Modelos Estadísticos para la Regresión y la Clasificación

Práctico 4 - Regresión lineal simple

Micaela Long

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

13 de setiembre de 2024

Práctico 3
Probabilidad

Material para hacer práctico 4 (disponible en EVA):

- Presentación Regresión lineal (Teórico 4/9)
- Presentación Regresión lineal 2da parte (Teóricos 9/9 y 11/9)
- Laboratorio Práctico 4 en Python/Laboratorio Práctico 4 en R

Objetivo: establecer una relación entre una variable dependiente Y y una variable independiente x para poder hacer predicciones sobre Y cuando se conoce x.

$$y = f(x) + \epsilon$$

Modelo regresión lineal simple:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Donde:

- y es la variable dependiente.
- x es la variable independiente.
- β_0, β_1 son parámetros desconocidos.
- ϵ es el error aleatorio.

Dado un conjunto de datos (x_i, y_i) para $i = 1, 2, \ldots, n$, el objetivo es encontrar los coeficientes β_0 y β_1 de la mejor recta:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

¿Cómo definimos mejor? Vamos a medir qué tan lejos está nuestra predicción $\hat{y}_i$ del valor verdadero y_i

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\beta_0 + \beta_1 x_i) \end{aligned}$$

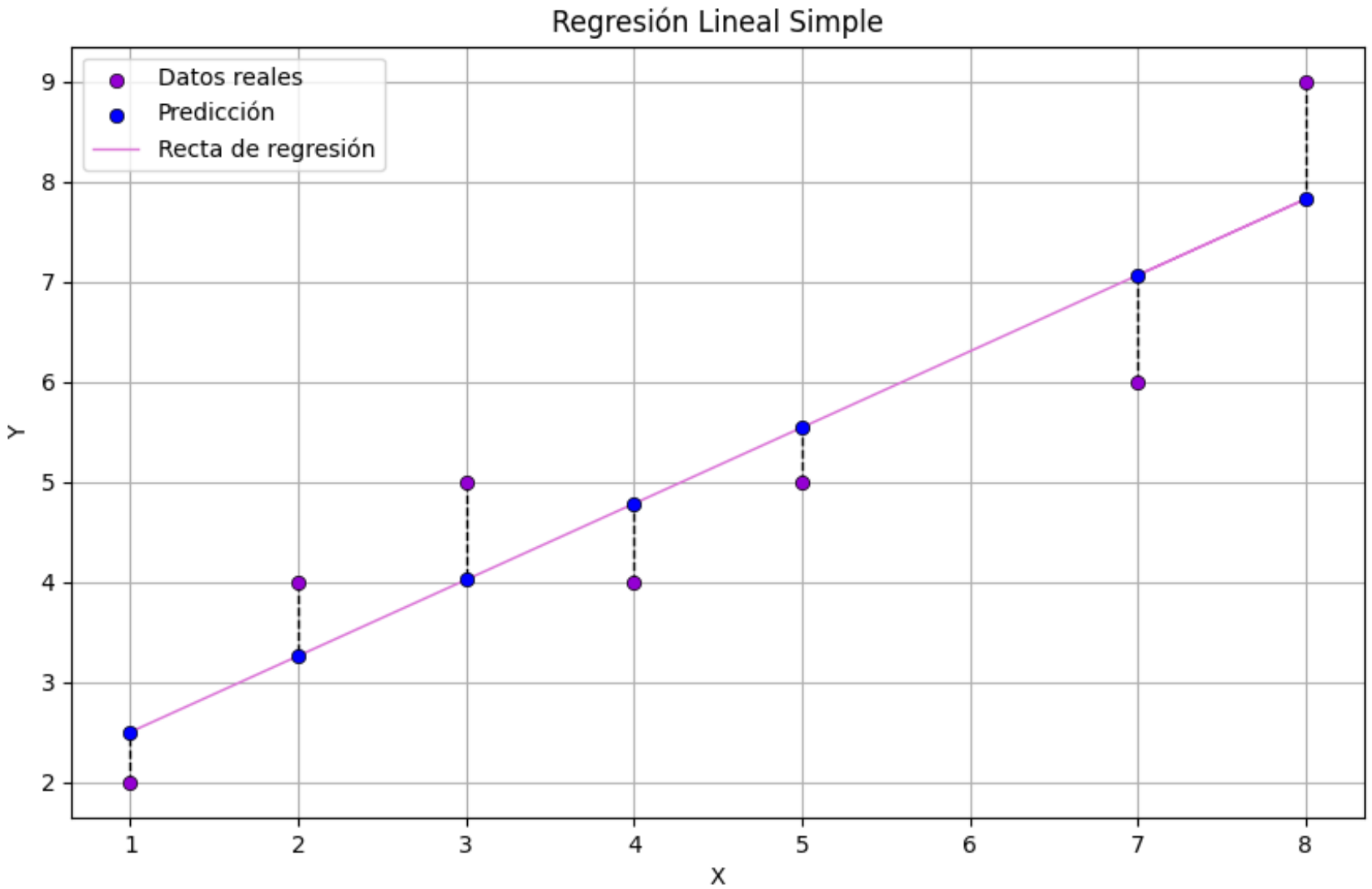
Queremos a **minimizar la suma de los cuadrados residuales (SCR)**

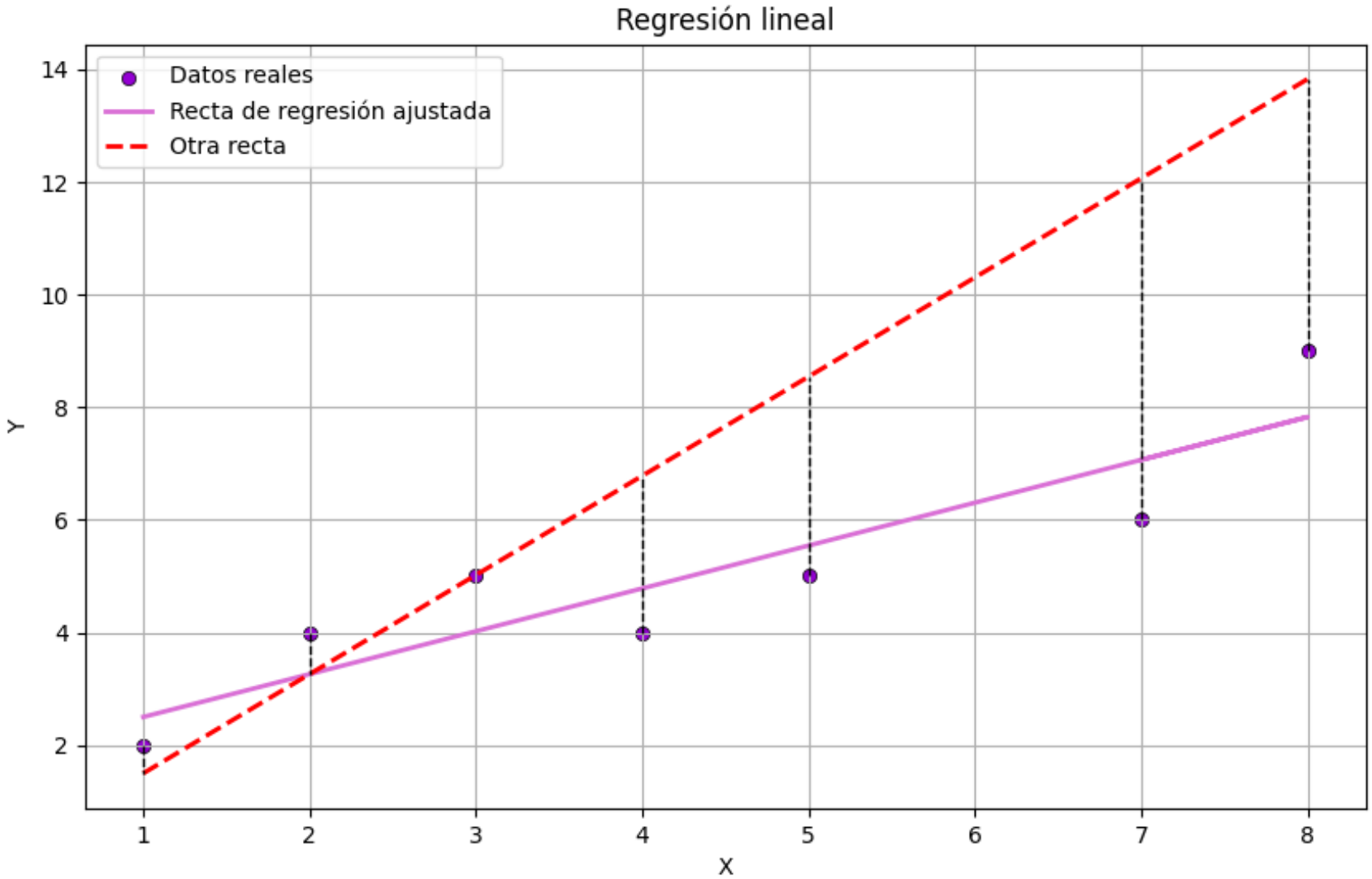
$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Aplicando el **método de mínimos cuadrados** obtenemos:

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

$$\hat{\beta}_2 = \bar{y} - \hat{\beta}_1 \bar{x}$$


Regresión Lineal Simple
Datos reales
Predicción
Recta de regresión
Y
X
1
2
3
4
5
6
7
8
9


Regresión lineal
Datos reales
Recta de regresión ajustada
Otra recta
Y
X
14
12
10
8
6
4
2
1
2
3
4
5
6
7
8

Ejercicio 1: Mediciones apareadas y normal bi-variada

La siguiente tabla resume los puntajes de dos pruebas de matemática de un grupo de estudiantes:

Prueba	**Media**	**Desvío estándar**
1era prueba	120	10
2da prueba	130	9
Correlación: 0.6		

Asumir que los datos se distribuyen como una normal bi-variada.

1. ¿Qué nota pronosticarías para la 2da prueba a un estudiante que sacó 115 en la 1era?
2. ¿Cuál es la probabilidad de que tu pronóstico anterior sea erróneo por más de 5 puntos?

Ejercicio 1

Parte 1.

Recordar que en el teórico vimos que si (X, Y) tiene distribución normal bi-variada de parámetros $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, entonces

$$U = \rho V + \sqrt{1-\rho^2} Z$$

con

$$U = \frac{Y - \mu_Y}{\sigma_Y} \sim \mathcal{N}(0,1); \quad V = \frac{X - \mu_X}{\sigma_X} \sim \mathcal{N}(0,1)$$

y $Z \sim \mathcal{N}(0,1)$ independiente de V.

En este caso la función de regresión de Y sobre X es:

$$\hat{Y} = R(x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

Para hallar $R(115)$, sustituir los parámetros por los valores dados en la tabla.

Parte 2.

Nos interesa calcular $\mathbb{P}(|Y - \hat{Y}| > 5|X = 115)$:

$$\begin{aligned}
\mathbb{P}(|Y - \hat{Y}| > 5) &= \mathbb{P}\left(\left|Y - \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right| > 5|X = 115\right) \\
&= \mathbb{P}\left(\left|\frac{Y - \mu_Y}{\sigma_Y} - \rho\frac{X - \mu_X}{\sigma_X}\right| > \frac{5}{\sigma_Y}|X = 115\right) \\
&= \mathbb{P}\left(\sqrt{1-\rho^2}|Z| > \frac{5}{\sigma_Y}|X = 115\right) \\
&= \mathbb{P}\left(|Z| > \frac{5}{\sigma_Y\sqrt{1-\rho^2}}\right) \\
&= 2\left[1 - \mathbb{P}\left(Z \leq \frac{5}{\sigma_Y\sqrt{1-\rho^2}}\right)\right] \\
&\approx 0{,}48
\end{aligned}$$

La probabilidad de que el pronóstico sea erróneo por más de 5 puntos es 0.48.

Ejercicio 2: De tal palo, tal astilla

Datos de una población indican que la altura de madres e hijas se ajustan a la distribución normal bi-variada con correlación 0.5. Ambas variables tienen media 1.63m y desvío 0.05m.

1. Hallar las ecuaciones de las rectas de regresión.
2. La altura de dos amigas difiere en 5 cm. ¿Cuál es tu pronóstico para la diferencia de alturas de las madres?
3. Entre las hijas que miden más que la media, ¿qué porcentaje son más bajas que sus madres?
4. Una madre mide menos que la media. ¿Cuál es tu pronóstico para la altura de su hija?

Ejercicio 2

$$X = \text{altura de las hijas}$$

$$Y = \text{altura de las madres}$$

Parte 3: Entre las hijas que miden más que la media, ¿qué porcentaje son más bajas que sus madres?

Queremos calcular

$$\mathbb{P}(X < Y | X > \mu_X)$$

Como X, Y tiene distribución normal bivariada, tenemos

$$U = \rho V + \sqrt{1 - \rho^2} Z$$

donde

$$U = \frac{Y - \mu_Y}{\sigma_Y} \qquad V = \frac{X - \mu_X}{\sigma_X}$$

y Z tienen distribución normal estándar, y además V es independiente de Z.

Ahora bien, como $\mu_Y = \mu_X$ y $\sigma_Y = \sigma_X$

$$\mathbb{P}(X < Y | X > \mu_X) = \mathbb{P}(V < U | V > 0)$$

Ejercicio 2

$$\mathbb{P}(V < U | V > 0) = \frac{\mathbb{P}(V < U, V > 0)}{\mathbb{P}(V > 0)}$$

Como $V \sim \mathcal{N}(0, 1)$, es claro que

$$\mathbb{P}(V > 0) = \frac{1}{2}$$

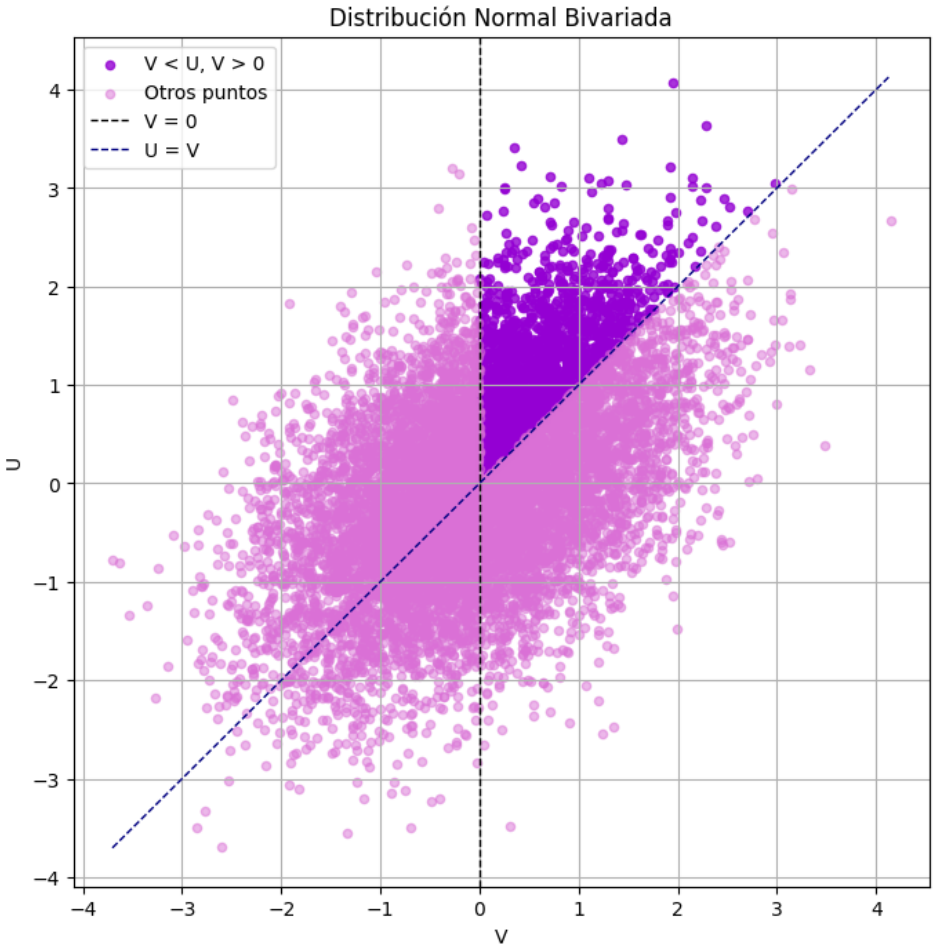
Pero $\mathbb{P}(V < U, V > 0)$ es difícil de calcular. Para enteder intuitivamente qué es, pensemos en el diagrama de dispersión de los puntos.

Para distintos valores de correlación entre dos normales:

Figura: Wikipedia

En nuestro caso $\rho = 0{,}5$.

Ejercicio 2



$\mathbb{P}(V < U, V > 0)$ es aproximadamente la proporción de puntos violetas.

Podemos aproximarla numéricamente!

`https://colab.research.google.com/drive/1pDrH1TsvlponOS-OPN6eVdJiR3NOu7pY#scrollTo=V6Eer_sL9jwM`

Entonces

$$\begin{aligned}\mathbb{P}(X < Y | X > \mu_X) = \mathbb{P}(V < U | V > 0) &= \frac{\mathbb{P}(V < U, V > 0)}{\mathbb{P}(V > 0)} \\ &= \frac{0{,}1667}{0{,}5} \\ &= 0{,}33\end{aligned}$$

Nota: La probabilidad $\mathbb{P}(V < U, V > 0)$ se puede aproximar analíticamente haciendo cuentas (calculando ángulos y usando propiedades de normales bivariadas).

Aquí vemos el poder de las simulaciones para aproximar (rápidamente) una probabilidad difícil de calcular.

Ejercicio 3

La cantidad de cierto producto químico que se disuelve en una cantidad de agua dada, a varias temperaturas, está dada por la siguiente tabla:

Y = cantidad disuelta	8	12	21	31	39	58
x = temperatura	0	10	20	30	40	60

Encontrar un intervalo de confianza para β_1 al 95 % al regresar Y sobre x.

Ejercicio 3

El modelo de regresión lineal es:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

De acuerdo a lo visto en el teórico, usando mínimos cuadrados obtenemos

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

El intervalo de confianza al $100(1-\alpha)\,\%$ para β_1 es:

$$[\hat{\beta}_1 - t_{n-2}(\alpha/2)s.e(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2}(\alpha/2)s.e(\hat{\beta}_1)]$$

donde $t_{n-2}(\alpha/2)$ proviene de la distribución t de Student con n-2 grados de libertad y un nivel de significancia del 5 % y

$$s.e(\hat{\beta}_1) = \sqrt{\frac{1}{n-2}\frac{SRC}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{1}{n-2}\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

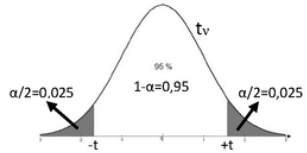
es el estimador del desvío estándar de $\hat{\beta}_1$.

Ejercicio 3

Para encontrar $t_{n-2}(\alpha/2)$ tenemos que ver la tabla de la distribución t de Student para $n-2=4$ grados de libertad y $\alpha/2=0{,}025$:

Distribución t de *Student*

Contiene los valores de t tales que $\frac{\alpha}{2} = P(t_\nu \geq t)$, donde ν son los Grados de Libertad



		α/2													
		0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,25	0,3	0,4	0,45	0,475
ν grados de libertad	1	636,619	318,309	63,657	31,821	12,706	6,314	3,078	1,376	1,000	0,727	0,325	0,158	0,079
	2	31,599	22,327	9,925	6,965	4,303	2,920	1,886	1,061	0,816	0,617	0,289	0,142	0,071
	3	12,924	10,215	5,841	4,541	3,182	2,353	1,638	0,978	0,765	0,584	0,277	0,137	0,068
	4	8,610	7,173	4,604	3,747	2,776	2,132	1,533	0,941	0,741	0,569	0,271	0,134	0,067
	5	6,869	5,893	4,032	3,365	2,571	2,015	1,476	0,920	0,727	0,559	0,267	0,132	0,066
	6	5,959	5,208	3,707	3,143	2,447	1,943	1,440	0,906	0,718	0,553	0,265	0,131	0,065
	7	5,408	4,785	3,499	2,998	2,365	1,895	1,415	0,896	0,711	0,549	0,263	0,130	0,065
	8	5,041	4,501	3,355	2,896	2,306	1,860	1,397	0,889	0,706	0,546	0,262	0,130	0,065
	9	4,781	4,297	3,250	2,821	2,262	1,833	1,383	0,883	0,703	0,543	0,261	0,129	0,064
	10	4,587	4,144	3,169	2,764	2,228	1,812	1,372	0,879	0,700	0,542	0,260	0,129	0,064
	11	4,437	4,025	3,106	2,718	2,201	1,796	1,363	0,876	0,697	0,540	0,260	0,129	0,064
	12	4,318	3,930	3,055	2,681	2,179	1,782	1,356	0,873	0,695	0,539	0,259	0,128	0,064
	13	4,221	3,852	3,012	2,650	2,160	1,771	1,350	0,870	0,694	0,538	0,259	0,128	0,064
	14	4,140	3,787	2,977	2,624	2,145	1,761	1,345	0,868	0,692	0,537	0,258	0,128	0,064
	15	4,073	3,733	2,947	2,602	2,131	1,753	1,341	0,866	0,691	0,536	0,258	0,128	0,064
	16	4,015	3,686	2,921	2,583	2,120	1,746	1,337	0,865	0,690	0,535	0,258	0,128	0,064
	17	3,965	3,646	2,898	2,567	2,110	1,740	1,333	0,863	0,689	0,534	0,257	0,128	0,064
	18	3,922	3,610	2,878	2,552	2,101	1,734	1,330	0,862	0,688	0,534	0,257	0,127	0,064
	19	3,883	3,579	2,861	2,539	2,093	1,729	1,328	0,861	0,688	0,533	0,257	0,127	0,064
	20	3,850	3,552	2,845	2,528	2,086	1,725	1,325	0,860	0,687	0,533	0,257	0,127	0,063
	21	3,819	3,527	2,831	2,518	2,080	1,721	1,323	0,859	0,686	0,532	0,257	0,127	0,063
	22	3,792	3,505	2,819	2,508	2,074	1,717	1,321	0,858	0,686	0,532	0,256	0,127	0,063
	23	3,768	3,485	2,807	2,500	2,069	1,714	1,319	0,858	0,685	0,532	0,256	0,127	0,063
	24	3,745	3,467	2,797	2,492	2,064	1,711	1,318	0,857	0,685	0,531	0,256	0,127	0,063
	25	3,725	3,450	2,787	2,485	2,060	1,708	1,316	0,856	0,684	0,531	0,256	0,127	0,063
	26	3,707	3,435	2,779	2,479	2,056	1,706	1,315	0,856	0,684	0,531	0,256	0,127	0,063
	27	3,690	3,421	2,771	2,473	2,052	1,703	1,314	0,855	0,684	0,531	0,256	0,127	0,063
	28	3,674	3,408	2,763	2,467	2,048	1,701	1,313	0,855	0,683	0,530	0,256	0,127	0,063
	29	3,659	3,396	2,756	2,462	2,045	1,699	1,311	0,854	0,683	0,530	0,256	0,127	0,063
	30	3,646	3,385	2,750	2,457	2,042	1,697	1,310	0,854	0,683	0,530	0,256	0,127	0,063
	31	3,633	3,375	2,744	2,453	2,040	1,696	1,309	0,853	0,682	0,530	0,256	0,127	0,063
	32	3,622	3,365	2,738	2,449	2,037	1,694	1,309	0,853	0,682	0,530	0,255	0,127	0,063
	33	3,611	3,356	2,733	2,445	2,035	1,692	1,308	0,853	0,682	0,530	0,255	0,127	0,063
	34	3,601	3,348	2,728	2,441	2,032	1,691	1,307	0,852	0,682	0,529	0,255	0,127	0,063
	35	3,591	3,340	2,724	2,438	2,030	1,690	1,306	0,852	0,682	0,529	0,255	0,127	0,063
	α	0,001	0,002	0,01	0,02	0,05	0,1	0,2	0,4	0,5	0,6	0,8	0,9	0,95

Figura: www.studocu.com

Ejercicio 3

Luego $t_{n-2}(\alpha/2) = 2{,}776$.

Podemos chequearlo en Python o R!

`https://colab.research.google.com/drive/1pDrH1TsvlponOS-OPN6eVdJiR3NOu7pY#scrollTo=V6Eer_sL9jwM`

Ejercicio 4

Los siguientes datos representan la altura x y el peso Y de 12 individuos. Aplicar un modelo de regresión lineal para predecir el peso a partir de la altura.

Altura (x)	Peso (Y)
163	62
163	63
165	65
166	67
168	68
169	69
170	70
170	72
171	71
172	70
172	72
174	74

Probar un TdH de que la pendiente es nula (i.e la altura no sirve para predecir el peso).

Ejercicio 4

La ecuación del modelo de regresión lineal es:

$$Y = \beta_0 + \beta_1 x$$

donde β_0 es el intercepto y β_1 es la pendiente.

Queremos estimar β_0 y β_1 a partir de los datos, para construir el modelo.

Vamos a hacer el siguiente test de hipótesis (TdH):

- $H_0 : \beta_1 = 0$ (la pendiente es cero, la altura no predice el peso).
- $H_1 : \beta_1 \neq 0$ (la pendiente no es cero, la altura sí predice el peso).

Sugerencia: Para generar algo de intuición conviene graficar los datos (diagrama de dispersión).

Ejercicio 4

Para el TdH necesitamos calcular el valor T_1 para la pendiente:

$$T_1 = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$$

donde $s.e(\hat{\beta}_1)$ es el error estándar de la pendiente:

$$s.e(\hat{\beta}_1) = \sqrt{\frac{1}{n-2}\frac{SRC}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{1}{n-2}\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Después comparamos T_1 con el valor crítico $t_{n-2}(\alpha/2)$:

- Si el valor T_1 calculado es mayor en valor absoluto que el valor crítico, **se rechaza la hipótesis nula**.
- Si el valor T_1 no es significativo, **no se puede rechazar la hipótesis nula**.

- Viernes 20/9: No hay práctico (asueto Universidad)
- Viernes 27/9: semana parciales FING