Regresión lineal

Regresión - ¿para qué se usa?

Regresión - ¿para qué se usa?

El análisis de regresión tiene múltiples usos y se encuentra en casi todas las áreas y disciplinas. Nos permite:

- Establecer relaciones entre dos o más variables
- Predecir qué valor tomará una variable en función de otra(s)
- Además de establecer relaciones entre variables, nos da indicios acerca de cómo es esa relación, es decir de cuánto el cambio de una variable independiente incide en la variable dependiente.
- A partir de modelar la relación entre variables (partiendo de una muestra, de datos históricos, de observaciones) es posible predecir ciertos resultados que permiten tomar decisiones informadas

Regresión

Habiendo estudiado la teoría de las variables aleatorias (por ejemplo cantidad de hijos) y analizado propiedades de las mismas (su esperanza, su varianza...), ahora nos ocuparemos de modelar la relación entre estas variables.

¿Cómo?

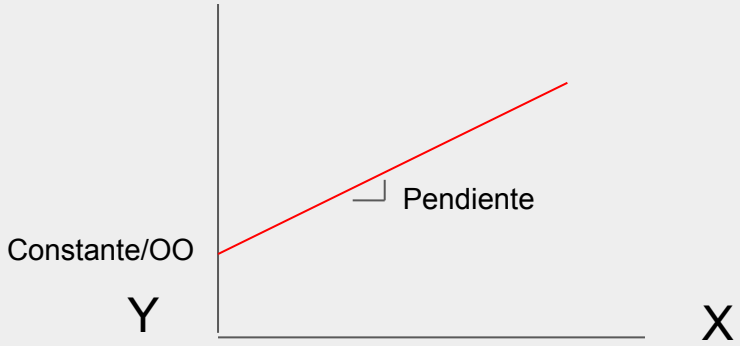
Buscando el modelo que mejor represente esa relación, y que nos permita estimar ciertos parámetros que la caractericen correctamente

Regresión

Comenzaremos con el modelo de regresión lineal, que es relativamente simple y cuya fórmula matemática es sencilla de interpretar.

Ecuación de la regresión lineal:

$$y = \beta_0 + \beta_1 * x + \varepsilon$$



Regresión lineal

Supuesto teórico del modelo lineal: la relación entre variables es lineal.
¿Cómo podemos indagar si este supuesto tiene sentido en relación a datos específicos?
Una buena primera aproximación es hacer un análisis de correlación entre las variables antes de estimar el modelo lineal

Regresión lineal - asociación entre variables y correlación

- Covarianza: medida que indica la asociación lineal entre dos variables aleatorias X e Y:

  $Cov(X, Y) = E((X_i - E(X) * (Y_i - E(Y))$

  Se analiza si ambas tienen variación conjunta. Variabilidad de cada variable con respecto a su valor esperado.

  $Cov(X, Y) > 0$ ⟶ relación positiva ↑ X ↑ Y (y lo mismo al revés)

  $Cov(X, Y) < 0$ ⟶ relación negativa ↑ X ↓ Y (y lo mismo al revés)

  $Cov(X, Y) = 0$ ⟶ ninguna relación

La covarianza está en unidades de medida que dependen de cada variable. Por eso se utiliza frecuentemente el coeficiente de correlación.

Regresión lineal - asociación entre variables y correlación

- Coeficiente de correlación: considera la varianza y la divide por el producto de los desvíos estándar de las variables

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma X * \sigma Y}$$

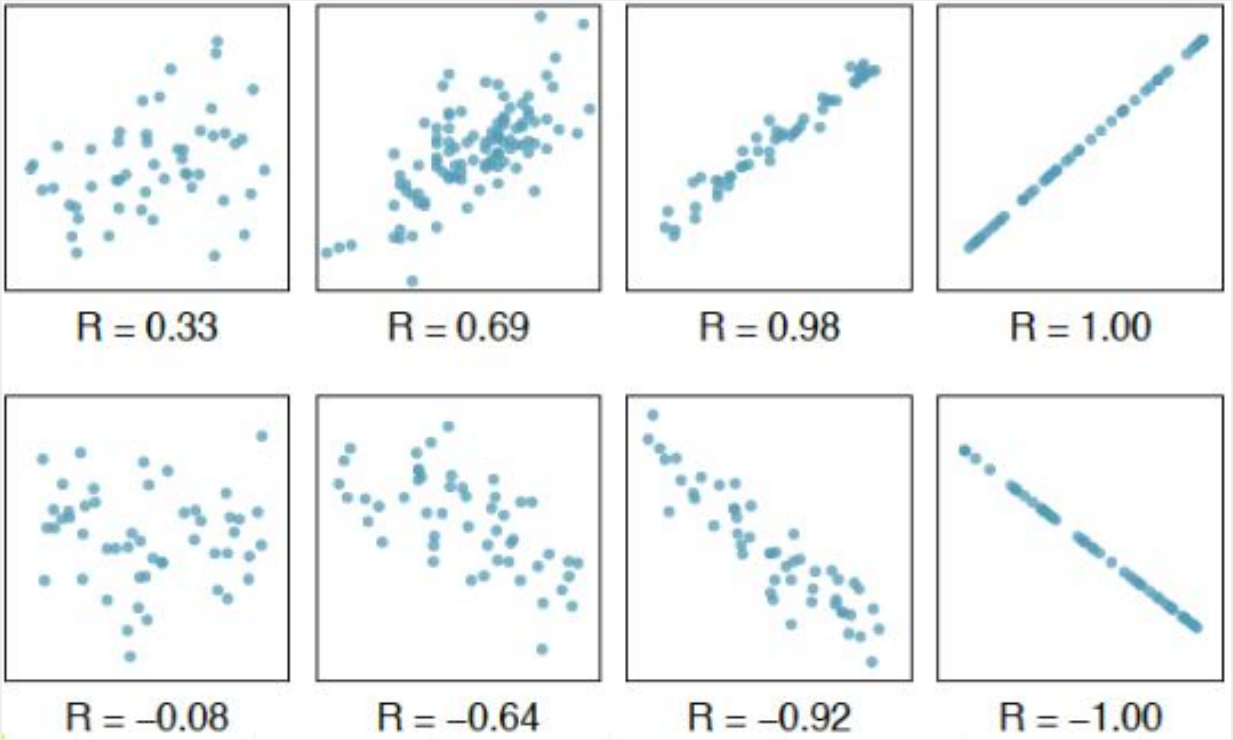
$$\sigma X = \sqrt{var(X)}$$

$$\sigma Y = \sqrt{var(Y)}$$

- Es un valor que va de -1 a 1 e indica tanto la fuerza como la dirección de la relación entre dos variables

$$-1 \leq \rho(X,Y) \leq 1$$

Regresión lineal - asociación entre variables y correlación



Dependiendo de cuál es la recta (en caso de que la haya) que mejor representa nuestros datos obtendremos distintos valores de R

Ejemplo práctico - experiencia laboral y salario

Ahora analicemos la relación entre dos variables específicas:
-Años de experiencia laboral
-Salario

¿Cómo creen que será esta relación?

Ejemplo práctico - experiencia laboral y salario

1. Estimamos la correlación entre ambas variables:

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Leemos el archivo csv con nuestros datos de salario y años de experiencia laboral

df = pd.read_csv('Salary_Data.csv')

# Observamos las primeras filas del dataset

print(df.head())
```

```
   YearsExperience   Salary
0              1.1  39343.0
1              1.3  46205.0
2              1.5  37731.0
3              2.0  43525.0
4              2.2  39891.0
```

Ejemplo práctico - experiencia laboral y salario

1. Estimamos la correlación entre ambas variables:

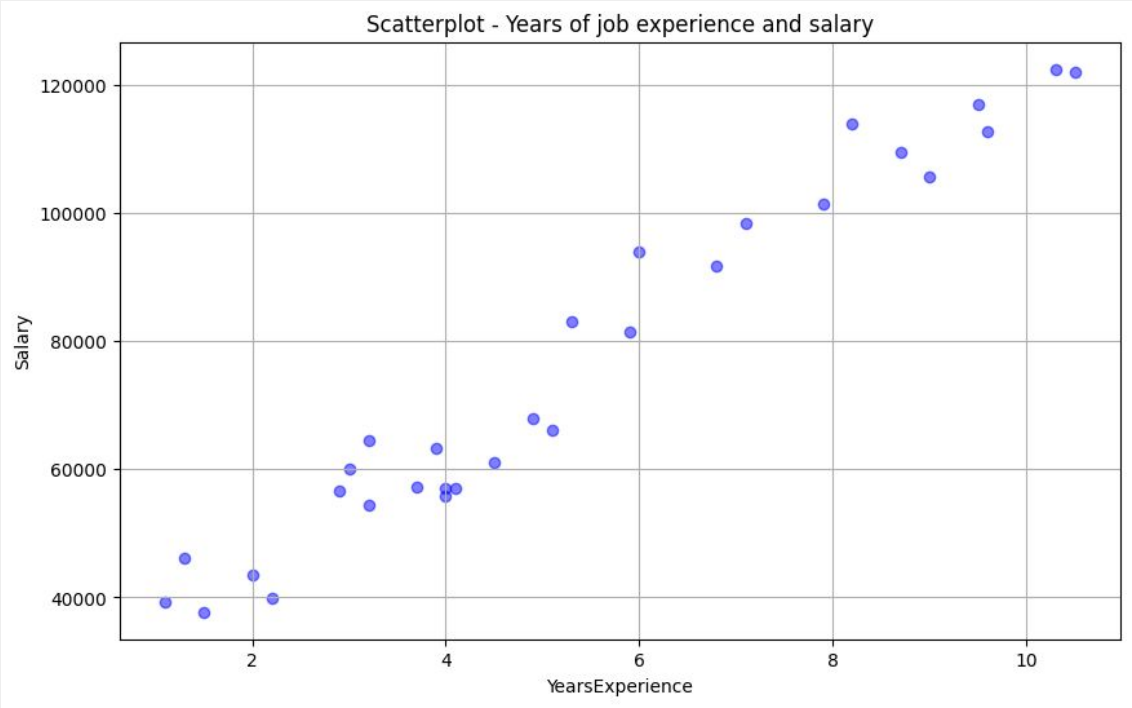
```
# Calculamos la correlación entre la var X (años de experiencia) e Y (Salario)

correlacion = df['YearsExperience'].corr(df['Salary'])
correlacion
```

```
0.9782416184887599
```

¿Cómo podemos ver gráficamente la relación lineal que estaría indicando este coeficiente?

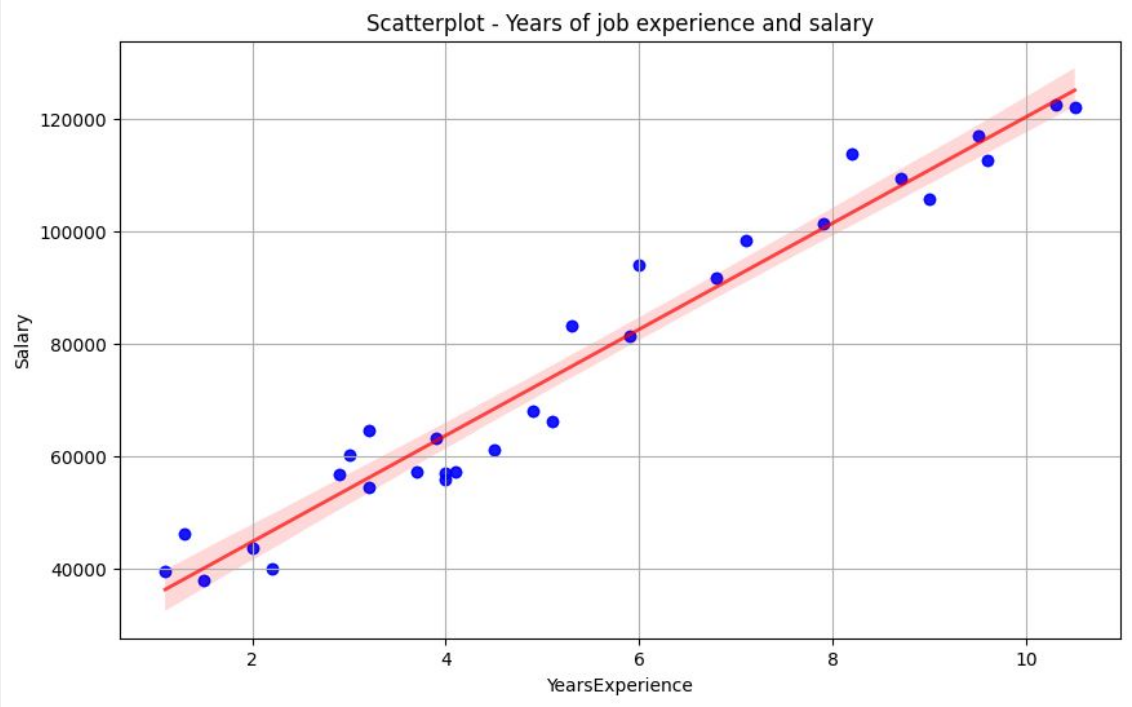
Ejemplo práctico - experiencia laboral y salario

2. Hacemos un gráficos de dispersión para ambas variables



Ejemplo práctico - experiencia laboral y salario

2. Hacemos un gráficos de dispersión para ambas variables



Al observar la relación entre nivel de experiencia laboral (medida en años) y salario, esta se presupone lineal. Podemos representarla a través de $\longrightarrow$ $y = \beta_0 + \beta_1 * x + \varepsilon$

Regresión lineal simple - elementos de la ecuación

Este modelo nos permite estimar el valor de una variable (Y) en función de otra (X), donde tenemos un único regresor X:

$$y = \beta_0 + \beta_1 * x + \varepsilon$$

Y = variable dependiente, endógena o de interés.

X = variable independiente, explicativa o regresor.

β0 = constante. Es el valor estimado de Y cuando X = 0

β1 = parámetro de una variable X_1

ε = término de error: captura todo aquello no incluido en x.

Regresión lineal simple - elementos de la ecuación

Años de experiencia y salario.

$$y = \beta_0 + \beta_1 * x + \varepsilon$$

Y = variable dependiente, endógena o de interés.
Y = salario
X = variable independiente, explicativa o regresor.
X = años de experiencia laboral
β0 = constante. Es el valor estimado de Y cuando X = 0
β0 = salario cuando años de experiencia laboral = 0
β1 = parámetro de una variable X1
β1 = captura la medida del cambio en el salario por cada unidad adicional de años de experiencia

Regresión lineal simple - estimación de parámetros

$\beta_0 = y - \beta_1 X_1 * X + \varepsilon \longrightarrow$ Salario(años de experiencia=0)

$\beta_1 \longrightarrow$ Salario(años.exp=n+1) - Salario(años.exp=n)

$= \beta_0 + \beta_1(n + 1) + \varepsilon - \beta_0 - \beta_1(n) - \varepsilon$

$= \beta_0 + \beta_1 n + \mathbf{\beta_1} + \varepsilon - \beta_0 - \beta_1 n - \varepsilon$

β1 captura el cambio que produce una unidad adicional de experiencia en el salario (en promedio).
Constituye el efecto marginal de la var. X en Y (por eso acompaña al regresor). Y también se lo suele definir como:

$$\beta_j = \frac{\partial Y}{\partial X_j} \qquad j = 1, 2, 3 \ldots k$$

Regresión lineal simple - estimación de parámetros

β0 y β1 son parámetros poblacionales desconocidos. A través de distintos modelos, como por ejemplo la regresión lineal, buscamos estimarlos.

Regresión lineal simple - estimación de parámetros

ε representa el término de error de la regresión
Es decir, en el error se incluye toda la variabilidad de Y que no se debe a X

ε en nuestro ejemplo = nivel educativo, industria/sector, género...

Nota: es importante que ε cumpla con ciertas características en una regresión lineal. Más adelante las veremos.

Regresión lineal simple - estimación de parámetros

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$\text{Salario} = \beta_0 + \beta_1 \text{Años.exp} + \varepsilon$$

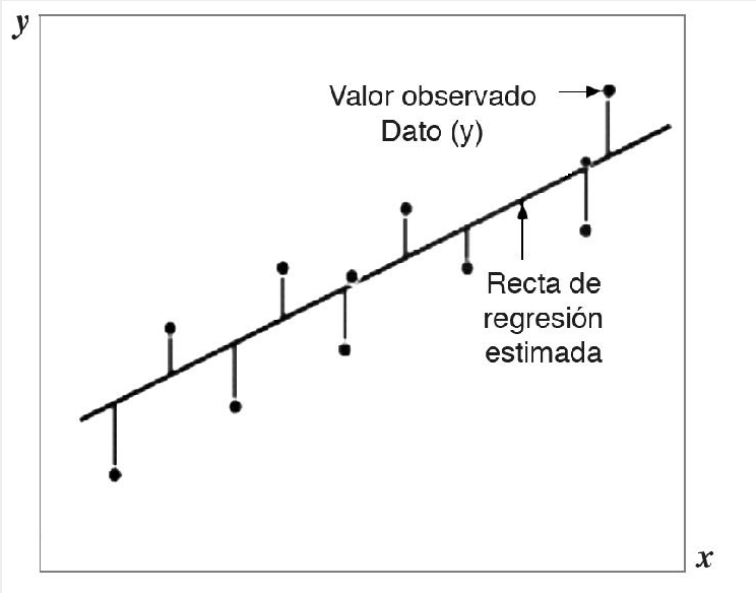
¿Cuál es el problema de estimar el valor de Y a partir de una recta?

Regresión lineal simple - estimación de parámetros

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$\text{Salario} = \beta_0 + \beta_1 \text{Años.exp} + \varepsilon$$

¿Cuál es el problema de estimar el valor de Y a partir de una recta?



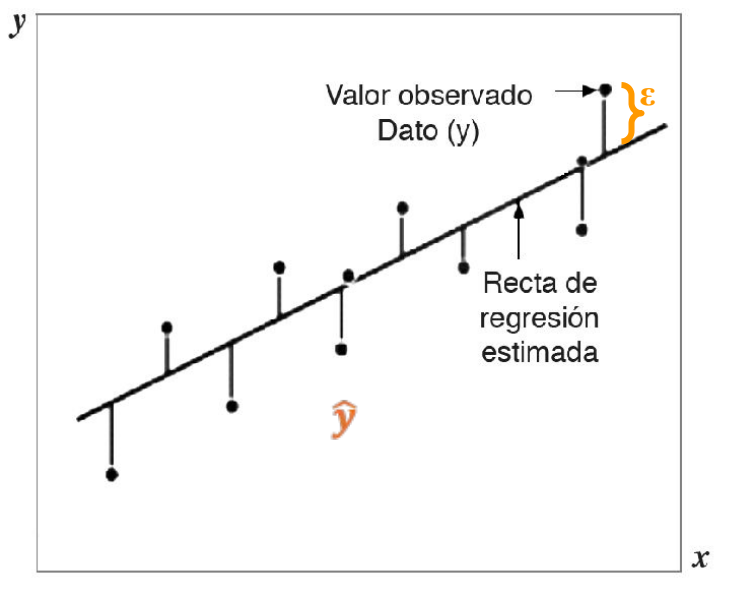
Regresión lineal simple - estimación de parámetros

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0
5	2.9	56642.0
6	3.0	60150.0
7	3.2	54445.0
8	3.2	64445.0
9	3.7	57189.0
10	3.9	63218.0
11	4.0	55794.0
12	4.0	56957.0
13	4.1	57081.0
14	4.5	61111.0
15	4.9	67938.0
16	5.1	66029.0
17	5.3	83088.0
18	5.9	81363.0

Veamos las primeras observaciones de la base de datos...

Para este nivel dado de años de experiencia, no tenemos un único valor de salario

Regresión lineal simple - estimación de parámetros



La regresión estima un cierto valor salarial dado un nivel de experiencia laboral x. Pero en la muestra se observan datos que se encuentran tanto por encima como por debajo. Una forma de dar con la mejor predicción de Y es minimizar los residuos. MCC minimiza la distancia vertical (ε) al cuadrado entre las observaciones y la recta.

Regresión lineal simple - estimación de parámetros

Ej. caso 7 (años.exp=3.2 y salario=54445)
β0 y β1 fueron estimados, siendo:
β0 = 2.58
β1 = 9449

Cuando los años de experiencia = 3.2, el modelo estimaría:

Ingreso = 2.58 + 9449*3.2 + ε

Ingreso = 30240 + ε

ε = ingreso − β0 + β1*X1

¿Qué ocurre con ε?

Regresión lineal simple - estimación de parámetros

Ej. caso 7 (años.exp=3.2 y salario=54445)
β0 y β1 fueron estimados, siendo:
β0 = 2.58
β1 = 9449

Cuando los años de experiencia =3.2, el modelo estimaría:

Ingreso = 2.58 + 9449*3.2 + ε

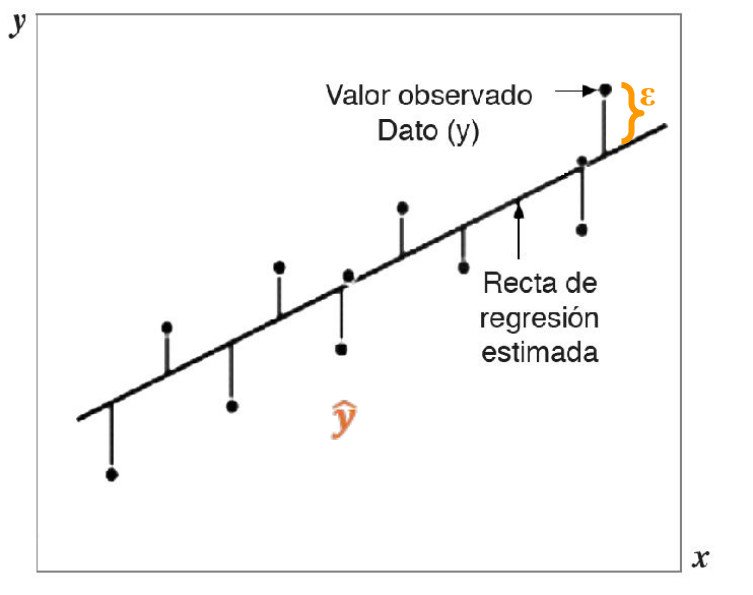
Ingreso = 30240 + ε

ε = ingreso − β0 + β1*X1

¿Qué ocurre con ε?

ε_i = +24205 → Se trata de un caso que presenta un salario por encima de la estimación para el nivel de experiencia que tiene (probablemente debido a otros factores explicativos que están en el error)

Regresión lineal simple - estimación de parámetros



A través del método de MCC/OLS buscamos minimizar los residuos, es decir la distancia general del valor de cada observación con respecto al valor estimado por la recta. En particular lo que hace este método es elegir estimadores β0 y β1 que minimicen la suma de los errores elevados al cuadrado. Objetivo:

εi ≈ 0 y en consecuencia:

$$\sum_{i=1}^{n} (\varepsilon i)^2 = \varepsilon 1^2 + \varepsilon 2^2 + \ldots + + \varepsilon n^2$$

¿Por qué elevamos al cuadrado la suma de los residuos?

Regresión lineal simple - estimación de parámetros

La suma de los cuadrados se eleva al cuadrado para evitar que los errores de las distintas observaciones se cancelen entre sí. Si la suma de los residuos al cuadrado tiene un valor pequeño, se supone que cada término de error también lo es, porque se trata de valores positivos en todos los casos.
MCC es una función que minimiza la suma de residuos al cuadrado considerando 2 variables = β0 y β1

$$\min_{\beta 0 \text{ y } \beta 1} \sum_{i=1}^{n} (\varepsilon i)^2 = \min_{\beta 0 \text{ y } \beta 1} \sum (Yi - \beta 0 - \beta 1 * xi)^2$$

$$= \min \sum (Y1 - \beta 0 - \beta 1 * x_1)^2 + (Y2 - \beta 0 - \beta 1 * x_2)^2$$

$$+ \quad \ldots (Yn - \beta 0 - \beta n * xn)^2$$

A menos que $\rho(X,Y) = 1$ o -1, no existen parámetros β0 y β1 tal que **$yi - \beta 0 - \beta i xi = 0$** para cualquier i

Regresión lineal simple - estimación de parámetros

La minimización de la función MCC se hace derivando con respecto a β0 y a β1 (en cada caso la otra variable se deja constante)

$$\frac{\partial \sum_{i=1}^{n} (Yi - \beta 0 - \beta 1 * xi)^2}{\partial \beta 0} = 0$$

$$\frac{\partial \sum_{i=1}^{n} (Yi - \beta 0 - \beta 1 * xi)^2}{\partial \beta 1} = 0$$

A partir de estas derivadas se obtienen las ecuaciones de que nos permiten despejar las 2 incógnitas (β0 y β1):

Regresión lineal simple - estimación de parámetros

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

β0 será la diferencia entre la estimación de Y y la parte explicada por el modelo:

$\beta_0 = Y - \beta_1 * X$

Regresión lineal simple - estimación de parámetros

Regresión lineal simple - estimación de parámetros

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

→ cov(x,y)

→ var(x)

esta división permite eliminar el peso de la unidad de medida de x

β0 será la diferencia entre la estimación de Y y la parte explicada por el modelo:

$\beta_0 = Y - \beta_1 * X$

Regresión lineal simple - variable explicativa binaria

¿Se puede utilizar MCC (y un modelo lineal en general) ante la presencia de variables explicativas no numéricas? Por ejemplo: **género**

Regresión lineal simple - variable explicativa binaria

¿Se puede utilizar MCC (y un modelo lineal en general) ante la presencia de variables explicativas no numéricas? Por ejemplo: **género**

Respuesta: sí.
Cuando un regresor X es una variable "indicadora" de la categoría hombre (por ejemplo), donde I(hombre) = 1 si se trata de un hombre e I(hombre)=0 si se trata de una mujer, los β miden la diferencia promedio entre la variable Y en cada categoría. Si Y=salario, β capturaría la diferencia (positiva o negativa) producto de que una persona sea hombre.
Si β = 5000, esto significa que los hombres ganan un salario de 5000 unidades más que las mujeres, en promedio.

Salario = $\beta_0 + \beta_1 I(\text{hombre}) + \varepsilon$

Si I(hombre)=0 $\longrightarrow$ Salario = $\beta_0 + \varepsilon$

Si I(hombre)=1 Salario = $\beta_0 + \beta_1 + \varepsilon$

Ingreso(I(hombre=1) - Ingreso(I(hombre=0) = $\cancel{\beta_0} + \beta_1 + \cancel{\varepsilon} - \cancel{\beta_0} + \cancel{\varepsilon}$ = **β_1**

Medidas de bondad de ajuste: R^2 (o coeficiente de determinación)

Las medidas de bondad de ajuste permiten evaluar qué tan bueno es un modelo, es decir estimar cuánta incertidumbre existe con respecto a los datos poblacionales.
Una de las más utilizadas dada su interpretación sencilla es el R^2. Varía entre 0 y 1 y puede interpretarse como un valor que indica qué medida de la varianza de Y es explicada por X.
Otra medida de bondad de ajuste que se utiliza frecuentemente es el error estándar.

Medidas de bondad de ajuste: R^2 (o coeficiente de determinación)

Su estimación consiste en la proporción de la variación de Y, es decir de la suma de los cuadrados totales SCT, que le corresponde a la variación de X (suma de los cuadrados explicados, SCE). Esto se calcula de la siguiente manera:

Siendo que:

$$SCT = \sum_{i=1}^{n} (Yi - \overline{Y})^2$$

$$SCR = \sum_{i=1}^{n} \varepsilon i^2$$

$$SCE = \sum_{i=1}^{n} [\beta 1(Xi - \overline{X})]^2$$

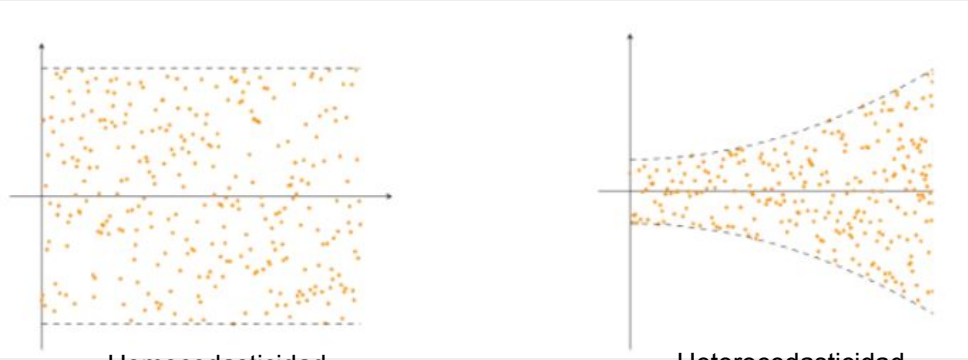
$$SCT = SCR + SCE \longrightarrow \mathbf{R^2 = SCE / SCT}$$

Supuestos de MCC para que los estimadores sean “buenos”

- Linealidad. Supone que una recta sea un buen ajuste para los datos

- Exogeneidad. $E(\varepsilon|X) = 0$. No debe haber nada en el término de error que dependa de X. Para cualquier nivel dado de X, el error debe ser en promedio = 0, valores de Y por encima y por debajo del valor esperado pero no una desviación sistemática. Ejemplo: relación entre la red de contactos (error) y los años de experiencia.

- Homocedasticidad. $Var(\varepsilon|Xi) = \sigma^2$. Ejemplo: gastos según nivel de ingreso (heterocedasticidad)



Homocedasticidad

Heterocedasticidad

Supuestos de MCC para que los estimadores sean “buenos”

- Ausencia de correlación serial entre errores. Cov(εi, εh) = 0. El error en una observación i no guarda relación con el de otra observación h.

- No multicolinealidad. No debe existir dependencia lineal entre variables explicativas. Por ejemplo: ingreso y salario.

La violación de estos supuestos produce que los estimadores de MCC no sean eficientes, consistentes e insesgados (“buenos”):
Eficiencia: su varianza es la menor de entre todos los estimadores.
Consistencia: a medida que aumenta el tamaño de la muestra, los estimadores de los parámetros se acercan a los valores de los parámetros poblacionales.
Insesgadez: el valor del parámetro estimado por el modelo coincide con el valor del parámetro estimado. $\mathbf{E(\hat{\beta}) = \beta}$

Supuestos de MCC para que los estimadores sean “buenos”

Estimación de una regresión lineal simple

¿Qué relación podemos esperar entre las siguientes variables?:

Autoubicación ideológica. Medida en una escala de izquierda a derecha que va de 1-10.
Grado de importancia asignado a la **corrupción** como problema a nivel país. Medida en una escala que va de 1-10.

¿Cuál podría ser Y y cuál X?

Estimación de una regresión lineal simple

Base de datos de encuestas a Diputados Nacionales (Argentina, 2022)

```python
import statsmodels.api as sm
```

Python

```python
file_path = 'BASEDATOS_ARGENTINA_122_ .sav'
df_diputados, meta = pyreadstat.read_sav(file_path)

# Ahora df_diputados es un DataFrame y meta es un diccionario de metadatos
print(type(df_diputados))
```

✓ 0.0s

Python

```
<class 'pandas.core.frame.DataFrame'>
```

```python
df_diputados
```

✓ 0.0s

Python

	Encuestado	País	legis	partido	departa	tipoelec	comision01	comision02	comision03	comision04	...	pcontacto2	pcontacto3	resultado1	resultado1
0	4.0	1.0	2022.0	20.0	12.0	2.0	7.0	16.0	31.0	NaN	...	NaN	NaN	1.0	
1	6.0	1.0	2022.0	7.0	2.0	2.0	2.0	3.0	15.0	25.0	...	NaN	NaN	1.0	
2	10.0	1.0	2022.0	2.0	15.0	2.0	1.0	2.0	3.0	4.0	...	NaN	NaN	1.0	
3	11.0	1.0	2022.0	7.0	2.0	2.0	5.0	15.0	27.0	35.0	...	NaN	NaN	1.0	
4	12.0	1.0	2022.0	29.0	9.0	2.0	37.0	38.0	NaN	NaN	...	NaN	NaN	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
106	252.0	1.0	2022.0	20.0	2.0	2.0	18.0	19.0	NaN	NaN	...	NaN	NaN	1.0	
107	256.0	1.0	2022.0	7.0	2.0	2.0	26.0	29.0	36.0	43.0	...	NaN	NaN	1.0	

Estimación de una regresión lineal simple

Base de datos de encuestas a Diputados Nacionales (Argentina, 2022)

```
df_diputados = df_diputados.loc[~((df_diputados['ID101'] == 98) | (df_diputados['ID101'] == 99) |
                                  (df_diputados['PRO102'] == 98) | (df_diputados['PRO112'] == 99))]
```

✓ 0.0s Python

```
print(df_diputados[['ID101','PRO109']])
```

✓ 0.0s Python

```
     ID101  PRO109
0      5.0     7.0
1      7.0    10.0
2      3.0     9.0
3      9.0     7.0
4      1.0    10.0
..     ...     ...
106    4.0     7.0
107    5.0    10.0
108    5.0    10.0
109    5.0    10.0
110    5.0     9.0

[110 rows x 2 columns]
```

```
valores_id101 = df_diputados['ID101'].unique()
valores_pro109 = df_diputados['PRO109'].unique()

print("Valores únicos en ID101:", valores_id101)
print("Valores únicos en PRO109:", valores_pro109)
```

✓ 0.0s Python

```
Valores únicos en ID101: [5. 7. 3. 9. 1. 6. 2. 4. 8.]
Valores únicos en PRO109: [ 7. 10.  9.  5.  8.  2.  3.  4.  6.]
```

Estimación de una regresión lineal simple
Base de datos de encuestas a Diputados Nacionales (Argentina, 2022)

```python
# Definir las variables independiente (X) y dependiente (Y)
X = df_diputados['PRO109']
Y = df_diputados['ID101']

# Agregar constante para la ordenada al origen (intercept, constante)
X = sm.add_constant(X)

# Ajustar el modelo de regresión
modelo = sm.OLS(Y, X).fit()

# Mostrar el resumen del modelo
print(modelo.summary())
```

✓ 0.0s Python

Estimación de una regresión lineal simple

Base de datos de encuestas a Diputados Nacionales (Argentina, 2022)

OLS Regression Results

Dep. Variable:	ID101	R-squared:	0.237
Model:	OLS	Adj. R-squared:	0.230
Method:	Least Squares	F-statistic:	33.64
Date:	Mon, 14 Oct 2024	Prob (F-statistic):	6.70e-08
Time:	20:48:33	Log-Likelihood:	-208.88
No. Observations:	110	AIC:	421.8
Df Residuals:	108	BIC:	427.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.4836	0.590	2.516	0.013	0.315	2.652
PRO109	0.4299	0.074	5.800	0.000	0.283	0.577

Omnibus:	2.441	Durbin-Watson:	2.410
Prob(Omnibus):	0.295	Jarque-Bera (JB):	1.914
Skew:	-0.301	Prob(JB):	0.384
Kurtosis:	3.235	Cond. No.	30.6