Módulo 3: Regresión II

Lección 4: Variables Cualitativas y Predicción

Variables cualitativas como regresores

Consumo	Cilindrada	Potencia	Peso	Aceleración	Origen
l/100Km	*cc*	*CV*	*kg*	*segundos*	
15	4982	150	1144	12	Europa
16	6391	190	1283	9	Japón
24	5031	200	1458	15	USA
9	1491	70	651	21	Europa
11	2294	72	802	19	Japón
17	5752	153	1384	14	USA
12	2294	90	802	20	Europa
17	6555	175	1461	12	USA
18	6555	190	1474	13	USA
12	1147	97	776	14	Japón
16	5735	145	1360	13	USA
12	1868	91	860	14	Europa
9	2294	75	847	17	USA
...	...	...	...	...	...

Variables cualitativas como regresores

Origen $\begin{cases} \text{Europa} \\ \text{Japón} \\ \text{USA} \end{cases}$

$$Z_{JAP\,i} = \begin{cases} 0 & \text{si } i \notin \text{JAPON} \\ 1 & \text{si } i \in \text{JAPON} \end{cases}$$

$$Z_{USA\,i} = \begin{cases} 0 & \text{si } i \notin \text{USA} \\ 1 & \text{si } i \in \text{USA} \end{cases}$$

$$Z_{EUR\,i} = \begin{cases} 0 & \text{si } i \notin \text{EUROPA} \\ 1 & \text{si } i \in \text{EUROPA} \end{cases}$$

$$\text{Consumo} = \beta_0 + \beta_1\,\text{CC} + \beta_2\,\text{Pot} + \beta_3\,\text{Peso} +$$

$$+ \beta_4\,\text{Acel} + \alpha_{JAP}\,Z_{JAP} + \alpha_{USA}\,Z_{USA} + \text{Error}$$

Variables cualitativas

Consumo	Cilindrada	Potencia	Peso	Aceleración	ZJAP	ZUSA	ZEUR
l/100Km	*cc*	*CV*	*kg*	*segundos*			
15	4982	150	1144	12	0	0	**1**
16	6391	190	1283	9	**1**	0	0
24	5031	200	1458	15	0	**1**	0
9	1491	70	651	21	0	0	**1**
11	2294	72	802	19	**1**	0	0
17	5752	153	1384	14	0	**1**	0
12	2294	90	802	20	0	0	**1**
17	6555	175	1461	12	0	**1**	0
18	6555	190	1474	13	0	**1**	0
12	1147	97	776	14	**1**	0	0
16	5735	145	1360	13	0	**1**	0
12	1868	91	860	14	0	0	**1**
9	2294	75	847	17	0	**1**	0
...	...	...	...	...	...	...	...

$$\text{Consumo} = \beta_0 + \beta_1 \text{CC} + \beta_2 \text{Pot} + \beta_3 \text{Peso} +$$

$$+ \beta_4 \text{Acel} + \alpha_{JAP} Z_{JAP} + \alpha_{USA} Z_{USA} + \text{Error}$$

Interpretación var. cualitativa

$$\text{Consumo} = \beta_0 + \beta_1 \text{CC} + \beta_2 \text{Pot} + \beta_3 \text{Peso} + \beta_4 \text{Acel} + \alpha_{JAP} Z_{JAP} + \alpha_{USA} Z_{USA} + \text{Error}$$

- Coches europeos: $Z_{JAP} = 0$ y $Z_{USA} = 0$ REFERENCIA

$$\text{Consumo} = \beta_0 + \beta_1 \text{CC} + \beta_2 \text{Pot} + \beta_3 \text{Peso} + \beta_4 \text{Acel} + \text{Error}$$

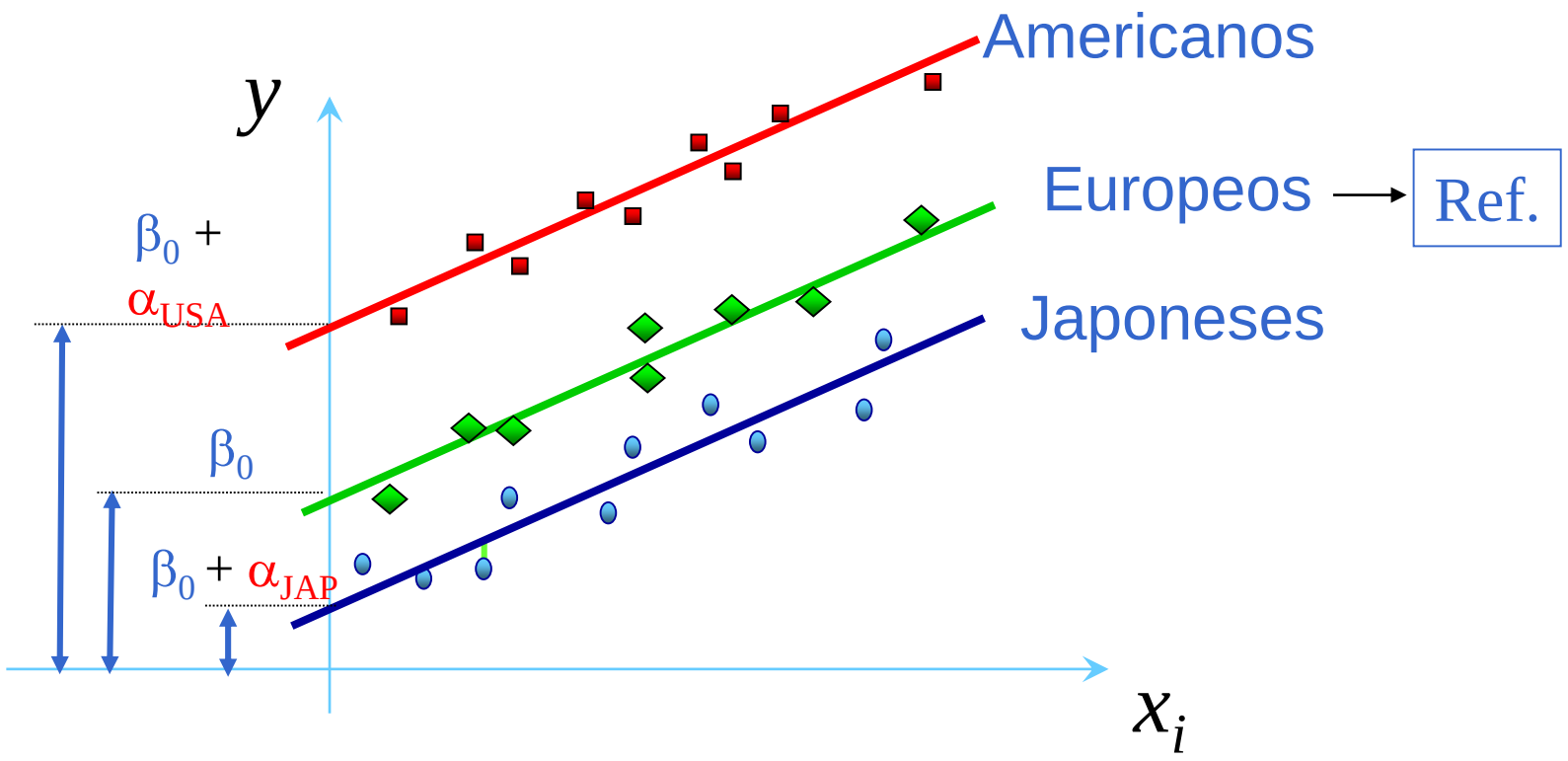
- Coches japoneses: $Z_{JAP} = 1$ y $Z_{USA} = 0$

$$\text{Consumo} = \underbrace{\beta_0 + \alpha_{JAP}} + \beta_1 \text{CC} + \beta_2 \text{Pot} + \beta_3 \text{Peso} + \beta_4 \text{Acel} + \text{Error}$$

- Coches americanos: $Z_{JAP} = 0$ y $Z_{USA} = 1$

$$\text{Consumo} = \underbrace{\beta_0 + \alpha_{USA}} + \beta_1 \text{CC} + \beta_2 \text{Pot} + \beta_3 \text{Peso} + \beta_4 \text{Acel} + \text{Error}$$

Interpretación del modelo



```
> ZUSA = Origen == 1
> ZEUR = Origen == 2
> ZJAP = Origen == 3
> mod_coches = lm(Consumo ~CC + CV + Peso + Acel + ZJAP  + ZUSA)
> summary(mod_coches)
```

```
Call:
lm(formula = Consumo ~ CC + CV + Peso + Acel + ZJAP + ZUSA)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0965 -0.9687  0.0213  0.9496  5.4896

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.4550414  1.0172459  -1.430   0.1534
CC           0.0003228  0.0001792   1.801   0.0724 .
CV           0.0422677  0.0067890   6.226 1.26e-09 ***
Peso         0.0055996  0.0009655   5.799 1.39e-08 ***
Acel         0.1108411  0.0496919   2.231   0.0263 *
ZJAPTRUE    -0.3617622  0.2790488  -1.296   0.1956
ZUSATRUE     0.0611229  0.2802356   0.218   0.8275
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.66 on 384 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.821,      Adjusted R-squared: 0.8182
F-statistic: 293.5 on 6 and 384 DF,  p-value: < 2.2e-16
```

Interpretación

- El *p-valor* del coeficiente asociado a Z_{JAP} es 0.1956>.05, se concluye que no existe diferencia significativa entre el consumo de los coches Japoneses y Europeos (manteniendo constante el peso, cc, pot y acel.)
- La misma interpretación para Z_{USA}.
- Comparando R^2 =0.821 de este modelo con el anterior R^2=0.8197, se confirma que el modelo con las variables de *Origen* no suponen una mejora sensible.

Modelo de regresión con variables cualitativas

- En general, para considerar una variable cualitativa con *r* niveles, se introducen en la ecuación *r-1* variables ficticias

$$z_{1i} = \begin{cases} 0 & i \notin \text{nivel1} \\ 1 & i \in \text{nivel1'} \end{cases} \quad z_{2i} = \begin{cases} 0 & i \notin \text{nivel2} \\ 1 & i \in \text{nivel2'} \end{cases} \quad \cdots, z_{r-1i} = \begin{cases} 0 & i \notin \text{nivel}r-1 \\ 1 & i \in \text{nivel}r-1 \end{cases}$$

Y el nivel *r* no utilizado es el que actúa de referencia

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \underbrace{+\alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_{r-1} z_{r-1,i}}_{\text{variable cualitativa}} + u_i$$

Ejemplo: Body

Nombre: Body (Cuerpo Humano) **Exploring Relationships in Body Dimensions**

507 Observaciones, 25 Variables

Descripción: Este ejemplo contiene 21 medidas del cuerpo humano, además de la edad, peso, altura y género (mujeres = 0, hombres =1) de 507 individuos de los que 247 son hombres y 260 mujeres. Los datos fueron recogidos entre personas que acudía frecuentemente al gimnasio en USA, la mayoría de ellos entre 20 y 40 años.

Fuente: Exploring Relationships in Body Dimensions**,** Grete Heinz,Louis J. Peterson,Roger W. Johnson , Carter J. Kerk, *Journal of Statistics Education* Volume 11, Number 2 (2003), www.amstat.org/publications/jse/v11n2/datasets.heinz.html

OBJETIVO: Relación entre el peso y altura diferenciando entre hombres y mujeres.

Body

	Estatura	Peso
Hombres	177.7cm	78.1 kg
Mujeres	164.9cm	60.6 kg
Diferencia	**12.8 cm**	**17.5 kg**

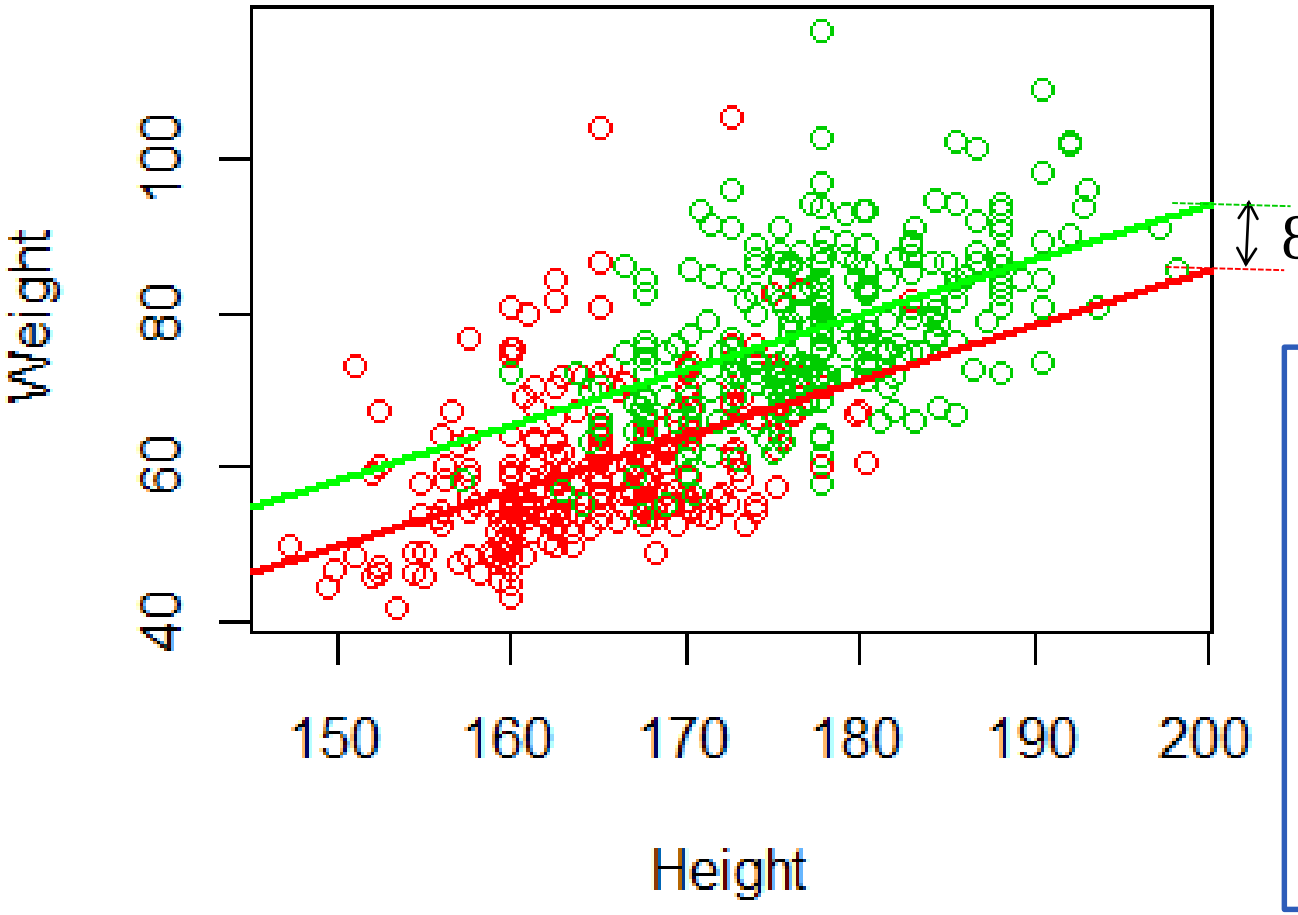
$$\text{Weight} = \beta_0 + \beta_1 \text{ Height} + \alpha_{HOM} Z_{HOM} + \text{Error}$$

$$\text{Weight} = -56.9 + 0.713 \text{ Height} + 8.366 Z_{HOM} + \text{Error}$$

	Pos	Gender	Age	Weight	Height
1	1	0	22	51.6	161.2
2	2	0	20	59.0	167.5
3	3	0	19	49.2	159.5
4	4	0	25	63.0	157.0
5	5	0	21	53.6	155.8
6	6	0	23	59.0	170.0
7	7	0	26	47.6	159.1
8	8	0	22	69.8	166.0
9	9	0	28	66.8	176.2
10	10	0	40	75.2	160.2
11	11	0	32	55.2	172.5
12	12	0	25	54.2	170.9

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -56.94949    9.42444  -6.043 2.95e-09 ***
Height        0.71298    0.05707  12.494  < 2e-16 ***
Gender        8.36599    1.07296   7.797 3.66e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 8.802 on 504 degrees of freedom
Multiple R-squared: 0.5668, Adjusted R-squared: 0.5651
F-statistic: 329.7 on 2 and 504 DF,  p-value: < 2.2e-16
```

Interpretación



A igualdad de ESTATURA, la diferencia de PESO entre un hombre y una mujer es 8.36 kg

FEV (Ejemplo 3)

Ejemplo "Fev" Forced Expiratory Volume (FEV)
654 observaciones, 5 variables

Descripción: Es una muestra de 654 jóvenes entre 3 y 19 años recogidos en Boston (USA) a finales de los 70. Se desea ver la relación entre la capacidad pulmonar (FEV) y fumar. En este primer análisis estudiaremos la relación entre FEV y la estatura. En la lección de regresión múltiple estudiaremos el efecto del tabaco.

Fuente:
Rosner, B. (1999), Fundamentals of Biostatistics, 5th Ed., Pacific Grove, CA: Duxbury

Variables

age	años del individuo
fev	variable continua en litros
ht	variable continua, estatura en pulgadas
sex	cualitativa (mujer=0, hombre=1)
smoke	cualitativa (No-fumador=0, fumador=1)

	age	fev	ht	sex	smoke
1	9	1.708	57.0	0	0
2	8	1.724	67.5	0	0
3	7	1.720	54.5	0	0
4	9	1.558	53.0	1	0
5	9	1.895	57.0	1	0
6	8	2.336	61.0	0	0
...					

Tabla 6.1

Modelo de regresión

$$\text{Log(fev)} = \beta_0 + \beta_1\, \text{ht} + \beta_2\, \text{age} + \alpha_{HOM}\, Z_{HOM} + \alpha_{HOM}\, Z_{HOM} + \text{Error}$$

$$\text{Log(fev)} = -1.9 + 0.042\text{ht} + 0.023\text{age} + 0.029\, Z_{HOM} - 0.046\, Z_{FUM} + \text{Error}$$

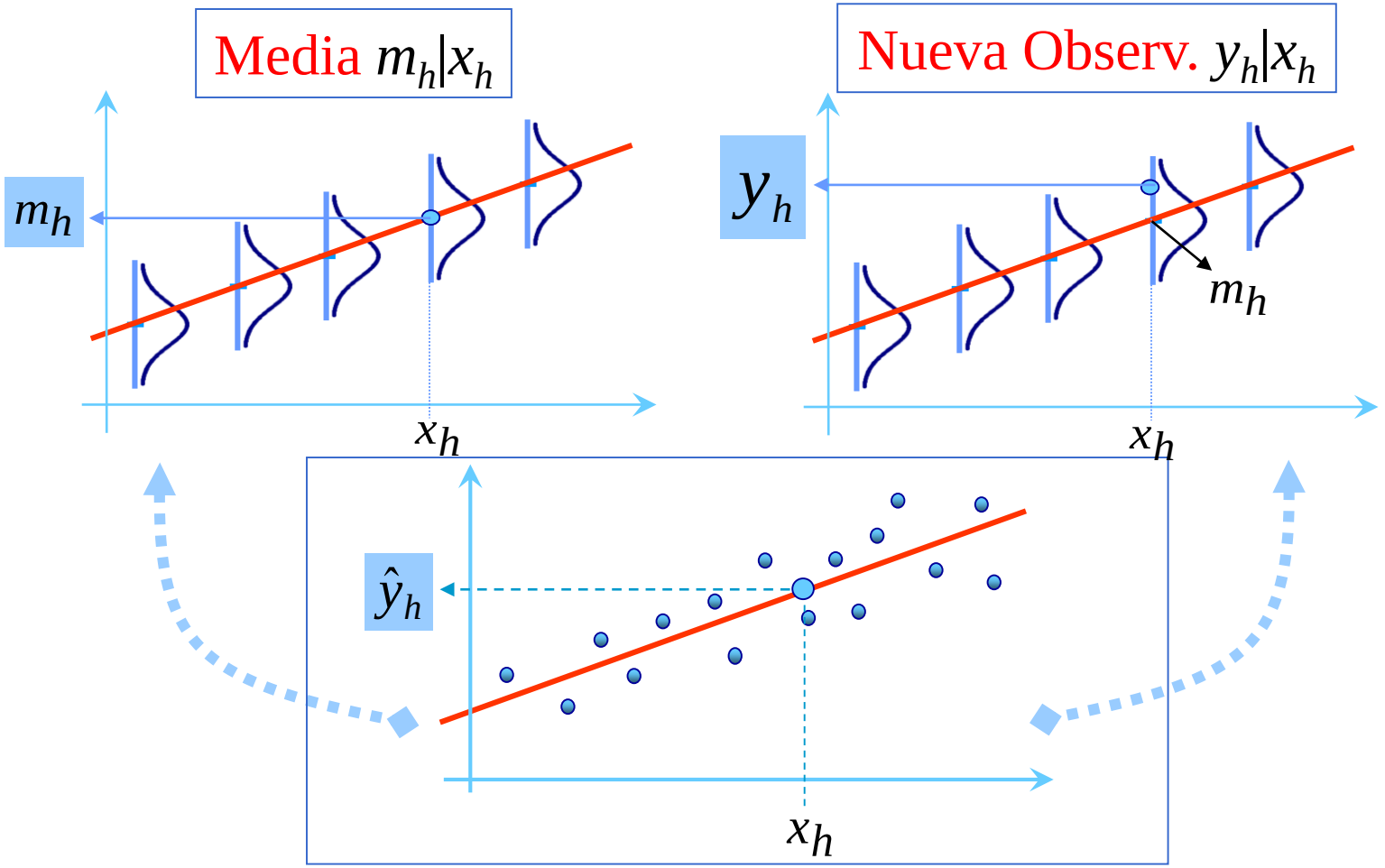
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
ht           0.042796   0.001679  25.489  < 2e-16 ***
age          0.023387   0.003348   6.984  7.1e-12 ***
sex          0.029319   0.011719   2.502   0.0126 *
smoke       -0.046068   0.020910  -2.203   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.1455 on 649 degrees of freedom
Multiple R-squared: 0.8106, Adjusted R-squared: 0.8095
F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

Interpretación

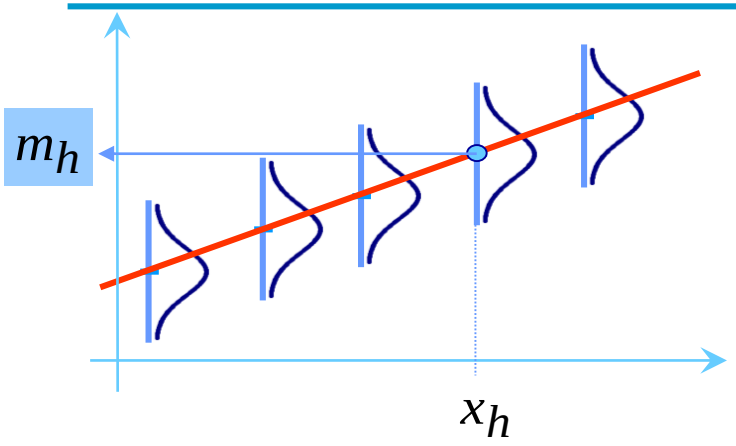
1. Todos los coeficientes son significativamente distintos de cero.
2. A igualdad del resto de las variables, un aumento de 1cm en la Estatura produce un incremento en fev del 4.2%
3. A igualdad del resto de las variables, un aumento de 1 año en la Edad produce un incremento en fev del 2.3%
4. A igualdad del resto de las variables, los hombres tienen un 2.9% más de fev que las mujeres.
5. **A igualdad del resto de las variables, los fumadores tienen un 4.6% menos de fev que los no-fumadores.**

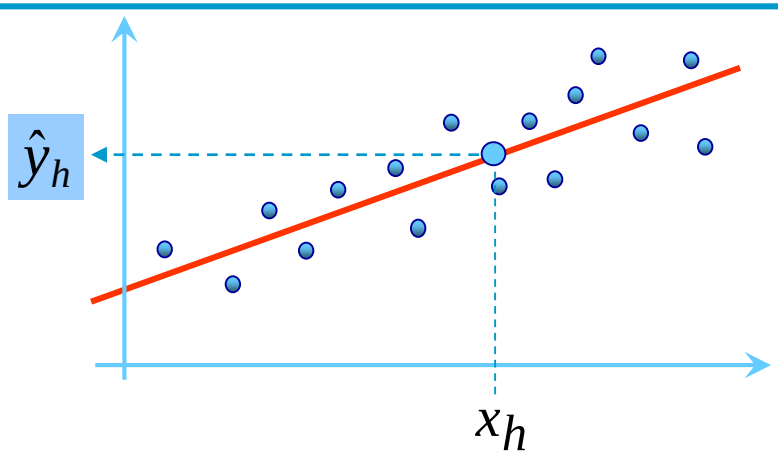
IMPORTANTE: El objetivo del estudio era cuantificar el efecto de fumar en la capacidad pulmonar de los jóvenes, el restos de las variables del modelo son necesarias (imprescindibles) para detectar el efecto, aunque juegan un papel secundario.

Predicción



Predicción de la media m_h (Regresión simple)



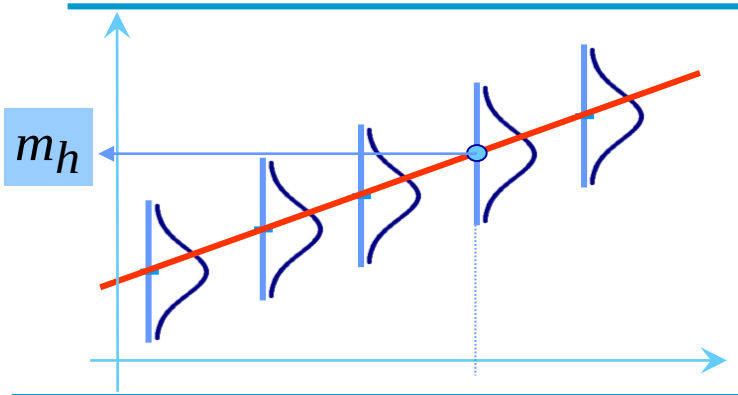


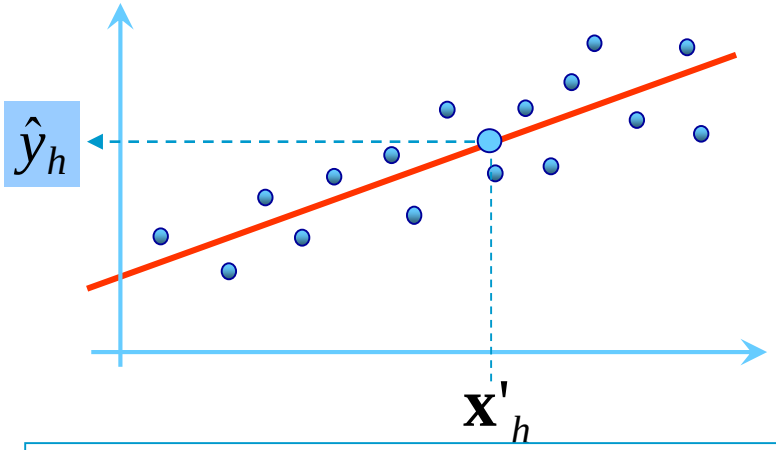
$$y_h \to N(\beta_0 + \beta_1 x_h, \sigma^2)$$
$$m_h = \beta_0 + \beta_1 x_h$$

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = \bar{y} + \hat{\beta}_1 (x_h - \bar{x})$$
$$E[\hat{y}_h] = E[\hat{\beta}_0 + \hat{\beta}_1 x_h] = \beta_0 + \beta_1 x_h = m_h$$
$$\text{var}[\hat{y}_h] = \text{var}[\bar{y} + \hat{\beta}_1 (x_h - \bar{x})]$$
$$= \text{var}[\bar{y}] + (x_h - \bar{x})^2 \, \text{var}[\hat{\beta}_1]$$
$$= \frac{\sigma^2}{n} + (x_h - \bar{x})^2 \frac{\sigma^2}{n s_x^2}$$

$$\hat{y}_h \to N\left(m_h, \frac{\sigma^2}{n}\left(1 + \frac{(x_h - \bar{x})^2}{s_x^2}\right)\right)$$

Predicción de la media m_h (Regresión múltiple)





$$y_h \to N(m_h, \sigma^2) \quad x_h$$

$$m_h = \beta_0 + \beta_1 x_{1h} + \cdots + \beta_k x_{kh} = \boldsymbol{\beta}^T \mathbf{x}'_h$$

$$\hat{y}_h \to N\left(m_h,\ \sigma^2 v_{hh} \right)$$

$$\hat{y}_h = \hat{\boldsymbol{\beta}}^T \mathbf{x}'_h, \quad \mathbf{x}'^T_h = (1, x_{1h}, x_{2h}, \cdots, x_{kh})$$

$$E[\hat{y}_h] = E[\hat{\boldsymbol{\beta}}^T \mathbf{x}'_h] = E[\hat{\boldsymbol{\beta}}^T]\mathbf{x}'_h = \boldsymbol{\beta}^T \mathbf{x}'_h$$

$$\text{var}[\hat{y}_h] = \text{var}[\hat{\boldsymbol{\beta}}^T \mathbf{x}'_h] = \mathbf{x}'^T_h \,\text{var}[\hat{\boldsymbol{\beta}}^T]\mathbf{x}'_h$$

$$= \mathbf{x}'^T_h (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}'_h \sigma^2 = v_{hh}\sigma^2$$

$$v_{hh} = \mathbf{x}'^T_h (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}'_h$$

Expresión alternativa para v_{hh}

$$\hat{y}_h = \bar{y} + \hat{\mathbf{b}}^T(\mathbf{x}_h - \bar{\mathbf{x}})$$

$$\text{var}[\hat{y}_h] = \text{var}[\bar{y} + \hat{\mathbf{b}}^T(\mathbf{x}_h - \bar{\mathbf{x}})] = \text{var}[\bar{y}] + (\mathbf{x}_h - \bar{\mathbf{x}})^T \,\text{var}[\hat{\mathbf{b}}](\mathbf{x}_h - \bar{\mathbf{x}})$$

$$= \frac{\sigma^2}{n} + (\mathbf{x}_h - \bar{\mathbf{x}})^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{x}_h - \bar{\mathbf{x}})\sigma^2, \qquad (\mathbf{S}_x = \frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n})$$

$$= \frac{\sigma^2}{n}(1 + (\mathbf{x}_h - \bar{\mathbf{x}})^T \mathbf{S}_x^{-1}(\mathbf{x}_h - \bar{\mathbf{x}}))$$

$$v_{hh} = \frac{1}{n}(1 + (\mathbf{x}_h - \bar{\mathbf{x}})^T \mathbf{S}_x^{-1}(\mathbf{x}_h - \bar{\mathbf{x}}))$$

$$\mathbf{x}_h = \bar{\mathbf{x}} \Rightarrow v_{hh} = 1/n$$

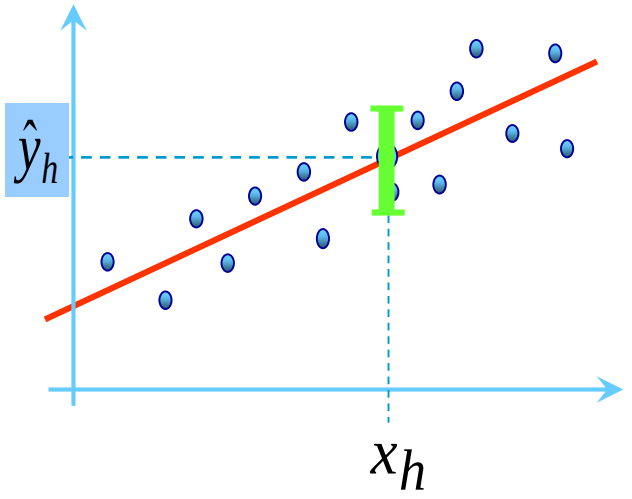
$$\mathbf{x}_h \neq \bar{\mathbf{x}} \Rightarrow v_{hh} > 1/n$$

Intervalos de confianza para la media m_h

$$\hat{y}_h \rightarrow N\left(m_h,\ \sigma^2 v_{hh}\right)$$

$$\frac{\hat{y}_h - m_h}{\sigma\sqrt{v_{hh}}} \rightarrow N(0,1)$$

$$\frac{\hat{y}_h - m_h}{\hat{s}_R\sqrt{v_{hh}}} \rightarrow t_{n-k-1}$$



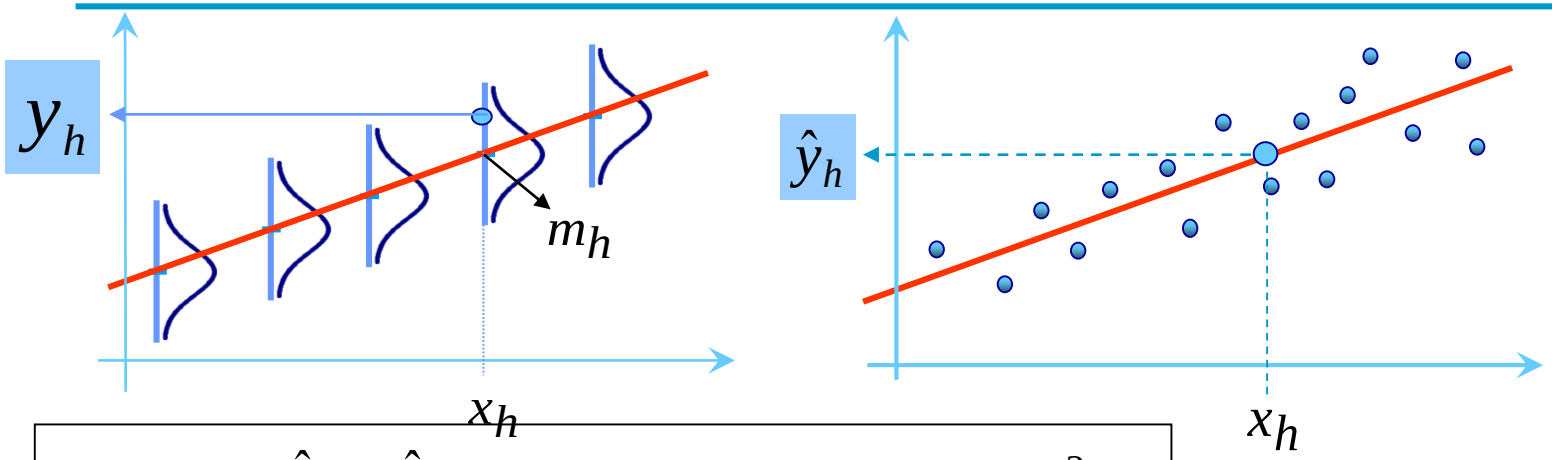
$$m_h \in \hat{y}_h \pm t_{\alpha/2}\hat{s}_R\sqrt{v_{hh}}$$

Regresión simple

$$v_{hh} = \frac{1}{n}\left(1+(\mathbf{x_h}-\bar{\mathbf{x}})^T\mathbf{S}_x^{-1}(\mathbf{x_h}-\bar{\mathbf{x}})\right)$$

$$v_{hh} = \frac{1}{n}\left(1+\frac{(x_h-\bar{x})^2}{s_x^2}\right)$$

Predicción de una nueva observación y_h (Reg.Simple)



$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h \qquad y_h \to N(m_h, \sigma^2)$$

$$\hat{y}_h \to N(m_h, \sigma^2 v_{hh}) \qquad m_h = \beta_0 + \beta_1 x_h$$

$$\tilde{e}_h = y_h - \hat{y}_h$$

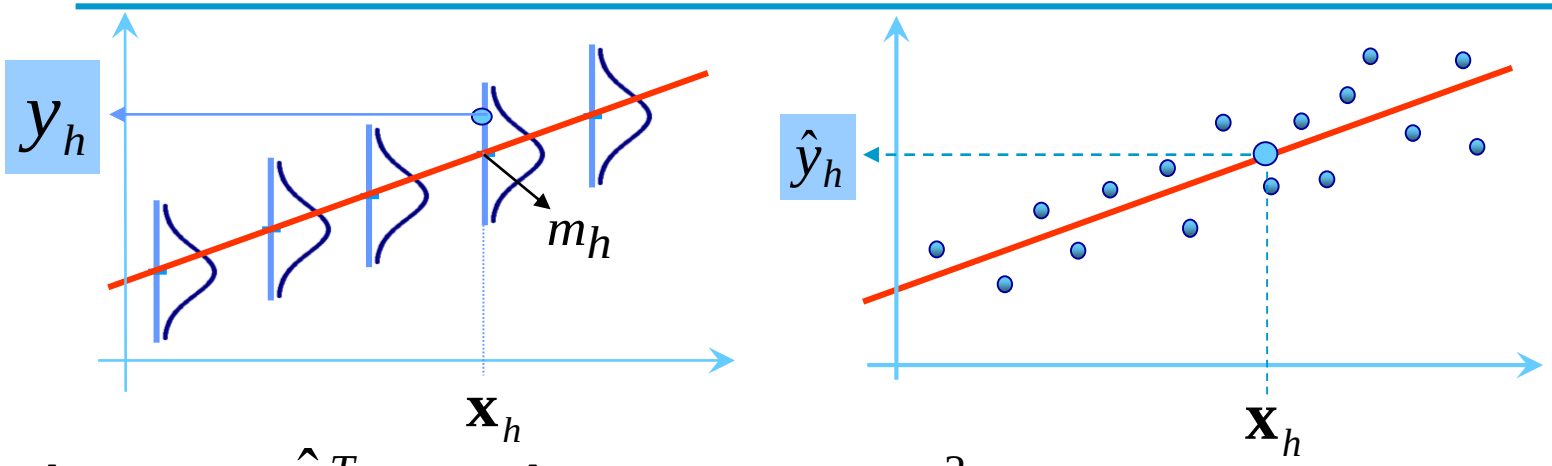
$$E[\tilde{e}_h] = E[y_h] - E[\hat{y}_h] = 0$$

$$\text{var}[\tilde{e}_h] = \text{var}[y_h] + \text{var}[\hat{y}_h]$$

$$= \sigma^2 + \sigma^2 v_{hh}$$

$$\tilde{e}_h \to N(0,\ \sigma^2(1+v_{hh}))$$

Predicción de una nueva observación y_h (Reg. Múltiple)



$$\hat{y}_h = \bar{y} + \hat{\mathbf{b}}^T \mathbf{x}_h \qquad \hat{y}_h \rightarrow N(m_h, \sigma^2 v_{hh})$$

$$\tilde{e}_h = y_h - \hat{y}_h \rightarrow \begin{cases} E[\tilde{e}_h] = E[y_h] - E[\hat{y}_h] = 0 \\ \text{var}[\tilde{e}_h] = \text{var}[y_h] + \text{var}[\hat{y}_h] = \sigma^2(1 + v_{hh}) \end{cases}$$

$$\tilde{e}_h \rightarrow N(\,0,\ \sigma^2(1+v_{hh}))$$

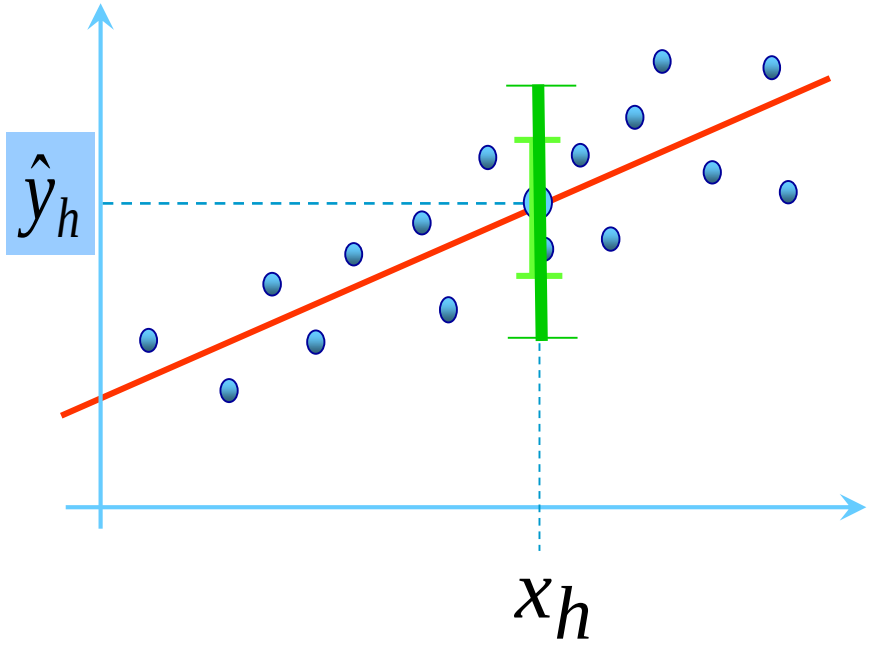
Intervalos de predicción para una nueva observación y_h

$$\tilde{e}_h \to N\left(0,\ \sigma^2(1+v_{hh})\right)$$

$$\tilde{e}_h = y_h - \hat{y}_h$$

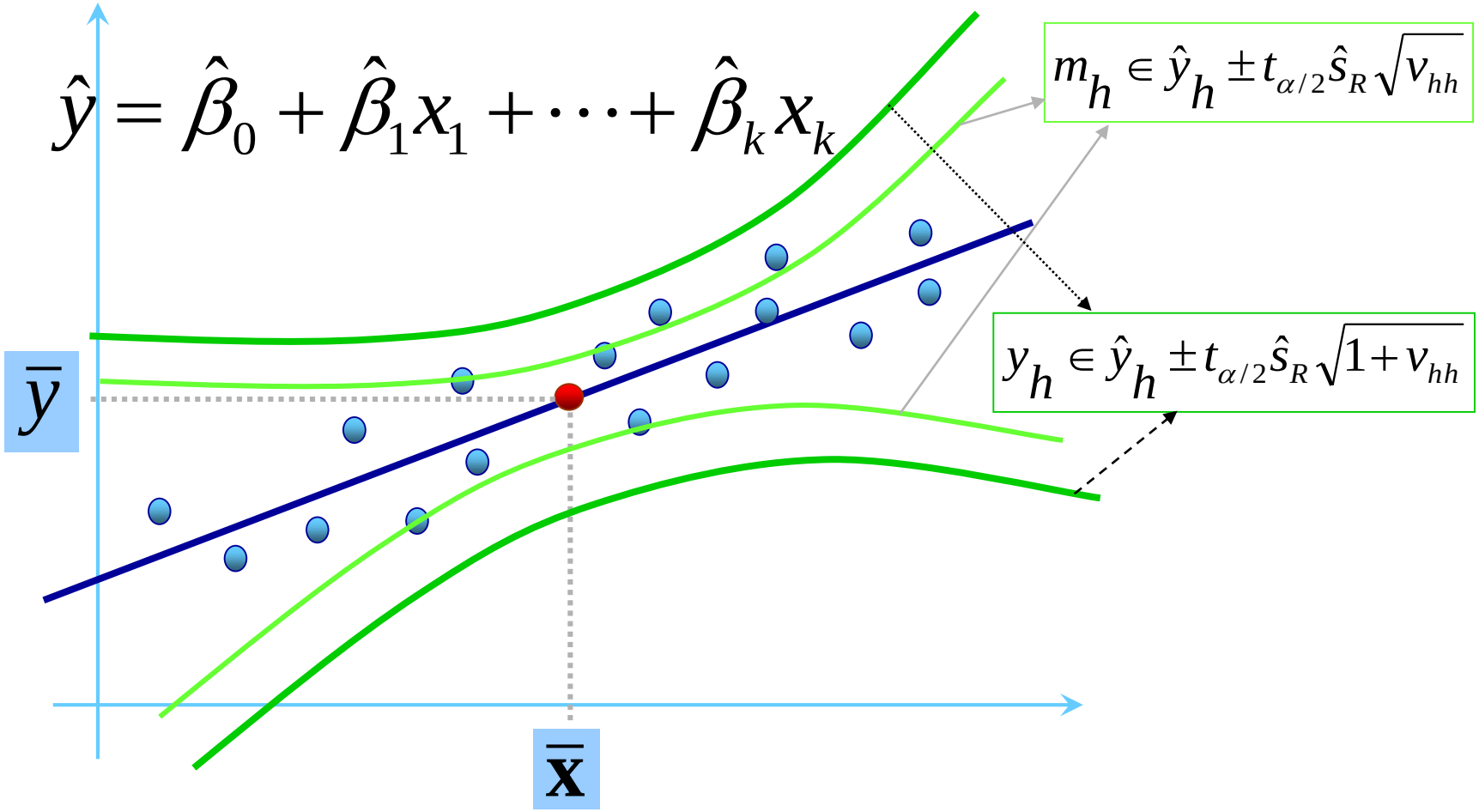
$$\frac{y_h - \hat{y}_h}{\sigma\sqrt{1+v_{hh}}} \to N(0,1)$$

$$\frac{y_h - \hat{y}_h}{\hat{s}_R\sqrt{1+v_{hh}}} \to t_{n-k-1}$$



$$y_h \in \hat{y}_h \pm t_{\alpha/2}\hat{s}_R\sqrt{1+v_{hh}}$$

Límites de predicción



Diagnosis: Residuos

$$\underbrace{\mathbf{Y}}_{\text{Observados}} = \underbrace{\mathbf{X}\hat{\boldsymbol{\beta}}}_{\text{Previstos}} + \underbrace{\mathbf{e}}_{\text{Residuos}}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki})$$

Distribución de los residuos

$$\mathbf{Y} \rightarrow N(\mathbf{X\beta}, \sigma^2\mathbf{I}) \qquad \mathbf{e} = (\mathbf{I} - \mathbf{V})\mathbf{Y}$$

$$\mathbf{V} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$$

$$\begin{cases} \mathbf{e} \rightarrow \text{Normal} \\ E[\mathbf{e}] = (\mathbf{I} - \mathbf{V})E[\mathbf{Y}] = (\mathbf{I} - \mathbf{V})\mathbf{X\beta} = \mathbf{0} \\ \text{var}[\mathbf{e}] = (\mathbf{I} - \mathbf{V})\,\text{var}(\mathbf{Y})(\mathbf{I} - \mathbf{V}) = \sigma^2(\mathbf{I} - \mathbf{V}) \end{cases}$$

$$\mathbf{e} \rightarrow N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{V}))$$

$$e_i \rightarrow N(0, \sigma^2(1 - v_{ii}))$$

Distancia de Mahalanobis

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}_x^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad \text{(Dist. de Mahalanobis)}$$

Mide la distancia de $\mathbf{x_i}$ a $\bar{\mathbf{x}} \Rightarrow \begin{cases} \mathbf{x_i} = \bar{\mathbf{x}} \Rightarrow D_i^2 = 0 \\ \mathbf{x_i} \neq \bar{\mathbf{x}} \Rightarrow D_i^2 > 0 \end{cases}$

$$v_{ii} = \mathbf{x'}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x'}_i = \frac{1}{n}(1 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}_x^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}))$$

v_{ii} son los elementos diagonales de la matriz V

$$\mathbf{V} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

$$v_{ii} = \sum_{j=1}^{n} v_{ij} v_{ji} = \sum_{j=1, j\neq i}^{n} v_{ij}^2 + v_{ii}^2 \Rightarrow v_{ii}(1 - v_{ii}) = \sum_{j=1, j\neq i}^{n} v_{ij}^2 \geq 0 \Rightarrow \frac{1}{n} \leq v_{ii} \leq 1$$

Residuos *estandarizados*

$$e_i \rightarrow N(0,(1-v_{ii})\sigma^2)$$

$$\text{var}(e_i) = (1-v_{ii})\sigma^2$$

Cuando $\mathbf{x}_i$ está próximo a $\mathbf{x} \Rightarrow \quad v_{ii} \approx 1/n \Rightarrow \text{var}(e_i) \approx \sigma^2$

Cuando $\mathbf{x}_i$ está lejos de $\mathbf{x} \Rightarrow \quad v_{ii} \approx 1 \quad \Rightarrow \text{var}(e_i) \approx 0 \Rightarrow e_i \approx 0$

Residuos estandarizados

$$r_i = \frac{e_i}{\hat{s}_R\sqrt{1-v_{ii}}}$$