Unidad 3. Modelo de RLS: Coeficiente de determinación (R^2)

Erika R. Badillo
erika.badilloen@unaula.edu.co

Facultad de Economía
Universidad Autónoma Latinoamericana

En este tema

- Coeficiente de determinación R^2
- Consideraciones sobre el R^2

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 2
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 3

Coeficiente de determinación R^2

- Intenta responder la siguiente pregunta:

  ¿Cómo evaluar qué tan bien nuestra regresión se ajusta a nuestra muestra?

- Hasta ahora hemos ajustado la línea de regresión muestral a los datos a través de mínimos cuadrados
- Se requiere un indicador que permita medir el grado de ajuste entre el modelo y los datos: R^2
- Por otro lado, en el caso de estimar varios modelos alternativos podría utilizarse medidas de bondad de ajuste para seleccionar el modelo más adecuado
- Aunque existen otras medidas de bondad de ajuste, la más conocida es el coeficiente de determinación (R^2)

Coeficiente de determinación R^2

- Podemos pensar en cada observación como compuesta por una parte explicada y una no explicada
- Construyendo el R^2

$$Y_i = \widehat{Y}_i + \widehat{u}_i$$

Restando a ambos lados $\bar{Y}$ y después elevando al cuadrado

$$(Y_i - \bar{Y})^2 = [(\widehat{Y}_i - \bar{Y}) + \widehat{u}_i]^2$$

Desagregando la anterior ecuación y sumando en i tenemos

$$\sum(Y_i - \bar{Y})^2 = \sum(\widehat{Y}_i - \bar{Y})^2 + \sum \widehat{u}_i^2$$

$\sum(Y_i - \bar{Y})^2 \Longrightarrow$ Suma total de cuadrados (SCT). Medida de la variación muestral total en las Y_i, es decir, mide qué tan dispersos están las Y_i en la muestra

$\sum(\widehat{Y}_i - \bar{Y})^2 \Longrightarrow$ Suma explicada de cuadrados (SEC) o suma de cuadrados del modelo (SCM). Qué parte de la variación total en Y es explicada por X (debida a la regresión)

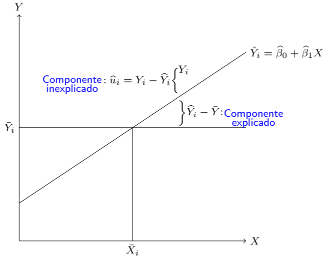
$\sum \widehat{u}_i^2 = \sum(Y_i - \widehat{Y})^2 \Longrightarrow$ Suma de cuadrados de los residuos (SCR). Qué parte de la variación total en Y no es explicada por X ("fuerzas aleatorias")

Así,

$$\text{SCT} = \text{SEC} + \text{SCR}$$

En otras palabras significa que

$$\text{Desviación total} = \text{componente explicado por } X + \text{componente inexplicado}$$



De aquí nace una medida de la proporción de la variación en Y explicada por X o por el modelo (medida de bondad de ajuste del modelo):

$$R^2 = \frac{SEC}{SCT} = 1 - \frac{SCR}{SCT}$$

Coeficiente de determinación R^2

- Es una medida entre 0 y 1

- Entre más cercano sea el R^2 a uno, más se acercará el modelo estimado $\hat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ a Y_i y mayor será la habilidad predictiva (precisión) de nuestro modelo

Si $R^2 = 0 \Longrightarrow$ SEC=0 el modelo no explica nada

Si $R^2 = 1 \Longrightarrow$ SEC=SCT el modelo explica todo

Coeficiente de determinación R^2

Una lectura atenta indicaría que R^2 compara dos modelos

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum \widehat{u}_i^2}{\sum (Y_i - \bar{Y}_i)^2} = 1 - \frac{\sum (Y_i - \widehat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$$

El primer modelo usa $\widehat{Y}_i$ como estimación (el RLS). Frente al que usa $\bar{Y}_i$ como estimación (el ingenuo)

La interpretación alterna indica cuanto se gana al incluir la variable X_i frente a sólo usar el intercepto

Modelo de RLS

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$
$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

Modelo ingenuo

$$Y_i = \beta_0 + u_i$$
$$\widehat{Y}_i = \bar{Y}$$
$$\widehat{u}_i = Y_i - \bar{Y}_i$$

$$R^2 = 1 - \frac{\sum (Y_i - \widehat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$$

Compara la suma de cuadrado de residuos de dos modelos: RLS versus Ingenuo

Consideraciones sobre el R^2

- R^2 en un modelo de RLS sin intercepto
  En este tipo de modelos la ecuación vista del R^2 ya no es apropiada, ya que sin intercepto

$$\sum(Y_i - \bar{Y}_i)^2 \neq \sum(\widehat{Y}_i - \bar{Y}_i)^2 + \sum(\widehat{u}_i)^2$$

$$\text{SCT} \neq \text{SCM} + \text{SCR}$$

$$\text{variación total} \neq \text{variación explicada por la regresión} + \text{variación del error}$$

  En estas circunstancias no tiene sentido hablar de la proporción de variación total que es explicada por la regresión. Dos alternativas son posibles:
  - no reportar el valor del R^2
  - medir la variación alrededor de cero en lugar de la media muestral $\bar{Y}$. Se tendría la siguiente formula para el R^2

$$R_*^2 = 1 - \frac{\sum(\widehat{u}_i)^2}{\sum Y_i^2}$$

Consideraciones sobre el R^2

- El R^2 sirve para elegir entre modelos
  - Ya que el coeficiente de determinación nos informa sobre la proporción de variación en la variable dependiente explicada por las variables explicativas, una forma de escoger entre dos modelos que compiten es elegir aquel modelo con mayor R^2. Este procedimiento nos da el modelo con mayor poder explicatorio
  - Usar el R^2 para comparar modelos es inválido cuando:
    - los modelos tienen diferentes variables dependientes
    - los modelos tienen diferentes número de regresores: el R^2 siempre aumenta cuando se adicionan Xs incluso si el regresor es irrelevante
    - uno de los modelos no tiene un término constante
  - Escoger el modelo con mayor R^2 ajustado, el cual es definido como

    $$\bar{R}^2 = 1 - \frac{SCR/(N-K)}{SCT/(N-1)}$$

    donde K es el número de coeficientes en el modelo

    Este R^2 ajustado no siempre aumenta cuando se adicionan regresores. Este R^2 ajustado se suele utilizar para comparar modelos con diferente número de regresores

Coeficiente de determinación R^2

Ejemplo - Stata

```
reg wage educ
```

Source	SS	df	MS
Model	1179.73204	1	1179.73204
Residual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

Number of obs	=	526
F(1, 524)	=	103.36
Prob > F	=	0.0000
R-squared	=	0.1648
Adj R-squared	=	0.1632
Root MSE	=	3.3784

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

Ejemplo - Stata

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(1, 524)     =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

```
* Calculando R2 = 1 - (SCR/SCT)
* Calculando SCT = SEC + SCR
* e(mss): es la varianza de la SCM, pero la varianza de
* SCM es igual al SCM, ya que Var(SCM) = SCM/gdl, pero gdl=1
* e(rss): es la SCR
scalar SCT = e(mss) + e(rss)
dis SCT
7169.4143

scalar R2 = 1-(e(rss)/SCT)
dis R2
.16475751

* Calculando el R2 ajustado = 1 - (SCR/(N-K)/SCT/(N-1))
scalar R2_ajust = 1 - ((e(rss)/(526-2))/(SCT/(526-1)))
dis R2_ajust
.16316354
```

```
* Qué es Root MSE en la salida de Stata
* Root MSE: desviación estándar del Var(U)
scalar var_U = e(rss)/(526-2)
dis var_U

scalar root_mse = sqrt(var_U)
dis root_mse
3.3783895

twoway (scatter wage educ) ///
       (lfit wage educ), ///
    xtitle("Educ") ytitle("Wage") legend(off)
```