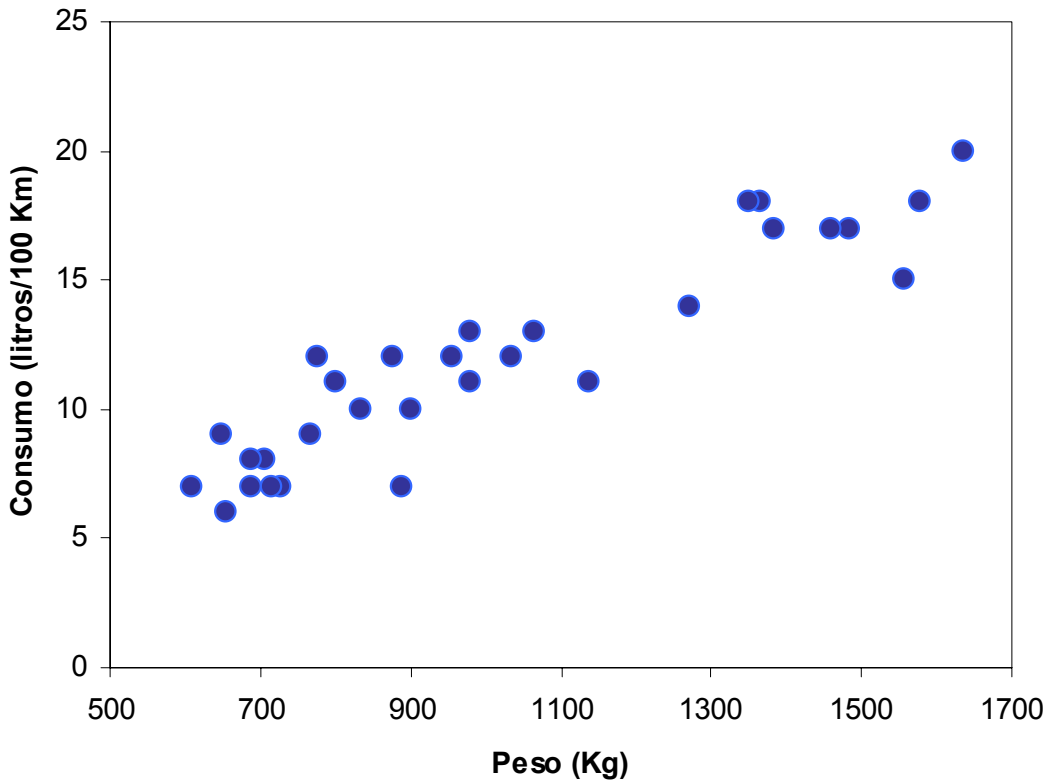
Regresión lineal

Estadística 2002-2003

Regresión simple

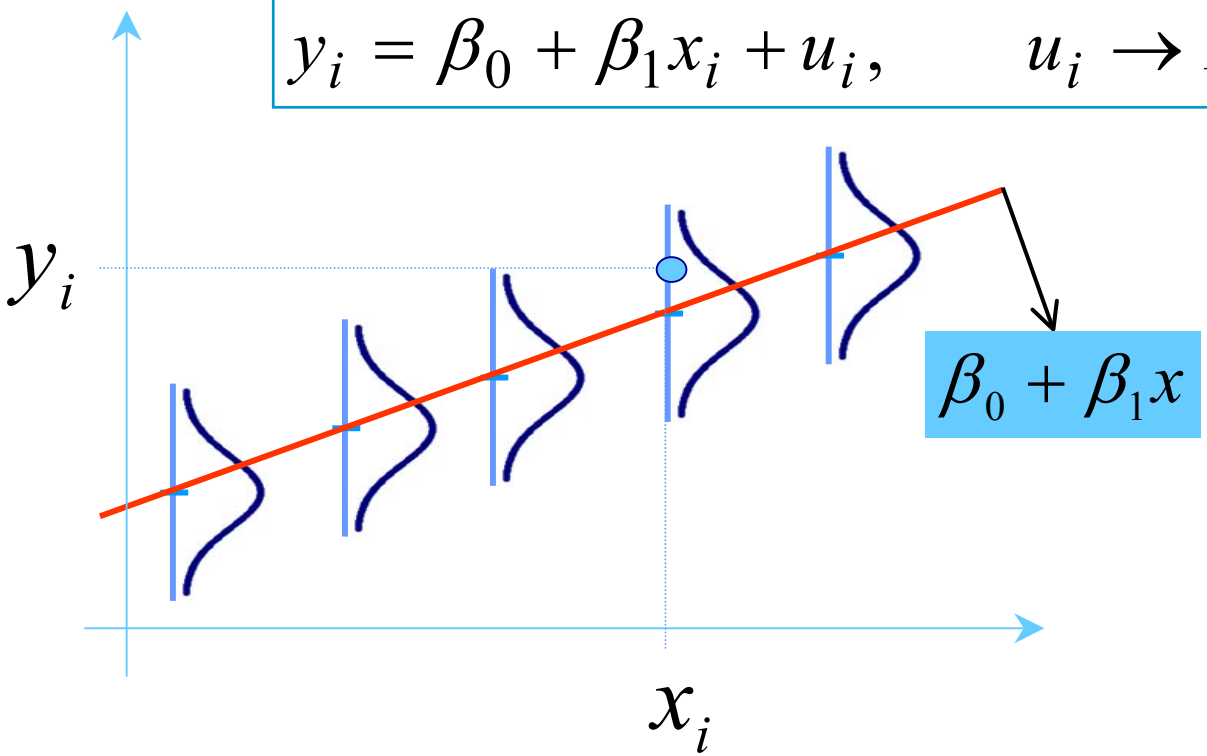
consumo y peso de automóviles

Núm. Obs. (i)	Peso kg	Consumo litros/100 km
1	981	11
2	878	12
3	708	8
4	1138	11
5	1064	13
6	655	6
7	1273	14
8	1485	17
9	1366	18
10	1351	18
11	1635	20
12	900	10
13	888	7
14	766	9
15	981	13
16	729	7
17	1034	12
18	1384	17
19	776	12
20	835	10
21	650	9
22	956	12
23	688	8
24	716	7
25	608	7
26	802	11
27	1578	18
28	688	7
29	1461	17
30	1556	15



Modelo

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad u_i \to N(0, \sigma^2)$$



$\beta_0, \beta_1, \sigma^2$: parámetros desconocidos

Hipótesis del modelo

- Linealidad
  - $y_i = \beta_0 + \beta_1 x_i + u_i$
- Normalidad
  - $y_i | x_i \Rightarrow N(\beta_0 + \beta_1 x_i, \sigma^2)$
- Homocedasticidad
  - $Var[y_i | x_i] = \sigma^2$
- Independencia
  - $Cov[y_i, y_k] = 0$

Parámetros

$$\beta_0$$

$$\beta_1$$

$$\sigma^2$$

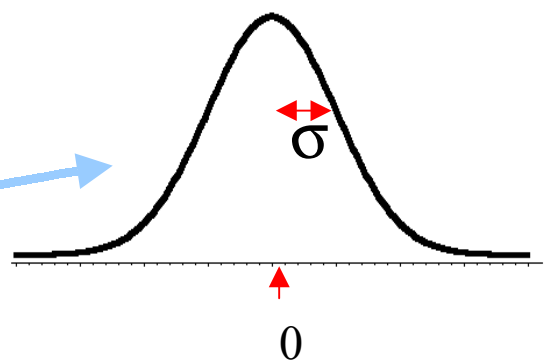
Modelo

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad u_i \rightarrow N(0, \sigma^2)$$

y_i : Variable dependiente

x_i : Variable independiente

u_i : Parte aleatoria



Estimación

$$M(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{dM}{d\beta_0} = -\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$$\frac{dM}{d\beta_0} = -\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \qquad \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

$$\left.\begin{array}{l} \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ \sum_{i=1}^{n} x_i y_i \Big/ n = \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \sum x_i^2 \Big/ n \end{array}\right\} \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n} = \hat{\beta}_1 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$\hat{\beta}_1 = \frac{\mathrm{cov}(x_i, y_i)}{\mathrm{var}(x_i)} \quad ; \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimación: máxima verosimilitud

$$L(\beta_0,\beta_1,\sigma^2)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\right]$$

$$l(\beta_0,\beta_1,\sigma^2)=\log L(\beta_0,\beta_1,\sigma^2)$$

$$=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

$$\frac{dl}{d\beta_0}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1x_i)=0 \qquad \sum y_i=n\hat\beta_0+\hat\beta_1\sum x_i$$

$$\frac{dl}{d\beta_0}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1x_i)x_i=0 \qquad \sum x_iy_i=\hat\beta_0\sum x_i+\hat\beta_1\sum x_i^2$$

$$\left.\begin{array}{l}\bar y=\hat\beta_0+\hat\beta_1\bar x\\ \sum_{i=1}^{n}x_iy_i \Big/ n=\hat\beta_0\bar x+\hat\beta_1\sum x_i^2\Big/n\end{array}\right\}\quad \frac{\sum_{i=1}^{n}(y_i-\bar y)(x_i-\bar x)}{n}=\hat\beta_1\frac{\sum_{i=1}^{n}(x_i-\bar x)^2}{n}$$

$$\hat\beta_1=\frac{\mathrm{cov}(x_i,y_i)}{\mathrm{var}(x_i)} \quad ; \quad \hat\beta_0=\bar y-\hat\beta_1\bar x$$

Estimación σ^2 : máxima verosimilitud

$$l(\beta_0,\beta_1,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{dl}{d\sigma^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

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

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\left.\begin{aligned}\sum_{i=1}^{n} e_i &= 0\\ \sum_{i=1}^{n} e_i x_i &= 0\end{aligned}\right\} \quad \hat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Estimación

Máxima verosimilitud

$$Max\left\{\frac{1}{(2\pi)^{n/2}\sigma^{n}}\exp\left[-\frac{1}{2\sigma^{2}}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right]\right\}$$

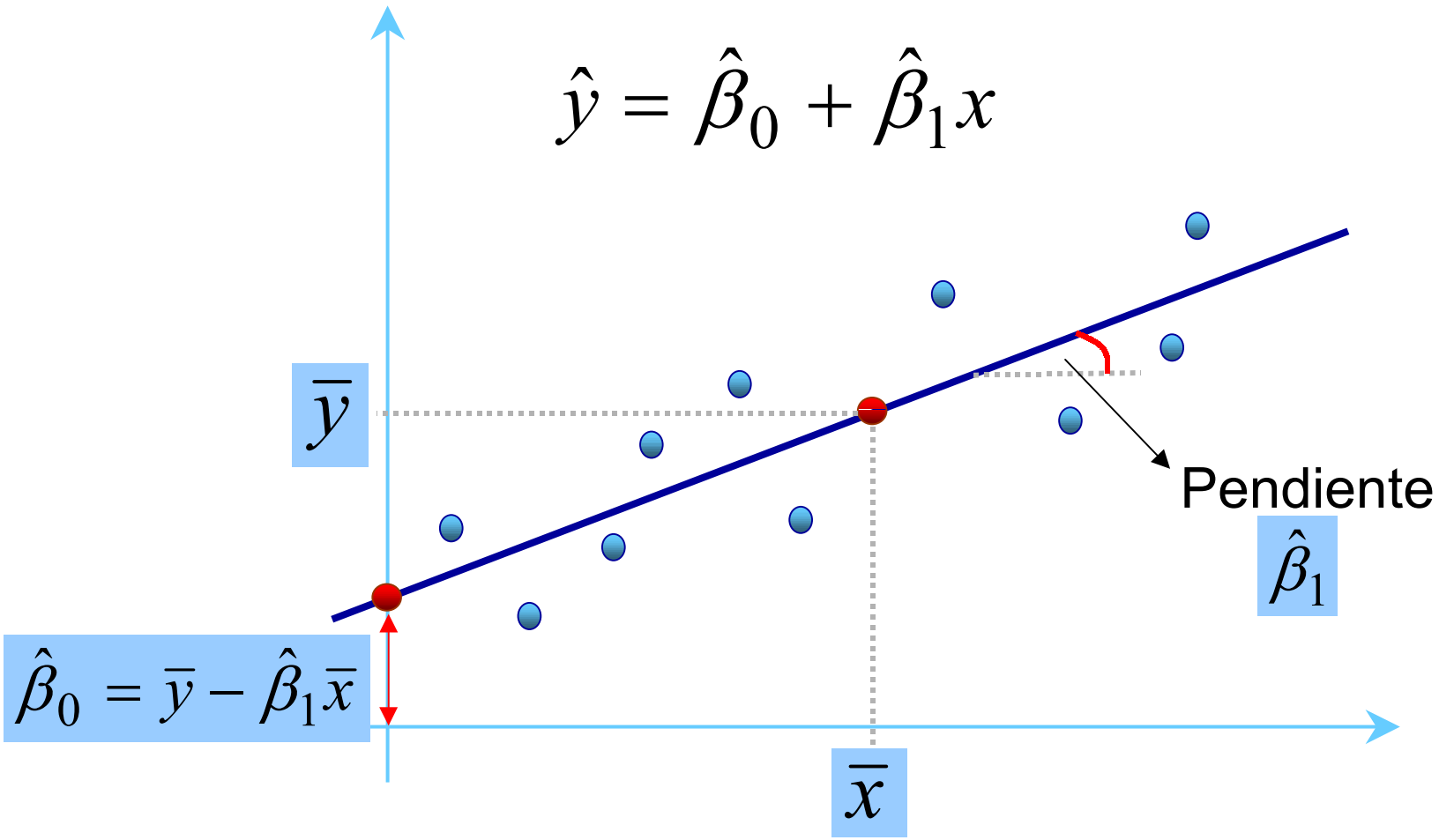
Mínimos cuadrados

$$\text{Mín}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

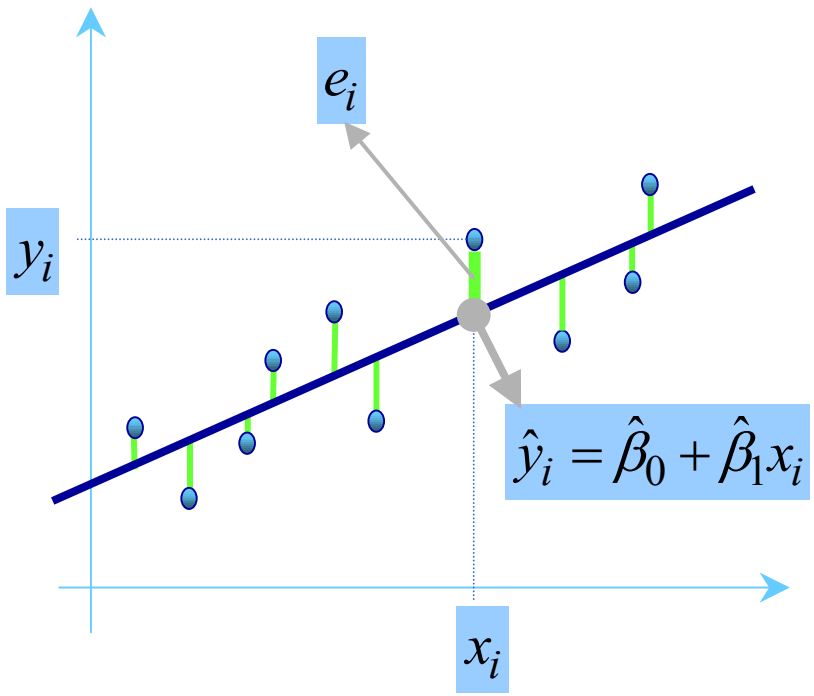
$$\hat{\beta}_1=\frac{\text{cov}(x_i,y_i)}{\text{var}(x_i)}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Recta de regresión



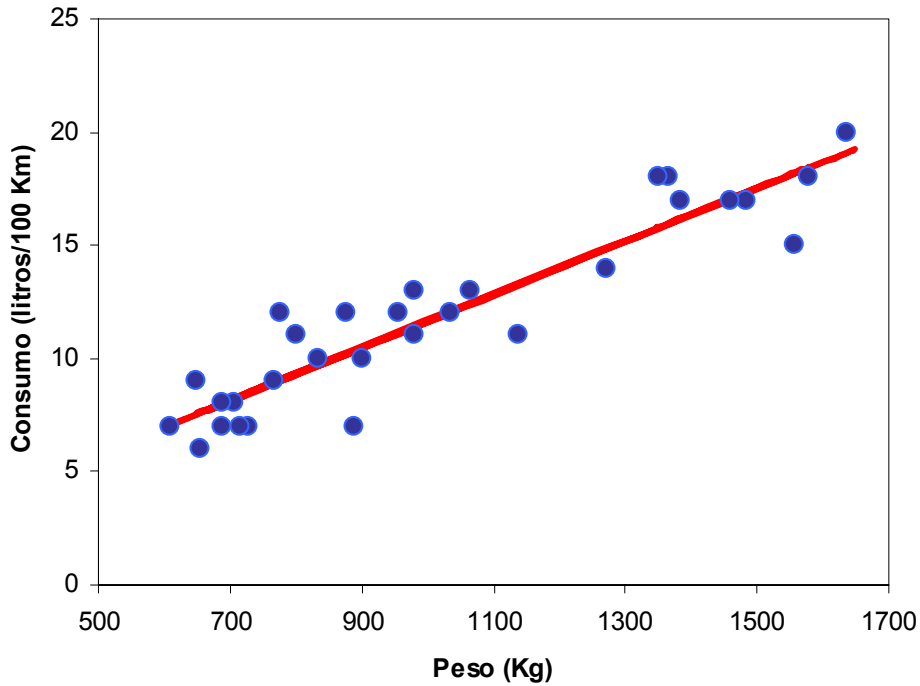
Residuos

$$\underbrace{y_i}_{\text{Valor observado}} = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\text{Valor Previsto}} + \underbrace{e_i}_{\text{Residuo}}$$



Ejemplo: estimación

Núm. Obs. (i)	Peso kg	Consumo litros/100 km	Predicción	Residuos
1	981	11	11,44	-0,44
2	878	12	10,23	1,77
3	708	8	8,23	-0,23
4	1138	11	13,28	-2,28
5	1064	13	12,41	0,59
6	655	6	7,61	-1,61
7	1273	14	14,86	-0,86
8	1485	17	17,35	-0,35
9	1366	18	15,95	2,05
10	1351	18	15,78	2,22
11	1635	20	19,11	0,89
12	900	10	10,49	-0,49
13	888	7	10,35	-3,35
14	766	9	8,91	0,09
15	981	13	11,44	1,56
16	729	7	8,48	-1,48
17	1034	12	12,06	-0,06
18	1384	17	16,16	0,84
19	776	12	9,03	2,97
20	835	10	9,72	0,28
21	650	9	7,55	1,45
22	956	12	11,14	0,86
23	688	8	8,00	0,00
24	716	7	8,33	-1,33
25	608	7	7,06	-0,06
26	802	11	9,34	1,66
27	1578	18	18,44	-0,44
28	688	7	8,00	-1,00
29	1461	17	17,07	-0,07
30	1556	15	18,18	-3,18



$$\hat{y}_i = -0.071 + 0.0117 x_i \quad ; \quad \hat{s}_R^2 = 2.38$$

Propiedades de $\hat{\beta}_1$

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

$$= \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})y_i - \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\bar{y} \quad \rightarrow 0$$

$$= \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{ns_x^2}\right)y_i = w_1 y_1 + w_2 y_2 + \cdots + w_n y_n$$

$$w_i = \frac{x_i - \bar{x}}{n\hat{s}_x^2}$$

- $\sum_{i=1}^{n} w_i = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- $\sum_{i=1}^{n} w_i x_i = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})x_i = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})x_i - \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1$
- $\sum_{i=1}^{n} w_i^2 = \left(\frac{1}{ns_x^2}\right)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{ns_x^2}$

$\bar{y}, \hat{\beta}_1$ son v.a. independientes

$$\left.\begin{aligned} \bar{y} &= \frac{1}{n}y_1 + \frac{1}{n}y_2 + \cdots + \frac{1}{n}y_n = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{a}^T\mathbf{Y} \\ \hat{\beta}_1 &= w_1y_1 + w_2y_2 + \cdots + w_ny_n = \begin{pmatrix} w_1 & w_2 & \cdots & w_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{w}^T\mathbf{Y} \end{aligned}\right\}$$

$$\operatorname{cov}(\bar{y}, \hat{\beta}_1) = \mathbf{a}^T \operatorname{var}(\mathbf{Y})\mathbf{w} = \frac{\sigma^2}{n} \sum_{i=1}^{n} w_i = 0$$

Distribución de $\hat{\beta}_1$

$$y_i \rightarrow N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\hat{\beta}_1 = w_1 y_1 + w_2 y_2 + \cdots + w_n y_n \rightarrow \text{Comb. lineal de normales}$$

$$\begin{aligned} E[\hat{\beta}_1] &= E[w_1 y_1 + w_2 y_2 + \cdots + w_n y_n] \\ &= w_1 E[y_1] + w_2 E[y_2] + \cdots + w_n E[y_n] \quad (E[y_i] = \beta_0 + \beta_1 x_i) \\ &= \beta_0 (\sum w_i) + \beta_1 (\sum w_i x_i) = \beta_1 \end{aligned}$$

$$\begin{aligned} Var[\hat{\beta}_1] &= Var[w_1 y_1 + w_2 y_2 + \cdots + w_n y_n] \\ &= w_1^2 Var[y_1] + w_2^2 Var[y_2] + \cdots + w_n^2 [y_n] \quad (Var[y_i] = \sigma^2) \\ &= (\sum_{i=1}^{n} w_i^2)\sigma^2 = \frac{\sigma^2}{n s_x^2} \end{aligned}$$

$$\hat{\beta}_1 \rightarrow N\left(\beta_1, \frac{\sigma^2}{n s_x^2}\right)$$

Modelo en diferencias a la media

$$\left.\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \\ \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \end{aligned}\right\} y_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}) + e_i$$

$$y_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) + e_i$$

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

Distribución de $\hat{\beta}_0$

- $\bar{y} \to N(\beta_0 + \beta_1 \bar{x}, \frac{\sigma^2}{n})$
- $\hat{\beta}_1 \to N(\beta_1, \frac{\sigma^2}{n s_x^2})$
- $\bar{y}, \hat{\beta}_1$ son independientes

$$\left.\begin{array}{c} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \to Normal \\ E[\hat{\beta}_0] = E[\bar{y}] - \bar{x} E[\hat{\beta}_1] = \beta_0 \\ \text{var}[\hat{\beta}_0] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right) \end{array}\right\} \quad \hat{\beta}_0 \to N\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right)\right)$$

Distribución de $\square^2_R$

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$u_i \to N(0, \sigma^2)$$

$$\frac{\sum_{i=1}^{n} u_i^2}{\sigma^2} \to \chi^2_n$$

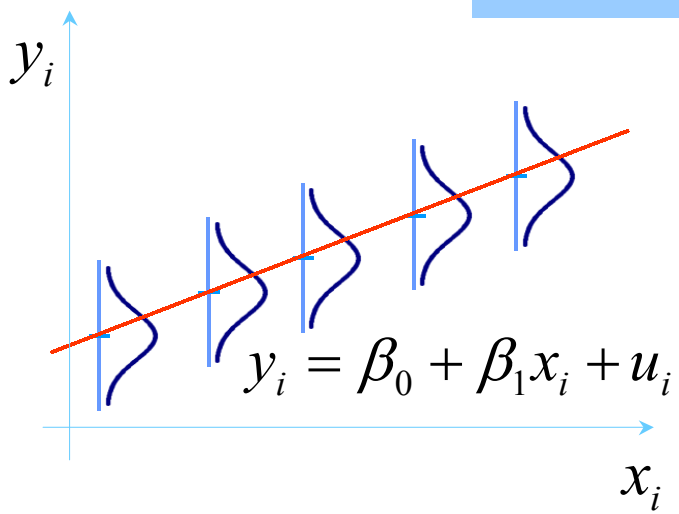
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \to \chi^2_{n-2} \quad \Longrightarrow \quad \begin{cases} \sum e_i = 0 \\ \sum e_i x_i = 0 \end{cases}$$

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \frac{(n-2)\hat{s}^2_R}{\sigma^2} \to \chi^2_{n-2}$$

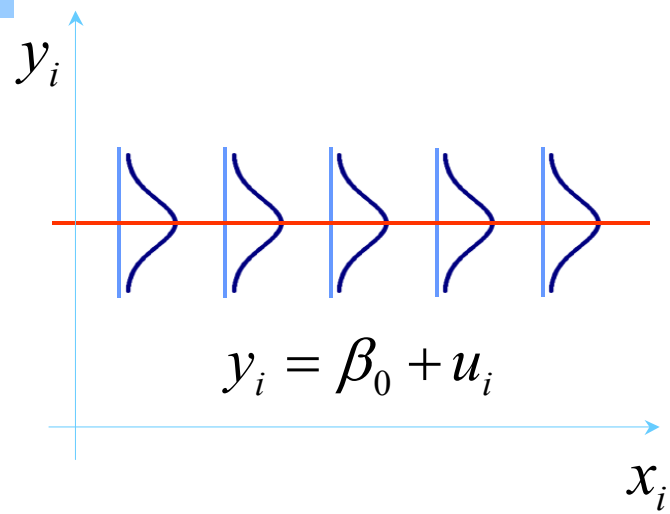
Contraste principal de regresión: ¿depende *y* de *x*?

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$



H_0 es falso

x e *y* están relacionados



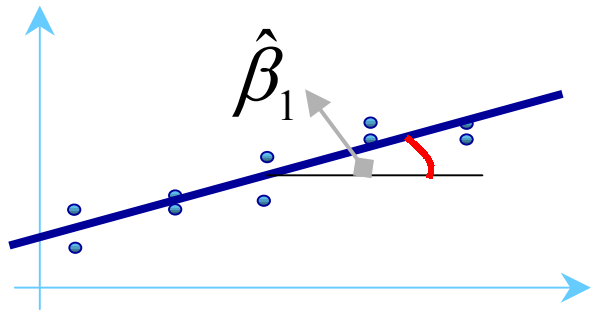
H_0 es cierto

x e *y* no están relacionados

Contraste sobre la pendiente

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

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



$$\hat{\beta}_1 \rightarrow N(\beta_1, \frac{\sigma^2}{ns_x^2})$$

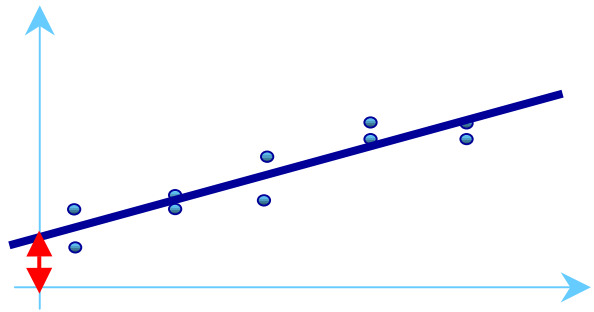
$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{n}s_x}} \rightarrow N(0,1) \quad \Rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{s}_R}{\sqrt{n}s_x}} \rightarrow t_{n-2}$$

$$t_1 = \frac{\hat{\beta}_1}{\frac{\hat{s}_R}{\sqrt{n}s_x}}; \quad |t_1| > t_{n-2;\alpha/2} \Rightarrow \text{Se rechaza Ho}$$

Contraste: ordenada en el origen

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

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



$$\hat{\beta}_0 \rightarrow N(\beta_0, \frac{\sigma^2}{n}(1 + \frac{\bar{x}^2}{s_x^2}))$$

$$t_0 = \frac{\hat{\beta}_0}{\frac{\hat{s}_R}{\sqrt{n}}\sqrt{1 + \frac{\bar{x}^2}{s_x^2}}};$$

$$|t_0| > t_{n-2;\alpha/2} \Rightarrow \text{Se rechaza Ho}$$

Descomposición de la variabilidad en regresión

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \underbrace{e_i}_{y_i - \hat{y}_i}$$

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) \quad \text{(restando } \bar{y}\text{)}$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad \text{(elevando al cuadrado y sumando)}$$

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

$$VT = VE + VNE$$

Coeficiente de determinación R^2

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

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

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

$$VT = VE + VNE$$

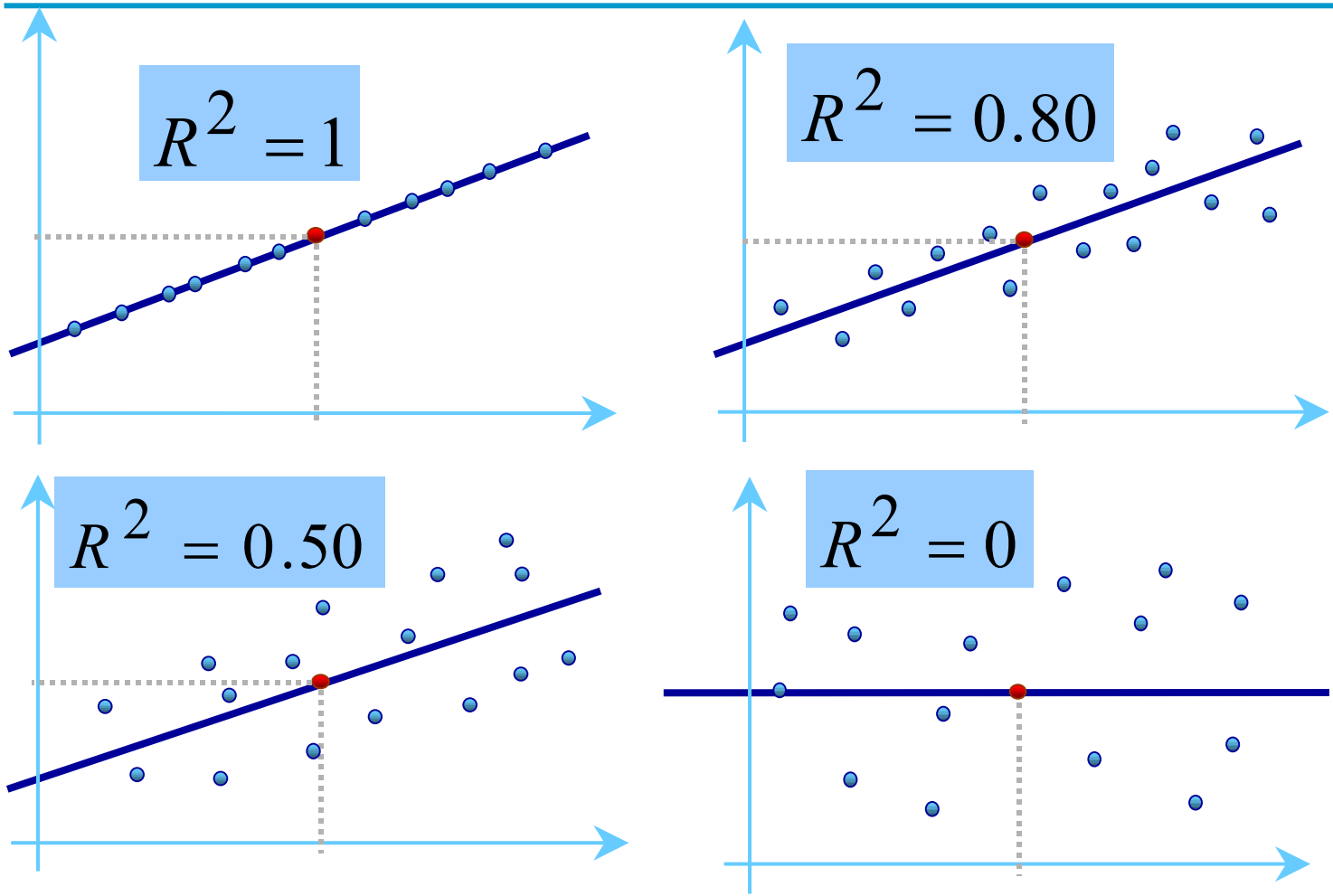
$$R^2 = \frac{VE}{VT}$$

$$0 \leq R^2 \leq 1$$

Mide el porcentaje de VT que está explicado por el regresor

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) :\Rightarrow VE = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 n s_x^2$$

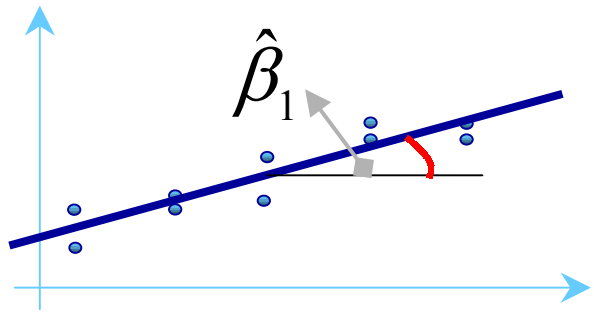
Coef. determinación



Contraste F

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

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



$$\frac{VE}{\sigma^2} \rightarrow \chi^2_1 \quad \text{(Si } H_o \text{ es cierto)}$$

$$\frac{VNE}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \frac{(n-2)\hat{s}_R^2}{\sigma^2} \rightarrow \chi^2_{n-2}$$

$$\frac{VE}{\sigma^2}, \frac{VNE}{\sigma^2} \text{ son independientes}$$

$$F = \frac{VE}{VNE/(n\text{-}2)} = \frac{VE}{\hat{s}_R^2} \rightarrow F_{1,n-2}$$

$$F > F_\alpha \Rightarrow \text{Se rechaza } H_0$$

Regresión con *Statgraphics*

Dependent variable: Consumo
Independent variable: Peso

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	-0,0712606	0,945148	-0,0753962	0,9404
Slope	0,0117307	0,000886531	13,2321	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Valu
Model	416,811	1	416,811	175,09	0,000
Residual	66,6559	28	2,38057		
Total (Corr.)	483,467	29			

Correlation Coefficient = 0,928509
R-squared = 86,2129 percent
Standard Error of Est. = 1,54291

Ejemplo regresión múltiple

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

Y **Consumo**	X1 **Cilindrada**	X2 **Potencia**	X3 **Peso**	X4 **Aceleración**
l/100Km	*cc*	*CV*	*kg*	*segundos*
15	4982	150	1144	12
16	6391	190	1283	9
24	5031	200	1458	15
9	1491	70	651	21
11	2294	72	802	19
17	5752	153	1384	14
...	...	...	...	...

Var. dependientes o respuesta

Var. Independientes o regresores

Modelo regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i,$$
$$u_i \rightarrow N(0, \sigma^2)$$

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k, \sigma^2$: parámetros desconocidos

- **Linealidad**

  $E[y_i] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$

- **Normalidad**

  $y_i | x_1, \ldots, x_k \Rightarrow Normal$

- **Homocedasticidad**

  $Var[y_i | x_1, \ldots, x_k] = \sigma^2$

- **Independencia**

  $Cov[y_i, y_k] = 0$

Notación matricial

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{a}} + \mathbf{U}$$

$$\mathbf{U} \rightarrow N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Estimación mínimo-cuadrática

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

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{a}} + \mathbf{e}$$

donde el vector $\mathbf{e}$ cumple

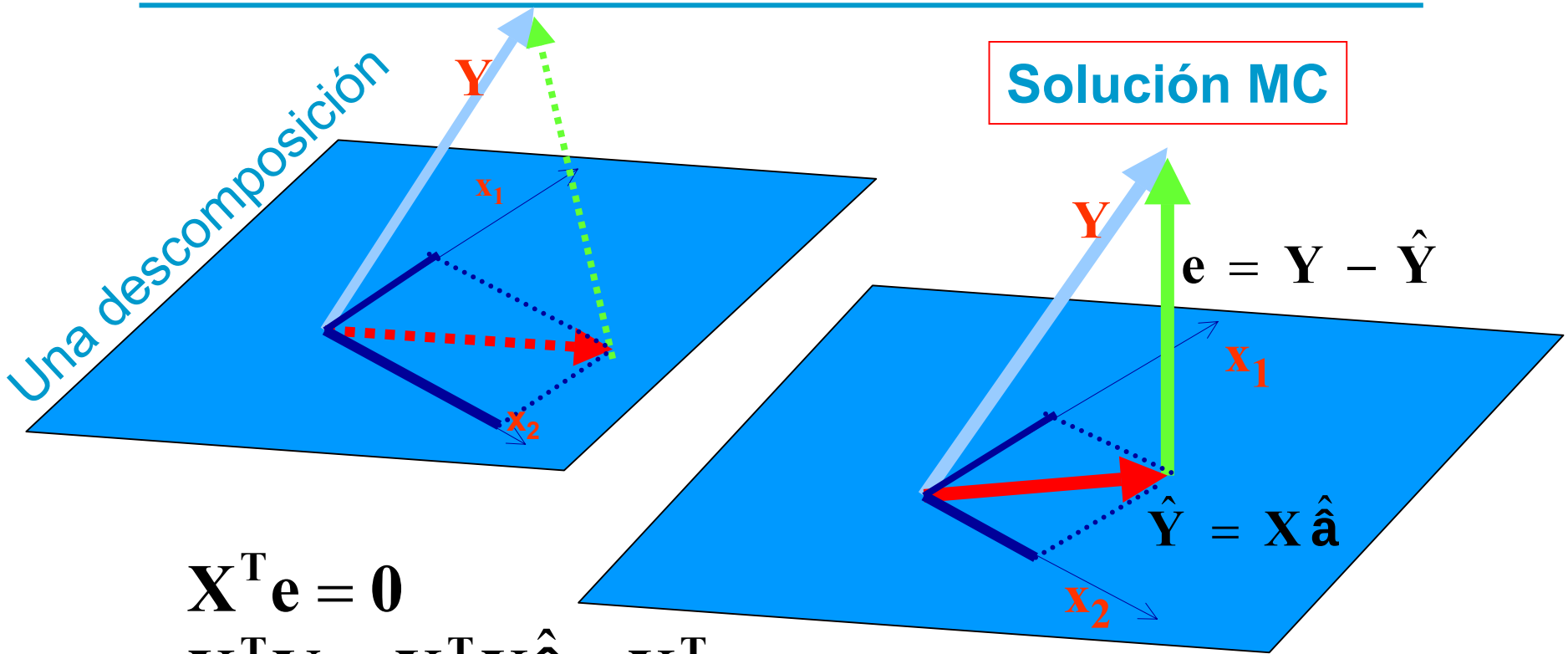
$$\|\mathbf{e}\|^2 = \sum_{i=1}^{n} e_i^2 \quad \text{es mínimo}$$

Para que $\|\mathbf{e}\|^2$ sea mínimo, $\mathbf{e}$ tiene que ser perpendicular al espacio vectorial generado las columnas de $\mathbf{X}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\Rightarrow \mathbf{X}^{\mathbf{T}}\mathbf{e} = \mathbf{0} \quad \begin{cases} \sum_1^n e_i = 0 \\ \sum_1^n e_i x_{1i} = 0 \\ \vdots \\ \sum_1^n e_i x_{ki} = 0 \end{cases}$$

Mínimos cuadrados

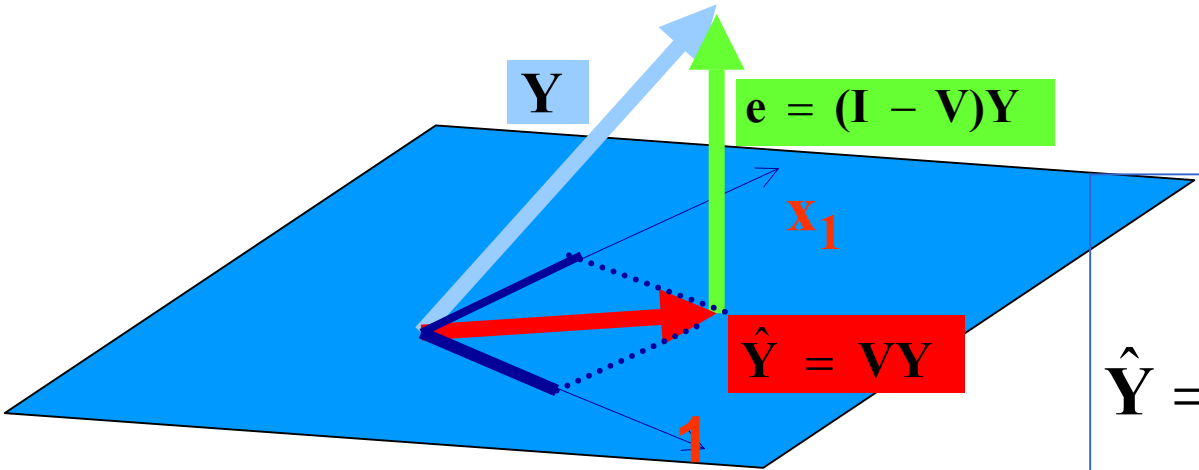


$$\mathbf{X}^T\mathbf{e} = \mathbf{0}$$

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\mathbf{a}} + \mathbf{X}^T\mathbf{e}$$

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\mathbf{a}} \quad \Rightarrow \quad \hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Matriz de proyección V



Val. Previstos

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\alpha}}$$
$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$
$$\hat{\mathbf{Y}} = \mathbf{V}\mathbf{Y}$$

Residuos

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\alpha}} = \mathbf{Y} - \mathbf{V}\mathbf{Y} = (\mathbf{I} - \mathbf{V})\mathbf{Y}$$

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

Simétrica $V=V^{T}$

Idempotente $VV=V$

Distribución de probabilidad de $\hat{\mathbf{a}}$

$$\mathbf{Y} \rightarrow N(\mathbf{X}\hat{\mathbf{a}}, \sigma^2\mathbf{I})$$

$$\hat{\mathbf{a}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y} = \mathbf{C}\mathbf{Y} \quad (\text{siendo } \mathbf{C} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T})$$

$$\hat{\mathbf{a}} \rightarrow Normal$$

$$E[\hat{\mathbf{a}}] = \mathbf{C}E[\mathbf{Y}] = \mathbf{C}\mathbf{X}\hat{\mathbf{a}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{X}\hat{\mathbf{a}} = \hat{\mathbf{a}}$$

$$Var[\hat{\mathbf{a}}] = Var[\mathbf{C}\mathbf{Y}] = \mathbf{C}Var[\mathbf{Y}]\mathbf{C}^\mathbf{T}$$

$$= ((\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T})(\sigma^2\mathbf{I})((\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T})^T$$

$$= \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$$

$$= \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$$

Distribución de probabilidad de $\hat{\mathbf{a}}$

$$\hat{\mathbf{a}} \rightarrow N(\mathbf{a}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

$$\hat{\beta}_i \rightarrow N(\beta_i, \sigma^2 q_{ii})$$

$$\hat{\mathbf{a}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \quad \mathbf{a} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \mathbf{Q} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} q_{00} & q_{01} & \cdots & q_{0k} \\ q_{10} & q_{11} & \cdots & q_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k0} & q_{k1} & \cdots & q_{kk} \end{pmatrix}$$

$$\dim(\mathbf{Q}) = (k+1) \times (k+1)$$

Residuos

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

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

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

Varianza Residual

$$\frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \rightarrow \chi^2_{n-k-1}$$

$$E\left[\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}\right] = n-k-1$$

$$E\left[\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}\right] = \sigma^2$$

$$\hat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

$$\frac{(n-k-1)\hat{s}_R^2}{\sigma^2} \rightarrow \chi^2_{n-k-1}$$

Contraste individual β_i

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

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

$$\hat{\beta}_i \rightarrow N(\beta_i, \sigma^2 q_{ii})$$

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{q_{ii}}} \rightarrow N(0,1) \quad \Rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{s}_R\sqrt{q_{ii}}} \rightarrow t_{n-k-1}$$

$$t_i = \frac{\hat{\beta}_i}{\hat{s}_R\sqrt{q_{ii}}}; \quad |t_i| > t_{n-k-1;\alpha/2} \Rightarrow \textbf{Se rechaza Ho}$$

Descomposición de la variabilidad en regresión

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

$$y_i = \hat{y}_i + e_i \quad \text{(Restando } \bar{y}\text{)}$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + e_i$$

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

$$VT = VE + VNE$$

Modelo en diferencias a la media

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

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_{1i} - \bar{x}_1) + \cdots + \hat{\beta}_k (x_{ki} - \bar{x}_k)$$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ki} + \underbrace{\sum_{i=1}^{n} e_i}_{0}$$

$$\begin{pmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \vdots \\ \hat{y}_n - \bar{y} \end{pmatrix} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

$$\hat{\mathbf{Y}} - \overline{\mathbf{Y}} = \tilde{\mathbf{X}}\hat{\mathbf{b}}$$

$$\mathbf{Y} - \overline{\mathbf{Y}} = \tilde{\mathbf{X}}\hat{\mathbf{b}} + \mathbf{e}$$

Modelo en diferencias a la media

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{b} + \mathbf{U}$$

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \vdots \\ \hat{y}_n - \bar{y} \end{pmatrix}, \quad \overline{\mathbf{Y}} = \begin{pmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{pmatrix}$$

$$\hat{\mathbf{b}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{Y}}$$

$$\hat{\mathbf{b}} \rightarrow N(\mathbf{b}, \acute{o}^2 (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1})$$

Contraste general de regresión.

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

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \text{alguno es distinto de } 0$$

$$\frac{VE}{\sigma^2} \to \chi^2_k \quad \text{(Si } H_o \text{ es cierto)}$$

$$\frac{VNE}{\sigma^2} = \frac{(n-k-1)\hat{s}^2_R}{\sigma^2} \to \chi^2_{n-k-1}$$

$$\frac{VE}{\sigma^2}, \frac{VNE}{\sigma^2} \text{ son independientes}$$

$$F = \frac{VE/k}{VNE/(n-k-1)} \to F_{k,n-k-1}$$

$$F > F_\alpha \Rightarrow \text{Se rechaza } H_0$$

Coeficiente de determinación R^2

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

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

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

$$VT = VE + VNE$$

$$R^2 = \frac{VE}{VT}$$

$$0 \leq R^2 \leq 1$$

Mide el porcentaje de VT que está explicado por los regresores

$$VE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\hat{\mathbf{Y}} - \overline{\mathbf{Y}})^T (\hat{\mathbf{Y}} - \overline{\mathbf{Y}}) = \hat{\mathbf{b}}^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}) \hat{\mathbf{b}} = \hat{\mathbf{b}}^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{Y}})$$

Coef. determinación corregido $\overline{R}^2$

$$R^2 = \frac{VE}{VT} = \frac{VT - VNE}{VT}$$

$$= 1 - \frac{VNE}{VT} = 1 - \frac{(n-k-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2}$$

$$\hat{s}_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$\overline{R}^2 = 1 - \frac{\hat{s}_R^2}{\hat{s}_y^2} = 1 - \frac{VNE/(n-k-1)}{VT/(n-1)}$$

Regresión con STATGRAPHICS

Multiple Regression Analysis

Dependent variable: consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-1,66958	0,983305	-1,69793	0,0903
cilindrada	0,000383473	0,0001625	2,35983	0,0188
potencia	0,0402844	0,00656973	6,13183	0,0000
peso	0,00578424	0,00095783	6,0389	0,0000
aceleracion	0,111501	0,0496757	2,24458	0,0254

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4845,0	4	1211,25	438,70	0,0000
Residual	1065,74	386	2,76099		
Total (Corr.)	5910,74	390			

R-squared = 81,9694 percent
R-squared (adjusted for d.f.) = 81,7826 percent
Standard Error of Est. = 1,66162

Interpretación (inicial)

- Contraste F=438 (p-valor=0.0000) $\Rightarrow$ Alguno de los regresores influye significativamente en el consumo.
- Contrastes individuales:
  - La potencia y el peso influyen significativamente (p-valor=0.0000)
  - Para α=0.05, la cilindrada y la aceleración también tienen efecto significativo (p-valor < 0.05)
- El efecto de cualquier regresor es “positivo”, al aumentar cualquiera de ellos aumenta la variable respuesta: consumo.
- Los regresores explican el 82 % de la variabilidad del consumo (R^2 = 81.969)

Multicolinealidad

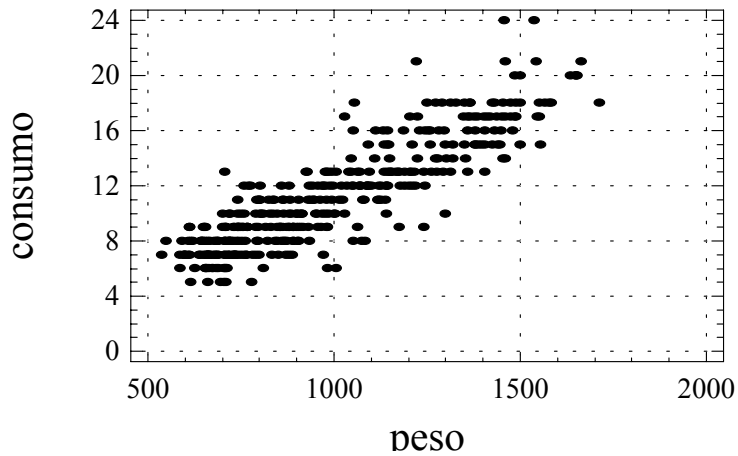
- Cuando la correlación entre los regresores es alta.
- Presenta graves inconvenientes:
  - Empeora las estimaciones de los efectos de cada variable β_i: aumenta la varianza de las estimaciones y la dependencia de los estimadores)
  - Dificulta la interpretación de los parámetros del modelo estimado (ver el caso de la aceleración en el ejemplo).

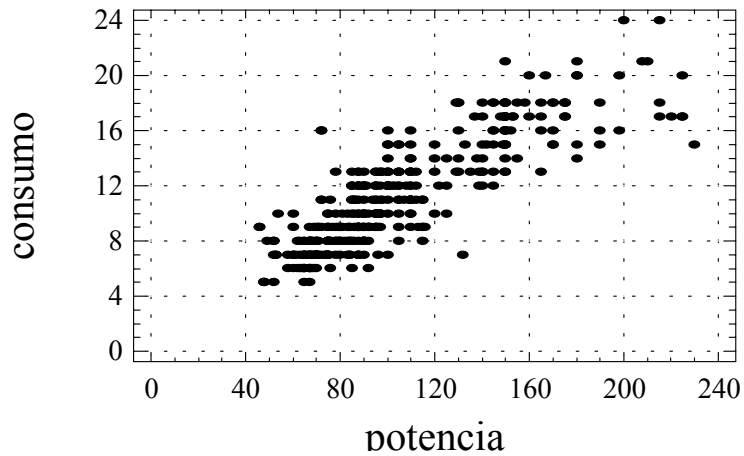
Identificación de la multicolinealidad:
Matriz de correlación de los regresores.

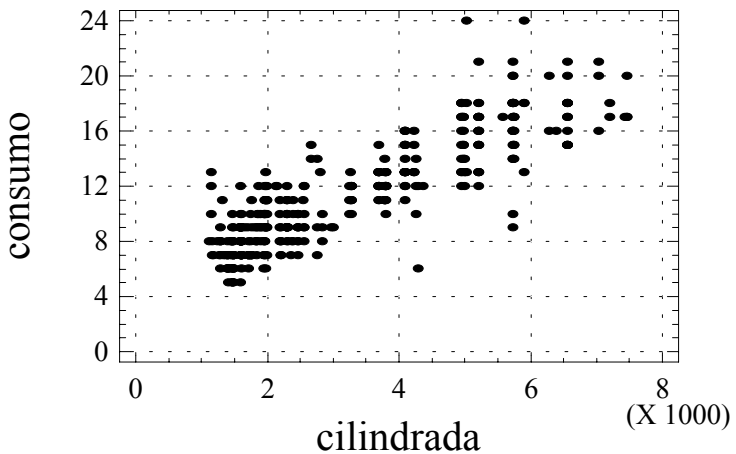
Correlations

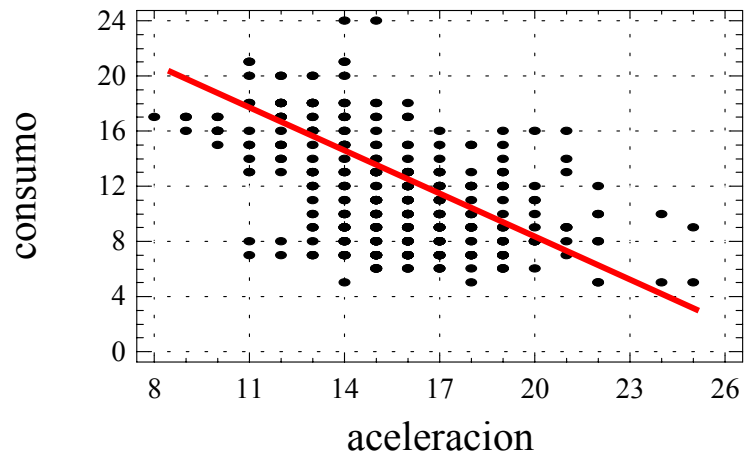
	cilindrada	potencia	peso	aceleraci
cilindrada		0,8984	0,9339	-0,5489
		(391)	(391)	(391)
		0,0000	0,0000	0,0000
potencia	0,8984		0,8629	-0,6963
	(391)		(391)	(391)
	0,0000		0,0000	0,0000
peso	0,9339	0,8629		-0,4216
	(391)	(391)		(391)
	0,0000	0,0000		0,0000
aceleracion	-0,5489	-0,6963	-0,4216	
	(391)	(391)	(391)	
	0,0000	0,0000	0,0000	

Gráficos consumo - x_i









Consumo y aceleración

R. simple

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: consumo
Independent variable: aceleracion

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	21,5325	1,00701	21,3827	0,0000
aceleracion	-0,657509	0,0632814	-10,3902	0,0000

R. múltiple

Multiple Regression Analysis

Dependent variable: consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-1,66958	0,983305	-1,69793	0,0903
cilindrada	0,000383473	0,0001625	2,35983	0,0188
potencia	0,0402844	0,00656973	6,13183	0,0000
peso	0,00578424	0,00095783	6,0389	0,0000
aceleracion	0,111501	0,0496757	2,24458	0,0254

Multicolinealidad: efecto en la varianza de los estimadores

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

$$\text{var}\left[\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix}\right] = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\sigma^2 \quad \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = n\mathbf{S}_x \quad \mathbf{S}_x = \begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix} = \begin{pmatrix} s_1^2 & r_{12}s_1s_2 \\ r_{12}s_1s_2 & s_2^2 \end{pmatrix}$$

$$|\mathbf{S}_x| = s_1^2 s_2^2 (1-r_{12}^2) \quad \mathbf{S}_x^{-1} = \begin{pmatrix} \dfrac{1}{s_1^2(1-r_{12}^2)} & \dfrac{-r_{12}}{s_1 s_2 (1-r_{12}^2)} \\ \dfrac{-r_{12}}{s_1 s_2 (1-r_{12}^2)} & \dfrac{1}{s_2^2(1-r_{12}^2)} \end{pmatrix}$$

$$\text{var}\left[\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix}\right] = \begin{pmatrix} \dfrac{\sigma^2}{ns_1^2(1-r_{12}^2)} & \dfrac{-r_{12}\sigma^2}{ns_1 s_2(1-r_{12}^2)} \\ \dfrac{-r_{12}\sigma^2}{ns_1 s_2(1-r_{12}^2)} & \dfrac{\sigma^2}{ns_2^2(1-r_{12}^2)} \end{pmatrix}$$

Consecuencias de la multicolinealidad

- Gran varianza de los estimadores β
- Cambio importante en las estimaciones al eliminar o incluir regresores en el modelo
- Cambio de los contrastes al eliminar o incluir regresores en el modelo.
- Contradicciones entre el contraste F y los contrastes individuales.