Módulo 3: Regresión

Lección 3: Regresión Múltiple I

Regresión Multiple

	log(fev)	ht	age	predict	residuo
1	0.535	57.0	9	0.715	-0.179
2	0.545	67.5	8	1.157	-0.612
3	0.542	54.5	7	0.565	-0.023
4	0.443	53.0	9	0.539	-0.095
5	0.639	57.0	9	0.715	-0.075
6	0.848	61.0	8	0.871	-0.022
7	0.652	58.0	6	0.699	-0.047
8	0.347	56.0	6	0.611	-0.264
9	0.687	58.5	8	0.761	-0.074
10	0.664	60.0	9	0.847	-0.183
11	0.471	53.0	6	0.479	-0.008
12	0.551	54.0	8	0.563	-0.012
13	0.785	58.5	8	0.761	0.024
14	0.750	60.5	8	0.849	-0.098

...

650	1.452	67.0	16	1.293	0.158
651	1.316	68.0	15	1.318	-0.002
652	1.048	60.0	18	1.025	0.023
653	1.028	63.0	16	1.117	-0.090
654	1.167	66.5	15	1.252	-0.085

Regresión múltiple

$$\log(\text{fev}_i) = \beta_0 + \beta_1 \text{ht}_i + \beta_2 \text{age}_i + u_i, \quad u_i \to N(0,\sigma)$$

$$\log(\text{fev}_i) = \underbrace{-1.97 + 0.0439\ \text{ht}_i + 0.0198\ \text{age}_i}_{\widehat{\log(\text{fev}_i)}} + e_i$$

$\hat{s}_R = 0.1476$

$$\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}{s_1}^2 & s_{12} \\ s_{12} & {s_2}^2\end{pmatrix}^{-1} \begin{pmatrix}s_{1y} \\ s_{2y}\end{pmatrix} = \begin{pmatrix}0.0439 \\ 0.0198\end{pmatrix}$$

$$\hat{\beta}_0 = \text{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 = -1.97$$

Ejemplo regresión múltiple

Consumo = β_0 + β_1 CC + β_2 Pot + β_3 Peso + β_4 Acel + Error

Y **Consumo**	X_1 **Cilindrada**	X_2 **Potencia**	X_3 **Peso**	X_4 **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

  $Cov[y_i, y_k] = 0$

Estimación

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

$$\boldsymbol{b} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \qquad \boldsymbol{b} = (S_{XX})^{-1}(S_{XY})$$

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

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

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

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

$$e_i = y_i - \hat{y}_i \rightarrow \hat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} \longrightarrow \text{g.l.} = n\text{-}k\text{-}1$$

Contrastes individuales

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

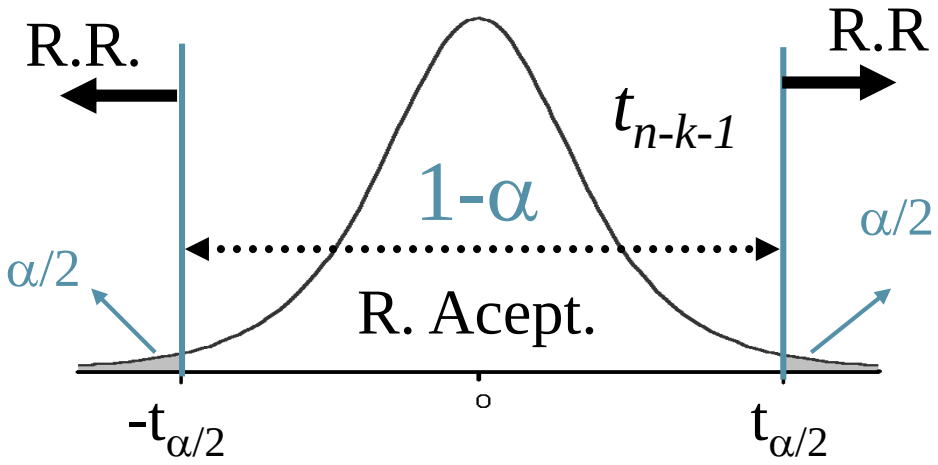
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

$$\beta_i \rightarrow \hat{\beta}_i, SE(\hat{\beta}_i)$$

$$t_i = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \rightarrow t_{n-k-1}$$

$$t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)};$$

$|t_1| > t_{n-k-1;\alpha/2} \Rightarrow$ Se rechaza Ho

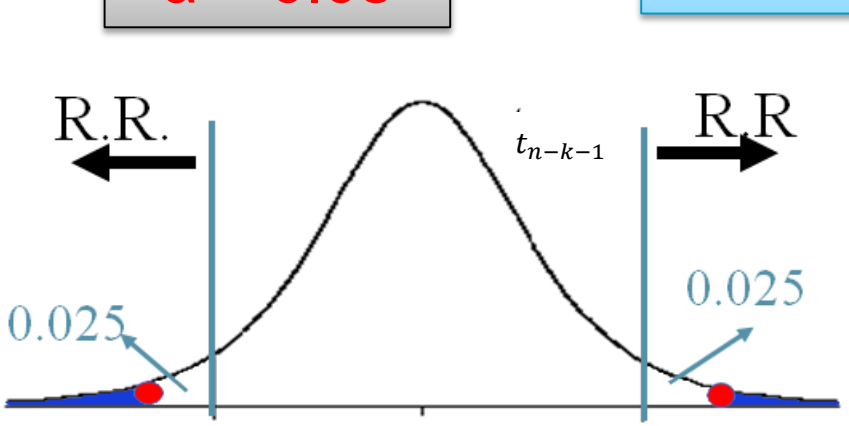


P-valor

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

α = 0.05

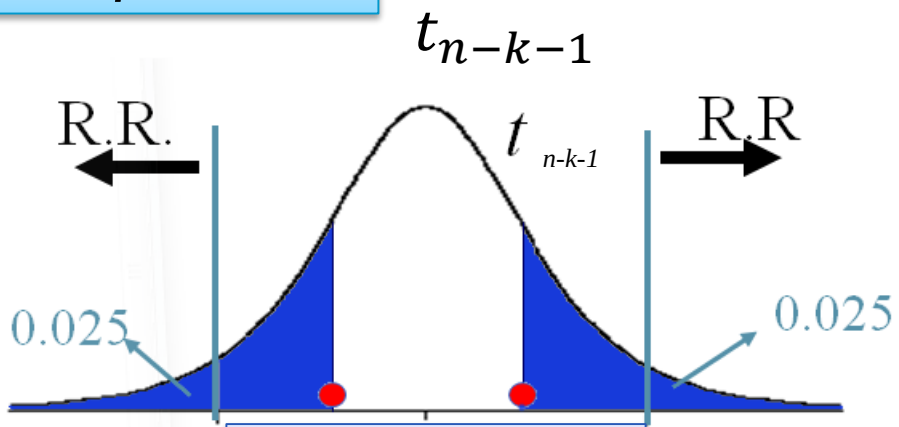
Area Azul = *p-valor*



P-valor ≤ α

Con α=0.05 "x" influye significativamente en "y"

$$H_1 : \beta_i \neq 0$$



P-valor > α

Con α=0.05 "x" NO influye significativamente en "y"

$$H_0 : \beta_i = 0$$

Modelo estimado y contrastes

Dependiente (**y**) ~ Independientes ($\mathbf{x_1}, \mathbf{x_2}, .., \mathbf{x_k}$)

	Estimate	**Stand Error**	**t value**	**Pr(>|t|)**
Intercept	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	$t_0 = \hat{\beta}_0 \,/\, SE(\hat{\beta}_0)$	p_0
X_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$t_1 = \hat{\beta}_1 \,/\, SE(\hat{\beta}_1)$	p_1
X_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$t_2 = \hat{\beta}_2 \,/\, SE(\hat{\beta}_2)$	p_2
…	…	…	…	…
X_k	$\hat{\beta}_k$	$SE(\hat{\beta}_k)$	$t_k = \hat{\beta}_k \,/\, SE(\hat{\beta}_k)$	p_k

Modelo estimado y contrastes

Dependiente (log(**fev**)) ~ Independientes (**ht** (estatura) **, age** (edad))

	Estimate	**Stand Error**	**t value**	**Pr(>|t|)**
Intercept	-1.9711	0.07833	-25.16	0.00000
ht	0.04399	0.001647	26.71	0.00000
age	0.01981	0.003181	6.23	0.00000

$$\widehat{\log(\text{fev})} = \underset{}{-1.97} + \underset{(0.078)}{0.0439}\ \text{ht} + \underset{(0.0016)}{0.0198}\ \text{age} \qquad \underset{(0.0031)}{}$$

$$\hat{s}_R = 0.1476$$

Descomposición de la variabilidad en regresión

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

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

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

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

$$VT = VE + VNE$$

Coeficiente de determinación R^2

$$\widehat{\log(\text{fev})} = \underset{(0.078)}{-1.97} + \underset{(0.0016)}{0.0439}\,\text{ht} + \underset{(0.0031)}{0.0198}\,\text{age}$$

$$VE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 58.536$$

$$VNE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 13.990$$

$$VT = 58.536 + 13.990 = 72.526$$

$$R^2 = \frac{VE}{VT} = \frac{58.536}{72.526} = 0.8071$$

$$0 \le R^2 \le 1$$

Mide el porcentaje de VT que está explicado por los regresores

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

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

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

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

$$= 1 - (1 - R^2) \times \frac{n-1}{n-k-1}$$

$$\bar{R}^2 = 1 - (1 - 0.8071) \times \frac{653}{651} = 0.8065$$

Contraste general de regresión.

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

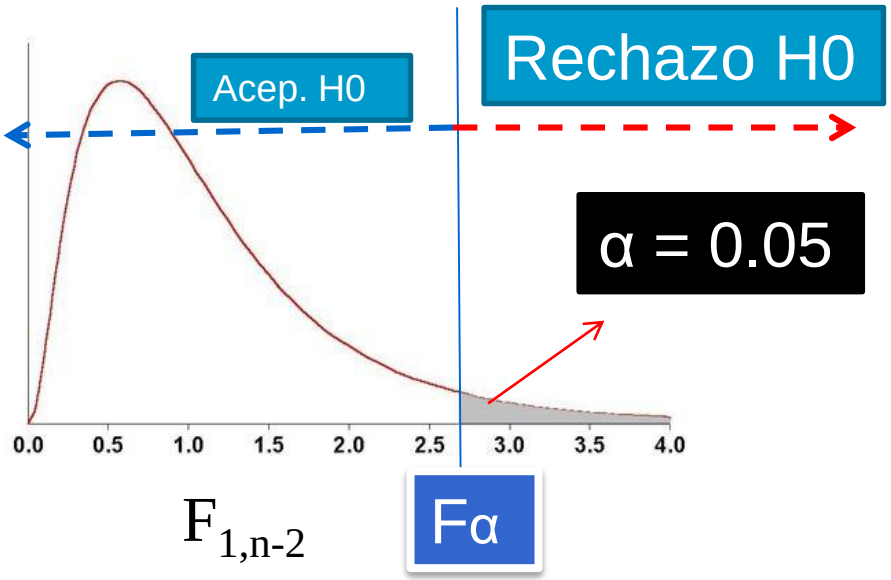
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \text{algunoes distintode } 0$$

$$\hat{s}_E^2 = \frac{VE}{k} \rightarrow \sigma^2 \quad \text{(Si } H_o \text{ es cierto)}$$

$$\hat{s}_R^2 \rightarrow \sigma^2$$

$$F = \frac{\hat{s}_E^2}{\hat{s}_R^2} \rightarrow F_{k,n-k-1}$$

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



Contraste F

$$\widehat{\log(\text{fev})} = \underset{(0.078)}{-1.97} + \underset{(0.0016)}{0.0439}\,\text{ht} + \underset{(0.0031)}{0.0198}\,\text{age}, \qquad \hat{s}_R = 0.1476$$

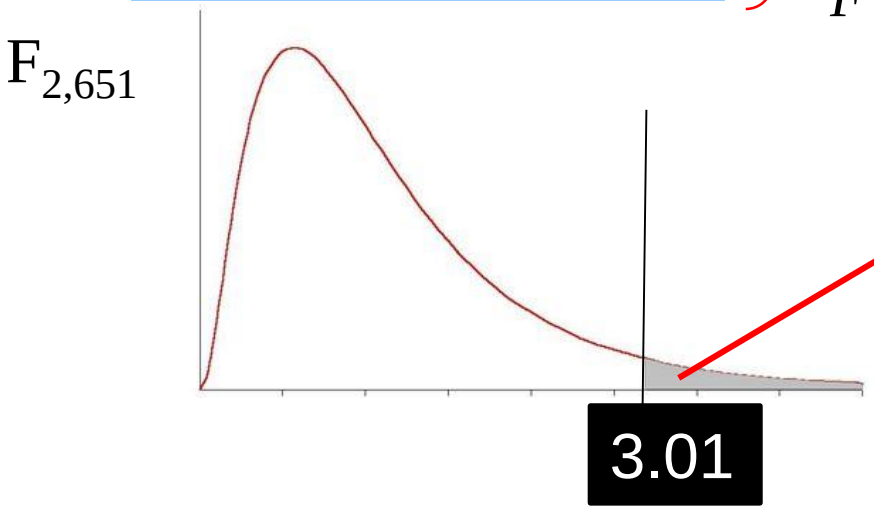
$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{algún } \beta_i \neq 0$$

$$\hat{s}_E^2 = \frac{VE}{k} = \frac{58.436}{2} = 29.268$$

$$\hat{s}_R^2 = 0.021$$

$$F = \frac{\hat{s}_E^2}{\hat{s}_R^2} = \frac{29.268}{0.021} = 1362$$



$1362 > 3.01 \Rightarrow$ Se rechaza H_0

P-valor = 0.00000…

Tabla de *Análisis de la Varianza*

FUENTES	Suma de Cuadrados	Grados de Libertad	Varianzas	F
Explicada (VE)	$\sum(\hat{y}_i - \bar{y})^2$	k	$\hat{s}_E^2$	$\hat{s}_E^2 / \hat{s}_R^2$
Residual (VNE)	$\sum(y_i - \hat{y}_i)^2$	$n-k-1$	$\hat{s}_R^2$	
Total (VT)	$\sum(y_i - \bar{y})^2$	$n-1$		

$$R^2 = \frac{VE}{VT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Tabla de *Análisis de la Varianza*

$$\widehat{\log(\text{fev})} = \underset{(0.078)}{-1.97} + \underset{(0.0016)}{0.0439}\ \text{ht} + \underset{(0.0031)}{0.0198}\ \text{age}, \qquad \hat{s}_R = 0.1476$$

FUENTES	Suma de Cuadrados	Grados de Libertad	Varianzas	F
Explicada (VE)	58.536	2	29.268	1362
Residual (VNE)	13.990	651	0.0215	
Total (VT)	72.526	653		

$$R^2 = \frac{58.536}{72.526} = 0.8071$$

Resumen de estimación con R

```
Call:
lm(formula = log(fev) ~ ht + age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64994 -0.08310  0.01055  0.09324  0.42156

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.971147   0.078332  -25.16  < 2e-16 ***
ht           0.043991   0.001647   26.71  < 2e-16 ***
age          0.019816   0.003181    6.23 8.35e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1466 on 651 degrees of freedom
Multiple R-squared: 0.8071, Adjusted R-squared: 0.8065
F-statistic:  1362 on 2 and 651 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: log(fev)
           Df Sum Sq Mean Sq F value    Pr(>F)
age         1 43.210  43.210  2010.8 < 2.2e-16 ***
ht          1 15.326  15.326   713.2 < 2.2e-16 ***
Residuals 651 13.990   0.021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Ejemplo 1: Cars

Depend | Regresores

Y | X_1 | X_2 | X_3 | X_4

	cons	engine	horse	weight	accel
1	16.79	340.0	160	3609	8.0
2	16.79	440.0	215	4312	8.5
3	15.67	390.0	190	3850	8.5
4	16.79	454.0	220	4354	9.0
5	15.67	400.0	150	3761	9.5
6	14.69	400.0	230	4278	9.5
7	16.79	455.0	225	4425	10.0
8	15.67	383.0	170	3563	10.0
9	15.67	429.0	198	4341	10.0
10	16.79	455.0	225	3086	10.0
11	13.83	302.0	140	3449	10.5
12	18.08	440.0	215	4735	11.0
...					
388	10.22	97.0	54	2254	23.5
389	5.42	90.0	48	2335	23.7
390	5.34	97.0	52	2130	24.6
391	8.64	141.0	71	3190	24.8

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0517457  0.9785989  -1.075   0.2832
engine       0.0058784  0.0026166   2.247   0.0252 *
horse        0.0369513  0.0065584   5.634 3.40e-08 ***
weight       0.0020182  0.0003132   6.443 3.49e-10 ***
accel        0.0813370  0.0493374   1.649   0.1000
```

$$\widehat{cons} = -1.05 + 0.0058\ \text{engine} + 0.0369\ \text{horse} + \\ + 0.0020\ \text{weight} + 0.0813\ \text{accel}$$

Valores Previstos y Residuos

$$\widehat{cons} = -1.05 + 0.0058\,\text{engine} + 0.0369\,\text{horse} + 0.0020\,\text{weight} + 0.0813\,\text{accel}$$

Datos | Resultados

Y X1 X2 X3 X4 $\hat{Y}$ e

	cons	engine	horse	weight	accel	cons_pred	residuo
1	16.79	340.0	160	3609	8.0	14.79	2.00
2	16.79	440.0	215	4312	8.5	18.87	-2.08
3	15.67	390.0	190	3850	8.5	16.72	-1.05
4	16.79	454.0	220	4354	9.0	19.27	-2.47
5	15.67	400.0	150	3761	9.5	15.21	0.47
6	14.69	400.0	230	4278	9.5	19.20	-4.51
7	16.79	455.0	225	4425	10.0	19.68	-2.89
8	15.67	383.0	170	3563	10.0	15.49	0.19
9	15.67	429.0	198	4341	10.0	18.36	-2.69
10	16.79	455.0	225	3086	10.0	16.98	-0.19
11	13.83	302.0	140	3449	10.5	13.71	0.12
12	18.08	440.0	215	4735	11.0	19.93	-1.85

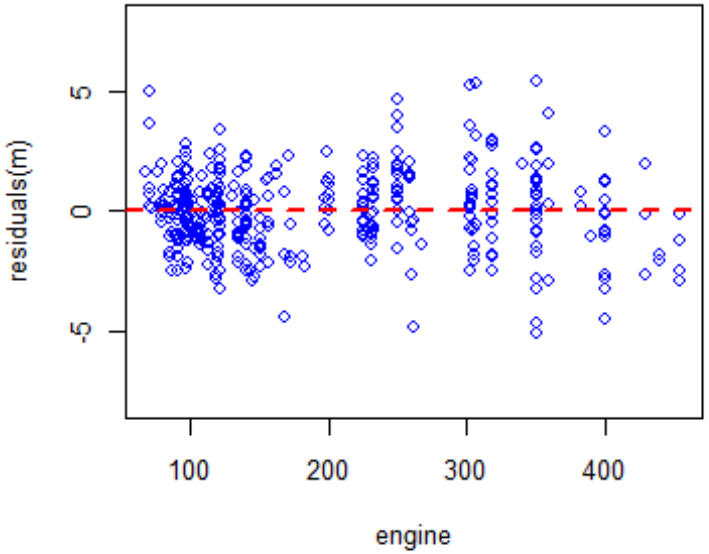
$$VE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 4725.0$$

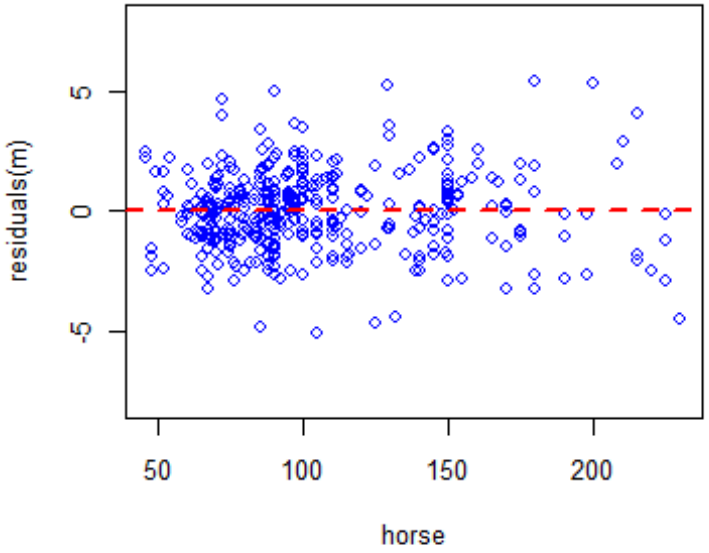
$$VNE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 1037.9$$

$$VT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 5762.9$$

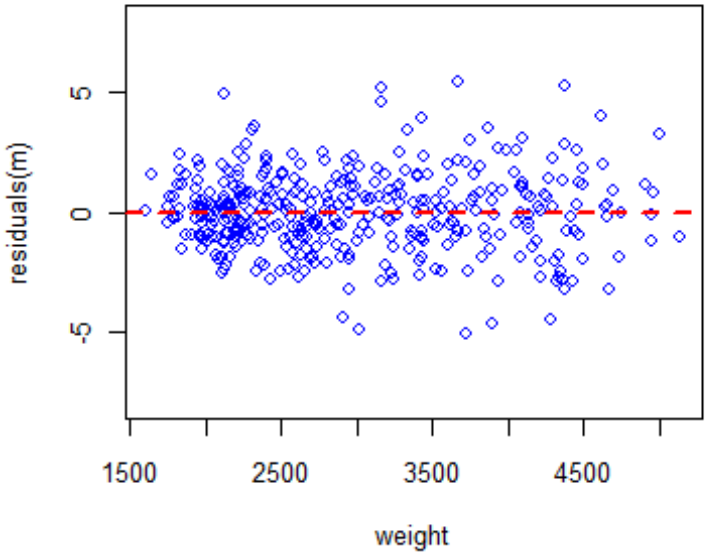
$$^2\hat{s}_R = \frac{VNE}{n-k-1} = \frac{1037.9}{386} = 2.7$$

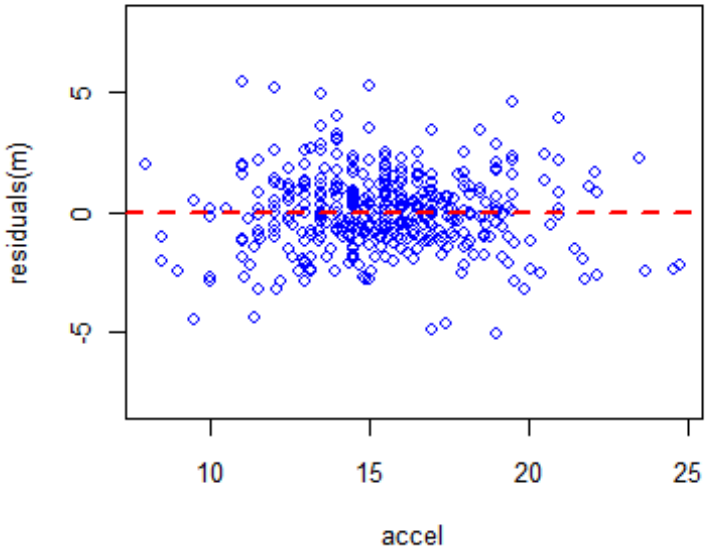
$$R^2 = \frac{VE}{VT} = \frac{4725}{5762.9} = 81.99$$





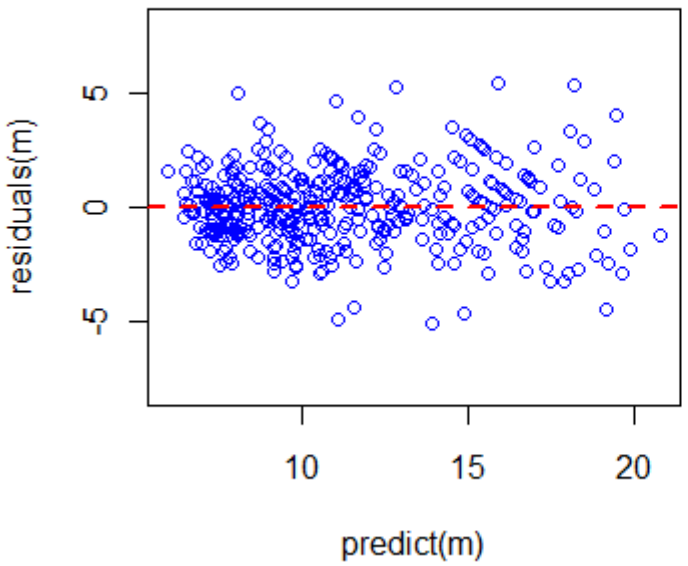
DIAGNOSIS: residuos ~ regresores



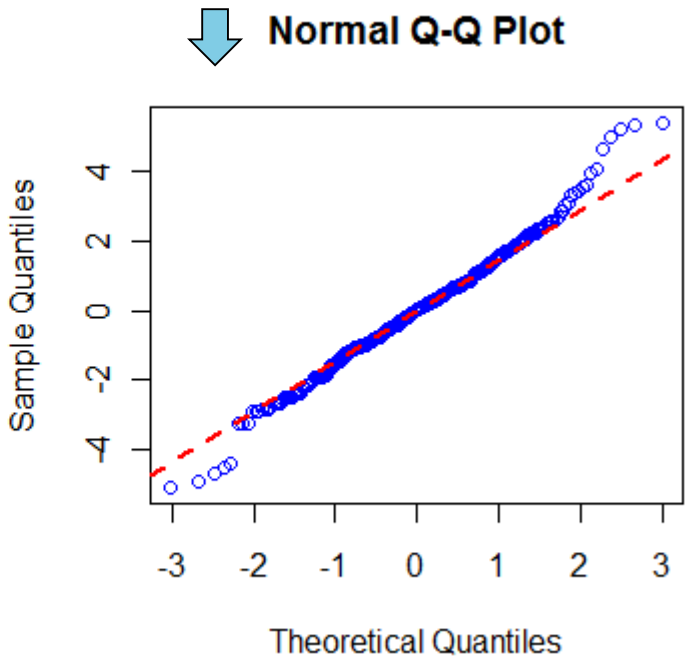


Diagnosis

Linealidad
Homocedasticidad
ok



Normalidad
ok



Resumen del modelo

```
lm(formula = cons ~ engine + horse + weight + accel)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1103 -1.0142  0.0174  0.9451  5.4266

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0517457  0.9785989  -1.075   0.2832
engine       0.0058784  0.0026166   2.247   0.0252 *
horse        0.0369513  0.0065584   5.634 3.40e-08 ***
weight       0.0020182  0.0003132   6.443 3.49e-10 ***
accel        0.0813370  0.0493374   1.649   0.1000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.64 on 386 degrees of freedom
Multiple R-squared: 0.8199, Adjusted R-squared: 0.818
F-statistic: 439.2 on 4 and 386 DF,  p-value: < 2.2e-16
```

Resumen del modelo

```
lm(formula = cons ~ engine + horse + weight)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8970 -1.0338  0.0543  0.8911  5.6837

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3631739  0.4711933   0.771   0.4413
engine      0.0052017  0.0025899   2.008   0.0453 *
horse       0.0299518  0.0050097   5.979 5.11e-09 ***
weight      0.0022568  0.0002784   8.107 6.89e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.643 on 387 degrees of freedom
Multiple R-squared: 0.8186, Adjusted R-squared: 0.8172
F-statistic: 582.1 on 3 and 387 DF,  p-value: < 2.2e-16
```

Conclusiones modelo final

$$\widehat{cons} = -1.05 + 0.0058\,\text{engine} + 0.0369\,\text{horse} + 0.0020\,\text{weight} + 0.0813\,\text{accel}$$

$\hat{s}_R = 1.64$ $\quad R^2 = 81.99$

1. No se aprecian desviaciones importantes de las hipótesis básicas del modelo: linealidad, homocedasticidad y normalidad.

2. Se observa relación lineal significativa entre el consumo de los coches y su peso (weight), potencia (horse) y centímetros cúbicos (engine). (Los p-valores son menores que 0.05 en elmodelos). Los coeficientes estimados son positivos, lo que significa que el aumento de cualquiera de las variables independientes incrementa el consumo del vehículo. Con las cuatro variables se explica el 81.99 % de la variabilidad del consumo.

Conclusiones modelo final (cont)

3. En el modelo de cuatro regresores el parámetro asociado a aceleración no es significativo. La inclusión de la variable "aceleración" no mejora significativamente el modelo. Eso no implica que no exista relación lineal entre aceleración y consumo (la regresión simple entre estas variables indican relación significativa con coeficiente negativo).

4. El coeficiente asociado al peso es 0.0020, es muy significativo. Para interpretarlo es necesario tener en cuenta las unidades: un aumento de una libra en el peso del coche manteniendo constante el resto de las variables produce un aumento del consumo de 0.002 litros/100 km. (Esto implica que un regresor se puede cambiar manteniendo el resto constante, lo que sólo es posible en los estudios experimentales.) El resto de los coeficientes se interpreta similarmente.

CARS: Todos los modelos

Modelo	1 engine	2 horse	3 weight	4 accel	$\hat{s}_R$	R^2	$\bar{R}^2$
1	**0,032** 0,0009				**1,874**	**76,28**	**76,22**
2		**0,085** 0,0026			**2,002**	**72,94**	**72,87**
3			**0,004** 0,0001		**1,780**	**78,55**	**78,49**
4				**-0,663** 0,062	**3,380**	**22,70**	**22,50**
12	**0,0202** 0,0019	**0,036** 0,0053			**1,775**	**78,78**	**78,67**
13	**0,01313** 0,0023		**0,00251** 0,0002872		**1,715**	**80,18**	**80,08**
14	**0,03215** 0,00108			**0,0048** 0,041	**1,877**	**76,28**	**76,16**
23		**0,0351** 0,00432	**0,0026** 0,00019		**1,650**	**81,67**	**81,58**
24		**0,1027** 0,0035		**0,336** 0,048	**1,892**	**75,90**	**75,78**
34			**0,00379** 0,0001147	**-0,1689** 0,0351	**1,734**	**79,75**	**79,65**
123	**0,0052** 0,0025	**0,0299** 0,005	**0,00225** 0,0002		**1,643**	**81,86**	**81,72**
124	**0,01765** 0,0019	**0,0539** 0,0063		**0,2282** 0,0459	**1,723**	**80,05**	**79,89**
134	**0,01006** 0,0026		**0,0027** 0,000298	**-0,0986** 0,039	**1,704**	**80,50**	**80,35**
234		**0,04113** 0,0063	**0,0025** 0,00022	**0,0639** 0,0489	**1,648**	**81,75**	**81,61**
1234	**0,00587** 0,0026	**0,03695** 0,0065	**0,002018** 0,00031	**0,0813** 0,049	**1,640**	**81,99**	**81,80**

Conclusiones Generales

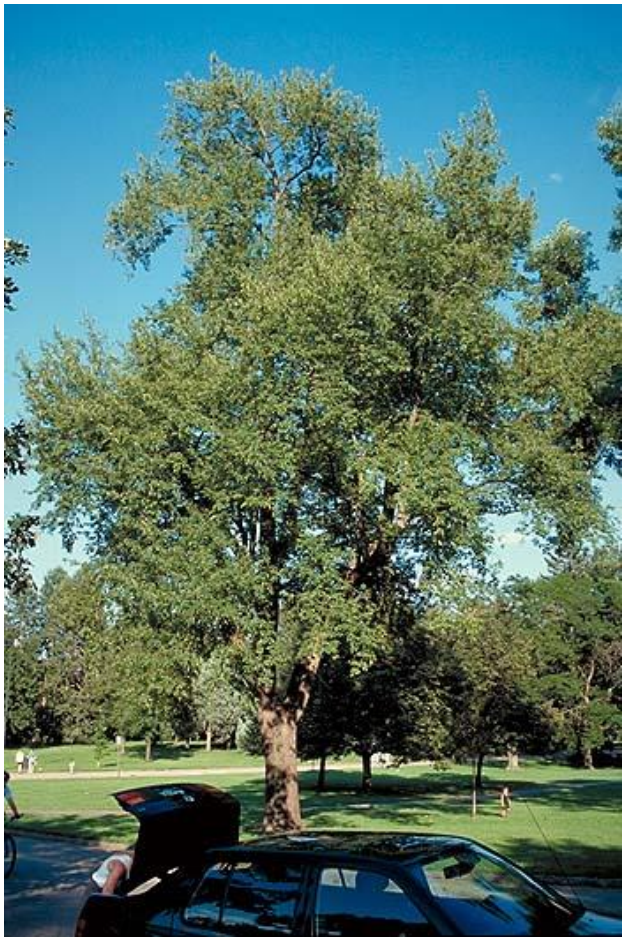
1. El que la relación lineal entre dos variables sea significativa **no implica** que exista relación de CAUSALIDAD entre las variables. Se debe interpretar como asociación entre las variables: los coches con más pesos presentan mayor consumo que los coches con menos peso.

2. Cuando se añaden o eliminan variables de un modelo los coeficientes del resto cambian. Eso es debido a la correlación entre los regresores. Cuando estas correlaciones son altas los coeficientes pueden cambiar mucho, incluso de signo. Esto se puede apreciar en el coeficiente de la variable *accel*, cuyo efecto sobre el cosnumo depende del resto de las variables en el modelo. La alta correlación entre los regresores hace muy difícil interpretar el significado de los coeficientes, a este problema se le denomina MULTICOLINEALIDAD.

Conclusiones (cont.)

7. La selección del modelo depende del objetivo. Siempre el modelo con más regresores tiene el mayor R2. Utilizando el "R2 corregido" hay tres modelos muy parecidos 23, 123 y 1234. El mejor modelo con un regresor es el 3, con R2 igual al 78.55%, al incluir la pontencia (horse) como nuevo regresor tenemos el modelo 23 cuyo R2 sólo aumenta un 3%, hasta 81.67%. El modelo 123, incluye además los cc del motor (engine) como regresor con un aumento en R2 despreciable (ahora 81.86%). En este modelo los tres coeficientes son significativos. Si añadimos la variable accel, llegamos al modelo completo con R2 igual a 81.99%. El coeficiente de la última variable no es significativo.
8. Al ir incluyendo regresores en un modelo los residuos van disminuyendo y con ello la variabilidad no explicada. La desviación típica residual también suele disminuir (hay que tener en cuenta que el denominador de la varianza residual también disminuye). Los modelos 23, 123 y 1234 tienen una desviación típica residual muy parecida y próxima a 1.64 litros/100km. La interpretación (aproximada) es la siguiente (con el modelo 1234): si nos proporcionan los datos del peso (weight), potencia (horse), cc (engine) y aceleración (accel) del coche la distribución de su consumo tiene media la proporcionada por el modelo y desviación típica 1.64 litros/100km.

Ejemplo 2: Cerezos Negros



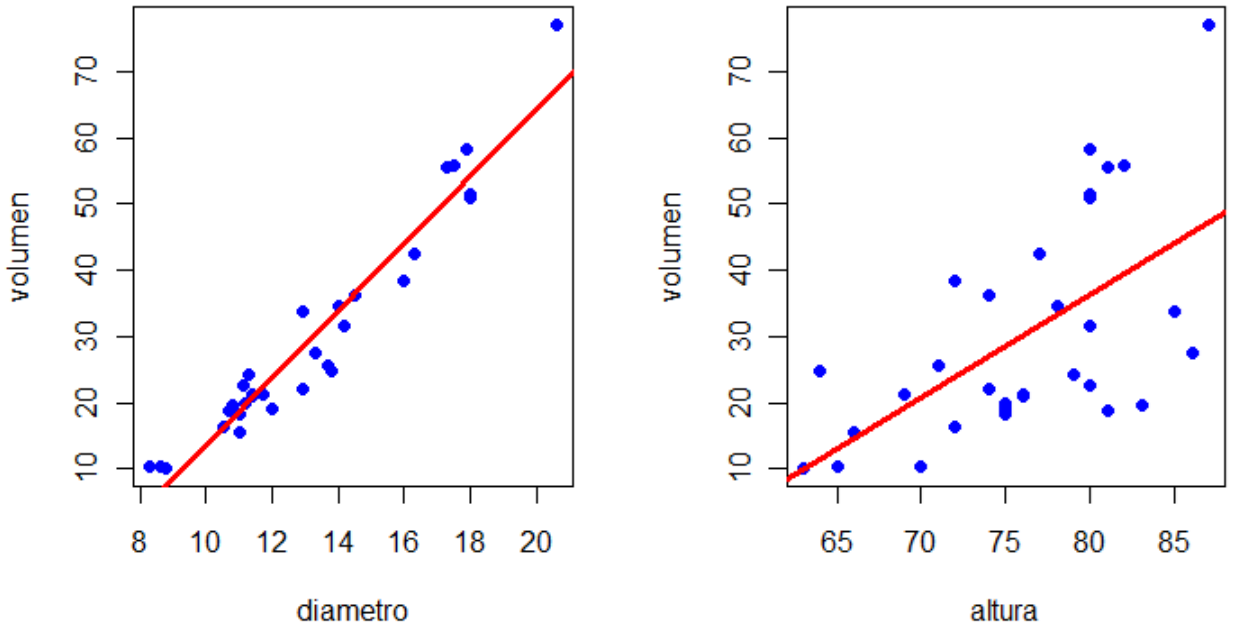
Se desea construir un modelo de regresión para obtener el **volumen** de madera de una "*cerezo negro*" en función de la altura del tronco y del diámetro del mismo a un metro sobre el suelo. Se ha tomado una muestra de 31 árboles. Las unidades de longitudes son *pies* y de volumen *pies cúbicos*.

Cerezos negros: Datos

Árbol	Diametro	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	Diametro	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



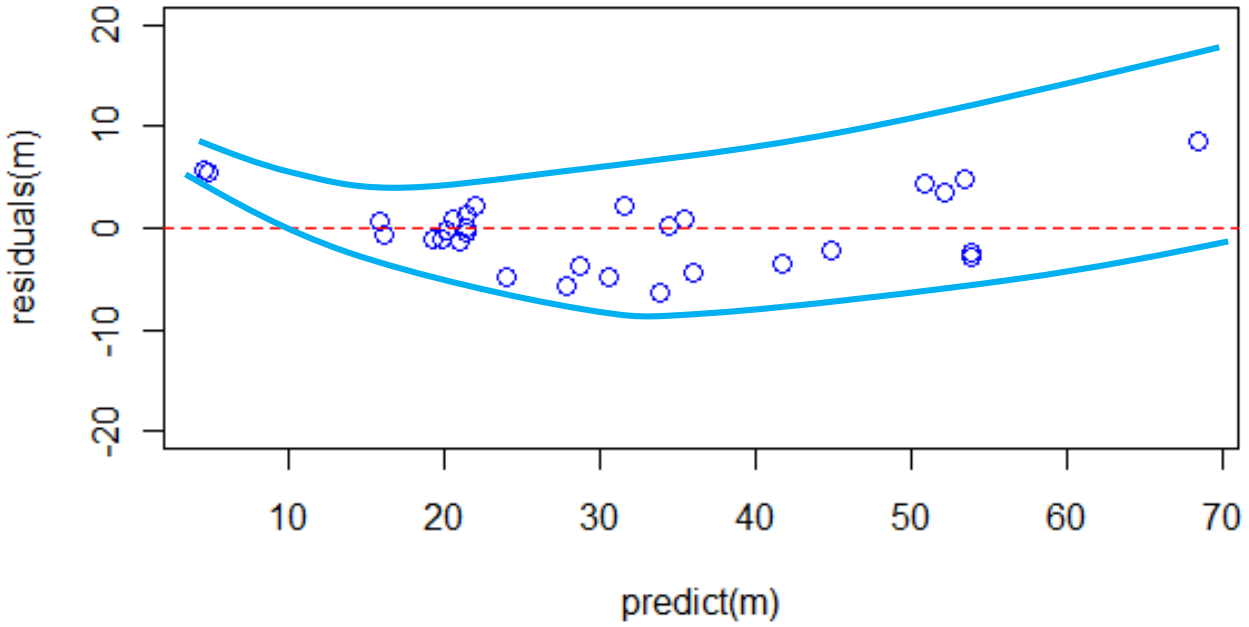
1. Se aprecia relación entre las dos variables y el volumen
2. El gráfico del volumen versus diámetro presenta ligera curvatura
3. El gráfico del volumen versus altura presenta clara heterocedasticidad

Primer modelo:cerezos negros

$$\text{Volumen} = \beta_0 + \beta_1 \text{Diametro} + \beta_2 \text{Altura} + \text{Error}$$

```
Call:
lm(formula = volumen ~ diametro + altura)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
diametro      4.7082     0.2643  17.816  < 2e-16 ***
altura        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Diagnosis



Indicios de falta de linealidad

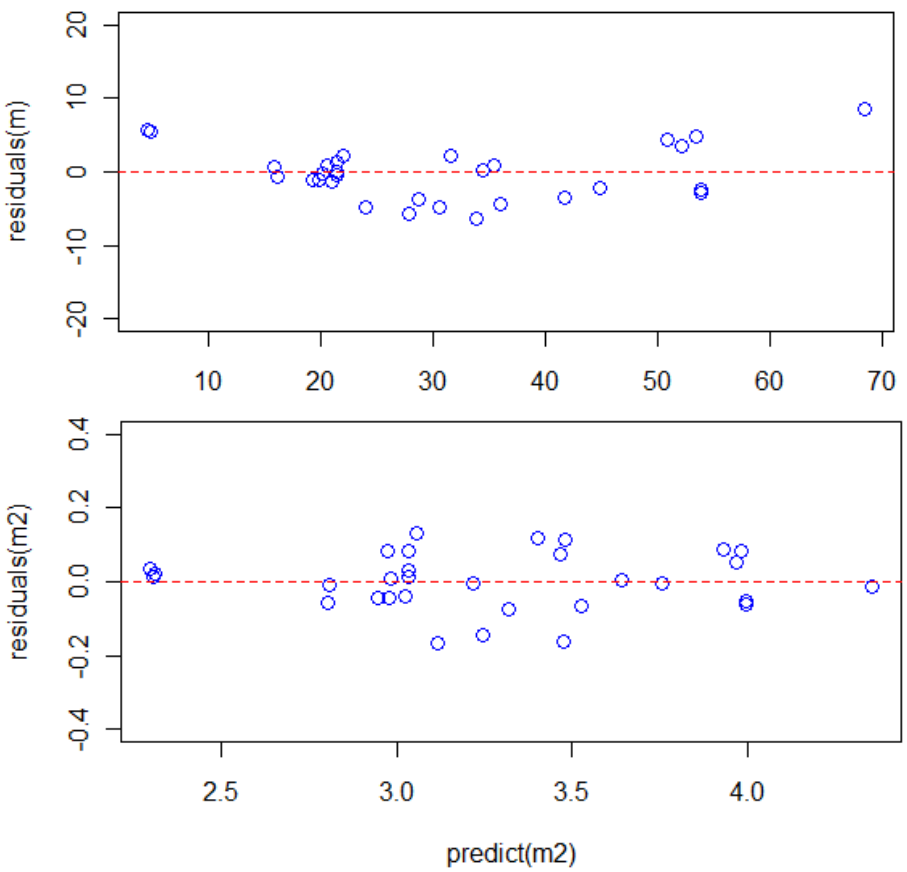
Transformación

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

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

```
> mod2<-lm(log(volumen)~log(diametro)+log(altura))
> summary(mod2)

Call:
lm(formula = log(volumen) ~ log(diametro) + log(altura))

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
log(diametro)  1.98265    0.07501  26.432  < 2e-16 ***
log(altura)    1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared: 0.9777,     Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

Diagnosis (modelo transformado)



Antes

Ahora

Interpretación

- Se comprueba gráficamente que la distribución de los residuos es compatible con las hipótesis de linealidad y homocedasticidad.
- El volumen está muy relacionada con la altura y el diámetro del árbol (R^2= 97.77%)
- 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 desviación típica residual es s_R=0.081 que indica que el error relativo del modelo en la predicción del volumen es del 8.1%.

Ejemplo 3: Tabaco

	marca	alq	nico	co
1	Alpine	14.1	0.86	13.6
2	Benson_&_Edges	16.0	1.06	16.6
3	Bull_Durham	29.8	2.03	23.5
4	Camel_lights	8.0	0.67	10.2
5	Carlton	4.1	0.40	5.4
6	Chesterfield	15.0	1.04	15.0
7	Golden_lights	8.8	0.76	9.0
8	Kent	12.4	0.95	12.3
9	Kool	16.6	1.12	16.3
10	L&M	14.9	1.02	15.4
11	Lark_lights	13.7	1.01	13.0
12	Marlboro	15.1	0.90	14.4
13	Merit	7.8	0.57	10.0
14	Multi_Filter	11.4	0.78	10.2
15	Newport_lights	9.0	0.74	9.5
16	Now	1.0	0.13	1.5
17	Old_Gold	17.0	1.26	18.5
18	Pall_Mall_lights	12.8	1.08	12.6
19	Raleigh	15.8	0.96	17.5
20	Salem_Ultra	4.5	0.42	4.9
21	Tareyton	14.5	1.01	15.9
22	True	7.3	0.61	8.5
23	Viceroy_Rich_light	8.6	0.69	10.6
24	Virginia_Slims	15.2	1.02	13.9
25	Winston_lights	12.0	0.82	14.9

Ejemplo "Tabaco" Monóxido de Carbono (CO)
25 observaciones, 3 variables

Descripción: Se proporciona la producción de monóxido de carbono (co) y el contenido de nicotina (nico) y alquitrán (alq) en 25 marcas diferentes de cigarrillos americanos.

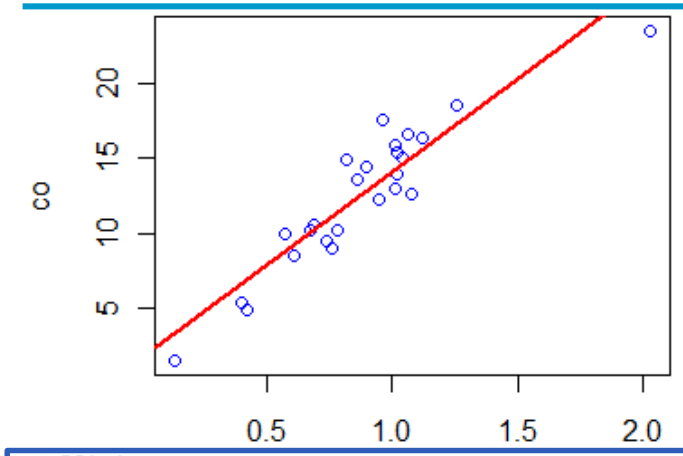
Fuente: Mendenhall, William, and Sincich, Terry (1992), Statistics for Engineering and the Sciences (3rd ed.), New York: (Original source: Federal Trade Commission, USA)

Variables

alq	contenido en alquitrán mg
nico	contenido en nicotina mg
co	monóxido de carbono CO mg

Objetivo: Estudiar la relación entre CO con alquitrán y nicotina

CO ~ nico

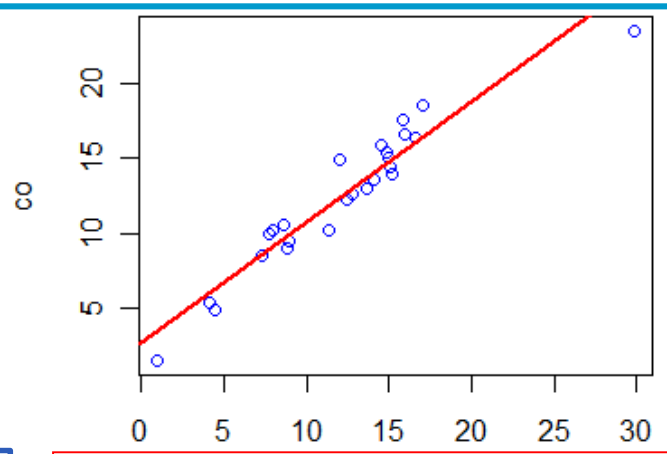



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.6647     0.9936   1.675    0.107
nico          12.3954     1.0542  11.759 3.31e-11 ***
```

$R^2 = 85.74$

$\hat{s}_R = 1.828$

CO ~ alq




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.74328    0.67521   4.063 0.000481 ***
alq          0.80098    0.05032  15.918 6.55e-14 ***
---
```

$R^2 = 91.68$

$\hat{s}_R = 1.397$

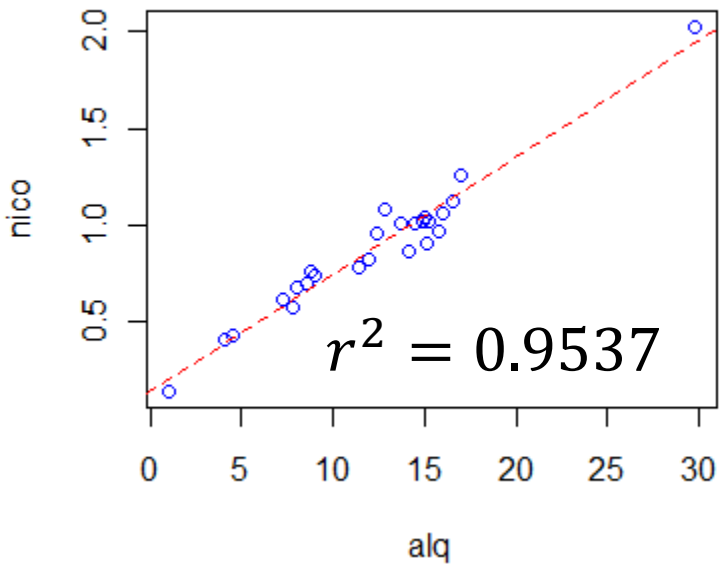
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0896     0.8438   3.662 0.001371 **
nico         -2.6463     3.7872  -0.699 0.492035
alq           0.9625     0.2367   4.067 0.000512 ***
```

$R^2 = 91.86$

$\hat{s}_R = 1.413$

Efecto de la multicolinealidad

(alta correlación entre nico y alq)



- El coeficiente de la variable “nico” cambia de 12.39 a -2.36.
- En el modelo con dos regresores, el coeficiente de la variable “nico” no es significativo.
- Los standard errors de los coeficientes en el modelo de dos regresores han aumentado considerablemente respecto a los de regresión simple. El de “nico” pasa de 1.05 a 3.78. El cambio para “alq” es mayor.
- Los estadísticos t se han reducido (debido al aumento de los standards errors)
- La desviación típica residual del modelo con dos regresores es mayor que en el modelo de regresión simple “CO ~ alq”

APÉNDICE

Notación matricial

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$$

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

Estimación mínimo-cuadrática

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

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

donde el vector $\mathbf{e}$ cumple

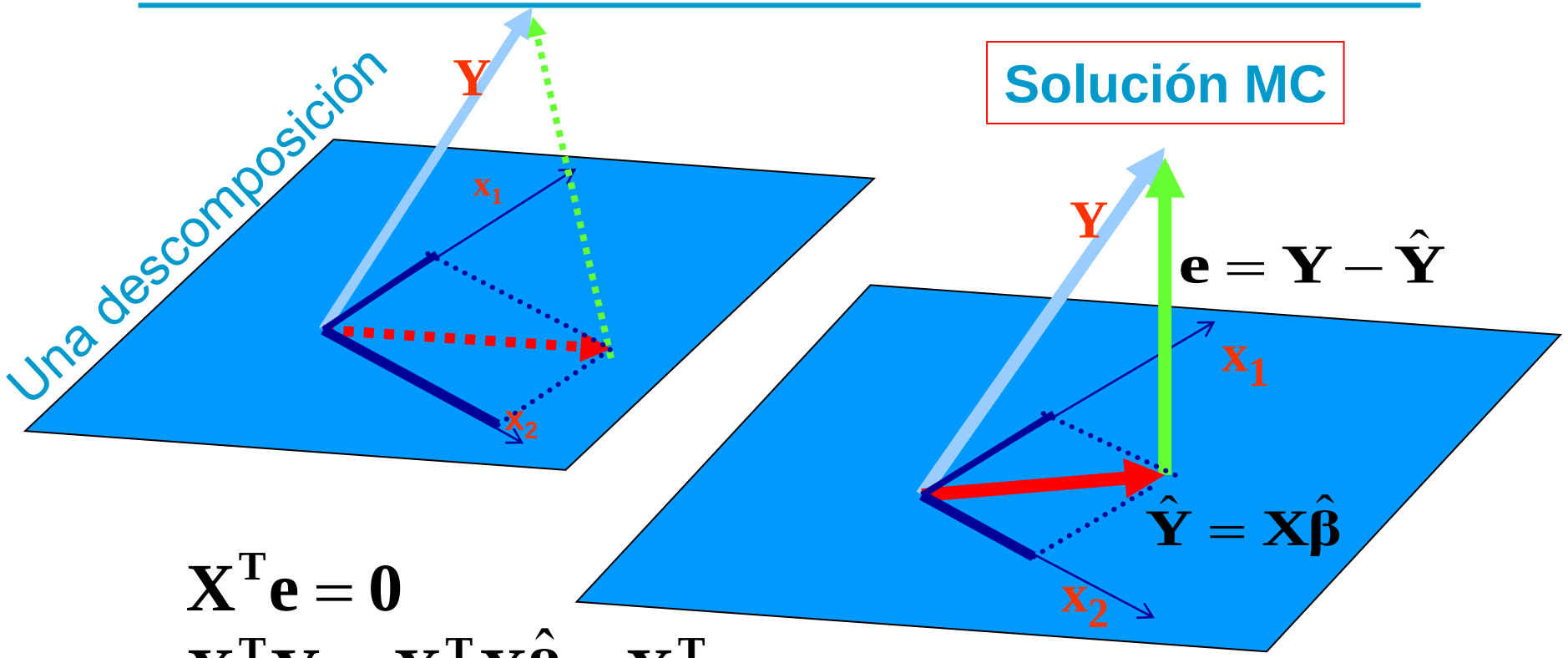
$$\|\mathbf{e}\|^2 = \sum_{i=1}^{n} e_i^2 \quad \text{es mínimo}$$

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

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

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

Mínimos cuadrados

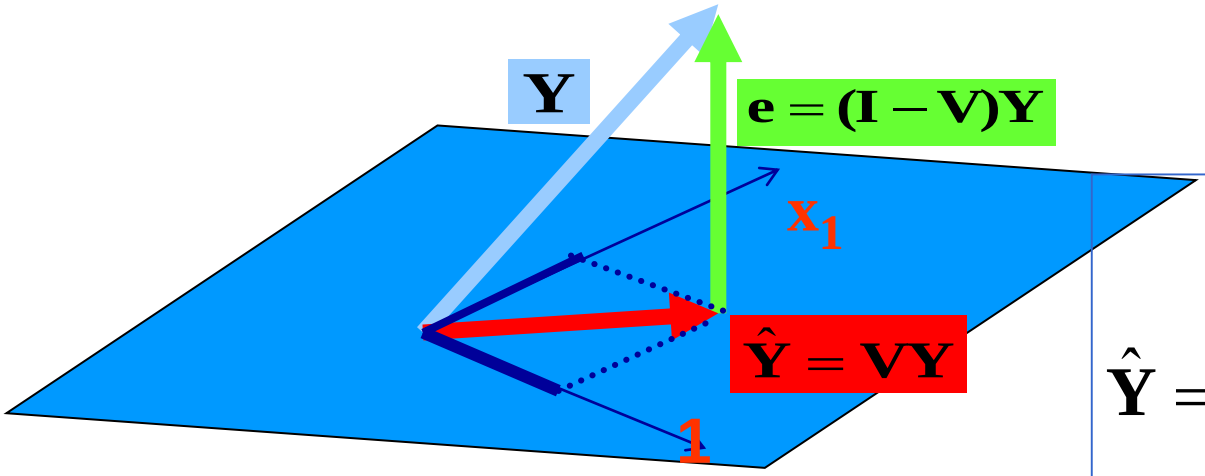


$$\mathbf{X}^T\mathbf{e} = \mathbf{0}$$

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

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

Matriz de proyección $\mathbf{V}$



Val.Previstos

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

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

$$\hat{\mathbf{Y}} = \mathbf{V}\mathbf{Y}$$

Residuos

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

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

Simétrica $\mathrm{V}=\mathrm{V}^{\mathrm{T}}$

Idempotente $\mathrm{VV}=\mathrm{V}$

Distribución de probabilidad de $\hat{\boldsymbol{\beta}}$

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

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

$$\hat{\boldsymbol{\beta}} \rightarrow Normal$$

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

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

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

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

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

Distribución de probabilidad de $\hat{\boldsymbol{\beta}}$

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

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

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

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

Residuos

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

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

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

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

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

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

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

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

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

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

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

$$\tilde{\mathbf{Y}} = \begin{pmatrix} y_1 - y \\ y_2 - y \\ \vdots \\ y_n - y \end{pmatrix}, \quad \overline{\mathbf{Y}} = \begin{pmatrix} y \\ y \\ \vdots \\ y \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - x_1 & x_{21} - x_2 & \cdots & x_{k1} - x_k \\ x_{12} - x_1 & x_{22} - x_2 & \cdots & x_{k2} - x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} - x_1 & x_{2n} - x_2 & \cdots & x_{kn} - x_k \end{pmatrix}$$

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

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

Regresión con R

```
> mod_coches<-lm(Consumo ~ CC + CV + Peso + Acel)
> summary(mod_coches)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.1095 -1.0180  0.0322  0.9906  5.5775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.6695833  0.9833053  -1.698   0.0903 .
CC           0.0003835  0.0001625   2.360   0.0188 *
CV           0.0402844  0.0065697   6.132 2.15e-09 ***
Peso         0.0057842  0.0009578   6.039 3.65e-09 ***
Acel         0.1115012  0.0496757   2.245   0.0254 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.662 on 386 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.8197,     Adjusted R-squared: 0.8178
F-statistic: 438.7 on 4 and 386 DF,  p-value: < 2.2e-16
```

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 = 0.8197)

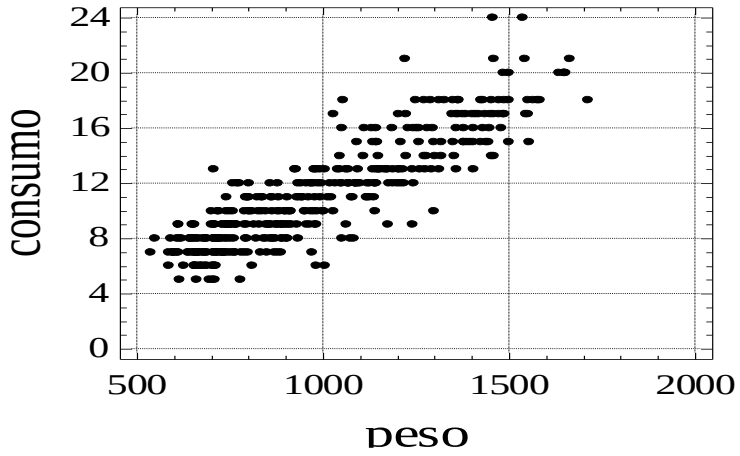
Multicolinealidad

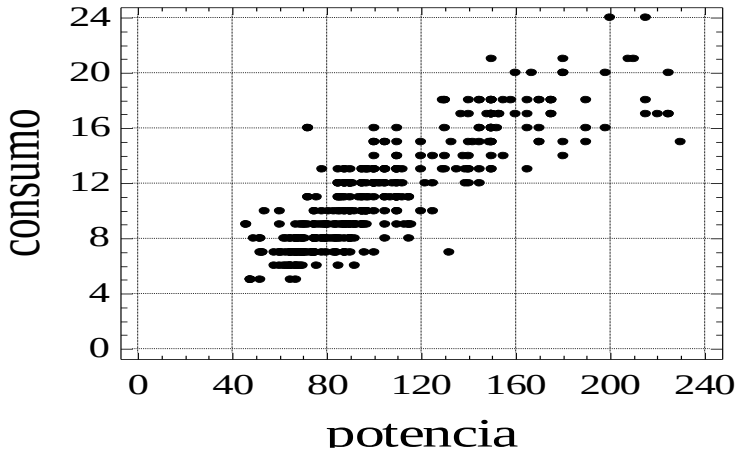
- Cuando la correlación entre los regresores es alta.
- Presenta graves inconvenientes:
  - Empeora las estimaciones de los efectos de cada variable β_i: aumenta la varianza de las estimaciones y la dependencia de los estimadores)
  - Dificulta la interpretación de los parámetros del modelo estimado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

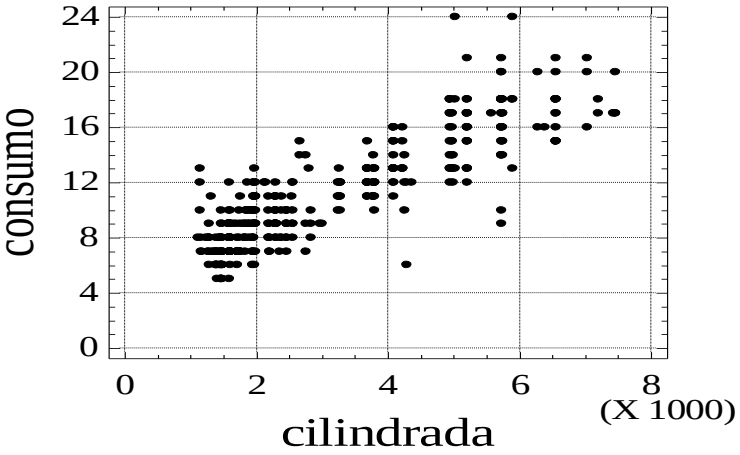
Identificación de la multicolinealidad:
Matriz de correlación de los regresores.

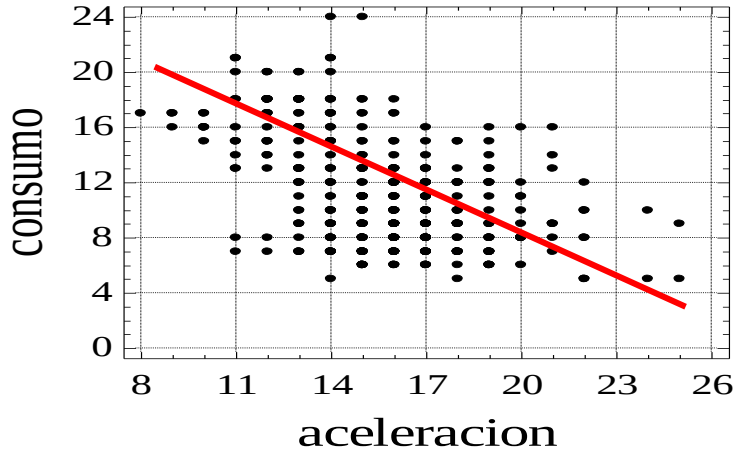
```
> cor(coches[,2:5],use='pair')
              CC         CV       Peso       Acel
CC     1.0000000  0.8984444  0.9336283 -0.5474669
CV     0.8984444  1.0000000  0.8628927 -0.6962805
Peso   0.9336283  0.8628927  1.0000000 -0.4212948
Acel -0.5474669 -0.6962805 -0.4212948  1.0000000
```

Gráficos consumo - x_i









Consumo y aceleración

R. simple

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.50322    1.00418   21.41   <2e-16 ***
Acel        -0.65591    0.06311  -10.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.442 on 391 degrees of freedom
Multiple R-squared: 0.2165,     Adjusted R-squared: 0.2145
F-statistic:   108 on 1 and 391 DF,  p-value: < 2.2e-16
```

R. múltiple

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.6695833  0.9833053  -1.698   0.0903 .
CC           0.0003835  0.0001625   2.360   0.0188 *
CV           0.0402844  0.0065697   6.132 2.15e-09 ***
Peso         0.0057842  0.0009578   6.039 3.65e-09 ***
Acel         0.1115012  0.0496757   2.245   0.0254 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.662 on 386 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.8197,     Adjusted R-squared: 0.8178
F-statistic: 438.7 on 4 and 386 DF,  p-value: < 2.2e-16
```

Multicolinealidad: efecto en la varianza de los estimadores

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

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

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

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

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.