Regresión Lineal Simple

Introducción a la Estadística

Fiona Franco Churruarín
fionafch96@gmail.com

UTDT

Febrero 2022

Hasta Ahora

Hasta ahora trabajamos definiendo un evento incierto, una variable aleatoria y ciertas propiedades como su esperanza o varianza.

Una vez hecho esto estudiamos herramientas con propiedades deseables para dar juicios de manera sistemática a los parámetros de alguna variable aleatoria.

Ahora modelaremos de manera explícita la relación entre dos variables aleatorias, e intentaremos adivinar ciertos parámetros que caracterizan esta relación.

Correlación

Recordemos algunas propiedades de la única medida de asociación entre dos variables aleatorias que vimos hasta ahora, la correlación:

- Un análisis de correlación expresa la relación entre dos series de datos utilizando un único número libre de unidades.
- El coeficiente de correlación es una medida de cuán estrechamente relacionadas con dos series de datos.
- El coeficiente de correlación mide la asociación lineal entre dos variables.
- El coeficiente de correlación nos dice si los puntos están todos cerca de una relación lineal, pero NO nos dice cual es esa relaci ón, ni puede indicar si existen relaciones no lineales

Motivación

Podríamos estar interesados en entender la relación entre años de educación y salario. Tenemos datos de ingresos (en $) y educación (en años) de 8 individuos:

Ingreso	**Educación**
8600	12
6000	7
8500	9
7600	3
11000	16
3000	5
7100	8
8500	11

Hasta aquí todo lo que sabemos decir es que, en esta muestra, el coeficiente de correlación entre años de educación y salario es 0.752.

Modelo de regresión lineal simple

Una forma mas sofisticada de entender la relación entre estas variables es **suponer** que a nivel poblacional se relacionan de la siguiente manera:

$$\text{Ingreso} = \beta_0 + \beta_1 \text{Educación} + \varepsilon \tag{1}$$

Donde ε representa todo lo que afecta al ingreso que no sean los años de educación (personalidad, contactos, experiencia, suerte, y demás). Ahora los **parámetros** que intentaremos aprender son β_0 y β_1

Una característica esencial que supondremos de ε es que no depende de los años de educación. Es decir, para cualquier nivel de años de educación supondremos que hay variables que hacen que las personas ganen un poco mas o un poco menos, representadas en ε, pero estos efectos 'en promedio' son cero. Es decir:

$$E(\varepsilon|\text{Educación}) = 0 \tag{2}$$

Modelo de regresión lineal simple

Notar que el modelo (1) junto con el supuesto (2) implican la siguiente relación:

$$E(\text{Ingreso}|\text{Educación}) = \beta_0 + \beta_1 \text{Educación} \tag{3}$$

Es decir, el ingreso esperado para un cierto nivel de años de educación esta dado por una constante β_0, mas otro parámetro β_1 multiplicado por los años de educación.

La ecuación (3) describe una **recta poblacional** que relaciona los años de educación con la **esperanza condicional** del ingreso, dados los años de educación.

Modelo de regresión lineal simple

Notemos la interpretación de los parámetros de este modelo:

$$E(\text{Ingreso}|\text{Educación} = 0) = \beta_0$$

El parámetro β_0 es el valor esperado ingreso condicional a tener 0 años de educación. Además:

$$\frac{\partial E(\text{Ingreso}|\text{Educación})}{\partial \text{Educación}} = \beta_1$$

El parámetro β_1 es el efecto marginal de años de educación sobre ingreso esperado. Si aumentan los años de educación en 1, el ingreso aumenta β_1 en valor esperado.

Modelo de regresión lineal simple

De manera mas general podemos escribir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde

- Y: variable dependiente o explicada
- X: variable explicativa o regresor
- β_0 y β_1: parámetros desconocidos
- ε: error aleatorio

Es importante destacar que estamos lejos de poder decir que X afecta a Y de manera *causal*. Todo lo que capturamos con este modelo son movimientos conjuntos.

No sabemos si una variable causa a la otra (o no), si hay una tercera que causa a las dos (o no), o si de casualidad estas dos variables tienen tendencias similares (recordar correlación espuria), o no.

Regresión lineal simple

Este modelo es bastante general y puede extenderse a todo tipo de variables, y podremos estimar sus parámetros de una forma no muy complicada siempre y cuando se mantenga la *linealidad* en los parámetros.

¿Qué representaría β_1 si la variable X fuera binaria? Por ejemplo considere $Y =$ Ingreso y X una variable binaria que toma el valor 1 para mujeres y 0 para hombres.

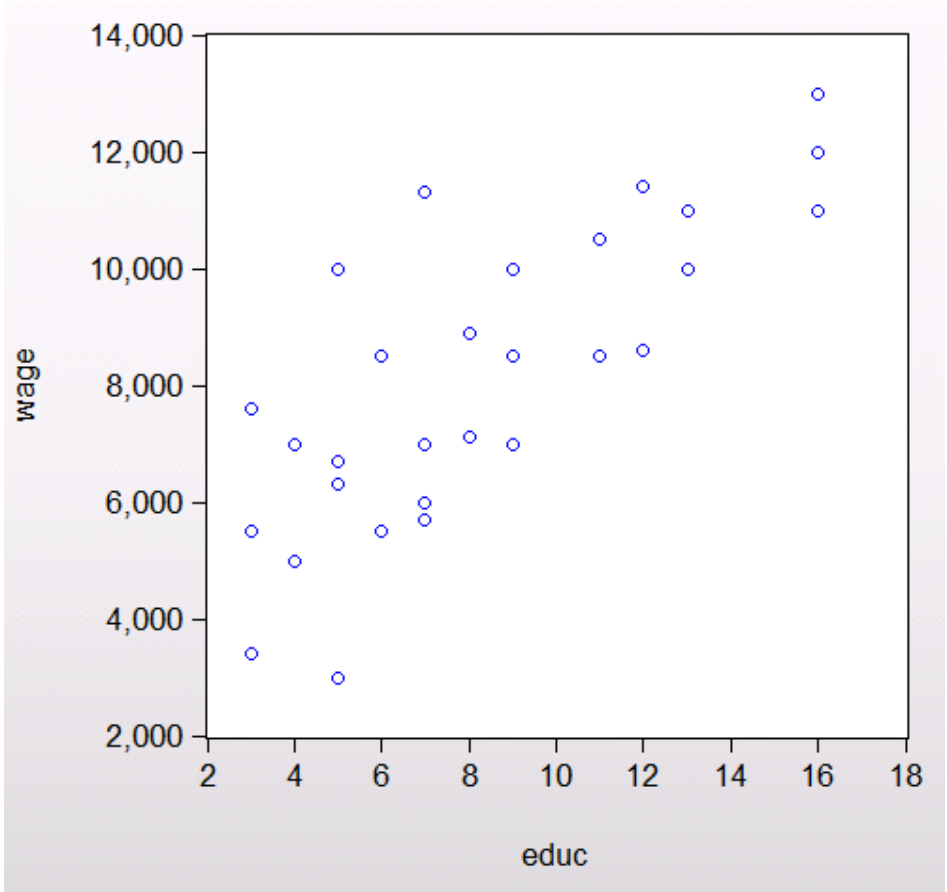
Regresión lineal simple

Algunas aplicaciones:

- Describir relaciones entre variables (no implica causalidad, sino cómo están relacionadas): cómo una variable puede ser explicada en función de otra(s).
- Para predecir (e.g. predecir cantidad demandada).
- Para estimar los parámetros (por ejemplo, un potencial efecto de tratamiento, o la elasticidad de una demanda).
- Para testear modelos (por ejemplo, para ver que modelo es más relevante en mis datos), y entender que variables tienen relaciones mas o menos fuertes que otras en términos de capacidad explicativa.

Población

Supongamos por un momento que la población está compuesta por 29 personas, y sus años de educación y salarios son:



Población

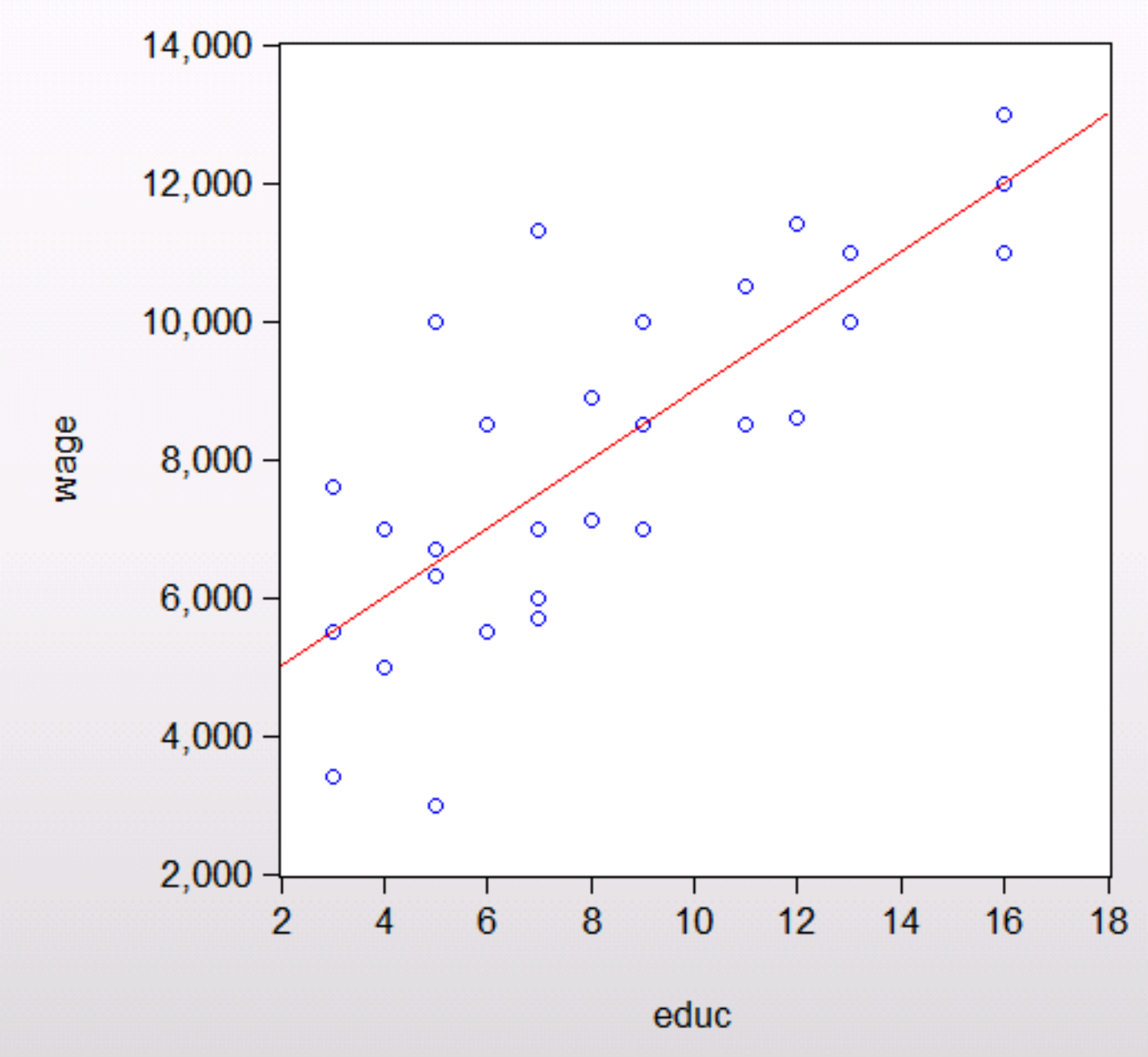
Por ejemplo, a nivel poblacional, para 9 años de educación hay 3 individuos con salarios de 7000, 8500, y 10000. Entonces:

$$E(\text{Ingreso}|\text{Educación}=9) = \frac{1}{3}7000 + \frac{1}{3}8500 + \frac{1}{3}10000 = 8500$$

En particular, a nivel poblacional, la relación entre esperanza condicional de ingreso y años de educación es:

$$\text{Ingreso} = 4000 + 500\text{Educación} + \varepsilon$$

Población


14,000
12,000
10,000
8,000
6,000
4,000
2,000
wage
2
4
6
8
10
12
14
16
18
educ

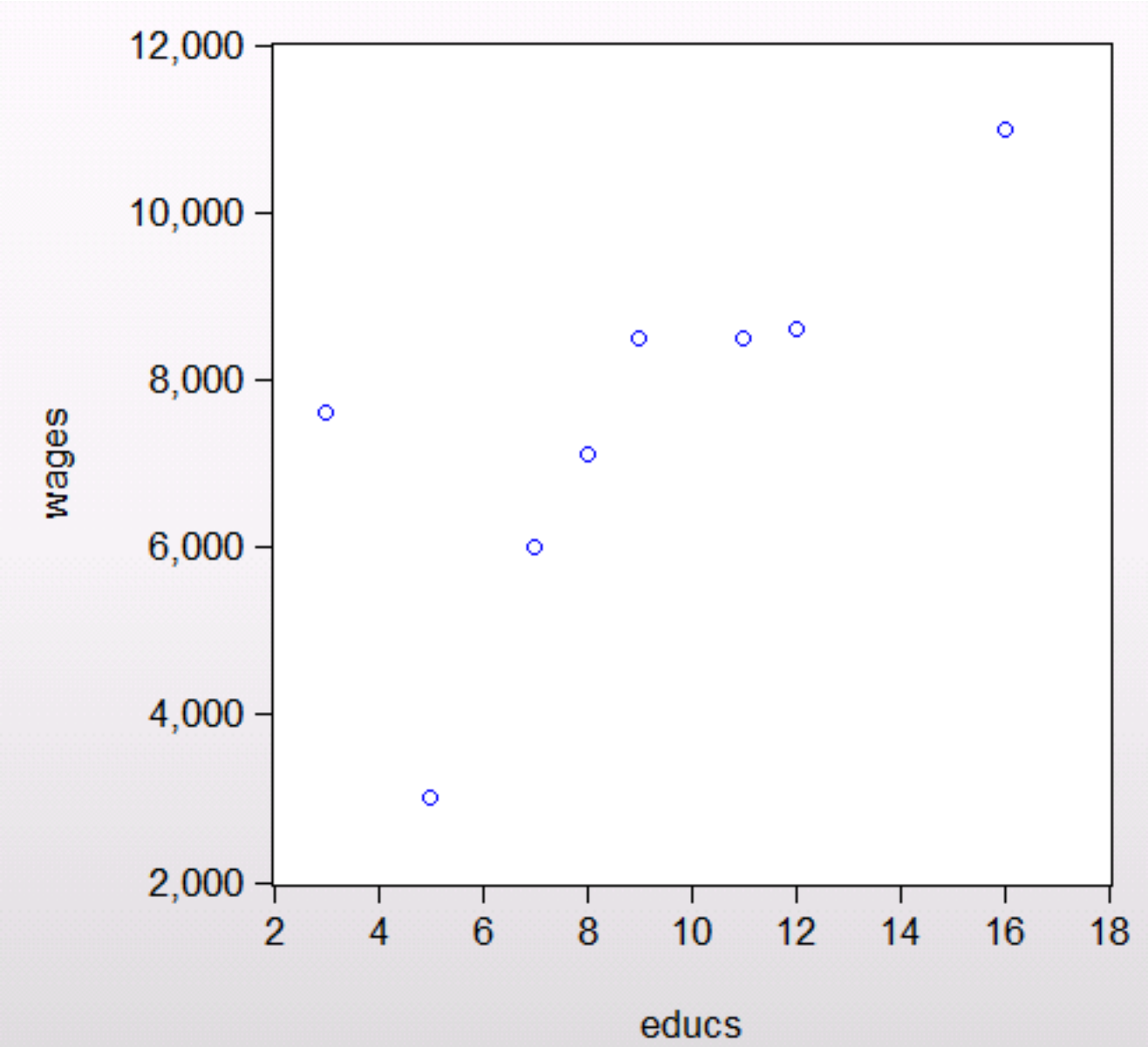
Muestra

Ahora bien, nosotros sólo trabajamos con una muestra de 8 observaciones:

Ingreso	**Educación**
8600	12
6000	7
8500	9
7600	3
11000	16
3000	5
7100	8
8500	11

La pregunta del millón es, ¿que tan bien podemos aprender la recta **poblacional** a partir de una **muestra**?

Muestra


12,000
10,000
8,000
6,000
4,000
2,000
wages
2
4
6
8
10
12
14
16
18
educs

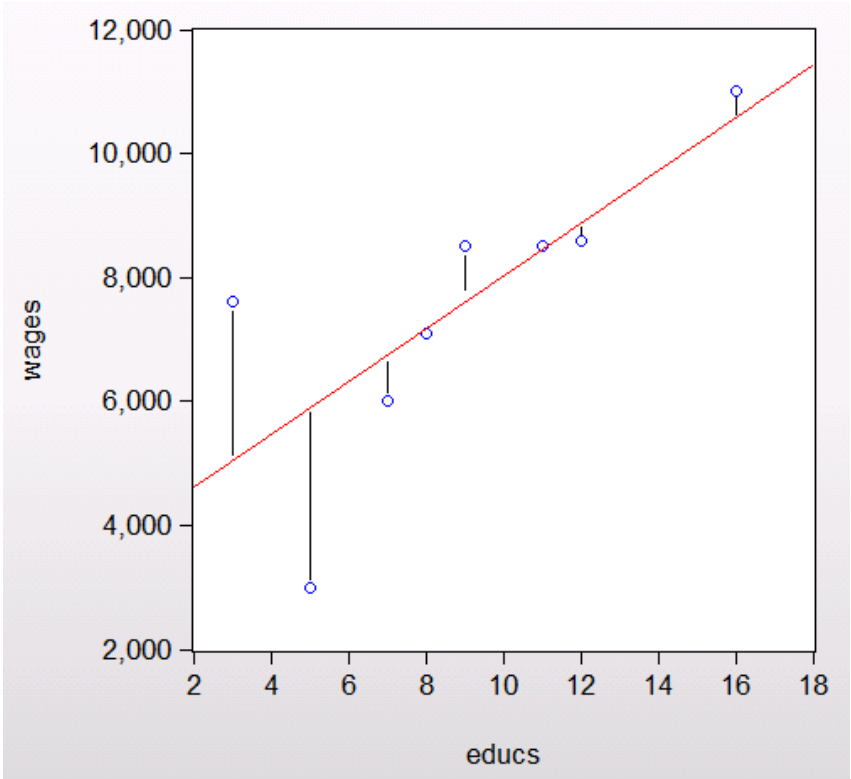
Muestra

Notar por ejemplo, que en la población $E(\text{Ingreso}|\text{Educación} = 5) = 6500$.

Pero en nuestra muestra no hay ninguna observación con 5 años de educación, por lo que con las herramientas que vimos hasta ahora no podríamos intentar adivinar la cantidad poblacional $E(\text{Ingreso}|\text{Educación} = 5)$.

Estimación

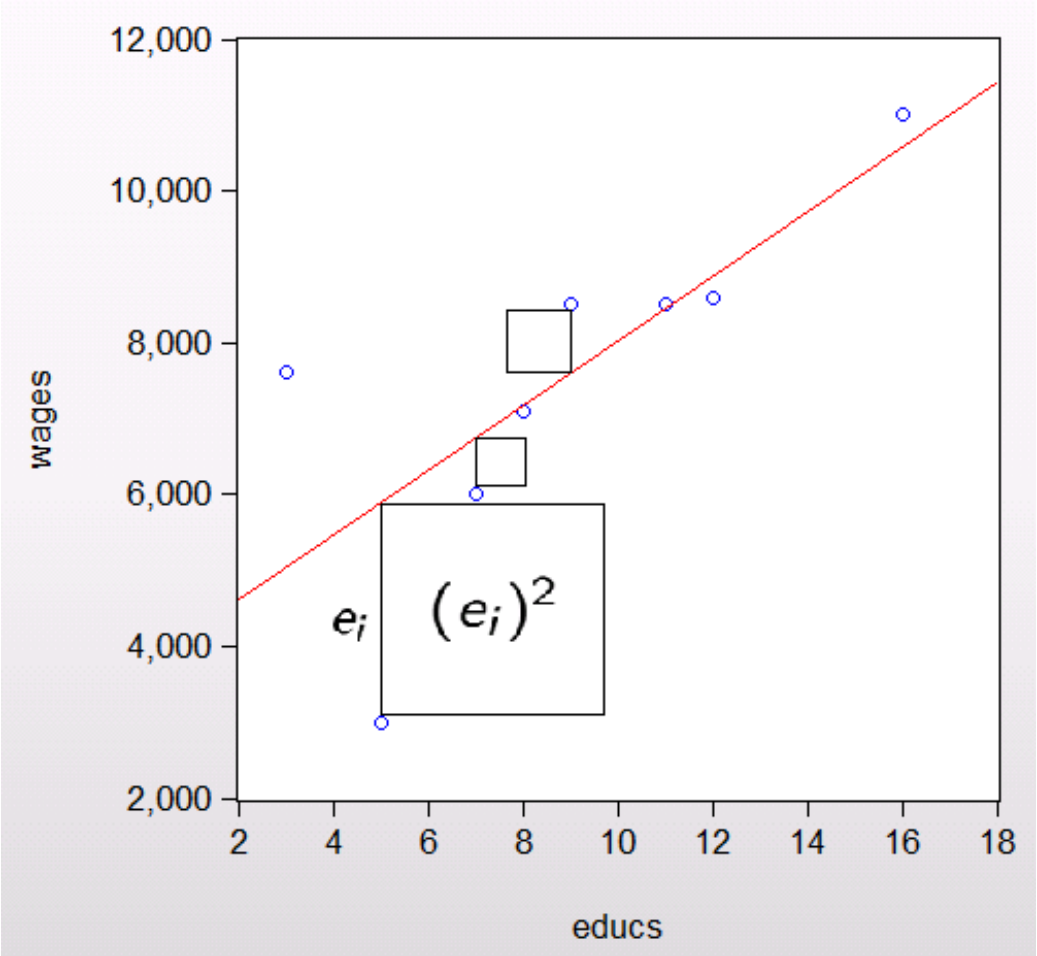
Nuevamente, debemos decidir como cuantificar el costo de errarle a algo que queremos adivinar. En este caso, miremos los datos de la muestra y propongamos una recta:



Al elegir una recta para estimar nos estamos comprometiendo a errarle a cada punto de la muestra por una distancia particular.

Criterio

El criterio que usaremos para estimar la recta poblacional será elegir la recta que haga las **distancias al cuadrado** entre la recta y las observaciones de la muestra lo mas pequeñas posibles:



Mínimos Cuadrados Ordinarios (OLS)

Entonces, elegiremos un 'intercepto' u ordenada al origen y una pendiente que resuelvan:

$$\min_{b_0,b_1} \sum_{i=1}^{n} (e_i)^2$$

O lo que es lo mismo:

$$\min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

MCO - OLS

Las condiciones de primer orden son:

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)X_i = 0$$

Reordenando y despejando se obtiene:

$$\hat{\beta}_1 = b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - \bar{X}\hat{\beta}_1$$

MCO - OLS

Para el caso de años de educación y salario podríamos realizar la estimación a mano:

Ingreso	**Educación**
8600	12
6000	7
8500	9
7600	3
11000	16
3000	5
7100	8
8500	11

MCO - OLS

obs	Y	X	$(Y-\overline{Y})$	$(X-\overline{X})$	$(X-\overline{X})^2$	$(Y-\overline{Y})(X-\overline{X})$
1	8600	12	1062.5	3.125	9.766	3320.3
2	6000	7	-1537.5	-1.875	3.516	2882.8
3	8500	9	962.5	0.125	0.016	120.3
4	7600	3	62.5	-5.875	34.516	-367.2
5	11000	16	3462.5	7.125	50.766	24670.3
6	3000	5	-4537.5	-3.875	15.016	17582.8
7	7100	8	-437.5	-0.875	0.766	382.8
8	8500	11	962.5	2.125	4.516	2045.3
sum	60300	71			118.875	50637.5
avg	7537.5	8.875				

Entonces $\hat{\beta}_1 = \frac{50637.5}{118.875} = 425.97$ y $\hat{\beta}_0 = 7537.5 - 8.875 * 425.97 = 3756.99$

MCO - OLS

Entonces nuestra **recta estimada** es:

$$\widehat{\text{Ingreso}}_i = 3756.99 + 425.97 \cdot \text{Educación}_i$$

Esto quiere decir que, según esta muestra, cuando años de educación aumenta en una unidad, el ingreso aumenta en \$425.97 en promedio.

Notar que no estamos declarando una relación causal de años de educación con ingreso, simplemente medimos correlaciones (o covarianzas). Ver modelo de Spence de señalización.

Bondad de ajuste

Una pregunta relevante es que tan bien el modelo ajusta a los datos. Es decir, esa ecuación que impusimos y relaciona a X con Y, ¿que tanto parece cumplirse en los datos?

Para responder a estas preguntas podemos calcular:

- el R^2
- el error estándar de la regresión (SER).

R^2

Una de las medidas (clásicas) más utilizadas respecto a cuán bueno es el ajuste del modelo es el R^2.
Se puede descomponer la variabilidad total de la variable dependiente en variabilidad **explicada** por el modelo y variabilidad **no explicada** por el modelo.

$$STC = SEC + SRC$$

Es la medida más utilizada, pero no la única y se calcula como la fracción de la varianza muestral de y que está siendo explicada por X.

$$R^2 = \frac{SEC}{STC} = \frac{\sum(\widehat{y_i} - \overline{y})^2}{\sum(y_i - \overline{y})^2} \tag{4}$$

$$= 1 - \frac{SRC}{STC} = 1 - \frac{\sum e_i^2}{\sum(y_i - \overline{y})^2} \tag{5}$$

El R^2 indica qué proporción de la variabilidad muestral de Y está siendo explicada por el modelo, es decir, por X. En el caso de la regresión simple, $R^2 = r_{XY}^2$. Está acotada entre 0 y 1.

El SER

El **SER (error estándar de la regresión)** es una estimación de la desviación estándar del error de la regresión (e).

Notar que Y y e tienen las mismas unidades de medida.

Entonces, el SER es una medida del desvío de las observaciones respecto de la recta estimada, medida en las unidades en que está medida la variable dependiente.

$$SER = S_e, \qquad S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

MCO - OLS

Volvamos al estimador de MCO...

Algunas propiedades deseables de los estimadores por MCO. Nos interesa que tengan:

- el menor sesgo posible
- la mayor precisión (menor varianza)

Para ver esto, se necesita de **supuestos** ademas de linealidad:

(1) $E(\varepsilon|\boldsymbol{X}) = 0 \rightarrow$ exogeneidad

(2) $E(\varepsilon^2|\boldsymbol{X}) = \sigma^2 \rightarrow$ homocedasticidad

(3) $E(\varepsilon_i \cdot u_j|\boldsymbol{X}) = 0 \, \forall i \neq j \rightarrow$ ausencia de autocorrelación

(4) $\varepsilon \sim \mathcal{N}(0, \sigma^2) \rightarrow$ normalidad de los errores

MCO - OLS

Bajo estos supuestos, se conoce la distribución de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ Conocer la distribución del estimador MCO nos permite hacer inferencia (pruebas de hipótesis e intervalos de confianza).

Además, si los supuestos se cumplen, el estimador MCO es el mejor estimador de todos los posibles estimadores lineales (MELI):

- Insesgado
- Eficiente
- Consistente

Supuestos clásicos

Bajo los supuestos clásicos se puede probar que:

1. MCO es un estimador **insesgado**:
   - $E(\hat{\beta}_1) = \beta_1$
   - $E(\hat{\beta}_0) = \beta_0$

2. Es un estimador **consistente:**
   - $plim(\hat{\beta}_1) = \beta_1$
   - $plim(\hat{\beta}_0) = \beta_0$

3. Tiene varianzas dadas por:
   - $Var(\hat{\beta}_1) = \sigma^2_\varepsilon \frac{1}{\sum(X_i - \overline{X})^2}$
   - $Var(\hat{\beta}_0) = \sigma^2_\varepsilon \frac{\sum X_i^2}{n \sum(X_i - \overline{X})^2}$

4. Es BLUE o MELI - Teorema de Gauss-Markov.

Supuestos clásicos

Las varianzas del estimador MCO son desconocidas porque no se conoce σ_ε^2.

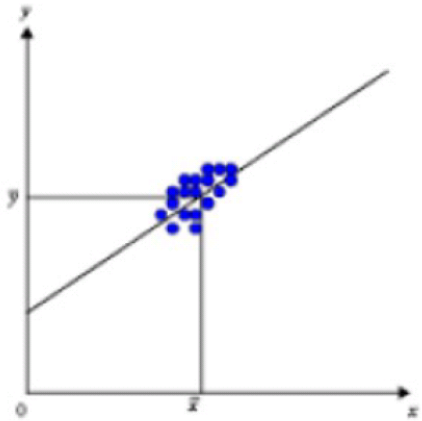
Pero como vimos antes, se puede estimar con:

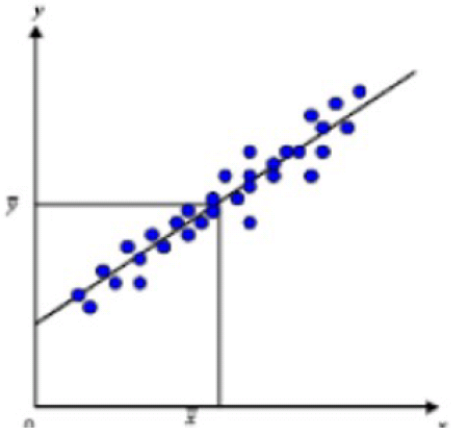
$$S_\varepsilon^2 = \frac{\sum e_i^2}{n-2}$$

Varianza del estimador

Notar que la varianza de los estimadores depende de la varianza del error y de la variabilidad muestral de X

- A mayor varianza del error, mayor varianza tendrán los estimadores (menor precisión). También se lo llama heterogeneidad no observada.
- A mayor varianza de X, menor varianza del estimador.
  - Idea: si X no tiene mucha variación nuestra estimación será poco precisa (se requiere variación para identificar).





En un caso extremo, ¿podríamos entender algún movimiento conjunto de años de educación y salario si en la muestra sólo tenemos personas con 6 años de educación?

Distribución de $\hat{\beta}$

Aún sin suponer normalidad de ε, si la muestra es lo suficientemente grande, por TCL:

$$\hat{\beta}_0 \overset{aprox}{\sim} N(\beta_0, \sigma^2_{\beta_0})$$

$$\hat{\beta}_1 \overset{aprox}{\sim} N(\beta_1, \sigma^2_{\beta_1})$$

Para hacer inferencia se necesita saber qué distribución siguen los estimadores.

¡Entonces podemos hacer intervalos de confianza y pruebas de hipótesis!

Inferencia

No estudiaremos aquí todos estos casos, pero para ilustrar la relevancia del resultado anterior, podemos testear:

- Significatividad individual: $H_0 : \beta_1 = 0$
- Significatividad conjunta: $H_0 : \beta_i = 0 \quad \forall i$
- Hipótesis que querramos testear respecto a los valores que desconocemos de los parámetros:

$$H_0 : \beta_1 \geq k$$

$$o$$

$$H_0 : \beta_1 \geq \beta_2$$

- Ver si los resultados obtenidos son consistentes con los supuestos realizados (validar el modelo) (e.g. ¿los residuos parecen normales?)

Tests para algún β

$$H_0 : \beta_1 = \beta_{1,H_0}$$
$$H_1 : \beta_1 \neq \beta_{1,H_0}$$

Estadístico:

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{\hat{S}_{\hat{\beta}_1}} \sim t_{n-k-1}$$

Sigue una distribución t de Student y no normal porque $\sigma^2_{\hat{\beta}}$ es desconocido entonces lo estimamos usando $S^2_{\hat{\beta}}$. Los grados de libertad son $n-k-1$ donde k es la cantidad de variables explicativas que utilizamos. En nuestro caso $k = 1$.

Nos referimos al **Test de Significatividad Individual** en el caso particular de $\beta_{1,H_0} = 0$. Es de los tests mas directos y utilizados en este contexto.

Intervalos de confianza: $\hat{\beta}_1 \pm t_c \times \hat{S}_{\hat{\beta}_1}$

Prueba de significatividad global

Cuando hay muchos regresores, podemos preguntarnos si todos los coeficientes son iguales a cero al mismo tiempo:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \text{algún } \beta_i \neq 0, \ \forall i = 1, ..., k$$

Hipótesis que se evalúa con un estadístico F (lo veremos más adelante en regresión múltiple).

Volviendo al primer ejemplo

En cualquier software estadístico (Stata, EViews, R) puede ejecutarse una regresión lineal simple para un conjunto de datos, y el output es de la forma:

Dependent Variable: WAGES
Method: Least Squares
Date: 02/19/20 Time: 15:32
Sample (adjusted): 1 8
Included observations: 8 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3756.993	1473.856	2.549090	0.0435
EDUCS	425.9727	152.3208	2.796550	0.0313

R-squared	0.565868	Mean dependent var	7537.500
Adjusted R-squared	0.493513	S.D. dependent var	2333.567
S.E. of regression	1660.751	Akaike info criterion	17.88025
Sum squared resid	16548559	Schwarz criterion	17.90011
Log likelihood	-69.52098	Hannan-Quinn criter.	17.74629
F-statistic	7.820689	Durbin-Watson stat	1.765233
Prob(F-statistic)	0.031306		

Bibliografía sugerida

- Greene - Econometric Analysis.
- Wooldridge - Cap. 2 (simple) y Cap. 3 (múltiple)
- Stock & Watson - Cap. 4 y 5.