IN4402 - Aplicaciones de Probabilidades y Estadística en Gestión
Profesor: Benjamín Villena
Profesores auxiliares: Ronald Leblebici, Angelo Muñoz

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Clase auxiliar # 3

Coeficiente de bondad de ajuste ($\mathbf{R^2}$): es el porcentaje de varianza que se explica gracias a los regresores. No es un criterio recomendable para elegir los regresores a incorporar en el modelo. $R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2}$. Este coeficiente siempre aumenta a medida que se agregan regresores, por lo que existe una versión "ajustada" que penaliza la incorporación de estos $R_a^2 = 1 - \frac{\text{SCR}}{\text{SCT}} \cdot \frac{N-1}{N-K}$.

Sesgo: sea $\widehat{\beta}$ el estimador del parámetro poblacional β. El sesgo de $\widehat{\beta}$ es la diferencia entre el valor esperado del estimador y el valor efectivo del parámetro. Es decir, $B(\widehat{\beta}) = \mathbb{E}(\widehat{\beta}) - \beta$.

Consistencia: un estimador consistente converge al valor del parámetro poblacional que busca estimar cuando el número de datos de la muestra tiende a infinito. Es decir, $\mathbb{P}(|\widehat{\beta}_N - \beta| > \varepsilon) = 0$.

Linealidad: sea Y un vector de observaciones. Un estimador $\widehat{\beta}$ es lineal en Y si se puede escribir como una combinación lineal de sus componentes. Es decir, $\widehat{\beta} = \sum_n \omega_n Y_n$.

Teorema de Gauss-Markov: si se cumplen los supuestos de MCO, entonces el estimador $\widehat{\beta}_{\text{MCO}}$ es el mejor estimador lineal insesgado (MELI) de los parámetros poblacionales β. ("Mejor" = "de mínima varianza").

Estimador de varianza de estimadores MCO: $\widehat{\mathbb{V}}(\widehat{\beta}_A) = \frac{\mathbb{V}(Y)(1-R^2)}{(N-K)\mathbb{V}(X_A)} \cdot \frac{1}{1-R^2_{A,-A}}$

Descomposición del error cuadrático medio: $ECM(\widehat{\beta}_k) = B(\widehat{\beta}_k)^2 + \mathbb{V}(\widehat{\beta}_k)$

1. Teorema de Gauss-Markov

1. **[P4, Control 1, 2017]** Comente la siguiente afirmación : "Si los supuestos del teorema de Gauss-Markov no se cumplen, entonces el modelo de regresión múltiple no es una buena aproximacion a la expectativa condicional de Y dado X, $\mathbb{E}(Y|X)$".

   Respuesta:

   Verdadero. Los supuestos del teorema de Gauss-Markov garantizan que $\hat{\beta}_{MCO}$ es el estimdor lineal insesgado de mínima varianza (menor incertidumbre). Por lo tanto, la estimación $\hat{Y}|X = X\beta_{MCO}$ representa la predicción más precisa de $E[Y|X]$. Si se viola(n) alguno(s) de los supuestos, se pierde esta propiedad, generando sesgo y/o ineficiencia.

2. **[P4, Control 1, 2014]** Suponga el modelo de regresión lineal simple $Y_n = \beta_1 + \beta_2 X_n + U_n$, y considere el siguiente estimador para el parámetro β_2: $\dot{\beta}_2 = \sum_n Y_n / \sum_n X_n$.

   a. Evalúe si el estimador $\dot{\beta}_2$ es lineal e insesgado. Explicite cualquier supuesto que deba realizar para demostrar o refutar esta propiedad.

Respuesta:

Es lineal pero sesgado. Comprobemos esto último:

$$\mathbb{E}[\dot{\beta}_2|X] = \mathbb{E}\left[\frac{\sum_n Y_n}{\sum_n X_n}\middle| X\right] = \mathbb{E}\left[\frac{\sum_n(\beta_1+\beta_2 X_n+U_n)}{\sum_n X_n}\middle| X\right] = \frac{\sum_n \beta_1 + \beta_2\sum_n X_n}{\sum_n X_n} = \frac{N\beta_1}{\sum_n X_n} + \beta_2 \neq \beta_2$$

La única excepción ante la cual el estimador sería insesgado es cuando $\beta_1 = 0$

b. Dados los supuestos realizados, ¿podría proponer usted un mejor estimador que $\dot{\beta}_2$?. ¿Qué estimador sería este? ¿En qué fundamenta esta propuesta?

Respuesta:

A priori no se puede proponer un estimador mejor que $\dot{\beta}_2$, ya que este estimador es lineal sesgado y por teorema visto en clases se sabe que el estimador obtenido por MCO es el mejor de los estimadores lineales insesgados, ante lo cual no se pueden comparar bajo este criterio. Esto ya que para comparar si un estimador es mejor que otro se recurre al error cuadrático medio (ECM), el cual aparte del sesgo, incorpora la varianza de éste, la cual no es conocida a priori para ninguno de los dos.

$$ECM(\widehat{\beta}_k) = B(\widehat{\beta}_k)^2 + \mathbb{V}(\widehat{\beta}_k)$$

c. Estime la varianza del estimador $\dot{\beta}_2$ y obtenga una expresión numérica para el nivel relativo de ineficiencia de dicho estimador respecto al MCO, es decir $\mathbb{V}(\dot{\beta}_2)/\mathbb{V}(\widehat{\beta}_2)$ ¿Hay alguna circunstancia en la que $\dot{\beta}_2$ sea tan bueno como $\widehat{\beta}_2$?

Respuesta:

$$\begin{aligned}
\mathbb{V}(\dot{\beta}_2|X) &= \mathbb{V}\left[\frac{\sum_n(\beta_1+\beta_2 X_n+U_n)}{\sum_n X_n}\middle| X\right] \\
&= \frac{1}{\left(\sum_n X_n\right)^2}\cdot\mathbb{V}\left[\sum_n(\beta_1+\beta_2 X_n+U_n)\middle| X\right] \\
&= \frac{1}{\left(\sum_n X_n\right)^2}\cdot\mathbb{V}\left[\sum_n \beta_1+\beta_2\sum_n X_n+\sum_n U_n\middle| X\right] \\
&= \frac{1}{\left(\sum_n X_n\right)^2}\cdot\mathbb{V}\left[\sum_n U_n\middle| X\right] \\
&= \frac{N\sigma^2}{\left(\sum_n X_n\right)^2}
\end{aligned}$$

La ineficiencia relativa con respecto al estimador MCO es:

$$\frac{\mathbb{V}(\dot{\beta}_2)}{\mathbb{V}(\widehat{\beta}_2)} = \frac{\frac{N\sigma^2}{\left(\sum_n X_n\right)^2}}{\frac{\sigma^2}{\sum_n X_n^2 - N\overline{X}^2}}$$

Para que un estimador sea mejor que otro, por lo general se recurre al criterio otorgado por el ECM, aquel que contenga un menor valor de éste será mejor, ya que lleva a un óptimo entre insesgo y eficiencia del estimador.

Luego se calculan los ECM para cada uno:

$$ECM(\widehat{\beta}_k) = B(\widehat{\beta}_k)^2 + \mathbb{V}(\widehat{\beta}_k)$$

$$ECM(\dot{\beta}_2) = \left(\frac{N\beta_1}{\sum_n X_n}\right)^2 + \frac{N\sigma^2}{(\sum_n X_n)^2}$$

$$ECM(\widehat{\beta}_2) = 0^2 + \frac{\sigma^2}{\sum_n X_n^2 - N\overline{X}^2}$$

2. Teorema Frisch-Waugh / Multicolinealidad

[P5, Control 1, 2014] Un empresario productor de leche desea estudiar si dos tipos de forraje (alimento) X y Z logran elevar la producción de leche Y de 250 vacas que tiene en su predio.

1. Suponga que a usted le pidieran diseñar el experimento para estudiar el efecto de los dos tipos de forraje X y Z sobre la producción de leche. Explique cómo asignaría cantidades de X y Z a las vacas para determinar los efectos de X y Z sobre la producción de leche con la mayor precisión.

   Respuesta:

   Se tiene que la varianza del estimador asociado a un coeficiente en una regresión múltiple está dado por:

   $$\widehat{\mathbb{V}}(\widehat{\beta}_A) = \frac{\mathbb{V}(Y)(1-R^2)}{(N-K)\mathbb{V}(X_A)} \cdot \frac{1}{1-R^2_{A,-A}}$$

   La mejor estimación posible se obtiene al diseñar el experimento de modo tal que la varianza del estimador se reduzca lo más posible. Esto puede hacerse de dos formas: (1) asignando cantidades de X y Z a las vacas de forma tal que el factor inflador de varianza sea 1 (o correlación entre X y Z sea 0) y (2) Elevando al máximo posible la variabilidad de X y Z. (En la práctica esto último puede tener algunos inconvenientes no estipulados con claridad, ya que una variabilidad demasiado extrema podría dejar sin comer a las vacas o tener un costo financiero demasiado alto.)

2. El empresario siguió su diseño experimental, pero encarga a un veterinario amigo que genere los resultados. El veterinario, solo sabe de regresion simple, por lo cual obtiene los siguientes resultados de dos regresiones donde las desviaciones estandar de los coeficientes están en el paréntisis bajo los respectivos coeficientes estimados. ¿Qué puede concluir sobre el verdadero impacto de X y Z sobre la producción de leche?

$\underset{(0,21)}{\widehat{Y}_n = 5,61} + \underset{(0,03)}{0,51X_n},\ R^2 = 0,15$	$\underset{(0,18)}{\widehat{Y}_n = 5,55} + \underset{(0,12)}{0,22Z_n},\ R^2 = 0,08$

   Respuesta:

   Como mediante el diseño experimental anterior se logró que las variables no estuviesen, el resultado que se hubiese obtenido por regresión múltiple sobre cada una de las variables es el mismo que se obtuvo al hacer una regresión bivariada. Por lo tanto, los efectos estimados de X y Z por el veterinario en dos regresiones bivariadas son efectivamente los estimadores MCO del modelo de regresión múltiple. El veterinario obtiene la respuesta correcta.

3. Suponga que en lugar de hacerle caso a ud., el empresario sigue el consejo de su amigo veterinario para diseñar el experimento. El veterinario sugiere que se asignen cantidades de X y Z aleatoriamente a cada vaca, completando exactamente 20 kilos de forraje en total. (Sugerencia: escriba la matriz de regresores). ¿Sería posible determinar los efectos individuales de X y Z bajo este diseño experimental? ¿Cómo cambia su respuesta si la cantidad de X más Z suman 20 con error aleatorio?

   Respuesta:

   Bajo este diseño experimental no es posible determinar los efectos individuales de X y Z, ya que X y Z tienen una correlación igual a -1, ya que $Z = 20 - X$, con lo que su correlación es perfecta y la varianza asociada al coeficiente será infinita de acuerdo a la fórmula ya mostrada, pero más aún el estimador MCO no está definido. No pudiendo verse así el efecto individual de cada variable. Mientras que si ambas suman 20 con un error aleatorio, será posible determinar los efectos de ambos, ya que habrá parte de la varianza sobre la producción de leche asociada tan sólo a cada variable (esto se deduce del teorema de Friesch-Waugh). Sin embargo existirá una correlación negativa entre X y Z, por lo cual el diseño experimental no será el ideal con $R^2_{-X} = 0$. Habrá mayor varianza que en el diseño anterior, y no se encontrarán respuestas adecuadas con regresión bivariada.

3. Bondad de ajuste y varianza de los estimadores

Usaremos la base de datos `Auxiliar 03 - BD.xls` subida a material docente. Usted está desarrollando su práctica profesional en una prestigiosa institución dedicada a la investigación económica. Su jefe le solicita regresionar el gasto en educación de los hogares en función de la edad del jefe de hogar (y su cuadrado). Sin embargo no está seguro si incluir el ingreso total de los hogares a la regresión debido al efecto que puede tener esto sobre la eficiencia de los estimadores de los otros regresores.

1. Ejecute ambas regresiones en *Stata* (con y sin el ingreso total del hogar) y analice el efecto de incluir esta variable sobre las desviaciones estándar de los estimadores y los coeficientes R^2 y R^2_a.

   Código:

```
// Limpiar variables
clear all

// Configurar directorio
cd "C:\Users\Toshiba\Desktop\Ronald\Universidad\Ramos (Equipo Docente)\Aplicaciones de probabilidades y estad sticas para la gesti n 2018-1\Auxiliares\Auxiliar 03\Auxiliar 03 - Pauta\Auxiliar 03 - Pauta P3\DTA"
gl dir "C:\Users\Toshiba\Desktop\Ronald\Universidad\Ramos (Equipo Docente)\Aplicaciones de probabilidades y estad sticas para la gesti n 2018-1\Auxiliares\Auxiliar 03\Auxiliar 03 - Pauta\Auxiliar 03 - Pauta P3"

// Importar base de datos
import excel "Auxiliar 03 - BD.xls", firstrow

// Regresionar lo pedido
reg gastoeduc ejh ejh2
reg gastoeduc ejh ejh2 ytoth
```

Source	SS	df	MS	Number of obs	=	15.727
				F(2, 15724)	=	48,39
Model	3,6907e+12	2	1,8454e+12	Prob > F	=	0,0000
Residual	5,9965e+14	15.724	3,8136e+10	R-squared	=	0,0061
				Adj R-squared	=	0,0060
Total	6,0335e+14	15.726	3,8366e+10	Root MSE	=	2,0e+05

gastoeduc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ejh	7175,608	1395,97	5,14	0,000	4439,346	9911,87
ejh2	-62,30282	15,13039	-4,12	0,000	-91,96012	-32,64552
_cons	-49417,79	31370,58	-1,58	0,115	-110907,7	12072,15

Figura 1: Resultados de regresionar el gasto en educación en función de la edad del jefe de hogar y su cuadrado.

Source	SS	df	MS	Number of obs	=	15.727
				F(3, 15723)	=	1723,59
Model	1,4932e+14	3	4,9772e+13	Prob > F	=	0,0000
Residual	4,5403e+14	15.723	2,8877e+10	R-squared	=	0,2475
				Adj R-squared	=	0,2473
Total	6,0335e+14	15.726	3,8366e+10	Root MSE	=	1,7e+0

gastoeduc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ejh	3122,962	1216,075	2,57	0,010	739,3141	5506,609
ejh2	-20,8402	13,179	-1,58	0,114	-46,67254	4,992148
ytoth	,0592616	,0008345	71,01	0,000	,0576259	,0608973
_cons	-37786,18	27298,32	-1,38	0,166	-91294,03	15721,67

Figura 2: Resultados de regresionar el gasto en educación en función del ingreso del hogar, la edad del jefe de hogar y su cuadrado.

Resultados:

Al comparar las figuras 1 y 2 se observan los siguientes efectos de incorporar la variable `ytoth` (ingreso total del hogar):

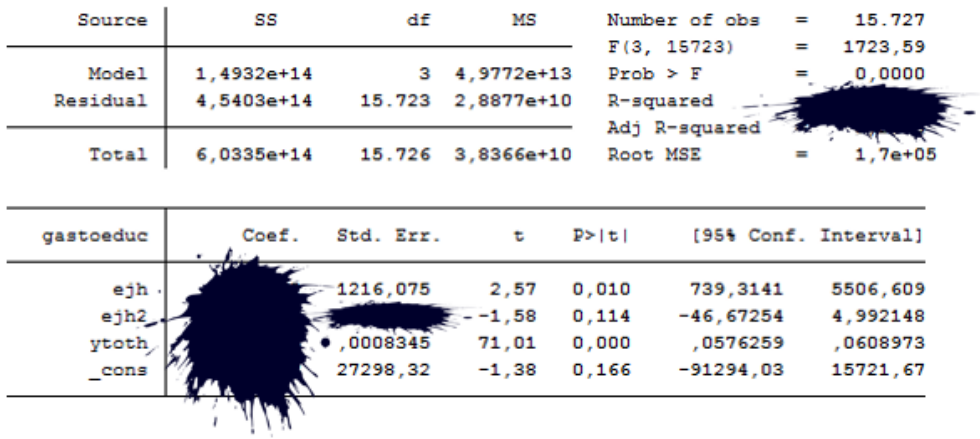
- Un aumento considerable $(0,0061 \rightarrow 0,2475)$ del coeficiente de ajuste R^2. Esto quiere decir que el modelo es capaz de explicar la disperción de los datos (varianza) relacionados al gasto en educación mucho mejor que antes de incorporar `ytoth` a la regresión.
- Un aumento considerable $(0,006 \rightarrow 0,2473)$ del coeficiente R^2 ajustado. La penalización por cantidad de regresores afectó muy levemente en comparación a la ganancia en términos de ajuste provocada por la incorporación de `ytoth`.
- La varianza (y desviación estándar, que es lo que vemos en la tabla) de todos los estimadores $\widehat{\beta}$ se vio reducida.

  Si observamos la fórmula del estimador de varianza de los estimadores MCO, podríamos intuir que esta reducción de varianza se debe a que: (1) el coeficiente de ajuste R^2 de la regresión aumentó considerablemente, (2) K solo aumentó en una unidad y (3) que `ytoth` no está suficientemente correlacionado

con los otros regresores, por lo que $R^2_{\text{ejh},-\text{ejh}}$ y $R^2_{\text{ejh}^2,-\text{ejh}^2}$ no aumentarían considerablemente al agregar `ytoth`.

$$\widehat{\mathbb{V}}(\widehat{\beta}_A) = \frac{\mathbb{V}(Y)(1-R^2)}{(N-K)\mathbb{V}(X_A)} \cdot \frac{1}{1-R^2_{A,-A}}$$

2. Es su último día de práctica usted ha llegado a casa feliz pues le fue bastante bien. Además, como usted es un(a) alumno(a) responsable empezó a hacer su informe de práctica con anticipación, por lo que está casi listo(a) para salir de vacaciones. Solo le falta reportar un par de resultados que tiene impresos en su mochila. Sin pensarlo dos veces usted la abre y se da cuenta que se le reventó un lapiz cuya tinta no le permite ver algunos datos. Usted no tiene *Stata* instalado en su computador, sin embargo sabe utilizar las fórmulas necesarias y cuenta con *Matlab*. Obtenga los datos faltantes para poder salir de vacaciones.



Source	SS	df	MS		
Model	1,4932e+14	3	4,9772e+13	Number of obs	= 15.727
Residual	4,5403e+14	15.723	2,8877e+10	F(3, 15723)	= 1723,59
Total	6,0335e+14	15.726	3,8366e+10	Prob > F	= 0,0000
				R-squared	
				Adj R-squared	
				Root MSE	= 1,7e+05

gastoeduc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ejh	1216,075		2,57	0,010	739,3141	5506,609
ejh2	-1,58		-1,58	0,114	-46,67254	4,992148
ytoth	,0008345		71,01	0,000	,0576259	,0608973
_cons	27298,32		-1,38	0,166	-91294,03	15721,67

Código:

```
clear all
close all
clc

cd 'C:\Users\Toshiba\Desktop'
bd = xlsread('Auxiliar 03 - BD.xls');

%% La primera variable corresponde a la mortalidad y la segunda al gasto
ejh = bd(:,1);
ejh = ejh(2:15727,1);

ejh2 = bd(:,2);
ejh2 = ejh2(2:15727,1);

ytoth = bd(:,4);
ytoth = ytoth(2:15727,1);

gastoeduc = bd(:,3);
gastoeduc = gastoeduc(2:15727,1);

%%
```

```
22
23 % Creamos un vector de unos, para incorporar intercepto a nuestro modelo
24 X_0 = ones(15726,1);
25
26 %%
27
28 % Creamos matriz X y llamamos Y al gasto en educacin
29 X = [X_0 ejh ejh2 ytoth];
30 Y = gastoeduc;
31
32 %% Calculamos estimador MCO
33 beta_MCO = (X'*X)^(-1)*X'*Y;
34 beta_0 = beta_MCO(1)
35 beta_ejh = beta_MCO(2)
36 beta_ejh2 = beta_MCO(3)
37 beta_ytoth = beta_MCO(4)
38
39 %% Calculamos el R^2
40 N = length(Y)
41 K = 4
42 Y_estimado = X*beta_MCO;
43 R2 = 1 - sum((Y - Y_estimado).^2)/sum((Y - mean(Y)).^2)
44 R2a = 1 - sum((Y - Y_estimado).^2)/sum((Y - mean(Y)).^2) * (N-1)/(N-K)
45
46 %% Calcular la varianza estimada de la edad del jefe de hogar
47
48 %Necesitamos realizar otra regresin  para obtener R^2_ejh2,-ejh2
49
50 y = ejh2;
51 x = [X_0 ejh ytoth];
52 %%
53 beta_mco = (x'*x)^(-1)*x'*y;
54 y_estimado = x*beta_mco;
55
56 R2_ejh2 = 1 - sum((y - y_estimado).^2)/sum((y - mean(y)).^2)
57
58 %% Finalmente calculamos los estimadores de la var. y desv. est. de
      beta_ejh2
59
60 var_beta_ejh2 = var(Y)*(1-R2)/((N-K)*var(ejh2)*(1-R2_ejh2))
61
62 sd_beta_ejh2 = sqrt(var_beta_ejh2)
```