Clase 2 – Curso series de tiempo en R - CEPAL

Modelo lineal y MCO

Contenidos de la clase

- **Modelo de regresión lineal**
- **Estimación por MCO**
- **Análisis del resultado de la regresión**
- **Métricas para la selección de modelos**
- **Ejercitación en R con datos de inflación**

Definición del modelo lineal

En el modelo de regresión más simple se establece una relación lineal entre dos variables, con la cual se desea explicar cómo varía una de ellas ($\boldsymbol{y}$) en función de la variación de la otra ($\boldsymbol{x}$).

$$y = \beta_0 + \beta_1 x + u$$

Donde:

$\boldsymbol{y}$ **:** Variable explicada o dependiente.

$\boldsymbol{x}$: Variable explicativa o independiente.

$\boldsymbol{u}$: Término de error (factores distintos a x que a afectan a y, y que son no observables).

$\boldsymbol{\beta_0}$: constante o intercepto.

$\boldsymbol{\beta_1}$: parámetro de la pendiente en la relación entre y y x, cuando todos los demás factores en u permanecen constantes.

Interpretación del modelo lineal

Si los demás factores en $\boldsymbol{u}$ permanecen constantes, de manera que el cambio en $\boldsymbol{u}$ sea cero, entonces $\boldsymbol{x}$ tiene un efecto lineal sobre $\boldsymbol{y}$:

$$\Delta y = \beta_1 \Delta \mathrm{x} \text{ si } \Delta \mathrm{u} = 0$$

Por tanto, **el cambio en $\boldsymbol{y}$ es $\boldsymbol{\beta_1}$ multiplicado por el cambio en $\boldsymbol{x}$**. Esto significa que β_1 es el parámetro de la pendiente en la relación entre y y x, cuando todos los demás factores en $\boldsymbol{u}$ permanecen constantes. **Este es el parámetro de interés primordial.**

La linealidad de la ecuación implica que todo cambio de x tiene siempre el mismo efecto sobre y, sin importar el valor inicial de x. En muchas aplicaciones de la economía esto no es realista. Por este motivo hay estrategias para abordar estas situaciones, que incluyen las interacciones y las estimaciones por umbrales.

Supuesto para asumir relación *ceteris paribus*: media condicional cero

Un supuesto de vital importancia para poder formular conclusiones *ceteris paribus* acerca de cómo afecta x a y (en una muestra aleatoria), es el de **media condicional cero**.

$$E(u|x) = 0$$

El supuesto de media condicional cero proporciona otra interpretación de β_1 :

$$E(y|x) = \beta_0 + \beta_1 x$$

La ecuación muestra que en la función de regresión poblacional (FRP), $E(y|x)$ es una función lineal de x. La linealidad significa que por cada aumento de una unidad en x el valor esperado de y se modifica en la cantidad β_1. Dado cualquier valor de x, la distribución de y está centrado en $E(y|x)$.

Importante: Esta ecuación dice cómo varía el valor promedio de y de acuerdo con la variación de x, y no dice que y sea igual a $\beta_0 + \beta_1 x$ para cada valor de la población.

Condiciones para asumir media condicional cero

1. La media del error en la población es cero. Esto se obtiene directamente de incluir el intercepto en la regresión, y no requiere de asumir una determinada relación entre las variables.

$$E(u) = 0$$

2. El valor esperado de u no depende del valor de x. Este indica que el valor promedio de los factores no observables es el mismo en todas las fracciones de la población determinados por los valores de x y que este promedio común es necesariamente igual al promedio de u en toda la población.

$$E(u|x) = E(u)$$

Estimación por Mínimos Cuadrados Ordinarios (MCO)

Para la estimación del intercepto y la pendiente de la función de regresión poblacional, el método de mínimos cuadrados ordinarios parte de los dos supuestos anteriores (condiciones de primer orden):

1. La esperanza del término de error es igual a cero

$$E(u) = 0$$

2. El término de error no está correlacionado con x en la población (lo que implica que la covarianza entre ellas es igual a cero).

$$Cov(x, u) = E(xu) = 0$$

Partiendo de estos supuestos es posible derivar estimadores de los parámetros poblacionales que minimicen la suma de residuales cuadrados.

Estimación por Mínimos Cuadrados Ordinarios (MCO)

A los fines de simplificar la interpretación, presentamos los estimadores para un modelo univariado:

Función de regresión poblacional (FRP) $$E(y|x) = \beta_0 + \beta_1 x$$

Función de regresión muestral (FRM) $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

<u>Parámetros de la FRM:</u>

Pendiente de la FRM $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Covarianza entre x e $\boldsymbol{y}$

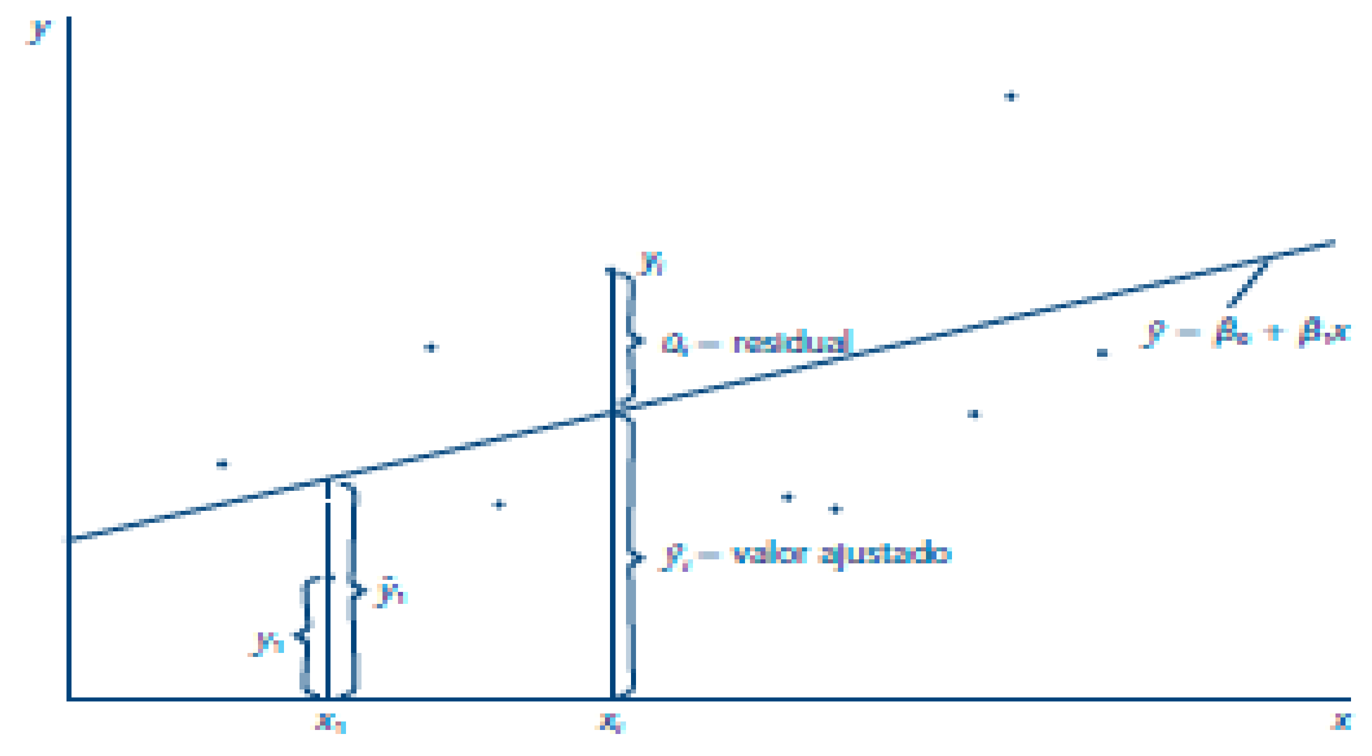
Varianza de x

Intercepto de la FRM $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimación por Mínimos Cuadrados Ordinarios (MCO)

FIGURA 2.4

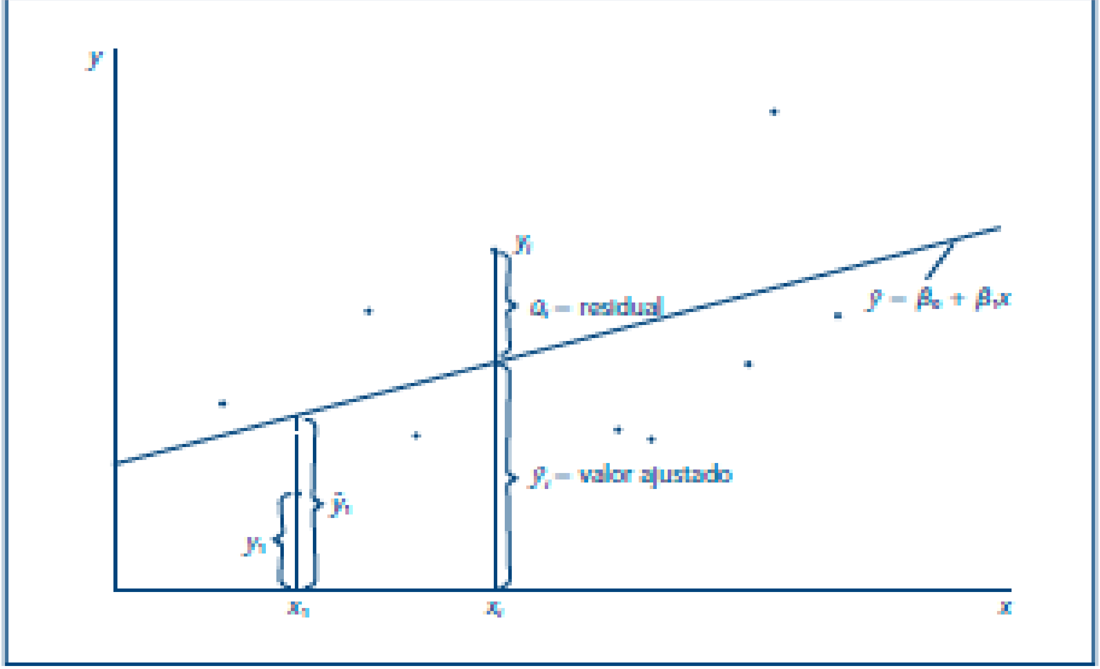
Valores ajustados y residuales.



Estimación por Mínimos Cuadrados Ordinarios (MCO)

FIGURA 2.4

Valores ajustados y residuales.



Los parámetros muestrales de MCO son tales que minimizan la suma de residuales cuadrados

$$\sum_{i=1}^{n} \hat{u}_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Esto solo es posible si se asumen dos condiciones de primer orden:

$$n^{-1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

Indicadores obtenidos de la regresión lineal por MCO

```
Call:
tslm(formula = dlIPC ~ dlTCN + l.dlIPC)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045121 -0.010965 -0.002918  0.010882  0.050161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.015942   0.004733   3.368  0.00128 **
dlTCN       0.163711   0.031208   5.246 1.88e-06 ***
l.dlIPC     0.610406   0.076558   7.973 3.60e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.017 on 64 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6734,    Adjusted R-squared:  0.6632
F-statistic: 65.97 on 2 and 64 DF,  p-value: 2.819e-16
```

Parámetros de la FRM

Desvío estándar de los parámetros

Desvío estándar del residuo

P-valor

R-cuadrado

Indicadores obtenidos de la regresión lineal por MCO

```
Call:
tslm(formula = dlIPC ~ dlTCN + l.dlIPC)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045121 -0.010965 -0.002918  0.010882  0.050161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.015942   0.004733   3.368  0.00128 **
dlTCN       0.163711   0.031208   5.246 1.88e-06 ***
l.dlIPC     0.610406   0.076558   7.973 3.60e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.017 on 64 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6734,    Adjusted R-squared:  0.6632
F-statistic: 65.97 on 2 and 64 DF,  p-value: 2.819e-16
```

Parámetros de la FRM

Desvío estándar de los parámetros

Desvío estándar del residuo

P-valor

R-cuadrado

Desvío estándar de los parámetros y del residuo

- **En el error se prefiere menor variabilidad:** cuanto mayor sea la varianza del error ($\hat{\boldsymbol{\sigma}}^2$), mayor es $Var(\hat{\beta}_1)$. Es decir que, una mayor variación en los factores no observables que afectan a y, más difícil estimar β_1 con precisión.
- **En la variable independiente se prefiere mayor variabilidad:** a medida que aumenta la variabilidad de las x_i la varianza de $\hat{\beta}_1$ disminuye. También esto es intuitivamente correcto, ya que entre más dispersa sea la muestra de las variables independientes, más fácil será hallar la relación entre $E(y|x)$ y x. Es decir, será más sencillo estimar β_1. Si hay poca variación en las x_i, entonces puede ser difícil hallar cómo varía $E(y|x)$ con la variación en x. A medida que se incrementa el tamaño de la muestra, también aumenta la variación total de las x_i. Por tanto, un tamaño de muestra mayor da como resultado una varianza menor de las $\hat{\beta}_1$.

$$\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}_1} = \boldsymbol{Var}(\hat{\boldsymbol{\beta}}_1) = \frac{\hat{\boldsymbol{\sigma}}^2}{\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^2} = \hat{\boldsymbol{\sigma}}^2/\boldsymbol{STC}_{\boldsymbol{x}}$$

$$\hat{\boldsymbol{\sigma}}^2 = \frac{\mathbf{1}}{(\boldsymbol{n}-\mathbf{2})}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \hat{\boldsymbol{u}}_i^2 = \boldsymbol{SRC}/(\boldsymbol{n}-\mathbf{2})$$

¿Qué utilidad tiene conocer los desvíos estándar?

1. **Estimar intervalos de confianza:** ejemplo para modelo univariado con un nivel de confianza al 95% (percentil 97,5% de una distribución t de Student).

Intervalo de confianza para los estimadores $\boldsymbol{\widehat{\beta}_1 \pm 1,96\, \hat{\sigma}_{\beta_1}}$

Intervalo de confianza para la variable dependiente $\boldsymbol{\hat{y} \pm 1,96\, \hat{\sigma}}$

"Regla del pulgar": como el estadístico con un nivel de confianza al 95% es 1,96, a veces se suele hacer un cálculo más sencillo que consiste en sumar y restar 2 desvíos estándar a la estimación puntual para obtener los intervalos de confianza.

¿Qué utilidad tiene conocer los desvíos estándar?

2. Test de hipótesis: Los estimadores poblacionales son desconocidos. Aun así, pueden plantearse hipótesis sobre su valor y probarlas a través de la inferencia estadística.

En la mayoría de las aplicaciones, el interés principal reside en probar la hipótesis nula:

$$H_0\colon\ \beta_1 = 0$$

El estadístico que se emplea para probar la hipótesis nula se llama el estadístico t o el coeficiente t de β_j y se define como:

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$$

La regla de rechazo para $H_0\colon\ \beta_j = 0$ es (donde c es un valor crítico elegido de manera aproximada. Para determinar c, se especifica un nivel de significancia, por ejemplo 5%)

$$\left|t_{\beta}\right| > c$$

Indicadores obtenidos de la regresión lineal por MCO

```
Call:
tslm(formula = dlIPC ~ dlTCN + l.dlIPC)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045121 -0.010965 -0.002918  0.010882  0.050161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.015942   0.004733   3.368  0.00128 **
dlTCN       0.163711   0.031208   5.246 1.88e-06 ***
l.dlIPC     0.610406   0.076558   7.973 3.60e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.017 on 64 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6734,    Adjusted R-squared:  0.6632
F-statistic: 65.97 on 2 and 64 DF,  p-value: 2.819e-16
```

Parámetros de la FRM

Desvío estándar de los parámetros

Desvío estándar del residuo

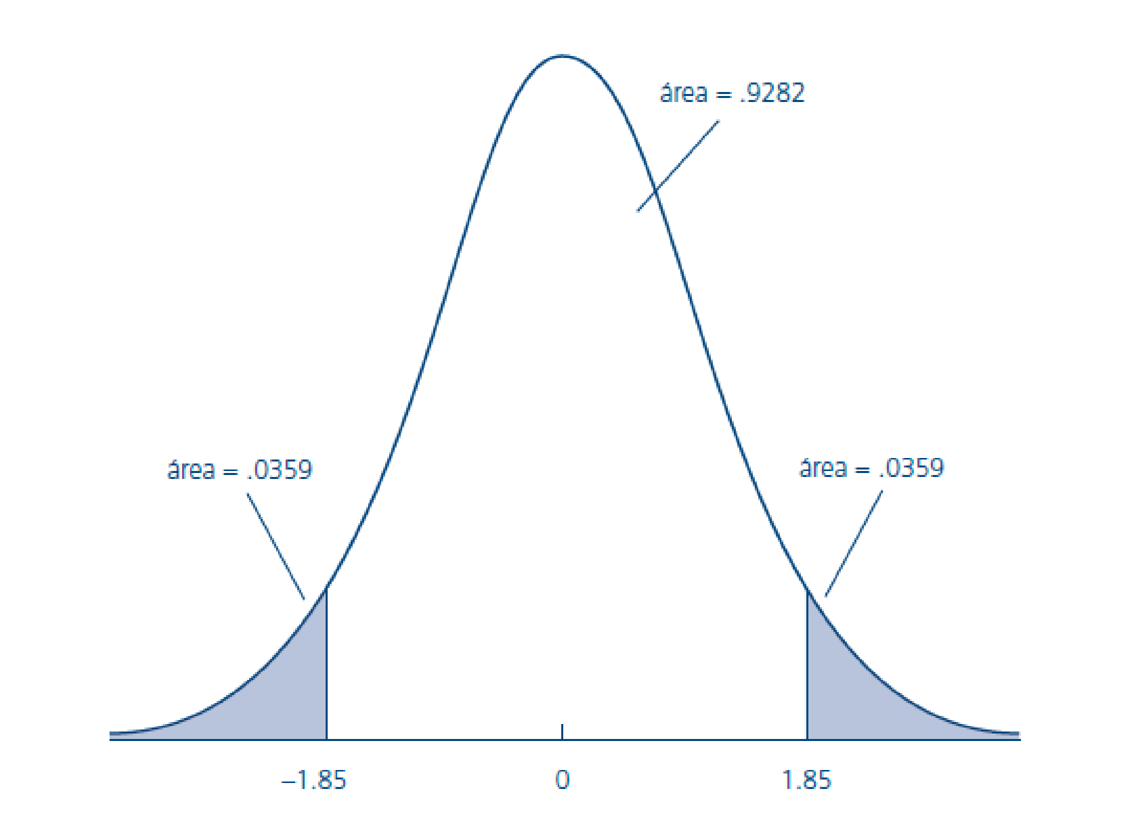
P-valor

R-cuadrado

P-valor

FIGURA 4.6

Obtención del valor-*p* en el caso de una alternativa de dos colas, siendo $t = 1.85$ y $gl = 40$.



- El **procedimiento clásico** consiste en elegir un valor crítico con un nivel de significancia arbitrario y compararlo con el valor del estadístico t. Con eso se determina el rechazo o no de la hipótesis nula.
- Un **procedimiento alternativo** consiste en responder a la siguiente pregunta: dado el valor observado del estadístico t, ¿cuál es el menor nivel de significancia al que se habría rechazado la hipótesis nula? Este nivel se conoce como el valor-p.
- Para obtenerlo se resta 100 al percentil correspondiente al valor t y se lo multiplica por 2 (para el test de dos colas).

Indicadores obtenidos de la regresión lineal por MCO

```
Call:
tslm(formula = dlIPC ~ dlTCN + l.dlIPC)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045121 -0.010965 -0.002918  0.010882  0.050161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.015942   0.004733   3.368  0.00128 **
dlTCN       0.163711   0.031208   5.246 1.88e-06 ***
l.dlIPC     0.610406   0.076558   7.973 3.60e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.017 on 64 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6734,	Adjusted R-squared:  0.6632
F-statistic: 65.97 on 2 and 64 DF,  p-value: 2.819e-16
```

Parámetros de la FRM

Desvío estándar de los parámetros

Desvío estándar del residuo

P-valor

R-cuadrado

R2 cuadrado o coeficiente de determinación

El R2 es el cociente de la variación explicada entre la variación total; por tanto, se interpreta como la proporción de la variación muestral de y que es explicada por x.

Suma total de cuadrados (STC): $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Suma explicada de los cuadrados (SEC): $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Suma residual de los cuadrados (SRC): $\sum_{i=1}^{n} \hat{u}_i^2$

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

Se lo conoce como R cuadrado porque es igual al cuadrado del coeficiente de correlación muestral entre y_i y $\hat{y}_i$, y al coeficiente de correlación se lo suele denotar con la letra R.

Propiedades en muestras finitas de MCO bajo los supuestos clásicos

Consistencia de los parámetros

1. **Linealidad y dependencia débil.** El modelo es lineal en sus parámetros, es estacionario y débilmente dependiente. En particular, la ley de los grandes números y el teorema del límite central pueden aplicarse a los promedios muestrales.
2. **Ausencia de multicolinealidad.** No hay variables independientes que sean iguales entre sí, combinaciones lineales de las demás o iguales a una constante.
3. **Media condicional cero.** Dadas las variables explicativas para todos los periodos, el valor esperado del error es cero. Las variables explicativas no se correlacionan contemporáneamente con el error (tampoco con los valores pasados si la serie es débilmente dependiente). Supone que no hay problema de endogeneidad derivada, por ejemplo, de variables omitidas o error de medición. Para este supuesto es clave la especificación de la forma funcional del modelo.

Propiedades en muestras finitas de MCO bajo los supuestos clásicos

4. Homocedasticidad. Los errores son contemporáneamente homocedásticos

5. Ausencia de correlación serial. Los errores en dos períodos distintos no están correlacionados.

Bajo los supuestos 1 a 5, los estimadores de MCO tienen distribuciones asintóticamente normales. Además, los errores estándar usuales de MCO, los estadísticos t, los estadísticos F y los estadísticos ML son asintóticamente válidos.

Los modelos con variables explicativas con tendencia satisfacen los supuestos 1 a 5, siempre y cuando sean estacionarios con tendencia determinística. Mientras las tendencias determinísticas se incluyan en las ecuaciones cuando sea necesario, los procedimientos de inferencia usuales son asintóticamente válidos.

Métricas para la sección de modelos

Cuando hay muchos predictores posibles, se necesita una estrategia para elegir a los mejores predictores para usar en el modelo de regresión. Una estrategia común es hacer múltiples regresiones lineales sobre todos los predictores y descartar todas las variables cuyos p-valores son mayores a 0,05. Esto no es correcto porque la significatividad estadística no siempre indica valor predictivo, y porque los p-valores no son buenos indicadores cuando dos o más predictores están correlacionados uno con el otro (problema de multicolinealidad).

En su lugar, vamos a usar tres medidas de precisión predictiva:

1. R cuadrado ajustado

2. Criterio de Akaike

3. Criterio de información bayesiano de Schwarz

1. R2 ajustado

- El R2 convencional tiene como debilidad que no permite controlar por grados de libertad, entonces siempre que agreguemos más variables su valor va a aumentar, sin importar que tan relevante sea esa variable.
- Una alternativa pensada para superar estos problemas es el R2 ajustado (también llamado "R-bar-squared")

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T - 1}{T - k - 1}$$

Donde T es la cantidad de observaciones y k es la cantidad de predictores. Esto es una mejora del R2, en la medida que no va a aumentar con la incorporación de nuevos predictores. Usando esta medida, el mejor modelo será aquel con el valor más alto.

2. Criterio de información de Akaike

Otro método es el criterio de información de Akaike, que definimos como:

$$AIC = T \log\left(\frac{SCR}{T}\right) + 2(k + 2)$$

Donde T es la cantidad de observaciones usadas para la estimación y k es la cantidad de predictores en el modelo. Los diferentes paquetes computacionales usan definiciones levemente distintas para AIC, a pesar de que todos deberían llevar a seleccionar el mismo modelo.

La parte de $k + 2$ en la ecuación ocurre porque hay $k + 2$ parámetros en el modelo: los k coeficientes de los predictores, el intercepto y la varianza de los residuos.

La idea es penalizar el ajuste del modelo (SSE) con la cantidad de parámetros que necesitan ser estimados.

3. Criterio de información bayesiano de Schwarz

Una medida relacionada es el criterio de información bayesiano de Schwarz (usualmente abreviado como BIC, SBIC or SC):

$$BIC = T\ log\left(\frac{SCR}{T}\right) + (k+2)\log(T)$$

Al igual que con el AIC, minimizar el BIC tiene como objetivo dar el mejor modelo. El modelo elegido por el BIC es el mismo que el elegido por el AIC o uno con menos términos. Esto es porque el BIC penaliza más la cantidad de parámetros que el AIC.

¿Qué métrica utilizar para la selección de un modelo?

- Si bien **R2 ajustado** se usa ampliamente y ha existido durante más tiempo que las otras medidas, su tendencia a seleccionar demasiadas variables predictoras lo hace menos adecuado para la predicción.
- A muchos estadísticos les gusta usar el **BIC** porque tiene la característica de que si existe un verdadero modelo subyacente, el BIC seleccionará ese modelo si se reciben suficientes datos. Sin embargo, en realidad, rara vez, o nunca, hay un modelo subyacente verdadero, e incluso si hubiera un modelo subyacente verdadero, seleccionar ese modelo no necesariamente dará los mejores pronósticos (porque las estimaciones de los parámetros pueden no ser precisas).
- En consecuencia, se suele recomendar utilizar el **AIC** que tiene como objetivo la predicción.
- Si el valor de T es lo suficientemente grande, todos conducirán al mismo modelo.

En resumen…

- El **método de MCO** es la estrategia más utilizada para estimar los parámetros poblacionales de un modelo lineal. La fórmula de cálculo de los estimadores muestrales se deriva del supuesto de media condicional cero.

- El **test de hipótesis** es un procedimiento que sirve para juzgar si el valor de un estimador muestral es compatible con el valor del parámetro poblacional.

- El **p-valor** es una alternativa al test de hipótesis que tiene como ventaja que brinda más información que el método convencional en el que se elige un nivel de significancia arbitrario. Esta métrica indica cuál es el menor nivel de significancia al que se habría rechazado la hipótesis nula.

- Para la **selección del mejor modelo** entre varias alternativas se debe utilizar una métrica que penalice por la cantidad de parámetros. Algunas opciones son el R2 ajustado, el AIC o el BIC. El R cuadrado convencional no es buen indicador porque siempre aumenta con un mayor número de parámetros, y nos puede llevar a elegir un modelo que sobreajuste a los datos muestrales.

Ejercitación en R

1. Preparar los datos según vimos en la primera clase.
2. Estimar un modelo lineal por MCO.
3. Analizar el resultado de la regresión.
4. Estimar modelos con distintas variables y elegir el que ajusta mejor.
5. Realizar lo mismo incorporando otras variables a la ecuación de inflación.

Bibliografía

- **Enders**, W. (2008). Applied econometric time series. John Wiley & Sons.
- **Hyndman**, R.J., & **Athanasopoulos**, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.
- **Wooldridge**, J. M. (2015). Introductory econometrics: A modern approach. Cengage learning.