CAPÍTULO 6

Multicolinealidad

1. Introducción

Como es sabido, la regresión lineal es una herramienta estadística ampliamente usada para analizar cómo influyen (si es que lo hacen) un conjunto de variables (independientes o explicativas) en otra (dependiente o explicada), permitiendo la estimación numérica de los signos y magnitudes de los coeficientes en una relación lineal previamente establecida. Es decir, el modelo de regresión múltiple es una de las técnicas más usadas cuando se desea establecer relaciones lineales entre variables. Así, dado el modelo de regresión lineal con n observaciones y k variables independientes de la expresión (7.1) que, recordemos, responde a la siguiente expresión:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times k} \cdot \boldsymbol{\beta}_{k\times 1} + \mathbf{u}_{n\times 1},$$

donde la primera columna de $\mathbf{X}$ está formada por unos representando a la constante (lo cual se notará como $\mathbf{1}_{n\times 1} = (1, \ldots, 1)^t$) y $\mathbf{u}$ representa a la perturbación aleatoria (que se presupone centrada y esférica[1], el objetivo es el de estimar los coeficientes de las variables independientes, $\boldsymbol{\beta}$, para a partir de los valores obtenidos establecer el sentido de las relaciones (con el signo) y cuantificar las mismas (con el número).

Para obtener las estimaciones comentadas el método más usado es el de Mínimos Cuadrados Ordinarios (MCO), el cual proporciona[2] la estimación $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, por lo que es necesario que exista la inversa de la matriz $\mathbf{X}^t\mathbf{X}$, es decir, se asume la independencia lineal entre las variables independientes presentes en $\mathbf{X}$. Cuando esta condición no se verifica se dice que en el modelo hay multicolinealidad. Por tanto, la multicolinealidad describe la situación de ausencia de ortogonalidad entre las variables independientes del modelo de regresión.

Si la multicolinealidad es exacta, lo cual ocurre cuando una de las variables independientes es combinación lineal exacta de algunas o todas las demás, no es posible obtener la inversa de la matriz $\mathbf{X}^t\mathbf{X}$ y, en tal caso, el objetivo marcado sería inalcanzable ya que no existiría una estimación única para $\widehat{\boldsymbol{\beta}}$. Sin embargo, si la multicolinealidad es aproximada, lo cual ocurre

[1] Esto es, $E[\mathbf{u}_{n\times 1}] = \mathbf{0}_{n\times 1}$ y $var(\mathbf{u}_{n\times 1}) = \sigma^2 \cdot \mathbf{I}_{n\times n}$, donde $\mathbf{0}$ es un vector de ceros, σ^2 es la varianza de la perturbación aleatoria e $\mathbf{I}$ es la matriz identidad.

[2] Como es bien sabido esta expresión coincide con la dada por el Método de Máxima Verosimilitud.

Variable	Estimación	Desviación típica
Constante	0.22471	0.0397
Tipos de interés a 3 meses	-0.62891	0.06582
Tipos de interés a 6 meses	1.59334	0.06394
R^2	0.9965	
$\widehat{\sigma}^2$	0.03330625	
$F_{2,121}$	17371.66	

Tabla 6.1: Estimación por MCO de los tipos de interés a 12 meses en función de los tipos de interés a 3 y 6 meses

Variable	Estimación	Desviación típica
Constante	0.4404	0.09556
Tipos de interés a 3 meses	1.00569	0.01343
R^2	0.9787	
$\widehat{\sigma}^2$	0.2025	
$F_{2,122}$	5611	

Tabla 6.2: Estimación por MCO de los tipos de interés a 12 meses en función de los tipos de interés a 3 meses

cuando una de las variables independientes es aproximadamente igual a una combinación lineal de las restantes, sí es posible calcular dicha inversa.

En el primer caso (multicolinealidad perfecta), el modelo no satisface la condición de rango completo y, como se ha comentado, conduce a infinitas estimaciones de los coeficientes del modelo de regresión, mientras que en el segundo caso (multicolinealidad aproximada), aunque dicha condición es satisfecha, la estimación será inestable y se pueden presentar problemas relacionados con la estimación del modelo y el análisis estadístico del mismo.

Luego, el segundo caso es el problemático ya que pone en entredicho las conclusiones obtenidas en el análisis realizado. Además, una de las características de las variables económicas es la posible correlación entre ellas debido a la existencia de determinantes comunes, por lo que la multicolinealidad entre variables explicativas en una regresión lineal múltiple debe considerase una situación habitual.

Ejemplo 6.1 *Consideremos los datos disponibles en Wooldridge [148] sobre datos trimestrales (desde primer trimestre de 1950 al cuarto trimestre de 1980, 124 datos) sobre tipos de interés a 12, 6 y 3 meses. Si se estima por MCO el modelo que analiza los tipos de interés a 12 meses a partir del resto, se obtienen los resultados mostrados en la Tabla 6.1.*

Se observa que todos los coeficientes son significativamente[3] *distintos de cero y que el modelo es válido globalmente, sin embargo, no es esperable el signo negativo obtenido en los tipos de interés a 3 meses ya que, entre otros motivos, el coeficiente de correlación*[4] *entre los rendimientos a 3 y 12 meses es* 0.9893021*. ¿A qué se debe esta relación no esperada entre los rendimientos a 3 y 12 meses?*

Si se estima por MCO el modelo de regresión lineal simple (modelo (7.1) para $k = 2$) que analiza los tipos de interés a 12 meses a partir de los rendimientos a 3 meses se obtienen los resultados mostrados en la Tabla 6.2. En este caso, la relación entre los tipos de interés a 3 meses y 12 meses si tiene el signo esperado. Es decir, la introducción de los tipos de interés

[3] Mientras que no se especifique lo contrario, toda la inferencia se realiza al 5 % de significación.

[4] Adviértase que en este caso es ingenuo pensar que se verifica el conocido *ceteris paribus*, es decir, cuando varían los tipos de interés a 3 meses es claro que también van a cambiar los tipos de interés a 6 meses.

Variable	Estimación	Desviación típica
Constante	-0.1174	6.4764
Consumo	-2.3429	3.33507
Ingresos	2.8562	1.91234
R^2	0.92202	
$\widehat{\sigma}^2$	0.8228	
$F_{2,14}$	82.77	

Tabla 6.3: Estimación por MCO del crédito en Estados Unidos

a 6 meses en el modelo de regresión lineal simple provoca la relación estimada no esperada comentada debido a la relación lineal existente entre ambas variables. □

Ejemplo 6.2 *Para analizar la influencia del consumo e ingresos personales,* $\mathbf{C}$ *e* $\mathbf{I}$*, respectivamente, sobre la deuda pendiente de hipoteca,* $\mathbf{D}$*, en los años 1996 a 2012 (datos disponibles en la Tabla C.1 del Apéndice 1.1) se obtienen resultados mostrados en la Tabla 6.3.*

Se puede observar que ningún coeficiente es significativamente distinto de cero aunque sí lo es el modelo globalmente, lo cual es contradictorio. ¿Qué está ocurriendo en este modelo? □

Como se ilustrará a lo largo del capótulo, en los modelos de los ejemplos anteriores existe un grado de multicolinealidad aproximada preocupante. Por tanto, la existencia de un alto grado de multicolinealidad no es una cuestión baladí cuando se estima el modelo de regresión por el método de MCO, de ahí la importancia de detectarla y tratarla de forma adecuada.

2. Multicolinealidad exacta o perfecta

La multicolinealidad exacta o perfecta hace referencia a la existencia de una relación lineal exacta entre dos o más variables independientes.

Dicho tipo de multicolinealidad se traduce en que la matriz $\mathbf{X}$ no es de rango completo por columnas, esto es, el rango de $\mathbf{X}$ es menor que k. Como se ha comentado, el incumplimiento de esta condición no permite invertir la matriz $\mathbf{X}^t\mathbf{X}$, por lo que el sistema de ecuaciones normales, $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y}$, es compatible indeterminado, es decir, es imposible obtener una solución única para $\boldsymbol{\beta}$ (hay infinitas).

¿Qué hacer ante esta situación? Evidentemente no se podrán estimar los coeficientes de las variables independientes, sin embargo, si se podrá estimar una combinación lineal de los mismos. Aunque en tal caso no se tiene garantizado que se puedan recuperar a partir de éstas las estimaciones de los parámetros originales.

Ejemplo 6.3 *Considerando el modelo (7.1) para* $k = 3$ *donde* $\mathbf{X}_2 - \mathbf{X}_3 = 1$*, sin más que sustituir* $\mathbf{X}_2 = 1 + \mathbf{X}_3$ *en el modelo original* $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3\mathbf{X}_3 + \mathbf{u}$ *se tiene que:*

$$\mathbf{y} = \beta_1 + \beta_2 \cdot (1 + \mathbf{X}_3) + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u} = (\beta_1 + \beta_2) + (\beta_2 + \beta_3) \cdot \mathbf{X}_3 + \mathbf{u}.$$

Entonces, es posible estimar las combinaciones lineales de los parámetros originales $\beta_1 + \beta_2$ *y* $\beta_2 + \beta_3$*.*

Si se sustituye $\mathbf{X}_3 = \mathbf{X}_2 - 1$ *las combinaciones lineales estimables de los parámetros originales son* $\beta_1 - \beta_3$ *y* $\beta_2 + \beta_3$ *ya que:*

$$\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot (\mathbf{X}_2 - 1) + \mathbf{u} = (\beta_1 - \beta_3) + (\beta_2 + \beta_3) \cdot \mathbf{X}_2 + \mathbf{u}.$$

Variable	Estimación	Desviación típica
Constante	-25.88	17.96
$\mathbf{X}_2$	32.77	17.96
$\mathbf{X}_3$	-34.76	17.95
R^2	0.9961	
$\widehat{\sigma}^2$	1.343281	
$F_{2,47}$	6011	

Tabla 6.4: Estimación por MCO de los datos del Ejemplo 12.1

En esta situación no es posible obtener la estimación de los coeficientes originales a no ser que se disponga a priori de algún tipo de información sobre algunos de ellos. □

Ejemplo 6.4 *Con el objetivo de ilustrar la situación del ejemplo anterior, supongamos que se generan*[5] *50 observaciones para una variable* $\mathbf{X}_2$ *a partir de una normal de media 10 y desviación típica 10 y una variable* $\mathbf{p}$ *como una normal de media 1 y desviación típica 0.01, a partir de las cuales se calcula* $\mathbf{X}_3 = \mathbf{X}_2 - \mathbf{p}$. *Puesto que los valores de* $\mathbf{p}$ *son prácticamente iguales a 1, lo que se pretende es ilustrar la situación anterior en la que* $\mathbf{X}_3 = \mathbf{X}_2 - 1$.

Al mismo tiempo, se genera una variable $\mathbf{u}$ *(que va a representar el término perturbación) como una normal de media 0 y varianza 1 y se construye* $\mathbf{y} = 5 + 2 \cdot \mathbf{X}_2 - 4 \cdot \mathbf{X}_3 + \mathbf{u}$. *De esta forma se conocen los verdaderos valores de los coeficientes y se puede saber cuánto se ajusta la estimación por MCO de la misma a la realidad.*

Estimando por MCO la regresión de $\mathbf{y}$ *en función de* $\mathbf{X}_2$ *y* $\mathbf{X}_3$, *se obtienen los resultados mostrados en la Tabla 12.2. Se observa una vez más que los coeficientes de* $\mathbf{X}_2$ *y* $\mathbf{X}_3$ *no son significativamente distintos de cero al mismo tiempo que el modelo es válido en su conjunto. Además, las estimaciones de cada uno de los coeficientes se alejan de los verdaderos valores con los que se ha generado la variable* $\mathbf{y}$. *Sin embargo, las combinaciones lineales de los parámetros* $\beta_1 - \beta_3 = 5 + 4 = 9$ *y* $\beta_2 + \beta_3 = 2 - 4 = -2$ *son muy parecidas a las estimaciones de las mismas ya que* $\widehat{\beta}_1 - \widehat{\beta}_3 = -25.88 + 34.76 = 8.88$ *y* $\widehat{\beta}_2 + \widehat{\beta}_3 = 32.77 - 34.76 = -1.99$.

Si las estimaciones de $\beta_1 - \beta_3$ *y* $\beta_2 + \beta_3$ *son fiables, también lo serán las de* $\beta_1 - \beta_3$ *y* $(\beta_2 + \beta_3) \cdot a$ *siendo* a *cualquier número real, al igual que la estimación de* $(\beta_1 - \beta_3) + (\beta_2 + \beta_3) \cdot a$. *Puesto que esta expresión es equivalente a* $\beta_1 + \beta_2 \cdot a + \beta_3 \cdot (1 - a)$, *es esperable que la estimación del modelo* $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \mathbf{u}$ *sea también fiable. Es decir, aunque las estimaciones individuales de los parámetros no se aproximen a la realidad, sí se obtiene una buena estimación de las combinaciones lineales que provocan la multicolinealidad y, por tanto, una buena estimación de los valores de la variable dependiente.*

Esta situación se traduce en una suma de cuadrados de los residuos pequeña y, por tanto, en un coeficiente de determinación elevado. Adviértase que ante un coeficiente de determinación elevado no puede inferirse automáticamente que exista un grado de multicolinealidad grave, ya que éste puede deberse a otros factores, como una buena especificación del modelo. Es decir, sin que se presenten algunos de los otros síntomas comentados, considerar la posibilidad de que exista un problema de multicolinealidad grave debido a un coeficiente de determinación alto no tiene sentido. □

Observación 6.1 *De los resultados mostrados en el ejemplo anterior puede concluirse que la existencia multicolinealidad aproximada grave en un modelo de regresión como el dado en (7.1) no necesariamente es un problema si el interés radica exclusivamente en la predicción.*

[5] El código usado en el entorno de programación **R** para generar los datos de este ejemplo puede consultarse en el Apéndice 2.2.

Es decir, si la presencia de este problema induce coeficientes de determinación altos, se tiene que el ajuste lineal realizado es bueno y, por tanto, las predicciones también lo serían.

Ahora bien, se ha de tener presente que para que la afirmación anterior se mantenga, la estructura de dependencia lineal presente en el modelo estimado se ha de mantener también en la muestra usada para la predicción. Es decir, si los datos usados para estimar el modelo $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \mathbf{u}$ *del Ejemplo 12.1 verifican la relación lineal* $\mathbf{X}_3 = \mathbf{X}_2 - 1$*, los datos para los que se desea realizar predicción también han de verificar dicha relación.* ♢

3. Multicolinealidad aproximada o no perfecta

La multicolinealidad aproximada hace referencia a la existencia de una relación lineal aproximada entre dos o más variables independientes.

En este caso, no se incumplirá la hipótesis básica de que la matriz $\mathbf{X}$ sea completa por columnas (rango de $\mathbf{X}$ igual a k), por lo que se podrá invertir $\mathbf{X}^t\mathbf{X}$ y obtener los estimadores por MCO. Sin embargo, aunque exista $(\mathbf{X}^t\mathbf{X})^{-1}$, ésta será inestable y, en consecuencia, cuando existe un problema de multicolinealidad no perfecta se presentan los siguientes problemas:

a) Los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos.

b) El signo de los coeficientes puede ser contrario al esperado (consecuencia del punto anterior).

c) Las varianzas de los estimadores son muy grandes y, como consecuencia, al efectuar contrastes de significación individual no se rechazará la hipótesis nula,

d) Tendencia a rechazar la hipótesis nula al realizar contrastes conjuntos.

El problema b) ha sido ilustrado mediante el Ejemplo 6.1, mientras que los problemas c) y d) lo han sido en los Ejemplos 6.2 y 12.1. El problema a) se ilustra con los siguientes ejemplos.

Ejemplo 6.5 *Supongamos que se desea analizar el gasto mensual en transporte público (*$\mathbf{G}$*, medido en euros) a partir de la edad de los individuos encuestados (*$\mathbf{E}$*, medida en años) y su número de familiares con trabajo (*$\mathbf{F}$*). Considerando los siguientes datos observados para dichas variables en los alumnos de Econometría 2:*

$$\mathbf{X}_1 = [\mathbf{1}\ \mathbf{E}\ \mathbf{F}] = \begin{pmatrix} 1 & 21 & 2 \\ 1 & 22 & 2 \\ 1 & 21 & 1 \\ 1 & 22 & 2 \\ 1 & 22 & 2 \\ 1 & 21 & 1 \\ 1 & 21 & 1 \end{pmatrix}, \quad \mathbf{X}_2 = [\mathbf{1}\ \mathbf{E}\ \mathbf{F}_p] = \begin{pmatrix} 1 & 21 & 2 \\ 1 & 22 & 2 \\ 1 & 21 & 1 \\ 1 & 22 & 2 \\ 1 & 22 & 2 \\ 1 & 21 & 1 \\ 1 & 21 & \boldsymbol{2} \end{pmatrix}, \quad \mathbf{y} = \mathbf{G} = \begin{pmatrix} 40 \\ 35 \\ 20 \\ 38 \\ 30 \\ 20 \\ 20 \end{pmatrix},$$

se obtienen las siguientes estimaciones:

$$\widehat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1^t\mathbf{X}_1\right)^{-1}\mathbf{X}_1^t\mathbf{y} = \begin{pmatrix} 119 \\ -5.667 \\ 20 \end{pmatrix}, \quad \widehat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2^t\mathbf{X}_2\right)^{-1}\mathbf{X}_2^t\mathbf{y} = \begin{pmatrix} 81 \\ 4.333 \\ 10 \end{pmatrix},$$

Variable	Estimación	Desviación típica
Constante	-3.5181	5.2296
Consumo	-0.7001	2.7459
Ingresos	1.9456	1.6073
R^2	0.9161	
$\widehat{\sigma}^2$	0.8851	
$F_{2,14}$	76.47	

Tabla 6.5: Estimación por MCO de la deuda pendiente de hipoteca en función de las versiones perturbadas del consumo y de los ingresos

los cuales son las soluciones de los siguientes sistemas de ecuaciones (normales):

$$\begin{pmatrix} 7 & 150 & 11 \\ 150 & 3216 & 237 \\ 11 & 237 & 19 \end{pmatrix} \cdot \begin{pmatrix} \beta_{1,1} \\ \beta_{1,2} \\ \beta_{1,3} \end{pmatrix} = \begin{pmatrix} 203 \\ 4366 \\ 346 \end{pmatrix},$$

$$\begin{pmatrix} 7 & 150 & 12 \\ 150 & 3216 & 258 \\ 12 & 258 & 22 \end{pmatrix} \cdot \begin{pmatrix} \beta_{2,1} \\ \beta_{2,2} \\ \beta_{2,3} \end{pmatrix} = \begin{pmatrix} 203 \\ 4366 \\ 346 \end{pmatrix}.$$

Se puede observar que las matrices de diseño, $\mathbf{X}_1$ y $\mathbf{X}_2$, sólo difieren en el último elemento ya que en la primera matriz se tiene un 1 y en la segunda un 2. Sin embargo, este leve cambio se traduce en estimaciones muy distintas tanto en signo como en magnitud ya que a) en el primer caso se tiene para la edad un signo negativo y positivo en el segundo y b) el efecto del número de familiares que trabajan sobre el gasto en transporte público es la mitad en el segundo modelo.

Finalmente, adviértase que los determinantes de las matrices $\mathbf{X}_1^t\mathbf{X}_1$ y $\mathbf{X}_2^t\mathbf{X}_2$ son, respectivamente, iguales a 9 y 12. Esto es, asociar la inestabilidad de la inversa de $\mathbf{X}^t\mathbf{X}$ a un determinante pequeño no es del todo acertado. □

Ejemplo 6.6 *Considerando los datos sobre el crédito en Estados Unidos del Ejemplo 6.2, se introduce*[6] *una perturbación de un 1 % en las dos variables independientes consideradas dando lugar a los datos mostrados en las dos últimas columnas de la Tabla C.1.*

Estimando por MCO la deuda pendiente de hipoteca en función de las versiones perturbadas del consumo y de los ingresos, se observa de nuevo (ver Tabla 6.5) que ninguno de los coeficientes de las variables independientes es significativamente distinto de cero mientras que el modelo sí es válido en su conjunto.

Además, comparando las estimaciones obtenidas con las dadas en el Ejemplo 6.2 se observa importantes diferencias. Más concretamente, se ha producido una variación del 105.1103 % en las mismas ya que:

$$\begin{aligned} \frac{||\boldsymbol{\beta} - \boldsymbol{\beta}_p||}{||\boldsymbol{\beta}||} &= \frac{\sqrt{(-0.1174 + 3.5181)^2 + (-2.3429 + 0.7001)^2 + (2.8562 - 1.9456)^2}}{\sqrt{0.1174^2 + 2.3429^2 + 2.8562^2}} \\ &= \sqrt{\frac{15.09274}{13.66084}} = \sqrt{1.104818} = 1.051103. \end{aligned}$$

□

[6] El código usado en el entorno de programación **R** para generar esta perturbación puede consultarse en el Apéndice 2.3.

Ejemplo 6.7 *Consideremos los datos disponibles en Ramanathan [112] sobre el salario mensual,* ***S*** *(en dólares), años de educación,* ***Ed****, experiencia laboral,* ***Ex*** *(años de experiencia en el último trabajo), y edad,* ***E*** *(en años) de 49 individuos. A priori, se espera que exista relación lineal entre la edad del individuo y la experiencia laboral del mismo, sin embargo, la forma de medir ésta última variable, años de experiencia en el último trabajo en lugar de considerar toda la vida laboral, podría evitar la previsible relación lineal entre ambas.*

Si se estima por MCO el modelo en el que se analiza el salario en función del resto de variables independientes, se obtiene la estimación $\widehat{\boldsymbol{\beta}} = (632.244,\ 142.51,\ 43.225,\ -1.913)^t$*. Mientras que si se perturban un 1 % las variables independientes, la estimación obtenida es* $\widehat{\boldsymbol{\beta}} = (623.335,\ 143.213,\ 43.299,\ -2.052)^t$*, lo cual supone una variación del 1.375995 %.*

Como se verá en las próximas secciones, en este modelo el grado de multicolinealidad aproximada existente no es grave. □

Por otro lado, se puede observar que los problemas anteriores se pueden agrupar en dos bloques:

- Aquellos que afectan a la estimación numérica de $\boldsymbol{\beta}$ (apartados a) y b)), y
- aquellos que afectan al análisis estadístico del modelo (apartados c) y d)).

Los primeros se deben al mal condicionamiento de $\mathbf{X}^t\mathbf{X}$ que hace que su inversa sea inestable, mientras que para los segundos tengamos en cuenta que la varianza de los coeficientes estimados se puede expresar como:

$$var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{SCR_j} = \frac{\sigma^2}{SCT_j \cdot (1 - R_j^2)}, \tag{6.1}$$

donde σ^2 es la varianza de la perturbación aleatoria, SCT_j es la suma de cuadrados totales de la variable independiente $\mathbf{X}_j$ y R_j^2 es el coeficiente de determinación de la regresión auxiliar en la que se analiza la variable independiente $\mathbf{X}_j$ en función del resto de variables independientes:

$$\mathbf{X}_j = \alpha_1 + \alpha_2 \cdot \mathbf{X}_2 + \cdots + \alpha_{j-1} \cdot \mathbf{X}_{j-1} + \alpha_{j+1} \cdot \mathbf{X}_{j+1} + \cdots + \alpha_k \cdot \mathbf{X}_k + \mathbf{v}, \quad j = 2, \ldots, k.$$

Si la relación lineal entre $\mathbf{X}_j$ y el resto de variables independientes es alta (definición de multicolinealidad), es esperable que R_j^2 también lo sea y, por tanto, se tengan varianzas de los coeficientes estimados grandes. Puesto que estas varianzas se usan para obtener el valor experimental en los contrastes de significación individual dividiendo, habrá tendencia a obtener valores pequeños de la misma y, por tanto, a no rechazar la hipótesis nula en los contrastes de significación individual.

Finalmente, teniendo en cuenta lo expuesto en el Ejemplo 12.1, como el estadístico experimental de la prueba de significación conjunta se puede expresar en función del coeficiente de determinación:

$$F_{exp} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)},$$

es claro que un coeficiente de determinación alto se traduce en un valor también alto del estadístico experimental y, por tanto, tendencia a rechazar la hipótesis nula del contraste de significación conjunta.

3.1. Tipos de multicolinealidad aproximada

Las causas que producen multicolinealidad aproximada en un modelo son diversas, sin embargo, podrían dividirse en los siguientes bloques:

Multicolinealidad sistemática: debida a un problema estructural, es decir, a la alta correlación lineal de las variables independientes consideradas.

Multicolinealidad errática: debido a un problema puramente numérico, es decir, a un mal condicionamiento de los datos considerados por escasa variabilidad de las observaciones y/o reducido tamaño de la muestra.

Multicolinealidad no esencial: debida a la relación de la constante con el resto de variables independientes consideradas.

Multicolinealidad esencial: debida a la relación entre las variables independientes excluida la constante.

Dentro del primer grupo se podría considerar la relación existente entre la edad y experiencia laboral (siempre que ésta se defina como el número de años trabajados), a mayor edad es esperable una mayor experiencia laboral. El segundo caso se podría ilustrar mediante la posible relación existente entre la edad de un grupo de personas y el número de familiares directos con trabajo. En principio no se debe esperar ninguna relación entre dichas variables, sin embargo, si la muestra se reduce, por ejemplo, a los alumnos matriculados en Econometría 2, sí que podrían obtenerse cierta relación lineal entre ambas debido exclusivamente a la escasa variabilidad presente en la muestra (ver Ejemplo 6.5).

El tercer caso se presenta si una variable presenta poca variabilidad (varianza y/o coeficiente de variación bajos), como pueden ser los distintos índices usados en Economía (por ejemplo, el Índice de Precios al Consumo (IPC)), ya que indicaría que dicha variable es prácticamente constante y, por tanto, estaría relacionada con la constante del modelo. Por otro lado, si la variabilidad es alta esta relación con la constante no existiría y la multicolinealidad existente sólo se debería al resto de variables independientes (cuarto caso).

Finalmente, como se ilustrada en la Observación 6.2, es interesante distinguir el tipo de multicolinealidad existente en el modelo ya que al ser las causas que lo provocan distintas, su solución también debería serlo. Es más, como se verá en la sección 5, la solución idónea para una situación no tiene por qué funcionar en otra.

Observación 6.2 *Es evidente que los 4 bloques anteriores pueden combinarse para dar lugar a distintas situaciones de multicolinealidad aproximada:*

sistemática y no esencial	*sistemática y esencial*
errática y no esencial	*errática y esencial*

Así, por ejemplo, la multicolinealidad aproximada presentada en el Ejemplo 6.5 se podría encuadrar como errática y esencial ya que el reducido tamaño de la muestra supone una relación entre la edad de los individuos encuestados y su número de familiares con trabajo. También podría encuadrarse como errática y no esencial ya que el reducido tamaño de la muestra también hace que ambas variables tengan escasa variabilidad, lo que nos hace pensar que pudiera estar relacionada con la constante. En ambos casos, los síntomas de multicolinealidad aproximada esencial y no esencial detectados se podrían solventar aumentando la muestra y procurando obtener información con una mayor variabiliadad que la proporcionada por una muestra muy concreta (los alumnos de cierta asignatura).

Consideremos que se desea analizar el valor del euribor en función del IPC y de la balanza de pagos. En este caso, la poca variabilidad que suele presentar el IPC nos hace pensar que pudiera presentarse un problema de multicolinealidad sistemática no esencial ya que por mucho que se aumente la muestra es esperable que la relación del IPC con la constante no cambie. Luego la solución planteada en el ejemplo anterior no mitigaría el problema de multicolinealidad existente en este otro.

Finalmente, consideremos que deseamos analizar el precio de venta de un vehículo a partir de su longitud, anchura, altura, peso, cilindrada y tipo de transmisisón. Es evidente que existe una relación sistemática esencial entre prácticamente todas las variables independientes consideradas, ya que una mayor longitud, anchura o altura implican un mayor peso y, por tanto, una mayor cilindrada. Es más, si el análisis se realiza para un modelo concreto de vehículos (todoterrenos, gama alta/media/baja, etc) es posible que aparezca también multicolinealidad no esencial ya que vehículos de una misma gama van a tener una longitud, anchura, altura, peso y cilindrada similar. Es decir, es posible que en la muestra (aunque sea grande) exista poca variabilidad, por lo que las variables comentadas se relacionen linealmente con la constante del modelo. ♢

4. Detección de la multicolinealidad

Detectar si existe multicolinealidad perfecta es muy sencillo ya que es suficiente con calcular el determinante de $\mathbf{X}^t\mathbf{X}$ y ver si es o no igual a cero. Por lo que el presente apartado se centra en la detección de la multicolinealidad aproximada o no perfecta.

Por otro lado, de la definición proporcionada para esta última es claro que a no ser que las variables sean linealmente independientes, siempre existirá multicolinealidad aproximada en un modelo de regresión en el que se usan datos reales. Por tanto, el objetivo de este apartado no es determinar si hay o no multicolinealidad aproximada, sino si el grado de multicolinealidad aproximada existente tiene consecuencias negativas sobre el análisis realizado.

Por lo comentado hasta ahora, estar ante un modelo donde los coeficientes de las variables independientes no son significativamente distintos de cero mientras que sí lo es el modelo conjuntamente o haber obtenido etimaciones de los coeficientes de las variables independientes con signos no esperados, son síntomas que hacen pensar en la existencia de multicolinealidad aproximada grave. Sin embargo, basarse únicamente en dichos síntomas no es un método adecuado, por lo que se hace necesario el desarrollo de herramientas que permitan detectar si el grado de multicolinealidad aproximada existente es grave.

Algunas de las medidas más usadas con este objetivo son:

1. La existencia de coeficientes de correlación simple entre las variables explicativas con valores altos pueden denotar la presencia de multicolinealidad elevada. Sin embargo, si dichos valores son bajos esto no tiene por qué indicar la ausencia de multicolinealidad grave, por lo tanto, se puede decir que estos coeficientes son una condición suficiente pero no necesaria para la existencia de multicolinealidad grave.

2. Relacionado con el punto anterior, un valor próximo a cero del determinante de la matriz de correlaciones de las variables independientes es síntoma de que el grado de multicolinealidad aproximada es grave ya que indica que las correlaciones simples son próximas a uno. De igual forma, un determinante próximo a uno es síntoma de que el grado de multicolinealidad aproximada es bajo ya que indica que la matriz de correlaciones simples es próxima a la matriz identidad.

3. El número de condición es una medida ampliamente usada en Álgebra para medir como son de sensibles las soluciones de un sistema de ecuaciones a cambios en los datos iniciales. Puesto que el estimador por MCO se obtiene como solución del sistema de ecuaciones normales, aplicar esta medida al análisis de un modelo de regresión da una medida de los estables que son las estimaciones obtenidas. Atendiendo a los problemas comentados en la sección 3, parece claro que nos ayudaría a detectar aquellos problemas relacionados con el cálculo numérico del modelo de regresión.

4. Un valor elevado del coeficiente de determinación R_j^2 es un buen indicador de la presencia de multicolinealidad grave ya que indica una alta relación lineal entre $\mathbf{X}_j$, con $j = 2, \ldots, k$, y el resto de variables independientes. Hay otros dos instrumentos íntimamente relacionados con éste que son también calculados por algunos programas informáticos estadísticos como son la tolerancia, $TOL(j) = 1 - R_j^2$, y el Factor de Inflación de la Varianza (FIV), $FIV(j) = 1/TOL(j)$. Se observa que el FIV aparece en la expresión (6.1) sobre la varianza de los estimadores de los coeficientes, por tanto, es útil para detectar los problemas mostrados en la sección 3 relacionados con el análisis estadístico del modelo de regresión.

Por desgracia, estas herramientas no son contrastes estadísticos sino reglas prácticas que tratan de determinar si el grado de multicolinealidad aproximada existente en el modelo es o no preocupante.

4.1. Número de Condición (NC)

Decir que el NC mide la sensibilidad de la solución de un sistema de ecuaciones lineales con respecto a perturbaciones en los datos es equivalente a decir que mide la sensibilidad de la inversa de una matriz a perturbaciones en dicha matriz. Por tanto, puede usarse para medir la inestabilidad de $(\mathbf{X}^t\mathbf{X})^{-1}$.

Dado el modelo de regresión lineal múltiple (7.1), el NC se define como:

$$K(\mathbf{X}) = \sqrt{\frac{\xi_{max}}{\xi_{min}}}, \tag{6.2}$$

donde ξ_{max} y ξ_{min} son, respectivamente, los autovalores[7] máximo y mínimo de la matriz $\mathbf{X}^t\mathbf{X}$. Valores de $K(\mathbf{X})$ entre 20 y 30 suponen multicolinealidad aproximada moderada y valores superiores a 30 multicolinealidad aproximada grave.

En la práctica, los datos deben tener longitud unidad, es decir, las observaciones de cada variable han de ser divididas por la raíz cuadrada de la sumatoria de cada uno de sus elementos elevados al cuadrado. Esto supone usar $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ en lugar de $\mathbf{X}^t\mathbf{X}$ para calcular el NC donde $\widetilde{\mathbf{X}} = (\widetilde{x}_{ij})$ representa a la matriz $\mathbf{X} = (x_{ij})$ normalizada (para que tenga longitud unidad) de forma que cada elemento se obtiene como:

$$\widetilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}, \tag{6.3}$$

con $i = 1, \ldots, n$ y $j = 1, \ldots, k$.

La conveniencia de realizar dicha transformación se ilustra con el siguiente ejemplo.

[7] Los autovalores de una matriz se obtienen resolviendo en λ la ecuación de grado k resultante al igualar a cero el determinante de la matriz $\mathbf{X}^t\mathbf{X} - \lambda \cdot \mathbf{I}_{k\times k}$, donde $\mathbf{I}$ es la matriz identidad. Además, por ser $\mathbf{X}^t\mathbf{X}$ una matriz definida positiva, se tiene asegurado que las soluciones son reales positivos.

Ejemplo 6.8 *Supongamos que se considera el siguiente conjunto de datos y su transformación en longitud unidad asociada:*

$$\mathbf{X} = \begin{pmatrix} 1 & 200 \\ 1 & -200 \\ 1 & 20 \\ 1 & -20 \end{pmatrix}, \quad \widetilde{\mathbf{X}} = \begin{pmatrix} \frac{1}{\sqrt{4}} & \frac{200}{\sqrt{80800}} \\ \frac{1}{\sqrt{4}} & \frac{-200}{\sqrt{80800}} \\ \frac{1}{\sqrt{4}} & \frac{20}{\sqrt{80800}} \\ \frac{1}{\sqrt{4}} & \frac{-20}{\sqrt{80800}} \end{pmatrix},$$

de manera que:

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} 4 & 0 \\ 0 & 80800 \end{pmatrix}, \quad \widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Entonces, los Números de Condición a partir de los autovalores de cada una de las matrices anteriores serían:

$$K(\mathbf{X}) = \sqrt{\frac{80800}{4}} = 142.1267, \quad K(\widetilde{\mathbf{X}}) = 1.$$

Es decir, a partir de $\mathbf{X}$ *se diría que existe un problema de multicolinealidad aproximada grave cuando la realidad es que las columnas de la matriz* $\mathbf{X}$ *son linealmente independientes (la traspuesta de la primera por la segunda es igual a cero). Este hecho se recoge fielmente al calcular el NC a partir de* $\widetilde{\mathbf{X}}$*. En este caso, el NC es igual a su mínimo valor.* □

A continuación se calcula el NC en algunos de los ejemplos mostrados hasta el momento.

Ejemplo 6.9 *Recuperando el Ejemplo 6.5 sobre la edad de una serie de individuos y su número de familiares con trabajo, se tendría*[8] *que las matrices asociadas con longitud unidad serían las siguientes*[9]*:*

$$\widetilde{\mathbf{X}}_1 = \begin{pmatrix} 0.378 & 0.3703 & 0.4588 \\ 0.378 & 0.3879 & 0.4588 \\ 0.378 & 0.3703 & 0.2294 \\ 0.378 & 0.3879 & 0.4588 \\ 0.378 & 0.3879 & 0.4588 \\ 0.378 & 0.3703 & 0.2294 \\ 0.378 & 0.3703 & 0.2294 \end{pmatrix}, \quad \widetilde{\mathbf{X}}_2 = \begin{pmatrix} 0.378 & 0.3703 & 0.4264 \\ 0.378 & 0.3879 & 0.4264 \\ 0.378 & 0.3703 & 0.2132 \\ 0.378 & 0.3879 & 0.4264 \\ 0.378 & 0.3879 & 0.4264 \\ 0.378 & 0.3703 & 0.2132 \\ 0.378 & 0.3703 & 0.4264 \end{pmatrix},$$

de forma que:

$$\widetilde{\mathbf{X}}_1^t\widetilde{\mathbf{X}}_1 = \begin{pmatrix} 1 & 0.9997334 & 0.953821 \\ 0.9997334 & 1 & 0.958768 \\ 0.953821 & 0.958768 & 1 \end{pmatrix},$$

$$\widetilde{\mathbf{X}}_2^t\widetilde{\mathbf{X}}_2 = \begin{pmatrix} 1 & 0.9997334 & 0.9669876 \\ 0.9997334 & 1 & 0.9699522 \\ 0.9669876 & 0.9699522 & 1 \end{pmatrix}.$$

Calculando los autovalores de las matrices $\widetilde{\mathbf{X}}_1^t\widetilde{\mathbf{X}}_1$ *y* $\widetilde{\mathbf{X}}_2^t\widetilde{\mathbf{X}}_2$ *(0.0001229386, 0.0581820186 y 2.9416950428 para la primera y 0.0001952811, 0.0419475912 y 2.9578571276 para la*

[8] El código usado en **R** para abordar este ejemplo se muestra en el Apéndice 2.4.

[9] Por ejemplo, la primera columna se obtiene dividiendo 1 entre la raíz cuadrada de 7. La segunda columna se obtendrá dividiendo cada elemento entre 56.7098, que es la raíz cuadrada de la suma de cada uno de los elementos de la segunda columna elevados al cuadrado.

segunda), se tiene que $K(\widetilde{\mathbf{X}}_1) = 154.6873$ *y* $K(\widetilde{\mathbf{X}}_2) = 123.0718$. *Puesto que son mayores que 30 la multicolinealidad presente en el modelo analizado es preocupante.*

Si se ignora la constante en el cálculo del NC se tienen valores iguales a 6.89246 y 8.096955. Es decir, la relación lineal entre las variables $\mathbf{E}$ *y* $\mathbf{F}$ *(multicolinealidad aproximada esencial) es muy baja, por tanto, la multicolinealidad aproximada presente en este modelo es del tipo no esencial.* □

Ejemplo 6.10 *Para el Ejemplo de 6.1 se tiene que el NC no teniendo en cuenta la constante es igual a 56.56906 mientras que si sí se tiene en cuenta es igual a 69.00941. En este caso el incremento producido es de un 18.02704 %, por lo que no parece relevante el papel de la constante en el grado de multicolinealidad preocupante (ya que se supera el umbral) detectado.*

Por otro lado, en el Ejemplo 6.2 se tiene el NC es igual a 207.6262, luego estaríamos ante un alto grado de multicolinealidad aproximada en el modelo. Además, si se excluye la constante del cálculo del NC se obtendría que es igual a 26.94713, es decir, de no incluir la constante a incluirla en el cálculo del NC se produce un incremento del 87.02133 % en éste. Si bien en ambos casos se supera el umbral establecido como preocupante, parece que el papel de la constante en el problema detectado es relevante. Por tanto, en primer lugar se debería mitigar la multicolinealidad no esencial existente y, posteriormente, la esencial.

En el Ejemplo 6.7 se tiene que el NC sin tener en cuenta la constante y teniéndola en cuenta son iguales, respectivamente, a 6.333185 y 12.83614. Puesto que están por debajo de los umbrales establecidos como preocupantes, la multicolinealidad aproximada existente en el modelo no es preocupante.

Finalmente, aunque se ha usado el incremento sufrido en el NC al pasar de no incluir a la constante en el cálculo del NC a sí hacerlo, por desgracia no disponemos de un umbral que nos indique que dicho incremento supone un problema de multicolinealidad no esencial. □

4.2. Factor de Inflación de la Varianza (FIV)

El FIV es una de las medidas más usadas para detectar si el grado de multicolinealidad presente en el modelo es preocupante. Para cada una de las variables independientes del modelo (7.1) se obtiene a partir de la expresión:

$$FIV(j) = \frac{1}{1 - R_j^2}, \quad j = 2, \ldots, k. \tag{6.4}$$

Como, siempre que haya término independiente en el modelo de regresión lineal múltiple, se verifica que $0 \leq R_j^2 \leq 1$ se tiene que $FIV(j) \geq 1$, para todo j.

Puesto que el FIV se obtiene a partir del coeficiente de determinación, R_j^2, de la regresión auxiliar que ajusta la variable independiente j-ésima, $\mathbf{X}_j$, en función del resto de variables independientes entonces se tiene que conforme mayor sea la relación de $\mathbf{X}_j$ con el resto de variables independientes, es decir, cuanto mayor sea la relación lineal entre las variables independientes (o, lo que es lo mismo, mayor grado de multicolinealidad aproximada) mayor será el valor del FIV asociado.

Finalmente, es comúnmente aceptado que valores del FIV superiores a 10 (lo cual es equivalente a un R_j^2 superior a 0.9 o a una $TOL(j)$ menor que 0.1) indicarían que el grado de multicolinealidad presente en el modelo es preocupante. Es decir, una vez calculados los FIVs asociados a cada una de las variables independientes, se diría que la multicolinealidad no es grave si todos son inferiores a 10.

Variable	Estimación	Desviación típica
Constante	-32.3903	27.3395
$\mathbf{X}_2$	-1.6727	0.2758
$\mathbf{X}_3$	2.0286	0.0259
R^2	0.9845	
$\widehat{\sigma}^2$	6.996025	
$F_{2,97}$	3080	

Tabla 6.6: Estimación por MCO del modelo simulado en el Ejemplo 7.3

Ejemplo 6.11 *Para un modelo de consumo,* $\mathbf{C}$*, en función del ingreso,* $\mathbf{I}$*, y la riqueza,* $\mathbf{R}$*, de 50 familias, se ha obtenido la siguiente información para la única regresión auxiliar posible*[10]*:*

$$\widehat{\mathbf{I}}_i = 1.65 + 0.96 \cdot R_i, \quad R^2 = 0.986.$$

En este caso, como $FIV(2) = FIV(3) = \frac{1}{1-0.986} = 71.4286$*, es mayor que 10, se podría decir que el grado de multicolinealidad existente en el modelo de consumo inicial es preocupante.* □

Finalmente, destacar que el FIV ignora el papel de la constante[11], es decir, es incapaz de detectar la multicolinealidad no esencial, por lo que sólo es útil para detectar la multicolinealdiad esencial.

Los siguientes ejemplos ilustran este hecho.

Ejemplo 6.12 *Se generan*[12] *una variable* $\mathbf{X}_2$ *como una normal de media 100 y varianza 1, una variable* $\mathbf{X}_3$ *a partir de una normal de media 100 y varianza 100 y otra variable* $\mathbf{u}$ *a partir de una normal de media 0 y varianza 3. Se observa que* $\mathbf{X}_2$ *se genera con poca variabilidad (su varianza es igual a 0.9467001 y su coeficiente de variación a 0.009731967) por lo que se espera que esté relacionada con la constante del modelo. Finalmente se construye* $\mathbf{y} = 3 - 2 \cdot \mathbf{X}_2 + 2 \cdot \mathbf{X}_3 + \mathbf{u}$ *y se estima el modelo por MCO obteniéndose los resultados mostrados en la Tabla 6.6.*

En este caso no se observan anomalías en los contrastes de significación individual y conjunta. Por otro lado, comparando las estimaciones de los coeficientes con los verdaderos valores, se observa que el coeficiente de $\mathbf{X}_3$ *se acerca enormemente a la realidad, siendo las estimaciones de los otros coeficientes los que sufren los problemas de multicolinelidad grave. Lo cual era esperable ya que son las dos variables independientes relacionadas linealmente.*

Si se calculan los FIVs y el NC se obtienen los siguientes resultados:

$$FIV(2) = FIV(3) = 1.019434, \quad K(\mathbf{X}) = 253.7548.$$

Se observa que el FIV es próximo a su mínimo valor, mientras que el número de condición supera ampliamente el umbral establecido como preocupante. Como se ha comentado, esta

[10] Puesto que el modelo inicial tiene dos variables independientes además de la constante, las regresiones auxiliares posibles son aquellas en las que el ingreso está en función de la riqueza y en la que la riqueza está en función del ingreso. Sin embargo, al tratarse de dos regresiones simples, tienen el mismo coeficiente de determinación, el cual a su vez coincide con el cuadrado del coeficiente de correlación entre ambas variables. Por lo expuesto, el FIV asociado a cada una de las variables será el mismo.

[11] Observando la expresión (6.4) se tiene que el FIV no se define para $j = 1$. En este caso la variable dependiente de la regresión auxiliar es la constante y, por tanto, se estaría ante una suma de cuadrados totales igual a cero. Por otro lado, se tiene que si se calcula el FIV en el modelo de regresión lineal simple siempre se obtiene un valor igual a 1 independientemente de quiénes sean los datos considerados.

[12] El código usado en el entorno de programación **R** para generar los datos de este ejemplo puede consultarse en el Apéndice 2.5.

aparente contradicción se sustenta en que el FIV no detecta la relación de la constante con el resto de variables independiente mientras que el número de condición sí lo hace.

Finalmente, si se calcula el NC eliminado la constante de la matriz $\mathbf{X}$*, se obtiene un valor igual a 19.46892 (por debajo de los umbrales establecidos como preocupantes). Es decir, considerar la constante en el cálculo del NC supone un incremento del 92.32767% en el mismo. Se confirma pues que en este modelo la multicolinealidad aproximada preocupante es del tipo no esencial.* □

Ejemplo 6.13 *El cálculo del FIV en los Ejemplos 6.1 y 6.2 proporciona, respectivamente, unos valores iguales a 146.1685 y 262.858. Luego parece confirmarse la existencia de multicolinealidad esencial en ambos casos.*

Por otro lado, el FIV para el Ejemplo 6.9 es igual a 2.285714 para $\mathbf{X}_1$ *y a 1.428571 para* $\mathbf{X}_2$*. Puesto que son muy próximos a su mínimo valor, se tiene que la multicolinealidad aproximada esencial existente en el modelo no es preocupante. Confirmando la sospecha de que la multicolinealidad aproximada presente en este modelo es del tipo no esencial.*

Finalmente, los FIVs del Ejemplo 6.7 son 1.084993, 1.265056 y 1.192164. Es claro que indican que la multicolinealidad aproximada esencial existente en el modelo no es preocupante. □

Por tanto, unos valores del FIV por debajo de los umbrales establecidos como preocupantes y un NC por encima de los mismos o un gran incremento en el NC al pasar de no considerar la constante en su cálculo a si considerarla, es indicativo de un problema de multicolinealidad no esencial en el modelo. Es conocido que para resolver este problema se han de centrar las variables que se considera que provocan el problema.

Observación 6.3 *Puesto que el FIV se basa en el cálculo de un coeficiente de determinación, no es adecuado usarlo cuando la variable independiente a la que va asociado es binaria ya que es sabido que en este caso se tiene un modelo no lineal y su estimación por MCO plantea diversos problemas, entre ellos la falta de representatividad del coeficiente de determinación.*

Otra propiedad del FIV a destacar, es que es invariante a los cambios de origen y escala en los datos. ♢

4.3. Matriz de correlaciones de las variables independientes y su determinante

El uso de la matriz de correlaciones de las variables independientes y su determinante como medida de detección de la multicolinealidad aproximada es menos habitual que la anterior, seguramente debido a la ausencia de umbrales (tanto para las correlaciones simples como para el determinante) que indiquen que el grado de multicolinealidad aproximada existente es preocupante.

Sin embargo, teniendo en cuenta que la matriz de correlaciones del modelo (7.1) para $k = 3$ es:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

donde ρ es la correlación simple entre $\mathbf{X}_2$ y $\mathbf{X}_3$ y su FIV es:

$$FIV = \frac{1}{1-\rho^2},$$

es claro que un valor del FIV superior a 10 implica un valor de ρ superior a $\sqrt{0.9} = 0.948683$ y del $det(\mathbf{R}) = 1 - \rho^2$ inferior a 0.1 (donde $det(\mathbf{R})$ denota al determinante de la matriz $\mathbf{R}$).

Es decir, un coeficiente de correlación simple superior a 0.95 y un determinante de la matriz de correlaciones inferior a 0.1 serían indicativos de que el grado de multicolinealidad aproximada existente en el modelo es preocupante.

Aunque la intuición indica que estos umbrales se deben mantener para cualquier valor de k, es necesario estudiar más profundamente dicha afirmación tal y como se pone de manifiesto en el siguiente ejemplo.

Ejemplo 6.14 *Supongamos que se dispone de la siguiente matriz de correlaciones:*

$$\mathbf{R} = \begin{pmatrix} 1 & 0.85 & 0.85 \\ 0.85 & 1 & 0.85 \\ 0.85 & 0.85 & 1 \end{pmatrix}.$$

El determinante de esta matriz es igual a 0.06074, mientras que el mayor FIV es 4.56. Es decir, puesto que el determinante es menor que 0.1, usando el umbral anterior se determinaría que la multicolinealidad aproximada existente es preocupante, lo cual se contradice con la información proporcionada por el mayor FIV. □

Por otro lado, destacar que puesto que la matriz de correlaciones de las variables independientes ignora a la constante, la multicolinealidad aproximada que se puede detectar como preocupante a partir de estas medidas es sólo de tipo esencial. Es más, sólo podría detectar la multicolinealidad esencial dos a dos, ya que las correlaciones simples no captan las relaciones lineales en las que hay involucradas más de dos variables. Al mismo tiempo, no queda claro qué tipo de multicolinealidad aproximada se detecta a partir de su determinante.

Ejemplo 6.15 *Para los datos mostrados en las Tablas C.2 y C.3, se tienen las siguientes matrices de correlaciones simples:*

$$\begin{pmatrix} 1 & 0.9431118 & 0.8106989 \\ 0.9431118 & 1 & 0.7371272 \\ 0.8106989 & 0.7371272 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0.8746162 & 0.5827324 & 0.6772081 \\ 0.8746162 & 1 & 0.6949208 & 0.7704504 \\ 0.5827324 & 0.6949208 & 1 & 0.1004228 \\ 0.6772081 & 0.7704504 & 0.1004228 & 1 \end{pmatrix}.$$

En ninguno de los casos se tiene una correlación simple superior a 0.948683, especialmente en el segundo caso, poniendo de manifiesto que en el caso de existir multicolinealidad aproximada preocupante del tipo esencial (dos a dos), ésta involucra a más de dos variables independientes.

Si se calculan los FIVs se obtiene para el primer caso 12.296544, 9.230073 y 2.976638, mientras que para el segundo 4.423212, 57.242325, 20.175817 y 25.540824. Puesto que el máximo FIV es superior a 10, se concluiría que la multicolinealidad aproximada existente es preocupante. □

Ejemplo 6.16 *En los Ejemplos 6.1 y 6.2 se obtienen, respectivamente, valores de ρ iguales a 0.9965734 y 0.998096, y del determinante iguales a 0.00681421 y 0.003804335. Al estar por encima y debajo de los umbrales comentados, supondrían que el grado de multicolinealidad aproximada esencial es preocupante y se estaría en consonancia con las conclusiones comentadas hasta el momento.*

Mientras para el Ejemplo 6.9 se obtendrían valores de ρ y del determinante iguales a 0.75 y 0.4375 para $\mathbf{X}_1$ y a 0.5477226 y 0.7 para $\mathbf{X}_2$. En este caso, teniendo los umbrales anteriores, se concluiría que no existe un grado de multicolinealidad aproximada esencial preocupante. Sin embargo, nada puede afirmarse sobre la posible existencia de multicolinealidad aproximada preocupante del tipo no esencial.

Finalmente, para el Ejemplo 6.7 se tiene la siguiente mariz de correlaciones:

$$R = \begin{pmatrix} 1 & -0.2738121 & -0.1357008 \\ -0.2738121 & 1 & 0.4005713 \\ -0.1357008 & 0.4005713 & 1 \end{pmatrix}.$$

Puesto que las correllaciones simples (en valor absoluto) son inferiores a 0.948683, no parece existir un grado de multicolinealidad aproximada (dos a dos) preocupante. Similar conclusión se obtiene a partir de su determinante, que es igual a 0.7759225. □

Finalmente, adviértase que en la diagonal principal de $\mathbf{R}^{-1}$ se tienen a los FIVs. Este hecho es claro en el caso en el que $k = 3$ ya que:

$$R^{-1} = \begin{pmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}.$$

También es importante destacar que puesto que no es adecuado calcular el coeficiente de correlación simple para la variables no cuantitativas, para utilizar la matriz de correlaciones de las variables independientes o su determinante para detectar si el grado de multicolinealidad aproximada es preocupante habría que dejar fuera las variables binarias. Por tanto, en el caso de que existan variables independientes de esta naturaleza se ha de usar el NC.

5. Soluciones al problema de la multicolinealidad

Si la multicolinealidad es exacta es claro que a no ser que se disponga de algún tipo de información que permita recuperar las estimaciones de los parámetros originales a partir de las combinaciones lineales estimadas de los parámetros (ver sección 2) la única solución posible es la de eliminar alguna de las variables responsables del problema. De hecho, todos los paquetes informáticos aplican automáticamente esta solución ante una situación de multicolinealidad perfecta.

Si la multicolinealidad es aproximada, entre las diversas soluciones que se suelen barajar se tienen aquellas relacionadas con los datos que se disponen (como, por ejemplo, mejora del diseño muestral, incorporar más observaciones a la muestra o el uso de información a priori), el centrado de variables, la aplicación de técnicas de estimación alternativas a MCO (como, por ejemplo, la regresión cresta, regresión alzada, regresión de componentes principales, regresión con variables ortogonales, regresión LASSO o máxima entropía) o incluso la eliminación del modelo de las variables que se consideran provocan el problema.

Sin embargo, dependiendo de las causas del problema se deberían usar unas u otras soluciones, por lo que es importante tener claro la información proporcionada por las medidas de detección presentadas en la sección 4. Así:

- En el caso de la multicolinealidad errática una posible solución pudiera ser el aumento de la muestra en el caso de que ésta sea pequeña. Por ejemplo, en el ejemplo sobre la edad de un conjunto de individuos y el número de familiares con trabajo, seguramente el grado de multicolinealidad aproximada existente disminuya si se ampliase la muestra

incorporando información de individuos con distintos rangos de edad. Sin embargo, por desgracia, mejorar la calidad de los datos disponibles no siempre es posible. Además, si la nueva información recopilada es "más de lo mismo", aunque se haya comentado en la Observación 6.1 que este hecho es bueno para la predicción, en este caso no se tiene garantizado la mitigación del problema.

- En el caso de multicolinealidad no esencial es conocido que la solución al problema es centrar (restarle la media) la o las variables con poca variabilidad. Sin embargo, esta solución es inoperativa en el caso de multicolinealidad aproximada esencial, donde quizás sea más adecuado utilizar técnicas alternativas de estimación como las comentadas anteriormente. Estas técnicas pueden resultar también útiles ante la multicolinealidad sistemática.

- Si en el modelo analizado no se rechaza la hipótesis nula en los contrastes de significación individual, una manera de mitigar estos problemas es el aumento del número de observaciones o añadir variables independientes que se consideren relevantes ya que observando la estimación de la expresión (6.1) reescrita como:

$$\widehat{var}\left(\widehat{\beta}_j\right) = \frac{\widehat{\sigma}^2}{n \cdot var\left(\mathbf{X}_j\right) \cdot \left(1 - R_j^2\right)}, \quad j = 2, \ldots, k, \tag{6.5}$$

se tiene (claramente) que el aumento de la muestra (aunque exista multicolinealidad sistemática y la nueva información siga presentando relación lineal) hace disminuir a $\widehat{var}\left(\widehat{\beta}_j\right)$. Al mismo tiempo, la inclusión de nuevas variables en el modelo disminuirá la suma de cuadrados de los residuos (en mayor medida cuanto más relevantes sean estas nuevas variables independientes), por lo que la estimación de la varianza de la perturbación aleatoria será menor y, por tanto, también disminuirá $\widehat{var}\left(\widehat{\beta}_j\right)$.

A continuación se profundiza en algunas de las soluciones propuestas.

5.1. Aumento de la muestra

Puesto que la multicolinealidad aproximada es un problema inherente a la muestra disponible para realizar el análisis del modelo de regresión lineal múltiple, la primera solución que se propone es el de mejorar la calidad de la muestra y/o ampliar la misma. Ahora bien, ¿por qué no se han tomado estas medidas desde un inicio? Seguramente por que no es posible o tiene un elevado coste (material, económico, etc) que lo imposibilita.

En cualquier caso, supongamos que es posible ampliar el tamaño muestral. Si inicialmente se disponía de una muestra excesivamente sesgada (como es el caso de la multicolinealidad errática detectada en el Ejemplo 6.5) este aumento podría suponer una mitigación del problema de multicolinealidad aproximada grave. Ahora bien, si no es así, si la nueva información es "más de lo mismo", se tiene que si bien esta situación es favorable a la predicción (ver Observación 6.1) no tiene por qué mitigar el problema de multicolinealidad aproximada grave.

Para ilustrar esta afirmación se considera el siguiente ejemplo.

Ejemplo 6.17 *Un caso extremo de ampliar la muestra con información similar a la disponible es que coincidiera plenamente, es decir, que se repitieran exactamente los mismos datos. Considerando esta situación para el Ejemplo 6.2 sobre el crédito en Estados Unidos,*

Variable	Estimación	Desviación típica
Constante	-0.1174	2.1012
Consumo	-2.3429	1.0820
Ingresos	2.8562	0.6204
R^2	0.92202	
$\widehat{\sigma}^2$	0.6928898	
$F_{2,133}$	786.3	

Tabla 6.7: Estimación por MCO del crédito en Estados Unidos en el modelo aumentado

Variable	Estimación	Desviación típica
Constante	-4.74586	1.62326
Consumo	0.02951	0.81048
Ingresos	1.49682	0.4625
R^2	0.9158	
$\widehat{\sigma}^2$	0.748571	
$F_{2,133}$	723	

Tabla 6.8: Estimación por MCO del crédito en Estados Unidos en el modelo aumentado perturbado

si se aumenta la muestra repitiéndola 7 veces más (136 observaciones en total), se tendrían los resultados mostrados en la Tabla 6.7.

Se puede observar que las estimaciones de los coeficientes de las variables independientes no han cambiado al igual que el coeficiente de determinación mientras que la desviación típica estimada de los coeficientes estimados ha disminuido y el valor experimental del contraste de significación conjunta ha aumentado. En el Apéndice 3 se muestra el desarrollo teórico que permite afirmar que esta situación se va a repetir siempre que se aumente la muestra de esta forma.

¿Qué ha ocurrido? Tal y como se razonaba anteriormente al observar la expresión (6.5), un aumento de la muestra implica una disminución de la desviación típica estimada, lo cual supone un aumento en el estadístico experimental de los contrastes de significación conjunta. Es decir, se revierte la tendencia a no rechazar la hipótesis nula de nulidad a sí rechazarla. Por tanto, se habría solventado la contradicción que encontrabámos en el Ejemplo 6.2.

Ahora bien, si se calcula el FIV y el NC en este modelo aumentado se obtiene que son iguales, respectivamente, a 262.858 y 207.6262. Es decir, coinciden con los del modelo original y siguen estando por encima de los umbrales establecidos marcando la existencia de multicolinealidad aproximada preocupante. ¿Es contradictorio con la mejora experimentada en el modelo en referencia a la significación individual?

En la sección 3 se agrupaban los problemas relacionados con la multicolinealidad aproximada en dos bloques, uno referente al cálculo numérico de las estimaciones de los coeficientes del modelo y otro al análisis estadístico del mismo. Hasta ahora se tiene que se habría mitigado el segundo, pero nada se puede afirmar del primero.

Para estudiar la primera cuestión, se perturban las variables independientes referentes al consumo e ingresos un 1 % repitiendo el proceso planteado en el Ejemplo 6.6. La estimación por MCO del modelo aumentado perturbado ofrece los resultados mostrados en la Tabla 6.8.

Se tiene que una perturbación en las variables independientes iguales a un 1 % supone un cambio en las estimaciones de sus coeficientes de un 145.4431 %. Además, este cambio numérico viene acompañado de cambios "estadísticos", ya que el coeficiente de la constante

Variable	Estimación	Desviación típica
Constante	5.469264	13.016791
Consumo	-4.252429	5.135058
Ingresos	3.120395	2.035671
Crédito Pendiente	0.002879	0.005764
R^2	0.8695563	
$\widehat{\sigma}^2$	0.8228	
$F_{2,13}$	52.3	

Tabla 6.9: Estimación por MCO sobre el crédito en Estados Unidos incorporando el crédito pendiente

pasa de no ser significativamente distinto de cero a sí serlo y el del consumo de serlo a no serlo. □

Por tanto, tal y como se ha ilustrado en el anterior ejemplo, la solución ampliamente admitida de aumentar el tamaño de la muestra es idónea siempre y cuando este aumento suponga mejorar la ya existente en términos de variabilidad o diversidad de datos, ya que en caso contrario (la nueva información es similar a la existente) sólo se mitigaría la parte relacionada con la inferencia del modelo y no con el cálculo numérico del mismo. En tal caso, si las estimaciones son inestables, sería desaconsejable sacar conclusiones a partir de los contrastes de significación conjunta aunque se haya conseguido disminuir el valor de la desviación típica estimada de los coeficientes estimados.

5.2. Uso de información a priori

En el presente apartado se propone mitigar el problema de multicolinealidad aproximada preocupante usando información a priori incorporada a la estimación del modelo mediante la estimación del mismo por el método de Mínimos Cuadrados Restringidos (MCR). Puesto que este método proporciona estimaciones de las varianzas de los coeficientes estimados menores que el método de MCO, se estaría mitigando uno de los principales problemas producidos por una multicolinealidad aproximada grave. Para poder aplicar MCR, es necesario plantear una restricción a considerar como hipótesis nula dado un modelo concreto de manera que no se rechace.

Por otro lado, como se ha puesto de manifiesto en el Ejempo 12.1, bajo un grado dde multicolinealidad aproximada alto se pueden obtener una estimación bastante fiable de combinaciones lineales de los coeficientes del modelo de regresión lineal. Parece lícito entonces usar dichas relaciones lineales, estimadas en las regresiones auxiliares para calcular el FIV, como las restricciones a usar en los MCR.

El siguiente ejemplo ilustra la metodología anterior.

Ejemplo 6.18 *En el modelo planteado en el Ejemplo 6.2 se incorpora como variable independiente el crédito pendiente (***CP***, medido en billones de dólares), obteniéndose los resultados mostrados en la Tabla 6.9 al estimarlo por MCO.*

Se observa que sigue verificándose que ninguno de los coeficientes estimados es significativamente distinto de cero cuando el modelo sí que es globalmente significativo. Los FIVs son iguales a 589.754, 281.8862 y 189.4874, mientras que el NC es igual a 332.3. En ambos casos se superan de manera clara los umbrales establecidos como preocupantes, por lo que se considera que la multicolinealidad aproximada existente es preocupante. Además, si se perturban un 1 % las variables independientes, se obtienen las estimaciones $\widehat{\boldsymbol{\beta}} =$

Variable	Estimación	Desviación típica
Constante	1.38298	6.58191
Consumo	-2.60447	2.65614
Ingresos	2.57108	0.879181
Crédito Pendiente	0.0048358	0.00212491
$\widehat{\sigma}^2$	0.7147739	

Tabla 6.10: Estimación por MCR sobre el crédito en Estados Unidos incorporando el crédito pendiente

$(-10.525897,\ 1.846969,\ 1.364484,\ -0.003063)^t$, *lo cual supone una variación del 226.4807 % con respecto a las estimaciones iniciales.*

En el Ejemplo 12.1 se mostró que en situaciones de multicolinealidad aproximada grave se pueden estimar con cierta fiabilidad combinaciones lineales de los coeficientes del modelo, las cuales vienen determinadas por las relaciones lineales existentes entre las variables independientes del mismo, que a su vez son estimadas en las regresiones auxiliares usadas para calcular los FIVs.

La regresión auxiliar usada para calcular el primer FIV[13]*, porporciona la siguiente estimación:*

$$\widehat{\mathbf{C}}_t = 2.478 + 0.331 \cdot \mathbf{I}_t + 0.0008 \cdot \mathbf{CP}_t.$$

Si se sustituye la relación anterior en el modelo inicial se tiene:

$$\begin{aligned} \mathbf{D}_t &= \beta_1 + \beta_2 \cdot (2.478 + 0.331 \cdot \mathbf{I}_t + 0.0008 \cdot \mathbf{CP}_t) + \beta_3 \cdot \mathbf{CP}_t + \mathbf{u}_t \\ &= (\beta_1 + 2.478 \cdot \beta_2) + (0.331 \cdot \beta_2 + \beta_3) \cdot \mathbf{I}_t + (0.0008 \cdot \beta_2 + \beta_4) \cdot \mathbf{CP}_t + \mathbf{u}_t. \end{aligned}$$

Por otro lado, la estimación por MCO del modelo $\mathbf{D}_t = \alpha_1 + \alpha_2 \cdot \mathbf{I}_t + \alpha_3 \cdot \mathbf{CP}_t + \mathbf{v}_t$ *porporciona*[14] *las estimaciones* $\widehat{\alpha}_1 = -5.0709$, $\widehat{\alpha}_2 = 1.709$ *y* $\widehat{\alpha}_3 = -0.0006$. *Es decir, se verifican las relaciones siguientes:*

$$\beta_1 + 2.478 \cdot \beta_2 = -5.0709, \quad 0.331\dot{\beta}_2 + \beta_3 = 1.709, \quad 0.0008 \cdot \beta_2 + \beta_4 = -0.0006. \qquad (6.6)$$

Considerando las relaciones lineales anteriores como restricciones del modelo, se tiene que no se rechazan al obtenerse un p-valor igual a 0.984135 en el correspondiente contraste. Estimando[15] *entonces el modelo por MCR se obtienen los resultados mostrados en la Tabla 6.10.*

Se observa que, como era esperable, la desviación típica estimada ha disminuido en todos los casos, llegándose a verificar que el coeficiente de los ingresos es significativamente distinto de cero. Es decir, se ha mitigado en cierta medida el problema planteado en el Ejemplo 6.2.

En cualquier caso, no se tiene asegurado la mitigación completa del problema. Por ejemplo, considerando la misma perturbación del 1 % anterior en las variables independientes y estimando por MCR se obtienen las estimaciones $\widehat{\boldsymbol{\beta}} = (-3.47212,\ -0.64592,\ 1.92256,\ -0.000083)^t$, *lo cual supone una variación del 134.845 % con respecto a las estimaciones iniciales. Si bien ha disminuido con respecto a la obtenida al estimar por MCO, sigue siendo grande.* □

[13] Se usa esta relación lineal ya que es la que conduce al mayor FIV, es decir, es la relación lineal existente más fuerte entre las variables independientes, por lo que es esperable que proporciones mejores resultados. En cualquier caso es interesante repetir el procedimiento de este ejemplo para las otras regresiones auxiliares.

[14] Adviértase que en este modelo también existe un grado de multicolinealidad aproximada preocupante ya que los valores del FIV y NC son, respectivamente, iguales a 84.45602 y 82.49228.

[15] En este caso se ha usado el software econométrico **GRETL** para estimar por MCR ya que el comando existente en **R**, *rsl()* de la librería *lrmest*, presenta problemas con la inferencia individual.

Variable	Estimación	Desv. Típica
Constante	-199.622	2.5835
$\widetilde{\mathbf{X}}_2$	-1.6727	0.2758
$\mathbf{X}_3$	2.0286	0.0259
R^2	0.9845	
$\widehat{\sigma}^2$	6.996025	
$F_{2,97}$	3080	
FIV	1.019434	
$K(\mathbf{X}_{-1})$	1.014372	
$K(\mathbf{X})$	19.48899	

Variable	Estimación	Desv. Típica
Constante	168.9159	27.5726
$\mathbf{X}_2$	-1.6727	0.2758
$\widetilde{\mathbf{X}}_3$	2.0286	0.0259
R^2	0.9845	
$\widehat{\sigma}^2$	6.996025	
$F_{2,97}$	3080	
FIV	1.019434	
$K(\mathbf{X}_{-1})$	1.001338	
$K(\mathbf{X})$	208.5459	

Tabla 6.11: Estimación por MCO del modelo del Ejemplo 7.3 centrando $\mathbf{X}_2$ o $\mathbf{X}_3$

Observación 6.4 *Téngase en cuenta que si en el modelo (7.1) se verifica que* $\sum_{i=1}^{n} X_{ij} = 0$ *para* $j = 2, \ldots, k$*, se tiene que no existe multicolinealidad aproximada del tipo no esencial, ya que la constante no está relacionada linealmente con el resto de variables independientes. En tal caso, se verifica*[16] *que* $\widehat{\beta}_1 = \overline{\mathbf{y}}$ *y, por tanto, esta igualdad prodría ser usada como restricción en los MCR.*

Así, en el Ejemplo 6.18 se verifica que $\overline{\mathbf{D}} = 6.762459$*, de manera que considerando que* $\widehat{\beta}_1 = 6.762459$ *en el sistema de ecuaciones (6.6), se obtendrían los valores:*

$$\widehat{\beta}_2 = -4.775367, \quad \widehat{\beta}_3 = 3.289646, \quad \widehat{\beta}_4 = 0.003220294.$$

Aunque los valores obtenidos son muy parecidos a los proporcionados incialmente por MCO (diferencia del 18.49323 %), sin embargo, puesto que en este caso se verifica que $\overline{\mathbf{C}} = 6.222153$*,* $\overline{\mathbf{I}} = 7.512735$ *y* $\overline{\mathbf{CP}} = 1496.902$*, se puede afirmar que se está lejos de la premisa inicial y que, por tanto, la solución obtenida al introducir la restricción sobre* β_1 *no debería tenerse en consideración.* ♢

5.3. Centrado de variables

Cuando la multicolinealidad aproximada preocupante es del tipo no esencial, es bien sabido que la solución a la misma es la de centrar (restar la media) aquellas variables que se sospechan provocan el problema. Para ilustrar esta solución, se usará el Ejemplo 7.3 donde la multicolinealidad existente es del tipo no esencial.

Ejemplo 6.19 *Dados los datos del Ejemplo 7.3 y considerando que* $\widetilde{\mathbf{X}}_2 = \mathbf{X}_2 - \overline{\mathbf{X}}_2$ *y* $\widetilde{\mathbf{X}}_3 = \mathbf{X}_3 - \overline{\mathbf{X}}_3$ *son, respectivamente, las versiones centradas de* $\mathbf{X}_2$ *y* $\mathbf{X}_3$*, a continuación se muestran*[17] *los resultados obtenidos al estimar por MCO los modelos donde se analiza* $\mathbf{y}$ *en función de* $\widetilde{\mathbf{X}}_2$ *y* $\mathbf{X}_3$ *y de* $\mathbf{X}_2$ *y* $\widetilde{\mathbf{X}}_3$ *se muestran en la Tabla 6.11.*

Observando los resultados obtenidos se tiene que:

[16] Es claro que, en este caso, el estimador por MCO se obtiene como:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} n & \mathbf{0}^t \\ \mathbf{0} & \mathbf{X}_{-1}^t\mathbf{X}_{-1} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \mathbf{X}_{-1}^t\mathbf{y} \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{y}} \\ \left(\mathbf{X}_{-1}^t\mathbf{X}_{-1}\right)^{-1} \cdot \mathbf{X}_{-1}^t\mathbf{y} \end{pmatrix},$$

donde $\mathbf{0}$ es un vector de ceros de dimensión $(k-1) \times 1$ y $\mathbf{X}_{-1}$ denota a la matriz $\mathbf{X}$ sin tener en cuenta la primera columna referente a la constante (esto es, $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_{-1}]$).

[17] Por $K(\mathbf{X}_{-1})$ se denota el cálculo del NC sin tener en cuenta la constante del modelo.

Variable	Estimación	Desviación típica
Constante	1.6842	0.2645
$\widetilde{\mathbf{X}}_2$	-1.6727	0.2758
$\widetilde{\mathbf{X}}_3$	2.0286	0.0259
R^2	0.9845	
$\widehat{\sigma}^2$	6.996025	
$F_{2,97}$	3080	
FIV	1.019434	
$K(\mathbf{X}_{-1})$	1.149077	
$K(\mathbf{X})$	1.149077	

Tabla 6.12: Estimación por MCO del modelo del Ejemplo 7.3 centrando $\mathbf{X}_2$ y $\mathbf{X}_3$

- *Cuando se centra la variable $\mathbf{X}_2$, que es la variable que provoca el problema de multicolinealidad aproximada no esencial grave, se observa que ha cambiado la estimación de la constante y de su desviación típica estimada, experimentándose una importante reducción en ésta última. Las estimaciones de los coeficientes de las restantes variables no se ven afectadas al igual que las características globales del modelo (coeficiente de determinación, estimación de la varianza de la perturbación aleatoria, etc). También se observa que el FIV coincide con el del modelo original y los NCs calculados quedan por debajo de los umbrales establecidos como preocupantes*[18].

- *Cuando se centra la variable $\mathbf{X}_3$, que no es la variable que provoca el problema de multicolinealidad aproximada no esencial grave, se observa que ha cambiado la estimación de la constante y de su desviación típica estimada, sin embargo, en este segundo caso no se produce reducción alguna. Al igual que antes, las estimaciones de los coeficientes de las restantes variables no se ven afectadas al igual que las características globales del modelo. También se observa que el FIV coincide con el del modelo original mientras que el NC sigue siendo superior a los umbrales establecidos como preocupantes. Es decir, centrar las variables que no provocan el problema es del todo innecesario ya que no tienen efecto alguno sobre el grado de multicolinealidad existente.*

Adviértase que el hecho de que el FIV se mantenga invariante se debe a que, tal y como se dijo en la Observación 6.3, es invariante a los cambios de origen y escala.

Si se estima el modelo que analiza $\mathbf{y}$ en función de las dos variables centradas se obtienen los resultados mostrados en la Tabla 6.12. Se vuelven a observar las características comentadas anteriores. Además, destaca especialmente que la estimación de la constante coincide con la media de la variable dependiente, $\overline{\mathbf{y}}$, y que el cálculo del NC con y sin constante coincide. Esta última cuestión es síntoma de que se ha eliminado la relación de $\mathbf{X}_2$ y $\mathbf{X}_3$ con la constante.

Finalmente, comentar que aunque algunas de las estimaciones de los coeficientes no han cambiado, sí lo hace su interpretación, ya que en los casos donde se ha transformado las variables los incrementos se producen con respecto a su valor medio. □

Aunque en el ejemplo anterior se atisba que el centrado de variables es una solución totalmente inoperante ante la multicolinealidad aproximada esencial, este hecho se ilustra claramente con el siguiente ejemplo.

[18] Puesto que en este caso se ha eliminado la relación lineal de $\mathbf{X}_2$ con la constante, el NC recoge la relación lineal de la constante con $\mathbf{X}_3$ y de $\mathbf{X}_2$ con $\mathbf{X}_3$.

Variable	Estimación	Desviación típica
Constante	6.92694	0.01639
Tipos de interés a 3 meses (centrados)	-0.62891	0.06582
Tipos de interés a 6 meses (centrados)	1.59334	0.06394
R^2	0.9965	
$\widehat{\sigma}^2$	0.03330625	
$F_{2,121}$	17371.66	
FIV	146.1685	
$K(\mathbf{X}_{-1})$	24.1386	
$K(\mathbf{X})$	24.1386	

Tabla 6.13: Estimación por MCO del modelo del Ejemplo 6.1 centrando los tipos de interés a 3 y 6 meses

Ejemplo 6.20 *Dados los datos del Ejemplo 6.1, si se centran las variables referentes a los tipos de interés a 3 y 6 meses y se estima por MCO el modelo que tiene como variable dependiente los tipos de interés a 12 meses e independientes las variables centradas se obtienen los resultados mostrados en la Tabla 6.13.*

Se observa que si bien ha cambiado la estimación y su desviación típica estimada, se sigue manteniendo el signo no esperado en la estimación del coeficiente de los rendimientos a 3 meses.

Por otro lado, si bien se ha reducido el valor obtenido en el NC, sigue estando por encima de los umbrales establecidos. Puesto que la multicolinealidad aproximada no esencial ha sido eliminada (se han centrado todas las variables independientes excluida la constante), la multicolinealidad aproximada que queda es del tipo esencial. Este tipo de multicolinealidad aproximada es claramente detectada por el FIV, que coincide con el del modelo original. Por tanto, se pone de manifiesto que el centrado de variables no ayuda a mitigar la multicolinealidad aproximada esencial existente en un modelo de regresión lineal múltiple. □

5.4. Estimador cresta

A pesar de que diversos autores han manifestado ciertas anomalías en el uso del estimador cresta, se trata de la técnica de estimación alternativa a los MCO más usada. Puesto que es un estimador sesgado de $\boldsymbol{\beta}$, surge con el objetivo de proporcionar un menor error cuadrático medio que el de MCO cuando el grado de multicolinealidad existente es preocupante.

Dado el modelo (7.1), su expresión responde a:

$$\widehat{\boldsymbol{\beta}}(l) = \left(\mathbf{X}^t\mathbf{X} + l \cdot \mathbf{I}\right)^{-1} \cdot \mathbf{X}\mathbf{y}, \tag{6.7}$$

donde $\mathbf{I}$ es la matriz identidad de orden $k \times k$ y l es un parámetro no negativo que tradicionalmente se considera que toma valores en el intervalo $[0, 1]$. Es claro que cuando l es positivo el estimador cresta es un estimador sesgado[19] de $\boldsymbol{\beta}$ y que para $l = 0$ coincide con el estimador por MCO. Por otro lado, su matriz de varianzas-covarianzas es:

$$var\left(\widehat{\boldsymbol{\beta}}(l)\right) = \sigma^2 \cdot \left(\mathbf{X}^t\mathbf{X} + l \cdot \mathbf{I}\right)^{-1} \cdot \mathbf{X}^t\mathbf{X} \cdot \left(\mathbf{X}^t\mathbf{X} + l \cdot \mathbf{I}\right)^{-1}.$$

[19] Como el estimador cresta se puede expresar como $\widehat{\boldsymbol{\beta}}(l) = \mathbf{Z}(l) \cdot \widehat{\boldsymbol{\beta}}$ donde $\mathbf{Z}(l) = \left(\mathbf{X}^t\mathbf{X} + l \cdot \mathbf{I}\right)^{-1} \cdot \mathbf{X}^t\mathbf{X}$, es claro que $E[\widehat{\boldsymbol{\beta}}(l)] = \mathbf{Z}(l) \cdot \boldsymbol{\beta}$. Por tanto, $E[\widehat{\boldsymbol{\beta}}(l)] \neq \boldsymbol{\beta}$ si $l \neq 0$. Al mismo tiempo, su matriz de varianzas-covarianzas se puede obtener teniendo en cuenta que $var\left(\widehat{\boldsymbol{\beta}}(l)\right) = var\left(\mathbf{Z}(l) \cdot \widehat{\boldsymbol{\beta}}\right) = \mathbf{Z}(l) \cdot var\left(\widehat{\boldsymbol{\beta}}\right) \cdot \mathbf{Z}(l)^t$.

Como se puede observar en la expresión (6.7), con el objetivo de mejorar el condicionamiento de la matriz $\mathbf{X}^t\mathbf{X}$ y conseguir así que su inversa sea más estable, se introduce un elemento constante en su diagonal principal.

Adviértase que la expresión (6.7) se puede obtener también por MCO el modelo ampliado $\mathbf{y}_a = \mathbf{X}_a \cdot \boldsymbol{\beta} + \boldsymbol{\eta}$ donde:

$$\mathbf{y}_a = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{k\times 1} \end{pmatrix}, \quad \mathbf{X}_a = \begin{pmatrix} \mathbf{X} \\ \sqrt{l}\cdot \mathbf{I}_{k\times k} \end{pmatrix},$$

donde $\mathbf{0}$ es un vector de ceros e $\mathbf{I}$ es la matriz identidad, ya que $\widehat{\boldsymbol{\beta}}_a = (\mathbf{X}_a^t\mathbf{X}_a)^{-1}\cdot\mathbf{X}_a^t\mathbf{y}_a = \widehat{\boldsymbol{\beta}}(l)$. Si bien:

$$var\left(\widehat{\boldsymbol{\beta}}_a\right) = \sigma^2\cdot\left(\mathbf{X}_a^t\mathbf{X}_a\right)^{-1} = \sigma^2\cdot\left(\mathbf{X}^t\mathbf{X}+l\cdot\mathbf{I}\right)^{-1} \neq var\left(\widehat{\boldsymbol{\beta}}(l)\right),$$

es decir, ninguna característica más de este modelo ampliado coincide con las del estimador cresta.

Observación 6.5 *Como se acaba de mostrar, la aplicación del estimador cresta supone la modificación arbitraria de los datos usados para estimar el modelo, sienda ésta una de las principales objeciones a la aplicación de este método ya que no queda muy claro cómo se han de interpretar las estimaciones obtenidas.* ♢

Finalmente, la elección idónea de este parámetro constante ha sido ampliamente estudiado, siendo la elección más común:

$$l = k\cdot\frac{\widehat{\sigma}^2}{\widehat{\boldsymbol{\beta}}^t\widehat{\boldsymbol{\beta}}}, \tag{6.8}$$

la cual tiene una probabilidad superior a 0.5 de proporcionar un error cuadrático medio menor que el de MCO y donde $\widehat{\sigma}^2$ y $\widehat{\boldsymbol{\beta}}$ son, respectivamente, las estimaciones por MCO de los parámetros σ^2 y $\boldsymbol{\beta}$. Sin embargo, en ningún momento se tiene asegurado que el problema de multicolinealidad aproximada grave haya sido mitigado.

Ejemplo 6.21 *Considerando el modelo del Ejemplo 6.2 sobre el crédito en Estados Unidos y calculando*[20] *el estimador cresta para* $l \in \{0, 0.1, 0.2, 0.3, \cdots, 0.9, 1\}$ *(lo cual se denomina traza) se obtienen los valores mostrados en la Figura 6.1, donde mediante círculos se han resaltado las estimaciones por MCO y cuadrados la del estimador cresta para* $l = 0.1807054$ *(valor obtenido al calcular la expresión (6.8)).*

Se puede observar que las estimaciones decrecen hacia cero (lo cual se prueba teóricamente), siendo la estimación de los coeficientes de las variables independientes para el valor de l seleccionado la siguiente:

$$\widehat{\boldsymbol{\beta}}(0.1807054) = (-1.341494,\ -1.550298,\ 2.364572)^t.$$

Para comprobar si el problema de multicolinealidad ha sido mitigado habría que calcular el valor del FIV o NC para $l = 0.1807054$. □

[20] El código creado en *R* para realizar los cálculos de este ejemplo se muestran en el Apéndice 2.7.

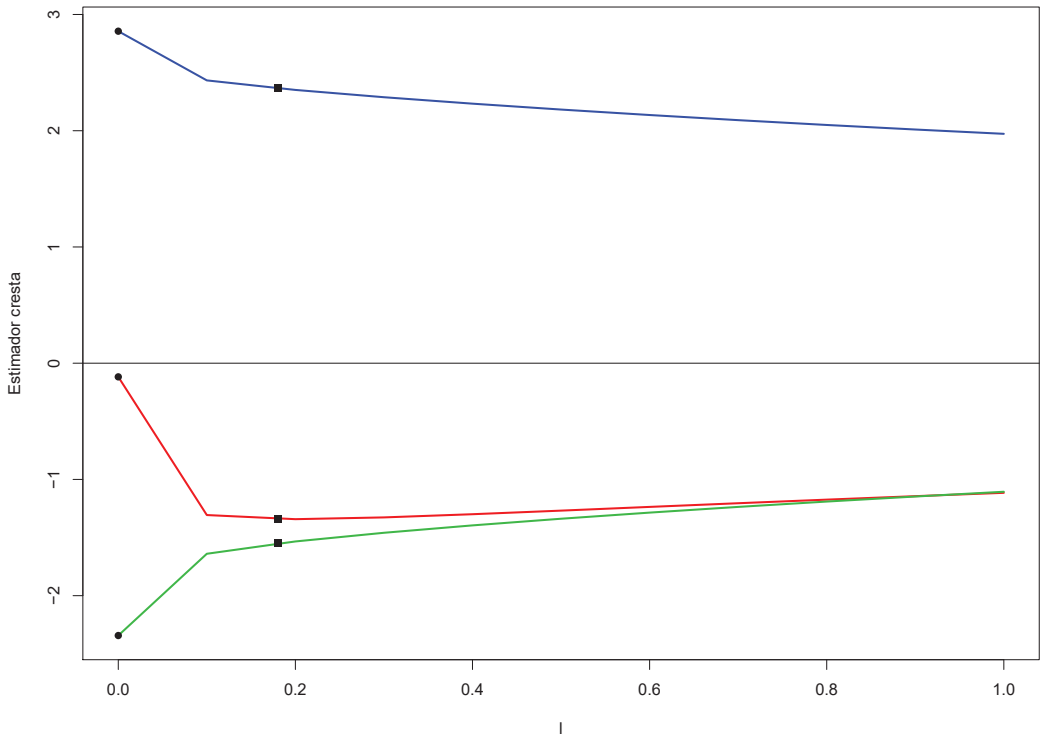

Figura 6.1: Traza del estimador cresta para los datos del Ejemplo 6.2 sobre el crédito en Estados Unidos

5.5. Variables ortogonales

El método con variables ortogonales trabaja con los residuos, $\mathbf{e}_j$, de la regresión auxiliar de la variable independiente j-ésima, $\mathbf{X}_j$, en función del resto, es decir, de la misma regresión auxiliar usada para calcular el FIV. A continuación, se sustituye la variable independiente j-ésima del modelo original por $\mathbf{e}_j$ obteniéndose el modelo con variables ortogonales siguiente:

$$\mathbf{y}_{n\times 1} = \mathbf{X}^O_{n\times k} \cdot \boldsymbol{\beta} + \mathbf{w}, \tag{6.9}$$

donde $\mathbf{X} = [\mathbf{1}, \mathbf{X}_2, \ldots, \mathbf{X}_{j-1}, \mathbf{e}_j, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_k]$. Puesto que $\mathbf{e}_j$ es ortogonal al resto de variables presentes en $\mathbf{X}^O$, se tiene de inmediato que el grado de multicolinealidad aproximada (esencial y no esencial) presente en el modelo se mitiga.

Por otro lado, se tiene que $\mathbf{e}_j$ se puede interpretar como la parte de $\mathbf{X}_j$ no relacionada con el resto de variables independientes, es decir, se obtiene una interpretación distinta para la estimación de su coeficiente asociado. Al mismo tiempo, esta nueva interpretación supone que se tenga que tener cuidado a la hora de elegir la regresión auxiliar más adecuada.

Ejemplo 6.22 *A continuación se ilustra el método de variables ortogonales mediante su aplicación al modelo de regresión del Ejemplo 6.1 sobre tipos de interés a 3, 6 y 12 meses (*$\mathbf{r3}$*,* $\mathbf{r6}$ *y* $\mathbf{r12}$*, respectivamente). Considerando las dos regresiones auxiliares posibles en dicho ejemplo:*

$$\mathbf{r3}_t = \alpha_1 + \alpha_2 \cdot \mathbf{r6}_t + \boldsymbol{\epsilon}_t, \quad \mathbf{r6}_t = \alpha_1 + \alpha_2 \cdot \mathbf{r3}_t + \boldsymbol{\epsilon}_t,$$

se tiene que en la primera, los residuos se interpretan como la parte de los rendimientos a 3 meses que no tienen relación con los de 6 meses, lo cual no tiene sentido. Sin embargo, en la

Variable	Estimación	Desviación típica
Constante	0.4404	0.0387
Tipos de interes a 3 meses (primer trimestre)	1.00569	0.00569
Tipos de interes del segundo trimestre	1.59334	0.06394
R^2	0.9965	
$\widehat{\sigma}^2$	0.03330625	
$F_{2,121}$	17371.66	

Tabla 6.14: Estimación por MCO del modelo sobre tipos de interés tras ortogonalizar los tipos de interes a 6 meses

segunda los residuos, $\mathbf{e}_{r6}$, representan la parte de los rendimientos a 6 meses no relacionados con los rendimientos a 3 meses, es decir, se estaría interpretando como los meses 4, 5 y 6 (segundo trimestre).

Considerando la segunda regresión auxiliar, la estimación[21] *por MCO del modelo con variables ortogonales se muestra en la Tabla 6.14. Se puede observar que:*

- *La estimación e inferencia de la variable ortogonalizada no cambia, si bien sí lo hace su interpretación: influencia del segundo trimestre sobre los rendimientos a 12 meses.*
- *La suma de cuadrados de los residuos de ambos modelos coinciden, por lo que se obtiene la misma estimación de la varianza de la perturbación aleatoria y los mismos valores para el coeficiente de determinación y del estadístico experimental del contraste de significación conjunta.*
- *Cambia la estimación e inferencia de las variables inalteradas. En este caso, se ha corregido la relación inversa entre los rendimientos a tres y doce meses.*
- *Se puede comprobar que las estimaciones obtenidas para las variables inalteradas coinciden con las obtenidas estimando el modelo original tras eliminar la variable ortogonalizada, es decir, con las del modelo $\mathbf{r12}_t = \delta_1 + \delta_2 \cdot \mathbf{r6}_t + \boldsymbol{\nu}_t$ (ver Tabla 6.2).*
- *Puesto que $\mathbf{e}_{r6}$ y $\mathbf{r3}$ son ortogonales entre sí, el problema de multicolinealidad esencial aproximada ha sido eliminada del modelo (FIV = 1). Además, en este caso se verifica el ceteris paribus ya que cuando aumenta una variable realmente la otra permanece constante.*
- *El NC es igual a 4.507993 y puesto que $\mathbf{e}_{r6}$ es ortogonal a la constante del modelo y a $\mathbf{r3}$, recoge la relación lineal de $\mathbf{r3}$ con la constante, es decir, mide el grado de multicolinealidad no esencial existente en el modelo con variables ortogonales (que en este caso no es preocupante).*

□

6. Conclusiones

En un modelo de regresión lineal múltiple como el dado en la expresión (7.1) siempre existe cierto grado de multicolinealidad aproximada, por lo que el objetivo ha de ser determinar si este grado existente es preocupante o no, en el sentido de si el análisis númerico (estabilidad

[21] El código usado en **R** para abordar la estimación mediante variables ortogonales se muestra en el Apéndice 2.8.

de las estimaciones de los coeficientes de las variables independientes) y estadístico (inferencia individual de las variables independientes) del modelo se ve afectado al estimarlo por Mínimos Cuadrados Ordinarios. Para lograr dicho objetivo, es interesante cuantificar cuánto cambian las estimaciones del modelo ante (pequeñas) perturbaciones en las variables independientes o si éstas difieren (en signo, al menos) por las obtenidas en las regresiones lineales simples o coeficientes de correlación lineal simples.

Por otro lado, es importante determinar el tipo de multicolinealidad aproximada presente, ya que en función de ésta se ha de poner en práctica una solución u otra para mitigarla. Para distinguirla es impotante tener en cuenta factores como el tamaño de la muestra, que el Factor de Inflación de la Varianza ignora el papel de la constante en las relaciones lineales de las variables independientes, el incremento que se experimenta en el Número de Condición al pasar de no tener en cuenta la constante a sí tenerla o posibles contradicciones en los valores obtenidos en las dos medidas anteriores.

Sin embargo, un grado de multicolinealidad grave no siempre es un problema, ya que precisamente las relaciones lineales conducen a un ajuste lineal alto (coeficientes de determinación grandes) y, por tanto, a predicciones de calidad (siempre y cuando las nuevas observaciones respeten la estructura de multicolinealidad existente en las observaciones usadas para la estimación inicial).

E incluso, en el caso que de estar interesados en la inferencia del modelo, es importante dirimir si aquella variable en la que se está interesado se ve afectada o no por el problema. Así, en el Ejemplo 6.18, se tiene que la estimación de la variable ingresos por MCR es igual a 2.57108, mientras que la de su versión perturbada es 1.92256. Es decir, aunque el resto de estimaciones son inestables, la de esta variable no parece serlo. Igual ocurre con la significación individual, ya que en ambos casos se rechaza la hipótesis nula de que el coeficiente es cero. Una situación similar nos encontramos en el Ejemplo 7.3, donde la estimación del coeficiente de la tercera variable (no afectada por el problema de multicolinealdiad no esencial detectado) se acerca mucho al verdadero valor con el que se han simulado los datos.

En definitiva, cada modelo es único y, por tanto, el problema de multicolinealidad aproximada ha de estudiarse de manera individual alejándose de soluciones universales. En este sentido, a continuación se muestra un modelo en el que se determina que el grado de multicolinealidad aproximada existente es grave y, sin embargo, parece no afectar al análisis del modelo, por lo que (quizás) la mejor solución sea no hacer nada.

Ejemplo 6.23 *En el siguiente ejemplo se simulan*[22] *datos sobre el consumo e ingresos mensuales (ambas medidas en euros) y género (codificada como 1 si se es hombre y 0 en caso contrario) de 50 individuos residentes en Alemania. Con el objetivo de provocar un problema de multicolinealidad grave se generan los datos con el objetivo de que los hombres tengan más ingresos y gastos que las mujeres, lo cual se ve reflejado en los siguientes valores medios:*

Variable	*Hombre*	*Mujer*
Consumo	*2199.24*	*1799.84*
Ingresos	*2999.96*	*2200.4*

La estimación por MCO del consumo en función de los ingresos y el género en su versión original y perturbada se muestra en la Tabla 6.15. En ambos casos, el NC indicaría un problema de multicolinealidad aproximada grave. Además, el NC excluyendo la constante del modelo, indica que la multicolinealidad existente es del tipo no esencial. Este hecho, se contradice con los valores obtenidos para el FIV, que en ambos casos indicarían la existencia

[22] El código usado en **R** está disponible en el Apéndice 2.9.

Variable	Estimación	Desv. Típica
Constante	1864.59259	354.942
Ingresos	-0.029443	0.1613
Género	422.92917	128.982
R^2	0.9995	
$\widehat{\sigma}^2$	23.12648	
$F_{2,47}$	43100	
FIV	8990.384	
$K(\mathbf{X}_{-1})$	3.054154	
$K(\mathbf{X})$	1347.898	

Variable	Estimación	Desv. Típica
Constante	1753.487	72.93
Ingresos_p	0.020902	0.03289
Género	382.7287	26.26
R^2	0.9995	
$\widehat{\sigma}^2$	22.95368	
$F_{2,47}$	43440	
FIV	375.7167	
$K(\mathbf{X}_{-1})$	3.047254	
$K(\mathbf{X})$	277.6645	

Tabla 6.15: Estimación por MCO del consumo e ingresos mesuales

de multicolinealidad aproximada esencial en el modelo (ya que el FIV no detecta la no esencial).

En este caso, debido a que una de las dos variables para las que se calcula el FIV es binaria, tal y como se comenta en la Observación 6.3, el coeficiente de determinación de la regresión auxiliar no es representativo. Luego, el FIV queda en entredicho y es más adecuado usar el NC para determinar si la multicolinealdiad aproximada existente es preocupante.

Adviértase que, por la misma razón, no es posible calcular el coeficiente de correlación simple entre los ingresos y el género y, por tanto, tampoco el determinante de la matriz de correlaciones.

Una vez determinada que la multicolinealidad existente es preocupante, ¿afecta ésta al análisis del modelo? Desde el punto de vista numérico, se tiene que una perturbación de un 1 % en los salarios supone un cambio del 6.179769 % en las estimaciones. Mientras que desde el punto de vista de la inferencia, en ambos casos, la estimación de la constante y del coeficiente del género son ignificativamente distintos de cero y el modelo es globalmente válido. Es decir, las desviaciones típicas estimadas infladas (en teoría) parecen no afectar a la inferencia del modelo.

Por tanto, parece que la multicolinealidad aproximada existente en el modelo aún siendo grave, según el NC, no afecta al análisis del modelo ni desde el punto de vista numérico ni estadístico. Además, dicha multicolinealidad puede favorecer la predicción a realizar a partir del mismo. Luego, aunque parezca contradictorio, quizás la mejor solución sea no hacer nada. □

7. Ejercicios Propuestos

1. Especificar el código necesario para obtener los resultados mostrados en el Ejemplo 6.1 sobre la regresión lineal simple.

2. El director de Marketing de cierta empresa láctea desea analizar la posible relación que puede tener el gasto en publicidad, $\mathbf{G}$, sobre el número de ventas anuales de leche para bebés, $\mathbf{V}$, realizadas en los últimos 15 años. Con tal objetivo analiza el siguiente modelo de regresión lineal $\mathbf{V} = \alpha + \beta \cdot \mathbf{G} + \mathbf{u}$, donde $\mathbf{u}$ representa a la perturbación aleatoria la cual se supone esférica. ¿Es posible que exista un grado de multicolinealidad aproximada preocupante en este modelo? En caso de existir, ¿qué tipo de multicolinealidad sería? ¿Cómo resolvería el problema?

3. En el modelo de regresión $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \mathbf{u}$ se verifica que $\mathbf{X} = 0.5 \cdot \mathbf{Z}$. ¿Qué parámetros son estimables? ¿Y qué combinaciones lineales de los parámetros?

4. En el modelo de regresión $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \mathbf{u}$ se verifica que $\mathbf{X} = 2 \cdot \mathbf{Z}$. ¿Qué parámetros son estimables si se sabe que $\beta_3 = 1$?

5. Supongamos que en el modelo de regresión $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$ se dispone de las siguientes matrices de diseño:

$$a)\ \mathbf{X} = \begin{pmatrix} 1 & 4 & -2 \\ 1 & -2 & 1 \\ 1 & 10 & -5 \\ 1 & 0 & 0 \\ 1 & -1 & 0.5 \end{pmatrix}, \quad b)\ \mathbf{X} = \begin{pmatrix} 1 & 4 & -2 \\ 1 & 3.9 & 1 \\ 1 & 4.1 & -5 \\ 1 & 4 & 0 \\ 1 & 3.95 & 0.5 \end{pmatrix}, \quad c)\ \mathbf{X} = \begin{pmatrix} 1 & 4 & 3.85 \\ 1 & 3.9 & 3.71 \\ 1 & 4.1 & 3.9 \\ 1 & 4 & 3.8 \\ 1 & 3.95 & 3.7 \end{pmatrix}.$$

   Indique qué tipo de multicolinealidad existe en cada caso, qué parámetros son estimables y cómo mitigaría el problema, si es que es posible.

6. Dado el modelo $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$ donde se verifica que $2 \cdot \mathbf{X}_2 + \mathbf{X}_3 = 1$, se pide contestar razonadamente a las siguientes cuestiones:

   a) ¿Es posible obtener una estimación de β_1, β_2 y β_3 por Mínimos Cuadrados Ordinarios?

   b) ¿Cuál es la estimación de β_1 y β_2 sabiendo que $\widehat{\mathbf{y}} = 3.25 - 1.5 \cdot \mathbf{X}_2$ y $\beta_3 = 0.25$?

7. En el modelo de regresión $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \beta_4 \cdot \mathbf{W} + \mathbf{u}$ se verifica que $\mathbf{X} = \mathbf{Z} - \mathbf{W}$. ¿Qué parámetros son estimables? ¿Y qué combinaciones lineales de los parámetros? ¿Y si la relación fuese $\mathbf{Z} = \mathbf{W} + 2$ o $\mathbf{X} = 5 \cdot \mathbf{W}$?

8. Repetir el análisis realizado en el Ejemplo 6.2 sobre el crédito en Estados Unidos incorporando al modelo la variable independiente referente al crédito pendiente al consumidor, **CP**. ¿Siguen apareciendo en el modelo síntomas de multicolinealidad aproximada? Usar el código mostrado en el Apéndice 2.1.

9. Adaptar el código mostrado en el Apéndice 2.2 para reproducir el Ejemplo 12.1 en el caso en el que se sustituye $\mathbf{X}_2 = 1 + \mathbf{X}_3$ en el modelo inicial y se simulan 200 observaciones. ¿La generación de un mayor número de observaciones tiene alguna consecuencia en la estimación por MCO?

10. Adaptar el código mostrado en el Apéndice 2.2 para reproducir el Ejemplo 12.1 en el caso de que la relación lineal entre $\mathbf{X}_2$ y $\mathbf{X}_3$ tengan una baja relación lineal. Comprobar, usando el código mostrado en el Apéndice 2.3 sobre el Ejemplo 6.6, que en este caso pequeños cambios en los datos no afecta sustancialmente a la estimación por MCO del modelo.

11. Especificar el código necesario para obtener los resultaados mostrados en el Ejemplo 6.7.

12. Dados los datos de la Tabla C.1 sobre el crédito en Estados Unidos, calcular el Número de Condición especificando la versión normalizada de la matriz de diseño. Usar el código mostrado en el Apéndice 2.4.

13. Especificar el código necesario para obtener los resultaados mostrados en los Ejemplos 6.8 y 6.10.

14. Usar el código del Apéndice 2.5 para calcular los Factores de Inflación de la Varianza del Ejemplo 6.13.

15. Para analizar el volumen de consumo textil per cápita, **C**, en los años 1923 a 1939 se considera el ingreso pér cápita, **I**, y el índice de precios de los textiles, **P**, obteniéndose la siguiente estimación por Mínimos Cuadrados Ordinarios:

$$\widehat{\mathbf{C}} = 130.707 + 1.06171 \cdot \mathbf{I} - 1.383 \cdot \mathbf{P}, \quad R^2 = 0.9513.$$

Teniendo en cuenta que los Factores de Inflación de la Varianza y el Número de Condición son iguales, respectivamente, a 1.033 y 48.953. ¿Existe un grado de multicolinealidad aproximada preocupante en este modelo? En caso de existir, ¿qué tipo de multicolinealidad aproximada sería? ¿Cómo resolvería el problema?

16. En el modelo en el que se explica el reparto de dividendos de una empresa, **D**, a partir del endeudamiento a corto plazo de la misma, **EC**, del endeudamiento a largo plazo, **EL**, y del número de ventas anuales, **V**, se sospecha que pueda existir multicolinealidad aproximada preocupante debido a la relación entre las variables **EC** y **EL**. Por tal motivo se realizan las siguientes regresiones:

- regresión de la variable **EC** sobre el resto de variables independientes del modelo, obteniéndose un coeficiente de determinación de 0.990727.
- regresión de la variable **EL** sobre el resto de variables independientes del modelo, obteniéndose un coeficiente de determinación de 0.9907107.
- regresión de la variable **V** sobre el resto de variables independientes del modelo, obteniéndose un coeficiente de determinación de 0.01864573.

¿Existe multicolinealidad aproximada preocupante en el modelo? En caso afirmativo, ¿cómo la solucionaría?

17. En el modelo $\mathbf{C} = \beta_1 + \beta_2 \cdot \mathbf{R} + \beta_3 \cdot \mathbf{H} + \mathbf{u}$ donde **C** es el consumo familiar, **R** es la renta familiar y **H** el número de hijos, se ha obtenido que el autovalor más grande de $\mathbf{X}^t\mathbf{X}$ convenientemente transformada es 143.08, mientras que el más pequeño es 2.2. ¿Existe multicolinealidad aproximada preocupante en el modelo?

18. Si al modelo del ejercicio anterior se le añade una nueva variable que mida el número de miembros de la familia con trabajo, el autovalor máximo pasa a ser 243.7 y el mínimo a 0.15. ¿Ha cambiado el grado de multicolinealidad aproximada? Indique cuáles son las consecuencias de la presencia de multicolinealidad aproximada grave y cómo resolvería este problema.

19. Suponga que dado el modelo $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$ se tiene que la matriz de correlaciones de las variables independientes es:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.95 & 0.3 \\ 0.95 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{pmatrix}.$$

Usando el FIV, ¿se puede considerar que la multicolinealidad aproximada existente en dicho modelo es preocupante? ¿Qué tipo de multicolinealidad aproximada sería?

20. Suponga que se ha estimado por Mínimos Cuadrados Ordinarios un modelo de regresión en el que se desea explicar el consumo a partir del ingreso y la riqueza. Detecte la posible presencia de multicolinealidad aproximada grave sabiendo que para dichas variables se tiene la siguiente matriz de corralaciones:

	Consumo	Ingreso	Riqueza
Consumo	1	0.9851	0.985
Ingreso	0.9851	1	0.9644
Riqueza	0.985	0.9644	1

21. Indicar a partir de las siguientes matrices de correlaciones de variables independientes si el grado de multicolinealidad aproximada existente en el modelo correspondiente es preocupante:

$$R = \begin{pmatrix} 1 & -0.8 & 0.8 \\ -0.8 & 1 & -0.75 \\ 0.8 & -0.75 & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0.9 & -0.75 \\ 0.9 & 1 & -0.9 \\ -0.75 & -0.9 & 1 \end{pmatrix}.$$

¿Qué tipo de multicolinealidad aproximada sería? Especificar la ecuación del modelo de regresión lineal múltiple.

22. A partir del código del Apéndice 2.6 del Ejemplo 6.14, especificar el código necesario para obtener los resultados de los Ejemplos 6.15 y 6.16.

23. Se ha estimado por Mínimos Cuadrados Ordinarios un modelo que analiza la inflación, **I**, en función del desempleo, **D**, y el porcentaje de cambio de los salarios, **S**, entre los años 1980 y 2019, obteniéndose los siguientes resultados:

$$\widehat{\mathbf{I}} = -2.76 + 0.405 \cdot \mathbf{D} + 1.07 \cdot \mathbf{S}, \quad R^2 = 0.814.$$

Analice si la multicolinealidad existente es preocupante teniendo en cuenta que:

$$\widehat{\mathbf{D}} = 5.81 + 0.0608 \cdot \mathbf{S}, \quad SCT = 78.7432, \quad \widehat{\sigma}^2 = 2.2349.$$

24. Suponga que los autovalores de la matriz $\mathbf{X}^t\mathbf{X}$ (debidamente transformada) asociada al modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ son 0.0015, 0.0687, 1.256 y 3.023. ¿Existe algún problema en el modelo que desaconseje su estimación por MCO?.

25. Suponga que en el modelo $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \mathbf{u}$ se verifica que la correlación simple entre **X** y **Z** es igual a cero y el Número de Condición a 55. ¿Cuánto vale el Factor de Inflación de la Varianza? ¿Existe multicolinealidad aproximada preocupante en el modelo? En caso de existir, ¿de qué tipo de multicolinealidad aproximada se trata? ¿Cómo la mitigaría?

26. Suponga que en el modelo $\mathbf{y}_t = \beta_1 + \beta_2 \cdot \mathbf{X}_t + \beta_3 \cdot \mathbf{Z}_t + \mathbf{u}_t$ se verifica que el Número de Condición calculado considerando la constante y sin considerarla coincide, y es igual a 55. ¿Existe multicolinealidad aproximada preocupante en el modelo? En caso de existir, ¿de qué tipo de multicolinealidad aproximada se trata?

27. Reproducir el Ejemplo 6.18 partiendo de las restantes regresiones auxiliares posibles para calcular los FIVs.

28. Calcular el porcentaje de variación sufrida en las estimaciones proporcionadas en el Ejemplo 6.18 ante una perturbación del 1 % en las variables independientes.

29. Calcular el porcentaje de variación sufrida en las estimaciones proporcionadas en la Observación 6.18 con respecto a las obtenidas en la estimación por Mínimos Cuadrados Ordinarios inicial.

30. Generar el código necesario para obtener los resultados de los Ejemplos 6.19 y 6.20.

31. Comprobar que aplicar el método de regresión con variables ortogonales al modelo de regresión lineal simple es equivalente al centrado de variables presentado en la subsección 5.3.

32. Dado el modelo $\mathbf{y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$, para el cual se dispone de la siguiente información muestral:

$$\mathbf{y} = \begin{pmatrix} 4 \\ -3 \\ 3 \\ 2 \\ -4 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 4 & 2 & 6 \\ 1 & -4 & -2 & -6 \\ 1 & 5 & 2.5 & 7.5 \\ 1 & -3 & -1.5 & -4.5 \\ 1 & -2 & -1 & -3 \end{pmatrix},$$

proporcionar una estimación de β_1, β_2, β_3 y β_4 sabiendo que $\widehat{\mathbf{y}} = 5 + 2 \cdot \mathbf{X}_2 + 4 \cdot \mathbf{X}_3$.

33. Dadas las matrices de diseño siguientes:

$$a)\ \mathbf{X} = \begin{pmatrix} 1 & 4 & 1.99 \\ 1 & 3 & 1.51 \\ 1 & -4 & -2.05 \\ 1 & -3 & -1.45 \end{pmatrix}, \quad b)\ \mathbf{X} = \begin{pmatrix} 1 & 4 & 7 \\ 1 & 3 & -7 \\ 1 & -4 & -13 \\ 1 & -3 & +13 \end{pmatrix},$$

indicar, en cada caso, qué tipo de multicolinealidad existe y si es o no preocupante.

34. En la Tabla 10.1 se muestran datos sobre volumen de consumo textil per cápita, **C**, ingreso real per cápita con base 1925, **I**, y el precio relativo de los textiles con base 1925, **P**, en Estados Unidos en los años 1923 a 1939. Se pide:

    a) Estimar por MCO el modelo $\mathbf{C} = \beta_1 + \beta_2 \cdot \mathbf{I} + \beta_3 \cdot \mathbf{P} + \mathbf{u}$.

    b) Comprobar que en el modelo anterior el grado de multicolinealidad aproximada existente es preocupante.

    c) ¿De qué tipo de multicolinealidad aproximada se trata? ¿Cómo la resolvería?

35. Analizar los conjuntos de datos mostrados en las Tablas C.2 y C.3 teniendo en cuenta que la variable dependiente es la de la segunda columna.

Año	**C**	**I**	**P**
1923	99.2	96.7	101
1924	99	98.1	100.1
1925	100	100	100
1926	111.6	104.9	90.6
1927	122.2	104.9	86.5
1928	117.6	109.5	89.7
1929	121.1	110.8	90.6
1930	136	112.3	82.8
1931	154.2	109.3	70.1
1932	153.6	105.3	65.4
1933	158.5	101.7	61.3
1934	140.6	95.4	62.5
1935	136.2	96.4	63.6
1936	168	97.6	52.6
1937	154.3	102.4	59.7
1938	149	101.6	59.5
1939	165.5	103.8	61.3

Tabla 6.16: Datos sobre consumo textil en Estados Unidos

APÉNDICE C

Apéndice Capítulo 6

1. Conjuntos de datos usados

1.1. Crédito en los Estados Unidos

En la Tabla C.1 se tienen los datos disponibles para la deuda pendiente de hipoteca (**D**, en billones de dólares), del consumo personal (**C**, en billones de dólares), de los ingresos personales (**I**, en billones de dólares) y del crédito pendiente al consumidor (**CP**, en billones de dólares) para los años 1996 a 2012. Estos datos referentes al crédito en los Estados Unidos han sido tomados de Wissell [146].

También se muestran (dos últimas columnas) los datos de **C** e **I** perturbados 1 % siguiendo el código del apartado 2.3. Se tiene que las medias de $\mathbf{C} - \mathbf{C}_p$ e $\mathbf{I} - \mathbf{I}_p$ son, respectivamente, 0.05755971 y 0.06612741. Es decir, ambos conjuntos de datos son muy parecidos.

1.2. Consumo e ingresos salariales en los Estados Unidos

En la Tabla C.2 se muestran los datos del modelo usado por Klein y Goldberger [73] sobre el consumo e ingresos salariales en los Estados Unidos para los años 1936 a 1952 (los datos de 1942 a 1944 no están disponibles por estar en guerra). Más concretamente, se tiene el consumo, **C**, son los ingresos salariales, **I**, los ingresos no agrícolas, **InA**, y los ingresos agrícolas, **IA**.

1.3. Consumo e ingresos salariales en los Estados Unidos

En la Tabla C.3 se muestran los datos recogidos por el Banco de Thailandia durante los años 1989 a 2005 sobre el dinero depositado en bancos comerciales, **D**, el crédito privado, **C**, el ingreso nacional, **IN**, el Producto Interno Bruto, **PIB**, la inversión, **I**, y el tipo de cambio con el dólar estadounidense, **TC**.

Año	**D**	**C**	**I**	**CP**	$\mathbf{C}_p$	$\mathbf{I}_p$
1996	3.80510	4.7703	4.8786	808.23	4.815389	4.987343
1997	3.94580	4.7784	5.0510	798.03	4.878906	5.032297
1998	4.05790	4.9348	5.3620	806.12	4.928561	5.461873
1999	4.19130	5.0998	5.5585	865.65	5.164388	5.592996
2000	4.35850	5.2907	5.8425	997.30	5.256441	5.899322
2001	4.54530	5.4335	6.1523	1140.70	5.49978	6.197405
2002	4.81490	5.6194	6.5206	1253.40	5.66251	6.619613
2003	5.12860	5.8318	6.9151	1324.80	5.931252	6.901483
2004	5.61510	6.1258	7.4230	1420.50	6.16048	7.560622
2005	6.22490	6.4386	7.8024	1532.10	6.505051	7.792404
2006	6.78640	6.7394	8.4297	1717.50	6.82464	8.456612
2007	7.49440	6.9104	8.7241	1867.20	6.980757	8.781423
2008	8.39930	7.0993	8.8819	1974.10	7.081545	8.759557
2009	9.39510	7.2953	9.1636	2078.00	7.375371	9.260236
2010	10.68000	7.5614	9.7272	2191.30	7.603095	9.616988
2011	12.07100	7.8036	10.3010	2284.90	7.87653	10.34256
2012	13.44821	8.0441	10.9830	2387.50	8.093913	10.93781

Tabla C.1: Datos referentes al crédito en los Estados Unidos

Año	**C**	**I**	**InA**	**IA**
1936	62.8	43.41	17.1	3.96
1937	65	46.44	18.65	5.48
1938	63.9	44.35	17.09	4.37
1939	67.5	47.82	19.28	4.51
1940	71.3	51.02	23.24	4.88
1941	76.6	58.71	28.11	6.37
1945	86.3	87.69	30.29	8.96
1946	95.7	76.73	28.26	9.76
1947	98.3	75.91	27.91	9.31
1948	100.3	77.62	32.3	9.85
1949	103.2	78.01	31.39	7.21
1950	108.9	83.57	35.61	7.39
1951	108.5	90.59	37.58	7.98
1952	111.4	95.47	35.17	7.42

Tabla C.2: Datos referentes al salario mensual en los Estados Unidos

Año	C	IN	PIB	I	TC
1989	118.7	1107.6	1440.1	651.18	25.7
1990	1426	1479	1672.9	902.98	25.59
1991	1730.6	1789.7	1910.4	1073.9	25.52
1992	2010.6	2161.7	2145.7	1131.3	25.4
1993	1397.3	2662.9	2402.8	1266.4	25.32
1994	2710.6	3463.3	2740.6	1460.9	25.15
1995	3203.6	4300.9	3149.9	1762.2	24.92
1996	3643.3	4911.4	3394	1928.2	25.34
1997	4224.7	6060.9	3437.7	1593.2	31.37
1998	4595.9	5472.7	3311	945.97	41.37
1999	4575	5248.3	3334.8	950.61	37.84
2000	4816	4723.7	3628.7	1117.6	40.16
2001	5009.1	4447.9	3776.16	1237.09	44.48
2002	5132	4779.9	3983.53	1297.33	43
2003	5358.1	4954.3	4306.84	1477.48	41.53
2004	5497	5284.3	4794.92	1739.75	40.27
2005	5956.6	5710.3	5182.75	2231.75	40.27

Tabla C.3: Datos referentes al depósito de dinero en bancos comerciales de Thailandia

2. Código en R de las simulaciones realizadas

A lo largo del capítulo se han simulado distintos conjuntos de datos así como realizado varios análisis estadísticos. Si bien la cartera de programas informáticos disponibles para realizarlos es amplia, se ha optado por usar el entorno de programación **R** (`https://www.r-project.org/`). A continuación se muestra el código usado en los distintos ejemplos considerados. Si no quedasen claras las características de algún comando, escribir en la ventanas de comandos *help()* (por ejemplo, *help(mean)*).

2.1. Ejemplo 6.2

En el siguiente código, en primer lugar (línea 1), se eliminan todas las variables existentes en **R** para evitar posibles valores previos existentes en la memoria que pudieran intervenir en los cálculos, a continuación se incorporan los datos mediante el comando *read.table()* (línea 3) y se carga el nombre de las variables a la memoria de **R** (línea 5) para poder ser referenciadas al realizar cualquier análisis. Finalmente, en la línea 7 se estima el modelo por MCO mediante el comando *lm()* y, a continuación (línea 8), se presentan los resultados en pantalla mediante el comando *summary()*.

```
rm(list = ls())
2
datos = read.table("Wissel.txt", header = T, sep=";")
4 head(datos)
attach(datos)
6
reg = lm(D~C+I)
8 summary(reg)

10 detach(datos)
```

2.2. Ejemplo 12.1

En el siguiente código se generan 50 observaciones (línea 1) para tres variables aleatorias normales (líneas 3, 4 y 7, consultar el comando *rnorm()*). La estimación por MCO y presentación de resultados se realiza en las líneas 10 y 11. En la línea 13 se almacenan en la variable *beta* las estimaciones de los coeficientes de las variables independientes y en las dos líneas finales se calculan las estimaciones de las combinaciones lineales de los coefcientes de las variables independientes indicadas en el Ejemplo 12.1.

```
T = 50

X2 = rnorm(T, 10, 10)
p = rnorm(T, 1, 0.01)
X3 = X2 - p

u = rnorm(T, 0, 1)
y = 5 + 2*X2 - 4*X3 + u

reg = lm(y~X2+X3)
summary(reg)

beta = as.double(reg$coefficients)

beta[1]-beta[3]
beta[2]+beta[3]
```

2.3. Ejemplo 6.6

A continuación se muestra el código en **R** generado para el Ejemplo 6.6. En primer lugar, se tiene una función creada para perturbar un conjunto de datos $\mathbf{x}$, de tamaño n, un $tol\,\%$ mediante la expresión:

$$\mathbf{x}_p = \mathbf{x} + tol \cdot \mathbf{p} \cdot \frac{||\mathbf{x}||}{||\mathbf{p}||},$$

donde $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$ (esta norma se calcula mediante el comando *norm()*), para lo cual se genera un conjunto de datos aleatorios $\mathbf{p}$ mediante una normal para una media y varianza concreta. De esta forma, para $tol = 1$, se obtiene que:

$$||\mathbf{x}_p|| = ||\mathbf{x}|| + 0.01 \cdot ||\mathbf{p}|| \cdot \frac{||\mathbf{x}||}{||\mathbf{p}||} = 1.01 \cdot ||\mathbf{x}||.$$

El código es el siguiente:

```
perturb <- function(x, media, dv, tol){
    p = rnorm(length(x), media, dv)
    x.p = x + p*tol*(norm(x,"2")/norm(p,"2"))
    return(x.p)
}
```

A continuación, se perturban las variables referentes al consumo e ingresos (líneas 11 y 12), para a continuación estimar el modelo con las variables independientes perturbadas. En las líneas 17 y 18 se calcula el porcentaje de perturbación introducido en los datos y en la línea 19 el porcentaje de cambio producido en las estimaciones del modelo. Ambos cambios se calculan a partir de la tasa de variación:

$$\frac{||\mathbf{x} - \mathbf{x}_p||}{||\mathbf{x}||}, \quad \frac{||\boldsymbol{\beta} - \boldsymbol{\beta}_p||}{||\boldsymbol{\beta}||},$$

donde el subíndice hace referencia al vector perturbado.

```
datos = read.table("Wissel.txt", header = T, sep=";")
head(datos)
attach(datos)

reg = lm(D~C+I)
beta = as.double(reg$coefficients)

tol = 0.01
media = 10
dv = 10
Cp = perturb(C, media, dv, tol)
Ip = perturb(I, media, dv, tol)
reg.p = lm(D~Cp+Ip)
summary(reg.p)
beta.p = as.double(reg.p$coefficients)

(norm(C-Cp,"2")/norm(C,"2"))*100
(norm(I-Ip,"2")/norm(I,"2"))*100
(norm(beta-beta.p,"2")/norm(beta,"2"))*100
```

2.4. Ejemplo 6.9

A continuación se implementan en **R** las funciones necesarias para calcular el Número de Condición. En la primera función (líneas 1 a 7) se implementa la expresión (6.3) para obtener la versión normalizada (longitud unidad) de una matriz, mientras que en la segunda se calcula el NC según la expresión (6.2) a partir de la matriz $\mathbf{X}$ convenientemente transformada (línea 10). Los comandos *crossprod()* (línea 11) permite calcular $\mathbf{X}^t\mathbf{X}$ y *eigen()* (línea 12) los autovalores y autovectores de una matriz cuadrada. Con los comandos *max()* y *min()* se obtienen, respectivamente, el máximo y mínimo de un vector. Finalmente, *sqrt()* calcula la raíz cuadrada. En ambos casos, mediante *return()* se especifica el valor a devolver por la función.

```
lu <- function(X){
        Xlu = matrix( , nrow=dim(X)[1], ncol=dim(X)[2])
        for(i in 1:dim(X)[2]){
                Xlu[,i] = X[,i]/norm(X[,i],"2")
        }
        return(Xlu)
}

NC <- function(X){
     X = lu(X)
     XX = crossprod(X)
     landas = eigen(XX)[[1]]
     nc = sqrt(max(landas)/min(landas))
     return(nc)
}
```

Para obtener los resultados mostrados en el Ejemplo 6.9 para $\mathbf{X}_1$ se ha usado el código:

```
X1 = matrix(c(1,21,2,1,22,2,1,21,1,1,22,2,1,22,2,1,21,1,1,21,1),
          nrow=7,ncol=3,byrow=T)
NC(X1)
```

Mediante el comando *matrix()* se crea una matriz con número de filas y columnas iguales a las especificadas en *nrow* y *ncol*. Con *byrow* se consigue que los valores se ordenen por filas.

2.5. Ejemplo 7.3

En este caso, en primer lugar se tiene una función que permite calcular los FIVs basándose en un algoritmo recursivo que obtiene los coeficientes de todas las posibles regresiones auxiliares. Adviértase que la entrada de esta función es la matriz de diseño **X** tras haber eliminado la primera columna de unos correspondiente a la constante.

```
1   FIV <- function(X)
    {
3       fiv=array(0,dim(X)[2])
        for (i in 1:dim(X)[2]) {
5           reg_aux = lm(X[,i]~X[,-i])
            R2 = summary(reg_aux)[[8]]
7           fiv[i] = 1/(1-R2)
        }
9       return(fiv)
    }
```

A continuación se generan en **R** las variables del Ejemplo 7.3 (líneas 3 a 6), se estima el modelo y se muestran los resultados en pantalla (líneas 8 y 9), se calcula el FIV (línea 12) y el NC (línea 14 sin tener en cuenta la constante y en la línea 16 teniéndola en cuenta) y, finalmente, se calcula el incremento experimentado en el NC. La constante se genera mediante el comando *array()*.

```
    T = 100
2
    X2 = rnorm(T, 100, 1)
4   X3 = rnorm(T, 100, 10)
    u = rnorm(T, 0, 3)
6   y = 3 - 2*X2 + 2*X3 + u

8   reg = lm(y~X2+X3)
    summary(reg)
10
    x = cbind(X2,X3)
12  FIV(x)

14  NC(x)
    X = cbind(array(1,T),x)
16  NC(X)

18  ((NC(X)-NC(x))/NC(X))*100
```

2.6. Ejemplo 6.14

A continuación se muestra el código usado en el Ejemplo 6.14. En la línea 1 se define la matriz de correlaciones, en la 3 se calcula su determinante, en la 4 su inversa y en la 5 se selecciona la diagonal principal de dicha inversa (es decir, los FIV).

```
    R = matrix(c(1, 0.85, 0.85, 0.85, 1, 0.85, 0.85, 0.85, 1),
2                   ncol=3, nrow=3, byrow=T)
    det(R)
4   solve(R)
    diag(solve(R))
```

2.7. Ejemplo 6.21

A continuación se muestra el código usado en el Ejemplo 6.21. En la línea 8 se crea la matriz de variables independientes para a partir de sus dimensiones (líneas 9 y 10) obtener

el número de observaciones y variables ondependientes. En las líneas 12 a 18 se calcula el estimador cresta según la expresión (6.7) para $l \in \{0, 0.1, 0.2, 0.3, \cdots, 0.9, 1\}$, en la línea 20 se representan los valores obtenidos de forma conjunta (indispensable usar el comando *ts()*) y en la 21 se añade una asíntota horizontal que representa al eje de abscisas.

En las líneas 24 a 29 se estima el modelo por MCO, se almacenan los residuos y estimación de los coeficientes de las variables independientes y se agregan éstas últimas a la representación gráfica mediante el comando *points()*. En las líneas 31 y 32 se calculan la estimación de la varianza de la perturbación aleatoria y el valor de l según la expresión (6.8). Finalmente, en las líneas 34 a 37 se obtiene el estimador cresta para el valor concreto de l calculado anteriormente y se incorpora a la representación gráfica.

```
1   rm(list = ls())
    source("_funciones.txt")
3
    datos = read.table("Wissel.txt", header = T, sep=";")
5   head(datos)
    attach(datos)
7
    X = cbind(array(1,length(D)),C,I)
9   n = dim(X)[1]
    k = dim(X)[2]
11
    betal = matrix(,k,length(seq(0,1,0.1)))
13  i = 1
    for (l in seq(0,1,0.1))
15  {
      betal[,i] = solve(crossprod(X)+l*diag(k))%*%crossprod(X,D)
17    i = i + 1
    }
19
    plot(ts(t(betal), start=0, frequency=10), plot.type="single",
21             col=2:4, lwd=2, ylab="Estimador cresta", xlab="l")
    abline(h=0)
23
    reg = lm(D~C+I)
25  e = reg$residuals
    beta = reg$coefficients
27  points(0, beta[1], lwd=2, pch=16)
    points(0, beta[2], lwd=2, pch=16)
29  points(0, beta[3], lwd=2, pch=16)

31  sigma2 = crossprod(e)/(n-k)
    l = as.numeric(k*(sigma2/crossprod(beta)))
33
    beta.l = solve(crossprod(X)+l*diag(k))%*%crossprod(X,D)
35  points(l, beta.l[1], lwd=2, pch=15)
    points(l, beta.l[2], lwd=2, pch=15)
37  points(l, beta.l[3], lwd=2, pch=15)

39  detach(datos)
```

2.8. Ejemplo 6.22

A continuación se muestra el código usado en el Ejemplo 6.22. En las líneas 5 y 6 se obtienen los resultados mostrados en el Ejemplo 6.1 mientras que en las líneas 8 a 14 se calcula el FIV y NC (con y sin constante).

En la línea 16 se estima la regresión auxiliar, de la que se almacenan sus residuos (línea 17). Finalmente, en las líneas 19 y 20 se estima y muestran en pantalla los resultados del

modelo con variables ortogonales y en las líneas 22 a 27 el FIV y NC (con y sin constante).

```
1   datos = read.table("INTQRT.txt", header=T, sep = ";")
    head(datos)
3   attach(datos)

5   reg = lm(r12 ~ r3 + r6)
    summary(reg)
7
    x = cbind(r3,r6)
9   FIV(x)

11  NC(x)
    X = cbind(array(1,length(r12)),x)
13  NC(X)
    ((NC(X)-NC(x))/NC(X))*100
15
    reg.aux = lm(r6~r3)
17  e.r6 = reg.aux$residuals

19  reg.r6 = lm(r12 ~ r3 + e.r6)
    summary(reg.r6)
21
    x = cbind(r3,e.r6)
23  FIV(x)

25  NC(x)
    X = cbind(array(1,length(r12)),x)
27  NC(X)
```

2.9. Ejemplo 6.23

En este apartado se muestra el código usado en el Ejemplo 6.23. En las líneas 4 a 8 se generan los datos suponiendo que hay 25 hombres y 25 mujeres, en las líneas 10 a 15 se obtienen los FIVs y NCs y en las líneas 17 y 18 se estima el modelo y presentan los resultados. En la línea 19 se almacenan las estimaciones obtenidas por MCO para ser comparadas (línea 35) con las obtenidas (línea 32) tras perturbar las variables independientes (línea 21).

En las líneas 23 a 31 se obtienen los valores de los FIVs, NCs y estimación del modelo para el modelo perturbado.

```
1   T1 = 25
    T2 = 25
3
    consumo = round(c(rnorm(T1, 2200, sqrt(20)),
5               rnorm(T2, 1800, sqrt(20))), digits = 0)
    salario = round(c(rnorm(T1, 3000, sqrt(20)),
7               rnorm(T2, 2200, sqrt(20))), digits = 0)
    genero = c(array(1, T1), array(0, T2))
9
    x = cbind(salario, genero)
11  FIV(x)

13  NC(x)
    X = cbind(array(1,length(consumo)),x)
15  NC(X)

17  reg = lm(consumo~salario+genero)
    summary(reg)
19  beta = reg$coefficients
```

```
21          salario.p = perturb(salario, 5, 5, 0.01)

23          x.p = cbind(salario.p, genero)
            FIV(x.p)
25
            NC(x.p)
27          X.p = cbind(array(1,length(consumo)),x.p)
            NC(X.p)
29
            reg.p = lm(consumo~salario.p+genero)
31          summary(reg.p)
            beta.p = reg.p$coefficients
33
            (norm(salario-salario.p,"2")/norm(salario,"2"))*100
35          (norm(beta-beta.p,"2")/norm(beta,"2"))*100
```

3. Aumento del tamaño muestral en el modelo de regresión lineal múltiple

Dadas las observaciones, $(\mathbf{y}, \mathbf{X})$, de las variables del modelo (7.1), se plantea el modelo aumentado:

$$\mathbf{y}_A = \mathbf{X}_A \cdot \boldsymbol{\beta} + \mathbf{v}, \tag{C.1}$$

donde $\mathbf{v}$ es la perturbación aleatoria (que se presupone esférica) e:

$$\mathbf{y}_A = \begin{pmatrix} \mathbf{y} \\ \vdots \\ \mathbf{y} \end{pmatrix}_{n\cdot h \times 1}, \quad \mathbf{X}_A = \begin{pmatrix} \mathbf{X} \\ \vdots \\ \mathbf{X} \end{pmatrix}_{n\cdot h \times k},$$

es decir, $(\mathbf{y}_A, \mathbf{X}_A)$ se obtiene repitiendo h veces los datos iniciales $(\mathbf{y}, \mathbf{X})$.

Es fácil comprobar que el estimador de $\boldsymbol{\beta}$ del modelo (C.1), que denotaremos[1] $\widehat{\boldsymbol{\beta}}_A$, verifica:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_A &= \left(\mathbf{X}_A^t \mathbf{X}_A\right)^{-1} \mathbf{X}_A^t \mathbf{y}_A = \frac{1}{h} \cdot \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \cdot h \cdot \mathbf{X}^t \mathbf{y} = \widehat{\boldsymbol{\beta}}, \\
SCR_A &= \mathbf{y}_A^t \mathbf{y}_A - \widehat{\boldsymbol{\beta}}_A^t \mathbf{X}_A^t \mathbf{y}_A = h \cdot \mathbf{y}^t \mathbf{y} - \widehat{\boldsymbol{\beta}}^t \cdot h \cdot \mathbf{X}^t \mathbf{y} = h \cdot SCR, \\
SCT_A &= \mathbf{y}_A^t \mathbf{y}_A - n \cdot h \cdot \overline{\mathbf{y}}_A^2 = h \cdot \mathbf{y}^t \mathbf{y} - n \cdot h \cdot \overline{\mathbf{y}}^2 = h \cdot SCT, \\
R_A^2 &= 1 - \frac{SCR_A}{SCT_A} = R^2, \qquad\qquad (\text{C.2}) \\
\overline{R}_A^2 &= 1 - (1 - R_A^2) \cdot \frac{n \cdot h - 1}{n \cdot h - k} = 1 - (1 - \overline{R}^2) \cdot \frac{n-k}{n-1} \cdot \frac{n \cdot h - 1}{n \cdot h - k}, \\
\widehat{\sigma}_A^2 &= \frac{SCR_A}{n \cdot h - k} = h \cdot \frac{n-k}{n \cdot h - k} \cdot \widehat{\sigma}^2, \\
\widehat{var}\left(\widehat{\boldsymbol{\beta}}_A\right) &= \widehat{\sigma}_A^2 \cdot \left(\mathbf{X}_A^t \mathbf{X}_A\right)^{-1} = \frac{n-k}{n \cdot h - k} \cdot \widehat{var}\left(\widehat{\boldsymbol{\beta}}\right), \qquad\qquad (\text{C.3})
\end{aligned}$$

[1] En general, se usa el subíndice A para hacer referencia a las cantidades relacionadas con el modelo (C.1) y sin el subíndice a las del modelo (7.1).

$$t_{exp,A}\left(\beta_i\right) = \left|\frac{\widehat{\beta}_{i,A}}{\sqrt{\widehat{var(\widehat{\beta}_{i,A})}}}\right| = \left|\frac{\widehat{\beta}_i}{\sqrt{\frac{n-k}{n\cdot h-k}}\cdot\sqrt{\widehat{var(\widehat{\beta}_i)}}}\right| = \sqrt{\frac{n\cdot h-k}{n-k}}\cdot t_{exp}\left(\beta_i\right), \quad \text{(C.4)}$$

$$F_{exp,A} = \frac{\frac{R_A^2}{k-1}}{\frac{1-R_A^2}{n\cdot h-k}} = \frac{n\cdot h-k}{k-1}\cdot\frac{R^2}{1-R^2} = \frac{n\cdot h-k}{n-k}\cdot F_{exp}. \quad \text{(C.5)}$$

Como $h > 1$, se tiene que $\frac{n\cdot h-k}{n-k} > 1$ y entonces a partir de las expresiones (C.3), (C.4) y (C.5) se tiene que:

$$\widehat{var\left(\widehat{\boldsymbol{\beta}}_A\right)} < \widehat{var\left(\widehat{\boldsymbol{\beta}}\right)}, \quad t_{exp,A}\left(\beta_i\right) > t_{exp}\left(\beta_i\right), \quad F_{exp,A} > F_{exp},$$

es decir, en el modelo (C.1) se reduce con respecto al modelo (7.1) la varianza estimada de los coeficientes estimados y aumentan los estadísticos experimentales de los contrastes de significación individual y conjunta.

Por otro lado, para rechazar la hipótesis nula en los contrastes de significación individual en el modelo (C.1) se ha de verificar que[2]:

$$t_{exp,A}\left(\beta_i\right) > t_{n\cdot h-k}(1-\alpha/2),$$

lo cual es equivalente a que se verifique la condición $h_i > cota_i$ donde:

$$cota_i = \frac{1}{n}\left(\left(\frac{t_{n\cdot h-k}(1-\alpha/2)}{t_{exp}\left(\beta_i\right)}\right)^2\cdot(n-k)+k\right). \quad \text{(C.6)}$$

Es decir, se rechazará la hipótesis nula de que el coeficiente i-ésimo es nulo si se aumenta la muestra h_i veces.

Observando la expresión (C.6) se tiene que para calcular h_i es necesario conocer h para determinar el valor de $t_{n\cdot h-k}(1-\alpha/2)$. Esta contradicción se salva aproximando el valor de $t_{n\cdot h-k}(1-\alpha/2)$ por 1.96 usando la relación existente entre la normal tipificada y la t de Student. Por tanto, considerando $h = max\{h_1,\ldots,h_k\}$ se tiene asegurado que se rechaza la hipótesis nula en todos los contrastes de significación individual del modelo (C.1). En el Ejemplo 6.17 se tiene que $h = 7$ si se deja fuera la constante.

De esta forma, es claro que se mitiga uno de los síntomas de la existencia de multicolinealidad aproximada preocupante, rechazar la hipótesis nula en los contrastes de significación individual y no hacerlo en el contraste de significación conjunta.

Ahora bien, si se calculan las medidas de detección de la multicolinealidad presentadas en la sección 4 en el modelo (C.1) se tiene que:

- El cálculo del FIV se basa en el coeficiente de determinación de la regresión auxiliar que tiene como variable dependiente a la variable j-ésima de $\mathbf{X}_A$ en función del resto de variables de dicha matriz. Sin embargo, esta regresión auxiliar es la versión aumentada de la regresión auxiliar usada al calcular el FIV en el modelo (7.1). Por lo expuesto en la expresión (C.2), los coeficientes de determinación de ambas regresiones auxiliares coinciden y, por tanto, los FIVs del modelo inicial y aumentado también serán los mismos.

[2] Adviértase que en comparación con el modelo (7.1) también disminuye el valor del estadístico teórico.

- Puesto que $\mathbf{X}_A^t\mathbf{X}_A = h \cdot \mathbf{X}^t\mathbf{X}$, se tiene que los autovalores de $\mathbf{X}_A^t\mathbf{X}_A$ y $\mathbf{X}^t\mathbf{X}$ son proporcionales, es decir, si ξ es autovalor de $\mathbf{X}^t\mathbf{X}$, $h \cdot \xi$ lo es de $\mathbf{X}_A^t\mathbf{X}_A$. En tal caso:

$$K(\mathbf{X}_A) = \sqrt{\frac{h \cdot \xi_{max}}{h \cdot \xi_{min}}} = K(\mathbf{X}),$$

  es decir, el NC en ambos modelos también coincide. Adviértase que esta relación también se verifica al transformar la matriz $\mathbf{X}_A$ para que tenga longitud unidad.

Es decir, si inicialmente estas medidas indican la existencia de multicolinealidad preocupante en el modelo (7.1), también lo seguirán indicando en el modelo (C.1).