3, 11, 15, 21.
- El bloque $\vec{Y}_4$ conformado por los registros 8, 9, 13, 17.
- El bloque $\vec{Y}_5$ conformado por los registros 1, 7, 15, 13.

Tabla 4.2: Compendio de p-valores para los bloques conformados en la prueba de la hipótesis nula H_0 : El bloque $\vec{Y}_i$ no es influyente

Bloque	$\mathcal{P}$	**Valor de** p	**Decisión**
$\vec{Y}_1$	3.280285	0.003936	Influyente
$\vec{Y}_2$	3.319434	0.003604	Influyente
$\vec{Y}_3$	1.700168	0.105411	No es influyente
$\vec{Y}_4$	2.985039	0.007610	Influyente
$\vec{Y}_5$	1.377395	0.184401	No es influyente

Se concluye de los p-valores calculados según la estadística $\mathcal{P}$ y presentados en la tabla 4.2 que ella bien sirve a nuestros propósitos ya que detecta a un nivel de significancia

del 1 %, los tres bloques influyentes $\vec{Y}_1, \vec{Y}_2$ y $\vec{Y}_4$ en los cuales se incluyó el único registro influyente.

4.3.3.2. Modelo de regresión múltiple con dos variables

Como ilustración y aplicación del criterio de influencia definido en 4.3.2. se usan los datos muestrales citados en Eld et al. (1954) en los cuales sobre una investigación en suelos calcáreos se deseaba conocer el efecto de las fuentes de donde el cultivo de maíz tomaba el fósforo, para lo cual se midieron las siguientes variables

X_1 : concentración de fósforo inorgánico mediante el método de Bray y Kurtz para estimar el "fósforo aprovechable por la planta" (ppm) (PINORGBK)

X_2 : concentración de fósforo orgánico soluble en K_2CO_3 e hidrolizado mediante hipobromito (PORGKH)

Y : fósforo disponible por la planta sembrada en un suelo con temperaturas de $20°C$ (ppm). Esta variable se trata como una variable dependiente de las variables independientes X_1 y X_2.

En la tabla 4.3 se presentan los datos referidos en este ejemplo junto con la estadística Q_k de cada observación, la respectiva variación porcentual que está genera en la suma de cuadrados residual (SSE) cuando es eliminada, la estadística $\mathcal{P}$ de cada observación y su p-valor correspondiente. En el análisis individual de influencia resulta ser el registro número 17 el único influyente según el criterio definido en 4.3.2.

Tabla 4.3: Datos de Eld et al. (1954) con el análisis individual de influencia

OBS	X_1	X_2	Y	Q_k	**Variación**	$\mathcal{P}$	**valor de** p
1	0.4	53	64	8.0782	0.1259	0.1374	0.8925
2	0.4	23	60	1.3343	0.0208	0.0559	0.9562
3	3.1	19	71	73.6730	1.1486	0.4151	0.6840
4	0.6	34	61	0.6105	0.0095	0.0378	0.9704
5	4.7	24	54	193.3561	3.0146	0.6725	0.5115
6	1.7	65	77	269.7407	4.2055	0.7942	0.4394
7	9.4	44	81	18.0710	0.2817	0.2056	0.8399
8	10.1	31	93	283.1688	4.4149	0.8138	0.4285
9	11.6	29	93	206.9421	3.2264	0.6957	0.4973
10	12.6	58	51	1271.3244	19.8213	1.7243	0.1052
11	10.9	37	76	9.3981	0.1465	0.1483	0.8841
12	23.1	46	96	35.8066	0.5583	0.2894	0.7763
13	23.1	50	77	711.7091	11.09630	1.2901	0.2165
14	21.6	44	93	37.0250	0.5773	0.2943	0.7726
15	23.1	56	95	64.9544	1.0127	0.3898	0.7022
16	1.9	36	54	86.7416	1.3524	0.4504	0.6589
17	26.8	58	168	4312.6515	67.2387	3.1758	0.0063
18	29.9	51	99	303.9667	4.7392	0.8431	0.4124

En la tabla 4.4 se presenta la estadística $\mathcal{P}$ calculada para $\vec{Y}_1, \vec{Y}_2, \vec{Y}_3, \vec{Y}_4, \vec{Y}_5$ diferentes bloques que incluyen o excluyen el registro 17, único clasificado como influyente en el análisis individual de influencia, con una variación porcentual de la suma de cuadrados de los residuales del 67.24 %. Los bloques se conformaron siguiendo la misma metodología del ejemplo 4.3.3.1, incluyendo forzosamente el registro influyente en los bloques $\vec{Y}_1$, $\vec{Y}_2$ y $\vec{Y}_4$

- El bloque $\vec{Y}_1$ esta conformado por los registros 10, 2, 13, 17.
- El bloque $\vec{Y}_2$ esta conformado por los registros 8, 7, 3, 17.
- El bloque $\vec{Y}_3$ esta conformado por los registros 5, 11, 18, 13.
- El bloque $\vec{Y}_4$ esta conformado por los registros 3, 14, 17, 18.

- El bloque $\vec{Y}_5$ esta conformado por los registros 8, 6, 15, 12.

Tabla 4.4: Compendio de p-valores para los bloques conformados en la prueba de la hipótesis nula H_0 : El bloque $\vec{Y}_i$ no es influyente

Bloque	$\mathcal{P}$	**Valor de** p	**Decisión**
$\vec{Y}_1$	3.8375	0.00161	Influyente
$\vec{Y}_2$	3.3110	0.00475	Influyente
$\vec{Y}_3$	1.6880	0.11208	No es influyente
$\vec{Y}_4$	3.3250	0.00462	Influyente
$\vec{Y}_5$	1.2364	0.23532	No es influyente

Se concluye de los p-valores calculados según la estadística $\mathcal{P}$ y presentados en la tabla 4.4 que ella bien sirve a nuestros propósitos ya que detecta a un nivel de significancia del 1 %, los tres bloques influyentes $\vec{Y}_1, \vec{Y}_2$ y $\vec{Y}_4$ en los cuales se incluyó el único registro influyente.

4.3.3.3. Modelo de regresión múltiple con tres variables

Como ilustración y aplicación del criterio de influencia definido en 4.3.2. se usan los datos citados en Weisberg (1980) presentados en Cook and Weisberg (1982) en los cuales se buscaba estudiar la cantidad de droga retenida en el hígado de un ratón, para lo cual se midieron las siguientes variables

X_1 : Peso del ratón en gramos.

X_2 : Peso del hígado del ratón en gramos.

X_3 : Dosis relativa.

Y : Cantidad de droga retenida en el hígado del ratón. Esta variable se trata como una variable dependiente de las variables independientes $X_1,\ X_2$ y X_3.

En la tabla 4.5 se presentan los datos referidos en este ejemplo junto con la estadística Q_k de cada observación, la respectiva variación porcentual que está genera en la suma de cuadrados residual (SSE) cuando es eliminada, la estadística $\mathcal{P}$ de cada observación y su p-valor correspondiente. En el análisis individual de influencia resultan ser los registros 1 y 19 los únicos influyentes según el criterio dado en 4.3.2.

Tabla 4.5: Datos de Weisberg (1980) con el análisis individual de influencia

OBS	X_1	X_2	X_3	Y	Q_k	**Variación**	$\mathcal{P}$	**valor de** p
1	176	6.5	0.88	0.42	0.0186	20.7928	1.7660	0.0977
2	176	9.5	0.88	0.25	0.0097	10.8042	1.2730	0.2224
3	190	9.0	1.00	0.56	0.0039	4.3433	0.8072	0.4322
4	176	8.9	0.88	0.23	0.0113	12.6451	1.3772	0.1886
5	200	7.2	1.00	0.23	0.0075	8.4090	1.1231	0.2791
6	167	8.9	0.83	0.32	0.0001	0.0677	0.1007	0.9211
7	188	8.0	0.94	0.37	0.0037	4.1391	0.7880	0.4430
8	195	10.0	0.98	0.41	0.0033	3.6762	0.7426	0.4692
9	176	8.0	0.88	0.33	0.0002	0.1813	0.1649	0.8712
10	165	7.9	0.84	0.38	0.0000	0.0103	0.0392	0.9692
11	158	6.9	0.80	0.27	0.0073	8.1412	1.1051	0.2865
12	148	7.3	0.74	0.36	0.0022	2.4193	0.6024	0.5559
13	149	5.2	0.75	0.21	0.0141	15.7217	1.5357	0.1454
14	163	8.4	0.81	0.28	0.0008	0.9465	0.3768	0.7116
15	170	7.2	0.85	0.34	0.0011	1.2074	0.4256	0.6765
16	186	6.8	0.94	0.28	0.0044	4.9176	0.8589	0.4039
17	146	7.3	0.73	0.30	0.0004	0.4671	0.2647	0.7948
18	181	9.0	0.90	0.37	0.0043	4.8273	0.8509	0.4082
19	149	6.4	0.75	0.46	0.0221	24.6283	1.9220	0.0738

En la tabla 4.6 se presenta la estadística $\mathcal{P}$ calculada para $\vec{Y}_1, \vec{Y}_2, \vec{Y}_3, \vec{Y}_4, \vec{Y}_5$ diferentes bloques que incluyen o excluyen los registros 1 y 19 que son los únicos clasificados como influyentes en el análisis individual de influencia, con una variación porcentual de la suma de cuadrados de los residuales del 20.79 % y 24.63 %. Se conforman los bloques siguiendo la misma metodología de los ejemplos anteriores incluyendo forzosamente uno de los registros influyentes en los bloques $\vec{Y}_2$, $\vec{Y}_3$ y $\vec{Y}_5$

- El bloque $\vec{Y}_1$ esta conformado por los registros 8, 9, 11, 4.
- El bloque $\vec{Y}_2$ esta conformado por los registros 7, 19, 15, 16.
- El bloque $\vec{Y}_3$ esta conformado por los registros 8, 19, 3, 11.
- El bloque $\vec{Y}_4$ esta conformado por los registros 4, 17, 13, 9.
- El bloque $\vec{Y}_5$ esta conformado por los registros 3, 11, 19, 17.

Tabla 4.6: Compendio de p-valores para los bloques conformados en la prueba de la hipótesis nula H_0 : El bloque $\vec{Y}_i$ no es influyente

Bloque	$\mathcal{P}$	**Valor de** p	**Decisión**
$\vec{Y}_1$	1.9226	0.07372	Influyente
$\vec{Y}_2$	2.2878	0.03709	Influyente
$\vec{Y}_3$	2.4735	0.02582	Influyente
$\vec{Y}_4$	2.0862	0.05445	Influyente
$\vec{Y}_5$	2.3742	0.03136	Influyente

Se concluye de los p-valores calculados según la estadística $\mathcal{P}$ y presentados en la tabla 4.6 que ella bien sirve a nuestros propósitos ya que detecta a un nivel de significancia del 10 %, los tres bloques influyentes y aunque en los otros bloques no se introdujo una de las observaciones influyentes dichos bloques también resultaron ser influyentes.

5 Metodología de imputación en la variable respuesta

En este capítulo se presenta la metodología para imputar un bloque no influyente en un modelo de regresión lineal múltiple con información faltante en la variable respuesta, principal objetivo de este trabajo. La metodología propuesta ofrece dos características importantes que la hacen novedosa:

1. Los valores imputados son **invariantes**, es decir, resultan ser siempre **los mismos sin importar los valores iniciales** usados en el método de imputación.

2. El bloque de observaciones imputado no es influyente según la definición dada en 4.3.2. y el criterio expuesto en 4.3.3.

5.1. Metodología de imputación

Para el modelo $Y = X\vec{\beta} + \vec{\epsilon}$ con algunas observaciones faltantes en la variable respuesta Y, la metodología de imputación propuesta en este trabajo sigue los siguientes pasos:

1. Organice los datos de tal manera que $\vec{Y}_1$ sea el bloque conformado por las observaciones consideradas faltantes, particionando el modelo según:

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}$$

2. Calcule la estadística Q_k de cada observación para el modelo

$$\vec{Y_2} = X_2\vec{\beta_c} + \vec{\epsilon} \qquad \text{con} \qquad \hat{\beta}_c = (X_2'X_2)^{-1}X_2'\vec{Y_2}.$$

3. Ordene en forma descendente el bloque $\left[\vec{Y_2},\, X_2\right]$ tomando como criterio la estadística Q_k, es decir ubique las r-observaciones menos influyentes al final del bloque.

4. Use el ordenamiento hecho en el paso 3. y organice nuevamente los datos en el modelo

$$\begin{bmatrix}\vec{Y_1}\\ \vec{Y_2}\end{bmatrix} = \begin{bmatrix}X_1\\ X_2\end{bmatrix}\vec{\beta} + \begin{bmatrix}\vec{\epsilon_1}\\ \vec{\epsilon_2}\end{bmatrix}$$

5. Introduzca un conjunto A arbitrario de valores iniciales en el bloque $\vec{Y_1}$,

$$\vec{Y_c} = \begin{bmatrix}A\\ \vec{Y_2}\end{bmatrix} = \begin{bmatrix}X_1\\ X_2\end{bmatrix}\vec{\beta} + \begin{bmatrix}\vec{\epsilon_1}\\ \vec{\epsilon_2}\end{bmatrix}$$

y con el modelo completado $\vec{Y_c} = X\vec{\beta} + \vec{\epsilon}$, calcule los residuales $\hat{\epsilon}$.

6. Con las observaciones que incluyen los valores iniciales que componen el bloque $\vec{Y_1}$ y las observaciones completas de mayor influencia, conforme un vector $\vec{Y_3}$ tal que la variación porcentual que genera el bloque $\left[\vec{Y_3},\, X_3\right]$ sea superior al 95 %. Y con este bloque calcule el vector $\hat{\gamma}$ según

$$\hat{\gamma} = -(I - H_3)^{-1}\hat{\epsilon}$$

con $\vec{\epsilon}$ los residuales correspondientes al bloque $\left[\vec{Y_3},\, X_3\right]$ calculados en el paso 5., y H_3 el bloque de la matriz H correspondiente al bloque $\left[\vec{Y_3},\, X_3\right]$.

7. Calcule el bloque $\vec{Y_1}$ a imputar ajustando los valores iniciales del conjunto A, con los valores $\hat{\gamma}_1$ correspondientes al bloque $\vec{Y_1}$ obtenidos del vector $\hat{\gamma}$ calculado en el paso 6., es decir

$$\vec{Y_1} = A + \hat{\gamma}_1 \tag{5.1}$$

5.2. Ejemplos

Como ilustración y aplicación de la metodología de imputación propuesta en las siguientes secciones se relacionan tres ejemplos que poseen un número de observaciones específico, y diferente número de variables regresoras. Del número de observaciones tomadas se seleccionaron observaciones que conformaran el bloque $\vec{Y}_1$ de observaciones faltantes correspondientes al 5 %, 10 % y 20 % de n el número de registros simulados en cada caso y se realizó la imputación propuesta utilizando para ello el paquete estadístico SAS en su procedimiento IML.

En cada uno de los ejemplos del número de registros que poseen se seleccionaron las primeras observaciones como las componentes de los bloques a utilizar de información faltante en tres escenarios cuando el porcentaje de información faltante es del 5 %, 10 % y 20 %, es decir

- **Escenario 1.** Correspondiente al 5 % del total de registros que se tomaron.
- **Escenario 2.** Correspondiente al 10 % del total de registros que se tienen.
- **Escenario 3.** Correspondiente al 20 % del total de registros que se tomaron.

En cada uno de los escenarios se presenta la desviación estándar en los parámetros estimados la cual se calcula como la raíz cuadrada de los elementos de la diagonal de la matriz de covarianza, ésta matriz se define como

$$\mathrm{Var}\big(\hat{\beta}\big) = (X'X)^{-1}\hat{\sigma}^2$$

donde $\hat{\sigma}^2$ es la estimación de la varianza determinada para cada uno de los modelos como

$$\hat{\sigma}^2 = \frac{SSE}{n-r} = CME \tag{5.2}$$

Cada una de las situaciones propuestas se plantean para los modelos que se definen a continuación:

- **Básico**: El modelo sin ajustar, es decir, con sus datos originales.
- **Completo**: El modelo ajustado eliminando el registro o registros en los cuales se presenta información faltante.
- **Valor inicial**: El modelo ajustado cuando la información faltante se completa con el promedio de $\vec{Y}_2$, el bloque completo, como valor inicial para realizar la imputación.
- **Propuesto**: El modelo ajustado cuando la imputación se realiza con el método propuesto en este trabajo, es decir, se calcularon los vectores $\vec{\gamma}$ definidos en el punto 3 del método de imputación propuesto y con ellos se obtuvo el valor o valores a imputar.

5.2.1. Modelo de regresión simple

Se usan los datos de lluvias y filtraciones asociadas al río Monocacy en Puente Jug, Maryland citados en Behar and Yepes (1996) en los cuales se midieron las siguientes variables

X: Filtración en pulg.

Y: Precipitación en pulg.

Los datos se presentan en la tabla 5.1, junto con las estadísticas Q_k, la variación porcentual que cada observación genera en la suma de cuadrados residual (SSE), cuando es eliminada, la estadística $\mathcal{P}$ de cada observación y el p-valor del criterio definido 4.3.2.

Tabla 5.1: Datos de lluvias y filtraciones, con el análisis individual de influencia

OBS	X	Y	Q_k	**Variación**	$\mathcal{P}$	**valor de** p
1	0.52	1.11	0.28477	3.83800	0.93954	0.35722
2	0.40	1.17	0.05958	0.80304	0.42977	0.67136
3	0.97	1.79	0.48123	6.48576	1.22136	0.23432
4	2.92	5.62	0.52721	7.10552	1.27839	0.21386
5	0.17	1.13	0.02516	0.33905	0.27925	0.78255
6	0.19	1.54	0.30001	4.04334	0.96435	0.34490
7	0.76	3.19	1.28981	17.38345	1.99955	0.05750
8	0.66	1.73	0.02710	0.36527	0.28985	0.77453
9	1.24	2.75	0.05230	0.70482	0.40263	0.69094
10	0.39	1.20	0.03766	0.50762	0.34169	0.73569
11	0.30	1.01	0.04763	0.64196	0.38425	0.70432
12	1.74	5.11	1.69925	22.90175	2.29508	0.03118
13	0.56	1.52	0.03584	0.48307	0.33333	0.74191
14	1.12	2.93	0.03496	0.47119	0.32920	0.74498
15	0.64	1.16	0.50309	6.78039	1.24880	0.22430
16	0.70	1.64	0.11058	1.49033	0.58547	0.56393
17	0.77	1.57	0.28823	3.88470	0.94524	0.35436
18	0.59	1.54	0.05102	0.68768	0.39770	0.69452
19	0.95	2.09	0.12164	1.63945	0.61406	0.54520
20	0.45	2.57	1.21090	16.32001	1.93742	0.06507
21	1.02	3.54	1.00281	13.51540	1.76311	0.09117
22	0.39	1.17	0.05063	0.68237	0.39616	0.69564
23	0.23	1.15	0.00388	0.05234	0.10972	0.91358
24	1.59	3.57	0.00321	0.04324	0.09973	0.92142
25	0.78	2.09	0.00063	0.00851	0.04424	0.96509

En las tablas 5.2, 5.3 y 5.4 se presentan las estimaciones del R^2, de los parámetros del modelo, el CME definido en (5.2) y la desviación estándar de los estimadores de los parámetros para los modelos **Completo**, **Valor inicial**, **Propuesto**.

Tabla 5.2: Resultados de las estimaciones del escenario 1 para los datos de lluvias y filtraciones.

Modelo	**Valor a imputar**	**Parámetros estimados**		$\mathbf{R^2}$	**CME**	**Desviación**	
		$\hat{\beta}_0$	$\hat{\beta}_1$			$\hat{\beta}_0$	$\hat{\beta}_1$
Básico		0.6618	1.8627	0.8008	0.3226	0.1924	0.1937
Completo	1.6573	0.6980	1.8447	0.8024	0.3243	0.1968	0.1952
Valor inicial	2.1992	0.7340	1.8269	0.7947	0.3224	0.1924	0.1936
Propuesto	1.6149	0.6952	1.8461	0.8042	0.3103	0.1887	0.1900

Como se observa en la tabla 5.2 el modelo propuesto es el que presenta el CME más pequeño a pesar de que el modelo completo tiene un menor número de observaciones, por lo anterior la desviación de los estimadores de los parámetros tambien resulta ser más pequeña en el modelo propuesto.

Tabla 5.3: Resultados de las estimaciones del escenario 2 para los datos de lluvias y filtraciones.

Modelo	**Valor a imputar**	**Parámetros estimados**		$\mathbf{R^2}$	**CME**	**Desviación**	
		$\hat{\beta}_0$	$\hat{\beta}_1$			$\hat{\beta}_0$	$\hat{\beta}_1$
Básico		0.6618	1.8627	0.8008	0.3226	0.1924	0.1937
Completo	1.6731 1.4534	0.7211	1.8309	0.7983	0.3362	0.2062	0.2008
Valor inicial	2.2439 2.2439	0.8202	1.7753	0.7741	0.3439	0.1987	0.2000
Propuesto	1.6149 1.3957	0.7127	1.8355	0.8039	0.3072	0.1878	0.1890

Un resultado importante del método propuesto es que el valor imputado en el escenario 1, cuando se asume faltante el 5 % y el valor imputado cuando se asume faltante el 10 %, es el mismo, es decir que dicho valor permanece invariante y no depende del valor

inicial, mientras que en el modelo con valor inicial estos valores son diferentes, además nuevamente el CME en el modelo propuesto es menor que en los otros modelos y, por consiguiente, la desviación de los estimadores de los parámetros en este modelo es también la más pequeña.

Tabla 5.4: Resultados de las estimaciones del escenario 3 para los datos de lluvias y filtraciones.

Modelo	**Valor a imputar**	**Parámetros estimados**		$\mathbf{R^2}$	**CME**	**Desviación**	
		$\hat{\beta}_0$	$\hat{\beta}_1$			$\hat{\beta}_0$	$\hat{\beta}_1$
Básico		0.6618	1.8627	0.8008	0.3226	0.1924	0.1937
Completo	1.6449 1.3835 2.6251 6.8724 0.8826	0.5123	2.1781	0.7325	0.3269	0.2664	0.3102
Valor inicial	2.1535 2.1535 2.1535 2.1535 2.1535	1.4634	0.8605	0.2894	0.6797	0.2793	0.2812
Propuesto	1.6149 1.3957 2.4372 6.0001 0.9754	0.6533	1.9532	0.8407	0.2703	0.1762	0.1773

Nuevamente se observa que en el modelo propuesto el valor imputado en el escenario 1, cuando se asume faltante el 5 %; el valor imputado en el escenario 2, cuando se asume faltante el 10 %, y el valor que imputa cuando se considera faltante el 20 % es el mismo, es decir, que dicho valor permanece invariante y no depende del valor inicial, mientras que en el modelo con valor inicial estos valores son diferentes, además nuevamente el CME en el modelo propuesto es menor que en los otros modelos y, por consiguiente, la desviación de los estimadores de los parámetros en este modelo es también la más pequeña.

5.2.2. Modelo de regresión con dos regresores

Se usan los datos sobre la Gravedad específica, contenidos de nitrógeno y fósforo de 20 muestras de papa tomados de Little and Hills (1972) en los cuales se midieron las siguientes variables

X_1 : Contenido de nitrógeno. X_2 : Contenido de fósforo.

Y : Gravedad específica de la papa.

Los datos se presentan en la tabla 5.5, junto con las estadísticas Q_k, la variación porcentual que cada observación genera en la suma de cuadrados residual (SSE), cuando es eliminada, la estadística $\mathcal{P}$ de cada observación y el p-valor del criterio definido 4.3.2.

Tabla 5.5: Datos de Little and Hills (1972) con el análisis individual de influencia

OBS	X_1	X_2	Y	Q_k	**Variación**	$\mathcal{P}$	**valor de** p
1	82	36	14	229.00014	5.65036	0.98008	0.34079
2	88	42	15	62.89612	1.55190	0.51364	0.61411
3	100	28	16	125.59109	3.09884	0.72581	0.47783
4	114	26	27	159.97137	3.94714	0.81915	0.42404
5	14	10	179	377.19729	9.30698	1.25785	0.22545
6	94	26	54	297.76924	7.34717	1.11759	0.27929
7	74	15	58	836.80630	20.64738	1.87351	0.07829
8	36	35	68	215.47509	5.31664	0.95070	0.35508
9	121	30	15	330.04351	8.14350	1.17660	0.25557
10	36	25	82	518.30232	12.78861	1.47447	0.15863
11	58	26	91	199.48282	4.92204	0.91474	0.37312
12	31	25	97	167.04423	4.12165	0.83707	0.41417
13	38	24	98	36.36191	0.89719	0.39054	0.70099
14	56	11	101	190.06983	4.68979	0.89290	0.38438
15	24	22	128	20.95036	0.51693	0.29644	0.77049
16	37	11	140	38.64986	0.95365	0.40264	0.69223
17	10	24	163	1010.64736	24.93674	2.05894	0.05516
18	73	15	83	8.49458	0.20960	0.18876	0.85252
19	96	40	2	3.67918	0.09078	0.12423	0.90259
20	71	33	48	0.90960	0.02244	0.06177	0.95147

En las tablas 5.6, 5.7 y 5.8 se presentan las estimaciones del R^2, de los parámetros del modelo, el CME definido en (5.2) y la desviación estándar de los estimadores de los parámetros para los modelos **Completo**, **Valor inicial**, **Propuesto**.

Tabla 5.6: Resultados de las estimaciones del escenario 1 para los datos de lluvias y filtraciones.

Modelo	**Valor a imputar**	**Parámetros estimados**			**R^2**	**CME**	**Desviación**		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Básico		199.9667	-1.0998	-2.2663	0.9208	238.4026	10.5159	0.1176	0.4204
Completo	30.1363	198.0765	-1.0974	-2.1653	0.9193	238.9903	10.7045	0.1178	0.4334
Valor inicial	77.1053	192.5746	-1.0904	-1.8712	0.8784	339.0629	12.5410	0.1403	0.5014
Propuesto	27.9267	198.3354	-1.0978	-2.1791	0.9230	225.1846	10.2203	0.1143	0.4086

Como se observa en la tabla 5.6 el modelo propuesto es el que presenta el CME más pequeño a pesar de que el modelo completo tiene un menor número de observaciones, por lo anterior la desviación de los estimadores de los parámetros tambien resulta ser más pequeña en el modelo propuesto.

Tabla 5.7: Resultados de las estimaciones del escenario 2 para los datos de lluvias y filtraciones.

Modelo	**Valor a imputar**	**Parámetros estimados**			**R^2**	**CME**	**Desviación**		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Básico		199.9667	-1.0998	-2.2663	0.9208	238.4026	10.5159	0.1176	0.4204
Completo	29.068 9.092	199.4040	-1.0973	-2.2321	0.9123	253.1741	11.7641	0.1213	0.4920
Valor inicial	80.556 80.556	178.8289	-1.0928	-1.1829	0.7935	526.1202	15.6220	0.1748	0.6246
Propuesto	27.927 7.884	199.7837	-1.0975	-2.2515	0.9249	223.4976	10.1819	0.1139	0.4071

Un resultado importante del método propuesto es que el valor imputado en el escenario 1, cuando se asume faltante el 5 % y el valor imputado cuando se asume faltante el 10 %, es el mismo, es decir que dicho valor permanece invariante y no depende del valor inicial, mientras que en el modelo con valor inicial estos valores son diferentes, además nuevamente el CME en el modelo propuesto es menor que en los otros modelos y, por consiguiente, la desviación de los estimadores de los parámetros en este modelo es también la más pequeña.

Tabla 5.8: Resultados de las estimaciones del escenario 3 para los datos de lluvias y filtraciones.

Modelo	**Valor a imputar**	**Parámetros estimados**			**R^2**	**CME**	**Desviación**		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Básico		199.9667	-1.0998	-2.2663	0.9208	238.4026	10.5159	0.1176	0.4204
Completo	29.2105 9.3739 26.8102 15.6691	199.2705	-1.1096	-2.1965	0.8996	273.4846	12.3743	0.1446	0.5136
Valor inicial	87.9375 87.9375 87.9375 87.9375	171.9129	-0.7769	-1.4008	0.6131	805.9182	19.3347	0.2163	0.7730
Propuesto	27.927 7.884 21.9828 8.5603	200.1214	-1.1396	-2.1848	0.9307	212.0353	9.9174	0.1110	0.3965

Un resultado importante del método propuesto es que el valor imputado en el escenario 1, cuando se asume faltante el 5 % y el valor imputado cuando se asume faltante el 10 %, es el mismo, es decir que dicho valor permanece invariante y no depende del valor inicial, mientras que en el modelo con valor inicial estos valores son diferentes, además nuevamente el CME en el modelo propuesto es menor que en los otros modelos y por consiguiente la desviación de los estimadores de los parámetros en este modelo es también la más pequeña.

5.2.3. Modelo de regresión simple con tres regresores

Se usan los datos sobre 21 diferentes días, las mediciones corresponden a la corriente de aire (X_1), la temperatura del agua fría (X_2), la concentración de ácido (X_3) y la cantidad de amoniaco que se escapa cuando se oxida (Y) tomadas de Brownlee (1965). Los datos se presentan en la tabla 5.9, junto con los estadísticos Q, la variación porcentual que cada observación genera en la suma de cuadrados residual SSE, cuando es eliminada, la estadística $\mathcal{P}$ de cada observación y el p-valor del criterio definido 4.3.2.

Tabla 5.9: Datos de Brownlee (1965) con el análisis individual de influencia

OBS	X_1	X_2	X_3	Y	Q_k	**Variación**	$\mathcal{P}$	**valor de** p
1	58	17	88	13	9.86863	2.75926	0.68489	0.50265
2	62	24	87	28	37.25166	10.41554	1.33065	0.20088
3	80	27	89	42	14.98026	4.18847	0.84382	0.41049
4	62	22	87	18	3.09118	0.86429	0.38331	0.70624
5	75	25	90	37	25.14327	7.03004	1.09321	0.28955
6	62	23	87	18	9.80115	2.74039	0.68254	0.50409
7	80	27	88	37	5.38987	1.50700	0.50615	0.61925
8	58	18	89	14	8.22525	2.29977	0.62527	0.54010
9	62	24	93	19	7.31293	2.04469	0.58957	0.56323
10	62	24	93	20	2.47282	0.69140	0.34284	0.73592
11	70	20	91	15	73.21724	20.47148	1.86552	0.07946
12	58	23	87	15	11.49911	3.21514	0.73931	0.46981
13	56	20	82	15	2.16800	0.60617	0.32101	0.75212
14	58	18	82	11	2.42239	0.67730	0.33932	0.73852
15	50	19	72	8	3.92982	1.09877	0.43219	0.67104
16	50	18	89	8	6.88827	1.92595	0.57220	0.57468
17	58	18	80	14	2.00734	0.56125	0.30889	0.76116
18	50	18	86	7	0.94268	0.26357	0.21168	0.83487
19	50	20	80	9	0.43359	0.12123	0.14356	0.88754
20	50	19	79	8	0.24673	0.06899	0.10829	0.91503
21	58	19	93	12	0.00321	0.00090	0.01235	0.99029

En las tablas 5.10, 5.11 y 5.12 se presentan las estimaciones del R^2, de los parámetros del modelo, el CME y la desviación estándar de los estimadores de los parámetros para los modelos **Completo**, **Valor inicial**, **Propuesto**.

Tabla 5.10: Resultados estimaciones del escenario 1 para los datos de Brownlee (1965).

Modelo	**Valor a imputar**	**Parámetros estimados y desviación**				**R^2**	**CME**
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		
Básico		-39.920 (11.896)	0.761 (0.135)	1.295 (0.368)	-0.152 (0.156)	0.914	10.519
Completo	9.45	-39.784 (11.920)	0.684 (0.139)	1.453 (0.403)	-0.172 (0.158)	0.917	10.560
Valor inicial	17.75	-40.101 (13.281)	0.758 (0.151)	1.085 (0.411)	-0.125 (0.174)	0.891	13.112
Propuesto	10.00	-39.805 (11.571)	0.689 (0.131)	1.428 (0.358)	-0.169 (0.152)	0.920	9.953

Como se observa en la tabla 5.10 el modelo propuesto es el que presenta el CME más pequeño a pesar de que el modelo completo tiene un menor número de observaciones, por lo anterior la desviación de los estimadores de los parámetros tambien resulta ser más pequeña en el modelo propuesto.

Tabla 5.11: Resultados estimaciones del escenario 2 para los datos de Brownlee (1965).

Modelo	**Valor a imputar**	**Parámetros estimados y desviación**				**R^2**	**CME**
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		
Básico		-39.920 (11.896)	0.761 (0.135)	1.295 (0.368)	-0.152 (0.156)	0.914	10.519
Completo	9.96 21.68	-38.520 (10.998)	0.738 (0.131)	1.228 (0.388)	-0.173 (0.145)	0.931	8.959
Valor inicial	17.21 17.21	-37.888 (12.226)	0.838 (0.139)	0.764 (0.378)	-0.135 (0.161)	0.902	11.111
Propuesto	10.00 19.00	-37.976 (10.550)	0.760 (0.120)	1.141 (0.326)	-0.174 (0.139)	0.929	8.273

Un resultado importante del método propuesto es que el valor imputado en el escenario 1, cuando se asume faltante el 5 % y el valor imputado cuando se asume faltante el

10 %, es el mismo, es decir que dicho valor permanece invariante y no depende del valor inicial, mientras que en el modelo con valor inicial estos valores son diferentes, además nuevamente el CME en el modelo propuesto es menor que en los otros modelos y por consiguiente la desviación de los estimadores de los parámetros en este modelo es también la más pequeña.

Tabla 5.12: Resultados estimaciones del escenario 3 para los datos de Brownlee (1965).

Modelo	**Valor a imputar**	**Parámetros estimados y desviación**				**R^2**	**CME**
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		
Básico		-39.920 (11.896)	0.761 (0.135)	1.295 (0.368)	-0.152 (0.156)	0.914	10.519
Completo	10.20 21.22 36.74 18.89	-37.371 (10.914)	0.683 (0.135)	1.166 (0.388)	-0.135 (0.146)	0.911	8.760
Valor inicial	15.71 15.71 15.71 15.71	-31.670 (18.724)	0.536 (0.212)	0.555 (0.579)	0.038 (0.246)	0.655	26.061
Propuesto	10.00 19.00 31.00 17.00	-35.524 (10.546)	0.634 (0.120)	1.045 (0.326)	-0.098 (0.139)	0.909	8.268

Nuevamente se observa que en el modelo propuesto el valor imputado en el escenario 1, cuando se asume faltante el 5 %; el valor imputado en el escenario 2, cuando se asume faltante el 10 %, y el valor que imputa cuando se considera faltante el 20 % es el mismo, es decir que dicho valor permanece invariante y no depende del valor inicial, mientras que en el modelo con valor inicial estos valores son diferentes, además nuevamente el CME en el modelo propuesto es menor que en los otros modelos y por consiguiente la desviación de los estimadores de los parámetros en este modelo es también la más pequeña.

6 Conclusiones

Para finalizar este trabajo se presenta en este capítulo algunos de los resultados más notables en el desarrollo de esta propuesta

1) La generalización de la estadística $DFBeta(\vec{Y}_1)$ es verdaderamente útil cuando se quiere controlar la estimación de los parámetros previamente conocidos, por ejemplo en control de calidad, pero no resulta muy útil para medir influencia desde el punto de vista de ajuste del modelo, en los ejemplos desarrollados se observaron bloques no influyentes en la estimación de los parámetros pero si influyentes en el ajuste del modelo.

2) El estadístico Q_k calculado para un conjunto de k observaciones, resulta ser más eficaz al determinar si un registro o grupo de registros, son influyentes, cuando en el ajuste interesa únicamente la influencia de dicho registro o grupo de registros en la estimación del modelo.

3) En el ejemplo 4.3.3.3., es interesante observar que un grupo de registros puede resultar influyente como bloque y, sin embargo, las componentes que lo conforman no ser influyentes en el análisis individual.

4) Un aspecto importante a considerar de la metodología de imputación propuesta en este trabajo es que se garantiza que los valores imputados **son únicos, es decir, no dependen del valor inicial base de los cálculos**.

5) Como era de esperarse después del análisis individual de las componentes del grupo imputado se cumple que son no influyentes.

7 Recomendaciones hacia el futuro

En este capítulo, se presenta una alternativa para realizar imputación en las variables explicativas X la cual se desarrolla en las secciones 7.1. el marco teórico y en la 7.2. la metodología; además una alternativa para realizar imputación en ambas variables tanto en la respuesta $\vec{Y}$ como en las explicativas X procedimiento que se desarrolla en las secciones 7.3. el marco teórico correspondiente y en la 7.4. la metodología a seguir y en la sección 7.5. se plantean las conclusiones pertinentes

7.1. Imputación de valores en las variables explicativas

Para el modelo de regresión lineal múltiple

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}$$

definido en (3.1), se plantea el modelo

$$\vec{Y} = X^*\vec{\beta}^* + \vec{\epsilon}^* \tag{7.1}$$

siendo $X^* = X + Z$, para Z una matriz de tamaño $n \times r$, en la cual la primera columna esta compuesta de ceros; interesa establecer las expresiones de los nuevos estimadores, en función de Z, X y de los estimadores descritos para el modelo (3.1).

Si se reemplaza X^* en el modelo (7.1), se tiene

$$\vec{Y} = (X + Z)\vec{\beta}^* + \vec{\epsilon}^*$$

de donde

$$\begin{aligned}\vec{\epsilon^*} &= \vec{Y} - (X+Z)\vec{\beta^*} \\ &= \vec{Y} - X\vec{\beta^*} - Z\vec{\beta^*}\end{aligned}$$

para obtener la estimación del vector $\vec{\beta^*}$ por el método de mínimos cuadrados ordinarios, se considera

$$\frac{\partial(\vec{\epsilon^*})'(\vec{\epsilon^*})}{\partial\vec{\beta^*}} = \frac{\partial}{\partial\vec{\beta^*}}(\vec{Y} - X\vec{\beta^*} - Z\vec{\beta^*})'(\vec{Y} - X\vec{\beta^*} - Z\vec{\beta^*})$$

de la cual se obtiene las ecuaciones normales

$$\begin{aligned}X'\vec{Y} =& X'X\hat{\beta^*} + X'Z\hat{\beta^*} - Z'\vec{Y} + Z'X\hat{\beta^*} + Z'Z\hat{\beta^*} \\ =& X'X\hat{\beta^*} + X'Z\hat{\beta^*} - Z'[\vec{Y} - X\hat{\beta^*} - Z\hat{\beta^*}] \\ =& X'X\hat{\beta^*} + X'Z\hat{\beta^*} - Z'\hat{\epsilon^*}\end{aligned}$$

si se multiplica por $(X'X)^{-1}$ se llega a

$$\hat{\beta} = \hat{\beta^*} + CZ\hat{\beta^*} - (X'X)^{-1}Z'\hat{\epsilon^*} \tag{7.2}$$

siendo $C = (X'X)^{-1}X'$, si se despeja $\hat{\beta^*}$ se tiene

$$\hat{\beta^*} = (I + CZ)^{-1}[\hat{\beta} + (X'X)^{-1}Z'\hat{\epsilon^*}]$$

si se multiplica (7.2) por $-X$, se obtiene

$$-X\hat{\beta} = -X\hat{\beta^*} - HZ\hat{\beta^*} + C'Z'\hat{\epsilon^*}.$$

El vector de errores estimado del modelo (7.1), se obtiene sumando a ambos lados de la expresión anterior $\vec{Y}$

$$\begin{aligned}\vec{Y} - X\hat{\beta} =& \vec{Y} - X\hat{\beta^*} - HZ\hat{\beta^*} + C'Z'\hat{\epsilon^*} \\ \hat{\epsilon} =& (\vec{Y} - X\hat{\beta^*} - Z\hat{\beta^*}) + Z\hat{\beta^*} - HZ\hat{\beta^*} + C'Z'\hat{\epsilon^*} \\ =& \hat{\epsilon^*} + C'Z'\hat{\epsilon^*} + [I - H]Z\hat{\beta^*} \\ =& (I + C'Z')\hat{\epsilon^*} + (I - H)Z\hat{\beta^*}.\end{aligned} \tag{7.3}$$

Si se reescribe esta última expresión se llega a

$$(I + C'Z')\hat{\epsilon}^* = \hat{\epsilon} - (I - H)Z\hat{\beta}^*$$

$$\hat{\epsilon}^* = (I + C'Z')^{-1}[\hat{\epsilon} - (I - H)Z\hat{\beta}^*]$$

La expresión (7.3) nos permite calcular la suma de cuadrados de los residuales del modelo (7.1) como

$$\begin{aligned}(\hat{\epsilon})'(\hat{\epsilon}) =&[(I + C'Z')\hat{\epsilon}^* + (I - H)Z\hat{\beta}^*]'[(I + C'Z')\hat{\epsilon}^* + (I - H)Z\hat{\beta}^*] \\ =&[(\hat{\epsilon}^*)'(I + ZC) + (\hat{\beta}^*)'Z'(I - H)][(I + C'Z')\hat{\epsilon}^* + (I - H)Z\hat{\beta}^*] \\ =&(\hat{\epsilon}^*)'(I + ZC)(I + C'Z')\hat{\epsilon}^* + (\hat{\beta}^*)'Z'(I - H)(I + C'Z')\hat{\epsilon}^* \\ &+ (\hat{\epsilon}^*)'(I + ZC)(I - H)Z\hat{\beta}^* + (\hat{\beta}^*)'Z'(I - H)Z\hat{\beta}^* \end{aligned} \tag{7.4}$$

pero como

$$\begin{aligned}(I - H)(I + C'Z) &= I - H + C'Z' - HC'Z' \\ &= I - H + C'Z' - [X(X'X)^{-1}X'][X(X'X)^{-1}]Z' \\ &= I - H\end{aligned}$$

de manera análoga se tiene que

$$(I + ZC)(I - H) = I - H$$

y puesto que

$$(\hat{\beta}^*)'Z'(I - H)(I + C'Z')\hat{\epsilon}^* = (\hat{\epsilon}^*)'(I + ZC)(I - H)Z\hat{\beta}^*$$

estas expresiones reemplazadas en (7.4) proporcionan la siguiente

$$\hat{\epsilon}'\hat{\epsilon} = (\hat{\epsilon}^*)'(I + ZC)(I + C'Z')\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'(I - H)\hat{\epsilon}^* + (\hat{\beta}^*)'Z'(I - H)Z\hat{\beta}^* \tag{7.5}$$

por otra parte

$$(\hat{\epsilon}^*)'(I + ZC)(I + C'Z')\hat{\epsilon}^* = (\hat{\epsilon}^*)'\hat{\epsilon}^* + (\hat{\epsilon}^*)'ZC\hat{\epsilon}^* + (\hat{\epsilon}^*)'C'Z'\hat{\epsilon}^* + (\hat{\epsilon}^*)'ZCC'Z'\hat{\epsilon}^*$$

pero

$$\begin{aligned} C\hat{\epsilon}^* =&(X'X)^{-1}X'(\vec{Y} - X\hat{\beta}^* - Z\hat{\beta}^*) \\ =&(X'X)^{-1}X'\vec{Y} - (X'X)^{-1}X'X\hat{\beta}^* - (X'X)^{-1}X'Z\hat{\beta}^* \\ =&\hat{\beta} - \hat{\beta}^* - CZ\hat{\beta}^* \end{aligned}$$

por lo tanto, se tiene que

$$\begin{aligned} (\hat{\epsilon}^*)'(I + ZC)(I + C'Z')\hat{\epsilon}^* =&(\hat{\epsilon}^*)'\hat{\epsilon}^* + (\hat{\epsilon}^*)'Z[\hat{\beta} - \hat{\beta}^* - CZ\hat{\beta}^*] + [\hat{\beta} - \hat{\beta}^* \\ &- CZ\hat{\beta}^*]'Z'\hat{\epsilon}^* + (\hat{\epsilon}^*)'ZCC'Z'\hat{\epsilon}^* \end{aligned}$$

si se reemplaza (7.2) se obtiene

$$\begin{aligned} (\hat{\epsilon}^*)'(I + ZC)(I + C'Z')\hat{\epsilon}^* =&(\hat{\epsilon}^*)'\hat{\epsilon}^* + (\hat{\epsilon}^*)'Z[CZ\hat{\beta}^* - (X'X)^{-1}Z'\hat{\epsilon}^* - CZ\hat{\beta}^*] \\ &+ [(CZ\hat{\beta}^*)' - [(X'X)^{-1}Z'\hat{\epsilon}^*]' - (\hat{\beta}^*)'Z'C']Z'\hat{\epsilon}^* \\ &+ (\hat{\epsilon}^*)'Z[(X'X)^{-1}X'][X(X'X)^{-1}]Z'\hat{\epsilon}^* \\ =&(\hat{\epsilon}^*)'\hat{\epsilon}^* + (\hat{\epsilon}^*)'Z[-(X'X)^{-1}Z'\hat{\epsilon}^*] + \\ &[-(\hat{\epsilon}^*)'Z(X'X)^{-1}]Z'\hat{\epsilon}^* + (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* \\ =&(\hat{\epsilon}^*)'\hat{\epsilon}^* - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* \end{aligned}$$

la cual al ser reemplazada en (7.5) proporciona la siguiente

$$\hat{\epsilon}'\hat{\epsilon} = (\hat{\epsilon}^*)'\hat{\epsilon}^* - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'(I - H)\hat{\epsilon}^* + (\hat{\beta}^*)'Z'(I - H)Z\hat{\beta}^*$$

pero como

$$\begin{aligned} (I - H)\hat{\epsilon}^* =&(I - H)(\vec{Y} - X\hat{\beta}^* - Z\hat{\beta}^*) \\ =&(I - H)\vec{Y} - (I - H)X\hat{\beta}^* - (I - H)Z\hat{\beta}^* \\ =&\hat{\epsilon} - (I - H)Z\hat{\beta}^* \end{aligned}$$

por lo tanto, se tiene que

$$\begin{aligned}\hat{\epsilon}'\hat{\epsilon} =&\hat{\epsilon}^{*'}\hat{\epsilon}^* - \hat{\epsilon}^{*'}Z(X'X)^{-1}Z'\hat{\epsilon}^* + 2\hat{\beta}^{*'}Z'[\hat{\epsilon} - (I-H)Z\hat{\beta}^*] + \hat{\beta}^{*'}Z'(I-H)Z\hat{\beta}^* \\ =&(\hat{\epsilon}^*)'\hat{\epsilon}^* - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'\hat{\epsilon} - (\hat{\beta}^*)'Z'(I-H)Z\hat{\beta}^*\end{aligned}$$

Esta expresión se puede reescribir como

$$SSE - SSE^* = 2(\hat{\beta}^*)'Z'\hat{\epsilon} - (\hat{\beta}^*)'Z'(I-H)Z\hat{\beta}^* - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^*$$

Sin pérdida de generalidad, los resultados anteriores se pueden particularizar para el modelo (3.1) particionado

$$\begin{bmatrix}\vec{Y}_1 \\ \vec{Y}_2\end{bmatrix} = \begin{bmatrix}X_1 \\ X_2\end{bmatrix}\vec{\beta}^* + \begin{bmatrix}Z_1 \\ 0\end{bmatrix}\vec{\beta}^* + \begin{bmatrix}\hat{\epsilon}_1^* \\ \hat{\epsilon}_2^*\end{bmatrix} \tag{7.6}$$

con Z_1 de dimensión $k \times r, k < n$ y en tal caso

$$\begin{aligned}\hat{\beta} =&\hat{\beta}^* + (X'X)^{-1}X'Z\hat{\beta}^* - (X'X)^{-1}Z'\hat{\epsilon}^* \\ =&\hat{\beta}^* + (X'X)^{-1}\begin{bmatrix}X_1' & X_2'\end{bmatrix}\begin{bmatrix}Z_1 \\ 0\end{bmatrix}\hat{\beta}^* - (X'X)^{-1}\begin{bmatrix}Z_1' & 0\end{bmatrix}\begin{bmatrix}\hat{\epsilon}_1^* \\ \hat{\epsilon}_2^*\end{bmatrix} \\ =&\hat{\beta}^* + (X'X)^{-1}X_1'Z_1\hat{\beta}^* - (X'X)^{-1}Z_1'\hat{\epsilon}_1^*\end{aligned} \tag{7.7}$$

Bajo esta partición la ecuación (7.3) se expresa como

$$\begin{aligned}\begin{bmatrix}\hat{\epsilon}_1 \\ \hat{\epsilon}_2\end{bmatrix} &= \begin{bmatrix}\hat{\epsilon}_1^* \\ \hat{\epsilon}_2^*\end{bmatrix} + \begin{bmatrix}X_1 \\ X_2\end{bmatrix}(X'X)^{-1}\begin{bmatrix}Z_1' & 0\end{bmatrix}\begin{bmatrix}\hat{\epsilon}_1^* \\ \hat{\epsilon}_2^*\end{bmatrix} + \begin{bmatrix}I-H_{11} & -H_{12} \\ -H_{21} & I-H_{22}\end{bmatrix}\begin{bmatrix}Z_1 \\ 0\end{bmatrix}\hat{\beta}^* \\ &= \begin{bmatrix}I + C_1'Z_1' & 0 \\ C_2'Z_1' & I\end{bmatrix}\begin{bmatrix}\hat{\epsilon}_1^* \\ \hat{\epsilon}_2^*\end{bmatrix} + \begin{bmatrix}I-H_{11} & -H_{12} \\ -H_{21} & I-H_{22}\end{bmatrix}\begin{bmatrix}Z_1\hat{\beta}^* \\ 0\end{bmatrix}\end{aligned}$$

con $C_i = (X'X)^{-1}X_i'$ y $H_{ij} = X_i(X'X)^{-1}X_j'$.

De tal manera que el vector $Z_1\hat{\beta}^*$ que hace $\hat{\epsilon}_1^* = 0$ está dado por

$$Z_1\hat{\beta}^* = (I - H_{11})^{-1}\hat{\epsilon}_1 \tag{7.8}$$

para despejar la matriz Z_1 se multiplica a ambos lados de (7.8) por $(\hat{\beta}^*)'$, es decir

$$Z_1\hat{\beta}^*(\hat{\beta}^*)' = (I - H_{11})^{-1}\hat{\epsilon}_1(\hat{\beta}^*)'$$

y luego se multiplica por $[\hat{\beta}^*(\hat{\beta}^*)']^{-1}$, la cual puede ser una inversa generalizada, para finalmente obtener que

$$Z_1 = (I - H_{11})^{-1}\hat{\epsilon}_1(\hat{\beta}^*)'[\hat{\beta}^*(\hat{\beta}^*)']^{-1}$$

7.2. Metodología de imputación en las variables explicativas

Para el modelo $Y = X\vec{\beta} + \vec{\epsilon}$ con algunas observaciones faltantes en las variables explicativas X, la metodología de imputación que se recomienda en este trabajo sigue los siguientes pasos:

1. Organice los datos de tal manera que X_1 sea la matriz conformada por las observaciones consideradas faltantes, particionando el modelo según:

$$\begin{bmatrix}\vec{Y}_1\\ \vec{Y}_2\end{bmatrix} = \begin{bmatrix}X_1\\ X_2\end{bmatrix}\vec{\beta} + \begin{bmatrix}\vec{\epsilon}_1\\ \vec{\epsilon}_2\end{bmatrix}$$

2. Calcule la estadística Q_k de cada observación para el modelo

$$\vec{Y}_2 = X_2\vec{\beta}_c + \vec{\epsilon} \qquad \text{con} \qquad \hat{\beta}_c = (X_2'X_2)^{-1}X_2'\vec{Y}_2.$$

3. Ordene en forma descendente el bloque $\left[\vec{Y}_2,\, X_2\right]$ tomando como criterio la estadística Q_k, es decir, ubique las r-observaciones menos influyentes al final del bloque.

4. Use el ordenamiento hecho en 3. y organice nuevamente los datos en el modelo

$$\begin{bmatrix}\vec{Y}_1\\ \vec{Y}_2\end{bmatrix} = \begin{bmatrix}X_1\\ X_2\end{bmatrix}\vec{\beta} + \begin{bmatrix}\vec{\epsilon}_1\\ \vec{\epsilon}_2\end{bmatrix}$$

5. Introduzca un conjunto B arbitrario de valores iniciales en la matriz X_1,

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} B \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}$$

con el modelo completo $\vec{Y} = X_c\vec{\beta} + \vec{\epsilon}$, donde $X_c = \begin{bmatrix} B \\ X_2 \end{bmatrix}$, calcule los residuales $\hat{\epsilon}$.

6. Con las observaciones que incluyen los valores iniciales que componen la matriz X_1 y las observaciones completas de mayor influencia, conforme una matriz X_3 tal que la variación porcentual que genera el bloque $\left[\vec{Y}_3,\, X_3\right]$ sea superior al 95 %. Y con este bloque calcule el vector $Z\hat{\beta}^*$ según

$$Z\hat{\beta}^* = (I - H_3)^{-1}\hat{\epsilon} \tag{7.9}$$

donde $\vec{\epsilon}$ son los residuales correspondientes al bloque $\left[\vec{Y}_3,\, X_3\right]$ calculados en el paso 5., y H_3 el bloque de la matriz H correspondiente al bloque $\left[\vec{Y}_3,\, X_3\right]$; para despejar la matriz Z se multiplica a ambos lados de (7.9) por $(\hat{\beta}^*)'$, es decir

$$Z\hat{\beta}^*(\hat{\beta}^*)' = (I - H_3)^{-1}\hat{\epsilon}(\hat{\beta}^*)'$$

y luego se multiplica por $[\hat{\beta}^*(\hat{\beta}^*)']^{-1}$, la cual puede ser una inversa generalizada, para finalmente obtener que

$$Z = (I - H_3)^{-1}\hat{\epsilon}(\hat{\beta}^*)'[\hat{\beta}^*(\hat{\beta}^*)']^{-1}$$

7. Calcule la matriz X_1 a imputar ajustando los valores iniciales del conjunto B, con los valores Z_1 correspondientes a la matriz X_1 obtenidos de la matriz Z calculado en el paso 6., es decir

$$X_1 = B + Z_1. \tag{7.10}$$

7.3. Imputando valores simultáneamente en ambas variables

Para el modelo de regresión lineal múltiple

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}$$

definido en (3.1), se plantea el modelo

$$\vec{Y}^* = X^*\vec{\beta}^* + \vec{\epsilon}^* \tag{7.11}$$

siendo $\vec{Y}^* = \vec{Y} + \gamma$ y $X^* = X + Z$, para $\vec{\gamma}$ un vector de constantes conocido y Z una matriz de tamaño $n \times r$, interesa establecer las expresiones de los nuevos estimadores, en función de $\vec{Y}$, $\vec{\gamma}$, Z, X y de los estimadores descritos para el modelo (3.1).
si se reemplaza $\vec{Y}^*$ y X^* en el modelo (7.11), se tiene

$$\vec{Y} + \vec{\gamma} = (X + Z)\vec{\beta}^* + \vec{\epsilon}^*$$

de donde

$$\begin{aligned}\vec{\epsilon}^* &= \vec{Y} + \vec{\gamma} - (X + Z)\vec{\beta}^* \\ &= \vec{Y} + \vec{\gamma} - X\vec{\beta}^* - Z\vec{\beta}^*\end{aligned}$$

para obtener la estimación del vector β^* por el método de mínimos cuadrados ordinarios, se considera

$$\frac{\partial(\vec{\epsilon}^*)'(\vec{\epsilon}^*)}{\partial\vec{\beta}^*} = \frac{\partial}{\partial\vec{\beta}^*}(\vec{Y} + \vec{\gamma} - X\vec{\beta}^* - Z\vec{\beta}^*)'(\vec{Y} + \vec{\gamma} - X\vec{\beta}^* - Z\vec{\beta}^*)$$

de la cual se llega a las ecuaciones normales

$$\begin{aligned}X'Y =& X'X\hat{\beta}^* + X'Z\hat{\beta}^* - Z'\vec{Y} + Z'X\hat{\beta}^* + Z'Z\hat{\beta}^* - X'\vec{\gamma} - Z'\vec{\gamma} \\ =& X'X\hat{\beta}^* + X'Z\hat{\beta}^* - X'\vec{\gamma} - Z'[\vec{Y} + \vec{\gamma} - X\hat{\beta}^* - Z\hat{\beta}^*] \\ =& X'X\hat{\beta}^* + X'Z\hat{\beta}^* - X'\vec{\gamma} - Z'\hat{\epsilon}^*\end{aligned}$$

si se multiplica por $(X'X)^{-1}$ se llega a

$$\hat{\beta} = \hat{\beta}^* + CZ\hat{\beta}^* - C\vec{\gamma} - (X'X)^{-1}Z'\hat{\epsilon}^* \tag{7.12}$$

al multiplicar esta expresión por X, se obtiene

$$\begin{aligned} X\hat{\beta} &= X\hat{\beta}^* + HZ\hat{\beta}^* - H\vec{\gamma} - C'Z'\hat{\epsilon}^* \\ \hat{Y} &= [X\hat{\beta}^* + Z\hat{\beta}^*] - Z\hat{\beta}^* + HZ\hat{\beta}^* - H\vec{\gamma} - C'Z'\hat{\epsilon}^* \\ \hat{Y} &= \hat{Y}^* - (I-H)Z\hat{\beta}^* - H\vec{\gamma} - C'Z'\hat{\epsilon}^*. \end{aligned} \tag{7.13}$$

Luego, el nuevo vector de predicciones $\hat{Y}^*$ está dado por

$$\hat{Y}^* = \hat{Y} + (I-H)Z\hat{\beta}^* + H\vec{\gamma} + C'Z'\hat{\epsilon}^*$$

Para calcular el vector de errores estimado del modelo (7.11), se multiplica la expresión (7.13) por (-1) y se suma a ambos lados $\vec{Y}$

$$\begin{aligned} \vec{Y} - \hat{Y} =& \vec{Y} - \hat{Y}^* + (I-H)Z\hat{\beta}^* + H\vec{\gamma} + C'Z'\hat{\epsilon}^* \\ \hat{\epsilon} =& (\vec{Y} + \vec{\gamma} - \hat{Y}^*) - \vec{\gamma} + H\vec{\gamma} + (I-H)Z\hat{\beta}^* + C'Z'\hat{\epsilon}^* \\ =& \hat{\epsilon}^* + C'Z'\hat{\epsilon}^* - (I-H)\vec{\gamma} + (I-H)Z\hat{\beta}^* \\ =& (I + C'Z')\hat{\epsilon}^* + (I-H)[Z\hat{\beta}^* - \vec{\gamma}]. \end{aligned} \tag{7.14}$$

Si se reescribe esta última expresión se llega a

$$\begin{aligned} (I + C'Z')\hat{\epsilon}^* &= \hat{\epsilon} - (I-H)[Z\hat{\beta}^* - \vec{\gamma}] \\ \hat{\epsilon}^* &= (I + C'Z')^{-1}[\hat{\epsilon} - (I-H)(Z\hat{\beta}^* - \vec{\gamma})] \end{aligned}$$

La expresión (7.14) nos permite calcular la suma de cuadrados de los residuales del modelo (7.11) como

$$\begin{aligned} (\hat{\epsilon})'(\hat{\epsilon}) =& [(I + C'Z')\hat{\epsilon}^* + (I-H)(Z\hat{\beta}^* - \vec{\gamma})]'[(I + C'Z')\hat{\epsilon}^* + (I-H)(Z\hat{\beta}^* - \vec{\gamma})] \\ =& [\hat{\epsilon}^{*'}(I + ZC) + \hat{\beta}^{*'}Z'(I-H) - \vec{\gamma}'(I-H)][(I + C'Z')\hat{\epsilon}^* + (I-H)Z\hat{\beta}^* \\ & - (I-H)\vec{\gamma}] \end{aligned}$$

$$
\begin{aligned}
(\hat{\epsilon})'(\hat{\epsilon}) =&(\hat{\epsilon}^*)'(I+ZC)(I+C'Z')\hat{\epsilon}^* + (\hat{\beta}^*)'Z'(I-H)(I+C'Z')\hat{\epsilon}^* + \vec{\gamma}'(I-H)\vec{\gamma} - \\
&\hat{\epsilon}^{*'}(I+ZC)(I-H)\vec{\gamma} + \hat{\epsilon}^{*'}(I+ZC)(I-H)Z\hat{\beta}^* + \hat{\beta}^{*'}Z'(I-H)Z\hat{\beta}^* \\
&-(\hat{\beta}^*)'Z'(I-H)\vec{\gamma} - \vec{\gamma}'(I-H)(I+C'Z')\hat{\epsilon}^* - \vec{\gamma}'(I-H)Z\hat{\beta}^* \qquad (7.15)
\end{aligned}
$$

pero como

$$
\begin{aligned}
(I-H)(I+C'Z') &= I - H + C'Z' - HC'Z' \\
&= I - H + C'Z' - [X(X'X)^{-1}X'][X(X'X)^{-1}]Z' \\
&= I - H
\end{aligned}
$$

de manera análoga se tiene que

$$
(I+ZC)(I-H) = I - H
$$

y puesto que

$$
(\hat{\beta}^*)'Z'(I-H)(I+C'Z')\hat{\epsilon}^* = (\hat{\epsilon}^*)'(I+ZC)(I-H)Z\hat{\beta}^*
$$

estas expresiones reemplazadas en (7.15) proporcionan la siguiente

$$
\begin{aligned}
\hat{\epsilon}'\hat{\epsilon} =&(\hat{\epsilon}^*)'(I+ZC)(I+C'Z')\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'(I-H)\hat{\epsilon}^* + (\hat{\beta}^*)'Z'(I-H)Z\hat{\beta}^* \\
&-(\hat{\epsilon}^*)'(I-H)\vec{\gamma} - (\hat{\beta}^*)'Z'(I-H)\vec{\gamma} - \vec{\gamma}'(I-H)\hat{\epsilon}^* - \vec{\gamma}'(I-H)Z\hat{\beta}^* \\
&+\vec{\gamma}'(I-H)\vec{\gamma}
\end{aligned}
$$

además

$$
\begin{aligned}
(\hat{\epsilon}^*)'(I-H)\vec{\gamma} &= \vec{\gamma}'(I-H)\hat{\epsilon}^* \\
(\hat{\beta}^*)'Z'(I-H)\vec{\gamma} &= \vec{\gamma}'(I-H)Z\hat{\beta}^*
\end{aligned}
$$

por lo tanto

$$
\begin{aligned}
\hat{\epsilon}'\hat{\epsilon} =&(\hat{\epsilon}^*)'(I+ZC)(I+C'Z')\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'(I-H)\hat{\epsilon}^* + (\hat{\beta}^*)'Z'(I-H)Z\hat{\beta}^* \\
&-2\vec{\gamma}'(I-H)\hat{\epsilon}^* - 2\vec{\gamma}'(I-H)Z\hat{\beta}^* + \vec{\gamma}'(I-H)\vec{\gamma} \qquad (7.16)
\end{aligned}
$$

por otra parte

$$(\hat{\epsilon}^*)'(I+ZC)(I+C'Z')\hat{\epsilon}^* = (\hat{\epsilon}^*)'\hat{\epsilon}^* + (\hat{\epsilon}^*)'ZC\hat{\epsilon}^* + (\hat{\epsilon}^*)'C'Z'\hat{\epsilon}^* + (\hat{\epsilon}^*)'ZCC'Z'\hat{\epsilon}^*$$

pero

$$\begin{aligned} C\hat{\epsilon}^* =&(X'X)^{-1}X'(\vec{Y}+\vec{\gamma}-X\hat{\beta}^*-Z\hat{\beta}^*) \\ =&(X'X)^{-1}X'\vec{Y}+(X'X)^{-1}X'\vec{\gamma}-(X'X)^{-1}X'X\hat{\beta}^*-(X'X)^{-1}X'Z\hat{\beta}^* \\ =&\hat{\beta}-\hat{\beta}^*+C\vec{\gamma}-CZ\hat{\beta}^* \end{aligned}$$

por consiguiente

$$\begin{aligned} (\hat{\epsilon}^*)'(I+ZC)(I+C'Z')\hat{\epsilon}^* =&(\hat{\epsilon}^*)'\hat{\epsilon}^*+(\hat{\epsilon}^*)'Z[\hat{\beta}-\hat{\beta}^*+C\vec{\gamma}-CZ\hat{\beta}^*]+[\hat{\beta}-\hat{\beta}^*+ \\ &C\vec{\gamma}-CZ\hat{\beta}^*]'Z'\hat{\epsilon}^*+(\hat{\epsilon}^*)'ZCC'Z'\hat{\epsilon}^* \end{aligned}$$

si se reemplaza (7.12) se obtiene

$$\begin{aligned} (\hat{\epsilon}^*)'(I+ZC)(I+C'Z')\hat{\epsilon}^* =&(\hat{\epsilon}^*)'\hat{\epsilon}^*+(\hat{\epsilon}^*)'Z[CZ\hat{\beta}^*-C\vec{\gamma}-(X'X)^{-1}Z'\hat{\epsilon}^*+C\vec{\gamma}- \\ &CZ\hat{\beta}^*]+[(CZ\hat{\beta}^*)'-(C\vec{\gamma})'-[(X'X)^{-1}Z'\hat{\epsilon}^*]'+\vec{\gamma}'C' \\ &-(\hat{\beta}^*)'Z'C']Z'\hat{\epsilon}^*+(\hat{\epsilon}^*)'Z[(X'X)^{-1}X'][X(X'X)^{-1}]Z'\hat{\epsilon}^* \\ =&(\hat{\epsilon}^*)'\hat{\epsilon}^*+(\hat{\epsilon}^*)'Z[-(X'X)^{-1}Z'\hat{\epsilon}^*]+ \\ &[-(\hat{\epsilon}^*)'Z(X'X)^{-1}]Z'\hat{\epsilon}^*+(\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* \\ =&(\hat{\epsilon}^*)'\hat{\epsilon}^*-(\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* \end{aligned}$$

y como

$$\begin{aligned} (I-H)\hat{\epsilon}^* &= (I-H)(\vec{Y}+\vec{\gamma}-X\hat{\beta}^*-Z\hat{\beta}^* \\ &= \hat{\epsilon}+(I-H)\vec{\gamma}-(I-H)X\hat{\beta}^*-(I-H)Z\hat{\beta}^*) \\ &= \hat{\epsilon}+(I-H)\vec{\gamma}-(I-H)Z\hat{\beta}^* \end{aligned}$$

Estas expresiones reemplazadas en (7.16) proporcionan la siguiente

$$\begin{aligned}\hat{\epsilon}'\hat{\epsilon} =&(\hat{\epsilon}^*)'\hat{\epsilon}^* - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'[\hat{\epsilon} + (I-H)\vec{\gamma} - (I-H)Z\hat{\beta}^*] + \\ &(\hat{\beta}^*)'Z'(I-H)Z\hat{\beta}^* - 2\vec{\gamma}'[\hat{\epsilon} + (I-H)\vec{\gamma} - (I-H)Z\hat{\beta}^*] - 2\vec{\gamma}'(I-H)Z\hat{\beta}^* + \\ &\vec{\gamma}'(I-H)\vec{\gamma} \\ SSE =&SSE^* - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^* + 2(\hat{\beta}^*)'Z'\hat{\epsilon} - (\hat{\beta}^*)'Z'(I-H)Z\hat{\beta}^* - \\ &2\vec{\gamma}'\hat{\epsilon} + 2\vec{\gamma}'(I-H)Z\hat{\beta}^* - \vec{\gamma}'(I-H)\vec{\gamma}\end{aligned}$$

Esta expresión se puede reescribir como

$$SSE - SSE^* = 2(Z\hat{\beta}^* - \vec{\gamma})'\hat{\epsilon} - (Z\hat{\beta}^* - \vec{\gamma})'(I-H)(Z\hat{\beta}^* - \vec{\gamma}) - (\hat{\epsilon}^*)'Z(X'X)^{-1}Z'\hat{\epsilon}^*$$

Sin pérdida de generalidad, los resultados anteriores se pueden particularizar para el modelo (3.1) particionado

$$\begin{bmatrix}\vec{Y}_1\\ \vec{Y}_2\\ \vec{Y}_3\end{bmatrix} + \begin{bmatrix}\vec{\gamma}_1\\ \vec{0}\\ \vec{0}\end{bmatrix} = \begin{bmatrix}X_1\\ X_2\\ X_3\end{bmatrix}\vec{\beta}^* + \begin{bmatrix}0\\ Z_2\\ 0\end{bmatrix}\vec{\beta}^* + \begin{bmatrix}\vec{\epsilon}_1^*\\ \vec{\epsilon}_2^*\\ \vec{\epsilon}_3^*\end{bmatrix} \tag{7.17}$$

con $\vec{Y}_1$ de dimensión $k_1 \times 1$ y X_2 de dimensión $k_2 \times r, k_1 + k_2 < n$ y en tal caso

$$\begin{aligned}\hat{\beta} =&\hat{\beta}^* + (X'X)^{-1}X'Z\vec{\beta}^* - (X'X)^{-1}X'\vec{\gamma} - (X'X)^{-1}Z'\vec{\epsilon}^* \\ =&\hat{\beta}^* + (X'X)^{-1}\begin{bmatrix}X_1' & X_2' & X_3'\end{bmatrix}\begin{bmatrix}0\\ Z_2\\ 0\end{bmatrix}\hat{\beta}^* - (X'X)^{-1}\begin{bmatrix}X_1' & X_2' & X_3'\end{bmatrix}\begin{bmatrix}\vec{\gamma}_1\\ \vec{0}\\ \vec{0}\end{bmatrix} \\ &(X'X)^{-1}\begin{bmatrix}0 & Z_2' & 0\end{bmatrix}\begin{bmatrix}\hat{\epsilon}_1^*\\ \hat{\epsilon}_2^*\\ \hat{\epsilon}_3^*\end{bmatrix} \\ =&\hat{\beta}^* + (X'X)^{-1}X_2'Z_2\hat{\beta}^* - (X'X)^{-1}X_1'\vec{\gamma}_1 - (X'X)^{-1}Z_2'\hat{\epsilon}_2^*\end{aligned} \tag{7.18}$$

Bajo esta partición la ecuación (7.14) se expresa como

$$\begin{bmatrix}\hat{\epsilon}_1\\ \hat{\epsilon}_2\\ \hat{\epsilon}_3\end{bmatrix} = \begin{bmatrix}\hat{\epsilon}_1^*\\ \hat{\epsilon}_2^*\\ \hat{\epsilon}_3^*\end{bmatrix} + \begin{bmatrix}X_1\\ X_2\\ X_3\end{bmatrix}(X'X)^{-1}\begin{bmatrix}0 & Z_2' & 0\end{bmatrix}\begin{bmatrix}\hat{\epsilon}_1^*\\ \hat{\epsilon}_2^*\\ \hat{\epsilon}_3^*\end{bmatrix} +$$

$$\begin{bmatrix}I-H_{11} & -H_{12} & -H_{13}\\ -H_{21} & I-H_{22} & -H_{23}\\ -H_{31} & -H_{32} & I-H_{33}\end{bmatrix}\left[\begin{bmatrix}0\\ Z_2\\ 0\end{bmatrix}\hat{\beta}^* - \begin{bmatrix}\vec{\gamma}_1\\ 0\\ 0\end{bmatrix}\right]$$

$$= \begin{bmatrix}I & C_1'Z_2' & 0\\ 0 & I+C_2'Z_2' & 0\\ 0 & C_3'Z_2' & I\end{bmatrix}\begin{bmatrix}\hat{\epsilon}_1^*\\ \hat{\epsilon}_2^*\\ \hat{\epsilon}_3^*\end{bmatrix} + \begin{bmatrix}I-H_{11} & -H_{12} & -H_{13}\\ -H_{21} & I-H_{22} & -H_{23}\\ -H_{31} & -H_{32} & I-H_{33}\end{bmatrix}\begin{bmatrix}0\\ Z_2\hat{\beta}^*\\ 0\end{bmatrix} -$$

$$\begin{bmatrix}I-H_{11} & -H_{12} & -H_{13}\\ -H_{21} & I-H_{22} & -H_{23}\\ -H_{31} & -H_{32} & I-H_{33}\end{bmatrix}\begin{bmatrix}\vec{\gamma}_1\\ 0\\ 0\end{bmatrix}$$

con $C_i = (X'X)^{-1}X_i'$ y $H_{ij} = X_i(X'X)^{-1}X_j'$.

De tal manera que el vector $\vec{\gamma}_1$ que hace $\hat{\epsilon}_1^* = 0$ está dado por

$$\hat{\gamma}_1 = -(I - H_{11})^{-1}\hat{\epsilon}_1 \tag{7.19}$$

y el vector $Z_2\hat{\beta}^*$ que hace $\hat{\epsilon}_2^* = 0$ esta dado por

$$Z_2\hat{\beta}^* = (I - H_{22})^{-1}\hat{\epsilon}_2 \tag{7.20}$$

para despejar la matriz Z_2 se multiplica a ambos lados de (7.20) por $(\hat{\beta}^*)'$, es decir

$$Z_2\hat{\beta}^*(\hat{\beta}^*)' = (I - H_{22})^{-1}\hat{\epsilon}_2(\hat{\beta}^*)'$$

y luego se multiplica por $[\hat{\beta}^*(\hat{\beta}^*)']^{-1}$, la cual puede ser una inversa generalizada, para finalmente obtener que

$$Z_2 = (I - H_{22})^{-1}\hat{\epsilon}_2(\hat{\beta}^*)'[\hat{\beta}^*(\hat{\beta}^*)']^{-1}.$$

7.4. Metodología de imputación en ambas variables

Para el modelo $Y = X\vec{\beta} + \vec{\epsilon}$ con observaciones faltantes en la variable respuesta Y y en las variables explicativas X, la metodología de imputación que se recomienda en este trabajo sigue los siguientes pasos:

1. Organice los datos de tal manera que $\vec{Y}_1$, X_2 sean respectivamente el bloque y la matriz conformados por las observaciones consideradas faltantes, particionando el modelo según:
$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \\ \vec{Y}_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \\ \vec{\epsilon}_3 \end{bmatrix}$$

2. Calcule la estadística Q_k de cada observación para el modelo
$$\vec{Y}_3 = X_3\vec{\beta}_c + \vec{\epsilon} \qquad \text{con} \qquad \hat{\beta}_c = (X_3'X_3)^{-1}X_3'\vec{Y}_3.$$

3. Ordene en forma descendente el bloque $\left[\vec{Y}_3,\, X_3\right]$ tomando como criterio la estadística Q_k, es decir ubique las r-observaciones menos influyentes al final del bloque.

4. Use el ordenamiento hecho en 3. y organice nuevamente los datos en el modelo
$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \\ \vec{Y}_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \\ \vec{\epsilon}_3 \end{bmatrix}$$

5. Introduzca un conjunto A arbitrario de valores iniciales en el bloque $\vec{Y}_1$ y un conjunto B arbitrario de valores iniciales en la matriz X_2,
$$\vec{Y}_c = \begin{bmatrix} A \\ \vec{Y}_2 \\ \vec{Y}_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ B \\ X_3 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \\ \vec{\epsilon}_3 \end{bmatrix}$$

con el modelo completo $\vec{Y}_c = X_c\vec{\beta} + \vec{\epsilon}$, donde $X_c = \begin{bmatrix} X_1 \\ B \\ X_3 \end{bmatrix}$, calcule los residuales $\hat{\epsilon}$.

6. Con las observaciones que incluyen los valores iniciales que componen el bloque $\vec{Y}_1$, en la matriz X_2 y las observaciones completas de mayor influencia, conforme un vector $\vec{Y}_4$ y matriz X_4 tal que la variación porcentual que genera el bloque $\left[\vec{Y}_4,\, X_4\right]$ sea superior al 95 %. Y con este bloque calcule el vector $Z\hat{\beta}^* - \vec{\gamma}$ según

$$Z\hat{\beta}^* - \vec{\gamma} = (I - H_4)^{-1}\hat{\epsilon}$$

con $\vec{\epsilon}$ los residuales correspondientes al bloque $\left[\vec{Y}_4,\, X_4\right]$ calculados en el paso 5., y H_4 el bloque de la matriz H correspondiente al bloque $\left[\vec{Y}_4,\, X_4\right]$.

7. Calcule el bloque $\vec{Y}_1$ a imputar ajustando los valores iniciales del conjunto A, con los valores $\hat{\gamma}_1$ correspondientes al bloque $\vec{Y}_1$ obtenidos del vector $Z\hat{\beta}^* - \hat{\gamma}$ calculado en el paso 6., es decir

$$\vec{Y}_1 = A + \hat{\gamma}_1 \tag{7.21}$$

y la matriz X_2 a imputar ajustando los valores iniciales del conjunto B, con los valores Z_2 correspondientes a la matriz X_2 obtenidos del vector $Z\hat{\beta}^* - \hat{\gamma}$ calculado en el paso 6., es decir

$$X_2 = B + Z_2. \tag{7.22}$$

Bibliografía

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Supplement to the Journal of the Royal Statistical Society*, 4(2):137–183.

Behar, R. and Yepes, M. (1996). *Estadística: Un Enfoque Descriptivo*. Universidad del Valle.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. John Wiley, New York.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*. John Wiley & Sons, New York, 2nd edition.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Monographs on statistics and applied probability. Chapman & Hall, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society*, 39(1):1–38.

Draper, N. R. and John, J. A. (1981). Influential observations and outliers in regression. *Technometrics*, 23(1):21–26.

Eld, M. T., Black, C. A., Kempthorne, O., and Zoellner, J. A. (1954). Significance of soil organic phosphorus to plant growth. *Research Bulletin (Iowa Agriculture and Home Economics Experiment Station)*, 31(406).

Espinosa, Y. (1998). Un criterio de convergencia para algoritmos iterativos en la estimación de modelos de regresión lineal simple con información incompleta. Tesis de grado, Facultad de Ciencias. Departamento de Estadística. Universidad Nacional de Colombia. Sede Bogotá.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York.

Little, T. M. and Hills, F. J. (1972). *Statistical Methods in Agricultural Research*. AXT/ Agricultural Extension, University of California. University of California, Public Services Offices.

Mickey, M. R., Dunn, O. J., and Clark, V. (1967). Note on the use of stepwise regression in detecting outliers. *Computers and Biomedical Research*, 1(2):105–111.

Rincón, L. F. and López, L. A. (1997). Una generalización de la estadística DFbeta en modelos de regresión lineal simple. *Revista Colombiana de Estadística*, 18(35):27–39.

Rincón, T. (1999). Una propuesta para caracterizar observaciones influyentes en modelos de regresión lineal múltiple. Tesis de grado, Facultad de Ciencias. Departamento de Estadística. Universidad Nacional de Colombia. Sede Bogotá.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York.

Tukey, J. (1977). *Exploratory Data Analyse*. Addison Wesley, New York.