Anales de la Real Academia de Doctores
Volumen 4, pp. 267-297, 2000

EFECTO DE LA ELIMINACIÓN PROGRESIVA DE INDIVIDUOS ATÍPICOS EN LA REGRESIÓN POR ETAPAS

FRANCISCO JAVIER DÍAZ-LLANOS Y SAINZ-CALLEJA
Y CARMEN CERMEÑO CARRASCO

RESUMEN

Esta nota describe, de manera sucinta, el principio de la **regresión por etapas con retro-eliminación reiterada de individuos atípicos**. La ilustración, mediante ejemplos numéricos de dimensiones pequeñas lo cual permite seguir el proceso metodológico con una simple calculadora y las tablas de las funciones de distribución de las variables aleatorias: T_n Student-Fisher (1908) y χ^2_n de Helmert (1875) —muestra que: **la eliminación progresiva de individuos atípicos** uno a uno (12)— y no en grupos (13), puesto que conllevaría un tiempo de cálculo prohibitivo- **conduce, en ocasiones, a resultados muy distintos en cuanto a la ecuación de regresión lineal estimada**.

INTRODUCCIÓN

Entre los cinco métodos[1] que nos permiten retener las variables más correlacionadas con la variable a explicar por una parte y, las menos correlacionadas entre ellas por otra, el más difundido para la elaboración de un modelo de regresión lineal múltiple —a partir de un conjunto de variables cuantitativas— es el de la **regresión paso a paso (stepwise regression, régression pas à pas)**. Este método, basado en la noción del coeficiente de correlación parcial y en la contribución marginal de cada variable explicativa, está expuesto de forma didáctica, en BAILLARGEON (1985, pp. 182-199). Aunque éste es el más difundido, sin embargo, hemos preferido retener el de la **regresión por etapas (stagewise regression, régression par étage)** dado que, éste último, presenta la ventaja frente a los otros de que, en el proceso de elección de variables permite **minimizar las intercorrelaciones de las series explicadas por estudio del residuo**. En este sentido. PALM (27) indica que, las técnicas clásicas tales como: la selección progresiva, regresiva o paso a paso tienen en común el **«no garantizar jamás la obtención del mejor subconjunto de variables para un número dado de variables»**.

[1] Todas las regresiones posibles, eliminación progresiva, selección progresiva, regresión paso a paso y regresión por etapas.

Aunque el método de la **clásica regresión por etapas** contemplado en DRAPER y SMITH (1981, pp. 337-341), BOURBONNAIS y USUNIER (1992, pp. 119-121) y BOURBONNAIS (1998, pp. 105-109) ya se ha aplicado y comparado con el de la **regresión paso a paso** a principios de los años setenta por LUND (24) sin embargo, su difusión en cuanto a su exposición e implementación en un paquete de programas de estadística no ha sido la misma. Por tal motivo, **esta nota se articula en dos partes fundamentales**: la primera, consiste en la presentación de una exposición —lo más didáctica posible— de la **regresión por etapas con retro-eliminación inicial reiterada de individuos atípicos**. La segunda, consiste en la presentación de un CONJUNTO DE EJERCICIOS destinados a ilustrar el principio de dicha regresión.

PROCESO METODOLÓGICO

El proceso metodológico que proponemos en este artículo en el cual está incluido **la clásica regresión por etapas**, aunque aún **no está implementado íntegramente** en ningún paquete de programas de estadística, «SE PUEDE RECONSTRUIR» —pero ya no de forma automática— **haciendo uso de programas existentes** en el mercado tales como: el STATlab (9) y el STAT-ITCF (1).

La **clásica regresión por etapas** (6) permite **minimizar las intercorrelaciones de las variables exógenas de manera optimal**.

Partimos de, **una variable** (explicada, respuesta, exógena, dependiente) y, de **p variables** (explicativas, de control, endógenas, independientes, regresoras) cuantitativas.

FASES DEL PROCESO METODOLÓGICO

1. Construcción de la tabla de datos originales.

La tabla de datos debe guardar la siguiente forma:

y_1	x_{i1}	x_{i2}	x_{i3}	.	.	x_{ip}
y_1	x_{11}	x_{12}	x_{13}	.	.	x_{1p}
y_2	x_{21}	x_{22}	x_{23}	.	.	x_{2p}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
y_n	x_{n1}	x_{n2}	x_{n3}	.	.	x_{np}

2. Construcción de la matriz de correlaciones de BRAVAIS-PEARSON entre las variables explicativas.

La matriz de correlaciones viene definida de la siguiente manera:

$$R = D_p^{-\frac{1}{2}} P D_p^{-\frac{1}{2}}$$

donde,

$$P = X^T X - \frac{1}{n} X^T 1_n 1_n^T X$$

D_p: matriz constituida por los elementos de la diagonal de P.

$$1_n = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix} \in R^n$$

X: matriz que contiene las variables explicativas de dimensión nxp.

Una interpretación detallada del significado del coeficiente de correlación lineal de BRAVAIS-PEARSON se encuentra en DROESBEKE (1988, pp. 335-341)

Conviene evitar correlaciones entre variables explicativas con valor absoluto cercano a 1 y tratar de evitar situaciones en las que, la correlación entre las variables explicativas sea mayor que, la correlación de esas variables con la variable explicada.

3. **Test de detección de una posible presunción de multicolinealidad.**

Entre el test de KLEIN y el test de FARRAR y GLAUBER existen diferencias, en cuanto a la detección de una posible presunción de multicolinealidad (6). El principal motivo por el cual hemos retenido, en nuestro proceso metodológico, el test de FARRAR y GLAUBER es que, éste, **se basa en una ley de probabilidad mientras el de KLEIN no. La ley de probabilidad es la χ^2 de Helmert** (1875). El desarrollo de estos dos test, así como, un ejemplo en el cual, se muestran diferencias —en cuanto a sus resultados— se encuentran, de forma didáctica, en BOURBONNAIS (1998, pp. 100-103). Una exposición exhaustiva del test de FARRAR y GLAUBER se encuentra en (19). Con respecto al test de FARRAR y GLAUBER, el resultado negativo de una posible presunción de multicolinealidad, puede verse cambiado por una diminución en el nivel de significación, como se observa más adelante, en esta nota.

La **regla general de decisión del test de FARRAR y GLAUBER** extraída de BOURBONNAIS (1998, pp. 101-103) y adaptada a nuestra nomenclatura es la siguiente,

Si $\left[(n-1)-\frac{[2(p+1)+5]}{6}\right]\log_e[\mathrm{Det}(R)] \leq F^{-1}_{\chi^2_{\frac{p(p+1)}{2}}}(1-\alpha)$ se acepta H_o, no hay presunción de multicolinealidad.

Si $\left[(n-1)-\frac{[2(p+1)+5]}{6}\right]\log_e[\mathrm{Det}(R)] > \overset{-1}{F}_{\chi^2_{\frac{p(p+1)}{2}}}(1-\alpha)$ se rechaza H_o, hay presunción de multicolinealidad.

Donde,

n: es el número de individuos activos en el modelo de regresión lineal.

p: es el número de variables explicativas.

Det (R): es el determinante de la matriz de correlaciones entre las variables explicativas.

$\overset{-1}{F}_{\chi^2_{\frac{p(p+1)}{2}}}$:es la función inversa de la función de distribución de la variable aleatoria χ^2 de Helmert (1875) con p(p+1)/2 grados de libertad y para un área de (1-α).

4. **Estudio de la matriz simétrica definida positiva**.

$$\begin{pmatrix} 1_n^T\, 1_n & 1_n^T\, X_{(r)} \\ X_{(r)}^T\, 1_n & X_{(r)}^T\, X_{(r)} \end{pmatrix}$$

donde,

$$1_n = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix} \in R^n$$

$X_{(r)}$:es la matriz —no aleatoria— de dimensión (n, r+1) que contiene las variables explicativas retenidas en la r-ésima etapa.

Una matriz de esta forma va a ser —desde un punto de vista operativo— **el eje central del proceso** de **elección de variables**. Tendremos que invertirla tantas veces como variables retenidas, aumentando así su orden sucesivamente desde 2 hasta las variables retenidas más 1.

Cuando esta matriz sea singular es aconsejable llevar a cabo alguna de estas dos situaciones:

1. La eliminación de **una o más variables explicativas**.

2. La aplicación de alguna de las **técnicas de regresión puestas a punto para atenuar los efectos de la multicolinealidad**.

Entre estas técnicas, vamos a exponer —tan sólo— los principios de cinco de ellas. Tres de ellos, están basados en el **cálculo de nuevas variables explicativas**, siendo las siguientes:

- La regresión en función de las componentes principales.
- La regresión propuesta por WESTER, GUNST y MASSON (1974).
- La regresión por los mínimos cuadrados parciales propuesta por Harald MARTENS, Herald WOLD y Svante WOLD (1983).

Y las dos restantes en los estimadores «estrechos», siendo estas:

- La regresión pseudo-ortogonal (1970)
- La regresión utilizando los estimadores de JAMES y STEIN (1961)

Estas cinco técnicas están expuestas —de forma didáctica— en el artículo de PALM y IEMMA (28).

Una exposición exhaustiva de otras técnicas, derivadas de la **regresión con restricciones lineales y no lineales**, se encuentran contempladas en los artículos de CAZES (10,11).

Fases del procedimiento metodológico

1. Elegir como primera variable explicativa aquella cuyo coeficiente de correlación lineal con la variable sea el **máximo**.

	x_1	x_2	x_3	.	.	x_p
y	r_{y,x_1}	r_{y,x_2}	r_{y,x_3}	.	.	r_{y,x_p}

2. Consideremos que la variable elegida ha sido la x_1.

3. Contrastar, si el coeficiente de correlación poblacional entre la variable explicada y la variable explicativa elegida es —significativamente— diferente a cero.

4. El modelo lineal, bajo la forma matricial, a una variable explicativa y,n observaciones —en la primera etapa— puede escribirse de la siguiente forma:

$$\begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ . \\ . \\ y_n^{(1)} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ . & . \\ . & . \\ 1 & x_{n1} \end{pmatrix} \begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} + \begin{pmatrix} \epsilon_1^{(1)} \\ \epsilon_2^{(1)} \\ . \\ . \\ \epsilon_n^{(1)} \end{pmatrix}$$

donde,

$\begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ . \\ . \\ y_n^{(1)} \end{pmatrix} \in$: es un vector aleatorio observable con valores en R^n.

$\begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ . & . \\ . & . \\ 1 & x_{n1} \end{pmatrix} = \left(1_n \middle| X_{(1)}\right)$: matriz de datos de dimensión (n,2) que contiene:

$$1_n = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix} \in R^n$$

$X_{(1)}$: matriz de datos no aleatoria observada que contiene la primera variable retenida en la primera etapa.

$\begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} \in R^2$: es un vector no aleatorio no observable con valores en que contiene los parámetros desconocidos del modelo en la primera etapa.

$\begin{pmatrix} \epsilon_1^{(1)} \\ \epsilon_2^{(1)} \\ . \\ . \\ \epsilon_n^{(1)} \end{pmatrix} \in$: es un vector aleatorio no observable con valores en R^n.

A partir de este modelo, pretendemos estimar el vector no aleatorio —no observable— que contiene los parámetros del modelo lineal: $\beta^{(1)}$

$$\beta^{(1)} = \begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix}$$

Para ello, aplicamos el método de **mínimos cuadrados ordinarios** (MCO) que consiste en **minimizar la suma de cuadrados de los errores** (Bourbonnaís, 1998, pp. 49-50). Diferenciando esta **suma de cuadrados** con respecto a $\beta^{(1)}$ e igualando a cero obtenemos el vector aleatorio $\hat{\beta}^{(1)}$.

$$\hat{\beta}^{(1)} = \begin{pmatrix} \hat{\beta}_0^{(1)} \\ \hat{\beta}_1^{(1)} \end{pmatrix} = \begin{pmatrix} 1_n^T 1_n & 1_n^T X_{(1)} \\ X_{(1)}^T 1_n & X_{(1)}^T X_{(1)} \end{pmatrix}^{-1} \begin{pmatrix} 1_n^T \\ X_{(1)}^T \end{pmatrix} y^{(1)}$$

Dado que $\hat{\beta}^{(1)}$ es función de $y^{(1)}$, éste es un vector aleatorio y, por lo tanto, sus componentes $\hat{\beta}_0^{(1)}$ y $\hat{\beta}_1^{(1)}$, son variables aleatorias. Si en esta expresión sustituimos el vector aleatorio —observable— $y^{(1)}$ por un vector concreto, obtendremos las estimaciones de los parámetros $\hat{\beta}_0^{(1)}$ y $\hat{\beta}_1^{(1)}$. Estas estimaciones las expresamos por: $\hat{\beta}_0^{*(1)}$ y $\hat{\beta}_1^{*(1)}$.

Finalmente el modelo estimado —en la primera etapa— se expresa de la siguiente manera:

$$y^{*(1)} = \hat{\beta}_0^{*(1)} + \hat{\beta}_1^{*(1)} x_1$$

5. Cálculo de los siguientes indicadores:

a: residuos
b: residuos normalizados
c: residuos estudentizados
d: distancia de Dennis R. COOK

Estos cuatro indicadores —recogidos en MONTGOMERY y RUNGER (1996, pp. 571-579)— se encuentran adaptados a nuestra nomenclatura de la siguiente manera.

a: residuos $e^{(r)}, r = 1,2,..p$

El vector aleatorio residuo, correspondiente a la primera etapa, viene definido de la siguiente forma:

$$e^{(1)} = \begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ . \\ . \\ y_n^{(1)} \end{pmatrix} - \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ . & . \\ . & . \\ 1 & x_{n1} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0^{(1)} \\ \hat{\beta}_1^{(1)} \end{pmatrix}$$

donde,

$\begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ . \\ . \\ y_n^{(1)} \end{pmatrix} \in R^n$: es un vector aleatorio observable con valores en R^n asociado a la primera etapa.

$\begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ . & . \\ . & . \\ 1 & x_{n1} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0^{(1)} \\ \hat{\beta}_1^{(1)} \end{pmatrix} \in R^n$: es un vector aleatorio con vaores en R^n asociado a la primera etapa

cuya i-ésima observación la podemos expresar por, $e_i^{(1)}$

El vector residuo estimado —en la primera etapa— lo expresamos con la siguiente notación, $e^{*(1)}$

b: residuos normalizados: $e_s^{(r)}, r = 1,\ldots,p$

Los residuos normalizados se calculan a partir de los residuos.

El **vector residuo normalizado estimado** —en la primera etapa— lo expresamos con la siguiente notación, $e_s^{*(1)}$.

La i-ésima observación de este vector la calculamos utilizando la fórmula:

$$e_{s_i}^{*(1)} = \frac{e_i^{*(1)}}{\sqrt{\frac{\sum_{i=1}^{i=n}\left(e_i^{*(1)}\right)^2}{\left[n-(p+1)\right]}}}$$

c: *residuos estudentizados:* $e_t^{(r)}, r = 1,\ldots,p$

Los residuos estudentizados se calculan a partir de los residuos normalizados.

El **vector residuo estudentizado estimado** —en la primera etapa— lo expresamos con la siguiente notación, $e_t^{*(1)}$.

La i-ésima observación de este vector la calculamos utilizando la fórmula:

$$e_{t_i}^{*(1)} = \frac{e_{s_i}^{*(1)}}{\sqrt{\left[1-h_{ii}\right]}}$$

d: *Distancia de Dennis R. COOK:* $D^{(r)}, r = 1,\ldots,p$

La distancia de Dennis R. COOK se calcula a partir de los residuos estudentizados.

El vector distancia de COOK estimado —en la primera etapa— lo expresamos así: $D^{*(1)}$.

La i-ésima observación de este vector la calculamos utilizando la fórmula siguiente:

$$D_i^{*(1)} = \frac{\left[e_{t_i}^{*(1)}\right]^2 h_{ii}}{(p+1)\left[(1-h_{ii})\right]}$$

donde:

p: es el número de variables explicativas y,

h_{ii}: es el i-ésimo elemento que está en la diagonal principal de la matriz H

$$H = X(X^T X)^{-1} X^T$$

Siendo X la matriz del modelo lineal de la dimensión (n, p+1).

Entre estos cuatro indicadores —ya definidos— hemos de indicar que, en el proceso metodológico, hemos retenido el de la **distancia de COOK** (12). Un individuo se considera atípico cuando la **distancia de COOK** es mayor que 1 tal como, indican MONTGOMERY y RUNGER (1996, pp. 571-579). En este caso, procedemos a su eliminación e **reiniciaremos** el proceso metodológico con un individuo menos ya que, si no, **podría afectar en la elección del modelo lineal retenido**. Este punto está desarrollado en uno de los dos ejercicios propuestos en esta nota.

La distancia de COOK según indican PALM y IEMMA (27), es una función creciente del cuadrado del residuo y de una medida del alejamiento del individuo en relación al, centro de gravedad —de la nube de puntos— en el espacio de las variables explicadas, siquiera si su regresión comporta un término independiente. Esta distancia está directamente relacionada con la **distancia de MAHALANOBIS** entre un individuo y el vector de las medias (30). Una exposición didáctica de la **distancia de MAHALANOBIS** se encuentra en DAGNELIE (1977, pp. 227-250). Para una exposición exhaustiva ver el artículo de MAHALANOBIS (25).

Detección de una eventual dependencia de los errores.

Entre el test de DURBIN y WATSON y el test de BREUSCH-GODFREY, hemos retenido este último dado que, mientras el primero sólo permite detectar una autocorrelación de orden 1 el segundo, permite contrastar una autocorrelación de orden superior a 1. El test de BREUSCH-GODFREY se basa en la ley de probabilidad de la χ^2 de Helmert (1985). El desarrollo de este test, así como, una comparación con el DURBIN y WATSON, se encuentra —de forma didáctica— en BOURBONNAIS (1998, pp. 116-120). Para una exposición exhaustiva del test de BREUSCH-GODFREY revisar (8,22).

Test para detectar la normalidad de los errores.

Entre los tests de asimetría y kurtosis, y el test de JARQUE y BERA, retenemos el de JARQUE y BERA dado que, esta fundamentado en la noción de asimetría y de kurtosis. Este test, permite verificar la normalidad de una distribución estadística. El test de JARQUE y BERA, al igual que el test de FARRAR y GLAUBER, se basan en la ley de probabilidad de la χ^2 de Helmert (1985). El desarrollo de este test, se encuentra —de forma didáctica— en BOURBONNAIS (1998, p. 220). Para una exposición exhaustiva del test ver los artículos de JARQUE y BERA (4,5).

El proceso metodológico continúa en el caso que estos dos test den negativos.

6. Elegir como segunda variable explicativa, aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual estimada —en la primera etapa— **sea máximo**.

	x_1	x_2	x_3	.	.	x_p
$e^{*(1)}$	$r_{e^{*(1)},x_1}$	$r_{e^{*(1)},x_2}$	$r_{e^{*(1)},x_3}$	.	.	$r_{e^{*(1)},x_p}$

6. Consideremos que la variable elegida ha sido la x_2.

7. Contrastar, si el coeficiente de correlación pblacional entre la variable residual estimada —en la primera etapa— y la variable explicada elegida es —significativamente— diferente de cero.

La regla general de decisión asociada al test de hipótesis bilateral —para el coeficiente de correlación poblacional— que a continuación presentamos,

$$H_0: \rho_{e^{*(1)},x_2} = 0$$

$$H_1: \rho_{e^{*(1)},x_2} \neq 0$$

adaptada a nuestra nomenclatura es la siguiente:

$$\text{Si } \overset{-1}{F}_{T_{n-2}}\left(\frac{\alpha}{2}\right) \leq \sqrt{n-2}\,\frac{r_{e^{*(1)},\,x_2}}{\sqrt{1-r^2_{e^{*(1)},\,x_2}}} \leq \overset{-1}{F}_{T_{n-2}}\left(1-\frac{\alpha}{2}\right) \text{ se acepta } H_0$$

$$\text{Si } \overset{-1}{F}_{T_{n-2}}\left(\frac{\alpha}{2}\right) > \sqrt{n-2}\,\frac{r_{e^{*(1)},\,x_2}}{\sqrt{1-r^2_{e^{*(1)},\,x_2}}} > \overset{-1}{F}_{T_{n-2}}\left(1-\frac{\alpha}{2}\right) \text{ se acepta } H_1$$

donde,

$$\overset{-1}{F}_{T_{n-2}}\left(\frac{\alpha}{2}\right) \text{ y } \overset{-1}{F}_{T_{n-2}}\left(1-\frac{\alpha}{2}\right)$$

representan las funciones inversas de la función de distribución de la variable aleatoria T de Student-Fisher con n-2 grados de libertad para un área de

$$\frac{\alpha}{2} \text{ y } 1-\frac{\alpha}{2} \text{ respectivamente.}$$

El proceso de elección de variables continúa mientras que, se vaya rechazando la hipótesis nula. Consideraremos que se rechaza la hipótesis nula para que el proceso continúe.

8. El modelo lineal, bajo la forma matricial, a dos variables explicativas y, n observaciones —en la segunda etapa— puede escribirse de la siguiente forma:

$$\begin{pmatrix} y_1^{(2)} \\ y_2^{(2)} \\ . \\ . \\ y_n^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ & . & \\ & . & \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0^{(2)} \\ \beta_1^{(2)} \\ \beta_2^{(2)} \end{pmatrix} + \begin{pmatrix} \epsilon_1^{(2)} \\ \epsilon_2^{(2)} \\ . \\ . \\ \epsilon_n^{(2)} \end{pmatrix}$$

Para evitar posibles reiteraciones, el significado de las matrices que aparecen en este fórmula es el mismo que en la primera etapa. Lo único que varía son: las dimensiones de la matriz de datos y la de parámetros.

A partir de este modelo, pretendemos estimar el vector no aleatorio —no observable— que contiene los parámetros del modelo lineal a estimar $\beta^{(2)}$:

$$\beta^{(2)} = \begin{pmatrix} \beta_0^{(2)} \\ \beta_1^{(2)} \\ \beta_2^{(2)} \end{pmatrix}$$

Para ello, aplicamos el método de mínimos cuadrados ordinarios (MCO) y obtenemos,

$$\hat{\beta}^{(2)} = \begin{pmatrix} \hat{\beta}_0^{(2)} \\ \hat{\beta}_1^{(2)} \\ \hat{\beta}_2^{(2)} \end{pmatrix} = \begin{pmatrix} 1_n^T 1_n & 1_n^T X_{(2)} \\ X_{(2)}^T 1_n & X_{(2)}^T X_{(2)} \end{pmatrix}^{-1} \begin{pmatrix} 1_n^T \\ X_{(2)}^T \end{pmatrix} y^{(2)}$$

Dado que $\hat{\beta}^{(2)}$ es función de $y^{(2)}$, éste, es un vector aleatorio y, por lo tanto, sus componentes $\hat{\beta}_0^{(2)}$, $\hat{\beta}_1^{(2)}$ y $\hat{\beta}_2^{(2)}$, son variables aleatorias. Si en esta expresión sustituimos el vector aleatorio —observable— $y^{(2)}$ por un vector concreto, obtendremos las estimaciones de los parámetros. $\beta_0^{(2)}$, $\beta_1^{(2)}$ y $\beta_2^{(2)}$. Estas estimaciones las expresamos por: $\hat{\beta}^{*(2)}_0$, $\hat{\beta}^{*(2)}_1$ y $\hat{\beta}^{*(2)}_2$.

Finalmente el modelo estimado —en la segunda etapa— se expresa en la siguiente manera,

$$y^{*(2)} = \hat{\beta}_0^{*(2)} + \hat{\beta}_1^{*(2)} x_1 + \hat{\beta}_2^{*(2)} x_2$$

9. Cálculo de los siguientes indicadores

   * residuos

   * residuos normalizados

   * residuos estudentizados

   * distancia de Dennis R. COOK

Para la i-ésima observación el residuo correspondiente —en la segunda etapa— se expresa de la siguiente manera: $e_i^{(2)}$

Dado que, en esta etapa se aplican las mismas fórmulas que en la primera, **omitimos** las fórmulas para el cálculo de los nuevos indicadores.

Si en esta etapa, hubiese habido al menos un individuo atípico cuya distancia de COOK es mayor que 1, habría que eliminarlo y, reiniciar el proceso metodológico con un individuo menos.

10. Elegir como tercera variable explicativa, aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual estimada —en la segunda etapa— **sea máximo**.

	x_1	x_2	x_3	x_4	.	x_p
$e^{*(2)}$	$r_{e^{*(2)},x_1}$	$r_{e^{*(2)},x_2}$	$r_{e^{*(2)},x_3}$	$r_{e^{*(2)},x_4}$	.	$r_{e^{*(2)},x_p}$

11. Consideremos que la variable elegida ha sido x_3.

12. Contrastar, si el coeficiente de correlación poblacional entre la variable residual estimada en la segunda etapa y la variable explicada elegida es —significativamente— diferente a cero.

Igualmente, dado que, en esta etapa se aplican las mismas fórmulas que en la primera, omitimos las fórmulas que nos permiten tomar la decisión de seguir rechazando la hipótesis nula.

El proceso de elección de variables continuará mientras se vaya rechazando la hipótesis nula. Consideremos que se rechaza la hipótesis nula para que el proceso continúe.

13. El modelo lineal, bajo la forma matricial, a tres variables explicativas y, n observaciones —en la tercera etapa— puede escribirse de la siguiente forma:

$$\begin{pmatrix} y_1^{(3)} \\ y_2^{(3)} \\ y_3^{(3)} \\ . \\ . \\ y_n^{(3)} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ & & . & \\ & & . & \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0^{(3)} \\ \beta_1^{(3)} \\ \beta_2^{(3)} \\ \beta_3^{(3)} \end{pmatrix} + \begin{pmatrix} \in_1^{(3)} \\ \in_2^{(3)} \\ . \\ . \\ \in_n^{(3)} \end{pmatrix}$$

A partir de éste modelo, pretendemos estimar el vector no aleatorio —no observable— que contiene los parámetros del modelo lineal a estimar: $\beta_0^{(3)}$

$$\beta^{(3)} = \begin{pmatrix} \beta_0^{(3)} \\ \beta_1^{(3)} \\ \beta_2^{(3)} \\ \beta_3^{(3)} \end{pmatrix}$$

Para ello, aplicamos el método de mínimos cuadrados ordinarios (MCO) y obtenemos,

$$\hat{\beta}^{(3)} = \begin{pmatrix} \hat{\beta}_0^{(3)} \\ \hat{\beta}_1^{(3)} \\ \hat{\beta}_2^{(3)} \\ \hat{\beta}_2^{(3)} \end{pmatrix} = \begin{pmatrix} 1_n^T 1_n & 1_n^T X_{(3)} \\ X_{(3)}^T 1_n & X_{(3)}^T X_{(3)} \end{pmatrix}^{-1} \begin{pmatrix} 1_n^T \\ X_{(3)}^T \end{pmatrix} y^3$$

Dado que $\hat{\beta}^{(3)}$ es función de y^3, éste, es un vector aleatorio y, por tanto, sus componentes $\hat{\beta}_0^{(3)}$, $\hat{\beta}_1^{(3)}$, $\hat{\beta}_2^{(3)}$ y $\hat{\beta}_3^{(3)}$, son variables aleatorias. Si en esta expresión sustituimos el vector aleatorio —observable— y^3 por un vector concreto, obtendremos las estimaciones de los parámetros $\beta_0^{(3)}$, $\beta_1^{(3)}$, $\beta_2^{(3)}$, y $\beta_3^{(3)}$. Estas estimaciones las expresamos por: $\hat{\beta}_0^{*(3)}$, $\hat{\beta}_1^{*(3)}$, $\hat{\beta}_2^{*(3)}$ y $\hat{\beta}_3^{*(3)}$.

Finalmente el modelo estimado —en la tercera etapa— se expresa de la siguiente manera:

$$y^{*(3)} = \hat{\beta}_0^{*(3)} + \hat{\beta}_1^{*(3)} x_1 + \hat{\beta}_2^{*(3)} x_2 + \hat{\beta}_3^{*(3)} x_3$$

14. Cálculo de los siguientes indicadores:

* residuos

* residuos normalizados

* residuos estudentizados

* distancia de Dennis R. COOK

Para la i-ésima observación el residuo corrspondiente —en la tercera etapa— viene dado mediante la siguiente expresión $e_i^{(3)}$:

Dado que, en esta etapa se aplican las mismas fórmulas que en la primera etapa, omitimos las fórmulas para el cálculo de los nuevos indicadores.

Si en esta etapa hubiese habido al menos un individuo cuya distancia de COOK es mayor que 1 habría que eliminarlo y reiniciar el proceso metodológico con un individuo menos.

15. Elegir como cuarta variable explicativa aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual estimada —en la tercera etapa— sea **máximo**.

	x_1	x_2	x_3	x_4	.	x_p
$e^{*(3)}$	$r_{e^{*(3)},x_1}$	$r_{e^{*(3)},x_2}$	$r_{e^{*(3)},x_3}$	$r_{e^{*(3)},x_4}$	.	$r_{e^{*(3)},x_p}$

16. Consideramos que la variable elegida ha sido la x_4.

17. Contrastar, si el coeficiente de correlación poblacional entre la variable residual estimada —en la tercera etapa— y la variable explicada elegida es —significativamente— diferente de cero.

Dado que, en esta primera etapa se aplican las mismas fórmulas que en la primera etapa, omitimos las fórmulas que nos permiten tomar la decisión de seguir rechazando la hipótesis nula.

El proceso de elección de variables se detiene dado que, aceptamos la hipótesis nula.

Así pues, la ecuación de regresión lineal estimada retenida, se expresa de la siguiente manera:

$$y^{*(3)} = \hat{\beta}_0^{*(3)} + \hat{\beta}_1^{*(3)} x_1 + \hat{\beta}_2^{*(3)} x_2 + \hat{\beta}_3^{*(3)} x_3$$

Consideraciones sobre el número de variables explicativas

Estudios realizados por FURNIVAL y WILSON (1971, 1974) recomiendan que, cuando el número de variables explicativas es superior a 40, es aconsejable utilizar la técnica de la **regresión en componentes principales** que a su vez, actúa contra las variables que están —significativamente— correlacionadas entre si. Independientemente del número de variables explicativas LEBART, MORINEAU y PIRON (1995, pp. 233-237) recomiendan la realización de la regresión considerando, como variables explicativas, las componentes principales más significativas.

Consideraciones sobre la validez de los resultados de una regresión

Entre los dos criterios que **miden la calidad de una ecuación de regresión: 1, la desviación típica residual y 2, el coeficiente de determinación múltiple ajustado**, el segundo, puede conducir a errores de interpretación ya que, la ecuación que tiene el coeficiente de determinación múltiple más grande, no tiene —necesariamente— la varianza residual más débil. El valor más elevado del coeficiente puede ser debido, en efecto, a una varianza **más** grande de las variables explicadas.

Para evitar la situación de que la varianza residual estimada pudiera dar —en la práctica— una idea demasiado optimal de la calidad de la ecuación de regresión-cuando los mismos datos sirvan para establecer la ecuación y cifrar su previsión recomendamos, «técnicas basadas en la separación de los datos en dos o más conjuntos y en la validación cruzada». La solución más directa consiste en, repartir los datos en dos grupos: un grupo servirá para determinar el modelo y, el otro para apreciar la calidad del modelo, calculando —por ejemplo— el cuadrado medio del error de predicción, si el objetivo fuera el de definir la calidad predictiva del modelo. Esta validación es aceptada si el tamaño de la muestra es suficiente. SNEE (1977) considera que, el

tamaño debe ser superior a 2p+20, siendo p el mayor gran número de coeficientes, susceptibles de ser introducidos en la ecuación.

Nota: el modelo es correcto si los residuos reducidos estimados se encuentran entre -2 y 2 tal como indican TOMASSONE y colaboradores (1992, p. 24).

Ejercicios destinados a ilustrar el proceso metodológico

Los dos ejercicios que tratamos, no constituyen más que, un soporte —práctico— destinado a ilustrar los principios expuestos.

Primer ejercicio

En este ejercicio aplicamos el proceso metodológico propuesto y observamos que, el test de FARRAR y GLAUBER conduce a resultados distintos cuando variamos ligeramente el nivel de signifcación.

Proceso preparatorio al procedimiento metodológico

1. Construcción de la tabla de datos originales

x_i	x_{i1}	x_{i2}	x_{i3}
15	1	2	3
31	2	5	6
37	3	6	7
49	4	7	10
57	5	9	11

2. Construcción de la matriz de correlaciones de BRAVAIS-PEARSON entre las variables explicativas.

	x_1	x_2	x_3
x_1	1	0,9774	0,9853
x_2	—	1	0,9751
x_3	—	—	1

3. Test de detección de una posible presunción de multicolinealidad.

Como resultado de la aplicación de la regla general de decisión del test de FARRAR y GLAUBER concluimos que, para:

a) un nivel de significación de 0,05

$$12{,}3944 < 12{,}592$$

y por tanto, **aceptamos la hipótesis nula**; es decir, **no hay presunción de multicolinealidad.**

b) un nivel de significación de 0,01

$$12{,}3944 > 10{,}645$$

y por tanto, **rechazamos la hipótesis nula**; es decir, **hay presunción de multicolinealidad**.

Fases del procedimiento metodológico

1. Elegir como primera variable explicativa, aquella cuyo coeficiente de correlación lineal con la variable explicada sea el **máximo**.

	x_1	x_2	x_3
y	0,9903	0,9893	0,9969

2. La variable elegida ha sido la x_3.

3. Contrastar, si el coeficiente de correlación poblacional entre la variable explicada y la variable explicativa elegida, es —significativamente— diferente a cero.

El test de hipótesis que hay que realizar es el siguiente,

$$H_0: \rho_{y,x_3} = 0$$

$$H_3: \rho_{y,x_3} \neq 0$$

Como resultado de la apliación de la regla general de decisión de dicho test concluimos que, como:

$$-3{,}182 > 21{,}9459 > 3{,}182$$

Para un nivel de significación de 0,05 **rechazamos la hipótesis nula** y, por consiguiente, **el proceso de elección de variables continúa.**

4. El vector aleatorio $\hat{\beta}^{(1)}$ en nuestro caso concreto, está definido de la siguiente manera:

$$\begin{pmatrix}\hat{\beta}_0^{(1)}\\ \hat{\beta}_3^{(1)}\end{pmatrix}=\begin{pmatrix}5 & 37\\ 37 & 315\end{pmatrix}^{-1}\begin{pmatrix}1 & 1 & 1 & 1 & 1\\ 3 & 6 & 7 & 10 & 11\end{pmatrix}\begin{pmatrix}y_1^{(1)}\\ y_2^{(1)}\\ y_3^{(1)}\\ y_4^{(1)}\\ y_5^{(1)}\end{pmatrix}$$

Haciendo las operaciones de inversión y multiplicación de matrices deducimos fácilmente que,

$$\begin{pmatrix}\hat{\beta}_0^{(1)}\\ \hat{\beta}_3^{(1)}\end{pmatrix}=\frac{1}{206}\begin{pmatrix}204 & 93 & 56 & -55 & -92\\ -22 & -7 & -2 & 13 & 18\end{pmatrix}\begin{pmatrix}y_1^{(1)}\\ y_2^{(1)}\\ y_3^{(1)}\\ y_4^{(1)}\\ y_5^{(1)}\end{pmatrix}$$

Por consiguiente, la estimación de dicho vector se calcula sustituyendo el vector aleatorio —observable— $y^{(1)}$ por el vector realización.

$$\begin{pmatrix}\hat{\beta}_0^{*(1)}\\ \hat{\beta}_3^{*(1)}\end{pmatrix}=\frac{1}{206}\begin{pmatrix}204 & 93 & 56 & -55 & -92\\ -22 & -7 & -2 & 13 & 18\end{pmatrix}\begin{pmatrix}15\\ 31\\ 37\\ 49\\ 57\end{pmatrix}$$

Llegando al resultado final:

$$\begin{pmatrix}\hat{\beta}_0^{*(1)}\\ \hat{\beta}_3^{*(1)}\end{pmatrix}=\begin{pmatrix}0{,}3689\\ 5{,}0583\end{pmatrix}$$

De lo que se desprende que, la ecuación de regresión estimada, adopta la siguiente forma:

$$y^{*(1)} = 0{,}3689 + 5{,}0583\ x_3$$

5. El cálculo de las estimaciones de los cuatro indicadores expresados —en la primera etapa— se refleja en la siguiente tabla:

y_i	Residuos estimados	Residuos normalizados estimados	Residuos estudentizados estimados	Distancia de COOK estimada
15	–0,5438	–0,3650	–0,6353	0,4095
31	0,2813	0,1888	0,2177	0,0078
37	1,2230	0,8208	0,9199	0,1084
49	–1,9519	–1,3100	–1,6427	0,7725
57	0,9898	0,6643	0,9534	0,4818

El hecho de que, no se detecte ninguna distancia de COOK superior a 1 nos indica que **«no hay ningún individuo atípico»** y, por consiguiente, no es necesario **reiniciar** el proceso metodológico.

6. Elegimos como segunda variable explicativa, aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual estimada —en la primera etapa— sea **máximo**.

	x_1	x_2	x_3
$e^{*(1)}$	0,1016	0,2172	0,0000

7. La variable elegida ha sido la x_2.

8. Contrastar, si el coeficiente de correlación poblacional entre la variable residual estimada —en la primera etapa— y la variable explicada elegida es —significativamente— diferente de cero.

El test de hipótesis que hay que realizar es el siguiente:

$$H_0: \rho_{e^{*(1)},x_2} = 0$$

$$H_1: \rho_{e^{*(1)},x_2} \neq 0$$

Como resultado de la aplicación de la regla general de decisión de dicho test concluimos que, como:

$$-3{,}182 < 0{,}3854 < 3{,}182$$

Para un nivel de significación de 0,05 **aceptamos la hipótesis nula** y, por consiguiente, el **proceso de elección de variables se para**.

Así pues, el modelo estimado se expresa por,

$$y^{*(1)} = 0{,}3689 + 5{,}0583\, x_3$$

Segundo ejercicio

En este ejercicio, aplicamos el proceso metodológico bajo dos situaciones.

Primera: consideramos —«a priori»— que, no hay ningún individuo atípico.

Segunda: consideramos que en el supuesto de existencia de atipicidad individual se **reiniciará** el proceso.

Primera situación: en este ejercicio, vamos a mostrar —tan sólo— el proceso metodológico, sin la inclusión de los tres tests de hipótesis, contenidos en el proceso

metodológico: test de FARRAR y GLAUBER, test de BREUSCH-GODFREY y test de JARQUE y BERA. El cálculo de la distancia de COOK, también la hemos omitido ya que, hemos considerado «a priori» que no hay ningún individuo atípico.

Proceso preparatorio al procedimiento metodológico

1. Construcción de la tabla de datos originales

y_i	x_{i1}	x_{i2}	x_{i3}
1	–3	5	–1
0	–2	0	1
0	–1	–3	1
1	0	–4	0
2	1	–3	–1
3	2	0	–1
3	3	5	1

2. Construcción de la matriz de correlaciones de BRAVAIS-PEARSON entre las variables explicativas.

	x_1	x_2	x_3
x_1	1	0	0
x_2	—	1	0
x_3	—	—	1

Fases del proceso metodológico

1. Elegir como primera variable explicativa, aquella cuyo coeficiente de correlación lineal con la variable explicada sea el **máximo**.

	x_1	x_2	x_3
y	0,8489	0,3501	–0,3930

2. La variable elegida ha sido la x_1.

3. Contrastar, si el coeficiente de correlación poblacional entre la variable explicada y la variable explicativa elegida es —significativamente— diferente de cero.

El test de hipótesis que hay que realizar es el siguiente.

$$H_0: \rho_{y,x_1} = 0$$

$$H_1: \rho_{y,x_1} \neq 0$$

Como resultado de la aplicación de la **regla general de decisión** de dicho test concluimos que, como:

$$-2{,}015 > 3{,}5913 > 2{,}015$$

Para un nivel de significación del 0,1 **rechazamos la hipótesis nula** y, por consiguiente, **el proceso de elección de variables continúa**.

4. El vector aleatorio $\hat{\beta}^{(1)}$, en nuestro caso concreto, está definido de la siguiente manera:

$$\begin{pmatrix} \hat{\beta}_0^{(1)} \\ \hat{\beta}_1^{(1)} \end{pmatrix} = \begin{pmatrix} 7 & 0 \\ 0 & 28 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}^T \begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ y_3^{(1)} \\ y_4^{(1)} \\ y_5^{(1)} \\ y_6^{(1)} \\ y_7^{(1)} \end{pmatrix}$$

Haciendo uso de las operaciones de inversión y de multiplicación de matrices deducimos fácilmente que,

$$\begin{pmatrix} \hat{\beta}_0^{(1)} \\ \hat{\beta}_1^{(1)} \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 1 & -\frac{3}{4} \\ 1 & -\frac{2}{4} \\ 1 & -\frac{1}{4} \\ 1 & 0 \\ 1 & \frac{1}{4} \\ 1 & \frac{2}{4} \\ 1 & \frac{3}{4} \end{pmatrix}^T \begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ y_3^{(1)} \\ y_4^{(1)} \\ y_5^{(1)} \\ y_6^{(1)} \\ y_7^{(1)} \end{pmatrix}$$

Por consiguiente, la estimación de dicho vector se calcula sustituyendo el vector aleatorio —observable— $y^{(1)}$ por el vector realización.

$$\begin{pmatrix} \hat{\beta}_0^{*(1)} \\ \hat{\beta}_1^{*(1)} \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 1 & -\frac{3}{4} \\ 1 & -\frac{2}{4} \\ 1 & -\frac{1}{4} \\ 1 & 0 \\ 1 & \frac{1}{4} \\ 1 & \frac{2}{4} \\ 1 & \frac{3}{4} \end{pmatrix}^T \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 2 \\ 3 \\ 3 \end{pmatrix}$$

Llegando al resultado final,

$$\begin{pmatrix} \hat{\beta}_0^{*(1)} \\ \hat{\beta}_1^{*(1)} \end{pmatrix} = \begin{pmatrix} 1{,}4286 \\ 0{,}5000 \end{pmatrix}$$

De lo que se desprende que, la ecuación de regresión estimada —en la primera etapa— adopta la siguiente forma,

$$y^{*(1)} = 1{,}4286 + 0{,}5000\ x_1$$

5. Cálculo del vector residuos estimado —en la primera etapa— mediante la ecuación de regresión.

$$e^{*(1)} = \begin{pmatrix} 1{,}0714 \\ -0{,}4286 \\ -0{,}9286 \\ -0{,}4286 \\ 0{,}0714 \\ 0{,}5714 \\ 0{,}0714 \end{pmatrix}$$

6. Elegir como segunda variable explicativa, aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual estimada —en la primera etapa— sea **máximo**.

	x_1	x_2	x_3
$e^{*(1)}$	0,0000	0,6623	–0,7406

7. La variable elegida ha sido x_3.

8. Contrastar, si el coeficiente de correlación poblacional entre la variable residual estimada —en la primera etapa— y la variable explicativa elegida es —significativamente— diferente de cero.

El test de hipótesis que hay que realizar es el siguiente.

$$H_0: \rho_{e^{*(1)},x_3} = 0$$
$$H_1: \rho_{e^{*(1)},x_3} \neq 0$$

Como resultado de la aplicación de la **regla general de decisión** de dicho test concluimos que, como:

$$-2{,}015 > 2{,}4645 > 2{,}015$$

Para un nivel de significación del 0,1 **rechazamos la hipótesis nula** y, por consiguiente, **el proceso de elección de variables continúa**.

9. El vector aleatorio $\hat{\beta}^{(2)}$, en este caso concreto, está definido de la siguiente manera,

$$\begin{pmatrix} \hat{\beta}_0^{(2)} \\ \hat{\beta}_1^{(2)} \\ \hat{\beta}_3^{(2)} \end{pmatrix} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 28 & 0 \\ 0 & 0 & 6 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -3 & -1 \\ 1 & -2 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & -1 \\ 1 & 2 & -1 \\ 1 & 3 & 1 \end{pmatrix}^{T} \begin{pmatrix} y_1^{(2)} \\ y_1^{(2)} \\ y_2^{(2)} \\ y_3^{(2)} \\ y_4^{(2)} \\ y_5^{(2)} \\ y_6^{(2)} \\ y_7^{(2)} \end{pmatrix}$$

Haciendo uso de las operaciones de inversión y de multiplicación de matrices deducimos fácilmente que,

$$\begin{pmatrix} \hat{\beta}_0^{(2)} \\ \hat{\beta}_1^{(2)} \\ \hat{\beta}_3^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{7} & -\frac{3}{28} & -\frac{1}{6} \\ \frac{1}{7} & -\frac{2}{28} & \frac{1}{6} \\ \frac{1}{7} & -\frac{1}{28} & \frac{1}{6} \\ \frac{1}{7} & 0 & 0 \\ \frac{1}{7} & \frac{1}{28} & -\frac{1}{6} \\ \frac{1}{7} & \frac{2}{28} & -\frac{1}{6} \\ \frac{1}{7} & \frac{3}{28} & \frac{1}{6} \end{pmatrix}^T \begin{pmatrix} y_1^{(2)} \\ y_2^{(2)} \\ y_3^{(2)} \\ y_4^{(2)} \\ y_5^{(2)} \\ y_6^{(2)} \\ y_7^{(2)} \end{pmatrix}$$

Por consiguiente, la estimación de dicho vector se calcula sustituyendo el vector aleatorio —observable— $y^{(2)}$ por el vector realización.

$$\begin{pmatrix} \hat{\beta}_0^{*(2)} \\ \hat{\beta}_1^{*(2)} \\ \hat{\beta}_3^{*(2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{7} & -\frac{3}{28} & -\frac{1}{6} \\ \frac{1}{7} & -\frac{2}{28} & \frac{1}{6} \\ \frac{1}{7} & -\frac{1}{28} & \frac{1}{6} \\ \frac{1}{7} & 0 & 0 \\ \frac{1}{7} & \frac{1}{28} & -\frac{1}{6} \\ \frac{1}{7} & \frac{2}{28} & -\frac{1}{6} \\ \frac{1}{7} & \frac{3}{28} & \frac{1}{6} \end{pmatrix}^T \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 2 \\ 3 \\ 3 \end{pmatrix}$$

llegando el resultado final,

$$\begin{pmatrix} \hat{\beta}_0^{*(2)} \\ \hat{\beta}_1^{*(2)} \\ \hat{\beta}_3^{*(2)} \end{pmatrix} = \begin{pmatrix} 1{,}4286 \\ 0{,}5000 \\ -0{,}5000 \end{pmatrix}$$

De lo que se desprende que, la ecuación de regresión estimada —en la segunda etapa— adopta la siguiente forma.

$$y^{*(2)} = 1{,}4286 + 0{,}5000\ x_1 - 0{,}5000\ x_3$$

10. Cálculo del vector residuo estimado —en la segunda etapa— mediante la ecuación de regresión.

$$e^{*(2)} = \begin{pmatrix} 0{,}5714 \\ 0{,}0714 \\ -0{,}4286 \\ -0{,}4286 \\ -0{,}4286 \\ 0{,}0714 \\ 0{,}5714 \end{pmatrix}$$

11. Elegir como tercera variable explicativa, aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual —en la segunda etapa— sea **máximo**.

	x_1	x_2	x_3
$e^{*(2)}$	0,0000	0,9901	0,0000

12. La variable elegida ha sido la x_2

13. Contrastar, si el coeficiente de correlación poblacional entre la variable residual estimada —en la segunda etapa— y la variable explicativa elegida es ·significativamente— diferente de cero.

El test de hipótesis que hay que realizar es el siguiente:

$$H_0: \rho_{e^{*(2)},x_2} = 0$$
$$H_1: \rho_{e^{*(2)},x_2} \neq 0$$

Como resultado de la aplicación de la **regla general de decisión** de dicho test concluimos que, como:

$$-2{,}015 > 15{,}6926 < 2{,}015$$

Para un nivel de significación del 0,1 **rechazamos la hipótesis nula** y, por consiguiente, **el proceso de elección de variables continúa**.

14. El vector aleatorio $\hat{\beta}^{(3)}$, en nuestro caso concreto, está definido por la siguiente manera.

$$\begin{pmatrix} \hat{\beta}_0^{(3)} \\ \hat{\beta}_1^{(3)} \\ \hat{\beta}_3^{(3)} \\ \hat{\beta}_2^{(3)} \end{pmatrix} = \begin{pmatrix} 7 & 0 & 0 & 0 \\ 0 & 28 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 84 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -3 & -1 & 5 \\ 1 & -2 & 1 & 0 \\ 1 & -1 & 1 & -3 \\ 1 & 0 & 0 & -4 \\ 1 & 1 & -1 & -3 \\ 1 & 2 & -1 & 0 \\ 1 & 3 & 1 & 5 \end{pmatrix} \begin{pmatrix} y_1^{(3)} \\ y_2^{(3)} \\ y_3^{(3)} \\ y_4^{(3)} \\ y_5^{(3)} \\ y_6^{(3)} \\ y_7^{(3)} \end{pmatrix}$$

haciendo uso de las operaciones de inversión y de multiplicación de matrices deducimos fácilmente que,

$$\begin{pmatrix} \hat{\beta}_0^{(3)} \\ \hat{\beta}_1^{(3)} \\ \hat{\beta}_3^{(3)} \\ \hat{\beta}_2^{(3)} \end{pmatrix} = \begin{pmatrix} \frac{1}{7} & -\frac{3}{28} & -\frac{1}{6} & \frac{5}{54} \\ \frac{1}{7} & -\frac{2}{28} & \frac{1}{6} & 0 \\ \frac{1}{7} & -\frac{1}{28} & \frac{1}{6} & -\frac{3}{84} \\ \frac{1}{7} & 0 & 0 & -\frac{4}{84} \\ \frac{1}{7} & \frac{1}{28} & -\frac{1}{6} & -\frac{3}{84} \\ \frac{1}{7} & \frac{2}{28} & -\frac{1}{6} & 0 \\ \frac{1}{7} & \frac{3}{28} & \frac{1}{6} & \frac{5}{84} \end{pmatrix}^T \begin{pmatrix} y_1^{(3)} \\ y_2^{(3)} \\ y_3^{(3)} \\ y_4^{(3)} \\ y_5^{(3)} \\ y_6^{(3)} \\ y_7^{(3)} \end{pmatrix}$$

Por consiguiente, la estimación de dicho vector se calcula sustituyendo el vector aleatorio —observable— $y^{(3)}$ por el vector realización.

$$\begin{pmatrix} \hat{\beta}_0^{*(3)} \\ \hat{\beta}_1^{*(3)} \\ \hat{\beta}_3^{*(3)} \\ \hat{\beta}_2^{*(3)} \end{pmatrix} = \begin{pmatrix} \frac{1}{7} & -\frac{3}{28} & -\frac{1}{6} & \frac{5}{54} \\ \frac{1}{7} & -\frac{2}{28} & \frac{1}{6} & 0 \\ \frac{1}{7} & -\frac{1}{28} & \frac{1}{6} & -\frac{3}{84} \\ \frac{1}{7} & 0 & 0 & -\frac{4}{84} \\ \frac{1}{7} & \frac{1}{28} & -\frac{1}{6} & -\frac{3}{84} \\ \frac{1}{7} & \frac{2}{28} & -\frac{1}{6} & 0 \\ \frac{1}{7} & \frac{3}{28} & \frac{1}{6} & \frac{5}{84} \end{pmatrix}^T \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 2 \\ 3 \\ 3 \end{pmatrix}$$

llegando al resultado final,

$$\begin{pmatrix} \hat{\beta}_0^{*(3)} \\ \hat{\beta}_1^{*(3)} \\ \hat{\beta}_3^{*(3)} \\ \hat{\beta}_2^{*(3)} \end{pmatrix} = \begin{pmatrix} 1{,}4286 \\ 0{,}5000 \\ -0{,}5000 \\ 0{,}1190 \end{pmatrix}$$

de lo que se desprende que, la ecuación de regresión estimada —en la tercera etapa— adopta la siguiente forma,

$$y^{*(3)} = 1{,}4286 + 0{,}5000\ x_1 - 0{,}5000\ x_3 + 0{,}1190\ x_2$$

El vector residuo estimado —en la tercera etapa— mediante la ecuación de regresión es:

$$e^{*(3)} = \begin{pmatrix} -0{,}0236 \\ 0{,}0714 \\ -0{,}0716 \\ 0{,}0474 \\ -0{,}0716 \\ 0{,}0714 \\ -0{,}0236 \end{pmatrix}$$

Comentario

La entrada progresiva de las variables en la ecuación de regresión, es función del nivel de significación. De tal manera que, cuando el nivel de significación aumenta, es posible la entrada de más variables. en concreto, cuando:

— el nivel de significación es 0,01 no entra ninguna variable.

— el nivel de significación es 0,05 entra la x_1.

— el nivel de significación es 0,1 entran las tres en el siguiente orden: 1° x_1, 2° x_3 y 3° x_2.

Segunda situación: consideramos que hay individuos atípicos y, por consiguiente, **reiniciaremos el proceso metodológico**.

Proceso preparatorio al proceso metodológico

Omitimos este proceso ya que es el mismo que en la **primera situación**.

Fases de proceso metodológico

Omitimos las cuatro primeras fases ya que son las mismas que en la **primera situación**.

1. La ecuación de regresión estimada —en la primera etapa— adopta la siguiente forma:

$$y^{*(1)} = 1{,}4286 + 0{,}5000\ x_1$$

2. El cálculo de las estimaciones de los cuatro indicadores expresados —en la primera etapa— se refleja en la siguiente tabla.

y_i	Residuos estimados	Residuos normalizados estimados	Residuos estudentizados estimados	Distancia de COOK estimada
1	1,0714	1,4541	1,9887	1,7104*
0	–0,4286	–0,5817	–0,6883	0,0948
0	–0,9286	–1,2603	–1,3905	0,2102
1	–0,4286	–0,5817	–0,6283	0,0329
2	0,0714	0,0969	0,1110	0,0013
3	0,5714	0,7755	0,9176	0,1684
3	0,0714	0,0969	0,1324	0,0076

Teniendo en cuenta el criterio adoptado sobre la distancia de COOK, eliminamos el primer individuo.

Reinicialización del proceso metodológico

3. Construcción de la nueva matriz de datos

y_1	x_{i1}	x_{i2}	x_{i3}
0	–2	0	1
0	–1	–3	1
1	0	–4	0
2	1	–3	–1
3	2	0	–1
3	3	5	1

4. Construcción de la matriz de correlaciones de BRAVAIS-PEARSON entre las variables explicativas.

	x_1	x_2	x_3
x_1	1,0000	0,5649	–0,3806
x_2	—	1,0000	0,3583
x_3	—	—	1,0000

Fase del procedimiento metodológico

1. Elegir como primera variable explicativa, aquella cuyo coeficiente de correlación lineal con la variable explicada sea el **máximo**.

	x_1	x_2	x_3
y	0,9695	0,5477	–0,5165

2. La variable elegida ha sido la x_1

3. Contrastar, si el coeficiente de correlación poblacional entre la variable explicad y la variable explicativa elegida es —significativamente— diferente de cero

El test de hipótesis que hay que realizar es el siguiente:

$$H_0: \rho_{y,x_1} = 0$$
$$H_1: \rho_{y,x_1} \neq 0$$

Como resultado de la aplicación de la regla general de decisión de dicho test concluimos que, como:

$$-2,132 > 7,9113 > 2,132$$

Para un nivel de significación del 0,1 **rechazamos la hipótesis nula** y, por consiguiente, **el proceso de elección de variables continúa**.

4. La ecuación de regresión estimada —en la primera etapa— adopta la siguiente forma:

$$y^{*(1)} = 1,1429 + 0,7143\ x_1$$

5. El cálculo de las estimaciones de los cuatro indicadores expresados —en la primera etapa— se refleja en la siguiente tabla.

y_i	Residuos estimados	Residuos normalizados estimados	Residuos estudentizados estimados	Distancia de COOK estimada
0	0,2857	0,7559	1,0443	0,5998
0	–0,4286	–1,1340	–1,3506	0,3821
1	–0,1429	–0,3781	–0,4177	0,0193
2	0,1428	0,3778	0,4174	0,0192
3	0,4285	1,1337	1,3502	0,3819
3	–0,2858	–0,7562	–1,0956	0,6602

Teniendo en cuenta el criterio aportado por la distancia de COOK, no eliminamos ningún individuo y, por consiguiente seguimos el proceso metodológico.

6. Elegimos como segunda variable explicativa, aquella cuyo coeficiente de correlación lineal de BRAVAIS-PEARSON con la variable residual estimada —en la primera etapa— sea **máximo**.

	x_1	x_2	x_3
$e^{*(1)}$	–0,0001	0,0000	–0,6017

7. La variable elegida ha sido la x_3.

8. Contrastar, si el coeficiente de correlación poblacional entre la variable residual estimada —en la primera etapa— y la variable explicativa elegida es —significativamente— diferente a cero.

El test de hipótesis que hay que realizar es el siguiente,

$$H_0: \rho_{e^{*(1)},x_3} = 0$$
$$H_3: \rho_{e^{*(1)},x_3} \neq 0$$

Como resultado de la aplicación de la **regla general de decisión** de dicho test concluimos que, como:

$$-2,132 < 1,5067 < 2,132$$

Para un nivel de significación del 0,1 **aceptamos la hipótesis nula** y, por consiguiente, **el proceso de elección de variables se para**.

Por lo tanto, la ecuación de regresión estimada es,

$$y^{*(1)} = 1,1429 + 0,7143\ x_1$$

Comentarios

El hecho de que, hayamos **reiniciado** el proceso de elección de variables por la eliminación de individuos atípicos, nos conduce a un resultado de interés, en cuanto a las variables retenidas en la estimación de la ecuación de regresión. En lugar de haber retenido tres variables: x_1, x_2 y x_3, hemos retenido una variable: x_1. Aunque, sin duda alguna, con las tres variables obtenemos una **varianza residual** menor que con una, sim embargo, consideraremos el caso de una ya que, **esta nos aporta resultados satisfactorios, en cuanto a la estimación se refiere.**

«La estimación de la varianza residual con una variable habiendo eliminado individuos atípicos es al menos dos veces menor que, la estimación de la varianza residual con dos variables, sin haber eliminado individuos atípicos».

BIBLIOGRAFÍA

(1) AGERSON T., DENIS J. B., FOUCART T., JOLIVET E., PRUVOT F., ROUX M., TOMASSONE R. (1991). STAT-ITCF versión 4.0 Institut Technique des Céréales et des Fourrages. 8 avenue du Président Wilson 75116 Paris.

(2) AUGRAIN S.- LESQUOY-de-TURCKHEIM, MILLIER C., TOMASSONE R. (1992). La Régression nouveaux regards sur une ancienne méthode statistique. Masson. 2ª édition réviseé.

(3) BAILLARGEON G. (1985). Méthodes Statistiques. Méthodes d'analyse de régression linéaire simple et régression multiple avec applications dans différentes secteurs de l'entreprise. Volume 2. Les éditions SMG.

(4) BERA A. K., JARQUE C. M. (1980). Efficient test for normality, homoskedasticity and serial independance of regression residual. Econometrics Letters, 6, 225-259.

(5) BERA A. K., JARQUE C. M. (1984). Testing the normality assumption in limited dependent variable model. International Economic Review, vol. 25, nº 3.

(6) BOURBONNAIS R. (1998). Econometrie. 2ª édition. Dunod.

(7) BOURBONNAIS R., USUNIER J-CL. (1992). Pratique de la prévision des ventes. Conception de Systémes. Ed. Economica. 49, rue Héricat, 75017 Paris.

(8) BREUSCH T. (1978). Testing for autocorrelation in dynamic linear models. Australian Economic. Papers, vol. 17.

(9) CAMPOS SÁNCHEZ L., DIAZ-LLANOS Y SAINZ-CALLEJA Fco. J. (1997) Procedimientos de gestión informática utilizando el STATlab y sus aplicaciones en la Estadística Exploratoria Multidimensinal. Oficina provincial del registro de la propiedad intelectual. Madrid. Solicitud número 63. 787.

(10) CAZES P. (1975). Protectin de la regression par utilisarion de contraintes linéaires et non linéaires. Rev. Stat. Appl. 23, 37-57.

(11) CAZES P. (1978). Méthodes de régression: la régression sous contraintes. Cah. Anal. Données 3, 147-165.

(12) COOK R. D. (1997). Detection of influential observations in linear regression. Technometrics 19, 15-18.

(13) COOK R. D., WEISBERG S. (1982). Residuals and influence in regression. Chapman et Hall, New York, 230 p.

(14) DAGNELIE P. (1977). Analyse statistique à plusieurs variables. Les Presses Agronomiques de Gembloux.

(15) DRAPER N. R., SMITH H. (1981). Applied regression analysis. Second Edition. John Wiley & Sons.

(16) DROESBEKE J-J. (1988). Elements de Statistique. Editions de l'Université de Bruxelles.

(17) DURBIN J., WATSON G. S. (1950). Testing for serial correlation in least-squares regression. Biometrika, vol 37.

(18) DURBIN J. WATSON G. S. (1951). Testing for serial correlation in least-squares regression. Biometrika, vol 38.

(19) FARRAR D. E., GLAUBER R. R. (1967). Multicolinearity in regression analysis. Review of Economics and Statistics, vol 49.

(20) FURNIVAL G. M. (1971). All possible regressions with les computation. Technometrics, 13, p 403-408.

(21) FURNIVAL G. M., WILSON R. W. (1974). Regression by leaps and bounds. Technometrics, 16, p 403-511.

(22) GODFREY L. G.(1978). Testing for higher order serial correlation in regression equation when the regressors contain lagget dependant variables. Econometrica, vol. 46.

(23) LEBART L., MORINEAU A., PIRON M. (1995). Statistique exploratoire multidimensionnelle. Dunod.

(24) LUND I. A. (1971). An application of stagewise and strepwise regression procedure to a problem of stimating precipitation in California. J. Appl. Meterol., 10, 892-902.

(25) MAHALANOBIS P. C. (1936). On the generalized distance in statistics. Proc. Nat. Inst. Sci. India, 12, p 49-55.

(26) MONTGOMERY D. C., RUNGER G. C. (1996). Probabilidad y Estadística aplicadas a la Ingeniería. McGRAW-HILL Interameticana editores, S.A.

(27) PALM R. (1994) La régression: un probléme complexe à partir d'une idée simple. Biom. Praxim, 34, 109-123.

(28) PALM R., IEMMA A. F. (1995). Quelques alternatives à la regression classique dans le cas de la colinéarité. Revue. Statist. Appl., 43 (2), p 5-23.

(29) SNEE R. D. (1977). Validarion of regression models: methods and examples. Technometrics 19, 415-428.

(30) WEISBERG S. (1985). Applied linear regression. New York, Wiley, 324.