Unidad II

Análisis del modelo clásico de regresión lineal general

Dr. Roger Alejandro Banegas Rivero, *Ph.D.*

Regresión múltiple y el término constante

- El Modelo clásico de regresión lineal múltiple se escribe:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{kt} + u_t \quad , t=1,2,\ldots,T$$

- ¿Dónde está x_1? Es el término constante. De hecho está representado por una constante de valor 1 sujeto al número de observaciones T:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

β_1 el el coeficiente de término constante.

Diferentes formas de expresar el MCRLG

- Se puede reescribir cada ecuación individual acorde al período *t*:

$$y_1 = \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + u_1$$
$$y_2 = \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + u_2$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$
$$y_T = \beta_1 + \beta_2 x_{2T} + \beta_3 x_{3T} + \ldots + \beta_k x_{kT} + u_T$$

- Se reestable en expresión matricial

$$y = X\beta + u$$

Donde : y es $T \times 1$
X es $T \times k$
β es $k \times 1$
u es $T \times 1$

Dentro de las matrices del MCRLG

- si *k* es 2, se tienen 2 regresores, una de las columnas es uno:

$$\underset{T\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}} = \underset{T\times 2}{\begin{bmatrix} 1 & x_{21} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{2T} \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{T\times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}}$$

- Las matrices hacen más fácil el análisis.

¿Cómo se estiman los parámetros (los β) en un caso generalizado?

- De forma previa, se toman los residuos cuadrados, para luego minimizarlos.
- En la notación matricial, se tiene:

$$\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix}$$

- La SRC está dada por:

$$\hat{u}'\hat{u} = \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & \dots & \hat{u}_T \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \\ \hat{u}_T \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_T^2 = \sum \hat{u}_t^2$$

Estimadores de MCO para el MCRLG

- Para obtener los parámetros estimados β_1, β_2,..., β_k, se deben minimizar los errores con relación a los parámetrosβs.

- Para ello, se puede demostrar que:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1} X' y$$

Calculando los errores estándares en el MCRLG

- Verificar la dimensión $\hat{\beta}$ es $k \times 1$ como requerimiento.
- ¿Cómo se calculan los errores estándares de lo parámetros?
- Previamente, se estima la varianza de los errores, σ^2, se usa $s^2 = \frac{\sum \hat{u}_t^2}{T-2}$.
- Ahora con notación matricial, se emplea : $s^2 = \frac{\hat{u}'\hat{u}}{T-k}$
- Donde k es el número de regresores. Se puede provar que los parámetros estimados $\hat{\beta}$ está dado por los elementos diagnoales de $s^2(X'X)^{-1}$, así que la varaizanza del primer elemento, $\hat{\beta}_1$,is the first element, la varianza del segundo elemento,$\hat{\beta}_2$, y …, y la varianza de $\hat{\beta}_k$ es el k^{th} elemento diagonal.

Calculando parámetros y errores estándares para un MCRLM: un ejemplo

- Ejemplo: un modelo con k=3 se estima a partir de 15 observaciones.

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Los siguientes datos se han calculado a partir de las X's.

$$(X'X)^{-1} = \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}, (X'y) = \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}, \hat{u}'\hat{u} = 10.96$$

Se pide calcular los coeficientes y los errores estándares.

- Para calcular los coeficientes, simplemente se múltiplica. $(X'X)^{-1}X'y$
- Para calcular los errores estándares, se necesita la varianza σ^2.

$$s^2 = \frac{SRC}{T-k} = \frac{10.96}{15-3} = 0.91$$

Calculando parámetros y errores estándares para un MCRLM: un ejemplo

- La matriz varianza- covarianza de $\hat{\beta}$ está dada por:

$$s^2(X'X)^{-1} = 0.91(X'X)^{-1} = \begin{bmatrix} 1.83 & 3.20 & -0.91 \\ 3.20 & 0.91 & 5.94 \\ -0.91 & 5.94 & 3.93 \end{bmatrix}$$

- Las varianzas se encuentran en la diagonal principal:

$$Var(\hat{\beta}_1) = 1.83 \qquad EE(\hat{\beta}_1) = 1.35$$
$$Var(\hat{\beta}_2) = 0.91 \Leftrightarrow \; EE(\hat{\beta}_2) = 0.96$$
$$Var(\hat{\beta}_3) = 3.93 \qquad EE(\hat{\beta}_3) = 1.98$$

- Se puede reescribir: $\hat{y} = 1.10 - 4.40x_{2t} + 19.88x_{3t}$

$$(1.35) \quad (0.96) \quad (1.98)$$

Pruebas de hipótesis múltiples: la prueba *F*

- Se usan pruebas *t* para evaluar hipótesis simples: un solo coeficiente. ¿Qué sucede si queres evaluar simultáneamente más de un coeficiente a la vez?

- Se emplea la prueba *F* que incluye la estimación de dos regresiones

- La regresión irrestricta es aquella donde los coeficientes están determinados por los datos

- La regresión restricta es aquella donde los coeficientes se encuentran restringidos: ciertas restricciones son impuestras sobre algunos βs.

La prueba *F*: Regresiones restrictas e irrestrictas

- Ejemplo

  La regresión general es:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t \qquad (1)$$

- Se quiere probar la restricción qeu $\beta_3 + \beta_4 = 1$ (se tiene alguna información con base en teoría que indica la hipótesis). La regresión irrestricta es la ecuación (1) arriba, ¿cuál es la ecuación restricta?

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t \quad s.a.\ \beta_3 + \beta_4 = 1$$

- Se sustituye la restricción ($\beta_3 + \beta_4 = 1$) dentro de la regresión que se incluyen sobre los datos:

$$\beta_3 + \beta_4 = 1 \Rightarrow \beta_4 = 1 - \beta_3$$

La prueba *F*-: estimando la regresión restricta

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + (1-\beta_3) x_{4t} + u_t$$
$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + x_{4t} - \beta_3 x_{4t} + u_t$$

- Se factorizarn los términos de β y reorganizar.

$$(y_t - x_{4t}) = \beta_1 + \beta_2 x_{2t} + \beta_3 (x_{3t} - x_{4t}) + u_t$$

- Es la llamada regresión restricta. Se crean dos nuevas variables, por ejemplo, P_t y Q_t.

$$P_t = y_t - x_{4t}$$
$$Q_t = x_{3t} - x_{4t}$$

Así que:

$P_t = \beta_1 + \beta_2 x_{2t} + \beta_3 Q_t + u_t$ es el modelo restricto sobre el que se realiza la estimación.

Calculado el estadístico y la prueba F

- El estadístico esta dado por:

$$Estadístico = \frac{SRC\ 2 - SRC\ 1}{SRC\ 1} \times \frac{T - k}{m}$$

where $SRC\ 1$ = SRC de la regresión irrestrictra

$SRC\ 2$ = SRC de la regresión restricta.

m = número de restricciones

T = número de observaciones

k = número de regresores en la regresión irrestricta incluyendo una constante (o el número total de parámetros a estimarse de forma irrestricta).

La distribución *F*

- La prueba sigue la distribución *F*, el cual tiene 2 parametrós de G.L.
- El valor de los parámetros de GL son *m* y (*T*-*k*) de firna respectiva (el orden de los G.L. es importante).
- Los valores críticos apropiados estarán en columna *m* y fila (*T*-*k*).
- La distribución *F* sólo tiene valores positivos y no es simétrica. En consecuencia, se rechaza Ho si los valores del estadístico F > valores críticos *F*.

Determinando el número de restricciones en una prueba *F*

- Ejemplos :

H_0: hipótesis	No. De restricciones, *m*
$\beta_1 + \beta_2 = 2$	1
$\beta_2 = 1$ y $\beta_3 = -1$	2
$\beta_2 = 0$, $\beta_3 = 0$ y $\beta_4 = 0$	3

- Si el modelo es $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$,

  Luego, las hipótesis nulas

  H_0: $\beta_2 = 0$, y $\beta_3 = 0$ y $\beta_4 = 0$ se prueban por la regresión F. Se prueba la hipótesis nulas que todos los coeficiente, excepto el intercepto, son cero.

- La forma de la hipótesis alternativas es H_1: $\beta_2 \neq 0$, o $\beta_3 \neq 0$ o $\beta_4 \neq 0$

Que no se puede probar con las pruebas *F* o *t*

- No se pueden probar hipótesis multiplicativas o no lineales:

$H_0: \beta_2 \beta_3 = 2$ o $H_0: \beta_2^{\,2} = 1$

La relación en distribuciones *t* y *F*

- Cualquier hipótesis que puede probarse con la prueba *t* , se puede probar al emplear la prueba *F* pero no en sentido contrario.

  Por ejemplo, considere la siguiente hipótesis

  H_0: $\beta_2 = 0.5$

  H_1: $\beta_2 \neq 0.5$

  Se puede emplear la prueba *t* tradicional: $Estad. = \dfrac{\hat{\beta}_2 - 0.5}{EE(\hat{\beta}_2)}$

  Puede probarse también con la prueba *F*.

- La distribución *t* brinda los mismos resultados, debido a que es una distribución especial de la distribución *F*.
- Por ejemplo, si se tiene una variable aleatoria *Z*, y $Z \sim t\,(T\text{-}k)$ también $Z^2 \sim F(1,T\text{-}k)$

Ejemplo de la prueba *F*

- Pregunta: Suponga que un investigador desea investigar si el retorno accionario de una compañia (y) muestra sensibilidad individual unitaria a dos factores (factor x_2 and factor x_3) dentro de tres considerados. La regresión se estima a partir de 144 observaciones mensuales. La regresión es: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$

  - ¿Cuáles es la regresión restricta e irrestricta?

  - Si las dos SRC son 436.1 y 397.2 de forma respectiva, aplique la prueba.

- Solución:

  Sensibilidad unitaria implica que H_0: $\beta_2=1$ and $\beta_3=1$. La regresión restricta es la primera pregunta. La regresión restricta es: $(y_t - x_{2t} - x_{3t}) = \beta_1 + \beta_4 x_{4t} + u_t$ *o*
  $z_t = y_t - x_{2t} - x_{3t}$, la ecuación restricta es: $z_t = \beta_1 + \beta_4 x_{4t} + u_t$
  Para la prueba *F*, $T=144$, $k=4$, $m=2$, *SRCRes*=436.1, *SRCIrr*=397.2
  Estadístico $F = 6.68$. Valor crítico $F(2,140) = 3.07$ (5%) and 4.79 (1%).
  Conclusión: se rechaza H_0.

Minería de datos

- Minería de datos consiste en buscar varias relaciones estadísticas entre variables sin justificativo teórico.
- Por ejemplo, supóngase que se genera una variable dependiente y veinte variables explicativas independientes unas de otras.
- Si se regresa la variable dependiente en función de cada variable explicativa, de forma separada, en promedio el parámetro pendiente será significativo al nivel del 5%.
- Si la minería de datos ocurre, la verdadera significancia será mayor que la significancia nominal.

Estadísiticos de Bondad de ajuste

- Se desea una medición sobre la bondad de ajuste del modelo a los datos.
- El coeficiente de determinación R^2 es el más común. Una manera de definir R^2 es decir la correlación al cuadrado entre y y $\hat{y}$.
- De forma alternativa, se requiere explicar la varianza de la variable endógena, $\bar{y}$, i.e. Por la suma total de cuadrados , STC o *TSS*:

$$STC = \sum_t (y_t - \bar{y})^2$$

- Se puede dividir la *STC* en dos partes: 1) la explicada por el modelo (conocido como suma de cuadrados explicada, *SEC*) y la parte que no explica el modelo (*SRC*).

Definiendo R^2

- Esto es, $STC = SEC + SRC$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2$$

- La bondad de ajuste es:

$$R^2 = \frac{SEC}{STC}$$

- Pero como $STC = SEC + SRC$, se puede escribir:

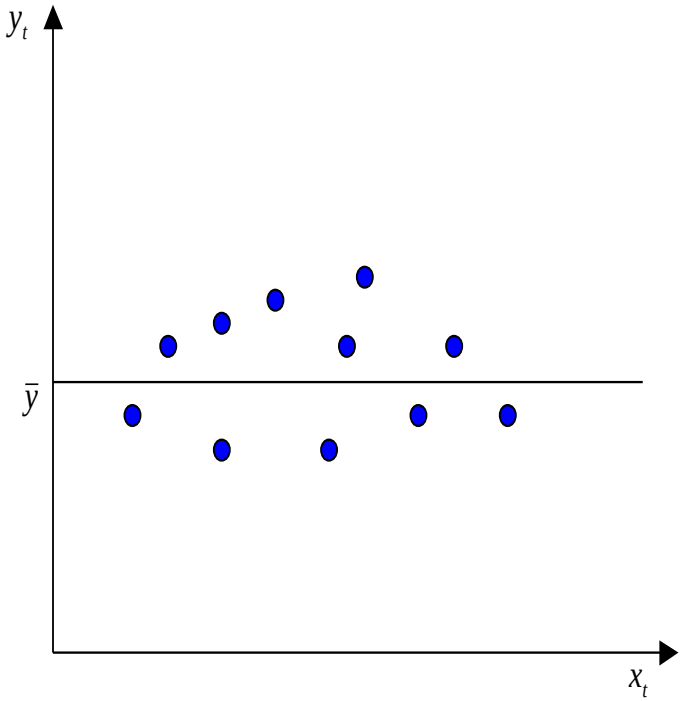
$$R^2 = \frac{SEC}{STC} = \frac{STC - SRC}{STC} = 1 - \frac{SRC}{STC}$$

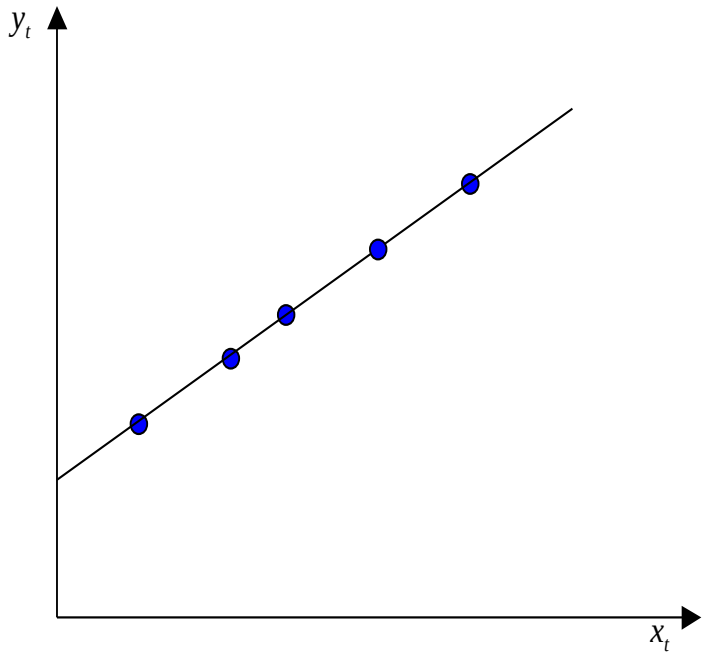
- R^2 siempre estará entre 0 y 1. Considere los dos extremos:

  $SRC = STC$ i.e. $SEC = 0$ así que $R^2 = SEC/STC = 0$

  $SEC = STC$ i.e. $SRC = 0$ así que $R^2 = SEC/STC = 1$

Casos y el límite: $R^2 = 0$ y $R^2 = 1$





Problemas con el R^2 como medida de bondad de ajuste

- Hay un número de problemas:

1. R^2 está definido en términos de la media de y así que el modelo está reparamétrizado (reacomodado) y la variable dependiente siempre cambiará, R^2 cambiará.

2. R^2 nunca cae si se añaden más regresores:

Regresión1: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

Regresión 2: $y = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$

R^2 siempre será más alto en la regresión 2 relativa a la regresión1.

3. R^2 es sospecho para series de tiempo mayores a 0.9 en modelos de series de tiempo: puede existir problemas de endogenidad, simultaneidad o operaciones de suma y resta.

R^2 Ajustado

- Para tratar los poblemas previos, se utiliza el R2 ajustado. Se conoce como $\overline{R}^2$ o $R^{2:}$ ajustado.

$$\overline{R}^2 = 1 - \left[\frac{T-1}{T-k} (1 - R^2) \right]$$

- Si se añade un regresor extra, k se incrementa y al menos que R^2 se incremente en la misma proporción, $\overline{R}^2$ caerá.

- Siguen existiendo algunos problemas con este criterio:
  1. Una regla suave.
  2. No hay distribuciones para $\overline{R}^2$ o $R^{2.}$

Una regresión de ejemplo: Modelo Hedonic para valoración inmobiliaria

- Modelos Hedonic se utilizan para valuar activos reales, especialmente en el sector inmobiliario.
- Des Rosiers y Thérialt (1996) consideraron el efecto de varios factores determinantes para la valoración inmobiliaria de la renta de casas y departamentos (alquiler) en 5 sub-mercados de Quebec, Canada.
- El valor de la renta mensual estaba medida en dólares canadienses (la variable dependiente) es una función de 9 a 12 variables exógenas (depende del aréa de consideración). El documento emplea datos de1990 para la ciudad de Quebec, hay 13,378 observaciones, y las 12 variables explicativos son:

  LnAGE - log de la aparante edad del inmueble.

  NBROOMS - número de habitaciones.

  AREABYRM - área por habitación (en metros cuadrados)

  ELEVATOR -una variable dicotómica = 1 si tiene elevador; 0 en otro caso.

  BASEMENT - una variable dicotómica = 1 si la unidad está localizada en la base; 0 en otro caso.

Modelo Hedonic para valuación de rentas inmobiliarias
Definiciones de variables

OUTPARK - número de espacios de parqueo externo.

INDPARK - número de espacios de parqueo interno.

NOLEASE - una variable dicotómica = 1 si la unidad no tiene contrato de arrendamiento; 0 en otro caso.

LnDISTCBD - log de distancia al centro de la ciudad.

SINGLPAR - porcentaje de familias monoparientales en el edificio.

DSHOPCNTR- distancia en Km del centro comercial más próximo.

VACDIFF1 - diferencia de vacancia entre lo ocupado y lo rentado.

- Examine los signos y los coefcientes.
  - Los coeficientes estimados muestran los efectos individuales, ceteris paribus, de las variables exógenas sobre el renta mensual inmobiliaria.

Resultados del Modelo Hedonic

Variable dependiente: Renta mensual inmobilliliaria en dólares canadienses.

Variable	Coefficient	*t*-ratio	*A priori* sign expected
Intercept	282.21	56.09	+
LnAGE	-53.10	-59.71	-
NBROOMS	48.47	104.81	+
AREABYRM	3.97	29.99	+
ELEVATOR	88.51	45.04	+
BASEMENT	-15.90	-11.32	-
OUTPARK	7.17	7.07	+
INDPARK	73.76	31.25	+
NOLEASE	-16.99	-7.62	-
LnDISTCBD	5.84	4.60	-
SINGLPAR	-4.27	-38.88	-
DSHOPCNTR	-10.04	-5.97	-
VACDIFF1	0.29	5.98	-

Notes: Adjusted R^2 = 0.65l; regression F-statistic = 2082.27. Source: Des Rosiers and Thérialt
(1996). Reprinted with permission of the American Real Estate Society.

Pruebas de hipótesis no anidadas

- Un contexto de hipótesis anidadas.
- ¿Qué sucede si se quieren comparar los siguientes modelos?

$$\text{Model 1: } y_t = \alpha_1 + \alpha_2 x_{2t} + u_t$$

$$\text{Model 2: } y_t = \beta_1 + \beta_2 x_{3t} + v_t$$

- ¿Se emplea el R^2 o R^2 ajustado? Pero que sucedería si el número de variables explicativas fuera distinto.
- Un enfoque alternativo es un modelo híbrido:

$$\text{Model 3: } y_t = \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{3t} + w_t$$

Pruebas de hipótesis no anidadas (cont')

- Hay cuatro posibles resultados en el modelo 3:
  - γ_2 es significativo pero γ_3 no es.
  - γ_3 es significativo pero γ_2 no es.
  - γ_2 y γ_3 son ambas estadísticamente significativas.
  - Ni γ_2 ni γ_3 son significativas.

- Problemas con modelos híbridos.
  - Puede carecer de sentido.
  - Posible alta correlación entre x_2 y x_3.

Biblográfía

Brooks, Chr. (2008) Chapter 3: Further development and analysis of the classical linear regression model in *Introductory Econometrics for Finance*, Cambrigdge University Press.