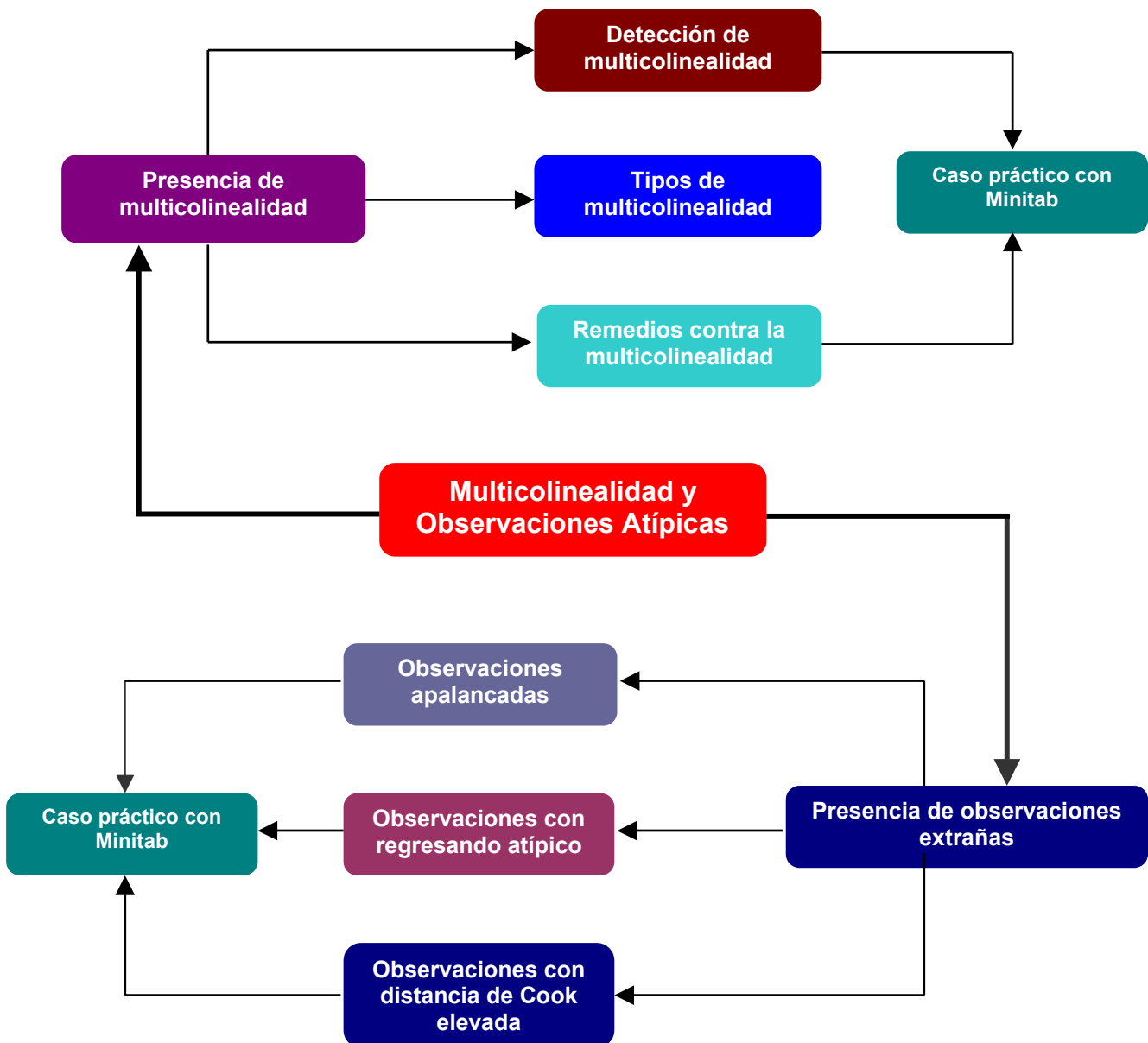


MULTICOLINEALIDAD Y OBSERVACIONES ATÍPICAS

Autores: Renatas Kizys (rkizys@uoc.edu), Ángel Alejandro Juan Pérez (ajuanp@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

En el *math-block* relativo al Análisis de Especificación se contemplan las consecuencias para la estimación MCO ante el incumplimiento de una de las hipótesis básicas del modelo de regresión lineal, como la especificación correcta de la parte determinista del modelo.

Sin embargo, los problemas del MRLM no se limitan a una incorrecta especificación del modelo, pues las deficiencias muestrales presentan otra cara de la moneda. En este tema, analizaremos problemas que pueden surgir a partir del conjunto de observaciones de las variables del modelo de regresión y que afectan a los resultados de estimación. En concreto, consideraremos la multicolinealidad y las observaciones atípicas. La multicolinealidad aparece cuando las variables explicativas de un modelo econométrico están correlacionadas entre sí, mientras que nos referimos a las observaciones atípicas cuando hay evidencia de que éstas distan sistemáticamente del resto de las observaciones. Nuestro objetivo, por tanto, es detectar y solventar los errores muestrales.

OBJETIVOS

- Conocer las consecuencias negativas asociadas a los modelos con diferentes grados de multicolinealidad.
- Saber detectar si existe multicolinealidad o no, así como cuales son las variables que la generan.
- Seleccionar la mejor alternativa (en cuanto a especificación del modelo que presenta multicolinealidad) para alcanzar los objetivos inicialmente deseados del modelo econométrico.
- Saber detectar cuando una observación es atípica, presenta apalancamiento, tiene una influencia en el ajuste mayor que el resto o es un outlier.
- Conocer las características de los distintos tipos de observaciones mencionadas en el punto anterior, así como las consecuencias que producen sobre la estimación del modelo.

CONOCIMIENTOS PREVIOS

Aparte de estar iniciado en el uso del paquete estadístico Minitab, resulta muy conveniente haber leído con profundidad los siguientes *math-blocks* relacionados con Estadística e Introducción a la Econometría:

- Análisis de regresión y correlación lineal
- Modelo de Regresión Lineal Múltiple
- Restricciones lineales

CONCEPTOS FUNDAMENTALES

□ Multicolinealidad

En este tema analizaremos problemas que pueden surgir a partir del conjunto de observaciones de las variables del modelo y que afectan a los resultados de estimación. En concreto, consideraremos la multicolinealidad y las observaciones atípicas. Nuestro objetivo es detectar y solventar los errores muestrales.

Ya hemos visto que el MRLM estándar tenía que cumplir, entre otros, la hipótesis de ausencia de multicolinealidad perfecta.

La situación de ausencia de multicolinealidad perfecta ocurre cuando no hay ninguna variable explicativa que se pueda obtener como combinación lineal del resto de las variables. Dicho de otra manera, no hay ninguna variable explicativa que presente una correlación perfecta respecto a una o varias variables explicativas.

La presencia de multicolinealidad en un modelo tiene consecuencias negativas sobre la estimación del modelo y, por consiguiente, sobre el resto del análisis econométrico.

Tipos de multicolinealidad

Así, en un modelo con k variables explicativas nos podemos encontrar con tres tipos de situaciones [1]:

Presencia de multicolinealidad perfecta. En este caso, el rango de la matriz \mathbf{X} será de orden menor que k , es decir, $\rho(\mathbf{X}) < k$, lo cual quiere decir que alguna variable explicativa puede obtenerse como combinación lineal de las demás variables explicativas.

Ausencia total de multicolinealidad. El rango de \mathbf{X} será igual a k , es decir, $\rho(\mathbf{X}) = k$, con lo cual las variables explicativas del modelo están incorrelacionadas.

Presencia de un cierto nivel de multicolinealidad. Al igual que en el caso de ausencia total de multicolinealidad, el rango de \mathbf{X} será igual a k . No obstante, existe una correlación distinta de cero entre algunas o todas las variables explicativas. Este caso suele ocurrir con más frecuencia en la práctica que los demás casos y, por otra parte, es más complejo de estudiar.

Bajo *multicolinealidad perfecta*, como ya hemos comentado, la matriz \mathbf{X} no tiene rango completo, con lo que no podemos estimar el modelo por MCO. Las consecuencias de multicolinealidad perfecta pueden describirse mediante un ejemplo.

Ejemplo 1. Multicolinealidad perfecta. Sea un modelo de regresión siguiente:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i, \forall i = 1, \dots, n.$$

Supongamos que existe relación entre las variables explicativas definida por $X_{4i} = X_{2i} - 2X_{3i}$. Sustituyendo la variable X_{4i} por su expresión en el modelo y reordenando los términos, queda:

$$Y_i = \beta_1 + (\beta_2 + \beta_4) X_{2i} + (\beta_3 - 2\beta_4) X_{3i} + u_i, \forall i = 1, \dots, n.$$

O bien, reescribiendo, obtenemos:

$$Y_i = \delta_1 + \delta_2 X_{2i} + \delta_3 X_{3i} + u_i, \forall i = 1, \dots, n,$$

que permite obtener estimaciones MCO de los parámetros Δ (que son una combinación lineal de los parámetros \mathbf{B}), pero no de los \mathbf{B} .

El caso es que *ausencia total de multicolinealidad* en la práctica casi no ocurre. La correlación nula entre los regresores implica que la estimación del vector de parámetros poblacionales es la misma tanto si estimamos el modelo de regresión lineal múltiple, como el modelo de regresión lineal simple.

Consecuencias

Las consecuencias de *un cierto nivel de multicolinealidad* pueden ser las siguientes [1]:

- Cuanto más grande sea la correlación, más próximo a cero será el determinante de la matriz $\mathbf{X}'\mathbf{X}$, lo cual incrementará las varianzas y covarianzas del vector de los parámetros estimados.

- Las elevadas varianzas hacen que los parámetros estimados sean imprecisos e inestables. En efecto, cuanto mayor es la varianza, menor será el estadístico de contraste t , lo cual a menudo nos llevará a la conclusión que una variable explicativa es irrelevante, cuando en realidad no lo es.
- Un cierto nivel de multicolinealidad, sin embargo, no afecta el estadístico de contraste de significación global, F . La idea es que el estadístico F considera toda la explicación de la variabilidad de la variable endógena, mientras que las variables explicativas comparten una parte de la variabilidad con las demás variables implicadas.

Técnicas de detección de multicolinealidad

Una vez contempladas las consecuencias de la presencia de multicolinealidad, pasaremos a estudiar, en cada caso, si hay multicolinealidad y, en tal caso, determinar en qué grado se presenta. A tales efectos, nos interesa disponer de instrumentos que nos indiquen la intensidad con que se presenta, para saber hasta que punto afecta los resultados del modelo.

No obstante, econometría no dispone de contrastes contruidos expresamente para detectar la multicolinealidad, puesto que el problema descansa más bien en la muestra que en la población. De modo que la literatura propone métodos algo menos formales que a continuación pasaremos a contemplar.

- Analizar los coeficientes de correlación simple entre los regresores de dos en dos. Este método consiste en detectar la presencia de multicolinealidad.
- Analizar el determinante de la matriz de correlaciones, R_x . Si hay multicolinealidad perfecta, el valor del determinante será 0 , mientras que en ausencia total de multicolinealidad será igual a 1 . En comparación con el método anterior, este es el preferible, puesto que tiene en cuenta la correlación que se produce entre cualquier número de variables explicativas conjuntamente, mientras que el anterior solamente presenta la correlación de dos en dos.
- Estudiar los coeficientes de determinación de las regresiones en las cuales figura como variable endógena, sucesivamente, cada una de las variables explicativas del modelo. Un coeficiente de determinación elevado implica la presencia de multicolinealidad.
- El problema multicolinealidad también puede detectarse estudiando los elementos del vector de los parámetros estimados. Si las estimaciones se ven significativamente alteradas al efectuar pequeños cambios en los datos, entonces hay indicios de multicolinealidad en el modelo.
- Si las estimaciones obtenidas contradicen a la teoría económica, entonces puede que ello sea a causa de las elevadas varianzas en la presencia de multicolinealidad.
- Adicionalmente, podemos analizar los coeficientes de determinación de sucesivos MRLM en que se elimina un regresor. Si al eliminar una variables explicativa del modelo, R^2 no cambia sustancialmente, entonces, bien esta variable es poco relevante para explicar la variabilidad de la variable endógena, o bien, lo que explica la variable eliminada ya queda explicado por otros regresores, en cuyo caso hay indicios de multicolinealidad en el modelo.
- El último método consiste en calcular el factor de incremento de la varianza (**FIV**) de cada una de las variables explicativas. Se calcula de la siguiente manera:

$FIV_j = \frac{Var(\hat{\beta}_j)}{Var(\beta_j^*)} = 1/(1 - R_j^2)$, donde $Var(\beta_j^*) = \frac{\sigma^2}{\sum_{i=1}^N (X_{ji} - \bar{X}_j)}$ es la varianza óptima en el caso de

ausencia de correlación entre los regresores, $Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^N (X_{ji} - \bar{X}_j)(1 - R_j^2)}$ es la varianza de

un estimador cualquiera y R_j^2 el coeficiente de determinación de la regresión entre X_j y el resto de las variables explicativas del modelo inicial. Valores del $FIV_j > 5$ están asociados a $R_j^2 > 0,8$ en cuyo caso se puede considerar que las consecuencias sobre el MRLM ya pueden ser relevantes.

Remedios contra la multicolinealidad

Los métodos remediales para multicolinealidad dependen de cómo se utilice posteriormente el modelo estimado y no siempre consisten en eliminarla por completo. Si se trata de realizar predicciones en el caso de multicolinealidad, cabe esperar que la multicolinealidad detectada en el período muestral, también se dé en el período de predicción. En cambio, si queremos realizar un análisis de cambio estructural, entonces resulta vital eliminar el problema por completo.

Una vez identificado un problema de multicolinealidad en el modelo, hay varias formas para tratar de solventarlo. Las técnicas más utilizadas para solucionar el problema de multicolinealidad son las siguientes:

1. Incorporación de nueva información [1]. Sin embargo, hemos de discernir entre
 - a) el aumento del tamaño muestral de modo que se reduzca el problema de correlación entre las variables explicativas; y
 - b) utilización de otra clase de información extramuestral que se lleva a cabo mediante restricciones sobre los parámetros del modelo inicial.
2. Reespecificación del modelo que puede llevarse a cabo mediante [1]:
 - a) eliminación algunas variables explicativas de la regresión (especialmente si esto afecta poco a la bondad del ajuste);
 - b) transformación de las variables del modelo; por ejemplo, en el caso de modelos cuadráticos, extrayendo la media del regresor antes de considerar su cuadrado.
3. Estimación Ridge [1]. Este método se basa en la premisa de que el valor reducido del determinante de la matriz $X'X$ puede causar problemas a la hora de aplicar el método de mínimos cuadrados ordinarios. Así pues, solventar este tipo de problemas se procede a la estimación Ridge que consiste en sumar una determinada cantidad a los elementos de la diagonal principal de $X'X$. El estimador Ridge es

$$B_{\text{Ridge}} = [X'X + cI_k]^{-1} \cdot X'Y,$$

siendo c una constante arbitraria positiva. El inconveniente es que el estimador Ridge presenta un sesgo que aumenta con el valor de la constante c . Sin embargo, ante una elección entre el estimador MCO y el estimador Ridge, optamos por el último en el caso si el error cuadrático medio del mismo es menor que el de MCO.

□ Presencia de valores extraños

Las técnicas de solucionar la multicolinealidad no siempre acaban con todos los errores muestrales. En ocasiones, es útil ver con más detalle de cómo se ha generado la muestra que ya tenemos. Un análisis particularizado, sobre todo cuando disponemos de una muestra pequeña, nos permite detectar observaciones que han sido generadas por un proceso distinto. Las observaciones atípicas hacen que la recta de regresión tienda a desplazarse en su dirección, o bien cambie de pendiente, así causando alteraciones inesperadas en los resultados de estimación en comparación de lo que predice la teoría económica.

Las observaciones atípicas, extrañas o influyentes pueden tener formas diferentes; distinguimos entre las siguientes clases de observaciones anómalas [1]:

- a) Atípicas con respecto al eje de abscisas.
- b) Atípicas en relación de eje de ordenadas;
- c) Atípicas respecto tanto a las abscisas como a las ordenadas.

A fin de determinar si una observación es atípica en relación de las variables explicativas, estudiamos el grado de apalancamiento (leverage) de esa observación. La intuición nos sugiere que las observaciones atípicas o extrañas, siendo alejadas del resto de las observaciones, pueden también presentar un cierto grado de apalancamiento.

Apalancamiento (leverage)

Una observación presenta apalancamiento si está muy alejada del resto de observaciones. Hay un leverage (de aquí a delante lo llamaremos el lever y denotaremos h_{ii}) asociado a cada observación y representa el elemento i -ésimo de la diagonal principal de la matriz \mathbf{H} cuya expresión es la siguiente:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

El lever tiene las siguientes propiedades:

- El lever de una observación será tanto más grande cuanto sea diferente, en términos de las variables explicativas, del resto de observaciones.
- El lever está acotado inferiormente por $1/n$ y superiormente por 1 : $1/n \leq h_{ii} \leq 1$.

Cuando lever coincide con la cota inferior, apalancamiento no se presenta. En este caso, la observación coincide con la media de las variables explicativas, por lo que la distancia entre dicha observación y la media de \mathbf{X} es cero. Si, por el contrario, el lever alcanza la cota superior, entonces la observación presenta el máximo apalancamiento posible. El caso más habitual es cuando el lever cae dentro del intervalo $[1/n, 1]$.

En concreto, diremos que una observación presenta apalancamiento si su lever es más grande que dos veces la media de los levers, $h_{ii} = 2 \cdot \bar{h}$.

Por otro lado, para poder detectar si una observación es atípica con respecto a la variable endógena, estudiamos *el residuo* correspondiente a esta observación, así como *el residuo estandarizado*, *residuo estudentizado* o *residuo estudentizado con omisión*.

Residuo, residuo estandarizado, residuo estudentizado y residuo estudentizado con omisión

En primer lugar, podemos examinar los residuos de una estimación inicial. Podemos pensar que si una observación es influyente, generará un residuo de alto valor absoluto. No obstante, si pretendemos comparar los residuos entre sí, hemos de normalizarlos adecuadamente antes de proceder a la comparación. Consideremos el residuo estandarizado:

$$z_i = \frac{e_i}{\sigma} = \frac{e_i}{\sqrt{\frac{e'e}{n-k}}}$$

En la práctica se utiliza la siguiente regla:

- Si $|z_i| \geq 2 \Rightarrow$ la observación i -ésima puede considerarse un outlier.
- Si $|z_i| < 2 \Rightarrow$ la observación i -ésima no puede considerarse un outlier.

Para evitar interferencias entre las observaciones, se utiliza *el residuo estudentizado* que denotaremos por r_i . El residuo estudentizado, a diferencia del residuo estandarizado, pondera el error de ajuste MCO asociado a la observación i -ésima por su desviación estándar. Para ello, recordemos que si se denota el vector de residuos MCO se tiene:

$$\text{Var}(e) = \sigma^2 M = \sigma^2 (I_n - X(X'X)^{-1}X') = \sigma^2 (I_n - H).$$

Por lo que cada residuo, normalizado por su desviación típica, sería:

$$r_i = \frac{e_i}{\sqrt{\sigma^2 \cdot (1 - h_{ii})}}$$

El residuo estudentizado se distribuye asintóticamente con una distribución **t de Student** con $n - k$ grados de libertad. Por tanto, utilizaremos la siguiente regla:

- Si $|r_i| \geq t_{n-k; \alpha/2} \Rightarrow$ la observación i -ésima puede considerarse un outlier.
- Si $|r_i| < t_{n-k; \alpha/2} \Rightarrow$ la observación i -ésima no puede considerarse un outlier.

Habitualmente se utiliza un método, denominado *residuo estudentizado con omisión*. El método descansa en un procedimiento más complejo, que consiste en estimar por MCO omitiendo la observación i -ésima:

$$B(i) = [X(i)'X(i)]^{-1}X(i)'Y(i),$$

donde (i) denota el índice de la observación omitida. A partir de la estimación MCO omitiendo la i -ésima observación, generamos un vector de residuos $e(i)$:

$$e_i(i) = Y_i - X_i B(i) = u_i + X_i [B - B(i)].$$

Como la observación i -ésima no se ha utilizado para obtener $B(i)$, ambos sumandos son independientes, y se tiene:

$$\text{Var}(e_i(i)) = \sigma^2 [1 + X_i (X(i)'X(i))^{-1} X_i']$$

El residuo con omisión se estudentiza de la siguiente manera:

$$r_i^* = \frac{e_i(i)}{\sqrt{\sigma^2 \cdot (1 - h_{ii}(i))}}$$

Para ver si es outlier, aplicamos la siguiente regla:

- Si $|r_i^*| \geq t_{n-k-1; \alpha/2} \Rightarrow$ la observación i -ésima puede considerarse un outlier.
- Si $|r_i^*| < t_{n-k-1; \alpha/2} \Rightarrow$ la observación i -ésima no puede considerarse un outlier.

Una vez contempladas las medidas de anomalía de las observaciones con respecto al eje de abscisas y al eje de ordenadas por separado, consideremos **la Distancia de Cook**, una medida compleja que combina las dos clases de medidas anteriores.

La Distancia de Cook

La distancia de Cook es una medida que permite detectar las observaciones atípicas en combinación del apalancamiento y la concordancia del proceso generador de dichas observaciones con el proceso generador del resto de observaciones. La distancia de Cook, para la observación i -ésima, se calcula de la siguiente manera:

$$d_i = r_i^2 \frac{h_{ii}}{k \cdot (1 - h_{ii})}$$

Recordemos que el lever indica si una determinada observación tiene regresores atípicos, mientras que el residuo estudentizado nos dice si una observación tiene un regresando atípico.

La diferencia de Cook se distribuye con una distribución **F de Snedecor** con k grados libertad en el numerador y $n - k$, en el denominador. Así pues, diremos que la observación i -ésima tiene una influencia más grande sobre el ajuste del modelo de regresión que el resto si el valor del estadístico de contraste supera el valor crítico de las tablas. La regla puede resumirse de la manera siguiente:

- Si $d_i \geq F_{k, n-k; \alpha} \Rightarrow$ la observación i -ésima puede considerarse atípica en relación a la variable endógena y las variables explicativas conjuntamente.
- Si $d_i < F_{k, n-k; \alpha} \Rightarrow$ la observación i -ésima no puede considerarse atípica en relación a la variable endógena y las variables explicativas conjuntamente.

CASOS PRÁCTICOS CON SOFTWARE

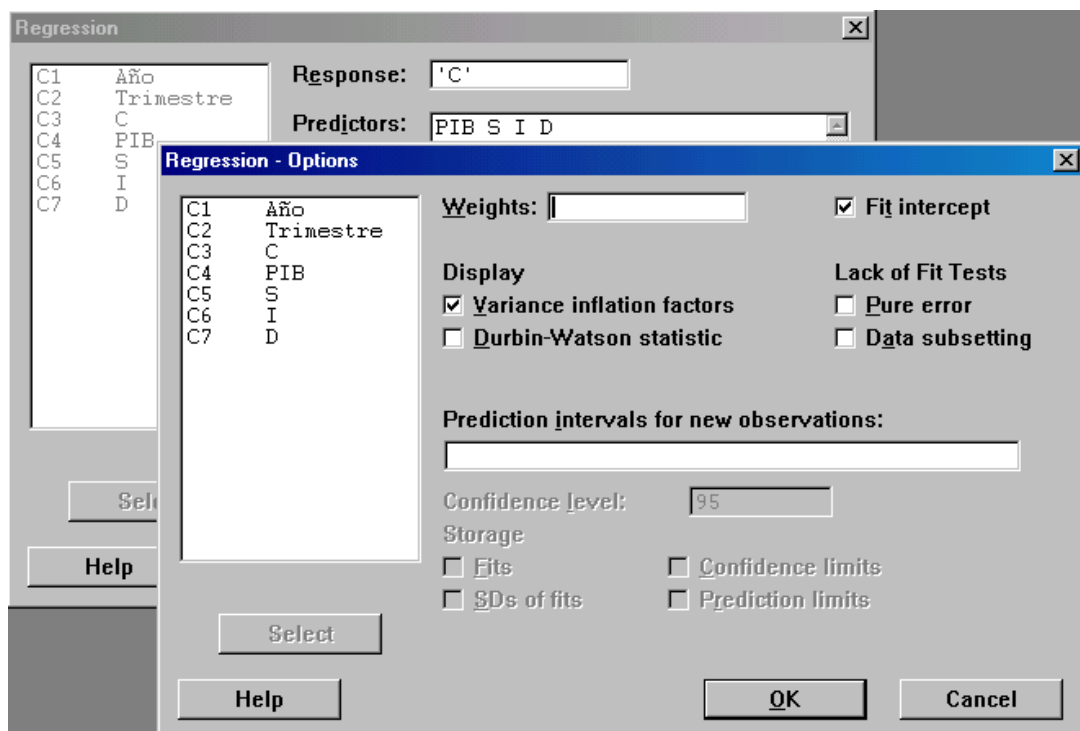
□ Multicolinealidad

Ejemplo 1. Se desea estudiar la dinámica del consumo en los hogares de España (**C**), en relación del Producto Interior Bruto (**PIB**), la remuneración de asalariados (**S**), el tipo de interés legal de dinero establecido por Banco de España, **I** y la tasa de desempleo, **D**. Disponemos de 80 datos trimestrales, correspondientes a 1980:1T – 1999:4T e expresados

en billones de pesetas constantes a precios de mercado (excepto el tipo de interés y la tasa de desempleo que vienen expresados en tanto por ciento). Los datos se encuentran en el fichero **Consumo.mtw**. Suponiendo, que el modelo satisface las hipótesis del modelo de regresión lineal, construiremos el siguiente modelo para explicar el consumo de los hogares españoles:

$$C_t = \beta_1 + \beta_2 \text{PIB}_t + \beta_3 \text{S}_t + \beta_4 \text{I}_t + \beta_5 \text{D}_t + u_t; t = 1, \dots, T.$$

A priori, parece razonar esperar que las variables PIB y S estén correlacionados. Intuitivamente, la remuneración de los asalariados (S) forma parte del producto interior bruto (PIB). **Minitab** dispone de una opción, llamada **Variance Inflation Factors (VIF)**, la cual nos permite identificar la multicolinealidad entre los las variables explicativos del modelo. Para hallar los mencionados diagnósticos, seleccionamos: **> Stat > Regresión > Regresión... > Options > Variance inflation factors**:

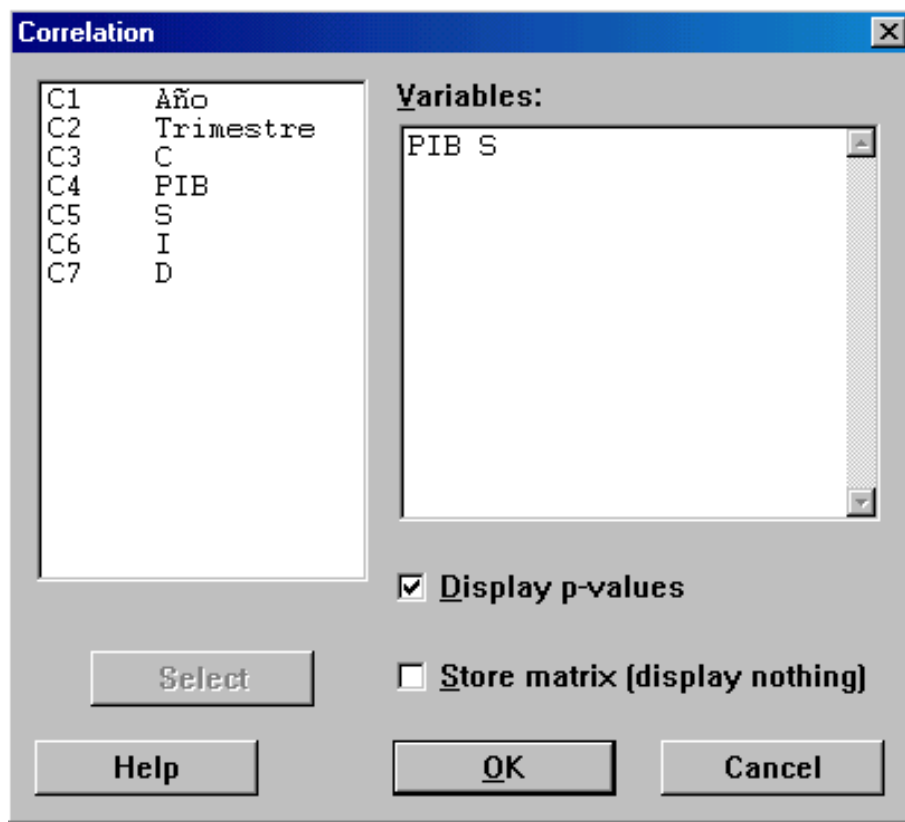


Los resultados se presentan en el cuadro siguiente:

Regression Analysis						
The regression equation is						
C = 0,712 + 0,410 PIB + 0,305 S + 0,0104 I - 0,0146 D						
Predictor	Coef	StDev	T	P	VIF	
Constant	0,7116	0,1321	5,39	0,000		
PIB	0,41020	0,01929	21,26	0,000	31,3	
S	0,30464	0,03568	8,54	0,000	30,1	
I	0,010417	0,003423	3,04	0,003	2,1	
D	-0,014624	0,003133	-4,67	0,000	1,3	
S = 0,07934 R-Sq = 99,7% R-Sq(adj) = 99,7%						

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	4	162,475	40,619	6453,19	0,000
Residual Error	75	0,472	0,006		
Total	79	162,947			

El análisis de los factores de inflación (o de incremento) de la varianza (**VIF**) nos sirve para confirmar de que los parámetros asociados a las variables **PIB** y **S** están afectados por un alto grado de multicolinealidad. En efecto, $VIF_{PIB} = 31,3$ y $VIF_S = 30,1$, con lo cual son muy superiores al valor 5. Por este motivo, es muy probable que este modelo presente dificultades estadísticas y computacionales. En el análisis de multicolinealidad es conveniente calcular el coeficiente de correlación entre los predictores en cuestión. A tales efectos, seleccionamos la opción **> Stat > Basic Statistics > Correlation**:



A continuación presentamos el siguiente valor para el coeficiente de correlación:

Correlations (Pearson)	
Correlation of PIB and S	= 0,980; P-Value = 0,000

Una vez más, los diagnósticos nos sugieren una multicolinealidad intensa, lo cual era de esperar. La variable **PIB** comparte con la variable **S** una parte de la variabilidad de la variable endógena; los cierto es que dicha variabilidad podría ser perfectamente explicada por cualquiera de las dos variables. Dado que el estadístico t asociado al **PIB** resulta superior al

de **S**, parece razonable eliminar **S**. Estimamos, pues el modelo de regresión lineal múltiple sin **S**, obteniendo los siguientes resultados:

Regression Analysis					
The regression equation is					
C = 0,811 + 0,569 PIB + 0,00752 I - 0,0246 D					
Predictor	Coef	StDev	T	P	VIF
Constant	0,8107	0,1836	4,42	0,000	
PIB	0,569072	0,007088	80,28	0,000	2,2
I	0,007515	0,004751	1,58	0,118	2,1
D	-0,024610	0,004054	-6,07	0,000	1,1
S = 0,1107 R-Sq = 99,4% R-Sq(adj) = 99,4%					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	3	162,017	54,006	4409,40	0,000
Residual Error	76	0,931	0,012		
Total	79	162,947			

Observamos que el coeficiente de determinación, al eliminar un regresor, no ha cambiado sustancialmente. Podemos concluir, por tanto, que lo que explicaba la variable eliminada ya quedaba explicado por otros regresores, y en concreto, por el Producto Interior Bruto.

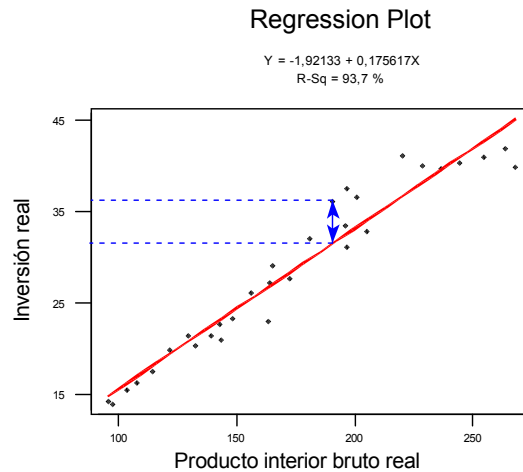
□ Observaciones atípicas

Ejemplo 2. Consideremos los siguientes datos anuales correspondientes al período 1960-1990 de la economía de los Estados Unidos:

Observación	Año	Y (inversión real)	X (PIB real)
1	1960	14,2226	95,065
2	1961	13,9336	97,281
3	1962	15,5040	103,159
4	1963	16,3105	107,607
5	1964	17,4936	113,860
6	1965	19,8906	121,153
7	1966	21,4803	129,102
8	1967	20,4046	132,340
9	1968	21,4776	138,663
10	1969	22,6821	142,856
11	1970	20,9722	143,120
12	1971	23,3538	147,928
13	1972	26,1040	155,955
14	1973	29,1101	164,946
15	1974	27,2418	163,921
16	1975	23,0096	163,426
17	1976	27,6116	172,485
18	1977	32,1111	180,519
19	1978	36,1788	190,509
20	1979	37,5671	196,497
21	1980	33,5069	196,024
22	1981	36,6088	200,832

23	1982	31,1554	196,769
24	1983	32,7752	205,341
25	1984	41,1886	220,230
26	1985	39,9715	228,703
27	1986	39,6866	236,500
28	1987	40,2991	244,560
29	1988	40,9538	254,771
30	1989	41,9323	263,683
31	1990	39,8393	268,304

En el math-block MRLM vimos que ajuste MCO del modelo de regresión lineal simple puede representarse de la siguiente forma:



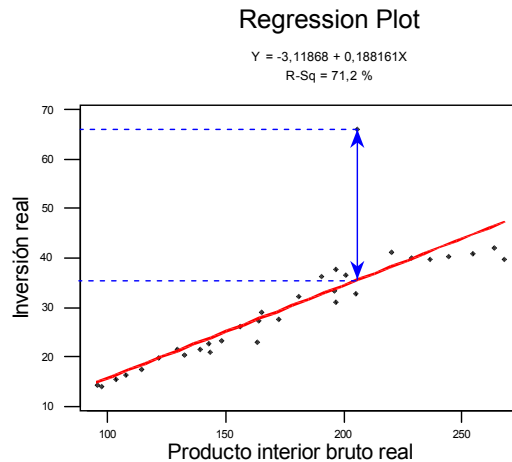
Los resultados de estimación de la función de regresión simple aparecen en el siguiente cuadro:

Regression Analysis							
The regression equation is							
Rinv = - 1,92 + 0,176 Rpib							
Predictor	Coef	StDev	T	P			
Constant	-1,921	1,525	-1,26	0,218			
Rpib	0,175617	0,008452	20,78	0,000			
S = 2,348	R-Sq = 93,7%	R-Sq(adj) = 93,5%					
Analysis of Variance							
Source	DF	SS	MS	F	P		
Regression	1	2380,1	2380,1	431,76	0,000		
Residual Error	29	159,9	5,5				
Total	30	2539,9					
Unusual Observations							
Obs	Rpib	Rinv	Fit	StDev Fit	Residual	St Resid	
19	191	36,179	31,535	0,446	4,643	2,01R	
20	196	37,567	32,587	0,465	4,980	2,16R	
31	268	39,839	45,198	0,906	-5,358	-2,47R	
R denotes an observation with a large standardized residual							

La salida de la estimación nos indica que hay tres observaciones, **20** y **31** que presentan un comportamiento atípico, pues tienen un residuo estudentizado mayor que $t_{29;0,025} = 2,0452$. Estaremos, además, interesados en replicar una observación que sea claramente *atípica con respecto del eje de ordenadas*, pero no respecto a las abscisas. A fin de representar tal situación gráficamente, añadimos a la muestra una observación adicional, correspondiente al año 1991, de características respectivas:

Observación	Año	Y (inversión real)	X (PIB real)
32	1991	65,84	205,35

Al trazar la recta de regresión sobre la nube de puntos, obtenemos el siguiente gráfico:



Los resultados de estimación se presentan en el siguiente cuadro:

Regression Analysis							
The regression equation is							
Rinv = - 3,12 + 0,188 Rpib							
Predictor	Coef	StDev	T	P			
Constant	-3,119	3,964	-0,79	0,438			
Rpib	0,18816	0,02186	8,61	0,000			
S = 6,113	R-Sq = 71,2%	R-Sq(adj) = 70,2%					
Analysis of Variance							
Source	DF	SS	MS	F	P		
Regression	1	2767,2	2767,2	74,06	0,000		
Residual Error	30	1120,9	37,4				
Total	31	3888,1					
Unusual Observations							
Obs	Rpib	Rinv	Fit	StDev Fit	Residual	St Resid	
32	205	65,84	35,52	1,27	30,32	5,07R	
R denotes an observation with a large standardized residual							

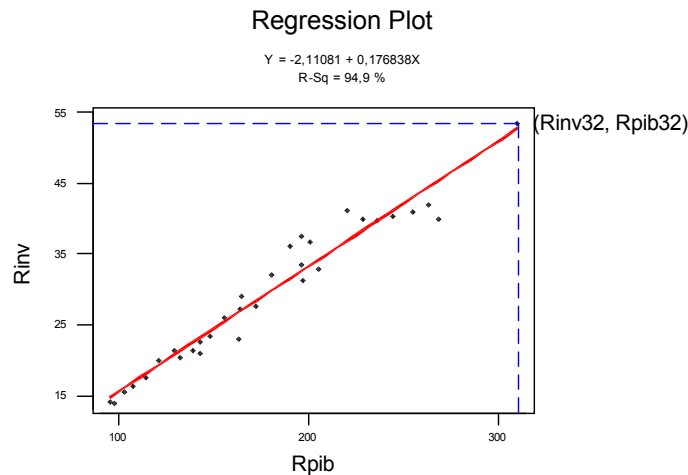
Tal y como muestra los resultados del ajuste de regresión, la observación adicional correspondiente al año 1991 puede caracterizarse como extraña respecto al eje de

ordenadas y no respecto al de abscisas. De hecho, el residuo estudentizado relativo a la observación adicional, r_{32} es de **5,07**, superando sobradamente al valor del estadístico $t_{30;0,025} = 2,0423$. Se aprecia una enorme distancia vertical entre la observación en cuestión y la recta de regresión. Comparando con el gráfico de regresión anterior, la observación adicional ha arrastrado hacia arriba la pendiente de la recta de regresión y, por otro lado, ha hecho disminuir el ajuste global del modelo. Podemos, además, decir que el dato atípico ha incrementado la suma cuadrática de errores.

Ahora procedemos a analizar las consecuencias que puede tener una observación *atípica con respecto a las abscisas*, pero no atípica respecto a las ordenadas, sobre la estimación de la recta de regresión. Para replicar tal situación, añadimos una observación adicional, de características siguientes:

Observación	Año	Y (inversión real)	X (PIB real)
32	1991	53,457	310,34

Al ajustar la recta de regresión sobre la nube de puntos, obtenemos la siguiente representación gráfica:



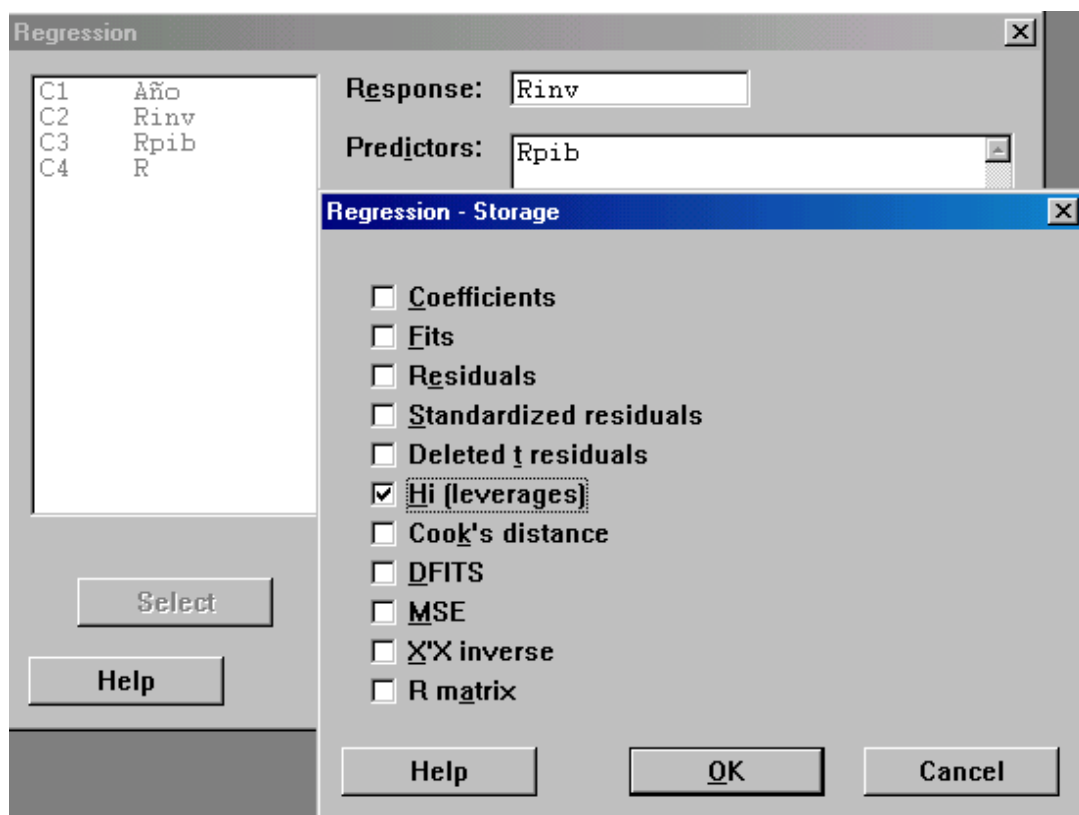
Los resultados de estimación correspondientes pueden resumirse en el siguiente cuadro:

Regression Analysis							
The regression equation is							
Rinv = - 2,11 + 0,177 Rpib							
Predictor	Coef	StDev	T	P			
Constant	-2,111	1,392	-1,52	0,140			
Rpib	0,176838	0,007490	23,61	0,000			
S = 2,313	R-Sq = 94,9%	R-Sq (adj) = 94,7%					
Analysis of Variance							
Source	DF	SS	MS	F	P		
Regression	1	2981,2	2981,2	557,35	0,000		
Residual Error	30	160,5	5,3				
Total	31	3141,6					
Unusual Observations							
Obs	Rpib	Rinv	Fit	StDev Fit	Residual	St Resid	

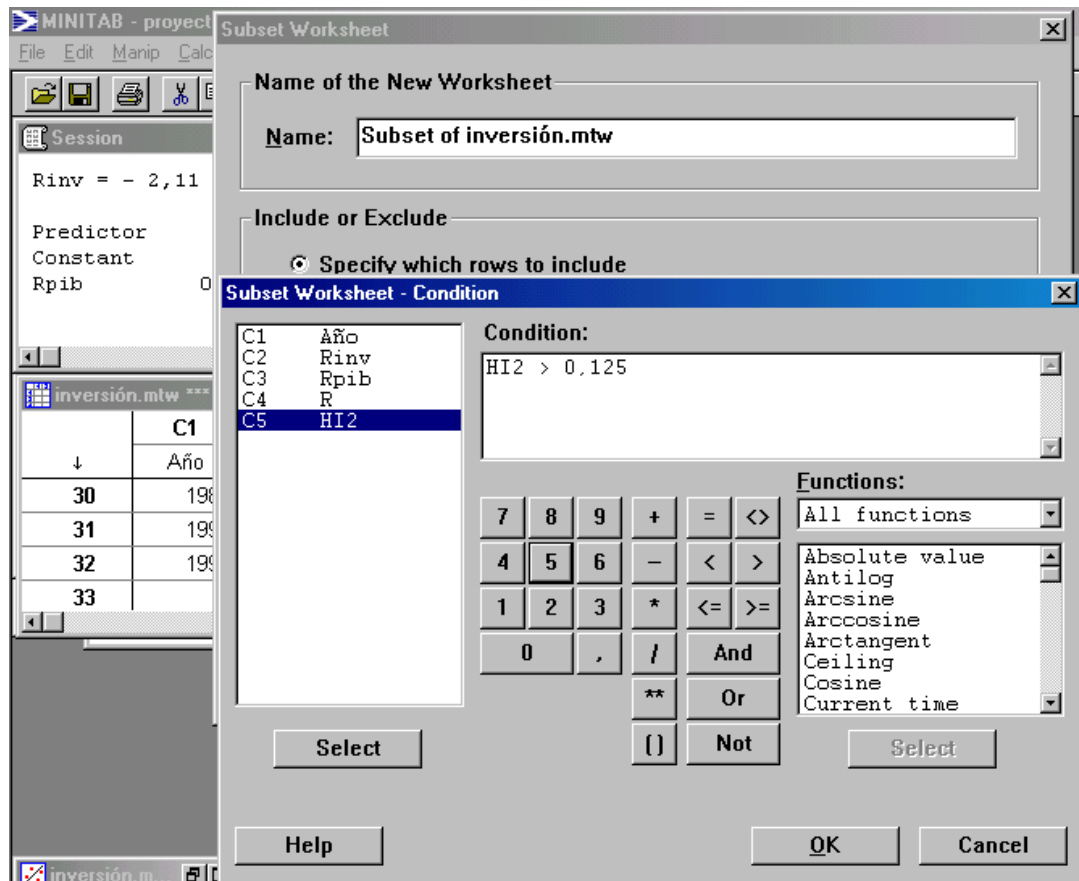
19	191	36,179	31,578	0,420	4,600	2,02R
20	196	37,567	32,637	0,432	4,930	2,17R
31	268	39,839	45,336	0,792	-5,496	-2,53R
32	310	53,457	52,769	1,074	0,688	0,34 X

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

Dentro de la tabla de estimación, la última observación, marcada con una **X**, identifica el período caracterizado por un valor de apalancamiento, h_{ii} , mayor que dos veces el número de parámetros dividido por el tamaño muestral, i.e.: $2 \cdot 2 / 32 = 0,125$. Dicha observación presenta el valor atípico de las variables explicativas. En el entorno del Minitab, los valores del apalancamiento pueden visualizarse simultáneamente estimando el modelo, vía **> Stat > Regresión > Regresión > Storage > Hi (leverages)**:



Esta opción genera automáticamente, en la misma hoja de cálculo, una columna de levers. Para detectar aquellos valores extremos que pueden influir notablemente sobre el modelo (i.e., que tienen un apalancamiento elevado), utilizamos la opción **> Manip > Subset Worksheet...**:



El Minitab selecciona las siguientes observaciones:

Subset of inversión.mtw ***							
	C1	C2	C3	C4	C5	C6	C7
↓	Año	Rinv	Rpib	HI2			
1	1991	53,457	310,34	0,215796			
2							
3							
4							
5							
6							
7							
8							

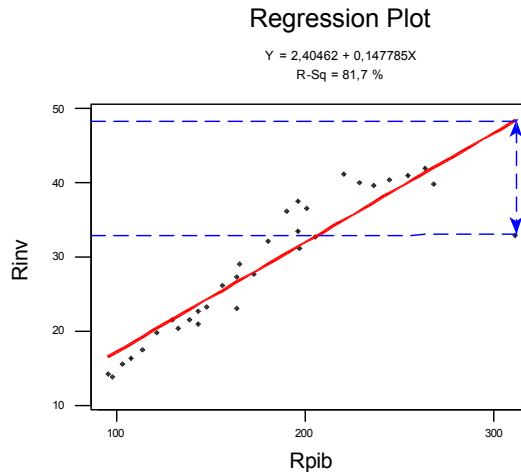
Así pues, la única observación apalancada es la que hemos simulado.

Tal y como se muestra en el análisis anterior, la observación adicional es, en efecto, atípica en relación a un conjunto de las variables explicativas. Sin embargo, al añadir dicha observación ha aumentado notablemente la suma cuadrática de regresión y, por consiguiente, ha mejorado la bondad del ajuste del modelo. Sin embargo, si comparamos los parámetros estimados con la observación adicional y sin ella, pues los cambios no han sido sustanciales al respecto. De hecho, mientras que la pendiente del modelo haya permanecido constante e estadísticamente significativa, el intercepto ha sufrido una ligera reducción.

Por último, consideraremos una observación *atípica o extraña respecto tanto a las abscisas como a las ordenadas*. A tales efectos, generamos una observación adicional de la siguiente manera:

Observación	Año	Y (inversión real)	X (PIB real)
32	1991	15,690	281,42

La recta de regresión viene representada en el siguiente gráfico:

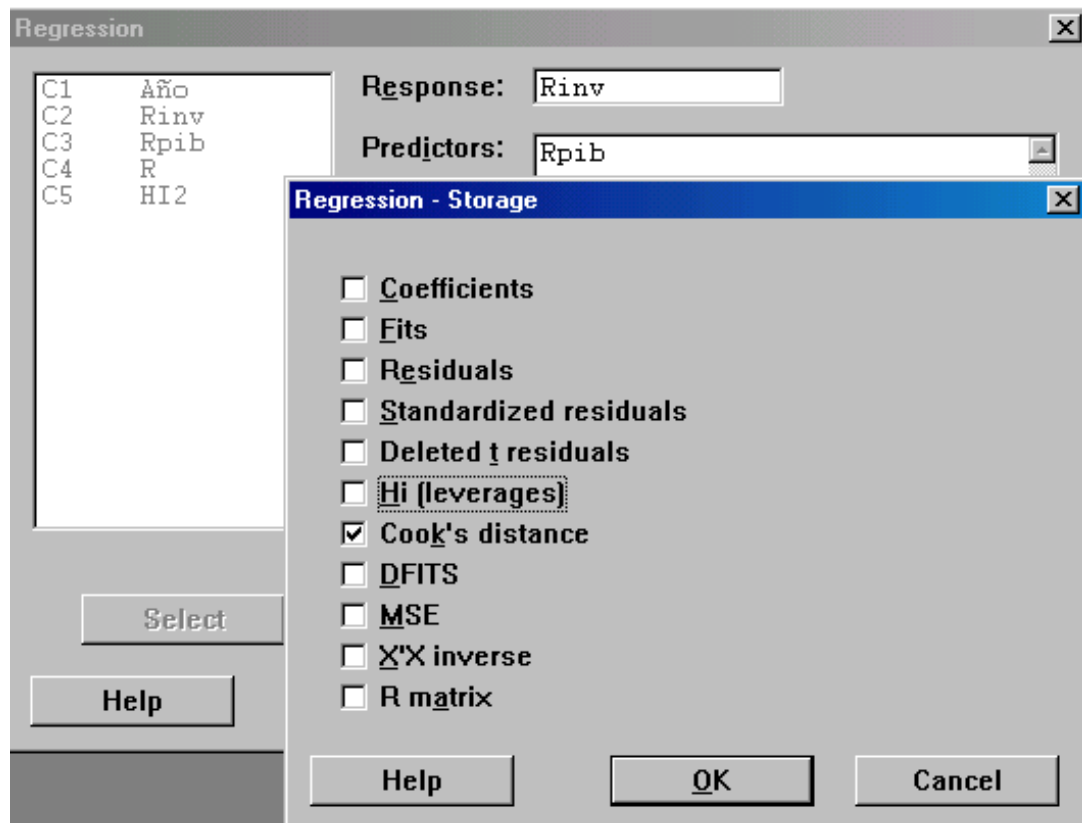


Asimismo, presentamos los resultados de estimación:

Regression Analysis							
The regression equation is							
Rinv = 2,40 + 0,148 Rpib							
Predictor	Coef	StDev	T	P			
Constant	2,405	2,375	1,01	0,319			
Rpib	0,14779	0,01277	11,57	0,000			
S = 3,951		R-Sq = 81,7%		R-Sq(adj) = 81,1%			
Analysis of Variance							
Source	DF	SS	MS	F	P		
Regression	1	2090,5	2090,5	133,89	0,000		
Residual Error	30	468,4	15,6				
Total	31	2558,9					
Unusual Observations							
Obs	Rpib	Rinv	Fit	StDev Fit	Residual	St Resid	
32	312	32,958	48,481	1,849	-15,523	-4,45RX	
R denotes an observation with a large standardized residual							
X denotes an observation whose X value gives it large influence.							

Sin embargo, el cuadro de estimación no conlleva ninguna información acerca de los valores de **la Distancia de Cook**. Recordemos que distancia de Cook combina residuos estandarizados y levers en una medida común, para determinar el grado de anomalía de la variable endógena y las variables explicativas conjuntamente. En concreto, estaremos

interesados en observaciones que tengan asociada una Distancia de Cook superior al valor $F_{k;n-k;\alpha}$, siendo $F_{k;n-k;\alpha}$ aquel valor que, en una distribución F con k grados de libertad en el numerador y n-k grados de libertad en el denominador, deje a su derecha un área determinado por un nivel de significación α . Aquí, k es el número de predictores del modelo, y n es el tamaño muestral. El Minitab permite visualizar los valores del apalancamiento mediante **> Stat > Regresión > Regresión > Storage > Cook's distance**:



Una vez generados los valores de la Distancia de Cook en la hoja de cálculo, procedemos a detectar aquellos valores que superen el valor $F_{k;n-k;\alpha}$. En este caso, $k = 2$ y $n = 32$ y sea $\alpha = 0,05$. Por tanto, estaremos interesados en aquellas observaciones cuya distancia de Cook sea mayor a $F_{2;30;0,05} = 3,3158$. Sin embargo, no hay ninguna observación que sea mayor al mencionado estadístico F para $\alpha = 0,05$. De hecho, el valor máximo de la distancia de Cook es de 2,3177 que corresponde a la observación 32. Sin embargo, en la literatura se han propuesto otros criterios menos restrictivos, como la mediana de la distribución $F_{k;n-k}$ que equivale al nivel de significación $\alpha = 0,5$. Si adoptamos este criterio, obtendremos $F_{2;30;0,5} = 0,7094$, cifra inferior a $d_{32} = 2,3177$. Así pues, sobre la base de este criterio, podemos concluir que la observación adicional presenta una distancia de Cook elevada y, consecuentemente tiene influencia pronunciada sobre la variable endógena y sobre las variables explicativas conjuntamente.

A partir de análisis anterior, apreciamos un cambio importante en los parámetros estimados, comparándolos con los obtenidos con las 31 observaciones. Por un lado, la pendiente de la recta de regresión en la última estimación ha tomado un valor más pequeño que en el caso inicial. Eso quiere decir que la observación atípica ha hecho disminuir la significación económica del producto interior bruto real. Por otro lado, el intercepto de la regresión que ahora es positivo, en lugar de ser negativo, como se ha visto en los casos anteriores. Además, el estadístico t nos indica que el PIB real ha sufrido una reducción también en la significación estadística. Por último, tanto la suma cuadrática de los errores, como la varianza estimada del término de error se ven incrementadas notablemente, mientras que los estadísticos R^2 y F nos indican una disminución sustancial en el ajuste del modelo.

BIBLIOGRAFÍA

- [1] Artís, M.; Suriñach, J.; et al (2001): "Introducción a la Econometría". Ed. Fundació per a la Universitat Oberta de Catalunya. Barcelona.
- [2] Doran, H. (1989): "Applied Regression Analysis in Econometrics". Ed. Marcel Dekker, Inc. ISBN: 0-8247-8049-3
- [3] Green, W. H. (1999): "Análisis Económico". Prentice Hall Iberia. Madrid. ISBN: 84-8322-007-5
- [4] Gujarati, D. (1997): "Econometría básica". McGraw-Hill. ISBN 958-600-585-2
- [5] Johnston, J. (2001): "Métodos de econometría". Ed. Vicens Vives. Barcelona. ISBN 84-316-6116-X
- [6] Kennedy, P. (1998): "A Guide to Econometrics". Ed. MIT Press. ISBN: 0262611406
- [7] Novalés, A. (1993): "Econometría". McGraw-Hill. ISBN 84-481-0128-6
- [8] Pulido, A. (2001): "Modelos econométricos". Ed. Pirámide. Madrid. ISBN 84-368-1534-3
- [9] Uriel, E. (1990): "Econometría: el modelo lineal". Ed. AC. Madrid. ISBN 84-7288-150-4
- [10] Wooldridge, J. (2001): "Introducción a la conometría: un enfoque moderno". Ed. Thomson Learning. ISBN: 970-686-054-1

ENLACES

- ❑ <http://www.feweb.vu.nl/econometriclinks/index.html>
The Econometrics Journal On-Line
- ❑ <http://www.elsevier.com/hes/books/02/menu02.htm>
Libro on-line: Handbook of Econometrics Vols. 1-5
- ❑ <http://elsa.berkeley.edu/users/mcfadden/discrete.html>
Libro on-line: Structural Analysis of Discrete Data and Econometric Applications
- ❑ http://www.oswego.edu/~kane/econometrics/stud_resources.htm
Online Resources for Econometric Students
- ❑ <http://www.econ.uiuc.edu/~morillo/links.html>
Econometric Sources: a collection of links in econometrics and computing. University of Illinois
- ❑ <http://www.econometrics.net/>
Econometrics, Statistics, Mathematics, and Forecasting
- ❑ <http://ideas.uqam.ca/EDIRC/ectrix.html>
Economics Departments, Institutes and Research Centers in the World: Econometrics, Mathematical Economics