ASTURIAS
Corporación Universitaria

Multicolinealidad, Heterocedasticidad, Autocorrelación

© Asturias Corporación Universitaria

Índice

Red SUMMA ©

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

Objetivos:

- Conocer los problemas más habituales de la correlación lineal al tratar con datos observacionales.
- Conocer los elementos básicos de las técnicas utilizadas para resolver las limitaciones del método de estimación por Mínimos Cuadrados Ordinarios.

1 Multicolinealidad

"Situación en la que dos o más variables explicativas se parecen mucho y resulta difícil medir sus efectos individuales sobre la variables explicada"

El término **colinealidad** (o **multicolinealidad**) en Econometría se refiere a una situación en la que dos o más variables explicativas se parecen mucho y, por tanto, resulta **difícil medir sus efectos individuales** sobre la variable explicada.

Este fenómeno puede presentarse con frecuencia en un contexto de **series temporales** y con **series macroeconómicas**. **Por ejemplo**, la población y el PIB en general suelen estar altamente correlacionados.

Podemos encontrar:

- **Multicolinealidad exacta**: Se da cuando los valores de una variable explicativa se obtienen como combinación lineal exacta de otras.
- **Multicolinealidad de grado**: Se da cuando los valores de diferentes variables están tan correlacionados que se hace casi imposible estimar con precisión los efectos individuales de cada uno de ellos.

Red SUMMA ©

2 Multicolinealidad Exacta

En el caso de la **multicolinealidad exacta**, el determinante:

$$|X'X| = 0$$

Lo que significa que el **sistema de ecuaciones de los estimadores MCO**,

$$X'X\,\hat{\mathrm{B}} = X'Y$$

tiene **infinitas soluciones**.

2.1 Multicolinealidad de Grado

Cuando dos o más variables explicativas en un modelo están altamente correlacionadas en la muestra, es muy difícil separar el efecto parcial de cada una de estas variables sobre la variable dependiente. La información muestral que

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

incorpora una de estas variables es casi la misma que la del resto de las correlacionadas con ella.

En este caso, el determinante $|X'X| \cong 0$. Matemáticamente, existirá una **solución única al problema de la mínima suma de cuadrados**, pero también existirán **muchas soluciones casi iguales a ella**.

Los **síntomas de este problema** que podemos encontrar son fundamentalmente:

- Las estimaciones de los parámetros MCO son muy sensibles a la muestra: **pequeños cambios en los datos o en la especificación provocan grandes cambios en las estimaciones de los coeficientes**.
- Las **estimaciones de los coeficientes presentan signos distintos a los esperados** o **magnitudes poco razonables**.
- El efecto más pernicioso de la existencia de un alto grado de multicolinealidad es el de **incrementar las varianzas de los coeficientes estimados por MCO**. Como consecuencia, **los test de significatividad de los parámetros individuales no son fiables** (se tiende a concluir que las variables no son significativas individualmente).
- Se obtienen **valores altos del** R^2 aun cuando los valores de los estadísticos t de significatividad individual son bajos. El problema reside en la **identificación del efecto individual de cada variable explicativa, no tanto en su conjunto**. Por eso, si se realiza un contraste de significatividad conjunta de las variables explicativas, se concluirá normalmente que la**s variables son significativas en conjunto, aunque individualmente cada una de ellas no lo sea**.

2.2 Diagnóstico de la Multicolinealidad

"La colinealidad no es problema si nuestro objetivo es predecir, pero es un problema grave si queremos interpretar las estimaciones de los parámetros"

Para decidir si la colinealidad de grado constituye un problema debemos tener en cuenta los objetivos de nuestro análisis concreto. **Por ejemplo**, la colinealidad no nos preocupa demasiado si nuestro objetivo es predecir, pero es un problema muy grave si el análisis se centra en interpretar las estimaciones de los parámetros.

Una primera aproximación para diagnosticarla consiste en obtener los **coeficientes de correlación muestral simples** para cada par de variables explicativas y ver **si el grado de correlación entre estas variables es alto**. Pero se puede dar el caso de tener una relación lineal casi perfecta entre tres o más variables y sin embargo, las correlaciones simples entre pares de variables no ser mayores que 0,5.

Otro método consiste en realizar la **regresión de cada variable explicativa** sobre el resto. Se realizan j regresiones, y se obtienen los coeficientes de determinación R_j^2. Si alguno de ellos es alto, podemos sospechar la existencia de colinealidad.

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

2.3 Soluciones para la Multicolinealidad

El problema de colinealidad se reduce a que la muestra no contiene suficiente información para estimar todos los parámetros. Por ello, resolver el problema requiere añadir nueva información, sea muestral o extramuestral, o cambiar la especificación. Algunas posibles soluciones en esta línea son:

- **Añadir nuevas observaciones**. Si realmente es un problema muestral, una posibilidad es cambiar de muestra porque puede ser que con nuevos datos el problema se resuelva, aunque esto no siempre ocurre. La idea consiste en conseguir datos menos correlacionados que los anteriores, bien cambiando toda la muestra o simplemente incorporando más datos en la muestra inicial. No siempre resulta fácil obtener mejores datos por lo que muy probablemente debamos convivir con el problema teniendo cuidado con la inferencia realizada y las conclusiones de la misma.
- **Restringir parámetros**. Si la Teoría Económica o la experiencia sugieren algunas restricciones sobre los parámetros más afectados por la colinealidad, imponerlas permitirá reducir el problema. Obviamente, se corre el riesgo de imponer restricciones que no son ciertas.
- **Suprimir variables**. Si se suprimen variables que están correlacionadas con otras, la pérdida de capacidad explicativa será pequeña y la colinealidad se reducirá. Esta medida puede provocar otro tipo de problemas, ya que si la variable que eliminamos del modelo realmente sí es significativa, estaremos omitiendo una variable relevante, lo que hará que los estimadores de los coeficientes del modelo y de su varianza sean sesgados por lo que la inferencia realizada no sería válida.
- **Transformar las variables del modelo**. Si la colinealidad se debe a que se están relacionando series temporales con tendencia, puede convenir transformar las variables para eliminar esta tendencia.

Red SUMMA®

3 Heterocedasticidad

Uno de los supuestos del modelo de regresión lineal es la **homocedasticidad de la perturbación aleatoria**, es decir, que **todos los términos de la perturbación se distribuyen de la misma forma alrededor de la recta de regresión**: **tienen la misma varianza** (varianza constante):

$$Var[u_i] = Var[u_j] = \sigma^2, \forall \mathrm{i} \neq \mathrm{j}$$

"Se da cuando la dispersión de los términos de perturbación es diferente para diferentes valores de la variable explicativa"

Cuando no se cumple esta condición, es decir, cuando la dispersión de los términos de perturbación es diferente para diferentes valores de la variable explicativa, nos encontramos con la **heterocedasticidad**.

La heterocedasticidad tiene importantes consecuencias en el **método de estimación MCO**. Los estimadores de los coeficientes siguen siendo insesgados,

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

pero la estimación de los errores estándar de esos parámetros no es válida (que denotaremos **SE estimado**). Por esta razón, no podemos construir intervalos de confianza ni hacer pruebas de hipótesis correctas, pues para hacerlos se utiliza el error estándar.

Tratar con la heterocedasticidad no es fácil, porque puede seguir muchos patrones diferentes. **Un Ejemplo**, es el caso en el que la dispersión de los términos de perturbación alrededor de la línea de regresión va creciendo a medida que el valor de la variable explicativa X crece, como se muestra en el siguiente gráfico:

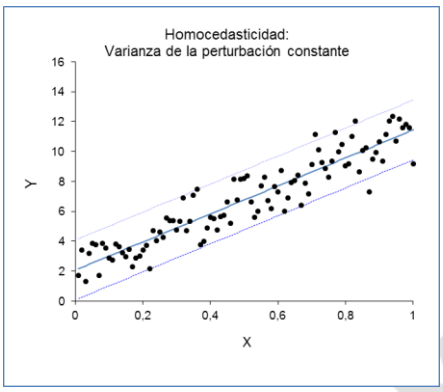

Figura 5. 1

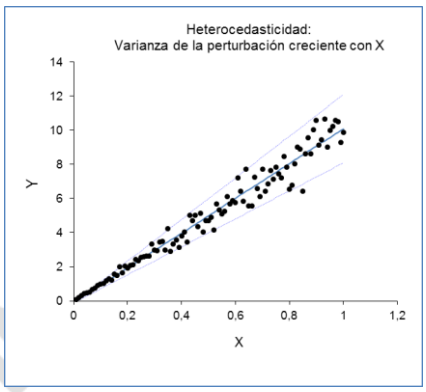

Figura 5. 2

Ejemplo:

Supongamos que tenemos datos del nivel de renta y el gasto en alimentación para un número grande de familias. Si representamos en un gráfico el gasto en alimentación frente a la renta, es de esperar que encontremos heterocedasticidad, ya que, probablemente, la dispersión en el gasto en alimentación para diferentes niveles de renta aumente con la renta.

- Las familias con nivel de renta bajo tienen menos flexibilidad en su nivel de gasto en alimentación, de manera que veremos poca dispersión en el gasto de alimentación para dichas familias.
- Por el contrario, entre las familias con nivel de renta alto, encontraremos unas que gasten mucho en alimentación y otras con preferencias diferentes que gasten menos en alimentación, destinando su renta a otros usos.

4 Tratamiento de la Heterocedasticidad

Una solución utilizada habitualmente para resolver el problema de la heterocedasticidad consiste en utilizar los estimadores calculados mediante el **método de mínimos cuadrados ordinarios (MCO)**, pero no sus Errores Estándar

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

(SE), sino en su lugar los llamados **Errores Estándar Robustos** (o errores estándar de Eicker-White, que denotaremos **RSE**). Esta técnica tiene la ventaja de que puede aplicarse sin necesidad de conocer el patrón concreto que sigue la heterocedasticidad en cada caso.

Los **RSE** son **estimadores de los errores estándar de los coeficientes estimados** que tienen en cuenta la heterocedasticidad de la muestra de datos, de forma que pueden utilizarse para realizar inferencia estadística inmune a la heterocedasticidad. Lo vemos con más detalle en el caso del modelo econométrico de dos variables:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

En el que cada término de perturbación aleatoria u_i tiene una desviación típica $SD(u_i)$.

El cálculo del estimador MCO de la pendiente $\hat{\beta}_1$ se puede expresar como una suma ponderada:

$$\hat{\beta}_1 = \sum_{i=1}^{n} peso_i \, y_i$$

Donde cada peso viene dado por la fórmula:

$$peso_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Esta fórmula dice que el peso de cada observación es la desviación del valor de x correspondiente a esa observación respecto de la media de los valores de x, dividida por la varianza de x (recordamos que $Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$) por el número de observaciones).

Dado que asumimos que todos los u_i son independientes, y no están correlacionados con los valores de x, que son fijos en todas las muestras, podemos calcular la varianza (que es el cuadrado del Error Estándar) de $\hat{\beta}_1$, como:

$$Var(\hat{\beta}_1) = \sum_{i=1}^{n} peso_i^2 \cdot SD\,(u_i)^2$$

Y el Error Estándar del estimador $\hat{\beta}_1$ es la raíz cuadrada:

$$SE(\hat{\beta}_1) = \sqrt{\sum_{i=1}^{n} peso_i^2 \cdot SD\,(u_i)^2}$$

Red SUMMA ©

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

En caso de homocedasticidad, la $SD(u_i)$ es la misma en todas las observaciones, y podemos expresarla simplificadamente $SD(u)$.

En el método MCO, se utiliza el valor llamado **RMSE** (raíz del error cuadrático medio, o *Root-mean-square error*, también llamado **error estándar de la regresión**) para medir la dispersión de los datos observados respecto de la línea de regresión:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n-k-1}}$$

En otras palabras, el **RMSE** es una **medida del tamaño típico de los residuos**, y se utiliza como estimador de la desviación estándar $SD(u_i)$, ya que el verdadero valor de la SD es desconocido.

Así pues, el RMSE nos sirve para calcular un estimador del SE de los coeficientes:

$$SEestimado(\hat{\beta}_1) = \sqrt{\sum_{i=1}^{n} peso_i^2 \cdot RMSE^2}$$

Nuestro problema es que el RMSE es un valor fijo, y si hay heterocedasticidad, no podemos tomar un valor fijo para estimar la $SD(u_i)$, pues precisamente en cada observación la desviación será diferente, y la estimación que resulte no será válida.

Para estimar el Error Estándar Robusto, no utilizaremos como estimador de cada $SD(u_i)$ un valor del residuo típico, sino el residuo $\hat{u}_i$ correspondiente a cada observación, y de esta forma sí tenemos en cuenta las diferencias:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$RSEestimado(\hat{\beta}_1) = \sqrt{\sum_{i=1}^{n} peso_i^2 \cdot \hat{u}_i{}^2}$$

4.1 Contraste de Significatividad Individual de los Parámetros

Una vez que tenemos estimadores del Error Estándar de los parámetros que son inmunes a la heterocedasticidad (el RSE), se puede demostrar (la demostración queda fuera del alcance de este curso) que el estadístico definido como:

$$t = \frac{\hat{\beta}_1 - \beta_1}{RSEestimado(\hat{\beta}_1)}$$

Real SUMMA ©

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

tiene una distribución t-student, que tiende a una distribución normal al crecer el tamaño de la muestra. Por tanto, podemos usarlo como **estadístico de contraste** para verificar la hipótesis de que el valor del parámetro estimado $\hat{\beta}_1$ sea un cierto dado β_1 para un nivel de confianza dado.

Una de las pruebas más comunes es verificar la hipótesis de partida (H_0) de que el parámetro β_1 sea igual a cero (y si es así, entonces la variable x correspondiente no interviene para explicar y, por lo que ha de ser eliminada del modelo), o bien sea diferente de cero (y entonces se descarta la hipótesis de partida H_0, y se concluye que sí es una variable significativa para explicar y).

El procedimiento consiste en calcular el estadístico de prueba:

$$t_0 = \frac{\hat{\beta}_1 - 0}{RSEestimado(\hat{\beta}_1)}$$

y compararlo con el valor en las tablas de la distribución t-student para un nivel de significatividad (α) dado (por ejemplo $\alpha = 0{,}05$, que corresponde a un nivel de confianza del 95%), y para $n - k$ grados de libertad (siendo n el tamaño de la muestra y k el número de variables explicativas).

Si $|t_0| > t_{n-k}^{\alpha/2}$, entonces rechazo la hipótesis H_0, y la acepto en caso contrario. Nótese, que hemos tomado el valor en tablas correspondiente a $\alpha/2$, porque el valor puede estar fuera de la zona de aceptación por la izquierda o por la derecha, como se muestra en la figura:

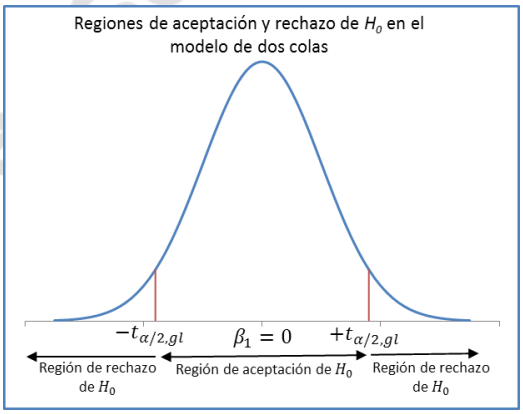

Figura 5. 3

Por otro lado, es importante advertir que en presencia de heterocedasticidad no tiene sentido considerar el error estándar de la regresión (RMSE) o el R^2 como medidas de bondad del ajuste.

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

5 Autocorrelación

Hay situaciones, como con frecuencia ocurre al tratar con datos de series temporales, en las que no se cumple el supuesto del modelo de que los términos de perturbación aleatoria son independientes unos de otros. Al contrario, hay una **correlación entre la perturbación de un período y la del período anterior** (denotaremos el período con el subíndice t).

Ejemplo:

Supongamos que tenemos datos anuales de la cantidad de cigarrillos demandada (C) y su precio en ese año, en el período 1960 – 1990, según el modelo:

$$C_t = \beta_0 + \beta_1 Precio_t + u_t, \qquad t = 1960, 1961, ..., 1990$$

Los factores que influyen en la demanda recogidos en u_t, como la moda de fumar o el gasto en publicidad cambian lentamente, de forma que los de 1960 serán similares a los de 1961, y los de 1985 a los de 1986. Si esto es cierto, y esos factores influyen significativamente en la demanda de cigarrillos, entonces los términos u_t no serán independientes entre sí.

"La autorregresión se refiere a un modelo de regresión en el que hay autocorrelación"

Se utiliza el término **Autorregresión** para referirse a un modelo de regresión en el que hay autocorrelación, esto es, un modelo en el que una variable se expresa en términos de ella misma. El **orden** de la autorregresión indica el número de observaciones utilizadas:

Autorregresión de primer orden: $Z_t = \beta_0 + \beta_1 Z_{t-1} + u_t$

Autorregresión de segundo orden: $Z_t = \beta_0 + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + u_t$

Nos centraremos en el modelo de autorregresión de primer orden (AR1), en el que están correlacionadas las perturbaciones aleatorias. El modelo AR1 más simple se puede expresar con las dos ecuaciones:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

siendo $u_t = \rho\, u_{t-1} + v_t, \quad \text{con } -1 < \rho < 1$

En las que v_t son términos aleatorios independientes, y podemos interpretar ρ como un coeficiente de correlación entre el término de perturbación u de un período y el del período anterior. Así expresamos el significado de la autocorrelación: que los valores actuales están influidos por los valores pasados.

Red SUMA ©

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

6 Consecuencias de la Autocorrelación

Si tenemos **autocorrelación de la perturbación aleatoria**:

- Los **estimadores MCO** siguen siendo **insesgados** (su valor medio esperado sigue siendo el verdadero valor del parámetro).
- Los **errores estándar de los estimadores MCO** son **inconsistentes** (si la correlación es positiva, se estimarán sistemáticamente demasiado bajos, y a la inversa): Este sesgo no se soluciona con tamaños muestrales mayores.
- Los **estimadores MCO ya no serán los mejores posibles**. Veremos que se pueden encontrar otros con menor error estándar.

En definitiva, la **inferencia estadística se ve afectada**, y los **contrastes de hipótesis**, al depender de los errores estándar de los estimadores, **no serán fiables**, con lo que las pruebas pierden validez.

7 Mínimos Cuadrados Generalizados (MCG)

Podemos expresar el modelo AR1 con una sola ecuación, sustituyendo la expresión de la perturbación en la primera ecuación:

$$y_t = \beta_0 + \beta_1 x_t + \rho\, u_{t-1} + v_t$$

Para resolver los problemas que causa la autocorrelación, deseamos eliminar el término $\rho\, u_{t-1}$, de forma que nos quedemos solo con v_t, y así poder volver a contar con la validez de los supuestos del modelo econométrico clásico.

Todos los y_t se estiman con la misma ecuación, de modo que el valor de la variable en el instante t-1 es:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + u_{t-1}$$

Que multiplicada por ρ, es:

$$\rho\, y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho u_{t-1}$$

Y, restando esta ecuación de la primera:

$$y_t - \rho\, y_{t-1} = \beta_0(1-\rho) + \beta_1 x_t - \rho\beta_1 x_{t-1} + v_t$$

Definimos ahora las nuevas variables y_t^* y x_t^* de la siguiente forma:

$$y_t^* = y_t - \rho\, y_{t-1}$$

$$x_t^* = x_t - \rho\, x_{t-1}$$

RESUMMA ©

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

Y obtenemos finalmente la ecuación:

$$y_t^* = \beta_0(1-\rho) + \beta_1 x_t^* + v_t$$

La estimación de mínimos cuadrados sobre este nuevo modelo transformado se conoce como **Mínimos Cuadrados Generalizados**. El nombre viene del hecho de que es un caso más general, del que el MCO es un caso particular para $\rho = 0$.

En la práctica, como no conocemos el valor de ρ, tendremos también que estimarlo, lo que da lugar al método de **Mínimos Cuadrados Generalizados Factible (MCGF)**. Hay diversas formas de estimar ρ, por lo que resultan varios estimadores conocidos como MCFG.

Una de las formas de estimar ρ es la de Prais y Winsten. En este método, calculamos la regresión lineal auxiliar de los residuos en el periodo t sobre los residuos en el periodo t-1 en el origen (es decir sin término independiente).

El estimador $\hat{\rho}$ de ρ será la pendiente de esta regresión auxiliar. Con este valor de $\hat{\rho}$ podemos calcular el modelo transformado, y hallar los estimadores de los parámetros $\hat{\beta}$ (que ya hemos visto que son los mismos que los obtenidos por MCO) y sus errores estándar correspondientes (que son necesarios para poder realizar inferencias y contrastes válidos). Una vez que tenemos estos resultados, ya no necesitamos utilizar los datos transformados x_t^* e y_t^*, sino los originales.

Red SUMMA ©

8 Resumen

- El término colinealidad (o multicolinealidad) en Econometría se refiere a una situación en la que dos o más variables explicativas se parecen mucho y, por tanto, resulta difícil medir sus efectos individuales sobre la variable explicada.
- Uno de los supuestos del modelo de regresión lineal es la homocedasticidad de la perturbación aleatoria, es decir, que todos los términos de la perturbación se distribuyen de la misma forma alrededor de la recta de regresión: tienen la misma varianza.
- Hay situaciones, como con frecuencia ocurre al tratar con datos de series temporales, en las que no se cumple el supuesto del modelo de que los términos de perturbación aleatoria son independientes unos de otros. Al contrario, hay una correlación entre la perturbación de un período y la del período anterior (denotaremos el período con el subíndice t).

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.

9 Bibliografía

- Goldberger, A.S.: *Introducción a la econometría*, Barcelona, Ariel, 2001.
- Wooldridge, F.M: *Introducción a la econometría: un enfoque moderno*, Madrid : Thomson, 2006

Red SUMMA ©

ASTURIAS CORPORACIÓN UNIVERSITARIA®
Nota Técnica preparada por Asturias Corporación Universitaria. Su difusión, reproducción o uso total o parcial para cualquier otro propósito queda prohibida. Todos los derechos reservados.