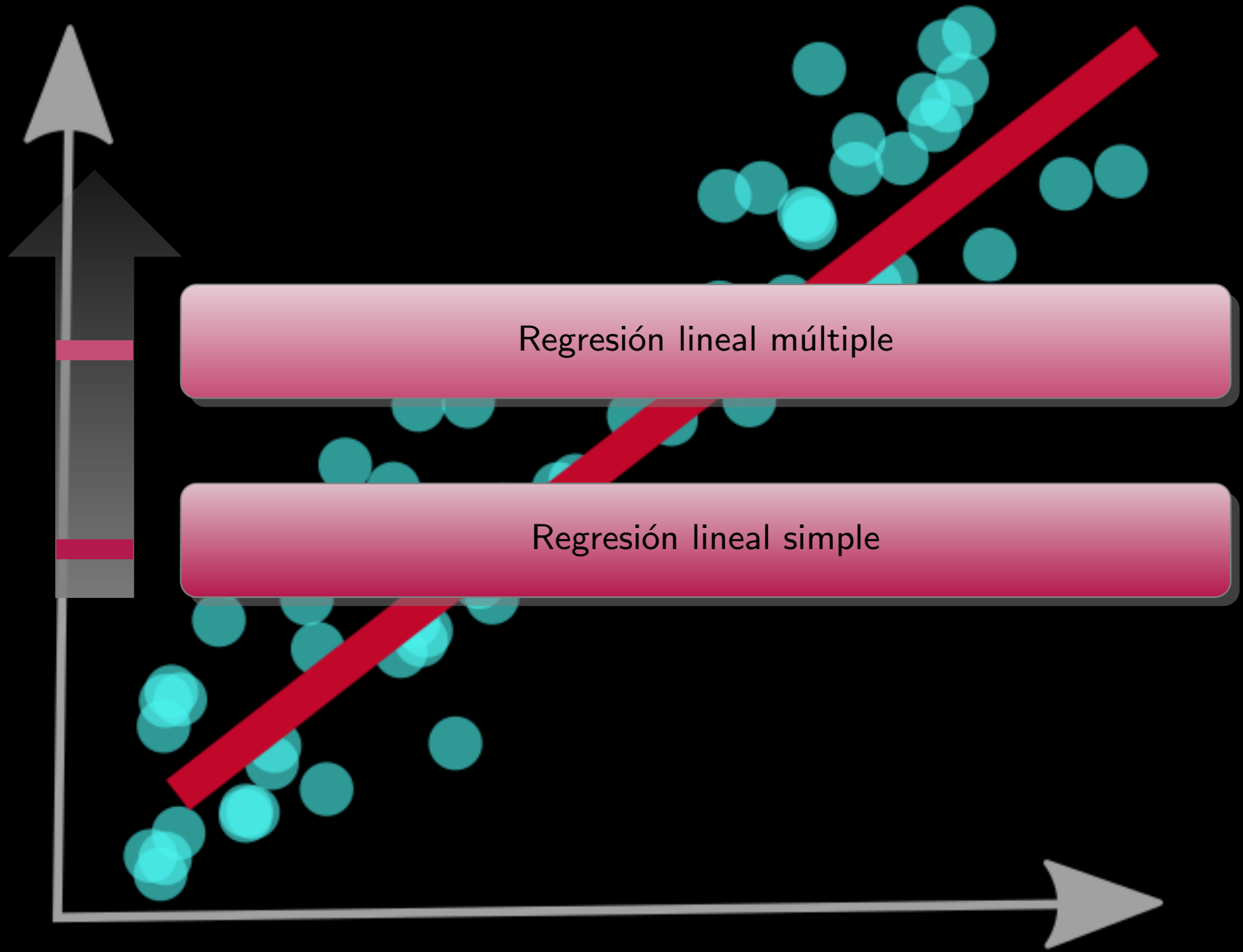
Fundamentos de Análisis de Datos
Regresión Lineal

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA


Regresión lineal múltiple
Regresión lineal simple

Tipo de variables

Variables predictoras

Son las variables que explican el fenómeno, usualmente denotadas por $x_1, x_2, \ldots, x_k$, o simplemente x en el caso de que sea una sola.
También se las llama variables explicativas.

Variable respuesta

Es la variable explicada por el fenómeno, usualmente denotada por y.

Modelos

Regresión lineal simple

Existe una única variable predictora y el modelo es de la forma:

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

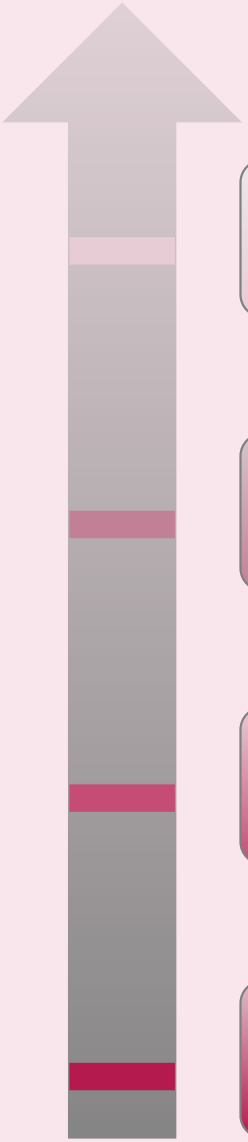
Regresión lineal múltiple

Existen al menos dos variables predictoras y el modelo es de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

Los β_i son los coeficientes del modelo y ϵ es el error bajo el cual se realizan ciertos supuestos.

REGRESIÓN LINEAL SIMPLE



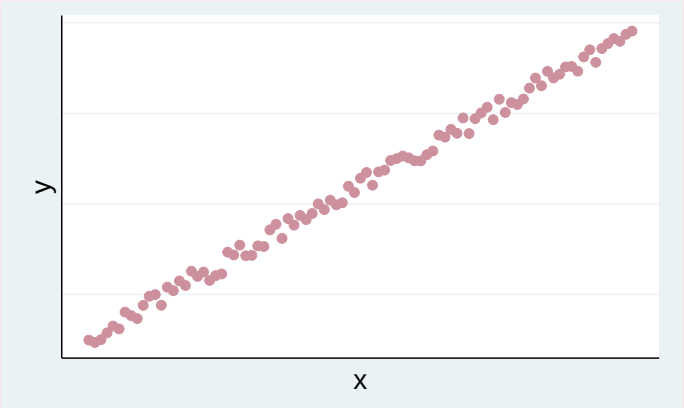
Se ajustan los puntos mediante una recta

Se grafican los puntos en un scatterplot

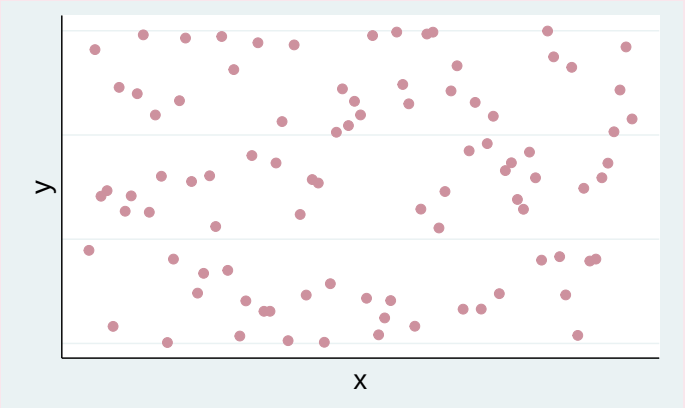
Una es la variable respuesta a la otra variable

Estudio simultáneo de dos variables continuas

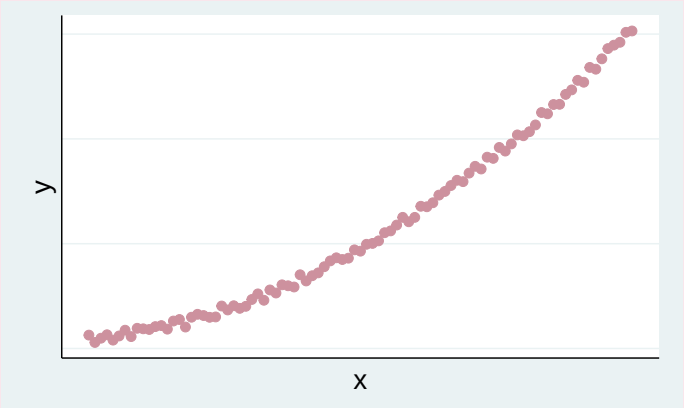
Scatterplots



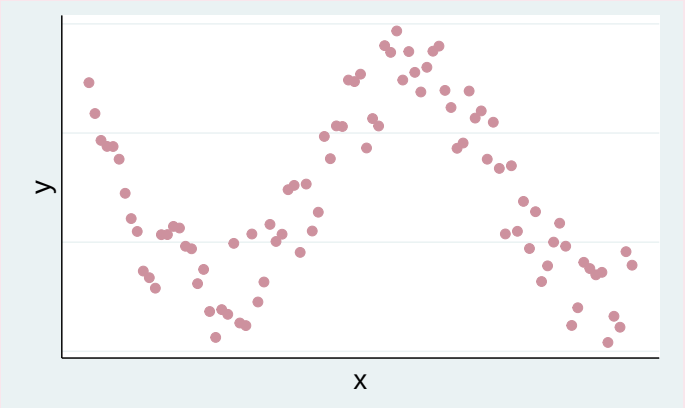
Datos relacionados linealmente



Datos no relacionados



Datos relacionados no linealmente



Datos relacionados no linealmente

Relación lineal

Aún cuando x e y están relacionadas linealmente, esta relación rara vez es perfecta.

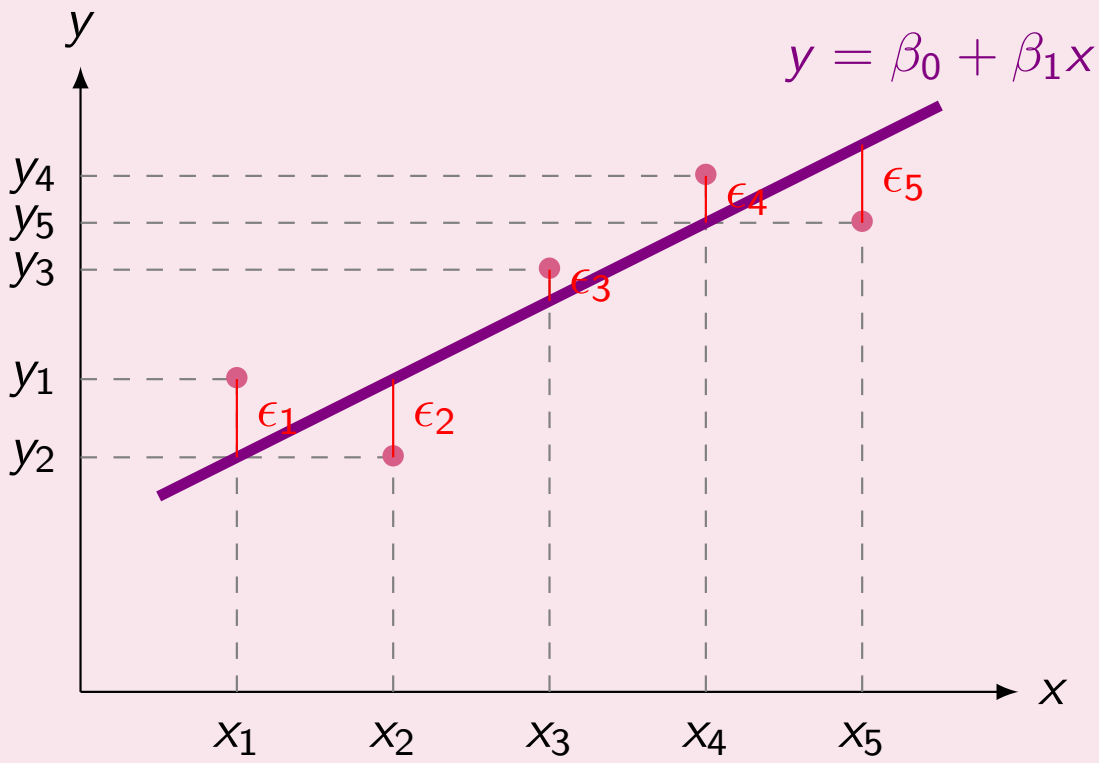
Pregunta

¿Cómo podemos encontrar la recta que mejor modela los datos?

Uno de los modelos más utilizados es la **regresión lineal** que consiste en una técnica simple y de fácil interpretación.

Modelo de regresión lineal simple

Representación gráfica para 5 puntos:



Modelo de regresión lineal simple

Dados los puntos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, modelamos la recta de ajuste como:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ para } i = 1, 2, \ldots, n,$$

donde ϵ_i son independientes y satisfacen:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Esto significa que estamos asumiendo que la variable respuesta está relacionada con la variable predictora más una componente aleatoria normalmente distribuida.

Modelo de regresión lineal simple

Coeficientes de regresión

β_0 es la ordenada al origen

β_1 es la pendiente

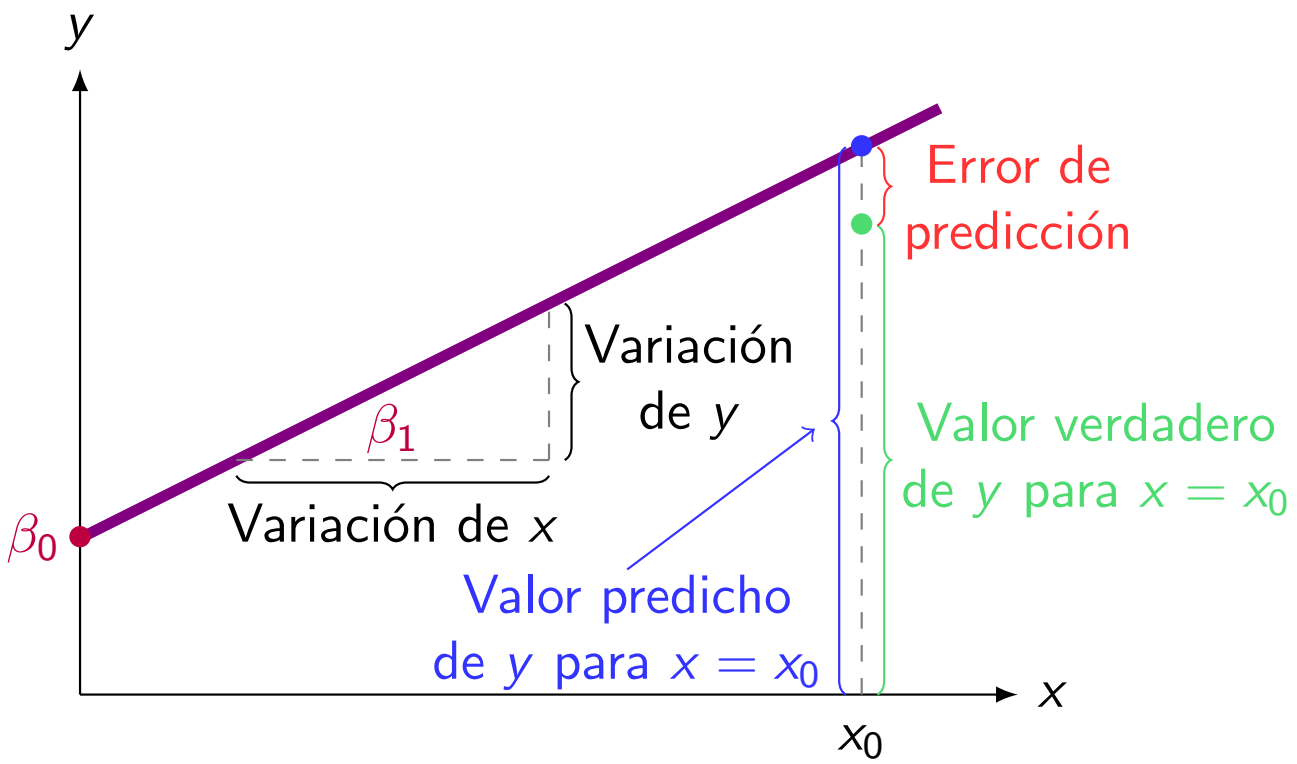
Objetivo

Estimar los coeficientes de regresión a partir de los datos.

Observar que la varianza σ^2 que se introduce al suponer que los errores son independientes y normales, es un tercer parámetro desconocido por lo que también debe ser estimada.

Modelo de regresión lineal simple

Interpretación gráfica



Método de cuadrados mínimos

Recta de cuadrados mínimos

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Residuos

$$e_i = y_i - \hat{y}_i \text{ para } i = 1, 2, \ldots, n$$

Problema de optimización

Buscamos hallar $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimicen la suma de los cuadrados de los residuos:

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

Método de cuadrados mínimos

Puesto que la función a minimizar es una función de β_0 y β_1, debemos hallar las derivadas parciales e igualarlas a cero:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0,$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = -2 \sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0.$$

Deducimos el siguiente sistema lineal:

$$\begin{cases} \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - n\beta_0 & = 0, \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 & = 0. \end{cases}$$

Método de cuadrados mínimos

Despejando β_0 de la primera ecuación:

$$\beta_0 = \frac{1}{n}\left[\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} \frac{y_i}{n} - \beta_1 \sum_{i=1}^{n} \frac{x_i}{n} = \bar{y} - \beta_1 \bar{x}.$$

Método de cuadrados mínimos

Reemplazando en la segunda ecuación:

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\beta_1 = \frac{\bar{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i}{\bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}$$

$$\beta_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i}{\frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}$$

$$\beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Método de cuadrados mínimos

Realicemos el siguiente cálculo:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}[x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}] \\
&= \sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n} \bar{x}y_i + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\frac{1}{n}\sum_{i=1}^{n} x_i\frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} x_iy_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} x_iy_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i.
\end{aligned}$$

Si multiplicamos esta expresión por n obtenemos el numerador de β_1.

Método de cuadrados mínimos

Realicemos el siguiente cálculo:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}[x_i^2-2x_i\bar{x}+\bar{x}^2] \\
&= \sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\bar{x}+\sum_{i=1}^{n}\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\frac{1}{n}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 \\
&= \sum_{i=1}^{n}x_i^2-\frac{2}{n}\left(\sum_{i=1}^{n}x_i\right)^2+\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \\
&= \sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 .
\end{aligned}$$

Si multiplicamos esta expresión por n obtenemos el denominador de β_1.

Método de cuadrados mínimos

Luego, podemos concluir lo siguiente:

Coeficientes de regresión por mínimos cuadrados

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

Notación para las sumas

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Método de cuadrados mínimos: Ejemplo

Estamos interesados en analizar la relación entre la cantidad de horas que un grupo de estudiantes dedicó a realizar un trabajo práctico y la calificación obtenida.

Observaciones

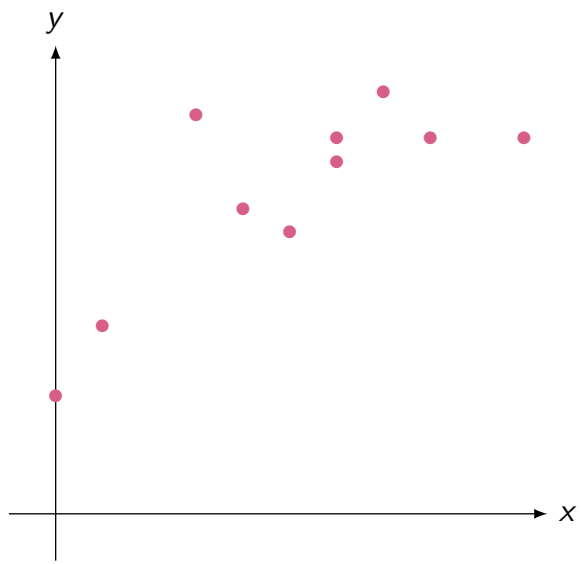
	Cantidad de horas	Calificación
	6	8.0
	10	8.0
	3	8.5
	4	6.5
	6	7.5
	7	9.0
	0	2.5
	1	4.0
	8	8.0
	5	6.0
Promedio	5	6.8

Observar que $n = 10$.

Representación gráfica

x: Cantidad de horas

y: Calificación



Método de cuadrados mínimos: Ejemplo

Calculamos la tabla para mínimos cuadrados para obtener los coeficientes estimados:

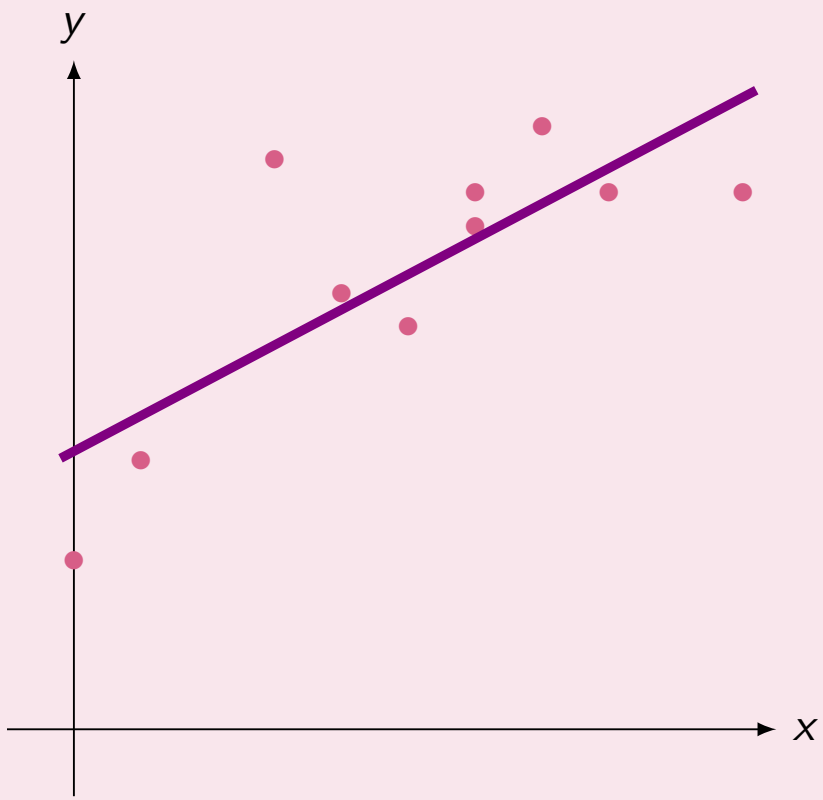
	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
	6	8.0	1	1.2	1.2	1
	10	8.0	5	1.2	6.0	25
	3	8.5	-2	1.7	-3.4	4
	4	6.5	-1	-0.3	0.3	1
	6	7.5	1	0.7	0.7	1
	7	9.0	2	2.2	4.4	4
	0	2.5	-5	-4.3	21.5	25
	1	4.0	-4	-2.8	11.2	16
	8	8.0	3	1.2	3.6	9
	5	6.0	0	-0.8	0.0	0
Totales	50	68	0	0	45.5	86

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{45.5}{86} = \boxed{0.529} \Longrightarrow \hat{\beta}_0 = 6.8 - 0.529 \cdot 5 = \boxed{4.155}$$

Método de cuadrados mínimos: Ejemplo

Luego:

$$\boxed{\hat{y} = 4.155 + 0.529x}.$$



Propiedades de la recta de regresión

- Consideramos el ordenamiento de las observaciones de la variable predictiva dado por $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. La interpretación de la recta de regresión es significativa en el intervalo $[x_{(1)}, x_{(n)}]$. Esto quiere decir que una predicción a partir de un valor fuera de este intervalo, podría no ser válido.
- Otra manera de escribir la recta de regresión es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \Longrightarrow \boxed{\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})}.$$

- De la expresión anterior, resulta claro que la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$.

Propiedades de la recta de regresión

- Usando la expresión anterior, tenemos que los residuos:

$$e_i = y_i - \hat{y}_i = y_i - \left(\bar{y} + \hat{\beta}_1(x_i - \bar{x})\right) = y_i - \bar{y} - \hat{\beta}_1 x_i + \hat{\beta}_1 \bar{x}.$$

Aplicando la suma:

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} - \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} \bar{x} \\
&= \sum_{i=1}^{n} y_i - n\bar{y} - \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_1 n\bar{x} \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i \Longrightarrow \boxed{\sum_{i=1}^{n} e_i = 0}.
\end{aligned}$$

Propiedades de la recta de regresión

- Calculando el promedio de las predicciones:

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}[\bar{y} + \hat{\beta}_1(x - \bar{x}_i)]$$

$$= \frac{1}{n}\underbrace{\sum_{i=1}^{n}\bar{y}}_{n\bar{y}} + \frac{\hat{\beta}_1}{n}\underbrace{\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\right)}_{=0} \Longrightarrow \boxed{\bar{\hat{y}} = \bar{y}}.$$

- El estimador $\hat{\beta}_1$ está directamente relacionado con el coeficiente de correlación de Pearson r:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \underbrace{\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}}_{r}\sqrt{\frac{S_{yy}}{S_{xx}}} \Longrightarrow \boxed{\hat{\beta}_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}}.$$

Ejemplo*

*Base de datos `father.son` del paquete `UsingR` de `R`

REGRESIÓN LINEAL MÚLTIPLE

Modelo de regresión lineal múltiple

Supongamos que tenemos n observaciones:

$$\begin{aligned}
&(x_{11}, x_{12}, \ldots, x_{1k}, y_1)\\
&(x_{21}, x_{22}, \ldots, x_{2k}, y_2)\\
&\vdots\\
&(x_{n1}, x_{n2}, \ldots, x_{nk}, y_n)
\end{aligned}$$

que satisfacen el modelo; es decir:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \text{ para } i = 1, 2, \ldots, n,$$

donde ϵ_i son independientes y satisfacen:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Modelo de regresión lineal múltiple

Usando notación matricial, el modelo puede escribirse como:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

La matriz $\boldsymbol{X}$ se llama **matriz de diseño**.

Modelo de regresión lineal múltiple

Al igual que en el caso de regresión lineal simple, recurrimos al método de cuadrados mínimos para obtener:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Interpretación de los estimadores de los coeficientes del modelo

- $\hat{\beta}_j$ representa el cambio parcial de y cuando el valor de x_j varía pero el resto de las variables predictoras no cambian.
- $\hat{\beta}_0$ puede interpretarse como el valor medio esperado de la variable respuesta cuando todas las variables predictoras se anulan. Sin embargo, muchas veces esta interpretación no tiene sentido en la realidad donde el experimento no permite la anulación de todas las variables explicativas.

Modelo de regresión lineal múltiple

Propiedades de los estimadores por mínimos cuadrados

- $\hat{\beta}$ es **insesgado**; es decir, $\mathrm{E}(\hat{\beta}) = \beta$.
- $\hat{\sigma}^2$ es **insesgado**; es decir, $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.
- $\hat{\beta}$ es **consistente**; es decir, $\hat{\beta}$ converge a β cuando n tiende a infinito.
- $\hat{\beta}$ está **asintóticamente normalmente distribuido**.
- $\hat{\beta}$ es el mejor estimador insesgado de β; es decir, el que tiene menor varianza.

Análisis de la regresión

Estimación de la varianza

La varianza poblacional de los residuos se estima mediante la varianza muestral como:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

La raíz de la varianza muestral, s, suele llamarse **error estándar residual**.

Análisis de la regresión

Usando los supuestos de independencia y normalidad sobre los errores, se pueden deducir las siguientes fórmulas:

Varianza de $\hat{\beta}_j$

La varianza de la estimación del coeficiente j-ésimo ($j = 1, 2, \ldots, k$) es:

$$\sigma^2_{\hat{\beta}_j} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}\sigma^2.$$

Cuando las varianzas de los coeficientes son estimadas mediante el error estándar residual, notamos $\hat{\sigma}^2_{\hat{\beta}_j}$.

Análisis de la regresión

En el caso particular de regresión lineal simple, se cumple lo siguiente:

Varianza de $\hat{\beta}_0$

La varianza de la estimación de la ordenada al origen es:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

Varianza de $\hat{\beta}_1$

La varianza de la estimación de la pendiente es:

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

Análisis de la regresión

Observación

Si existe un j tal que $\beta_j = 0$, el modelo no contiene a la variable x_j. Esto significa que la variable j-ésima no contribuye en la explicación de la variable y.

Pregunta

¿Cómo podemos probar si una variable independiente está asociada con la variable respuesta, en el sentido de que ayuda a explicar sus variaciones?

Intervalos de confianza

Los intervalos de confianza $(1-\alpha)100\%$ para los coeficientes del modelo son:

$$\hat{\beta}_j \pm t_{n-k-1;1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_j}.$$

Esto significa que existe un $(1-\alpha)100\%$ de posibilidades de que el intervalo contenga al parámetro correspondiente.

Regla de decisión

Si el intervalo de confianza no contiene al 0 podemos concluir que β_j no se anula y, por lo tanto, la variable independiente j-ésima está asociada con y. Sin embargo, si el 0 pertenece al intervalo, no podemos concluir que existe una asociación entre estas variables.

Test de hipótesis

Hipótesis

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

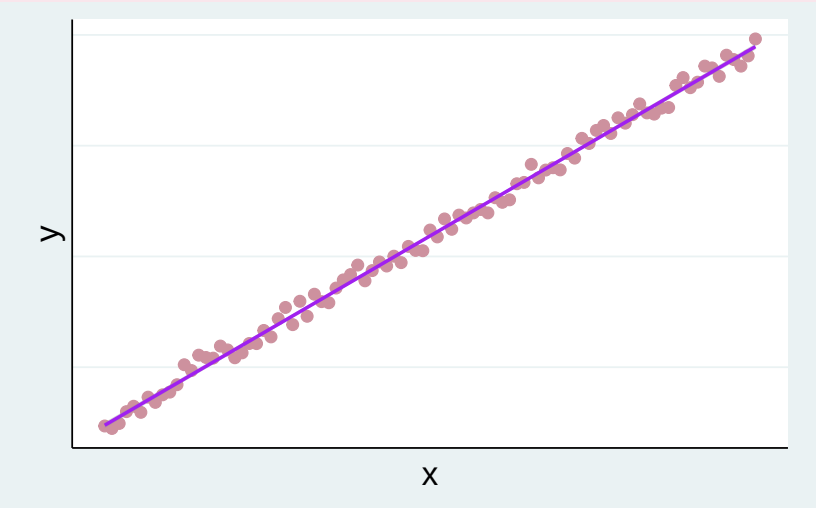
Estadístico

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim T_{n-k-1}$$

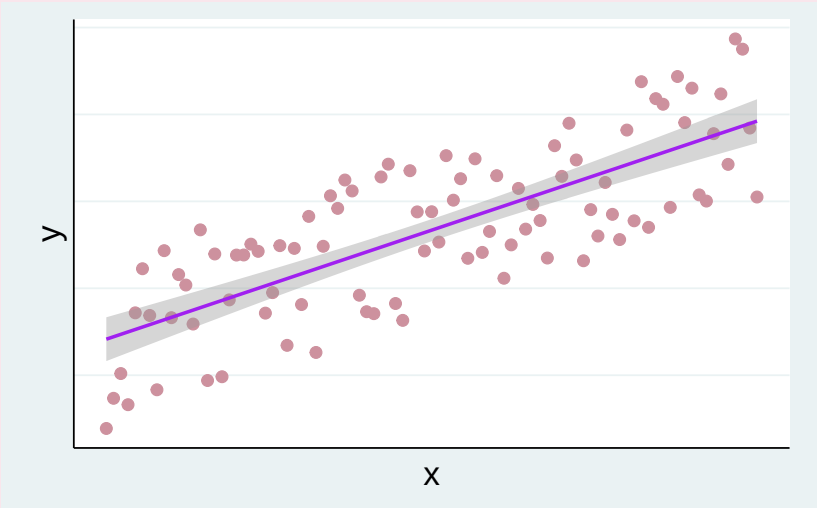
Regla de decisión

Si $|T| > t_{t-k-1;1-\frac{\alpha}{2}}$ rechazamos H_0 con significancia α.

Bondad de ajuste



Relación lineal fuerte



Relación lineal débil

Pregunta

¿Cómo podemos evaluar cuantitativamente la calidad del modelo de ajuste?

Bondad de ajuste

Descomposición de la varianza

La variación total de y puede descomponerse como sigue:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE}}.$$

- SST indica la suma de cuadrados totales que representa la variabilidad total de y respecto de su media y es proporcional a la varianza muestral.
- SSR indica la suma de cuadrados del modelo de regresión que es la variabilidad explicada por el modelo de regresión.
- SSE indica la suma de cuadrados del error que refleja la variación ocasionada por los errores aleatorios involucrados en el modelo.

Bondad de ajuste

Valores grandes de SSE indican que las desviaciones de las observaciones con respecto a la recta de regresión son grandes, implicando un mal ajuste de los datos a partir del modelo.

Lo deseable es minimizar el valor de SSE.

Para juzgar la bondad del ajuste, se puede estudiar el valor de SSE en relación con SST como sigue.

- En la situación ideal en que SSE se anula, tenemos que SST=SSR lo que indica una bondad de ajuste es óptima.
- Si el valor de SSE es muy grande, el valor de SSR es más pequeño lo que indica una bondad del ajuste pobre.
- En el caso en que SSR se anule, el ajuste del modelo es el peor posible.

Coeficiente de determinación

Definición

Es la proporción de variación en la variable respuesta que es explicada por la regresión. Se define como:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Interpretación

- Cuanto más cercano a 1 sea el valor de R^2, mejor será la bondad de ajuste.
- Cuanto más cercano a 0 sea el valor de R^2, peor será la bondad de ajuste.
- Si $R^2 = a$, significa que sólo el $100a\%$ de la variación de los datos es explicado por el modelo.

Coeficiente de determinación

Propiedades

- Si el término independiente del modelo de regresión no es nulo (*with intercept*), entonces $0 \leq R^2 \leq 1$.
- Si el término independiente del modelo de regresión es nulo (*without intercept*):

$$R^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2}.$$

- Si el modelo tiene una única variable predictora, $R^2 = r^2$.

Coeficiente de determinación

Calculemos R^2 para nuestro ejemplo anterior:

	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$\hat{y}_i - \bar{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
	6	8.0	7.329	1.2	0.529	1.44	0.2798
	10	8.0	9.445	1.2	2.645	1.44	6.9960
	3	8.5	5.742	1.7	-1.058	2.89	1.1194
	4	6.5	6.271	-0.3	-0.529	0.09	0.2798
	6	7.5	7.329	0.7	0.529	0.49	0.2798
	7	9.0	7.858	2.2	1.058	4.84	1.1194
	0	2.5	4.155	-4.3	-2.645	18.49	6.9960
	1	4.0	4.684	-2.8	-2.116	7.84	4.4775
	8	8.0	8.387	1.2	1.587	1.44	2.5186
	5	6.0	6.800	-0.8	0.000	0.64	0.0000
Totales	50	68	68	0.00	0.00	39.60	24.0663

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{24.0663}{39.60} \Longrightarrow \boxed{R^2 = 0.6077}$$

Concluimos que el 60.77% de las variaciones en las calificaciones está explicado por la cantidad de horas dedicadas al trabajo práctico.

Coeficiente de determinación

Problema

El coeficiente de determinación no penaliza la inclusión de variables explicativas no significativas. Es decir, si se añaden nuevas variables explicativas al modelo, tales que la relación que guardan con la variable respuesta es poca, el valor de R^2 aumentará. Es por ello que algunos expertos se oponen al uso de esta medida como un indicador representativo de la bondad del ajuste real.

Coeficiente de determinación ajustado

Si k denota la cantidad de variables explicativas del modelo:

$$R^2_{\text{ajustado}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2).$$

Observar que $R^2_{\text{ajustado}} < R^2$ y que la igualdad sólo se da si $R^2 = 1$.

Coeficiente de determinación

El coeficiente de determinación corregido para nuestro ejemplo es:

$$
\begin{aligned}
R^2_{\text{ajustado}} &= 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2) \\
&= 1 - \left(\frac{10-1}{10-1-1}\right)(1-0.6077) \\
&= 1 - \frac{9}{8}\,0.3923 \Longrightarrow \boxed{R^2_{\text{ajustado}} = 0.5587}.
\end{aligned}
$$

Validación del modelo

Test de hipótesis

H_0 : No existe una relación entre $\boldsymbol{x}$ e y.

H_1 : Existe algún tipo de relación entre $\boldsymbol{x}$ e y.

Las hipótesis pueden reescribirse matemáticamente como sigue.

Test de hipótesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \text{ para algún } j = 1, 2, \ldots, k$$

Observar que el supuesto se realiza sobre los coeficientes de las variables predictoras, ya que no tiene mucho sentido hacer supuestos sobre la ordenada al origen.

Validación del modelo

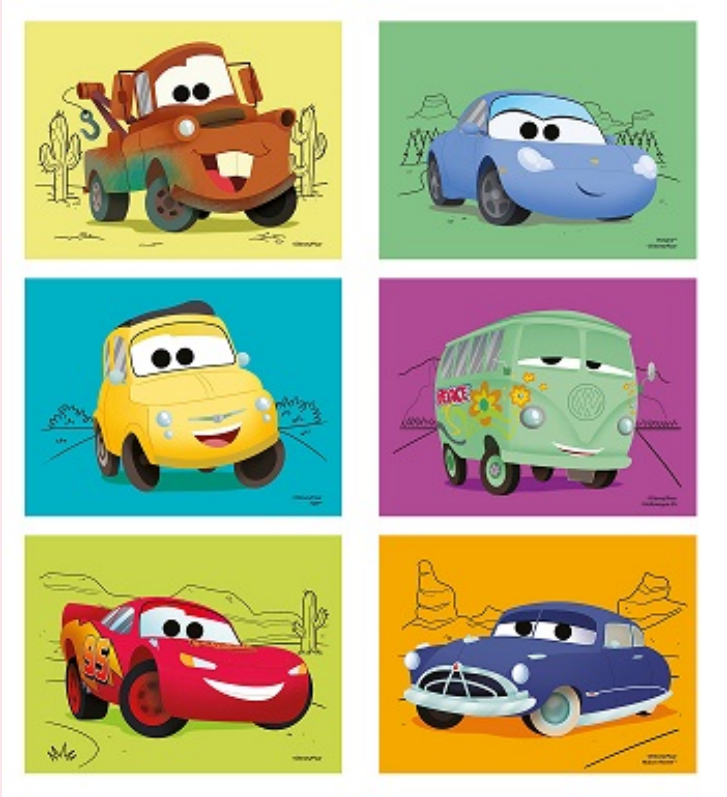
Estadístico

$$F = \frac{n-k-1}{k}\frac{\text{SSR}}{\text{SSE}} \sim \mathcal{F}_{k,n-k-1}$$

Regla de decisión

Si $F > F_{k,n-k-1;1-\alpha}$ rechazamos H_0 con significancia α.

Ejemplo*



*Base de datos `mtcars` de `R`

Regresión log-lineal

Suele usarse en situaciones donde el crecimiento o la caída de la variable respuesta es muy acelerado en un comienzo con una tendencia a suavizarse con el correr del tiempo.

Modelo

$$\ln(\boldsymbol{y}) = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_k \boldsymbol{x}_k + \boldsymbol{\epsilon},$$

donde el vector $\boldsymbol{y}$ tiene entradas positivas y los errores son independientes y normales.

Vemos que esto es un modelo de regresión lineal con variable dependiente $\ln(y)$.

Pregunta

¿Cómo podemos interpretar su resultado?

Regresión log-lineal

Observemos que al aplicar la función exponencial:

$$\begin{aligned} \boldsymbol{y} &= \exp\left(\beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_k \boldsymbol{x}_k + \boldsymbol{\epsilon}\right) \\ &= e^{\beta_0 \mathbf{1}} \cdot e^{\beta_1 \boldsymbol{x}_1} \cdot e^{\beta_2 \boldsymbol{x}_2} \cdot \ldots \cdot e^{\beta_k \boldsymbol{x}_k} \cdot e^{\boldsymbol{\epsilon}}. \end{aligned}$$

Supongamos que la variable j-ésima se incrementa en un una unidad, entonces el valor de la variable respuesta se multiplica por e^{β_j}. Luego, debe interpretarse el efecto de este incremento como e^{β_j} en lugar de β_j. Es decir, para un incremento de una unidad en la variable independiente, la variable dependiente se incrementa en un $(e^{\hat{\beta}_j} - 1)100\%$.

Otras transformaciones logarítmicas

Modelo de regresión linear-log

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \ln(\boldsymbol{x}_1) + \beta_2 \ln(\boldsymbol{x}_2) + \cdots + \beta_k \ln(\boldsymbol{x}_k) + \boldsymbol{\epsilon},$$

donde las variables predictoras tienen entradas positivas y los errores son independientes y normales.

Modelo de regresión log-log

$$\ln(\boldsymbol{y}) = \beta_0 \mathbf{1} + \beta_1 \ln(\boldsymbol{x}_1) + \beta_2 \ln(\boldsymbol{x}_2) + \cdots + \beta_k \ln(\boldsymbol{x}_k) + \boldsymbol{\epsilon},$$

donde tanto las variables predictoras como la variable respuesta tienen entradas positivas y los errores son independientes y normales.

Ejemplo*

*Base de datos simulada


¿Esto no lo vimos?

ANOVA

La manera clásica de presentar el ANOVA de un factor para k niveles se basa en la introducción de las variables "dummy":

$$x_{ij} = \begin{cases} 1 & \text{si la } i\text{-ésima observación corresponde al nivel } j, \\ 0 & \text{si la } i\text{-ésima observación no corresponde al nivel } j. \end{cases}$$

Modelo

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_k x_{ik} + \epsilon_i \text{ para } i = 1, 2, \ldots, n,$$

donde ϵ_i son independientes y normales.

Con el fin de que μ coincida con $\bar{y}$, se impone la restricción $\boxed{\sum_{j=1}^{k} \alpha_j = 0}$.

El coeficientes α_j mide el desvío con respecto a la media de $\boldsymbol{y}$ debido a este nivel del factor.

ANOVA

Si las variable independientes son categóricas, la codificación de las variables predictoras puede extenderse a situaciones más generales con más de un factor y con la posibilidad de interacciones entre los factores.

Esto permite escribir el ANOVA multifactorial como un modelo de regresión lineal múltiple.

ANOVA versus regresión lineal

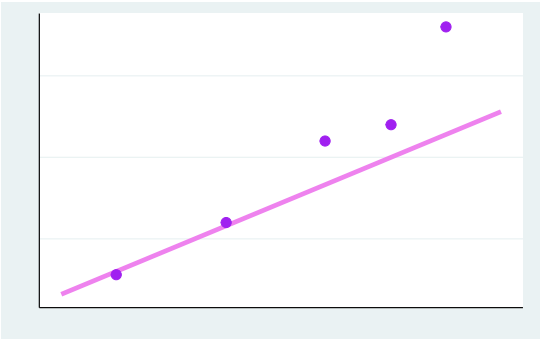
- Desde un punto de vista matemático, estas técnicas son idénticas.
- Ambos separan la varianza total de los datos en porciones diferentes y verifican la igualdad de estas "subvarianzas" en términos de un F-test.
- En ambos casos, la variable dependiente es continua. La diferencia radica en las variables independientes, las cuales pueden ser categóricas sólo en el ANOVA.
- La mayor diferencia que los distingue es la manera en que los datos son presentados.
- El ANOVA también puede usarse para evaluar un modelo creado por una regresión lineal.

Ejemplo*

*Base de datos `chickwts` de `R`

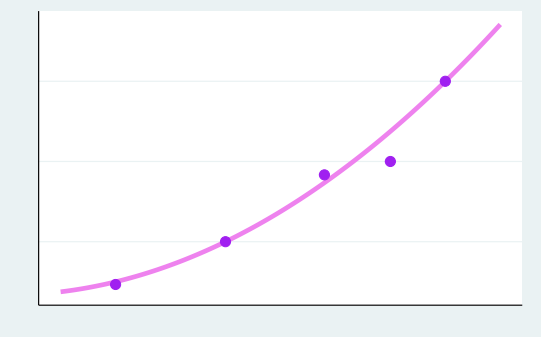
Tipos de ajuste

Underfitting



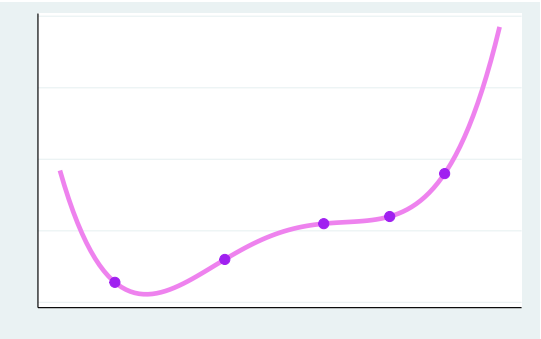
Modelo demasiado simple que no puede generalizar el conocimiento.

Balanceado



Modelo adecuado.

Overfitting



Modelo demasiado complejo con demasiado detalle del conocimiento.

Limitaciones de la regresión lineal

- Es muy sensible a la presencia de outliers.
- Es propenso al *underfitting*, proporcionando un ajuste insuficiente cuando un modelo no puede extraer adecuadamente la estructura subyacente de los datos.
- Si el modelo es demasiado complejo y depende de muchos parámetros o el conjunto de datos es muy pequeño, el modelo puede incurrir en un *overfitting*.
- En muchos ejemplos reales las variables no están linealmente relacionadas, por lo que el ajuste no será bueno.

Regresión lineal

Pregunta

¿Cómo podemos hacer predicciones usando un modelo de regresión lineal?

Coming Soon