RESUMEN ECONOMETRÍA

Resumen basado en Stock & Watson y Wooldridge

Alexis Deppeler y Yamila Sonder
Editado por Nahuel Saucedo en 2025
Transcripción en LaTeX

2024 original, 2025 edición

Índice general

Capítulo 1

Cuestiones Económicas y Datos

La econometría es, desde un punto de vista amplio, la ciencia y el arte de utilizar la teoría económica y las técnicas estadísticas para analizar los datos económicos.

1.1 Preguntas económicas a examen

Las preguntas requieren de una respuesta númerica. La teoría económica proporciona las claves sobre la respuesta , pero el valor númerico efectivo debe averiguarse empíricamente mediante el análisis de los datos, el metodo conceptual utilizado en este libro es el modelo de regresión múltiple, este se utiliza para aislar el efecto del cambio de una variable respecto al cambio en los demás factores.

1.2 Efectos causales y experimentos ideales

Causalidad: La causalidad implica que una acción específica conlleva una consecuencia específica, medible. Para medir los efectos causales se realizan experimentos aleatorios.

Experimento aleatorios: Un experimento aleatorizado controlado se controla porque existe un grupo de control, que no recibe tratamiento, y otro de tratamiento que si lo recibe. Se aleatoriza porque el tratamiento se asigna aleatoriamente. Esto elimina la posibilidad de una relación sistemática con otra variable, por lo que la única diferencia sistemática entre los grupos es el tratamiento. El efecto causal es el efecto sobre un resultado de una acción dada o tratamiento medido en un experimento aleatorizado controlado ideal.

Cuasi experimentos: En la práctica no es posible llevar a cabo experimentos ideales. Y son escasos los experimentos en econometría porque son inmorales, imposibles de ejecutar satisfactoriamente o son prohibitivamente caros. Aunque si se pueden realizar cuasiexperimentos.

1.3 Tipos de datos

- **Datos experimentales y observacionales:** Los datos experimentales provienen de experimentos diseñados para evaluar un tratamiento o política o investigar un efecto causal. Los datos observacionales (más habituales) se obtienen mediante la obser- vación del comportamiento real fuera de un marco experimental. Estos se recopilan utilizando encuestas y registros administrativos. Este tipo de datos posee el desafío de que los niveles de tratamiento no fueron asignados aleatoriamente.
- **Datos de sección cruzada:** datos de individuos o entidades diferentes para un único periodo de tiempo. El número de observaciones es un número asignado arbitrariamente que sirve para organizar los datos. Con este tipo de datos se puede aprender sobre las relaciones entre variables estudiando las diferencias entre personas, empresas u otras entidades económicas durante un único periodo de tiempo.

- **Datos de series temporales:** son datos para un único individuo o entidad recogidos para múlti- ples periodos. El número de observaciones en estos casos se expresa como observaciones. Estos datos pueden utilizarse para estudiar la evolución de las variables en el tiempo y predecir valores futuros de esas variables.
- **Datos de panel:** o datos longitudinales, son datos sobre varios individuos en los que cada individuo se observa durante uno, dos o más periodos de tiempo. El número de individuos es y el número de periodos es . Por lo que tendremos observaciones.

Capítulo 2

Repaso de Probabilidad

2.1 Variables aleatorias y distribuciónes de probabilidad

Probabilidades, espacio muestral y variables aleatorias

Los resultados potenciales mutuamente excluyentes de un proceso aleatoria se denominan **resultados**. La **probabilidad** de un resultado es la proporción de veces que el resultado ocurre en el largo plazo. El conjunto de todos los posibles resultados se denomina **espacio muestral**. Un **suceso** es un subconjunto del espacio muestral. Una **variable aleatoria** es un resumen númerico de un resultado aleatorio, las discretas toman balores sobre un conjunto discreto y las continuas sobre un continuo de posibles valores.

Distribución de probabilidad de una variable aleatoria discreta

La distribución de probabilida de una variable aleatoria discreta es una relación de todos los valores posibles de la variable junto con la probabilidad de que ocurra cada valor. Esas probabilidades suman 1.

La probabilidad de un suceso puede calcularse a partir de la distribución de probabilidad. La distribución de probabilidad acumulada es la probabilidad de que la variable aleatoria sea menor o igual a un valor concreto, esta distribución de probabilidad se conoce además como función de distribución acumulada f.d.a, o distribución acumulada.

Un caso particular es cuando la variable aleatoria es binaria, en este caso se denomina variable aleatoria Bernouilli

Distribución de probabilidad de una variable aleatoria continua

Debido a que una variable aleatoria continua puede tomar sus valores posibles en un continuo, la probabilidad es recogida por la función de densidad de probabilidad. Una función de densidad de probabilidad se denomina asimismo como f.d.p., función de densidad, o simplemente densidad.

2.2 Esperanza, media y varianza

La esperanza de una variable aleatoria Y se define como el valor medio de largo plazo de la variable aleatoria, para las variables aleatorias discretas se calcula como la media ponderada de los posibles resultados, donde las ponderaciones son las probabilidades de esos resultados, para las continuas se toma la integral.

$$E(Y) = \mu_Y = \sum_{i=1}^{k} y_i p_i$$

$$E(Y) = \mu_Y = \int y f_Y(y) dy$$

La varianza y la desviación típica miden la dispersión o «difusión» de una distribución de probabilidad. La varianza de una variable aleatoria Y (que viene expresada por $var(Y)$, es el valor esperado del cuadrado de la desviación de Y respecto de su media. A causa de que la varianza incluye el cuadrado de Y, las unidades de la varianza son las unidades de Y al cuadrado.

$$\sigma_Y^2 = var(Y) = E[(Y-\mu_Y)^2] = \sum_{i=1}^{k}(y_i-\mu_Y)^2 p_i \tag{2.1}$$

- **Asimetría:** roporciona un método matemático para describir cuánto se desvía una distribución de la simetría. si la distribución es simétrica se neutralizan los valores positivos y negativos y por ende el numerador. La asimetría es el tercer momento.
$$\text{Asimetría} = \frac{E[(Y-\mu_Y)^3]}{\sigma_Y^3}$$
- **Curtosis:** es una medida de cuánta masa probabilística se encuentra en sus colas, por tanto, es una medida de cuánta varianza de Y proviene de los valores extremos. Un valor extremo de Y se denomina atípico (outlier). Cuanto mayor es la curtosis de una distribución, más probables son los atípicos.
$$\text{Curtosis} = \frac{E[(Y-\mu_Y)^4]}{\sigma_Y^4}$$
- **Momentos:** La media de Y se denomina el momento primero de Y, el valor esperado del cuadrado de Y se denomina el momento segundo de Y y así con los demás.

2.3 Dos variables aleatorias

La **Distribución de probabilida conjunta** de dos variables aleatorias discretas, X e Y , es la probabilidad de que las dos variables aleatorias tomen valores concretos de forma simultánea x e y (probabilidad de que pasen ambos eventos al mismo tiempo). Las probabilidades de todas las posibles combinaciones (x, y) suman 1. La distribución de probabilidad puede describirse como la función $Pr(X = x, Y = y)$.

La **Distribución marginal** de una variable aleatoria Y es solo otro nombre para su distribución de probabilidad. Este término se utiliza para distinguir la distribución de Y en solitario (la distribución marginal) de la distribución conjunta de Y con otra variable aleatoria. La distribución marginal de Y puede calcularse a partir de la distribución conjunta de X e Y sumando todas las probabilidades de todos los resultados posibles para los cuales Y toma un valor particular. Si X puede tomar i diferentes valores $x1, x2, ..., xi$, entonces la probabilidad marginal de que Y tome el valor y es:

$$Pr(Y = y) = \sum_{i=1}^{k} Pr(X = x_i, Y = y)$$

(x toma distintos valores e y queda fija)

La **Distribución condicional** de una variable aleatoria Y a que otra variable aleatoria X tome un valor específico se denomina distribución condicional de Y dado X. La probabilidad condicional de que Y tome el valor y cuando X toma el valor x se expresa como $Pr(Y = y|X = x)$. En general, la distribución condicional de Y dado $X = x$ es:

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$$

La **esperanza condicional** de Y dado X, asimismo denominada media condicional de Y dado X, es la media de la distribución condicional de Y dado X. Es decir, la esperanza condicional es el valor esperado de Y, calculado mediante la distribución condicional de Y dado X. Si Y toma k valores $y1, y2, ..., yk$, entonces la media condicional de Y dado $X = x$ es:

$$E(Y|X = x) = \sum_{i=1}^{k} y_i Pr(Y = y_i|X = x)$$

La **Ley de esperanzas iteradas:** La media de Y es la madia ponderada de la esperanza condicional de Y dado X, ponderada por la distribución de probabilidad de X, si X toma l valores $x1, ..., xl$, entonces:

$$E(Y) = \sum_{i=1}^{l} E(Y|X = x_i)\Pr(X = x_i)$$

Expresado de otro modo, la esperanza de Y es la esperanza de la esperanza condicional de Y dado X, donde la esperanza interior se calcula utilizando la distribución condicional de Y dado X y la esperanza exterior se calcula utilizando la distribución marginal de X.

$$E(Y) = E[E(Y|X)]$$

Por ejemplo, el número medio de averías M; es la media ponderada de la esperanza condicional de M dado que es viejo y la esperanza condicional de M dado que es nuevo, por lo que $E(M) = E(M \mid A = 0) \times Pr(A = 0) + E(M \mid A = 1) \times Pr(A = 1) = 0,56 \times 0,50 + 0,14 \times 0,50 = 0,35$. Esta es la media de la distribución marginal de M.

La ley de esperanzas iteradas implica que si la media condicional de Y dado X es cero, entonces la media de Y es cero

La **varianza de** Y **condicionada a** X es la varianza de la distribución condicional de Y dado X, es decir:

$$var(Y|X = x) = \sum_{i=1}^{k} [y_i - E(Y|X = x)]^2 Pr(Y = y_i|X = x)$$

Dos variables aleatorias X e Y están **independientemente distribuidas**, o son independientes, si el conocimiento del valor de una de las variables no proporciona información sobre la otra. En concreto, X e Y son independientes si la distribución condicional de Y dado X es igual a la distribución marginal de Y. Es decir, X e Y están independientemente distribuidas si, para todos los valores de x e y, $Pr(Y = y|X = x) = Pr(Y = y)$ entonces $Pr(X = x, Y = y) = Pr(X = x) \cdot Pr(Y = y)$. Es decir, la distribución conjunta de dos variables aleatorias independientes es el producto de sus distribuciones marginales.

Covarianza y Correlación

La **covarianza** entre dos variables X e Y es una medida del grado al que dos variables aleatorias evolucionan conjuntamente. La covarianza solo mide relaciones lineales. Es un subconjunto de dependencia, por lo que podemos tener $cov = 0$ pero dependencia (relación no lineal)

$$cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

Esto es igual a:

$$\sum_{i=1}^{k} \sum_{j=1}^{l} (x_j - \mu_X)(y_i - \mu_Y) Pr(X = x_j, Y = y_i)$$

Para interpretar esta fórmula, supongamos que cuando X es mayor que su media (por tanto, $X - \mu$ es positivo), entonces Y tiende a ser mayor que su media (por lo que $Y - \mu$ es positivo), y cuando X es menor que su media, entonces Y tiende a ser menor que su media. En ambos casos, el producto $(X - \mu)(Y - \mu)$ tiende a ser positivo, por lo que la covarianza es positiva. Si X e Y tienden a evolucionar en sentido opuesto (si X es grande cuando Y es pequeña, y viceversa), la covarianza es negativa. Finalmente, si X e Y son independientes, entonces la covarianza es cero.

Correlación: El coeficiente de correlación:Dado que la covarianza consiste en multiplicar las desviaciones de X y de Y , podríamos estar multipli- cando unidades distintas, lo que complica su interpretación. La correlación es una medida alternativa de la dependencia entre X e Y que resuelve el problema de las «unidades» de la covarianza. En concreto, la correlación entre X e Y es la covarianza entre X e Y dividida por sus desviaciones típicas.

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Al ser las unidades del numerador las mismas que las del denominador, las unidades se cancelan y la correlación no tiene unidades. Las variables aleatorias X e Y se dice que están incorrelacionadas si la correlación es igual a cero. La correlación siempre toma valores entre -1 y 1. (la covarianza no está acotada a la inequalidad, mientras que la correlación solamente es lineal)

Si la media condicional de Y no depende de X, entonces la correlación es cero. Aunque podría darse que a la inversa si haya correlación.

La media y la varianza de la suma de variables aleatorias

Concepto clave 2.3: Medias, varianzas y covarianzas de la suma de variables aleatorias

$$
\begin{aligned}
E(a+bX+cY) &= a+b\mu_X+c\mu_Y \\
var(a+bY) &= b^2\sigma_Y^2 \\
var(aX+bY) &= a^2\sigma_X^2+2ab\sigma_{XY}+b^2\sigma_Y^2 \\
E(Y^2) &= \sigma_Y^2+\mu_Y^2 \\
cov(a+bX+cV,Y) &= b\sigma_{XY}+c\sigma_{VY} \\
E(XY) &= \sigma_{XY}+\mu_X\mu_Y \\
|corr(X,Y)| \le 1 \quad \text{y} \quad |\sigma_{XY}| &\le \sqrt{\sigma_X^2\sigma_Y^2} \quad \text{(desigualdad de Cauchy-Schwarz)}
\end{aligned}
$$

La media de la suma de dos variables aleatorias, X e Y , es la suma de sus medias:

$$E(X+Y) = E(X)+E(Y) = \mu_X+\mu_Y \tag{2.2}$$

La varianza de la suma de X e Y es la suma de sus varianzas más dos veces su covarianza:

$$\text{var}(X+Y) = \text{var}(X)+\text{var}(Y)+2\text{cov}(X,Y) = \sigma_X^2+\sigma_Y^2+2\sigma_{XY}$$

y si son independientes solo será la suma de sus varianzas.

2.4 Las distribucioens normal, chi-cuadrado, t de Student y F

Distribución normal: es acampanada, y su función de densidad tiene media μ y varianza σ^2, es simétrica respecto de su media y tiene el 95 % de su probabilidad entre $\mu \pm 1.96\,\sigma$. La distribución normal estándar es una distribución normal con media 0, varianza 1 y curtosis 3. Para buscar probabilidades de una variable normal con cualquier media μ y varianza, debemos estandarizar la variable restando primero la media, y posteriormente dividiendo el resultado por la desviación típica. La distribución normal puede generalizarse para describir la distribución conjunta de un conjunto de variables. En este caso, la distribución se denomina distribución normal multivariante, o bien si solo se están considerando dos variables, distribución normal bivariante.

La distribución normal multivariante: presenta cuatro propiedades importantes. Si X e Y presentan una distribución normal bivariante, entonces aX + bY posee distribución normal:

$aX+bY$ está distribuida $N(\mu_X+b\mu_Y, a^2\sigma_X^2+b^2\sigma_Y^2+2ab\sigma_{XY})$. (X; Y normal bivariante)

- Si n variables aleatorias presentan una distribución normal multivariante, entonces cualquier com- binación lineal de esas variables se distribuye normalmente.
- Si un conjunto de variables posee una distribución normal multivariante, entonces la distribución marginal de cada una de esas variables es normal.
- Si las variables que presentan una función distribución normal multivariante tienen covarianzas iguales a cero, entonces las variables son independientes.

- Si X e Y presentan una distribución normal multivariante, entonces la esperanza condicional de Y dado X es lineal en X, es decir $E(Y|X) = a + bX$. La normalidad conjunta implica linealidad de las esperanzas condicionales, pero la linealidad de las esperanzas condicionales no implica normalidad conjunta.

Distribución Chi-Cuadrado: es la distribución de la suma de m variables aleatorias normales estándar independientes al cuadrado (al estar al cuadrado no hay valores negativos). Esta distribución depende de m grados de libertad.

Distribución t de Student: Con m grados de libertad se define como la distribución del cociente entre una variable aleatoria normal estándar y la raíz cuadrada de una variable aleatoria chi-cuadrado independientemente distribuida con m grados de libertad dividida por m. Es decir, sea Z una variable aleatoria normal estándar, sea W una variable aleatoria con distribución chi-cuadrado con m grados de libertad, y sean Z y W independientemente distribuidas. Entonces la variable aleatoria $\frac{Z}{\sqrt{W/m}}$ presenta una distribución t de Student con m grados de libertad.

Distribución F: Con m y n grados de libertad, se define como la distribución del cociente entre una variable aleatoria chi-cuadrado con m grados de libertad, dividida por m, y una variable aleatoria chi-cuadrado independientemente distribuida con n grados de libertad, dividida por n. $\frac{W/m}{V/n}$ posee una distribución F con m grados de libertad en el numerador y con n grados de libertad en el denominador.

2.5 Muestreo aleatorio y distribución de la media muestral

En el **muestreo aleatorio simple** se seleccionan aleatoriamente n objetos a partir de una población. Las n observaciones de la muestra se expresan mediante Y1, Y2, . . . , Yn. El hecho de seleccionar n mues- tras aleatoriamente significa que Y1, Y2, . . . , Yn pueden ser tratados como variables aleatorias. Todos tienen la misma probabilidad de salir y son independientes. Antes del muestreo, Y1, Y2, . . . , Yn pueden tomar muchos valores diferentes; tras haber sido seleccionados, se registra un valor específico para cada observación.

Al ser Y1, Y2, . . . , Yn extracciones aleatorias de una misma población, la distribución marginal de Yi es la misma distribución de la población para cada i, entonces se dice que Y1, Y2, . . . , Yn están **identicamente distribuidas**.

La **media muestral** o promedio muestral es:

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

La media de la media muestral de una muestra aleatoria produce el efecto de hacer que la media muestral $\bar{Y}$ sea una variable aleatoria. Si $\bar{Y}$ presenta una distribución de probabilidad, la distribución muestral, que está asociada a los posibles valores de $\bar{Y}$ que podrían obtenerse para diferentes muestras.

La media de $\bar{Y}$ es igual a:

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y$$

Su varianza es:

$$\begin{aligned}\text{var}(\bar{Y}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq 1}^{n}\text{cov}(Y_i, Y_j) \\ &= \frac{\sigma_Y^2}{n}\end{aligned}$$

Y su desviación típica es la raíz cuadrada de la varianza:

$$\frac{\sigma_y}{\sqrt{n}}$$

2.6 Aproximación para muestras grandes de las distribuciones muestrales

Existen dos métodos para la caracterización de las distribuciones muestrales: el método exacto y el método aproximado.

Método exacto: conlleva la obtención de una fórmula para la distribución muestral que se cumpla con exactitud para cualquier valor de n. La distribución muestral que se describe exactamente la distribución de $\bar{Y}$ para cualquier n se llama la distribución exacta o distribución en muestras finitas de $\bar{Y}$. Desafor- tunadamente, si la distribución de Y no es normal, en general, la distribución muestral exacta de $\bar{Y}$ será muy complicada y dependerá de la distribución de Y.

Método aproximado: utiliza aproximaciones para la distribución muestral que se basan en el hecho de que el tamaño muestral es grande. La aproximación para muestras grandes de la distribución muestral a menudo se denomina distribución asintótica. Estas aproximaciones pueden ser muy precisas incluso si el tamaño muestral es solamente de n = 30 observaciones.

Para aproximar distribuciones muestrales cuando n es grande se utiliza la ley de los grandes números y el teorema del límite central.

Ley de los grandes números: dice que, cuando el tamaño muestral es elevado, $\bar{Y}$ estará cerca de $\mu_Y$con probabilidad muy elevada. Esto a veces se denomina la ley de promedios. Cuando un número elevado de variables aleatorias con la misma media se promedian conjuntamente, los valores más cercanos a la media muestral están cerca de la media común.

La propiedad de que $\bar{Y}$ esté cerca de μ_Y con probabilidad creciente cuando n aumenta se denomina convergencia en probabilidad o consistencia. La ley de los grandes números establece que, bajo ciertas condiciones, $\bar{Y}$ es consistente. Los supuestos para esta son que $Y\,i$ son iid (independientes e idénticamente distribuidas) y que $E(Y_i^2)$ es finito, es decir, que la varianza de Y_i es finita.

Teorema central del límite: dice que, cuando el tamaño muestral es elevado, la distribución muestral de la media muestral estandarizada es aproximadamente normal. La distribución de $\bar{Y}$ se aproxima a la normal cuando n se hace grande, si se cumple la condición de que Y tenga media y varianza finitas.

Capítulo 3

Repaso de Estadística

3.1 Estimación de la media poblacional

La media muestral Y1 es la manera natural de estimar $\bar{Y}$, pero no es el único método. La distribución muestral de un estimdor debe estar tan estrechamente centrada sobre el vlor desconocido como fuer posible. Las tres características desebles de un estimador son insesgadez, consistenci y eficiencia

- **Insesgadez:** la esperanza de la estimación de la media pobacional a partir de los datos de la muestra, es decir la media de su distribución muestral debe ser igual a la verdadera media poblacional, es decir: $E(\bar{\mu_Y}) = \mu_Y$.
- **Consistencia:** cuando el tamaño muestral sea grande, la incertidumbre acerca del valor del verdadero parámetro, proveniente de las variaciones aleatorias de la muestra, sea muy pequeña.
- **Eficiencia:** también se debe considerar cuál es el de menor varianza, ya que esto hace que sea más eficiente, ya que utiliza la información de los datos de mejor manera que el otro estimador con mayor varianza.

$\bar{Y}$ es un estimador ELIO de μY . Esto significa que es un Estimador Lineal Insesgado Óptimo. El estimador m que minimiza la suma de las distancias al cuadrado de los errores de predicción se denomina estimador de mínimos cuadrados. Tal que min $\sum_{i=1}^{n}(Y_i - m)^2$. El muestreo no aleatorio puede originar que $\bar{Y}$ sea sesgado.

Demostración: La distribución muestral de $\bar{Y}$ ha sido ya analizada en las Secciones 2.5 y 2.6 (2.5). Como se muestra en la Sección 2.5, $E(\bar{Y}) = \mu_Y$, por lo que $\bar{Y}$ es un estimador insesgado de μ_Y. De forma similar, la ley de los grandes números establece que $\bar{Y} \xrightarrow{p} \mu_Y$; es decir, $\bar{Y}$ es consistente.

Con respecto a la eficiencia, comparando el estimador de la media de Y con Y, dado que $VAR(Y) = \sigma_Y^2$ y que $VAR(\bar{Y}) = \sigma_Y^2/n$, entonces al aumentar n la varianza de la media de Y es menor a la varianza de Y, por ende el primero es más eficiente.

3.2 Contrastes de hipótesis sobre la media poblacional

El contraste de hipótesis consiste en especificar la hipótesis a contrastar, denominada hipótesis nula. El contraste de hipótesis implica la utilización de datos para comparar la hipótesis nula con la segunda hipótesis, denominada hipótesis alternativa. Se asume inicialmente que la hipótesis nula es cierta, y se la rechaza si el valor observado es tan extremo bajo esa hipótesis alternativa asociada a que se deriva si la nula no lo es. Puede ser bilateral o unilateral.

Si la hipótesis nula es “aceptada”, esto no significa que el estadístico declare que es cierta, sino que es aceptada provisionalmente reconociendo que puede ser rechazada más tarde en base a la evidencia adicional. Por esta razón, el contraste estadístico de hipótesis puede plantearse en términos de tomar la hipótesis nula como si no hubiera. No es conclusivo porque es información muestral.

El P-valor

Denominado asimismo probabilidad de significación, es la probabilidad de obtener un valor del estadístico al menos tan extremo para la hipótesis nula como el calculado en la muestra, bajo la suposición de que la

hipótesis nula es cierta

Suponiendo que la hipótesis nula es cierta. En el ejemplo, el p-valor es la probabilidad de obtener una $\bar{Y}$ al menos tan alejada en lo que respecta a las colas de su distribución, bajo la hipótesis nula, como la media muestral calculada realmente

Matemáticamente, para establecer la definición del p-valor, sea $\bar{Y}^{act}$ la expresión del valor de la media muestral calculada realmente con los datos en cuestión, y sea $p_r n$ la probabilidad calculada bajo la hipótesis nula. El p-valor es $P-valor = p_{rn}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$.

Es decir, el p-valor es el área de las colas de la distribución de $\bar{Y}$ la hipótesis nula más allá de $\bar{Y}^{act}$. Si el p-valor es pequeño, entonces el valor observado de $\bar{Y}$ es coherente con la hipótesis nula, pero no lo es si el p-valor es grande. Para calcular el p-valor, es necesario conocer la distribución muestral de $\bar{Y}$ bajo la hipótesis nula. De todos modos, el cálculo depende de si la varianza es conocida.

Cálculo del p-valor con varianza conocida

El p-valor es la probabilidad de observar un valor de $\bar{Y}$ calculado de $\mu_{Y,0}$ que juntos bajo la hipótesis nula es, de hecho, la probabilidad de que el valor muestral de $\bar{Y}$ calculado sea al menos tan extremo como el valor solución.. Expresada matemáticamente, es decir, el p-valor es:

$$P-valor = p_{rn}\left(\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right) = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right)$$

donde Φ es la función de distribución normal estándar acumulada. Es decir, el p-valor es el área de las colas de una distribución normal estándar más allá de $\pm(\bar{Y}^{act} - \mu_{Y,0})/\sigma_{\bar{Y}}$.

la varianza muestral, desviación típica muestral y error estándar

La varianza muestral es un estimador de la varianza poblacional, la desviación típica muestral es un estimador de la desviación típica poblacional, y el error estándar de la media muestral es un estimador de la desviación típica de la distribución muestral de $\bar{Y}$. La varianza muestral es:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Dividir por n-1 en lugar de n se denomina corrección de los grados de libertad; la estimación de la media consume parte de la información (introduce un pequeño sesgo a la baja, este se corrige al dividir por n-1) , es decir, consume 1 grado de libertad de los datos.

La varianza muestral es un estimador consistente de la varianza poblacional. En otras palabras, la varianza muestral está cerca a la varianza poblacional con elevada probabilidad cuando n es grande.

El error estandar de $\bar{Y}$: como la desviación típica es $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$, por la consistencia de la varianza muestral el error estándar es $\hat{\sigma}_{\bar{Y}} = s_Y/\sqrt{n} = ES(\bar{Y})$.

Cálculo del p-valor con varianza desconocida

Al ser s_Y^2 un estimador consistente de la varianza, el p-valor puede calcularse reemplazando el desvío por el error estándar, $ES(\bar{Y}) = \sigma$. Tal que:

$$P\text{-valor} = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}\right|\right)$$

La media muestral estandarizada con varianza desconocida se denomina estadístico t:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

y se utiliza para contrastar hipótesis. La fórmula del p-valor puede escribirse en términos del estadístico t, **cuando n es grande**, ya que se aproxima a una distribución normal. Es decir, al ser n grande, la varianza

muestral es cercana a la varianza poblacional, por lo que la distribución del estadístico t es aproximadamente la misma que $(\bar{Y} - \mu_{Y,0})/\sigma_{\bar{Y}}$, que a su vez se aproxima bien a la normal estándar con n grande por teorema central del límite.

En este caso el p-valor se puede reescribir en términos del estadístico t tal que:

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{ES(\bar{Y})}$$

y el p-valor puede calcularse como:

$$p\text{-valor} = 2\Phi(-|t^{act}|)$$

Contrastes de hipótesis con nivel de significación preestablecido

Se puede llevar a cabo el contraste de hipótesis sin calcular el valor p si se especifica previamente la probabilidad que se está dispuesto a tolerar de cometer error tipo I (rechazar la hipótesis nula cuando es verdadera). Se rechaza la hipótesis nula si y solo si el p-valor es menor al valor de significancia elegido. El contraste de hipótesis mediante un nivel de significancia preestablecido no necesita el cálculo de p valores.

Cuanto menor es el nivel de significancia, mayor es el valor crítico y más difícil se convierte rechazar la nula cuando la nula es falsa. Cuanto menor sea el nivel de significancia, menor será la potencia del contraste. A menudo se considera un 5 % de nivel de significación como una convención razonable

Concepto calve 3.5: Terminología del contraste de hipótesis

Error de tipo I: se rechaza la hipótesis nula cuando en realidad es cierta
Error de tipo II: la hipótesis nula no se rechaza cuando en realidad es falsa
El **nivel de significanción** es la probabilidad de rechazar la nula cuando es cierta , el **valor crítico** del estadístico de contraste es el valor del estadístico para el cual el contraste rechaza la hipótesis nula a un nivel de significancia dado, la **región de rechazo** es el conjunto de valores del estadístico de contraste para los que se rechaza la nula, de forma inversa la región de aceptación, el **tamaño del contraste** es la probabilidad de que el contraste rechace de forma incorrecta la nula cuando es verdadera y la **potencia del contraste** es la probabilidad de rechazar la nula cuando en realidad es falsa.

Alternativas unilaterales

El análisis se ha centrado en los intervalos de confianza bilaterales. Se podría construir además un intervalo de confianza unilateral como el conjunto de valores de μ que no queden por rechazadas mediante un contraste de hipótesis unilateral. Aunque los intervalos de confianza unilaterales se aplican en algunas ramas de la estadística, son poco comunes en el análisis econométrico aplicado.

El p-valor para una alternativa unilateral derecha es:

$$p\text{−valor} = \Pr_{H_0}(Z > t^{act}) = 1 - \Phi(t^{act})$$

Reglas para rechazar H_0

El criterio del p-valor se compar el p-valor directamente con el nivel de significancia α, si el p-valor es menor que α, se rechaza H_0, es decir, si la probabiliad de observar en tu muestra (asumiendo H_0 verdadera) un estadístico es menor al 5 % se rechaza la hipótesis nula.

El criterio del estadístico se compara el estadístico de contraste (calculado con la muestra) con el estadístico crítico (se obtiene a partir de las tablas). Si el estadístico de contraste es mayor al estadístico crítico, se rechaza la hipótesis nula. Es decir, si el estadístico de contraste cae dentro de la zona de rechazo (las colas de la distribución definidas por el valor crítico) se rechaza la hipótesis nula.

3.3 Intervalos de confianza para la media poblacional

Debido al error de muestreo aleatorio es imposible saber el valor exacto de la media poblacional usando solo la información muestral. Pero es posible utilizar los datos de la muestra aleatoria para construir un conjunto de valores que contengan la verdadera media poblacional con un cierta probabilidad preestablecida. Esto es un intervalo de confianza.

Este contiene las hipótesis que pueden o no rechazarse, y posee $1-\alpha$ de probabilidad de que contenga el verdadero valor de la media poblacional.

Para calcular los intervalos de confianza, dado el nivel de significancia, el valor del estadístico y el error estándar del mismo, a partir del nivel de significancia se obtiene el z-valor (para normales estándar) y a este se lo multiplica por el error estándar, el intervalo resulta igual a el valor del estadístico + y - el producto recién calculado.

La probabilidad de cobertura de un intervalo de confianza para la media poblacional es la probabil- idad, calculada sobre todas las posibles muestras aleatorias, de que contenga el verdadero valor de la media poblacional.

3.4 Comparación de medias de diferentes poblaciones

También se puede realizar un contraste de hipótesis para la diferencia entre dos medias $\mu_m - \mu_w$ se puede estimar mediante $\bar{Y}_m - \bar{Y}_w$. Y para esto debemos conocer la distribución de este último. Si las varianzas un conocidas y la distribución es aproximadamente normal se puede proceder con el p-valor para contrastar la hipótesis. Pero si la varianza es desconocida el error estándar es:

$$ES(\bar{Y}_m - \bar{Y}_w) = \sqrt{\left(\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}\right)}$$

$$t = \frac{(\bar{Y}_m - \bar{Y}_w) - d_0}{ES(\bar{Y}_m - \bar{Y}_w)}$$

Hay que tener en cuenta que se utiliza una versión modificada del estadístico t para la diferencia de medias, basado en una fórmula para el error estándar agrupado tiene una distribución exacta t de Student cuando Y está normalmente distribuida; sin embargo, la fórmula del error estándar agrupado es aplicable solamente en el caso particular de que los dos grupos tengan la misma varianza o de que ambos grupos tengan el mismo número de observaciones. El estimador de la varianza agrupada es:

$$s^2_{\text{agrupada}} = \frac{1}{n_m + n_w - 2}\left[\sum_{i=1}^{n_m}(Y_i - \bar{Y}_m)^2 + \sum_{i=1}^{n_w}(Y_i - \bar{Y}_w)^2\right]$$

Por lotanto el error estandar de la diferencia de medias es:

$$s_{agrupada}\sqrt{\frac{1}{n_m} + \frac{1}{n_w}}$$

Entonces utilizando este error estándar agrupado se calcula el estadístico t con la fórmula ya presentada para el contraste de diferencia de medias. Si las varianzas son iguales y tenemos una distribución t con $n_m + n_w - 2$ grados de libertad. Pero si los n y las varianzas son diferentes, el estimador de varianza agrupada es sesgado e inconsistente.

De la misma forma puede construirse un intervalo de confianza para la diferencia de medias, en este caso el intervalo de confianza es:

$$(\bar{Y}_m - \bar{Y}_w) \pm 1.96 \cdot ES(\bar{Y}_m - \bar{Y}_w)$$

3.5 Estimación de la difeencia de medias de los efectos causales mediante datos experimentales

El **efecto causal** de un tratamiento es el efecto previo sobre los resultados de interés del tratamiento de acuerdo con lo medido en un experimento aleatorizado controlado ideal. Este efecto puede ser expresado como la

diferencia de dos esperanzas condicionales. En concreto, el efecto causal sobre Y de un nivel de tratamiento x es la diferencia de las esperanzas condicionales, $E(Y|X = x) - E(Y|X = 0)$, donde $E(Y|X = x)$ es el valor esperado de Y para el grupo de tratamiento (que recibe el nivel de tratamiento X=x) en un experimento aleatorio controlado ideal y $E(Y|X = 0)$ es el valor esperado de Y para el grupo de control. Si solamente existen dos niveles de tratamiento es decir, si el tratamiento es binario, entonces podemos hacer que X=0 exprese el grupo de control y X=1 refiera el grupo de tratamiento.

Si tratamiento en un experimento aleatorizado controlado es binario, entonces el efecto causal puede ser estimado por la diferencia en los resultados medios muestrales entre los grupos de tratamiento y control.

La hipótesis de que el tratamiento es ineficaz es equivalente a la hipótesis de que ambas medias son iguales, lo cual puede contrastarse utilizando el estadístico t para comparar dos medias

Debido a que los experimentos en economía son escasos, a veces se analizan experimentos naturales, o cuasi experimentos, en los que algún suceso relacionado con las características del tratamiento o del sujeto tiene el efecto de asignar los diferentes tratamientos a diferentes sujetos, como si hubieran sido parte de un experimento aleatorizado controlado

3.6 Utilización del estadístico t cuando el tamaño muestral es pequeño

Consideremos el estadístico t para la diferencia de medias, cuando el tamaño muestral es pequeño no se puede recurrir al teorema central del límite, por ende, la distribución exacta del estadísitco t depende de la distribución de Y, y puede ser muy complicada. Existe un caso donde la distribución exacta del estadístico t es sencilla, si Y esta normalmente distribuida, entonces t presenta una distribución t de student (2.4) con n-1 grados de libertad, en esta Z y W están independientemente distribuidas. Cuando $Y_1, ..., Y_n$ son iid y la distribución poblacional de Y es normal, t puede escribirse como un cociente.

Donde $Z = \frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\sigma_Y^2/n}}$ y $W = \frac{(n-1)s_Y^2}{\sigma_Y^2}$. Partiendo de t, si se multiplica y divide por $\sqrt{\sigma_Y^2}$ y se agrupan términos se llega a:

$$
\begin{aligned}
t &= \frac{\bar{Y} - \mu_{Y,0}}{\sqrt{s_Y^2/n}} \\
&= \frac{(\bar{Y} - \mu_{Y,0})}{\sqrt{\sigma_Y^2/n}} \div \sqrt{\frac{s_Y^2}{\sigma_Y^2}} \\
&= \frac{(\bar{Y} - \mu_{Y,0})}{\sqrt{\sigma_Y^2/n}} \div \sqrt{\frac{(n-1)s_Y^2/\sigma_Y^2}{n-1}} \\
&= Z \div \sqrt{W/(n-1)}
\end{aligned}
$$

3.7 Diagramas de dispersión, covarianza muestral y correlación muestral

Tres formas de recoger relaciones entre variables son: el diagrama de dispersión, la covarianza muestral y el coeficiente de correlación muestral.

Un **diagrama de dispersión** es una gráfica de n observaciones sobre X_i e Y_i, en la que cada observación está representada por el punto (X_i, Y_i). Cada punto corresponde a un par (X,Y) de observaciones.

La **covarianza y correlación muestrales** son estimadores de la covarianza y correlación poblacionales. Se calculan sustituyendo la media poblacional (la esperanza), por la media muestral. La covarianza muestral, expresada mediante s_{XY} , es:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

El coeficiente de correlación muestral, o correlación muestral, se expresa mediante r_{XY} y es la: ratio entre la

covarianza muestral y las desviaciones típicas muestrales.

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

La correlación muestral mide la intensidad de la asociación lineal entre X e Y en una muestra de n observaciones. Como la correlación poblacional, la correlación muestral no tiene unidades de medida y toma valores entre -1 y 1. Está cerca igual a 1 si X e Yi e igual a -1 si X e Yi. Un coeficiente de correlación elevado no significa necesariamente que la línea tenga una pendiente pronunciada; más bien significa que los puntos del diagrama de dispersión se encuentran muy cerca de una línea recta.

Capítulo 4

Regresión Lineal con Regresor Único

4.1 El modelo de regresión lineal

El modelo de regresión lineal relaciona linealmente una variable X con otra variable Y. la pendiente de la recta que relaciona X con Y es el efecto de la variación de una unidad de X sobre Y, y es una característica desconocida de la distribución poblacional conjunta de X e Y. La pendiente y la constante de la recta que relaciona X con Y pueden estimarse mediante el método de los mínimos cuadrados ordinarios. El modelo poblacional se expresa como:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{4.1}$$

Donde:

- Y_i: Variable dependiente.
- X_i: Regresor o variable independiente.
- β_0, β_1: Coeficientes o parámetros.
- u_i: Término de error.

La primera parte de la ecuación es la recta de regresión poblacional o función de regresión poblacional. Y representa la relación entre X e Y que se cumple en promedio para la población. El valor en el origen y la pendiente son los coeficientes o parámetros de la recta de regresión poblacional. No siempre el intercepto tiene una interpretación económica significativa. Aun así, es el coeficiente que determina el nivel de la recta de regresión.

El término u_i recoge todos los factores responsables de la diferencia entre el verdadero valor y el predicho por la recta de regresión poblacional. Tendremos n valores de u ya que este es la diferencia entre lo observado y la recta de regresión poblacional.

Los parámetros del modelo de regresión dependen de la población y por tanto son desconocidos. Pero se pueden estimar a partir de una muestra de datos extraídos de esa población. Los factores que afectan a la variable, pero no se incluyen como tal en el modelo, debilitan el ajuste al modelo y la correlación entre las variables.

4.2 Estimación de los coeficientes del modelo de regresión lineal

Para elegir la recta que proporcione un mejor ajuste al modelo debemos utilizar el estimador de mínimos cuadrados ordinarios. Este elige los coeficientes de regresión de tal forma que la recta de regresión estimada se encuentre lo más cercana posible a los datos observados, y la cercanía está medida por la suma de los errores al cuadrado que se cometen con la predicción de Y dado X

Estimador de Mínimos Cuadrados Ordinarios (MCO)

El estimador MCO minimiza la suma de los cuadrados de los residuos:

$$\text{mín} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Los estimadores del término constante y de la pendiente que minimizan la suma de los cuadrados de los errores se denominan estimadores de mínimos cuadrados ordinarios. $\rightarrow min \sum_{i=1}^{n} u_1^2$

Concepto clave 4.2: El estimador MCO, valores estimados y residuos

Los estimadores MCO de la pendiente y del intercepto son:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

El valor de predicción MCO y los residuos son:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n$$
$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n$$

Los estimadores MCO tienen propiedades teóricas deseables. Son insesgados y consistentes, y bajo ciertas condiciones, eficientes.

4.3 Medidas de ajuste

El R^2 y el error estándar de la regresión miden la bondad del ajuste de la recta de regresión MCO a los datos. El R^2 oscila entre 0 y 1 y mide la proporción de la varianza de Y_i explicada por X_i. El error estándar de la regresión mide la distancia que habitualmente separa a Y_i de su valor esperado. El R^2 de la regresión es la proporción de la varianza muestral de Y_i explicada por (o predicha por) X_i.

El R^2 puede escribirse como el cociente entre la suma explicada (de cuadrados) y la suma total (de cuadrados). La suma explicada SE es la suma de las desviaciones al cuadrado de los valores de predicción de YI respecto de su media y la suma total ST es la suma de los cuadrados de las desviaciones de Y_i. respecto de su media, tal que:

$$SE = \sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2 \tag{4.2}$$

$$ST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{4.3}$$

Entonces $R^2 = SE/ST$ se puede expresar en términos del cociente entre la varianza de Y_i no explicada por X_i. La suma de los cuadrados de los residuos, suma residual, o SR, es la suma de los residuos MCO al cuadrado.

$$SR = \sum_{i=1}^{n} \hat{u_i}^2$$

Y como $ST = SE + SR$ entonces R^2 puede expresarse asimismo como $R^2 = 1 - SR/ST$.

error estándar de la regresión (ESR) es un estimador de la desviación típica del error de regresión u_i. (mide la bondad de ajuste, ya que mide la variabilidad entre los puntos y la línea de regresión, es decir, los residuos) Las unidades de u_i e Y_i son las mismas, por lo que el ESR es una medida de la dispersión de las observaciones en torno a la recta de regresión, medida en las unidades de la variable dependiente. Debido a que los errores de regresión no son observables, el ESR se calcula mediante sus homólogos muestrales, los residuos MCO $\hat{u_1}, \ldots, \hat{u_n}$.

$$ESR = s_{\hat{u}}$$

Donde:

$$s_{\hat{u}}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{SR}{n-2}$$

El n-2 corrige el sesgo a la baja introducido al estimar dos coeficientes de regresión (es una corrección por grados de libertad).

El hecho de que el R^2 de una regresión sea bajo (y el ESR sea grande) no implica que esta sea buena o mala, sino que expresa es que otros factores importantes influyen en la variable. Indica por tanto que X solo explica una pequeña parte de Y.

4.4 Los supuestos de mínimos cuadrados

Concepto clave 4.3: Supuestos de MCO

1. $E(u_i|X_i) = 0$ (Media condicional del error es cero).
2. (X_i, Y_i) son i.i.d. (Independientes e idénticamente distribuidas).
3. Los valores atípicos son improbables (curtosis finita).

El primer supuesto implica que la variable independiente X contiene toda la información relevante sobre la variable dependiente Y, establece que los otros factores contenidos en u_i no están correlacionados con X_i. Si este supuesto no se cumple, el estimador MCO estará sesgado.

El segundo supuesto implica que las observaciones muestrales son independientes entre sí y que todas ellas se extraen de la misma población.

El tercer supuesto implica que los valores atípicos son improbables, esto se precisa suponiendo que X e Y tienen momento de cuarto orden que existen y son finitos. Si existen valores atípicos, el estimador MCO puede ser muy sensible a ellos.

Si se cumplen estos supuestos el estimador MCO tiene distribución muestral normal cuando n es grande. Permitiendo desarrollar inferencias estadísticas (contrastes de hipótesis e intervalos de confianza) sobre los coeficientes de regresión.

4.5 Distribución muestral de los estimadores MCO

Debido a que los estimadores se calculan a partir de una muestra seleccionada aleatoriamente, los estimadores en sí mismos son variables aleatorias con una distribución de probabilidad muestral, que describe los valores que podrían tomar en las diferentes muestras aleatorias posibles. En muestras pe- queñas, estas distribuciones son complicadas, pero en muestras grandes, son aproximadamente normales por el teorema central del límite.

Por lo tanto, las distribuciones muestrales de los coeficientes estimados de la regresión MCO son insesgados bajo los supuestos de mínimos cuadrados.

$$E(\widehat{\beta}_0) = \beta_0 \quad \text{y} \quad E(\widehat{\beta}_1) = \beta_1$$

Con muestra grande, por teorema central del límite, las distribuciones muestrales de los coeficientes se aproximan bien a la normal bivariante, técnicamente el teorema refiere a medias, pero si se examina el númerador de los coeficientes se verá que son un tipo de media.

Concepto clave 4.4: Distribución para muestras grandes de $\hat{\beta}_0$ y $\hat{\beta}_1$

Bajo los supuestos, la distribución normal para muestras grandes de $\hat{\beta}_1$ es $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, donde la varianza de esta distribución es:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}$$

La distribución normal para muestras grandes de $\hat{\beta}_0$ es $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, donde:

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{\left[E(H_i^2)\right]^2}$$

donde $H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right] X_i$.

Cuanto mayor es la varianza de X_i menor es la varianza de β_1. Y cuanto menor es la varianza de u, menor es la varianza de β_1, ya que, si los errores son menores, entonces los datos presentarán una menor dispersión alrededor de la recta de regresión poblacional, por lo que su pendiente se estimará de manera más precisa.

4.6 Apéndice

4.2 Obtención de los estimadores

Se parte de la minimización de $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$, tomando derivadas prciales respecto a b_0 y b_1:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 = -2\sum_{i=1}^n (Y_i - b_0 - b_1 X_i) \quad \textbf{y} \tag{4.4}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 = -2\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)X_i \tag{4.5}$$

igualando estas derivadas a cero, agrupando términos y dividiendo por n, se obtiene:

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0 \tag{4.6}$$

$$\frac{1}{n}\sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^n X_i^2 = 0 \tag{4.7}$$

resolviendo queda:

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{XY}}{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \tag{4.8}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{4.9}$$

4.3 Distribución muestral del estimador MCO

Representación de β_1 en términos de los regresores y los errores

dado que $Y_i = \beta_0 + \beta_1 X_i + u_i$, $\quad Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + u_i - \bar{u}$, ahora podemos reescribir el número rador de la fórmula de β_1:

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})] \tag{4.10}$$

$$= \beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u}). \tag{4.11}$$

ahora trabajando el segundo miembro queda: $\sum_{i=1}^{n}(X_i-\bar{X})(u_i-\bar{u}) = \sum_{i=1}^{n}(X_i-\bar{X})u_i - \sum_{i=1}^{n}(X_i-\bar{X})\bar{u} = \sum_{i=1}^{n}(X_i-\bar{X})u_i$, donde la última igualdad se obtiene a partir de la definición de $\bar{X}$, esto implica que $\sum_{i=1}^{n}(X_i-\bar{X})\bar{u} = (\sum_{i=1}^{n}X_i - n\bar{X})\bar{u} = 0$, por lo tanto nos queda:

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \beta_1\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(X_i-\bar{X})u_i$$

Sustituyendo esta expresión en la fórmula de β_1 queda:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Prueba de que β_1 es insesgado

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] \tag{4.12}$$

$$= \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})E(u_i|X_1,...,X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] = \beta_1 \tag{4.13}$$

la segunda igualdad de la ecuación se obtiene utilizando la ley de esperanzas iteradas. Por el segundo supuesto de mínimos cuadrados, u_i se distribuye independientemente de X para todas las demás observaciones distintas de i, por lo que $E(u_i|X_1,...,X_n) = E(u_i|X_i)$. Por el primer supuesto de mínimos cuadrados, sin embargo, $E(u_i|X_i) = 0$. Por lo tanto β_1 es insesgado.

Algunas propiedades algebraicas adicionales acerca de MCO

Los residuos MCO y los valores estimados satisfacen:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0, \tag{4.14}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}, \tag{4.15}$$

$$\sum_{i=1}^{n}\hat{u}_i X_i = 0 \quad \text{y} \quad s_{\hat{u}X} = 0, \text{y} \tag{4.16}$$

$$ST = SR + SE \tag{4.17}$$

Demostración de 4.14 se tiene en cuenta la definición de $\hat{\beta}_0$ que nos permite escribir los residuos MCO como $u_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$ por lo tanto:

$$\sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}(Y_i-\bar{Y}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i-\bar{X})$$

Las definiciones de $\bar{Y}$ y $\bar{X}$ implican que $\sum_{i=1}^{n}(Y_i-\bar{Y}) = 0$ y $\sum_{i=1}^{n}(X_i-\bar{X}) = 0$ por lo que $\sum_{i=1}^{n}\hat{u}_i = 0$.

Demostración de 4.15 se tiene en cuenta que $Y_i = \hat{Y} + \hat{u}$ por lo que $\sum_{i=1}^{n}Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i$ por la condición anterior.

Demostración de 4.16 se tiene en cuenta que $\sum_{i=1}^{n} \hat{u}_i = 0$, esto implica que $\sum_{i=1}^{n} \hat{u}_i X_i = \sum_{i=1}^{n} \hat{u}_i (X_i - \bar{X})$, esto se cumple ya que si en el lado derecho distribuimos y luego sacamo a $\bar{X}$ por ser constante se cumple la igualdad, por lo tanto se tiene que:

$$\sum_{i=1} \hat{u}_i X_i = \sum_{i=1} [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})](X_i - \bar{X}) \tag{4.18}$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0, \tag{4.19}$$

A partir de este resultado y del anterior se llega a que $s_{\hat{u}X} = 0$. Concretamente se parte de la definición de la misma $s_{\hat{u}X} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(\hat{u}_i - \bar{\hat{u}})$ y del uso de que:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \rightarrow \bar{\hat{u}} = 0$$

$$\sum_{i=1}^{n} X_i \hat{u}_i = 0$$

Demostración 4.17 se obtiene de los resultados previos:

$$\text{ST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \tag{4.20}$$

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \tag{4.21}$$

$$= SR + SE + 2 \sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = SR + SE, \tag{4.22}$$

Donde la última igualdad se obtiene a partir de $\Sigma_{i=1}^{n} \hat{u}_i \hat{Y}_i = \Sigma_{i=1}^{n} \hat{u}_1 (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 \Sigma_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \Sigma_{i=1}^{n} \hat{u}_i X_i = 0$, por los resultados anteriores

Capítulo 5

Regresión Lineal Simple: Test de Hipótesis e Intervalos de Confianza

5.1 Contraste de hipótesis acerca de uno de los coeficientes

El contraste de H0 frente a la alternativa bilateral se realiza siguiendo tres pasos: El primero consiste en calcular el error estándar, que es un estimador de la desviación típica de la distribución poblacional de la VA. El segundo paso consiste en calcular el estadístico t, que presenta la forma general

$$t = \frac{\text{estimador - valor en la hipótesis nula}}{\text{error estándar del estimador}}$$

El tercer paso consiste en calcular el p-valor, que es el menor nivel de significación con el que la hipótesis nula puede ser rechazada, en base al estadístico de contraste observado en realidad; de forma equivalente, el p-valor es la probabilidad de obtener un estadístico, debido a la variabilidad del muestreo aleatorio, al menos tan diferente del valor de la hipótesis nula como el estadístico observado en realidad, suponiendo que la hipótesis nula es cierta. El tercer paso puede ser sustituido por la mera comparación entre el estadístico t con el valor crítico apropiado para el contraste con el nivel de significación deseado.

Contraste de hipótesis sobre la pendiente β_1

Seguimos los pasos mencionados mencionados anteriormente, donde calculamos primero el error estándar del estimador:

$$ES(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\beta_1}}$$

Donde la varianza es:

$$\hat{\sigma}^2_{\beta_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

Luego obtenemos el estadístico t:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{ES(\hat{\beta}_1)}$$

Por último calculamos el p-valor teniendo en cuenta como ejemplo a un test bilateral y que la pendiente se distribuye aproximadamente como una normal en muestras grandes, y por lo tanto t se distribuye aproximadamente como una VA normal estándar:

$$\begin{aligned}
\text{P} - \text{Valor} &= \Pr_{\mathbf{H_0}}\left[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{\mathbf{act}} - \beta_{1,0}|\right] \\
&= \Pr_{\mathbf{H_0}}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{\text{ES}(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{\mathbf{act}} - \beta_{1,0}}{\text{ES}(\hat{\beta}_1)}\right|\right] \\
&= \Pr_{\text{H}_0}\left(|t| > |t^{\text{act}}|\right) \\
&= \Pr(|Z| > |t^{act}|) = 2\Phi(-|T^{act}|)
\end{aligned}$$

Para el caso unilaateral (cola izquierda) el p-valor es:

$$p - valor = Pr(Z < t^{act}) = \Phi(t^{act})$$

También pueden realizarse test de hipótesis para el intercepto siguiendo los mismos pasos.

hay muchas veces en que ninguna hipótesis sencilla sobre un coeficiente de la regresión es dominante, y en su lugar a uno le gustaría conocer un rango de valores del coeficiente que sea consistente con los datos. Lo cual reclama la construcción de un intervalo de confianza

5.2 Intervalos de confianza para un coeficiente de la regresión

Un intervalo al 95 % para β_1 es el conjunto de valores que no pueden rechazarse mediante un contraste de hipótesis bilateral con un nivel de significación del 5 %. Se trata de un intervalo que presenta una probabilidad del 95 % de contener el verdadero valor de β_1. Debido a que este intervalo contiene el valor real en el 95 % de todas las muestras, se dice que tiene un nivel de confianza del 95 %

$$\text{intervalo de confianza al 95 \% para } \beta_1 = [\hat{\beta}_1 - 1.96 \cdot \text{ES}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \cdot \text{ES}(\hat{\beta}_1)]$$

Puede construirse un intervalo de confianza al 95 % para la predicción del efecto de una variación general en X. Consideremos la variación de X en una cantidad dada, Δx. La variación predicha en Y asociada a la variación de X es $\Delta x \beta_1$. La pendiente poblacional β_1 es desconocida, pero como se puede construir un intervalo de confianza para β_1, se puede construir un intervalo de confianza para el efecto esperado $\Delta x \beta_1$.

$$\text{Intervalo de confianza al 95 \% para } \beta_1 \Delta x = [\hat{\beta}_1 \Delta x - 1.96 ES(\hat{\beta}_1) \times \Delta x, \hat{\beta}_1 \Delta x + 1.96 ES(\hat{\beta}_1) \times \Delta x]$$

5.3 Regresión cuando X es binaria

El análisis de regresión también puede ser utilizado cuando el regresor es binario, es decir, cuando solamente toma dos valores, 0 o 1. Una variable binaria se denomina asimismo variable indicador o a veces variable ficticia o variable dummy.

$$D_i = \begin{cases} 1 & \text{si la ratio estudiantes-maestros del distrito } i\text{-ésimoes} < 20 \\ 0 & \text{si la ratio estudiantes-maestros del distrito } i\text{-ésimoes} \geq 20 \end{cases}$$

El modelo de regresión poblacional con D_i como variable explicativa es:

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \quad i = 1, \ldots, n.$$

La regresión en este caso se realiza de la misma forma que con un regresor continuo X, salvo que D no es continua y no resultaría útil interpretar la pendiente, pues no existe una línea recta. β_1 simplemente será el coeficiente que multiplica a D. Este parámetro será la diferencia entre las medias poblacionales condicionales cuando D = 0 y D = 1.

Si las dos medias poblacionales son iguales, entonces β_1 es cero. Por tanto, se puede contrastar la hipótesis nula de que las dos medias poblacionales son iguales frente a la hipótesis alternativa de que son distintas. (se puede utilizar para decidir si es significativa la diferencia entre las medias poblacionales de dos grupos al nivel del 5 %)

5.4 Heterocedasticidad y homocedasticidad

El único supuesto realizado sobre la distribución de u_i condicionada a X_i es que tiene una media igual a cero. Si, además, la varianza de esta distribución condicional no depende de X_i, entonces se dice que los errores son **homocedásticos**.

Cuando la distribución de u_i condicionada a x_i tiene media igual a cero y la varianza da la distribución condicional depende de x_i los errores son **heterocedásticos**.

Implicaciones matemáticas de la homocedasticidad

Los estimadores MCO siguen siendo insesgados y asintóticamente normales dado que los supuestos de MCO no establecen restricciones sobre la varianza condicional. Siendo aplicables las consecuencias de los supuestos para heterocedasticidad y homocedasticidad, es decir, el estimador MCO es insesgado, consistente y asintóticamente normal.

Concepto clave 5.1: Formulas para las varianzas

1. Varianza de β_1 válida para heterocedasticidad:

$$\hat{\sigma}^2_{\beta_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

2. Varianza de β_0 válida para heterocedasticidad:

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2 \hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}$$

   Donde: $\hat{H}_i = 1 - (\bar{X} / \frac{1}{n}\sum_{i=1}^{n} X_i^2) X_i$

3. Varianza de β_1 válida para homocedasticidad:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma_u^2}{n\sigma_X^2}$$

4. Varianza de β_0 válida para homocedasticidad:

$$\sigma^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma_X^2}\sigma_u^2$$

Debido a que la homocedasticidad es un caso particular de la heterocedasticidad, los estimadores de las varianzas válidos para heterocedasticidad dan inferencias estadísticas válidas tanto si los errores son heterocedásticos como si son homocedásticos. No se cumple lo contrario.

5.5 Fundamentos teóricos de mínimos cuadrados ordinarios

Si se cumplen los tres supuestos de MCO y si el error es homocedástico, entonces el estimador MCO tiene la menor varianza condicionada a $X_1, ..., X_n$, de entre todos los estimadores de la clase de estimadores lineales condicionalmente insesgados, es decir es ELIO. Esto se conoce como el teorema de **Gauss-Markov**.

Este teorema tiene 2 limitaciones:

1. Las condiciones podrían no cumplirse en la práctica si el término de error es heterocedástico como sucede a menudo en las aplicaciones económicas, en este caso el estimador MCO ya no es ELIO. Por ende la heterocedásticidad no amenaza la inferencia estadística pero si la característica de ELIO.
2. Aún cuando las condiciones se cumplen, existen otros posibles estimadores que no son lineales y condicionalmente insesgados, que bajo ciertas conndiciones son más eficientes que MCO. Estos se tratan a continuación.

Estimadores de regresión alternativos a MCO

Existe un estimador alternativo a MCO cuando existe heterocedasticidad con forma conocida, con un factor constante de proporcionalidad, llamado estimador de **mínimos cuadrados ponderados**. El MCP pondera la i-ésima observación por la inversa de la raíz cuadrada de la varianza condicional de u_i dado x_i. Debido a esta

ponderación, los errores de esta regresión ponderada son homocedásticos, por lo que MCO, cuando se aplican a los datos ponderados, es ELIO.

El problema práctico de los mínimos cuadrados ponderados es que es necesario conocer cómo la varianza condicional de u_i depende de x_i, algo que raramente se conoce en las aplicaciones econométricas. Por tanto, los mínimos cuadrados ponderados se utilizan con mucha menos frecuencia que MCO.

Otro estimador que se utiliza dado que MCO puede ser sensible a valores atípicos es el **estimador de mínima desviación absoluta**, donde los coeficientes se obtienen minimizando el error de predicción solo que se utiliza el valor absoluto del mismo en lugar del cuadrado , es decir: $\sum_{i=1}^{n} |Y_i - b_0 - b_1 X_i|$.

5.6 La utilización del estadístico t en regresión para muestras pequeñas

Cuando el tamaño de la muestra es pequeño, la distribución exacta del estadístico t es compleja y depende de la distribución poblacional de los datos que es desconocida. No obstante, si los tres supuestos de mínimos cuadrados se cumplen, los errores de regresión son homocedásticos, y además los errores de regresión se distribuyen normalmente, entonces el estimador MCO se distribuye normalmente y el estadístico t válido con homocedasticidad presenta una distribución t de Student. Estos cinco supuestos, los tres supuestos de mínimos cuadrados, que los errores son homocedásticos, y que los errores se distribuyen normalmente, se conocen colectivamente como los supuestos de la regresión normal homocedástica.

Si los errores de la regresión son homocedásticos y se distribuyen normalmente y si se utiliza el estadístico t válido con homocedasticidad, entonces los valores críticos que deben tomarse son los de la distribución t de Student en lugar de los de la distribución normal estándar. Debido a que la diferencia entre la distribución t de Student y la distribución normal es insignificante si n es mediano o grande, esta distinción solo es relevante si el tamaño de la muestra es pequeño.

Si asumimos normalidad exacta, porque tenemos una muestra grande, la distribución será exactmente una t. pero si la muestra no es grande es más complejo

5.7 Apéndice

Apéndice 5.1: Fórmulas de los errores estándar MCO

Errores estándar heterocedástico-robustos

El estimador $\hat{\sigma}^2_{\hat{\beta}_1}$ definido en la sección 5.1 se obtiene mediante la sustitución de varianzas poblacionales por las varianzas muestrales. En el caso de $\hat{\sigma}^2_{\hat{\beta}_0}$ se obtiene lo siguiente:

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{H}_i^2 \hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n} \hat{H}_i^2\right)^2}$$

Donde $\hat{H} = 1 - (\bar{X}/\frac{1}{n}\sum_{i=1}^{n} X_i^2)X_i$. El error estándar de $\hat{\beta}_0$ es $ES(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_0}}$.

Varianzas válidas con homocedasticidad

Con homocedasticidad la varianza condicional de u_i dado X_i es una constante, por lo tanto las fórmulas de las varianzas se simplifican (5.4).

Para la obtención de la varianza de $\hat{\beta}_1$ se expresa su numerador como:

$$\text{var}[(X_i - \mu_x)u_i] = E(\{(X_i - \mu_x)u_i - E[(X_i - \mu_x)u_i]\}^2) = E\{[(X_i - \mu_x)u_i]^2\} = E[(X_i - \mu_x)^2 u_i^2] = \quad (5.1)$$

$$= E[(X_i - \mu_x)^2 \text{var}(u_i|X_i)] \quad (5.2)$$

Donde la segunda igualdad se obtiene prque $E[(X_i - \mu_X)u_i] = 0$ por el primer supuesto de mínimos cuadrados, y la última igualdad se desprende de la ley de esperanzas iteradas. Si u_i es homocedástico, su varianza es

constante por lo que $E[(X_i-\mu_x)^2\text{var}(u_i|X_i)] = E[(X_i-\mu_x)^2]\sigma_u^2 = \sigma_X^2\sigma_u^2$. Sustituyendo esta expresión en el número se llega a la varianza de $\hat{\beta}_1$ válida para homocedásticidad.

Para el caso de $\hat{\beta}_0$ se comienza con el número donde puede sacarse $\hat{u}_i^2$ de la sumatoria dado que ahora su varianza es constante, y el término $\hat{H}_i^2$ se cancela con el denominador, es decir:

$$\frac{\hat{\sigma_u^2}\sum_{i=1}^n \hat{H}_i^2}{(\sum_{i=1}^n \hat{H}_i^2)^2}$$

Luego resta trabajar sobre $\sum_{i=1}^n \hat{H}_i^2$, se desarrolla el cuadrado, luego se aplica la sumatoria a cada término, luego se simplifica un poco y se restan los términos semejantes para factorizar n, es decir:

$$\begin{aligned}\sum_{i=1}^n \hat{H}_i^2 &= \sum_{i=1}^n \left(1-\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n X_i^2}X_i\right)^2 \\ &= \sum_{i=1}^n \left(1-2\left(\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)X_i+\left(\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)^2 X_i^2\right)\end{aligned}$$

$$\begin{aligned}\ldots &= \sum_{i=1}^n 1-2\left(\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)\sum_{i=1}^n X_i+\left(\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)^2\sum_{i=1}^n X_i^2 \\ &= n-2\left(\frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right)(n\bar{X})+\frac{\bar{X}^2}{\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^2}\left(\sum_{i=1}^n X_i^2\right) \\ &= n-\frac{2n\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2}+\frac{n\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2} \\ &= n-\frac{n\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2}\end{aligned}$$

$$\begin{aligned}\ldots &= n\left(1-\frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right) \\ &= n\left(\frac{\frac{1}{n}\sum_{i=1}^n X_i^2-\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2}\right) \\ &= \frac{n\left[\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\right]}{\frac{1}{n}\sum_{i=1}^n X_i^2} \\ &= \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\frac{1}{n}\sum_{i=1}^n X_i^2}\end{aligned}$$

Sustituyendo este resultado la formula de $\hat{\sigma}^2_{\hat{\beta}_0}$ nos queda como:

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma_X^2}\sigma_u^2$$

Errores estándar válidos con homocedasticidad

Se obtienen mediante la sustitución de las medias muestrales por las medias y las varianzas poblacionales tal que:

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^n (X_i-\bar{X})^2}$$

$$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)s^2_{\hat{u}}}{\sum_{i=1}^n (X_i-\bar{X})^2}$$

Apéndice 5.2: Las condiciones de Gauss-Markov y la demostración del teorema de Gauss-Markov

Las condiciones Gauss-Markov

Las 3 condiciones son:

(i) $E(u_i|X_1,\ldots,X_n)=0$
(ii) $\text{var}(u_i|X_1,\ldots,X_n)=\sigma_u^2, \quad 0<\sigma_u^2<\infty$
(iii) $E(u_iu_j|X_1,\ldots,X_n)=0, \quad i\neq j$

Las condiciones Gauss-Markov estan implícitas en los supuestos de MCO y el supuesto de que los errores son homocedásticos.

El estimador MCO $\hat{\beta}_1$ es un estimador lineal condicionalmente insesgado

Primero reescribimos $\hat{\beta}_1$, donde $\sum_{i=1}^n(X_i\bar{X})=0$ por definición de $\bar{X}$, entonces $\sum_{i=1}^n(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^n(X_i-\bar{X})Y_i-\bar{Y}\sum_{i=1}^n(X_i-\bar{X})=\sum_{i=1}^n(X_i-\bar{X})Y_i$ sustituyendo esta expresión en la fórmula de $\hat{\beta}_1$ se tiene:

$$\hat{\beta}_1=\frac{\sum_{i=1}^n(X_i-\bar{X})Y_i}{\sum_{j=1}^n(X_j-\bar{X})^2}=\sum_{i=1}^n\hat{a}_iY_i \tag{5.3}$$

Dado que las ponderaciones de $\hat{a}_i$ dependen de X_i pero no de Y_i, entonces $\hat{\beta}_1$ es un estimador lineal. Bajo Gauss-Markov es condicionalmente insesgado (demostrado en apéndice de capítulo 4) y la varianza de su distribución condicional es:

$$\text{var}(\hat{\beta}_1|X_1,...,X_n)=\frac{\sigma_u^2}{\sum_{i=1}^n(X_i-\bar{X})^2}$$

Prueba del teorema de Gauss-Markov

Para todos los estimadores lineales y condicionalmente insesgados que satisfacen:

$$\tilde{\beta}_1=\sum_{i=1}^n a_iY_i \quad \rightarrow \text{Lineal}$$
$$E(\tilde{\beta}_1\mid X_1,\ldots,X_n)=\beta_1 \quad \rightarrow \text{Insesgado}$$

Sustituyendo $Y_i=\beta_0+\beta_1X_i+u_i$ en $\tilde{\beta}_1=\sum_{i=1}^n a_iY_i$ y agrupando términos se obtiene:

$$\tilde{\beta}_1=\beta_0\left(\sum_{i=1}^n a_i\right)+\beta_1\left(\sum_{i=1}^n a_iX_i\right)+\sum_{i=1}^n a_iu_i \tag{5.4}$$

Por la primera condición de Gauss-Markov $E(\sum_{i=1}^n a_iu_i|X_1,...,X_n)=\sum_{i=1}^n a_iE(u_i|X_1,...,X_n)=0$, entonces tomando esperanza condicionada a ambos lados se obtiene $E(\tilde{\beta}_1|X_1,...,X_n)=\beta_0(\Sigma_{i=1}^n a_i)+\beta_1(\Sigma_{i=1}^n a_iX_i)$. Por hipótesis debe ocurrir que $\beta_0(\Sigma_{i=1}^n a_i)+\beta_1(\Sigma_{i=1}^n a_iX_i)=\beta_1$ para que esto se cumpla debe ocurrir que:

$$\sum_{i=1}^n a_i=0 \quad \text{y} \quad \sum_{i=1}^n a_iX_i=1.$$

Al sustituir estas condiciones en (5.4) se obtiene $\tilde{\beta}_1-\beta_1=\sum_{i=1}^n a_iu_i$. Por lo tanto la $var(\tilde{\beta}_1\mid X_1,\ldots,X_n)=var(\sum_{i=1}^n a_iu_i|X_1,\ldots,X_n)=\sum_{i=1}^n\sum_{j=1}^n a_ia_j\,\text{cov}(u_iu_j|X_1,...,X_n)$ aplicando las condiciones segunda y tercera de Gauss-Markov, los términos cruzados del doble sumatorio desaparecen y la expresión se simplifica a:

$$\text{var}(\tilde{\beta}_1|X_1,...,X_n)=\sigma_u^2\sum_{i=1}^n a_i^2 \tag{5.5}$$

Ahora demostraremos que esta varianza es mayor que la de $\hat{\beta}_1$. Sea $a_i=\hat{a}_i+d_i$, por lo que $\sum_{i=1}^n a_i^2=\sum_{i=1}^n(\hat{a}_i+d_i)^2=\sum_{i=1}^n\hat{a}_i^2+2\sum_{i=1}^n\hat{a}_id_i+\sum_{i=1}^n d_i^2$.

Utilizando la definición de $\hat{a}_i$ (5.3)se tiene que:

$$\sum_{i=1}^{n} \hat{a}_i d_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})d_i}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = \frac{\sum_{i=1}^{n} d_i X_i - \bar{X}\sum_{i=1}^{n} d_i}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

$$= \frac{\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} \hat{a}_i X_i\right) - \bar{X}\left(\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \hat{a}_i\right)}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = 0,$$

La penultima igualdad se deduce de que $d_i = a_i - \hat{a}_i$ y la última a partir de las condiciones que especificamos antes sobre $\sum_{i=q}^{n} a_i$ y $\sum_{i=q}^{n} \hat{a}_i$. por lo tanto $\sigma_u^2 \Sigma_{i=1}^{n} a_i^2 = \sigma_u^2 \Sigma_{i=1}^{n} \hat{a}_i^2 + \sigma_n^2 \Sigma_{i=1}^{n} d_i^2 = \text{var}(\hat{\beta}_1 | X_1, ..., X_n) + \sigma_u^2 \Sigma_{i=1}^{n} d_i^2$. sustituyendo este resultado en (5.5) se tiene que:

$$\text{var}(\tilde{\beta}_1 | X_1, ..., X_n) - \text{var}(\hat{\beta}_1 | X_1, ..., X_n) = \sigma_n^2 \sum_{i=1}^{n} d_i^2.$$

Por lo tanto $\tilde{\beta}_1$ tiene una mayor varianza condicional que $\hat{\beta}_1$ si d_i es distinto de cero para cualquier $i, \ldots, n$.

El teorema de Gauss-Markov si X es no aleatoria

Es aplicable a las variables explicativas que no cambian sus valores en muestras repetidas. En concreto , si el segundo supuesto de mínimos cuadrados se sustituye por el supuesto de que $X_1, ..., X_n$ son no aleatorias (fijas en muestras repetidas) y $u_1, ..., u_n$ son i.i.d., entonces lo anteriormente definido, así como la prueba del teorema de Gauss-Markov, son aplicables directamente, salvo que todo lo definido como «condicionado a» resulta innecesario, debido a que $X_1, ..., X_n$ toman los mismos valores de una muestra a otra.

La media muestral es el estimador lineal eficiente de E(Y)

Una consecuencia del teorema de Gauss-Markov es que la media muestral, $\bar{Y}$, es el estimador lineal más eficiente de $E(Y_i)$ si $Y_1, ..., Y_n$ son i.i.d. Para comprobarlo, consideremos el caso de la regresión sin una «X», por lo que el único regresor es la variable constante $X_{0i} = 1$. Entonces el estimador MCO $\hat{\beta}_0 = \bar{Y}$. De ello se desprende que, bajo los supuestos de Gauss-Markov $\bar{Y}$, es ELIO. Téngase en cuenta que el requisito de Gauss-Markov de que el error sea homocedástico se satisface automáticamente en este caso porque no hay regresores, por lo que se deduce que $\bar{Y}$ es ELIO si $Y_1, ..., Y_n$ son i.i.d.

Capítulo 6

Teoría de Regresión Lineal con Regresor Único

6.1 Los supuestos ampliacos de mínimos cuadrados y el estimador MCO

Los primeros tres son los que mencionamos en el capítulo 4: que la media condicional de u_i dando x_i es igual a cero: que x_i, y_i son extracciones independientes e idénticamente distribuidas de su distribución conjunta: y que x_i y u_i tienen momentos de orden cuatro finitos. Bajo estos supuestos, el estimador MCO es insesgado, consistente y tiene una distribución muestral asintóticamente normal.

Si estos tres supuestos se cumplen, entonces la contrastación mediante el estadístico t y la construcción de intervalos de confianza están justificados si el tamaño de la muestra es grande. No obstante, para desarrollar una teoría de estimación eficiente mediante MCO o para caracterizar la distribución muestral exacta del estimador MCO, son necesarios unos supuestos más fuertes.

El cuarto supuesto ampliado de mínimos cuadrados es que u_i es homocedástico; es decir, $\text{var}(u_i|x_i) = \sigma_u^2$ es una constante. Si este supuesto adicional se cumple, el estimador MCO es eficiente entre todos los estimadores lineales e insesgados, condicionado a x_i

El quinto supuesto ampliado de mínimos cuadrados es que la distribución condicional de u_i, dado x_i, es normal.

Bajo todos estos supuestos u_i es normal con media en cero y varianza σ_u^2 y la distribución condicional de $u_i|x_i$ es normal con media en cero y varianza en $\text{var}(u_i|x_i) = \sigma_u^2$, la cual es σ_u^2 por el cuarto supuesto. u_i y x_i están distribuidas independientemente.

Si se cumplen los cinco supuestos ampliados de mínimos cuadrados, el estimador MCO tiene una distribución muestral exacta normal y el estadístico t válido con homocedasticidad tiene una distribución exacta t de Student.

6.2 Fundamentos de la teoría de distribución asintótica

Es la teoría de la distribución de cuando el tamaño muestral es grande. La teoría es asintótica en el sentido de que caracteriza el comportamiento del estadístico en el límite, a medida que n tiende a infinito. Aunque las muestras grandes no son nunca infinitas, la teoría de distribución asintótica interpreta un papel central en econometría y en estadística por dos razones. El límite asintótico puede proporcionar una aproximación de alta calidad a la distribución en muestras finitas y porque son más sencillas, y por tanto más fáciles de utilizar en la práctica, que las distribuciones exactas en muestras finitas. Las dos piedras angulares de la teoría de distribución asintótica son la ley de los grandes números y el teorema central del límite.

La convergencia en pobabilidad y la ley de los grandes números

Sea $(s_1, s_2, ..., s_n, ...)$ una secuencia de variables aleatorias. Por ejemplo, s_n podría ser la media muestral $\bar{y}$ de una muestra de n observaciones de la variable aleatoria Y. La secuencia de variables aleatorias s_n se dice que converge en probabilidad a un límite, μ, si la probabilidad de que s_n se encuentre a una distancia menor o igual a σ de μ tiende a 1 a medida que n tiende a infinito, siendo σ positivo. Es decir:

$$S_n \xrightarrow{P} \mu \quad \text{si y solo si} \quad \Pr(|S_n - \mu| \geq \delta) \to 0$$

Si $n \to \infty$ para todo $\delta > 0$. Si $S_n \xrightarrow{P} \mu$, entonces se dice que S_n es un **estimador consistente** de μ.

Ley de los grandes números: establece que, si $Y_1, ..., Y_n$ son i.i.d., $E(Y) = \mu_Y$ y $var(Y) < \infty$, entonces $\bar{Y} \xrightarrow{P} \mu_Y$. Una característica de la distribución muestral es que la varianza de $\bar{Y}$ disminuye al aumentar el tamaño muestral; otra característica es que la probabilidad de que $\bar{Y}$ esté más allá de un cierto distancia $\pm\delta$ de su esperanza μ_Y decrece a medida que n aumenta. El vínculo entre estas dos características lo proporciona la desigualdad de Chebyshev:

$$\Pr(|\bar{Y} - \mu_Y| \geq \delta) \leq \frac{var(\bar{Y})}{\delta^2} = \frac{var(Y)}{n\delta^2}$$

Para cualquier constante $\delta > 0$, a medida que n tiende a infinito, el lado derecho de la desigualdad tiende a cero, y por tanto el lado izquierdo también tiende a cero. Por lo tanto, $\bar{Y} \xrightarrow{P} \mu_Y$.
Se desprende de esto que esta desigualdad tenderá a cero cuando n tienda a infinito.

El teorema central del límite y la convergencia en distribución

Una sucesión de variables aleatorias s_n se dice que converge en distribución a una variable aleatoria S si la función de distribución acumulada (FDA) de s_n tiende a la FDA de S en todos los puntos donde la FDA de S es continua. Es decir:

$$S_n \xrightarrow{d} S \quad \text{si y solo si} \quad F_{S_n}(x) \to F_S(x)$$

O de forma similar:

$$S_n \xrightarrow{d} S \quad \text{si y solo si} \quad \lim_{n\to\infty} F_n(x) = F_S(x)$$

El límite se cumple en todos los instantes x donde la distribución límite F es continua, la distribución de F se denomina **distribución asintótica** de la secuencia s_n.
El teorema central del límite dice que, bajo ciertas condiciones generales, la media muestral estandarizada converge en distribución a una normal estándar. Es decir, si $Y_1, ..., Y_n$ son i.i.d., $E(Y_i) = \mu_Y$ y $var(Y_i) = \sigma_Y^2 < \infty$, entonces: la distribución asintótica de $(\bar{Y} - \mu_Y)/\sigma_{\bar{Y}}$ es N(0,1), y como $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$, entonces:

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Que puede expresarse como:

$$\sqrt{n}(\bar{Y} - \mu_Y) \xrightarrow{d} \sigma_Y^2 Z$$

Donde Z es una variable aleatoria N(0,1). Por lo tanto $\sqrt{n}(\bar{Y} - \mu_Y)$ converge a N(0,σ_Y^2).

Teorema de Slutsky: combina la consistencia y la convergencia en distribución. Supongamos que $a_n \xrightarrow{P} a$, $b_n \xrightarrow{d} b$, y $c_n \xrightarrow{P} c$. Entonces:

1. $a_n + b_n \xrightarrow{d} a + b$
2. $a_n b_n \xrightarrow{d} ab$
3. $b_n/c_n \xrightarrow{d} b/c$, siempre que c $\neq 0$

Estos tres resultados se denominan de forma conjunta teorema de Slutsky.

El **teorema de la función continua** se refiere a las propiedades asintóticas de una función continua, g, de una sucesión de variables aleatorias, S_n. El teorema tiene dos partes. La primera es que si S_n converge en probabilidad a una constante a, entonces g(S_n) converge en probabilidad a g(a); la segunda es que si S_n converge en distribución a S, entonces g(S_n) converge en distribución a g(S). Es decir, si g es una función continua, entonces:

1. Si $S_n \xrightarrow{P} a$, entonces g(S_n) $\xrightarrow{P}$ g(a)
2. Si $S_n \xrightarrow{d} S$, entonces g(S_n) $\xrightarrow{d}$ g(S)

Puede utilizarse el teorema central del límite junto con el teorema de Slutsky y el teorema de la función continua para derivar la distribución asintótica de una amplia variedad de estadísticos, como por ejemplo el estadísitco t basado en la media muestral.

$$t = \frac{\bar{Y} - \mu_0}{s_Y/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma_Y} \div \frac{s_Y}{\sigma_Y}$$

Debido a que $Y_1, ..., Y_n$ tienen momentos de segundo orden (implícito por el supuesto de varianza finita), el estimador, y dado que son i.i.d., cumple el teorema central del límite, por lo que: $\sqrt{n}(\bar{Y} - \mu_0)/\sigma_Y \xrightarrow{d} N(0,1)$. Además, $s_Y^2 \xrightarrow{P} \sigma_Y^2$ por la ley de los grandes números, por lo que $s_Y/\sigma_Y \xrightarrow{P} 1$. Aplicando el teorema de Slutsky al cociente, obtenemos que:

$$t \xrightarrow{d} N(0,1) \quad y \quad a_n = \frac{s_Y}{\sigma_Y} \xrightarrow{p} 1$$

Deduciendose que el estadístico t tiene una distribución asintótica N(0,1).

6.3 Distribución asintótica del estimador MCO y del estadístico t

La distribución para muestras grandes del estimador MCO $\hat{\beta}_1$ puede derivarse aplicando el teorema central del límite a la expresión del estimador MCO:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$$

Dado que los datos son i.i.d. y que $E(u_i|X_i) = 0$, el numerador es una media muestral de variables aleatorias i.i.d. con media cero. Por el teorema central del límite, la distribución asintótica del numerador es normal con media cero y varianza:

$$\text{var}[(X_i - \mu_X)u_i] = E[(X_i - \mu_X)^2 u_i^2]$$

Por la ley de los grandes números, el denominador converge en probabilidad a σ_X^2. Aplicando el teorema de Slutsky, la distribución asintótica de $\hat{\beta}_1$ es normal con media β_1 y varianza:

$$\sigma_{\hat{\beta}_1}^2 = \frac{\text{var}[(X_i - \mu_X)u_i]}{n\sigma_X^4}$$

Expresado como esta en el libro:

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\text{var}(v_i)}{[\text{var}(X_i)]^2}\right) \tag{6.1}$$

Donde $v_i = (X_i - \mu_X)u_i$.

Bajo los 3 primeros supuestos de MCO, los errores estándar heterocedástico-robustos para $\hat{\beta}_1$ constituyen la base para realizar inferencias estadísticas válidas. En concreto:

$$\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} \xrightarrow{p} 1$$

Donde: $\sigma_{\hat{\beta}_1}^2$ es la varianza del estimador $\hat{\beta}_1$ y $\hat{\sigma}_{\hat{\beta}_1}^2$ es el estimador heterocedástico-robusto de dicha varianza. Siendo este último igual a:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right]^2}$$

Para demostrar que el coeficiente de más arriba converge en probabilidad a 1 se utilizan en primer lugar las definiciones y se reescribe como:

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} = \left[\frac{n}{n-2}\right]\left[\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\hat{u}_i^2}{\text{var}(v_i)}\right] \div \left[\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2}{\text{var}(X_i)}\right]^2$$

Cada uno de estos términos convergen en probabilidad a 1 cuando n tiende a infinito.

El primer término converge a 1 trivialmente. El segundo término converge a 1 por la propiedad de consistencia de la varianza muestral y la ley de los grandes números. Solo nos resta que el segundo término convergá a 1, es decir $\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$.

Esto se demuestra en 2 pasos, el primero implica que $\frac{1}{n}\sum_{i=1}^n v_i^2 \xrightarrow{p} \text{var}(v_i)$; el segundo que $\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^n v_i^2 \xrightarrow{p} 0$.

Se parte de suponer que X_i y u_i tienen momentos de orden ocho, para el primer paso se debe demostrar que $\sum_{i=1}^n v_i^2$ cumple la ley de los grandes números, para esto v_i^2 debe ser i.i.d. y tener varianza finita. Son i.i.d. por supuesto de MCO y su varianza es finita por la desigualdad de Cauchy-Schwarz:

$$var(v_i^2) \leqslant E(v_i^4) = E[(x_i-\mu_X)^4 u_i^4] \leqslant [E[x_i-\mu_X]^8 E[u_i^8]]^{\frac{1}{2}}$$

Entonces si tienen momentos de orden ocho, la varianza es finita y se satisface la ley de los grandes números.

Para el segundo paso, se tiene que demostrar que: $\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^n v_i^2 \xrightarrow{p} 0$, por la definición de $\hat{v}_i$ se tiene:

$$\frac{1}{n}\sum_{i=1}^n [(X_i-\bar{X})^2\hat{u}_i^2 - (X_i-\mu_X)^2 u_i^2] \xrightarrow{p} 0$$

Como $\hat{u}_i = u_i - (\hat{\beta}_0-\beta_0) - (\hat{\beta}_1-\beta_1)X_i$, se puede reescribir la expresión anterior como:

$$\frac{1}{n}\sum_{i=1}^n [(X_i-\bar{X})^2[u_i - (\hat{\beta}_0-\beta_0) - (\hat{\beta}_1-\beta_1)X_i]^2 - (X_i-\mu_X)^2 u_i^2] \xrightarrow{p} 0$$

Como las estimaciones β son consistentes se simplifica a:

$$\frac{1}{n}\sum_{i=1}^n [(X_i-\bar{X})^2 u_i^2 - (X_i-\mu_X)^2 u_i^2] \xrightarrow{p} 0$$

Dado que la media de X es consistente, a medida que aumenta n este término tiende en probabilidad a cero, demostrandose así que $\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} \xrightarrow{p} 1$.

Demostración de que el estadístico t heterocedástico-robusto tiene una distribución asintótica N(0,1) si se cumplen los 3 supuestos de MCO:

$$t = \frac{\widehat{\beta}_1-\beta_{1,0}}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{\sqrt{n}(\widehat{\beta}_1-\beta_{1,0})}{\sqrt{n\sigma^2_{\widehat{\beta}_1}}} \div \sqrt{\frac{\widehat{\sigma}^2_{\widehat{\beta}_1}}{\sigma^2_{\widehat{\beta}_1}}}$$

Se deduce de los resultados anteriores (6.1) que el primer término seguido a la igualdad converge en distribución a N(0,1), además por consistencia del error heterocedástico robusto, el segundo término converge en probabilidad a 1. Por teorema de Slutsky, el estadístico t heterocedástico-robusto converge en distribución a N(0,1).

6.4 Distribuciones muestrales exactas con errores normalmente distribuidos

si los cinco supuestos ampliados de mínimos cuadrados se cumplen, entonces el estimador MCO tiene una distribución muestral normal, condicionada a x_i. Además, el estadístico t tiene una distribución t de Student.

Se presentan a continuación estos resultados para β_1:

Si los errores son i.i.d. con distribución normal e independientes de los regresores, entonces la distribución de β_1 condicionada a $X_1,\ldots,X_n$, es $N(\beta_1,\sigma^2_{\hat{\beta}_1|X})$, donde

$$\sigma^2_{\hat{\beta}_1|X} = \frac{\sigma^2_u}{\sum_{i=1}^n (X_i-\bar{X})^2}$$

Para verificar este debe establecerse primero que la distribución es normal , para esto debe tenerse en cuenta que condicionado a las x_i, $\hat{\beta}_1-\beta_1$ es una media ponderada de los errores u_i:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})u_i}{\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2}$$

Por los supuestos 1, 2 , 4 y 5, u_i es N(0,σ^2_u) e i.i.d., entonces u_i y x_i estan independientemente distribuidas, dado que las medias ponderadas de las variables normales son normales, se deduce que $\hat{\beta}_1$ se distribuye normal condicionada a x_i.

En segundo lugar se toman las esperanzas condicionadas a ambos lados de la expresión anterior para llegar a que $\hat{\beta}_1$ es condicionalmente insesgada:

$$E(\hat{\beta}_1|X_1,...,X_n) = \beta_1 + \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})E(u_i|X_1,...,X_n) = \beta_1$$

En tercer lugar, se usa el hecho de que los errores se distribuyen de forma independiente de x_i para calcular la varianza condicional de $\hat{\beta}_1$:

$$\begin{aligned}
\text{var}(\widehat{\beta}_1|X_1,...,X_n) &= \text{var}\left[\frac{\sum_{i=1}^n (X_i-\bar{X})u_i}{\sum_{i=1}^n (X_i-\bar{X})^2}|X_1,...,X_n\right] \\
&= \frac{\sum_{i=1}^n (X_i-\bar{X})^2 \,\text{var}(u_i|X_1,...,X_n)}{\left[\sum_{i=1}^n (X_i-\bar{X})^2\right]^2} \\
&= \frac{\sum_{i=1}^n (X_i-\bar{X})^2\sigma^2_u}{\left[\sum_{i=1}^n (X_i-\bar{X})^2\right]^2}
\end{aligned}$$

Distribución del estadístico t válido con homocedasticidad

Se parte de:

$$t = \frac{\widehat{\beta}_1-\beta_{1,0}}{\widehat{\sigma}_{\widehat{\beta}_1}}$$

Utilizando los errores estándar válidos con homocedasticidad y reemplazando se tiene:

$$t = \frac{\widehat{\beta}_1-\beta_{1,0}}{\sqrt{s^2_{\hat{u}}/\sum_{i=1}^n (X_i-\bar{X})^2}} = \frac{\widehat{\beta}_1-\beta_{1,0}}{\sqrt{\sigma^2_u/\sum_{i=1}^n (X_i-\bar{X})^2}} \div \sqrt{\frac{s^2_u}{\sigma^2_u}}$$

Que es igual a:

$$\frac{(\widehat{\beta}_1-\beta_{1,0})/\sigma_{\widehat{\beta}_{1|X}}}{\sqrt{W/(n-2)}}$$

Donde $s^2_u = \frac{1}{n-2}\sum_{i=1}^n \hat{u}^2_i$ y $W = \sum_{i=1}^n \hat{u}^2_i/\sigma^2_u$. bajo la hipótesis nula, el $\hat{\beta}_1$ se distribuye como N($\beta_{1,0}$,$\sigma^2_{\hat{\beta}_1|X}$) condicionado a x_i, por lo que el numerador se distribuye como N(0,1). W se distribuye como una chi-cuadrado con n-2 grados de libertad y es independiente del numerador. Por lo tanto, el estadístico t tiene una distribución t de Student con n-2 grados de libertad, condicionada a x_i.

6.5 Mínimos cuadrados ponderados

El estimador de mininos cuadrados ponderados (MCP) es más eficiente que el MCO cuando los errores son heterocedástico. Este método requiere conocer acerca de la función de varianza condicional $var(u_i|x_i)$, la cual cambia para cada x. Existen dos casos, cuando esa varianza es conocida con un factor de proporcionalidad y MCP es ELIO, y cuando la forma funcional de esa varianza es conocida pero contiene algunos parámetros desconocidos que deben ser estimados. Bajo ciertas condiciones adicionales, la distribución asintótica de MCP en el segundo caso es la misma que si los parámetros de la función de la varianza condicional fueran en realidad conocidos, y en este sentido el estimador MCP es asintóticamente ELIO.

Cuando se trata del primer caso, $var(u_i|x_i) = \lambda h(x_i)$ donde λ es una constante y h es una función conocida. En este caso, el estimador MCP es el estimador obtenido dividiendo en primer lugar la variable dependiente y el regresor por la raíz cuadrada de h y posteriormente realizando una regresión MCO de esta variable dependiente transformada sobre el regresor transformado. En concreto, se dividen ambos lados del modelo de una sola variable independiente por raíz $h(x_i)$ para obtener:

$$\tilde{Y}_i = \beta_0 \tilde{X}_{0i} + \beta_1 \tilde{X}_{1i} + \tilde{u}_i$$

Donde $\tilde{Y}_i = Y_i/\sqrt{h(x_i)}$, $\tilde{X}_{0i} = 1/\sqrt{h(x_i)}$, $\tilde{X}_{1i} = X_i/\sqrt{h(x_i)}$ y $\tilde{u}_i = u_i/\sqrt{h(x_i)}$.

El estimador MCP será entonces el estimador MCO obtenido de la ecuación anterior. Este será ELIO bajo los 3 primeros supuestos más el supuesto de heterocedasticidad conocida. La razón de que el estimador MCP sea ELIO es que la ponderación de las variables hace que el término de error $\tilde{u}_i$ de la regresión ponderada sea homocedástico. Es decir:

$$var(\tilde{u}_i|X_i) = var\left(\frac{u_i}{\sqrt{h(x_i)}}|X_i\right) = \frac{var(u_i|X_i)}{h(x_i)} = \frac{\lambda h(x_i)}{h(x_i)} = \lambda$$

por lo que la varianza condicional de u (a X_i) es constante, porque los primeros cuatro supuestos son aplicables, y si bien el teorema de Gauss-Márkov como viene siendo utilizado no es aplicable por la incorporación de la variable X_{0i} en el término independiente, si lo es la generalización de este teorema para regresión múltiple, y en consecuencia es estimador MCP es ELIO. En la práctica la función h suele ser desconocida y debe estimarse.

Si la heterocedasticidad tiene una forma funcional conocida, entonces la función de heterocedasticidad h puede ser estimada y el estimador MCP puede calcularse utilizando la función estimada.

Si la varianza es conocida pero con parámetros desconocidos no es posible construir las variables ponderadas (las de $\tilde{u}_i$), pero sí es posible estimar los parámetros para estimar la varianza y luego calcular las variables explícitas ponderadas con la varianza estimada.

$$\hat{\tilde{Y}}_i = Y_i/\sqrt{\text{var}(u_i|X_i)}, \quad \hat{\tilde{X}}_{0i} = 1/\sqrt{\text{var}(u_i|X_i)}, \quad \hat{\tilde{X}}_{1i} = X_{1i}/\sqrt{\text{var}(u_i|X_i)}$$

Debido a que este método de MCP se puede llevar a cabo mediante la estimación de los parámetros desconocidos de la función de la varianza condicional, este método a veces se denomina MCP factibles o MCP estimados.

En resumen , el método de los MCP factibles consta de 5 pasos:

1. Regresión de Y_i sobre X_i mediante MCO y obtención de los residuos MCO, $\hat{u}_i$, $i = 1, \ldots, n$.
2. Estimación de un modelo para la función de la varianza condicional $var(u_i|X_i)$, esto implica la regresión de $\hat{u}_i^2$ sobre X_i^2. En general, este paso implica la estimación de una función para la varianza condicional $var(u_i|X_i)$.
3. Utilización de la función estimada para calcular los valores esperados de la función de la varianza condicional, $var(u_i|X_i)$.
4. Ponderación de la variable dependiente y el regresor (incluido el término independiente) por la inversa de la raíz cuadrada de la función de la varianza condicional estimada.
5. Estimación de los coeficientes de la regresión ponderada mediante MCO; los estimadores resultantes son los estimadores MCP.

Existen dos maneras de actuar en presencia de heterocedasticidad: estimar los parámetros mediante MCP o mediante MCO y utilizar los errores estándar heterocedástico-robustos. La decisión acerca de qué método utilizar en la práctica requiere sopesar las ventajas y desventajas de cada uno de ellos.

La ventaja de MCP consiste en que es más eficiente que el estimador MCO de los coeficientes de los regresores originales, al menos asintóticamente. La desventaja de MCP consiste en que es necesario conocer la función de la varianza condicional y estimar sus parámetros.

La ventaja de utilizar errores estándar heterocedástico-robustos es que dan lugar a inferencias asintóticamente válidas incluso si no se conoce la forma de la función de la varianza condicional. Una ventaja adicional es que los errores estándar heterocedástico-robustos se calculan fácilmente como una opción dentro de los paquetes informáticos modernos de regresión, por lo que no es necesario ningún esfuerzo adicional para protegerse frente a esa amenaza. La desventaja de los errores estándar heterocedástico-robustos consiste en que el estimador MCO tendrá una mayor varianza que el estimador MCP (basado en la verdadera función de la varianza condicional), al menos asintóticamente.

6.6 Apéndice

Apéndice 17.2: Dos desigualdades

La desiguldad de Chebychev

La desigualdad de Chebychev utiliza la varianza de la variable aleatoria V con el fin de acotar la probabilidad de que V se encuentre a más distancia que $\pm\delta$ respecto de su media, donde δ es una constante positiva:

$$\Pr(|V-\mu_V| \geqslant \delta) \leqslant \frac{\text{var}(V)}{\delta^2}$$

La demostración es la siguiente, sea $W = V - \mu_V$, sea f la f.d.p. de W, y sea δ cualquier número positivo. Ahora

$$\begin{aligned}
E(W^2) &= \int_{-\infty}^{\infty} w^2 f(w)dw && (6.2)\\
&= \int_{-\infty}^{-\delta} w^2 f(w)dw + \int_{-\delta}^{\delta} w^2 f(w)dw + \int_{\delta}^{\infty} w^2 f(w)dw && (6.3)\\
&\geqslant \int_{-\infty}^{-\delta} w^2 f(w)dw + \int_{\delta}^{\infty} w^2 f(w)dw && (6.4)\\
&\geqslant \delta^2 \left[\int_{-\infty}^{-\delta} f(w)dw + \int_{\delta}^{\infty} f(w)dw\right] && (6.5)\\
&= \delta^2 \Pr(|W| \geqslant \delta), && (6.6)
\end{aligned}$$

La primera desigualdad se cumple debido a que el término que no es tenido en cuenta es no negativo, la segunda desigualdad se cumple debido a que $w^2 \geqslant \delta^2$ a lo largo de todo el rango de integración y la última igualdad se cumple por la definición de $Pr(|\ W\ | \geqslant \delta)$. ustituyendo $W = V - \mu_V$ en la última expresión, teniendo en cuenta que $E(W^2) = E[(V-\mu_V)^2] = var(V)$, y reordenando se obtiene la desigualdad. Si V es discreta, esta prueba es aplicable con sumatorios en sustitución de las integrales.

La desigualdad de Cauchy-Schwarz

La desigualdad de Cauchy-Schwarz es una generalización de la desigualdad de la correlación $|\ \rho_{XY}\ | \leqslant 1$, para incorporar medias distintas de cero. La desigualdad de Cauchy-Schwarz es:

$$|E(XY)| \leqslant \sqrt{E(X^2)E(Y^2)}$$

La demostración es similar a la prueba de la desigualdad de la correlación del Apéndice 2.1. Sea $W = Y + bX$, donde b es una constante. Entonces $E(W^2) = E(Y^2) + 2bE(XY) + b^2E(X^2)$. Ahora sea $b = -E(XY)/E(X^2)$ por lo que (después de la simplificación) la expresión se convierte en $E(W^2) = E(Y^2) - [E(XY)]^2/E(X^2)$. Debido a que $E(W^2) = 0$ (debido a que $W^2 \geqslant 0$), debe ocurrir que $[E(XY)]^2 = E(X^2)E(Y^2)$, y la desigualdad de Cauchy-Schwarz se deduce tomando la raíz cuadrada.

Capítulo 7

Regresión Lineal con Varios Regresores

Los factores omitidos en un modelo de regresión pueden ocasionar que el estimador de mínimos cuadrados ordinarios sea sesgado.

La idea clave de la regresión múltiple es que si se dispone de datos sobre las variables omitidas, entonces se pueden incluir como regresores adicionales y por tanto calcular el efecto de un regresor mientras se mantienen constantes las otras variables.

7.1 Sesgo de variable omitida

Los coeficientes del modelo de regresión múltiple se pueden estimar a partir de los datos utilizando MCO; los estimadores MCO de regresión múltiple son variables aleatorias porque dependen de los datos de una muestra aleatoria; y en muestras grandes las distribuciones muestrales de los estimadores MCO son aproximadamente normales.

En un modelo de único regresor los demás factores quedan recopilados en el término de error. Si el regresor está correlacionado con una variable que ha sido omitida en el análisis y esta determina, en parte, la variable independiente, el estimador MCO presentará sesgo de variable omitida. Este se produce cuando se cumplen dos condiciones: cuando variable omitida está correlacionada con los regresores incluidos en la regresión y cuando la variable omitida es un factor determinante de la variable dependiente.

El sesgo de variable omitida implica que el primer supuesto de MCO, $E(u_i|X_i) = 0$, no se cumple. Esto es porque el término de error u_i en el modelo de regresión lineal con un único regresor representa todos los factores, distintos de X_i, que son determinantes de Y_i. Si uno de esos otros factores está correlacionado con X_i, esto significa que el término de error (que contiene a este factor) está correlacionado con X_i. Esto causa que el estimador MCO sea sesgado. Este sesgo no desaparece incluso en muestras muy grandes, y el estimador MCO es inconsistente.

Sea ρ^2_{xu} la correlación entre X_i y u_i, y que se cumplen el 2do y 3er supuesto de MCO, entonces el estimador MCO tiene el límite:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X} \tag{7.1}$$

Esto ocurre a medida que aumenta el tamaño de la muestra, con probabilidad creciente. El Sesgo de variable omitida es un problema tanto si el tamaño de la muestra es grande como si es pequeño. Debido a que $\hat{\beta}_1$ no converge en probabilidad al verdadero valor β_1, es sesgado e inconsistente. En la formula, el segundo término representa el sesgo de variable omitida, el hecho de que el mismo sea pequeño o grande depende de ρ_{Xu}. La dirección del sesgo depende del signo de la correlación entre X_i y u_i.

7.2 El modelo de regresión múltiple

El modelo de regresión múltiple extiende el modelo de regresión simple para incluir variables adicionales como regresores. Este modelo permite estimar el efecto sobre Y_i de la variación de una variable (X_{1i}) manteniendo constantes el resto de regresores $(X_{2i}, X_{3i}, \ldots, X_{ki})$.

Supongamos por el momento que solo hay dos variables independientes, X_{1i} y X_{2i}. En el modelo de regresión lineal múltiple, la relación promedio entre estas dos variables independientes y la variable dependiente, Y, está dada por la función lineal $E(Y_i|X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Esta es la recta de regresión poblacional o función de regresión poblacional en el modelo de regresión múltiple. El coeficiente β_0 es el intercepto, término independiente o término constante; el coeficiente β_1 es el coeficiente de la pendiente de X_{1i}; y el coeficiente β_2 es el coeficiente de la pendiente de X_{2i}. A una o más variables independientes del modelo de regresión múltiple se les denomina a veces variables de control.

β_1 es el efecto sobre Y de la variación en una unidad de X_{1i}, manteniendo constante X_{2i} o teniendo en cuenta X_{2i}, o controlado por X_{2i}. Esta interpretación de β_1 se deriva de la definición según la cual el efecto esperado sobre Y de un cambio en X_{1i}, ΔX_{1i}, manteniendo X_{2i} constante, es la diferencia entre el valor esperado de Y cuando las variables independientes toman los valores $X_{1i} + \Delta X_{1i}$ y X_{2i} y el valor esperado de Y cuando las variables independientes toman los valores X_{1i} y X_{2i}.

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

Se obtiene una ecuación para ΔY en términos de ΔX_1 restando la ecuación $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ la anterior Ecuación, por lo que se obtiene $\Delta Y = \beta_1 \Delta X_1$. Es decir:

$$Expresion \quad \beta_1 = \frac{\Delta Y}{\Delta X_1} \quad manteniendo \quad X_2 \quad constante$$

El coeficiente β_1 es el efecto sobre Y (la esperanza de la variación de Y_{1i}) de un cambio unitario en X_{1i}, manteniendo fija X_{2i}. Describir β_1 es el efecto parcial sobre Y de X_{1i}, manteniendo constante X_{2i}.

La interpretación del término independiente en el modelo de regresión múltiple β_0 es el valor esperado de Y_i cuando X_{1i} y X_{2i} son iguales a cero.

Al igual que en el caso de la regresión con un único regresor, los factores que determinan Y_i además de X_{1i} y X_{2i} se incorporan en forma de un término de error u_i. Este término de error es la desviación de una observación concreta respecto de la relación poblacional promedio. En consecuencia, se tiene que:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \ldots, n,$$

En la regresión con regresores binarios, puede ser útil considerar β_0 como el coeficiente de un regresor que es siempre igual a 1; piénsese en β_0 como el coeficiente de X_0, siendo $X_{0i} = 1$ para todo i. Por tanto, el modelo de regresión múltiple poblacional puede escribirse como

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad \text{donde} \quad X_{0i} = 1, i = 1, \ldots, n.$$

La variable X_{0i} se denomina a veces regresor constante, ya que toma el mismo valor, el valor 1, para todas las observaciones. De este modo, el intercepto β_0, a veces se denomina término constante de la regresión.

El término de error u_i en el modelo de regresión múltiple es homocedástico si la varianza de la distribución condicional de u_i dados X_1, var$(u_i|X_{1i}, \ldots, X_{ki})$, es constante para todo i, y por tanto no depende de X. En cualquier otro caso, el término de error es heterocedástico.

7.3 El estimador MCO en regresión múltiple

Para estimar los coeficientes del MCO para regresión múltiple puede calcularse del mismo modo que para regresión simple. La suma de los cuadrados de los errores de predicción para las n observaciones es:

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$$

Los coeficientes que minimizan esto son los estimadores de mínimos cuadrados de $b_0, b_1, ..., b_k$. Las fórmulas para estos regresores se obtienen igual que antes, aunque resulta más simple expresarlas y analizarlas mediante notación matricial.

7.4 Medidas de ajuste

Tres estadísticos que se utilizan habitualmente en la regresión múltiple son el error estándar de la regresión, el R^2 de la regresión, y el R^2 ajustado. Los tres estadísticos miden la bondad de la estimación MCO de la recta de regresión múltiple, es decir, en qué medida la recta describe, o "se ajusta" a los datos.

El error estándar de la regresión (ESR) estima la desviación típica del término de error u_i. Por tanto, el ESR es una medida de la dispersión de la distribución de Y alrededor de la recta de regresión. En regresión múltiple, el ESR es:

$$ESR = s_{\hat{u}} \quad \text{donde} \quad s^2_{\hat{u}} = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SR}{n-k-1}$$

Donde SR es la suma de los residuos al cuadrado, k es el número de regresores independientes (excluyendo el término independiente) y n es el tamaño de la muestra. La única diferencia con este nuevo ESR es que se divide por $n-k-1$ (en vez de $n-2$), es decir, se ajusta a la baja por los k + 1 coeficientes (ajuste por grados de libertad). Cuando el n es grande, el efecto del ajuste por los grados de libertad es insignificante.

El R^2 de la regresión es la proporción de la varianza muestral de Y_i que está explicada (o predicha) por los regresores. De manera equivalente, el R^2 es 1 menos la proporción de la varianza de Y_i no explicada por las variables explicativas, es decir:

$$R^2 = 1 - \frac{SR}{ST} = \frac{ST-SR}{ST} = \frac{SE}{ST}$$

En regresión múltiple, el R^2 aumenta cada vez que se añade un regresor, a menos que el coeficiente estimado del regresor adicional sea exactamente cero. Para comprobarlo, comenzamos con un único regresor y posteriormente añadimos un segundo regresor. Cuando se utiliza MCO para estimar el modelo con ambas variables explicativas, MCO halla aquellos valores de los coeficientes que reduzcan al mínimo la suma de los cuadrados de los residuos. Si resulta que MCO elige un coeficiente para el nuevo regresor que sea exactamente cero, entonces SR será la misma tanto si se incluye la segunda variable en la regresión como si no. Pero si MCO escoge cualquier otro valor distinto de cero, entonces debe ocurrir que ese valor reduzca la SR de la regresión que excluye a este regresor. En la práctica, es extremadamente inusual que un coeficiente estimado sea igual a cero, por lo que en general la SR disminuye al añadirse un nuevo regresor. Pero esto significa que el R^2 en general aumenta (y nunca disminuye) al añadirse un nuevo regresor.

Ese aumento de R^2 no significa que la adición de una variable mejore el modelo, de hecho, este estimador proporciona una estimación exagerada de la bondad de ajuste. Para corregirlo lo deflactamos o reducimos mediante algún factor, de modo que obtenemos **el R^2 ajustado**:

$$\bar{R}^2 = 1 - \frac{SR/(n-k-1)}{ST/(n-1)} = 1 - \frac{n-1}{n-k-1}\frac{SR}{ST} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2) = 1 - \frac{s^2_{\hat{u}}}{s^2_Y} \tag{7.2}$$

La diferencia con la fórmula anterior es que agrega un factor de corrección multiplicando. Este coeficiente de ajuste siempre será mayor a 1, por lo que el R^2 ajustado será menor al R^2 y el R^2 ajustado puede ser negativo, lo que ocurre si todos los regresores, considerados de forma conjunta, reducen la suma de los cuadrados de los residuos en una cantidad tan pequeña que no pueda compensar al factor de corrección incorporado.

El R^2 ajustado es útil porque cuantifica la medida en que los regresores explican la variabilidad de la variable dependiente. De todos modos no debemos fiarnos demasiado de estos estadísticos, ya que la decisión sobre incluir o no una variable debe basarse en si esta permite estimar mejor el efecto causal de interés.

7.5 Los supuestos de mínimos cuadrados en regresión múltiple

Concepto Clave 6.4: Supuestos de MCO en regresión múltiple

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, ..., n$$

Donde:

1. La distribución condicional de u_i dado los regresores $X_{1i}, X_{2i}, \ldots, X_{ki}$ tiene media cero, es decir:

$$E(u_i|X_{1i}, X_{2i}, ..., X_{ki}) = 0$$

2. $x_{1i}, x_{2i}, \ldots, x_{ki}; \quad con \quad i = 1, \ldots, n$ son extracciones i.i.d. de su distribución conjunta.
3. Los valores atípicos elevados son poco probables: $x_{1i}, x_{2i}, \ldots, x_{ki}$ e y_i presentan momentos de cuarto orden finitos y distintos de cero.
4. No existe multicolinealidad perfecta.

El cuarto supuesto se incorpora para la regresión múltiple. Descarta la multicolinealidad perfecta, bajo la cual es imposible calcular el estimador MCO. Los regresores presentan multicolinealidad perfecta si uno de ellos es función lineal perfecta del resto. La razón de que no se pueda calcular MCO en ese caso es que produce un cociente con divisor igual a cero en las fórmulas de MCO.

A nivel intuitivo implica que, cuando se presenta multicolinealidad perfecta, el coeficiente de uno de los regresores no puede estimarse manteniendo constantes los demás regresores, porque al estar linealmente relacionados, no varían cuando los demás se mantienen constantes.

7.6 La distribución de los estimadores MCO en regresión múltiple

Si se cumplen los supuestos de mínimos cuadrados del Concepto clave 6.4 (7.5), entonces en muestras grandes los estimadores MCO $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ están distribuidos normalmente de forma conjunta y cada $\hat{\beta}_j$ se distribuye $N(\beta_j, \sigma^2_{\hat{\beta}_j}), j = 0, ..., k$

En regresión lineal con único regresor, con muestras grandes, los estimadores se aproximan a una distribución normal bivariante. Para el análisis de regresión múltiple, esto se hace extensible para la distribución normal multivariante. El teorema central del límite se aplica del mismo modo, los estimadores son promedios de una muestra aleatoria, y si esta es suficientemente grande, la distribución muestral de esos promedios se convierte en normal.

7.7 Multicolinealidad

La multicolinealidad perfecta puede surgir por incorporar la variable nuevamente a modo de porcentaje , que sea una variable binaria y no existan valores iguales a cero, que se incluya la "inversa" de una variable, que se considere como variable binaria a cada subcategoría de una variable (esto es la **trampa de la variable ficticia**), o que se incluyan todas las variables ficticias de una variable categórica. Para evitar la trampa de la variable ficticia, se debe eliminar una de las variables ficticias de cada conjunto de variables ficticias que representen una variable categórica.

La multicolinealidad imperfecta implica que dos o más de los regresores están altamente correlacionados. Esto no resulta un problema para la teoría de los MCO. Si los regresores presentan multicolinealidad imperfecta, entonces los coeficientes de al menos un regresor individual se estimarán de forma imprecisa. Estos seguiran siendo insesgados y consistentes, pero sus varianzas serán elevadas. En consecuencia, los errores estándar serán grandes, y los intervalos de confianza para los coeficientes serán amplios. A diferencia de la multicolinealidad perfecta, la multicolinealidad imperfecta no impide la estimación de la regresión, ni implica un problema lógico en la selección de los regresores.

7.8 Apéndice

Apéndice 6.1: Obtención de la ecuación 6.1 (7.1)

Se parte de la ecuación:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)u_i}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

Bajo los 2 últimos supuestos del concepto clave 4.3 (3). $(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2 \xrightarrow{p} \sigma_X^2$, y $(1/n)\sum_{i=1}^{n}(X_i - \bar{X})u_i \xrightarrow{p} cov(u_i, X_i) = \rho_{Xu}\sigma_u\sigma_X$. Mediante la sustitución de estos límites en la Ecuación recién presentada se llega a la ecuación 6.1 (7.1).

Apéndice 6.2: Distribución de los estimadores MCO en presencia de dos regresores y errores homocedásticos

Aunque la fórmula general para la varianza de los estimadores MCO en regresión múltiple es complicada, con dos variables explicativas (k = 2) y si los errores son homocedásticos, entonces la fórmula se simplifica lo suficiente como para proporcionar alguna información sobre la distribución de los estimadores MCO.

Debido a que los errores son homocedásticos, la varianza condicional de u_i dado X_{1i} y X_{2i} es σ_u^2. Cuando hay dos variables explicativas X_1 y X_2 y el término de error u_i es homocedástico, en muestras grandes la distribución muestral de $\hat{\beta}_1$ es $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, donde la varianza de esta distribución, $\sigma_{\hat{\beta}_1}^2$, es:

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n}\left(\frac{1}{1-\rho_{X_1,X_2}^2}\right)\frac{\sigma_u^2}{\sigma_{X_1}^2} \tag{7.3}$$

donde ρ_{X_1,X_2} es la correlación poblacional entre las dos variables explicativas X_1 y X_2, y $\sigma_{X_1}^2$ es la varianza poblacional de X_1.

La varianza $\sigma_{\hat{\beta}_1}^2$ de la distribución muestral de $\hat{\beta}_1$ depende del cuadrado de la correlación entre los regresores. Si X_1 y X_2 están altamente correlacionadas, ya sea positiva o negativamente, entonces ρ_{X_1,X_2}^2 se acerca a 1 y por tanto el término $1 - \rho_{X_1,X_2}^2$ en el denominador de la varianza recién mostrada es pequeño y la varianza de $\hat{\beta}_1$ es mayor de lo que lo sería si ρ_{X_1,X_2} acercara a cero.

Otra característica de la distribución normal conjunta para muestras grandes de los estimadores MCO es $\hat{\beta}_1$ y $\hat{\beta}_2$ están en general correlacionados. Cuando los errores son homocedásticos, la correlación entre los estimadores MCO $\hat{\beta}_1$ y $\hat{\beta}_2$ es el opuesto (cambia de signo) de la correlación entre los dos regresores:

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\rho_{X_1,X_2}$$

Apéndice 6.3: El teorema de Frisch-Waugh

El estimador MCO en regresión múltiple se puede calcular mediante una serie de regresiones más cortas. Consideremos el modelo de regresión múltiple con k regresores. El estimador MCO de β_1 se puede calcular en tres etapas:

1. Regresión de X_{1i} sobre los demás regresores $X_{2i}, X_{3i}, \ldots, X_{ki}$ y obtención de los residuos $\hat{v}_i$.
2. Regresión de Y_i sobre los demás regresores $X_{2i}, X_{3i}, \ldots, X_{ki}$ y obtención de los residuos $\hat{w}_i$.
3. Regresión de $\hat{w}_i$ sobre $\hat{v}_i$ sin término independiente. El coeficiente estimado en esta regresión es igual a $\hat{\beta}_1$.

donde las regresiones incluyen un término constante (intercepto). El teorema de Frisch-Waugh establece que el coeficiente MCO de la etapa 3 es igual al coeficiente de MCO de X_1 del modelo de regresión múltiple.

Este resultado proporciona una formulación matemática de la forma en que el coeficiente de regresión múltiple $\hat{\beta}_1$ estima el efecto neto Y de X_1 controlado por las dos primeras regresiones (etapas 1 y 2) eliminan de Y y $_1$ la variación asociada con $X_2, \ldots, X_k$ y por lo tanto el efecto de X_1 se Y se utiliza para quedar despejado de eliminar (controlar por) el efecto de las otras X. El teorema de Frisch-Waugh se demuestra en el Ejercicio 18.17.

Este teorema sugiere de qué manera la Ecuación (7.3) se puede deducir a partir de la varianza de β_1 válida con homocedásticidad (5.4). Debido a que $\hat{\beta}_1$ es el coeficiente de regresión MCO de la regresión de $\tilde{Y}$ sobre $\tilde{X}_1$ la Ecuación (5.4) sugiere que la varianza válida con homocedasticidad de $\hat{\beta}_1$ es $\sigma^2_{\tilde{Y}}/\sigma^2_{\tilde{X}_1}$ n donde $\sigma^2_{\tilde{X}_1}$ es la varianza de $\tilde{X}_1$. Debido a que $\tilde{X}_1$ es el residuo de la regresión de X_1 sobre X_2 (recordemos que la Ecuación (7.3) se refiere al modelo con k = 2 regresores), la Ecuación (7.2) implica que $\sigma^2_{\tilde{X}_1} = (1 - R^2_{X_2,X_1})\sigma^2_{X_1}$ donde $R^2_{X_2,X_1}$ es el R2 ajustado de la regresión de X_1 sobre X_2. La Ecuación (7.3) se deduce de $\sigma^2_{\tilde{Y}} \to \sigma^2_{\tilde{X}_1}$, $R^2_{X_2,X_1} \to \rho^2_{X_1,X_2}$ y $\sigma^2_{\tilde{X}_1} \to \sigma^2_{X_1}$.

Capítulo 8

Test de Hipótesis e IC para Regresión Múltiple

Presenta los métodos para la cuantificación de la incertidumbre de muestreo del estimador MCO a través de la utilización de errores estándar, contrastes de hipótesis estadísticos e intervalos de confianza. Una nueva posibilidad que aparece en regresión múltiple es una hipótesis que involucra simultáneamente a dos o más coeficientes de regresión. El método general para contrastar esas hipótesis «conjuntas» incluye un nuevo estadístico de contraste, el estadístico F.

8.1 Constrastes de hipótesis e IC para un único coeficiente

Para el caso de único regresor, era posible estimar la varianza del estimador MCO mediante sustitución de las medias muestrales por las esperanzas, lo que conducía a un estimador $\hat{\sigma}^2$. Bajo mínimos cuadrados, la ley de grandes números implica que las medias muestrales convergen a las poblacionales, siendo que $\hat{\sigma}^2/\sigma^2$ converge en probabilidad a σ^2. La raíz cuadrada de σ^2 es el error estándar del parámetro estimado, y es un estimador de la desviación típica de la distribución muestral del estimador. Todo esto es extensible a la regresión múltiple. Las ideas clave, la normalidad de los estimadores en muestras grandes y la posibilidad de estimar consistentemente la desviación típica de su distribución muestral, son las mismas, ya sea con uno, dos, o 12 regresores.

Si se desea contrastar la hipótesis de que el verdadero coeficiente β_j toma un valor específico, el cual proviene de la teoría o una decisión tomada en el contexto del caso.

para el p-valor utilizo la distribución normal estándar acumulada (y para el conreaste de valor critico el estadístico t). Los fundamentos teóricos de este procedimiento son que el estimador MCO tiene una distribución normal en muestras grandes que, bajo la hipótesis nula, tiene como media el verdadero valor bajo la hipótesis nula y que la varianza de esta distribución puede estimarse de modo consistente. La distribución muestral de $\hat{\beta}_j$ es aproximadamente normal. Bajo la hipótesis nula, la media de esta distribución es $\beta_{j,0}$. La varianza de esta distribución puede estimarse consistentemente.

Concepto Clave 7.1: Contraste de hipótesis para un único coeficiente en regresión múltiple

$$H_0 : \beta_j = \beta_{j,0} \quad vs \quad H_a : \beta_j \neq \beta_{j,0}$$

Donde el estadístico de contraste es:

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{ES(\hat{\beta}_j)}$$

En muestras grandes, bajo H_0, t se distribuye aproximadamente como una distribución normal estándar. El p-valor para este contraste bilateral es:

$$p - \text{valor} = 2Pr(Z > |t^{act}|) = 2\Phi(-|t^{act}|)$$

Donde Z es una variable aleatoria con distribución normal estándar, Φ es la función de distribución acumulada de la normal estándar y t^{act} es el valor observado del estadístico t.
Se rechaza la hipótesis nula al nivel de significación α si el p-valor es menor que α o, equivalentemente, si $|t^{act}| > t_{\alpha/2}$.

Concepto Clave 7.2: Intervalo de confianza para un único coeficiente en regresión múltiple

Un intervalo de confianza del 100(1 - α) % para β_j es:

$$\hat{\beta}_j \pm z_{\alpha/2} ES(\hat{\beta}_j)$$

Donde $z_{\alpha/2}$ es el valor crítico de la distribución normal estándar tal que $Pr(Z > z_{\alpha/2}) = \alpha/2$.

8.2 Contrastes de hipótesis conjuntas

El contraste acerca de dos o más coeficientes parte de una hipótesis nula conjunta que impone dos o mas restricciones sobre los coeficientes de regresión.

$$H_0 : \beta_j = \beta_{j,0}, \beta_m = \beta_{m,0}, ..., \quad para \quad un \quad total \quad de \quad q \quad restricciones \tag{8.1}$$

$$H_1 : una \quad o \quad mas \quad de \quad las \quad q \quad restricciones \quad bajo \quad H_0 \quad no \quad se \quad cumplen \tag{8.2}$$

Si alguna de las igualdades bajo la hipótesis nula H_0 es falsa, entonces la hipótesis nula conjunta en sí misma es falsa. Por tanto la hipótesis alternativa es que al menos una de las igualdades de la hipótesis nula no se cumple.

No se puede contrastar los coeficientes individuales de uno en uno mediante el estadístico t habitual porque no es un procedimiento fiable. Este método rechaza la hipótesis nula con demasiada frecuencia, debido a que se le dan demasiadas oportunidades: si no se rechaza mediante el primer estadístico t, se intenta otra vez mediante el segundo. Si los regresores están correlacionados, la situación es aún más complicada. El tamaño del procedimiento «una a una» depende del valor de la correlación entre los regresores. Debido a que el procedimiento de contraste «una a una» tiene el tamaño erróneo, es decir, su tasa de rechazo bajo la hipótesis nula no es igual al nivel de significación deseado, se necesita un nuevo método. Un método consiste en modificar el procedimiento «una a una» para lo que se utilizan diferentes valores críticos que aseguran que su tamaño sea igual a su nivel de significación. Este método denominado **método de Bonferroni**. La ventaja de este es que es aplicable de forma muy general. Su desventaja es que puede tener baja potencia: con frecuencia no rechaza la hipótesis nula cuando en realidad la hipótesis alternativa es verdadera.

Existe otro método para contrastar hipótesis conjuntas con mayor potencia, particularmente cuando los regresores se encuentran altamente correlacionados. Este método se basa en el estadístico F. **El estadístico F se utiliza para contrastar hipótesis conjuntas** sobre los coeficientes de regresión. Para el caso de dos restricciones, combina los estadísticos t_1 y t_2 mediante la formula:

$$F = \frac{1}{2} \frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}}$$

donde $\hat{\rho}_{t_1,t_2}$ es un estimador de la correlación entre los dos estadísticos t. si no están correlacionados este término se puede eliminar.

Si el estadístico F se calcula utilizando la formula general heterocedástico-robusta, su distribución en muestras grandes bajo la hipótesis nula es $F_{q,\infty}$ independientemente de si los errores son homocedásticos o heterocedástico.

El p-valor del estadístico F puede calcularse utilizando la aproximación de su distribución para mues- tras grandes, tal que:

$$p-\text{valor} = Pr(F_{q,\infty} > F^{act})$$

El estadístico F general contrasta la hipótesis conjunta de que todos los coeficientes de las pendientes son cero. Es decir, la hipótesis nula y la hipótesis alternativa son: $H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$ vs. $H_1 : \beta_j \neq 0$ para al menos una $j; j = 1, ..., k$ hipótesis nula, ninguna de las variables explicativas explica nada de la variabilidad de Y_i , aunque el término independiente (que bajo la hipótesis nula es la media de Y_i) puede ser distinto de cero.

- Cuando q=1, el estadístico F contrasta una única restricción. Entonces la hipótesis nula conjunta se reduce a la hipótesis nula sobre un solo coeficiente de regresión y el estadístico F es el estadístico t.
- Un estadístico F elevado debería estar asociado con un aumento sustancial en el R2. Si el término de error es homocedástico, el estadístico F puede expresarse en términos de mejora en el ajuste de la regresión, medida ya sea por la disminución de la suma de los cuadrados de los residuos o bien por el aumento del R^2 de la regresión. El estadístico F resultante se conoce como el estadístico F válido con homocedasticidad, porque solamente es válido si el término de error es homocedástico.

Por el contrario, el estadístico F heterocedástico-robusto válido tanto si el término de error es homocedástico como si es heterocedástico. A pesar de esta limitación significativa del estadístico F válido con homocedasticidad, su sencilla fórmula puede ser calculada utilizando los resultados estándar de la regresión. El estadístico F válido con homocedasticidad se calcula utilizando una fórmula sencilla basada en la suma de los cuadrados de los residuos de dos regresiones. En la primera regresión, denominada regresión restringida, se impone el cumplimiento de la hipótesis nula. Cuando la hipótesis nula es aquella en la que todos los valores de la hipótesis son cero, los regresores relevantes se excluyen de la regresión. En la segunda regresión, denominada regresión sin restringir, la hipótesis alternativa se considera cierta. Si la suma de los cuadrados de los residuos es lo suficientemente más pequeña en la regresión sin restringir, libre, que en la regresión restringida, entonces el contraste rechaza la hipótesis nula. El estadístico F válido con homocedasticidad está dado por la fórmula:

$$F = \frac{\left(SR_{restringida} - SR_{sinrestringir}\right)/q}{SR_{sinrestringir}/(n - k_{sinrestringir} - 1)}$$

Alternativamente esta la fórmula en términos del R^2:

$$F = \frac{\left(R^2_{sinrestringir} - R^2_{restringida}\right)/q}{(1 - R^2_{sinrestringir})/(n - k_{sinrestringir} - 1)}$$

Si los errores son homocedásticos, entonces la diferencia entre el estadístico F válido con homocedasticidad y el estadístico F heterocedástico robusto se desvanece cuando el tamaño de la muestra, n, aumenta. Por tanto, si los errores son homocedásticos, la distribución muestral del estadístico F válido con homocedasticidad bajo la hipótesis nula es $F_{q,\infty}$ para muestras grandes.

Si los errores son homocedásticos y se distribuyen normales i.i.d., entonces el estadístico F válido con homocedasticidad tiene una distribución exacta $F_{q,n-k_{sinrestringir}-1}$ bajo la hipótesis nula. Este converge a una distribución $F_{q,\infty}$ cuando n tiende a infinito.

8.3 Constraste de una sola restricción sobre varios coeficientes

Por ejemplo si tenemos el caso de que $\beta_1 = \beta_2$, es decir:

$$H_0 : \beta_1 = \beta_2 \quad vs \quad H_1 : \beta_1 \neq \beta_2$$

En este caso se pueden aplicar dos métodos, el primero implica contrastar la rescricción directamente utilizando el estadístico F con una distribución $F_{1,\infty}$. El segundo método implica transformar el modelo de regresión para que la restricción se refiera a un único coeficiente. En este caso, podemos restar y sumar $\beta_2 X_{1i}$ de ambos lados de la ecuación de regresión, obteniendo:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i && (8.3)\\ &= \beta_0 + \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{2i} + \beta_2 X_{1i} + u_i && (8.4)\\ &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i && (8.5)\\ &= \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i && (8.6) \end{aligned}$$

Ahora puede plantearse la hipótesis nula como $H_0 : \gamma_1 = 0$ y contrastarse utilizando el estadístico t habitual. En la práctica esto se realiza primero construyendo el nuevo regresor $W_i = X_{1i} + X_{2i}$ y posteriormente estimando la regresión:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

En general, es posible tener q restricciones bajo la hipótesis nula en las que algunas o todas estas restricciones implican a varios coeficientes. El estadístico F es extensible a este tipo de hipótesis conjun- tas, y puede calcularse por cualquiera de los dos métodos que acabamos de mencionar. La mejor manera de hacer esto en la práctica depende del software de regresión que en concreto se utilice

8.4 Conjuntos de confianza para varios coeficientes

la generalización a dos o más coeficientes de un intervalo de confianza para un único coeficiente. La fórmula del conjunto de confianza para un número arbitrario de coeficientes se basa en la fórmula para el estadístico F. Cuando hay dos coeficientes, los conjuntos de confianza resultantes son elipses.

8.5 Especificación del modelo en regresión múltiple

Para determinar que variables incluir en regresión múltiple debemos considerar las posibles fuentes del sesgo de variable omitida. Se debe centrar en obtener una estimación insesgada del efecto causal de interés y no basarse únicamente en estadísticas de ajuste, como R^2 o R^2 ajustado.

Concepto Clave 7.3: Sesgo de variable omitida en regresión múltiple

El sesgo de variable omitida es el sesgo en el estimador MCO que aparece cuando uno o más regresores incluidos están correlacionados con una variable omitida. Para que surja el sesgo de variable omitida deben cumplirse dos cosas:

1. Al menos uno de los regresores incluidos debe estar correlacionado con la variable omitida.
2. La variable omitida debe ser un factor determinante de la variable dependiente, Y_i.

Si se cumplen las dos condiciones para el sesgo de variable omitida, entonces al menos uno de los regresores esta correlacionado con el termino de error. Esto significa que la esperanza condicional de u_i dados X_i es distinta de cero, por lo que se viola el primer supuesto de mínimos cuadrados. Como consecuencia, el sesgo de variable omitida persiste incluso si el tamaño de muestra es grande, por ende, los estimadores MCO son inconsistentes), y los errores estándar son incorrectos.

Una variable de control no es el objeto de interés del estudio; sino que es un regresor incluido para mantener constantes los factores que, si se descuidan, podrían llevar a que la estimación del efecto causal de interés presente sesgo de variable omitida. Los supuestos de mínimos cuadrados de la regresión múltiple consideran los regresores simétricamente. En este apartado, se presenta una alternativa a los supuestos de mínimos cuadrados en la que la distinción entre una variable de interés y una variable de control es explícita. Si se cumple este supuesto alternativo, el estimador de MCO del efecto de interés es insesgado, pero los coeficientes MCO de las variables de control serán, en general, sesgados y no tendrán una interpretación causal.

La distinción entre variables de interés y variables de control puede ser establecida de forma matemáticamente precisa reemplazando el primer supuesto de mínimos cuadrados, es decir, el supuesto de esper- anza condicional igual a cero, por un supuesto denominado independencia de la media condicional. Consideremos una regresión con dos variables, en la cual $X_{1,i}$ es la variable de interés y $X_{2,i}$ es la vari- able de control. La independencia en media condicional requiere que la esperanza condicional de u_i dados $X_{1,i}$ y $X_{2,i}$ no dependa de $X_{1,i}$, aunque pueda depender de $X_{2,i}$. Es decir, $E(u_i|X_{1,i}, X_{2,i}) = E(u_i|X_{2,i})$ independencia en media condicional.

La idea de la **independencia en media condicional** es que una vez que se controla $X_{2,i}$, $X_{1,i}$ puede ser tratada como si estuviera asignada al azar, en el sentido de que la media condicional del término de error ya no depende de $X_{1,i}$. La inclusión de $X_{2,i}$ como variable de control hace que $X_{1,i}$ no esté correlacionada con el término de error por lo que MCO puede estimar el efecto causal sobre $Y_{1,i}$ de un cambio en $X_{1,i}$. La variable de control, sin embargo, sigue estando correlacionada con el término de error, por lo que el coeficiente de la variable de control está sujeto al sesgo de variable omitida y no tiene una interpretación causal.

La variable de control $X_{2,i}$ se incluye debido a que tiene en cuenta (controla) los factores omitidos que afectan a Y_i y están correlacionados con $X_{1,i}$ y debido a que podría (aunque no necesariamente) tener un efecto causal por sí misma. Por tanto, el coeficiente de $X_{1,i}$ es el efecto sobre Y_i de $X_{1,i}$, utilizando la variable de control $X_{2,i}$ tanto para mantener constante el efecto directo de $X_{2,i}$ como para controlar por los factores correlacionados con $X_{2,i}$. lo habitual es simplemente decir que el coeficiente de $X_{1,i}$ es el efecto sobre Y_i, controlando por $X_{2,i}$.

En teoría, cuando se dispone de datos sobre la variable omitida, la solución para el sesgo de variable omitida es incluirla. En la práctica, la decisión de incluir una variable en particular puede ser difícil y requiere una valoración. Nuestro sistema para el problema del sesgo potencial de variable omitida es doble. En primer lugar, debería elegirse un conjunto central o conjunto base de variables explicativas mediante una combinación de una opinión experta, la teoría económica, y el conocimiento de cómo fueron recogidos los datos, siendo esta regresión la especificación base. Esta debería contener las variables de interés principal y las variables de control sugeridas por la opinión experta y la teoría económica. Sin embargo, rara vez resultan decisivas. Por lo tanto, el siguiente paso es desarrollar una lista de especificaciones alternativas candidatas, es decir, conjuntos alternativos de regresores. Si las estimaciones de los coeficientes de interés son numéricamente similares entre las especificaciones alternativas, esto proporciona evidencia de que las estimaciones de la especificación base son fiables. Si, por otro lado, las estimaciones de los coeficientes de interés varían sustancialmente, indica que la original presenta sesgo de variable omitida.

Nota: Si bien estos estadísticos (F y R^2) aumentan cada vez que se agrega un regresor, no implica que este sea significativo. Y que estos estadísticos sean elevados, no implica verdadera causalidad de la variable dependiente, ni ausencia de sesgo de variable omitida. Tampoco indica que las variables elegidas sean las mejores.

Concepto Clave 7.4: R^2 y R^2 ajustado, qué nos dicen y qué no

El R^2 y el R^2 ajustado **nos dicen** si los regresores son buenos para predecir, o «explicar» los valores de la variable dependiente en la muestra de datos disponible. Si el R^2 (o el R^2 ajustado) está cerca de 1, entonces los regresores proporcionan buenas predicciones sobre la varia- ble dependiente en esa muestra, en el sentido de que la varianza de los residuos MCO es pequeña comparada con la varianza de la variable dependiente. Si el R^2 (o el R^2 ajustado) está cercano a 0, es cierto todo lo contrario.
El R^2 y el R^2 ajustado **no nos dicen** si:

1. Una variable incluida es estadísticamente significativa.
2. Los regresores son la verdadera causa de los movimientos de la variable dependiente.
3. Existe un sesgo de variable omitida: que estos indicadores sean elevados no implica que no exista sesgo de variable omitida.
4. Se ha elegido el conjunto más adecuado de regresores: tanto si el valor es alto como si es bajo, no implica que se hayan elegido las mejores variables explicativas.

Capítulo 9

Funciones de Regresión no Lineales

Hasta ahora suponíamos que la función de regresión poblacional era lineal, por lo que la pendiente era constante (el efecto de los cambios en x sobre y no dependían del valor de x).

En este capítulo se desarrollan dos grupos de métodos para la detección y modelización de funciones de regresión poblacionales no lineales. Los métodos del primer grupo son útiles cuando el efecto sobre Y de un cambio en una variable X_1, depende del valor de X_1 en sí misma. Los métodos del segundo grupo resultan útiles cuando el efecto sobre Y de un cambio en X_1 depende del valor de otra variable independiente, digamos X_2.

En estos casos donde la función de regresión poblacional es una función no lineal de las variables independientes, la esperanza condicional $E(Y_i|X_{1i},\ldots,X_{ki})$ es una función no lineal de una o más de las X. A pesar de que son no lineales en las X, estos modelos son funciones lineales de los coeficientes desconocidos (o parámetros) del modelo de regresión poblacional, y por tanto son versiones del modelo de regresión múltiple. Por tanto, los parámetros desconocidos de estas funciones de regresión no lineales pueden estimarse y contrastarse utilizando MCO.

9.1 Estrategia general para la modelización de funciones de regresión no lineales

Ejemplo: Un modelo de regresión poblacional cuadrática que relacione las calificaciones en los exámenes y la renta puede expresarse matemáticamente como:

$$\text{CalificaciónExamen}_i = \beta_0 + \beta_1 \text{Renta}_i + \beta_2 \text{Renta}_i^2 + u_i$$

donde β_0, β_1 y β_2 son coeficientes, Renta_i es la renta del distrito i-ésimo, Renta_i^2 es el cuadrado de la renta del distrito i-ésimo, y u_i es un término de error que, como es habitual, representa todos los otros factores que determinan las calificaciones en los exámenes. La Ecuación se denomina modelo de regresión cuadrática porque la función de regresión poblacional, $E(\text{CalificaciónExamen}_i|\text{Renta}_i) = \beta_0 + \beta_1 \text{Renta}_i + \beta_2 \text{Renta}_i^2$, es una función cuadrática de la variable independiente Renta.

Si se conociesen los coeficientes poblacionales β_0, β_1 y β_2 de la Ecuación, se podría predecir la calificación en el examen de un distrito en base a su renta media. Sin embargo, estos coeficientes poblacionales son desconocidos, y por lo tanto, deben estimarse a partir de una muestra de datos.

Dado que esta ecuación en realidad es una versión del modelo de regresión múltiple con dos variables explicativas sus coeficientes poblacionales desconocidos se pueden estimar y contrastar mediante los métodos MCO descritos en los Capítulos anteriores.

Contrastes de hipótesis: Para contrastar si la función esta especificada correctamente se puede probar la hipótesis nula de $\beta_2 = 0$ contra su alternativa, mediante el estadístico t.

Los modelos de regresión poblacional no lineales con los que trabajamos son de forma:

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{ki}) + u_i \tag{9.1}$$

donde $f(X_{1i}, X_{2i}, \ldots, X_{ki})$ es la función de regresión no lineal poblacional.

Concepto clave 8.1: El efecto esperado en Y de un cambio en X_1 en un modelo de regresión no lineal

La variación esperada en Y, ΔY, asociada con una variación en X_1, ΔX_1, manteniendo constantes X_2, ..., X_k, es la diferencia entre el valor de la función de regresión poblacional antes y después de la variación de X_1, manteniendo constantes X_2, ..., X_k. Es decir, la variación esperada en Y es la diferencia:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$$

El estimador de esta diferencia poblacional desconocida es la diferencia entre los valores esperados para estos dos casos. Sea $\hat{f}(X_1, X_2, \ldots, X_k)$ el valor esperado de Y basado en el estimador $\hat{f}$ de la función de regresión poblacional. Entonces la variación esperada en Y es:

$$\Delta \bar{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k)$$

El estimador del efecto sobre Y de un cambio en X_1 depende del estimador de la función de regresión poblacional, $\hat{f}$, que varía de una muestra a otra. Por tanto, el efecto estimado contiene un error de muestreo. Una forma de cuantificar la incertidumbre en el muestreo asociada al efecto estimado es calcular un intervalo de confianza para el verdadero efecto poblacional. Para hacerlo, es necesario calcular el error estándar de $\Delta \hat{Y}$.

Existen dos métodos para hacerlo utilizando el software de regresión habitual, que se corresponden con los dos métodos del capítulo 8 para contrastar una restricción única con varios coeficientes. El primer método es el de utilizar el «Método 1», que consiste en calcular el estadístico F para contrastar la hipótesis. El error estándar de $\Delta \hat{Y}$ está dado por:

$$ES(\Delta \hat{Y}) = \frac{|\Delta \hat{Y}|}{\sqrt{\frac{F}{v}}}$$

donde v es el número de restricciones (en este caso, v = 1).

El segundo método consiste en utilizar el «Método 2», lo que implica la transformación de las variables explicativas de modo que, en la regresión transformada.

En los modelos no lineales la función de regresión se interpreta mejor mediante su representación gráfica y mediante el cálculo del efecto esperado sobre Y de la variación de una o más variables indepen- dientes. La interpretación no suele ser la natural y habitual de los modelos lineales.

El método general para modelizar las funciones de regresión no lineales consta de cinco pasos:

1. Identificación de una posible relación no lineal: basado en la teoría económica y los conocimientos sobre el caso.
2. Especificación de una función no lineal y estimación de sus parámetros por MCO
3. Determinación de si el modelo no lineal mejora el modelo lineal: debe determinarse empíricamente, pueden utilizarse los estadísticos t y F para contrastar la hipótesis nula de que la función de regresión poblacional es lineal frente a la alternativa de que no lo es
4. Pre presentación de la función de regresión no lineal: bosquejo sobre el diagrama de dispersión
5. Estimación del efecto sobre Y de un cambio en X: se utiliza la regresión estimada para calcular el efecto sobre Y de una variación en uno o más regresores X.

9.2 Funciones no lineales de una sola variable independiente

Polinomios

se especifica la función de regresión mediante un polinomio en X. el modelo de regresión polinomial de grado r es:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

Cuando $r = 2$ es cuadrática, si $r = 3$ es cúbica. Aquí los regresores son potencias de la misma variable independiente X. se aplican las mismas técnicas de regresión múltiple. Para contrastar la hipótesis nula de que la regresión poblacional es lineal haríamos:

$$H_0 : \beta_2 = 0, \beta_3 = 0, ..., \beta_r = 0 \quad vs \quad H_1 : alguna \quad \beta_j \quad distinta \quad de \quad 0$$

La hipótesis nula tendrá $q = r - 1$ restricciones y utilizaremos el estadístico F.

Para determinar el grado del polinomio se puede hacer una prueba de hipótesis secuencial, que consiste en elegir el valor máximo para r y utilizar el estadístico t para contrastar $\beta_r = 0$. Si no se rechaza, eliminamos X^r de la regresión y contrastamos β_{r-1}. Continuamos de este modo hasta que el coeficiente de mayor potencia sea estadísticamente significativo.

Logaritmos

Otra forma es con logaritmos naturales de Y y/o X. estos convierten las variaciones de las variables en cambios porcentuales. Se debe tener en cuenta que la función logarítmica solo está definida para valores positivos de x.

El vínculo entre el logaritmo y los porcentajes se basa en que cuando el incremento de x es pequeño, la diferencia entre el logaritmo de $x + \Delta x$ y el logaritmo de x es aproximadamente la variación porcentual entre 100.

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x} \quad \text{si } \frac{\Delta x}{x} \text{ es pequeño}$$

Existen tres casos distintos en los que pueden utilizarse logaritmos: cuando se transforma X tomando sus logaritmos pero no Y, cuando se transforma Y tomando su logaritmo pero no X, y cuando tanto Y como X se transforman en sus logaritmos. La interpretación de los coeficientes de la regresión es diferente en cada caso.

En el primer caso $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$, $i = 1, ..., n$. Debido a que Y no está expresada en logaritmos pero X si lo está, se lo conoce como el modelo lineal-log. En este una variación del 1 % en X está asociada con un cambio en Y de $0,01\beta_1$. Para comprobarlo:

$$\begin{aligned}[\beta_0 + \beta_1 \ln(X - \Delta X)] - [\beta_0 + \beta_1 \ln(X)] &= \beta_1 [\ln(X + \Delta X) - \ln(X)] \\ &\approx \beta_1 (\Delta X / X) = \beta_1 (0,01) \text{ si } \Delta X / X = 0,01\end{aligned}$$

Para estimar los coeficientes β_0 y β_1 primero se calcula una nueva variable, $\ln(x)$, y luego se estiman mediante la regresión MCO de la variable Y_i sobre $\ln(x)$. las hipótesis se pueden contrastar utilizando el estadístico t.

En el segundo caso $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$, $i = 1, ..., n$. Debido a que Y está expresada en logaritmos pero X no lo está, se lo conoce como el modelo log-lineal. En este una variación de una unidad en X está asociada con un cambio porcentual en Y de aproximadamente $100\beta_1$. Para comprobarlo:

$$[\ln(Y + \Delta Y) - \ln(Y)] = \beta_1 \Delta X \quad \Rightarrow \quad \ln(Y + \Delta Y) - \ln(Y) \approx \frac{\Delta Y}{Y} \approx \beta_1 \Delta X \text{ si } \Delta X = 1$$

En porcentajes un cambio unitario en X está asociado con un cambio porcentual en Y de aproximadamente $100\beta_1$.

Para estimar los coeficientes β_0 y β_1 primero se calcula una nueva variable, $\ln(Y)$, y luego se estiman mediante la regresión MCO de la variable $\ln(Y_i)$ sobre X. las hipótesis se pueden contrastar utilizando el estadístico t.

En el tercer caso $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$, $i = 1, ..., n$. Debido a que tanto Y como X están expresadas en logaritmos, se lo conoce como el modelo log-log. En este una variación del 1 % en X está asociada con un cambio porcentual en Y de aproximadamente $100\beta_1$. Para comprobarlo:

$$\begin{aligned}[\ln(Y + \Delta Y) - \ln(Y)] &= \beta_1 [\ln(X + \Delta X) - \ln(X)] \approx \beta_1 (\Delta X / X) \\ \frac{\Delta Y}{Y} &\approx \beta_1 (\Delta X / X) \\ \beta_1 &\approx \frac{\Delta Y / Y}{\Delta X / X} = \frac{100(\Delta Y / Y)}{100(\Delta X / X)} = \text{elasticidad de Y con respecto a X}\end{aligned}$$

Concepto clave 8.2: Logaritmos en la regresión: 3 casos

Los logaritmos pueden utilizarse para transformar la variable dependiente Y, una variable independiente X, o ambas (pero la variable que se transforme debe ser positiva). La siguiente tabla resume estos tres casos, así como la interpretación del coeficiente de regresión b1. En cada caso, se puede estimar b1 mediante la aplicación de MCO tras haber tomado logaritmos de la variable dependiente y/o las independientes.

Caso	Ecuación	Interpretación de β_1
1	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	Una variación del 1 % en X está asociada con un cambio en Y de $0,01\beta_1$.
2	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	Una variación de una unidad en X está asociada con un cambio porcentual en Y de aproximadamente $100\beta_1$.
3	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	Una variación del 1 % en X está asociada con un cambio porcentual en Y de aproximadamente β_1.

Se puede usar el R^2 ajustado para comparar los modelos log-lineal y log-log, así como el lineal-log con la regresión lineal de Y sobre x. sin embargo, no se puede usar para comparar el lineal-log con el log-log tal que sus variables dependientes son diferentes. Debido a este problema, lo mejor que se puede hacer en cada caso particular es decidir si tiene sentido especificar Y en logaritmos, de acuerdo con la teoría económica y según el propio conocimiento previo del problema en cuestión, así como el de otros expertos.

Si la variable dependiente Y se ha transformado tomando logaritmos, puede utilizarse la regresión estimada para calcular directamente el valor de predicción de $\ln(Y)$. Sin embargo, es un poco más difícil de calcular el valor esperado de Y en sí mismo.

9.3 Interacciones entre variables independientes

Las interacciones entre variables independientes también se incorporan al modelo de regresión, para considerar el efecto sobre Y de un cambio en una variable independiente la cual depende del valor de otra variable independiente.

Interacción entre dos variables binarias: Consideremos la regresión poblacional del logaritmo de los ingresos salariales Y_i, donde $Y_i = \ln(\text{Ingresos}_i)$, sobre dos variables binarias: si un trabajador tiene un título universitario D_{1i}, donde $D_{1i} = 1$ si la i-ésima persona es graduada universitaria, y el género del trabajador D_{2i}, donde $D_{2i} = 1$ si la i-ésima persona es de sexo femenino. La regresión lineal poblacional de Y_i sobre estas dos variables binarias es:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

Esta tiene la limitación de que no considera cómo afecta el género en el caso de tener o no título al salario, de manera que el valor en el mercado de un grado universitario podría ser diferente para hombres y mujeres. Para corregir esto se puede modificar la regresión introduciendo otro regresor que sea el producto de las variables binarias, $D_{1i} \times D_{2i}$. La regresión resultante es:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

El nuevo regresor se denomina **término de interacción** o regresor de interacción y el modelo de regresión que lo incorpora se denomina modelo de regresión con interacciones de variables binarias.

Para medir el efecto poblacional sobre el logaritmo de los ingresos salariales Yi de tener un título universitario dependiendo del género, debemos calcular la esperanza condicional de Yi para D1i = 0, dado un valor de D2i, y luego la esperanza condicionada a D1i = 1 para el mismo valor de D2i. Luego hacemos la diferencia de estas:

$$\begin{aligned} E(Y_i|D_{1i} = 1, D_{2i}) - E(Y_i|D_{1i} = 0, D_{2i}) &= [\beta_0 + \beta_1(1) + \beta_2 D_{2i} + \beta_3(1 \times D_{2i})] \\ &\quad - [\beta_0 + \beta_1(0) + \beta_2 D_{2i} + \beta_3(0 \times D_{2i})] \\ &= \beta_1 + \beta_3 D_{2i} \end{aligned}$$

Esto arroja como resultado que si la persona es de sexo masculino ($D_{2i} = 0$) el efecto del título es β_1, y si es femenino es $\beta_1 + \beta_3$.

Interacciones entre una variable continua y una variable binaria: Consideremos ahora la regresión poblacional del logaritmo de los ingresos $Y_i = \ln(\text{Ingresos}_i)$ sobre una variable continua, los años de experiencia laboral de una persona X_i, y una variable binaria, si el trabajador tiene un título universitario D_i, donde $D_i = 1$ si la i-ésima persona es graduada universitaria. La recta de regresión poblacional que relaciona Y y la variable continua X puede depender de la variable binaria D de tres formas diferentes:

1. Puede ser que solo difieran en sus interceptos, donde el modelo será:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

   Cuando $D_i = 0$ es solo $\beta_0 + \beta_1 X_i$, pero cuando $D_i = 1$ la función es $\beta_0 + \beta_1 X_i + \beta_2$, por lo que el intercepto es $\beta_0 + \beta_2$.
2. Puede ser que difieran en pendientes e intercepto. Para esto se añade un término de interacción:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

   Entonces cuando $D_i = 0$, la función es $\beta_0 + \beta_1 X_i$, mientras que si $D_i = 1$ es $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$.
3. Por último, puede ser que tengan diferentes pendientes pero el mismo intercepto. En ese caso:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$

Las tres especificaciones son versiones del modelo de regresión múltiple, y una vez que se ha creado una nueva variable Xi · Di, los coeficientes de todos ellos pueden estimarse mediante MCO.

Concepto clave 8.4: Interacciones entre variables binarias y continuas

Mediante el uso del término de interacción $X_i \times D_i$, la recta de regresión poblacional que relaciona Y_i con la variable continua X_i puede tener una pendiente que dependa de la variable binaria D_i. Existen tres posibilidades:

1. Diferentes interceptos pero misma pendiente:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

2. Diferentes interceptos y diferentes pendientes:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

3. Mismo intercepto pero diferentes pendientes:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$

Interacción entre dos variables continuas: Supongamos ahora que ambas variables independientes (X_{1i} y X_{2i}) son continuas. Un ejemplo de ello es cuando Yi es el logaritmo de los ingresos salariales del trabajador i-ésimo, X_{1i} son sus años de experiencia laboral, y X_{2i} es el número de años que él o ella fueron a la escuela. La interacción entre las dos variables se puede modelizar incluyendo el término de interacción:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

El término de interacción permite que el efecto de un cambio unitario en X_{1i} dependa de X_{2i}, el efecto sobre Y de un cambio en X_{1i}, manteniendo constante X_{2i}, es $\beta_1 + \beta_3 X_{2i}$.

El efecto sobre Y de un cambio en X_{2i}, manteniendo constante X_{1i}, es $\beta_2 + \beta_3 X_{1i}$.

Colocando juntos estos dos efectos se muestra que el coeficiente β_3 del término de interacción es el efecto de un aumento unitario en X_{1i} y en X_{2i}, mucho más allá de los efectos de un cambio unitario solamente en X_{1i}

y un aumento unitario en X_{2i} en solitario. Es decir, si X_{1i} cambia en ΔX_{1i} y X_{2i} cambia en ΔX_{2i}, entonces el cambio esperado en Y es:

$$\Delta Y = (\beta_1 + \beta_3 X_{2i})\Delta X_{1i} + (\beta_2 + \beta_3 X_{1i})\Delta X_{2i} + \beta_3(\Delta X_{1i} \times \Delta X_{2i})$$

El primer término es el efecto que proviene del cambio en X_{1i} manteniendo constante X_{2i}; el segundo término es el efecto que proviene del cambio en X_{2i} manteniendo constante X_{1i}; y el último término es el efecto extra del cambio tanto en X_{1i} como en X_{2i}.

Cuando las interacciones se combinan con transformaciones logarítmicas, pueden utilizarse para estimar las elasticidades precio cuando las elasticidades precio dependen de las características del bien.

Capítulo 10

Evaluación de Estudios Basados en Regresión Múltiple

Un estudio es válido internamente si sus inferencias estadísticas acerca de los efectos causales son válidas para la población y el escenario estudiados; es válido externamente si sus inferencias pueden generalizarse a otras poblaciones y escenarios.

10.1 Validez interna y externa

La población estudiada es la población de individuos de los cuales se extrajo la muestra. La población para la cual los resultados se generalizan, o población de interés, es la población de entidades individuales para la que se van a aplicar las inferencias causales del estudio. Con «escenario», nos referimos al entorno institucional, legal, social y económico.

La validez interna tiene 2 componentes . En primer lugar, el estimador del efecto causal debe ser insesgado y consistente. En segundo lugar, los contrastes de hipótesis deben tener el nivel de significación deseado (la tasa de rechazo efectiva del contraste bajo la hipótesis nula debe ser igual al nivel de significación deseado), y los intervalos de confianza deben tener el nivel de confianza deseado.

En un estudio basado en la regresión MCO, los requisitos para la validez interna son que el estimador MCO sea insesgado y consistente, y que los errores estándar se calculen de una manera que haga que los intervalos de confianza presenten el nivel de confianza deseado. Por diferentes razones estos requisitos podrían no cumplirse, y estas razones constituyen **amenazas a la validez interna**. Estas amenazas conducen a incumplimientos de uno o más de los supuestos de mínimos cuadrados.

Las posibles amenazas a la validez externa surgen de las diferencias entre la población y el escenario estudiado y la población y el escenario de interés. el verdadero efecto causal puede no ser el mismo en la población estudiada y en la población de interés. Esto podría deberse a que la población fue elegida de una manera que la hace diferente de la población de interés, por las diferencias en las características de la población, las diferencias geográficas, o bien debido a que el estudio no está actualizado. Incluso aunque la población estudiada y la población de interés sean la misma, tal vez no sea posible generalizar los resultados del estudio si los escenarios son distintos, podría haber diferencias en el entorno institucional, diferencias en las leyes o diferencias en el entorno físico. Cuanto más cercanos a la población y al escenario del estudio se encuentren la población y escenario de interés, más fuertes serán las razones para la validez externa.

La validez externa debe ser juzgada mediante el conocimiento específico de las poblaciones y los escenarios estudiados y los de interés. Las diferencias importantes entre ellos pondrán en tela de juicio la validez externa del estudio. A veces existen dos o más estudios sobre poblaciones diferentes, pero relacionadas. Si es así, la validez externa de ambos estudios se puede comprobar mediante la comparación de sus resultados. En general, las conclusiones similares en dos o más estudios impulsan las razones para la validez externa, mientras que las diferencias en sus resultados que no resulten fácilmente explicables ponen en duda su validez externa.

10.2 Amenazas a la validez interna del análisis de regresión múltiple

Los estudios basados en el análisis de regresión son internamente válidos si los coeficientes de regresión estimados son insesgados y consistentes, y si sus errores estándar proporcionan intervalos de confianza con el nivel de confianza deseado. Un estimador MCO puede ser sesgado incluso en muestras grandes por varias razones, siendo estas variables omitidas, errores de especificación de la forma funcional de la función de regresión, medición imprecisa de las variables independientes, selección muestral y causalidad simultánea.

Sesgo de variable omitida

El sesgo de variable omitida se produce cuando se omite de la regresión una variable que determina Y y que esta correlacionada con uno o mas de los regresores incluidos en esa misma regresión.

- *Soluciones para el sesgo de variable omitida cuando la variable es observable o bien existen variables de control adecuadas:* si se dispone de la información, se puede incluir la variable en la regresión múltiple solucionando el problema. si se dispone de datos sobre una o más variables de control, y si esas variables de control son adecuadas en el sentido de que conducen a la independencia en media condicional, entonces la inclusión de las variables de control elimina el posible sesgo en el coeficiente de la variable de interés. La adición de otra variable presenta costes y beneficios. Y si se incluye una variable cuando no corresponde, se reducirá la precisión de los estimadores de los otros coeficientes de la regresión. En la práctica, existen cuatro pasos que pueden ayudar a decidir si se incluye una variable o un conjunto de variables en una regresión.
- *Soluciones al sesgo de variable omitida cuando no se dispone de variables de control adecuadas:* no es opción si no hay datos o no es adecuado. Podemos solucionar el sesgo de variable omitida de 3 formas: utilizando datos en los que se observa la misma unidad observacional en diferentes momen- tos del tiempo (datos de panel), utilizando la regresión de variables instrumentales, o utilizando un diseño de estudio en el que el efecto de interés se estudie mediante un experimento aleatorizado controlado.

Concepto clave 9.2: Sesgo de variable omitida; ¿deberían incluirse más variables en la regresión?

Si se incluye otra variable en la regresión múltiple, se eliminará la posibilidad del sesgo de variable omitida que pueda surgir al excluir esa variable, pero la varianza de los estimadores de los coeficientes de interés puede aumentar. Se ofrecen aquí algunas pautas que pueden ayudar a decidir si se debe incluir una variable adicional:

1. Ser específico acerca del coeficiente o coeficientes de interés.
2. Utilizar un razonamiento a priori para identificar las fuentes potenciales más impor- tantes de sesgo de variable omitida, lo que lleva a una especificación base y a algunas variables «cuestionables».
3. Contrastar si otras variables de control «cuestionables» tienen un coeficiente distinto de cero.
4. Proporcionar tablas que representen los resultados «de divulgación completa» para que otros puedan ver el efecto de la inclusión de las variables cuestionables sobre el(los) coeficiente(s) de interés. ¿Cambian los resultados si se incluye una variable de control cuestionable?

Error en la especificación de la forma funcional de la regresión

Si se estima una regresión lineal pero la verdadera función de regresión poblacional no es línea estaremos ante el error de especificación de la forma funcional, que provoca que el estimador MCO sea sesgado.

- *Soluciones para el error de especificación de la forma funcional:* Cuando la variable dependiente es continua se puede resolver mediante los métodos del capítulo 9.

Sesgo de errores de medida y por errores en las variables

El sesgo por errores en las variables se origina en un error de medición en las variables independientes. Este sesgo persiste incluso en muestras muy grandes, por lo que el estimador MCO es inconsistente si existe error de medición. Existen muchas fuentes posibles de error de medición. Si los datos son recogidos a través de una encuesta, el encuestado puede dar una respuesta equivocada. Si en su lugar los datos se obtienen de los

registros administrativos informatizados, podría haber habido errores tipográficos, cuando se introdujeron los datos, etc.

Para comprobar que los errores en las variables pueden dar lugar a la existencia de correlación entre el regresor y el término de error, supongamos que existe un único regresor X_i pero que X_i está medido de forma imprecisa mediante $\tilde{X}_i$. Debido a que se observa $\tilde{X}_i$ pero no X_i, la ecuación de regresión estimada en realidad es la que se basa en $\tilde{X}_i$.

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + (u_i + \beta_1(X_i - \tilde{X}_i)) = \beta_0 + \beta_1 \tilde{X}_i + v_i$$

Donde $v_i = u_i + \beta_1(X_i - \tilde{X}_i)$ es el nuevo término de error. Si el error de medición $X_i - \tilde{X}_i$ está correlacionado con $\tilde{X}_i$, entonces $\tilde{X}_i$ estará correlacionado con v_i, y el estimador β_1 será sesgado e inconsistente, y la cuantía del sesgo dependerá de el error de medición.

Supongamos que el valor medido es igual al verdadero valor no medible más un componente puramente aleatorio W_i, que tiene media igual a cero y varianza σ_w^2 Debido a que el error es puramente aleatorio, podríamos suponer que w_i no esta correlacionado con X_i ni con el error de la regresión u_i. Este supuesto es el modelo clásico de error de medición en el que $\tilde{X}_i = X_i + w_i$, en el que $corr(w_i, X_i) = 0$ y β_1 tiene el límite de probabilidad:

$$\widehat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\beta_1 \tag{10.1}$$

Demostración de este límite de probabilidad:

1. Definición del Modelo y el Error

$$\begin{aligned} \text{Modelo Verdadero:} \quad & Y_i = \beta_0 + \beta_1 X_i + u_i \\ \text{Medición Observada:} \quad & \tilde{X}_i = X_i + w_i \quad (\text{donde } w_i \text{ es ruido blanco}) \end{aligned}$$

2. Sustitución en la Ecuación Estimada

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(\tilde{X}_i - w_i) + u_i \\ Y_i &= \beta_0 + \beta_1 \tilde{X}_i + \underbrace{(u_i - \beta_1 w_i)}_{v_i} \end{aligned}$$

3. Límite de Probabilidad de MCO Sabemos que $\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{Cov(\tilde{X}_i, v_i)}{Var(\tilde{X}_i)}$. Calculamos los componentes:

$$\begin{aligned} Var(\tilde{X}_i) &= \sigma_X^2 + \sigma_w^2 \quad (\text{por independencia de } X \text{ y } w) \\ Cov(\tilde{X}_i, v_i) &= Cov(X_i + w_i, u_i - \beta_1 w_i) \\ &= -\beta_1 Cov(w_i, w_i) = -\beta_1 \sigma_w^2 \end{aligned}$$

4. Resultado Final

$$\begin{aligned} \hat{\beta}_1 &\xrightarrow{p} \beta_1 + \frac{-\beta_1 \sigma_w^2}{\sigma_X^2 + \sigma_w^2} \\ \hat{\beta}_1 &\xrightarrow{p} \beta_1 \left(1 - \frac{\sigma_w^2}{\sigma_X^2 + \sigma_w^2}\right) \\ \hat{\beta}_1 &\xrightarrow{p} \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\right) \end{aligned}$$

Entonces, dado que el cociente de las varianzas es menor a 1, si el error de medición es muy grande converge a cero. Y si es muy pequeño converge al valor de la variable.

El efecto del error de medición en Y es diferente del error de medición en X. Si Y presenta un error de medición clásico, entonces este error de medición aumenta la varianza de la regresión y de β_1, pero no induce sesgo en β_1.

Error de medición en Y: Si Y presenta un error de medición clásico, este aumenta la varianza de la regresión y de β_1, pero no induce sesgo en β_1.

Supongamos que la medida de Y_i es $\tilde{Y}_i = Y_i + w_i$, donde w_i es un error de medición con media cero, varianza σ^2_w, entonces el modelo de regresión estimado es:

$$\tilde{Y}_i = \beta_0 + \beta_1 X_i + (u_i + w_i) = \beta_0 + \beta_1 X_i + v_i$$

Si w_i y X_i se distribuyen de forma independiente, entonces $E(w_i|X_i) = 0$, y por lo tanto $E(v_i|X_i) = E(u_i|X_i) = 0$, lo que implica que no hay sesgo en β_1. Sin embargo, dado que la varianza de v_i es mayor que la de u_i, la varianza de β_1 también es mayor.

Concepto clave 9.4: Sesgo por errores en las variables

El sesgo por errores en las variables en el estimador MCO se produce cuando una variable independiente se mide de forma imprecisa. Este sesgo depende de la naturaleza del error de medida y persiste incluso si el tamaño de la muestra es grande. Si la variable medida es igual al valor real, más un error de medición con media igual a cero, que está independientemente distribuido, entonces el estimador MCO en una regresión con una sola variable en su parte derecha está sesgado hacia cero, y su límite de probabilidad está dado por la Ecuación (10.1).

- *Soluciones para el sesgo por errores en las variables:* a mejor manera de resolver el problema de los errores en las variables consiste en obtener una medida precisa de X. No obstante, si esto es imposible, se pueden utilizar métodos econométricos para mitigar el sesgo de errores en las variables. Uno de estos es la regresión de variables instrumentales. Y otro seria desarrollar un modelo matemático para el error de medición, y utilizar las formulas resultantes para ajustar las estimaciones.

Datos perdidos y selección muestral

Datoa perdidos: representan una amenaza a la validez interna dependiendo el motivo por el cual están perdidos, siendo posibles tres casos: cuando faltan de forma aleatoria, cuando la perdida de datos se basa en X y cuando faltan debido a un proceso de selección que esta relacionado con Y además de depender de X. cuando se trata del primer caso, el efecto es una reducción de la muestra, pero no introduce sesgo. Cuando la perdida de datos se basa en el valor de un regresor, el efecto será el mismo.

Por el contrario, si los datos se perdieron debido a un proceso de selección que esta relacionado con el valor de la variable dependiente, además de depender de los regresores, entonces esto puede introducir correlación entre el termino de error y los regresores. El sesgo resultante en el estimador MCO se denomina **sesgo de selección muestral**. (Soluciones al sesgo de selección: no se dan en este libro, se presentan en el capitulo 13)

Causalidad simultanea

La causalidad simultanea va hacia atrás y hacia adelante, desde la variable dependiente hacia una o mas variables. Si existe causalidad simultánea, una regresión MCO recogerá ambos efectos, por lo que el estimador MCO será sesgado e inconsistente. La causalidad simultánea conduce a la correlación entre el regresor y el término de error. Esto se puede precisar matemáticamente mediante la introducción de una ecuación adicional que describa el vínculo causal inverso. Con 2 variables:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

La primera ecuación es la habitual, donde β_1 es el efecto sobre Y de una variación en X, donde u son los otros factores. La segunda ecuación es el efecto causal inverso de Y sobre X. La causalidad simultánea conduce a la correlación entre Xi y el término de error ui en la primera ecuación. Debido a que esto puede expresarse matemáticamente mediante dos ecuaciones simultáneas, el sesgo de causalidad simultánea a veces se denomina **sesgo de ecuaciones simultáneas**.

- *Soluciones al sesgo de causalidad simultánea:* existen dos maneras, una mediante la regresión por variables instrumentales y otra diseñando y llevando a cabo un experimento aleatorizado controlado en el que se anule el canal de la causalidad inversa (temas del capitulo 14 y 15 respectivamente)

Concepto clave 9.6: Sesgo por causalidad simultánea

El sesgo por causalidad simultánea, asimismo denominado sesgo de ecuaciones simultá- neas, aparece en una regresión de Y sobre X, cuando, además del vínculo causal de interés que va desde X hacia Y, existe un vínculo causal desde Y hacia X. Esta causalidad inversa provoca que X esté correlacionado con el término de error en la regresión pobla- cional de interés.

Origen de la inconsistencia de los errores estándar MCO

Los errores estándar inconsistentes representan una amenaza diferente para la validez interna. Incluso aunque el estimador MCO sea consistente y la muestra sea grande, la inconsistencia de los errores estándar origina que los contrastes de hipótesis presenten un tamaño distinto del nivel de significación deseado, así como que los intervalos de confianza al «95 %» no incluyan al verdadero valor en el 95 % de las muestras repetidas. Existen dos razones principales para la inconsistencia de los errores estándar: un tratamiento no adecuado de la heterocedasticidad y la correlación del término de error entre observaciones.

En cuanto a la heterocedasticidad, se deben utilizar los errores estándar heterocedástico robustos y construir estadísticos F utilizando un estimador de la varianza heterocedástico robusto.

En cuanto a evitar la correlación del termino de error entre observaciones, se puede extraer los datos de una población mediante muestreo aleatorio, debido a que esto asegura que los errores esten distribuidos de forma independiente entre una observación y la siguiente.

A veces, sin embargo, el muestreo tan solo es aleatorio parcialmente. La circunstancia más común es cuando los datos son observaciones repetidas del mismo individuo en el tiempo. Si las variables omitidas que forman parte del error de regresión son persistentes, entonces se induce correlación «serial» en el error de regresión a lo largo del tiempo. La correlación serial en el término de error puede aparecer en los datos de panel y de series temporales.

Otra situación en la que el término de error puede estar correlacionado entre las distintas observaciones es cuando el muestreo está basado en una unidad geográfica. Si existen variables omitidas que reflejan las influencias geográficas, estas variables podrían dar lugar a la correlación entre los errores de regresión para observaciones adyacentes. La correlación del error de regresión entre las distintas observaciones no hace que el estimador MCO sea sesgado o inconsistente, pero viola el segundo supuesto de mínimos cuadrados. La consecuencia es que los errores estándar MCO, tanto los válidos con homocedasticidad como los heterocedástico-robustos, son incorrectos en el sentido de que no dan lugar a intervalos de confianza con el nivel de confianza deseado. En muchos casos, este problema se puede solucionar mediante el uso de una fórmula alternativa para los errores estándar.

Concepto clave 9.7: Amenazas a la validez interna de un estudio de regresión múltiple

Existen cinco amenazas principales a la validez interna de un estudio de regresión múltiple:
1. Sesgo de variable omitida.
2. Error de especificación de la forma funcional de la regresión.
3. Sesgo por errores en las variables (errores de medición en las variables explicativas).
4. Sesgo de selección muestral.
5. Sesgo por causalidad simultánea.

Cada uno de ellos, si está presente, se traduce en el incumplimiento del primer supuesto de mínimos cuadrados, $E(u_i|X_{1i},...,X_{ki}) \neq 0$, lo que a su vez significa que el estimador MCO es sesgado e inconsistente.

El cálculo incorrecto de los errores estándar representa asimismo una amenaza a la validez interna. Los errores estándar válidos con homocedasticidad no son válidos en presencia de heterocedasticidad. Si las variables no son independientes entre distintas observaciones, lo cual puede ocurrir en datos de panel y en datos de series temporales, entonces se necesita un nuevo ajuste en la fórmula de los errores estándar a fin de obtener errores estándar válidos.

La aplicación de esta lista de amenazas a un estudio de regresión múltiple constituye un método sistemático de evaluar la validez interna del estudio.

10.3 Validez interna y externa cuando la regresión se utiliza para la predicción

Cuando los modelos de regresión se utilizan para predicción, la preocupación acerca de la validez externa es muy importante, pero la preocupación acerca de la estimación insesgada de los efectos causales no lo es. Su aplicabilidad dependerá de los objetivos que se tengan, aunque el sesgo de variable omitida hace que ciertas regresiones no tengan valor para responder la cuestión de la causalidad, todavía puede ser útil para fines de pronóstico. De manera más general, los modelos de regresión pueden originar previsiones fiables, aunque sus coeficientes no tengan una interpretación causal. Este reconocimiento se encuentra detrás de la utilización de la mayoría de los modelos de regresión con fines predictivos.

Capítulo 11

Teoría de Regresión Múltiple

11.1 El modelo lineal de regresión múltiple y el estimador MCO en forma matricial

El modelo de regresión múltiple poblacional tiene la forma:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

Con la finalidad de expresar el modelo en forma matricial, se definenen los vectores y matrices:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Por lo tanto Y es $n \times 1$, X es $n \times (k+1)$, β es $(k+1) \times 1$ y u es $n \times 1$. A lo largo del capítulo se expresan las matrices y los vectores en negrita donde:

- **Y** es el vector de dimensión $n \times 1$ de las observaciones de la variable dependiente.
- **X** es la matriz de dimensión $n \times (k+1)$ de las observaciones de las variables independientes, incluyendo una columna de unos para el intercepto.
- El vector columna $\mathbf{X}_i$ de dimensión $(k+1) \times 1$ es la observación i-ésima de los k+1 regresores, es decir, $\mathbf{X}_i' = (1, X_{1i}, X_{2i}, \dots, X_{ki})$ la traspuesta de $\mathbf{X}_i$.
- $\boldsymbol{\beta}$ es el vector de dimensión $(k+1) \times 1$ de los coeficientes de regresión poblacionales.
- **u** es el vector de dimensión $n \times 1$ de los términos de error.

El modelo en notación matricial para la observación i-ésima es:

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta} + u_i$$

Y el modelo recopilando todas las observaciones es:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Los supuestos ampliados de mínimos cuadrados

Muy similares a los que vimos en el capítulo de regresión múltiple solo que ahora se enuncian con notación matricial habiendo mínimos cambios.

Concepto clave 18.1: Supuestos ampliados de mínimos cuadrados para el modelo de regresión múlt

Donde el modelo de regresión múltiple poblacional es:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Los supuestos ampliados de mínimos cuadrados son:

1. $E(u_i \mid \mathbf{X}_i) = 0$ (u_i tiene media condicional igual a cero).
2. $(\mathbf{X}_i, Y_i)$, $i = 1, 2, \ldots, n$ son i.i.d. a partir de su distribución conjunta.
3. $\mathbf{X}_i$ y u_i tienen momentos de cuarto orden finitos y distintos de cero.
4. No existe multicolinealidad perfecta. La matriz $\mathbf{X}$ tiene rango completo, es decir, el rango de $\mathbf{X}$ es igual a $k+1$.
5. $var(u_i|\mathbf{X}_i) = \sigma_u^2$ (homocedasticidad).
6. La distribución condicional de u_i dado $\mathbf{X}_i$ es normal (errores normales).

El supuesto de homocedasticidad es útil cuando se estudia la eficiencia del estimador MCO, y el de normalidad se utiliza cuando se estudia la distribución muestral exacta del estimador MCO y de los estadísticos de contraste.

En cuarto supuesto, recordemos que la multicolinealidad perfecta surge cuando un regresor se puede escribir como combinación lineal perfecta del resto de los regresores. En notación matricial significa que una columna de $\mathbf{X}$ es una combinación lineal perfecta del resto de las otras columnas de $\mathbf{X}$, si esto es cierto entonces $\mathbf{X}$ no tiene rango de columnas completo.

El primer y el segundo supuesto implican que:

$$\begin{aligned} E(u_i|\mathbf{X}) &= E(u_i|\mathbf{X}_i) = 0 \\ cov(u_i, u_j|\mathbf{X}) &= E(u_i u_j|\mathbf{X}) = E(u_i u_j|\mathbf{X}_i\mathbf{X}_j) = E(u_i|\mathbf{X}_i)E(u_j|\mathbf{X}_j) = 0 \\ &\text{para } i \neq j \end{aligned}$$

Los supuestos primero, segundo y quinto implican que:

$$E(u_i^2|\mathbf{X}) = E(u_i^2|\mathbf{X}_i) = \sigma_u^2$$

Los supuestos primero, segundo, quinto y sexto implican que la distribución condicional del vector aleatorio n-dimensional, $\mathbf{U}$ condicionada a $\mathbf{X}$, es la distribución normal multivariante.

Combinando estos 3 resultados se obtiene que:

1. Bajo los supuestos 1 y 2, $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}_n$.
2. Bajo los supuestos 1, 2 y 5, $E(\mathbf{UU'}|\mathbf{X}) = \sigma_u^2\mathbf{I}_n$.
3. Bajo los supuestos 1, 2, 5 y 6, la distribución condicional de $\mathbf{U}$ dada $\mathbf{X}$ es $N(\mathbf{0}_n, \sigma_u^2\mathbf{I}_n)$.

Donde $\mathbf{0}_n$ es el vector n-dimensional de ceros e $\mathbf{I}_n$ es la matriz de identidad $n \times n$.

El estimador MCO

El estimador minimiza la suma de los errores de predicción al cuadrado:

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$$

Preguntar bien al profe hasta que parte entre de este capítulo

Capítulo 12

Regresión con Datos de Panel

12.1 Datos de panel

Introducción a los datos de panel

Los **datos de panel**, también conocidos como datos longitudinales, se refieren a la observación de múltiples entidades individuales a lo largo de diferentes períodos de tiempo. En econometría, estos datos permiten analizar tanto las diferencias entre entidades como los cambios dentro de cada entidad a lo largo del tiempo. (Combinacion de datos de series de tiempo con datos de corte transversal)

Una característica fundamental de los datos de panel es que combinan aspectos de las series de tiempo y los datos de corte transversal. Específicamente, si n representa el número de entidades (por ejemplo, estados, individuos, empresas) y T representa el número de períodos de tiempo en los que se observan esas entidades, un conjunto de datos de panel puede describirse mediante $n \times T$ observaciones. Cada observación puede incluir múltiples variables. Existen algunos términos adicionales asociados con los datos de panel para indicar si existen algunas observaciones perdidas. Un **panel equilibrado** dispone de todas sus observaciones. Un panel al que le faltan algunos datos perdidos para al menos un periodo de tiempo o para al menos una entidad individual se denomina **panel incompleto** (Los métodos que se presentan en este capítulo están descritos para un panel equilibrado).

Los datos de panel observan n entidades a lo largo de T periodos.

$$(X_{it}, Y_{it}), \quad i = 1, \ldots, n \quad t = 1, \ldots, T$$

12.2 Periodos temporales: comparaciones antes y despúes

Cuando se dispone de datos de panel con dos periodos de tiempo T = 2, es posible realizar comparaciones “antes y después” entre los valores de la variable dependiente en ambos periodos. Este tipo de análisis permite mantener constantes los factores no observables que varían entre las entidades (por ejemplo, estados) pero que no cambian con el tiempo dentro de cada entidad.

Consideremos Z_i como una variable que influye en la tasa de mortalidad en el estado i, y que no cambia con el tiempo. Un ejemplo de tal variable podría ser la actitud cultural local hacia beber y conducir, la cual puede ser considerada constante entre 1982 y 1988. La ecuación de regresión lineal poblacional que relaciona Z_i y el impuesto real sobre la cerveza con la tasa de mortalidad es:

$$\text{TasaMortalidad}_{it} = \beta_0 + \beta_1 \text{ImpuestoCerveza}_{it} + \beta_2 Z_i + u_{it}$$

donde u_{it} es el término de error, $i = 1, \ldots, n$ representa los estados y $t = 1, \ldots, T$ los periodos de tiempo.

Dado que Z_i no cambia en el tiempo, su efecto sobre la tasa de mortalidad entre 1982 y 1988 se puede eliminar al analizar la variación de la tasa de mortalidad entre estos dos periodos.

Consideremos la ecuación de regresión para cada uno de los dos años, 1982 y 1988:

$$\text{TasaMort}_{i,1982} = \beta_0 + \beta_1 \text{ImpCerve}_{i,1982} + \beta_2 Z_i + u_{i,1982}$$
$$\text{TasaMort}_{i,1988} = \beta_0 + \beta_1 \text{ImpCerve}_{i,1988} + \beta_2 Z_i + u_{i,1988}$$

Al restar la ecuación de 1982 de la ecuación de 1988, se elimina el efecto de Z_i, obteniendo:

$$\text{TasaMort}_{i,1988} - \text{TasaMort}_{i,1982} = \beta_1(\text{ImpCerve}_{i,1988} - \text{ImpCerve}_{i,1982}) + (u_{i,1988} - u_{i,1982})$$

Este enfoque intuitivo muestra que los cambios en la tasa de mortalidad a lo largo del tiempo son atribuibles a factores como cambios en el impuesto sobre la cerveza y otros factores capturados por el término de error. Las variables no observables que son constantes en el tiempo, como Z_i, se eliminan mediante esta especificación de diferencias.

Este método de diferencias es útil cuando se disponen de datos para solo dos periodos. Sin em- bargo, cuando $T > 2$, es recomendable utilizar el método de efectos fijos, que permite analizar todas las observaciones disponibles sin descartar información valiosa.

12.3 Regresión de Efectos Fijos

La regresión de efectos fijos es una técnica utilizada en análisis de datos de panel, que permite controlar las variables no observadas que son constantes a lo largo del tiempo pero que pueden variar entre diferentes entidades. Esto se logra al incluir un término de efecto fijo en el modelo, que representa estas características inobservables específicas para cada entidad.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

Donde Z_i es una variable no observable, que varía de un estado a otro, pero que no cambia en el tiempo (por ejemplo, Z_i representa las actitudes culturales hacia la bebida y la conducción). Se pretende estimar β_1, el efecto sobre Y de X manteniendo constantes la características no observables del estado Z. Debido a que Z_i varía de un estado a otro, pero es constante en el tiempo, se puede interpretar que el modelo de regresión poblacional de la Ecuación. Contiene n interceptos, uno para cada estado. En concreto:

$$\text{Sea:} \quad \alpha_i = \beta_0 + \beta_2 Z_i$$
$$\text{Entonces:} \quad Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

Es el modelo de regresión de efectos fijos, en el que $\alpha_1, \alpha_2, \ldots, \alpha_n$ se tratan como interceptos desconocidos a estimar, uno para cada estado. Donde el coeficiente de la pendiente de la recta de regresión poblacional, β_1, es el mismo para todos los estados, pero el intercepto de la recta de regresión poblacional varía de un estado a otro. $\alpha_1, \alpha_2, \ldots, \alpha_n$ se conocen como efectos fijos individuales cuya variación proviene de las variables omitidas.

Los interceptos específicos de cada estado en el modelo de regresión de efectos fijos pueden asimismo expresarse utilizando variables binarias que expresen los estados individuales. Para desarrollar el modelo de regresión de efectos fijos mediante variables binarias, sea D_{1i} una variable binaria que es igual a 1 cuando i=1 y es igual a 0 en caso contrario, sea D_{2i} igual a 1 cuando i=2 y es igual a 0 en caso contrario, y así sucesivamente. No pueden incluirse las n variables binarias además de un intercepto común porque si se hace, los regresores serán perfectamente multicolineales (esta es la "trampa de la variable ficticia")

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \cdots + \gamma_n D_{ni} + u_{it} \tag{12.1}$$

Estimación e inferencia

En principio la especificación con variables binarias del modelo de regresión de efectos fijos se puede estimar mediante MCO, no obstante con muchos regresores su calculo es tedioso por ello se resuelve con utilizacion de software que obtienen mediante el álgebra de la regresión de efectos fijos.

Nota: aunque parezcan diferentes las especificaciones tanto con como sin variables binarias son equivalentes por lo que el beta sera el mismo. Ademas, si se cumplen un conjunto de supuestos —denominados supuestos de la regresión de efectos fijos—, entonces la distribución muestral del estimador MCO de efectos fijos es normal en muestras grandes, la varianza de esta distribución puede estimarse a partir de los datos.

Algoritmo MCO en desviaciones respecto de su media El estimador CO de efectos fijos se calcula utilizando las variables "en desviaciones respecto de su media". En el caso de un solo regresor, se parte de la ecuación de efectos fijos y se toma diferencia respecto a la media en ambos lados de la ecuación. Así, para $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, $\bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$, y $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Esto implica que:

$$Y_{it} - \bar{Y}_i = (\alpha_i + \beta_1 X_{it} + u_{it}) - (\alpha_i + \beta_1 \bar{X}_i + \bar{u}_i)$$
$$Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)$$

Definiendo las variables en desviaciones respecto de sus medias: $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$, $\tilde{X}_{it} = X_{it} - \bar{X}_i$ y $\tilde{u}_{it} = u_{it} - \bar{u}_i$, entonces:

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$$

Por lo tanto, β_1 puede estimarse mediante regresión MCO con las variables en desviaciones respecto de su media. Este estimador es idéntico al obtenido en el modelo de efectos fijos de la Ecuación (12.1) con $n-1$ variables binarias.

Comparaciones entre especificaciones: Cuando T = 2, existen tres formas equivalentes de estimar β_1 mediante MCO:

- La especificación "antes y después"(sin intercepto).
- La especificación con variable binaria.
- La especificación "en desviaciones respecto de su media".

En estos casos, los estimadores de β_1 son idénticos.

Distribución muestral y errores estándar: En regresión múltiple con datos de panel, si se cumplen los supuestos de la regresión de efectos fijos, el estimador MCO de efectos fijos tiene una distribución normal en muestras grandes. La varianza de esta distribución puede estimarse y su raíz cuadrada, el error estándar, se utiliza para la inferencia estadística, como el contraste de hipótesis (con el estadístico t) y la construcción de intervalos de confianza, de forma similar a los datos de sección cruzada.

12.4 Regresión son efectos fijos temporales

La regresión con **efectos fijos temporales** permiten tener en cuenta las variables que son constantes entre las entidades individuales, pero que evolucionan en el tiempo, como estándares de seguridad na- cionales. El modelo seria:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

Donde el efecto de la seguridad del automóvil, se expresa mediante "S".

Solamente efectos temporales

Aunque S no sea observable, su influencia puede eliminarse debido a que varía en el tiempo, pero no entre los estados, del mismo modo que es posible eliminar el efecto de Z_i , que varía entre los estados, pero no en el tiempo. De tal forma la presencia de S lleva a un modelo de regresión en el que cada periodo de tiempo tiene su propio intercepto. Estos interceptos se puede considerar como el efecto sobre Y del año t por lo que se conocen como efectos fijos temporales. El modelo de efectos fijos temporales se representa mediante la inclusión de $T-1$ indicadores binarios temporales junto con un intercepto, de la siguiente manera:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B_{2t} + \cdots + \delta_T B_{Tt} + u_{it}$$

donde $\delta_2, \ldots, \delta_T$ son coeficientes desconocidos y $B_{2t}, \ldots, B_{Tt}$ son variables binarias que indican los distintos periodos de tiempo.

Efectos fijos individuales y temporales

Cuando las variables omitidas incluyen tanto efectos constantes en el tiempo entre entidades como efec- tos constantes entre entidades a lo largo del tiempo, es adecuado utilizar un modelo de efectos fijos individuales y temporales. Este modelo se expresa como:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

donde α_i representa los efectos fijos individuales (entre entidades) y λ_t los efectos fijos temporales. Este modelo puede equivaler a incluir $n-1$ indicadores binarios para las entidades y $T-1$ indicadores binarios temporales, junto con un intercepto:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_{2i} + \cdots + \gamma_n D_{ni} + \delta_2 B_{2t} + \cdots + \delta_T B_{Tt} + u_{it}$$

donde $\gamma_2, \ldots, \gamma_n$ y $\delta_2, \ldots, \delta_T$ son coeficientes desconocidos.

Estimación

El modelo de efectos fijos temporales, al igual que el modelo combinado de efectos fijos individuales y tempo- rales, puede ser estimado mediante mínimos cuadrados ordinarios (MCO) incluyendo las variables binarias adicionales. Alternativamente, en un panel equilibrado, los coeficientes pueden calcularse expresando primero las variables dependiente e independiente en términos de desviaciones respecto a sus medias individuales y temporales, y luego estimando la ecuación de regresión múltiple utilizando estas desviaciones.

12.5 Supuestos de la regresión de efectos fijos y los errores estándar

En los modelos de datos de panel, el error de regresión puede estar correlacionado en el tiempo dentro de una entidad individual. Si bien esta correlación no introduce sesgo en el estimador de efectos fijos, sí afecta a la varianza del estimador, lo que a su vez impacta en el cálculo de los errores estándar. Los errores estándar utilizados en las regresiones de efectos fijos que se presentan en este capítulo son los llamados errores estándar agrupados, que son robustos ante la heterocedasticidad y la correlación temporal dentro de una entidad individual. Cuando el número de entidades individuales (n) es grande, las pruebas de hipótesis y los intervalos de confianza se pueden calcular utilizando los valores críticos estándar para muestras grandes, es decir, de las distribuciones normal y F.

Supuestos de la regresión de efectos fijos

Los supuestos de la regresión de efectos fijos extienden los supuestos clásicos de los mínimos cuadrados ordinarios (MCO) a los datos de panel. Bajo estos supuestos, el estimador de efectos fijos tiene una distribución asintóticamente normal cuando n es grande. El Supuesto 2 a diferencia de los datos de sección cruzada, el segundo supuesto para datos de panel permite que las variables estén correlacionadas dentro de una entidad individual a lo largo del tiempo. Por ejemplo, en datos de series temporales, es común que las variables estén autocorrelacionadas, lo que significa que las observaciones en un periodo tienden a estar correlacionadas con las observaciones en periodos anteriores. Es decir, esta autocorrelacionada (correlacionada consigo misma, en diferentes periodos) o **serialmente correlacionada**.

Concepto clave 10.3: Los supuestos de la regresión de efectos fijos

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{i1}, i = 1, ..., n, t = 1, ..., T$$

Donde:
1. u_{it} presenta media condicional igual a cero: $E(u_{it} \mid X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$.
2. $X_{i1}, \ldots, X_{iT}, \alpha_{i1}, \ldots, \alpha_{iT}$ son i.i.d. a partir de su distribución conjunta.
3. Los datos atípicos elevados son improbables: (X_{it}, uit) tienen momentos de cuarto orden finitos
4. No existe multicolinealidad perfecta.

Para regresores múltiples la X_{it} debería reemplazarse por la lista completa $X_{1,it}, \ldots, X_{k,it}$.

Errores estándar agrupados HAC

Cuando los errores de una regresión de efectos fijos están autocorrelacionados, la fórmula habitual para errores estándar heterocedástico-robustos no es válida. Esto es análogo a cómo en una regresión de sección cruzada, los errores estándar válidos bajo homocedasticidad no sirven si hay heterocedasticidad. De manera similar, si hay autocorrelación, los errores estándar tradicionales no son adecuados. Para abordar esto, se utilizan errores estándar consistentes a heterocedasticidad y autocorrelación (HAC). En particular, se emplean errores estándar agrupados que permiten heterocedasticidad y autocorrelación dentro de cada entidad individual, pero suponen que los errores no están correlacionados entre diferentes entidades. Esto es compatible con los supuestos de la regresión de efectos fijos. Estos errores estándar agrupados son válidos independientemente de la existencia de heterocedasticidad o autocorrelación. En muestras grandes, permiten realizar inferencias con los valores críticos habituales para los estadísticos. (El apendice 10.2 correspondiente al desarrollo de su formula no entra al examen)

12.6 Modelo de efectos aleatorios

El modelo de efectos inobservables se expresa como:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$$

Donde a_i **es el efecto inobservable** con media cero, y x_{itj} son las variables explicativas, que pueden incluir variables binarias temporales. En los modelos de efectos fijos o primeras diferencias, a_i se elimina bajo el supuesto de que está correlacionado con una o más x_{itj}. No obstante, si a_i no está correlacionado con ninguna variable explicativa en todos los periodos, eliminar a_i (como en modelos de efectos fijos) resulta en estimadores ineficientes.

El modelo de efectos aleatorios asume que el efecto inobservable a_i no está correlacionado con ninguna de las variables explicativas:

$$\text{Cov}(x_{itj}, a_i) = 0, \quad \forall t = 1, 2, \ldots, T \quad \text{y} \quad \forall j = 1, 2, \ldots, k.$$

Este modelo incluye todos los supuestos de los efectos fijos, con el requisito adicional de que a_i es independiente de todas las variables explicativas en todos los periodos. Si a_i está correlacionado con alguna variable explicativa, deben utilizarse las primeras diferencias o efectos fijos.

Bajo el supuesto de la ecuación (12.6), las β_j pueden estimarse consistentemente utilizando un solo corte transversal, aunque esto ignora información útil de otros periodos (incrementando el HAC). También es posible usar un procedimiento combinado de mínimos cuadrados ordinarios (MCO), pero este método no considera la correlación serial en los errores compuestos $v_{it} = a_i + u_{it}$:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + v_{it}$$

Dado que a_i está presente en el término de error en cada periodo, los vit están serialmente correlacionados:

$$\text{Corr}(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}, \quad \text{para} \quad t \neq s,$$

donde $\sigma_a^2 = Var(a_i)$ y $\sigma_u^2 = Var(u_{it})$. Esta correlación positiva en los errores puede ser considerable, y los errores estándar de MCO combinados serán incorrectos.

La transformación de mínimos cuadrados generalizados (MCG) que elimina la correlación serial en los errores se define como:

$$\phi = \left[\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}\right]^{1/2}$$

donde ϕ está entre cero y uno. La ecuación transformada resulta en:

$$y_{it} - \phi\bar{y}_i = \beta_0(1-\phi) + \beta_1(x_{it1} - \phi\bar{x}_{i1}) + \cdots + \beta_k(x_{itk} - \phi\bar{x}_{ik}) + (v_{it} - \phi\bar{v}_i)$$

donde la barra indica promedios a lo largo del tiempo. Esta transformación permite incluir variables explicativas que son constantes en el tiempo, una ventaja de los efectos aleatorios (EA) sobre los efectos fijos o primeras diferencias.

El parámetro ϕ no se conoce en la práctica, pero puede estimarse de manera consistente. El estimador de MCG factibles que usa $\hat{\phi}$ en lugar de ϕ se denomina estimador de efectos aleatorios. Bajo los supuestos de efectos aleatorios, el estimador es consistente y asintóticamente normal cuando N crece con T fijo.

La ecuación (12.6) relaciona el estimador de EA con los estimadores combinados de MCO y de efectos fijos.

- Cuando $\phi = 0$, el estimador de EA coincide con el estimador de MCO combinados.
- cuando $\phi = 1$, coincide con el estimador de efectos fijos.

En la práctica, $\hat{\phi}$ nunca es exactamente cero ni uno, pero si está cerca de cero, las estimaciones de EA serán próximas a las de MCO combinados. Si σ_a^2 es grande respecto a σ_u^2, $\hat{\phi}$ estará cerca de uno, haciendo que las estimaciones de EA se acerquen a las de efectos fijos.

Finalmente, es útil comparar las estimaciones de MCO combinados, efectos fijos y efectos aleatorios para entender la naturaleza de los sesgos introducidos al dejar el efecto inobservable a_i en el término de error. Sin embargo, incluso si a_i no se correlaciona con ninguna variable explicativa, los errores estándar de MCO combinados suelen ser inválidos debido a la correlación serial en $v_{it} = a_i + u_{it}$. Los paquetes estadísticos modernos permiten calcular errores estándar robustos frente a correlación serial y heterocedasticidad en $v_i t$.

Efectos aleatorios o efectos fijos

Los efectos fijos (EF) permiten una correlación arbitraria entre a_i y las x_{itj} , mientras que los efectos aleatorios (EA) no. Por esta razón, se considera que los EF son una herramienta más convincente para la estimación de efectos ceteris paribus. Sin embargo, los EA se aplican en ciertas situaciones. Por ejemplo, si la variable explicativa clave es constante en el tiempo, no es posible usar EF para estimar su efecto sobre y. En estos casos, se recurre a EA o a MCO combinados, siempre y cuando se acepte que a_i no está correlacionado con ninguna de las variables explicativas.

Cuando se utilizan EA, es preferible incluir tantos controles constantes en el tiempo como sea posible entre las variables explicativas, algo innecesario en un análisis de EF. EA suele ser más eficiente que MCO combinados.

Si la variable explicativa de interés cambia con el tiempo, es común que EA se utilice en vez de EF solo cuando se puede asumir $Cov(x_{itj}, a_i) = 0$, lo cual es raro. Un caso donde EA es apropiado es cuando la variable de política clave se establece de forma experimental, como en la asignación aleatoria de niños a diferentes tamaños de clase cada año. Sin embargo, generalmente, los regresores son resultado de procesos de elección y tienden a correlacionarse con a_i.

Es común que los investigadores apliquen tanto EF como EA y luego realicen la prueba de Hausman para detectar diferencias estadísticamente significativas en los coeficientes de las variables que cambian con el tiempo. La prueba de Hausman sugiere usar EA a menos que esta los rechace. Si no hay rechazo, significa que las estimaciones de EA y EF son suficientemente cercanas o que la variación de muestreo en EF es tan grande que no se pueden detectar diferencias significativas. Un rechazo en la prueba indica que el supuesto clave de EA es falso, por lo que se utilizan las estimaciones de EF.

Finalmente, algunos autores prefieren EF sobre EA dependiendo de si consideran a_i como parámetros o como variables aleatorias. Sin embargo, estas consideraciones suelen ser desatinadas. El aspecto crucial para elegir entre EF y EA es la posibilidad de suponer convincentemente que a_i no se correlaciona con x_{itj} . En

aplicaciones con datos de panel, especialmente cuando la unidad de observación es una unidad geográfica grande, tiene sentido considerar a_i como un intercepto separado para cada unidad de corte transversal, lo que justifica el uso de EF.

Capítulo 13

Regresión con Variable Dependiente Binaria

13.1 Variable dependiente binaria y modelo de probabilidad lineal

Variables dependientes binarias

El modelo con una variable dependiente binaria consiste en interpretar la regresión como la modelización de la probabilidad de que la variable dependiente sea igual a 1. la función de regresión poblacional es el valor esperado de Y dados los regresores, condicionado al valor de los regresores.

$$E(Y \mid X_1, \ldots, X_k) = Pr(Y = 1 \mid X_1, \ldots, X_k)$$

El modelo de probabilidad lineal

El modelo de probabilidad lineal es una extensión del modelo de regresión múltiple, pero se aplica cuando la variable dependiente es binaria, es decir, toma valores de 0 o 1. En este contexto, la función de regresión poblacional representa la probabilidad de que la variable dependiente sea igual a 1, dado un conjunto de variables independientes $X_1, \ldots, X_k$. El coeficiente de un regresor X en este modelo refleja cómo cambia la probabilidad de que $Y = 1$ cuando X varía en una unidad. Del mismo modo, el valor predicho por el método de mínimos cuadrados ordinarios (MCO) es la probabilidad estimada de que $Y = 1$. Aunque muchas de las herramientas usadas en la regresión múltiple estándar son aplicables aquí (como el cálculo de intervalos de confianza y la prueba de hipótesis), el modelo de probabilidad lineal tiene limitaciones importantes. Una de las principales es que los errores del modelo son heterocedásticos (En definitiva la Y binaria es una bernulli cuya varianza no es constante, es $P(1-P)$), lo que requiere el uso de errores estándar robustos para la inferencia. Otra limitación es que el R^2, una medida común de ajuste en modelos lineales con variables dependientes continuas, no es útil en este contexto. Esto se debe a que es imposible que todos los datos se ajusten perfectamente a la recta de regresión cuando la variable dependiente es binaria, a menos que los regresores también sean binarios. Además, la linealidad del modelo puede llevar a predicciones de probabilidad que no tienen sentido, como valores inferiores a 0 o superiores a 1. Este es un defecto inherente del modelo de probabilidad lineal, ya que las probabilidades deben estar siempre en el rango de 0 a 1. Para abordar esta limitación, se utilizan modelos no lineales como el probit y el logit, que están diseñados específicamente para variables dependientes binarias.

Concepto clave 11.1: El modelo de probabilidad lineal

El modelo de probabilidad lineal es el modelo lineal de regresión múltiple,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

aplicado a una variable dependiente binaria Y_i. Debido a que Y es binaria, $E(Y \mid X_1, X_2, \ldots, X_k) = Pr(Y = 1 \mid X_1, X_2, \ldots, X_k)$, por lo que el modelo de probabilidad lineal,

$$\Pr(Y = 1|X_1, X_2, ..., X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

El coeficiente de regresión β_1 es la variación de la probabilidad de que $Y = 1$ asociadacon una variación unitaria en X_1, manteniendo constantes las otras variables explicativas, y sucesivamente para $\beta_2, \ldots, \beta_k$. Los coeficientes de la regresión se pueden estimar por MCO, y se pueden utilizar los errores MCO habituales (heterocedástico-robustos) para los intervalos de confianza y los contrastes de hipótesis.

13.2 Regresión probit y logit

Las regresiones probit y logit son modelos de regresión no lineales diseñados específicamente para variables dependientes binarias. Debido a que una regresión con una variable dependiente binaria Y modeliza la probabilidad de que Y = 1, tiene sentido adoptar una formulación no lineal que obligue a que los valores estimados estén entre 0 y 1.

Regresión Probit

El modelo de regresión probit con un único regresor X se define como:

$$\Pr(Y = 1 \mid X) = \Phi(\beta_0 + \beta_1 X)$$

donde Φ es la función de distribución acumulada (FDA) de una distribución normal estándar. Esta función transforma la combinación lineal de los regresores en una probabilidad. Por ejemplo, si Y es una variable binaria que indica la denegación de una solicitud de hipoteca (denegar) y X es la proporción de pagos-ingresos (ratio P/I), y los coeficientes son β_0 = -0.2 y β_1 = 3, la probabilidad de denegación; cuando ratio P/I es 0.4 se calcula como:

$$\Pr(\text{denegar} = 1 \mid \text{ratio P/I} = 0.4) = \Phi(-0.2 + 3 \times 0.4) = \Phi(1)$$

La interpretación de los coeficientes en un modelo probit no es directa, ya que no representan cambios marginales, el cambio esperado en Y que surge de un cambio en X es el cambio en la probabilidad de que Y = 1. Se calcula como la diferencia entre los valores esperados de Y entre ambos X. Los coeficientes probit se calculan mediante el método de máxima verosimilitud, que produce estimadores eficientes y de varianza mínima. Este método es consistente y, en muestras grandes, los estimadores se distribuyen normalmente, permitiendo la construcción de estadísticos (t) e intervalos de confianza de manera habitual.

Concepto clave 11.2: El modelo probit, probabilidades estimadas y efectos estimados

El modelo probit poblacional con varios regresores es

$$\Pr(Y = 1|X_1, X_2, ..., X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

donde la variable dependiente Y es binaria, Φ es la función de distribución normal estándar acumulada, y X_1 y X_2 , etc., son regresores. El modelo se interpreta mejor calculando las probabilidades esperadas y el efecto de un cambio en un regresor.
La probabilidad esperada de que $Y = 1$, dados los valores de $X_1, \ldots, X_k$, se calcula mediante el cómputo del z-valor, $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, y buscando luego este valor de z en la tabla de la distribución normal (Tabla 1 del Apéndice).
El coeficiente β_1 es el cambio en el z-valor derivado de un cambio unitario en X_1 , manteniendo constantes $X_2, \ldots, X_k$.
El efecto sobre la probabilidad esperada de un cambio en un regresor se calcula (1) calculando la probabilidad esperada para el valor inicial de las variables explicativas, (2) calculando la probabilidad esperada para el nuevo o modificado valor de los regresores, y (3) tomando su diferencia.

Regresión Logit

El modelo de regresión logit es similar al probit, pero utiliza una función de distribución acumulada logística en lugar de la normal. La especificación del modelo logit es:

$$\Pr(Y = 1 \mid X) = \Lambda(\beta_0 + \beta_1 X)$$

donde Λ es la función de distribución acumulada logística, definida como:

$$\Lambda(z) = \frac{1}{1 + e^{-z}}$$

Los coeficientes del modelo logit se estiman mediante el método de máxima verosimilitud, que es consis- tente y normalmente distribuido en muestras grandes, permitiendo la construcción de estadísticos (t) e intervalos de confianza de manera habitual. Las funciones de regresión logit y probit son muy similares, con diferencias mínimas. Históricamente, la regresión logística se prefería por su rapidez de cálculo, pero con los avances en computación, esta diferencia ya no es relevante.

Concepto clave 11.3: Regresión logit

El modelo de regresión logit poblacional de la variable dependiente binaria Y con varios regresores es:

$$\Pr(Y = 1|X_1, X_2, ..., X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k) \tag{13.1}$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \tag{13.2}$$

La regresión logit es similar a la regresión probit excepto porque la función de distribución acumulada es diferente.

Comparación entre Probit y Logit

Aunque los modelos probit y logit son muy similares en su funcionalidad y en los resultados que producen, existen algunas diferencias clave. Históricamente, la regresión logística fue preferida debido a la simplicidad computacional de la función logística comparada con la función normal. Sin embargo, con los avances en el poder de cómputo, esta ventaja ha perdido relevancia. Ambos modelos son consistentes y convergen en muestras grandes, lo que significa que en la práctica, las diferencias entre ellos suelen ser mínimas. Las funciones de regresión probit y logit son casi indistinguibles cuando se grafican, como se observa en la Figura

11.3 del texto, que compara ambas funciones para la probabilidad de denegación de hipotecas en función de la variable ratio P/I. La principal diferencia radica en las colas de las distribuciones: la distribución logística tiene colas ligeramente más pesadas, lo que puede influir en la estimación de probabilidades extremas.

13.3 Estimación e inferencia en los modelos logit y probit

los modelos probit y logit son no lineales en los coeficientes, lo que impide su estimación por MCO. Los coeficientes probit y logit se estiman mediante máxima verosimilitud, un método más complejo pero implementado en software estadístico moderno

El método de mínimos cuadrados no lineales

Se utiliza para estimar parámetros desconocidos en una función de regresión no lineal. Este método extiende el estimador de mínimos cuadrados ordinarios (MCO) a funciones no lineales de los parámetros. Selecciona los valores de los parámetros que minimizan la suma de los errores de predicción al cuadrado. Para el modelo probit, la esperanza condicional de (Y) dadas las (X) se ajusta a una función no lineal de los parámetros. El estimador de mínimos cuadrados no lineales minimiza la suma de los errores de predicción al cuadrado para obtener los coeficientes probit. Es decir minimiza:

$$\sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \right]^2$$

Aunque este estimador es consistente y normalmente distribuido en muestras grandes, es ineficiente comparado con otros métodos. Por esta razón, en la práctica, se prefiere la estimación por máxima verosimilitud para los coeficientes probit.

Estimación máximo verosímil (*Este metodo usamos*)

La función de verosimilitud es la distribución de probabilidad conjunta de los datos (Dado que las suponemos independientes es su productorio), considerada como una función de los coeficientes descono- cidos. Para n observaciones independientes, la función de verosimilitud es:

$$L(\theta; x_1, \ldots, x_n) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$$

Para simplificar los cálculos, se utiliza el logaritmo de la función de verosimilitud:

$$l(\theta; x_1, \ldots, x_n) = \ln[f(x_1;\theta)] + \ln[f(x_2;\theta)] + \cdots + \ln[f(x_n;\theta)]$$

El objetivo del EMV es encontrar los valores de los parámetros que maximizan esta función de log- verosimilitud. Debido a que el estimador de máxima verosimilitud (EMV) es consistente y se distribuye normalmente en muestras grandes, la inferencia estadística sobre los coeficientes probit y logit basada en el EMV se realiza de manera similar a la inferencia sobre los coeficientes de la función de regresión lineal basada en el estimador de mínimos cuadrados ordinarios (MCO)

13.4 Apéndice

Apéndice 11.2: Casos particulares

En este apéndice se ofrece una breve introducción sobre la estimación de máxima verosimilitud (EMV) en el contexto de los modelos de respuesta binaria. Se comienza por obtener la probabilidad de éxito p para n observaciones i.i.d. de una variable aleatoria de Bernoulli. Luego, se abordan los modelos probit y logit, y el análisis del pseudo R^2. Finalmente, se estudian los errores estándar de las probabilidades estimadas.

EMV de n variables aleatorias i.i.d. de Bernoulli El primer paso en el cálculo del EMV es obtener la distribución de probabilidad conjunta. Para n observaciones i.i.d. de una variable aleatoria de Bernoulli, la distribución de probabilidad conjunta es:

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

Esto se puede expresar de manera compacta como:

$$\Pr(Y_1 = y_1, \ldots, Y_n = y_n) = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i}$$

La función de verosimilitud, considerando la distribución de probabilidad conjunta como función del parámetro desconocido p, es:

$$\mathscr{L}(p; Y_1, \ldots, Y_n) = p^S(1-p)^{n-S}$$

donde $S = \sum_{i=1}^{n} Y_i$. El EMV de p se obtiene maximizando esta función de verosimilitud. En lugar de maximizar directamente la verosimilitud, es conveniente maximizar su logaritmo:

$$\ln \mathscr{L}(p) = S\ln(p) + (n-S)\ln(1-p)$$

La derivada de la función log-verosimilitud respecto a p es:

$$\frac{d}{dp}\ln \mathscr{L}(p) = \frac{S}{p} - \frac{n-S}{1-p}$$

Igualando la derivada a cero y resolviendo para p, se obtiene el estimador de máxima verosimilitud $\hat{p} = \frac{S}{n}$.

EMV del modelo Probit En el modelo probit, la probabilidad de que $Y_i = 1$, condicionada a $X_{1i}, \ldots, X_{ki}$, es $p_i = \phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$, donde $\phi(\cdot)$ es la función de distribución acumulativa de la normal estándar. La distribución de probabilidad condicional para la observación i-ésima es:

$$\Pr(Y_i = y_i \mid X_{1i}, \ldots, X_{ki}) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Asumiendo que las observaciones son i.i.d., la función $\ln L(\beta_0 + \cdots + \beta_k; Y_1, \ldots, Y_n \mid X_{1i}, \ldots, X_{ki})$ de log-verosimilitud es:

$$\sum_{i=1}^{n} [Y_i \ln(\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})) +$$
$$(1-Y_i)\ln(1 - \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}))]$$

Dado que no existe una fórmula cerrada para el EMV en este caso, la maximización de la función de verosimilitud se realiza mediante métodos numéricos.

EMV del modelo Logit a verosimilitud en el modelo logit se obtiene de manera análoga al modelo probit. La diferencia radica en que la probabilidad condicional de éxito p_i se define como:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}$$

El logaritmo de la función de verosimilitud es similar al del modelo probit, con la sustitución de $\phi(\cdot)$ por p_i.

Errores estándar de las probabilidades estimadas Considerando un único regresor en el modelo probit, la probabilidad estimada $\hat{p}(x)$ dado un valor fijo del regresor x es $\hat{p}(x) = \phi(\hat{\beta}_0 + \hat{\beta}_1 x)$. La varianza de esta probabilidad estimada se obtiene mediante una expansión de Taylor de primer orden:

$$\hat{p}(x) \approx \Phi(\beta_0 + \beta_1 x) + a_0(\hat{\beta}_0 - \beta_0) + a_1(\hat{\beta}_1 - \beta_1)$$

donde a_0 y a_1 son las derivadas parciales. La varianza de $\hat{p}(x)$ es:

$$\mathrm{Var}[\hat{p}(x)] \approx a_0^2 \mathrm{Var}(\hat{\beta}_0) + a_1^2 \mathrm{Var}(\hat{\beta}_1) + 2a_0 a_1 \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

Finalmente, el error estándar de $\hat{p}(x)$ se calcula a partir de esta varianza.

Medidas de ajuste

En la Sección 11.1, se mencionó que el R^2 es una medida de ajuste deficiente para el modelo de probabilidad lineal. Esto es válido igualmente para las regresiones probit y logit. Existen dos medidas de ajuste para los modelos con variable dependiente binaria: la proporción correctamente estimada y el pseudo-R2. La proporción correctamente estimada utiliza la regla siguiente: si $Y_i = 1$ y la probabilidad estimada supera el 50 % o si $Y_i = 0$ y la probabilidad estimada es inferior al 50 %, entonces se dice que Y_i está correctamente estimada. De lo contrario, se dice que Y_i está incorrectamente estimada. La proporción correctamente estimada es la proporción de las n observaciones $Y_1, \ldots, Y_n$ que están correctamente estimadas. Una ventaja de esta medida de ajuste es que resulta fácil de comprender. Una desventaja es que no refleja la calidad de la predicción: si $Y_i = 1$, la observación se considera como correctamente estimada si la probabilidad estimada es del 51 % o del 90 %.

El **pseudo-**R^2 mide el ajuste del modelo mediante la función de verosimilitud. Debido a que el EMV maximiza la función de verosimilitud, la adición de otro regresor a un modelo probit o logit aumenta el valor de la verosimilitud maximizada, al igual que la adición de un regresor necesariamente reduce la suma de los cuadrados de los residuos en la regresión lineal por MCO. Este hecho sugiere medir la calidad de ajuste de un modelo probit mediante la comparación del valor de la función de verosimilitud maximizada con todas las variables explicativas con el valor de la función de verosimilitud sin regresores. Es decir, lo que hace el pseudo-R^2. Se proporciona una fórmula para el pseudo-R^2 en el Apéndice 11.2.

$$\text{Pseudo}-R^2 = 1 - \frac{\ln(\mathscr{L}_{\text{máx}}^{\text{probit}})}{\ln(\mathscr{L}_{\text{máx}}^{\text{Bernoulli}})}$$

13.5 Modelos de variables dependiente limitada y correcciones a la selección muestral (wooldridge)

Variable dependiente limitada (VDL) se define en sentido amplio como una variable dependiente cuyo rango de valores está restringido de alguna manera. Esto significa que no puede tomar cualquier valor posible, sino que está sujeta a ciertas limitaciones

Modelo Tobit

Un tipo importante de variable dependiente limitada es una respuesta de solución de esquina. Tal variable es cero para una fracción no trivial de la población, pero tiene una distribución aproximadamente continua a través de valores positivos. Un ejemplo es la cantidad que un individuo gasta en alcohol en un mes determinado. En la población de personas de más de 21 años en Estados Unidos, esta variable asume un amplio rango de valores. Para alguna fracción importante, la cantidad gastada es de cero.

El modelo Tobit es una herramienta adecuada para estas situaciones. Este modelo expresa la respuesta observada y en términos de una variable latente subyacente y^*:

$$\begin{aligned} y^* &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \\ y &= \text{máx}(0, y^*) \end{aligned}$$

donde u es un término de error que sigue una distribución normal con media cero y varianza σ^2, es decir, u $Normal(0, \sigma^2)$. La variable latente y^* satisface los supuestos del modelo lineal clásico; en particular, tiene una distribución normal y es homocedástica con una media condicional lineal.

Dado que la variable observada y es igual a y^* cuando $y^* > 0$ y a cero cuando $y^* \leq 0$, la distribución de y se compone de dos partes:

- Para $y > 0$, la densidad de y dada x es:

$$f(y \mid x) = \frac{1}{\sigma}\phi\left(\frac{y - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_k}{\sigma}\right)$$

  donde ϕ es la función de densidad de la normal estándar.

- Para $y = 0$, la probabilidad de que y sea cero dada x es:

$$P(y = 0 \mid x) = \Phi\left(\frac{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}{\sigma}\right)$$

donde Φ es la función de distribución acumulada de la normal estándar

La función de log-verosimilitud para una observación i es:

$$\ell_i(\beta, \sigma^2) = 1(y_i = 0) \log\left[1 - \Phi\left(\frac{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}{\sigma}\right)\right] +$$
$$1(y_i > 0)\left(\log\frac{1}{\sigma} - \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2}{2\sigma^2} - \log\sqrt{2\pi}\right)$$

La log-verosimilitud para una muestra aleatoria de tamaño n se obtiene al sumar la expresión anterior a través de todas las observaciones i. Las estimaciones de máxima verosimilitud de β y σ se obtienen maximizando la función de log-verosimilitud, lo que generalmente se realiza mediante métodos numéricos.

Como en el caso de los modelos logit y probit, cada estimación Tobit se acompaña con un error estándar, el cual se puede utilizar para construir estadísticos t para cada $\hat{\beta}_j$. La fórmula matricial para hallar los errores estándar es complicada y no se presenta aquí. Para probar restricciones de exclusión múltiples, se pueden usar la prueba de Wald o la prueba de razón de verosimilitudes. La prueba de Wald tiene una forma similar en el caso logit o probit, mientras que la prueba de razón de verosimilitudes utiliza las funciones de log-verosimilitud Tobit para los modelos restringido y no restringido.

Interpretación de las estimaciones Tobit Las estimaciones de máxima verosimilitud para los modelos Tobit generalmente no son mucho más difíciles de obtener que las estimaciones de MCO para un modelo lineal. Además, los resultados de Tobit y MCO suelen ser similares. Sin embargo, interpretar los coeficientes Tobit como si fueran estimaciones de una regresión lineal puede ser engañoso.

En el modelo Tobit, los coeficientes $\hat{\beta}_j$ miden los efectos parciales de las variables explicativas x_j sobre la variable latente y^*:

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

donde u sigue una distribución normal con media cero y varianza σ^2. La variable observable y se define como:

$$y = max(0, y^*)$$

Aunque y^* puede tener un significado económico en algunos casos, la variable de interés es y, ya que es la que se observa en la práctica. Por ejemplo, en el análisis de políticas, podríamos estar interesados en cómo las horas trabajadas responden a los cambios en las tasas marginales de impuestos.

Para obtener el valor esperado de y en función de x, se deben considerar dos expectativas importantes:

- La expectativa condicional $E(y \mid y > 0, x)$, que es el valor esperado de y dado que y es positivo.
- La expectativa no condicional $E(y|x)$, que es el valor esperado de y sin ninguna condición adicional.

La relación entre estas expectativas se expresa como:

$$E(y \mid x) = P(y > 0 \mid x) \cdot E(y \mid y > 0, x)$$

donde $P(y > 0|x)$ se calcula usando la función de distribución acumulada de la normal estándar Φ:

$$P(y > 0 \mid x) = 1 - \Phi\left(\frac{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}{\sigma}\right)$$

La expectativa condicional $E(y|y > 0, x)$ se puede calcular como:

$$E(y \mid y > 0, x) = x^T\beta + \sigma\phi\left(\frac{x^T\beta}{\sigma}\right) / \Phi\left(\frac{x^T\beta}{\sigma}\right)$$

donde ϕ es la función de densidad de la normal estándar y Φ es la función de distribución acumulada de la normal estándar. Simplificando, obtenemos:

$$E(y \mid y > 0, x) = x^T\beta + \sigma\phi\left(\frac{x^T\beta}{\sigma}\right) / \Phi\left(\frac{x^T\beta}{\sigma}\right)$$

a expectativa no condicional $E(y|x)$ se puede expresar como:

$$E(y \mid x) = \Phi\left(\frac{x^T\beta}{\sigma}\right) \cdot (x^T\beta + \sigma\phi\left(\frac{x^T\beta}{\sigma}\right) / \Phi\left(\frac{x^T\beta}{\sigma}\right))$$

Los efectos parciales de una variable continua x_j sobre $E(y|y > 0, x)$ están dados por:

$$\frac{\partial E(y \mid y > 0, x)}{\partial x_j} = \beta_j \cdot \left\{ 1 - \frac{\phi\left(\frac{x^T\beta}{\sigma}\right) \cdot \frac{x_j}{\sigma}}{\Phi\left(\frac{x^T\beta}{\sigma}\right)} \right\}$$

Para el cálculo de los efectos parciales en un modelo Tobit, el parámetro σ juega un papel crucial. Aunque no afecta el signo de los efectos parciales, sí influye en su magnitud. Por lo tanto, interpretar σ como un parámetro "auxiliar" puede ser engañoso, ya que afecta significativamente la importancia económica de las variables explicativas.

Para las variables discretas, como las binarias, el efecto parcial se calcula comparando los valores esperados de y para $x_j = 1$ y $x_j = 0$, manteniendo constantes las demás variables explicativas.

En resumen, para interpretar adecuadamente las estimaciones Tobit, es esencial comprender que los efectos parciales y la relación entre las variables explicativas y la variable dependiente son más complejos que en un modelo lineal simple. El parámetro σ es fundamental para estos cálculos y afecta la magnitud de los efectos parciales.

Problemas de especificación en los modelos Tobit El modelo Tobit y, en particular, las fórmulas para las expectativas en (13.3) y (13.4), dependen de manera crucial de la normalidad y la homocedasticidad en el modelo de la variable latente subyacente. Cuando el valor esperado de y dado x es lineal en x, es conocido que la normalidad condicional de y no afecta la insesgadez, consistencia o inferencia en muestras grandes. La heterocedasticidad no afecta el insesgamiento o consistencia de los estimadores de Mínimos Cuadrados Ordinarios (MCO), aunque es necesario calcular errores estándar y estadísticos de prueba robustos para realizar inferencia aproximada.

$$E(y|y > 0, x) = x\beta + \sigma\phi\left(\frac{x\beta}{\sigma}\right) / \Phi\left(\frac{x\beta}{\sigma}\right) \tag{13.3}$$

$$E(y|x) = \Phi\left(\frac{x\beta}{\sigma}\right)[x\beta + \sigma\phi\left(\frac{x\beta}{\sigma}\right)] \tag{13.4}$$

En el modelo Tobit, si cualquiera de los supuestos en falla, es difícil determinar qué está estimando la Estimación de Máxima Verosimilitud (EMV) Tobit. No obstante, para cambios moderados respecto a los supuestos, el modelo Tobit probablemente ofrecerá buenas estimaciones de los efectos parciales sobre las medias condicionales. Aunque es posible permitir supuestos más generales, tales modelos son más complejos de estimar e interpretar.

Una limitación importante del modelo Tobit es que el valor esperado condicional en $y > 0$ está estrechamente vinculado con la probabilidad de que $y > 0$. Esto se evidencia en las ecuaciones (13.5) y (13.6). Por ejemplo, considere la relación entre la cobertura del seguro de vida y la edad de una persona. Es menos probable que los jóvenes tengan seguro de vida, por lo que la probabilidad de que $y > 0$ aumenta con la edad (al menos hasta cierto punto). Sin embargo, dado que se tiene un seguro de vida, el valor de las pólizas puede disminuir con la edad, ya que el seguro se vuelve menos importante hacia el final de la vida. Esta posibilidad no se captura en el modelo Tobit.

Para evaluar informalmente si el modelo Tobit es adecuado, se puede estimar un modelo probit donde el resultado binario, por ejemplo, w, es igual a uno si $y > 0$ y $w = 0$ si $y = 0$. De acuerdo con la ecuación w sigue

un modelo probit, donde el coeficiente de x_j es γ_j/σ. Si el modelo Tobit es válido, la estimación probit, $\hat{\gamma}_j$, debe ser cercana a $\hat{\beta}_j/\hat{\sigma}$, donde $\hat{\beta}$ y $\hat{\sigma}$ son las estimaciones Tobit. Aunque estas estimaciones nunca serán idénticas debido al error de muestreo, se pueden buscar signos problemáticos. Por ejemplo, si $\hat{\gamma}_j$ es significativa y negativa, pero $\hat{\beta}_j$ es positiva, puede indicar que el modelo Tobit no es adecuado. De manera similar, si $\hat{\gamma}_j$ y $\hat{\beta}_j$ tienen el mismo signo, pero $\hat{\beta}_j/\hat{\sigma}$ es mucho mayor o menor que $\hat{\gamma}_j$, también podría señalar problemas.

Si se concluye que el modelo Tobit es inadecuado, existen modelos alternativos conocidos como modelos de dos partes o de obstáculos, que permiten que $P(y>0|x)$ y $E(y|y>0,x)$ dependan de diferentes parámetros. Estos modelos proporcionan flexibilidad adicional y pueden ser más apropiados si los determinantes de la censura y del valor esperado son diferentes.

$$P(y>0|x) = \Phi\left(\frac{x\beta}{\sigma}\right) \tag{13.5}$$

$$\frac{\partial E(y|x)}{\partial x_j} = \gamma_j \Phi\left(\frac{x\beta}{\sigma}\right) \tag{13.6}$$

Modelo Poisson

Otra clase de variable dependiente no negativa es una **variable de conteo**, la cual puede asumir valores enteros no negativos $0,1,2,\ldots$. Lo que más interesa aquí son los casos en los que y asume relativamente pocos valores, incluyendo el cero. Los ejemplos incluyen el número de hijos que ha tenido una mujer, el número de veces que alguien es arrestado en un año o el número de patentes que una empresa registra al año. Por las mismas razones discutidas en los modelos de respuestas binarias y Tobit, un modelo lineal para $E(y|x_1,\ldots,x_k)$ podría no proporcionar el mejor ajuste a lo largo de todos los valores de las variables explicativas.

Como no es posible tomar el logaritmo de una variable de conteo debido a que asume el valor cero, un método útil es modelar el valor esperado como una función exponencial:

$$\mathbb{E}(y \mid x_1, x_2, \ldots, x_k) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

Debido a que $exp(\beta)$ es siempre positivo, esta especificación asegura que los valores predichos para y también sean positivos. El logaritmo del valor esperado es lineal:

$$\log(\mathbb{E}(y \mid x_1, x_2, \ldots, x_k)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Por lo tanto, podemos interpretar los coeficientes como en un modelo lineal. En otras palabras, $100 \times \beta_j$ es el cambio porcentual aproximado en $E(y \mid x)$, dado un incremento de una unidad en x_j.

Para cambios discretos en el valor esperado, el cambio proporcional es:

$$\left[\frac{\exp(\beta_0 + x_k^{(1)}\beta_k)}{\exp(\beta_0 + x_k^{(0)}\beta_k)}\right] - 1 = \exp(\beta_k \Delta x_k) - 1$$

Si $\Delta x_k = 1$, entonces el cambio es $exp(\beta_k) - 1$. Dada una estimación $\hat{\beta}_k$, se puede calcular $exp(\hat{\beta}_k) - 1$ y multiplicar por 100 para obtener el cambio porcentual.

Cuando una variable explicativa es el logaritmo de alguna variable positiva, es decir, $x_j = \log(z_j)$, entonces el coeficiente β_j se interpreta como una elasticidad.

Como la ecuación es no lineal en sus parámetros, no se pueden usar métodos de regresión lineal ordinaria. Aunque se podrían usar mínimos cuadrados no lineales, las distribuciones estándar de datos de conteo suelen mostrar heterocedasticidad, por lo que aquí se prefiere la estimación por máxima verosimilitud (MV) o cuasi máxima verosimilitud (CMV).

La función de log-verosimilitud de la regresión de Poisson es:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \exp(x_i \beta) \right]$$

Los errores estándar se obtienen fácilmente tras maximizar esta función. Sin embargo, como las probabilidades y momentos mayores de la distribución de Poisson están determinados por su media, la varianza es igual a la media:

$$\text{Var}(y \mid x) = \mathbb{E}(y \mid x)$$

Este supuesto puede ser restrictivo y, en muchas aplicaciones, se observa sobredispersión, donde $Var(y \mid x) > \mathbb{E}(y \mid x)$. Para ajustar los errores estándar en presencia de sobredispersión, se utiliza un estimador consistente de la varianza ajustada:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{\hat{u}_i^2}{\hat{y}_i}$$

donde $\hat{u}_i = y_i - \hat{y}_i$ son los residuales y $\hat{y}_i = exp(x_i\hat{\beta})$ son los valores ajustados. Multiplicando los errores estándar nominales por $\hat{\sigma}$ se obtiene un ajuste adecuado para la sobredispersión.

Modelo de regresión censurada y truncada

Capítulo 14

Regresión con Variables Instrumentales (VI)

Si X_i y u_i están correlacionadas, el estimador MCO es inconsistente; es decir, puede no estar cercano al verdadero valor del coeficiente de regresión, incluso cuando la muestra es muy grande, esta correlacion puede provenir de variables omitidas, errores en las variables y causalidad simultánea. Cualquiera que sea el origen de la correlación entre X y u, si existe una variable instrumental válida, Z, el efecto sobre Y de un cambio unitario en X puede estimarse utilizando el estimador de variables instrumentales.

14.1 El estimador VI con regresor único e instrumento único

El modelo VI y los supuestos

El modelo de regresión poblacional que relaciona la variable dependiente Y y la variable independiente X es:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n \tag{14.1}$$

Donde u_i es el término de error que representa los factores omitidos que determinan Y_i. Si X_i y u_i están correlacionados, el estimador de mínimos cuadrados ordinarios (MCO) es inconsistente. La estimación con variables instrumentales (VI) utiliza una variable instrumental Z adicional para aislar la parte de X que no está correlacionada con u.

Endogeneidad y Exogeneidad

- **Endogeneidad:** Estan correlacionadas con el término de error poblacional, $corr(X, u) \neq 0$.
- **Exogeneidad:** No estan correlacionadas con el término de error poblacional, $corr(X, u) = 0$.

Condiciones para un intrumento válido: na variable instrumental válida debe cumplir dos condi- ciones:

- **Relevancia del intrumento:** $corr(Z, X) \neq 0$.
- **Exogeneidad del instrumento:** $corr(Z, u) = 0$.

Si un instrumento es relevante y exógeno, puede captar la variación exógena de X, que a su vez se utiliza para estimar el coeficiente poblacional β_1.

El estimador de mínimos cuadrados en 2 etapas

Si el instrumento Z cumple los requisitos de relevancia y exogeneidad, el coeficiente β_1 puede ser estimado mediante un estimador VI denominado de mínimos cuadrados en dos etapas (MC2E). Como el nombre sugiere, el estimador de mínimos cuadrados en dos etapas se calcula en dos fases.

Primera etapa La primera etapa descompone X en dos componentes: una componente problemática que puede estar correlacionada con el error de la regresión y otra componente sin problemas que no está correlacionada con el error. La primera etapa comienza con una regresión poblacional que liga a X con Z:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

Esta regresión proporciona la necesaria descomposición de X_i. Una componente es $\pi_0 + \pi_1 Z_i$, la parte de X_i que puede predecirse mediante Z_i. Debido a que Z_i es exógena, esta componente de X_i está incorrelacionada con u_i. El otro componente de X_i es v_i, que es la componente problemática de X_i que está correlacionada con u_i. la primera etapa de MC2E consiste en aplicar MCO y utilizar los valores de predicción de la regresión MCO, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$.

Segunda etapa La segunda etapa de MC2E es sencilla: la regresión de Y_i sobre $\hat{X}_i$ por MCO. Los estimadores resultantes de la regresión de la segunda etapa son los estimadores MC2E, $\hat{\beta}_{0,MC2E}$ y $\hat{\beta}_{1,MC2E}$.

Distribución muestral del estimador MC2E

La distribución exacta del estimador MC2E para muestras pequeñas es complicada. No obstante, como la del estimador MCO, su distribución en muestras grandes es muy sencilla: el estimador MC2E es consistente y se distribuye normalmente. (desarrollado en el apéndice 3).

Fórmula del estimador MC2E A pesar de que las dos etapas de MC2E hacen que el estimador parezca complicado, cuando hay una sola X y un único instrumento Z, existe una fórmula sencilla para el estimador MC2E. Sea s_{ZY} la covarianza muestral entre Z e Y y sea s_{ZX} la covarianza muestral entre Z y X. Como se muestra en el Apéndice 12.2, el estimador MC2E con un único instrumento es:

$$\hat{\beta}_{1,\text{MC2E}} = \frac{s_{ZY}}{s_{ZX}}$$

Es decir, el estimador MC2E de β_1 es el cociente entre la covarianza muestral entre Z e Y y la covarianza muestral entre Z y X.

Inferencia estadística mediante la distribución para muestras grandes La varianza $\sigma^2_{\hat{\beta}_1,MC2E}$ se puede estimar mediante la estimación de los términos de varianza y covarianzas que aparecen en la Ecuación (14.2), y la raíz cuadrada de la estimación de $\sigma^2_{\hat{\beta}_1,MC2E}$ es el error estándar del estimador VI. Esto se obtiene automáticamente mediante los comandos de la regresión MC2E de los paquetes de software econométrico.

$$\sigma^2_{\hat{\beta}_1\,MC2E} = \frac{1}{n}\frac{Var[(Z_i - \mu_z)\mu_i]}{[Cov(Z_i, X_i]^2} \tag{14.2}$$

Debido a que $\hat{\beta}_{1,\text{MC2E}}$ se distribuye normalmente en muestras grandes, los contrastes de hipótesis acerca de β_1 se pueden realizar mediante el cálculo del estadístico t, y un intervalo de confianza al 95 % para muestras grandes viene dado por:

$$\hat{\beta}_{1,\text{MC2E}} \pm 1.96 \cdot \text{ES}(\hat{\beta}_{1,\text{MC2E}})$$

14.2 El modelo general de regresión VI

El modelo general de regresión de variables instrumentales (VI) incluye cuatro tipos de variables:

- **Variable dependiente** (Y)
- **Regresores endógenos** (X): Variables problemáticas que están correlacionadas con el término de error, como el precio de los cigarrillos.
- **Regresres exógenos incluidos** (W): Variables adicionales que no están correlacionadas con el término de error.

- **Variables instrumentales** (Z): Se utilizan para instrumentar los regresores endógenos.

En general, puede haber varios regresores endógenos (X), varios regresores exógenos incluidos (W) y varias variables instrumentales (Z). Para que la regresión VI sea posible, debe haber al menos tantas variables instrumentales (Z) como regresores endógenos (X). En la Sección 12.1, se discutió el caso con un único regresor endógeno y un único instrumento. Tener al menos un instrumento para cada regresor endógeno es esencial para calcular el estimador de variables instrumentales, ya que sin el instrumento no existiría la regresión de la primera etapa del método de mínimos cuadrados en dos etapas (MC2E). La relación entre el número de instrumentos y el número de regresores endógenos se describe con la siguiente terminología:

- **Exactamente identificads**: Si el número de instrumentos (m) es igual al número de regresores endógenos (k), es decir, m = k.
- **Sobreidentificados**: Si el número de instrumentos supera al número de regresores endógenos, es decir, m >k.
- **Subidentificados**: Si el número de instrumentos es menor que el número de regresores endógenos, es decir, m <k.

Concept clave 12.1: El modelo general de regresión de variables instrumentales y su terminología

El modelo general de regresión VI es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i \tag{14.3}$$

Donde:
1. Y_i es la variable dependiente.
2. $X_{1i}, X_{2i}, \ldots, X_{ki}$ son regresores endógenos.
3. $W_{1i}, W_{2i}, \ldots, W_{ri}$ son regresores exógenos.
4. $Z_{1i}, Z_{2i}, \ldots, Z_{mi}$ son variables instrumentales.

Los coeficientes deben estar o bien exactamente identificados o bien sobreidentificados para ser estimados mediante la regresión VI. Las variables W en la Ecuación (14.3) pueden ser:

- **Variables exógenas**
- **Variables de control:** No necesitan tener una interpretación causal, sino que se incluyen para garantizar que el instrumento no esté correlacionado con el término de error.

Condiciones para Variables de Control Efectivas Si W es una variable de control efectiva en la regresión VI, entonces su inclusión hace que el instrumento no esté correlacionado con u, y el estimador MC2E del coeficiente de X es consistente. Sin embargo, si W está correlacionada con u, el coeficiente MC2E de W está sujeto a un sesgo de variable omitida y no tiene una interpretación causal. La lógica de las variables de control en la regresión VI es paralela a la lógica de las variables de control en MCO.

Condición matemática La condición matemática para que W sea una variable de control efectiva en la regresión VI es similar a la condición sobre las variables de control en MCO. En concreto, la inclusión de W debe asegurar que la media condicional de u no dependa de Z, cumpliendo así la independencia de la media condicional:

$$E(u_i \mid Z_i, W_i) = E(u_i \mid W_i)$$

Para mayor claridad, en la parte principal de este capítulo nos centramos en el caso en el que las variables W son exógenas, por lo que $E(u_i|W_i) = 0$. En el Apéndice 12.6 se explica cómo extender los resultados para el caso en el que W sea una variable de control, sustituyendo la condición de media condicional igual a cero por la condición de independencia en media condicional.

14.3 MC2E en el modelo general VI

Cuando existe un único regresor endógeno X y algunas variables exógenas incluidas adicionales, la ecuación de interés es:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

donde Xi podría estar correlacionada con el término de error, pero $W_{1i}, \ldots, W_{ri}$ no lo están.

Primera Etapa de MC2E La regresión poblacional de la primera etapa de MC2E relaciona X con las variables exógenas, es decir, las W y los instrumentos (Z):

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+r} W_{ri} + v_i$$

donde $\pi_0, \pi_1, \ldots, \pi_{m+r}$ son los coeficientes de regresión desconocidos y vi es un término de error. Esta ecuación se denomina a veces ecuación de la forma reducida para X. En la primera etapa de MC2E, los coeficientes desconocidos d se estiman por MCO, y los valores de predicción de esta regresión son $\hat{X}_1, \ldots, \hat{X}_n$.

Segunda Etapa de MC2E En la segunda etapa de MC2E, los coeficientes desconocidos se estiman por MCO, excepto que X_i se sustituye por su valor estimado en la primera etapa. Es decir, Y_i se regresa sobre $\hat{X}_i, W_{1i}, \ldots, W_{ri}$ mediante MCO. El estimador resultante de $\beta_0, \ldots, \beta_{1+r}$ es el estimador MC2E.

Extensión a múltiples regresores endógenos Cuando existen varios regresores endógenos, el algoritmo MC2E es similar, excepto que cada regresor endógeno requiere su propia regresión en la primera etapa. En conjunto, estas regresiones de la primera etapa dan lugar a valores de predicción para cada uno de los regresores endógenos. En la segunda etapa se estima MCO sustituyendo X por sus estimados predictivos $\hat{X}$. El estimador resultante de $\beta_0, \ldots, \beta_{k+r}$ es el estimador MC2E.

Relevancia y Exogeneidad de los Instrumentos as condiciones de relevancia y exogeneidad de instrumentos necesitan ser modificadas para el modelo de regresión VI general.

- **Relevancia:**Cuando existe una única variable endógena incluida pero varios instrumentos, la condición para la relevancia de los instrumentos es que al menos una Z sea útil para predecir X, dado W . Cuando existen varias variables endógenas incluidas, esta condición es más complicada porque hay que descartar multicolinealidad perfecta en la regresión poblacional de la segunda etapa.
- **Exogeneidad:** La condición general del requisito de exogeneidad del instrumento es que cada instrumento debe estar incorrelacionado con el término de error u_i.

Concepto clave 12.3: Las 2 condiciones para la validez de los intrumentos

Un conjunto de m instrumentos $Z_{1i}, ..., Z_{mi}$ debe cumplir las dos condiciones siguientes para ser válido:

1. **Relevancia:**
   - En general, sea $\hat{X}^*_{1i}$ el valor de predicción de X_{1i} a partir de la regresión poblacional de X_{1i} sobre los instrumentos (Z) y los regresores exógenos incluidos (W), y sea «1» la expresión del regresor constante que toma el valor 1 para todas las observaciones. Entonces $(\hat{X}^*_{1i}, ..., hatX^*_{ki}, W_{1i}, W_{ri}, 1)$ no son perfectamente multicolineales.
   - Si solo hay una X, entonces para que se cumpla la condición anterior, al menos una Z debe tener un coeficiente distinto de cero en la regresión poblacional de X sobre las Z y las W.
2. **Exogeneidad:** Los instrumentos no están correlacionados con el término de error; es decir, $corr(Z_{1i}, u_i) = 0, ..., corr(Z_{mi}, u_i) = 0$.

Supuestos de la Regresión VI y la Distribución Muestral del Estimador MC2E

Bajo los supuestos de la regresión VI, el estimador MC2E es consistente y tiene una distribución muestral que, en muestras grandes, es aproximadamente normal.

Supuestos de la regresión VI Son modificaciones de los supuestos de mínimos cuadrados para el modelo de regresión múltiple.

1. **Media condicional:** Se aplica solamente a las variables exógenas incluidas.
2. **Extracciones i.i.d.:** Las extracciones son independientes e idénticamente distribuidas, como en un muestreo aleatorio simple.
3. **Valores extremos:** Los valores extremos grandes son poco probables.
4. **Validez de los instrumentos:** Se satisfacen las dos condiciones para la validez de los instrumentos del Concepto clave 12.3.

La condición de relevancia del instrumento del Concepto clave 12.3 implica el cuarto supuesto de mínimos cuadrados del Concepto clave 4.6 (ausencia de multicolinealidad perfecta), suponiendo que las variables explicativas de la regresión de la segunda etapa no son perfectamente multicolineales.

Concepto clave 12.4: Los supuestos de la regresión VI

Las variables y los errores del modelo de regresión VI del Concepto clave 12.1 satisfacen lo siguiente:

1. $E(u_i \mid W_{1i}, ..., W_{ri}) = 0$;
2. $(X_{1i}, ..., X_{ki}, W_{1i}, ..., W_{ri}, Z_{1i}, ..., Z_{mi}, Y_i)$ son extracciones i.i.d. de su distribución conjunta;
3. Los valores extremos elevados son poco probables: las X, W, Z, e Y tienen momentos de cuarto orden finitos y distintos de cero; y
4. Se cumplen las dos condiciones para que un instrumento sea válido del Concepto clave 12.3.

Distribución Muestral del Estimador MC2E Bajo los supuestos de la regresión VI, el estimador MC2E es consistente y tiene una distribución normal en muestras grandes. Esto se muestra en la Sección 12.1 (y en el Apéndice 12.3) para el caso particular de un único regresor endógeno, un único instrumento, y sin variables exógenas incluidas. Conceptualmente, el razonamiento expuesto en la Sección 12.1 se traslada al caso general de varios instrumentos y varias variables endógenas incluidas. Sin embargo, las expresiones para el caso general son complicadas y se trasladan al Capítulo 18.

Inferencia mediante el estimador MC2E Debido a que la distribución muestral del estimador MC2E es normal en muestras grandes, los procedimientos generales para la inferencia estadística (contrastes de hipótesis e intervalos de confianza) de los modelos de regresión se extienden a la regresión MC2E. Del mismo modo, las hipótesis conjuntas sobre los valores de los coeficientes poblacionales se pueden contrastar mediante el estadístico F.

Cálculo de los Errores Estándar MC2E Hay dos cuestiones a tener en cuenta sobre los errores estándar MC2E:

1. **Errores estándar incorrectos:** Los errores estándar obtenidos mediante la estimación MCO de la regresión de la segunda etapa son incorrectos porque no se tiene en cuenta que es la segunda etapa de un proceso de dos etapas. Las fórmulas de los errores estándar que realizan los ajustes necesarios están incorporadas en los comandos de la regresión MC2E del software econométrico (Se utilizan automáticamente).
2. **Heterocedasticidad:** El error u podría ser heterocedástico. Por ello, es importante utilizar las versiones de los errores estándar heterocedástico-robustos, exactamente por la misma razón que es importante la utilización de errores estándar heterocedástico-robustos para los estimadores MCO del modelo de regresión múltiple.

14.4 Verificación de la validez de los instrumentos

El hecho de que la regresión de variables instrumentales resulte útil en un caso concreto depende de si los instrumentos son válidos: los instrumentos no válidos dan lugar a resultados que carecen de sentido. Los instrumentos que explican una pequeña proporción de la variación de X se denominan instrumentos débiles.

Comprobación de la Debilidad de los Instrumentos

Una forma de comprobar los instrumentos débiles cuando existe un único regresor endógeno consiste en calcular el estadístico F para el contraste de la hipótesis de que todos los coeficientes de los instrumentos son iguales a cero en la regresión de la primera etapa de MC2E. Este estadístico F de la primera etapa proporciona una medida del contenido de la información incluida en los instrumentos: cuanta más información contengan, mayor es el valor esperado del estadístico F.

Una regla práctica sencilla es que no es necesario preocuparse de los instrumentos débiles si el es- tadístico F de la primera etapa es mayor que 10. (Para más detalles, véase el Apéndice 12.5). Esto se resume en el Concepto clave 12.5.

Concepto clave 12.5: Una regla práctica para la verificación de instrumentos débiles

El estadístico F de la primera etapa es el estadístico F para contrastar la hipótesis de que los coeficientes de los instrumentos $Z_{1i}, ..., Z_{mi}$ son iguales a cero en la primera etapa de los mínimos cuadrados en dos etapas. Cuando existe un único regresor endógeno, un estadístico F en la primera etapa menor que 10 indica que los instrumentos son débiles, en cuyo caso el estimador MC2E es sesgado (incluso en muestras grandes) y los estadísticos t MC2E y los intervalos de confianza son poco fiables.

¿Qué hacer si se tiene instrumentos débiles? Si se tienen muchos instrumentos, probablemente algunos de esos instrumentos sean más débiles que otros. Si se tiene un número pequeño de instrumentos fuertes y muchos débiles, será mejor descartar el más débil de los instrumentos y utilizar el subconjunto de los más relevantes para el análisis MC2E. Los errores estándar MC2E podrían aumentar cuando se quitan los instrumentos débiles, pero es necesario tener en cuenta que los errores estándar originales no eran significativos.

Sin embargo, si los coeficientes están exactamente identificados, no se pueden descartar los instrumen- tos débiles. Aun cuando los coeficientes estén sobreidentificados, puede que no se disponga de suficientes instrumentos fuertes para lograr la identificación, por lo que desechar algunos instrumentos débiles no ayudará. En este caso, existen dos opciones:

1. **Encontrar instrumentos adicionales, fuertes:** sto requiere un conocimiento profundo del problema en cuestión y puede implicar el rediseño del conjunto de datos y de la naturaleza del estudio empírico.
2. **Continuar el análisis empírico con los instrumentos débiles:** mpleando métodos distintos de MC2E. Algunos otros métodos de análisis de variables instrumentales son menos sensibles a los instrumentos débiles que MC2E, y algunos de estos métodos se tratan en el Apéndice 12.5.

Supuestos

1. **Relevancia de los instrumentos** l papel de la condición de relevancia de los instrumentos en la regresión VI es sutil. Una forma de entender la relevancia de los instrumentos es compararla con el tamaño de la muestra: cuanto más relevantes sean los instrumentos —es decir, cuanta más variación de X se explique por medio de los instrumentos— más información está disponible para su uso en la regresión VI. Un instrumento más relevante da lugar a un estimador más preciso, al igual que un tamaño muestral más grande da lugar a un estimador más preciso.
   Por otra parte,la **inferencia estadística mediante MC2E** se basa en que el estimador MC2E tenga una distribución muestral normal. De acuerdo con el teorema central del límite, la distribución normal es una buena aproximación para uestras grandes, pero no necesariamente para muestras pequeñas. Si disponer de una mayor relevancia de los instrumentos es como disponer de un tamaño de muestra mayor, esto sugiere que cuanto más relevante sea el instrumento, mejor es la aproximación normal para la distribución muestral del estimador MC2E y su estadístico t.
2. **Exogeneidad de los instrumentos** Si los instrumentos no son exógenos, entonces MC2E es inconsistente: el estimador MC2E converge en probabilidad a algo distinto del coeficiente poblacional de la regresión. La idea de la regresión con variables instrumentales es que el instrumento contenga información sobre la variación de Xi que no esté correlacionada con el término de error ui. Si el instrumento no es exógeno, no se puede

identificar esta variación exógena en Xi, y la regresión VI no proporcionará un estimador consistente. Las matemáticas que respaldan este argumento están resumidas en el Apéndice 12.4.

¿Es posible contrastar estadísticamente la hipótesis de que los instrumentos son exógenos?

- **Exactamente identificados:** No es posible contrastar la hipótesis de que los instrumentos son exógenos cuando los coeficientes están exactamente identificados. No se puede utilizar la evidencia empírica para resolver si estos instrumentos satisfacen el requisito de exogeneidad. En este caso, la única forma de evaluar si los instrumentos son exógenos es recurrir a una opinión experta y al conocimiento personal de los problemas empíricos que se están analizando.
- **Sobreidentificados:** Si los coeficientes están sobreidentificados, es posible contrastar la sobrei- dentificación de las restricciones, es decir, contrastar la hipótesis de que los instrumentos extras son exógenos bajo el cumplimiento del supuesto de que existen suficientes instrumentos válidos para identificar los coeficientes de interés.

Evaluación de la exogeneidad Para evaluar si los instrumentos son exógenos se requiere necesariamente un criterio técnico basado en el conocimiento personal del caso concreto. Por ejemplo, el conocimiento de Philip Wright sobre la oferta y la demanda agrícolas le llevó a sugerir que las lluvias por debajo de la media posiblemente desplazarían la curva de oferta de la mantequilla, pero que no desplazarían directamente la curva de demanda. (Si hay más instrumentos que regresores endógenos, puede ser útil en este proceso el conocido como contraste de sobreidentificación de restricciones).

Contraste de Sobreidentificación de Restricciones Supongamos que se dispone de un único re- gresor endógeno y de dos instrumentos. Se pueden calcular dos estimadores MC2E diferentes: uno que utilice el primer instrumento y otro que utilice el segundo. Estos dos estimadores no serán iguales debido a la variación muestral, pero si ambos instrumentos son exógenos, tenderán a estar cerca uno del otro. Si estos dos instrumentos dan lugar a estimaciones muy diferentes, se puede concluir que hay algo incorrecto en uno o ambos instrumentos, es decir, que no son exógenos.

El contraste de sobreidentificación de restricciones realiza implícitamente esta comparación sin calcular todas las posibles estimaciones VI diferentes. La exogeneidad de los instrumentos implica que no están correlacionados con u_i. Esto sugiere que los instrumentos deberían estar aproximadamente incorrelacionados con $\hat{u}_{MC2E_i}$, donde:

$$\hat{u}_{MC2E_i} = Y_i - (\hat{\beta}_{MC2E_0} + \hat{\beta}_{MC2E_1} X_{1i} + \ldots + \hat{\beta}_{MC2E_k} X_{ki})$$

es el residuo de la regresión estimada MC2E utilizando todos los instrumentos. Si los instrumentos son exógenos, los coeficientes de los instrumentos en una regresión de $\hat{u}_{MC2E_i}$ sobre los instrumentos y las variables exógenas deberían ser iguales a cero, y esta hipótesis se puede contrastar.

Este método se resume en el Concepto clave 12.6. El estadístico se calcula utilizando el estadístico F válido con homocedasticidad y se denomina comúnmente estadístico J. En muestras grandes, si los instrumentos no son débiles y los errores son homocedásticos, bajo la hipótesis nula de que los instrumentos son exógenos, el estadístico J presenta una distribución chi-cuadrado con $m-k$ grados de libertad (χ^2_{m-k}). Aunque el número de restricciones que se contrastan sea m, los grados de libertad de la distribución asintótica del estadístico J son $m-k$ porque solo es posible contrastar las restricciones sobreidentificadas, de las que hay $m-k$. La modificación del estadístico J para errores heterocedásticos se ofrece en la Sección 18.7.

Concepto clave 12.6: Contraste de sobreidentificación de restricciones

Sea $\hat{u}_i^{MC2E}$ el residuo de la estimación MC2E de la Ecuación (12.12). Se utiliza MCO para la estimación de los coeficientes de regresión en:
Donde e_i es el término de error de la regresión. Sea F la expresión del estadístico F válido con homocedasticidad para el contraste de la hipótesis de que $\delta_1 = \cdots = \delta_n = 0$. El estadístico para el contraste de sobreidentificación de restricciones es $J = mF$. Bajo la hipótesis nula de que todos los instrumentos son exógenos, si e_i es homocedástico, en muestras grandes J se distribuye χ^2_{m-k}, donde $m-k$ es el «grado de sobreidentifica- ción», es decir, el número de instrumentos menos el número de regresores endógenos.

Para comprobar que no se puede contrastar la exogeneidad de los regresores cuando los coeficientes están exactamente identificados ($m = k$), consideremos el caso de una sola variable endógena incluida ($k = 1$). Si hay dos instrumentos, se pueden calcular dos estimadores MC2E, uno por cada instrumento, y compararlos. Pero si solo se dispone de un instrumento, se puede calcular un solo estimador MC2E y no se dispone de otro con el que compararlo. Si los coeficientes están exactamente identificados ($m = k$), el estadístico J de contraste de sobreidentificación es exactamente igual a cero (seríaN una $\chi^2_{m-k=0}$).

14.5 ¿De donde provienen los instrumentos válidos

Procedencia de los instrumento válidos

En la práctica, el aspecto más difícil de la estimación VI es encontrar instrumentos que sean relevantes y exógenos. Existen dos métodos principales para identificar estos instrumentos, reflejando dos perspectivas diferentes sobre la modelización en Econometría y Estadística.

Método basado en la teoría económica

El primer método utiliza la teoría económica para sugerir instrumentos. Por ejemplo, el conocimiento de Philip Wright sobre la economía de los mercados agrícolas le llevó a buscar un instrumento que desplazara la curva de oferta pero no la curva de demanda, considerando las condiciones climáticas de las regiones agrícolas. Este enfoque ha sido particularmente exitoso en la economía financiera, donde algunos modelos económicos sobre el comportamiento de los inversores incluyen variables que están incorrelacionadas con el término de error. Estos modelos a veces son no lineales en los datos y parámetros, requiriendo una extensión de los métodos de VI a los modelos no lineales, denominada estimación del método generalizado de momentos (GMM). Sin embargo, las teorías económicas son abstracciones que a menudo no consideran los matices y detalles necesarios para el análisis de una base de datos específica, por lo que este método no siempre es efectivo.

Método basado en la variación exógena

El segundo método consiste en buscar una fuente exógena de variación en X que surja de un fenómeno aleatorio que induce cambios en el regresor endógeno. Por ejemplo, en un caso hipotético, los daños de un terremoto pueden aumentar el tamaño promedio de las clases en algunos distritos escolares, y esta variación no estaría correlacionada con las variables que afectan al rendimiento estudiantil. Este método requiere un conocimiento profundo del problema estudiado y una cuidadosa atención a los detalles de los datos, lo cual se explica mejor a través de ejemplos.

14.6 Apéndice

Apéndice 12.2: obtención de la fórmula del estimador MC2

Esta fórmula es válida cuando hay un solo regresor y un solo instrumento. La fórmula para el estimador MC2E, expresada en términos del valor de predicción $\hat{X}_i$, es similar a la fórmula del estimador MCO (Concepto clave

4.2), pero con $\hat{X}_i$ sustituyendo a X_i:

$$\hat{\beta}_{\text{MC2E}} = \frac{s_{\hat{X}Y}}{s^2_{\hat{X}}}$$

Donde $s^2_{\hat{X}_i}$ es la varianza muestral de $\hat{X}_i$ y $s_{\hat{X}Y}$ es la covarianza muestral entre Y_i y $\hat{X}_i$.

Debido a que $\hat{X}_i$ es el valor de predicción de X_i en la primera etapa, es decir, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, se tiene que:

$$s_{\hat{X}Y} = \hat{\pi}_1 s_{ZY} \quad \text{y} \quad s^2_{\hat{X}} = \hat{\pi}^2_1 s^2_Z$$

Por lo tanto, el estimador MC2E puede expresarse como:

$$\hat{\beta}_{\text{MC2E}} = \frac{s_{ZY}}{\hat{\pi}_1 s^2_Z}$$

Finalmente, como $\hat{\pi}_1$ es el coeficiente MCO de la regresión de X_i sobre Z_i, se cumple que:

$$\hat{\pi}_1 = \frac{s_{ZX}}{s^2_Z}$$

Sustituyendo esta expresión en la fórmula anterior, se obtiene la fórmula del estimador MC2E:

$$\hat{\beta}_{\textbf{MC2E}} = \frac{s_{ZY}}{(s_{ZX}/s^2_Z)s^2_Z} = \frac{s_{ZY}}{s_{ZX}}$$

Apéndice 12.3: Distribución para muestras grandes

Este apéndice analiza la distribución del estimador MC2E para muestras grandes, bajo el supuesto de un único instrumento y una única variable endógena, sin variables exógenas. Se parte de una expresión del estimador MC2E en términos de los errores.

Partiendo de la Ecuación (14.1), se tiene:

$$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$$

De aquí, la covarianza muestral entre Z y Y se puede expresar como:

$$s_{ZY} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y}) = \beta_1 s_{ZX} + \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(u_i - \bar{u})$$

Donde $s_{ZX} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})$. Al sustituir estas expresiones en la fórmula del estimador MC2E y reescalar por $(n-1)/n$, obtenemos:

$$\hat{\beta}_{\text{MC2E}} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})} \tag{14.4}$$

Cuando se cumplen los supuestos de la regresión de VI La Ecuación (14.4) presenta el estimador MC2E de una forma similar al estimador MCO (mínimos cuadrados ordinarios). La principal diferencia es que, en lugar de aparecer X en el numerador, ahora aparece Z (el instrumento), y en lugar de la varianza de X en el denominador, aparece la covarianza entre Z y X.

Debido a que Z es exógeno (es decir, no está correlacionado con el término de error u), el mismo argumento que muestra que el estimador MCO se distribuye normalmente en muestras grandes se puede aplicar al estimador MC2E.

En detalle, cuando la muestra es grande, la covarianza muestral entre Z y X (s_{ZX}) es un buen estimador de la verdadera covarianza poblacional $cov(Z_i, X_i)$. Dado que el instrumento es relevante (es decir, $cov(Z_i, X_i) \neq 0$), podemos escribir la distribución del estimador MC2E como una distribución normal.

El numerador del estimador MC2E se aproxima a una suma de $q_i = (Z_i - \hat{Z})u_i$, donde $E(q_i) = 0$ debido a la exogeneidad de Z. Por el teorema central del límite, esta suma se distribuye normalmente en grandes muestras.

Por lo tanto, el estimador MC2E se distribuye en muestras grandes aproximadamente como una normal $N(\beta_1, \sigma^2)$, donde la varianza del estimador está dada por:

$$\text{var}(\hat{\beta}_{\text{MC2E}}) = \frac{\text{var}[(Z_i - \bar{Z})u_i]}{[\text{cov}(Z_i, X_i)]^2 / n}$$

Esta es la misma expresión que aparece en la Ecuación (14.2) del libro.

Capítulo 15

Experimentos y Cuasi Experimentos

En muchas áreas como la psicología y la medicina, los efectos causales se estiman mediante la utilización de experimentos. Existen tres razones para estudiar los experimentos aleatorizados controlados en un curso de econometría:

1. Un experimento aleatorizado controlado ideal proporciona un punto de referencia conceptual para juzgar las estimaciones de efectos causales realizadas con datos observacionales.
2. Los resultados de estos experimentos pueden ser muy influyentes, por lo que es importante entender sus limitaciones y amenazas a la validez, así como sus puntos fuertes.
3. Las circunstancias externas a veces originan situaciones que parecen aleatorias, dando lugar a cuasi experimentos o experimentos naturales, donde muchos métodos desarrollados para analizar experimentos aleatorizados pueden aplicarse con algunas modificaciones.

15.1 Variables de respuesta, efectos causales y experimentos ideales

Los datos provenientes de un *experimento aleatorizado controlado* pueden ser analizados mediante la comparación de las diferencias en las medias, o bien mediante una regresión que incluya una variable indicadora del tratamiento junto con otras variables de control adicionales. Esta última especificación, conocida como **estimador de las diferencias con regresores adicionales**, puede aplicarse también en esquemas de aleatorización más complejos, donde las probabilidades de aleatorización dependen de covariables observables.

El estimador de las diferencias

El **estimador de las diferencias** se refiere a la diferencia en las medias muestrales entre los grupos de tratamiento y de control. Esto se puede calcular mediante la regresión de la variable respuesta Y sobre un indicador binario de tratamiento X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n. \tag{15.1}$$

Como se mencionó en la Sección 4.4, si X se asigna al azar, entonces $E(u_i \mid X_i) = 0$, y el estimador de Mínimos Cuadrados Ordinarios (MCO) de β_1 en la ecuación anterior es un estimador insesgado y consistente del efecto causal.

El estimador de las diferencias con variables explicativas adicionales

La eficiencia del estimador de las diferencias puede mejorarse incluyendo algunas variables de control W en la regresión, lo cual da lugar al estimador de las diferencias con regresores adicionales:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i, \quad i = 1, \ldots, n. \tag{15.2}$$

Si W ayuda a explicar la variación en Y , la inclusión de W reduce el error estándar de la regresión y, por lo tanto, el error estándar de $\hat{\beta}_1$. Para que el estimador $\hat{\beta}_1$ sea insesgado, las variables de control W deben cumplir la condición de independencia en media condicional, es decir, $E(u_i \mid X_i, W_i) = E(u_i \mid W_i)$. Esta condición se

cumple si W_i son características pretratamiento, tales como el género. Es importante destacar que los regresores W no deben incluir resultados del experimento, ya que X_i no es asignado al azar dado un resultado experimental. El coeficiente de la variable de control en este contexto no tiene una interpretación causal.

Estimación de efectos causales que dependen de variables observables

Como se discutió en el Capítulo 8, la variación en los efectos causales que dependen de variables observ- ables puede estimarse incluyendo funciones no lineales apropiadas o interacciones con X_i. Por ejemplo, si $W1i$ es un indicador binario que representa el género, los efectos causales diferentes para hombres y mujeres pueden estimarse mediante la inclusión de la variable de interacción $W_{1i} \times X_i$ en la regresión de la ecuación anterior.

Aleatorización basada en las covariables

La aleatorización basada en las covariables ocurre cuando la probabilidad de asignación al grupo de tratamiento depende de una o más variables observables W . En este caso, el estimador de las diferencias basado en la ecuación anterior puede presentar sesgo de variable omitida. Un ejemplo de esto se discute en el Apéndice 7.2, donde se describe un experimento hipotético para estimar el efecto causal de las tareas obligatorias frente a las optativas en un curso de econometría. En dicho experimento, los estudiantes de economía ($W_i = 1$) fueron asignados al grupo de tratamiento (tareas obligatorias, $X_i = 1$) con una probabilidad más alta que los estudiantes de otras titulaciones ($W_i = 0$). Sin embargo, si los estudiantes de economía tienden a tener un mejor desempeño que los estudiantes de otras titulaciones, existe un sesgo de variable omitida debido a la correlación entre el tratamiento y la variable omitida, ser estudiante de economía.

Dado que X_i fue asignado aleatoriamente condicionado a W_i, este sesgo de variable omitida puede eliminarse utilizando el estimador de las diferencias con la variable de control adicional W_i. La asignación aleatoria de X_i dado W_i (combinada con el supuesto de una función de regresión lineal) implica que, dado W_i, X_i es independiente de ui en la ecuación anterior. Esta independencia condicional implica la independencia en media condicional, es decir, $E(u_i \mid X_i, W_i) = E(u_i \mid W_i)$. Por lo tanto, el estimador MCO $\hat{\beta}_1$ en la ecuación es un estimador insesgado del efecto causal cuando X_i se asigna aleatoriamente basado en W_i.

15.2 Amenazas a la validez interna de los experimentos

Amenazas a la validez de los experimentos Un estudio estadístico es **internamente válido** si las inferencias estadísticas acerca de los efectos causales son válidas para la población estudiada. Es externamente válido si esas inferencias pueden generalizarse a otras poblaciones y escenarios. Sin embargo, en el mundo real, diversos problemas pueden amenazar tanto la validez interna como externa de los análisis estadísticos de los experimentos con seres humanos.

Amenazas a la validez interna

Las principales amenazas a la validez interna en experimentos aleatorizados controlados incluyen la ausencia de aleatoriedad, el incumplimiento del protocolo de tratamiento, la deserción, los efectos exper- imentales y los tamaños muestrales pequeños.

1. **Ausencia de aleatoriedad:**Si el tratamiento no se asigna al azar y depende de las características o preferencias del sujeto, los resultados experimentales pueden reflejar tanto el efecto del tratamiento como el de la asignación no aleatoria. Esto introduce un sesgo en el estimador del efecto del tratamiento. Para contrastar la aleatoriedad del tratamiento, se puede realizar una regresión de la variable de tratamiento sobre características pretratamiento W y realizar un test F sobre los coeficientes de W.
2. **Incumplimiento del protocolo de tratamiento:** En muchos experimentos, no todos los sujetos siguen el protocolo asignado. Por ejemplo, algunos sujetos asignados al grupo de tratamiento pueden no recibir el tratamiento y algunos del grupo de control podrían recibirlo. Este cumplimiento parcial del protocolo introduce un sesgo en el estimador MCO debido a la correlación entre X_i y u_i. En casos donde se dispone de datos tanto del tratamiento recibido como del asignado, se puede utilizar una estrategia de variables instrumentales para estimar el efecto del tratamiento.

3. **Deserción o abandono:**La deserción ocurre cuando sujetos asignados aleatoriamente al grupo de tratamiento o control abandonan el estudio. Si la deserción está relacionada con el tratamiento, introduce un sesgo en el estimador MCO debido a la correlación entre X_i y ui para los que permanecen en la muestra. Esto conduce a un sesgo de selección.
4. **Efectos experimentales:**En experimentos con seres humanos, el comportamiento de los sujetos puede cambiar simplemente por saber que están en un experimento, conocido como el **efecto Hawthorne**. Un protocolo de doble ciego puede mitigar este efecto, pero es inviable en experimentos económicos donde tanto los sujetos como los instructores saben quién está en el grupo de tratamiento.
5. **Muestras pequeñas:**Un tamaño de muestra pequeño no sesga los estimadores, pero puede llevar a estimaciones impre- cisas del efecto causal. También plantea problemas para la validez de los intervalos de confianza y los contrastes de hipótesis, especialmente si los errores no siguen una distribución normal.

Amenazas a la validez externa

Las amenazas a la validez externa comprometen la capacidad de generalizar los resultados a otras pobla- ciones y escenarios. Las principales amenazas incluyen:

1. **Muestra no representativa:** Si la población estudiada no es representativa de la población de interés, los resultados del estudio podrían no ser generalizables. Por ejemplo, un programa evaluado con exreclusos podría no ser generalizable a trabajadores que nunca han cometido un delito.
2. **Programa o política no representativa:**La política o programa de interés debe ser sufi- cientemente similar al programa estudiado para permitir la generalización de los resultados. La implementación a gran escala de un programa experimental podría no reflejar los mismos resultados debido a diferencias en control de calidad, financiación o duración.
3. **Efectos de equilibrio general:**Ampliar un programa experimental pequeño a gran escala puede cambiar el entorno económico, afectando la generalización de los resultados. Por ejemplo, un pro- grama de formación laboral podría complementar la capacitación proporcionada por empleadores a pequeña escala, pero desplazarse a gran escala podría reducir los beneficios netos del programa.

15.3 Cuasi experimentos

Los métodos estadísticos y las intuiciones de los experimentos aleatorizados controlados pueden aplicarse a contextos no experimentales a través de los cuasi experimentos, también conocidos como experimentos naturales. En estos, la aleatoriedad es introducida por variaciones en las circunstancias individuales, como las normativas legales, la ubicación, el calendario de políticas, la aleatoriedad natural (por ejemplo, fechas de nacimiento), la lluvia u otros factores no relacionados con el efecto causal bajo estudio.

Tipos de cuasi experimentos

Existen dos tipos de cuasi experimentos:

1. Aquellos en los que el tratamiento se percibe como determinado aleatoriamente. En estos casos, el efecto causal puede ser estimado mediante Mínimos Cuadrados Ordinarios (MCO) utilizando el tratamiento, Xi, como regresor.
2. Aquellos en los que la variación “como si fuera aleatoria” solo determina en parte el tratamiento. En estos casos, el efecto causal se estima mediante regresión de variables instrumentales, donde la variación “como si fuera” aleatoria proporciona la variable instrumental.

El estimador de diferencias en diferencias

El estimador de diferencias en diferencias (DiD) se utiliza en cuasi experimentos donde el tratamiento es “como si” fuera asignado al azar, condicionado a ciertas variables observadas W . Aunque la aleatoriedad no es controlada por el investigador, pueden existir diferencias entre los grupos de tratamiento y control. Para ajustar estas diferencias, en lugar de comparar los resultados Y , se compara la variación en los resultados pre y

post-tratamiento, ajustando por las diferencias en los valores pre-tratamiento de Y entre ambos grupos. Este método se denomina estimador de diferencias en diferencias.

Un ejemplo de aplicación del estimador DiD es el estudio de Card (1990) sobre el efecto de la inmigración en los salarios de trabajadores poco cualificados en Miami. Card comparó la variación en los salarios en Miami con la variación en otras ciudades de EE.UU.

Sea $Y_1^{\text{tratamiento, antes}}$ y $Y_1^{\text{tratamiento, después}}$ las medias muestrales de Y para el grupo de tratamiento antes y después del experimento, respectivamente. Sean $Y_1^{\text{control, antes}}$ y $Y_1^{\text{control, después}}$ las medias correspondientes para el grupo de control. La variación promedio en Y durante el experimento para el grupo de tratamiento es $Y_1^{\text{tratamiento, después}} - Y_1^{\text{tratamiento, antes}}$, y para el grupo de control es $Y_1^{\text{control, después}} - Y_1^{\text{control, antes}}$.

El estimador de diferencias en diferencias es:

$$\hat{\beta}_{\text{DiD}} = \left(Y_1^{\text{tratamiento, después}} - Y_1^{\text{tratamiento, antes}}\right) - \left(Y_1^{\text{control, después}} - Y_1^{\text{control, antes}}\right) 4$$

Este estimador mide la diferencia en la variación promedio en Y entre el grupo de tratamiento y el grupo de control. Si el tratamiento es asignado aleatoriamente, $\hat{\beta}_{\text{DiD}}$ es un estimador insesgado y consistente del efecto causal.

Notación en regresión El estimador de diferencias en diferencias puede expresarse en notación de regresión. Sea ΔY_i la diferencia en Y antes y después del experimento para el individuo i. La regresión para el estimador DiD es:

$$\Delta Y_i = \beta_0 + \beta_1 X_i + u_i$$

Donde β_1 es el estimador de diferencias en diferencias.

Extensión con regresores adicionales El estimador de diferencias en diferencias puede extenderse para incluir regresores adicionales denotados por $W_{1i}, \ldots, W_{ri}$ que capturan características individuales antes del experimento. La regresión múltiple se define como:

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_r W_{ri} + u_i$$

Si X_i es "como si" fuera asignado aleatoriamente, condicionado a $W_{1i}, \ldots, W_{ri}$, entonces $\hat{\beta}_1$ es un estimador insesgado del efecto causal.

Aplicación en datos de panel y sección cruzada repetida

El estimador DiD también se aplica en contextos con datos de panel o de sección cruzada repetida. En el caso de datos de sección cruzada repetida, donde se observan individuos diferentes en cada periodo de tiempo, el modelo de regresión se expresa como:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \beta_4 W_{1it} + \cdots + \beta_r W_{rit} + u_{it}$$

Aquí, X_{it} es el tratamiento, Gi indica si el individuo está en el grupo de tratamiento, y Dt es un indicador binario del periodo. Si el cuasi experimento simula aleatoriedad en X_{it} condicionado a $W_{1it}, \ldots, W_{rit}$, entonces el efecto causal puede estimarse mediante el estimador MCO en esta ecuación.

Si existen más de dos periodos, el modelo se ajusta para incluir variables binarias adicionales que indiquen los diferentes periodos de tiempo.

Estimadores de variables instrumentales

Supongamos que estamos analizando un cuasi-experimento en el cual existe una variable Z_i que influye en la recepción del tratamiento, y se tiene acceso a los datos sobre Z_i y sobre el tratamiento recibido realmente X_i. Si Z_i es como si estuviera asignada aleatoriamente (quizás tras tener en cuenta algunas variables adicionales W_i), entonces Z_i es un instrumento válido para X_i.

Los coeficientes de la ecuación de regresión se pueden estimar mediante mínimos cuadrados en dos etapas (2SLS, por sus siglas en inglés). La ecuación a estimar es:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

Donde: Y_i es la variable dependiente, X_i es el tratamiento, W_i son las varialbes de control, y u_i es el términ de error.

En la primera etapa del 2SLS, se estima X_i como una función de Z_i y W_i:

$$X_i = \pi_0 + \pi_1 Z_i + \pi_2 W_i + v_i$$

Luego, en la segunda etapa, se reemplaza X_i por su predicción ajustada $\hat{X}_i$ en la ecuación original para estimar β_1:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_i + u_i$$

Diseños de discontinuidad en la regresión En algunos cuasi-experimentos, la recepción del tratamiento depende de si una variable observable W_i cruza un valor umbral w_0. Por ejemplo, supongamos que los estudiantes deben asistir a cursos de verano si su GPA cae por debajo de un umbral $w_0 = 2.0$. Una manera de estimar el efecto de los cursos de verano consiste en comparar los resultados de los estudiantes justo por debajo del umbral con los estudiantes justo por encima.

Si la función de regresión es lineal en W_i, con excepción de la discontinuidad inducida por el tratamiento, el efecto del tratamiento se puede estimar mediante la siguiente regresión:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

Este diseño es el conocido como **"brusco"**, en el cual el tratamient X_i es una binaria que dependerá del valor que tome W_i, otro tipo de diseño es el **"difuso"**, en el cual X_i es endógeno y se incluye el instrumento Z_i, para este caso el instrumento será un indicador binario que responderá al valor que asuma W_i, es decir, Z_i es un instrumento que será 0 o 1 si W_i pasa determinado umbral.

$$\text{Primera Etapa:} \quad X_i = \pi_0 + \pi_1 Z_i + \pi_2 W_i + v_i \tag{15.3}$$

$$\text{Segunda Etapa:} \quad Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_i + u_i \tag{15.4}$$

15.4 Problemas potenciales en cuasi experimentos

Los cuasi experimentos enfrentan amenazas a su validez interna y externa, similar a los experimentos aleatorizados, pero con algunas diferencias clave.

Amenazas a la validez interna

- **Ausencia de aleatoriedad:**La asignación "como si fuera" aleatoria puede no ser verdaderamente aleatoria, lo que introduce sesgos en los estimadores.
- **Incumplimiento del protocolo de tratamiento:**Ocurre cuando la asignación aleatoria no determina completamente el tratamiento, afectando la consistencia del estimador de variables in- strumentales.
- **Deserción o abandono:**La deserción puede inducir sesgos de selección muestral, haciendo que los estimadores sean sesgados e inconsistentes.
- **Validez de los instrumentos:**Aun cuando un instrumento esté asignado aleatoriamente, debe evaluarse cuidadosamente su exogeneidad para asegurar que no esté correlacionado con el término de error.

Amenazas a la validez externa

Las amenazas a la validez externa son similares a las de estudios basados en datos de observación. Las particularidades de los eventos que generan la asignación aleatoria pueden limitar la generalización de los resultados a otros contextos.

15.5 Apéndice

13.3: El marco de las variables respuesta para el análisis de datos procedentes de experimentos

En este apéndice se ofrece un tratamiento matemático del marco de análisis de las variables respuesta tratado en la Sección 13.1. El marco de las variables respuesta, en combinación con un efecto del tratamiento constante, implica el modelo de regresión de la Ecuación (15.1). Si la asignación es aleatoria condicionada a las covariables, el marco de las variables respuesta lleva a la Ecuación (15.2) y a la independencia en media condicional. Consideramos un tratamiento binario con $X_i = 1$ que indica la recepción del tratamiento.

Sea $Y_i(1)$ la variable respuesta del individuo i si recibe el tratamiento y sea $Y_i(0)$ la variable respuesta si no recibe tratamiento, por lo que el efecto del tratamiento para el individuo i es $Y_i(1) - Y_i(0)$. Debido a que el individuo es tratado o no lo es, solamente es observable una de los dos posibles resultados posibles o variables respuesta. El resultado observado, Y_i, esta relacionado con las variables respuesta mediante

$$Y_i = Y_i(1)X_i + Y_i(0)(1 - X_i)$$

Si algunos individuos reciben el tratamiento y otros no, la diferencia esperada en los resultados observados entre los dos grupos es $E[Y_i \mid X_i = 1] - E[Y_i \mid X_i = 0] = E[Y_i(1) \mid X_i = 1] - E[Y_i(0) \mid X_i = 0] = E[Y_i(1) - Y_i(0)]$. Esto es cierto sin importar como se determine el tratamiento y dice simplemente que la diferencia esperada es la media del resultado del tratamiento para el tratado, menos la media del resultado de la ausencia de tratamiento para el no tratado. Si además los individuos son asignados aleatoriamente a los grupos de tratamiento y control, entonces X_i se distribuye independientemente de todos los atributos personales y, en particular, es independiente de $[Y_i(1), Y_i(0)]$. Con asignación aleatoria, la diferencia de medias entre los grupos de tratamiento y de control es,

$$E[Y_i|X_i = 1] - E[Y_i|X_i = 0] = E[Y_i(1)|X_i = 1] - E[Y_i(0)|X_i = 0] = E[Y_i(1) - Y_i(0)].$$

donde la segunda igualdad utiliza el hecho de que $[Y_i(1), Y_i(0)]$ son independientes de X_i por asignación aleatoria y la linealidad de las esperanzas (2.2). Por tanto, si X_i se asigna aleatoriamente, la diferencia de medias de los resultados experimentales entre los dos grupos es el efecto promedio del tratamiento en la población de la cual se extrajeron los sujetos.

Capítulo 16

Predicción con muchos Regresores y Big Data

En este capitulo se retoma el ejemplo de la escuela, los maestros y los alumnos. Ahora nos posicionaremos del lado del padre que debia elegir donde vivir y por lo tanto, deseaba predecir cual era el distrito con mayores calificaciones.

Objetivo de la predicción estadística

La predicción estadística implica usar datos para estimar un modelo y aplicarlo a nuevas observaciones fuera de la muestra. El objetivo es lograr una predicción precisa fuera de la muestra, en contraste con el análisis causal, que busca identificar efectos específicos.

16.1 Métodos de predicción

Mínimos Cuadrados Ordinarios (MCO) Cuando se cuenta con pocos predictores, el método de Mínimos Cuadrados Ordinarios (MCO) suele funcionar bien bajo ciertos supuestos (Apéndice 6.4). Sin embargo, en presencia de muchos predictores, como en el ejemplo de las 817 características escolares, MCO tiende a sobreajustar los datos, resultando en predicciones deficientes fuera de la muestra.

Estimadores de Contracción (Shrinkage) Para mejorar la precisión de la predicción en conjuntos de datos con numerosos predictores, se pueden emplear estimadores de contracción. Estos estimadores introducen un sesgo intencionado para reducir la varianza y mejorar la precisión general de la predicción fuera de la muestra. Algunos ejemplos de estos estimadores incluyen el Ridge Regression y el Lasso.

Datos transversales vs. series temporales - Predicción y prnóstico

Este capítulo se enfoca en la predicción usando datos transversales para generalizar a una población mayor. Por otro lado, cuando la predicción se refiere a eventos futuros se denomina pronostico y se usan datos de series temporales, que requieren técnicas y notación adicionales, abordadas en el capitulo anterior.

Grandes Conjuntos de Datos y Big Data La disponibilidad de grandes volúmenes de datos permite el uso de muchos predictores, acercándonos a campos como el aprendizaje automático y la ciencia de datos. En contextos de big data, los métodos tradicionales como MCO son insuficientes, y técnicas avanzadas de contracción y regularización se vuelven esenciales para mejorar el rendimiento predictivo. Para problemas de predicción en econometría, especialmente en presencia de múltiples predictores, es fundamental aplicar técnicas que reduzcan el sobreajuste y optimicen la predicción fuera de la muestra. Esto incluye el uso de estimadores de contracción y enfoques de big data, que han sido esenciales en aplicaciones modernas de predicción estadística.

Grandes conjuntos de datos en econometría

Los conjuntos de datos pueden considerarse grandes no solo por el número de observaciones, sino tam- bién por la cantidad de predictores, o incluso ambos. Los conjuntos de datos puede no ser estándar, conteniendo variables como texto o imágenes, lo cual amplía las aplicaciones posibles.

Predicción con Alta Dimensión Cuando el número de predictores k es grande en comparación con el número de observaciones n, los métodos tradicionales de regresión como MCO suelen ser insuficientes. Este tipo de problema es común cuando se utilizan predictores no lineales o múltiples transformaciones de los predictores originales, generando cientos o miles de regresores. Para abordar estos casos, se emplean **Métodos de Contracción y Regularización:** Técnicas como Ridge Regression y Lasso permiten manejar la alta dimensionalidad reduciendo el sobreajuste y mejorando la capacidad predictiva.

Categorización y Clasificación Otra aplicación importante de los macrodatos es la categorización. En problemas de clasificación, como la aprobación o rechazo de préstamos, los modelos de logit y probit son útiles para predecir la probabilidad de eventos binarios. Estos modelos pueden aplicarse en problemas de toma de decisiones automatizadas:

- En la industria de préstamos, el aprendizaje automático se emplea para clasificar solicitudes de préstamos basándose en la probabilidad de aprobación o rechazo, replicando el proceso de decisión de un oficial de préstamos.

Pruebas de Múltiples Hipótesis El análisis de grandes conjuntos de datos también permite realizar pruebas de múltiples hipótesis, especialmente cuando hay varios tratamientos potenciales. La estadística F tradicional no es adecuada para estos problemas. En su lugar, se emplean técnicas que permiten identificar qué efectos de tratamiento son significativos sin inflar los errores de tipo I.

Manejo de Datos no Estándar Los macrodatos incluyen frecuentemente información en formatos no estándar, como texto o imágenes. Para analizarlos, es necesario convertir estos datos en formatos numéricos. Estas técnicas se utilizan para extraer características clave que luego pueden ser tratadas como predictores en modelos econométricos.

Reconocimiento de Patrones El aprendizaje profundo, una técnica avanzada de aprendizaje au- tomático, permite reconocer patrones complejos en los datos, como en el reconocimiento facial o la traducción de idiomas. Estos modelos no lineales aprovechan grandes volúmenes de datos para detectar estructuras intrínsecas.

Desafíos Computacionales en Big Data

El manejo de grandes volúmenes de datos presenta varios desafíos computacionales, como el almace- namiento y el acceso eficiente a la información. Además, la estimación de modelos complejos requiere algoritmos rápidos y eficientes. Aunque estos aspectos computacionales son fundamentales, no son el foco de este capítulo y se abordan en otras disciplinas, como la informática.

Aplicaciones Cotidianas del Big Data Los resultados del aprendizaje automático aplicados a big data tienen impacto en numerosas áreas de la vida cotidiana. Algunos ejemplos incluyen:

- Asistencia en diagnósticos médicos.
- Publicidad en línea personalizada.
- Algoritmos de reconocimiento facial para la seguridad.
- Estimación de ingresos locales mediante imágenes satelitales.
- Predicción de ventas empresariales basadas en datos detallados de clientes.

En econometría, el análisis de datos no estándar, especialmente los de texto, es cada vez más común y esencial en aplicaciones prácticas.

16.2 El problema de los predictores multiples y MCO(OLS)

Respecto al problema de predecir los resultados de las pruebas escolares utilizando variables que describen tanto las características de la escuela como las de sus estudiantes y su comunidad. El conjunto de datos, obtenido en 2013, contiene información sobre 3932 escuelas primarias en el estado de California. La tarea principal es desarrollar un modelo de predicción capaz de generalizar fuera de la muestra, es decir, proporcionar predicciones precisas para escuelas que no se encuentran en el conjunto de datos de entrenamiento.

Para abordar este problema de predicción fuera de la muestra, se divide el conjunto de datos en dos: la mitad de las observaciones (n = 1966) se utiliza para estimar los modelos de predicción, mientras que la otra mitad se reserva como un conjunto de prueba que permitirá evaluar el rendimiento del modelo más adelante en la Sección 14.6.

El conjunto de datos incluye 817 variables distintas que describen características de las escuelas y sus comunidades. Si se usaran únicamente las variables principales habría 38 regresores.

No Linealidades e Interacciones en los Datos

El análisis de los puntajes de las pruebas escolares en la Sección 8.4 mostró varias no linealidades e interacciones importantes. La inclusión de estas interacciones, cuadrados y cubos aumenta el número de predictores a 817, como se indica en la Tabla 14.1. En la Sección 14.6 se considera un conjunto de datos aún mayor, con 2065 predictores, superando las 1966 observaciones en la muestra de estimación.

Limitaciones del Método de Mínimos Cuadrados Ordinarios

Aunque el método de mínimos cuadrados ordinarios (MCO) es un punto de partida natural, su aplicación puede generar predicciones deficientes cuando el número de predictores es grande en relación con el tamaño de la muestra. Existen, sin embargo, estimadores alternativos al MCO que pueden producir predicciones más confiables en estos casos. Este resultado puede parecer contradictorio con el teorema de Gauss-Markov, el cual establece que el estimador MCO tiene la varianza más baja entre los estimadores insesgados, siempre que se cumplan sus supuesto. La clave de esta aparente contradicción es que estos estimadores alternativos son sesgados, lo cual permite reducir la varianza total del estimador, resultando en predicciones más precisas a pesar del sesgo introducido.

Medida de Precisión de los Modelos de Predicción

Para comparar los modelos de predicción, es necesario utilizar una medida cuantitativa de la precisión de las predicciones. En este contexto, emplearemos el cuadrado del error, específicamente el error de las predicciones fuera de la muestra, como se ha hecho a lo largo de este texto. Utilizar el cuadrado del error de predicción implica que los errores pequeños tienen poco peso, mientras que los errores grandes reciben mucho peso. Esta ponderación resulta adecuada en muchos problemas de predicción, donde los errores pequeños no tienen un impacto significativo, pero los errores grandes pueden disminuir la utilidad y la credibilidad de las predicciones.

El **error cuadrático medio de predicción** (MSPE, por sus siglas en inglés) es el valor esperado del cuadrado del error de predicción cuando se emplea el modelo para hacer predicciones para observaciones que no están en el conjunto de datos de entrenamiento. Matemáticamente, el MSPE se define como:

$$\text{MSPE} = E[(Y_{\text{oos}} - \hat{Y}(X_{\text{oos}}))^2] \tag{16.1}$$

Donde X_{oos} y Y_{oos} representan observaciones fuera de la muestra (“oos”), y $\hat{Y}(X_{\text{oos}})$ es el valor predicho de Y para un valor dado de los predictores X.

Es importante señalar que la observación fuera de la muestra no se utiliza en la estimación del modelo de predicción. Desde la perspectiva de minimizar el MSPE, la mejor predicción posible es la media condicional, es decir, $E[Y_{\text{oos}} \mid X_{\text{oos}}]$. Esta predicción ideal se conoce como la predicción de oráculo.

Predicción de Oráculo y Fuentes de Error en el MSPE

Aunque la media condicional es desconocida, la predicción de oráculo sirve como un parámetro ideal con el que se pueden comparar todas las predicciones factibles. En el modelo de regresión, la predicción de oráculo equivale a la predicción que se obtendría utilizando los verdaderos coeficientes de regresión de la población, los cuales son desconocidos en la práctica.

El MSPE incorpora dos fuentes de error en las predicciones. En primer lugar, incluso si se conociera la media condicional, la predicción no sería perfecta: la predicción de oráculo conlleva el error de predicción $Y_{\text{oos}} - E[Y_{\text{oos}} \mid X_{\text{oos}}]$. En segundo lugar, debido a que $E[Y_{\text{oos}} \mid X_{\text{oos}}]$ es desconocido, la estimación de los parámetros (es decir, la estimación de los coeficientes del modelo de predicción $\hat{Y}(X)$) introduce una fuente adicional de error.

Primera Suposición de Mínimos Cuadrados para la Predicción

La primera suposición de mínimos cuadrados para la predicción establece que la observación fuera de la muestra proviene de la misma distribución que las observaciones dentro de la muestra utilizadas para estimar el modelo:

Primera suposición de mínimos cuadrados para la predicción: Las observaciones fuera de la muestra $(X_{\text{oos}}, Y_{\text{oos}})$ son tomadas aleatoriamente de la misma distribución poblacional que la muestra de estimación (X_i, Y_i), $i = 1, \ldots, n$.

Dado que las observaciones dentro y fuera de la muestra provienen de la misma distribución, la media condicional, $E[Y \mid X]$, es la predicción de oráculo tanto para las observaciones dentro de la muestra como para las fuera de ella. Esta suposición es una declaración sobre la validez externa: el modelo estimado en la muestra puede generalizarse a la observación fuera de la muestra de interés.

El Modelo de Regresión Predictiva con Regresores Estandarizados

Este capítulo emplea una versión modificada del modelo de regresión lineal en el que todos los regresores están estandarizados;

Sea $(X_{1i}^*, \ldots, X_{ki}^*, Y_i^*)$, $i = 1, \ldots, n$, el conjunto de datos tal como se recolectó originalmente, donde X_{ji}^* representa la i-ésima observación sobre el j-ésimo regresor original. Los regresores estandarizados se definen como $X_{ji} = \frac{X_{ji}^* - \mu_{X_j^*}}{\sigma_{X_j^*}}$, donde $\mu_{X_j^*}$ y $\sigma_{X_j^*}$ son, respectivamente, la media y la desviación estándar poblacional de X_j^* . La variable dependiente transformada (centrada) es $Y_i = Y_i^* - \mu_{Y^*}$, donde μ_{Y^*} es lamedia poblacional de Y^*.

Con esta notación, el modelo de regresión predictiva estandarizado es la regresión de Y , con media 0, sobre los k regresores X estandarizados:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{16.2}$$

El intercepto se excluye porque todas las variables tienen media 0. Debido a que los regresores están estandarizados, los coeficientes de regresión tienen las mismas unidades: β_j representa la diferencia en el valor predicho de Y asociada a una diferencia de una desviación estándar en X_j^* , manteniendo constantes los otros X.

Dado que el enfoque de este capítulo es la predicción, adoptamos a lo largo del mismo la interpretación predictiva del modelo de regresión en el Apéndice 6.4; es decir, $E[Y \mid X] = \sum_{j=1}^{k} \beta_j X_j$ y $E[u \mid X] = 0$.

Como es habitual, la estructura lineal en la Ecuación (16.2) implica que las predicciones son lineales en los coeficientes; sin embargo, la función de regresión puede ser no lineal en los predictores, ya que X puede incluir términos no lineales, como cuadrados o interacciones.

El MSPE en el Modelo de Regresión Predictiva Estandarizado

En el modelo de regresión estandarizado de la Ecuación (14.2), la predicción para el valor fuera de la muestra de los predictores es:

$$\hat{Y}(X_{\text{oos}}) = \hat{\beta}_1 X_{\text{oos},1} + \cdots + \hat{\beta}_k X_{\text{oos},k}.$$

El error de predicción es:

$$Y_{\text{cos}} - (\hat{\beta}_1 X_{\text{cos},1} + \cdots + \hat{\beta}_k X_{\text{cos},k}) = u_{\text{cos}} - [(\hat{\beta}_1 - \beta_1) X_{\text{cos},1} + \cdots + (\hat{\beta}_k - \beta_k) X_{\text{cos},k}] \tag{16.3}$$

Donde la expresión final se obtiene usando la Ecuación (16.2) y uoos es el valor del error u para la observación fuera de la muestra. Debido a que uoos es independiente de los datos usados para estimar los coeficientes y no está correlacionado con Xoos, el MSPE en la Ecuación (16.1) para el modelo de regresión predictiva estandarizado puede escribirse como la suma de dos componentes:

$$\text{MSPE} = \sigma_u^2 + E\left[\sum_{j=1}^{k} (\hat{\beta}_j - \beta_j) X_{\text{oos},j}\right]^2$$

El primer término es la varianza del error de predicción del oráculo, es decir, el error de predicción realizado usando la media condicional verdadera (desconocida) $E[Y \mid X]$.

El segundo término es la contribución al error de predicción que surge de los coeficientes de regresión estimados. Representa el costo, medido en términos de incremento del error cuadrático medio de predicción, de tener que estimar los coeficientes en lugar de usar la predicción del oráculo.

Como el error cuadrático medio es la suma de la varianza y el cuadrado del sesgo, el segundo término es la suma de la varianza de la predicción derivada de la estimación de β y el sesgo cuadrado de la predicción. Al elegir un estimador, el objetivo es hacer este segundo término tan pequeño como sea posible.

Estandarización usando las Medias y Varianzas de la Muestra

En la práctica, las medias y desviaciones estándar poblacionales de las variables originales no son cono- cidas. Por lo tanto, se utilizan las medias y varianzas de la muestra para estandarizar los regresores, y se resta la media de la muestra de la variable dependiente.

Dado que los regresores están estandarizados y la variable dependiente está centrada, se requiere un paso adicional para producir la predicción para una observación fuera de la muestra. Específicamente, la observación fuera de la muestra de los predictores debe estandarizarse usando la media y la desviación estándar de la muestra, y la media de la variable dependiente de la muestra debe sumarse nuevamente a la predicción.

El MSPE de MCO y el Principio de Reducción (Shrinkage)

En el caso especial en que el error de regresión u en la Ecuación (16.2) es homocedástico, el MSPE de Mínimos Cuadrados Ordinarios (MCO) se aproxima por:

$$\text{MSPE}_{\text{MCO}} \approx \left(1 + \frac{k}{n}\right)\sigma_u^2$$

Esta expresión tiene una interpretación sencilla. Como se discutió en la Ecuación (16.3), el MSPE de la predicción del oráculo (que usa el valor verdadero de β) es σ_u^2. Cuando los k coeficientes de regresión se estiman por MCO, el MSPE aumenta en un factor de $1 + k/n$ en relación con el MSPE óptimo. Así, el costo de usar MCO, medido en términos de MSPE, depende de la proporción de regresores con respecto al tamaño de la muestra.

Dado que MCO es insesgado bajo la interpretación predictiva de la Ecuación (16.2), el factor de inflación $1 + k/n$ surge únicamente de la varianza del estimador MCO. Según las condiciones de Gauss-Markov, el estimador MCO tiene la varianza más baja entre los estimadores lineales insesgados. Esto podría desalentar, a primera vista, encontrar mejoras cuando k/n es grande. Sin embargo, un avance conceptual clave de los años 60 fue descubrir que si se permite un sesgo en los estimadores, la varianza del estimador puede reducirse tanto que el MSPE puede ser menor que el de MCO.

El Principio de Reducción (Shrinkage)

Un estimador de reducción introduce sesgo al "reducir" el estimador MCO hacia un valor específico, reduciendo así su varianza. Como el error cuadrático medio es la suma de la varianza y el sesgo cuadrado, si la varianza del estimador se reduce lo suficiente, esta disminución puede compensar el aumento en el sesgo cuadrado, resultando en un estimador con un error cuadrático medio menor que MCO.

James y Stein (1961) desarrollaron el primer estimador que lograba reducir el error cuadrático medio del estimador al introducir sesgo. Cuando los regresores son no correlacionados, el estimador de James- Stein puede expresarse como $\hat{\beta}^{JS} = c\hat{\beta}$, donde $\hat{\beta}$ es el estimador MCO y c es un factor menor que 1 que depende de los datos. Dado que c es menor que 1, el estimador de James-Stein reduce el estimador MCO hacia 0, introduciendo así un sesgo hacia 0. Este estimador muestra un error cuadrático medio menor que el de MCO cuando los verdaderos valores de β son pequeños. Sin embargo, James y Stein demostraron que, si los errores son normalmente distribuidos, su estimador tiene un error cuadrático medio menor que el de MCO, independientemente del valor verdadero de β, siempre que $k \geq 3$.

El notable resultado de James y Stein es la base de muchos métodos de predicción para grandes cantidades de predictores en el contexto de big data, que incluye la regresión ridge y el estimador Lasso.

Estimación del MSPE

El MSPE es una expectativa poblacional, por lo que es desconocido en la práctica. Sin embargo, se puede estimar a partir de una muestra de datos. Existen dos métodos para estimar el MSPE: el primero es la estimación con muestras divididas (split-sample), que utiliza la definición directa del MSPE; el segundo es la validación cruzada m-fold, que extiende esta idea usando los datos de forma simétrica y más eficiente al dividir la muestra en m submuestras.

Estimación del MSPE con Muestra Dividida (Split-Sample)

Recuerda que el MSPE es la varianza del error de predicción para un valor de X elegido aleatoriamente, en el que la observación no se utiliza para estimar β. Esta definición sugiere estimar el MSPE dividiendo el conjunto de datos en dos partes: una submuestra para estimación y una submuestra de "prueba" para simular la predicción fuera de la muestra. La primera se usa para estimar β, generando el estimador $\tilde{\beta}$, que podría obtenerse por MCO u otro método. Luego, este estimador se usa para hacer una predicción Y_n para cada una de las ntest observaciones en la submuestra de prueba. El MSPE se estima con los errores de predicción resultantes:

$$\text{MSPE}_{\text{split-sample}} = \frac{1}{n_{\text{test}}} \sum \left(Y_i - Y_n^i \right)^2 \tag{16.4}$$

Estimación del MSPE por Validación Cruzada m-Fold

El procedimiento de muestra dividida trata los datos de manera asimétrica al dividir las observaciones en dos submuestras que se usan con diferentes propósitos. Esta estimación puede mejorarse tratando los datos de manera simétrica. Específicamente, las dos submuestras pueden intercambiarse para producir dos estimadores diferentes del MSPE al cambiar cuál submuestra se usa para estimar β y cuál para estimar el MSPE.

Esta idea se extiende a m submuestras diferentes seleccionadas al azar. El procedimiento resultante se llama **validación cruzada m-fold**. En la validación cruzada m-fold, se generan m estimaciones separadas del MSPE, cada una calculada dejando fuera secuencialmente una de las m submuestras cuando se estima β y usando esa submuestra reservada para estimar el MSPE. El estimador del MSPE en la validación cruzada m-fold es el promedio de los m estimadores de submuestras del MSPE, tal como se resume en el Concepto Clave 14.1. (Recordar que los datos son de corte transversal, no son series de tiempo, por lo que su orden en la muestra no es relevante).

Un aspecto a definir en la validación cruzada m-fold es cómo elegir m. Un valor más grande de m produce estimaciones más eficientes de β porque se usan más observaciones cada vez que se estima β. Sin embargo, un valor mayor de m implica que β debe estimarse m veces, lo que puede ser computacionalmente costoso,

especialmente cuando k es grande, lo cual requiere considerable tiempo de cálculo. En consecuencia, la elección de m debe considerar las limitaciones prácticas del tiempo disponible y la capacidad computacional.

El estimador de validación cruzada m-fold puede utilizarse para estimar el MSPE en contextos muy generales, independientemente de cómo se estime β. Incluso funciona para modelos que solo pueden expresarse como algoritmos y no en términos de parámetros, lo que lo hace ampliamente utilizado en trabajos empíricos con big data.

Concepto clave 14.1: Validación cruzada m-fold

El estimador del MSPE de validación cruzada m-fold se determina de acuerdo a los siguientes 6 pasos:

1. Dividir la muestra de prueba en m submuestras elegidas aleatoriamente de igual tamaño.
2. Usar las sumbmuestras $2,\ldots,m$ para obtener $\tilde{\beta}$, una estimación de β.
3. Usar $\tilde{\beta}$ y la ecuación (16.2) para obtener los valores predecidos de $\hat{Y}$ y los errores de predicción $Y-\hat{Y}$ para las observaciones en la submuestra 1.
4. Usar la submuestra 1 como la submuestra de entrenamiento, estimar el MSPE con los valores predecidos de la submuestra 1 y la ecuación (16.4). Lo llamaremos $\hat{MSPE}_1$.
5. Repetir los pasos 2 a 4 usando la submuestra 2 como la submuestra apartada para entrenamiento, luego la 3, y así sucesivamente hasta obtener m estimaciones de MSPE.
6. Luego el estimador del MSPE de validación cruzada m-fold es el promedio de los m estimadores de MSPE:

$$\widehat{\text{MSPE}}_{\text{m-fold cross validation}} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{n_i}{n/m}\right)\widehat{\text{MSPE}}_i,$$

Donde n_i es el tamaño de observaciones en la submuestra i y el factor en el paréntesis permite diferentes números de observaciones en las diferentes submuestras.

16.3 Regresión Ridge

Las Secciones anteriores presentan dos estimadores de shrinkage diseñados para manejar situaciones con muchos predictores. En esta sección se discute la regresión ridge, una técnica que reduce los parámetros estimados hacia 0 mediante la adición de una penalización en la suma de residuos al cuadrado, la cual aumenta con el cuadrado de los parámetros estimados. Al minimizar la suma de estos dos términos, llamada suma de residuos penalizada, la regresión ridge introduce sesgo en el estimador, pero reduce su varianza. En algunas aplicaciones, la regresión ridge puede mejorar notablemente el MSPE en comparación con el MCO.

Shrinkage mediante Penalización y Regresión Ridge

Una forma de reducir los coeficientes estimados hacia 0 es penalizar los valores grandes del estimador. El estimador de regresión ridge se basa en esta idea, minimizando la suma de residuos penalizada, que es la suma de los residuos al cuadrado más un factor de penalización que aumenta con la suma de los coeficientes al cuadrado:

$$S_{\text{Ridge}}(\beta;\lambda_{\text{Ridge}}) = \sum_{i=1}^{n}\left(Y_i-\beta_1X_{1i}-\cdots-\beta_kX_{ki}\right)^2+\lambda_{\text{Ridge}}\sum_{j=1}^{k}\beta_j^2 \tag{16.5}$$

Donde $\lambda_{\text{Ridge}}\geq 0$. Este parámetro λ_{Ridge} es el parámetro de shrinkage de ridge, y el estimador de regresión ridge es el valor de β que minimiza $S_{\text{Ridge}}(\beta;\lambda_{\text{Ridge}})$.

El primer término del lado derecho de la ecuación es la suma habitual de residuos al cuadrado para un valor dado de β. Si solo estuviera este término, los estimadores ridge y MCO serían iguales. Sin embargo, el segundo término, que es nuevo, aumenta con la suma de los coeficientes al cuadrado y penaliza al estimador por elegir un coeficiente grande. Este término de penalización, cuando se escala por el parámetro de shrinkage y se agrega a la suma de residuos al cuadrado, se llama suma de residuos penalizada.

El término de penalización hace que el estimador de regresión ridge se acerque a 0. Sin la penalización, se minimiza la suma de residuos al cuadrado, lo que produce el estimador MCO. Al agregar la penalización, el

mínimo de la función penalizada se desplaza hacia 0, haciendo que el coeficiente estimado de ridge esté más cerca de 0 en comparación con el estimador MCO; es decir, el estimador de regresión ridge se "encoge" hacia 0.

La magnitud del shrinkage depende del parámetro λ_{Ridge}. Si $\lambda_{\text{Ridge}} = 0$, no hay shrinkage y el estimador de regresión ridge es igual al estimador MCO. Cuanto mayor sea λ_{Ridge}, mayor será la penalización para un valor dado de β y mayor será el shrinkage del estimador hacia 0. Dado que estamos usando el modelo de regresión predictiva estandarizado, todos los coeficientes tienen las mismas unidades, por lo que un solo parámetro de shrinkage λ_{Ridge} se puede usar para todos los coeficientes.

Fórmula para el estimador de regresión Ridge

La suma de residuos penalizada en la ecuación se puede minimizar usando cálculo para obtener una expresión simple del estimador de regresión ridge.

En el caso especial en el que los regresores no están correlacionados, el estimador de regresión ridge es:

$$\hat{\beta}_{\text{Ridge},j} = \frac{1}{1 + \frac{\lambda_{\text{Ridge}}}{\sum_{i=1}^{n} X_{ji}^2}} \hat{\beta}_j$$

Donde $\hat{\beta}_j$ es el estimador MCO de β_j . En este caso, el estimador de regresión ridge encoge el estimador MCO hacia 0, de manera similar al estimador de James-Stein. Cuando los regresores están correlacionados, el estimado ridge a veces puede ser mayor que el estimado MCO, aunque en general los estimadores de regresión ridge están reducidos hacia 0.

Cuando hay multicolinealidad perfecta, como cuando k >n, el estimador MCO no puede calcularse, pero el estimador ridge sí.

Estimación del Parámetro de Shrinkage Ridge mediante Validación Cruzada

El estimador de regresión ridge depende del parámetro de shrinkage λ_{Ridge}. Aunque podría elegirse un valor arbitrario para λ_{Ridge}, una estrategia más eficaz es elegir un valor que permita al estimador de regresión ridge ajustarse bien a los datos. En este caso, minimizar $S_{\text{Ridge}}(\beta; \lambda_{\text{Ridge}})$ en la Ecuación (16.5) para un valor de prueba de β llevaría a establecer $\lambda_{\text{Ridge}} = 0$. Sin embargo, cuando $\lambda_{\text{Ridge}} = 0$, el estimador de regresión ridge coincide con el estimador MCO, el cual proporciona el mejor ajuste dentro de la muestra. El objetivo de la predicción, en cambio, es obtener un buen ajuste fuera de la muestra (esto es, minimizar el MSPE).

Esta idea sugiere que λ_{Ridge} debería seleccionarse minimizando el MSPE estimado, lo cual puede implementarse usando el estimador del MSPE mediante m-fold validación cruzada. Supongamos que tenemos dos valores candidatos para λ_{Ridge}, seleccionamos el valor que proporcione el MSPE estimado más bajo.

Repitiendo estos pasos para varios valores de λ_{Ridge}, obtenemos un estimador de λ_{Ridge} que minimiza el MSPE de validación cruzada de m-fold. Aunque este estimador podría ser 0, lo que haría que el mejor estimador de ridge fuera el estimador MCO, típicamente el mejor parámetro de shrinkage no será 0, y el estimador de ridge diferirá del estimador MCO.

Dado que $\hat{\lambda}_{\text{Ridge}}$ se elige para minimizar el MSPE validado cruzadamente, el MSPE en $\hat{\lambda}_{\text{Ridge}}$ ya no es un estimador insesgado del MSPE.

16.4 Regresión lasso

En los métodos de MCO y regresión ridge, ninguno de los coeficientes estimados es exactamente 0, por lo que todos los regresores se utilizan para hacer la predicción. Sin embargo, en algunas aplicaciones, solo unos pocos predictores pueden ser útiles, mientras que el resto son irrelevantes.

Un modelo de regresión en el cual solo una pequeña fracción de los predictores tiene coeficientes distintos de cero se denomina **modelo sparse**. Si el modelo es sparse, las predicciones pueden mejorarse estimando muchos de los coeficientes exactamente en 0.

El estimador analizado en esta sección, el Lasso (*Least Absolute Shrinkage and Selection Operator*), está diseñado para modelos sparse. Al igual que la regresión ridge, el Lasso reduce los coeficientes estimados a 0. Sin embargo,

a diferencia de la regresión ridge, el Lasso establece muchos de los coeficientes estimados exactamente en 0, eliminando esos regresores del modelo. Además, los regresores que se mantienen están sujetos a una menor reducción que en la regresión ridge. Así, el Lasso ofrece una forma de seleccionar un subconjunto de los regresores y luego estimar sus coeficientes con un nivel moderado de shrinkage.

Al igual que la regresión ridge, el Lasso puede usarse cuando k >n. También como la regresión ridge, el Lasso cuenta con un parámetro de shrinkage que puede estimarse minimizando el MSPE validado cruzadamente.

Shrinkage usando el Lasso

El estimador del Lasso minimiza una suma de cuadrados penalizada, donde la penalización aumenta con la suma de los valores absolutos de los coeficientes:

$$S_{\text{Lasso}}(\beta;\lambda_{\text{Lasso}}) = \sum_{i=1}^{n}(Y_i - \beta_1 X_{1i} - \cdots - \beta_k X_{ki})^2 + \lambda_{\text{Lasso}} \sum_{j=1}^{k} |\beta_j| \tag{16.6}$$

Donde λ_{Lasso} es el parámetro de shrinkage del Lasso. El Lasso estima β minimizando la fórmula antes obtenida, $S_{\text{Lasso}}(\beta;\lambda_{\text{Lasso}})$. Similar a la regresión ridge, si $\lambda_{\text{Lasso}} = 0$, el Lasso se reduce al estimador MCO. El segundo término en la ecuación penaliza los valores grandes de β, llevando las estimaciones hacia cero.

La primera parte del nombre Lasso—least absolute shrinkage—refleja la naturaleza de la penalización, la cual depende de los valores absolutos de los coeficientes, en contraste con la penalización cuadrática en la regresión ridge. La segunda parte del nombre—selection operator —se debe a que el Lasso tiende a estimar algunos coeficientes exactamente en cero, eliminando así predictores del modelo.

Cuando el valor de β en el estimador MCO es grande, la penalización de la regresión ridge es mayor que la del Lasso. Por lo tanto, el Lasso reduce menos que ridge en estos casos, pero para valores pequeños de MCO, el Lasso aplica una reducción mayor, incluso llevando el estimador a cero. Si hay múltiples predictores, el Lasso usualmente reduce las estimaciones hacia cero, pero en algunos casos podría estimar algunos coeficientes mayores que en MCO.

Cálculo del estimador del Lasso

A diferencia de MCO y la regresión ridge, no existe una expresión sencilla para el estimador del Lasso cuando k >1, por lo que su minimización debe realizarse numéricamente. Actualmente, existen algoritmos especializados en software econométrico que facilitan el uso del Lasso.

Estimación del parámetro de shrinkage mediante validación cruzada

Al igual que en la regresión ridge, el parámetro de ajuste del Lasso puede estimarse minimizando una estimación del MSPE mediante validación cruzada, siguiendo el mismo procedimiento que en ridge.

Advertencia sobre los estimadores ridge y Lasso

Los estimadores ridge y Lasso difieren de otros estimadores como MCO en que el ajuste y las predicciones dependen de la combinación lineal específica de regresores utilizada. Para Lasso, los valores de los coeficientes poblacionales cambian al modificar las combinaciones lineales de los regresores. Por ejemplo, en una especificación con intercepto y una variable dummy de género (hombre o mujer), el Lasso podría eliminar el predictor de hombre en una combinación pero no en otra, lo que genera predicciones distintas.

En el caso de ridge, la dependencia de la combinación lineal se debe a que diferentes combinaciones pueden tener distintas correlaciones entre sí.

16.5 componentes principales

Cuando los regresores son perfectamente colineales, al menos uno de ellos se puede eliminar del conjunto de datos sin perder información, ya que el regresor eliminado se puede reconstruir perfectamente a partir de

los restantes. Esto sugiere que puede haber poca pérdida de información al eliminar una variable que esté altamente, pero no perfectamente, correlacionada con otros regresores. Este es el fundamento del análisis de componentes principales, que reduce el número de regresores reteniendo la mayor cantidad de información original posible. La reducción de los regresores permite la estimación y predicción mediante MCO.

Componentes Principales con Dos Variables

os componentes principales de un conjunto de variables estandarizadas X son combinaciones lineales de estas variables, construidas de manera que sean mutuamente incorrelacionados y capturen la mayor cantidad posible de información original. Específicamente, los pesos de la primera combinación lineal (primer componente principal) se eligen para maximizar su varianza, de modo que explique la mayor parte de la variabilidad de X. El segundo componente principal se construye para ser incorrelacionado con el primero y maximizar la varianza remanente, y así sucesivamente.

Por ejemplo, si consideramos dos variables X_1 y X_2, ambas distribuidas como normales estándar con una correlación de r = 0.7, el primer componente principal se expresa como una combinación lineal ponderada:

$$\text{PC}_1 = w_1 X_1 + w_2 X_2,$$

donde los pesos w_1 y w_2 maximizan la varianza de PC_1. En este caso, la dirección de máxima dispersión se encuentra a lo largo de la diagonal de 45º, lo que implica que $w_1 = w_2 = 1/\sqrt{2}$, resultando en:

$$\text{PC}_1 = \frac{X_1 + X_2}{\sqrt{2}}.$$

El segundo componente principal, ortogonal al primero, minimiza la varianza en su dirección, quedando definido como:

$$\text{PC}_2 = \frac{X_1 - X_2}{\sqrt{2}}$$

correspondiendo a la diagonal de 45º descendente. Las varianzas de estos componentes principales están dadas por:

$$\text{Var}(\text{PC}_1) = 1 + |r|, \quad \text{Var}(\text{PC}_2) = 1 - |r|.$$

Esto confirma que, cuando las variables están correlacionadas, PC_1 captura más varianza que PC_2. Además, la suma de las varianzas de PC_1 y PC_2 es igual a la suma de las varianzas de X_1 y X_2, proporcionando una interpretación del R^2 en componentes principales: la fracción de la varianza total explicada por PC_1 es:

$$\frac{\text{Var}(\text{PC}_1)}{\text{Var}(X_1) + \text{Var}(X_2)},$$

mientras que la fracción explicada por PC_2 es la complementaria. En el caso de $r = 0.7$, el primer componente principal explica aproximadamente el 85 % de la varianza total, mientras que el segundo explica el 15 % restante.

Aunque para solo dos variables el uso de componentes principales no es necesario, este método resulta útil cuando se tienen muchas variables correlacionadas, ya que permite reducir la dimensionalidad capturando gran parte de la variabilidad en un menor número de componentes.

Componentes Principales con k Variables

Los componentes principales de k variables $X_1, \ldots, X_k$ son combinaciones lineales de estas variables que son mutuamente incorrelacionadas, tienen pesos al cuadrado que suman 1, y maximizan la varianza de la combinación lineal controlando por los componentes principales previos. Suponiendo que no hay multicolinealidad perfecta entre las variables, el número de componentes principales de X es el mínimo entre n y k.

Concepto clave 14.2: El componente principal de X

Los componentes principales de las k variables $X_1, \ldots, X_k$ son combinaciones lineales de X que tienen las siguiente propiedades:

1. Los cuadrados ponderados de las combinaicones lineales suman 1.
2. El primer componente principal maximiza la varianza de su combinación lineal.
3. El segundo componente principal maximiza la varianza de su combinaicón lineal, sujeta a estar incorrelacionado con el primero; y así sucesivamente.
4. De forma general, el j-esimo componente principal maximiza la varianza de su combinación lineal, sujeta a estar incorrelacionado con los componentes principales anteriores.

- Asumiento que no hay multicolinealidad perfecta en X, el número de componentes principales es el mínimo entre n y k.
- La suma de las varianzas muestrales de los componentes principales iguala la suma de las varianzas muestrales de $X's$.

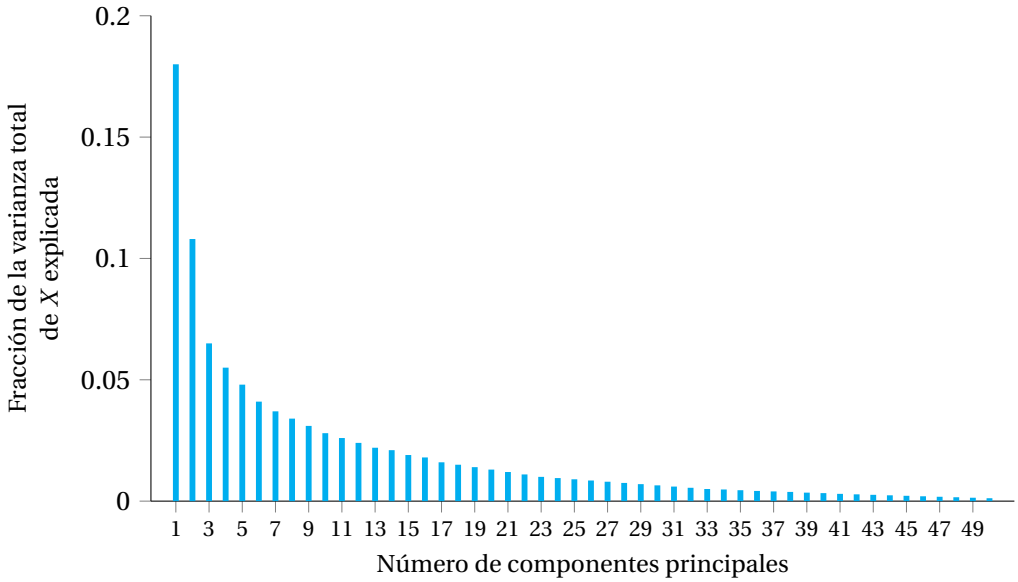
$$\sum_{j=1}^{\text{mín}(n,k)} \text{var}(PC_j) = \sum_{j=1}^{k} \text{var}(X_j).$$

- El ratio $var(PC_j)/\sum_{j=1}^{k} var(X_j)$ es la fracción de la varianza total muestral de $X's$ explicada por el j-ésimo componente principal. Esta medida is como un R^2 para la varianza total de $X's$.

El scree plot

La igualdad en la Ecuación (14.10) permite construir un gráfico útil, el scree plot, que visualiza la cantidad de variación en X capturada por cada componente principal. Un scree plot es un gráfico de la varianza muestral del j-ésimo componente principal en relación con la varianza total muestral de X, es decir, la relación $var(PC_j)/\sum_{j=1}^{k} var(X_j)$ frente al número del componente principal j.

FIGURE 14.6 Scree Plot para las 817-Variables del set de datos de School (Primeros 50 componentes principales)



Los valores gráficados son la fracción total de la varianza de los 817 regresores explicados por el componente principal indicado. El primer componente principal explica el 18 % de la varianza total de las 817 X's, y los primeros 10 componentes principales juntos explican el 63 % de la varianza total.

Esta relación tiene la interpretación de un R^2 para el j – ésimo componente principal, permitiendo ver la fracción de la varianza muestral explicada por cada componente. Dado que los componentes principales son

incorrelacionados, la suma acumulativa de estas razones hasta el componente p-ésimo representa la fracción de la varianza total explicada por los primeros p componentes.

Este comportamiento, donde los primeros componentes explican gran parte de la varianza, es común en datos con variables altamente correlacionadas, y da al scree plot su nombre, asemejando una pendiente rocosa.

Predicción Usando Componentes Principales

Un aspecto clave es determinar cuántos componentes principales p incluir en la regresión. Al igual que con los parámetros de ajuste de ridge y Lasso, el número de componentes p se puede estimar minimizando el MSPE, calculado mediante validación cruzada m-fold.

Para observaciones fuera de muestra, es necesario estandarizar cada predictor usando su media y varianza de la muestra. En regresión de componentes principales, los valores fuera de muestra de los componentes deben calcularse aplicando los pesos estimados w a los X estandarizados de la muestra.

El rendimiento predictivo se evalúa fuera de la muestra, utilizando un MSPE estimado a partir de validación cruzada m-fold, reservando la mitad de las observaciones para el conjunto de prueba reservado.

La conclusión más importante es que los métodos de muchos predictores tienen éxito donde OLS (MCO) falla. Esto se debe a que los métodos de muchos predictores permiten que las estimaciones de coeficientes estén sesgadas de una manera que reduce su varianza lo suficiente como para compensar el aumento del sesgo. Además, el MSPE estimado mediante validación cruzada se acerca al MSPE calculado utilizando el conjunto de prueba reservado.

Capítulo 17

Introducción a la Regresión de Series Temporales y Predicción

17.1 Utilización de los modelos de regresión para predicción

La regresión lineal es una herramienta fundamental en econometría, no solo para estimar efectos causales, sino también para realizar predicciones. Un ejemplo básico es el modelo que relaciona las calificaciones en exámenes con la ratio estudiantes-maestros (REM):

$$\text{CalificaciónExamen} = 989.9 - 2.28 \times \text{REM}$$

Aunque este modelo tiene limitaciones para estimar efectos causales, debido a la posible omisión de variables como características de la escuela y los estudiantes, puede ser útil en contextos de predicción. Por ejemplo, un padre que busca mudarse a un nuevo distrito escolar puede usar esta regresión para predecir calificaciones futuras en un área donde no se dispone de datos públicos.

En el ámbito de series temporales, la predicción se enfoca en eventos futuros utilizando datos pasados. Un modelo autorregresivo (AR), que se presenta en la Sección 14.3, utiliza valores pasados de una variable para predecir sus valores futuros. Este enfoque se puede extender con variables adicionales, como el tamaño de las clases, mejorando la precisión de las predicciones.

17.2 Introducción a los datos de series temporales y correlación serial

El análisis de series temporales se basa en la observación de datos a lo largo del tiempo. Por ejemplo, las tasas de inflación y desempleo en EE.UU. desde 1960 hasta 2004 muestran fluctuaciones significativas, que pueden analizarse mediante gráficos y modelos econométricos.

Retardos, Primeras Diferencias, Logaritmos y Tasas de Crecimiento

Dado un conjunto de datos de series temporales, la observación en el periodo t se denota como Y_t. El valor de Y en el periodo anterior $(t-1)$ se denomina primer retardo, Y_{t-1}, y el valor en el periodo $t-j$ es el j-ésimo retardo, Y_{t-j} . La variación entre los periodos $t-1$ y t, conocida como primera diferencia, se expresa como:

$$\Delta Y_t = Y_t - Y_{t-1}$$

Para series económicas, es común transformar los datos tomando logaritmos para analizar tasas de crecimiento exponenciales o variaciones proporcionales. La primera diferencia del logaritmo de Y_t es:

$$\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$$

La variación porcentual aproximada de Y_t entre los periodos $t-1$ y t es entonces:

$$\text{Variacion } \% \approx 100 \times \Delta \ln(Y_t)$$

Concepto clave 14.1: Retardos, primeras diferencias, logaritmos y tasas de crecimiento

- El primer retardo de una serie temporal Y_t es Y_{t-1} ; su j-ésimo retardo es Y_{t-j}.
- La primera diferencia de una serie, ΔY_t, es su variación entre los periodos $t-1$ y t; es decir, $\Delta Y_t = Y_t - Y_{t-1}$.
- La primera diferencia del logaritmo de Y_t es $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- La variación porcentual de una serie temporal Y_t entre los periodos $t-1$ y t es aproximadamente $100\Delta \ln(Y_t)$, siendo la aproximación más precisa cuando la variación porcentual es pequeña.

Autocorrelación

En series temporales, el valor de Y en un periodo suele estar correlacionado con su valor en periodos anteriores. Esta correlación se denomina autocorrelación. La autocorrelación de primer orden es la correlación entre Y_t y Y_{t-1}, y se denota como ρ_1. La j-ésima autocorrelación es la correlación entre Y_t y Y_{t-j} , denotada como ρ_j.
Las autocorrelaciones y autocovarianzas muestrales se calculan como:

$$\text{Autocovarianza muestral:} \quad \hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \bar{Y})(Y_{t-j} - \bar{Y}) \tag{17.1}$$

$$\text{Autocorrelación muestral:} \quad \hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0} \tag{17.2}$$

Estas medidas permiten identificar patrones de dependencia temporal en los datos, lo cual es crucial para construir modelos predictivos robustos.

Concepto clave 14.2: Autocorrelación (correlación serial) y autocovarianza

La j-ésima autocovarianza de una serie Y_t es la covarianza entre Y_t y su j-ésimo retardo Y_{t-j}, y el coeficiente de correlación j-ésimo es la correlación entre Y_t e Y_{t-j}. Es decir,

$$\text{j-ésima autocovarianza:} \quad = cov(Y_t, Y_{t-j}) \tag{17.3}$$

$$\text{j-ésima autocorrelación:} \quad = \rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}} \tag{17.4}$$

El coeficiente de autocorrelación j-ésimo a veces se denomina coeficiente de correlación serial j-ésimo

17.3 Modelos Autorregresivos (AR)

El Modelo Autorregresivo de Primer Orden

Los modelos autorregresivos son una clase de modelos de series temporales utilizados para predecir el valor futuro de una variable en función de sus valores pasados. Estos modelos son ampliamente utilizados en economía, por ejemplo, para predecir la inflación, una variable clave para los inversores, bancos centrales, empresas y gobiernos.

Definición del Modelo AR(1) Un modelo autorregresivo de primer orden, abreviado como AR(1), predice el valor de una serie temporal Y_t basándose en su valor inmediatamente anterior Y_{t-1}. Matemáticamente, el modelo AR(1) se define como:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

Donde u_t es un término de error que captura la variación no explicada por los valores pasados. En un ejemplo aplicado a la tasa de inflación trimestral en EE.UU., la ecuación estimada es:

$$\Delta \text{Inf}_t = 0.017 - 0.238\Delta \text{Inf}_{t-1}$$

Donde $\Delta Inft$ representa la variación en la tasa de inflación del trimestre t respecto al trimestre $t-1$. El coeficiente $\beta_1 = -0.238$ indica que un aumento en la inflación en el trimestre anterior está asociado con una disminución en la inflación en el trimestre siguiente.

Concepto clave 14.3: Modelos autorregresivos

El modelo autorregresivo de orden p [el modelo AR(p)] representa Y t como una función lineal de sus p primeros valores retardados:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t$$

donde $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$. El número de retardos p se denomina orden, o longitud de los retardos, de la autorregresión.

Predicción con el Modelo AR(1) Para realizar predicciones con el modelo AR(1), se utilizan los estimadores de Mínimos Cuadrados Ordinarios (MCO) $\hat{\beta}_0$ y $\hat{\beta}_1$ basados en datos históricos. La predicción para el valor de Y_{T+1} dada la información hasta el periodo T se expresa como:

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$$

El error de predicción es la diferencia entre el valor real observado Y_{T+1} y la predicción $\hat{Y}_{T+1} \mid T$:

$$\text{Error de predicción} = Y_{T+1} - \hat{Y}_{T+1|T}$$

Raíz del Error Cuadrático Medio de Predicción (RECMP) La raíz del error cuadrático medio de predicción (RECMP) es una medida de la magnitud del error de predicción. Se define como:

$$\text{RECMP} = \sqrt{\mathbb{E}[(Y_{T+1} - \hat{Y}_{T+1|T})^2]}$$

Este error tiene dos componentes: uno derivado de la incertidumbre sobre el futuro valor de u_t y otro de la incertidumbre en la estimación de β_0 y β_1. Si el tamaño de la muestra es grande, el primer componente suele dominar, y la RECMP puede aproximarse por la desviación estándar del término de error u_t, estimada mediante el error estándar de la regresión.

El Modelo Autorregresivo de Orden p (AR(p))

El modelo AR(1) puede extenderse al modelo autorregresivo de orden p, AR(p), en el que Y_t se predice utilizando sus primeros p valores retardados:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t$$

Aquí, p indica el número de retardos incluidos en el modelo. Este modelo es útil cuando se desea capturar la influencia de valores pasados más lejanos en la serie temporal.

Propiedades de la Predicción y del Término de Error en AR(p) Si Y_t sigue un proceso AR(p), la mejor predicción para Y_{T+1} basada en la historia completa de Y depende únicamente de los p valores más recientes:

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \hat{\beta}_2 Y_{T-1} + \cdots + \hat{\beta}_p Y_{T-p+1}$$

El término de error u_t en este modelo se asume que es serialmente incorrelacionado, lo que implica que no existe autocorrelación entre los errores en diferentes periodos.

17.4 Regresión de series temporales con predictores adicionales y modelo autorregresivo de retardos distribuidos

La teoría económica frecuentemente sugiere variables adicionales que podrían ayudar a predecir la variable de interés. Estas variables, o predictores, pueden incorporarse a un modelo de autorregresión para formar un modelo de regresión de series temporales con múltiples predictores. La adición de otras variables y sus retardos da lugar a un modelo autorregresivo de retardos distribuidos (ARD).

Predicción de la Variación de la Tasa de Inflación mediante Valores Pasados de la Tasa de Desempleo

Un valor alto de la tasa de desempleo suele estar asociado con una futura disminución de la tasa de inflación, lo que se conoce como la curva de Phillips de corto plazo. Por ejemplo, en 1982 la tasa de desempleo promedió un 9.7 %, mientras que la tasa de inflación cayó al 2.9correlación observada es 0.36.

Para evaluar si los valores pasados de la tasa de desempleo aportan información adicional sobre la inflación futura, se amplía el modelo AR(4) de la Ecuación (14.13) para incluir el primer retardo de la tasa de desempleo:

$$\Delta \text{Inf}_t = 1.28 - 0.31\Delta \text{Inf}_{t-1} - 0.39\Delta \text{Inf}_{t-2} - 0.09\Delta \text{Inf}_{t-3} - 0.08\Delta \text{Inf}_{t-4} - 0.21\text{Desemp}_{t-1}$$

Donde el estadístico t para la variable $Desemp(t-1)$ es -2.23, lo que indica significancia al nivel del 5 %. El R^2 de este modelo es 0.21, mejorando el R^2 del modelo AR(4) de 0.18.

Incorporando tres retardos adicionales de la tasa de desempleo, el modelo se expande a:

$$\begin{aligned}\Delta \text{Inf}_t =& 1.30 - 0.42\Delta \text{Inf}_{t-1} - 0.37\Delta \text{Inf}_{t-2} - 0.06\Delta \text{Inf}_{t-3} - 0.04\Delta \text{Inf}_{t-4} && (17.5)\\ & - 2.64\text{Desemp}_{t-1} - 3.04\text{Desemp}_{t-2} + 0.38\text{Desemp}_{t-3} - 0.25\text{Desemp}_{t-4} && (17.6)\end{aligned}$$

El estadístico F para los retardos de la tasa de desempleo es 10.76 (p-valor <0.001), indicando significancia conjunta. El R^2 es 0.34, mejorando significativamente respecto al modelo anterior, y el error estándar de la regresión es 1.36, comparado con 1.52 en el modelo AR(4).

Modelo Autorregresivo de Retardos Distribuidos (ARD)

Los modelos ARD combinan retardos de la variable dependiente con retardos distribuidos de un predictor adicional. Un modelo ARD(p, q) incluye p retardos de la variable dependiente y q retardos de un predictor adicional. Los modelos ARD de las Ecuaciones (14.16) y (14.17) se denominan ARD(4,1) y ARD(4,4) respectivamente.

Concepto clave 14.4: El modelo autorregresivo de retardos distribuidos

El modelo autorregresivo de retardos distribuidos con p retardos de Yt y q retardos de Xt, denominado ARD(p, q), es

$$\begin{aligned}Y_t =& \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} && (17.7)\\ & + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \cdots + \delta_q X_{t-q} + u_t, && (17.8)\end{aligned}$$

donde $\beta_0, \beta_1, ..., \beta_p, \delta_1, ..., \delta_q$ son coeficientes desconocidos y u_t es el término de error con $E(u_t \mid Y_{t-1}, Y_{t-2}, ..., X_{t-1}, X_{t-2}, ...) = 0$.

La estacionariedad es un supuesto clave en la regresión de series temporales, que implica que la distribución de la variable no cambia con el tiempo. En caso de no estacionariedad, la predicción puede estar sesgada o ser ineficiente. La no estacionariedad puede manifestarse en forma de tendencias o cambios estructurales, que se abordan en secciones posteriores.

Concepto clave 14.5: Estacionariedad

Una serie temporal Y_t es estacionaria si su distribución de probabilidad no varía en el tiempo, es decir, si la distribución conjunta de $(Y_{s+1}, Y_{s+2}, ..., Y_{s+T})$ no depende de s sea cual sea el valor de T; de lo contrario, se dice que Y_t es no estacionaria. Dos series temporales, X_t e Y_t, se dice que son conjuntamente estacionarias si la distribución conjunta de $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, ..., X_{s+T}, Y_{s+T})$ no depende de s, independientemente del valor de T. La estacionariedad requiere que el futuro sea como el pasado, al menos en un sentido probabilístico.

Regresión de Series Temporales con Múltiples Predictores

El modelo general con múltiples predictores se expande a:

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{t-i} + \sum_{j=1}^{q} \gamma_j Z_{t-j} + u_t$$

donde X_{t-i} son los predictores y Z_{t-j} son sus retardos. Los supuestos para este modelo incluyen:

- u_t tiene una media condicional igual a cero dado X y Z.
- Los datos provienen de una distribución estacionaria.
- Las variables aleatorias son independientes cuando están separadas por largos periodos.
- Los regresores no presentan multicolinealidad perfecta.

Concepto clave 14.6: Regresión de series temporales con varios predictores

El modelo general de regresión de series temporales permite k predictores adicionales, en el que se incluyen q_1 retardos del primer predictor, q_2 retardos del segundo predictor, y así sucesivamente:

$$\begin{aligned} Y_t = {} & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} && (17.9) \\ & + \delta_{11} X_{1t-1} + \delta_{12} X_{1t-2} + \cdots + \delta_{1q_1} X_{1t-q_1} && (17.10) \\ & + \cdots + \delta_{k1} X_{kt-1} + \delta_{k2} X_{kt-2} + \cdots + \delta_{kq_k} X_{kt-q_k} + u_t && (17.11) \end{aligned}$$

Donde:

1. $E(u_t \mid Y_{t-1}, Y_{t-1}, ..., X_{1t-1}, X_{1t-2}, ..., X_{kt-1}, X_{kt-2}, ...) = 0$.
2. (a) Las variables aleatorias $(Y_t, X_{1t}, ..., X_{kt})$ presentan una distribución estacionaria, y (b) $(Y_t, X_{1t}, ..., X_{kt})$ y $(Y_{t-j}, X_{1t-j}, ..., X_{kt-j})$ pasan a ser independientes cuando j se hace grande.
3. Los valores extremos elevados son poco probables: $X_{1t}, ..., X_{kt}, Y_t$ presentan momentos de cuarto orden finitos y distintos de cero.
4. No existe multicolinealidad perfecta.

Inferencia Estadística y Causalidad de Granger

La causalidad de Granger evalúa si un predictor tiene contenido predictivo adicional para una variable dependiente. El estadístico F para el contraste de causalidad de Granger verifica si los retardos de una variable aportan información predictiva significativa más allá de otras variables del modelo.

Concepto clave 14.7: Contraste de causalidad de Granger (contraste de contenido predictivo)

El estadístico para el contraste de causalidad de Granger es el estadístico F para el contraste de la hipótesis de que los coeficientes de todos los valores de una de las variables de la Ecuación (17.11) (por ejemplo, los coeficientes de $X_{1t-1}, X_{1t-2}, ..., X_{1t-q1}$) son iguales a cero. Esta hipótesis nula implica que estos regresores no tienen contenido predictivo para Y_t más allá del contenido en los otros regresores, y el contraste de esta hipótesis nula se denomina contraste de causalidad de Granger.

Incertidumbre de la Predicción e Intervalos de Predicción

La incertidumbre de la predicción incluye la variabilidad en la estimación de los coeficientes y el error futuro. La raíz del error cuadrático medio de predicción (RECMP) se calcula y se utiliza para construir intervalos de predicción. Bajo la suposición de normalidad, los intervalos de predicción pueden ser estimados y proporcionan una medida de la certeza de las predicciones.

$$RECMP = \sqrt{ECMP}$$

17.5 Selección de la longitud de los retardos mediante criterios de información

En el análisis de series temporales, determinar el número adecuado de retardos en una regresión es crucial. Este número afecta tanto la precisión del modelo como la complejidad de las estimaciones. Examinamos dos enfoques principales para la selección de retardos: la selección en autorregresiones y en modelos de regresión de series temporales con múltiples predictores.

Determinación del Orden de una Autorregresión

El orden de una autorregresión, denotado como p, debe equilibrar el beneficio marginal de incluir más retardos con el costo marginal de la incertidumbre adicional en las estimaciones.

Método del Estadístico F Un enfoque consiste en comenzar con un modelo con un número elevado de retardos y realizar contrastes de hipótesis sobre los coeficientes de los retardos más lejanos. Por ejemplo, se puede comenzar con un modelo AR(6) y verificar la significancia del coeficiente del sexto retardo. Si el coeficiente no es significativo al nivel del 5 %, se reduce el modelo a AR(5) y así sucesivamente. Este método puede resultar en la sobreestimación del número de retardos, ya que incluso cuando el verdadero orden es p, el contraste puede indicar incorrectamente la inclusión de un retardo adicional.

Criterio de Información de Bayes (BIC) El BIC, también conocido como Criterio de Información Schwarz (SIC), se define como:

$$\text{BIC}(p) = \ln\left(\frac{\text{SR}(p)}{T}\right) + \frac{(p+1)\ln(T)}{T}$$

donde SR(p) es la suma de los cuadrados de los residuos del modelo AR(p), y T es el número de observaciones. El valor de p que minimiza el BIC es considerado como la longitud óptima del retardo. El BIC penaliza los modelos con muchos retardos más fuertemente que el AIC, y por lo tanto, suele seleccionar modelos más simples.

Criterio de Información de Akaike (AIC) El AIC se define como:

$$\text{AIC}(p) = \ln\left(\frac{\text{SR}(p)}{T}\right) + \frac{2(p+1)}{T}$$

El AIC tiene un término de penalización más pequeño en comparación con el BIC, lo que significa que es menos severo en la penalización por complejidad del modelo. Aunque el AIC es menos consistente que el BIC para grandes muestras, se utiliza frecuentemente en la práctica debido a su menor penalización por agregar retardos.

17.6 Ausencia de estacionariedad I: Tendencias

En el concepto clave 14.6, se asumió que tanto la variable dependiente como los regresores eran esta- cionarios. Sin embargo, si estas variables no son estacionarias, los contrastes de hipótesis, los intervalos de confianza y

las predicciones pueden ser poco fiables. En esta sección y la siguiente, se examinan dos tipos importantes de ausencia de estacionariedad en datos de series temporales económicas: las tendencias y los cambios estructurales.

¿Qué es una tendencia?

Una tendencia es un movimiento persistente a largo plazo de una variable en el tiempo. La variable fluctúa en torno a su tendencia. Por ejemplo, la tasa de inflación en EE.UU. ha mostrado una tendencia creciente hasta 1982 y luego descendente. El tipo de cambio USD/JPY mostró una tendencia a la baja tras 1972, y el logaritmo del PIB de Japón mostró una tendencia complicada con crecimiento rápido al principio, moderado después y lento finalmente.

Tipos de tendencias

- **Determinísticas:** Una tendencia determinística es una función no aleatoria del tiempo, como una tendencia lineal.
- **Estocásticas:** Una tendencia estocástica es aleatoria y varía en el tiempo. Es más adecuada para modelar series temporales económicas debido a su complejidad y la imprevisibilidad de las fuerzas económicas.

Modelo de Paseo Aleatorio

El modelo más simple para una tendencia estocástica es el paseo aleatorio. Se define como:

$$Y_t = Y_{t-1} + u_t$$

donde u_t es i.i.d. (independiente e idénticamente distribuido) con media condicional igual a cero. En este modelo, el valor futuro de la serie es el valor presente más una variación impredecible. La extensión del modelo incluye un movimiento tendencial o "deriva":

$$Y_t = b_0 + Y_{t-1} + u_t$$

donde b_0 es la deriva. Si b_0 es positivo, la serie aumenta en promedio.

Propiedades del Paseo Aleatorio Un paseo aleatorio no es estacionario. La varianza de Y_t aumenta con el tiempo, y por lo tanto, la distribución de Yt cambia con el tiempo. La varianza de un paseo aleatorio es:

$$\text{var}(Y_t) = t \cdot \sigma_u^2$$

donde σ_u^2 es la varianza de u_t. Esto indica que la varianza depende del tiempo, y por lo tanto, Y_t es no estacionario.

Raíz Unitaria y Modelos AR

Un paseo aleatorio es un caso particular del modelo AR(1) donde $b_1 = 1$. Para que un proceso AR(p) sea estacionario, todas las raíces del polinomio característico deben ser mayores que 1 en valor absoluto. Si una raíz es igual a 1, la serie tiene una raíz unitaria y es no estacionaria.

Problemas con las Tendencias Estocásticas

- **Sesgo hacia cero en coeficientes autorregresivos:** El estimador MCO del coeficiente autor- regresivo en un modelo AR(1) está sesgado hacia cero si el verdadero valor es 1.
- **Distribuciones no normales del estadístico t:** El estadístico t MCO puede tener una distribu- ción distinta de la normal, afectando la validez de los intervalos de confianza y los contrastes de hipótesis.
- **Regresión espuria:** Dos series con tendencias estocásticas pueden parecer relacionadas cuando en realidad no lo están, lo que lleva a regresiones espurias.

Detección de Tendencias Estocásticas

El contraste de Dickey-Fuller se utiliza para detectar la presencia de una tendencia estocástica en una serie temporal. La hipótesis nula es que la serie tiene una raíz unitaria, mientras que la alternativa es que es estacionaria. La versión aumentada del contraste de Dickey-Fuller (ADF) se utiliza para modelos AR(p) y puede ser estimada mediante criterios de información.

Concepto clave 14.8: El contraste de Dickey-Fuller aumentado para raíz unitaria autorregresiva

El contraste de Dickey-Fuller aumentado (ADF) para una raíz unitaria autorregresiva contrasta la hipótesis nula $H_0 : \delta = 0$ frente a la hipótesis alternativa unilateral $H_1 : \delta < 0$ en la regresión

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_p \Delta Y_{t-p} + u_t. \tag{17.12}$$

Bajo la hipótesis nula, Y_t tiene una tendencia estocástica; bajo la hipótesis alternativa, Y_t es estacionaria. El estadístico ADF es el estadístico t MCO para contrastar $\delta = 0$ en la Ecuación (17.12).
Si en su lugar, la hipótesis alternativa es que Y_t es estacionaria en torno a una tendencia temporal lineal determinística, entonces debe añadirse esta tendencia, «t» (el número de observación), como regresor adicional, en cuyo caso la regresión de Dickey-Fuller se convierte en:

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_p \Delta Y_{t-p} + u_t, \tag{17.13}$$

Donde a es un coeficiente desconocido y el estadístico ADF es el estadístico t MCO para contrastar $\delta = 0$ en la Ecuación (17.13).
La longitud del retardo, p, se puede estimar utilizando el criterio BIC o AIC. Cuando $p = 0$, no se incluyen retardos de ΔY_t como regresores en las Ecuaciones (17.12) y (17.13), y el contraste ADF se simplifica al contraste de Dickey-Fuller en el modelo AR(1). El estadístico ADF no sigue una distribución normal, incluso en muestras grandes. Los valores críticos para el contraste ADF unilateral dependen de si el contraste está basado en la Ecuación (17.12) o en la (17.13) y se presentan en la Tabla 14.5.

Contraste de Raíz Unitaria en Modelos AR El contraste de Dickey-Fuller puede ser extendido al modelo AR(p) y se conoce como el estadístico ADF. La longitud del retardo p puede ser estimada utilizando criterios de información como el AIC.

Contraste Frente a una Tendencia Determinística Para series que muestran crecimiento a largo plazo, como el PIB japonés, se utilizan alternativas que consideran una tendencia determinística además de la estocástica.

17.7 Ausencia de estacionariedad II: Cambios estructurales

Un tipo de no estacionariedad ocurre cuando la función de regresión poblacional cambia durante la muestra. En economía, esto puede deberse a cambios en políticas, estructuras económicas o innovaciones que alteran industrias específicas. Estos cambios estructurales o rupturas pueden afectar la capacidad del modelo de regresión para hacer inferencias y predicciones precisas.

Detección de Cambios Estructurales

Contrastes de Hipótesis: Estrategias

- **Cambio Estructural con Punto de Ruptura Conocido:** Se puede detectar mediante la inclusión de una variable binaria que indica el período de cambio. El modelo estimado es:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + c_0 D_t(q) + c_1 [D_t(q) \cdot Y_{t-1}] + c_2 [D_t(q) \cdot X_{t-1}] + u_t \tag{17.14}$$

  Donde $D_t(q)$ es 0 antes del cambio estructural y 1 después. La hipótesis nula de ausencia de cambio estructural se prueba con el estadístico F para verificar si $c_0 = c_1 = c_2 = 0$.

- **Cambio Estructural con Punto de Ruptura Desconocido:** Se usa el estadístico de razón de verosimilitud de Quandt (QLR), que compara todos los posibles puntos de ruptura en un rango y usa el valor máximo de los estadísticos F para determinar el punto de ruptura. Los valores críticos para el estadístico QLR se deben obtener de una distribución particular, ya que el QLR tiene una distribución diferente a la del estadístico F individual.

Concepto clave 14.9: El contraste QLR para la estabilidad de los coeficientes

Sea F(q) la expresión del estadístico F para el contraste de la hipótesis de cambio estructural en los coeficientes de regresión en el momento τ; por ejemplo, en la regresión de la Ecuación (17.14), este es el estadístico F para contrastar la hipótesis nula de que $\gamma_0 = \gamma_1 = \gamma_2 = 0$. El estadístico de contraste QLR (o de sup Wald) es el mayor de los estadísticos dentro del rango $\tau_0 \leq \tau \leq \tau_1$:

$$\text{QLR} = \text{máx}[F(\tau_0), F(\tau_0 + 1), ..., F(\tau_1)].$$

1. Al igual que el estadístico F, el estadístico QLR puede ser utilizado para contrastar la existencia de un cambio estructural en todos o solo en algunos de los coeficientes de regresión.
2. En muestras grandes, la distribución del estadístico QLR bajo la hipótesis nula depende del número de restricciones que se contrasten, q, y de los extremos τ_0 y τ_1 como proporción de T. Los valores críticos están recogidos en la Tabla 14.6 para un 15 % de reducción ($\tau_0 = 0,15T$ y $\tau_1 = 0,85T$, redondeando al entero más cercano).
3. El contraste QLR puede detectar la existencia de un único cambio estructural discreto, varios cambios estructurales discretos, y/o la evolución lenta de la función de regresión.
4. Si existe un cambio estructural evidente en la función de regresión, el periodo en el que se registra el estadístico de Chow mayor es un estimador del punto de ruptura.

Ejemplos

- **Cambio Discreto:** Un evento como el colapso del sistema Bretton Woods en 1972 puede causar un cambio estructural evidente en las series temporales del tipo de cambio.
- **Cambio Gradual:** Evoluciones lentas en políticas económicas o en la estructura económica que afectan la regresión poblacional.

Problemas Los cambios estructurales pueden llevar a una estimación promedio en la regresión MCO, que puede no reflejar la verdadera relación al final de la muestra, resultando en malas predicciones.

Predicción Pseudo Fuera de la Muestra

La predicción pseudo fuera de la muestra es un método para evaluar el rendimiento predictivo de un modelo. Consiste en estimar el modelo con datos hasta un periodo cercano al final de la muestra y luego hacer predicciones usando ese modelo. Estas predicciones pueden ser comparadas con los datos reales futuros para evaluar la precisión del modelo.

Utilidades

- **Evaluación del Modelo:** Permite verificar si el modelo se comporta bien en periodos recientes y ajustar la confianza en la predicción.
- **Estimación de la RECMP:** Proporciona un estimador de la desviación típica muestral de los errores de predicción, útil para construir intervalos de predicción.
- **Comparación de Modelos:** Facilita la comparación de modelos candidatos para predicción basándose en su rendimiento pseudo fuera de la muestra.

Concepto clave 14.10: Predicciones pseudo fuera de la muestra

Las predicciones pseudo fuera de la muestra se calculan siguiendo los siguientes pasos:

1. Elegir un número de observaciones, P, para las que se van a generar las predicciones pseudo fuera de la muestra; por ejemplo, P podría ser el 10% o el 15% del tamaño de la muestra. Sea $s = T - P$.
2. Estimar la regresión de predicción con el conjunto de datos reducido para $t = 1, ..., s$.
3. Calcular la predicción para el primer periodo más allá de esta muestra reducida, $s + 1$; y denominarlo $\tilde{Y}_{s+1|s}$.
4. Calcular el error de predicción, $\tilde{u}_{s+1} = Y_{s+1} - \tilde{Y}_{s+1|s}$.
5. Repetir los pasos 2 a 4 para los periodos restantes, $s = T - P + 1$ hasta $T - 1$ (reestimando la regresión para cada periodo). Las predicciones pseudo fuera de la muestra son $\tilde{Y}_{s+1|s}$, $s = T - P, ..., T - 1$, y los errores de predicción pseudo fuera de la muestra son $\tilde{u}_{s+1}$, $s = T - P, ..., T - 1$.

Resolución de Problemas

Para problemas de cambios estructurales evidentes, se puede estimar una función de regresión ajustada para los periodos antes y después del cambio. Si el cambio es gradual, puede ser más complejo, requiriendo métodos adicionales no cubiertos aquí.

Este enfoque proporciona herramientas clave para manejar y detectar cambios estructurales en series temporales, permitiendo una mejor adaptación y predicción en modelos econométricos.

Capítulo 18

Estimación de efectos causales dinámicos

Este capítulo trata sobre la estimación de efectos causales dinámicos, como el impacto del clima frío en Florida sobre los precios del zumo de naranja concentrado. Se presenta el modelo de retardos distribuidos, que relaciona Y_t con los valores actuales y pasados de X_t, y se analizan métodos de estimación como los mínimos cuadrados ordinarios (MCO) y generalizados (MCG), destacando la importancia de la exogeneidad. También se examinan aplicaciones empíricas y casos prácticos en macroeconomía y finanzas, ofreciendo herramientas clave para estudiar impactos temporales en diversos contextos económicos.

18.1 Un «primer gusto en boca» de los datos del zumo de naranja

El estudio examina cómo las condiciones meteorológicas, en particular las heladas, afectan el precio del concentrado de jugo de naranja en la región productora de naranjas de Florida. En este contexto, las heladas disminuyen la oferta de naranjas, lo que, a su vez, reduce la oferta de jugo de naranja concentrado y eleva su precio. Sin embargo, dado que el concentrado es un bien almacenable, el precio no depende únicamente de la oferta actual, sino también de las expectativas sobre la oferta futura. Así, el precio actual del concentrado refleja no solo el estado de la oferta en el momento presente, sino también la anticipación de una menor oferta futura debido a las heladas.

La relación entre el precio del concentrado de jugo de naranja y las heladas se modela a través de una regresión lineal simple. La variable dependiente en este modelo es la variación porcentual del precio del concentrado de jugo de naranja, mientras que la variable independiente es el índice de heladas en la región. La fórmula básica de la regresión es la siguiente:

$$\%VPt = \beta_0 + \beta_1 IH_t + \epsilon_t$$

Donde:

- $\%VPt$ es la variación porcentual del precio del concentrado de jugo de naranja en el mes t.
- IH_t es el índice de heladas en el mes t.
- β_0 es el término constante.
- β_1 es el coeficiente de la variable independiente IH_t.
- ϵ_t es el error del modelo.

Los resultados de la regresión muestran que un aumento unitario en el índice de heladas (IHt) incrementa la variación porcentual del precio del concentrado de jugo de naranja en un 0.47 % . Esto implica que, en un mes con un índice de heladas de 4, como ocurrió en noviembre de 1950, se estima que el precio del concentrado aumentó en un 1.88 % en comparación con un mes sin heladas.

Para capturar los efectos persistentes de las heladas sobre los precios, se amplía el modelo básico incorporando los valores retardados del índice de heladas en los meses anteriores. Esto da lugar a una regresión de retardos distribuidos, en la que se incluyen las variables $IH_t, IH_{t-1}, IH_{t-2}, ..., IH_{t-6}$, correspondientes a los seis meses previos. La fórmula de este modelo es la siguiente:

$$\%VPt = \beta_0 + \beta_1 IH_t + \beta_2 IH_{t-1} + \beta_3 IH_{t-2} + \beta_4 IH_{t-3} + \beta_5 IH_{t-4} + \beta_6 IH_{t-5} + \beta_7 IH_{t-6} + \epsilon_t$$

En este caso, los coeficientes $\beta_1, \beta_2, ..., \beta_7$ estiman el impacto de las heladas sobre la variación porcentual del precio del concentrado de jugo de naranja en el mes t y en los meses posteriores. Los coeficientes estimados sugieren que un aumento unitario en el índice de heladas en el mes t incrementa el precio del concentrado de jugo de naranja en un 0.47 %. Además, un aumento en el índice de heladas en el mes $t-1$ tiene un efecto adicional del 0.14 %, y así sucesivamente con los valores retardados.

Este modelo permite comprender no solo los efectos inmediatos de las heladas sobre los precios, sino también cómo estos efectos persisten a lo largo del tiempo. En conclusión, las heladas no solo afectan el precio del concentrado de jugo de naranja de manera inmediata, sino que tienen efectos a largo plazo sobre los precios debido a la capacidad de almacenamiento y las expectativas de oferta futura.

18.2 Efectos causales dinámicos

Antes de conocer más acerca de las herramientas disponibles para la estimación de los efectos causales dinámicos, deberíamos parar un momento a pensar acerca de lo que, exactamente, se entiende por un efecto causal dinámico.

Efectos causales y datos de series temporales En la Sección 1.2 fue definido efecto causal como aquel resultado de un experimento aleatorizado controlado ideal. En las aplicaciones con series temporales, esta definición de los efectos causales necesita ser modificado.

Con datos de series temporales, resulta útil imaginar un experimento aleatorizado controlado que consista en someter al mismo sujeto a distintos tratamientos para diferentes momentos del tiempo (el mismo sujeto en diferentes momentos desempeña el papel tanto del grupo de tratamiento como del grupo de control). Debido a que los datos se recogen a lo largo del tiempo, es posible estimar el efecto causal dinámico, es decir, *la senda temporal de los efectos del tratamiento sobre los resultados de interés.*

Efectos dinámicos y modelo de retardos distribuidos Debido a que los efectos dinámicos nece- sariamente ocurren en el tiempo, resulta necesario que el modelo econométrico utilizado para estimar los efectos causales dinámicos incorpore retardos. Para hacerlo, se puede expresar Yt como un modelo de retardos distribuidos.

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \epsilon_t + \lambda_1 X_{t-r} + u_t \tag{18.1}$$

Implicaciones para el análisis empírico de series temporales Esta formulación de los efectos causales dinámicos en datos de series temporales como el resultado esperado de un experimento en el cual se aplican repetidamente diferentes niveles de tratamiento al mismo sujeto, tiene implicaciones.

- el efecto causal dinámico no debería cambiar a lo largo de la muestra sobre la que se dispone de datos.
- está implícito en los datos que son conjuntamente estacionarios.
- X no debe estar correlacionada con el término de error, y es en esta implicación en la que ahora se centra el análisis.

La hipótesis de que una función de regresión poblacional es estable en el tiempo puede contrastarse mediante el contraste QLR para un cambio estructural.

Dos tipos de exogeneidad

En la Sección 12.1 se definía como variable «exógena» a una variable que no estaba correlacionada con el término de error de la regresión y como variable «endógena» a una variable que estaba correlacionada con el término de error. Esta terminología sigue la senda de los modelos de varias ecuaciones, en los que una variable «endógena» se determina dentro del modelo mientras que una variable «exógena» se determina fuera del modelo. En términos generales, si han de estimarse los efectos causales dinámicos mediante el modelo de retardos distribuidos de la Ecuación (18.1), las variables explicativas (las X) deben estar incorrelacionadas con el término de error. Por lo tanto X debe ser exógena. Sin embargo, debido a que se trabaja con datos de series temporales, resulta necesario afinar las definiciones de exogeneidad. De hecho, existen dos conceptos diferentes de exogeneidad que aquí se utilizan.

El primer concepto de exogeneidad es que el término de error tiene una media condicional igual a cero, dados los valores actuales, y todos los anteriores de X_t, es decir, que $E(u_t \mid X_t, X_{t-1}, X_{t-2}, ...) = 0$. Esto modifica el supuesto habitual de media condicional para regresión múltiple con datos de sección cruzada [Supuesto 1 del Concepto clave 6.4(7.5)], que solo requiere que u_t tenga una media condicional igual a cero, dados los regresores incluidos, es decir, que $E(u_t \mid X_t, X_{t-1}, ..., X_{t-r}) = 0$. La inclusión de todos los valores retardados de X_t en la esperanza condicional implica que todos los efectos causales más distantes —todos aquellos efectos causales más allá del retardo r— son iguales a cero. Por tanto, bajo este supuesto, los coeficientes de los r retardos distribuidos de la Ecuación (18.1) constituyen todos los efectos causales dinámicos distintos de cero. Se puede denominar a este supuesto —que $E(u_t \mid X_t, X_{t-1}, ...) = 0$— como exogeneidad pasada y presente, pero debido a la similitud de esta definición con la definición de exogeneidad del Capítulo 12, solamente se utiliza el término **exogeneidad**.

El segundo concepto de exogeneidad es que el término de error tiene media igual a cero, dados todos los valores pasados, presentes y futuros de X_t, es decir, que $E(u_t \mid ..., X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, ...) = 0$. Esto se denomina **exogeneidad estricta**; para mayor claridad, puede denominarse asimismo exogeneidad pasada, presente, y futura. La razón de introducir el concepto de exogeneidad estricta es que, cuando X es estrictamente exógena, existen estimadores más eficientes de los efectos causales dinámicos que los estimadores MCO de los coeficientes de la regresión de retardos distribuidos de la Ecuación (18.1).

Una *manera de entender la diferencia entre ambos conceptos* es considerar las implicaciones de estas definiciones para las correlaciones entre X y u. Si X es exógena (pasada y presente), entonces u_t no está correlacionado con los valores actuales y pasados de X_t. Si X es estrictamente exógena, además u_t no está correlacionado con los valores futuros de X_t. Por ejemplo, si una variación en Y_t provoca variaciones en los valores futuros de X_t, entonces X_t no es estrictamente exógena a pesar de que podría ser exógena (pasada y presente).

Concepto clave 15.1: El modelo de retardos distribuidos y la exogeneidad

En el modelo de retardos distribuidos

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \cdots + \beta_{r+1} X_{t-r} + u_t, \tag{18.2}$$

existen dos tipos diferentes de exogeneidad, es decir, dos condiciones diferentes de exo- geneidad: Exogeneidad pasada y presente (exogeneidad):

$$E(u_t | X_t, X_{t-1}, X_{t-2}, ...) = 0;$$

Exogeneidad pasada, presente y futura (exogeneidad estricta):

$$E(u_t | ..., X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, ...) = 0$$

Si X es estrictamente exógena, es exógena, pero la exogeneidad no implica la exogenei- dad estricta.

18.3 Estimación de efectos causales dinámicos con regresores exóge- nos

Si X es exógena, entonces su efecto causal dinámico sobre Y se puede estimar mediante la estimación MCO de la regresión de retardos distribuidos En esta sección se recogen las condiciones bajo las cuales estos estimadores MCO dan lugar a inferencias estadísticas válidas.

Los supuestos del modelo de retardos distribuidos

son similares a los cuatro supuestos del modelo de regresión múltiple para datos de sección cruzada (Concepto clave 6.4(7.5)), modificados para los datos de series de temporales.

1. El primer supuesto es que X es exógena.

2. El segundo supuesto tiene dos partes: que las variables tengan una distribución estacionaria y requiere que pasen a ser independientemente distribuidas a medida que el espacio temporal que las separa aumente en gran medida.
3. El tercer supuesto es que los valores extremos muy grandes son poco probables (mediante el supuesto de que las variables tienen más de ocho momentos finitos y distintos de cero. este supuesto más fuerte es el que se utiliza en secciones anteriores tras el estimador de la varianza HAC).
4. El cuarto supuesto es que no exista multicolinealidad perfecta.

Concepto clave 15.2: Los supuestos del modelo de retardos distribuidos

El modelo de retardos distribuidos está recogido en el Concepto clave 15.1 [Ecuación (18.2)], donde:
1. X es exógena, es decir, $E(u_t \mid X_t, X_{t-1}, X_{t-2}, \ldots) = 0$.
2. a) Las variables aleatorias Y_t y X_t tienen una distribución estacionaria, y (b) (Y_t, X_t) y (Y_{t-j}, X_{t-j}) se hacen independientes a medida que j se hace grande.
3. Los valores extremos elevados son poco probables; Y_t y X_t tienen más de ocho momentos finitos distintos de cero.
4. No existe multicolinealidad perfecta.

u_t autocorrelacionados, errores estándar e inferencia

En el modelo de regresión de retardos distribuidos, el término de error u_t puede estar autocorrelacionado debido a factores omitidos. Si estos factores están correlacionados serialmente, u_t también lo estará.

La autocorrelación en ut no afecta la consistencia de MCO ni introduce sesgo, pero los errores estándar MCO serán inconsistentes. Esto es similar a la heterocedasticidad: utilizar errores estándar válidos para homocedasticidad en presencia de heterocedasticidad produce inferencias engañosas. La solución es usar errores estándar consistentes a heterocedasticidad y autocorrelación (HAC).

Multiplicadores dinámicos y multiplicadores dinámicos acumulativos

Los **multiplicadores dinámicos** miden el efecto de una variación unitaria en X sobre Y en el periodo h, siendo b_h el coeficiente de X_t en la regresión de retardos distribuidos. Por ejemplo, b_2 es el multiplicador dinámico de un periodo, b_3 de dos periodos, etc. El multiplicador dinámico del periodo cero es b_1, conocido como *efecto impacto*.

Los **multiplicadores dinámicos acumulativos** miden el efecto acumulado de una variación en X sobre Y en los primeros h periodos. Se calculan como la suma acumulada de los multiplicadores dinámicos: el acumulativo de un periodo es $b_1 + b_2$, el de dos periodos es $b_1 + b_2 + b_3$, y así sucesivamente. El multiplicador acumulativo de largo plazo es la suma de todos los multiplicadores dinámicos $b_1 + b_2 + \cdots + b_r$.

Una forma de estimar estos multiplicadores es a través de la regresión de retardos distribuidos modificada, dada por:

$$Y_t = d_0 + d_1 X_t + d_2 X_{t-1} + \cdots + d_r X_{t-r+1} + u_t$$

Los coeficientes $d_1, d_2, \ldots, d_r$ representan los multiplicadores dinámicos acumulativos, y su estimación por MCO permite obtener los errores estándar HAC correspondientes.

18.4 Errores estándar consistentes en presencia de Heterocedasticidad y autocorrelación

Si el término de error u_t está autocorrelacionado, entonces los estimadores MCO de los coeficientes son consistentes, pero en general los errores estándar MCO habituales para datos de sección cruzada no lo son. Esto significa que las inferencias estadísticas convencionales —los contrastes de hipótesis y los intervalos de confianza— basadas en los errores estándar MCO habituales, en general, inducen a error. Esta sección analiza los errores estándar HAC de la regresión con datos de series temporales.

Distribución del estimador MCO con errores autocorrelacionados

Por simplicidad, se considera el estimador MCO $\hat{\beta}_1$ del modelo de regresión de retardos distribuidos sin retardos, es decir, el modelo de regresión lineal con un único regresor X_t en el que se cumplen los supuestos del Concepto clave 15.2-

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$
$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})u_t}{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

La ecuación anterior es una reformulación de la ecuación (4.30) con un cambio de notación donde i y n se sustituyen por t y T . Para muestras grandes, $\hat{\beta}_1 - \beta_1$ se aproxima por:

$$\widehat{\beta}_1 - \beta_1 \cong \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \mu_X)u_t}{\sigma_X^2} = \frac{\frac{1}{T}\sum_{t=1}^{T} v_t}{\sigma_X^2} = \frac{\bar{v}}{\sigma_X^2}$$

Donde: $v_i = (X_i - \mu_X)u_i$ y $\bar{v} = \frac{1}{T}\sum_{i=1}^{T} v_i$.
La varianza de $\hat{\beta}_1$ se expresa como:

$$\text{var}(\widehat{\beta}_1) = \text{var}\left(\frac{\bar{v}}{\sigma_X^2}\right) = \frac{\text{var}(\bar{v})}{\left(\sigma_X^2\right)^2}$$

Si v_t es i.i.d. entonces $var(\hat{v}) = var(v_t)/T$ y es aplicable la fórmula de $\hat{\beta}_1$ del concepto clave 4.4 , sin embargo, si u_t y X_t no estan distribuidas de forma independiente a lo largo del tiempo, v_t estará serialmente correlacionada y $var(\hat{v}) \neq var(v_t)/T$, en ese caso la varianza de $\hat{v}$ es:

$$\begin{aligned}
\text{var}(v) &= \text{var}[(v_1 + v_2 + \cdots + v_T)/T] \\
&= [\text{var}(v_1) + \text{cov}(v_1, v_2) + \cdots + \text{cov}(v_1, v_T) \\
&+ \text{cov}(v_2, v_1) + \text{var}(v_2) + \cdots + \text{var}(v_T)]/T^2 \\
&= [T\,\text{var}(v_t) + 2(T-1)\,\text{cov}(v_t, v_{t-1}) \\
&+ 2(T-2)\,\text{cov}(v_t, v_{t-2}) + \cdots + 2\,\text{cov}(v_t, v_{t-T+1})]/T^2 \\
&= \frac{\sigma_v^2}{T} f_T
\end{aligned}$$

Donde:

$$f_T = 1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\rho_j \tag{18.3}$$

Y $\rho_j = corr(v_t, v_{t-j})$. Combinando estas fórmulas para la varianza de $\hat{\beta}_1$ cuando v_t esta autocorrelacionada es:

$$\text{var}(\widehat{\beta}_1) = \left[\frac{1}{T}\frac{\sigma_v^2}{\left(\sigma_X^2\right)^2}\right] f_T$$

Este resultado muestra que la varianza de $\hat{\beta}_1$ se ajusta por el factor f_T en presencia de correlación serial, lo que afecta el cálculo del error estándar de MCO.

Errores estándar HAC

Si el factor f_T , definido en la Ecuación (18.3), fuera conocido, la varianza de $\hat{\beta}_1$ podría estimarse multiplicando el estimador habitual de la varianza para sección cruzada por f_T . Sin embargo, este factor depende de las autocorrelaciones de v_t, que son desconocidas y deben ser estimadas. El estimador de la varianza de $\hat{\beta}_1$ que incorpora este ajuste es consistente tanto en presencia de heterocedasticidad como de autocorrelación. Este estimador se denomina **estimador de la varianza de $\hat{\beta}_1$ consistente a heterocedasticidad y autocorrelación (HAC)**, y su raíz cuadrada es el error estándar HAC de $\hat{\beta}_1$.

Fórmula HAC de la Varianza

El estimador HAC de la varianza de $\hat{\beta}_1$ es:

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1} \cdot \tilde{f}_T \tag{18.4}$$

donde $\tilde{\sigma}^2_{\hat{\beta}_1}$ es el estimador de la varianza de $\hat{\beta}_1$ en ausencia de correlación serial (Ecuación 5.4), y $\tilde{f}_T$ es un estimador del factor f_T de la Ecuación (18.3).

Estimación de f_T

La construcción de un estimador consistente $\tilde{f}_T$ es un desafío. Dos enfoques extremos son:

1. **Usar todas las autocorrelaciones muestrales:** Esto lleva a un estimador inconsistente debido al gran número de autocorrelaciones estimadas, lo que introduce un error de estimación significativo.
2. **Usar pocas autocorrelaciones muestrales:** Este enfoque ignora las autocorrelaciones de orden superior, lo que también resulta en un estimador inconsistente.

En la práctica, se busca un equilibrio entre estos extremos. El estimador $\tilde{f}_T$ se define como:

$$\widetilde{f_T} = 1 + 2\sum_{j=1}^{m-1}\left(\frac{m-j}{m}\right)\widetilde{\rho_j} \tag{18.5}$$

Donde:

$$\widetilde{\rho}_j = \frac{\sum_{t=j+1}^{T} \hat{v}_t \hat{v}_{t-j}}{\sum_{t=1}^{T} \hat{v}_t^2}$$

Y $\hat{v}_t = (X_t - \overline{X})\hat{u}_t$. El parámetro m se denomina **parámetro de truncamiento** y debe elegirse de manera que crezca con el tamaño de la muestra T , pero sea mucho menor que T.

Elección del Parámetro de Truncamiento m

Una regla común para elegir m es:

$$m = 0,75 \times T^{1/3} \tag{18.6}$$

redondeado al entero más cercano. Esta regla se basa en el supuesto de una autocorrelación moderada en v_t. Sin embargo, m puede ajustarse según el conocimiento de la serie: aumentar m si hay alta autocorrelación, o disminuirlo si la autocorrelación es baja.

Estimador de Newey-West

El estimador HAC de la Ecuación (18.4), con $\tilde{f}_T$ dado por la Ecuación (18.5), se conoce como **estimador de la varianza de Newey-West**. Este estimador es consistente bajo supuestos generales, siempre que m se elija adecuadamente (Newey y West, 1987).

Extensión a la Regresión Múltiple

Los conceptos anteriores se generalizan al modelo de regresión de retardos distribuidos y a la regresión múltiple con errores serialmente correlacionados. En estos casos, los errores estándar HAC son esenciales para la inferencia. El parámetro de truncamiento m puede elegirse usando la misma regla (Ecuación 18.6), independientemente del número de regresores.

Concepto clave 15.3: Errores estándar HAC

El problema: El término de error u_t en el modelo de regresión de retardos distribuidos del Concepto clave 15.1 puede estar serialmente correlacionado. Si es así, los estimadores MCO de los coeficientes son consistentes, pero en general los errores estándar MCO habituales no lo son, dando lugar a contrastes de hipótesis e intervalos de confianza erróneos.
La solución: Los errores estándar deberían calcularse a partir del estimador de la varianza consistente a heterocedasticidad y autocorrelación (HAC). El estimador HAC implica la estimación de $m-1$ autocovarianzas, así como de la varianza; en el caso de un único regresor, las fórmulas relevantes están recogidas por las Ecuaciones (18.4) y (18.5).
En la práctica, la utilización de los errores estándar HAC implica la elección del parámetro de truncamiento m. Para ello, se utiliza la fórmula de la Ecuación (18.6) como punto de referencia, y a continuación, se aumenta o se disminuye m, dependiendo de si los regresores y los errores presentan una correlación serial elevada o baja.

18.5 Estimación de efectos causales dinámicos con regresores estrictamente exógenos

Cuando X_t es estrictamente exógena, se dispone de dos estimadores alternativos para los efectos causales dinámicos. La estimación de un modelo autorregresivo de retardos distribuidos (ARD) ó estimar los coeficientes del modelo de retardos distribuidos, utilizando mínimos cuadrados generalizados (MCG) en lugar de MCO.

El modelo de retardos distribuidos con errores AR(1)

Supongamos que el efecto causal de X sobre Y tiene una duración de dos periodos, con un efecto inicial b_1 y un efecto en el siguiente periodo b_2. El modelo adecuado es un modelo de retardos distribuidos que considera los valores actuales y pasados de X:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t. \tag{18.7}$$

En general, u_t puede estar serialmente correlacionado. Si se estima mediante MCO, los errores estándar pueden ser engañosos, lo que requiere el uso de errores estándar HAC. Alternativamente, si X_t es estrictamente exógeno, se puede modelar u_t como un proceso autorregresivo de primer orden (AR(1)):

$$u_t = \phi_1 u_{t-1} + \tilde{u}_t, \tag{18.8}$$

donde $\tilde{u}_t$ no está serialmente correlacionado y no es necesario el término independiente porque $E(u_t) = 0$.

Representación ARD Un error serialmente correlacionado puede reescribirse como un modelo autorregresivo de retardos distribuidos con un error serialmente incorrelacionado. Para ello, se retarda cada lado de la Ecuación (18.7) y se resta ϕ_1 multiplicado por este retardo a cada uno de los lados:

$$Y_t - \phi_1 Y_{t-1} = (\beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t) - \phi_1(\beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_{t-1}) \tag{18.9}$$

$$= \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} - \phi_1\beta_0 - \phi_1\beta_1 X_{t-1} - \phi_1\beta_2 X_{t-2} + \tilde{u}_t \tag{18.10}$$

usando la ecuación (18.8)

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \tilde{u}_t, \tag{18.11}$$

Donde:

$$\alpha_0 = \beta_0(1-\phi_1), \quad \delta_0 = \beta_1, \quad \delta_1 = \beta_2 - \phi_1\beta_1, \quad \delta_2 = -\phi_1\beta_2. \tag{18.12}$$

Representación en cuasi-diferencias El modelo también puede expresarse en términos de cuasi- diferencias:

$$Y_t^* = Y_t - \phi_1 Y_{t-1}, \quad X_t^* = X_t - \phi_1 X_{t-1},$$

resultando en:

$$Y_t^* = \alpha_0 + \beta_1 X_t^* + \beta_2 X_{t-1}^* + \tilde{u}_t. \tag{18.13}$$

Ambas representaciones (ARD y cuasi-diferencias) son equivalentes, aunque sugieren diferentes estrategias de estimación.

Supuestos para la estimación Para estimar consistentemente β_1 y β_2, se requiere que X_t^* sea estrictamente exógeno:

$$\mathbb{E}(\tilde{u}_t \mid X_t^*, X_{t-1}^*, \ldots) = 0.$$

Dado que $X_t^* = X_t - \phi_1 X_{t-1}$, esta condición es equivalente a:

$$\mathbb{E}(u_t \mid X_t, X_{t-1}, \ldots) = 0, \tag{18.14}$$

Condicionar a X_t^* y a todos sus retardos es equivalente a condicionar a X_t, y a todos sus retardos. Lo cual implica que X_t debe ser estrictamente exógeno. Este supuesto es más fuerte que la mera exogeneidad contemporánea o pasada.

La condición de la Ecuación (18.14) se encuentra implícita en el hecho de que X_t es estrictamente exógena, pero no está implícita en la condición de que X_t sea exógena (pasada y presente). Por lo tanto, los supuestos de mínimos cuadrados para la estimación del modelo de retardos distribuidos de la Ecuación (18.13) se cumplen si X_t es estrictamente exógena, pero no es suficiente con que X_t sea exógena (pasada y presente).

Estimación MCO del modelo ARD

La primera estrategia para estimar los multiplicadores dinámicos consiste en utilizar MCO para los coe- ficientes del modelo ARD de la Ecuación (18.11). La deducción que conduce a dicha ecuación demuestra que la inclusión del retardo de Y_t y un retardo adicional de X_t como regresores hace que el término de error esté serialmente incorrelacionado, bajo el supuesto de que el error sigue un proceso autorregresivo de primer orden. Por lo tanto, no es necesario usar errores estándar HAC; se pueden emplear los errores estándar MCO habituales.

Aunque los coeficientes ARD estimados no son directamente estimaciones de los multiplicadores dinámicos, estos pueden calcularse a partir de dichos coeficientes. Para ello, se expresa la función de regresión estimada en términos de los valores actuales y pasados de X_t, eliminando Y_t mediante sustituciones sucesivas. La función de regresión estimada es:

$$\hat{Y}_t = \phi_1 \hat{Y_{t-1}} + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2},$$

donde el término independiente ha sido omitido, ya que no forma parte de los multiplicadores dinámicos. Retardando esta ecuación, se obtiene:

$$\hat{Y}_{t-1} = \phi_1 \hat{Y}_{t-2} + \delta_0 X_{t-1} + \delta_1 X_{t-2} + \delta_2 X_{t-3}$$

Sustituyendo $\hat{Y}_{t-1}$ en la ecuación original y agrupando términos, se llega a:

$$\hat{Y}_t = \delta_0 X_t + (\delta_1 + \phi_1 \delta_0) X_{t-1} + (\delta_2 + \phi_1 \delta_1) X_{t-2} + \phi_1 \delta_2 X_{t-3} + \phi_1^2 \delta_2 X_{t-4} + \ldots$$

Repitiendo el proceso de sustitución para los valores retardados de Y_t, se obtiene:

$$\hat{Y}_t = \delta_0 X_t + (\delta_1 + \phi_1 \delta_0) X_{t-1} + (\delta_2 + \phi_1 \delta_1 + \phi_1^2 \delta_0) X_{t-2} + \ldots$$

En esta forma, los coeficientes de $X_t, X_{t-1}, X_{t-2}, \ldots$ son los multiplicadores dinámicos estimados. Sin embargo, si se cumplieran exactamente las restricciones sobre los coeficientes de la Ecuación (18.12), todos los multiplicadores a partir del segundo serían iguales a cero. En la práctica, estas restricciones no se cumplen de forma exacta, por lo que los multiplicadores estimados a partir del segundo suelen ser distintos de cero.

Estimación MCG

La segunda estrategia para estimar los multiplicadores dinámicos cuando Xt es estrictamente exógena consiste en utilizar mínimos cuadrados generalizados (MCG), lo que implica la estimación de la Ecuación (18.13).

MCG Infactible

Suponiendo que ϕ_1 es conocido, las variables cuasi diferenciadas $\tilde{X}_t$ e $\tilde{Y}_t$ pueden calcularse directamente. Si X_t es estrictamente exógena, entonces $E(\tilde{u}_t \mid \tilde{X}_t, \tilde{X}_{t-1}, ...) = 0$. Por lo tanto, los coeficientes a_0, b_1, y b_2 de la Ecuación (18.13) se pueden estimar mediante la regresión MCO de $\tilde{Y}_t$ sobre $\tilde{X}_t$ y $\tilde{X}_{t-1}$ (incluyendo un término independiente). Estos estimadores MCO forman parte del estimador MCG infactible, el cual no es factible en la práctica porque ϕ_1 es desconocido.

MCG Factible

El estimador MCG factible utiliza un estimador preliminar de ϕ_1, denotado como $\tilde{\phi}_1$, para calcular las cuasi diferencias. Los estimadores MCG factibles de b_1 y b_2 son los estimadores MCO de la Ecuación (18.13), calculados mediante la regresión de $\tilde{Y}_t$ sobre $\tilde{X}_t$ y $\tilde{X}_{t-1}$ (con un término independiente), donde:

$$\tilde{X}_t = X_t - \hat{\phi}_1 X_{t-1} \quad \text{y} \quad \tilde{Y}_t = Y_t - \hat{\phi}_1 Y_{t-1}.$$

El estimador preliminar $\hat{\phi}_1$ se obtiene estimando primero la regresión de retardos distribuidos de la Ecuación (18.7) por MCO y luego utilizando MCO para estimar ϕ_1 en la Ecuación (18.8) con los residuos MCO $\hat{u}_t$. Este método se conoce como el estimador de Cochrane-Orcutt (1949).

Estimador Iterado de Cochrane-Orcutt

Una extensión del método de Cochrane-Orcutt es el proceso iterativo, donde se utilizan los estimadores MCG de b_1 y b_2 para calcular nuevos residuos, re-estimar ϕ_1, y repetir el proceso hasta que los estimadores converjan. Este enfoque se denomina estimador iterado de Cochrane-Orcutt.

Interpretación de Mínimos Cuadrados No Lineales (MCNL)

El estimador MCG también puede interpretarse como un estimador de mínimos cuadrados no lineales (MCNL) que impone restricciones no lineales sobre los parámetros del modelo ARD (18.11). Estas restricciones son funciones no lineales de b_0, b_1, b_2, y ϕ_1, por lo que la estimación no puede realizarse mediante MCO. Sin embargo, el estimador MCNL puede calcularse utilizando el algoritmo de Cochrane-Orcutt iterado.

Eficiencia de MCG

El estimador MCG es eficiente entre los estimadores lineales cuando X es estrictamente exógena y los errores transformados $\tilde{u}_t$ son homocedásticos. En muestras grandes, el estimador MCG factible tiene la misma varianza que el estimador MCG infactible, lo que lo convierte en el estimador lineal insesgado óptimo (ELIO). Además, el estimador MCG es más eficiente que el estimador MCO para los coeficientes de retardos distribuidos cuando X es estrictamente exógena.

Generalización de MCG

Los estimadores de Cochrane-Orcutt y Cochrane-Orcutt iterados son casos particulares de la estimación MCG. En general, la estimación MCG implica transformar el modelo de regresión para que los errores sean homocedásticos y serialmente incorrelacionados, y luego estimar los coeficientes del modelo transformado por MCO. El estimador MCG es consistente y ELIO en muestras grandes si X es estrictamente exógena, pero no lo es si X es solo exógena (pasada y presente).

El modelo de retardos distribuidos con retardos adicionales y errores AR(p)

El análisis del modelo de retardos distribuidos con un único retardo y un término de error AR(1) puede extenderse al caso general con varios retardos y un término de error AR(p).

Modelo General de Retardos Distribuidos con Errores AR(p)

El modelo general de retardos distribuidos con r retardos y un término de error AR(p) está dado por:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r} + u_t,$$

donde el término de error u_t sigue un proceso AR(p):

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \tilde{u}_t,$$

y $\tilde{u}_t$ es serialmente incorrelacionado. Este modelo puede reescribirse en forma de cuasi diferencias como:

$$\tilde{Y}_t = \alpha_0 + \beta_1 \tilde{X}_t + \beta_2 \tilde{X}_{t-1} + \cdots + \beta_{r+1} \tilde{X}_{t-r} + \tilde{u}_t, \tag{18.15}$$

Donde:

$$\tilde{Y}_t = Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} \text{ y } \tilde{X}_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}.$$

También puede escribirse de forma equivalente como ARD:

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \tilde{u}_t \tag{18.16}$$

Concepto clave 15.4: Estimación de multiplicadores dinámicos con exogeneidad estricta

El modelo general de retardos distribuidos con r retardos y término de error AR(p) es:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r} + u_t \tag{18.17}$$
$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \tilde{u}_t. \tag{18.18}$$

Si X_t es estrictamente exógena, entonces los multiplicadores dinámicos $\beta_1, ..., \beta_{r+1}$ se pueden estimar utilizando en primer lugar MCO para estimar los coeficientes del modelo ARD:

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \tilde{u}_t,$$

donde $q = r + p$ y posteriormente calculando los multiplicadores dinámicos utilizando el software de regresión. Por otra parte, los multiplicadores dinámicos se pueden estimar mediante la estimación de los coeficientes de los retardos distribuidos de la Ecuación (18.17) por MCG.

Condiciones para la Estimación de los Coeficientes ARD

Para estimar consistentemente los coeficientes ARD, se requiere que se cumpla la condición de media condicional igual a cero:

$$E(\tilde{u}_t | \tilde{X}_t, \tilde{X}_{t-1}, \ldots) = 0.$$

Esta condición implica que X_t debe ser estrictamente exógena, es decir:

$$E(u_t | X_{t+p}, X_{t+p-1}, X_{t+p-2}, \ldots) = 0.$$

Estimación del Modelo ARD mediante MCO

Los multiplicadores dinámicos pueden estimarse a partir de los coeficientes ARD de la Ecuación (18.16) utilizando MCO. Sin embargo, las fórmulas generales son más complejas que en el caso AR(1) y suelen implementarse en software especializado.

Estimación mediante MCG

Alternativamente, los multiplicadores dinámicos pueden estimarse mediante MCG (factibles), que implica estimar los coeficientes de la especificación en cuasi diferencias (Ecuación 18.16) utilizando estimaciones preliminares de $\phi_1, ..., \phi_p$. El estimador MCG es asintóticamente ELIO (Estimador Lineal Insesgado Óptimo) en muestras grandes.

¿MCO o MCG?

Ambos métodos tienen ventajas e inconvenientes:

- **Ventaja de MCO (ARD):** Proporciona una representación compacta de la distribución de retar- dos, reduciendo el número de parámetros a estimar. Esto es útil cuando la distribución de retardos es larga y compleja.
- **Ventaja de MCG:** Es más eficiente que MCO para una longitud de retardos dada r, al menos en muestras grandes.

En la práctica, la elección entre MCO y MCG depende de la estructura del modelo y de la disponibilidad de datos.