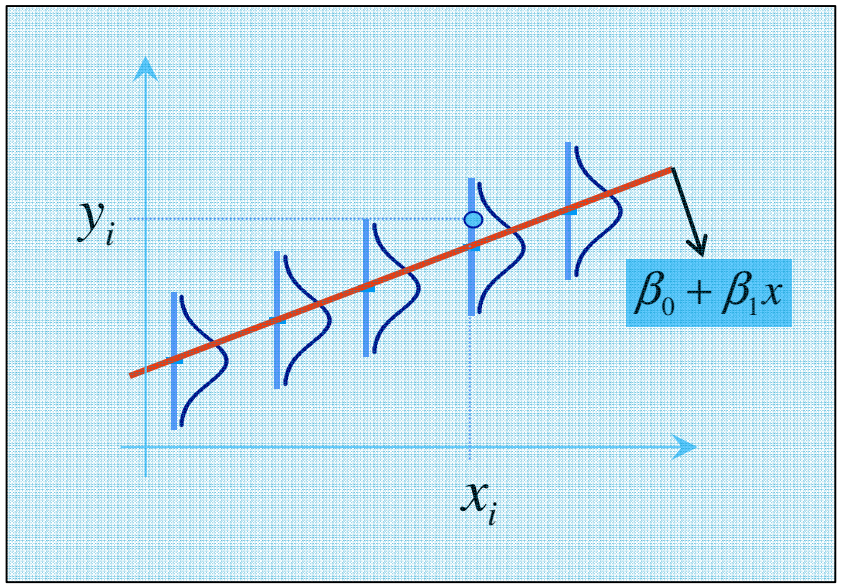
Módulo 3: Regresión

Lección 2: Regresión simple II (Diagnosis y Transformaciones)

Diagnosis del Modelo

La estimación está basada en las siguientes hipótesis:

- Linealidad
- Normalidad
- Homocedasticidad
- Independencia



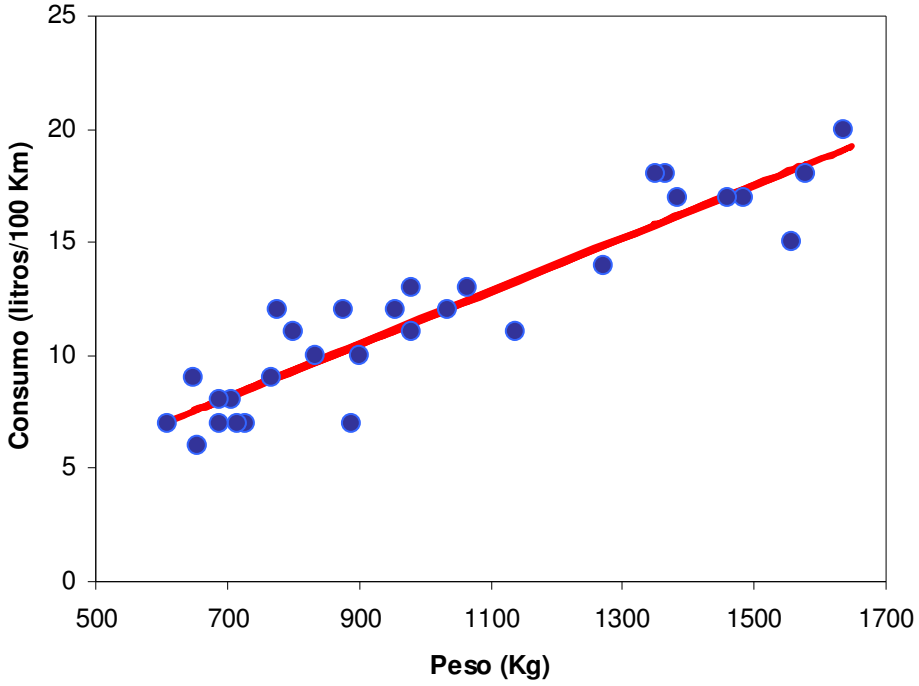
➢ Observaciones Atípicas (muy perjudiciales)

Las hipótesis se comprueban con los RESIDUOS

Análisis de los Residuos

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

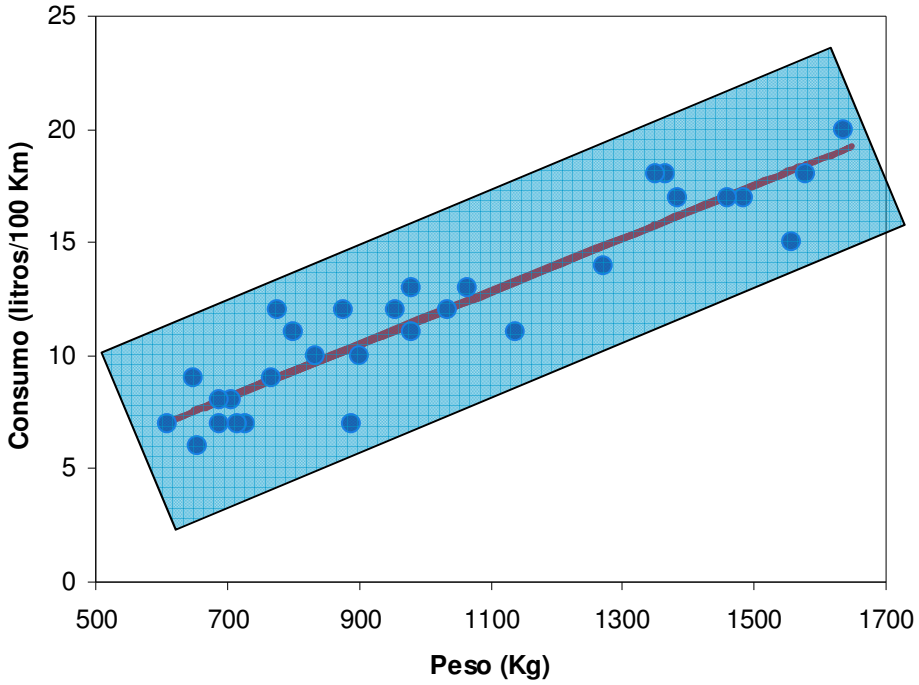
$$e_i = y_i - \hat{y}_i$$



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

Diagnosis del Modelo

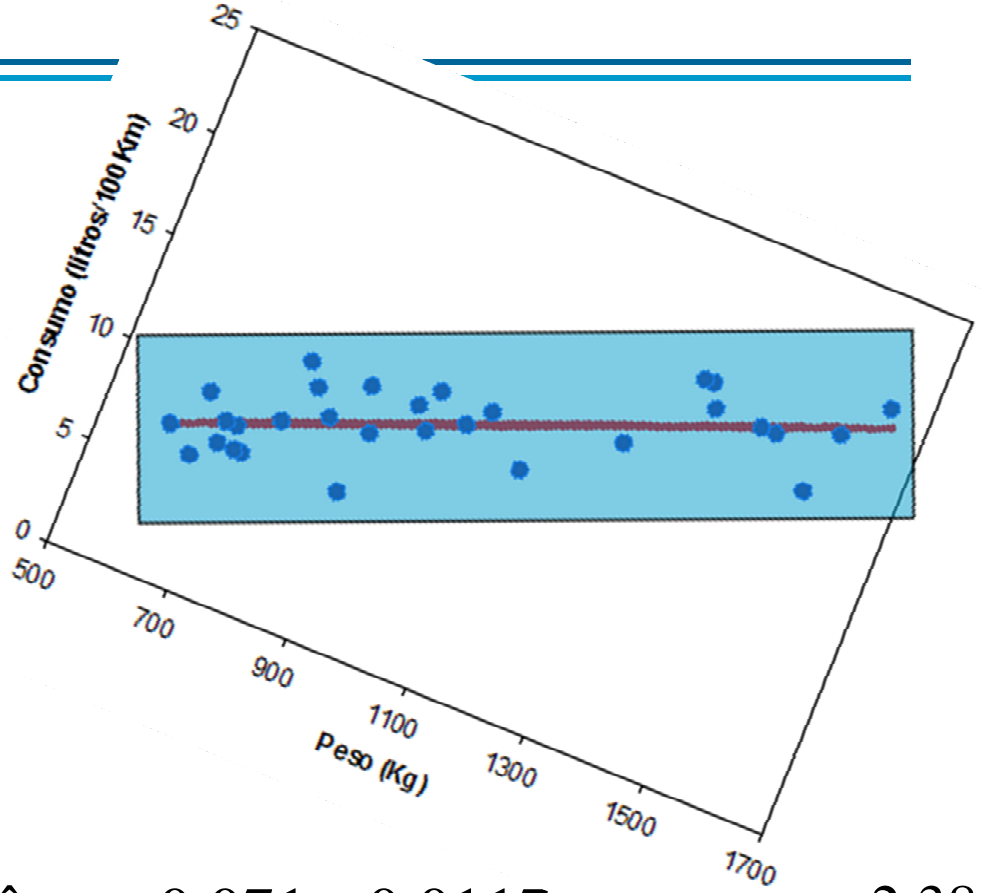
Núm. Obs. (i)	Peso kg	Consumo litros/100 km	Predicción	Residuos
1	981	11	11,44	-0,44
2	878	12	10,23	1,77
3	708	8	8,23	-0,23
4	1138	11	13,28	-2,28
5	1064	13	12,41	0,59
6	655	6	7,61	-1,61
7	1273	14	14,86	-0,86
8	1485	17	17,35	-0,35
9	1366	18	15,95	2,05
10	1351	18	15,78	2,22
11	1635	20	19,11	0,89
12	900	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



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

Diagnosis del Modelo

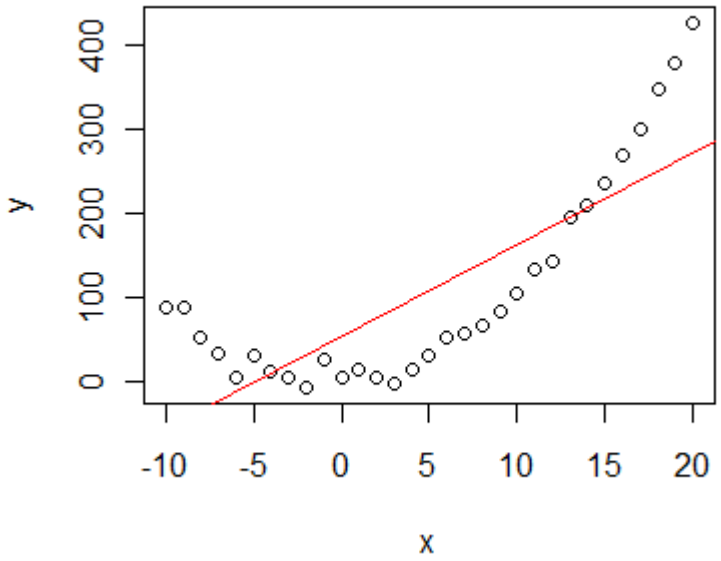
Núm. Obs. (i)	Peso kg	Consumo litros/100 km	Predicción	Residuos
1	981	11	11,44	-0,44
2	878	12	10,23	1,77
3	708	8	8,23	-0,23
4	1138	11	13,28	-2,28
5	1064	13	12,41	0,59
6	655	6	7,61	-1,61
7	1273	14	14,86	-0,86
8	1485	17	17,35	-0,35
9	1366	18	15,95	2,05
10	1351	18	15,78	2,22
11	1635	20	19,11	0,89
12	900	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	-1,4
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

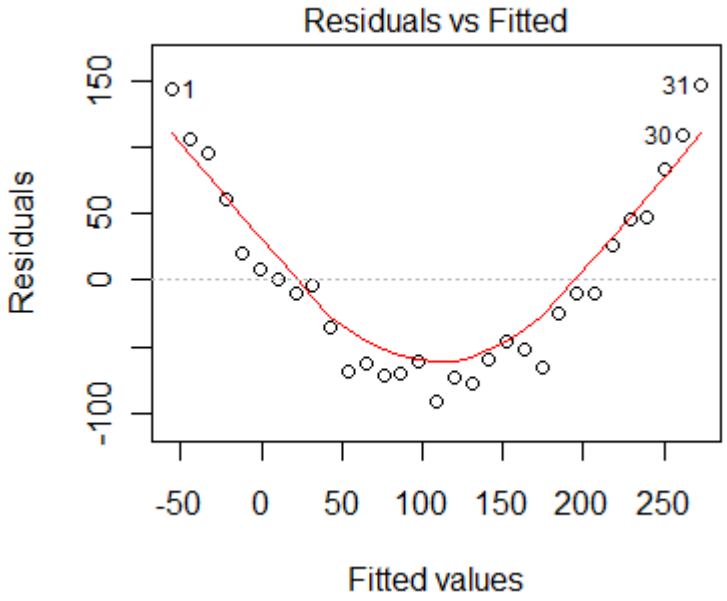


$$\hat{y}_i = -0.071 + 0.0117 x_i$$

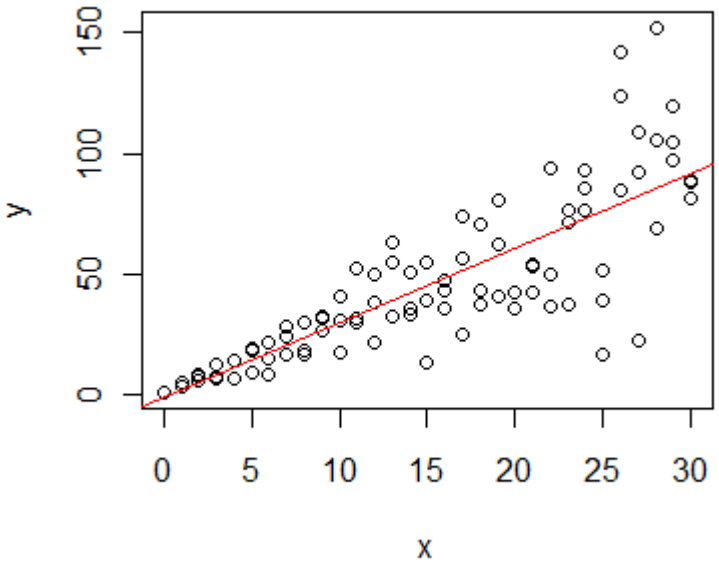
$= 2.38$

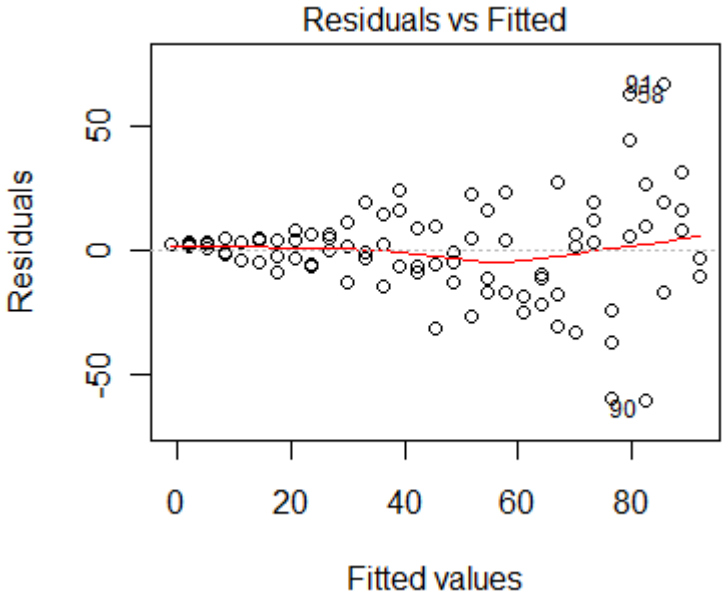
No linealidad



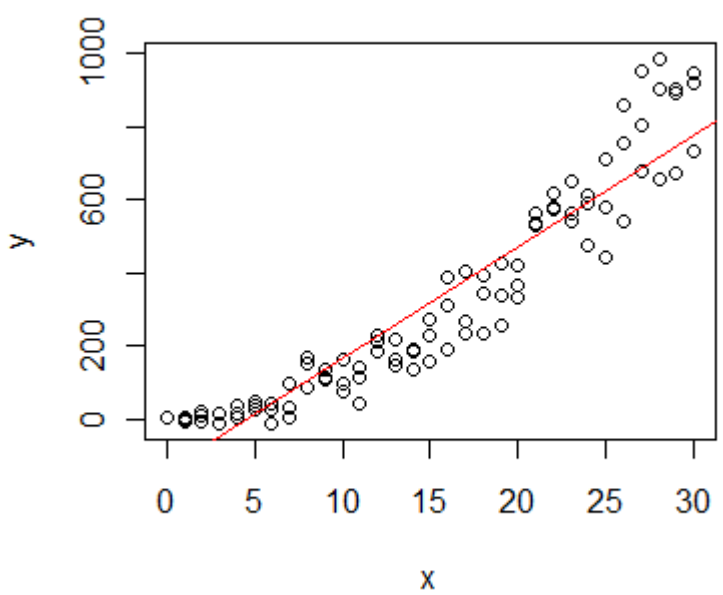


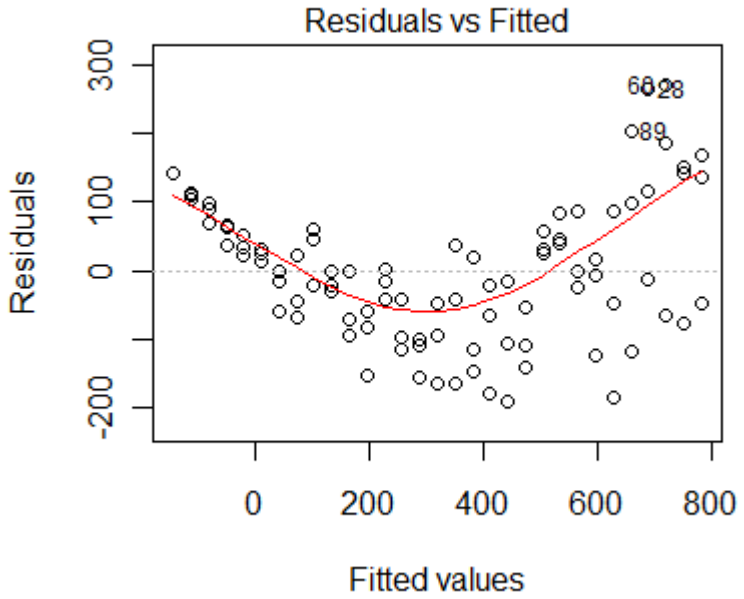
No homocedasticidad



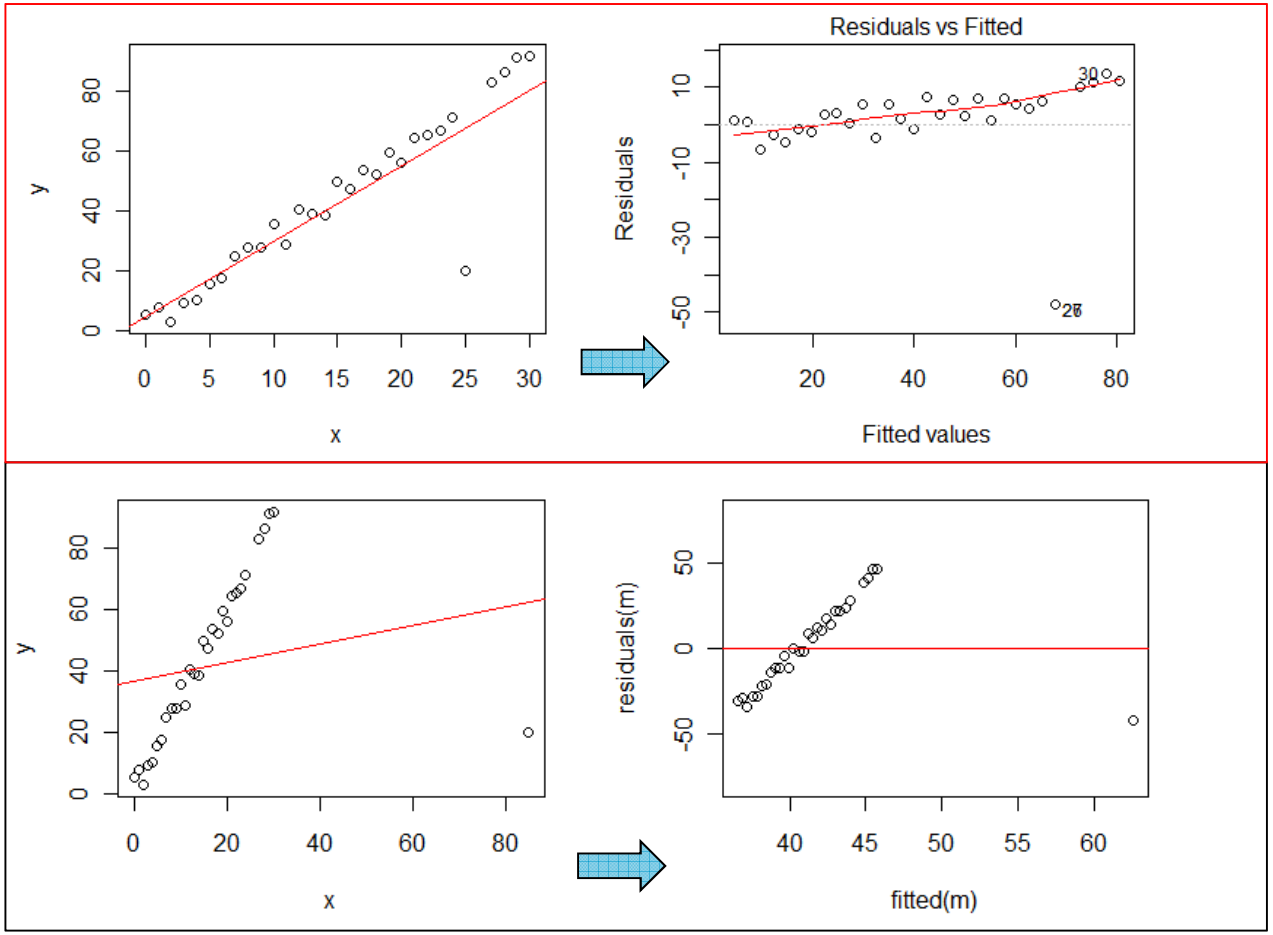


No homocedasticidad, ni linealidad

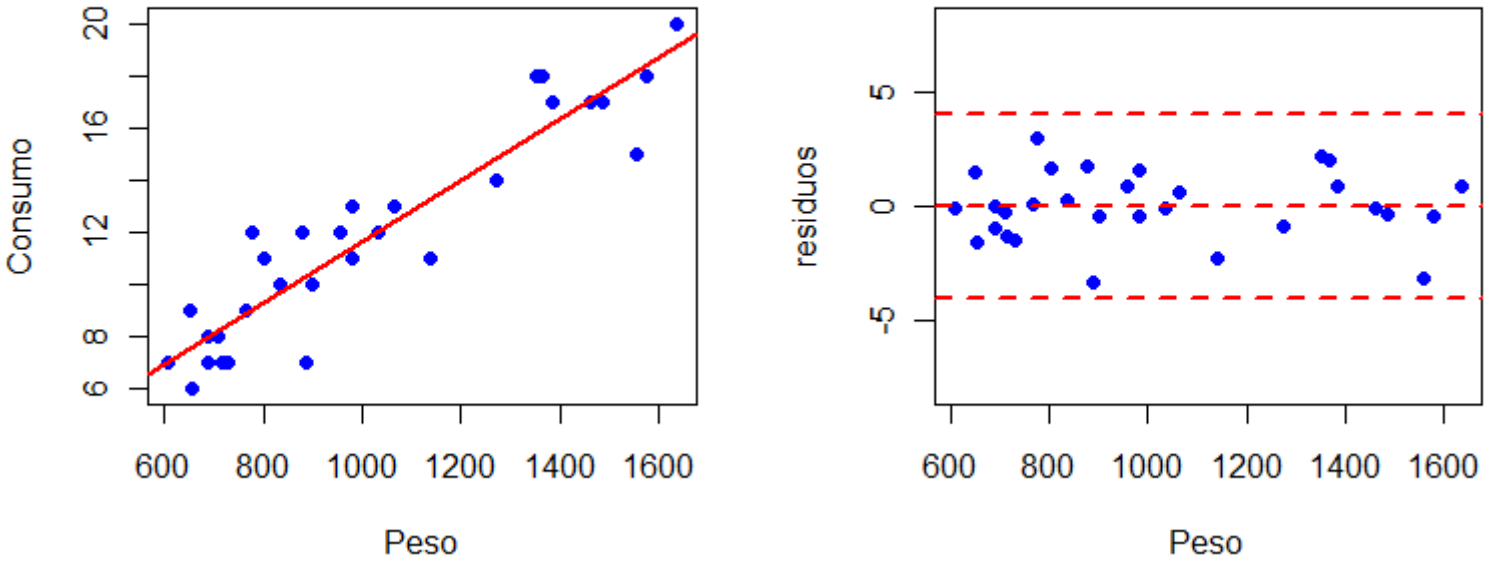




Observaciones atípicas



Residuos Aceptables

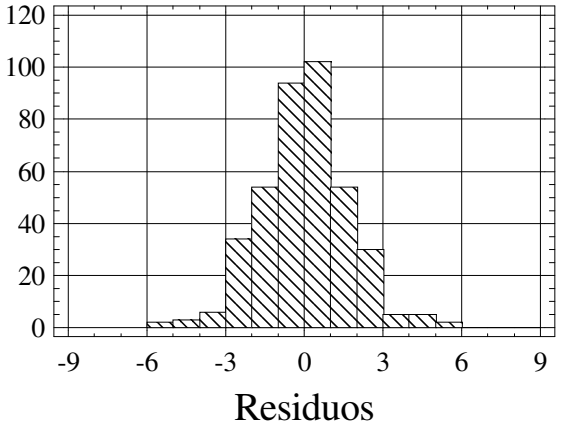


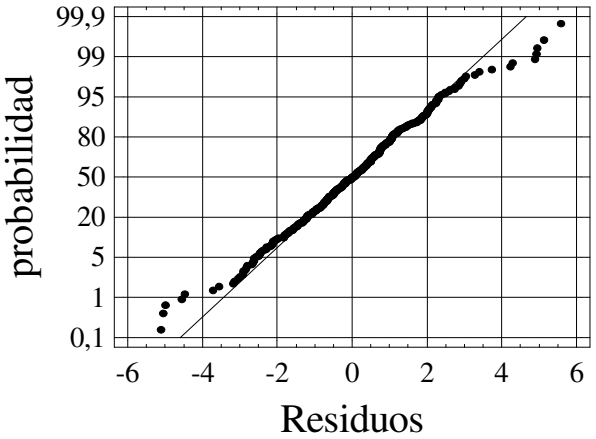
Normalidad de los Residuos

Herramientas de comprobación:

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

Ejemplo de coches





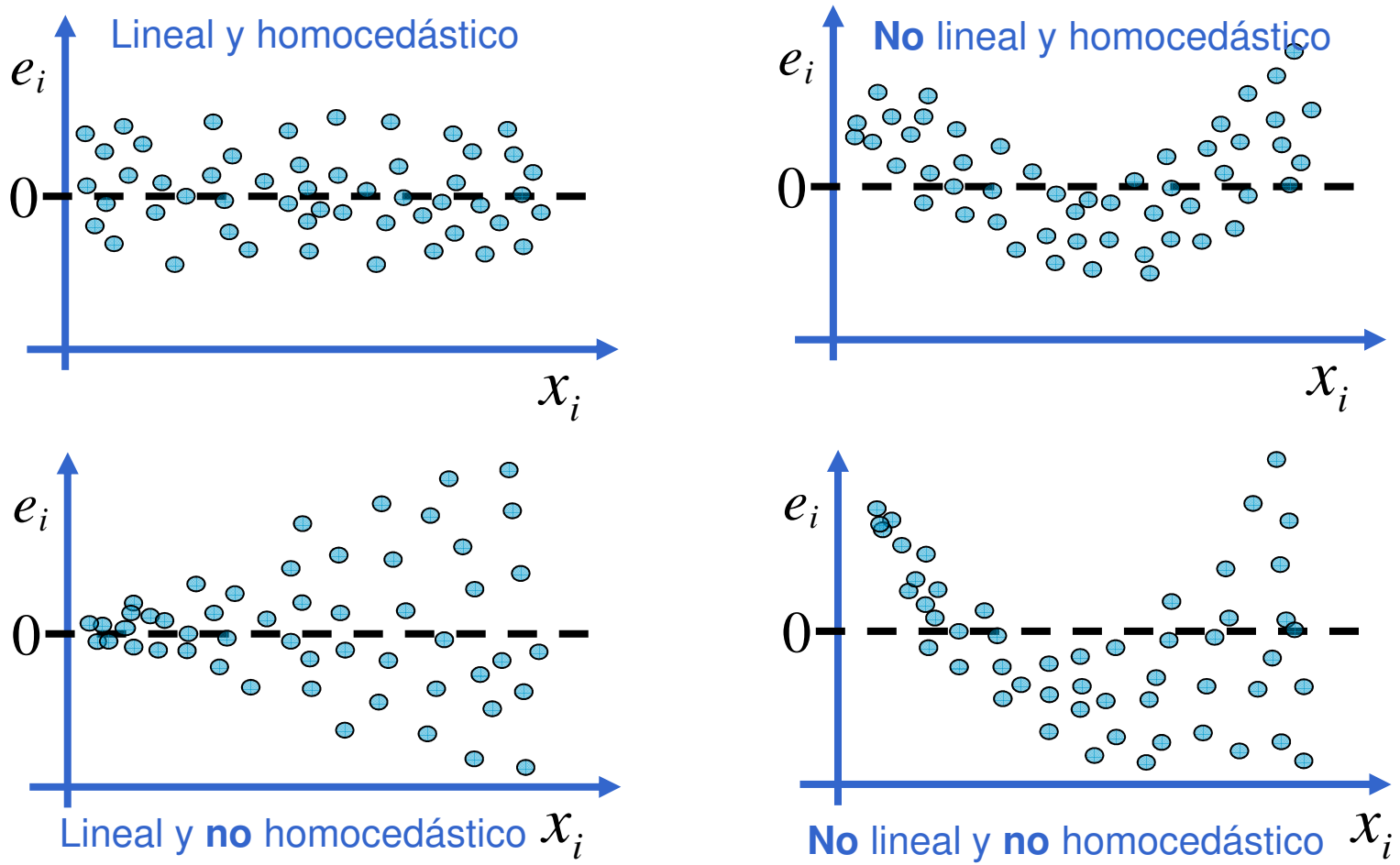
Comprobación de la linealidad y homocedasticidad

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

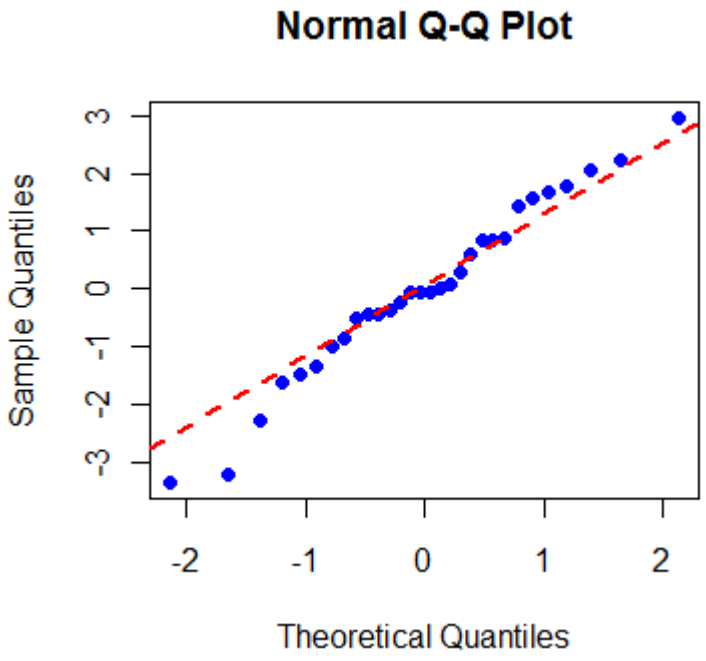
$$\log y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

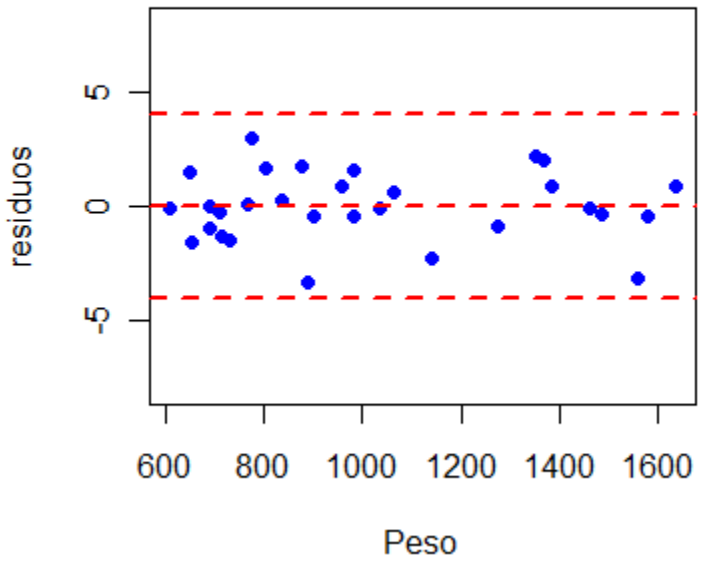
$$\log y_i = \beta_0 + \beta_1 \log x_{1i} + u_i$$

Residuos – Regresor o Val.Previstos



Coches (ejemplo 1): Consumo ~ Peso





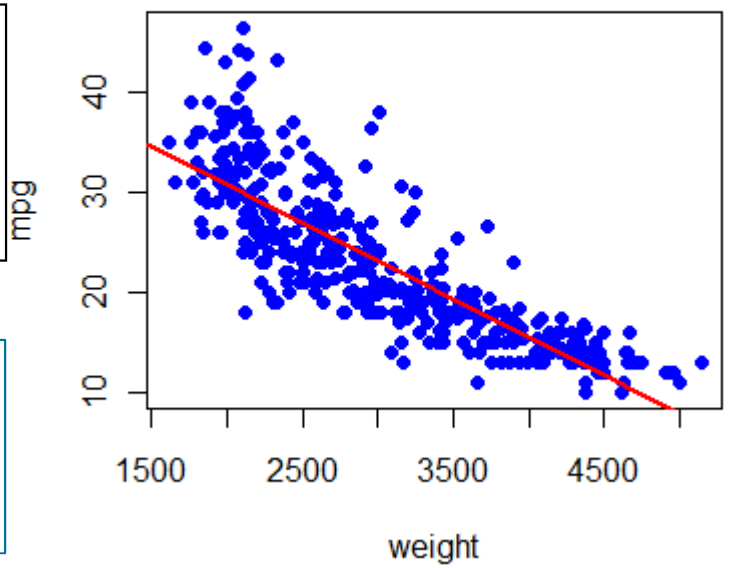
Normalidad **ok**

Linealidad **ok** y Homocedasticidad **ok**

Cars (Ejemplo 2): mpg ~ weight

DESCRIPCIÓN: Datos de 391 coches (archivo:cars.txt) con información del siete variables: consumo (*mpg*), cc (*engine*), potencia (*horse*), peso (*weight*), tiempo de aceleración (*accel*), origen del coche (*origin*, 1=USA, 2=UE, 3=Japón) y número de cilindros (*cylinders*)

	mpg	engine	horse	weight	accel	origin	cylinders
1	14	340	160	3609	8.0	1	8
2	14	440	215	4312	8.5	1	8
3	15	390	190	3850	8.5	1	8
4	14	454	220	4354	9.0	1	8
5	15	400	150	3761	9.5	1	8
6	16	400	230	4278	9.5	1	8

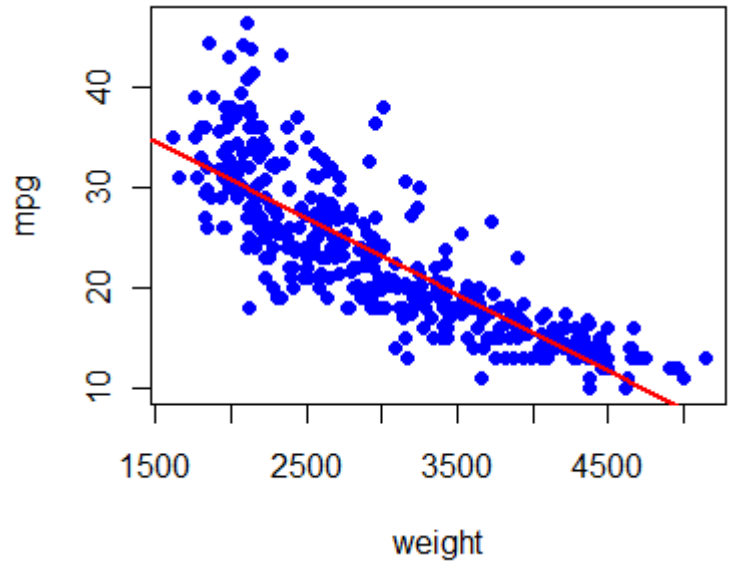


OBJETIVO: Estimar el modelo de regresión simple entre el consumo (*mpg*) y el peso (*weight*)

Cars: mpg ~ weight

$$\widehat{\text{mpg}} = \underset{(0.802)}{49.20} - \underset{(0.00025)}{0.0076}\ \text{weight}$$

$$R^2 = 0.69 \quad \hat{s}_R = 4.34$$



Cars: Figuras

No hay linealidad ni homocedasticidad

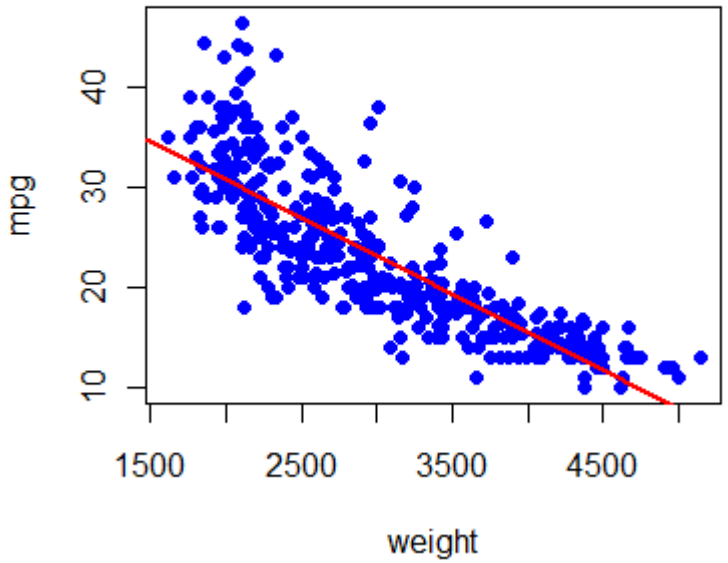


Figura 2.1

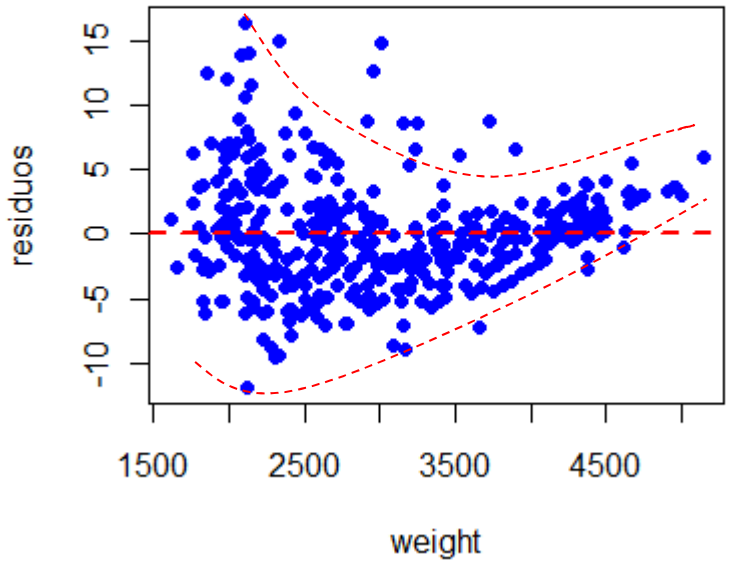


Figura 2.2

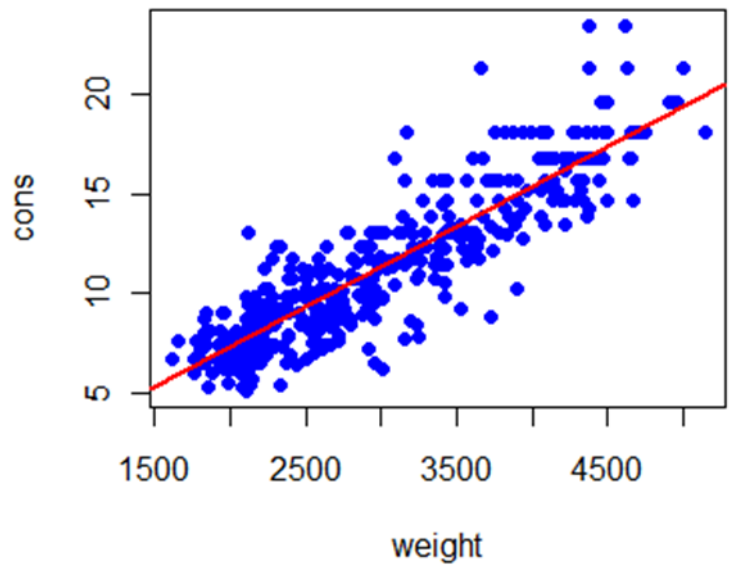
Cars: cons ~ weight

TRANSFORMACIÓN: En lugar de medir el consumo en millas por galón (*mpg*), vamos a cambiar a "*litros cada 100 km* (*cons*)"

$$cons = 235.1/mpg$$

	Y cons	X weight
1	16.79	3609
2	16.79	4312
3	15.67	3850
4	16.79	4354
5	15.67	3761
6	14.69	4278
7	16.79	4425
8	15.67	3563
9	15.67	4341
10	16.79	3086
11	13.83	3449
12	18.08	4735

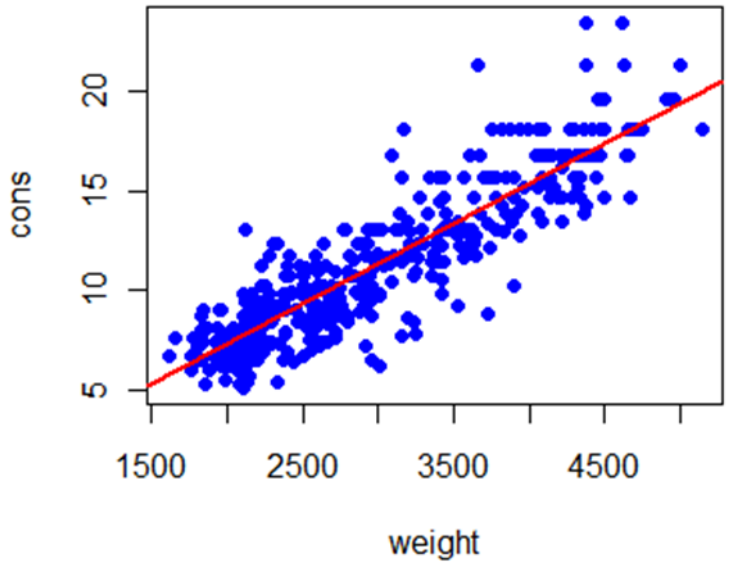
. . .



Cars: cons ~ weight

TRANSFORMACIÓN: En lugar de medir el consumo en millas por galón (*mpg*), vamos a cambiar a "*litros cada 100 km* (*cons*)"

cons = 235.1/*mpg*



$$\widehat{\text{cons}} = \underset{(0.3298)}{-0.7689} + \underset{(0.00011)}{0.0040}\ \text{weight}$$

$$R^2 = 0.79 \quad \hat{s}_R = 1.78$$

Cars: Cambio Variable

Mejora la linealidad y homocedasticidad

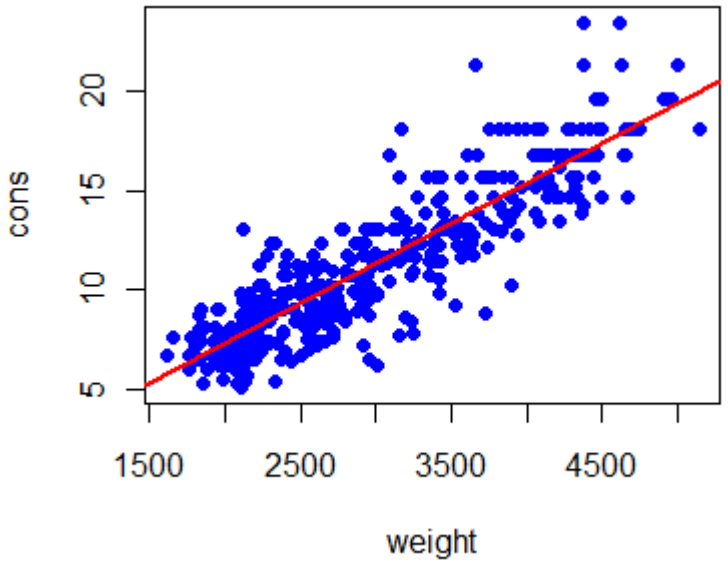


Figura 2.3

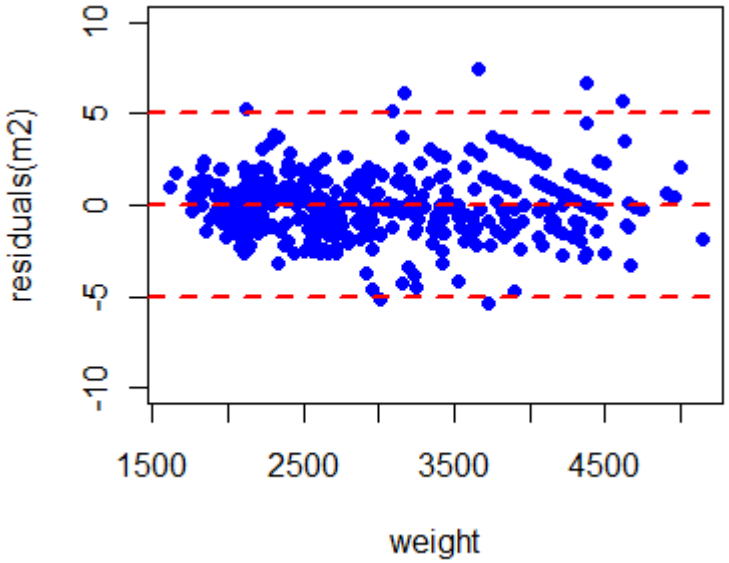


Figura 2.4

Cars: Normalidad

Normalidad no es problemática

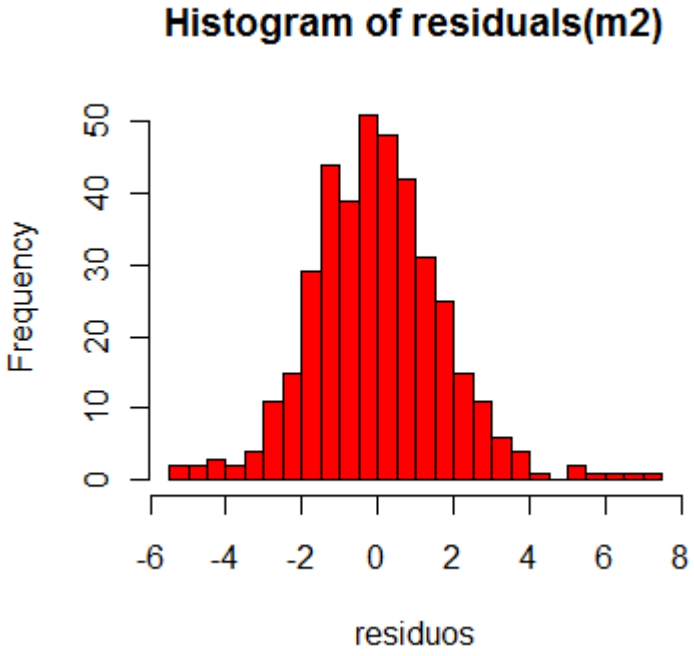


Figura 2.5

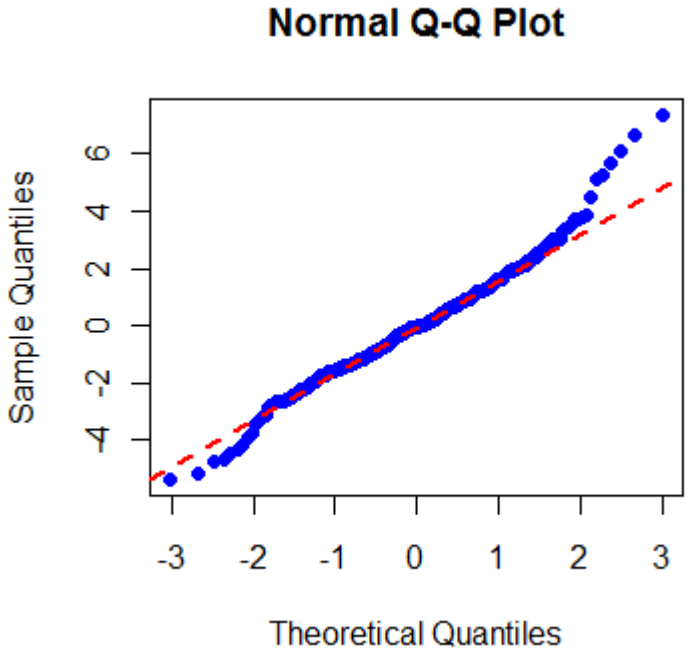


Figura 2.6

Cars: Instrucciones con R

```
> cars<-read.table("cars.txt",header=TRUE)  % LEE EL ARCHIVO CARS.TXT
> attach(cars) % AÑADE LAS VARIABLES DEL CONJUNTO DE DATOS cars A LA MEMORIA

> mod_cars<-lm(mpg ~ weight) % ESTIMA EL MODELO DE REGRESIÓN SIMPLE (MOD_CARS)

> par(mfrow=c(1,2)) % DIVIDE LA PANTALLA GRÁFICA EN 1 FILA Y 2 COLUMNAS (ver FIGURAs 2.1 2.2)

> plot(weight,mpg,pch=19,col="blue") % DIBUJA Figura 2.1
> abline(mod_cars,col="red",lwd=2) % AÑADE Linea roja A la figura 2.1

> plot(weight,residuals(mod_cars),pch=19,col="blue",ylab="residuos") % DIBUJA Figura 2.2
> abline(c(0,0),col="red",lty=2,lwd=2) > summary(mod_cars) % Línea roja de la figura 2.2

> summary(mod_cars) % MUESTRA Resumen del modelo de regresión
```

Cars: Instrucciones con R

```
Call:
lm(formula = mpg ~ weight, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9707  -2.7594  -0.3167   2.1455  16.5223

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.1998655  0.8026747   57.56   <2e-16 ***
weight      -0.0076409  0.0002597  -29.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.338 on 389 degrees of freedom
Multiple R-squared:  0.69, Adjusted R-squared: 0.6892
F-statistic: 865.6 on 1 and 389 DF,  p-value: < 2.2e-16
```

Tabla 2.1

Cars: Instrucciones con R

```
> cons <- 235.1/mpg                 % cambio variable
> m2 <- lm(cons ~ weight)           % nuevo modelo

> plot(weight,cons,pch=19,col="blue") % Figuras 2.3 y 2.4
> abline(m2,col="red",lwd=2)
> plot(weight,residuals(m2),pch=19,col="blue",ylim=c(-10,10))
> abline(c(0,0),col="red",lwd=2,lty=2)
> abline(c(5,0),col="red",lwd=2,lty=2)
> abline(c(-5,0),col="red",lwd=2,lty=2)

> hist(residuals(m2),xlab="residuos",col="red",nclas=20)  % figuras 2.5 y 2.6
> qqnorm(residuals(m2),col="blue",pch=19)
> qqline(residuals(m2),col="red",lwd=2,lty=2)

> summary(m2)                       % resumen del modelo m2 (tabla 2.2)
```

Cars: Instrucciones con R

```
Call: lm(formula = cons ~ weight)

Residuals: Min      1Q     Median   3Q    Max
         -5.3949 -1.1797 -0.0405 1.0152 7.3852

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.7689044 0.3298229 -2.331 0.0202 *
weight 0.0040274 0.0001067 37.741 <2e-16 ***

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.782 on 389 degrees of freedom
Multiple R-squared: 0.7855, Adjusted R-squared: 0.7849
F-statistic: 1424 on 1 and 389 DF, p-value: < 2.2e-16
```

Tabla 2.2

Forbes (Ejemplo 3)

Ejemplo "Forbes"

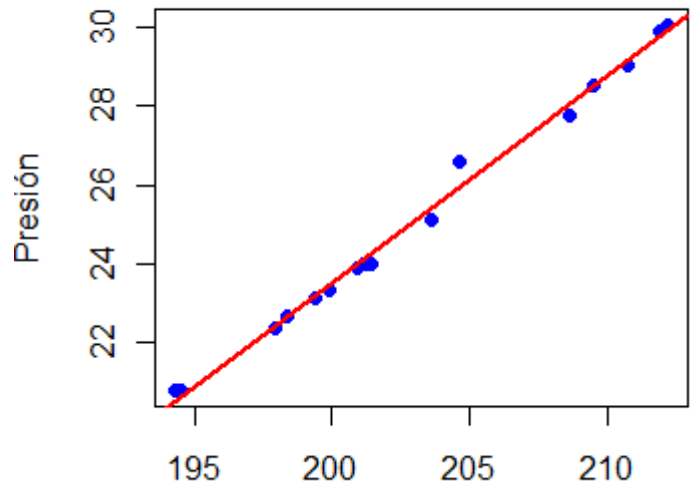
En un artículo de 1857 un físico escocés llamado James D. Forbes presentó una serie de experimentos realizados para estudiar la relación entre presión atmosférica y punto de ebullición del agua. Forbes sabía que la altitud podía ser determinada a partir de la presión atmosférica medida con un barómetro, con menores presiones a medida que aumenta la altitud. A mediados del siglo XIX los barómetros eran instrumentos muy frágiles y Forbes pensó que se podía sustituir la medidas de la presión con medidas de la temperatura de ebullición del agua. Recogió datos de 17 emplazamientos en los Alpes y los montes de Escocia. En cada lugar se midió con un barómetro la presión en pulgadas de mercurio (*Pres*) y la temperatura de ebullición del agua en grados Fahrenheit (*Temp*) empleando un termómetro. Los datos se encuentran en el archivo "forbes.txt"

"forbes.txt"

	Temp	Pres
1	194.5	20.79
2	194.3	20.79
3	197.9	22.40
4	198.4	22.67
5	199.4	23.15
6	199.9	23.35
7	200.9	23.89
8	201.1	23.99
9	201.4	24.02
10	201.3	24.01
11	203.6	25.14
12	204.6	26.57
13	209.5	28.49
14	208.6	27.76
15	210.7	29.04
16	211.9	29.88
17	212.2	30.06

Weisberg, S. (2005). Applied Linear Regression, 3rd edition. New York: Wiley.

Forbes: Modelo Inicial



$$\widehat{\text{Pres}} = \underset{(2.05)}{-81.06} + \underset{(0.010)}{0.523}\ \text{Temp}$$

$$R^2 = 0.994 \quad \hat{s}_R = 0.233$$

	Temp	Pres	Pred	Resid
1	194.5	20.79	20.639	0.1511552
2	194.3	20.79	20.534	0.2557337
3	197.9	22.40	22.417	-0.0166790
4	198.4	22.67	22.678	-0.0081252
5	199.4	23.15	23.201	-0.0510176
6	199.9	23.35	23.462	-0.1124638
7	200.9	23.89	23.985	-0.0953562
8	201.1	23.99	24.090	-0.0999347
9	201.4	24.02	24.247	-0.2268024
10	201.3	24.01	24.195	-0.1845131
11	203.6	25.14	25.397	-0.2571657
12	204.6	26.57	25.920	0.6499419
13	209.5	28.49	28.482	0.0077692
14	208.6	27.76	28.012	-0.2516277
15	210.7	29.04	29.110	-0.0697017
16	211.9	29.88	29.737	0.1428274
17	212.2	30.06	29.894	0.1659597

Tabla 3.1

Forbes: Conclusiones Modelo Inicial

- Según la figura y el valor R-cuadrado (0.994) el ajuste es muy bueno.
- Comparando los valores Previstos con los Observados (Pred) observamos que las diferencias (residuos) son pequeñas ($\hat{s}_R = 0.233$)
- Los dos parámetros del modelo son muy significativos (entre paréntesis se proporcionan las desv. típicas. estimadas de los parámetros estimados)

Forbes: Diagnosis

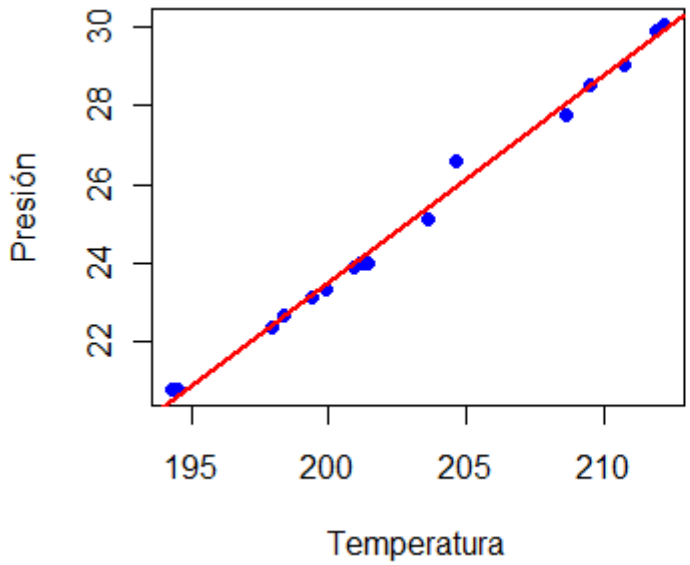


Figura 3.1

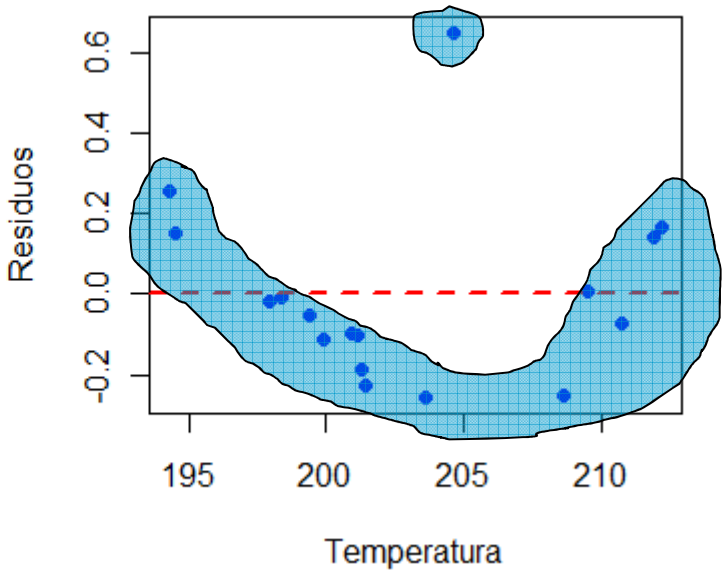


Figura 3.2

En el gráfico de residuos frente al regresor se observa:

- La mayoría de las observaciones muestran no-linealidad
- Existe una observación atípica

Forbes: Instrucciones R

```
> forbes <- read.table("forbes.txt",header=TRUE)
> attach(forbes)
> m <- lm(Pres ~ Temp)
> summary(m)
```

```
Call:
lm(formula = Pres ~ Temp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373    2.05182  -39.51   <2e-16 ***
Temp          0.52289    0.01011   51.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2328 on 15 degrees of freedom
Multiple R-squared: 0.9944, Adjusted R-squared: 0.9941
F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16
```

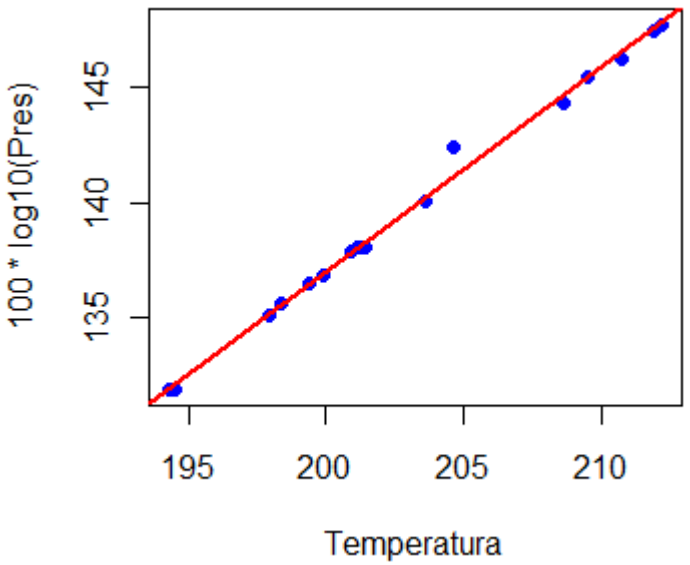
Forbes: Instrucciones R (cont)

```
> forbes$Pred <- predict(m)
> forbes$Resid <- residuals(m)
> print(forbes,digits=4,print.gap=3) % proporciona tabla 3.1

Figuras 3.1 y 3.2
> par(mfrow=c(1,2))
> plot(Temp,Pres,pch=19,col="blue",xlab="Temperatura",
       ylab="Presión")
> abline(m,col="red",lwd=2)
> plot(Temp,residuals(m),pch=19,col="blue",ylab="Residuos",
      xlab="Temperatura")
> abline(c(0,0),lty=2,lwd=2,col="red")
```

Forbes: Modelo 1

$$Lpres = 100 \times \log_{10} Pres$$



	Temp	Pres	Lpres	Pred	Resid
1	194.5	20.79	131.79	132.03	-0.2480225
2	194.3	20.79	131.79	131.85	-0.0688990
3	197.9	22.40	135.02	135.08	-0.0537700
4	198.4	22.67	135.55	135.53	0.0187713
5	199.4	23.15	136.46	136.42	0.0331010
6	199.9	23.35	136.83	136.87	-0.0411189
7	200.9	23.89	137.82	137.77	0.0561898
8	201.1	23.99	138.00	137.94	0.0584761
9	201.4	24.02	138.06	138.21	-0.1559337
10	201.3	24.01	138.04	138.12	-0.0844563
11	203.6	25.14	140.04	140.18	-0.1470658
12	204.6	26.57	142.44	141.08	**1.3599445**
13	209.5	28.49	145.47	145.47	0.0015070
14	208.6	27.76	144.34	144.66	-0.3197358
15	210.7	29.04	146.30	146.54	-0.2428181
16	211.9	29.88	147.54	147.62	-0.0791613
17	212.2	30.06	147.80	147.89	-0.0870083

$$\widehat{\text{Lpres}} = \underset{(3.34)}{-42.16} + \underset{(0.016)}{0.8956}\ \text{Temp}$$

Tabla 4.1

$$R^2 = 0.995 \quad \hat{s}_R = 0.379$$

Forbes : modelo 1

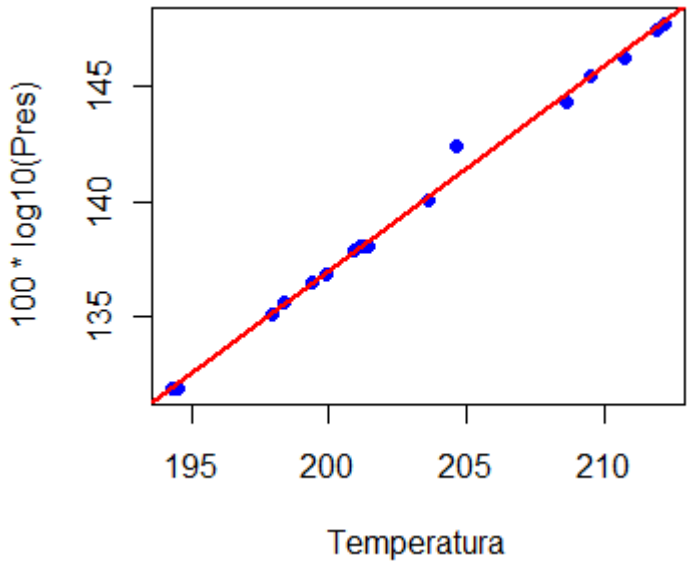


Figura 4.1

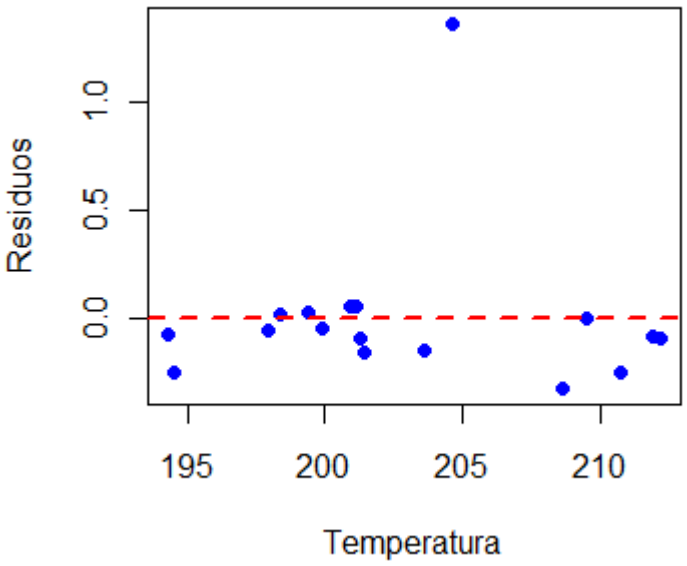


Figura 4.2

En el gráfico de residuos frente al regresor se observa:

- Existe una observación claramente atípica
- Se ha corregido la falta de linealidad en el resto de las observaciones.

Forbes: Modelo 1

- Se ha realizado la transformación logarítmica de la presión para corregir la falta de linealidad (da igual utilizar logaritmos neperianos o decimales, se ha multiplicado por 100 para evitar números muy pequeños en las estimaciones, no tiene efecto en el análisis)
- La observación atípica tiene mucha influencia en la estimación del modelo, se aprecia como los residuos del resto de las observaciones no tienen media cero.
- Por lo demás el ajuste es muy bueno como se ve en la gráfica y en la tabla 4.1, los valores previstos se parecen mucho a los observados (los residuos son pequeños)
- Conviene eliminar la observación atípica y recalcular.

Forbes: Instrucciones R

```
> forbes1 <- read.table("forbes.txt",header=TRUE)
> attach(forbes1)
> m1 <- lm(100*log10(Pres) ~ Temp)
> summary(m1)
```

```
Call:
lm(formula = 100 * log10(Pres) ~ Temp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.16418    3.34136  -12.62 2.17e-09 ***
Temp          0.89562    0.01646   54.42  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-squared: 0.995, Adjusted R-squared: 0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16
```

Forbes: Instrucciones R (cont)

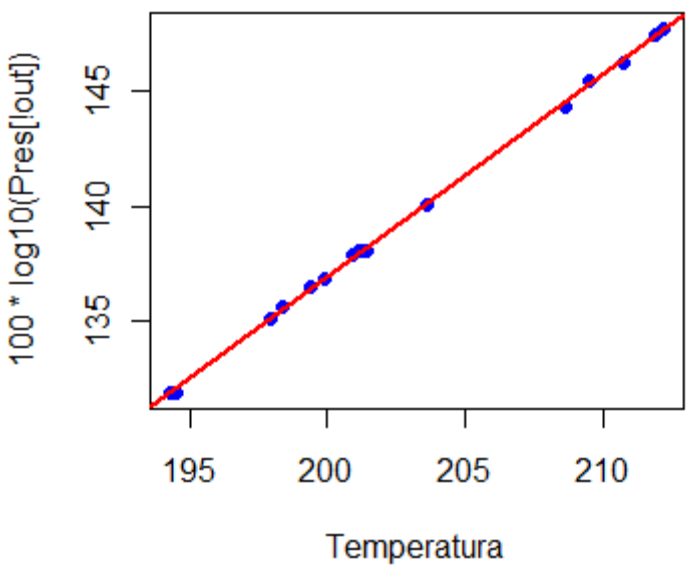
```
> forbes1$Lpres <- 100*log10(Pres)
> forbes1$Pred <- predict(m1)
> forbes1$Resid <- residuals(m1)
> print(forbes1,digits=4,print.gap=3) % proporciona tabla 4.1

Figuras 4.1 y 4.2
> par(mfrow=c(1,2))
> plot(Temp,100*log10(Pres),pch=19,col="blue",xlab="Temperatura“)
> abline(m1,col="red",lwd=2)
> plot(Temp,residuals(m1),pch=19,col="blue",ylab="Residuos",
       xlab="Temperatura")
> abline(c(0,0),lty=2,lwd=2,col="red")
```

Forbes: Modelo 2

(ELIMINANDO OBSERVACIÓN Nº 12)

$$Lpres = 100 \times \log_{10} Pres$$



	Temp	Pres	Lpres	Pred	Resid
1	194.5	20.79	131.79	131.99	-0.2006699
2	194.3	20.79	131.79	131.81	-0.0224480
3	197.9	22.40	135.02	135.02	0.0089107
4	198.4	22.67	135.55	135.46	0.0837061
5	199.4	23.15	136.46	136.35	0.1025441
6	199.9	23.35	136.83	136.80	0.0305783
7	200.9	23.89	137.82	137.69	0.1323953
8	201.1	23.99	138.00	137.87	0.1355832
9	201.4	24.02	138.06	138.13	-0.0774742
10	201.3	24.01	138.04	138.05	-0.0064475
11	203.6	25.14	140.04	140.10	-0.0586881
12*	204.6	26.57	142.44	140.99	1.4527324
13	209.5	28.49	145.47	145.35	0.1164833
14	208.6	27.76	144.34	144.55	-0.2088168
15	210.7	29.04	146.30	146.42	-0.1224318
16	211.9	29.88	147.54	147.49	0.0466349
17	212.2	30.06	147.80	147.76	0.0401403

Tabla 5.1

$$\widehat{\text{Lpres}} = \underset{(1.003)}{-41.33} + \underset{(0.0049)}{0.8911}\ \text{Temp}$$

$$R^2 = 0.9996 \quad \hat{s}_R = 0.1136$$

La obs. 12 no se ha utilizado en la estimación del modelo

Forbes : modelo 2

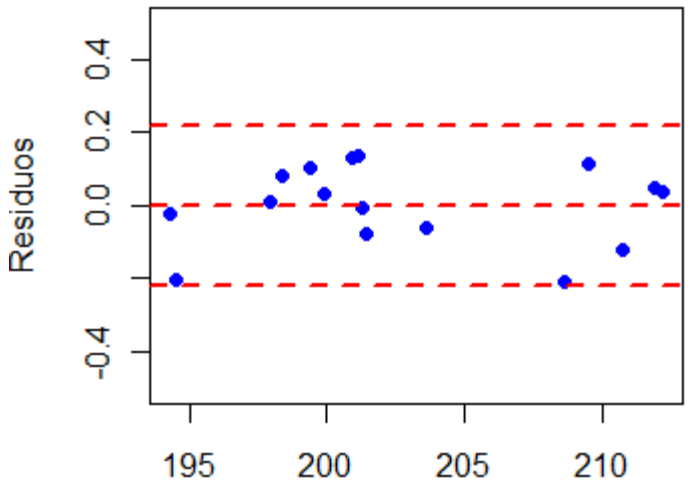


Figura 5.1

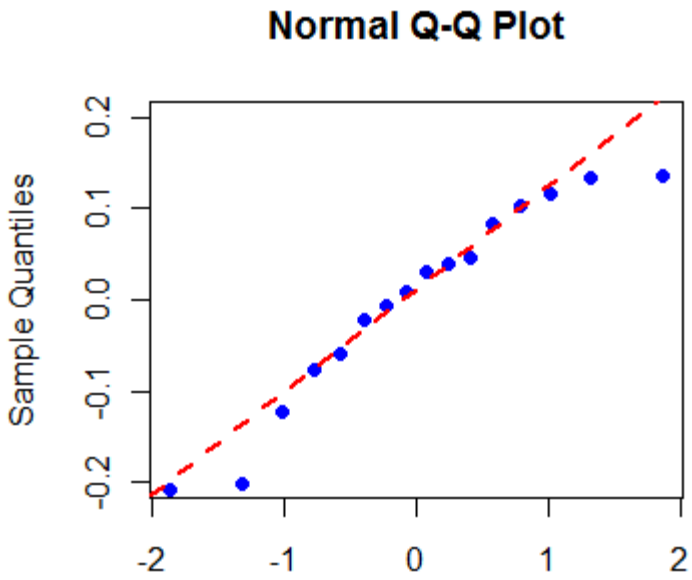


Figura 5.2

En el gráfico de residuos frente al regresor se observa:

- No existen observaciones atípicas (las líneas rojas se encuentran a $\pm 2\hat{s}_R$)
- No se observa ninguna anomalía grave en el qqplot..

Forbes: Modelo 2

- Se ha realizado la transformación logarítmica de la presión para corregir la falta de linealidad y se ha eliminado la observación 12 (el propio Forbes indica en su artículo que se trataba de un error de medida)
- Comparando el modelo 1 y 2, no se aprecian grandes cambios en los parámetros estimados $\hat{\beta}_0, \hat{\beta}_1$.
- La desviación típica residual se ha reducido considerablemente de uno a otro, pasando de 0.379 a 0.113, y como consecuencia las desviaciones típicas de los parámetros.
- El análisis de los residuos no indican ninguna desviación importante de las hipótesis del modelo

Forbes: Instrucciones R

```
> # Modelo m2 de Forbes
> out <- abs(residuals(m1)) > 3*0.3792
> m2 <- lm(100*log10(Pres[!out]) ~ Temp[!out])
> summary(m2)
```

```
Call:
lm(formula = 100 * log10(Pres[!out]) ~ Temp[!out])

Residuals:
     Min       1Q   Median       3Q      Max
-0.20882 -0.06338  0.01974  0.08842  0.13558

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.334683   1.003312   -41.2 5.16e-16 ***
Temp[!out]    0.891110   0.004944   180.2  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1136 on 14 degrees of freedom
Multiple R-squared: 0.9996, Adjusted R-squared: 0.9995
F-statistic: 3.249e+04 on 1 and 14 DF,  p-value: < 2.2e-16
```

Forbes 2: Instrucciones R (cont)

```
> # Tabla 5.1
> P_Lpres =c(predict(m2)[1:11],NA,predict(m2)[12:16])
> P_Lpres[12] = -41.334683 + 0.891110*Temp[12]
> forbes2 <- forbes1
> forbes2$Pred <- P_Lpres
> forbes2$Resid <- 100*log10(Pres)-P_Lpres
> print(forbes2,digits=5,print.gap=3)

> # Figuras 5.1 y 5.2
> par(mfrow=c(1,2))
> plot(Temp[!out],residuals(m2),pch=19,col="blue",ylab="Residuos",
+         xlab="Temperatura",ylim=c(-.5,.5))
> abline(c(0,0),lty=2,lwd=2,col="red")
> abline(c(-.22,0),lty=2,lwd=2,col="red")
> abline(c(+.22,0),lty=2,lwd=2,col="red")
>
> qqnorm(residuals(m2),ylim=c(-.2,.2),pch=19,col="blue")
> qqline(residuals(m2),col="red",lty=2,lwd=2)
```

FEV (Ejemplo 4)

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

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

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

Variables

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

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

Tabla 6.1

FEV: Modelo Inicial

- Tanto en el gráfico de dispersión de FEV y altura (ht) como en el de los residuos del modelo de regresión simple se observa la relación no-lineal entre las dos variables y la heterocedasticidad.

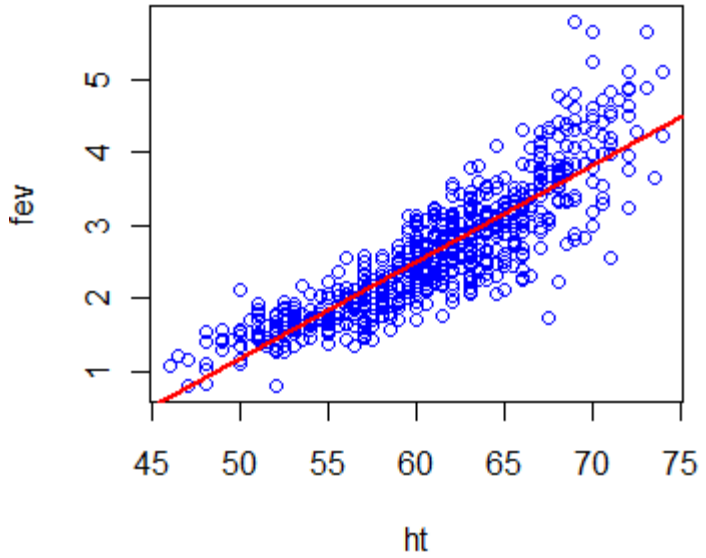


Figura 6.1

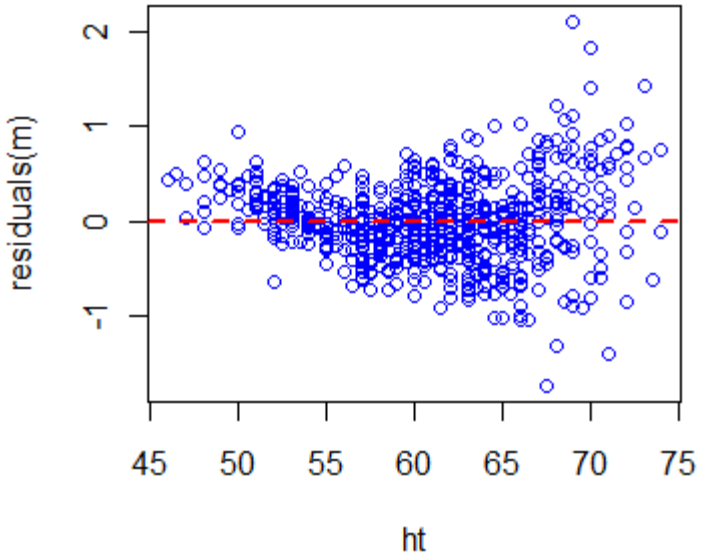


Figura 6.2

FEV: modelo 1

$$\widehat{\log(\text{fev})} = \underset{(0.063)}{-2.27} + \underset{(0.0010)}{0.052}\,\text{ht}$$

$$R^2 = 0.7956 \quad \hat{s}_R = 0.1508$$

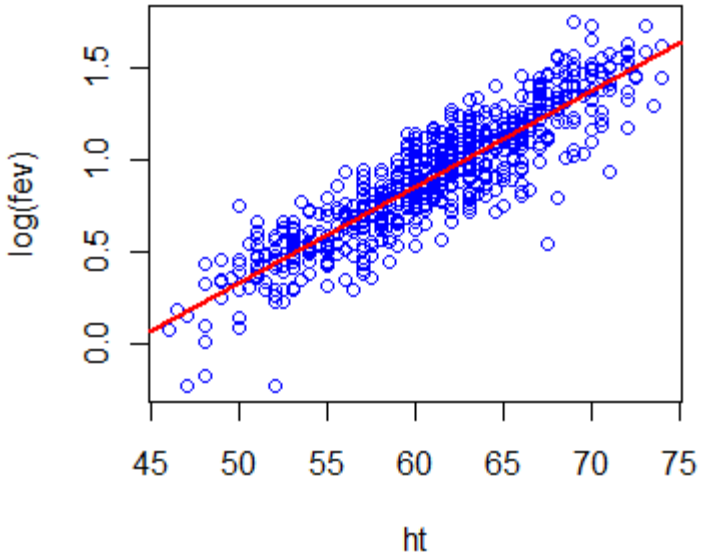


Figura 6.3

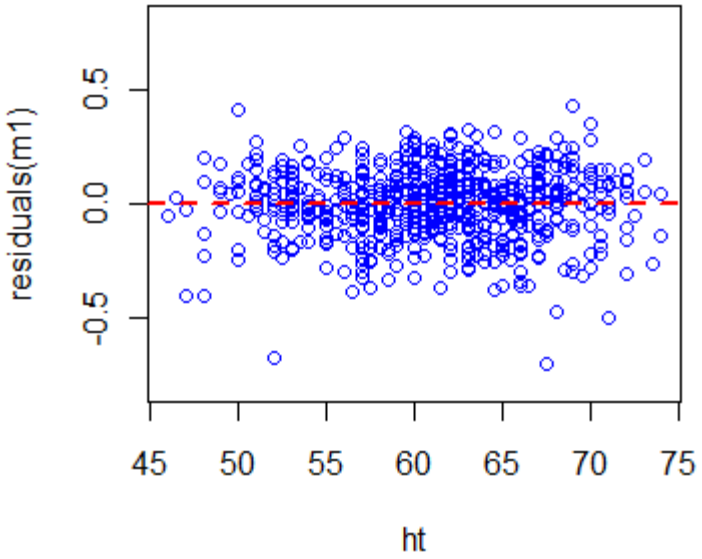


Figura 6.4

FEV: modelo 1

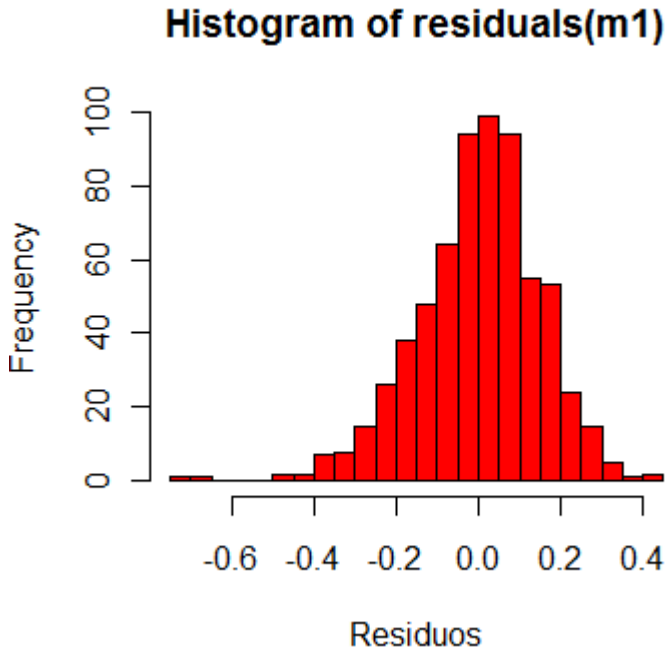


Figura 6.5

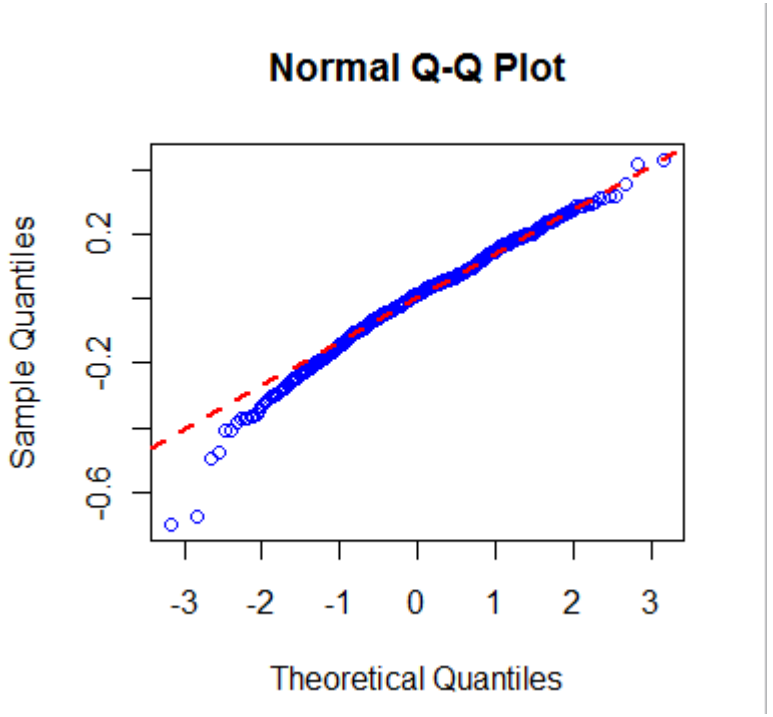


Figura 6.6

FEV: Modelo 1

$$\widehat{\log(\text{fev})} = \underset{(0.063)}{-2.27} + \underset{(0.0010)}{0.052}\,\text{ht}$$

$$R^2 = 0.7956 \quad \hat{s}_R = 0.1508$$

- Se ha realizado la transformación logarítmica de la variable respuesta (fev) y se ha corregido la falta de linealidad y la heterocedasticidad como se ve en las figuras 6.3 y 6.4
- El histograma y el qqplot (figura 6.5 y 6.6) no muestran grandes desviaciones de la normalidad.
- Existen algunas observaciones atípicas pero se puede comprobar que al eliminarlas los resultados no cambian sustancialmente.
- Existe una relación muy significativa entre log(fev) y ht (altura). Un incremento de un pulgada en la estatura supone un aumento de la capacidad pulmonar del 5% (este resultado cambiará al considerar otras variables)
- La altura explica un 79% (R2) de la variabilidad del log(fev).

FEV: Modelo m1 con R

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.70208 -0.08986  0.01190  0.09337  0.43174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.271312   0.063531  -35.75   <2e-16 ***
ht           0.052119   0.001035   50.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1508 on 652 degrees of freedom
Multiple R-squared: 0.7956, Adjusted R-squared: 0.7953
F-statistic:  2538 on 1 and 652 DF,  p-value: < 2.2e-16
```

Tabla 6.2

FEV : Instrucciones de R

```
> # FEV (ejemplo 4)
> dat <- read.table("fev.dat",header=TRUE)
> head(fev) #tabla 6.1
> attach(dat)
> m<-lm(fev~ht) # modelo m inicial
> par(mfrow=c(1,2))
> plot(ht,fev,col="blue") # figura 6.1
> abline(m,col="red",lwd=2)
> plot(ht,residuals(m),col="blue") # figura 6.2
> abline(c(0,0),col="red",lwd=2,lty=2)
> m1 <- lm(log(fev) ~ ht)
> summary(m1) # modelo estimado tabla 6.2
> plot(ht,log(fev),col="blue") # figura 6.3
> abline(m1,col="red",lwd=2) # figura 6.3
> plot(ht,residuals(m1),col="blue") # figura 6.4
> abline(c(0,0),col="red",lty=2,lwd=2)
> par(mfrow=c(1,2)) # figura 6.5 y 6.6
> hist(residuals(m1),col="red",nclass=20,xlab="Residuos")
> qqnorm(residuals(m1),col="blue")
> qqline(residuals(m1),col="red",lty=2,lwd=2)
```

Brains (ejemplo 5)

Ejemplo "Brains" Peso del cuerpo y cerebro de mamiferos
62 observaciones, 2 variables

Descripción:
Para 62 especies de mamíferos se proporciona el peso medio del cuerpo en kilogramos y del cerebro en gramos

Variables:
BrainWt Peso del cerebro (gramos)
BodyWt Peso del Cuerpo (kilogramos)

	BrainWt	BodyWt
Arctic_fox	44.500	3.385
Owl_monkey	15.499	0.480
Beaver	8.100	1.350
Cow	423.012	464.983
Gray_wolf	119.498	36.328
Goat	114.996	27.660

Tabla 7.1

OBJETIVO: Estudiar la relación entre peso del cerebro y peso del cuerpo.

Fuentes
Allison, T. and Cicchetti, D. (1976). Sleep in mammals: Ecology and constitutional correlates. Science, 194, 732-734.
Weisberg, S. (2005). Applied Linear Regression, 3rd edition. New York: Wiley

Brains: Transformación

- En la escala original (figura 7.1) no tiene sentido el modelo de regresión lineal.
- Haciendo las transformación logarítmica de las dos variables (figura 7.2) se aprecia una clara relación lineal

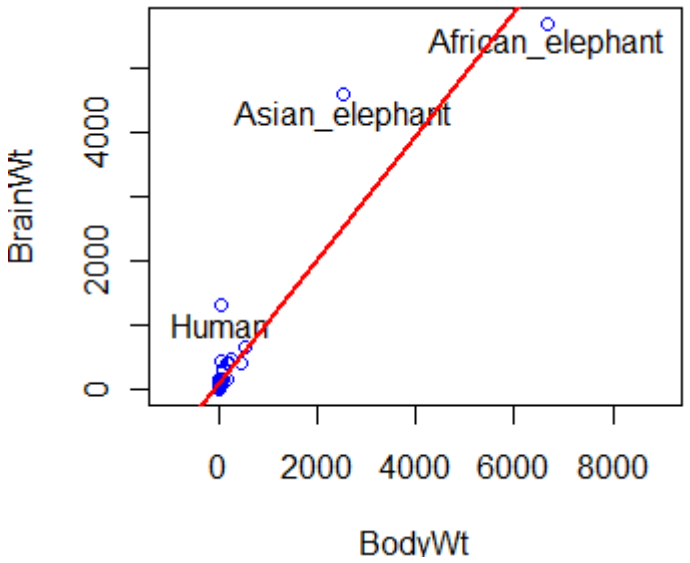


Figura 6.1

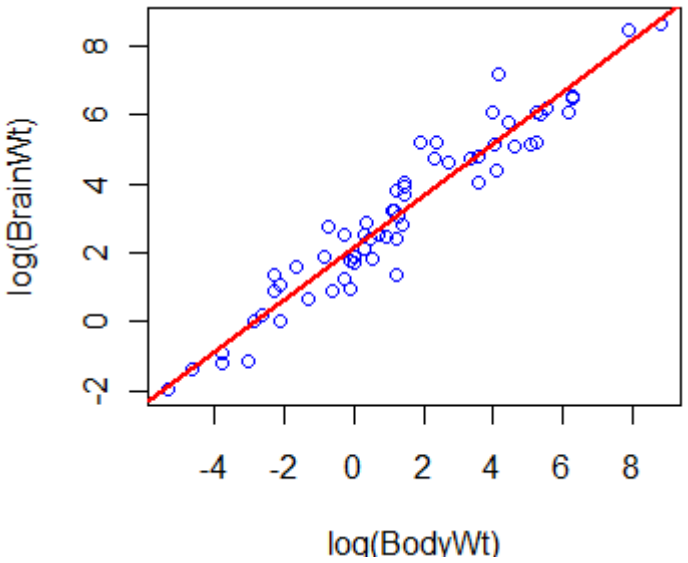


Figura 6.2

Brains: modelo 1

$$\widehat{\log(\text{BrainWt})} = \underset{(0.096)}{2.13} + \underset{(0.028)}{0.752}\log(\text{BodyWt})$$

$$R^2 = 0.9208 \quad \hat{s}_R = 0.6943$$

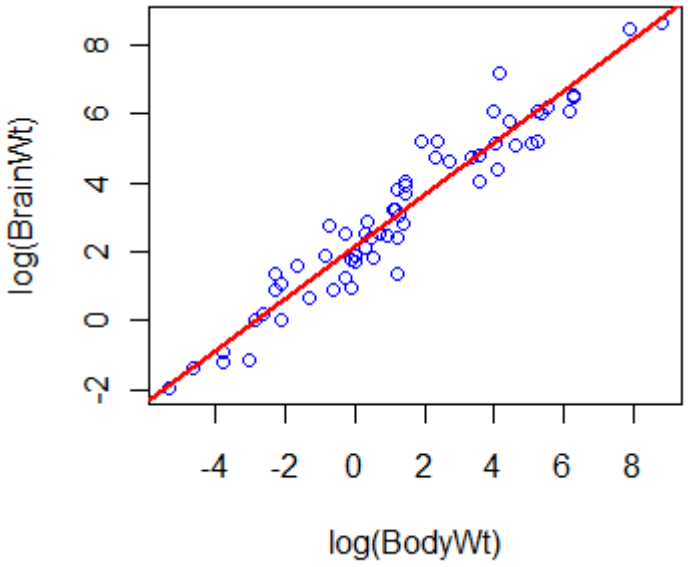


Figura 7.3

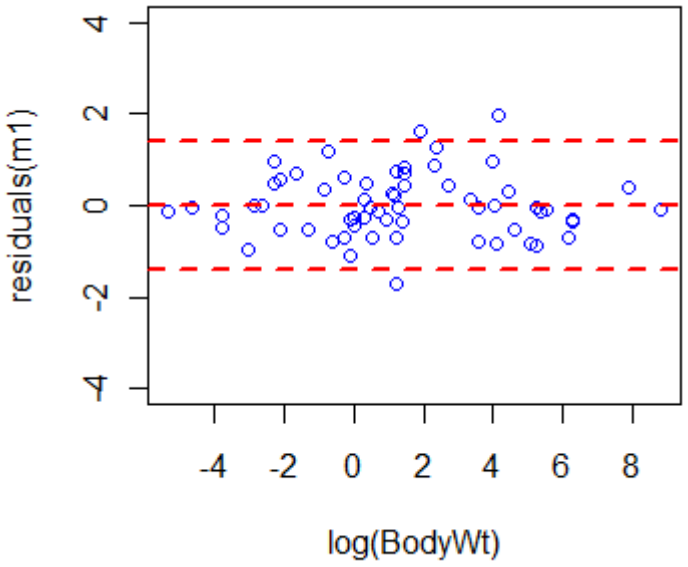


Figura 7.4

Brains

$$\widehat{\log(\text{BrainWt})} = \underset{(0.096)}{2.13} + \underset{(0.028)}{0.752}\log(\text{BodyWt})$$

$$R^2 = 0.9208 \quad \hat{s}_R = 0.6943$$

- La relación entre el logaritmo de peso del cuerpo y el logaritmo del peso del cerebro es lineal como se ve en las figuras 7.3 y 7.4
- Existen algunas observaciones atípicas pero se puede comprobar que al eliminarlas los resultados no cambian sustancialmente.
- El log del peso del cuerpo explica el 92% (R2) de la variabilidad del log del peso del cerebro.

Brains: Modelo m1 con R

```
Call:
lm(formula = log(BrainWt) ~ log(BodyWt))

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43598  1.94833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
log(BodyWt)  0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared: 0.9208, Adjusted R-squared: 0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Tabla 7.2

Brains : Instrucciones de R

```
> brains <- read.table("brains.txt",header=TRUE)
> head(brains) # tabla 7.1
> par(mfrow=c(1,2))
> plot(BodyWt,BrainWt,col="blue",xlim=c(-1000,9000)) # figura 7.1
> sel = BrainWt>1000 # selecciona observaciones con peso del cerebro >1000
> text(BodyWt[out],BrainWt[out]-300,labels=brains[out,1]) # etiquetas 7.1
> m <- lm(BrainWt  ~ BodyWt)
> abline(m,col="red",lwd=2) # figura 7.1
> plot(log(BodyWt),log(BrainWt),col="blue") # Figura 7.2 y 7.3
> m1 <- lm(log(BrainWt)  ~ log(BodyWt))
> abline(m1,col="red",lwd=2) # linea en figura 7.2 y 7.3
> summary(m1) # tabla 7.2
> plot(log(BodyWt),residuals(m1),col="blue",ylim=c(-4,4)) # figura 7.4
> abline(c(0,0),col="red",lty=2,lwd=2)
> abline(c(-2*.6943,0),col="red",lty=2,lwd=2)
> abline(c(+2*.6943,0),col="red",lty=2,lwd=2)
```

Funciones R para Regresión Simple

- **m <- lm(y~x)** Estima el modelo y (variable dependiente) y x (regresor). El modelo lo guarda en m
- **summary(m)** Modelo estimado
- **plot(m)** Diagnosis
- **coef(m)** Da los coeficientes
- **residuals(m)** Residuos del modelo
- **fitted(m)** Da los valores predichos
- **deviance(m)** Suma de residuos al cuadrado
- **predict(m)** Hace predicciones
- **anova(m)** Tabla ANOVA