

Econometría I

Mínimos Cuadrados Ordinarios



Carlos A. Yanes | Departamento de Economía | 2024-02-24





Preguntas de la sesión anterior?

Preliminar

La última vez:

1. Hasta el momento hemos hablado de estadísticas.
2. Hoy hablaremos mejor de las condiciones **MELI** de un estimador
3. Vamos a mirar algunas líneas de código en **R**
4. Para eso pensaremos en eventos con **muestras de datos**.

Modelo Poblacional vs Muestral



Modelo Poblacional vs Muestral

Podemos tener un modelo **Poblacional**

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Y uno **Muestral** de la siguiente forma

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

Un **modelo de regresión** produce un estimador por cada observación

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

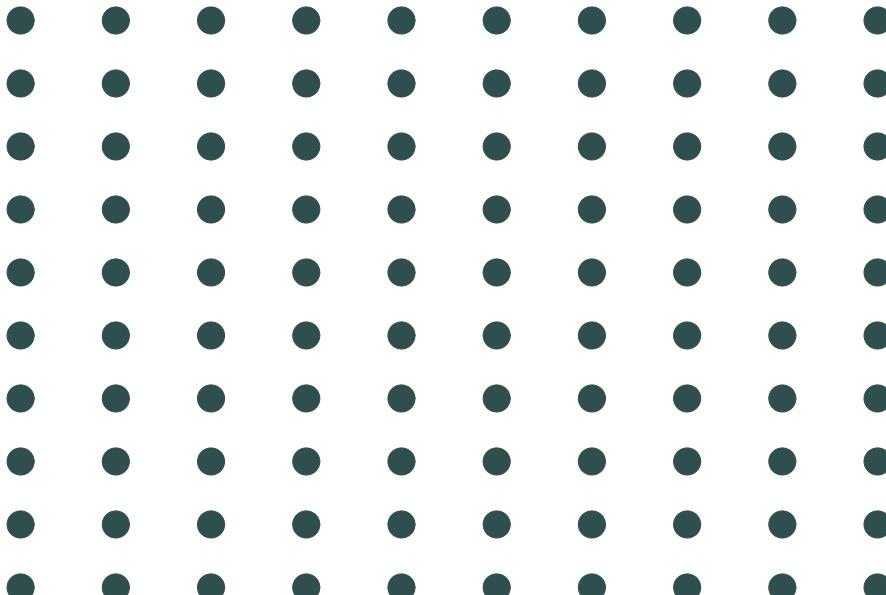
El cual nos dará el *mejor-ajuste* lineal a partir de nuestros datos.

Población vs. Muestra

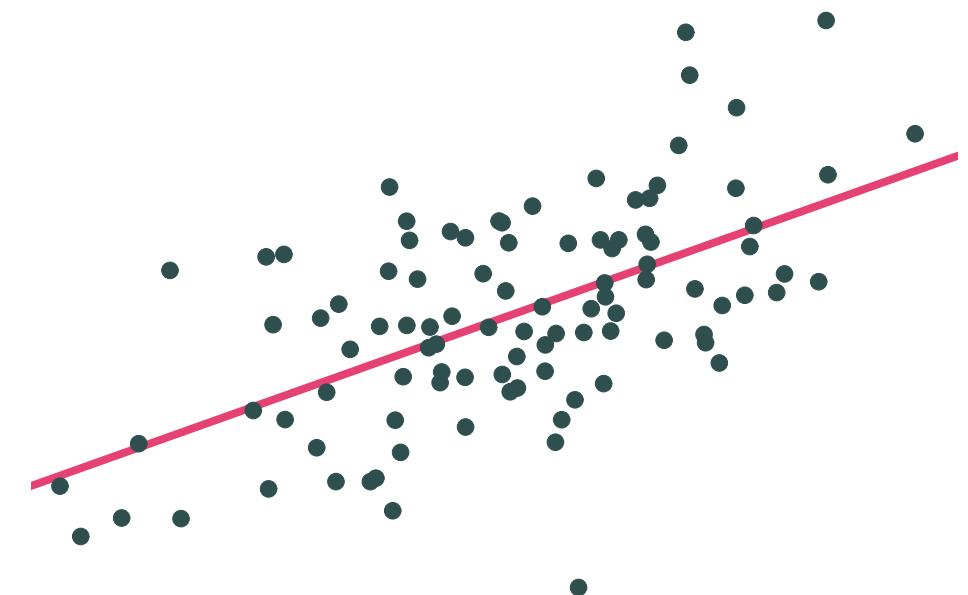


Población vs. Muestra

Pregunta: Por qué nos preocupa eso de la *población vs. muestra*?



Población

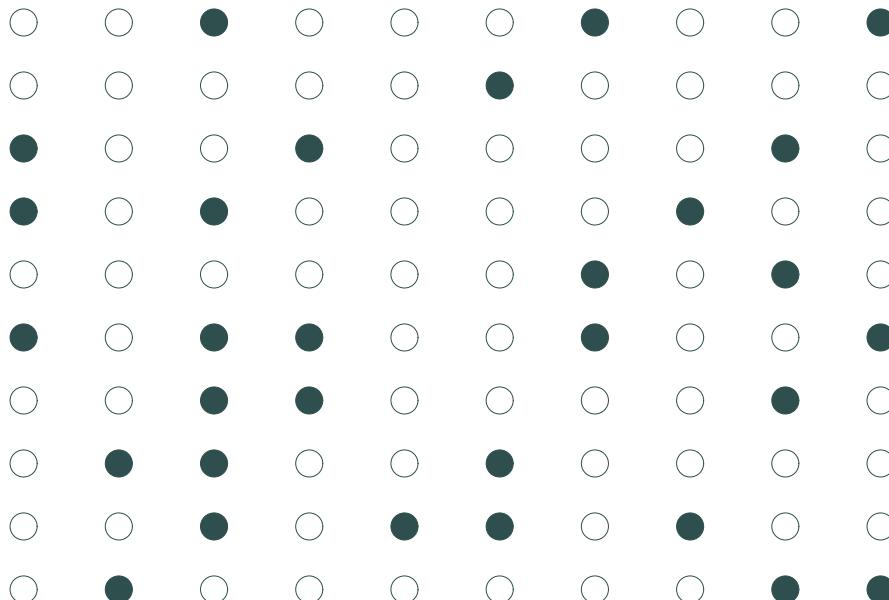


$$y_i = 2.53 + 0.57x_i + u_i$$

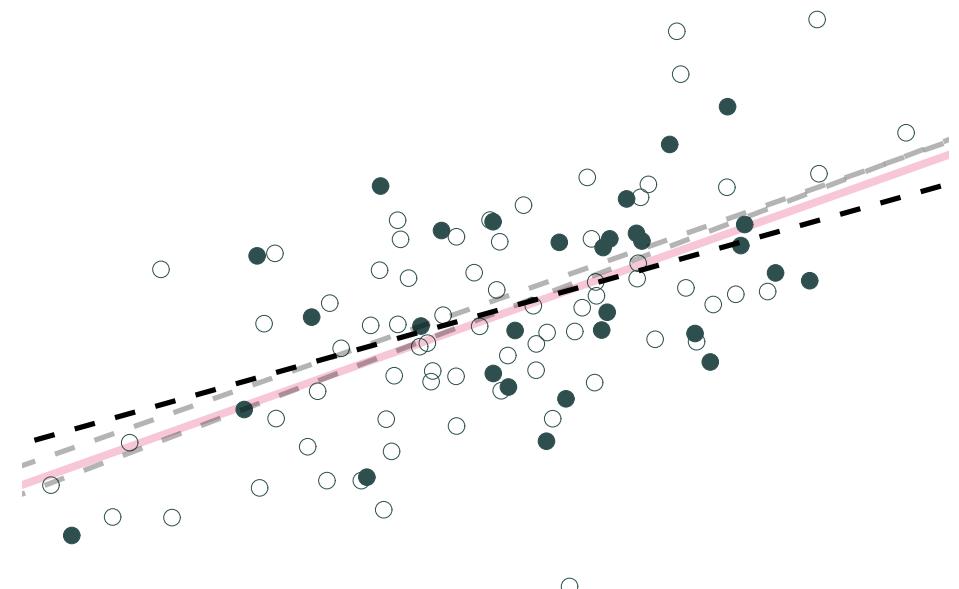
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Población vs. Muestra

Pregunta: Por qué nos preocupa eso de la *población vs. muestra*?



Muestra 3: 30 individuos aleatorios



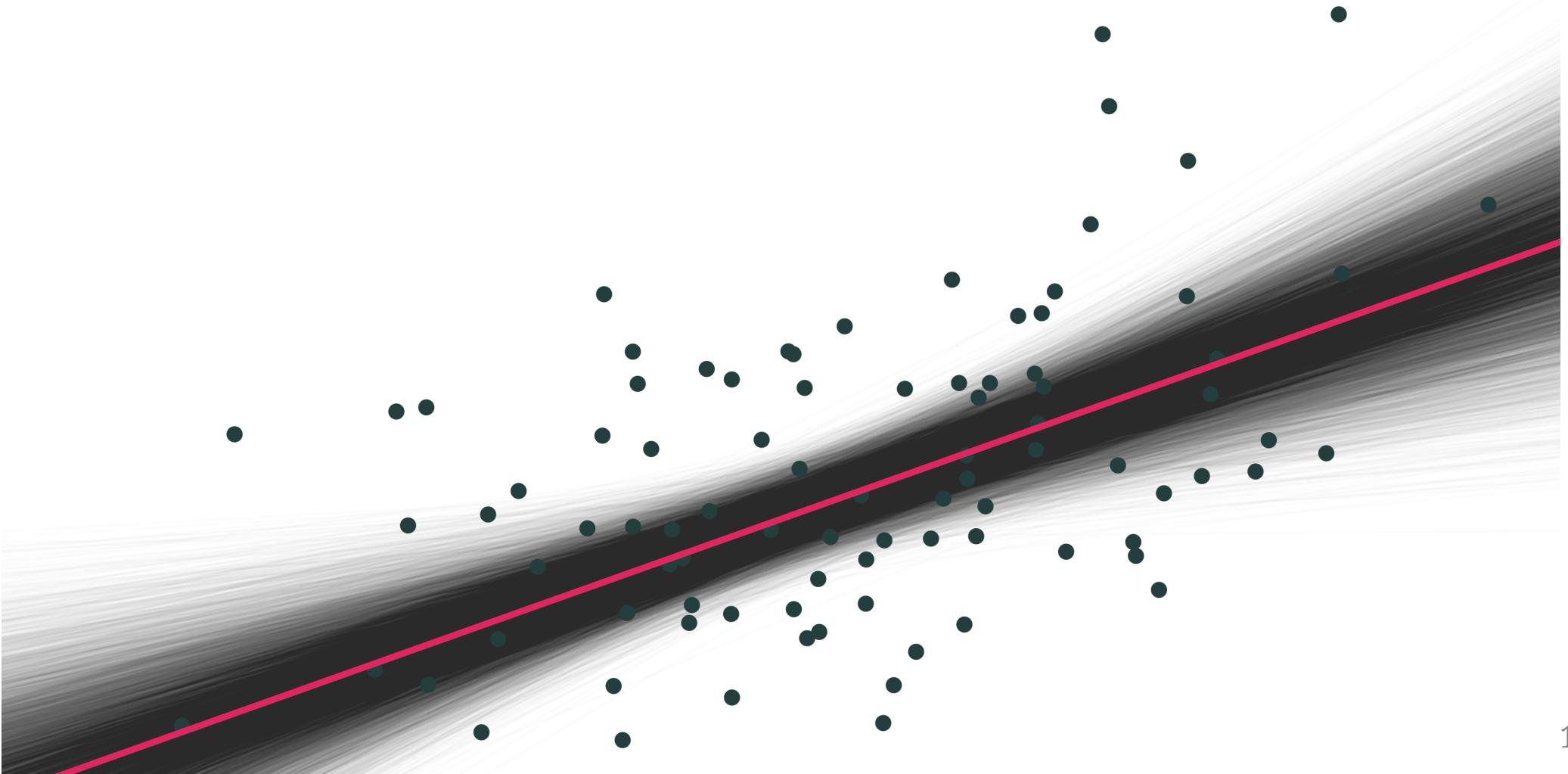
Relación Poblacional
 $y_i = 2.53 + 0.57x_i + u_i$

Relación Muestral
 $\hat{y}_i = 3.21 + 0.45x_i$

Podemos repetir esto **10,000 veces**.

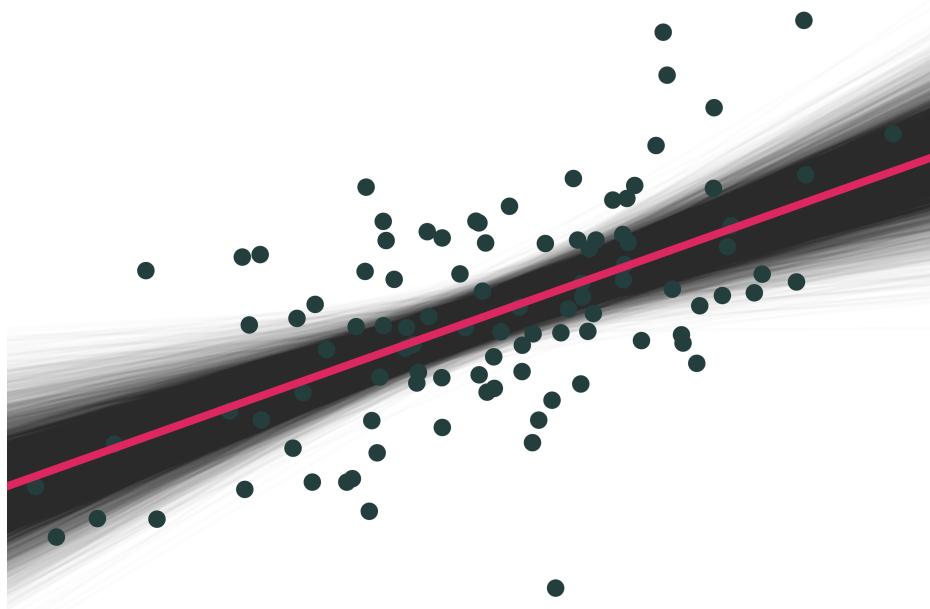
(Este ejercicio se llama simulación de (Monte Carlo))

Población vs. Muestra



Población vs. Muestra

Pregunta: Por qué nos preocupa eso de la *población vs. muestra*?



- En **promedio**, nuestras líneas de regresión coinciden con la línea de la población de forma correcta.
- Sin embargo, **Lineas individuales** (muestras) pueden fallar.
- Las diferencias entre las muestras individuales y de la población generan **incertidumbre** para el economista.

Población vs. Muestra

Pregunta: Por qué nos preocupa eso de la *población vs. muestra*?

Respuesta: La incertidumbre es importante.

- Se está interesado en **describir** y **evaluar** la relación entre una variable determinada (denominada *explicada* o *dependiente*) y una o más otras variables (comúnmente llamadas variables *explicativas* o *independientes*).
- Estableceremos como la variable *dependiente* por (y), mientras que las *independientes* por x_1, x_2, x_k .

✓ Si $k = 1$, solo hay una de las k -variables, por ende se estima una regresión simple.

✓ Si $k > 1$, hay más de las k -variables, tenemos entonces un modelo de regresión múltiple.

Modelos



Modelos

|||| Un ejemplo de modelos de regresión

y = Salario por horas

x = Años de educación

- Objetivo: **Determinar la relación entre (y) (Salario) y (x) (años de educación).**
- Un modelo mas general y con múltiples variables, como es el caso de los **Salario en función de la educación y otras características:**

y = Salario por horas

x_1 = Años de educación

x_2 = Edad

x_3 = Experiencia

- Objetivo: **Determinar la relación entre (y) (salario) y (x' s) (años de educación, edad y la experiencia).**

Modelos



Hay varios objetivos en estudiar este tipo de *relaciones*

- Analizar los **efectos** de políticas que envuelven cambiar los x' s individuales.
- Pronosticar **el valor** de y para un determinado conjunto de x' s.
- Examinar si alguno de los x' s tiene un **efecto** significativo en y .

¶ Comparaciones estadísticas y deterministicas

- En las relaciones **estadísticas** entre variables tratamos esencialmente con variables aleatorias (variables que tienen distribuciones de probabilidad).
- En la dependencia funcional o **determinística** también manejamos variables, pero no son aleatorias (ejemplo: leyes física).

Regresión vs. Causalidad



Regresión vs. Causalidad



- ⦿ A pesar de que el **análisis de regresión** tiene que ver con la *dependencia* de una variable respecto a otras variables, esto no implica causalidad necesariamente.
- ⦿ Para aducir **causalidad** se debe acudir a consideraciones a priori o teóricas.
- ⦿ **Ejemplo:** Un estudio de la dependencia existente entre el producto de una cosecha y la temperatura, lluvia, cantidad de sol y fertilizantes.

| No hay una relación estadística para suponer que la lluvia no depende del producto de la cosecha. El hecho que el producto de la cosecha se considere como dependiente de la lluvia (entre otros) es debido a otras consideraciones, como por ejemplo el *sentido común*.

Regresión vs. Causalidad 🏔

Estructura de un modelo 🐾

(X,Y) son dos variables *aleatorias*, que representan a alguna población, y estamos interesados en explicar Y en términos de X o en "estudiar como varia Y con cambios en X".

$$\underbrace{Y}_{\text{Variable dependiente}} = \underbrace{\beta_0}_{\text{Parámetro intercepto}} + \underbrace{\beta_1}_{\text{Parámetro pendiente}} \underbrace{X}_{\text{Variable independiente}} + \underbrace{\mu}_{\text{Término de error}}$$

- El parámetro μ es una variable aleatoria *no observable* que toma valores positivos o negativos, en términos generales representa *otros factores de X que afectan a Y*.
- La(s) variable(s) X tiene un efecto lineal en $Y \Rightarrow \Delta Y = \beta_1 \Delta X$ si y solo si $\Delta \mu = 0$.

Otro ejemplo



Piense en lo siguiente



|||| La directora de escuelas primarias de una localidad de Barranquilla quiere responder la siguiente pregunta:

- Si se reduce el **tamaño promedio** de las clases en dos (2) estudiantes, ¿cuál es el efecto en las calificaciones obtenidas por el resto del curso en un examen de cierta asignatura?

Una respuesta precisa a la *pregunta* exige una cuantificación de las *variaciones*: si la directora varía el número de alumnos por clase en cierta cantidad, ¿qué variación esperaría que sucediese sobre las puntuaciones de los exámenes?

- Una posible respuesta es:

$$\beta_i \equiv \beta_{\text{Tamaño clase}} = \frac{\text{Variación Calif Examen}}{\text{Variación Tamaño Clase}} = \frac{\Delta \text{Calificación Examen}}{\Delta \text{Tamaño Clase}}$$

Piense en lo siguiente



- Se podría responder a la pregunta real de la directora reorganizando la ecuación:

$$\Delta \text{Calificación Examen} = \beta_{\text{Tamaño Clase}} \times \Delta \text{Tamaño Clase}$$

- Si por alguna manera $\beta_{\text{Tamaño Clase}} = -0.6$, una reducción en dos alumnos da como variación de las calificaciones esperadas de $(-0.6) \times (-2) = 1.2$.

La **línea** recta que relaciona las *calificaciones* y el *Tamaño de la clase* puede escribirse como:

$$\text{Calificación examen} = \beta_0 + \beta_i \times \text{Tamaño Clase}$$

Recuerde que β_i es el **parámetro** del tamaño de la clase

Esta **ecuación** no se cumple con exactitud para todas las **localidades**. Una versión de esta *relación lineal* que se cumpliera en cada distrito debería incorporar otros factores que pueden influir en las calificaciones, incluyendo las características únicas de cada uno de los distritos (ejemplos: calidad maestros, características alumnos, fortuna estudiantes el día del examen, etc.)

Piense en lo siguiente



☞ Suponga que quisiéramos predecir la nota del examen de matemáticas dado *cierto tamaño de la clase*, entonces tendremos:

$$\text{Calificación examen} = 27 - 0.6 \times \text{Tamaño Clase} + \mu_i$$

Si colocamos como tamaño de clase el número de 40 estudiantes, entonces vamos a tener en promedio como resultado de nota 3.0. Observe que si el tamaño de la clase fuera ahora de 38. La **calificación** entonces estaría rondando una nota de 4.2.

Piense en lo siguiente



Un modelo completo 🍄

Es de pensar, que entonces un modelo más completo es:

$$\text{Calificación examen} = \beta_0 + \beta_{\text{Tamaño Clase}} \times \text{Tamaño Clase} + \text{Otros factores}$$

❑ Siempre es bueno tener en cuenta los supuestos del **Modelo de regresión**

Estos son:

Ⓐ Sea $\{(X_i, Y_i : i = 1, 2, 3, \dots, n)\}$ una muestra *aleatoria* de tamaño n de la población:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i \quad i = 1, 2, 3, \dots, n$$

Nuestro objetivo es tener estimado los **parámetros** desconocidos β_0 y β_1 dadas las n observaciones de (X, Y) . Para esto, tenemos algunos supuestos sobre μ .

Piense en lo siguiente



```
library(wooldridge)
library(tidyverse)
data("ceosal1")

mi_modelo<-lm(salary~roe, ceosal1)
summary(mi_modelo)
```

Qué interpretación tiene lo anterior?

$$\hat{\text{salary}} = 963.19 + 18501 \text{ roe}$$

- Lo que si el rendimiento del **capital** es cero $\text{roe} = 0$, el sueldo (intercepto), la parte de 963.191 es el salario promedio que recibe el gerente. Ya que el salario se mide en miles esto se interpreta así en términos de las unidades de \hat{y} .

```
#>
#> Call:
#> lm(formula = salary ~ roe, data = ceosal1)
#>
#> Residuals:
#>    Min     1Q   Median     3Q    Max 
#> -1160.2  -526.0  -254.0   138.8 13499.9
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)  963.19     213.24   4.517 1.05e-05 ***
#> roe          18.50      11.12   1.663   0.0978 .  
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
#>
#> Residual standard error: 1367 on 207 degrees of freedom
#> Multiple R-squared:  0.01319, Adjusted R-squared:  0.00777
```

- Lo que tenemos, el **cambio** que se predice para el sueldo en función del cambio en el roe se expresa tal que:

$$\Delta \text{salary} = 18,501(\Delta \text{roe})$$

Los supuestos de residuo y la estimación



Supuestos de los residuos $\mu \sim \epsilon$

1. **Media cero:** $E(\mu_i) = 0 \quad \forall i.$
2. **Varianza común:** $var(\mu_i) = \sigma^2 \quad \forall i.$
3. **Independencia (no correlación serial):** μ_i y μ_j son independientes para todo $i \neq j$. Dado (X_i) , las desviaciones de dos valores cualquiera de Y de su media no muestran valores *sistemáticos*.
4. **Independencia de X_j :** μ_i y X_j son independientes para todo i y j. Intuitivamente, si no se cumple entonces es difícil aislar la influencia de X y μ sobre Y.
5. **Normalidad:** μ_i está normalmente distribuida para todo i.

Regresión lineal



El estimador

Podemos estimar la regresión en R (`lm(y ~ x, my_data)`). Pero esas estimaciones de donde provienen?

Repasemos

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

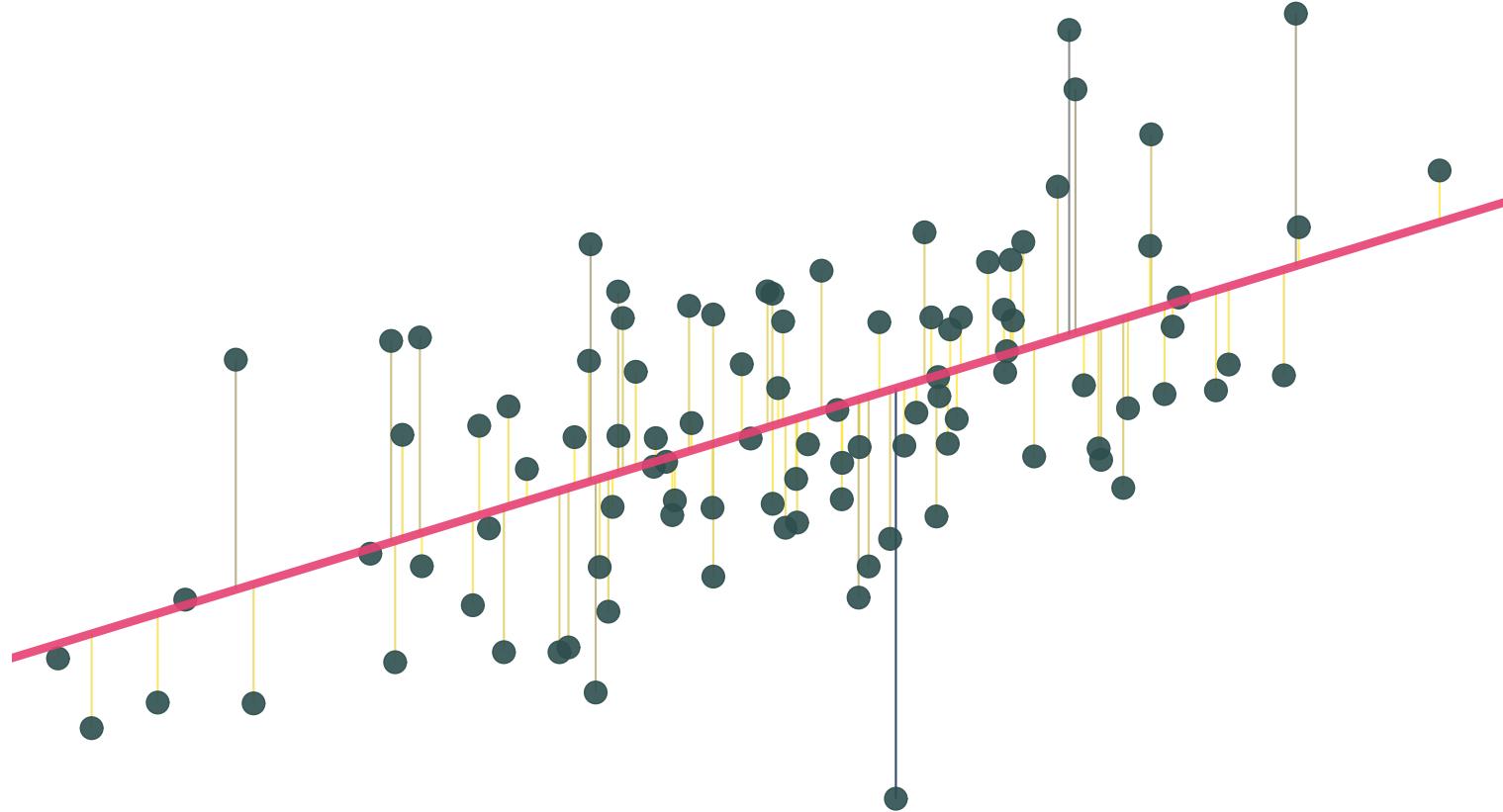
El cual nos da *mejor-ajuste* lineal de nuestros datos. Pero que significa eso de "Linea de mejor ajuste"?

- En (econometría), *mejor-ajuste* significa que la *linea* de los datos minimiza la suma del error al cuadrado (SSE):

$$\text{SSE} = \sum_{i=1}^n e_i^2 \quad \text{donde} \quad e_i = y_i - \hat{y}_i$$

- Mínimos **cuadrados ordinarios (MCO)** minimiza la suma de los errores al cuadrado.
- Basado en una serie de supuestos (en su mayoría aceptables), MCO:
 - Es insesgado (y consistente)
 - Es el *mejor* (mínima varianza) estimador lineal insesgado (MELI)

La estimación MCO busca tener un $\hat{\beta}_0$ y un $\hat{\beta}_1$ que minimiza a SSE.



El estimador

Formalmente

En el modelo de regresión simple, el estimador MCO vendrá a ser obtenido mediante $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimiza la suma de los residuos al cuadrado (SSE), *p.e.*,

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \text{SSE}$$

Pero ya sabemos que $\text{SSE} = \sum_i e_i^2$. Ahora definimos a los residuos e_i y el valor predicho de la dependiente \hat{y} .

$$\begin{aligned} e_i^2 &= (y_i - \hat{y}_i)^2 = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \\ &= y_i^2 - 2y_i \hat{\beta}_0 - 2y_i \hat{\beta}_1 x_i + \hat{\beta}_0^2 + 2\hat{\beta}_0 \hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2 \end{aligned}$$

Recuerde: Minimizar una función multivariada requiere (1) que la primera derivada (La condición de *1er-orden*) y (2) condición de segundo-orden o (concavidad).

El estimador



Nos estamos acercando. Tenemos que **minimizar la SSE**. Hemos mostrado cómo se relaciona el SSE con nuestra muestra (nuestros datos: x e y) y nuestras estimaciones (*p.e.*, $\hat{\beta}_0$ y $\hat{\beta}_1$).

$$\text{SSE} = \sum_i e_i^2 = \sum_i \left(y_i^2 - 2y_i\hat{\beta}_0 - 2y_i\hat{\beta}_1x_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1x_i + \hat{\beta}_1^2x_i^2 \right)$$

Para las condiciones de primer orden de minimización, tomamos ahora las primeras derivadas de SSE con respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$.

$$\begin{aligned} \frac{\partial \text{SSE}}{\partial \hat{\beta}_0} &= \sum_i \left(2\hat{\beta}_0 + 2\hat{\beta}_1x_i - 2y_i \right) = 2n\hat{\beta}_0 + 2\hat{\beta}_1 \sum_i x_i - 2 \sum_i y_i \\ &= 2n\hat{\beta}_0 + 2n\hat{\beta}_1\bar{x} - 2n\bar{y} \end{aligned}$$

donde $\bar{x} = \frac{\sum x_i}{n}$ y $\bar{y} = \frac{\sum y_i}{n}$ son las medias muestrales de x e y (tamaño n).

El estimador



Las condiciones de primer orden establecen que las derivadas son iguales a cero, por lo que:

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_0} = 2n\hat{\beta}_0 + 2n\hat{\beta}_1\bar{x} - 2n\bar{y} = 0$$

Lo cual implica

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

- ➊ Este **estimador** viene a ser la diferencia entre los promedios de nuestras variables dependientes e independientes teniendo presente el efecto de $\hat{\beta}_1$.

Ahora solo nos falta por hallar $\hat{\beta}_1$.

El estimador



Hay que tomar la derivada de SSE con respecto a $\hat{\beta}_1$

$$\begin{aligned}\frac{\partial \text{SSE}}{\partial \hat{\beta}_1} &= \sum_i \left(2\hat{\beta}_0 x_i + 2\hat{\beta}_1 x_i^2 - 2y_i x_i \right) = 2\hat{\beta}_0 \sum_i x_i + 2\hat{\beta}_1 \sum_i x_i^2 - 2 \sum_i y_i x_i \\ &= 2n\hat{\beta}_0 \bar{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2 \sum_i y_i x_i\end{aligned}$$

todo igual a cero (condición de primer-orden, de nuevo)

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_1} = 2n\hat{\beta}_0 \bar{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2 \sum_i y_i x_i = 0$$

y sustituimos $\hat{\beta}_0$, p.e., $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Así,

$$2n \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \bar{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2 \sum_i y_i x_i = 0$$

El estimador

De lo anterior

$$2n \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \bar{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2 \sum_i y_i x_i = 0$$

a multiplicar

$$2n\bar{y}\bar{x} - 2n\hat{\beta}_1\bar{x}^2 + 2\hat{\beta}_1 \sum_i x_i^2 - 2 \sum_i y_i x_i = 0$$

$$\implies 2\hat{\beta}_1 \left(\sum_i x_i^2 - n\bar{x}^2 \right) = 2 \sum_i y_i x_i - 2n\bar{y}\bar{x}$$

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

El estimador



Hecho!

Ahora tenemos estimadores OLS (encantadores) para la pendiente

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Para el intercepto o β_0

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

Y ahora **ya saben de dónde** viene la parte de *mínimos cuadrados* de MCO.

Otras condiciones



Propiedades de los estimadores de MCO

1. Los estimadores deben ser **lineales** sumado a las perturbaciones.
2. Nuestras variables \mathbf{X} son exogenas, p.e: $E[\mu | \mathbf{X}] = 0$
3. La relación entre las variables explicativas \mathbf{X} es inexistente, de lo contrario padecera de *multicolinealidad*.
4. La perturbación tiene media cero $E[\mu] = 0$ y varianza constante (σ^2), su distribución debe ser independiente e idénticamente distribuida.

Propiedades de los estimadores de MCO

$$E[\mu|X] = 0$$

Es una de las propiedades mas restrictivas. El cumplimiento de los supuestos 1-3 nos garantiza **insesgadez** en los estimadores. Ya se hace necesario tener 4 para decir que entonces es **mínima varianza**.

Un ejemplo

$$E[\mu|X = 10] = 0 \quad \text{de igual manera} \quad E[\mu|X = 100] = 0$$

Incluso con variables cualitativas, la condición debe mantenerse, esto es:

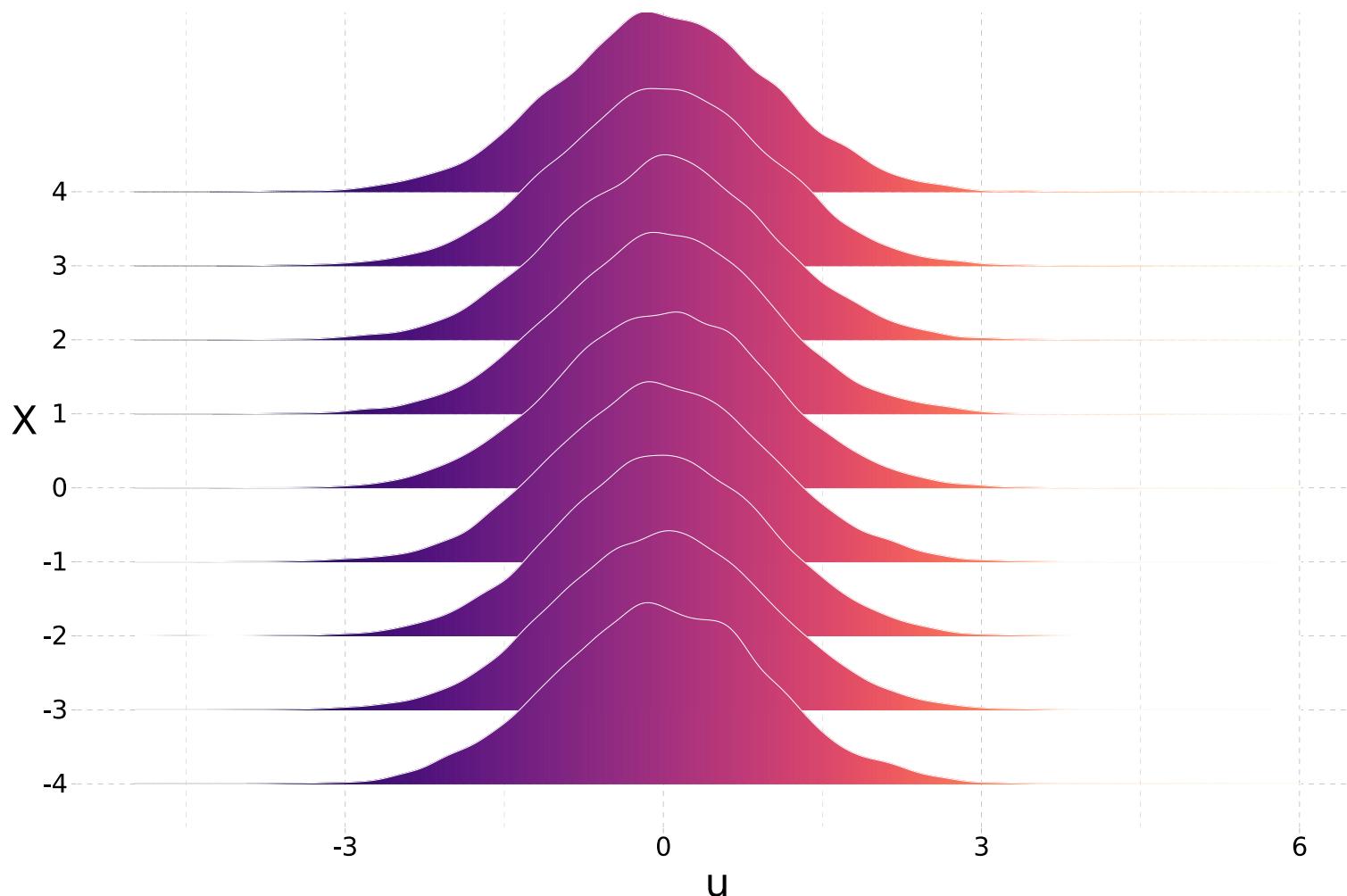
$$E[\mu|X = \text{mujer}] = 0 \quad \text{de igual manera} \quad E[\mu|X = \text{hombre}] = 0$$

Exogeneidad estricta

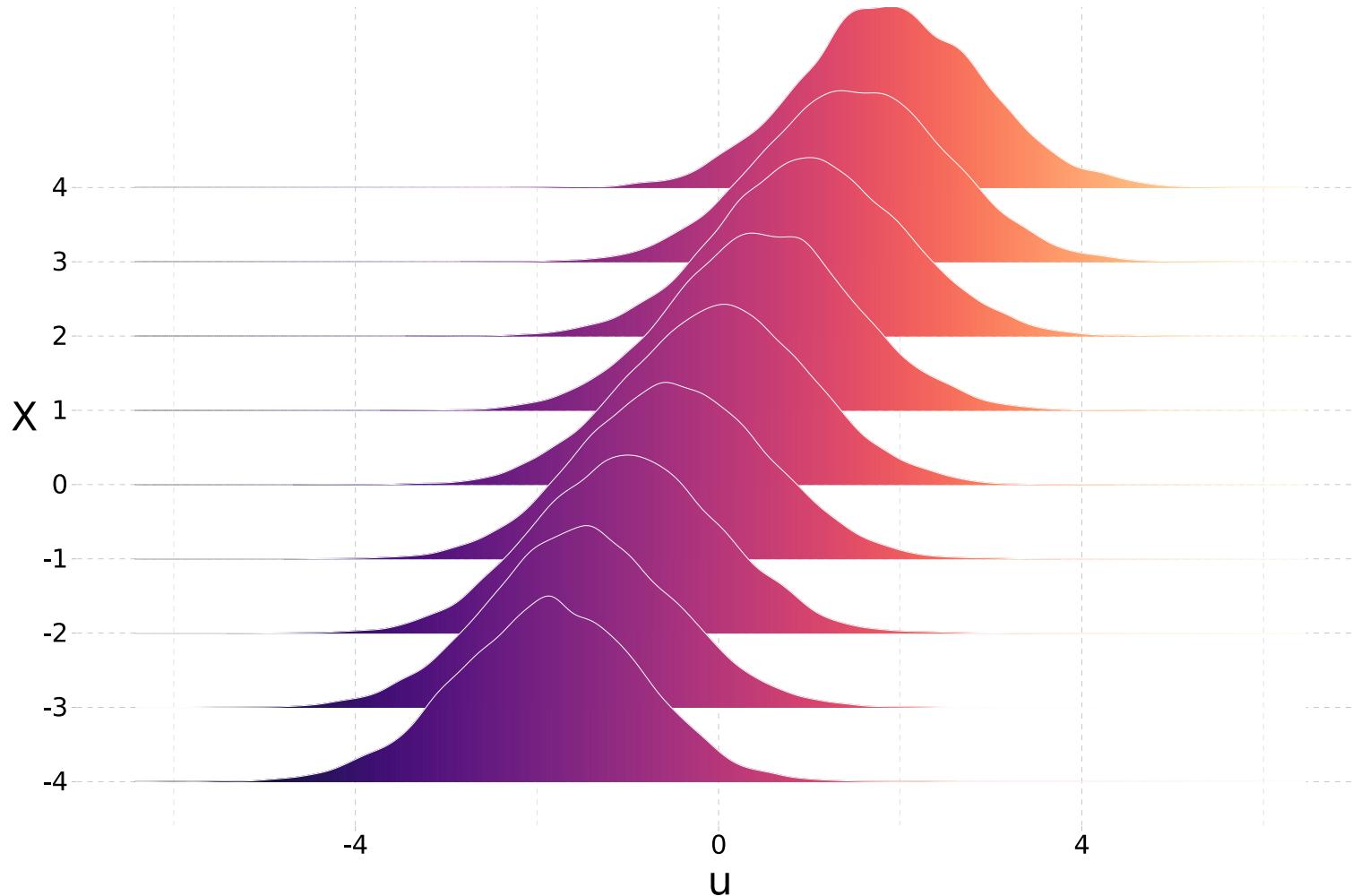


Exogeneidad estricta

Esa validez es, p.e., $E[u \mid X] = 0$



Esa validez no se da cuando, p.e., $E[u \mid X] \neq 0$



Estimación en

20
AÑO Pregrado en
ECONOMÍA



Estimación en R

La opción por default es: lm()

La forma de estimación en R para usar como base† para estimar los modelos de Regresión lineal es lm().

Puede hacerlo directamente

```
lm(y ~ x)
```

- Esto estima $y_i = \beta_0 + \beta_1 x_i + u_i$ (R lo hace automáticamente incluyendo el término del **intercepto**)^{††}
- Los datos se vinculan como objetos columna (y) (dependiente) y ademas (x) (independientes).

```
lm(y ~ x, data = bd_Dane)
```

- Estimamos $y_i = \beta_0 + \beta_1 x_i + u_i$
- Usando las columnas de y ademas de x del objeto bd_Dane.

† base es el formato por default del algoritmo

†† Puede remover el intercepto solo colocando -1 dentro de la formula, p.e., lm(y ~ -1 + x).

Estimación en

Ademas de lm()

Si necesita incluir mas variables? Pues... fácil

```
lm(y ~ x1 + x2 + x3, data = alguna_bd)
```

- Donde estima $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$
- La referencia de `alguna_bd` es para estipular la base de datos a usar.

Estimación en R

Algo más de lm()

Si requiere transformar/interactuar con variables? También es fácil: debe usar para eso I().

```
lm(y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2), data = bd_Dane)
```

- Esto estima $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i}x_{2i} + u_i$
- Utilizando las variables del objeto bd_Dane (donde están los datos)
- o se crean/generan vía I()

Nota: Los siguientes *ejemplos* son equivalentes:

- lm(y ~ x1 + x2 + I(x1*x2))
- lm(y ~ x1 + x2 + x1:x2)
- lm(y ~ x1*x2)

Estimación en R

Transformando variables con lm()

Observe lo siguiente:

```
lm(y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2), data = bd_Dane)
```

No necesitamos crear x_1^2 , x_2^2 , ademas de $x_1 \times x_2$ en el conjunto de datos.

El programa de R hace el calculo por nosotros (siempre y cuando x_1 y x_2 existan en la base de datos).

Cualquier **transformación** que quiera hace es posible

- Transformación Matemática/estadística: $I(x^2)$, $I(x/3)$, $I((x - mean(x))/sd(x))$
- Log/exponenenciales : $\log(x)$, $\exp(x)$
- Indicadores: $I(x < 100)$, $I(x == "Barranquilla")$

Bibliografía

- Álvarez, R. A. R., Calvo, J. A. P., Torrado, C. A. M., & Mondragón, J. A. U. (2013). *Fundamentos de econometría intermedia: teoría y aplicaciones*. Universidad de los Andes.
- Stock, J. H., Watson, M. W., & Larrión, R. S. (2012). *Introducción a la Econometría*.
- Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Gracias por su atención!

Alguna pregunta adicional?

Carlos Andres Yanes Guerra

 cayanes@uninorte.edu.co

 keynes37