Regresión Lineal Simple

Econometría

Camilo Vargas Walteros

1. Modelo de regresión

1.1 Lineal y simple

Lineal y simple

- Una **"regresión"** busca explicar los cambios de una variable (Y) que son generados por los cambios de otras variables (Xs).
- "Y" es una variable **"estocástica"** porque está asociada a una distribución de probabilidad.
- Las "Xs" son variables **"deterministas"**.

Nombres para "Y"	Nombres para "X"
Dependiente	Independiente
Regresando	Regresor
Variable consecuencia	Variable casual
Variable explicada	Variable explicativa

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 28, Box 2.1*

Lineal y simple

- Ejemplos de modelos determinísticos:

$$C = \frac{9F}{5} + 32$$

$$V = V_0 + at$$

$$V = \frac{k}{P}$$

$$F = P(1+r)^n$$

$$Y = C + I + G + X - M$$

$$V = H_0 D$$

$$R = Log\left(\frac{I}{I_0}\right)$$

$$A = P + E$$

$$IMC = \frac{m}{h^2}$$

Lineal y simple

- La **"correlación"** muestra el grado de **"asociación lineal"** entre dos variables sin realizar distinción entre las variables.
- En una **"regresión"** nos interesa conocer el valor **"promedio"** de una variable **"dependiente"** dados valores fijos de otras variables.
- Antes de iniciar el análisis de regresión es útil graficar la relación entre las variables.
- ¿Existe algún tipo de relación entre las variables?, ¿Es una relación lineal?, ¿Cuánto es su intercepto y pendiente?

Lineal y simple

Ingreso familiar (X) y Consumo familiar (Y)

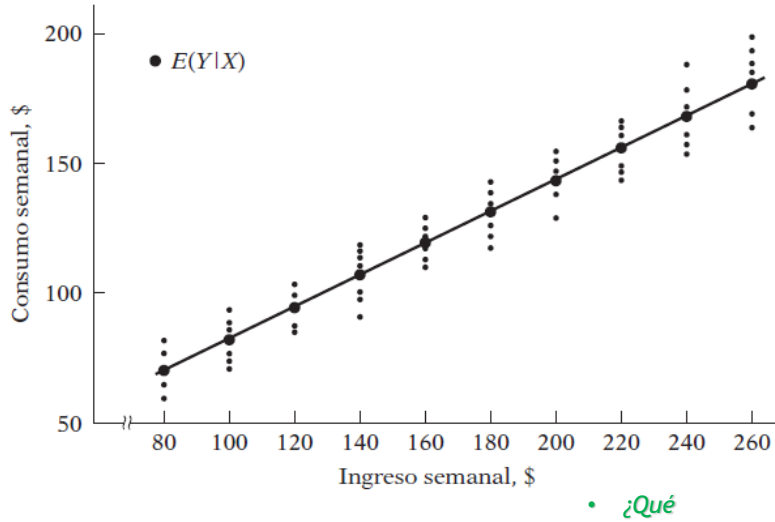
Y ↓ \ X →	80	100	120	140	160	180	200	220	240	260
Consumo familiar semanal Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1 043	966	1 211
Media condicional de Y, $E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

- *FUENTE: Gujarati (2010), Econometría, P 35, Tabla 2.1*

- *Realiza una gráfica colocando el ingreso familiar en el eje horizontal, el consumo familiar y su media condicional en el eje vertical.*
- *Traza una línea uniendo todas las medias condicionales.*

Lineal y simple

Ingreso familiar (X) y Consumo familiar (Y)

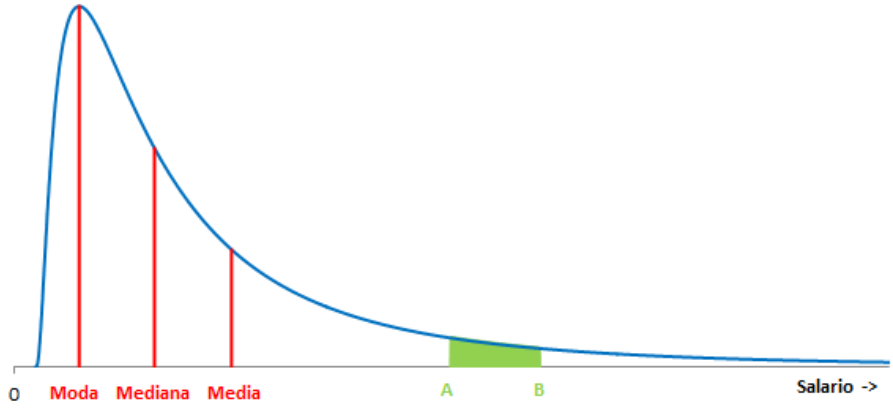


- *FUENTE: Gujarati (2010), Econometría, P 35, Figura 2.1*

- *¿Qué problemas presenta esta regresión?*

Lineal y simple

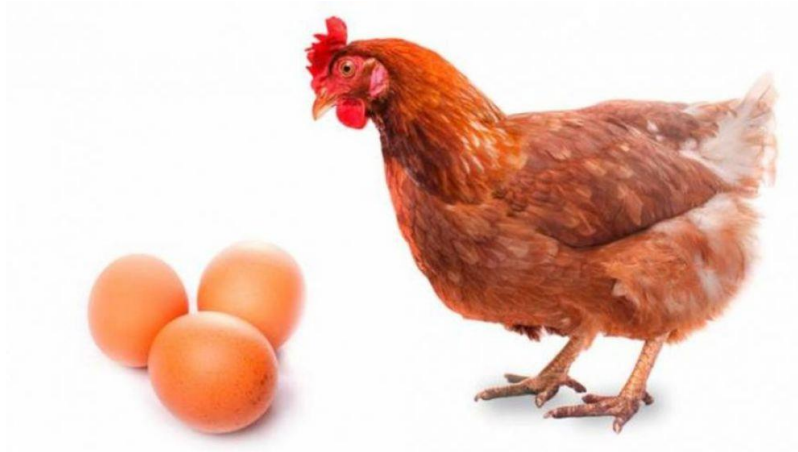
Problema 1: Lo que no dice la media



- *FUENTE: https://ssociologos.com/2014/05/13/la-omnipresente-media-estadistica-que-nos-dice-y-que-nos-oculta/*

Lineal y simple

Problema 2: Endogeneidad



- *¿Existe endogeneidad entre las horas de estudio y el rendimiento académico?*
- *FUENTE: www.lasexta.com/tecnologia-tecnoxplora/ciencia/descubrimientos/que-fue-antes-el-huevo-o-la-gallina_201809105b98ff200cf2e982a160a5b7.html*

Lineal y simple

- Se busca calcular el valor esperado de Y condicional a un valor de "X":

$$E(Y|X) = f(X_i)$$

- Asumiendo una forma lineal en los parámetros y en la variable independiente:

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

- El "modelo de regresión" adquiere su nombre de "**lineal**" porque debe ser lineal en los parámetros.

Lineal y simple

- Relaciones no lineales entre variables se pueden expresar mediante un modelo de regresión lineal:

$$q_i = AL_i^{\beta_2} e^{u_i}$$

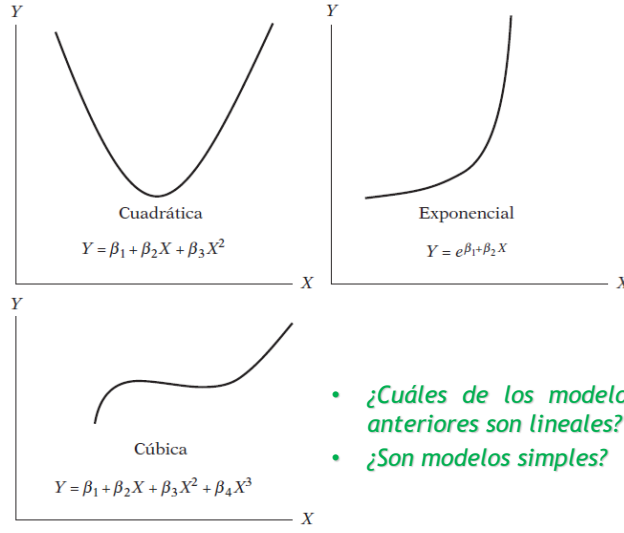
$$Ln(q_i) = Ln(A) + \beta_2 Ln(L_i) + u_i$$

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

- El "modelo de regresión" es "**simple**" porque únicamente depende de una variable independiente y tiene una pendiente.

Lineal y simple

Funciones lineales en los parámetros



- *¿Cuáles de los modelos anteriores son lineales?*
- *¿Son modelos simples?*

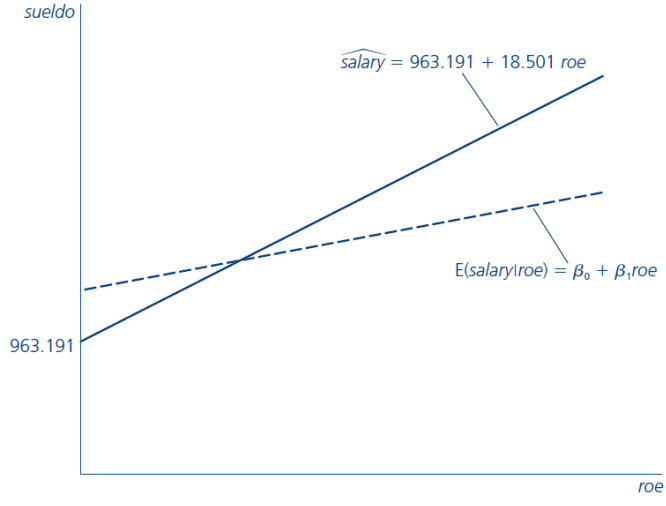
- *FUENTE: Gujarati (2010), Econometría, P 39, Figura 2.3*

1. Modelo de regresión

1.2 Mínimos Cuadrados Ordinarios

Mínimos Cuadrados Ordinarios

Líneas de regresión muestral y poblacional



- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 34, Figura 2.5*

Mínimos Cuadrados Ordinarios

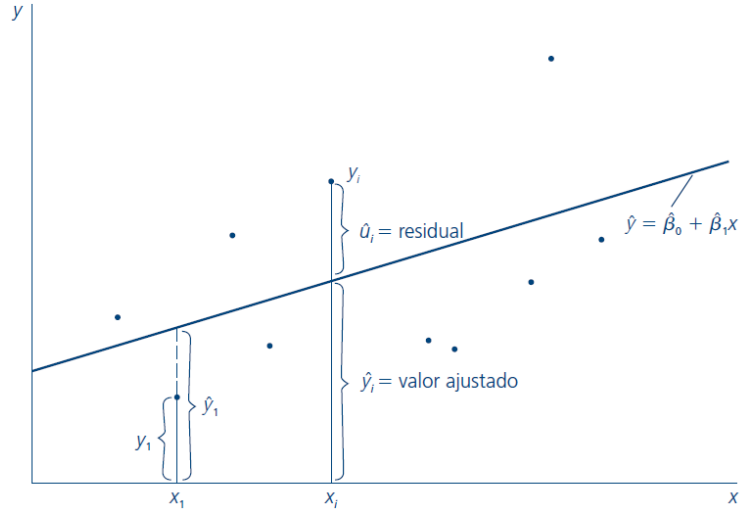
- Se busca estimar la siguiente regresión (muestral):

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

- Donde ($\hat{\beta}_1$) y ($\hat{\beta}_2$) son los estimadores.
- El objetivo es obtener una línea que mejor represente los datos a partir de los estimadores.
- El termino ($\hat{Y}_i$) representa la **"predicción"** de la variable dependiente dado un valor de la variable independiente (X_i).

Mínimos Cuadrados Ordinarios

Línea de mejor ajuste, dato observado y estimado



- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 31, Figura 2.4*

Mínimos Cuadrados Ordinarios

- Los estimadores pueden calcularse por varios métodos que buscan minimizar lo siguiente:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

- Esta distancia se define como la diferencia entre el valor de la variable dependiente (Y_i) y la predicción del modelo ($\hat{Y}_i$).
- Esta resta también muestra los **"residuos"** para cada uno de los datos de la regresión.

Mínimos Cuadrados Ordinarios

¿Por qué se debe incluir el termino del error?

- Error de medición en las variables: encuestas poblacionales (ingreso y distancia al hogar).
- Error de especificación (forma lineal o no lineal).
- En las ciencias sociales los experimentos no son controlados (el PIB no lo decide el investigador).
- Modelo más simple (guardar variables irrelevantes en el término del error).
- Variables no observables (creencia).
- Variables omitidas (vacíos en la teoría).
- Variables representativas inadecuadas (proxy).

Mínimos Cuadrados Ordinarios

- Por ejemplo, al tener 3 pares de datos para una variable X y una variable Y, se generan 3 pronósticos y 3 residuos:

$$\hat{Y}_1 = \hat{\beta}_1 + \hat{\beta}_2 X_1 \quad \Rightarrow \quad \hat{u}_1 = Y_1 - \hat{Y}_1$$

$$\hat{Y}_2 = \hat{\beta}_1 + \hat{\beta}_2 X_2 \quad \Rightarrow \quad \hat{u}_2 = Y_2 - \hat{Y}_2$$

$$\hat{Y}_3 = \hat{\beta}_1 + \hat{\beta}_2 X_3 \quad \Rightarrow \quad \hat{u}_3 = Y_3 - \hat{Y}_3$$

Mínimos Cuadrados Ordinarios

- El objetivo consiste en **"minimizar la suma de los errores"** por medio de varias formas (Ver Kennedy 2008, P 13):

$$Min: \sum_{i=1}^{3}\left(\frac{u_i}{3}\right) = \sum_{i=1}^{3}\left(\frac{Y_i - \hat{Y}_i}{3}\right)$$

$$Min: \sum_{i=1}^{3}|u_i| = \sum_{i=1}^{3}\left|Y_i - \hat{Y}_i\right|$$

Mínimos Cuadrados Ordinarios

- El método más utilizado para minimizar los errores consiste en realizar una suma cuadrática:

$$u_1^2 + u_2^2 + u_3^2 = \sum_{i=1}^{3} u_i^2$$

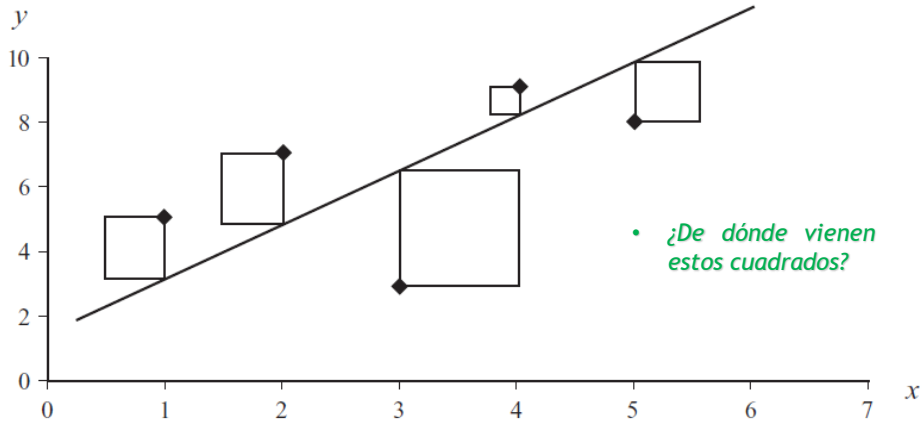
- *¿Por qué se elevan los errores al cuadrado?*

$$\sum_{i=1}^{3}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{3}\left(Y_i - \left[\hat{\beta}_1 + \hat{\beta}_2 X_i\right]\right)^2$$

$$\underset{\beta_1,\beta_2}{Min}: \sum_{i=1}^{3}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

Mínimos Cuadrados Ordinarios

Minimizando la distancia cuadrática (MCO)



- *¿De dónde vienen estos cuadrados?*

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 32, Figure 2.3*

$$\underset{\beta_1,\beta_2}{Min}: \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_2 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} \qquad m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$$\hat{\beta}_2 = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

Mínimos Cuadrados Ordinarios

- (Y) es el peso en libras y (X) la estatura en pulgadas.
- Encuentra el valor de los estimadores.
- Grafica la regresión.
- Calcula el Y estimado, los errores y la suma de los errores al cuadrado utilizando la información de la tabla.

Peso (Y)	Estatura (X)
132	68
108	64
102	62
115	65
128	66

- *FUENTE: Anderson (2008), Estadística para Economía y Administración, P 553.*

1. Modelo de regresión lineal

1.3 Clásico

Modelo clásico

- Los estimadores fueron calculados por (MCO) a partir de una **"muestra".**
- El modelo de regresión se puede utilizar para obtener conclusiones de toda la **"población"**.
- Se realizarán supuestos sobre la forma en como se genera las variables independientes (X) y el término del error (u).
- Estos supuestos fueron empleados por primera vez por Gauss en 1821 y desde ese entonces el modelo se considera un **"clásico"**.

- *FUENTE: http://es.wikipedia.org/wiki/Arist%C3%B3teles*

Modelo clásico

Supuesto 1: *Modelo de regresión lineal*

- El modelo de regresión es lineal en los parámetros.

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

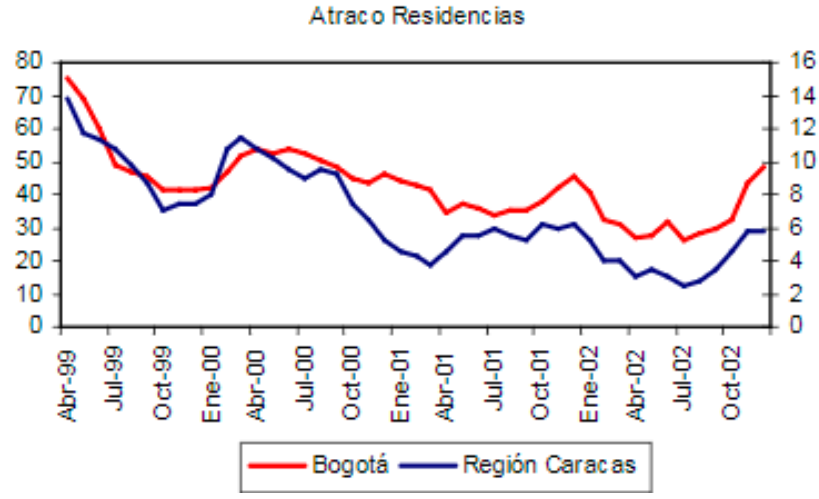
Supuesto 2: *Independencia entre las variables independientes y el término del error*

$$\text{cov}(X_i, u_i) = 0 \qquad \sum_{i=1}^{n} \hat{u}_i X_i = 0$$

- Las variables independientes se encuentran **"predeterminadas"**.
- No existe **"covarianza"** entre variables observadas y variables no observadas.
- No se presenta **"sesgo de selección"** (distancia al tablero y rendimiento).
- No tenemos **"endogeneidad"**.
- El cumplimiento de este supuesto permite la estimación de los betas por el **"Método Generalizado de Momentos"**.
- En la práctica el cumplimiento de este supuesto permite desarrollar ejercicios de **"evaluación de impacto"** (grupo de control y tratamiento).

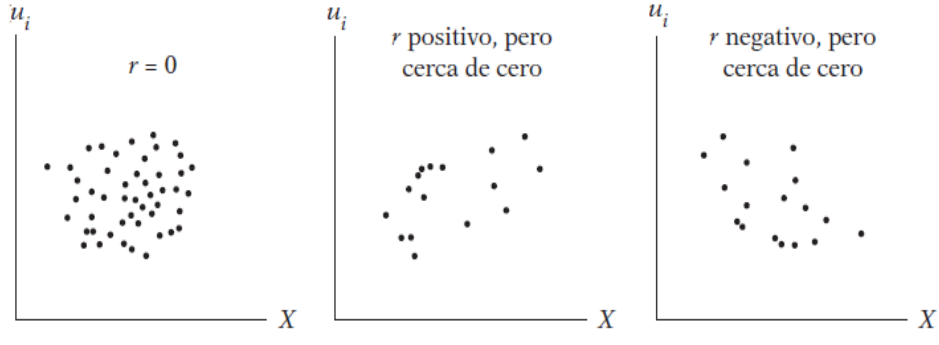
Modelo Clásico

Grupo de control y tratamiento (los efectos de Transmi)



- *FUENTE: Moreno (2005), Los efectos de Transmilenio en el crimen de la avenida Caracas y sus vecindades, Grafico 1, P11*

$$\text{cov}(X_i, u_i) = 0$$



- *FUENTE: Gujarati (2010), Econometría, P 78, Figura 3.10 (*)*

$$\underset{\beta_1,\beta_2}{Min}: L = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

$$\frac{\partial L}{\partial \hat{\beta}_2} = \sum_{i=1}^{n} 2\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)\left(-X_i\right) = 0$$

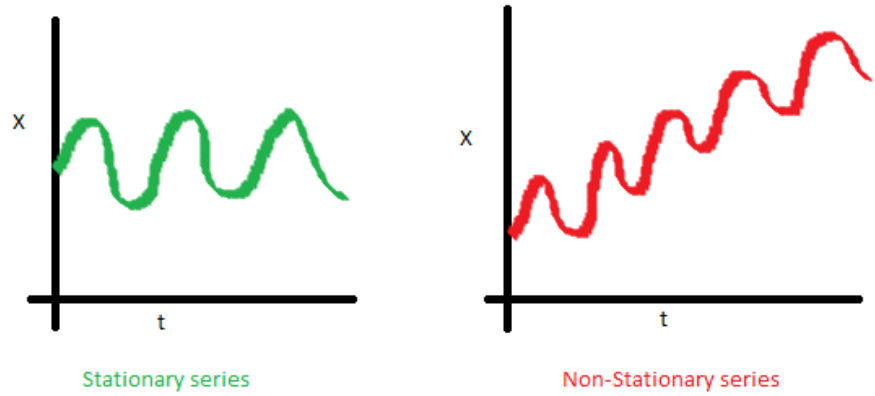
$$-2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)\left(X_i\right) = 0$$

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

- *Bajo el supuesto de independencia.*

Modelo clásico

Supuesto 3: *Errores estacionarios*



- *¿Cómo se comporta la media y la varianza en cada serie?*
- *FUENTE: https://estrategiastrading.com/series-estacionarias/*

Modelo clásico

Supuesto 3.1: *Valor Esperado del error es cero*

- La media del error es constante e igual a cero.

$$E(u_i) = 0 \qquad \sum_{i=1}^{n} \hat{u}_i = 0$$

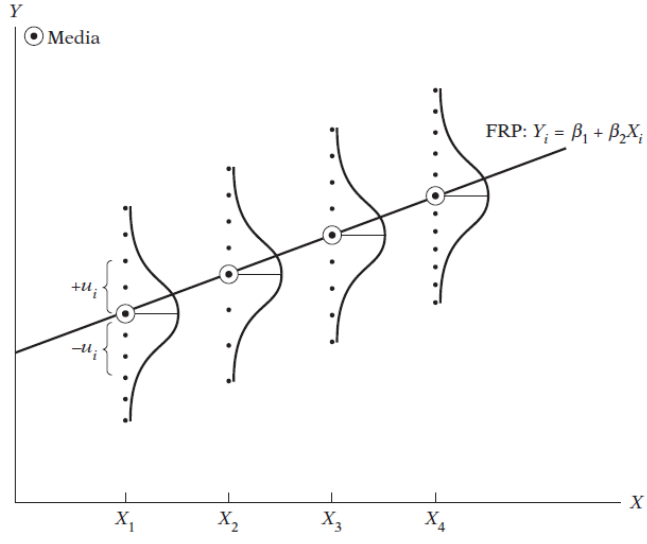
- *Realiza la suma de los errores para el ejemplo del peso y la estatura.*

Supuesto 3.2: *Errores homoscedásticos*

- La varianza de los errores es constante.

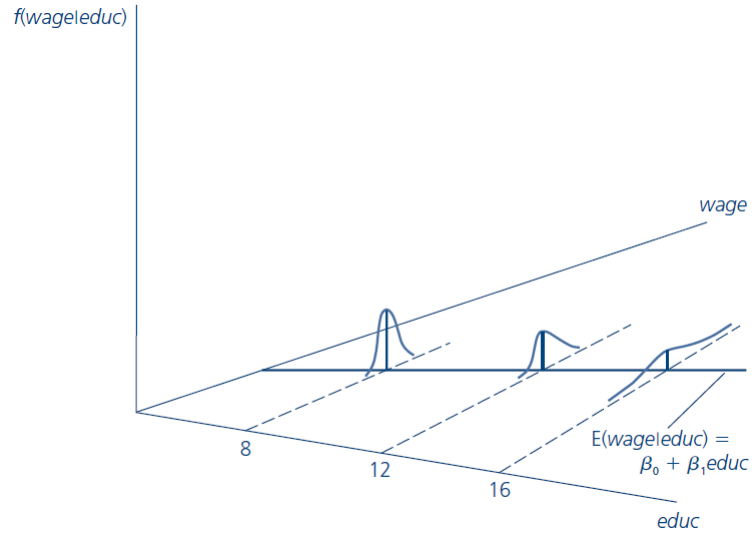
$$\text{var}(u_i) = \sigma^2$$

$$E(u_i) = 0 \qquad \text{var}(u_i) = \sigma^2$$



- *FUENTE: Gujarati (2010), Econometría, P 63, Figura 3.3*

$$\text{var}(u_i) = \sigma_i^2$$



- *FUENTE: Wooldridge (2010), Introducción a la Econometría, P 55, Figura 2.9*

Modelo clásico

Supuesto 4: *No auto correlación de los errores*

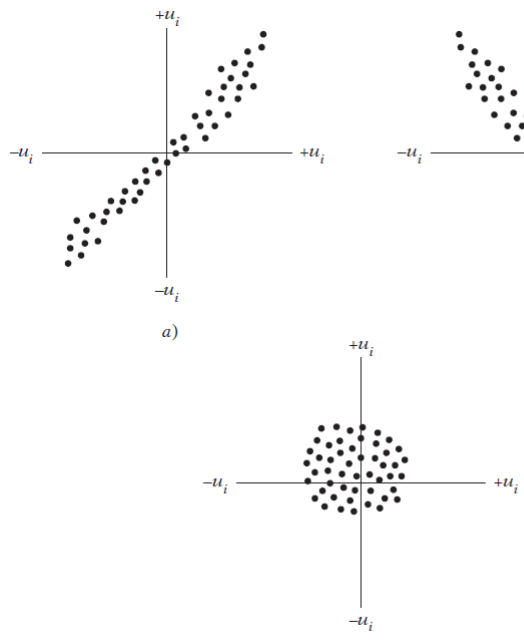
- No existe algún tipo de relación entre las variables no consideradas en el modelo.

$$\text{cov}(u_i, u_j) = 0$$

Supuesto 5: *El número de observaciones debe ser mayor al número de estimadores*

- No existe multicolinealidad perfecta.

$$n > k$$



$$\text{cov}(u_i, u_j) = 0$$

- *¿Cuál de las graficas cumple el supuesto de no autocorrelación?*

- *FUENTE: Gujarati (2010), Econometría, P 67, Figura 3.6*

1. Modelo de regresión lineal

1.4 Propiedades de los estimadores

Propiedades de los estimadores

Propiedades de los estimadores

Cuando se cumplen los supuestos del modelo los estimadores por MCO:

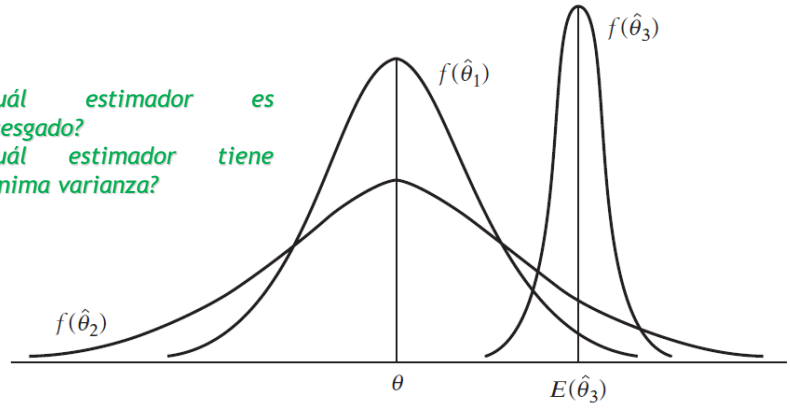
- Son **"insesgados"** porque en promedio los estimadores representan los verdaderos valores poblaciones (sin importar el tamaño de la muestra). Esto se cumple debido al supuesto 3.
- Son **"lineales"** en el sentido que sus fórmulas son combinaciones lineales de variables aleatorias.
- Son los **"mejores"** entre todos los estimadores lineales dado que tienen la menor varianza.

Propiedades de los estimadores

$$E\left(\hat{\theta}\right) = \theta$$

$$\text{var}\left(\hat{\theta}\right) = E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]^2$$

- *¿Cuál estimador es insesgado?*
- *¿Cuál estimador tiene mínima varianza?*



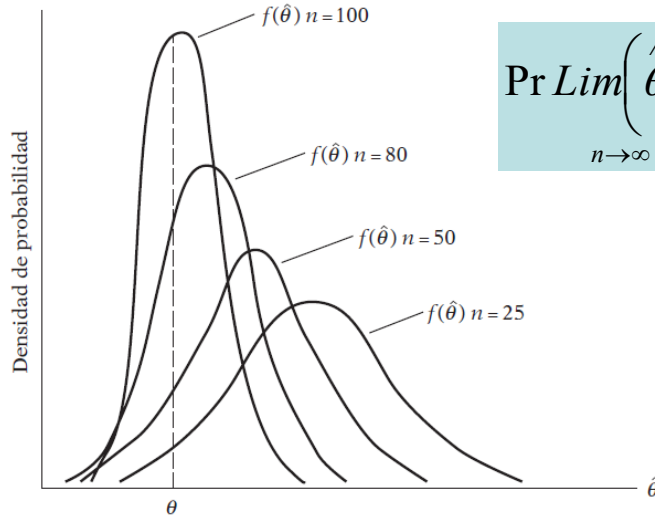
- *FUENTE: Gujarati (2010), Econometría, P 827, Figura A.9*

Propiedades de los estimadores

- Cuando se cumplen los supuestos del modelo el estimador por MCO se conoce por el nombre **"MELI"** es decir el **"Mejor Estimador Linealmente Insesgado"**.
- Cuando la muestra es muy grande el estimador no tienen sesgo y su varianza es cero por lo cual es **"consistente"**.
- La consistencia permite utilizar los estimadores porque estos convergen en probabilidad a sus contrapartes poblacionales.
- El **"teorema de Gauss Markov"** garantiza que los estimadores por MCO sean eficientes.

Propiedades de los estimadores

Consistencia



$$\Pr \underset{n\to\infty}{Lim}\left(\hat{\theta}\right)=\theta$$

- *FUENTE: Gujarati (2010), Econometría, P 829, Figura A.11*

Propiedades de los estimadores

Un ejemplo de la propiedad de **"consistencia"** se puede aplicar a la publicación de los boletines electorales (referendo Colombia 2016, Si = 1):

- N = 1.300.734 (15,19%). Promedio = 0,5111
- N = 3.228.870 (31,23%). Promedio = 0,5027
- N = 5.772.382 (50,13%). Promedio = 0,5009
- N = 8.421.244 (68,85%). Promedio = 0,5005
- N = 10.470.544 (83,08%). Promedio = 0,5000
- N = 11.598.295 (90,66%). Promedio = 0,4989
- N = 12.808.858 (99,98%). Promedio = 0,4978

  - *FUENTE: www.plebiscito.registraduria.gov.co*

1. Modelo de regresión lineal

1.5 Precisión de los estimadores

Precisión de los estimadores

Precisión de los estimadores por MCO

- En una población se pueden obtener diferentes muestras y como consecuencia estimadores distintos (edad de estudiantes).

¿Qué tan precisos son los estimadores utilizados?

¿Cuál es el grado de variabilidad de los estimadores al emplear muestras diferentes?

- Para responder estas preguntas es necesario calcular los errores estándar (EE) de los estimadores.

Precisión de los estimadores

Precisión de los estimadores por MCO

- Cuando se cumple el "supuesto 3" del modelo la varianza de los errores estimados es:

$$\hat{\sigma}^2 = \left(\frac{1}{n}\right)\sum \hat{u}_i^2$$

- Este parámetro es insesgado cuando:

$$\hat{\sigma}^2 = \left(\frac{1}{n-k}\right)\sum \hat{u}_i^2 = \left(\frac{1}{n-2}\right)\sum \hat{u}_i^2$$

Precisión de los estimadores

Precisión de los estimadores por MCO

$$\hat{\sigma} = \sqrt{\left(\frac{1}{n-2}\right)\sum \hat{u}_i^2}$$

- $(\hat{\sigma})$ es el error estándar de la regresión (EER).
- "EER" también se puede interpretar como la **"desviación"** estándar de la regresión.
- Entre menor sea "EER" más alta es la precisión de la regresión.

Precisión de los estimadores

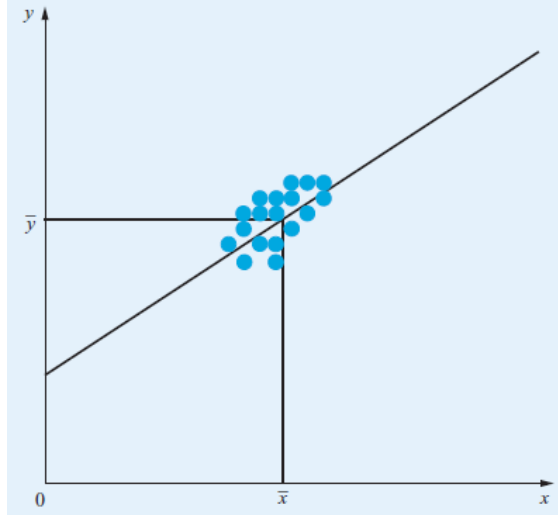
Precisión de los estimadores por MCO

$$ee\left(\hat{\beta}_1\right) = \hat{\sigma}\left(\sqrt{\frac{\sum X_i^2}{n\sum\left(X_i - \overline{X}\right)^2}}\right) = \hat{\sigma}\left(\sqrt{\frac{\sum X_i^2}{n\left[\sum X_i^2 - n\overline{X}^2\right]}}\right)$$

$$ee\left(\hat{\beta}_2\right) = \hat{\sigma}\left(\frac{1}{\sqrt{\sum\left(X_i - \overline{X}\right)^2}}\right) = \hat{\sigma}\left(\frac{1}{\sqrt{\sum X_i^2 - n\overline{X}^2}}\right)$$

Precisión de los estimadores

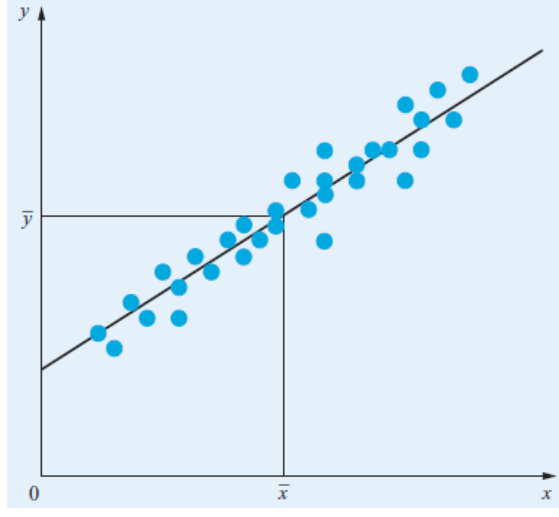
La pendiente y datos cercanos de la media



- *FUENTE: Brooks (2014), Introductory Econometrics for Finance, P 95, Figure 3.7*

Precisión de los estimadores

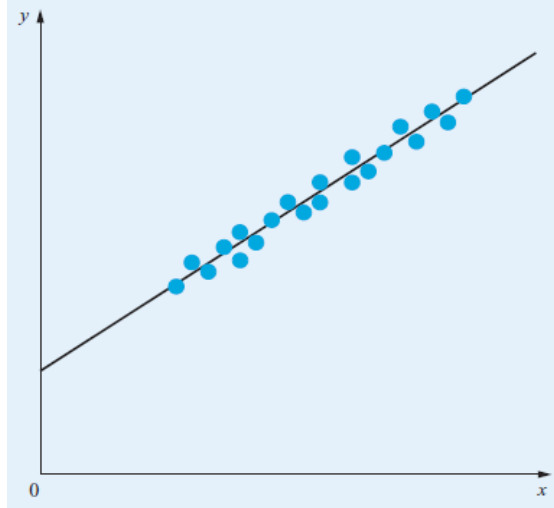
La pendiente y datos alejados de la media



- *FUENTE: Brooks (2014), Introductory Econometrics for Finance, P 96, Figure 3.8*

Precisión de los estimadores

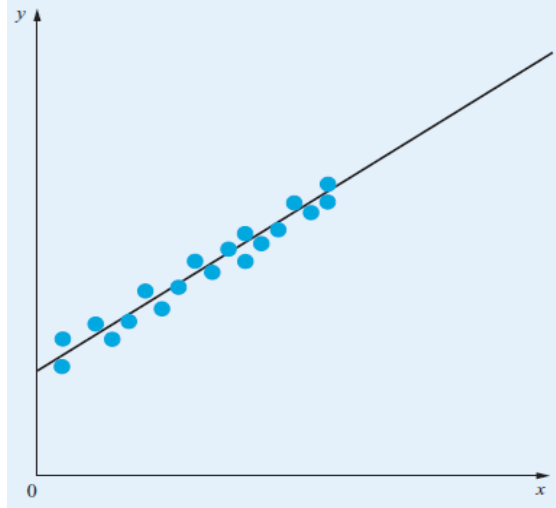
El intercepto y datos con suma X^2 grande



- *FUENTE: Brooks (2014), Introductory Econometrics for Finance, P 96, Figure 3.9*

Precisión de los estimadores

El intercepto y datos con suma X^2 pequeña



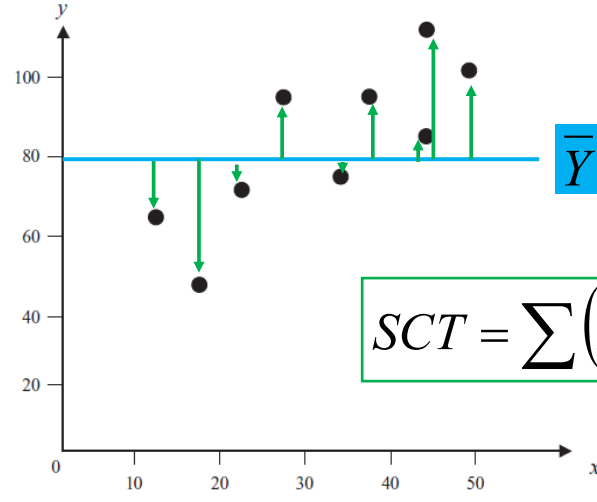
- *FUENTE: Brooks (2014), Introductory Econometrics for Finance, P 97, Figure 3.10*

1. Modelo de regresión lineal

1.6 Precisión del modelo

Precisión del modelo

Suma Total de Cuadrados (SCT)

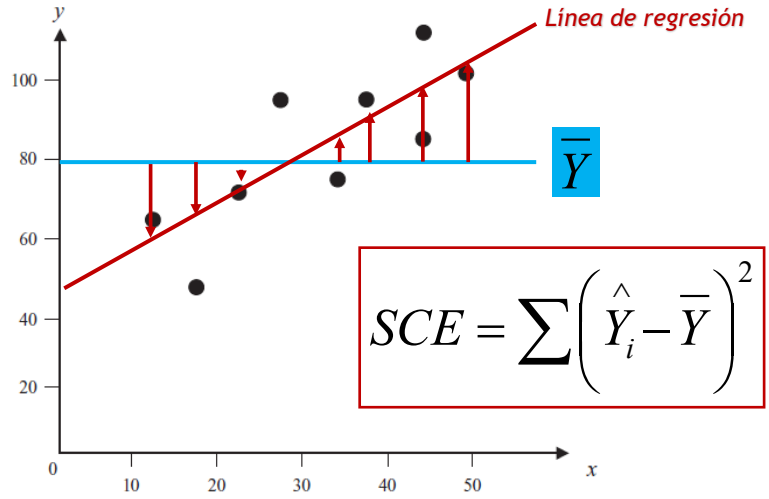


$$SCT = \sum \left(Y_i - \overline{Y}\right)^2$$

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 29, Figure 2.1 (adaptada con colores).*

Precisión del modelo

Suma Explicada de Cuadrados (SCE)

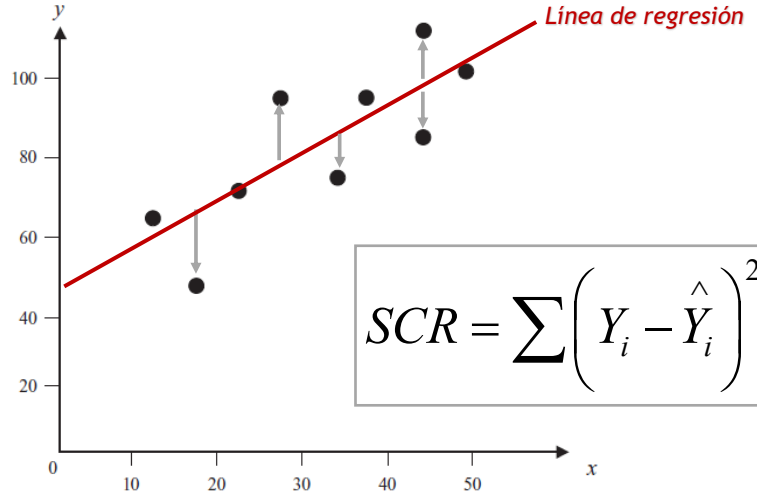


$$SCE = \sum \left(\hat{Y}_i - \overline{Y} \right)^2$$

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 29, Figure 2.1 (adaptada con colores).*

Precisión del modelo

Suma de Residuos al Cuadrado (SCR)



$$SCR = \sum \left(Y_i - \hat{Y}_i \right)^2$$

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 29, Figure 2.1 (adaptada con colores).*

Precisión del modelo

Medidas de bondad de ajuste (R cuadrado)

- Las variaciones de la muestra se generan de dos fuentes, la "suma explicada de cuadrados" (SCE) y la "suma de los residuos al cuadrado" (SCR):

$$Y_i = \left(\hat{\beta}_1 + \hat{\beta}_2 X_i \right) + \hat{u}_i$$

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

$$Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + \hat{u}_i$$

$$\left(Y_i - \overline{Y}\right)^2 = \left(\left[\hat{Y}_i - \overline{Y} \right] + \hat{u}_i \right)^2$$

$$\left(Y_i - \overline{Y}\right)^2 = \left(\left[\hat{Y}_i - \overline{Y} \right]^2 + 2\hat{u}_i \left[\hat{Y}_i - \overline{Y} \right] + \hat{u}_i^2 \right)$$

$$\sum \left(Y_i - \overline{Y}\right)^2 = \left(\sum \left[\hat{Y}_i - \overline{Y} \right]^2 + \sum 2\hat{u}_i \sum \left[\hat{Y}_i - \overline{Y} \right] + \sum \hat{u}_i^2 \right)$$

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum \left(\hat{Y}_i - \overline{Y} \right)^2 + \sum \left(Y_i - \hat{Y}_i \right)^2$$

$$SCT = SCE + SCR$$

Precisión del modelo

Medidas de bondad de ajuste (R cuadrado)

$$SCT = SCE + SCR$$

$$\frac{SCT}{SCT} = \frac{SCE}{SCT} + \frac{SCR}{SCT}$$

$$1 = \frac{SCE}{SCT} + \frac{SCR}{SCT}$$

Precisión del modelo

Medidas de bondad de ajuste (R cuadrado)

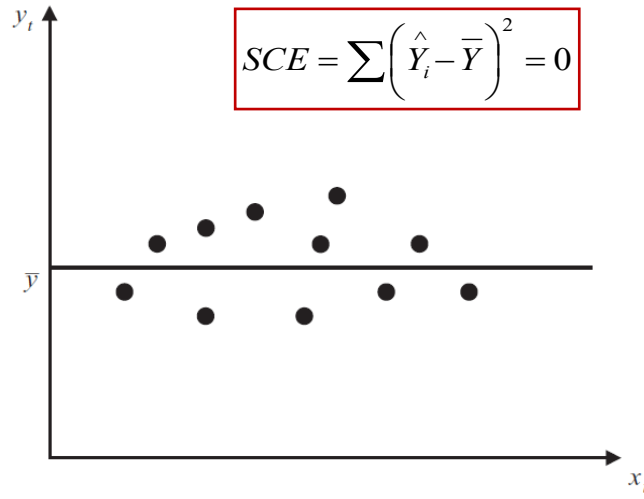
$$R^2 = 1 - \frac{SCR}{SCT} = \frac{SCE}{SCT}$$

- El R cuadrado mide el porcentaje de variación en "Y" explicada por las variaciones en "X".
- También se conoce como el **"coeficiente de determinación"** (no confundir con el coeficiente de correlación).
- La minimización de la suma de los errores al cuadrado implica la maximización del R cuadrado.

Precisión del modelo

Figure 3.1
$R^2 = 0$ demonstrated by a flat estimated line, i.e. a zero slope coefficient

$$\hat{Y}_i = \hat{\beta}_1 = \overline{Y}$$

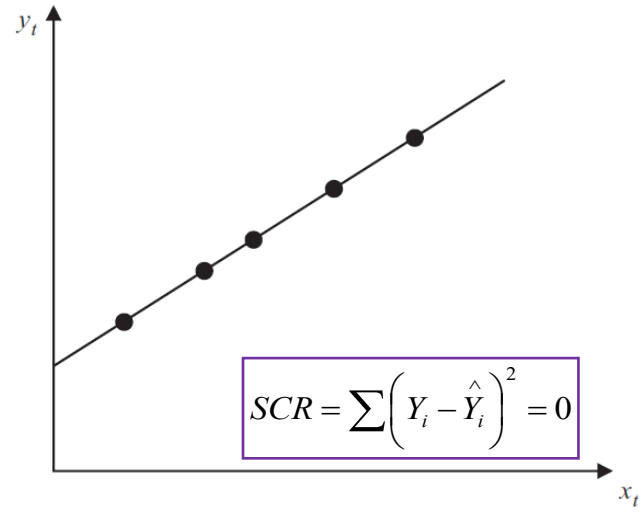


$$SCE = \sum \left(\hat{Y}_i - \overline{Y} \right)^2 = 0$$

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 109, Figure 3.1 (editada)*

Precisión del modelo

Figure 3.2
$R^2 = 1$ when all data points lie exactly on the estimated line



$$SCR = \sum \left(Y_i - \hat{Y}_i \right)^2 = 0$$

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 109, Figure 3.2 (editada)*

Precisión del modelo

- Encuentra el error estándar de la regresión, los errores estándar de los estimadores, la SCT, la SCE, la SCR y el R cuadrado.

Peso (Y)	Estatura (X)
132	68
108	64
102	62
115	65
128	66

- *FUENTE: Anderson (2008), Estadística para Economía y Administración, P 553.*

2. Inferencia Estadística

2.1 Pruebas de Hipótesis

Pruebas de Hipótesis

- ¿Los estimadores obtenidos por MCO reflejan el verdadero valor de los parámetros poblacionales?
- Por ejemplo, si beta es igual 0,2; ¿El beta poblacional en promedio también es igual a 0,2?
- Las pruebas de hipótesis evalúan estas afirmaciones asumiendo una **"distribución de probabilidad"** para los residuos.
- El modelo de regresión es **"normal"** porque se asume que los residuos siguen esa distribución.

Pruebas de Hipótesis

¿Por qué se asume la normalidad en los errores?

- Al aumentar el tamaño de muestra y sin importar el tipo de distribución que asuman las variables incluidas en el error estas siguen la normal.
- Lo anterior se cumple inclusive si el tamaño de muestra no es grande o las variables no son independientes (en algunos casos).
- Permite realizar pruebas de hipótesis con otras distribuciones (t-student, chi-cuadrado y F).
- Es importante realizar pruebas de hipótesis verificando la distribución normal de los errores.

Pruebas de Hipótesis

$$u_i \sim N(0, \sigma^2)$$

- Cuando los errores siguen la distribución normal entonces los estimadores por MCO también siguen esta distribución.
- Bajo este supuesto los estimadores por MCO son los Mejores Estimadores Insesgados **"MEI"** (entre los estimadores lineales y no lineales).
- Este resultado es más potente que el **"Teorema de Gauss Markov"**.

Pruebas de Hipótesis

- Cuando los errores siguen la distribución normal entonces los estimadores por MCO también siguen esta distribución:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right) \qquad \hat{\beta}_2 \sim N\left(\beta_2, \sigma^2_{\hat{\beta}_2}\right)$$

- Las variable normal estándar se puede construir a partir de estos estimadores al restar la media y dividir por la desviación estándar:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1) \qquad \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}2}} \sim N(0,1)$$

Pruebas de Hipótesis

- Dado que en la práctica es difícil encontrar la varianza poblacional se utiliza la **"distribución t"**, asumiendo que el error estándar es un buen estimador de la desviación estándar:

$$\frac{\hat{\beta}_1 - \beta_1}{ee\left(\hat{\beta}_1\right)} \sim t_{n-2} \qquad \frac{\hat{\beta}_2 - \beta_2}{ee\left(\hat{\beta}_2\right)} \sim t_{n-2}$$

- Las pruebas de hipótesis se pueden realizar bajo el enfoque del **"test de significancia"** o **"intervalos de confianza".**

Pruebas de Hipótesis

Pruebas de hipótesis

- Hipótesis Nula:

$$H_0 : \beta = 0{,}2$$

- Hipótesis Alternativa:

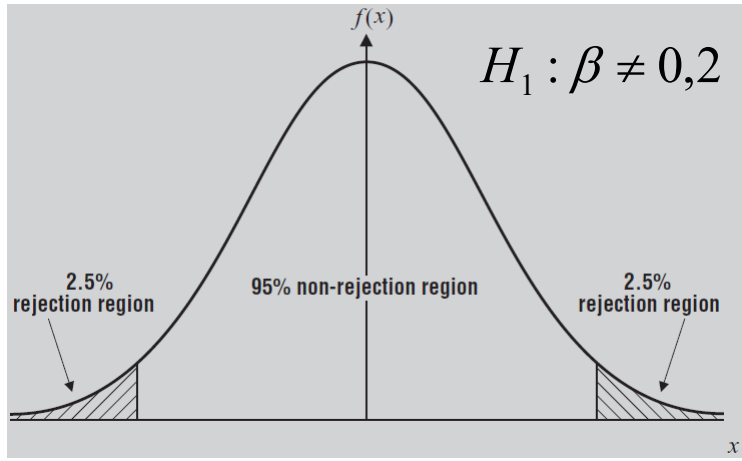
$$H_1 : \beta \neq 0{,}2$$

- En este caso se utiliza una prueba de dos colas dado que se tienen dos posibilidades:

$$\beta < 0{,}2 \qquad \beta > 0{,}2$$

Pruebas de Hipótesis

Regiones de rechazo al 5% con dos colas



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 57, Figure 2.13 (editada)*

Pruebas de Hipótesis

Pruebas de hipótesis

- Hipótesis Nula:

$$H_0 : \beta = 0,2$$

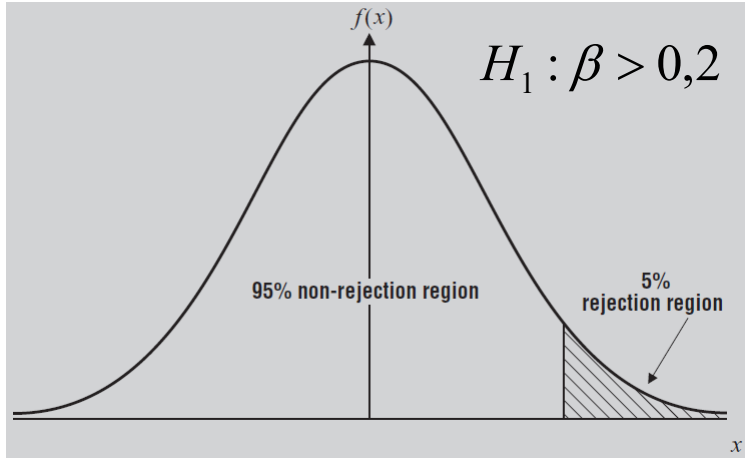
- Hipótesis Alternativa:

$$H_1 : \beta > 0,2$$

- En este caso se utiliza una prueba de una cola dado que se tiene una posibilidad.

Pruebas de Hipótesis

Región de rechazo al 5% con una cola



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 57, Figure 2.15 (editada)*

Pruebas de Hipótesis

Pruebas de hipótesis

- Hipótesis Nula:

$$H_0 : \beta = 0,2$$

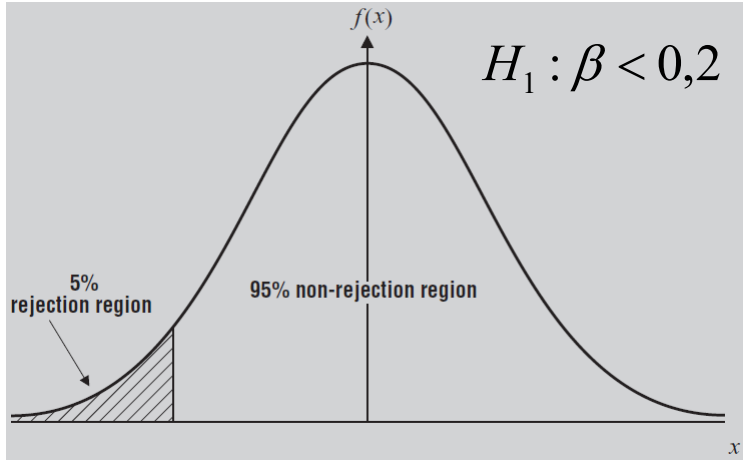
- Hipótesis Alternativa:

$$H_1 : \beta < 0,2$$

- En este caso se utiliza una prueba de una cola dado que se tiene una posibilidad.

Pruebas de Hipótesis

Región de rechazo al 5% con una cola



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 57, Figure 2.13 (editada)*

2. Inferencia Estadística

2.2 Test de significancia e Intervalos de Confianza

Test de significancia e Intervalos de Confianza

Test de significancia (pasos)

1. Estime los estimadores y sus errores estándar.
2. Calcule el valor del estadístico (observado).
3. Utilice una tabla de distribución del estadístico.
4. Seleccione un nivel de significancia.
5. Determine las regiones de rechazo.
6. Encuentre el valor crítico del estadístico.
7. Si el valor del estadístico es mayor o inferior a los valores críticos del estadístico rechace la hipótesis nula en caso contrario no rechace.

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 56, Box 2.5*

Test de significancia e Intervalos de Confianza

Intervalos de confianza (pasos)

1. Estime los estimadores y sus errores estándar.
2. Seleccione un nivel de significancia.
3. Utilice una tabla de distribución del estadístico.
4. Encuentre el valor crítico del estadístico.
5. Determine el intervalo de confianza.
6. Si el valor del estimador bajo la hipótesis nula se encuentra por fuera del intervalo rechace la hipótesis nula en caso contrario no rechace.

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 60, Box 2.6*

Test de significancia e Intervalos de Confianza

Equivalencia entre las pruebas

- Las conclusiones obtenidas por la prueba de significancia y los intervalos de confianza son equivalentes.
- Bajo el test de significancia la hipótesis nula no se rechaza si el estadístico se encuentra dentro de la región de no rechazo:

$$-t_{crt} \leq \frac{\hat{\beta}_2 - \beta_2}{ee\left(\hat{\beta}_2\right)} \leq t_{crt}$$

$$-t_{crt}\, ee\left(\hat{\beta}_2\right) \leq \hat{\beta}_2 - \beta_2 \leq t_{crt}\, ee\left(\hat{\beta}_2\right)$$

$$-\hat{\beta}_2 - t_{crt}\, ee\left(\hat{\beta}_2\right) \leq -\beta_2 \leq t_{crt}\, ee\left(\hat{\beta}_2\right) - \hat{\beta}_2$$

$$\hat{\beta}_2 + t_{crt}\, ee\left(\hat{\beta}_2\right) \geq \beta_2 \geq -t_{crt}\, ee\left(\hat{\beta}_2\right) + \hat{\beta}_2$$

$$\hat{\beta}_2 - t_{crt}\, ee\left(\hat{\beta}_2\right) \leq \beta_2 \leq \hat{\beta}_2 + t_{crt}\, ee\left(\hat{\beta}_2\right)$$

- Esta última expresión es el área de no rechazo de la hipótesis nula utilizando un intervalo de confianza.

2. Inferencia Estadística

2.3 Ejemplos

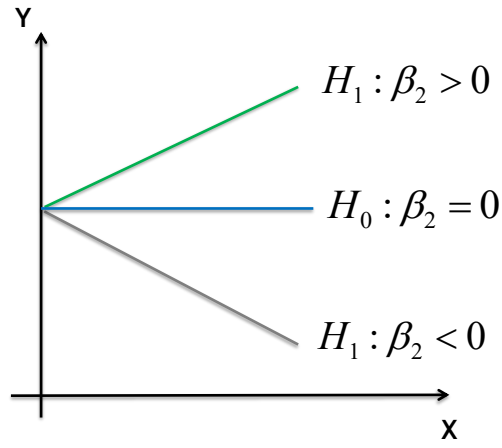
Ejemplo 1

Dada la siguiente regresión (Brooks 2008, P 60):

$$\underset{}{Y_i} = \underset{(14,38)}{20,3} + \underset{(0,2561)}{0,5091} X_i \qquad n = 22$$

- Hipótesis Nula: $H_0 : \beta_2 = 0$
- Hipótesis alternativa: $H_1 : \beta_2 \neq 0$
- Nivel de significancia = 10%
- Grados de libertad = 20
  - *¿Es una prueba de una cola o dos colas?, ¿Cuánto es el "t crítico"?*

Ejemplo 1



$$Y_i = \beta_1 + \beta_2 X_i$$

Ejemplo 1

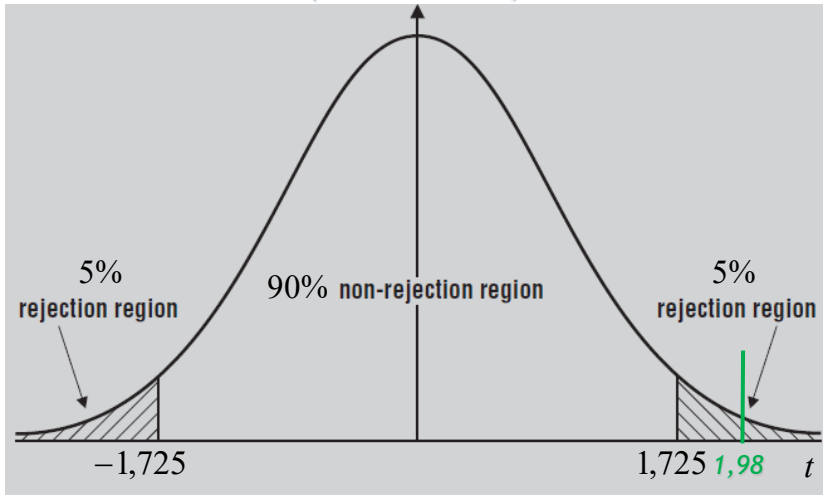
Test de significancia

$$t_{obs} = \frac{\hat{\beta}_2 - \beta_2}{ee\left(\hat{\beta}_2\right)} = \frac{0{,}5091 - 0}{0{,}2561} = 1{,}988$$

- Rechaza la hipótesis nula con un nivel de significancia del 10% (1,988 es superior a 1,725).

Ejemplo 1

Valores críticos y regiones de rechazo con dos colas (distribución t)



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 57, Figure 2.13 (editada)*

Ejemplo 1

$$t_{obs} = \frac{\hat{\beta}_2 - \beta_2}{ee\left(\hat{\beta}_2\right)} = \frac{0{,}5091 - 0}{0{,}2561} = 1{,}988$$

$t_{20,10\%} = \pm 1{,}725$ • Rechaza la hipótesis nula
(la variable es estadísticamente significativa)

$t_{20,5\%} = \pm 2{,}086$ • No rechaza la hipótesis nula

$t_{20,1\%} = \pm 2{,}845$ • No rechaza la hipótesis nula

Ejemplo 1

Intervalo de confianza

$$\hat{\beta}_2 - t_{crt}\, ee\left(\hat{\beta}_2\right) \leq \beta_2 \leq \hat{\beta}_2 + t_{crt}\, ee\left(\hat{\beta}_2\right)$$

$$0{,}5091 - (1{,}725)(0{,}2561) \leq \beta_2 \leq 0{,}5091 + (1{,}725)(0{,}2561)$$

$$0{,}06733 \leq \beta_2 \leq 0{,}9508$$

- Rechaza la hipótesis nula con un nivel de significancia del 10% ($\beta_2 = 0$ no se encuentra dentro de ese intervalo.)

Ejemplo 2

Dada la siguiente regresión (Gujarati 2010, P 79):

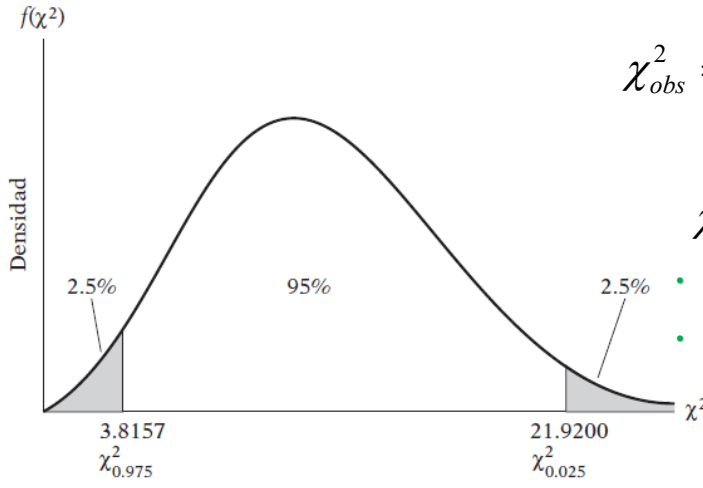
- X = nivel de escolaridad, Y = salario.

$$Y_i = \underset{(0{,}8746)}{-0{,}0144} + \underset{(0{,}069)}{0{,}7241} X_i \qquad n = 13$$

- Hipótesis Nula: $H_0 : \sigma^2 = 0{,}6$
- Hipótesis alternativa: $H_1 : \sigma^2 \neq 0{,}6$
- Nivel de significancia = 5%
- Grados de libertad = 11

Ejemplo 2

Valores críticos y regiones de rechazo con dos colas
Prueba Chi-cuadrado



- *FUENTE: Gujarati (2010), Econometría, P 112, Figura 5.1*

$$\chi^2_{obs} = (n-2)\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)$$

$$\chi^2_{CRT} \approx \chi^2_{n-2,\alpha/2}$$

- *Calcula el estadístico observado.*
- *¿Se rechaza la hipótesis nula?*

$$\hat{\sigma}^2 = 0{,}8811$$

Ejemplo 2

Intervalo de confianza

$$(n-2)\left(\frac{\hat{\sigma}^2}{\chi^2_{n-2,\alpha/2}}\right) \leq \sigma^2 \leq (n-2)\left(\frac{\hat{\sigma}^2}{\chi^2_{n-2,1-(\alpha/2)}}\right)$$

$$(11)\left(\frac{0{,}8811}{21{,}92}\right) \leq \sigma^2 \leq (11)\left(\frac{0{,}8811}{3{,}8157}\right)$$

$$0{,}4422 \leq \sigma^2 \leq 2{,}54$$

- No rechaza la hipótesis nula con un nivel de significancia del 5% ($\sigma^2 = 0{,}6$ se encuentra dentro de ese intervalo.)

2. Inferencia Estadística

2.4 Nivel de significancia

Nivel de significancia

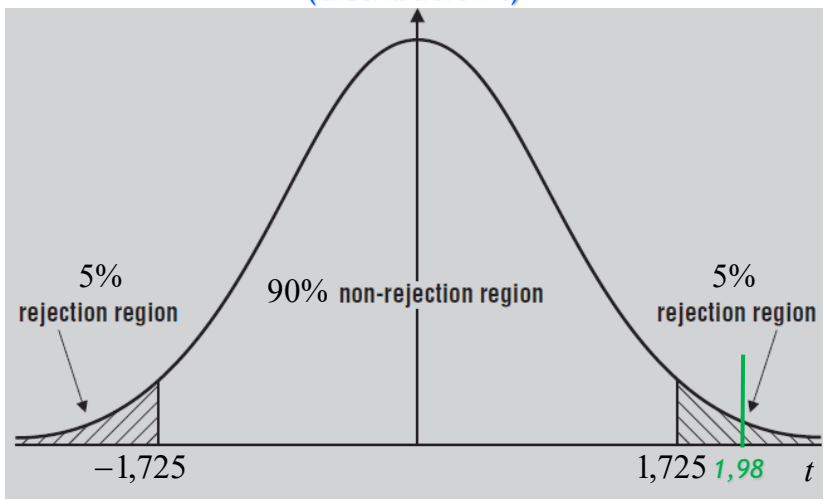
- No rechazar una hipótesis nula no implica aceptar la hipótesis nula porque en realidad cualquier otra hipótesis nula pudo haber sido aceptada (ejemplo de veredicto en juzgados).
- La hipótesis nula más utilizada en el análisis de regresión es $(H_0 : \beta_2 = 0)$ la cual busca establecer si existe una relación lineal entre la variable dependiente (Y) y la variable independiente (X).
- **Cuando los grados de libertad con superiores a 20, el nivel de significancia es del 5% y el estadístico "t" es superior a "2" rechace la hipótesis nula (regla del 2).**

Nivel de significancia

- La formulación de la hipótesis nula también depende del marco teórico existente (Gujarati 2010, P 121).
- El **"valor p"** encuentra el nivel de significancia más bajo para rechazar la hipótesis nula.
- **A medida que se incrementa el estadístico observado se reduce el valor p y se rechaza la hipótesis nula con mayor confianza (Gujarati 2010, P 122).**
- Los programas estadísticos muestran el valor del estadístico y el valor p.

Nivel de significancia

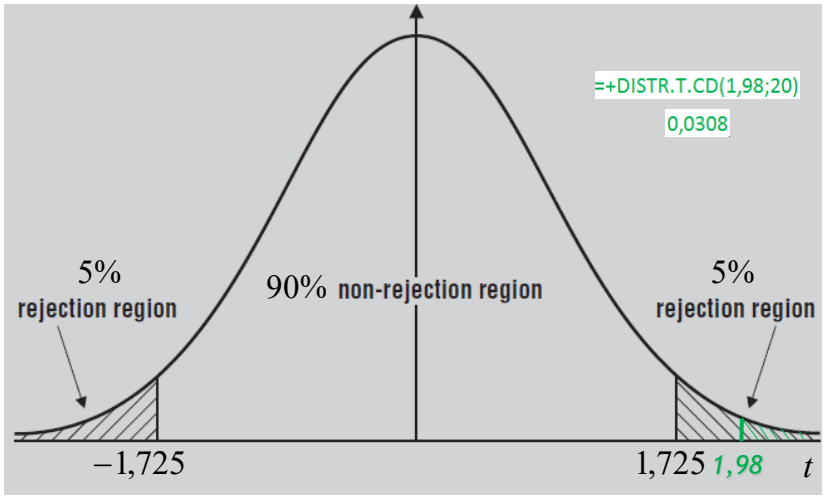
Valores críticos y regiones de rechazo con dos colas (distribución t)



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 57, Figure 2.13 (editada)*

Nivel de significancia

Rechazo de la hipótesis nula



- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 57, Figure 2.13 (editada)*

Nivel de significancia

- En esta situación:

$$t_{CRT} = t_{20,10\%} = \pm 1{,}725 \quad \Rightarrow \quad \frac{\alpha}{2} = 5\%$$

$$t_{obs} = \frac{\hat{\beta}_2 - \beta_2}{ee\left(\hat{\beta}_2\right)} = 1{,}988 \quad \Rightarrow \quad \text{ValorP} = 3{,}08\%$$

- *¿Se rechaza la hipótesis nula?, ¿Por qué?*

Nivel de significancia

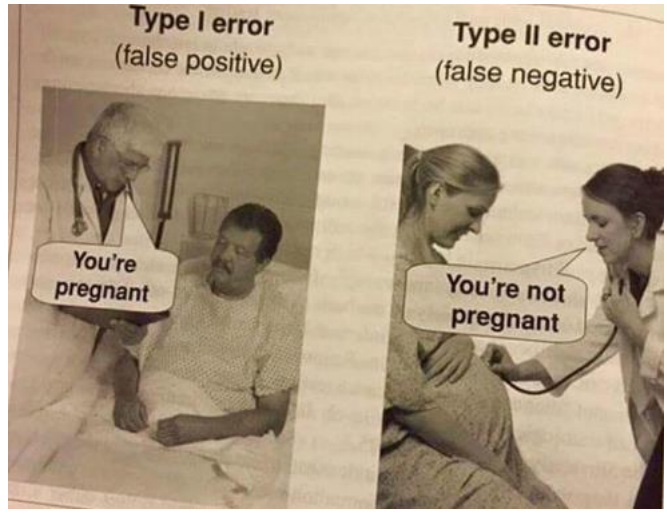
Clasificando los errores y aciertos al realizar pruebas de hipótesis

		Reality	
		H_0 is true	H_0 is false
Result of test	Significant (reject H_0)	Type I error = α	√
	Insignificant (do not reject H_0)	√	Type II error = β

- *FUENTE: Brooks (2008), Introductory Econometrics for Finance, P 64, Table 2.3*

Nivel de significancia

Error Tipo I y Error Tipo II



- *FUENTE: Ellis (2010), The essential guide to effect sizes, P 50, Figure 3.1*

Nivel de significancia

- Existe una disyuntiva entre los tipos de errores (es más grave el error tipo I).
- En la práctica se fija un valor para el error tipo I y se busca maximizar la potencia de la prueba ($1-\beta$).
- La única forma de reducir los dos tipos de errores en forma simultanea es utilizar una muestra más grande o con mayor varianza (Brooks 2008, P 65).

3. Análisis de varianza (*ANOVA*)

Análisis de varianza (ANOVA)

Medidas de bondad de ajuste (R cuadrado)

- Las variaciones de la muestra se generan de dos fuentes, la "suma explicada de cuadrados" (SCE) y la "suma de los residuos al cuadrado" (SCR):

$$Y_i = \left(\hat{\beta}_1 + \hat{\beta}_2 X_i \right) + \hat{u}_i$$

$$SCT = SCE + SCR$$

$$\sum \left(Y_i - \overline{Y} \right)^2 = \sum \left(\hat{Y}_i - \overline{Y} \right)^2 + \sum \left(Y_i - \hat{Y}_i \right)^2$$

Análisis de varianza (ANOVA)

Tabla ANOVA para regresión simple

Fuente de variación	SC	GL	SCP
Debido a la regresión (SCE)	$\sum \left(\hat{Y}_i - \overline{Y} \right)^2$	1	$\sum \left(\hat{Y}_i - \overline{Y} \right)^2$
Debido a los residuos (SCR)	$\sum \hat{u}_i^2$	n-2	$\sum \hat{u}_i^2 / n-2$
Suma Total de cuadrados (SCT)	$\sum \left(Y_i - \overline{Y} \right)^2$	n-1	$\sum \left(Y_i - \overline{Y} \right)^2 / n-1$

- *Encuentra la tabla ANOVA para el ejemplo del peso y la estatura.*
  - *FUENTE: Gujarati (2010), Econometría, P 125, Tabla 5.3*
  - *SC = Suma de Cuadrados.*
  - *GL = Grados de Libertad.*
  - *SCP = Suma de Cuadrados Promedio (SC dividido GL).*

Análisis de varianza (ANOVA)

Tabla ANOVA para regresión simple

Fuente de variación	SC	GL	SCP
Debido a la regresión (SCE)	605	1	605
Debido a los residuos (SCR)	51	3	17
Suma Total de cuadrados (SCT)	**656**	**4**	

Análisis de varianza (ANOVA)

Prueba F para regresión simple

$$F_{obs} = \frac{\frac{\chi_k^2}{k}}{\frac{\chi_m^2}{m}} = \frac{\text{Varianza 1}}{\text{Varianza 2}} = \frac{\text{GL Varianza 1}}{\text{GL Varianza 2}} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \frac{\frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{1}}{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

$$H_0 : \frac{\sigma_{SCE}^2}{\sigma_{SCR}^2} \leq 1 \qquad H_1 : \frac{\sigma_{SCE}^2}{\sigma_{SCR}^2} > 1$$

- *Encuentra el valor del "F obs" para el ejemplo del peso y la estatura.*

Fuente de variación	SC	GL	SCP
Debido a la regresión (SCE)	605	1	605
Debido a los residuos (SCR)	51	3	17
Suma Total de cuadrados (SCT)	**656**	**4**	

$$F_{obs} = \frac{605}{17} = 35{,}56$$

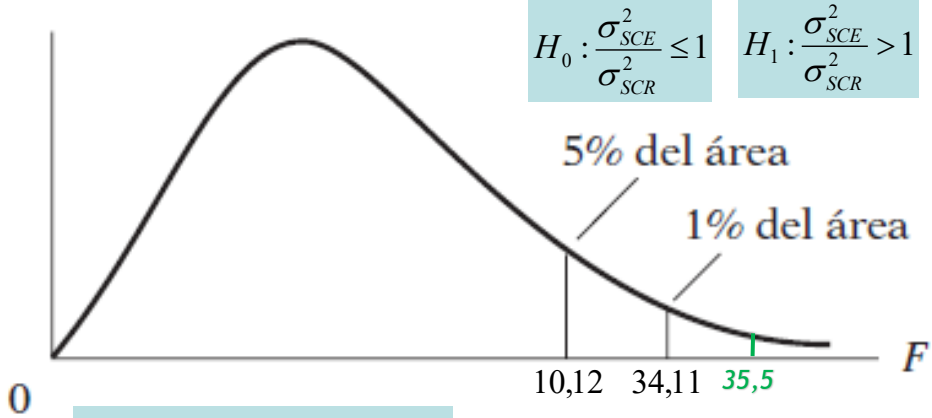
- SCE es 35 veces mayor a la SCR (descontando por GL), por tanto el modelo tiene alto poder explicativo (SCE pesa 92% en de la variabilidad de Y).

$$F_{obs} = \frac{R^2(n-k)}{(1-R^2)(k-1)} = \frac{(0{,}9222)(5-2)}{(1-0{,}9222)(2-1)} = 35{,}56$$

Distribución F

Valor crítico y región de rechazo con cola derecha

$$H_0 : \frac{\sigma^2_{SCE}}{\sigma^2_{SCR}} \leq 1 \qquad H_1 : \frac{\sigma^2_{SCE}}{\sigma^2_{SCR}} > 1$$



$$F_{CRT} = F_{1,n-2,\alpha} = F_{1,3,\alpha}$$

- *¿Se rechaza la hipótesis nula?*
- *¿El modelo tiene poder explicativo?*

- *FUENTE: Gujarati (2010), Econometría, P 880, Tabla D.3 (editada)*