Ciencia de Datos

Clase 4: Regresión Lineal

Roberto Ponce López

Tecnológico de Monterrey

rpl@tec.mx

3 de febrero del 2021

Agenda

Estimandos, Estimadores y Estimados

El objetivo de la inferencia estadística es parender acerca de distribuciones poblacionales no observadas, las cuales pueden ser caracterizadas por parámetros.

1. Estimandos son los parámetros poblacionales que buscamos estimar. Típicamente se escriben con letras griegas (μ, θ)
2. Estimadores son funciones de los datos muestrales (i.e. estadísticas) qu utilizamos para aprender acerca de los estimandos. Típicamente se denotan con un "sombrero"($\hat{\mu}$, $\hat{\theta}$)
3. Estimados son valores particulares de los estimadores que se obtienen a partir de una muestra específica

Ejemplo: ingreso de los hogares en 2008

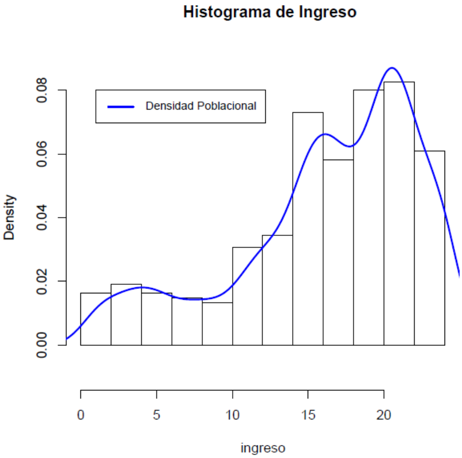


Figura: Distribución poblacional con media μ y varianza σ^2

Muestras aleatorias para estimar los parámetros poblacionales

Asume una población de una forma desconocida con media μ y varianza σ^2.
Ahora, veamos cuáles pueden ser los estimadores $\hat{\mu}$ para la media μ del ingreso en la población.

¿Cómo podemos utilizar datos para estimar μ?

Si pensamos que los datos fueron obtenidos de una muestra aleatoria de la distribución poblacional, entonces $Y_1, ..., Y_n$ son variables aleatorias independientes e identicamente distribuidas con $E[Y_i] = \mu$ y $V[Y_i] = \sigma^2$ para todas las $i \in 1, ..., n$

Nuestros estimadores, $\hat{\mu}$ y $\hat{\sigma}$ son funciones de $Y_1, ..., Y_n$ y por tanto serán variables aleatorias con sus propias distribuciones de probabilidad.

Propiedades de los Estimadores

Los estimadores son variables aleatorias. Esta aleatoriedad proviene del muestreo repetido de la población.

La distribución de un estimador debido al muestreo repetido se llama distribución muestral.

Las propiedades de un estimador se refieren a las características de su distribución muestral

Propiedades de los Estimadores (cont.)

Propiedades de muestras finitas (aplican para cualquier tamaño de muestra):

- Insesgado: ¿La probabilidad muestral de nuestro estimador está centrado en el verdadero valor del parámetro a estimar? $e[\hat{\mu}] = \mu$
- Eficiencia: ¿La varianza de la distribución muestral de nuestro estimador es razonablemente pequeña? $V[\hat{\mu}_1] < V[\hat{\mu}_2]$

Propiedades Asimptóticas (aplican para cualquier tamaño de muestra):

- Consistencia: conforme el tamaño de nuestra muestra aumenta hasta infinito, ¿la distribución muestral de nuestro estimador converge al valor verdadero del parámetro poblacional?
- Normalidad Asimptótica: conforme el tamaño de nuestra muestra aumenta, ¿la distribución muestral de nuestro estimador aproxima una distribución muestral?

Regresión Lineal

- La regresión lineal opera asumiendo una forma paramétrica lineal para la función de expectación condicional:
  $E[Y|X] = \beta_0 + X\beta_1$
- La expectación condicional definida por solamente dos coeficientes, los cuales son estimados de los datos:
  - β_0 es el intercepto o constante
  - β_1 es el coeficiente de la pendiente
- La función lineal impone una pendiente constante
- Supuesto: un cambio en $E[Y|X]$ es el mismo para todos los valores de X
- Geométricamente, la regresión lineal se observa como:
  - Una línea en casos con una sola variable X
  - Un plano en casos con dos variables X
  - Un hiperplano en casos con más de dos variables X

Ejemplo de Regresión Lineal

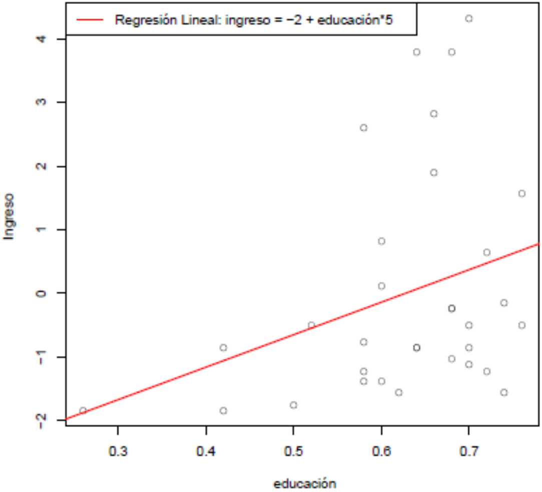


Figura: Ingreso regresado sobre años de educación

Interpretación de $\hat{\beta}_0$

Tenemos los coeficientes estimados de nuestra regresión:
$Y = \hat{\beta}_0 + \hat{\beta}_1 X$
$ingreso = -2 + 5 * educacion$

Sustituyendo 0 en X, tenemos que $E[Y|X] = \beta_0$

Nuestro estimado para el nivel promedio de ingreso es de -2 para individuos con educación igual a cero.

Interpretación de $\hat{\beta}_1$

Tenemos los coeficientes estimados de nuestra regresión:
$Y = \hat{\beta}_0 + \hat{\beta}_1 X$
$ingreso = -2 + 5 * educacion$ El signo del coeficiente de la pendiente

β_1 indica si $E[Y|X]$ crece o decrece con un cambio en X

- $\beta_1 > 0$: $E[Y|X]$ crece con X
- $\beta_1 < 0$: $E[Y|X]$ decrece con X
- $\beta_1 = 0$: $E[Y|X]$ no está linealmente relacionada con X

Interpretación de $\hat{\beta}_1$

Tenemos los coeficientes estimados de nuestra regresión:
$Y = \hat{\beta}_0 + \hat{\beta}_1 X$
$ingreso = -2 + 5 * educacion$ La magnitud del coeficiente de la pendiente β_1 indica que tan rápido $E[Y|X]$ crece o decrece con un cambio en X de una unidad

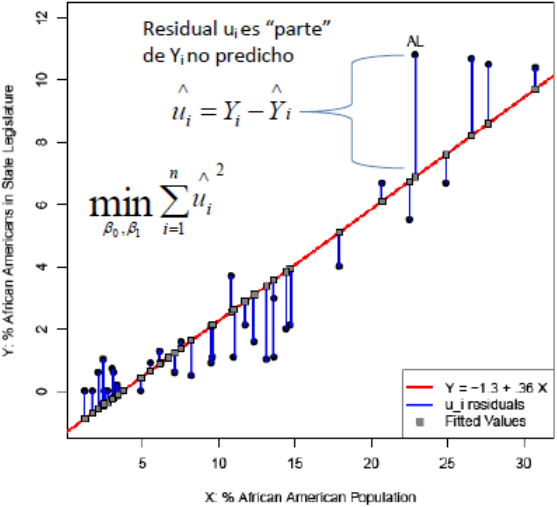
- El incremento de una unidad en X se asocia con un incremento de β_1 unidades en Y, en promedio
- El incremento de una unidad en educación se asocia con un incremento de 5 unidades en ingreso, en promedio

Predicción versus Causalidad

- La regresión lineal se puede utilizar para predecir nuevas observaciones.
  Ejemplo: Encontrar el nuevo valor esperado de Y para un punto no en la muestra con $X = x_{nuevo}$, calcula:
  $\hat{E}[Y|X = x_{nuevo}] = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
- La regresión lineal puede utilizarse para inferir causalidad, solamente si se cumple lo siguiente:
  - E[Y|X] es correctamente específicado como una función lineal (linealidad)
  - No hay otras variables que afecten ambas X y Y (exogeneidad)
  - Por ahora, pensemos en β como una cantidad a describir y predecir

Ordinary Least Squares (OLS)

Buscamos minimizar la suma del cuadrado de los residuales



Ordinary Least Squares (OLS)

- Tomemos $\hat{\beta}_0$ y $\hat{\beta}_1$ como posibles valores (estimadores) de β_0 y β_1
- La regresión OLS opera de la siguiente manera:

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \sum_{i=1}^{n} \hat{u}^2 = \underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x\hat{\beta}_1)^2$$

- ¿Por qué utilizar OLS?
  Fácil de derivar analíticamente. El resultado es insesgado y eficiente (bajo ciertos supuestos)

Derivación Analítica del OLS

- Tomemos $\hat{\beta}_0$ y $\hat{\beta}_1$ como posibles valores de β_0 y β_1
- La regresión OLS hace lo siguiente:

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}^2 = \arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x\hat{\beta}_1)^2$$

- ¿Por qué utilizar OLS?
  Fácil de derivar analíticamente. El resultado es insesgado y eficiente (bajo ciertos supuestos)

Derivación de los Estimadores

- Define la función objetivo de mínimos cuadrados:

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x\hat{\beta}_1)^2$$

- ¿Cómo derivamos los estimadores
  - Toma derivadas parciales de S con respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$
  - Iguala las derivadas parciales a cero
  - Resuelve para $\hat{\beta}_0, \hat{\beta}_1$

Derivación de los Estimadores

$$
\begin{aligned}
S(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x\hat{\beta}_1)^2 \\
&= \sum_{i=1}^{n} (y_i^2 - 2y_i\hat{\beta}_0 x_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2) \\
\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} &= \sum_{i=1}^{n} (-2y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 x_i) \\
\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} &= \sum_{i=1}^{n} (-2y_i x_i + 2\hat{\beta}_0 x_i + 2\hat{\beta}_1 x_i^2)
\end{aligned} \tag{1}
$$

Condiciones de Primer Orden

Las condiciones de primer orden son:

$$\begin{aligned}\sum_{i=1}^{n}(-2y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 x_i) = 0\\ \sum_{i=1}^{n}(-2y_i x_i + 2\hat{\beta}_0 x_i + 2\hat{\beta}_1 x_i^2) = 0\end{aligned} \tag{2}$$

Resolviendo para $\hat{\beta}_0$ y $\hat{\beta}_1$, arroja las siguientes ecuaciones (nota: $\sum_{i=1}^{n} y_i = N\bar{y}$):

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

$$N\hat{\beta}_0 = N\bar{y} - N\hat{\beta}_1\bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Condiciones de Primer Orden

Resolvemos para $\hat{\beta}_1$:

$$\begin{aligned}
\sum_{i=1}^{n}(-2y_ix_i + 2\hat{\beta}_0x_i + 2\hat{\beta}_1x_i^2) = 0 \\
\sum_{i=1}^{n}(y_ix_i - \hat{\beta}_0x_i - \hat{\beta}_1x_i^2) = 0 \\
\sum_{i=1}^{n}((y_ix_i) - (\bar{y} - \hat{\beta}_1\bar{x})x_i - \hat{\beta}_1x_i^2) = 0 \\
\sum_{i=1}^{n}(y_ix_i) - \bar{y}\sum_{i=1}^{n}x_i + \hat{\beta}_1\bar{x}\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\hat{\beta}_1x_i^2) = 0
\end{aligned} \tag{3}$$

Condiciones de Primer Orden

Finalmente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - N\bar{x}^2}$$

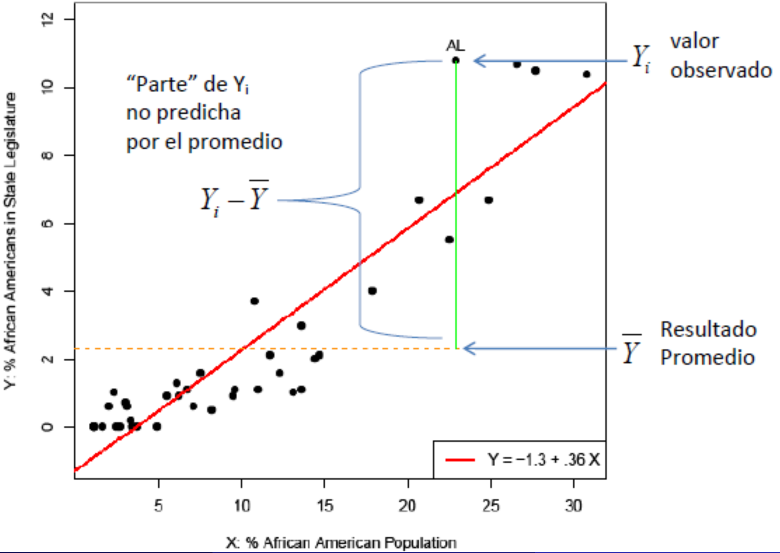
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\text{Covarianza muestral entre X y Y}}{\text{Varianza muestral de X}}$$

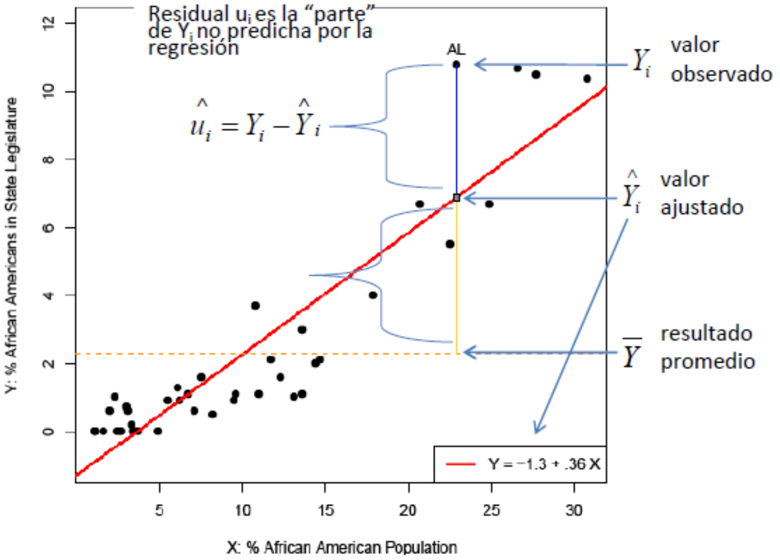
La definición del OLS implica las siguientes tres afirmaciones:

- $$\sum_{i=1}^{n} \hat{u}_i = 0$$
- $$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$
- $$\sum_{i=1}^{n} = \hat{y}_i \hat{u}_i = 0$$

Valores observados y valor promedio de los observados



Residuales y valores predichos



Análisis de la Varianza de la Regresión

- Total Sum of Squares:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = SST = Var[y]$$

- Explained/Model Sum of Squares

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SSE = Var[\hat{y}]$$

- Residual Sum of Squares

$$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{n}\hat{u}^2 = SSR = Var[\hat{u}]$$

- $SST = SSE + SSR$

Coeficiente de Determinación R^2

- Total Sum of Squares:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = SST = Var[y]$$

- Explained/Model Sum of Squares

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SSE = Var[\hat{y}]$$

- Residual Sum of Squares

$$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{n}\hat{u}^2 = SSR = Var[\hat{u}]$$

- $SST = SSE + SSR$

Coeficiente de Determinación R^2

Dado que $SST = SSE + SSR$, podemos dividir cada lado por SST:

$$\begin{aligned}\frac{SST}{SST} &= \frac{SSE}{SST} + \frac{SSR}{SST} \\ \frac{SST}{SST} &= 1 - \frac{SSR}{SST} = R^2\end{aligned} \tag{4}$$

Interpretación: Porcentaje de variación total en Y que es explicada por X.

Propiedades:

$$0 \leq R^2 \leq 1$$

- Si $R^2 = 1$, todos los puntos están sobre una linea recta (ajuste perfecto)
- Si $R^2 = 1$, no correlación entre Y y X
- *correlation is not causation*

¿Una R^2 mayor es mejor? ¡No!

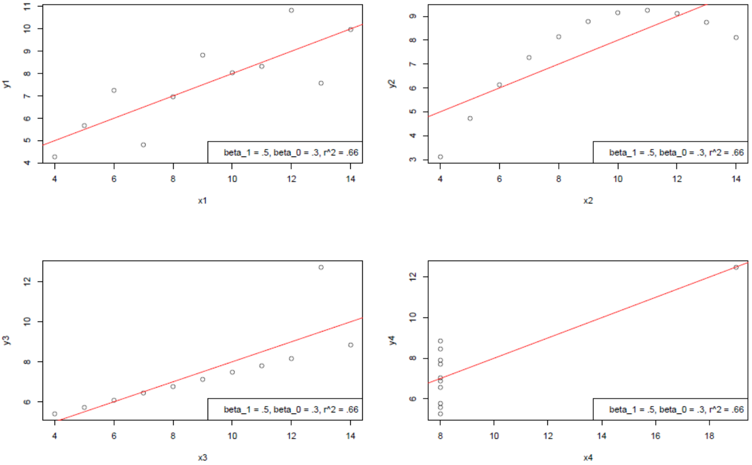


Figura: Cuarteto de Anscombe

Errores Estándar de $\hat{\beta}_0$ y $\hat{\beta}_1$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \Big[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\Big]$$

$$SE(\hat{\beta}_1)^2 = \sigma^2 \Big[\frac{\bar{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\Big]$$

dónde:

$$\sigma^2 = Var(u)$$

No conocemos σ, pero la podemos estimar a partir de los residuales:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{u}^2 \Big/ (n-2)$$

$$H0 : \text{No hay relación entre } X \text{ y } Y$$

$$H1 : \text{Hay alguna relación entre } X \text{ y } Y$$

Matemáticamente, corresponde a:

$$H0 : \beta_1 = 0$$

$$H1 : \beta_1 \neq 0$$

En la práctica, calculamos un estadístico t para realizar la prueba de hipótesis:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Resumen

- El modelo poblacional está dado por:

$$Y = \beta_0 + \beta_1 X + u$$

- Y es la variable dependiente
- X es la variable independiente
- β_0 es el intercepto y β_1 es la pendiente
- u es el término del error, una variable aletatoria que captura todos los elementos no observados que influyen en Y, además del regresor X
- El modelo estimado es:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u}$$

- Los residuales $\hat{u}$ son un estimado de u; es la variación en Y que no explica X

Supuestos del OLS

- Linealidad en los Parámetros: El modelo poblacional es lineal en sus parámetros y correctamente específicado
- Muestra Aleatoria: Los datos observados representan una muestra aleatoria de la población descrita en el modelo
- Variación en X: Hay variación en la variable independiente
- Media Condicional Cero: el valor esperado del termino error es cero, condicional sobre todos los valores de la variable independiente
- Homocedasticidad: el término del eerror tiene la misma varianza condicional sobre todos los valores de la variable independiente
- Normalidad: el término del error es independiente de las variables independientes y se distribuye de forma normal

Supuestos del OLS

Variación en X

x_i para $i = 1, ..., n$ no son el mismo valor

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

Supuestos del OLS

Media Condicional Cero

El valor esperado del término error es cero, condicional en cualquier valor de la variable X:

$$E[u|X] = 0$$

- $E[u|X] = 0$ implica la condición de $Cov(X, u) = 0$
- Asumiendo un muestreo aleatorio, $E[u|X] = 0$ también implica que $E[u_i|x_i]$ para todas las i

Violaciones al supuesto:

- u representa todas las variables no observadas que influyen en Y
- Si estos factores no observados también están correlacionados con X, entonces $Cov(X, y) \neq 0$
- Ejemplo: $ingreso = \beta_0 + \beta_1 * educacion + u$
- El supuesto se sostiene cuando hay una asignación aleatoria de X, es decir, con una asignación aleatoria del tratamiento.

Demostración de parámetros insesgados

Estimadores Insesgados del OLS

Tomando los supuestos del OLS: $E[\hat{\beta}_0] = \beta_0$ y $E[\hat{\beta}_1] = \beta_1$
Las distribuciones muestrales de los estimadores $\hat{\beta}_1$ y $\hat{\beta}_0$ están centradas en los parámetros poblacionales de β_0 y β_1

Prueba:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SST_x} \text{ (Nota: } \sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = 0 \text{)}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{SST_x}$$

$$= \frac{\beta_0 \sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$$

cont...

$$\hat{\beta}_1 = \frac{\beta_0 \cdot 0 + beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$$
$$= \frac{\beta_1 SST_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{STT_x}$$
$$= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$$

De esta forma,

$$E[\hat{\beta}_1|X] = E[\beta_1|X] + E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}|X\right]$$
$$= \beta_1 + \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{SST_x} \cdot E[u_i|X]$$
$$= \beta_1 \text{ (supuesto de media condicional de cero)}$$

...Continuación

Finalmente, por la ley de expectaciones iteradas:

$$E[\hat{\beta}_1] = E[E[\hat{\beta}_1|X]] = \beta_1$$

Supuestos del OLS

Homocedasticidad

La varianza condicional del término error es constante y no varía como función de la variable X:

$$Var[u|X] = \sigma_u^2$$

- Esto impica que $Var[u] = \sigma_u^2$. Todos los errores i tienen una varianza identica y constante $(\sigma_{ui}^2 = \sigma_u^2)$
- $E[Y1X] = \beta_0 + \beta_1 X$ y $Var[Y1X] = \sigma_u^2$
- Violaciones: $Var[u|X = x_1] \neq Var[u|X = x_2]$ se le llama heterocedasticidad
- Se puede corregir con el estimador sándwich

Heterocedasticidad

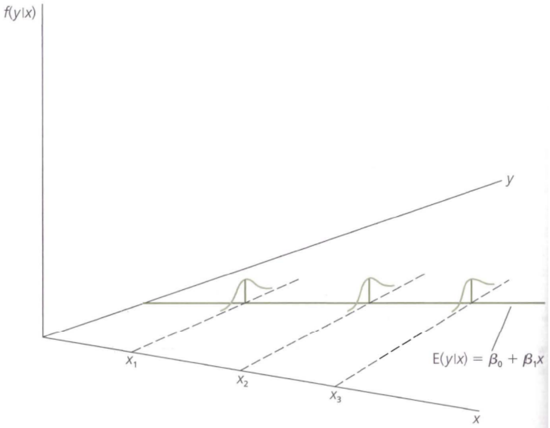


Figura: Ejemplo de Heterocedasticidad