Regresión lineal: *variables cualitativas, predicción y diagnosis.*

Estadística 2002-2003

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
...	...	...	...	...	...

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

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

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

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

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

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

Variables ficticias

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

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

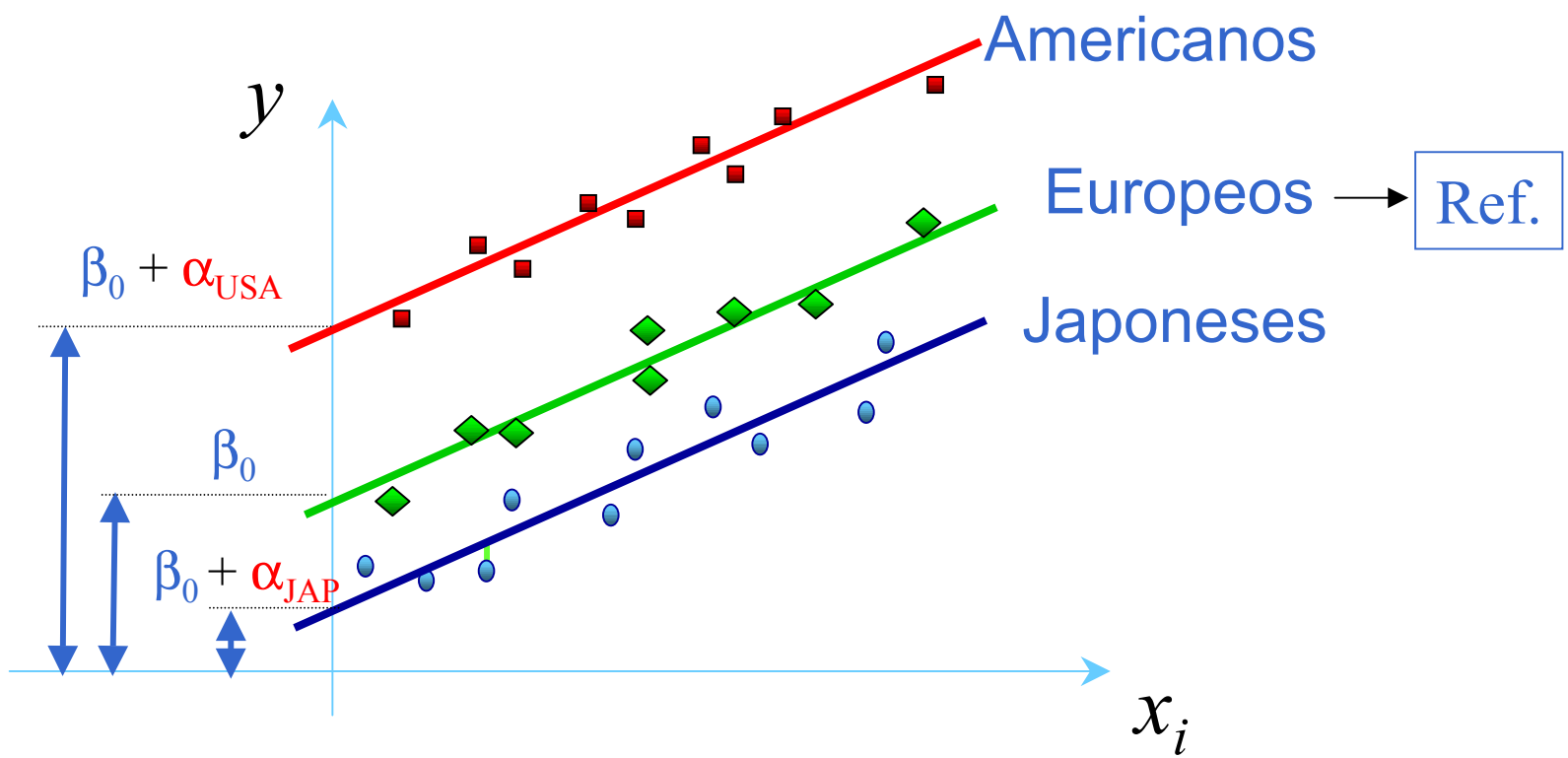
- Coches japoneses: $Z_{JAP} = 1$ y $Z_{USA} = 0$

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

- 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



Multiple Regression Analysis

Dependent variable: consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-1,45504	1,01725	-1,43037	0,1534
cilindrada	0,000322798	0,0001792	1,80133	0,0724
potencia	0,0422677	0,00678898	6,22592	0,0000
peso	0,00559955	0,000965545	5,79937	0,0000
aceleracion	0,110841	0,0496919	2,23057	0,0263
Zjap	-0,361762	0,279049	-1,29641	0,1956
Zusa	0,0611229	0,280236	0,218113	0,8275

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4852,53	6	808,756	293,48	0,0000
Residual	1058,21	384	2,75575		
Total (Corr.)	5910,74	390			

R-squared = 82,0969 percent
R-squared (adjusted for d.f.) = 81,8171 percent
Standard Error of Est. = 1,66005

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 =82.09 de este modelo con el anterior R^2=81.98, se confirma que el modelo con las variables de *Origen* no suponen una mejora sensible.

Modelo de regresión con variables cualitativas

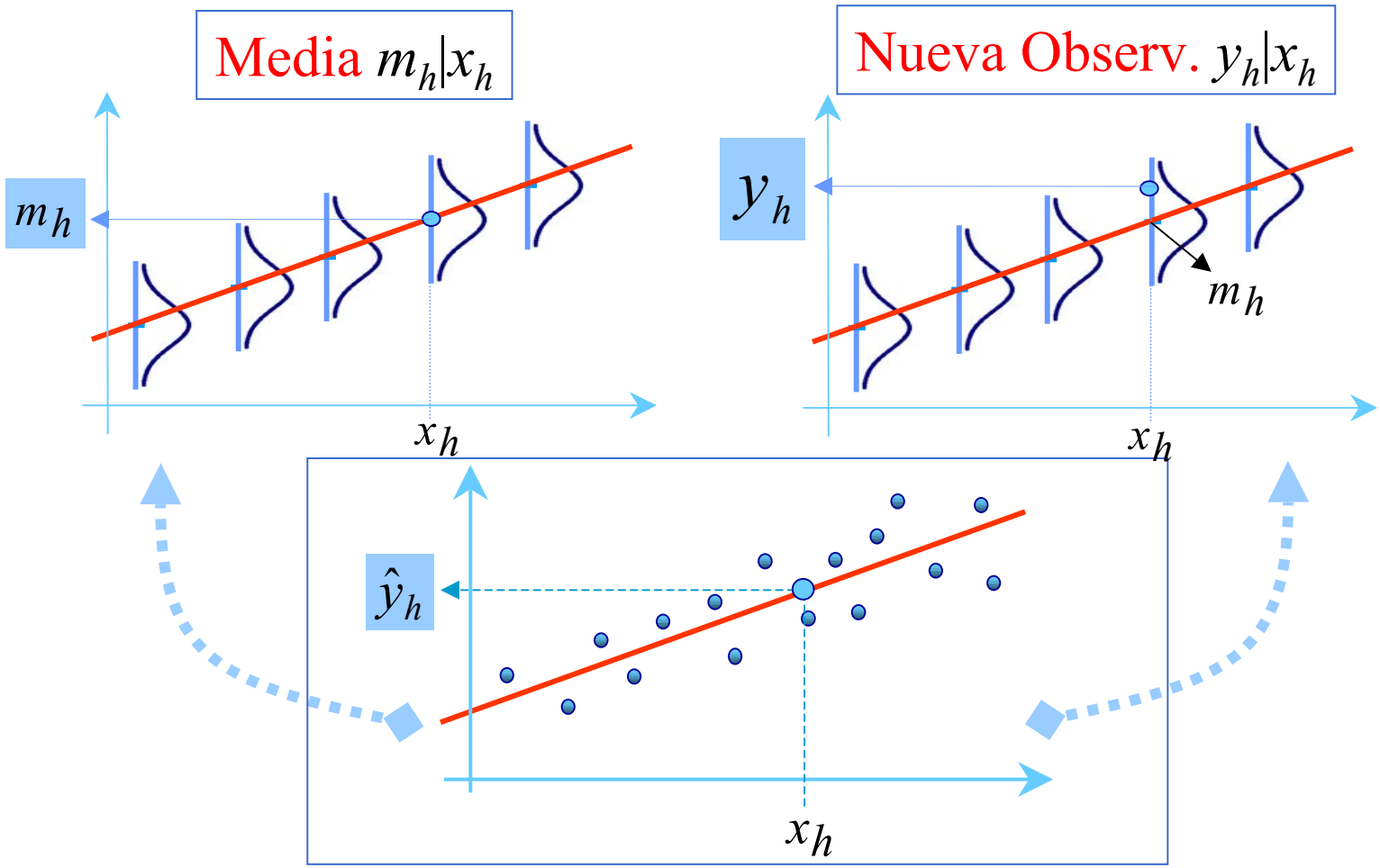
- En general, para considerar una variable cualitativa con r niveles, se introducen en la ecuación r-1 variables ficticias

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

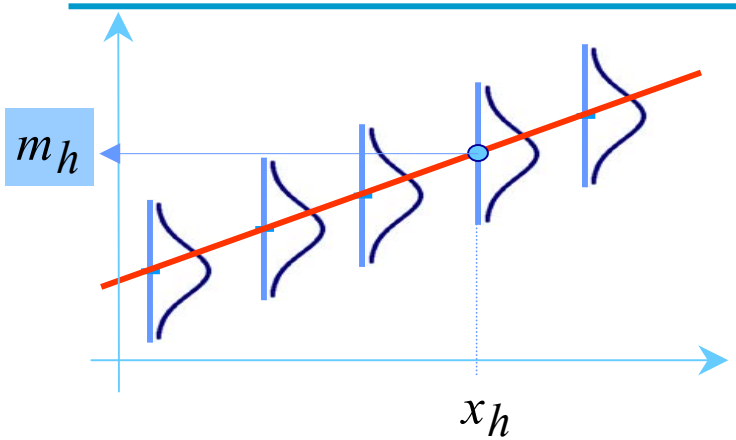
Y el nivel r no utilizado es el que actúa de referencia

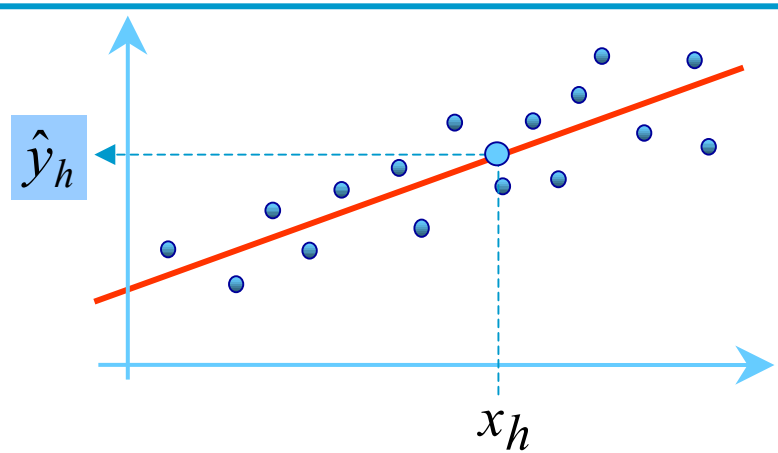
$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \underbrace{+ \alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_{r-1} z_{r-1,i}}_{\text{variable cualitativa}} + u_i$$

Predicción



Predicción de la media m_h (Regresión simple)



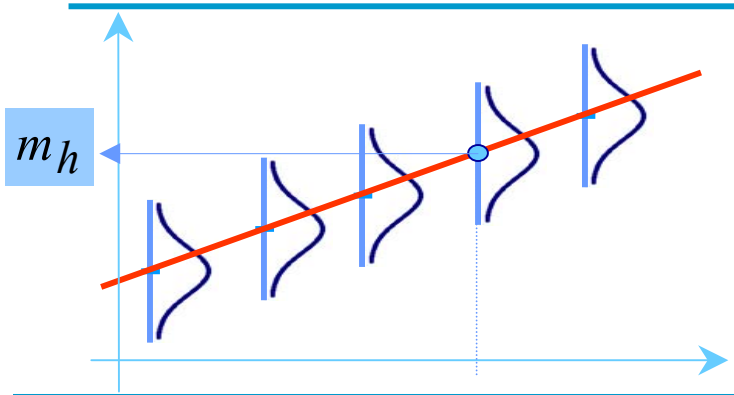


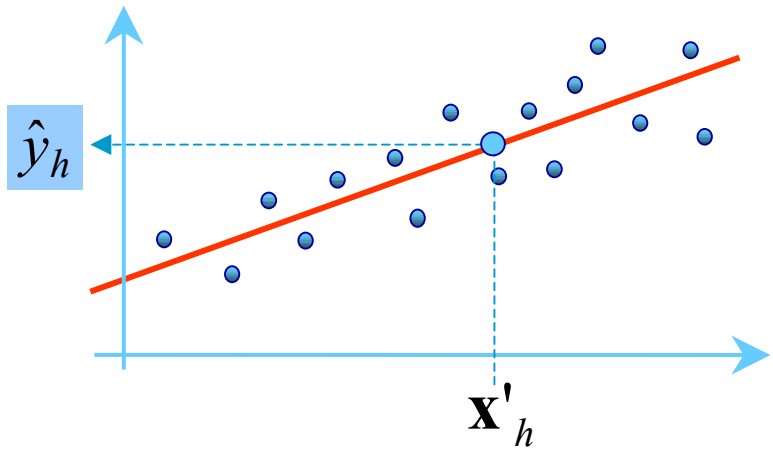
$$y_h \to N(\beta_0 + \beta_1 x_h, \sigma^2)$$
$$m_h = \beta_0 + \beta_1 x_h$$

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

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

Predicción de la media m_h (Regresión múltiple)





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

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

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

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

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

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

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

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

Expresión alternativa para v_{hh}

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

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

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

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

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

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

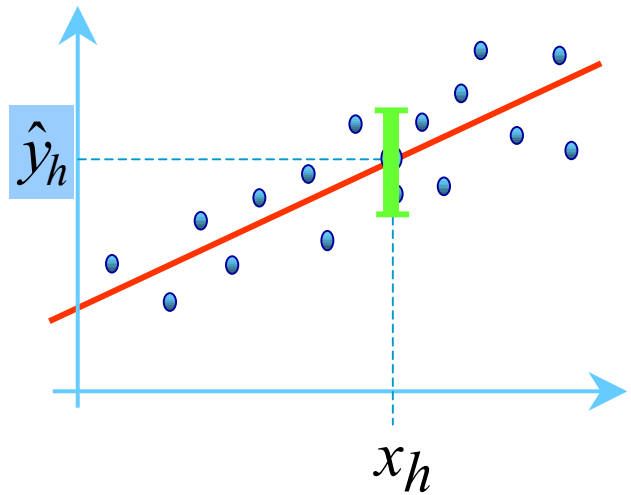
$$\mathbf{x}_h \neq \bar{\mathbf{x}} \Rightarrow v_{hh} > 1/n$$

Intervalos de confianza para la media m_h

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

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

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



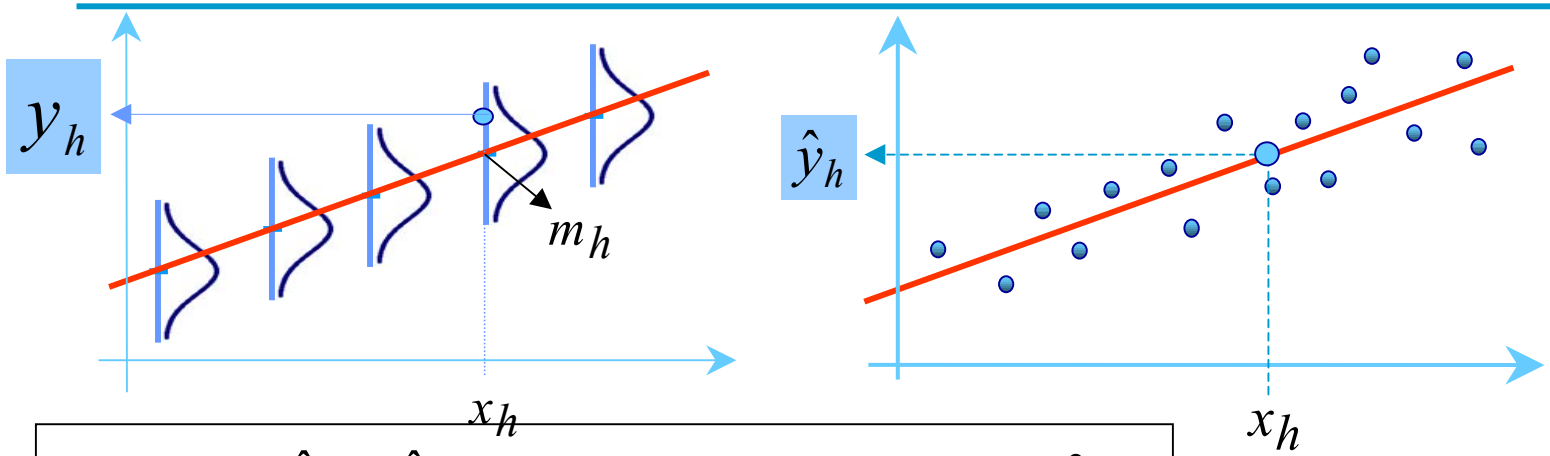
$$m_h \in \hat{y}_h \pm t_{\alpha/2}\hat{s}_R\sqrt{v_{hh}}$$

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)



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

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

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

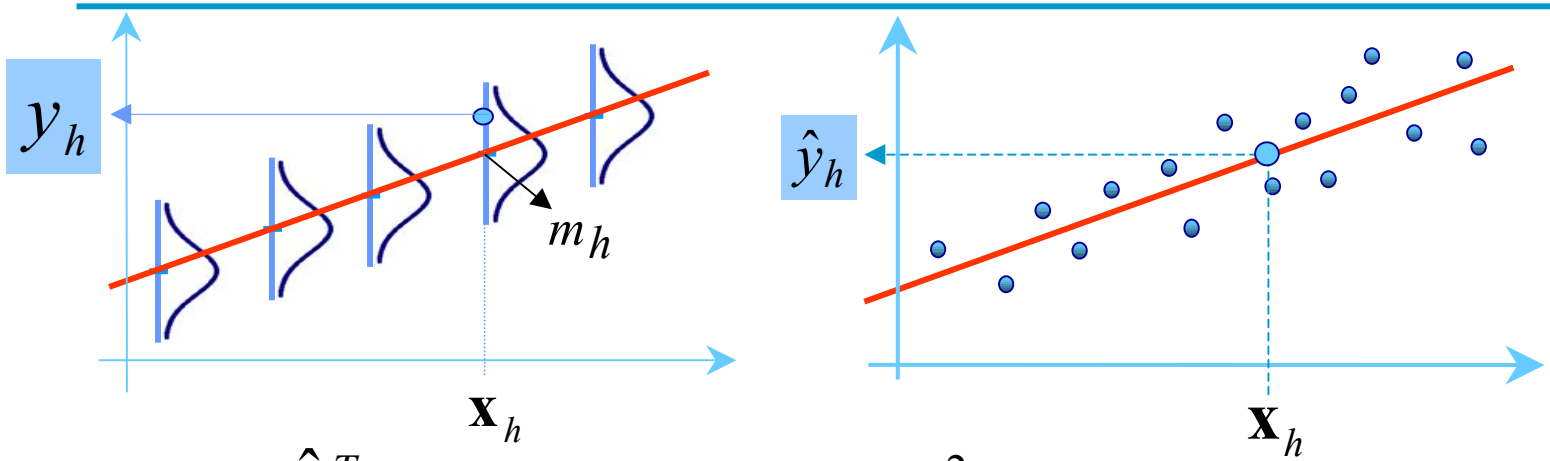
$$E[\tilde{e}_h] = E[y_h] - E[\hat{y}_h] = 0$$

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

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

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

Predicción de una nueva observación y_h (Reg. Múltiple)



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

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

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

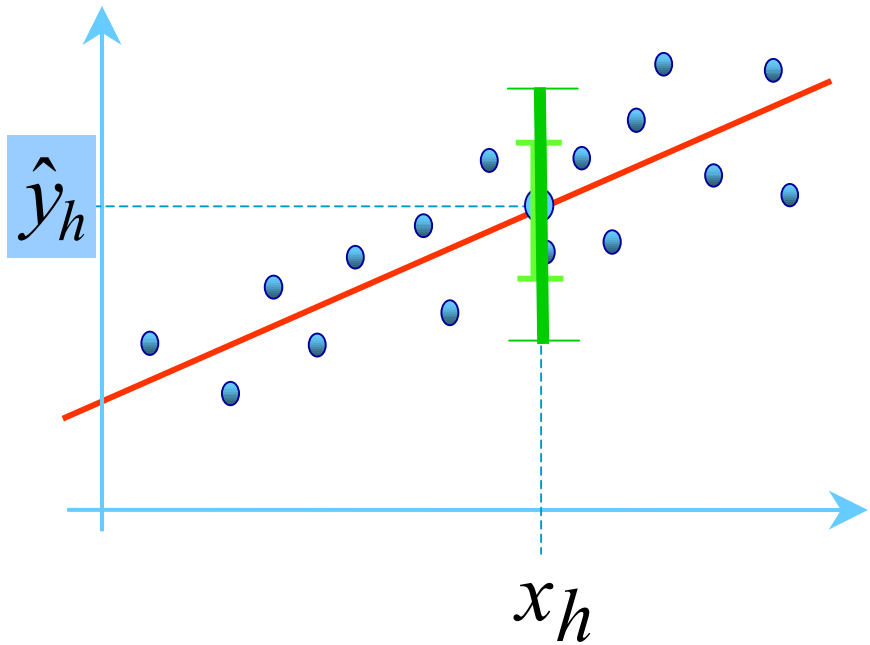
Intervalos de predicción para una nueva observación y_h

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

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

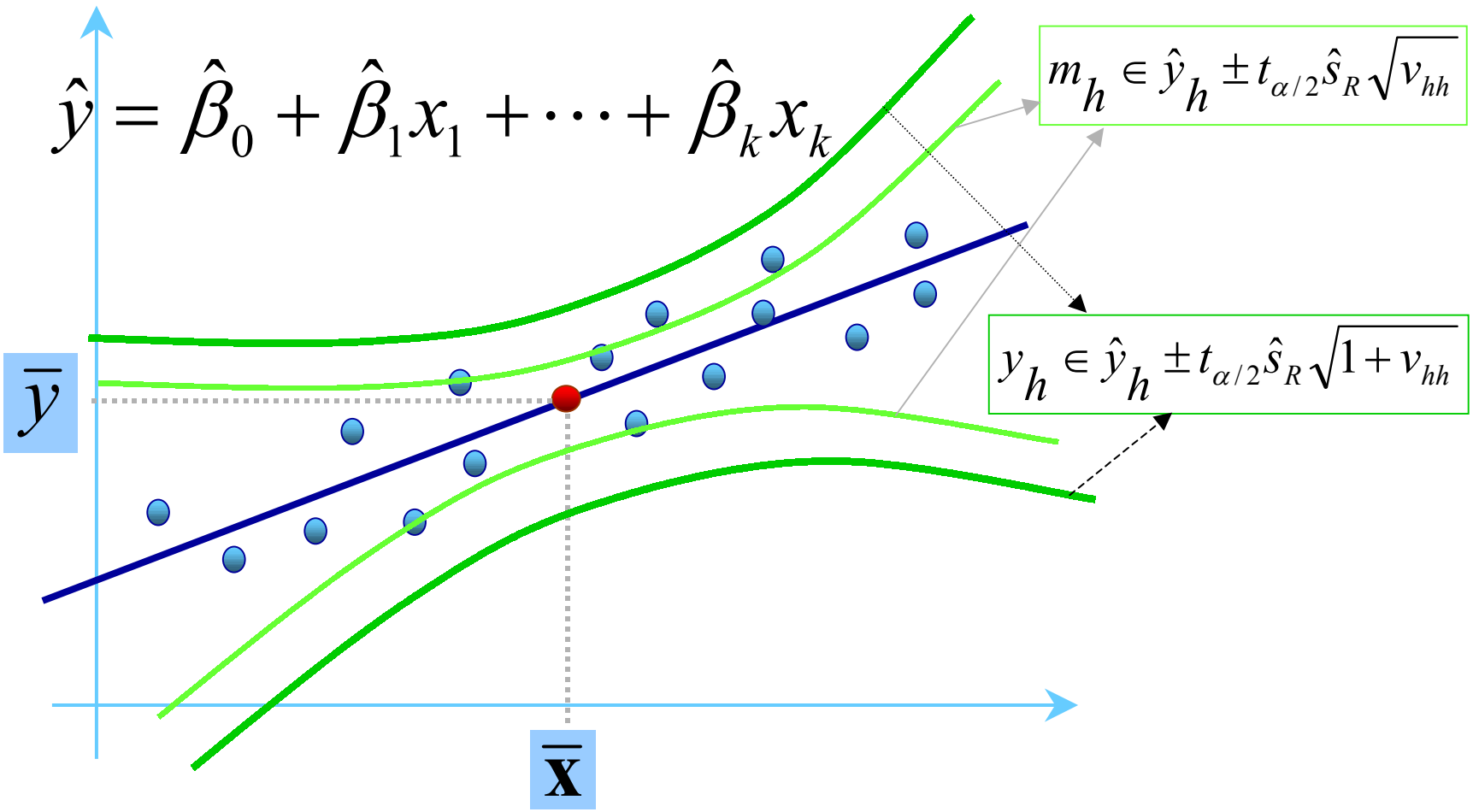
$$\frac{y_h - \hat{y}_h}{\sigma\sqrt{1+v_{hh}}} \to N(0,1)$$

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



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

Límites de predicción



Diagnosis: Residuos

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

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

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

Distribución de los residuos

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

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

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

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

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

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 *estudentizados*

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

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

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

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

Residuos estudentizados

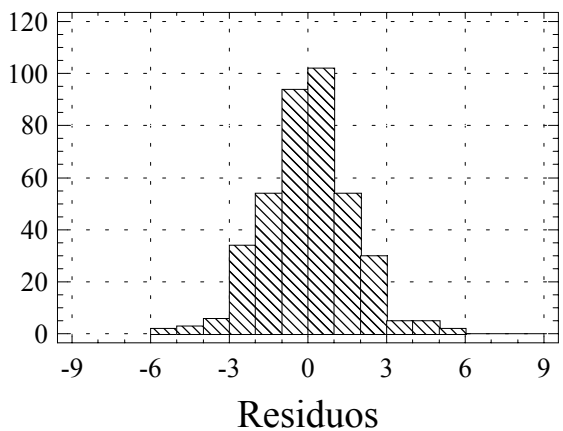
$$r_i = \frac{e_i}{\hat{s}_R\sqrt{1-v_{ii}}}$$

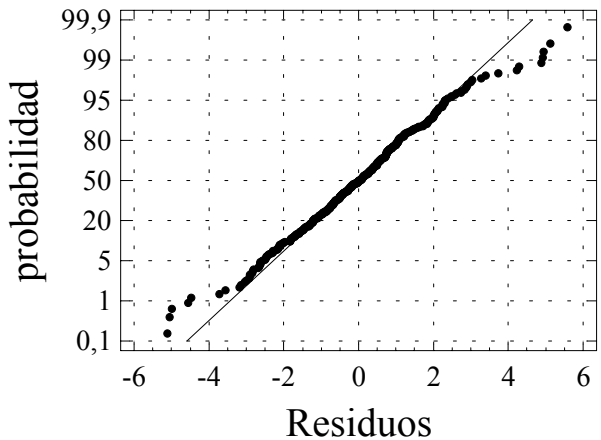
Hipótesis de normalidad

Herramientas de comprobación:

- Histograma de residuos
- Gráfico de probabilidad normal (Q-Q plot)
- Contrastes formales (Kolmogorov-Smirnov)

Ejemplo de coches





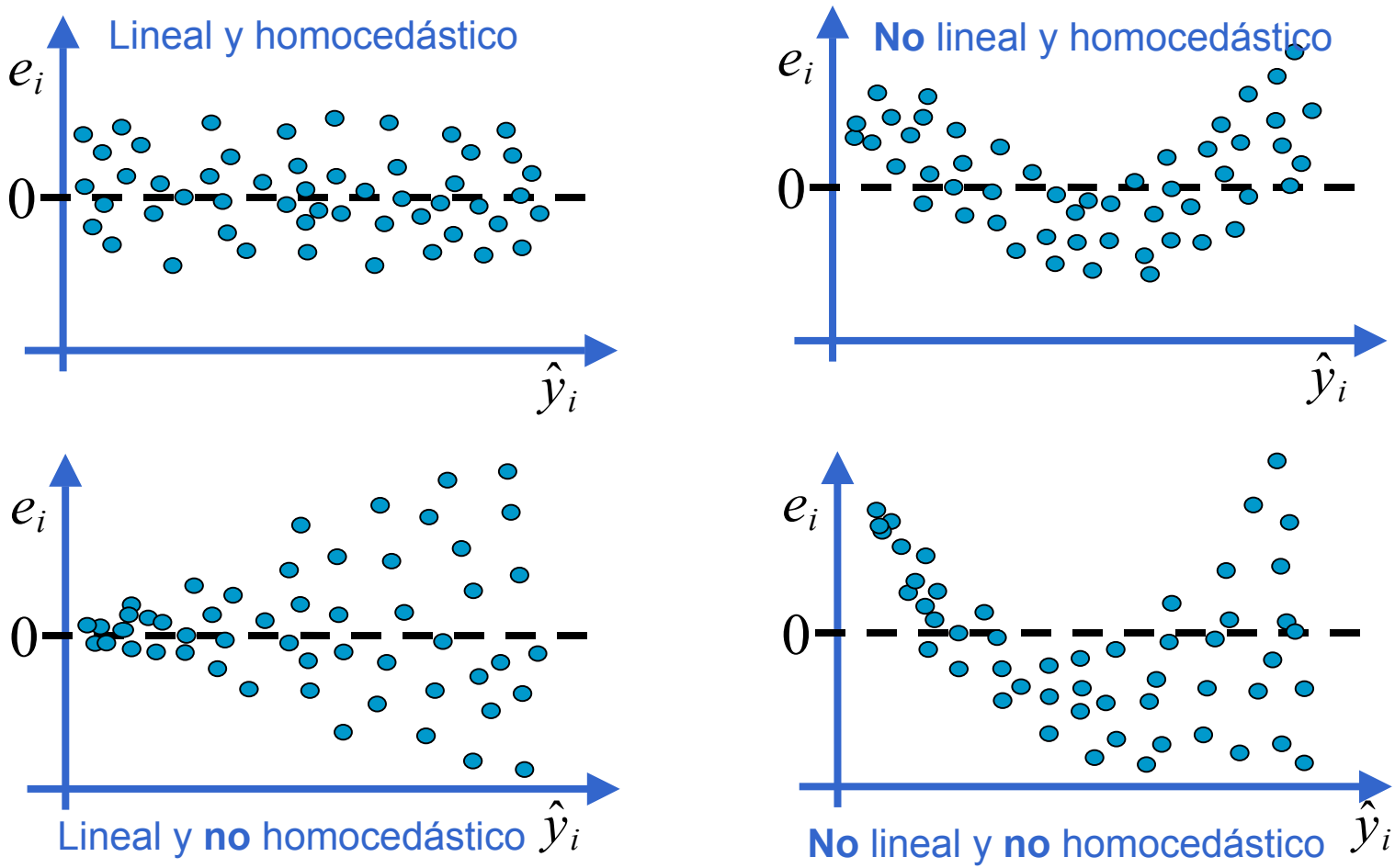
Comprobación de la linealidad y homocedasticidad

- Ambas hipótesis se comprueban conjuntamente mediante **gráficos de los residuos**
  - Frente a valores previstos
  - Frente a cada regresor.
- En muchas ocasiones se corrige la falta de linealidad y la heterocedasticidad mediante transformación de las variables.

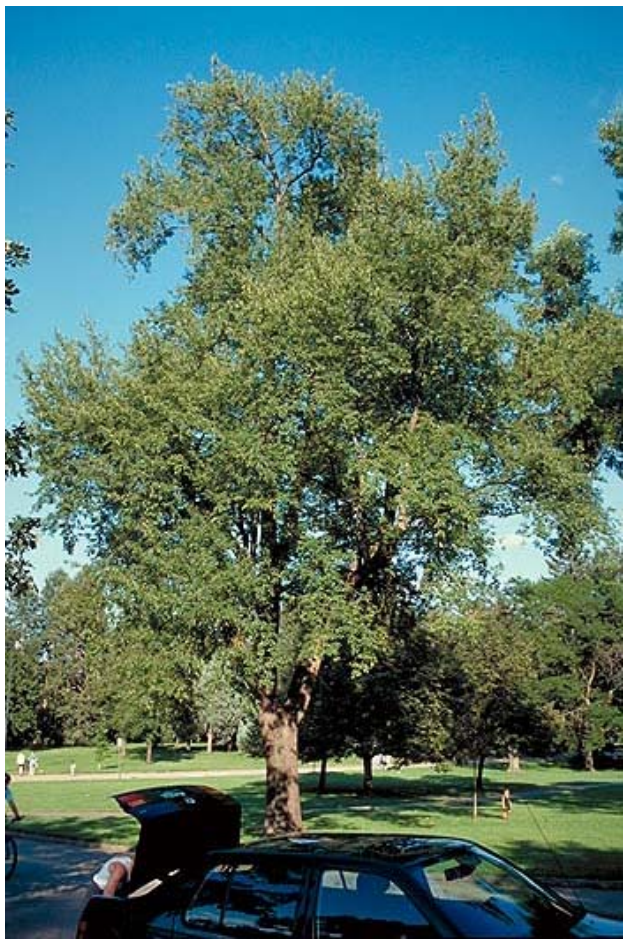
$$\log y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$$

$$\log y_i = \beta_0 + \beta_1 \log x_{1i} + \cdots + \beta_k \log x_{ki} + u_i$$

Residuos - Valores previstos



Ejemplo 1: Cerezos Negros



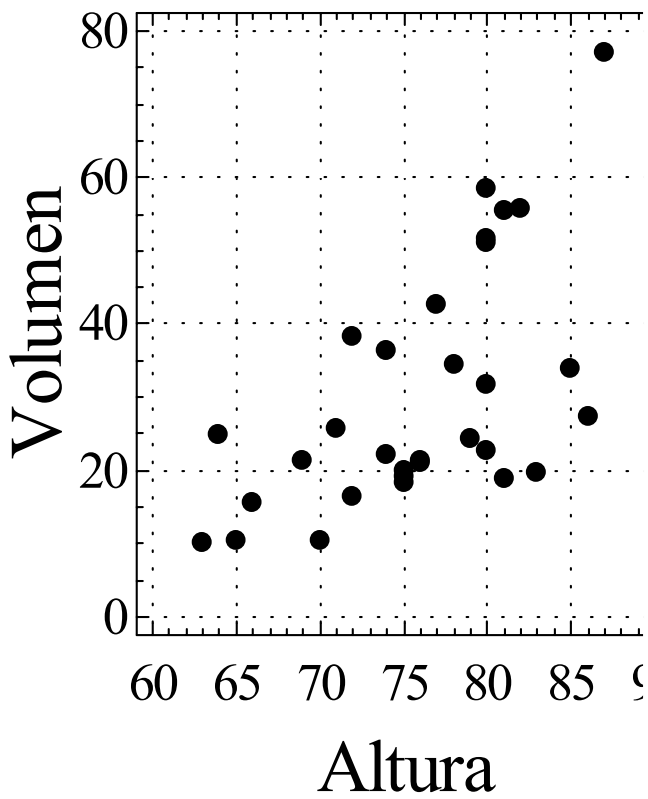
Se desea construir un modelo de regresión para obtener el **volumen** de madera de un "*cerezo negro*" en función de la altura del tronco y del diámetro del mismo a un metro sobre el suelo. Se ha tomado una muestra de 31 árboles. Las unidades de longitudes son *pies* y de volumen *pies cúbicos*.

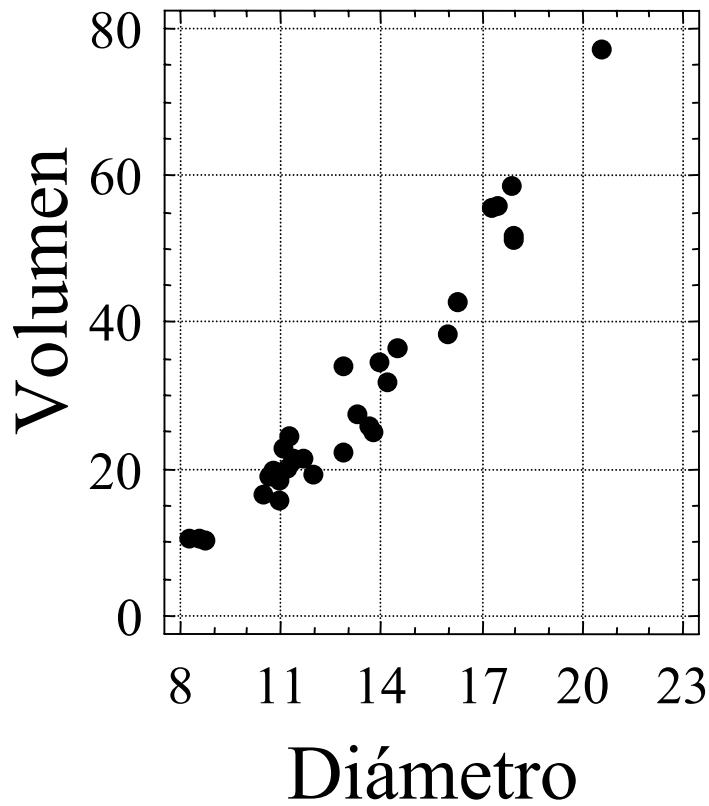
Cerezos negros: Datos

Árbol	Diámetro	Altura	Volumen
1	8,3	70	10,30
2	8,6	65	10,30
3	8,8	63	10,20
4	10,5	72	16,40
5	10,7	81	18,80
6	10,8	83	19,70
7	11,0	66	15,60
8	11,0	75	18,20
9	11,1	80	22,60
10	11,2	75	19,90
11	11,3	79	24,20
12	11,4	76	21,00
13	11,4	76	21,40
14	11,7	69	21,30
15	12,0	75	19,10
16	12,9	74	22,20

Árbol	Diámetro	Altura	Volumen
17	12,9	85	33,80
18	13,3	86	27,40
19	13,7	71	25,70
20	13,8	64	24,90
21	14,0	78	34,50
22	14,2	80	31,70
23	14,5	74	36,30
24	16,0	72	38,30
25	16,3	77	42,60
26	17,3	81	55,40
27	17,5	82	55,70
28	17,9	80	58,30
29	18,0	80	51,50
30	18,0	80	51,00
31	20,6	87	77,00

Gráficos x-y


Volumen
80
60
40
20
0
60
65
70
75
80
85
Altura


Volumen
80
60
40
20
0
8
11
14
17
20
23
Diámetro

Primer modelo:cerezos negros

$$\text{Volumen} = \hat{a}_0 + \hat{a}_1\ \text{Diámetro} + \hat{a}_2\ \text{Altura} + \text{Error}$$

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-57,9877	8,63823	-6,71291	0,0000
Altura	0,339251	0,130151	2,60659	0,0145
Diametro	4,70816	0,264265	17,8161	0,0000

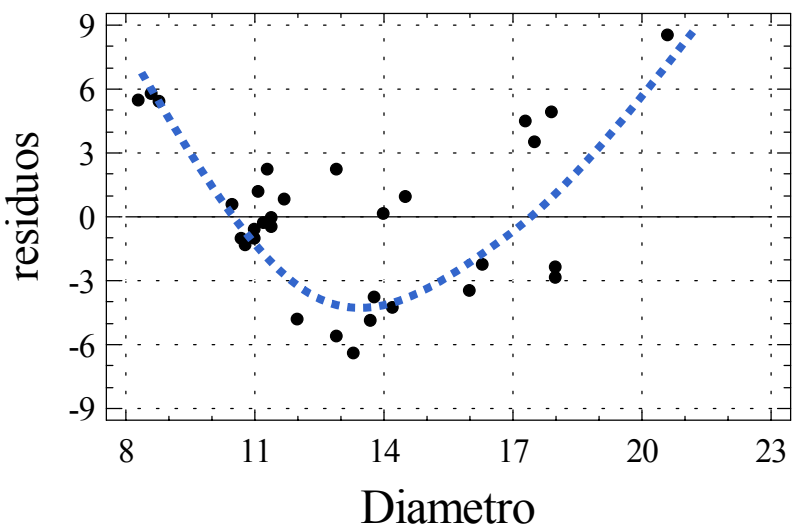
Analysis of Variance

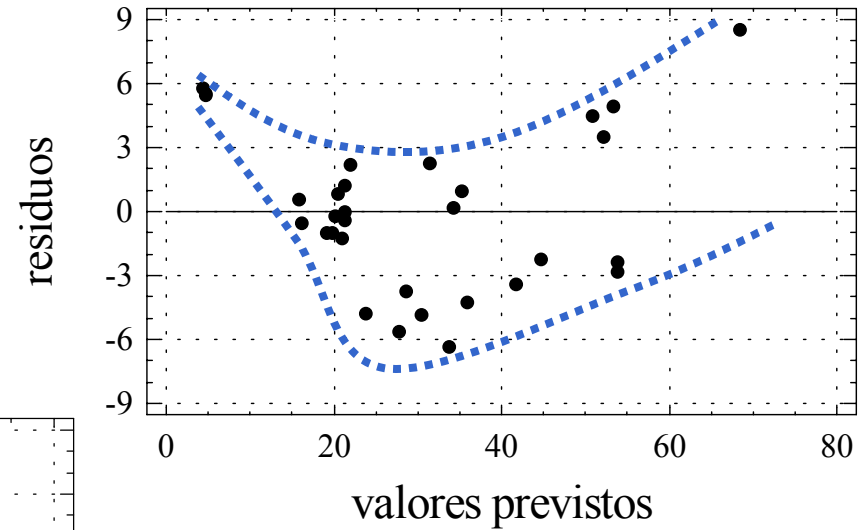
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	7684,16	2	3842,08	254,97	0,0000
Residual	421,921	28	15,0686		
Total (Corr.)	8106,08	30			

R-squared = 94,795 percent
R-squared (adjusted for d.f.) = 94,4232 percent

Diagnosis

Falta de linealidad





Falta de homocedasticidad

Transformación

$$\text{vol} \approx \text{k} \times \text{altura} \times \text{diámetro}^2$$

$$\log(\text{vol}) \approx \beta_0 + \beta_1 \log(\text{altura}) + \beta_2 \log(\text{diámetro}) + \text{error}$$

Dependent variable: log(Volumen)

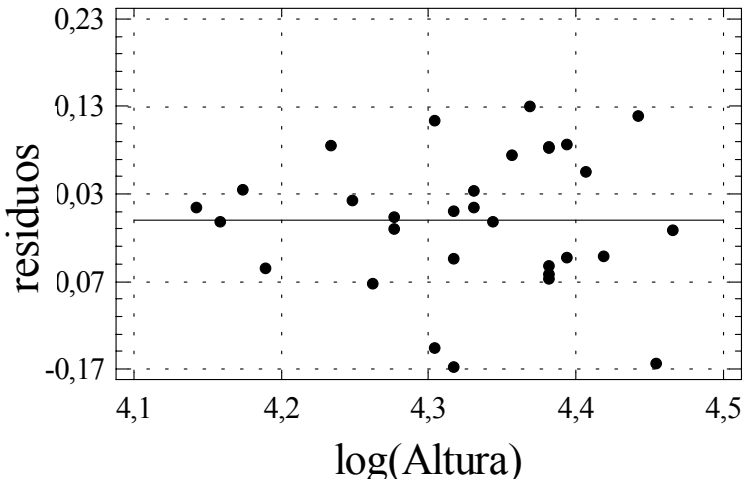
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-6,63162	0,79979	-8,2917	0,0000
log(Altura)	1,11712	0,204437	5,46439	0,0000
log(Diametro)	1,98265	0,0750106	26,4316	0,0000

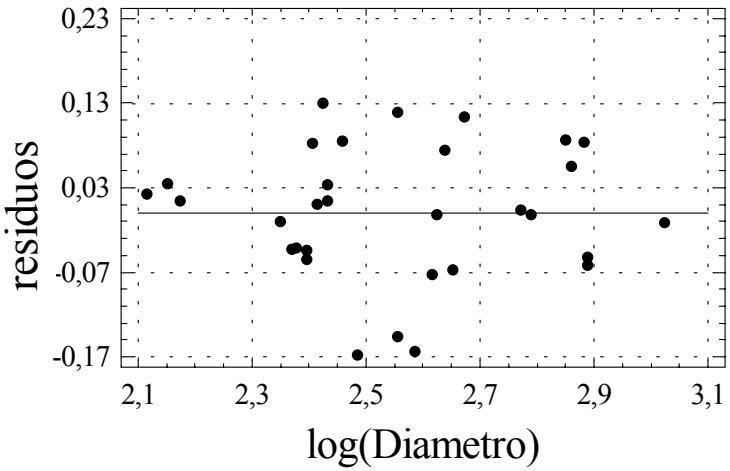
Analysis of Variance

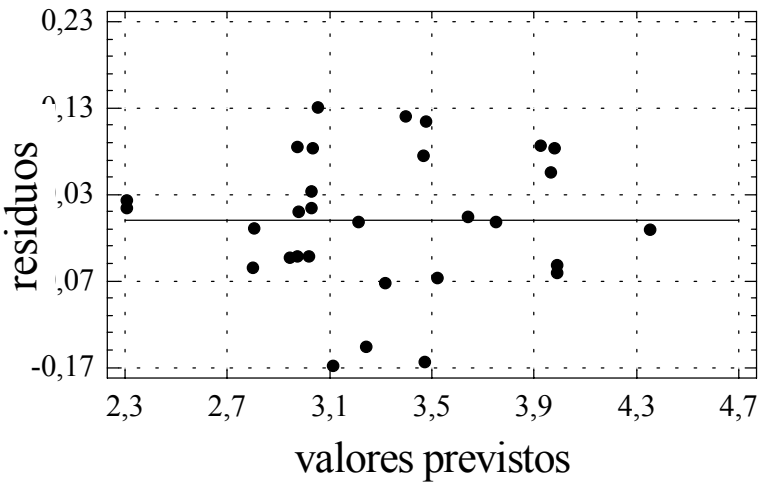
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	8,12323	2	4,06161	613,19	0,0000
Residual	0,185463	28	0,00662369		
Total (Corr.)	8,30869	30			

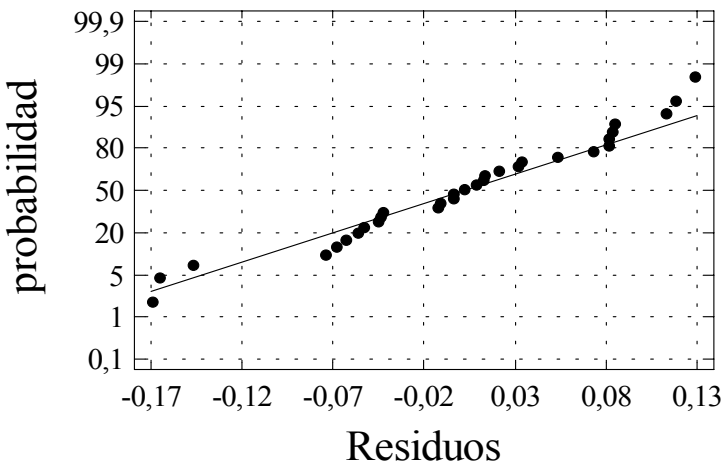
R-squared = 97,7678 percent
R-squared (adjusted for d.f.) = 97,6084 percent

Diagnosis (modelo transformado)









Interpretación

- Se comprueba gráficamente que la distribución de los residuos es compatible con las hipótesis de normalidad y homocedasticidad.
- El volumen está muy relacionada con la altura y el diámetro del árbol (R^2= 97.8%)
- El modelo estimado

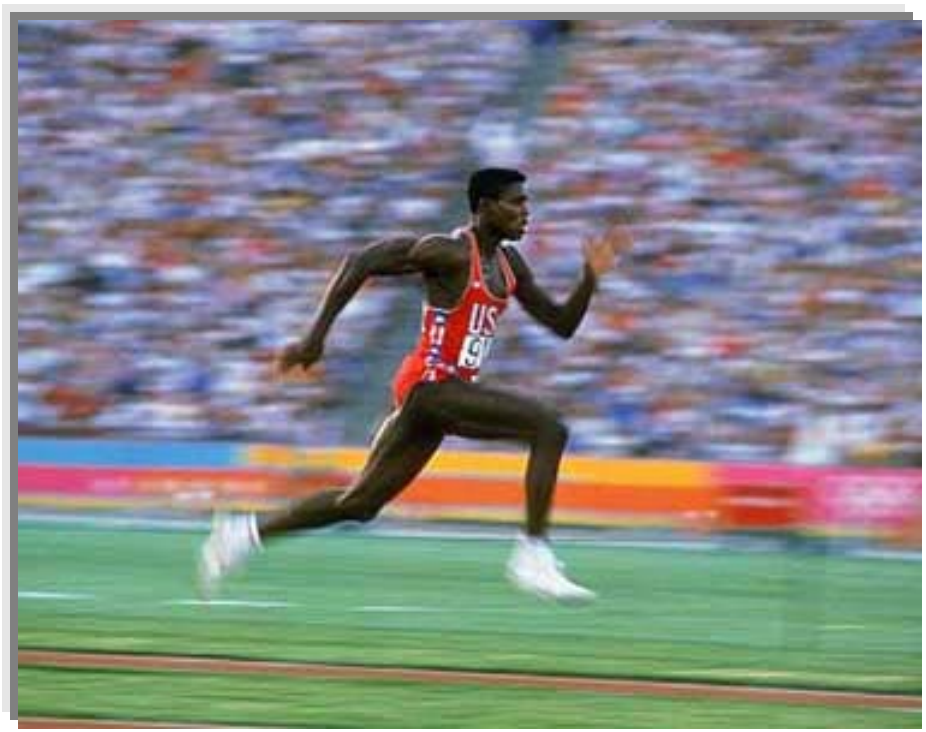
$$\log(Vol) = -6.6 + 1.12 \log(Alt) + 1.98 \log(Diam.) + Error$$

es compatible con la ecuación $vol = k \times Alt \times Diam^2$

- La varianza residual es 0.006623, es decir s_R=0.081 que indica que el error relativo del modelo en la predicción del volumen es del 8.1%.

Datos olímpicos

Tiempos de los campeones olímpicos en 200m, 400m, 800m y 1500m.



Se pretende construir un modelo de regresión con dos objetivos:

- Medir la evolución de estas marcas con el tiempo.
- Hacer una predicción del resultado en unas *futuras* olimpiadas.

Ejemplo: Carreras olímpicas

Ciudad	Altitud	Año	200 m	400 m	800 m	1500 m
París	79	1900	22,20	49,40	121,40	246,00
San Luis	138	1904	21,60	49,20	116,00	245,40
Londres	15	1908	22,40	50,00	112,80	243,40
Estocolmo	15	1912	21,70	48,20	111,90	236,80
Amberes	4	1920	22,00	49,60	113,40	241,80
París	79	1924	21,60	47,60	112,40	233,60
Amsterdan	-2	1928	21,80	47,80	111,80	233,20
Los Ángeles	100	1932	21,20	46,20	109,80	231,20
Berlín	50	1936	20,70	46,50	112,90	227,80
Londres	15	1948	21,10	46,20	109,20	225,20
Helsinki	25	1952	20,70	45,90	109,20	225,20
Melbourne	115	1956	20,60	46,70	107,70	221,20
Roma	15	1960	20,50	44,90	106,30	215,60
Tokyo	14	1964	20,30	45,10	105,10	218,10
Mexico	2220	1968	19,83	43,80	104,30	214,90
Munich	458	1972	20,00	44,66	105,90	216,30
Montreal	53	1976	20,23	44,26	103,50	219,20
Moscú	150	1980	20,19	44,60	105,40	218,40
Los Ángeles	100	1984	19,80	44,27	104,00	212,53
Seúl	34	1988	19,75	43,87	103,45	215,96
Barcelona	0	1992	20,01	43,50	103,66	220,12
Atlanta	320	1996	19,32	43,49	102,58	215,78

Tiempo = β_0 + β_1 Año + β_2 Distancia + Error

Dependent variable: Tiempo

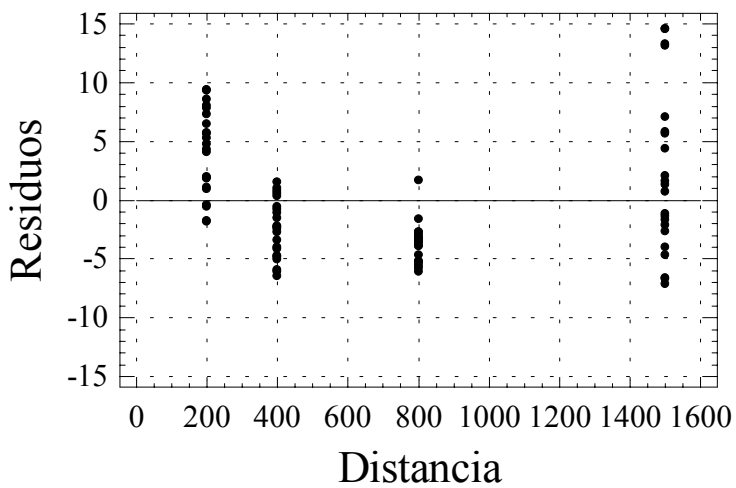
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	268,485	36,8179	7,29222	0,0000
Año	-0,145478	0,0188741	-7,70784	0,0000
Distancia	0,159578	0,00113405	140,715	0,0000

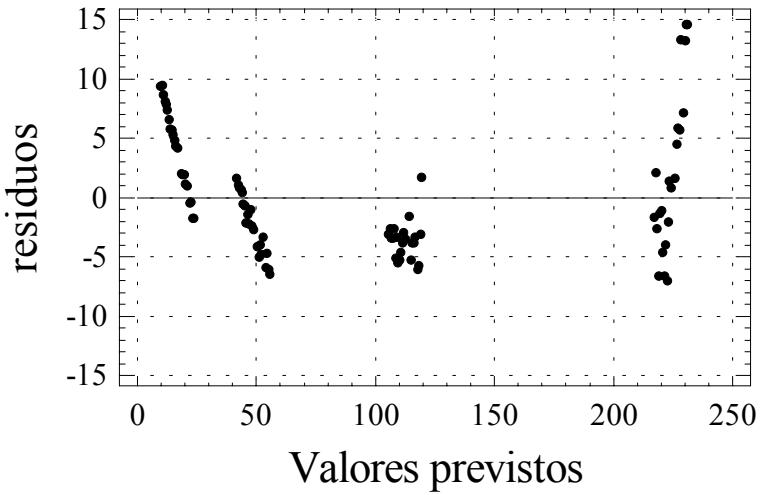
Analysis of Variance

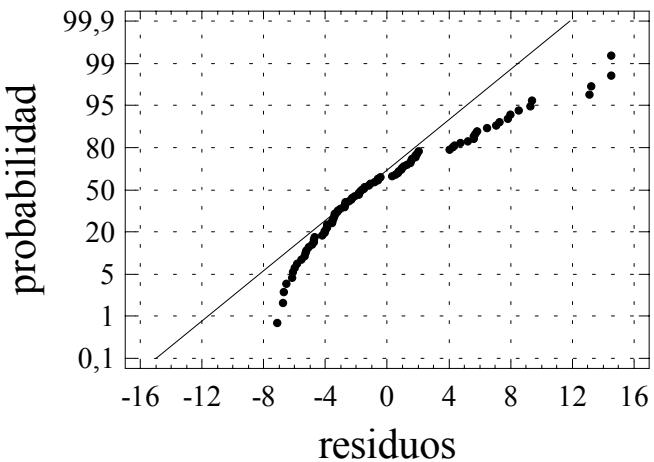
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	554892,0	2	277446,0	9930,11	0,0000
Residual	2374,89	85	27,9399		
Total (Corr.)	557267,0	87			

R-squared = 99,5738 percent
R-squared (adjusted for d.f.) = 99,5638 percent

Diagnosis


Residuos
15
10
5
0
-5
-10
-15
0
200
400
600
800
1000
1200
1400
1600
Distancia


residuos
15
10
5
0
-5
-10
-15
0
50
100
150
200
250
Valores previstos


probabilidad
99,9
99
95
80
50
20
5
1
0,1
-16
-12
-8
-4
0
4
8
12
16
residuos

Interpretación

- Los gráficos de los residuos con la distancia y con los valores previstos muestran falta de linealidad y heterocedasticidad (leve)
- El gráfico Q-Q muestra falta de normalidad
- La transformación 1/Tiempo puede servir para corregir el problema de heterocedasticidad. En este caso es más útil modelar la velocidad

$$\text{Velocidad}_i = \text{Distancia}_i / \text{Tiempo}_i$$

Velocidad = β_0 + β_1 Año + β_2 Dist. + Error

Dependent variable: Velocidad

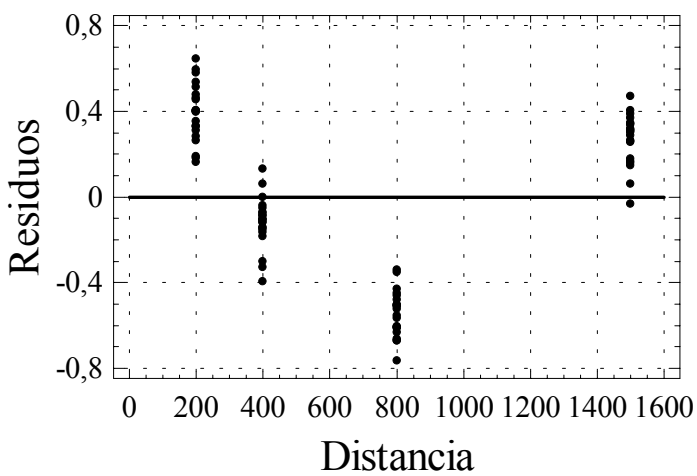
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-12,2153	2,73592	-4,46478	0,0000
Año	0,0112286	0,00140252	8,00603	0,0000
Distancia	-0,00220474	0,0000842706	-26,1627	0,0000

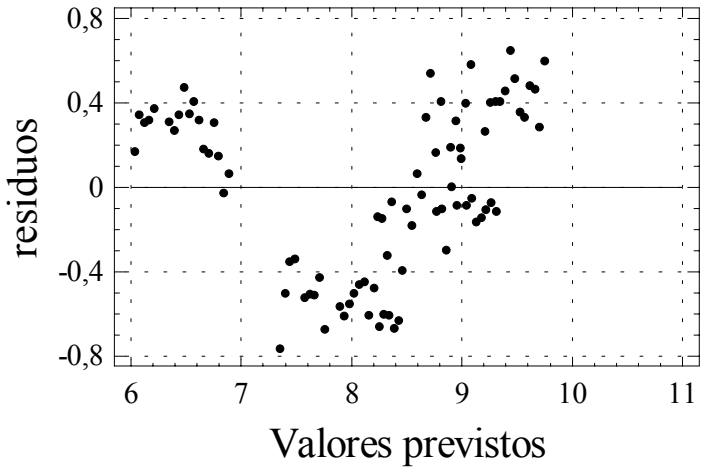
Analysis of Variance

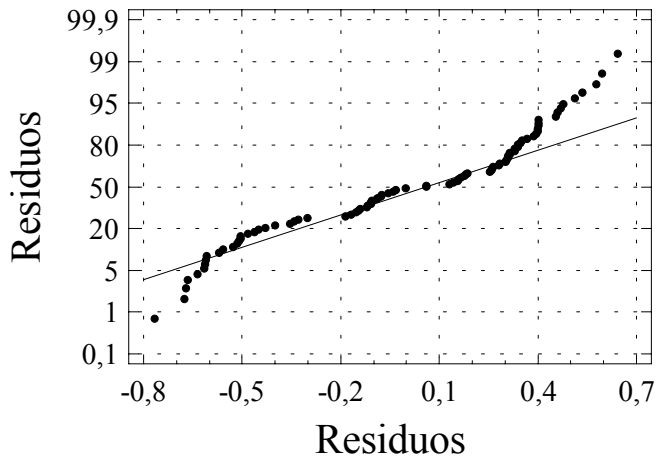
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	115,492	2	57,7459	374,29	0,0000
Residual	13,1139	85	0,154281		
Total (Corr.)	128,606	87			

R-squared = 89,803 percent
R-squared (adjusted for d.f.) = 89,5631 percent

Diagnosis


Residuos
0,8
0,4
0
-0,4
-0,8
0
200
400
600
800
1000
1200
1400
1600
Distancia


residuos
0,8
0,4
0
-0,4
-0,8
6
7
8
9
10
11
Valores previstos


Residuos
99,9
99
95
80
50
20
5
1
0,1
-0,8
-0,5
-0,2
0,1
0,4
0,7
Residuos

Velocidad = β_0 + β_1 Año + β_2 Dist. + β_3 Dist.^{2} + Error

Dependent variable: Velocidad

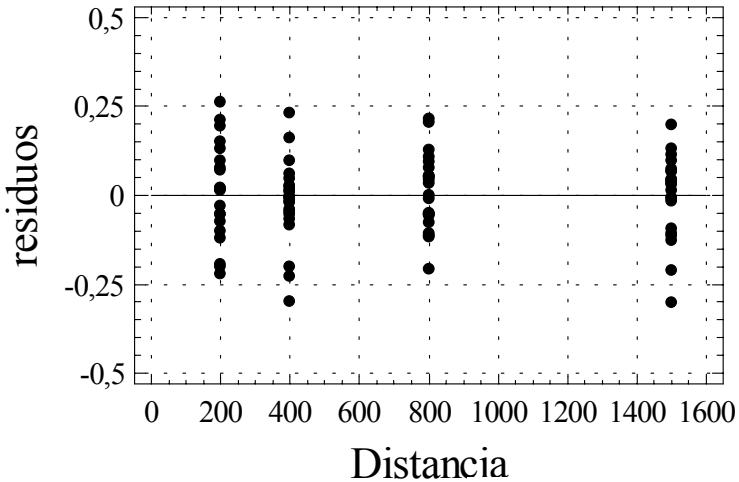
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-11,1792	0,834388	-13,3981	0,0000
Año	0,0112286	0,000427338	26,2758	0,0000
Distancia	-0,00588973	0,000130341	-45,1873	0,0000
Distancia^2	0,0000021172	7,34191E-8	28,8371	0,0000

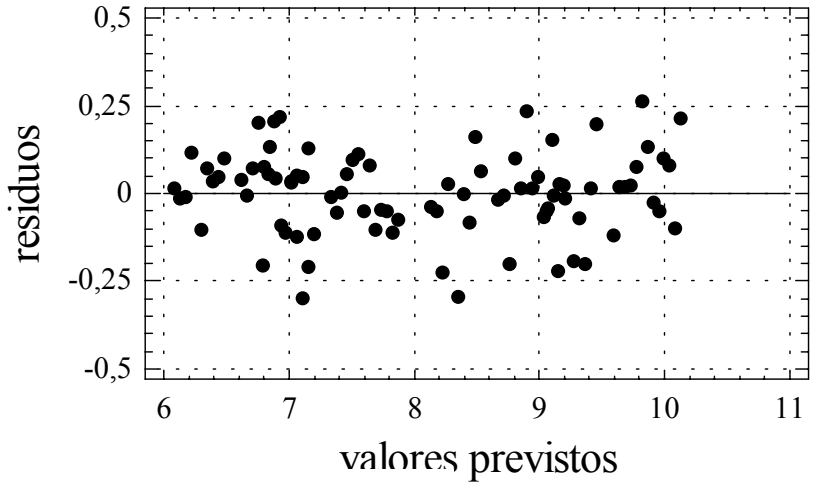
Analysis of Variance

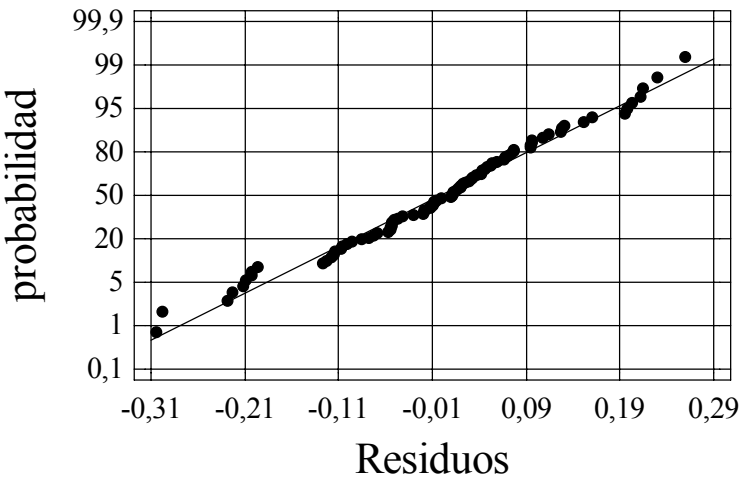
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	127,403	3	42,4675	2964,98	0,0000
Residual	1,20314	84	0,014323		
Total (Corr.)	128,606	87			

R-squared = 99,0645 percent
R-squared (adjusted for d.f.) = 99,0311 percent

Diagnosis


residuos
0,5
0,25
0
-0,25
-0,5
0
200
400
600
800
1000
1200
1400
1600
Distancia


residuos
0,5
0,25
0
-0,25
-0,5
6
7
8
9
10
11
valores previstos


probabilidad
99,9
99
95
80
50
20
5
1
0,1
-0,31
-0,21
-0,11
-0,01
0,09
0,19
0,29
Residuos

Interpretación

- El modelo cumple las condiciones de normalidad y homocedasticidad.
- El coeficiente de determinación R^2=99% da una medida de la bondad de ajuste del modelo.
- El coeficiente positivo del AÑO indica que conforme pasan los años se aumenta la velocidad (se mejoran las marcas).
- El término dominante de la variable DISTANCIA tiene coeficiente negativo que indica que la velocidad media disminuye al aumentar la distancia de la prueba.
- Se mejora ligeramente el modelo con una nueva variable ALTITUD de la ciudad donde se desarrolla las olimpiadas.

Vel. = β_0+β_1 Año+β_2 Dist. + β_3 Dist.^{2} + log(Alt)+Error

Dependent variable: Velocidad

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-10,6966	0,807542	-13,2459	0,0000
Año	0,0109342	0,000416677	26,2413	0,0000
Distancia	-0,00588973	0,000123874	-47,5461	0,0000
Distancia^2	0,0000021172	6,97766E-8	30,3425	0,0000
log(Altitud+3)	0,0237773	0,00751947	3,1621	0,0022

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	127,532	4	31,883	2464,46	0,0000
Residual	1,07378	83	0,0129371		
Total (Corr.)	128,606	87			

R-squared = 99,1651 percent
R-squared (adjusted for d.f.) = 99,1248 percent

Predicción Sydney 2000

Predicción para Velocidad - AÑO 2000 - SYDNEY

Row	Fitted Value	Stnd. Error for Forecast	Lower 95,0% CL for Forecast	Upper 95,0% CL for Forecast
200 m	10,1114	0,119833	9,87302	10,3497
400 m	9,18748	0,118783	8,95123	9,42374
800 m	7,84784	0,119901	7,60937	8,08632
1500 m	7,13371	0,120308	6,89442	7,373

Predicción del tiempo (segundos) y resultados Sydney 2000

Distancia	Intervalo de predicción (95%) Lím. Inf.	Lím. Sup.	Predicción	Resultado Sydney 2000	Error Absoluto	Error Relativo
200 m	19,32	20,26	19,78	20,09	0,31	2%
400 m	42,44	44,69	43,538	43,84	0,302	1%
800 m	98,93	105,13	101,939	**95,08**	**-6,859**	**-7%**
1500 m	203,44	217,57	210,269	212,07	1,801	1%