CAPITULO 17. Extensiones al modelo de regresión lineal

Capítulo 17. EXTENSIONES AL MODELO DE REGRESIÓN LINEAL

En el proceso de investigación econométrica, en la etapa del análisis de la información se relacionaron las variables a través de la especificación del modelo lineal general y se estimaron los parámetros. Luego de la estimación, pueden surgir problemas relacionados con la parte sistemática o la parte aleatoria del modelo. Entre los primeros se encuentran el cambio estructural, el error de especificación y la multicolinealidad; entre los segundos la heterocedasticidad y la autocorrelación. Este capítulo se ocupará del primer grupo de problemas. En este contexto, es necesario contrastar hipótesis sobre la especificación del modelo, con la finalidad de realizar el mejor ajuste de acuerdo a los datos disponibles sobre las variables involucradas, en un espacio y tiempo determinado. A modo de síntesis, el interés está centrado en contrastar las hipótesis de cambio estructural, linealidad, omisión de variables relevantes o inclusión de variables irrelevantes y multicolinealidad.

17.1 Introducción

Los problemas en la especificación pueden deberse a:

1. Cambio estructural

Inestabilidad de los parámetros. La posibilidad de que los parámetros varíen entre distintos sub períodos de tiempo o entre distintos grupos de individuos, dentro de la muestra considerada. Dado que uno de los supuestos, del modelo de regresión, es la constancia de los parámetros en todo el periodo de medición o para la totalidad de la muestra considerada, es interesante contrastar la existencia de cambios en los coeficientes del modelo; es decir, de un cambio en la estructura del mismo.

Categorías. La introducción en el modelo de variables explicativas que no son cuantificables por naturaleza -como el sexo, la profesión o el nivel de estudios, entre otras-; o bien, se expresan de forma categórica -como, la renta o la edad definida por intervalos, entre otras-.

2. Error de especificación

Elección errónea de variables. La elección del conjunto de variables explicativas del modelo y los efectos que puede tener sobre la estimación mínimo cuadrado ordinaria de los parámetros una mala elección de las mismas, bien sea porque se omiten variables que son relevantes (omisión de variables

relevantes) o porque se incluyen variables que no lo son (inclusión de variables irrelevantes).

No Linealidad. También es posible que la relación estimada no sea lineal; es decir, las variables incluidas en el modelo son las correctas pero la relación entre ellas no es la adecuada. La presencia de no linealidades hace que los residuos muestren tendencias que indican su falta de aleatoriedad.

3. Multicolinealidad

Dependencia lineal. Al especificar el modelo se supone que las variables exógenas son linealmente independientes; esta se denomina hipótesis de independencia y cuando no se cumple el modelo presenta multicolinealidad.

Identificación de los parámetros del modelo. Estos problemas surgen cuando no se puede estimar de forma única todos los parámetros del modelo y las características de la información muestral disponible no permite estimar con precisión los parámetros.

17.2. Variables ficticias y cambio estructural

Se denomina variable ficticia, en general, a una variable que se construye artificialmente para recoger, en el modelo, ciertos aspectos de carácter discreto o cualitativo que expliquen el comportamiento de la variable dependiente.

La incorporación de estas últimas en el modelo se realiza a través de variables dicótomas, que asumen el valor 1 si se presenta el evento y 0 si no se presenta y se ve reflejado en cambios en el término independiente o en los coeficientes de las variables explicativas.

El modelo de variables ficticias, debidamente especificado, permite estudiar el cambio estructural en la ordenada, el cambio estructural en la pendiente de alguna o todas las variables explicativas y cambio estructural en la ordenada y pendiente de alguna o todas las variables explicativas.

Una variable cualitativa puede tener *m* categorías pero en el modelo deben definirse *m-1* variables ficticias. Por ejemplo, dada una variable cualitativa XCL que tiene *m* categorías, se define:

$$F1\begin{cases}1 \text{ cuando } XCL = 1\\ 0 \text{ cuando } XCL \neq 1\end{cases}$$

$$F2\begin{cases}1 \text{ cuando } XCL = 2\\ 0 \text{ cuando } XCL \neq 2\end{cases}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$F(m-1)\begin{cases}1 \text{ cuando } XCL = m-1\\ 0 \text{ cuando } XCL \neq m-1\end{cases}$$

Si se definen tantas variables ficticias como categorías tenga la variable cualitativa a estudiar, se estaría en presencia de la *trampa de las variables ficticias*. Esto ocurre porque, en la matriz de regresores, la suma de las variables ficticias (F) da lugar a una combinación lineal exacta con el vector unitario, el cual está presente para estimar el término independiente.

Ejemplo 17.1. Se quiere estudiar las necesidades básicas insatisfechas en los departamentos de Córdoba.

La hipótesis principal a verificar es que el nivel de NBI en cada departamento se encuentra influenciado –positivamente- por los niveles de desempleo y –negativamente- por la población con nivel universitario.

El modelo se especifica:

$$NBI_i = \alpha + \beta De_i + \gamma U_i + \varepsilon_i \qquad \forall\, i = 1,2,\dots,26$$

donde

NBI, es la población con necesidades básicas insatisfechas, medida en porcentaje respecto de la población total

De es la población desempleada, medida en porcentaje respecto de la población económica activa

U es la población con educación universitaria completa, medida en porcentaje respecto de la población total

Se cuenta con observaciones para los 26 departamentos de la Provincia de Córdoba, ordenadas en la Tabla 17.1.

Se supone que los parámetros no varían para los distintos departamentos de la provincia; sin embargo, se sospecha que, al determinar las necesidades básicas insatisfechas, puede ser relevante tener en cuenta el nivel de ingreso de cada jurisdicción. Para cuantificar el nivel de ingreso se utiliza como variable proxi el producto bruto geográfico departamental (PBG).

Los efectos del nivel de ingreso en una región pueden introducir diferencias tanto en el NBI autónomo como en los aportes marginales de las variables explicativas.

Tabla 17.1 Necesidades Básicas Insatisfechas

Observaciones		NBI	De	U	PBG	F	F1
1	Calamuchita	10.70000	4.310000	6.487834	1095698.	0	0
2	Capital	9.800000	7.210000	15.95350	27994361	1	0
3	Colon	12.70000	6.120000	8.625939	30662333	1	0
4	Cruz del Eje	21.80000	6.950000	2.806930	761792.0	0	0
5	Gral Roca	9.100000	3.250000	5.776455	1358954.	0	0
6	Gral S.Martin	6.600000	3.940000	7.527124	3593824.	0	0
7	Ischillin	17.80000	6.640000	4.961305	520643.0	0	0
8	Juarez Celman	7.100000	3.630000	6.021916	3055536.	0	0
9	Marcos Juarez	4.700000	2.980000	6.190132	4148290.	0	0
10	Minas	30.80000	4.210000	0.935804	54187.00	0	1
11	Pocho	29.10000	5.190000	1.787449	73461.00	0	1
12	Pte.R.S.Peña	7.500000	3.290000	4.644607	1464628.	0	0
13	Punilla	9.400000	5.970000	8.926385	2657506.	0	0
14	Río Cuarto	7.000000	5.340000	11.61595	6373160.	0	0
15	Río Primero	12.60000	4.480000	3.973984	1650740.	0	0
16	Río Seco	21.50000	5.790000	1.552442	398182.0	0	0
17	Río Segundo	7.300000	4.220000	5.676485	2806007.	0	0
18	San Alberto	15.60000	5.030000	3.552508	458629.0	0	0
19	San Javier	15.80000	5.800000	4.505533	672730.0	0	0
20	San Justo	6.500000	3.470000	5.735176	6133137.	0	0
21	Santa María	12.80000	7.940000	7.721690	1677149.	0	0
22	Sobremonte	17.00000	5.190000	1.692216	88539.00	0	0
23	Tercero Arriba	5.600000	4.220000	6.244579	3122242.	0	0
24	Totoral	14.30000	4.940000	4.277393	752073.0	0	0
25	Tulumba	18.30000	3.940000	1.843992	559049.0	0	0
26	Unión	7.300000	3.680000	5.176065	4028922.	0	0

NBI: Población que tiene necesidades básicas insatisfechas (en % sobre el total de personas del Departamento), calculado a partir del Censo de Población de la Provincia de Córdoba para el año 2008 publicados por la Dirección General de Estadísticas y Censos de la Provincia de Córdoba.

De: Desempleo (en % de personas desempleadas respecto del total de personas ocupadas y desempleadas del Departamento), calculado a partir del Censo de Población de la Provincia de Córdoba para el año 2008 publicados por la Dirección General de Estadísticas y Censos de la Provincia de Córdoba.

U: Porcentaje de población que tiene el nivel universitario completo como máximo nivel educativo alcanzado (en % sobre el total de personas del Departamento), calculado a partir del Censo de Población de la Provincia de Córdoba para el año 2008 publicados por la Dirección General de Estadísticas y Censos de la Provincia de Córdoba.

PBG. Producto Bruto Geográfico en miles de pesos corrientes para el año 2008, en base a datos publicados por la Dirección General de Estadísticas y Censos de la Provincia de Córdoba.

F: Variable ficticia o dummy que asume el valor 1 cuando el departamento posee un nivel de **PBG** superior al percentil 0.95.

F1: Variable ficticia o dummy que asume el valor 1 cuando el departamento posee un nivel de **NBI** superior al percentil 0.95.

Sea **REGION 1** el conjunto de departamentos que tienen un **PBG** menor al percentil 0.95, y **REGION 2** el conjunto de regiones con **PBG** superior.

Este efecto diferenciador se puede recoger en una sola ecuación definiendo una variable ficticia (F) que distinga entre los dos tipos de regiones:

$$F_i = \begin{cases} 1 \text{ si la región } i \in REGION\ 2 \\ 0 \text{ en otro caso} \end{cases}$$

Esta variable es la que permitirá que los valores de los parámetros, de la función de NBI, varíen de unas regiones a otras; es decir, que la estructura de la función sea distinta para cada tipo de región.

Se pueden presentar los siguientes casos:

1) que las diferencias en el nivel de NBI debidas al nivel de ingreso, se reflejan solo en el intercepto de la ecuación, es decir, en el NBI autónomo.

2) que en las regiones con alto nivel de ingreso, el desempleo sobre las necesidades básicas insatisfechas de la población tiene un efecto diferente de aquellas con bajo nivel de ingreso.

3) que las diferencias en cada región se observan tanto en el nivel de NBI autónomo como en el efecto que causa el desempleo.

Cada una de estas situaciones se especifica y evalúa, teóricamente, a continuación.

1. Para que las diferencias en el nivel de NBI, debidas al nivel de ingreso, se reflejen solo en el intercepto de la ecuación -es decir, en el NBI autónomo- la ecuación de NBI se puede especificar como sigue:

$$NBI_i = \alpha_1 + \delta F_i + \beta De_i + \gamma U_i + \varepsilon_i \qquad \forall\, i = 1,2,\ldots,26$$

donde el coeficiente que acompaña a la variable ficticia F_i recoge la diferencia en el NBI autónomo entre los departamentos, de acuerdo a su nivel de ingreso.

El modelo estimado es:

$$\widehat{NBI}_i = \hat{\alpha}_1 + \hat{\delta} F_i + \hat{\beta} De_i + \hat{\gamma} U_i$$

La utilización de variables ficticias permite recoger cambios en la función. Para contrastar este posible *cambio estructural* -es decir, si existe evidencia de un cambio en el NBI autónomo de un grupo de regiones a otro dependiendo del nivel de ingreso-, la hipótesis de contraste, es:

$H_0 : \delta = 0$
$H_A : \delta \neq 0$

La hipótesis nula indica no influencia de los niveles de ingreso en las necesidades básicas insatisfechas; si la hipótesis nula se rechaza es porque existen diferencias significativas.

Si el modelo cumple los supuestos del modelo de regresión lineal general, los estimadores mínimo cuadráticos ordinarios de los coeficientes de regresión tienen buenas propiedades y el contraste de hipótesis, basado en el estadístico F, es válido.

Retomando el modelo estimado

$$\widehat{NBI}_i = \hat{\alpha}_1 + \hat{\delta}F_i + \hat{\beta}De_i + \hat{\gamma}U_i$$

Cuando $F_i = 0$:

$$\widehat{NBI}_i = \hat{\alpha}_1 + \hat{\beta}De_i + \hat{\gamma}U_i$$

el NBI autónomo viene dado por $\hat{\alpha}_1$ e indica el nivel de necesidades básicas insatisfechas para los departamentos de la **REGION 1**

Cuando $F_i = 1$:

$$\widehat{NBI}_i = \left(\hat{\alpha}_1 + \hat{\delta}\right) + \hat{\beta}De_i + \hat{\gamma}U_i$$

el NBI autónomo viene dado por $\hat{\alpha}_1 + \hat{\delta}$ e indica la estimación del NBI para la **REGION 2**

Renombrando $\hat{\alpha}_1 + \hat{\delta} = \hat{\alpha}_2$, es válido expresar que $\hat{\delta} = \hat{\alpha}_2 - \hat{\alpha}_1$. Donde α_1 es la ordenada o nivel de NBI autónomo de las regiones de bajos ingresos, α_2 es la correspondiente a las regiones de altos ingresos y δ informa la magnitud del cambio.

Esto conduce a una segunda manera de especificar el modelo:

$$NBI_i = \alpha_1 + (\hat{\alpha}_2 - \hat{\alpha}_1)F_i + \beta De_i + \gamma U_i + \varepsilon_i \qquad \forall\, i = 1,2,\ldots,26$$

Una tercer forma de recoger esta diferencia de comportamiento en el intercepto entre los dos grupos de regiones, equivalente a las anteriores, consiste en definir **dos variables ficticias**:

$$F_{1i} = \begin{cases} 1 \text{ si la región } i \in REGION\ 1 \\ 0 \text{ en otro caso} \end{cases}$$

$$F_{2i} = \begin{cases} 1 \text{ si la región } i \in REGION\ 2 \\ 0 \text{ en otro caso} \end{cases}$$

Y especificar el modelo de NBI como sigue:

$$NBI_i = \alpha_1 F_{1i} + \alpha_2 F_{2i} + \beta De_i + \gamma U_i + \varepsilon_i \qquad \text{i} = 1,2,\ldots,26$$

En este caso, los coeficientes que acompañan a las variables ficticias recogen, respectivamente, cada uno de los dos niveles de NBI.

El contraste de cambio de estructura en el NBI autónomo, se basa en contrastar la siguiente hipótesis:

$H_0 : \alpha_1 = \alpha_2$
$H_A : \alpha_1 \neq \alpha_2$

Se puede observar que los dos modelos dados son equivalentes, siendo el último modelo una reparametrización del primero.

Cuando se incluyen tantas variables ficticias como grupos o categorías tiene la variable cualitativa, no se ha de incluir el término constante.

En este ejemplo, si se especifica el modelo:

$$NBI_i = \alpha_0 + \alpha_1 F_{1i} + \alpha_2 F_{2i} + \beta De_i + \gamma U_i + \varepsilon_i \qquad \text{i} = 1,2,\ldots,26$$

Se cae en la trampa de las variables ficticias.

La primera columna de la matriz de regresores $\mathbf{X}$, es la suma de la segunda y tercera columna. Por lo tanto, el rango de la matriz $\mathbf{X}$ no es completo $\rho(\mathbf{X}) = 4 < 5$ y la matriz $\mathbf{X}'\mathbf{X}$ es singular, por lo que $(\mathbf{X}'\mathbf{X})^{-1}$ no existe.

El sistema de ecuaciones normales, tiene menos ecuaciones linealmente independientes que incógnitas y no se puede resolver de forma única.

2. Se supone ahora que en las regiones con alto nivel de ingreso, el desempleo tiene un efecto diferente sobre las necesidades básicas insatisfechas de la población respecto de aquellas con bajo nivel de ingreso.

En este caso, cambia el efecto de la variable explicativa desempleo sobre el NBI; estas diferencias pueden recogerse en una sola ecuación; utilizando la variable ficticia F_i se tiene:

$$NBI_i = \alpha + \beta_1 De_i + \delta F_i De_i + \gamma U_i + \varepsilon_i \qquad \text{i} = 1,2,\ldots,26$$

Si se supone que el modelo cumple los supuestos del modelo de regresión, los estimadores por mínimo cuadrado ordinario de $\alpha, \delta, \beta, \gamma$ son insesgados y eficientes; en este caso, el contraste de hipótesis basado en el estadístico F sigue siendo válido.

El contraste de cambio estructural, en el efecto del desempleo, se basa en contrastar la siguiente hipótesis:

$H_0 : \delta = 0$
$H_A : \delta \neq 0$

Si se rechaza la hipótesis nula, existen diferencias significativas en el efecto del desempleo en los niveles de NBI para las diferentes regiones.

Retomando el modelo estimado

$$\widehat{NBI}_i = \hat{\alpha} + \hat{\beta}_1 De_i + \hat{\delta} F_i De_i + \hat{\gamma} U_i \qquad \text{i} = 1,2,\ldots,26$$

Cuando $F_i = 0$

$$\widehat{NBI}_i = \hat{\alpha} + \hat{\beta}_1 De_i + \hat{\gamma} U_i$$

$\hat{\beta}_1$ muestra el efecto que tiene, sobre las necesidades básicas insatisfechas, una variación en la proporción de población desempleada en las regiones con bajo ingreso.

Cuando $F_i = 1$

$$\widehat{NBI}_i = \hat{\alpha} + \left(\hat{\beta}_1 + \hat{\delta}\right) De_i + \gamma U_i$$

El efecto de una variación en la proporción de población desempleada por departamento en las necesidades básicas insatisfechas es $\hat{\beta}_1 + \hat{\delta} = \hat{\beta}_2$.

Es válido expresar que $\hat{\delta} = \hat{\beta}_2 - \hat{\beta}_1$. Donde β_1 es el efecto de variaciones en el desempleo sobre el nivel de NBI de las regiones de bajos ingresos, β_2 es la correspondiente a las regiones de altos ingresos y $\hat{\delta} = \hat{\beta}_2 - \hat{\beta}_1$ recoge la diferencia del efecto del desempleo sobre las necesidades básicas insatisfechas, en regiones con diferentes niveles de ingreso.

Esto conduce a una segunda manera de especificar el modelo:

$$NBI_i = \alpha + \beta_1 De_i + \left(\hat{\beta}_2 - \hat{\beta}_1\right) F_i De_i + \gamma U_i + \varepsilon_i \qquad \mathrm{i} = 1,2,\ldots,26$$

3. Para contrastar un posible cambio estructural en todos los parámetros de la ecuación de NBI, tanto en el intercepto como en el efecto de todas las variables explicativas entre ambos grupos de regiones, se especifica el modelo como sigue:

$$NBI_i = \alpha + \delta_1 F_i + \beta De_i + \delta_2 F_i De_i + \gamma U_i + \delta_3 F_i U_i + \varepsilon_i \qquad \mathrm{i} = 1,2,\ldots,26$$

Bajo el supuesto de que este modelo cumple las hipótesis del modelo de regresión, los estimadores mínimos cuadrados ordinarios $\alpha, \beta, \gamma, \delta_1, \delta_2, \delta_3$ son insesgados y eficientes y los contrastes siguen siendo válidos.

La hipótesis nula de no existencia de *cambio estructural* en la función de NBI entre ambos grupos -es decir, que el nivel de ingresos no afecta a la función-, es:

$H_0: \delta_1 = 0$ $\quad H_0: \delta_2 = 0$ $\quad H_0: \delta_3 = 0$
$H_0: \delta_1 \neq 0$ $\quad H_0: \delta_2 \neq 0$ $\quad H_0: \delta_3 \neq 0$

Que se puede contrastar con el estadístico t en forma individual y con el estadístico F, de restricciones lineales, en forma conjunta

Con los datos de la tabla 17.1 y utilizando Eviews:

a) Realice la estimación del modelo de NBI.
b) Evalúe la normalidad de los residuos
c) Determine si hay cambio estructural en las regiones debido al nivel de ingreso de las mismas y a su propio nivel de NBI (variables ficticias F y F1 en la tabla 17.1)

17.3 Contraste de Chow

Cuando se tienen *datos temporales* y la variable ficticia particiona a la muestra en conjuntos continuos, para estudiar un posible *cambio estructural,* se utiliza el contraste de CHOW. El test de Chow aplica el estadístico:

$$F = \frac{\left(\mathbf{e}_R' \mathbf{e}_R - \left(\mathbf{e}_1' \mathbf{e}_1 + \mathbf{e}_2' \mathbf{e}_2\right)\right)/k}{\frac{\left(\mathbf{e}_1' \mathbf{e}_1 + \mathbf{e}_2' \mathbf{e}_2\right)}{T-2k}} \overset{H_0}{\sim} F\left(k, T-2k\right)$$

Donde, $\mathbf{e}_R'\mathbf{e}_R$ es la suma de cuadrados de los residuos provenientes del modelo restringido, $\mathbf{e}_1'\mathbf{e}_1$ es la suma de cuadrados de los residuos provenientes de una de las partes del modelo y $\mathbf{e}_2'\mathbf{e}_2$ es la suma de cuadrados de los residuos provenientes del complemento.

Se rechaza la hipótesis nula de no existencia de cambio estructural, si el valor del estadístico es mayor que la ordenada $F_\alpha(k,T-2k)$ de la distribución F de Snedecor con $(k, T-2k)$ grados de libertad.

Para poder llevar a cabo el contraste de cambio estructural, utilizando las sumas de cuadrados de residuos de las regresiones para cada sub muestra, es necesario disponer en cada grupo de un número suficiente de observaciones para estimar los parámetros de la ecuación.

En algunas ocasiones, es posible que los sub períodos no dispongan el número de observaciones necesarias.

Sin pérdida de generalidad, bajo el supuesto de que en el segundo sub período -o grupo- el número de observaciones t_2 es menor o igual que el número de coeficientes de regresión, k.

El contraste de cambio estructural se realiza modificando el estadístico como sigue:

$$F = \frac{\left(e_R' e_R - e_1' e_1\right)/T_2}{\frac{e_1' e_1}{T_1 - k}} \overset{H_0}{\sim} \mathrm{F}\left(\mathrm{T}_2, \mathrm{T}_1 - \mathrm{k}\right)$$

Bajo la hipótesis nula de no existencia de cambio estructural entre los dos sub períodos, este estadístico, conocido con el nombre de contraste predictivo de Chow, se distribuye como una F de Snedecor.

17.4. Multicolinealidad

La existencia de correlación entre las variables explicativas en la muestra se denomina *multicolinealidad*. La hipótesis nula a contrastar es

$$\mathrm{H}_0 : \mathrm{No\ Multicolinealidad}$$

Dada la especificación del modelo, si algún o algunos regresores se pueden expresar como una combinación lineal exacta de otros regresores, entonces se dice que existe *multicolinealidad perfecta*.

En este caso extremo, el rango de la matriz x no es completo; es decir, $\rho(\mathbf{X}) < k$.

Por lo tanto, la matriz $(\mathbf{X}'\mathbf{X})$ no es invertible y no existe una solución única para $\hat{\boldsymbol{\beta}}$ del sistema de ecuaciones normales, $\mathbf{(X'X)\hat{\beta} = X'y}$.

La multicolinealidad perfecta es un problema de identificación, distintos valores de los parámetros generan el mismo valor medio de la variable dependiente, $E(\mathbf{y}) = \mathbf{X\beta}$.

Por lo tanto, dada la muestra (Y, X), no se pueden identificar aquellos valores de los parámetros que la han generado porque la función criterio minimizada $[\mathbf{y} - \mathbf{E(y)}]'[\mathbf{y} - \mathbf{E(y)}]$ no discrimina entre distintos valores de $\boldsymbol{\beta}$.

Ejemplo 17.2. Matriz de regresores colineales

Se considera el modelo de regresión:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad i = 1, \cdots, n$$

Donde se satisface que $X_{2i} + X_{3i} = 3$.

Entonces, la suma de la segunda y tercera columna de la matriz de regresores $\mathbf{X}$, es igual a tres veces la primera, por lo que el rango de la matriz de regresores, 2, es menor que el número de parámetros, 3, y no existe solución única al sistema de ecuaciones normales.

En este caso, para cualquier observación:

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} =$$
$$= (\beta_1 + 3\beta_2) + (\beta_3 - \beta_2) X_{3i} =$$
$$= \gamma_1 + \gamma_2 X_{3i}$$

Se pueden obtener distintos valores de $\beta_1, \beta_2, y\ \beta_3$ para los que las combinaciones lineales $(\beta_1 + 3\beta_2)\, y\, (\beta_3 - \beta_2)$ permanecen invariantes y, por lo tanto, proporcionan el mismo valor de $E(Y_i)$.

No es posible discriminar entre todos esos valores y solamente se puede identificar o estimar de forma única $\gamma_1 = (\beta_1 + 3\beta_2)$ y $\gamma_2 = (\beta_3 - \beta_2)$, es decir, combinaciones lineales de los parámetros de interés.

Observación. Si el problema no es de multicolinealidad perfecta, sino de un alto grado de colinealidad entre las variables explicativas, los parámetros del modelo de regresión se pueden estimar de forma única por mínimos cuadrados ordinarios, y los estimadores serán lineales, insesgados y óptimos.

Se supone que en un modelo, la correlación entre los regresores X_{2i} y X_{3i} es muy alta, es decir, $r_{23}^2 \cong 1$.

Se puede demostrar que la varianza del estimador mínimo cuadrado de los coeficientes asociados a X_{2i} y X_{3i} está

directamente relacionado con el grado de correlación existente entre los regresores:

$$V(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_{2t} - \bar{X}_2)^2 (1 - r_{23}^2)}$$

Cuanto mayor sea la correlación muestral entre los regresores, mayor será la varianza de los estimadores y menor la precisión con la que se estiman los coeficientes individualmente.

En el caso extremo, si $r_{23}^2 = 1$, $V(\hat{\beta}_2) \to \infty$ lo que implica que cualquier valor para β_2 es admisible.

Detección de la Multicolinealidad

Como síntomas más comunes de multicolinealidad se tienen los siguientes:

- *Matriz de correlaciones*, $\mathbf{R}_{xx}$, de las variables explicativas en el intervalo [0.72;0.99].
- Poca significatividad individual con alta significatividad conjunta y buen R^2.
- Influencia en las estimaciones de la eliminación de una observación en el conjunto de datos.
- Factores de inflación de la varianza

$$VIF = \frac{1}{(1 - R_j^2)} > 10$$

donde R_j^2 es el coeficiente de determinación de la regresión auxiliar de la variable explicativa j en función de las demás variables explicativas.

- Valores propios λ_i de $\mathbf{X'X}$ cercanos a cero o *Índice de condición*

$$\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2} > 30$$

- Entre los estadísticos para detectar la multicolinealidad se encuentra el contraste de Farrar-Glauber:

$$G = -\ln|\mathbf{R}_{xx}|\left[(T-1)-\frac{(2k+5)}{6}\right] \overset{H_0}{\sim} \chi^2_{\alpha;\frac{k(k-1)}{2}}$$

donde $\ln|\mathbf{R}_{xx}|$ es el logaritmo natural del determinante de la matriz de correlación de las variables explicativas incluidas en la estimación y la hipótesis a contrastar es no multicolinealidad.

Consecuencias de la multicolinealidad

Un alto grado de multicolinealidad tiene consecuencias negativas sobre las estimaciones:

1) Aunque se obtenga un buen ajuste en base al R^2 y, por lo tanto, evidencia de que conjuntamente las variables explicativas son estadísticamente significativas, los coeficientes estimados pueden tener grandes desviaciones típicas y pueden resultar individualmente no significativos.

2) Las estimaciones son muy inestables ante pequeños cambios en la muestra.

3) Los coeficientes estimados, pueden presentar signos incorrectos o magnitudes poco esperadas a priori.

La multicolinealidad puede afectar mucho a la estimación de unos parámetros y nada la de otros. Los parámetros asociados a variables explicativas poco correlacionadas con las restantes, se podrán estimar con precisión.

Una vez detectado un posible problema de multicolinealidad, es difícil solucionarlo.

No es probable que se obtenga información nueva; es decir, otra muestra que no represente este problema, porque, de disponer de ella se utilizaría.

Una posible solución, pero no buena, es eliminar del modelo algunas de las variables que crean el problema. Sin embargo, si las variables omitidas son relevantes, proceder de esta manera puede introducir sesgos en la estimación y problemas en la validez de los contrastes.

La multicolinealidad no afecta a la predicción $\mathbf{y_p} = \mathbf{X_p'}\,\hat{\boldsymbol{\beta}}$, siempre que la misma estructura de colinealidad se mantenga fuera de la muestra.

Tampoco afecta al vector de residuos mínimo cuadrático ordinario, $\mathbf{e}$, que siempre está definido, ni crea problemas en la estimación de σ_ε^2.

Solución para modelos con multicolinealidad

Las soluciones se pueden clasificar en robustas y no robustas.

Las primeras son aquellas que suprimen la variable que genera la multicolinealidad con justificación estadística y económica.

Entre aquellas no robustas se encuentran las que no transforman las variables y las que si lo hacen. Entre los métodos que no transforman variables está la solución de ampliar la muestra de datos. Entre los segundos, existen varias alternativas:

- Usar el *modelo en diferencias* vigilando la autocorrelación

- Usar transformaciones de las variables exógenas usando ratios

- Usar la *regresión en cadena*, que ofrece como estimadores de los parámetros a

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

siendo c una constante, que en la práctica suele tomarse con valores en el intervalo [0,01;0,1]. En este modelo, la matriz de varianzas y covarianzas es $\sigma^2(\mathbf{X}'\mathbf{X} + c\mathbf{I})^2\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + c\mathbf{I})^{-1}$

- Usar la regresión sobre componentes principales

Se supone un modelo de regresión con T observaciones y k variables explicativas, el método consiste en sustituir el conjunto de k variables explicativas por sus k componentes principales $C_{1t}, C_{2t}, \ldots C_{kt}$, o por un subconjunto de éstas.

Así, en el modelo lineal,

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

sea

$$\mathbf{Z} = \begin{bmatrix} z_{11} & \cdots & z_{k1} \\ \cdots & \cdots & \cdots \\ z_{1T} & \cdots & z_{kT} \end{bmatrix}$$

las observaciones expresadas en forma de *variables tipificadas*, correspondiente a las k variables explicativas; de tal forma que

$$\mathbf{R} = \frac{1}{T-1}\mathbf{Z'Z}$$

será la matriz de correlaciones muestrales entre las k variables explicativas.

La naturaleza de las componentes principales puede enfocarse de distintas formas. Cuántas dimensiones existen en el conjunto de las k variables explicativas; es decir, hay suficiente correlación entre ellas que hagan pensar que dos o más representan la misma dimensión para el análisis. Para ello se plantea la transformación de las mismas en un nuevo conjunto de variables que, tomadas de dos en dos, no estén correlacionadas; este nuevo conjunto se denominará componentes principales.

Una de las características de estas nuevas variables es que la primera recogerá la mayor varianza del análisis, la segunda la mayor parte de la varianza residual y así siguiendo… A estas nuevas variables se las obtiene a partir de los vectores propios, estos son las direcciones principales de la nube de puntos.

Para calcular los vectores propios se necesita, primero, calcular los valores propios y eso se obtiene diagonalizando la matriz $\mathbf{R}$. Es posible demostrar que existen k números reales positivos $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ y k vectores asociados $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k$ que forman una nueva base ortonormal de $\Re^k$ y que verifican,

$$\mathbf{R}\mathbf{p}_k = \lambda_k \mathbf{p}_k; \quad \forall k$$

$$\mathbf{R}\mathbf{p}_k - \lambda_k \mathbf{p}_k = 0$$

$$\mathbf{p}_k (\mathbf{R} - \lambda_k \mathbf{I}_k) = 0$$

$$\mathbf{R} - \lambda_k \mathbf{I}_k = 0 \Rightarrow \left|\mathbf{R} - \lambda_k \mathbf{I}_k\right| = 0 = \left|\lambda_k \mathbf{I}_k - \mathbf{R}\right|$$

La solución a este sistema genera los k valores propios buscados. A partir de ellos se calculan los k vectores propios, formando la matriz ortogonal $\mathbf{k}\, x \mathbf{k}$

$$\mathbf{P} = \left[\mathbf{p_1} \quad \mathbf{p_2} \quad \cdots \quad \mathbf{p_k}\right]$$

De esta forma, se tiene que

$$C_{1t} = p_{11} z_{1t} + p_{21} z_{2t} + \cdots + p_{k1} z_{kt}; \quad t = 1, \ldots, T$$

Representa la primera de las nuevas variables.

En forma matricial

$$\mathbf{C_1} = \mathbf{Z}\mathbf{p_1}$$

En donde $\mathbf{C_1}$ es un vector de T elementos –T observaciones transformadas– y $\mathbf{p_1}$ un vector de k elementos –la primera columna de la matriz de vectores propios–.

Observación. La suma de cuadrados de $\mathbf{C}_1$ es

$\mathbf{C}_1'\mathbf{C}_1 = \mathbf{p}_1'\mathbf{Z}'\mathbf{Z}\mathbf{p}_1$ (ó también $\frac{\mathbf{C}_1'\mathbf{C}_1}{T-1} = \mathbf{p}_1'\frac{\mathbf{Z}'\mathbf{Z}}{T-1}\mathbf{p}_1$)

Se elige $\mathbf{p}_1$ que maximice $\mathbf{C}_1'\mathbf{C}_1$, pero hay que imponer alguna restricción, caso contrario la suma de cuadrados de $\mathbf{C}_1$ podrá hacerse infinitamente grande. Para ello, se normaliza haciendo

$\mathbf{p}_1'\mathbf{p}_1 = 1$

Ahora se trata de obtener un máximo sujeto a restricciones. Se define:

$$\theta = \mathbf{p}_1'\frac{\mathbf{Z}'\mathbf{Z}}{\mathrm{T}-1}\mathbf{p}_1 - \lambda_1(\mathbf{p}_1'\mathbf{p}_1 - 1)$$

En donde λ_1 es un multiplicador de Lagrange. De esta forma se tiene

$$\frac{\partial\theta}{\partial\mathbf{p}_1} = \frac{2}{T-1}\mathbf{Z}'\mathbf{Z}\mathbf{p}_1 - 2\lambda_1\mathbf{p}_1$$

Aplicando la condición de primer orden de máximo, se obtiene

$$\frac{1}{T-1}(\mathbf{Z}'\mathbf{Z})\mathbf{p}_1 = \lambda_1\mathbf{p}_1$$

De esta forma se demuestra que $\mathbf{p}_1$ es un *vector propio* de la matriz $\mathbf{R} = \frac{1}{T-1}\mathbf{Z}'\mathbf{Z}$, correspondiente al *valor propio* λ_1.

Además, se observa que

$\frac{1}{T-1}\mathbf{Z_1}'\mathbf{Z_1} = \lambda_1 \mathbf{p_1}'\mathbf{p_1} = \lambda_1$ ¿Por qué?

Por lo que se debe elegir como λ_1 al mayor de los valores característicos de **R** que, en ausencia de multicolinealidad perfecta, será definida positiva y por lo tanto sus valores propios serán positivos, es decir

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$$

La primera componente principal de $\mathbf{Z}$ es entonces $\mathbf{C_1}$.

Se define $\mathbf{C}_2 = \mathbf{Z}\mathbf{p}_2$

Se debe elegir elegir $\mathbf{p}_2$ tal que maximice $\mathbf{p}_2'\frac{\mathbf{Z'Z}}{\mathrm{T}-1}\mathbf{p}_2$, sujeto a que $\mathbf{p}_2'\mathbf{p}_2 = 1$ y $\mathbf{p_1}'\mathbf{p}_2 = 0$.

La razón de la segunda restricción es que $\mathbf{C}_2$ no debe estar correlacionada con $\mathbf{C_1}$.

La covarianza entre ellas viene dada por

$\mathbf{p}_1'\mathbf{Z}'\mathbf{Z}\mathbf{p}_2 = \lambda_1 \mathbf{p_1}'\mathbf{p_2} = 0$, siempre que $\mathbf{p_1}'\mathbf{p_2} = 0$

Se define

$$\theta = \mathbf{p}_2'\frac{\mathbf{Z'Z}}{T-1}\mathbf{p}_2 - \lambda_2(\mathbf{p}_2'\mathbf{p}_2 - 1) - \lambda^*(\mathbf{p}_1'\mathbf{p}_2)$$

En donde λ_2, λ^* son multiplicadores de Lagrange.

$$\frac{\partial \theta}{\partial \mathbf{p}_2} = \frac{2}{T-1}\mathbf{Z}'\mathbf{Z}\mathbf{p}_2 - 2\lambda_2\mathbf{p}_2 - \lambda^*\mathbf{p}_1 = 0$$

Premultiplicando por $\mathbf{p_1}'$, queda

$$\frac{2}{T-1}\mathbf{p}_1'\mathbf{Z}'\mathbf{Z}\mathbf{p}_2 - \lambda^* = 0$$

lo que a su vez, implica que

$$\lambda^* = \frac{2}{T-1}\mathbf{p}_1'\mathbf{Z}'\mathbf{Z}\mathbf{p}_2 \therefore \lambda^* = \left(\lambda^*\right)' = \frac{2}{T-1}\mathbf{p}_2'\mathbf{Z}'\mathbf{Z}\mathbf{p}_1$$

Pero conociendo que,

$$\frac{1}{T-1}(\mathbf{Z}'\mathbf{Z})\mathbf{p_1} = \lambda_1\mathbf{p_1}$$

$$\frac{1}{T-1}\mathbf{p}_2'(\mathbf{Z}'\mathbf{Z})\mathbf{p_1} = \lambda_1\mathbf{p}_2'\mathbf{p_1} = 0$$

Entonces, $\lambda^* = 0$ y se tiene que,

$$\frac{1}{T-1}\mathbf{Z}'\mathbf{Z}\mathbf{p}_2 = \lambda_2\mathbf{p}_2$$

Aquí se elige λ_2 tal que sea la segunda raíz característica más grande de $\mathbf{Z}'\mathbf{Z}$.

Se puede proceder de esta forma para cada una de las k raíces de $\mathbf{Z'Z}$ y con los vectores resultantes formar la matriz ortogonal $\mathbf{P} = [\mathbf{p_1} \quad \mathbf{p_2} \quad \cdots \quad \mathbf{p_k}]$.

De esta manera las k componentes principales de $\mathbf{Z}$ vienen dadas por la matriz $\mathbf{C}$ de orden $\mathbf{T}x\mathbf{k}$ definida como

$$\mathbf{C} = \mathbf{ZP}$$

Que verifican

$$\frac{1}{T-1}\mathbf{C'C} = \frac{1}{T-1}\mathbf{P'Z'ZP} = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

De esta manera se puede decir que las componentes principales son centradas, no correlacionadas y sus varianzas son los valores propios.

Para obtener las coordenadas del $t-ésimo$ individuo en el nuevo sistema de ejes, se procede de la siguiente forma,

$$C_{1t} = p_{11}z_{1t} + p_{21}z_{2t} + \cdots + p_{k1}z_{kt}; \quad t = 1,\ldots,T$$
$$C_{2t} = p_{12}z_{1t} + p_{22}z_{2t} + \cdots + p_{k2}z_{kt}; \quad t = 1,\ldots,T$$
$$\vdots$$
$$C_{kt} = p_{1k}z_{1t} + p_{2k}z_{2t} + \cdots + p_{kk}z_{kt}; \quad t = 1,\ldots,T$$

Ejemplo 17.3. La tabla 17.2 contiene información sobre 5 observaciones para tres variables explicativas (X_1, X_2, X_3). Estas variables presentan alta correlación por lo que se procede a calcular las componentes principales.

Tabla 17.2 Matriz de valores de X

Observaciones	X2	X3	X4
1	2	3	2
2	4	3	2
3	5	4	3
4	2	2	1
5	1	2	2

De acuerdo a lo analizado teóricamente se tienen que calcular los vectores propios ortogonales de la matriz $\mathbf{Z'Z}$, donde $\mathbf{Z}$ es la matriz de variables tipificadas de los valores originales de la matriz de variables independientes. Los elementos de $\mathbf{Z'Z}$ serán los siguientes, (¿por qué?)

$$\mathbf{Z'Z} = \begin{bmatrix} 4 & 3.4915 & 2.5820 \\ 3.4915 & 4 & 3.3806 \\ 2.5820 & 3.3806 & 4 \end{bmatrix}$$

Si se divide la matriz $\mathbf{Z'Z}$ por $T-1$ se obtiene la matriz de correlaciones, $\mathbf{R}$, de las variables explicativas X_{it} ¿Por qué?

Para calcular los vectores propios se necesita –primero- calcular los valores propios y eso se obtiene diagonalizando la matriz $\mathbf{Z'Z}$ Según los datos del ejemplo, existen $k=3$ números reales positivos $\lambda_1 \geq \lambda_2 \geq \lambda_3$ y $k=3$ vectores asociados $\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}$ que

forman una nueva base ortonormal de $\Re^3$ y que verifican, $\mathbf{R}\mathbf{p}_k = \lambda_k \mathbf{p}_k ; \quad \forall k = 1,2,3$

El lector deberá comprobar que la solución de este sistema genera 3 valores propios que, para los datos que se tienen en la tabla, son:

$\lambda_1 = 2{,}579783 \quad \lambda_2 = 0{,}355272 \quad \lambda_3 = 0{,}064945$

Y que forman la matriz diagonal correspondiente a $\mathbf{R}$

$$\Lambda = \begin{bmatrix} 2{,}579783 & 0 & 0 \\ 0 & 0{,}355272 & 0 \\ 0 & 0 & 0{,}064945 \end{bmatrix}$$

Demuestre que a estos valores propios le corresponden los vectores propios

$$\mathbf{p_1} = \begin{bmatrix} -0.564302 \\ -0.609621 \\ -0.556709 \end{bmatrix} \qquad \mathbf{p}_2 = \begin{bmatrix} 0.679974 \\ 0.039211 \\ -0.732187 \end{bmatrix} \qquad \mathbf{p}_3 = \begin{bmatrix} -0.468186 \\ 0.791723 \\ -0.392400 \end{bmatrix}$$

Se puede elegir $\mathbf{p}_k$ *ó* $(-\mathbf{p}_k) \forall k = 1,2,3$

Las componentes principales serán entonces,

$C_{1t} = -0.564302 z_{1t} - 0.609621 z_{2t} - 0.556709 z_{kt} ; \quad t = 1,\ldots,24$

$C_{2t} = 0.679974 z_{1t} + 0.039211 z_{2t} - 0.732187 z_{kt} ; \quad t = 1,\ldots,24$

$C_{3t} = -0.468186 z_{1t} + 0.791723 z_{2t} - 0.392400 z_{kt} ; \quad t = 1,\ldots,24$

Dada la matriz $\mathbf{Z}$

Observaciones	z_1	z_2	z_3
1	-0.4869	0.2390	0.0000
2	0.7303	0.2390	0.0000
3	1.3389	1.4343	1.4142
4	-0.4869	-0.9562	-1.4142
5	-1.0954	-0.9562	0.0000

Las coordenadas del $t-ésimo$ período en el nuevo sistema de ejes son

Observaciones	c_1	c_2	c_3
1	0.1290	-0.3217	0.4172
2	-0.5578	0.5060	-0.1527
3	-2.4172	-0.0688	-0.0462
4	1.6450	0.6669	0.0258
5	1.2011	-0.7824	-0.2442

Así, por ejemplo, la *primera coordenada* de la observación 3 en componentes principales, se obtuvo haciendo,

$$C_{1,3} = -0.564302 \cdot (1.3389) - 0.60962 \cdot (1.4343) - 0.556709 \cdot (1.4142)$$
$$= -2.4172$$

También se sugiere al lector que verifique que

$$\overline{C_1} = \overline{C_2} = \overline{C_3} = 0$$
$$V(C_1) = \lambda_1; V(C_2) = \lambda_2; V(C_3) = \lambda_3$$
$$Cov(C_1 C_2) = 0;\ Cov(C_1 C_3) = 0;\ Cov(C_2 C_3) = 0$$

Las componentes principales fueron obtenidas postmultiplicando la matriz de variables explicativas tipificadas por la matriz de vectores propios.

Teniendo en cuenta que $\mathbf{P'P} = \mathbf{I}_k$ y que los autovectores anteriores además de ortogonales se pueden elegir unitarios. El modelo original se puede transformar en

$$\mathbf{Y} = \mathbf{Z\beta} + \boldsymbol{\varepsilon} = \mathbf{ZPP'\beta} + \boldsymbol{\varepsilon} = \mathbf{C\alpha} + \boldsymbol{\varepsilon}$$

Los coeficientes de regresión $\boldsymbol{\alpha} = \mathbf{P'\beta}$ están asociados a k variables explicativas no correlacionadas pues las componentes principales son ortogonales.

Este modelo auxiliar

$$Y_t = \alpha_0 + \alpha_1 C_{1t} + \cdots + \alpha_k C_{kt} + \varepsilon_t; \qquad t = 1,\ldots,T$$

No estará afectado de multicolinealidad pues las variables $C_{1t},\ldots C_{kt}$ no están correlacionadas.

Si se eliminan las variables explicativas $C_{r+1},\ldots,C_k$, que son las $k-r$ últimas componentes cuya variabilidad es menor, se pierde poca información y el modelo resultante.

$$Y_t = \alpha_0^* + \alpha_1^* C_{1t} + \cdots + \alpha_r^* C_{rt} + \varepsilon_t^*; \qquad t = 1,\ldots,T$$

Será una aproximación al original, sin multicolinealidad, y a partir de sus estimaciones se obtiene el estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$.

Como, $\boldsymbol{\beta} = \mathbf{P}\boldsymbol{\alpha} = [\mathbf{P}_1 \left| \mathbf{P}_2 \right][\frac{\boldsymbol{\alpha}^*}{\boldsymbol{\alpha}^{**}}]$

Donde,

$\mathbf{P}_1$ es la matriz formada por las $r+1$ primeras columnas de $\mathbf{P}$

$\boldsymbol{\alpha}^* = (\alpha_0^* \alpha_1^* \cdots \alpha_r^*)'$.

Si las últimas $k-r$ componentes principales explican una pequeña parte de la variabilidad de las variables predeterminadas del modelo original, o sea si se puede considerar $\boldsymbol{\alpha}^{**} \cong \mathbf{0}$

Resulta que, $\boldsymbol{\beta} = \mathbf{P}_1 \boldsymbol{\alpha}^*$ con lo que el estimador de $\boldsymbol{\beta}$ será, $\hat{\boldsymbol{\beta}} = \mathbf{P}_1 \hat{\boldsymbol{\alpha}}^*$

Siendo $\hat{\boldsymbol{\alpha}}^*$ el estimador de los coeficientes $\boldsymbol{\alpha}^*$ en el modelo de las $r+1$ primeras componentes principales.

Ejemplo 17.3. (continuación). Para ilustrar esta segunda parte del análisis de las componentes principales con los datos del ejemplo se incluye una estimación al final del capítulo.

Por otra parte, la variación total de las variables tipificadas $\mathbf{Z}$ viene dada por

$$\sum_t z_{1t}^2 + \sum_t z_{2t}^2 + \cdots + \sum_t z_{kt}^2 = tr(\mathbf{Z}'\mathbf{Z})$$

Pero,

$\text{tr}(\mathbf{P'Z'ZP}) = \text{tr}(\mathbf{Z'ZPP'}) = \text{tr}(\mathbf{Z'Z})$, debido a que, $\mathbf{P'P} = \mathbf{I}_k$

Quiere decir que,

$$tr(\mathbf{P'Z'ZP}) = tr(\mathbf{Z'Z}) = tr\mathbf{\Lambda} = \sum_{i=1}^{k} \lambda_i = \\ = \mathbf{Z}_1'\mathbf{Z}_1 + \cdots + \mathbf{Z}_k'\mathbf{Z}_k$$

Pero como se ha trabajado con la matriz de variables tipificadas y diagonalizando la matriz de correlaciones, se tiene que esta última suma, igual a la traza de la matriz lambda, es igual a k. (Comprobar)

De esta forma,

$$\frac{\lambda_1}{\sum\lambda}, \frac{\lambda_2}{\sum\lambda}, \ldots, \frac{\lambda_k}{\sum\lambda}$$

Representa la proporción en que cada componente principal contribuye a la explicación de la varianza total de las $\mathbf{Z}$, y puesto que las componentes son ortogonales, estas proporciones suman la unidad (que el lector deberá comprobar).

Con frecuencia, la correlación entre los datos económicos y sociales significa que un número pequeño de componentes explicarán una gran proporción de la variación total y sería deseable poder realizar una prueba de hipótesis para evaluar cuál es el número de componentes que debe retenerse para un análisis posterior. Supongamos que hemos calculado las raíces $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ y que las

primeras r raíces $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r ; (r < k)$, parecen ser suficientemente grandes y diferentes como para retenerlas. En este caso, la pregunta es si las restantes $k-r$ raíces son lo suficientemente parecidas entre sí como para concluir que los verdaderos valores son iguales. Es decir, la hipótesis nula a corroborar es

$$H_0 : \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_k$$

Un contraste de hipótesis aproximado se basa en el estadístico

$$rho = T \ln\left[(\lambda_{r+1}\lambda_{r+2} \ldots \lambda_k)^{-1} \left(\frac{\lambda_{r+1}\lambda_{r+2} \ldots \lambda_k}{k-r} \right)^{k-r} \right] \overset{H_0}{\sim} \chi^2_{1/2(k-r-1)(k-r+2)}$$

En las aplicaciones prácticas (ver problema al final del capítulo) se espera que el número de componentes significativamente diferentes r que han de retenerse sea sustancialmente menor que el número k a partir de las cuales se obtienen las componentes.

Observación. El siguiente resultado muestra en forma conjunta las propiedades anteriores. Sea $\mathbf{c}$ un vector columna de k elementos y v una magnitud aleatoria escalar.

$$\underset{1x1}{v} = \underset{1xk}{\mathbf{c}'} \underset{kx1}{\boldsymbol{\beta}}$$

De tal manera que si se elige $\mathbf{c}' = [0 \quad 1 \quad 0 \quad \cdots \quad 0]$

Entonces, $\nu = [0 \quad 1 \quad 0 \quad \cdots \quad 0]\begin{bmatrix}\beta_1 \\ \beta_2 \\ \vdots \\ \beta_k\end{bmatrix} = \beta_2$

De esta forma, se puede usar $\nu = \mathbf{c'\beta}$ para seleccionar un elemento de $\boldsymbol{\beta}$.

Pero también, si $\mathbf{c'} = [1 \quad X_{2,n+1} \quad X_{3,n+1} \quad \cdots \quad X_{k,n+1}]$

Entonces, $\nu = E(Y_{n+1})$

Que es el *valor esperado* de la variable endógena Y en el período (u observación) $n+1$ condicionado a los valores de X en ese período.

Se considera una *clase* de estimadores *lineales e insesgados* de ν. Sea ω un escalar definido como combinación lineal de ν, tal que

$$\omega = \mathbf{a'y} = \mathbf{a'X\beta} + \mathbf{a'}\nu$$

Donde $\mathbf{a}$ es un vector columna de n elementos y donde $\mathbf{y, X, \beta}$ son los vectores y matriz definidas anteriormente. ω será un estimador insesgado de ν si y solamente si $\mathbf{a'X} = \mathbf{c'}$, se observa que

$$\begin{aligned} E(\omega) &= \mathbf{a'X\beta} + \mathbf{a'}E(\nu) = \\ &= \mathbf{a'X\beta} = \\ &= \mathbf{c'\beta} \Leftrightarrow \mathbf{a'X} = \mathbf{c'} \end{aligned}$$

Además,

$$
\begin{aligned}
V(\omega) &= E\{[\omega - E(\omega)]^2\} = \\
&= E\{[\mathbf{a'X\beta} + \mathbf{a'}\nu - \mathbf{a'X\beta}]^2\} = \\
&= E\{(\mathbf{a'}\nu)(\mathbf{a'}\nu)'\} = \\
&= E\{\mathbf{a'}\nu\nu'\mathbf{a}\} = \mathbf{a'}E(\nu\nu')\mathbf{a}
\end{aligned}
$$

Por tanto,

$$V(\omega) = \sigma^2 \mathbf{a'a}$$

Entonces el problema es elegir $\mathbf{a}$ para minimizar $\mathbf{a'a}$ sujeto a las k restricciones de que $\mathbf{a'X} = \mathbf{c'}$, esto es $\mathbf{a'X} - \mathbf{c'} = \mathbf{0}$.

Por lo que se tiene un problema de mínimo sujeto a restricciones. Utilizando los multiplicadores de Lagrange, se define

$$\underset{1x1}{\phi} = \underset{1x1}{\mathbf{a'a}} - 2\underset{1xk}{\boldsymbol{\lambda}'}(\underset{kxn}{\mathbf{X}'}\underset{nx1}{\mathbf{a}} - \underset{kx1}{\mathbf{c}})$$

Donde $\boldsymbol{\lambda}$ es el vector columna de los k multiplicadores de Lagrange (orden $\mathbf{kx1}$) y donde $\mathbf{a'X} - \mathbf{c'}$ se ha transpuesto para ser conformable.

Al diferenciar, para obtener la primera condición,

$$\frac{\partial \phi}{\partial \mathbf{a}} = 2\mathbf{a} - 2\mathbf{X}\boldsymbol{\lambda} = \mathbf{0} \Rightarrow \mathbf{a} - \mathbf{X}\boldsymbol{\lambda} = \mathbf{0}$$

$$\frac{\partial \phi}{\partial \boldsymbol{\lambda}} = 2(\mathbf{X'a} - \mathbf{c}) = \mathbf{0} \Rightarrow \mathbf{X'a} - \mathbf{c} = \mathbf{0}$$

De donde,

$$\mathbf{a} = \mathbf{X}\boldsymbol{\lambda} \Rightarrow \mathbf{X}'\mathbf{X}\boldsymbol{\lambda} = \mathbf{c} \Rightarrow \boldsymbol{\lambda} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$$
$$\therefore \mathbf{a} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$$

De forma tal que el estimador lineal e insesgado de varianza mínima deseado de $\nu = \mathbf{c}'\boldsymbol{\beta}$ es

$$\omega = \mathbf{a}'\mathbf{y}$$
$$\omega = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$\omega = \mathbf{c}'\hat{\boldsymbol{\beta}}$$

17.5. Error de especificación

Habitualmente se entiende por error de especificación al error que se comete en la especificación de la parte sistemática del modelo de regresión; es decir, qué variables explicativas se incluyen o se omiten, cuál es la forma funcional, etc.

A pesar de que pueden existir muchos problemas en la especificación del modelo, este apartado se ocupará solo de los relacionados con la selección del conjunto de variables explicativas -es decir, a las consecuencias de omitir variables relevantes o de incluir variables irrelevantes en el modelo- y a confirmar la forma funcional lineal.

Omisión de variables relevantes

Se supone que el modelo correctamente especificado es de la forma:

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Donde

- $X_1 (Txk_1)$ y $X_2 (Txk_2)$, son matrices de regresores no estocásticos.
- $E(\boldsymbol{\varepsilon}) = \mathbf{0}; E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_T$

Sin embargo, se especifica y se estima el siguiente modelo,

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}^*$$

Donde se han omitido (k_2) variables explicativas de la parte sistemática del modelo.

Dado que la perturbación del modelo es $\boldsymbol{\varepsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, se tiene

$$E(\boldsymbol{\varepsilon}^*) = \mathbf{X}_2\boldsymbol{\beta}_2 \qquad \text{y} \qquad E(\mathbf{X}_1'\boldsymbol{\varepsilon}^*) = \mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$$

Es interesante observar que, si al especificar el modelo de regresión se omiten variables explicativas relevantes para determinar la variabilidad de $\mathbf{y}$, el efecto de estas variables queda recogido en el término de error.

El comportamiento de la perturbación $\boldsymbol{\varepsilon}^*$ va a reproducir el funcionamiento de las variables $\mathbf{X}_2$ omitidas, por lo que, salvo casos

excepcionales, no va a cumplir los supuestos exigidos en el modelo de regresión lineal general.

Este resultado lleva a cuestionar las propiedades del estimador mínimo cuadrado ordinario de $\boldsymbol{\beta}_1$ en el modelo.

En este sentido, es fácil demostrar que el valor medio del estimador, es:

$$E\left(\hat{\boldsymbol{\beta}}_1\right) = \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1' E(\mathbf{y}) = \boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$$

El sesgo del estimador desaparece si: $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$.

Esta condición implica que las variables explicativas incluidas en el modelo y las omitidas, no están correlacionadas.

Por otro lado, el estimador habitual de la varianza de las perturbaciones:

$$S^2 = \frac{\mathbf{e}^{*'}\mathbf{e}^{*}}{T - k_1}$$

Será también sesgado, aunque se cumpla que $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, lo que implica que el estimador de $\mathbf{V}\left(\hat{\boldsymbol{\beta}}_1\right)$:

$$\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_1\right) = S^2\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}$$

No es insesgado y los contrastes de hipótesis habituales sobre el vector de coeficientes $\boldsymbol{\beta}_1$, no son válidos porque:

$I)$ $\dfrac{\mathbf{e}^{*'}\mathbf{e}^{*}}{\sigma^2}$ no se distribuye como una χ^2

$II)$ $\hat{\boldsymbol{\beta}}_1 \sim N\left(\boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\,\boldsymbol{\beta}_2,\ \sigma^2\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right)$

Inclusión de variables irrelevantes

Se supone que el modelo correctamente especificado es:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$

Donde $\mathbf{X}_1$ es una matriz $\mathbf{T}\,x\,\mathbf{k}$ de regresores no estocásticos y la perturbación, sigue una distribución normal con $E(\boldsymbol{\varepsilon}) = \mathbf{0}; E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right) = \sigma^2\mathbf{I}_T$.

Sin embargo, se incluyen (k_2) variables en el modelo de regresión que no son relevantes, de forma que se estima por mínimos cuadrados el siguiente modelo:

$$\mathbf{y} = \mathbf{X}_1\,\boldsymbol{\beta}_1 + \mathbf{X}_2\,\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Los estimadores mínimos cuadráticos ordinarios de los vectores de parámetros $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ obtenidos a partir del modelo, son:

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \left[\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\right]^{-1}\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}\mathbf{y} = \left[\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\right]^{-1}\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}\left(\mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}\right) =$$

$$= \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1\\ \mathbf{X}_2'\mathbf{X}_1\end{bmatrix}\boldsymbol{\beta}_1 + \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}\boldsymbol{\varepsilon}$$

Se puede demostrar que:

$$\begin{bmatrix} \mathbf{X}_1' \mathbf{X}_1 & \mathbf{X}_1' \mathbf{X}_2 \\ \mathbf{X}_2' \mathbf{X}_1 & \mathbf{X}_2' \mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1' \mathbf{X}_1 \\ \mathbf{X}_2' \mathbf{X}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{k_1} \\ \mathbf{0}_{k_2} \end{bmatrix}$$

Se obtiene que el valor medio de los estimadores mínimos cuadrático ordinario del modelo, es:

$$E\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0}_{k_2} \end{bmatrix} + \begin{bmatrix} \mathbf{X}_1' \mathbf{X}_1 & \mathbf{X}_1' \mathbf{X}_2 \\ \mathbf{X}_2' \mathbf{X}_1 & \mathbf{X}_2' \mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix} E(\boldsymbol{\varepsilon}) = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0}_{k_2} \end{bmatrix}$$

Por lo que se puede concluir que son insesgados; es decir, $E(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1$ y $E(\hat{\boldsymbol{\beta}}_2) = \boldsymbol{\beta}_2 = \mathbf{0}$ (dado que las variables $\mathbf{X}_2$ son irrelevantes).

Observación. Ahora bien, hay que señalar que a la hora de estimar los parámetros de interés de $\boldsymbol{\beta}_1$ no se incorpora toda la información disponible, ya que no se incluye la restricción cierta de que $\boldsymbol{\beta}_2 = \mathbf{0}$.

Por lo tanto, relativamente, se está perdiendo eficiencia al estimar $\boldsymbol{\beta}_1$ en el modelo mal especificado, que estimarlo en el modelo bien especificado. El estimador de la varianza de las perturbaciones en el modelo mal especificado:

$S^2 = \dfrac{\mathbf{e}'\mathbf{e}}{T-k}$ es un estimador insesgado de σ^2 y se mantiene la validez de los contrastes habituales de restricciones lineales sobre el vector de coeficientes $\boldsymbol{\beta}$.

Pruebas de errores de especificación.

1) Detección de la presencia de variables innecesarias: *data–mining*.

   Si un investigador desarrolla un modelo de *k* variables y va probando *una a una* la inclusión o no de variables, realiza lo que se conoce como *regresión por etapas.*

   Una de las consecuencias a la que se enfrenta es que estará modificando los niveles de significación.

   Lowel ha sugerido que si hay *c* candidatos a regresores de los cuales *k* son finalmente seleccionados (*k* < *c*) con base en la data–mining, entonces el verdadero nivel de significación (α^*) está relacionado con el nivel de significación nominal (α) de la siguiente manera:

   $$\alpha^* = (C/k)\cdot\alpha$$

   Por ejemplo, si *c = 15*, *k = 5 y* $\alpha = 5\%,$ el verdadero valor de significación es 15%.

   Por consiguiente, si un investigador extrae datos y selecciona 5 de 15 regresores y solamente informa los resultados al nivel de significación del 5% nominal y declara que estos resultados son estadísticamente significativos, esta conclusión se debe tomar con gran reserva.

2) Existen contrastes para observar si un modelo adolece de variables omitidas. El test de la razón de verosimilitud para variables omitidas permite añadir un conjunto de variables a una ecuación existente y contrastar si constituyen una

contribución significativa a la explicación de la variable dependiente. Este contraste tiene como hipótesis nula que el conjunto de regresores adicionales no son conjuntamente significativos.

También se puede aplicar el test de la razón de verosimilitud para variables redundantes que permite contrastar si un subconjunto de variables de una ecuación existente es conjuntamente significativo.

El test de Wald puede utilizarse para detectar cuando una variable es redundante. Basta comprobar cuando puede considerarse cero su coeficiente de modo formal a través de esta prueba.

Linealidad del modelo

La *linealidad* del modelo puede ser evaluada a partir de la prueba RESET de Ramsey. Partiendo de que cualquier función puede ser aproximada por polinomios del orden adecuado, en el modelo de regresión se pueden introducir términos con las potencias sucesivas de la variable endógena.

El contraste de Ramsey realiza una prueba para comprobar si los coeficientes de las potencias incluidas en el modelo se anulan; si se confirma esta hipótesis, se acepta la forma funcional lineal del mismo.

Para realizar el contraste RESET se debe decidir cuantas funciones de los valores ajustados se incluirán en la regresión ampliada. No hay una respuesta concreta a esta pregunta, pero los términos al cuadrado y al cubo suelen ser suficientes en la mayoría de los casos.

Sean $\hat{Y}_t$ los valores ajustados por MCO al estimar la ecuación

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

Se considera la ecuación ampliada

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \alpha_2 \hat{Y}^2 + \alpha_3 \hat{Y}^3 + \varepsilon_t$$

Obviamente no hay interés en los valores estimados de esta última ecuación, solo se quiere determinar la existencia de linealidad en el modelo estimado originalmente. Se debe recordar, al respecto, que $\hat{Y}^2, \hat{Y}^3$ son funciones no lineales de las variables exógenas.

La hipótesis nula es la linealidad. Formalmente, Ramsey establece,

$$H_0 : \boldsymbol{\varepsilon} \approx \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}); \quad H_1 : \boldsymbol{\varepsilon} \approx \mathbf{N}(\boldsymbol{\varepsilon}, \sigma^2 \mathbf{I}) \ \forall \boldsymbol{\varepsilon} \neq \mathbf{0}$$

El estadístico RESET es una F que, bajo hipótesis nula, tiene $2, T-k-2$ grados de libertad. ¿por qué?. En general, se pueden expresar los grados de libertad en función de la cantidad de regresores que se añaden, pero teniendo en cuenta que se deben dejar los suficientes grados de libertad para la estimación del modelo.

$$F = \frac{(R_n^2 - R_v^2)/k_n}{(1 - R_n^2)/(n - k_n)}$$

El estadístico se construye con los coeficientes de determinación de la ecuación original y la ampliada (R_n^2 y R_v^2, respectivamente) y los grados de libertad tienen en cuenta los parámetros adicionales en la ecuación ampliada.

CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

Problema 17.1: Componentes principales

Dada la siguiente tabla de datos, obtenga las componentes principales.

Tabla 17.3

Observación	X_2	X_3	X_4
1	7	15	4
2	6	12	3
3	4	10	1
4	3	11	-1
5	6	14	0
6	4	10	5

Caso 17.1: Determinantes del consumo

Dada las series de datos de PIB, Consumo y Tasa de Interés de Argentina para el periodo primer trimestre de 1993 a primer trimestre de 2009,

a) pruebe si hay cambio estructural en el modelo

$$\text{Consumo}_t = \alpha + \beta\,\text{PBI}_t + \gamma\,\text{Interés}_t + \varepsilon_t \qquad t = 1'93, 2'93, \cdots, 1'09$$

b) analice el gráfico de residuos que surge de estimar el modelo anterior a partir de los datos de la Tabla 17.4.

c) actualice la Tabla 17.4, realice nuevamente la estimación y compare los resultados.

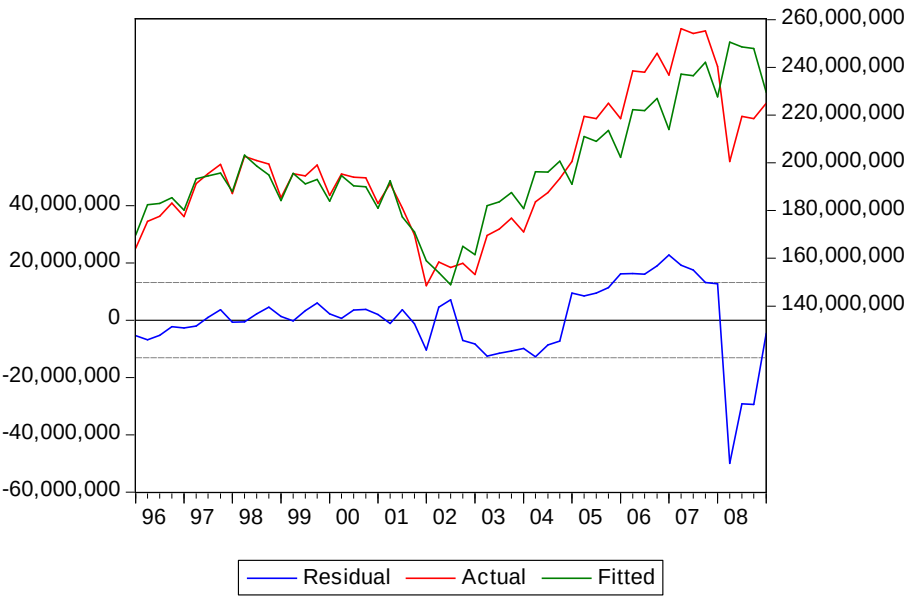

Tabla 17.4. Indicadores Macroeconómicos

Periodo		PIB	CONSUMO	F	INTERES
1993	I	216370111	152148446	1	
	II	241871858	166025867	1	
	III	242645522	166667550	1	
	IV	245132429	169860311	1	
1994	I	232945326	164965420	1	
	II	257476895	177234828	1	
	III	253467778	174510154	1	
	IV	257341544	177721808	1	
1995	I	237968103	164321480	1	
	II	248093639	166567449	1	
	III	242214699	164276737	1	
	IV	244467965	168866520	1	
1996	I	236566037	164311572	1	7.837
	II	260751925	175591878	1	6.773
	III	262166964	177726972	1	7.293
	IV	267020047	183153037	1	7.523
1997	I	256387857	177490019	1	7.007
	II	281769801	191310690	1	6.530
	III	284092268	195505523	1	6.410
	IV	287515346	199383506	1	7.920
1998	I	271702368	187196678	1	7.093
	II	301207598	202675183	1	6.667
	III	293315404	200922426	1	8.093
	IV	286267849	199434263	1	8.393
1999	I	265024636	185463056	1	8.110
	II	286412327	195463399	1	6.610
	III	278472694	194457732	1	7.780
	IV	283566399	199054269	1	9.687
2000	I	264555918	186315129	1	7.797
	II	285275176	195338736	1	7.630
	III	276767971	193972609	1	7.485
	IV	278091676	193703380	1	10.439
2001	I	259199874	182900187	1	8.678
	II	284795763	191297580	1	12.750
	III	263126505	181090983	1	22.867
	IV	248864555	169871185	1	20.359
2002	I	216849495	148507392	0	9.394
	II	246314633	158475554	0	60.913
	III	237416867	156093858	0	62.071
	IV	240361392	157992266	0	24.616

Tabla 17.4. Indicadores Macroeconómicos *Continuación*

Periodo		PIB	CONSUMO	F	INTERES
2003	I	228595882	153188337	0	18.277
	II	265402478	169567358	0	13.874
	III	261534523	172253988	0	4.578
	IV	268560967	176794330	0	3.913
2004	I	254330423	171056272	0	2.360
	II	284375611	183635133	0	2.330
	III	284392060	187557703	0	2.744
	IV	293467061	193373719	0	3.027
2005	I	274594503	200565514	0	2.782
	II	313927290	219462442	0	3.535
	III	310593081	218509900	0	4.125
	IV	319939241	224988560	0	4.607
2006	I	298695561	218515535	0	5.626
	II	338243727	238547451	0	6.518
	III	337741885	237975913	0	6.874
	IV	347578707	245923679	0	6.667
2007	I	322448871	236761556	0	7.189
	II	367492351	256321622	0	6.874
	III	367538727	254163194	0	8.331
	IV	379199661	255268779	0	9.493
2008	I	349945322	240312979	0	8.256
	II	396227240	200565514	0	10.237
	III	393039229	219462442	0	10.938
	IV	394564940	218509900	0	14.766
2009	I	357077664	224988560	0	12.515

PBI: Producto Bruto Interno a precios de mercado en miles de pesos a precios de 1993

Consumo: Consumo de los hogares con IVA en miles de pesos a precios de 1993

Interés: Tasa de interés trimestral a plazo fijo entre 30 y 59 días

FUENTE: Ministerio de Economía. República Argentina.

Caso 17.2: Regresión en componentes principales

La tabla 17.5 contiene información sobre 24 meses correspondientes a los gastos de comercialización (Gastos) de una empresa, el nivel de ventas (Ventas), su costo de personal (Personal) y los costos de materias primas (Insumos). El objetivo es estimar el nivel de ventas a partir de las restantes variables.

Primer Paso: Especificación del modelo

$$\text{Ventas} = \beta_1 + \beta_2 \text{Gastos} + \beta_3 \text{Insumos} + \beta_4 \text{Personal} + \mu$$

Segundo Paso: Estimación del Modelo

La tabla se encuentra en el archivo "ventas.xls". Esta información debe importarse en Eviews para realizar la estimación econométrica correspondiente. Los pasos a seguir consisten en

1) Generar en Eviews un archivo de trabajo (*workfile*) a partir de *File-New*, desde la ventana *Workfile frecuency* seleccionar *Undated or irregular dates*, en *End date* consignar la cantidad de observaciones que se tienen (en este caso 24).

2) Importar desde *File-Import-Read Text_Lotus_Excel* ubicando el archivo ventas.xls.

3) En la ventana de importación, en *Upper-left data cell*, consignar la celda donde se encuentra el primer dato. En *Names series or number of series if name in file*, especificar el nombre de las series o el número de series a importar.

Tabla 17.5				
Mes	Ventas	Gastos	Insumo	Personal
1	607	197	110	173
2	590	208	107	152
3	543	181	99	150
4	558	194	102	150
5	571	192	109	163
6	615	196	114	179
7	606	203	113	169
8	593	200	113	166
9	582	198	115	159
10	646	221	119	206
11	619	218	120	181
12	651	213	123	192
13	648	207	122	191
14	694	228	131	217
15	697	249	133	190
16	707	225	135	221
17	693	237	133	189
18	680	236	128	192
19	664	231	134	193
20	747	260	135	233
21	708	254	139	196
22	702	239	138	199
23	711	248	146	202
24	778	273	153	240

4) La estimación se realiza a partir de *Quick-Estimate Equation*, consignando la variable dependiente (*ventas*) seguida de la constante (*c*) y de las variables explicativas (*Gastos*, *Insumos*, *Personal*) de la siguiente manera: *ventas c gastos insumo personal*. Esto da lugar a la siguiente salida:

El modelo estimado es

$$\text{Ventas} = 107.444 + 0.923\,\text{Gastos} + 1.298\,\text{Insumos} + 0.950\,\text{Personal}$$
$$(18.058)\quad (0.223)\qquad (0.431)\qquad (0.156)$$
$$R^2 = 0.98 \qquad F = 323.64 \qquad DW = 1.30$$

donde los valores entre paréntesis indican el desvío estándar de los coeficientes estimados.

Tercer Paso: Análisis de la bondad del ajuste

a) **Nivel de explicación**: El $R^2 = 0.98$ indica que las variaciones del conjunto de variables explicativas determinan el 98% de las variaciones de la variable dependiente.

Estimación 17.2.1
Dependent Variable: VENTAS
Method: Least Squares
Date: 08/25/06 Time: 14:58
Sample: 1 24
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	107.4435	18.05749	5.950079	0.0000
GASTOS	0.922567	0.222733	4.142030	0.0005
INSUMO	1.297786	0.430729	3.012996	0.0069
PERSONAL	0.950177	0.155845	6.096928	0.0000

R-squared	0.979817	Mean dependent var	650.4167
Adjusted R-squared	0.976789	S.D. dependent var	62.39281
S.E. of regression	9.505570	Akaike info criterion	7.492645
Sum squared resid	1807.117	Schwarz criterion	7.688987
Log likelihood	-85.91174	F-statistic	323.6415
Durbin-Watson stat	1.299572	Prob(F-statistic)	0.000000

- **Nivel de significación individual de las variables**: La hipótesis nula es que el coeficiente que acompaña a la variable es nulo, de aceptarse esta hipótesis indica que la variable explicativa no está relacionada con la variable dependiente.

El conjunto de hipótesis a docimar es

$$H_0 = \beta_i = 0$$
$$H_1 = \beta_i \neq 0$$

La distribución teórica de probabilidades a utilizar para este contraste es la distribución t con *(n-k)* grados de libertad, con *k* igual al número de parámetros a estimar. Para un nivel de confianza del 95%, el valor crítico de la distribución t es de ± 2.086. El valor de prueba a utilizar para docimar la significatividad de la variable *Gastos* es:

$$t = \frac{\hat{\beta}_2 - \beta_2}{s_{\beta_2}} = \frac{0.922567 - 0}{0.222733} = 4.14$$

El estadístico empírico cae en la zona de rechazo de la hipótesis nula, (4.14>2.086) se concluye que la variable es significativa en el modelo.

Repitiendo el procedimiento para los demás coeficientes, se concluye que todas las variables son significativas individualmente.

- **Nivel de significación conjunta de las variables**: La hipótesis nula es que los coeficientes que acompañan a las variables son todos nulos, de aceptarse esta hipótesis indica que el conjunto de variables explicativas utilizado no determina el comportamiento de la variable dependiente.

El conjunto de hipótesis a docimar es

$$H_0 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 = \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$$

La distribución teórica de probabilidades a utilizar para este contraste es la distribución F con *k-1* y *n-k* grados de libertad, con *k* igual al número de parámetros a estimar. Para un nivel de confianza del 95%, el valor crítico de la distribución F es de $\pm 3{,}10$. El valor de prueba a utilizar es:

$$F = \frac{SCE/(k-1)}{SCR/(n-k)} = \frac{87728.72601/(4-1)}{1807.117/(24-4)} = 323.6415$$

El estadístico empírico cae en la zona de rechazo de la hipótesis nula, se concluye que el conjunto de variables explicativas determinan la variable dependiente.

¿Cómo se obtienen los componentes del estadístico *F*?

La suma de cuadrados explicada (SCE) es la diferencia entre la suma de cuadrados totales (SCT) y la suma de cuadrados de los residuos (SCR): SCE=SCT-SCR

El desvío de la variable dependiente es

$s_Y = \sqrt{\frac{SCT}{n-1}}$, de modo que $SCT = s_Y^2 * (n-1) = 62.393^2 * 23 = 89535.84301$

la *SCR*=1807.117, con lo cual

SCE=*SCT*-*SCR*=89535.84301-1807.117=87728.72601

- **Multicolinealidad**. El modelo se especifica y estima bajo el supuesto de que las variables explicativas no están relacionadas entre sí. A través del cálculo de la matriz de correlaciones se observa que la asociación estadística entre las

variables es alta. Los gastos de comercialización con respecto a gastos de personal y el costo de materias primas, muestran una correlación elevada 0.82 y 0.93; de igual modo, costo de materias primas y personal muestran una correlación de 0.86. Esta situación indica la existencia de multicolinealidad entre todas las variables

	GASTOS	INSUMO	PERSONAL
GASTOS	1.000000	0.931240	0.820452
INSUMO	0.931240	1.000000	0.857916
PERSONAL	0.820452	0.857916	1.000000

Otra manera de probar la existencia de multicolinealidad es regresionar las variables explicativas entre sí. De modo que la especificación de los modelos a estimar es

$$Gastos = \beta_1 + \beta_2 Insumos + \mu$$
$$Personal = \beta_1 + \beta_2 Insumos + \mu$$
$$Gastos = \beta_1 + \beta_2 Personal + \mu$$

Las respectivas estimaciones arrojan los siguientes resultados

$$Gastos = 20.82 + 1.618 Insumos \quad R^2 = 0.867$$
$$Gastos = 69.30 + 0.81 Personal \quad R^2 = 0.67$$
$$Personal = 0.53 + 1.51 Insumos \quad R^2 = 0.74$$

Los coeficientes de determinación de cada variable explicativa respecto de la otra indica nuevamente la existencia de multicolinealidad. La presencia de multicolinealidad provoca variabilidad en los coeficientes estimados. Para salvar este problema es necesario hallar las componentes principales de las

variables explicativas y estimar las ventas a partir de los factores resultantes.

Cuarto paso: Análisis de Componentes Principales

Con Eviews se realiza el ACP sobre el conjunto de variables explicativas

El primer eje factorial reúne el 91.35% de la varianza (inercia) de las variables explicativas y el primer plano (los dos primeros ejes, es decir, las dos primeras componentes) el 97.84%.

Correlation of GASTOS INSUMO PERSONAL

	Comp 1	Comp 2	Comp 3
Eigenvalue	2.740561	0.194568	0.064872
Variance Prop.	0.913520	0.064856	0.021624
Cumulative Prop.	0.913520	0.978376	1.000000

Eigenvectors:

Variable	Vector 1	Vector 2	Vector 3
GASTOS	-0.580238	-0.514175	-0.631623
INSUMO	-0.588138	-0.271946	0.761669
PERSONAL	-0.563399	0.813430	-0.144613

Ponderadores en la combinación lineal

Para cada observación, los ponderadores en la combinación lineal permiten calcular las coordenadas sobre cada eje factorial, determinando de esta manera las componentes principales.

obs	C1	C2	C3
1	1.506231	0.299516	-0.037624
2	1.853441	-0.576847	-0.366855
3	2.890413	0.090451	-0.086389
4	2.449552	-0.246221	-0.262575
5	1.900217	0.092529	0.097884
6	1.221876	0.442400	0.173072
7	1.326530	-0.020977	-0.006374
8	1.468150	-0.056916	0.090149
9	1.593096	-0.286881	0.294148
10	-0.216850	0.710388	-0.368121
11	0.390405	-0.078791	-0.086172
12	0.129487	0.335922	0.144863
13	0.339721	0.450410	0.253263
14	-1.149357	0.691960	0.043003
15	-1.117047	-0.696328	-0.238097
16	-1.339299	0.810884	0.318048
17	-0.804620	-0.473284	0.082786
18	-0.637431	-0.253687	-0.183776
19	-0.794810	-0.231346	0.271555
20	-2.460332	0.463226	-0.671708
21	-1.631015	-0.720918	-0.074868
22	-1.296164	-0.280755	0.246036
23	-1.922235	-0.530122	0.432073
24	-3.699960	0.065388	-0.064320

Quinto paso: Re especificación del modelo

El modelo inicial que presentaba multicolinealidad se reespecifica. Las ventas, ahora vienen explicadas por las componentes principales $\mathbf{C_1}, \mathbf{C_2}, \mathbf{C_3}$

$$\textbf{Ventas} = \alpha_1 + \alpha_2 \mathbf{C_1} + \alpha_3 \mathbf{C_2} + \alpha_4 \mathbf{C_3} + \mu$$

El resultado de la estimación muestra que la primera componente que reunía el 91.35% de la varianza de las variables exógenas es la que presenta un buen ajuste.

Estimación 17.3.2
Dependent Variable: VENTAS
Method: Least Squares
Date: 08/25/06 Time: 14:56
Sample: 1 24
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	650.4167	1.940316	335.2117	0.0000
C1	-36.51051	1.172068	-31.15050	0.0000
C2	2.551440	4.398832	0.580027	0.5684
C3	-3.687286	7.618078	-0.484018	0.6336

R-squared	0.979817	Mean dependent var	650.4167
Adjusted R-squared	0.976789	S.D. dependent var	62.39281
S.E. of regression	9.505570	Akaike info criterion	7.492645
Sum squared resid	1807.117	Schwarz criterion	7.688987
Log likelihood	-85.91174	F-statistic	323.6415
Durbin-Watson stat	1.299572	Prob(F-statistic)	0.000000

Se reespecifica nuevamente el modelo eliminando la tercera componente y se obtienen los resultados de la estimación 17.3.3.

La segunda componente no presenta un buen ajuste por lo que se reespecifica el modelo

$$\textbf{Ventas} = \alpha_1 + \alpha_2 \mathbf{C_1} + \mu$$

y se realiza la estimación 17.3.4

Estimación 17.3.3
Dependent Variable: VENTAS
Method: Least Squares
Date: 08/25/06 Time: 15:02
Sample: 1 24
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	650.4167	1.904613	341.4955	0.0000
C1	-36.51051	1.150501	-31.73444	0.0000
C2	2.551440	4.317890	0.590900	0.5609

R-squared	0.979580	Mean dependent var	650.4167
Adjusted R-squared	0.977636	S.D. dependent var	62.39281
S.E. of regression	9.330659	Akaike info criterion	7.420957
Sum squared resid	1828.285	Schwarz criterion	7.568214
Log likelihood	-86.05148	F-statistic	503.7120
Durbin-Watson stat	1.390091	Prob(F-statistic)	0.000000

Estimación 17.3.4
Dependent Variable: VENTAS
Method: Least Squares
Date: 08/25/06 Time: 15:03
Sample: 1 24
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	650.4167	1.876229	346.6617	0.0000
C1	-36.51051	1.133355	-32.21453	0.0000

R-squared	0.979241	Mean dependent var	650.4167
Adjusted R-squared	0.978297	S.D. dependent var	62.39281
S.E. of regression	9.191606	Akaike info criterion	7.354114
Sum squared resid	1858.684	Schwarz criterion	7.452285
Log likelihood	-86.24937	F-statistic	1037.776
Durbin-Watson stat	1.427324	Prob(F-statistic)	0.000000

El modelo estimado es: $\mathbf{Ventas} = 650.4167 - 36.51051\mathbf{C_1}$

$\mathbf{C_1}$ es la primer componente principal que se forma al hacer la suma ponderada, por los ponderadores de la combinación lineal, de las variables tipificadas para cada observación, es decir:

$$C_{1i} = -0.58\left(\frac{Gasto_i - \overline{Gasto}}{s_{Gasto}}\right) - 0.56\left(\frac{Personal_i - \overline{Personal}}{s_{Personal}}\right) - 0.59\left(\frac{Insumo_i - \overline{Insumo}}{s_{Insumo}}\right)$$

sustituyendo los respectivos valores de medias y desvíos para las variables

$$C_{1i} = -0.58\left(\frac{Gasto_i - 221.1667}{24.58}\right) - 0.56\left(\frac{Personal_i - 187.625}{24.92}\right) - 0.59\left(\frac{Insumo_i - 123.7917}{14.14}\right)$$

Reemplazando el valor de $\mathbf{C_{1i}}$ en el modelo estimado se tiene

$$V_i = 650.4167 - 36.51205\left[-0.58\left(\frac{G_i - 221.1667}{24.58}\right) - 0.56\left(\frac{P_i - 187.625}{24.92}\right) - 0.59\left(\frac{I_i - 123.7917}{14.14}\right)\right]$$

donde $V_i = \text{Ventas}$, $G_i = \text{Gastos}$, $P_i = \text{Personal}$, $I_i = \text{Insumos}$

Operando matemáticamente

$$V_i = 650.4167 - 36.51205\left[-0.0236G_i + 5.2187 - 0.0225P_i + 4.2163 - 0.0417I_i + 5.1653\right]$$

El modelo definitivo es:

$$V_i = 117.3298 + 0.8617G_i + 0.8215P_i + 1.5225I_i$$

donde se ha eliminado la multicolinealidad

BIBLIOGRAFIA

- Baronio, Alfredo Mario. "Metodología Estadística Para El Estudio Socioeconómico De Regiones," Secretaría de Posgrado. Río Cuarto: Universidad Nacional de Río Cuarto, 2001.
- Caridad, J.M. y Ocerin. Econometría: Modelos Econométricos Y Series Temporales. Barcelona: Editorial Reverté, S.A., 1998.
- Crivisqui, Eduardo. "Iniciación a Los Métodos Estadísticos Exploratorios Multivariados," Seminario de Métodos Estadísticos Multivariados Bruxelles, Belgique.: Laboratorio de Metodología de Tratamiento de datos. Université Libre de Bruxelles. , 2002.
- Gujarati, Damodar. Econometría. México: Mc.Graw Hill, 2004.
- Johnston, J. y Dinardo, J. Métodos De Econometría. Barcelona: Editorial Vicens Vives, 2001.
- Kendall, M. y Stuart, A. The Advanced Theory of Statistics. Londres: Charles Griffin, 1966.
- Lucas, Robert E. Studies in Business-Cycle Theory. Cambridge, Mass. [u.a.]: MIT Pr., 1995.
- Perez Lopez, C. Problemas Resueltos De Econometría. Madrid.: Thomson. , 2006.

  Pulido San Román, Antonio. Modelos Econométricos. Madrid: Editorial Pirámide, 1993.