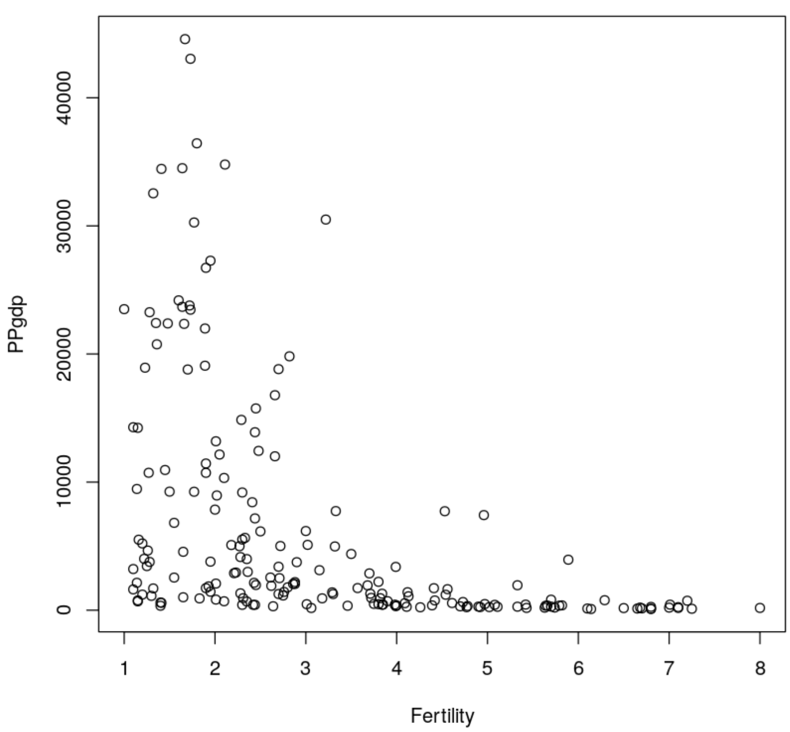
Trabajo 1 - Análisis de Regresión

Fabián Ramírez

```
# Incluyo la libreria de los datos
library('carData')
library('car')
library('alr3')
# Una función  útil
library('model')
# Se instala con el siguiente codigo
# if (!require('devtools')) install.packages('devtools')
# devtools::install_github('fhernanb/model', force=TRUE)
```

```
# Adjunto los datos con los nombres de sus variables.
attach(UN1)
```

```
#Cantidad de datos
n = length(Fertility)
n
# Un grafico de los datos
plot(Fertility,PPgdp)
```



Problema 2.6

2.6.1

Realizaremos una regresión del modelo:

$$\log_{10}(\texttt{Fertility}_i) = \beta_0 + \beta_1 \cdot \log_{10}(\text{PPgdp}_i) + u_i$$

Donde $u_i \sim \mathcal{N}(0, \sigma^2)$ e $i = 1, \ldots, 193$.

```
y = log(Fertility,10)
x = log(PPgdp,10)
reg1<-lm(y~x)
summary(reg1)
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.48587 -0.08148  0.03058  0.11327  0.39130 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.17399    0.05879   19.97   <2e-16 ***
x           -0.22116    0.01737  -12.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1721 on 191 degrees of freedom
Multiple R-squared:  0.4591,Adjusted R-squared:  0.4563 
F-statistic: 162.1 on 1 and 191 DF,  p-value: < 2.2e-16
```

Por tanto el modelo queda de la forma:

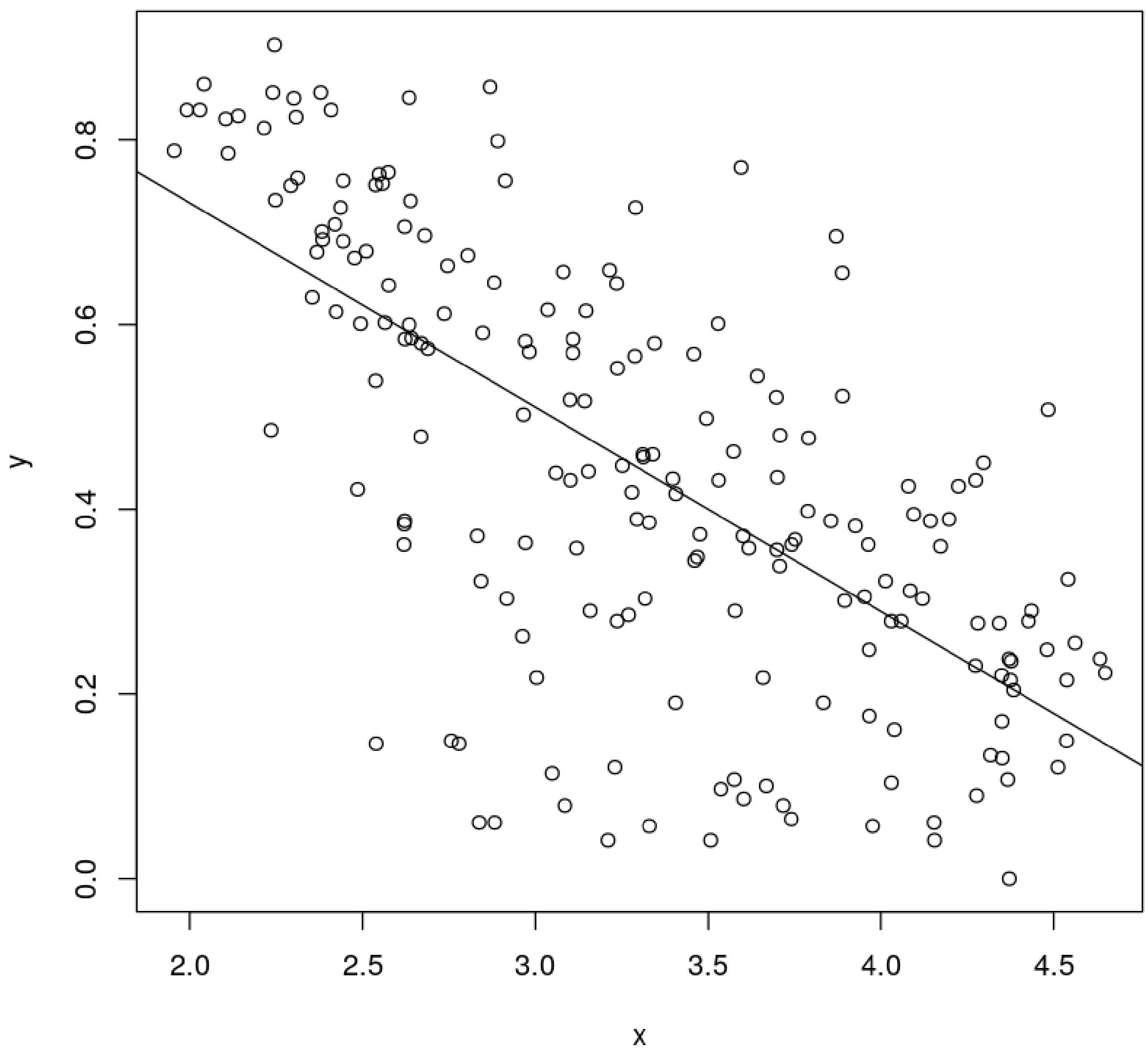
$$\log_{10}(\widehat{\texttt{Fertility}_i}) = 1{,}17399 - 0{,}22116 \cdot \log_{10}(\text{PPgdp}_i)$$

con $i = 1, \ldots, 193$

2.6.2

Realizar un gráfico de la regresión

```
plot(x,y)
abline(lm(y~x))
```



2.6.3

Queremos realizar la prueba:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 < 0 \end{cases}$$

```
t_test=beta_test(reg1,'less','x',0)[3]
```

```
   Estimate   Std.Err t value    Pr(>t)
x -0.221160  0.017368 -12.734 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notemos que el t^* visto en clases es -12,734 y el valor p es del orden de -16, prácticamente 0, por lo tanto la significancia mínima para rechazar H_0 es prácticamente 0, por tanto por ejemplo con un nivel de significancia de 0.05 se debe rechazar H_0.

2.6.4

Notemos que el coeficiente de determinación es:

```
summary(reg1)
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48587 -0.08148  0.03058  0.11327  0.39130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.17399    0.05879   19.97   <2e-16 ***
x           -0.22116    0.01737  -12.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1721 on 191 degrees of freedom
Multiple R-squared:  0.4591,Adjusted R-squared:  0.4563
F-statistic: 162.1 on 1 and 191 DF,  p-value: < 2.2e-16
```

Eso significa que un 45,91 % de la variabilidad (varianza) del logaritmo de `Fertility` puede ser explicada por el logaritmo de PPgdp

2.6.5

```
confint(reg1,level=0.95)
```

	2.5 %	97.5 %
(Intercept)	1.058022	1.289963
x	-0.255418	-0.186902

Donde el intervalo de confianza para β_1 viene dado por $[-0{,}255418; -0{,}186902]$ lo cual se interpreta con la frase 'al variar en una unidad el log(PPgdp), se espera que varié log(`Fertility`) en un número en el intervalo.

Mientras que para la variable original:

```
10^confint(reg1,level=0.95)
```

	2.5 %	97.5 %
(Intercept)	11.4293649	19.4968018
x	0.5553694	0.6502764

Entonces para β_1 el intervalo de confianza viene dado por $[0{,}5553694; 0{,}6502764]$ lo que significa que la tasa de fertilidad se multiplicará por un número entre aproximadamente 0,55 y 0,65, lo que equivale a una disminución de entre 45 % y 55 %

2.6.6

Realizaremos predicciones de $\log_{10}$ `Fertility` cuando $PPgdp = 1000$, junto con los intervalos de confianza con un 95 % de confiabilidad.

```
valor_a_predecir=data.frame(x=log(1000,10))
predict(reg1,valor_a_predecir,interval='prediction')
```

	fit	lwr	upr
1	0.5105127	0.1700834	0.8509421

Por tanto el intervalo de confianza para la predicción es $[0{,}1700834; 0{,}8509421]$.

Mientras que en las variables originales.

```
10^predict(reg1,valor_a_predecir,interval='prediction')
```

	fit	lwr	upr
1	3.239759	1.479392	7.094831

Por tanto el intervalo de confianza para la predicción es $[1{,}479392; 7{,}094831]$.

2.6.7

Notemos que:

```
# La localidad con mayor valor de Fertility es:
rownames(UN1)[Fertility == max(Fertility)]
```

'Niger'

```
# La localidad con menor valor de Fertility es:
rownames(UN1)[Fertility == min(Fertility)]
```

'Hong.Kong'

```
# Las dos localidades con mayores residuos positivos de la regresión son:
rownames(UN1)[order(resid(reg1),decreasing=TRUE)[c(1,2)]]
```

1. 'Equatorial.Guinea'
2. 'Oman'

```
# Las dos localidades con mayores residuos negativos de la regresión son:
rownames(UN1)[order(resid(reg1),decreasing=FALSE)[c(1,2)]]
```

1. 'Armenia'
2. 'Ukraine'

Problema 2.7

2.7.1

Notemos que:

$$\mathbb{E}\left[y|x\right] = y_i = \beta_1 x_i + u_i$$

con $i = 1, ..., n$ y $u_i \sim \mathcal{N}(0, \sigma^2)$ con $\operatorname{cov}\left(u_i, u_j\right) = 0$ para $i \neq j$. Si escribimos:

- $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$
- $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$
- $U = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$

Tenemos que el modelo es:

$$Y = X\beta_1 + U$$

Entonces buscamos un β_1 tal que los espacios generados sean ortogonales; es decir:

$$\begin{aligned}
\left\langle U, X\beta_1 \right\rangle = 0 &\Longrightarrow \left\langle Y - X\beta_1, X\beta_1 \right\rangle = 0 \\
&\Longrightarrow \left\langle Y, X\beta_1 \right\rangle - \left\langle X\beta_1, X\beta_1 \right\rangle = 0 \\
&\Longrightarrow \beta_1^T X^T Y - \beta_1 X^T X\beta_1 = 0 \\
&\Longrightarrow \beta_1 \left(X^T Y - X^T X\beta_1\right) = 0 \\
&\Longrightarrow X^T Y - X^T X\beta_1 = 0 \\
&\Longrightarrow \beta_1 = (X^T X)^{-1} X^T Y
\end{aligned}$$

Por lo tanto el estimador de β_1 es:

$$\begin{aligned}
\widehat{\beta_1} &= (X^T X)^{-1} X^T Y \\
&= \left(\sum_{i=1}^{n} x_i^2\right)^{-1} \sum_{i=1}^{n} x_i y_i \\
&= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

Por otro lado notemos que:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\beta_1}\right] &= \mathbb{E}\left[(X^TX)^{-1}X^TY\right] \\
&= (X^TX)^{-1}X^T\mathbb{E}\left[Y\right] \\
&= (X^TX)^{-1}X^T\left(\mathbb{E}\left[X\beta_1+U\right]\right) \\
&= (X^TX)^{-1}X^T\left(X\beta_1+\cancelto{\mathbf{0}}{\mathbb{E}[U]}\right) \\
&= \cancelto{\mathbf{1}}{\left[(X^TX)^{-1}X^TX\right]}\beta_1 \\
&= \beta_1
\end{aligned}
$$

Por tanto es insesgado. Además:

$$
\begin{aligned}
\mathbb{V}\left[\widehat{\beta_1}\right] &= \mathbb{V}\left[(X^TX)^{-1}X^TY\right] \\
&= \mathbb{V}\left[(X^TX)^{-1}X^T(X\beta_1+U)\right] \\
&= \mathbb{V}\left[(X^TX)^{-1}X^TX\beta_1+(X^TX)^{-1}X^TU\right] \\
&= \mathbb{V}\left[(X^TX)^{-1}X^TU\right] \\
&= (X^TX)^{-1}X^T\mathbb{V}\left[U\right]\left[(X^TX)^{-1}X^T\right]^T \\
&= (X^TX)^{-1}X^T\sigma^2 Id\cdot X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1} \\
&= \sigma^2\left(\sum_{i=1}^n x_i^2\right) \\
&= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}
\end{aligned}
$$

Por ultimo buscamos un estimador para σ^2, entonces procedemos a utilizar el estimador de suma de cuadrados del error; es decir:

$$
\begin{aligned}
SCE &= \|e_i\|^2 \\
&= \left\|Y-X\widehat{\beta_1}\right\|^2 \\
&= \left\langle Y-X\widehat{\beta_1},Y-\widehat{\beta_1}\right\rangle \\
&= \langle Y,Y\rangle-2\left\langle Y,X\widehat{\beta_1}\right\rangle+\left\langle X\widehat{\beta_1},X\widehat{\beta_1}\right\rangle \\
&= \sum_{i=1}^n y_i^2-2\widehat{\beta_1}\langle Y,X\rangle+\left\{\widehat{\beta_1}\right\}^2\langle X,X\rangle \\
&= \sum_{i=1}^n y_i^2-2\frac{\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n x_i^2}+\frac{\left\{\sum_{i=1}^n x_iy_i\right\}^2}{\sum_{i=1}^n x_i^2} \\
&= \sum_{i=1}^n y_i^2-\frac{\left(\sum_{i=1}^n x_iy_i\right)^2}{\sum_{i=1}^n x_i^2}
\end{aligned}
$$

Por otro lado notemos que:

$$\begin{aligned}
X^T(\widehat{Y}-Y) &= \sum_{i=1}^{n} x_i(\widehat{y}_i - y_i) \\
&= \sum_{i=1}^{n} \widehat{\beta_1} x_i^2 - x_i y_i \\
&= \widehat{\beta_1} S_{xx} - S_{xy} \\
&= \frac{S_{xy}}{S_{xx}} S_{xx} - S_{xy} \\
&= 0
\end{aligned}$$

Por tanto se obtiene solo una restricción:

$$\sum_{i=1}^{n} x_i e_i = 0$$

Por tanto el estimador de varianza tiene sólo $n-1$ grados de libertad. Por tanto tendremos que:

$$\widehat{\sigma^2} = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right]$$

2.7.2

Notemos que para un modelo de regresión simple estándar se tiene que:

$$SCE_1 = SCT - \frac{S_{xy}^2}{S_{xx}}$$

Mientras que para nuestro modelo sabemos que:

$$SCE = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}$$

Por tanto:

$$\begin{aligned}
SCR &= \sum_{i=1}^{n} \widehat{y}_i{}^2 - n\overline{\widehat{y}}^2 \\
&= \frac{S_{xy}^2}{S_{xx}} - n\overline{\widehat{y}}^2 + S_{yy} - S_{yy} \\
&= S_{yy} - n\overline{\widehat{y}}^2 - SCE \\
\implies SCR + SCE &= S_{yy} - n\overline{\widehat{y}}^2
\end{aligned}$$

Además:

$$SCT = S_{yy} - n\overline{y}^2$$

Por tanto:

$$SCR + SCE = SCT + \underbrace{n\overline{y}^2 - n\overline{\widehat{y}}^2}_{\neq 0}$$

Luego con tal de mantener la igualdad deseada se tiene que:

$$SCT^* = S_{yy}$$
$$SCR^* = \sum_{i=1}^{n} \widehat{y_i}^2$$

Luego SCT^* tiene n grados de libertad y SCR por construcción tiene 1 grado de libertad puesto que SCE tiene $n-1$ grados de libertad. Entonces la ANOVA viene dada por:

Fuente	grados de libertad	Suma de Cuadrados	Suma de Cuadrados Media	F
Regresion	1	SCR	MCR	
Error	n-1	SCE	σ^2	$\frac{MCR}{\sigma^2}$
Total	n	$\sum y_i^2$	MCT	

Para chequear que son numéricamente equivalente utilizaremos la información del problema siguiente 2.7.4

Notemos que:

```
reg_x = lm( snake$Y ~ snake$X )
```

```
(summary (reg_o) $ coefficients[ 3 ])^2
anova (reg_o ) $ 'F value' [ 1]
```

Imprime
1558.66110339189
1558.66110339189
Concluyendo lo solicitado.

2.7.3

La regresión por el origen viene dada por:

```
reg_o = lm( snake$Y ~ 0+snake$X )
summary(reg_o)
```

```
Call:
lm(formula = snake$Y ~ 0 + snake$X)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4207 -1.4924 -0.1935  1.6515  3.0771

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
snake$X  0.52039    0.01318   39.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.7 on 16 degrees of freedom
Multiple R-squared:  0.9898,Adjusted R-squared:  0.9892
F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16
```

Y el intervalo de confianza viene dado por:

```
confint(reg_o)
```

	2.5 %	97.5 %
snake$X	0.492451	0.548337

Por tanto el intervalo de confianza para β_1 es $[0{,}492451; 0{,}548337]$.

Ahora hacemos una regresión lineal simple para hacer el test:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

```
reg_x = lm( snake$Y ~ snake$X )
```

```
beta_test(reg_x,"two.sided",'(Intercept)',0)
```

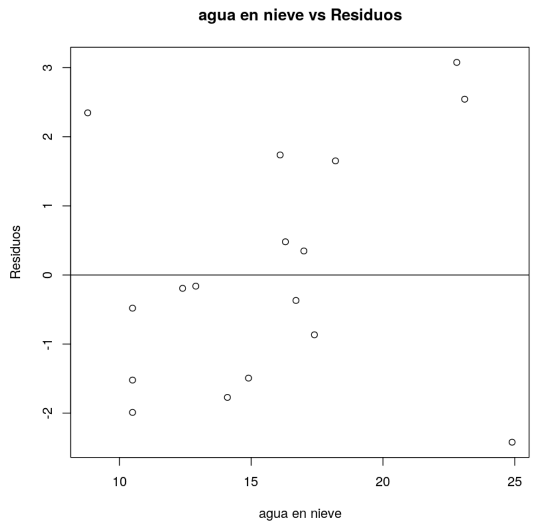
```
            Estimate Std.Err t value Pr(>t)
(Intercept)  0.72538 1.54882  0.4683 0.6463
```

Por tanto la significancia mínima para rechazar H_0 es de un 64.63 %. Por tanto por ejemplo para una significancia de 0.05 entonces no rechazo H_0, por ende podría pensar que el intercepto es 0.

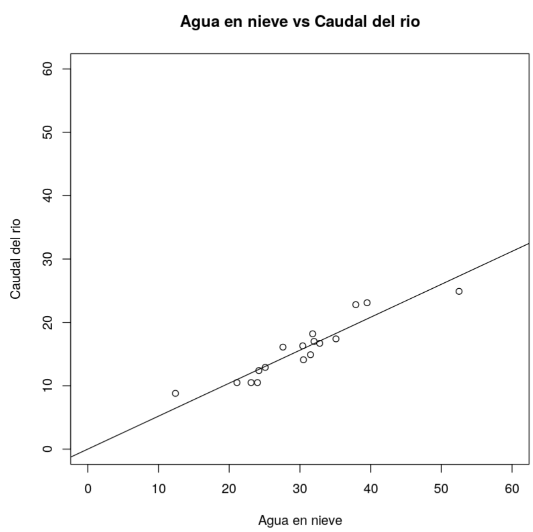
2.7.4

Realizamos los plots que nos solicitan.

```
plot(snake$Y,reg_o$residual ,xlab='agua en nieve',ylab='Residuos',main='agua en nieve vs␣
 ↪Residuos')
abline(h=0)
```




```
plot(snake$X,snake$Y,xlab='Agua en nieve',ylab='Caudal del rio' ,main='Agua en nieve vs␣
 ↪Caudal del rio',xlim=c(0,60),ylim=c(0,60))
abline(reg_o)
```



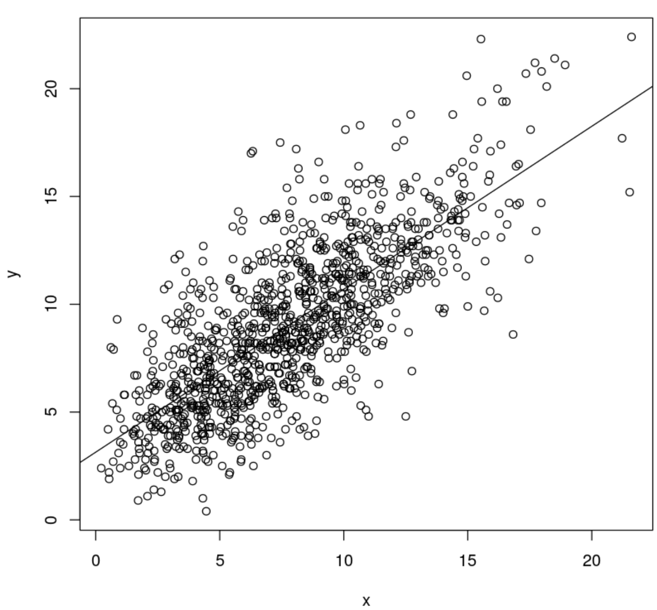
Problema 2.13

```
head(wm1)
attach(wm1)
```

	Date	CSpd	RSpd
	<fct>	<dbl>	<dbl>
1	2002/1/1/0	6.9	5.9666
2	2002/1/1/6	7.1	7.2176
3	2002/1/1/12	7.8	7.9405
4	2002/1/1/18	6.9	6.0174
5	2002/1/2/0	5.5	6.1646
6	2002/1/2/6	3.1	1.7687

2.13.1

```
y = CSpd
x = RSpd
reg2<-lm(y~x)
plot(x,y)
abline(lm(y~x))
```



Dada la estructura monótona de los datos y de crecimiento mas menos constante, es posible ajustar un modelo lineal para los datos.

2.13.2

```r
summary(reg2)
```

```

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7877 -1.5864 -0.1994  1.4403  9.1738 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.14123    0.16958   18.52   <2e-16 ***
x            0.75573    0.01963   38.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.466 on 1114 degrees of freedom
Multiple R-squared:  0.5709,Adjusted R-squared:  0.5705 
F-statistic:  1482 on 1 and 1114 DF,  p-value: < 2.2e-16
```

El valor del R^2 es 0.57, por tanto sólo la mitad de la variación en CSpd es explicada por RSpd. Por tanto podríamos decir que este modelo es de calidad mediana.

2.13.3

```r
valor_a_predecir=data.frame(x=7.4285)
predict(reg2,valor_a_predecir,interval='prediction')
```

	fit	lwr	upr
1	8.755197	3.914023	13.59637

Por tanto el intervalo de confianza para la predicción para un 95 % de confiabilidad es:

$$[3{,}914023; 13{,}59637]$$

2.13.4

En primer lugar notemos que:

$$\frac{1}{m}\sum_{i=1}^{m}\tilde{y}_{*i} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\beta}_0+\hat{\beta}_1 x_{*i}\right) = \hat{\beta}_0+\hat{\beta}_1\frac{1}{m}\sum_{i=1}^{m}x_{*i} = \hat{\beta}_0+\hat{\beta}_1\bar{x}_*$$

Por tanto el promedio de las predicciones es el mismo que la predicción del promedio.
En segundo lugar notemos que:

$$\mathbb{V}\left[\overline{\tilde{y}}_*\right] = \sigma^2\left[\frac{1}{n}+\frac{(\overline{x}-\overline{x}_*)^2}{S_{xx}}\right]$$

y

$$\begin{aligned}
\mathbb{V}\left[\overline{\tilde{y}}_*-\overline{y_*}\right] &= \sigma^2\left[\frac{1}{n}+\frac{(\overline{x}-\overline{x}_*)^2}{S_{xx}}\right]+\mathbb{V}\left[\overline{y_*}\right]-2\cancelto{0}{\mathrm{cov}\left(\overline{\tilde{y}}_*,\overline{y_*}\right)}\\
&= \sigma^2\left[\frac{1}{n}+\frac{(\overline{x}-\overline{x}_*)^2}{S_{xx}}\right]+\frac{1}{m^2}m\sigma^2\\
&= \sigma^2\left[\frac{1}{n}+\frac{(\overline{x}-\overline{x}_*)^2}{S_{xx}}\right]+\frac{1}{m}\sigma^2
\end{aligned}$$

Luego si el estimador de σ^2 es $\widehat{\sigma^2}$ se tiene que la desviación estándar estimada del error viene dada por:

$$\widehat{\sigma^2}\left[\frac{1}{n}+\frac{(\overline{x}-\overline{x}_*)^2}{S_{xx}}\right]+\frac{1}{m}\widehat{\sigma^2}$$

Observación 1: Este error estándar no es el promedio de los errores estándar para las predicciones individuales, ya que todas las predicciones están correlacionadas.
Observación 2: La covarianza es 0 pues $y_* = \beta_0+\beta_1 x_*+u_*$ y u_* es normal de media 0 y varianza σ^2 e independiente.

2.13.5

Notemos que nos estan diciendo que:

$$\bar{x}_* = 7{,}4285$$

y que:

$$m = 62039$$

Sabemos que el $S_e = 2{,}466$ es el estimador de σ, reemplazando tenemos que:

```
S_e2 = 2.466^2
n = length(x)
Sxx = sum(x^2)
x_ast = 7.4285
y_ast = 3.14123 + 0.75573*x_ast
m = 62039
```

```
Varianza_del_error=(S_e2*(1/n + (mean(x)-x_ast)^2/Sxx) + (1/m)*S_e2)
```

```
gamma = 0.95
valor_t= qt((1+gamma)/2,n-2)
limite_inferior = y_ast - sqrt(Varianza_del_error)*valor_t
limite_superior = y_ast + sqrt(Varianza_del_error)*valor_t
print(limite_inferior)
print(limite_superior)
```

```
[1] 8.608919
[1] 8.901422
```

Por tanto el intervalo de confianza para la predicción es:

$$[8{,}608919; 8{,}901422]$$

Bibliografía

[1] .Applied Linear Regression", Sanford Weisberg, Wiley Interscience.

[2] "The Elements of Statistical Learning", Hastie-Tibshirani-Friedman.