

# Análisis de datos e inferencia

9 de octubre de 2022

# Índice general

<b>1. Modelos de regresión lineal simple</b>	<b>2</b>
1.1. Introducción . . . . .	2
1.2. Modelo e hipótesis . . . . .	2
1.3. Estimación de los parámetros . . . . .	3
1.3.1. Método de máxima verosimilitud . . . . .	3
1.3.2. Estimación por mínimos cuadrados . . . . .	4
1.3.3. Estimación de la varianza . . . . .	4
1.4. Propiedades de los estimadores . . . . .	5
1.5. Intervalos de confianza para los parámetros . . . . .	6
1.6. Contraste de la regresión . . . . .	7
1.7. Evaluación del ajuste . . . . .	9
1.8. Predicción . . . . .	10
1.8.1. Estimación de las medias condicionadas . . . . .	10
1.8.2. Predicción de una observación futura . . . . .	11
1.9. Análisis de residuos y observaciones atípicas e influyentes . . . . .	11
1.10. Transformaciones . . . . .	12

# Capítulo 1

## Modelos de regresión lineal simple

### 1.1. Introducción

La regresión lineal es un modelo matemático que nos permite establecer la relación de dependencia entre una variable dependiente  $Y$  y una variable independiente  $X$ .

Nos interesan las relaciones de la forma  $y = f(x) + u$ , donde  $u$  es una variable aleatoria a la que llamamos perturbación. En el caso de la regresión lineal simple, el modelo será de la forma

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

con  $\beta_0$  y  $\beta_1$  parámetros. Llamamos intercepto a  $\beta_0$  y pendiente a  $\beta_1$ .

### 1.2. Modelo e hipótesis

Sea  $X$  una variable aleatoria cuantitativa,  $Y$  una variable aleatoria continua y  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  un conjunto de datos. Entonces el modelo de regresión lineal simple es

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \dots, n$$

#### Hipótesis del modelo

1.  $E(u_i) = 0, \quad \forall i = 1, \dots, n.$
2.  $Var(u_i) = \sigma^2, \quad \forall i = 1, \dots, n$  (homocedasticidad)
3.  $u_i \sim N(0, \sigma^2), \quad \forall i = 1, \dots, n$  (normalidad)
4.  $E(u_i u_j) = 0, \quad \forall i \neq j$  (independencia)

*Nota.* En realidad, la cuarta hipótesis es de incorrelación ( $Cov(u_i, u_j) = 0$ ).

$$Cov(u_i, u_j) = E(u_i u_j) - E(u_i)E(u_j) = E(u_i u_j)$$

Sin embargo, bajo normalidad la incorrelación y la independencia son equivalentes.

Podemos escribir las mismas hipótesis en términos de  $y_i$ ,  $\forall i = 1, \dots, n$ .

1.  $E(y_i|x_i) = E(\beta_0 + \beta_1 x_i + u_i) = \beta_0 + \beta_1 x_i$ ,  $\forall i = 1, \dots, n$  (linealidad)
2.  $Var(y_i|x_i) = Var(\beta_0 + \beta_1 x_i + u_i) = \sigma^2$ ,  $\forall i = 1, \dots, n$  (homocedasticidad)
3.  $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ,  $\forall i = 1, \dots, n$  (normalidad)
4.  $Cov(y_i, y_j) = 0$ ,  $\forall i \neq j$  (independencia)

Podemos dar un significado real a  $\beta_0$  y  $\beta_1$ :

- $\beta_0$  es el valor medio de la variable  $Y$  cuando  $x_i$  toma el valor 0.

$$E(y_i|x_i = 0) = \beta_0, \quad i = 1, \dots, n$$

- $\beta_1$  es la variación media que experimenta la variable  $Y$  cuando  $x_i$  aumenta en una unidad.

$$E(y_i|x_i + 1) - E(y_i|x_i) = \beta_1, \quad i = 1, \dots, n$$

### 1.3. Estimación de los parámetros

Queremos estimar  $\beta_0$ ,  $\beta_1$  y  $\sigma^2$ . Con los estimadores  $\hat{\beta}_0$  y  $\hat{\beta}_1$  podemos estimar

$$E(\hat{y}_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n$$

#### 1.3.1. Método de máxima verosimilitud

$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ,  $i = 1, \dots, n$ , así que podemos encontrar estimadores de máxima verosimilitud para los parámetros y para  $\sigma^2$ .

Usando el método de máxima verosimilitud llegamos las ecuaciones normales de la regresión:

$$\begin{cases} \frac{\partial \log L}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial \log(L)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

*Notación.*  $\hat{y}_i = E(\hat{y}_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Si definimos el error o residuo como  $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ , podemos escribir las ecuaciones normales de regresión de la siguiente forma:

$$\begin{cases} \sum_{i=1}^n e_i = 0 \\ \sum_{i=1}^n x_i e_i = 0 \end{cases}$$

Resolviendo este sistema, obtenemos los estimadores:

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{XY}}{s_X^2} \\ \hat{\beta}_0 &= \bar{y} - \frac{s_{XY}}{s_X^2} \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n e_i^2 \end{aligned}$$

La ecuación de la recta resultante es:

$$\hat{y}_i = \bar{y} + \frac{s_{XY}}{s_X^2} (x_i - \bar{x})$$

### 1.3.2. Estimación por mínimos cuadrados

Queremos minimizar la suma de los cuadrados de los errores  $\sum_{i=1}^n e_i^2$ , donde  $e_i = y_i - \hat{y}_i$ . Para ello minimizamos la función  $M(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ .

$$\begin{cases} \frac{\partial M}{\partial \beta_0}(\beta_0, \beta_1) = -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial M}{\partial \beta_1}(\beta_0, \beta_1) = -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

Simplificando obtenemos las ecuaciones normales de la regresión, como antes. Así que los estimadores de  $\beta_0$  y  $\beta_1$  por máxima verosimilitud coinciden con los estimadores por mínimos cuadrados.

### 1.3.3. Estimación de la varianza

Partiendo del estimador  $\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2$  obtenido previamente, podemos llegar a una expresión equivalente:

$$\hat{\sigma}^2 = s_Y^2 - \frac{s_{XY}^2}{s_X^2}$$

Veamos si este estimador es insesgado calculando su esperanza.

$$E(\hat{\sigma}^2) = E\left(\frac{\sum_{i=1}^n e_i^2}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^n e_i^2\right) = \frac{1}{n} \sigma^2 (n-2)$$

*Nota.*  $\frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi_{n-2}^2$ ,  $E\left(\frac{\sum_{i=1}^n e_i^2}{\sigma^2}\right) = n-2$

Observamos que este estimador no es insesgado. Consideramos entonces:

$$s_R^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

Este sí es un estimador insesgado de  $\sigma^2$  y le llamamos varianza residual. Tenemos la relación  $s_R^2 = \frac{n}{n-2} \hat{\sigma}^2$ .

## 1.4. Propiedades de los estimadores

Podemos escribir  $\hat{\beta}_1$  de la forma:

$$\hat{\beta}_1 = \sum_{i=1}^n w_i y_i, \quad w_i = \frac{x_i - \bar{x}}{ns_X^2}$$

Por las hipótesis del modelo,  $y_i$  son normales e independientes, luego  $\hat{\beta}_1 \sim N$ . Podemos calcular:

- $E(\hat{\beta}_1) = \beta_1$  (estimador insesgado)
- $V(\hat{\beta}_1) = \frac{\sigma^2}{ns_X^2}$

Por tanto,  $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{ns_X^2})$ .

De forma análoga, podemos escribir:

$$\hat{\beta}_0 = \sum_{i=1}^n \left( \frac{1}{n} - \bar{x} w_i \right)$$

Como las  $y_i$  son normales e independientes,  $\hat{\beta}_0 \sim N$ . Calculamos:

- $E(\hat{\beta}_0) = \beta_0$  (estimador insesgado)
- $V(\hat{\beta}_0) = \frac{\sigma^2}{n} \left( 1 + \frac{\bar{x}^2}{s_X^2} \right)$

Por tanto,  $\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n} (1 + \frac{\bar{x}^2}{s_X^2}))$ .

En cuanto a  $s_R^2$ , sabemos que  $\frac{1}{\sigma^2} \sum_{i=1}^n e_i^2 \sim \chi_{n-2}^2$ . Obtenemos que:

- $E(s_R^2) = \sigma^2$
- $V(s_R^2) = \frac{2}{n-2} (\sigma^2)^2$

## 1.5. Intervalos de confianza para los parámetros

### Intervalos de confianza para $\beta_1$

#### Caso 1: $\sigma^2$ conocida

Sabemos que  $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{ns_X^2})$ . Entonces:

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0, 1)$$

Por tanto, el intervalo de confianza para  $\beta_1$  a nivel de significación  $\alpha$  es:

$$IC_{1-\alpha}(\beta_1) = \left( \hat{\beta}_1 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{ns_X^2}}, \hat{\beta}_1 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{ns_X^2}} \right)$$

donde  $z_{1-\frac{\alpha}{2}}$  es el percentil de orden  $(1 - \frac{\alpha}{2})100\%$  de una variable aleatoria  $Z \sim N(0, 1)$ .

#### Caso 2: $\sigma^2$ desconocida

$$\left\{ \begin{array}{l} \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0, 1) \\ \frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2)s_R^2}{\sigma^2} \sim \chi_{n-2}^2 \end{array} \right. \Rightarrow \frac{\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}}}{\sqrt{\frac{(n-2)s_R^2}{\sigma^2} \frac{1}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}$$

Luego el intervalo de confianza para  $\beta_1$  a nivel de significación  $\alpha$  es:

$$IC_{1-\alpha}(\beta_1) = \left( \hat{\beta}_1 - t_{n-2, 1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{ns_X^2}}, \hat{\beta}_1 + t_{n-2, 1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{ns_X^2}} \right)$$

donde  $s_R = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$  y  $t_{n-2, 1-\frac{\alpha}{2}}$  es el percentil de orden  $(1 - \frac{\alpha}{2})100\%$  de una variable aleatoria  $T \sim t_{n-2}$ .

### Intervalos de confianza para $\beta_0$

#### Caso 1: $\sigma^2$ conocida

Sabemos que  $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} \left(1 + \frac{\bar{x}^2}{s_X^2}\right)\right)$ . Entonces, el intervalo de confianza para  $\beta_0$  a nivel de significación  $\alpha$  es:

$$IC_{1-\alpha}(\beta_0) = \left( \hat{\beta}_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{\bar{x}^2}{s_X^2}}, \hat{\beta}_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{\bar{x}^2}{s_X^2}} \right)$$

donde  $z_{1-\frac{\alpha}{2}}$  es el percentil de orden  $(1 - \frac{\alpha}{2})100\%$  de una variable aleatoria  $Z \sim N(0, 1)$ .

### Caso 2: $\sigma^2$ desconocida

Razonando de forma análoga al caso de  $\beta_1$ , tenemos que:

$$\frac{\hat{\beta}_0 - \beta_0}{\frac{s_R}{n} \sqrt{1 + \frac{\bar{x}^2}{s_X^2}}}$$

Por tanto, el intervalo de confianza para  $\beta_0$  a nivel de significación  $\alpha$  es:

$$IC_{1-\alpha}(\beta_0) = \left( \hat{\beta}_0 - t_{n-2, 1-\frac{\alpha}{2}} \frac{s_R}{\sqrt{n}} \sqrt{1 + \frac{\bar{x}^2}{s_X^2}}, \hat{\beta}_0 + t_{n-2, 1-\frac{\alpha}{2}} \frac{s_R}{\sqrt{n}} \sqrt{1 + \frac{\bar{x}^2}{s_X^2}} \right)$$

donde  $t_{n-2, 1-\frac{\alpha}{2}}$  es el percentil de orden  $(1 - \frac{\alpha}{2})100\%$  de una variable aleatoria  $T \sim t_{n-2}$ .

### Intervalos de confianza para $\sigma^2$

Sabemos que  $\frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2)s_R^2}{\sigma^2} \sim \chi_{n-2}^2$ . Queremos que  $P(a < \sigma^2 < b) = 1 - \alpha$ .

$$\begin{aligned} P(a < \sigma^2 < b) &= P\left(\frac{1}{b} < \frac{1}{\sigma^2} < \frac{1}{a}\right) = \\ &= P\left(\frac{(n-2)s_R^2}{b} < \frac{(n-2)s_R^2}{\sigma^2} < \frac{(n-2)s_R^2}{a}\right) \end{aligned}$$

Luego:

$$\begin{cases} \frac{(n-2)s_R^2}{b} = \chi_{n-2, \frac{\alpha}{2}}^2 \Rightarrow b = \frac{(n-2)s_R^2}{\chi_{n-2, \frac{\alpha}{2}}^2} \\ \frac{(n-2)s_R^2}{a} = \chi_{n-2, 1-\frac{\alpha}{2}}^2 \Rightarrow a = \frac{(n-2)s_R^2}{\chi_{n-2, 1-\frac{\alpha}{2}}^2} \end{cases}$$

Por tanto, el intervalo de confianza para  $\sigma^2$  a nivel de significación  $\alpha$  tiene por extremos  $a$  y  $b$ , es decir:

$$IC_{1-\alpha}(\sigma^2) = (a, b)$$

## 1.6. Contraste de la regresión

Consideramos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \beta_1 = 0 \Leftrightarrow E(y|x) = \beta_0 \\ H_1 : \beta_1 \neq 0 \Leftrightarrow E(y|x) = \beta_0 + \beta_1 x \end{cases}$$

Fijamos el nivel de significación  $\alpha$ . Podemos resolverlo de cuatro formas distintas.



## Intervalos de confianza

Sea  $IC_{1-\alpha}(\beta_1)$  el intervalo de confianza para  $\beta_1$  a nivel de significación  $\alpha$ . Entonces:

- Aceptamos  $H_0$  a nivel de significación  $\alpha$  si  $0 \in IC_{1-\alpha}(\beta_1)$ .
- Rechazamos  $H_0$  a nivel de significación  $\alpha$  en caso contrario.

## Estadístico $T$

Sabemos que  $\frac{\hat{\beta}_1 - \beta_1}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}$ . Entonces  $T = \frac{\hat{\beta}_1}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}$  si  $H_0$  es cierto.

Tomamos un  $t_{exp}$ .

- Si  $t_{exp} \in (-t_{n-2, 1-\frac{\alpha}{2}}, t_{n-2, 1-\frac{\alpha}{2}})$ , o equivalentemente  $|t_{exp}| \leq t_{n-2, 1-\frac{\alpha}{2}}$ , aceptamos  $H_0$  a nivel de significación  $\alpha$ .
- En caso contrario, rechazamos  $H_0$  a nivel de significación  $\alpha$ .

## Valor $p$

Sea  $p$  el valor  $p$  o  $p$ -valor de la distribución. Entonces:

- Si  $p \geq \alpha$ , aceptamos  $H_0$  a nivel de significación  $\alpha$ .
- En caso contrario, rechazamos  $H_0$  a nivel de significación  $\alpha$ .

## Tabla ANOVA

Partimos de que podemos escribir:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Definimos:

- Variabilidad total:

$$VT = \sum_{i=1}^n (y_i - \bar{y})^2 = ns_Y^2$$

- Variabilidad no explicada:

$$VNE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = (n-2)s_R^2 = n\hat{\sigma}^2 = n(s_Y^2 - \hat{\beta}_1^2 s_X^2)$$

- Variabilidad explicada:

$$VE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = VT - VNE = n\hat{\beta}_1^2 s_X^2$$

Observamos que:

$$\frac{VNE}{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi_{n-2}^2$$

Además, como  $\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0, 1)$ , entonces:

$$\frac{(\hat{\beta}_1 - \beta_1)^2}{\frac{\sigma^2}{ns_X^2}} \sim \chi_1^2$$

Luego  $\frac{\hat{\beta}_1^2}{\frac{\sigma^2}{ns_X^2}} = \frac{n\hat{\beta}_1^2 s_X^2}{\sigma^2} = \frac{VE}{\sigma^2} \sim \chi_1^2$  si  $H_0$  es cierta.

Consideramos ahora  $F = \frac{\frac{VE}{\sigma^2}/1}{\frac{VNE}{\sigma^2}/(n-2)} = \frac{VE}{s_R^2}$ . Observamos que  $F \sim F_{1,n-2}$  si  $H_0$  es cierta.

Tomamos un  $F_{exp}$ .

- Aceptamos  $H_0$  a nivel de significación  $\alpha$  si  $F_{exp} \leq F_{1,n-2,1-\alpha}$ .
- En caso contrario, rechazamos  $H_0$  a nivel de significación  $\alpha$ .

La tabla ANOVA es de la forma:

Fuentes	Suma de cuadrados	Grados de libertad	Cocientes
$VE$	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	1	$\frac{VE}{1}$
$VNE$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - 2$	$\frac{VNE}{n-2}$
$VT$	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	

También se incluyen columnas para  $F_{exp}$  y  $p$ -valor.

*Observación.* Existe la siguiente relación entre  $t_{exp}$  y  $F_{exp}$ :

$$t_{exp}^2 = F_{exp}$$

## 1.7. Evaluación del ajuste

Existen dos coeficientes para evaluar el ajuste del modelo: el coeficiente de correlación lineal y el coeficiente de determinación.

### Coeficiente de correlación lineal

El coeficiente de correlación lineal se define como:

$$r = \frac{s_{XY}}{s_X s_Y}, \quad -1 \leq r \leq 1$$

- Si  $r = 1$ , se tiene dependencia lineal exacta positiva.
- Si  $r = -1$ , se tiene dependencia lineal exacta negativa.
- Si  $r = 0$ , las variables están incorreladas linealmente.

Se dice que el ajuste es bueno si  $|r|$  es cercano a 1. Si por el contrario  $r$  se aproxima a 0, entonces las variables no tienen relación lineal.

## Coeficiente de determinación

El coeficiente de determinación se define como:

$$R^2 = \frac{VE}{VT}, \quad 0 \leq R^2 \leq 1$$

- Si  $R^2 = 1$  entonces  $VE = VT$  luego  $VNE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = 0$ . Por tanto,  $e_i = 0$  para todo  $i = 1, \dots, n$ , así que el ajuste lineal es exacto.
- Si  $R^2 = 0$  entonces  $VE = 0$ , luego  $VT = VNE$ . Así que el ajuste lineal es pésimo.

**Teorema 1.1.** *El coeficiente de determinación coincide con el coeficiente de correlación lineal al cuadrado. Es decir,*

$$r^2 = R^2$$

*Demostración.*

$$R^2 = \frac{VE}{VT} = \frac{n\hat{\beta}_1^2 s_X^2}{ns_Y^2} = \frac{\left(\frac{s_{XY}}{s_X^2}\right)^2 s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = r^2$$

□

## 1.8. Predicción

### 1.8.1. Estimación de las medias condicionadas

Llamamos  $m_0 = E(y|x = x_0) = \beta_0 + \beta_1 x_0$ . Observamos que  $m_0$  es un parámetro que podemos estimar de la forma:

$$\hat{m}_0 = E(y|\hat{x} = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

**Teorema 1.2.**

$$\hat{m}_0 \sim N\left(m_0, \frac{\sigma^2}{n} \left(1 + \frac{(x_0 - \bar{x})^2}{s_X^2}\right)\right)$$

### Intervalos de confianza para $m_0$

Podemos calcular los intervalos de confianza para  $m_0$  con nivel de confianza de  $100(1 - \alpha)\%$ .

Si  $\sigma^2$  es conocida,

$$\left(\hat{m}_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}}, \hat{m}_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}}\right)$$

Si  $\sigma^2$  es desconocida,

$$\left(\hat{m}_0 - t_{n-2, 1-\frac{\alpha}{2}} \frac{s_R}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}}, \hat{m}_0 + t_{n-2, 1-\frac{\alpha}{2}} \frac{s_R}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}}\right)$$

### 1.8.2. Predicción de una observación futura

Dado un conjunto de datos  $(x_1, y_1), \dots, (x_n, y_n)$  y dado  $x_0$  queremos predecir:

$$y_0 = \beta_0 + \beta_1 x_0 + u_0$$

$u_0$  es una variable aleatoria independiente de  $u_1, \dots, u_n$  con  $u_0 \sim N(0, \sigma^2)$ . Observamos que  $y_0$  es una variable aleatoria, a diferencia de la estimación

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{m}_0$$

Consideramos el error:

$$e_0 = y_0 - \hat{y}_0 = \beta_0 + \beta_1 x_0 + u_0 - (\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

que también es una variable aleatoria.

**Teorema 1.3.**

$$e_0 \sim \left( 0, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_X^2} \right) \right)$$

#### Intervalos de pronóstico para $y_0$

Podemos calcular los intervalos de pronóstico  $IP_{1-\alpha}(y_0)$  para  $y_0$  con contenido probabilístico  $1 - \alpha$ .

Si  $\sigma^2$  es conocida,

$$\left( \hat{y}_0 - z_{1-\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_X^2}}, \hat{y}_0 + z_{1-\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_X^2}} \right)$$

Si  $\sigma^2$  es desconocida,

$$\left( \hat{y}_0 - t_{n-2, 1-\frac{\alpha}{2}} s_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_X^2}}, \hat{y}_0 + t_{n-2, 1-\frac{\alpha}{2}} s_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_X^2}} \right)$$

## 1.9. Análisis de residuos y observaciones atípicas e influyentes

### Residuos

El residuo de un dato es la diferencia entre su valor y la predicción mediante el modelo.

$$e_i = y_i - \hat{y}_i, \quad \forall i = 1, \dots, n$$

El análisis de los residuos puede darnos información sobre el ajuste del modelo.

## Observaciones atípicas

Una observación atípica es un valor que es numéricamente distinto al resto de los datos. Visualmente, es un dato que se sale del patrón. Las observaciones atípicas pueden ser indicativas de errores de observación o errores en el modelo. Un error de observación se debe a datos que pertenecen a una población diferente del resto de muestras, mientras que un error en el modelo puede ser debido a que la muestra depende una variable desconocida que no se han tenido en cuenta.

## Observaciones influyentes

Una observación influyente  $(x_A, y_A)$  es una observación atípica cuya exclusión produce un cambio drástico en la recta de regresión. Puede ser causada por un error de observación o por un modelo incorrecto. Algunas posibles causas de que el modelo sea incorrecto son:

- La relación entre  $x$  e  $y$  no es lineal cerca de  $x_A$ .
- La varianza aumenta mucho con  $x$ .
- Una variable desconocida ha tomado un valor distinto en  $x_A$ .

## Puntos palanca

Los puntos palanca son observaciones con un valor alto de  $p_i$ . Estos tienen la capacidad de alterar en gran medida la recta de regresión.

## 1.10. Transformaciones

Cuando el diagrama de dispersión entre las dos variables o el de los residuos presenta indicios de incumplimiento de alguna hipótesis básica, entonces hay que abandonar el modelo inicial por uno menos simple o bien aplicar alguna transformación a los datos.