Estadística (1937)
Unidad 7.
Regresión Lineal

Marcelo Ruiz

Departamento de Matemática
FCEFQyN de la UNRC

Diciembre de 2023

Regresión lineal simple

Definiciones básicas

Una variable predictora X y una respuesta Y y asumimos que[1]

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

donde ϵ, el error, tiene una distribución normal de media cero y varianza σ^2.

La pendiente β_1 y la ordenada al origen β_0 son desconocidos como así también la varianza del error σ^2. Si X es aleatoria, es independiente de ϵ.

A β_0 y β_1 les denominamos parámetros del modelo

[1] Seguimos los textos de James et al. (2021) y Maronna (2021)

Regresión lineal simple

Estimación de los parámetros por mínimos cuadrados

Consideremos observaciones del modelo

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Si $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ son las medias muestrales,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

minimizan la suma de cuadrados de los residuos

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

donde $e_i = y_i - \hat{y}_i$ es el i-ésimo residuo.

Regresión lineal simple

Estimación de los parámetros por mínimos cuadrados

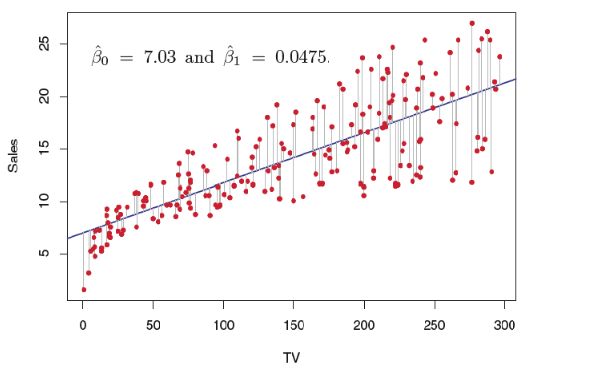


Figura 1: Datos “Advertising”. Ajuste por mínimos cuadrados. James et al. (2021)

Regresión lineal simple

Estimación de los parámetros por mínimos cuadrados

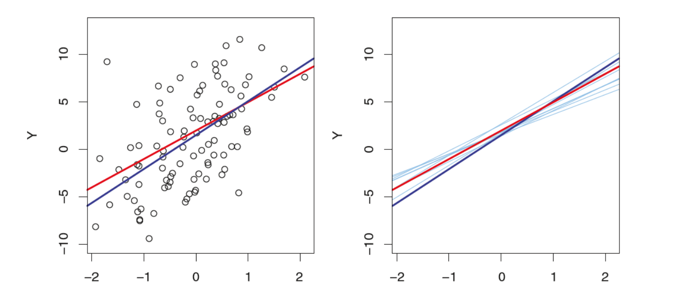


Figura 2: Datos simulados. A la izquierda: en azul la curva verdadera, en rojo la ajustada por OLS. A la derecha: en otros colores diferentes rectas incluida la ajustada por OLS. James et al. (2021)

Regresión lineal simple

Precisión de los estimadores

- $\mathrm{SE}\left(\hat{\beta}_1\right)$ y $\mathrm{SE}\left(\hat{\beta}_0\right)$ son insesgados.
- El error estándar (SE)de un estimador mide la variación del estimador.
  $\mathrm{SE}\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \mathrm{SE}\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right],$
  asumiendo que los ϵ_i son no-correlacionados con varianza común σ^2.
- A partir de ellos computamos intervalos de confianza (IC), que contendrán al verdadero valor del parámetro con cierto nivel. Un IC del 95 % de confianza es:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right).$$

Regresión lineal simple

Intervalos de confianza para los datos "Advertising"

- UN IC del 95 % para β_0 es [6.130,7.935] y para β_1, [0.042,0.053].
- Interpretación:
  - en ausencia de publicidad las ventas estarán en promedio en algún punto entre 6,130 y 7,940 unidades.
  - Y, por cada $\$1,000$ de incremento en publicidad televisiva, habrá en promedio un incremento en las ventas entre 42 y 53 unidades.

Regresión lineal simple

Prueba de hipótesis para la pendiente

- H_0 : $\beta_1 = 0$ o "No existe relación entre X e Y."
  versus
  H_A : $\beta_1 \neq 0$ ó "Y depende linealmente de X".
- El estadístico del test es

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}\left(\hat{\beta}_1\right)} \sim t_{n-2} \text{ si vale } H_0$$

- En R calculamos el valor p de la prueba (o probabilidad de error si rechazamos H_0).

Nota: la prueba de hipótesis para la ordenada al origen es similar.

Regresión lineal simple

Precisión global del modelo

- El error estándar residual

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

con $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ la suma de cuadrado residuales.

- R-cuadrado o fracción de la varianza explicada

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

con $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ la suma total de cuadrados.

- Se puede probar que para esta regresión lineal simple $R^2 = r^2$, donde r es el coeficiente de correlación estimado entre X e Y

Estudio de mercado

Los datos “Advertising.csv’: $n = 200$ con valor de venta e inversión en tres medios: tv, radio y diario impreso (James et al., 2021)
En R

```
rm(list = ls())
Advertising <- read.csv("Advertising.csv")
#descripción de los datos
names(Advertising)
[1] "X"          "TV"         "Radio"     "Newspaper" "Sales"
```

Estudio de mercado

Regresión lineal simple de Sales versus TV

```
attach (Advertising)
lm.fit =lm(Sales~TV)
Call:
lm(formula = Sales ~ TV)
Coefficients:
(Intercept)           TV
7.03259      0.04754
```

De donde

$$\widehat{y} = \underbrace{7.03259}_{\widehat{\beta}_0} + \underbrace{0.04754}_{\widehat{\beta}_1} x$$

Estudio de mercado

Pruebas de hipótesis y precisión global

```
> summary (lm.fit)

Call:
lm(formula = Sales ~ TV)

Residuals:
Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Estudio de mercado

De

	Coeficiente	Error estándar	estadístico-t	p-valor
Intercepto	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

concluimos que, con un valor-p inferior a 0.0001, Sales depende linealmente de la inversión publicitaria en TV y que el intercepto u ordenada al origen es no nula.

Estudio de mercado

La información de la tabla nos permite concluir que

Cantidad	Valor
Error estándar residual	3.26
R^2	0.612

- Los valores de venta reales en cada mercado se alejan de la verdadera recta de regresión en aproximadamente 3.26 unidades, en promedio[2].
- Dado que la media muestral de Sales es 14, el porcentaje de error es $3.26/14 = 23\,\%$
- R^2 =0.61 indica que menos de dos tercios de la variabilidad en las ventas es explicada por la regresión lineal sobre TV.

[2] El error estándar residual (RSE) es la cantidad promedio que la respuesta se aleja de la verdadera recta de regresión

Estudio de mercado

```
> predict (lm.fit , data.frame( TV=c(55 ,62 ,270) ),
 interval ="confidence")
fit        lwr       upr
1  9.647109  8.980040 10.31418
2  9.979865  9.339488 10.62024
3 19.867486 19.072436 20.66254
```

Nos provee de intervalos de predicción para tres valores de gastos publicitarios en TV

Cigüeñas y nacimientos

En Maronna (2021) se muestra la siguiente tabla que, para la ciudad alemana de Oldenburg, muestra los números de cigüenas (x) y de habitantes (en miles) (y) al final de cada año.

Año:	1930	1931	1932	1933	1934	1935	1936
Cigüeñas x :	130	148	175	185	247	263	255
Habitantes (miles) y :	55	65	63	66	68	72	75

Cuadro 1: Cigüeñas y habitantes

Cigüeñas y nacimientos

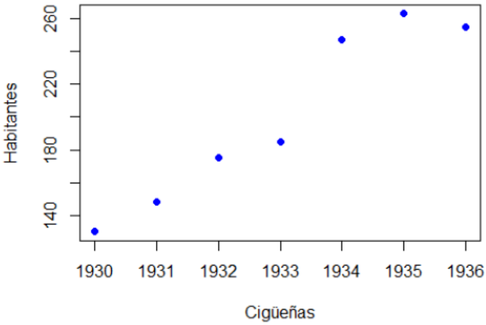


Figura 3: Cigüeñas y habitantes (Maronna, 2021)

Cigüeñas y nacimientos

- La figura muestra una relación lineal entre x e y.
- Si hacemos un ajuste lineal para predecir el número de habitantes en función del de cigüeñas, se obtienen

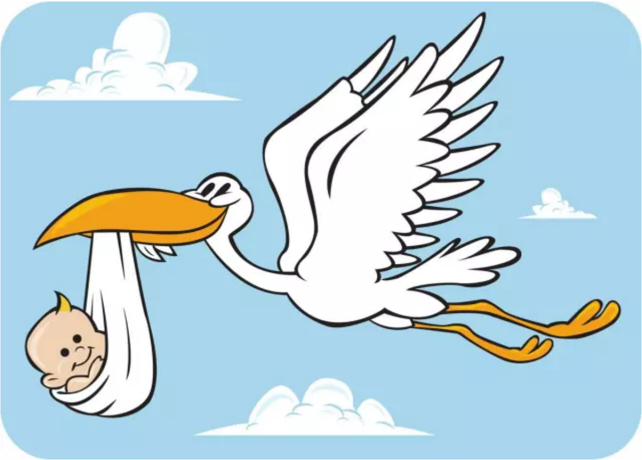
$$\hat{\beta}_0 = 36.9 \text{ y } \hat{\beta}_1 = 0.14$$

con una correlación alta:

$$\hat{\rho} = 0.81.$$

Cigüeñas y nacimientos

Las niñas y niños son traídos por cigüeñas [3]



[3] tomadodehttps://priceonomics.com/do-storks-deliver-babies/

¿Causalidad?

- Dado que el modelo ajustado es

$$y = 36.9 + 0.14x$$

  ¿se puede concluir que la importación de 10 cigüeñas implicaria un aumento medio de la población de 1400 habitantes?
- "La idea parece absurda, máxime que nosotros ya sabemos que dichas aves nada tienen que ver con el nacimiento de los niños (puesto que éstos nacen de un repollo)"
- ¿Cómo se explica entonces la alta correlación lineal? Es que las x como las y aumentan con el tiempo, y eso es simplemente la causa.
- "O sea: si dos variables estan muy correlacionadas, la causa puede ser una tercena variable que influye en ambas".

Cigüeñas y nacimientos

¡Texto recomendado!