Regresión Avanzada

Preliminares

Regresión Lineal Simple

Preliminares

La estadística busca extraer patrones y tendencias de datos que nos permitan comprender los fenómenos que hay detrás de ellos.

Aprendizaje

Supervisado busca predecir un el valor de una medida en base a información previa. Variable de respuesta.

No Supervisado no hay variable de respuesta. El objetivo es describir asociaciones y patrones entre datos.

Preliminares

La estadística busca extraer patrones y tendencias de datos que nos permitan comprender los fenómenos que hay detrás de ellos.

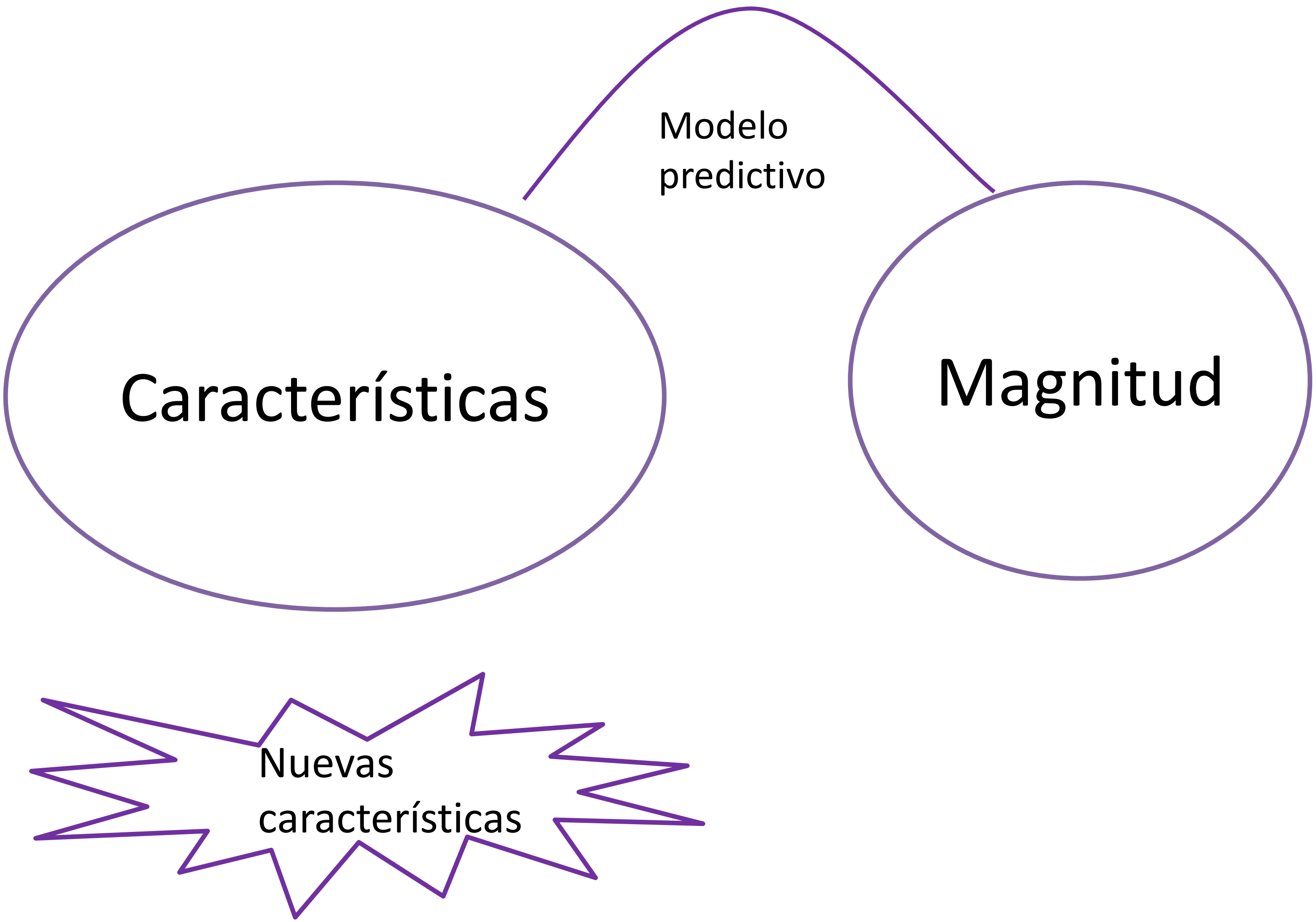
- Aprendizaje
  - Supervisado
    - Regresión
    - Clasificación
  - No Supervisado
    - Cluster

Preliminares

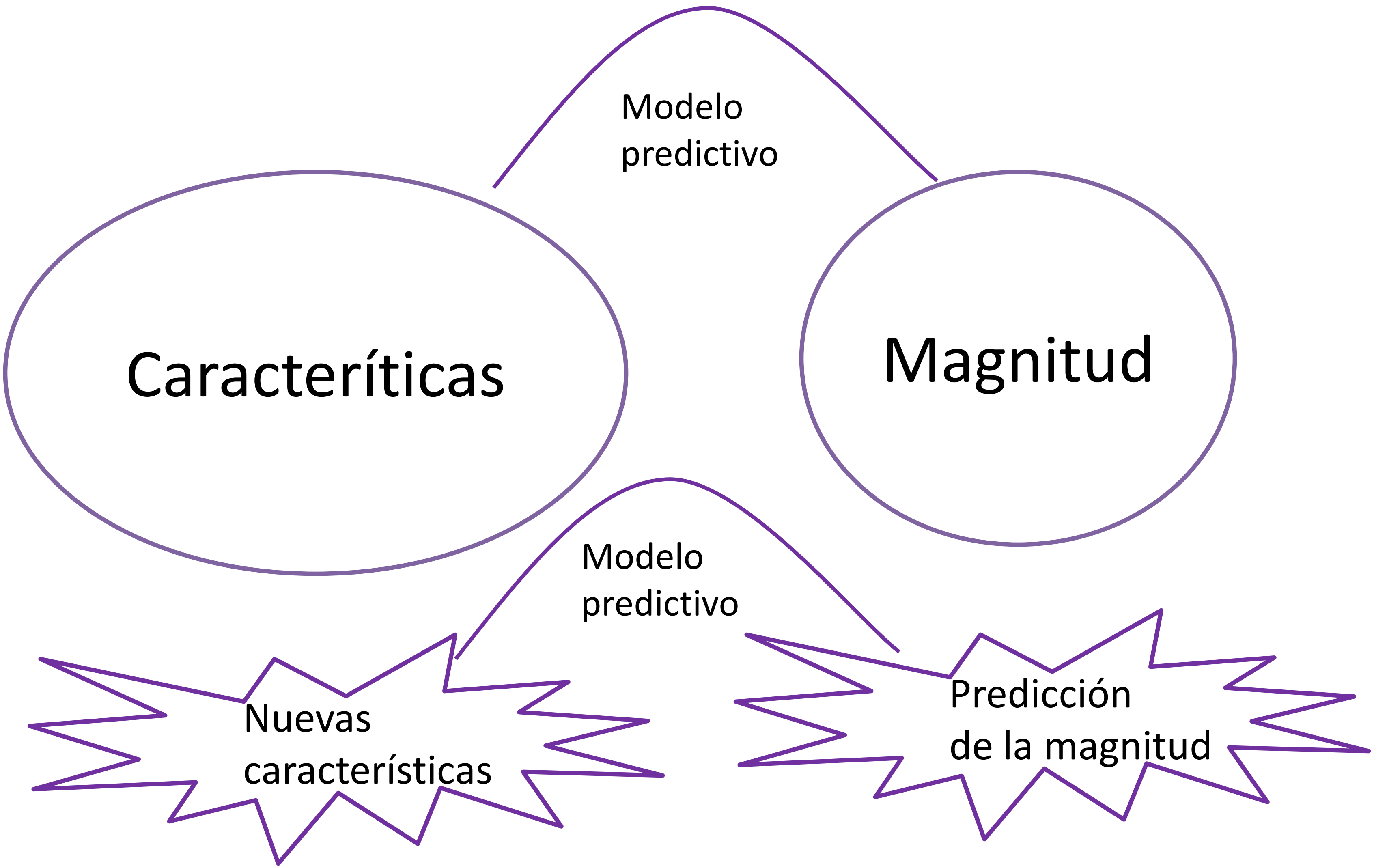
La estadística busca extraer patrones y tendencias de datos que nos permitan comprender los fenómenos que hay detrás de ellos.

- Aprendizaje
  - Supervisado
    - Regresión
    - Clasificación
  - No Supervisado
    - Cluster

Preliminares



Preliminares



Preliminares

Ejemplo

Y= ventas (en miles de unidades)

X= inversión publicitaria

TV

Radio

Diario

Preliminares

Ejemplo

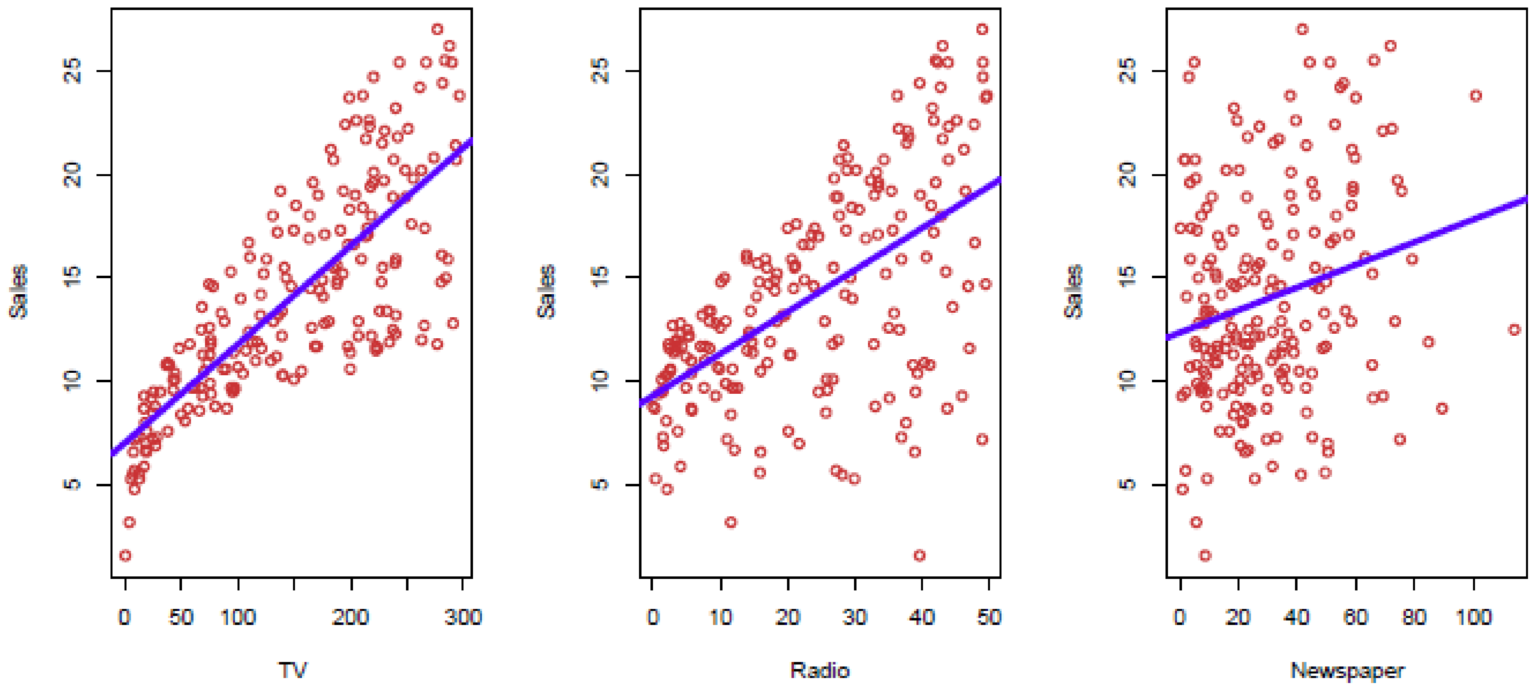
Y= ventas (en miles de unidades)

X_1= inversión publicitaria en televisión

X_2= inversión publicitaria en radio

X_3= inversión publicitaria en diario

Preliminares



"Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with a pplications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

Preliminares

Medidas resúmenes de los datos

	TV	Radio	Diario	Ventas
Media	147.04	23.64	30.55	14.02
Mediana	149.75	22.9	25.75	12.90
1er cuartil	74.38	9.975	12.75	10.38
3er cuartil	218.82	35.52	45.10	17.4

Preliminares

Variables regresoras

$X_1,\dots, X_p$

Variables independientes

Features (características)

$\xrightarrow{f}$

Variables de respuesta

Y

Variables dependientes

respuestas

En general, suponemos que observamos una cantidad Y (variable de respuesta) y p variables predictoras $X_1, X_2,\dots, X_p$. Asumimos que hay una relación entre Y y $X=(X_1, X_2,\dots, X_p)$ que de manera general se puede escribir como

$$Y=f(X)+e,$$

e es el término de error y f es la relación funcional entre X e Y que representa la Información sistemática que da X sobre Y.

Preliminares

Buscaremos estimar f.
Predicción: en muchas ocasiones se conoce el valor de X, pero el de Y es difícil de conocer. En esos casos se busca estimar Y a partir de X, teniendo $\widehat{Y}$,

$$\widehat{Y} = \widehat{f}(X),$$

donde $\widehat{f}$ es el estimador de f e $\widehat{Y}$ es la predicción de Y. La precisión de Y depende de dos fuentes de error el *error reducible* y el *irreducible*.
$\widehat{f}$ no va a ser una estimación perfecta de f, esta fuente de error es reducible, potencialmente se puede mejorar la estimación de f. Si la f fuese estimada sin error de todos modos $\widehat{Y} = f(X)$ tendría error ya que en el modelo Y también depende de e.

Preliminares

Asumiendo $\widehat{f}$ y X fija calculamos el error cuadrático medio entre Y y su predicción $\widehat{Y}$.

$$\begin{aligned} E(Y-\widehat{Y})^2 &= E(f(X)+e-\widehat{f}(X))^2 \\ &= E(f(X)-\widehat{f}(X))^2 + E(e^2) \\ &= \underbrace{(f(X)-\widehat{f}(X))^2}_{\text{reducible}} + \underbrace{var(e)}_{\text{irreducible}} \end{aligned}$$

El objetivo es reducir el error reducible.

Preliminares

Inferencia: en muchas ocasiones queremos encontrar el modo en que $X_1, \ldots, X_p$ afectan a Y. Es decir queremos comprender como los cambios en las variables regresoras afectan a la variable de respuesta. En esos casos necesitamos conocer explícitamente la forma funcional de Y. Queremos responder las siguientes preguntar:

- Cuáles son los predictores asociados a la variable de respuesta?
- Cuál es la relación entre cada variable predictora y la variable de respuesta?
- Se puede resumir la relacion entre cada variable predictora y la variable de respuesta o es necesaria una relación funcional más compleja?

Preliminares

En muchos casos los problemas que se nos presentan son de predicción, en otros de inferencia y en otros ambos.

Supongamos que queremos determinar el valor de una propiedad. Los factores que sin duda influyen, zona, tamaño, antigüedad, calidad del aire, distancia a colegio, nivel socioeconómico del barrio, etc.

Ej de problema de predicción: *determinar entre qué valores está el precio.*

Ej de problema de inferencia: *determinar cuánto aumenta el valor de una propiedad por tener pileta.*

Preliminares

El método que utilicemos para estimar *f* depende del propósito del problema.

- En problemas de predicción se hace foco en la precisión.
- En problemas de inferencia se hace foco en la simpleza e interpretación

Modelos lineales → Simples

→ Fáciles de interpretar

→ No muy precisos

A lo largo del curso veremos modelos lineales sencillos y fáciles de interpretar y otros no lineales y complejos pero que tienen mejores propiedades de a la hora de predecir.
Supongamos que tenemos un conjunto de n observaciones $(x_{i,1}, \ldots, x_{i,(p-1)}, y_i)$ con $i = 1, \ldots, n$. Donde $x_{i,j}$ representa el j-ésimo regresor correspondiente a la observación i-ésima, donde $i = 1, \ldots, n$ y $j = 1, \ldots, p-1$. y_i es la respuesta correspondiente a la i-ésima observación $i = 1, \ldots, n$. En forma resumida podemos escribir $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, donde $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p-1})^t$ donde ^{t} denota el vector traspuesto. Nuestro objetivo es estimar f, a partir de las observaciones que tenemos de forma tal que $Y \approx \widehat{f}(X)$ para cualquier observación dada (X, Y).

Moldelos paramétricos

- Se hacen supuestos sobre la forma funcional de f, por ejemplo una función lineal,

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}.$$

Para estimar esta función de p dimensiones alcanza con estimar p parámetros $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

- luego de seleccionar el modelo hay que *ajustarlo* con la muestra de entrenamiento, de formar tal que

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}.$$

Vent: se asume una forma sencilla para f se reduce el problema multidimensional a unas pocas variables.

Desvent: la forma funcional elegida puede no ajustarse bien a f.

Moldelos no paramétricos

- No se hacen supuestos sobre la forma funcional de f.
- Se busca estimar f de forma tal de tener el mejor ajuste posible sin sobresuavizar demasiado ni tener una función muy rugosa.

Vent: potencialmente tienen mejor poder predictivo porque el conjunto de posibles f es mucho mayor, más ductilidad.

Desvent: no reducen el problema de estimar f a un problema de dimensión baja, luego el tamaño muestral requerido es mucho más grande que en el caso paramétrico.

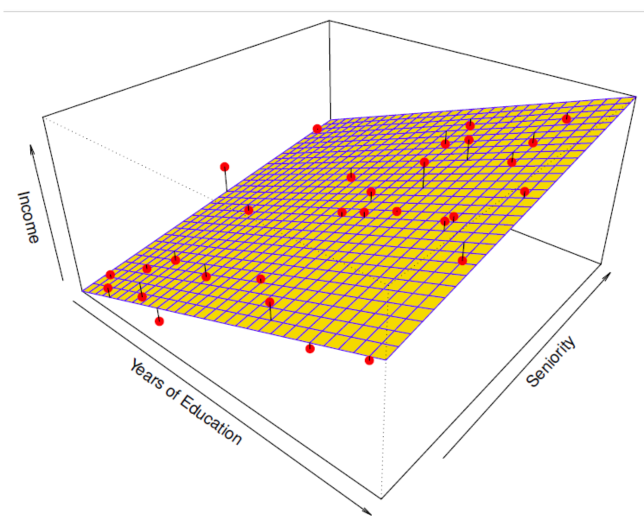
Preliminares

Consideremos el ejemplo, queremos estimar el *ingreso* (*income*) conociendo el *nivel educacional* (*education*) y la *antigüedad* de una persona (*seniority*)

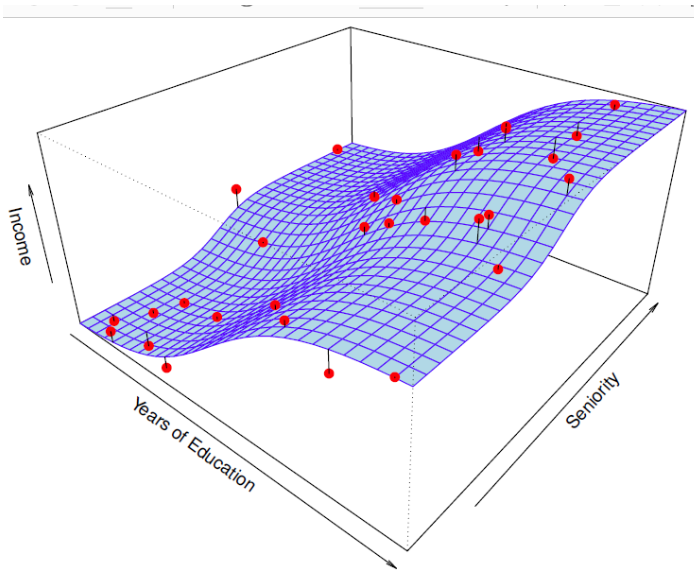
Modelo Paramétrico

income≈

$\beta_0+\beta_1$ *education*$+\beta_2$ *seniority*



Modelo No Paramétrico



Preliminares

Hay un compromiso entre el poder predictivo de un modelo y la capacidad de interpretación del mismo.

Modelos más flexibles ganan en precisión pero pierden en interpretación.

Modelos menos flexibles son más sencillos de interpretar pero pierden precisión.

Regresión lineal Simple

Y= variable de respuesta, variable dependiente.
Variable continua.

X= variable regresora o explicativa, independiente.
Variable continua.

El modelo

$$\boldsymbol{Y = \beta_0 + \beta_1 X + e}$$

Regresión lineal simple

Decimos que un modelo es lineal cuando es lineal en los parámetros:

- $y = \beta_0 + \beta_1 X$ es lineal ya que es lineal en los parámetros.
- $y = \beta_0 X^{\beta_1}$ es lineal ya que se puede escribir como $\log(y) = \log(\beta_0) + \beta_1 \log(X)$, aunque no es lineal como función de (X, y).
- $y = \beta_0 + \beta_1 X + \beta_2 X^2$ es lineal en los parámetros pero no como función de X.
- $y = \beta_0 + \frac{\beta_1}{\beta_1 + X}$ es no lineal en los parámetros y en las variables.
- $y = \beta_0 + \beta_1 X^{\beta_2}$ es no lineal en los parámetros y en las variables.

Regresión lineal Simple

- $\boldsymbol{Y = \beta_0 + \beta_1 X}$ es una **recta de regresión**.
- El parámetro $\boldsymbol{\beta_1}$es la **pendiente** de la recta. Indica la variación media de la variable de respuesta cuando **X** aumenta una unidad.
- El parámetro $\boldsymbol{\beta_0}$ es el **intercept**, el término independiente de la recta. Indica el valor medio de **Y** cuando **X**=0.
- ***e*** es el término de error. Es una variable aleatoria, con media cero, i.e. **E(*e*)=0**.

Regresión lineal Simple

Si los errores son pequeños,

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X}$$

Objetivos $\longrightarrow$ estimar los parámetros $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$

Cómo ? $\longrightarrow$ a través de una muestra.

Sean $(x_1, y_1), \ldots, (x_n, y_n)$ observaciones independientes de las variables $\mathbf{X}$ e $\mathbf{Y}$.

Regresión lineal simple

Ejemplo

X=presupuesto publicitario en TV.

Y= ventas (en miles de unidades).

$$\textbf{ventas} \approx \boldsymbol{\beta_0} + \boldsymbol{\beta_1 TV}$$

A partir de datos provenientes de ***n=200*** mercados $(x_1, y_1), \dots, (x_{200}, y_{200})$ buscamos estimar $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$.

Regresión lineal simple

En este caso el modelo sería

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + e_i},$$

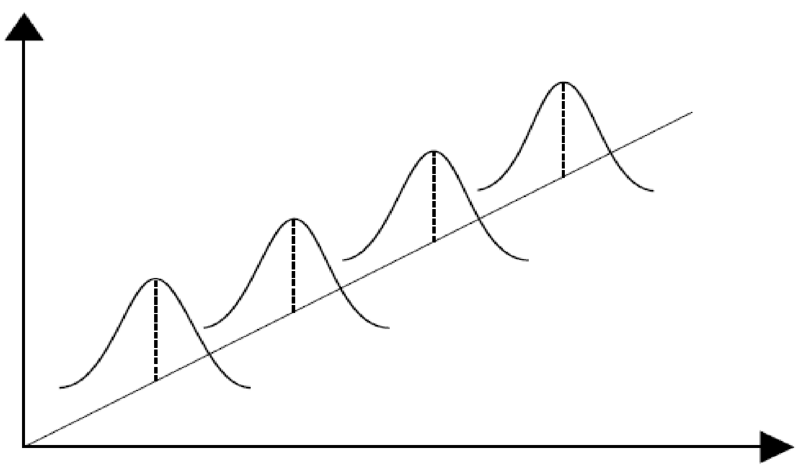
para **i=1,…, 200.**

Supuestos:

- Los errores son independientes entre si e independientes de las x's.
- Los errores tienen media cero, $\mathbf{E}(\boldsymbol{e_i})$ y tienen todos la misma varianza, que llamaremos $\boldsymbol{\sigma^2}$.
- Los errores tienen distribución normal.

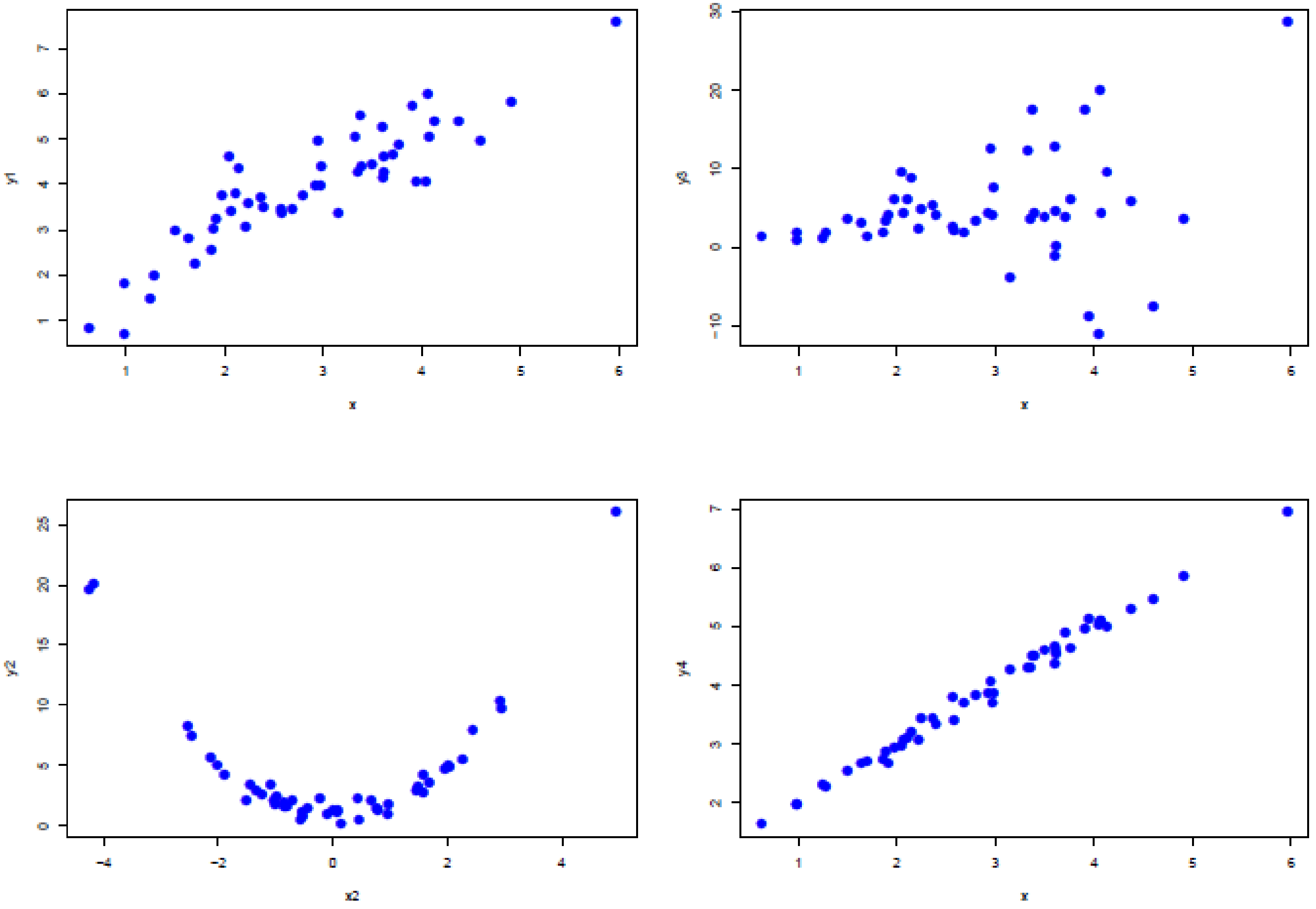
Regresión lineal simple

Qué quiere decir que los errores tengan distribución normal?



Regresión lineal simple

Dónde ajusta bien el modelo?



Regresión lineal Simple

- $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ es el estimador del intercept $\boldsymbol{\beta}_{\mathbf{0}}$.
- $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ es el estimador de la pendiente $\boldsymbol{\beta}_{\mathbf{1}}$.

- Si tuvieramos el valor de la inversión publicitaria **X=x**, podríamos predecir el valor de las ventas.

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\beta}}_{\mathbf{0}} + \hat{\boldsymbol{\beta}}_{\mathbf{1}}\mathbf{x},$$

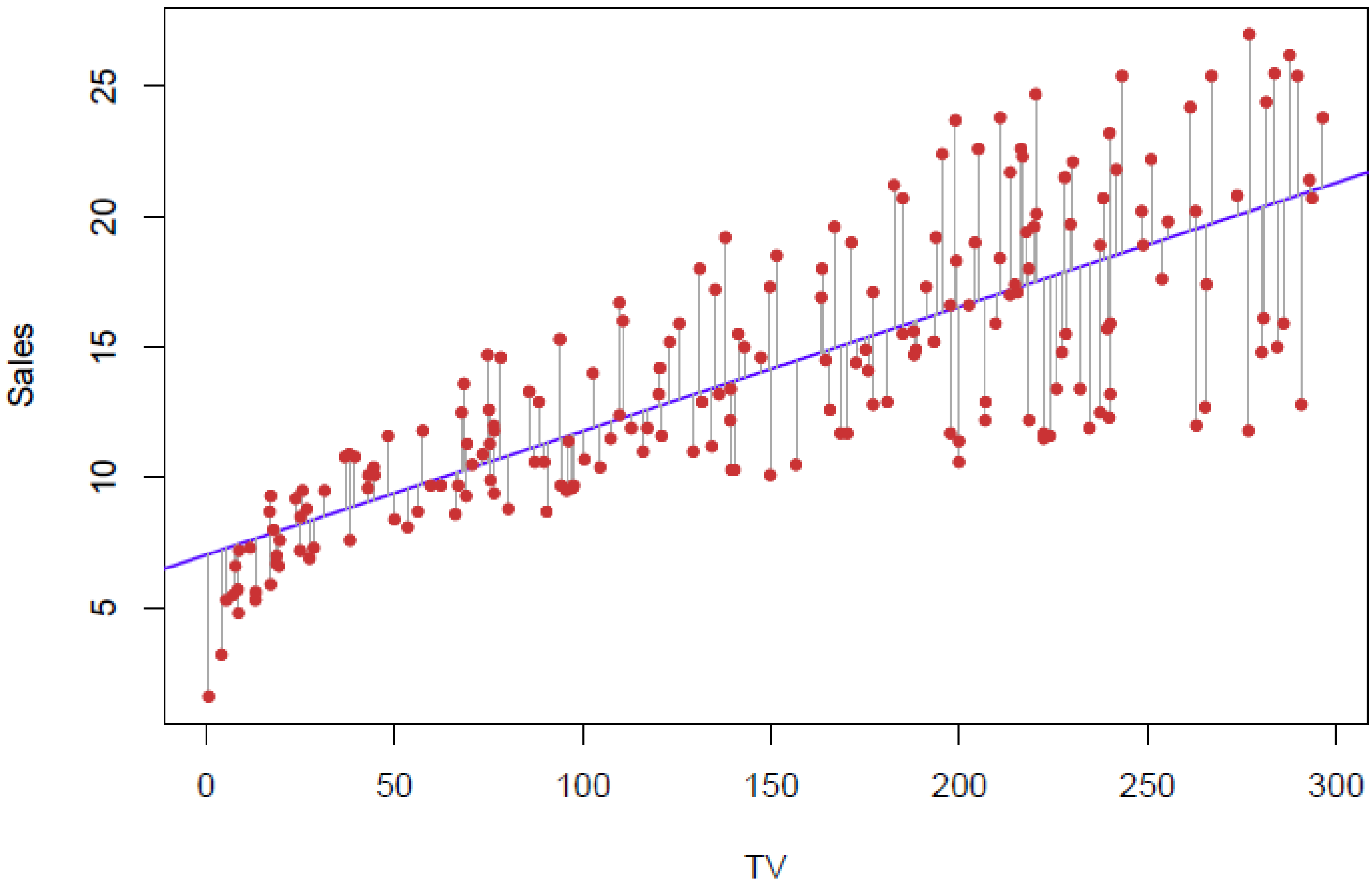
$\hat{\boldsymbol{y}}$ es el valor de ventas predicho para **x.**

Regresión lineal Simple

- $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$ no va a coincidir con y_i.
- Para cada observación tenemos el error de predicción $r_i = y_i - \widehat{y}_i$, se denomina **i-ésimo residuo**.
- Definimos la **Suma de cuadrados de los residuos** (RSS) como,

$$RSS = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Regresión lineal simple



El estimador de mínimos cuadrados minimiza la suma de los cuadrados de los residuos, es decir,

$$(\widehat{\beta_0}, \widehat{\beta_1}) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Derivando respecto de β_0 y β_1 respectivamente se obtienen las **ecuaciones normales**

$$\begin{aligned}
\frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\
\frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i.
\end{aligned}$$

Pendiente

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} = \widehat{\rho}_{xy}\frac{s_y}{s_x}.$$

Intercept

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}.$$

Los valores predichos o ajustados están dados por

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

```
adv.tv.lm=lm(sales~TV)
summary(adv.tv.lm)
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
```

Estimador del intercept

Estimador de la pendiente

En nuestro caso

$$\begin{aligned}\widehat{\beta}_0 &= 7.03\\ \widehat{\beta}_1 &= 0.0473.\end{aligned}$$

Es decir,

$$ventas \approx 7.03 + 0.0473tv.$$

Cada \$1000 que se invierten en publicidad se venden en promedio 47.3 unidades más del producto.
Si se invierten \$100 se espera vender

$$\widehat{ventas} = 7.03 + 0.0473 \times 100 = 11.76 \text{ unidades.}$$

Regresión Lineal Simple

$$\boldsymbol{E(Y|X) = \beta_0 + \beta_1 X}$$

En particular, si **X=0**, $\boldsymbol{E(Y|X = 0) = \beta_0}$. Es decir, que el intercept es la venta esperada sin inversión publicitaria.

La pendiente $\boldsymbol{\beta_1}$ indica el incremento medio de **Y** por cada incremento unitario de **X**.

El error captura todo aquello que no es capturado ni por **X** ni por la relación lineal que estamos imponiendo.

Regresión lineal simple

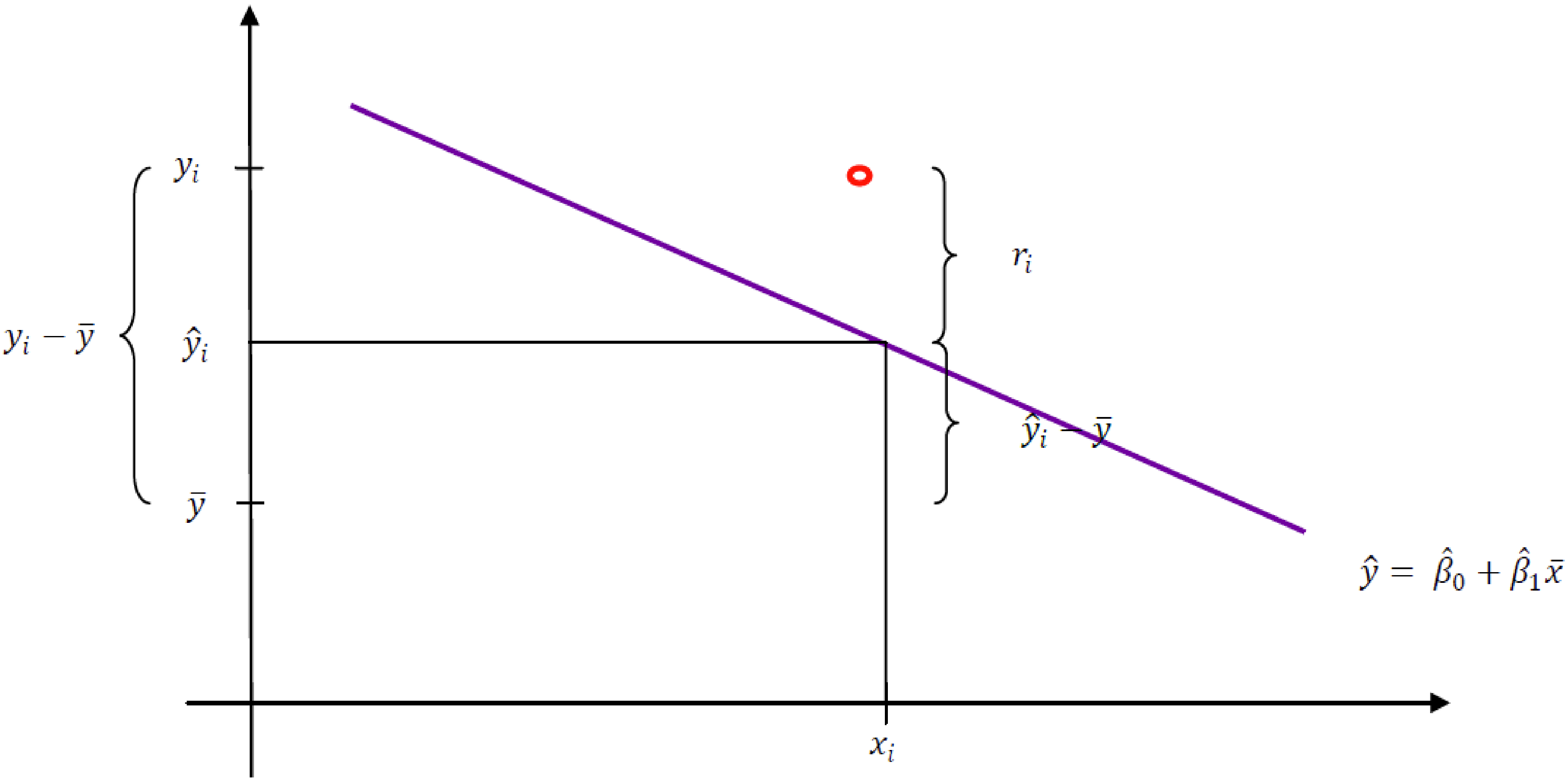
- La recta de mínimos cuadrados pasa por el punto $(\bar{x}, \bar{y})$.
- La suma de los residuos siempre vale 0.
- La recta para predecir Y en función de X no es la misma que para predecir X en función de Y.
- Si la variable regresora se incrementa en un desvío estándar s_x entonces la variable de respuesta se incrementa en $\hat{\rho}_{xy}$ desvíos estándares, $\hat{\rho}_{xy}\, s_Y$.

Regresión lineal simple

$$\begin{aligned}
y_i &= \widehat{y}_i + r_i \\
y_i - \overline{y} &= \widehat{y}_i - \overline{y} + r_i \\
\sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} r_i^2. \\
TSS &= ESS + RSS,
\end{aligned}$$

TSS mide la variabilidad total (tiene $n-1$ grados de libertad).
ESS mide la variabilidad explicada por el modelo (tiene 1 grados de libertad).
RSS mide la variabilidad no explicada o residual (tiene $n-2$ grados de libertad).

Regresión lineal simple



Regresión lineal simple

Tabla de anova

Fuente de variación	SS	gl	Cuad. medios	Estad.
Explicada (ESS)	$\sum_{i=n}^{n}(\widehat{y}_i - \overline{y})^2$	1	$\sum_{i=n}^{n}(\widehat{y}_i - \overline{y})^2$	F
Residual (RSS)	$\sum_{i=n}^{n} r_i^2$	$n-2$	$S_R^2 = \frac{\sum_{i=n}^{n} r_i^2}{n-2}$	
Total (TSS)	$\sum_{i=n}^{n}(y_i - \overline{y})^2$	$n-1$		

El estadístico $F = \frac{ESS}{s_R^2}$.

Si F es suficientemente grande (la variabilidad explicada es muy grande respecto a la no explicada), se debe rechazar H_0 : $\beta_1 = 0$.

Bajo $H_0 : \beta_1 = 0$, el estadístico F tiene distribución $f_{1,n-2}$. La región crítica de nivel α para el test de hipótesis es:

$$R = \{F > f_{1,n-2;1-\alpha}\}$$

Regresión lineal simple

Supongamos que $\sigma = 1, \beta_0 = 0 \ y \ \beta_1 = 1,$ es decir que nuestro modelo es

$$y_i = x_i + e_i$$

Donde los errores e_i son independientes con distribución normal estándar.

Fijemos x_i=1,2,…,10 (n=10) y generemos las respuestas correspondientes de acuerdo con este modelo.

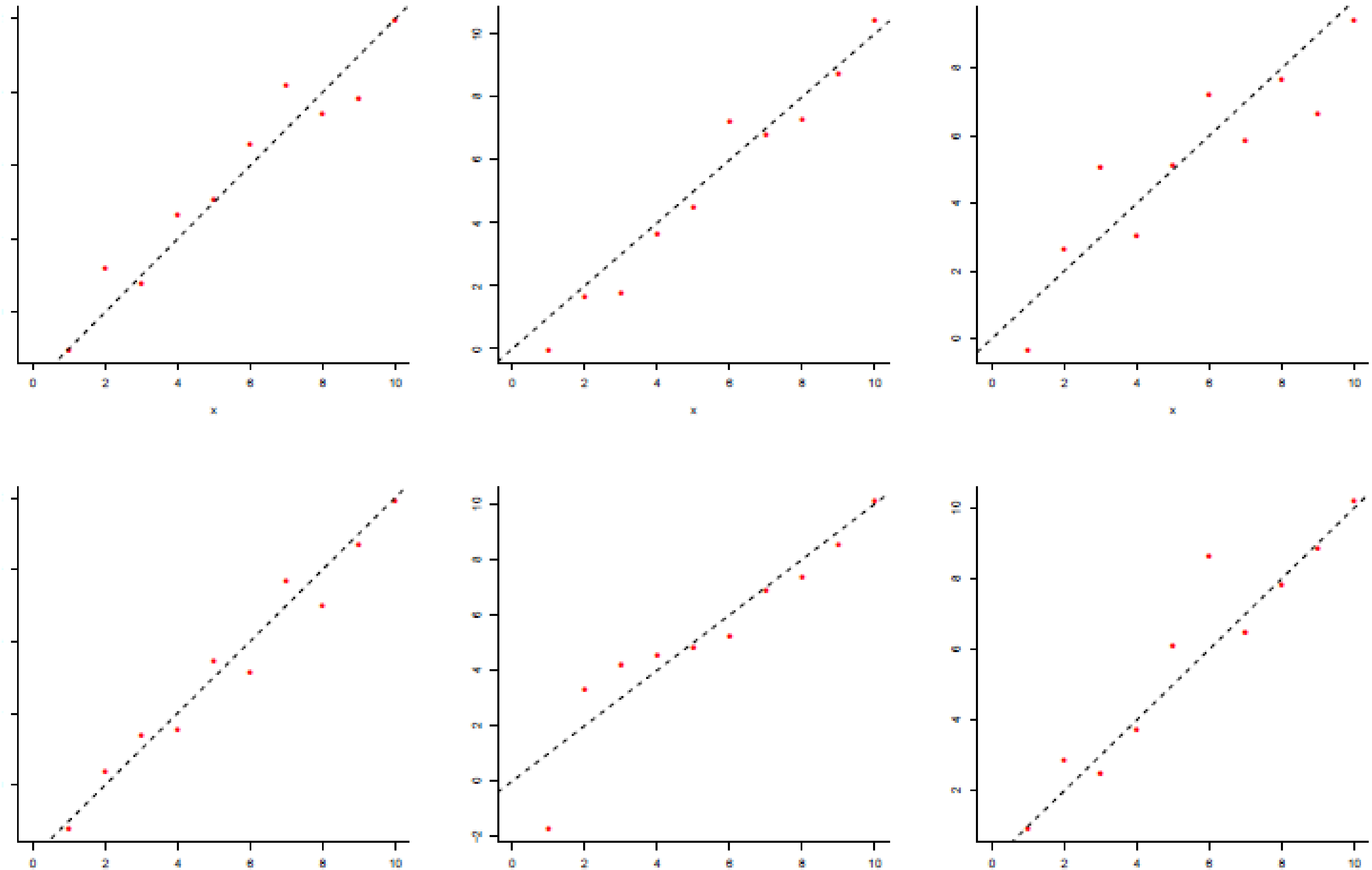
Regresión lineal simple

Posteriormente calculamos la recta de mínimos cuadrados y la representamos junto a la verdadera

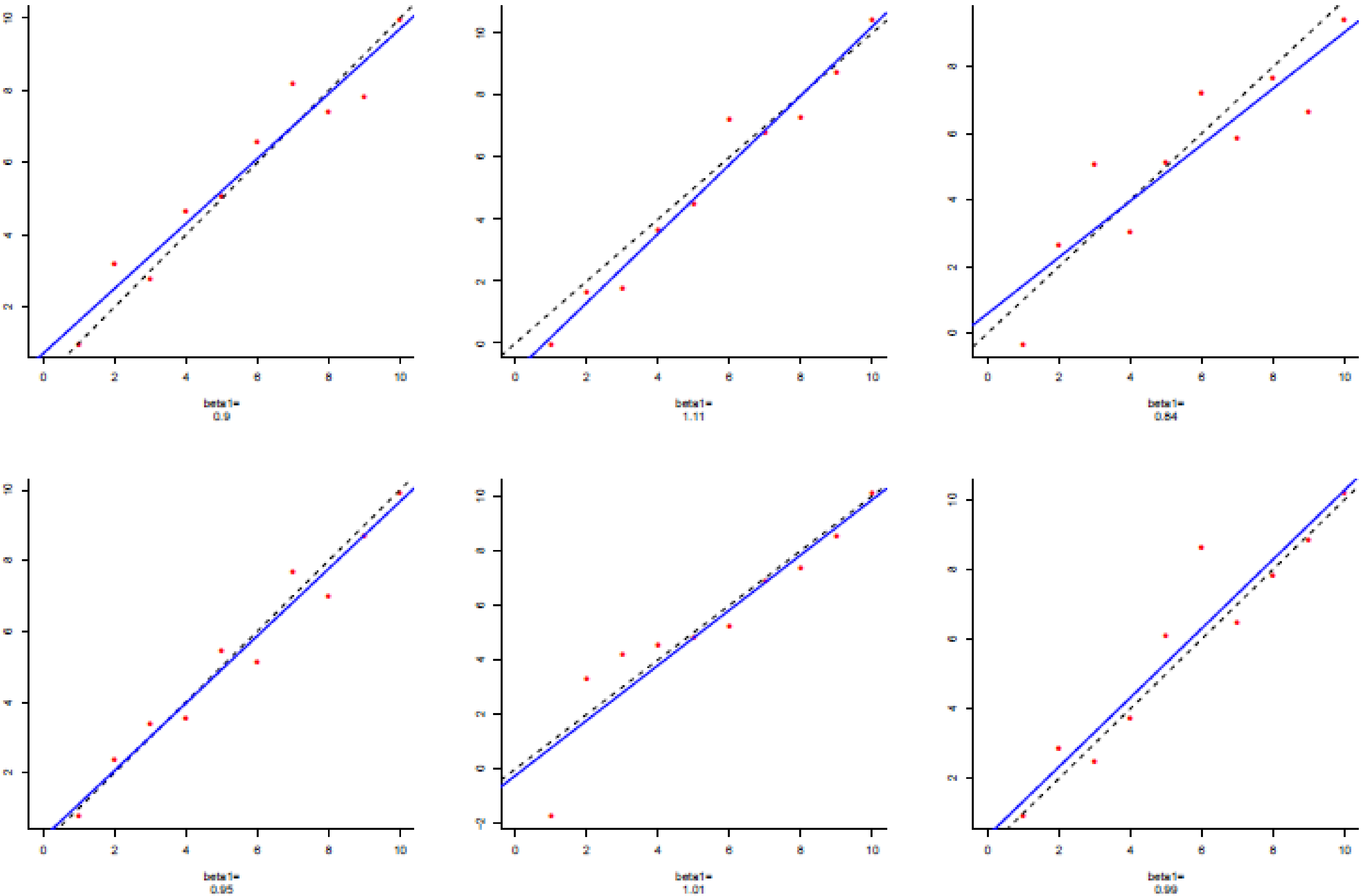
$$Y=X.$$

Repetimos este experimento seis veces.

Regresión lineal simple

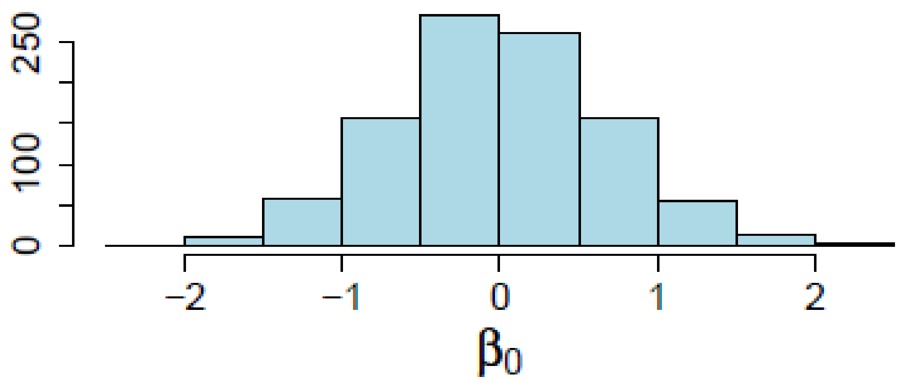


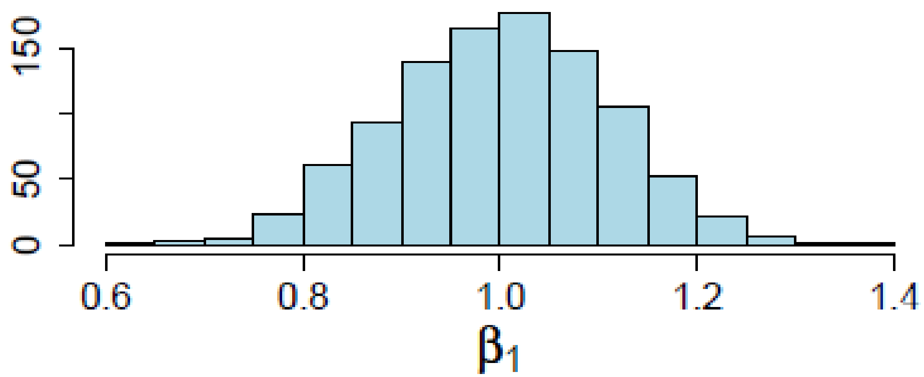
Regresión lineal simple



Regresión lineal simple

•





Realizamos este experimento 1000 veces, y graficamos los histogramas

Regresión lineal simple

- Los estimadores son insesgados y consistentes.
- Los estimadores siguen una distribución aproximadamente normal.
- Se puede medir su variabilidad y a partir de allí construir
  - Intervalos de confianza.
  - Test de Hipótesis.

Regresión lineal simple

Error estándar para el intercept.

$$var(\hat{\beta}_0) =$$

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

Donde $\boldsymbol{\sigma^2}$ es la varianza de los errores, hay que estimarla, más adelante veremos como hacerlo.

Si $\bar{x}^2$ es grande se estima con menos precisión el término independiente.

Regresión lineal simple

Error estándar para la pendiente

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

$$= \frac{\sigma^2}{(n-1){s_X}^2}$$

Al aumentar $(n-1){s_X}^2$ el error estándar disminuye.

Si se puede, conviene diseñar el experimento de manera tal que las x's tengan dispersión grande.

Como siempre aumentar el tamaño muestral da lugar

Regresión lineal simple

La varianza residual es un estimador insesgado de σ^2,

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2}$$

Se pierden dos grados de libertad ya que

- la media de los residuos es igual a cero.
- la covarianza entre los residuos y la variable regresora es cero.

Regresión lineal simple

Cómo estimar σ^2?

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

A $\hat{\sigma}$ lo llamaremos *Error Estándar de los Residuos (RSE)*.

Por lo tanto, el estimador del error estándar del intercept está dado por $\widehat{SE}\left(\hat{\beta}_0\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1){s_X}^2}}$

Regresión lineal simple

El estimador del error estándar de la pendiente está dado por

$$\widehat{SE}^2(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{(n-1){s_X}^2}.$$

Regresión lineal simple

A continuación podemos construir intervalos de confianza para los dos parámetros.

El intervalo de confianza de nivel $1-\alpha$ para β_0 es

$$\widehat{\beta}_0 \pm t_{n-2,1-\alpha/2}\widehat{SE}(\widehat{\beta}_0).$$

El intervalo de confianza de nivel $1-\alpha$ para β_1 es

$$\widehat{\beta}_1 \pm t_{n-2,1-\alpha/2}\widehat{SE}(\widehat{\beta}_1).$$

Regresión lineal simple

En el caso de la publicidad el intervalo de confianza de nivel 95% para el intercept será [6.13, 7.935].
Para calcular esto necesitamos encontrar el percentil $t_{198,0.975}$, que coincide con el mismo percentil que el de la normal estándar, 1.96.
Además $\hat{\sigma}^2, s_X$ y $\bar{x}$.

Regresión lineal simple

¿Cómo se interpreta?
Si la inversión publicitaria desaparece las ventas caerán hasta un nivel entre 6.13 y 7.935 (en miles de unidades)

Regresión lineal simple

El intervalo de confianza de nivel 95% para la pendiente está dado es [0.042, 0.053].

Para calcularlo tuvimos que hallar el mismo percentil que en el caso anterior y necesitamos conocer $\hat{\sigma}^2 y\ s_X$.

Cómo se interpreta?

Cada $1000 de inversión publicitaria en TV, las ventas aumenta en promedio entre 43 y 53 unidades.

Regresión lineal simple

A partir de los intervalos de confianza podemos construir test de hipótesis

$$H_0: \beta_1 = 0 \quad vs \quad H_A: \beta_1 \neq 0.$$

$$H_0: Hay\ relación\ lineal\ entre\ X e\ Y$$

$$vs$$

$$H_A: No\ hay\ relación\ lineal\ entre\ XeY.$$

Si $\beta_1 = 0$, el modelo queda $Y = \beta_0 + e$.

Regresión lineal simple

El estadístico para este test es

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}(\widehat{\beta}_1)} \sim t_{n-2}.$$

tiene distribución t_{n-2} bajo H_0.
Rechazamos H_0 a nivel α si $|t| > t_{n-2,1-\alpha/2}$.
El p-valor correspondiente es $p - valor = 2P(|t| > t_{obs})$.

Regresión lineal simple

- $$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 > 0$$

  Rechazo H_0 si $t > t_{n-2,1-\alpha}$.

- $$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 < 0$$

  Rechazo H_0 si $t < -t_{n-2,1-\alpha}$.

```
adv.tv.lm=lm(sales~TV)
summary(adv.tv.lm)
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
```

Desvío estándar estimado del intercept

Desvío estándar estimado de la pendiente

```
adv.tv.lm=lm(sales~TV)
summary(adv.tv.lm)
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
```

Estadístico del test donde se testea si el intercept es distinto de cero.

P-valor correspondiente

P-valor correspondiente

Estadístico del test donde se testea si la pendiente es distinta de cero.

Regresión lineal simple

Luego de haber rechazado las hipótesis nulas, sabiendo que intercept y pendiente son distintas de cero. Se quiere saber si **¿el modelo ajusta bien a los datos?**

- Error estándar residual (RSE)
- Estadístico R^2.
- Estadístico F.

Regresión lineal simple

Bondad de ajuste de los datos:
El error estándar residual (RSE), es una estimación de la desviación estándar de los errores (ϵ).

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

En nuestro caso el error estándar es $RSE = 3.26$. Interpretación: en cada mercado las ventas se desvían de la recta de regresión en aproximadamente 3260 unidades (porque las ventas están medidas en miles de unidades).
Aún conociendo el verdadero valor de β_0 y β_1 las ventas se desviarán del verdadero valor en aproximadamente 3260 unidades.

Regresión lineal simple

Es aceptable?
En nuestro caso la venta media es de \$14000 unidades, luego el porcentaje del error será de 23%

$$\frac{3260}{14000} \approx 0.23.$$

RSE, es una medida del desajuste del modelo, es bueno si es pequeño.

Regresión lineal simple

El estadístico R^2.

- Es una medida de bondad de ajuste del modelo.
- El coeficiente de correlación al cuadrado r entre las variables x e y .
- El coeficiente de determinación:

$$R^2 = \frac{\text{Variabilidad explicada}}{\text{Variabilidad total}} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

  En el modelo de regresión simple $R^2 = \widehat{\rho}^2_{xy}$, el coeficiente de determinación coincide con el coeficiente de correlación al cuadrado.
- No depende de las unidades de la variable de respuesta, ya que es una proporción.

Regresión lineal simple

Recordemos que **ESS** mide la variabilidad explicada luego de haber ajustado el modelo.

Por ente TSS-RSS, mide la varianza en la variable de respuesta, Y, que es removida al aplicar el modelo lineal.

R^2 mide la proporción de variabilidad de Y que puede ser explicada utilizando linealmente la X.

Valores cercanos a 1 indican que el ajuste es bueno.

Regresión lineal simple

En nuestro ejemplo R^2=0.61.

El 61% de la variabilidad en las ventas se pudo explicar vía la inversión publicitaria en TV.

adv.tv.lm=lm(sales~TV)
summary(adv.tv.lm)
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

RSE

Residual standard error: **3.259** on 198 degrees of freedom

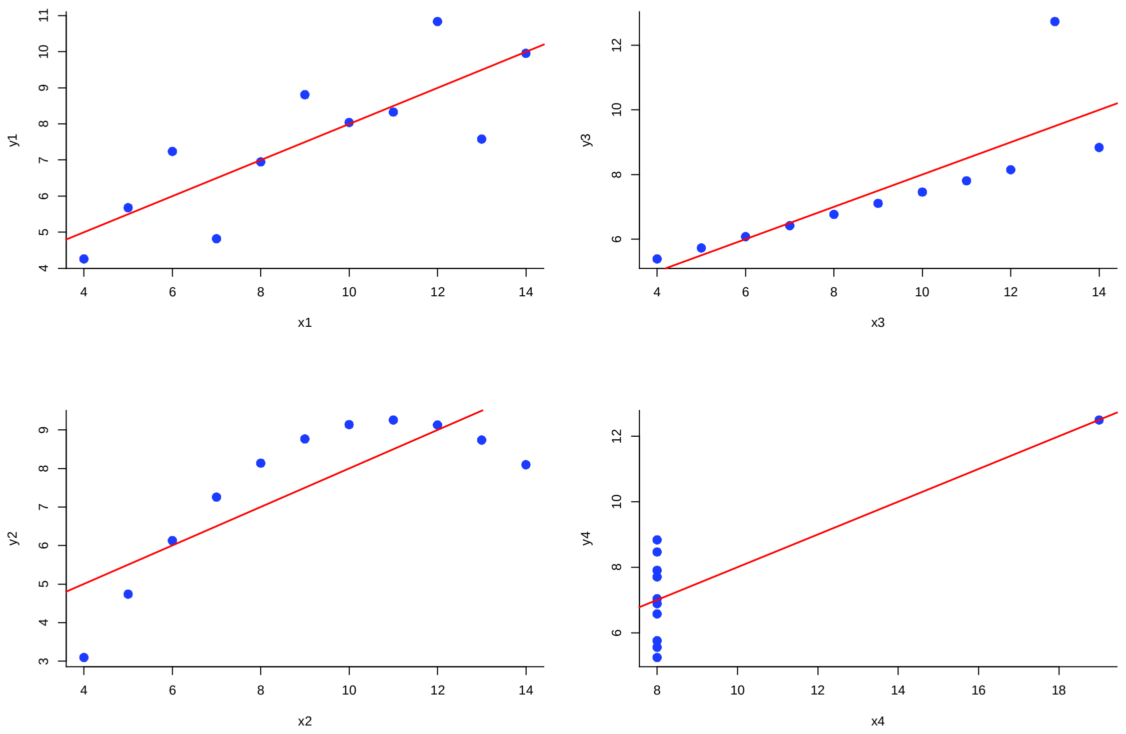
Multiple R-squared: **0.6119**, Adjusted R-squared: 0.6099

R^2

F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16

Regresión Lineal Simple

$$\hat{\beta}_0 \approx 3, \quad \hat{\beta}_1 \approx 0.5, \quad \mathrm{r}=0.7.$$



Bibliografía

- An introduction to Statistical Learning, 7th ed. Gareth, J., Witten, D., Hastie, T., Tibshirani, R., (2013), Springer. Capitulos 2 y 3.
- Linear Models and Generalizations: Least Squares and Alternatives, 3rd ed. Radhakrishna Rao C., Shalabh H. T., Heunaman (2008), Springer. Capitulos 1 y 2.