Trabajo Practico N6

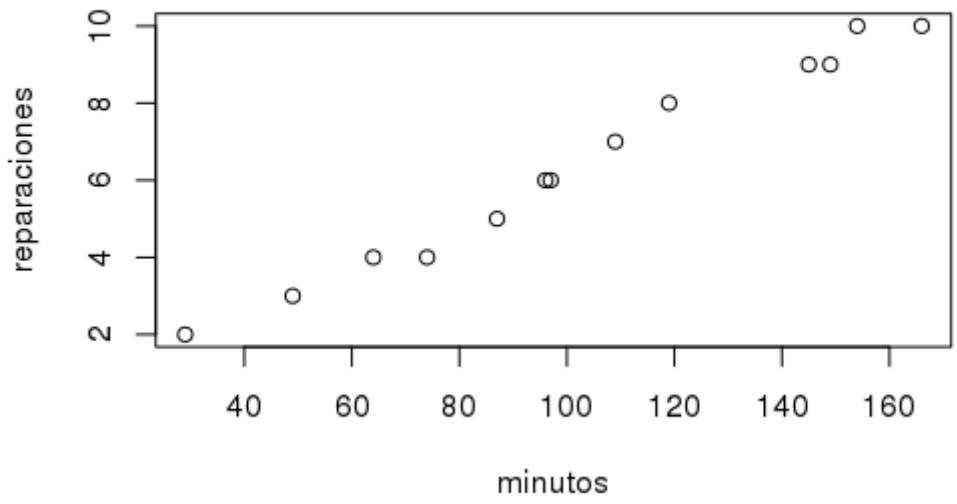
Regresion Lineal

Lucas Mufato

1. Regresión lineal simple. A partir del dataset presentado a continuación, el cual consiste en un pequeño relevamiento del tiempo que demandan las llamadas a servicio técnico de una empresa y la cantidad de unidades de hardware reparadas a partir de la misma, realice las siguientes operaciones:

a. Realice un scatter plot y comente brevemente que observa sobre la relación entre las variables.

Se puede observar una relacion entre todas la duracion de las llamadas y las reparaciones



b. Ahora, calcule el coeficiente de correlación, es consistente con su hipótesis del punto anterior?

```
> llam.cor <- cor(llamadas)
> print(llam.cor)
              minutos reparaciones
minutos      1.000000     0.992341
reparaciones 0.992341     1.000000
```

SI. es consistente.

c. Ahora verifique los supuestos de la regresión lineal simple a efectos de determinar si es posible avanzar sobre el análisis de regresión. ¿Qué conclusiones obtiene? ¿Cuál es la importancia de validar estos supuestos?

Si, es posible obtener una regresion lineal ya que las variables estan fuertemente relacionadas. De no ser asi, el modelo creado tendria muy poca precision para predecir la variable objetivo.

Los otros supuestos se validaran en los puntos siguientes.

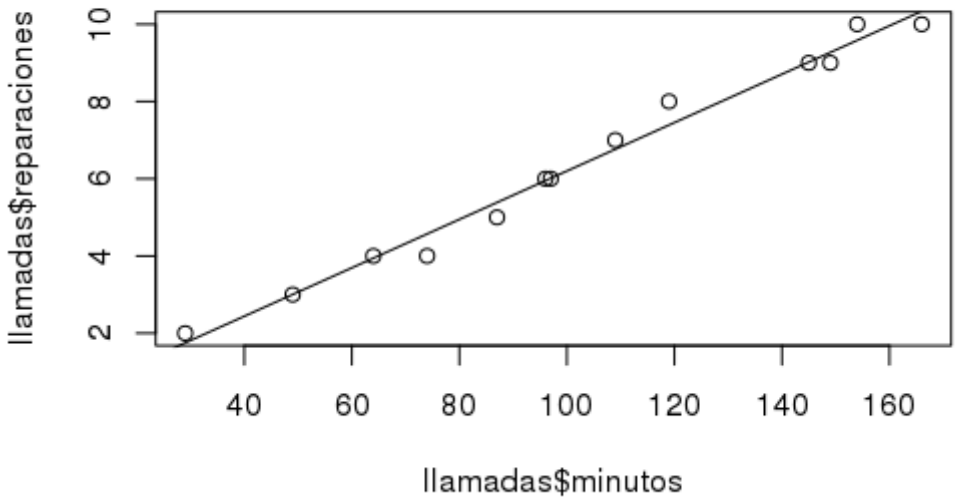
d. En caso de ser posible, estime los parámetros de la regresión manualmente, utilizando el método de mínimos cuadrados.

```
#C crear la recta de regresion

mediay <- mean(llamadas$X1.1)
mediax <- mean(llamadas$X23)

#para calcular B1
#la parte de arriba de la division
partearriba <- 0
for (i in 1:13) {
  partearriba <- partearriba + ((llamadas[i,3]-mediay)*(llamadas[i,2]-mediax))
}

#la parte de abajo de la division
parteabajo <- 0
for(i in llamadas$X23){
  parteabajo <- parteabajo + (i-mediax)^2
}

b1<- partearriba/parteabajo
b0<- mediay - (b1 * mediax)
```

de esto se obtienen los valores:
b0 = `-0.06493589`
b1 = `0.0626638`
Los cuales son iguales a los calculados por la funcion LM

e. Grafique la recta obtenida sobre el scatter plot del primer punto. ¿Que observa a simple vista sobre la bondad del ajuste?

A simple vista se puede observar que la regresion lineal explica bien los valores. Varios caen sobre la recta y otros bastante cerca.



f. Calcule el R2 y compare los resultados contra lo observado en el punto anterior.

```
dfllamadas <- data.frame(X23 =llamadas$X1.1)
prediccionesF <- predict(llam.reg,dfllamadas )
#ahora que tengo los valores predictos para cada valor
#uso la formulaa
R <- (cor(llamadas$X23,prediccionesF) )^2
print(R)
[1] 0.9847406
```

Al R^2 dar un numero tan cercano a 1 indica que la regresion lineal precide muy bien los atributos

g. Realice las predicciones para los valores [160, 25, 119]. Que observa en este problema en particular?

```
df <- data.frame(minutos =c(160,25,119))
> predicciones <- predict(llam.reg,df )
> print(predicciones)
       1        2        3
9.961272 1.501659 7.392057
```

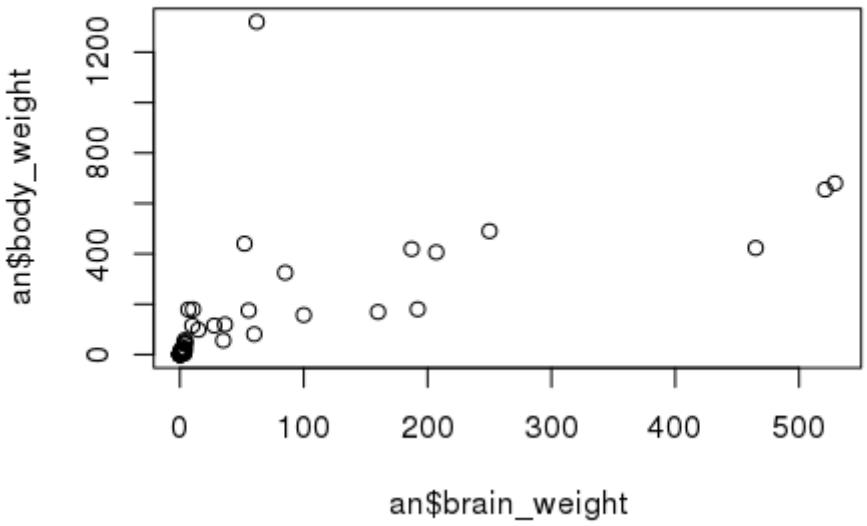
En este caso tendria una complicacion al para saber si con 25minutos de llamada se arreglan 1 o 2 equipos, dado que no se pueden arreglar 1,5.

2. Se cuenta con el dataset animals.csv con el registro del peso medio del cerebro y el cuerpo de un grupo de especies de mamíferos. Incorpore tal dataset en R y repita el procedimiento de la consigna anterior. Documente las conclusiones.

A) Mediante el calculo de la correlacion se puede observar que las variables estan fuertemente correlacionadas

```
             brain_weight body_weight
brain_weight    1.0000000   0.9341638
body_weight     0.9341638   1.0000000
```

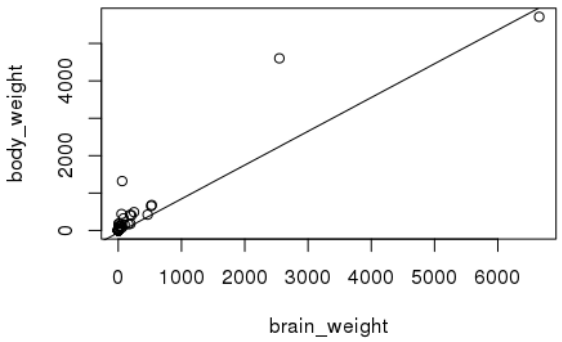
Esto tambien se muestra mediante un grafico(en el cual no se grafican 2 observaciones muy grandes, para que se puedan ver las otras variables mas pequeñas)



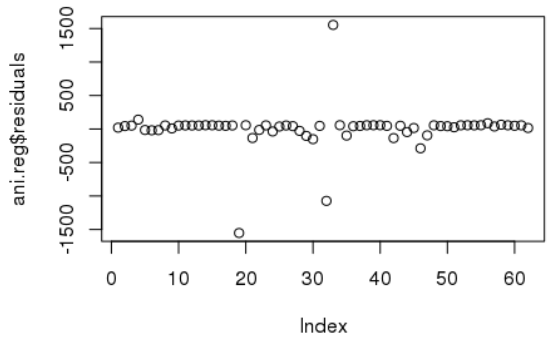
En el grafico se puede ver (con un poco de imaginacion) que hay una relacion positiva entre las variables.

Al crear la recta de regresion lineal queda con la forma:
Y = -56.85555 * 0.90291 X

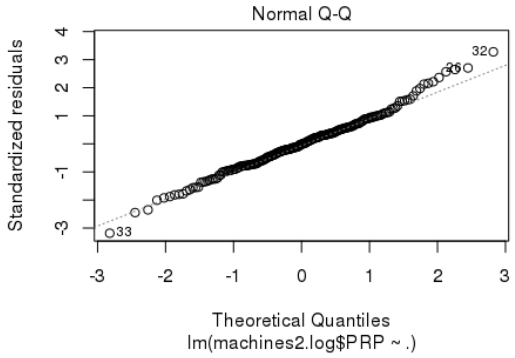
Dando un R= 0.8705 y un R^2=0.8727. Esto indica que el modelo explica bien los datos.
Al superponer la recta de regresion sobre los datos queda:



SUPUESTOS:
LINEALIDAD:
se demuestra en el grafico anterior ya que lo puntos forman una recta(se aprecia mejor en el grafico mas pequeño)
HOMOCEDASTICIDAD:
se demuestra el supuesto dado que los residuos son constantes (hay 3 casos particulares que llaman la atencion)



NORMALIDAD:
los datos se apegan a la recta, por lo que implica que se cumple el supuestos



INDEPENDENCIA:
la prueba durbin watson dio apenas sobre el limite:

```
DW = 2.548, p-value = 0.9861
alternative hypothesis: true autocorrelation is greater than 0

por lo tanto se puede considerar que no se cumple el supuesto.
```

3. Regresión lineal múltiple. Cargue el dataset machines.csv en R:

a. Explore las características de los datos, instancias, cantidad de atributos, su significado y el tipo de datos de cada uno.

El dataset cuenta con 209 instancias y 9 variables, 2 de estas variables son categoricas.
Su significado es:

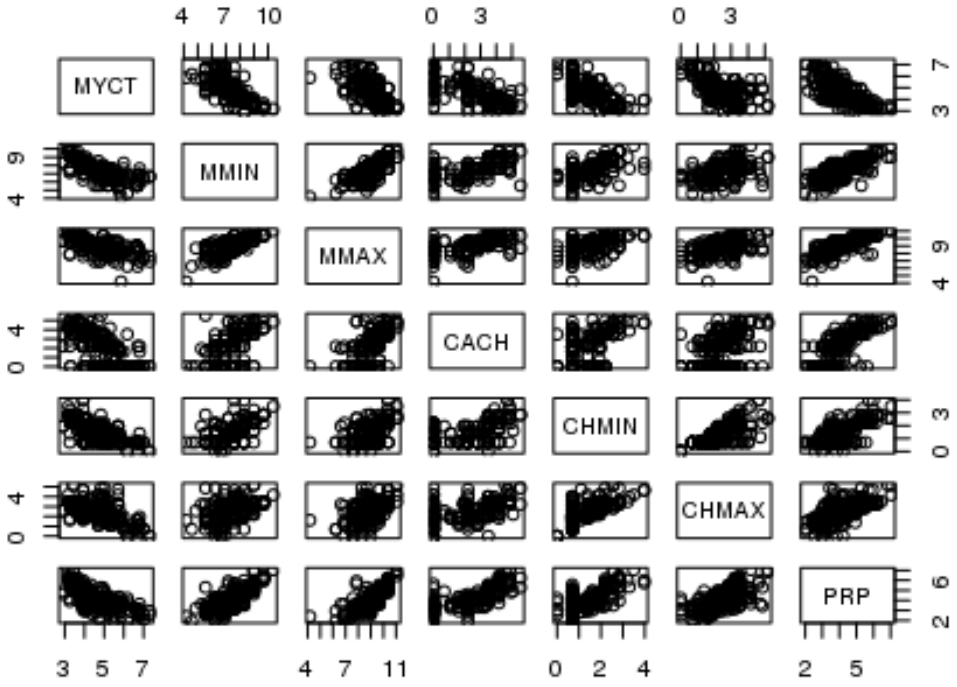
1. vendor name: 30 (adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang)
2. Model Name: many unique symbols
3. MYCT: machine cycle time in nanoseconds (integer)
4. MMIN: minimum main memory in kilobytes (integer)
5. MMAX: maximum main memory in kilobytes (integer)
6. CACH: cache memory in kilobytes (integer)

7. CHMIN: minimum channels in units (integer)
8. CHMAX: maximum channels in units (integer)
9. PRP: published relative performance (integer)

b. Realice un gráfico scatter plot en busca de la relación entre los diferentes atributos y la performance del CPU y concluya al respecto.

tuve que aplicar una transformacion logaritmica para observar de mejor manera la relacion entre las variables. Al aplicarle la transformacion se puede observar una relacion lineal entre las variables.

c. Ahora verifique los supuestos de la regresión lineal múltiple a efectos de determinar si es posible avanzar sobre el análisis de regresión.
¿Qué conclusiones obtiene? ¿Cuál es la diferencia con relación a los supuestos de la regresión lineal múltiple?



Despues de aplicar la transformacion logaritmica se observa que se cumple el supuesto de linealidad.

En el sub-intem siguiente se comprueba los otros supuestos.
La diferencia con los supuestos de la regresion lineal simple es que hay un supuesto mas, el de colinealidad, que indica que los datos independientes (las X) deben ser independientes entre ellos

d. Utilizando la herramienta, estime los parámetros de la regresión y grafique la recta ajustada sobre el scatter plot.
Creo la regresion:

machines2.reg <- lm(machines2.log$PRP~. , machines2.log)
summary(machines2.reg)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.172608   0.537865  -2.180  0.03040 *
MYCT         0.008777   0.047818   0.184  0.85455
MMIN         0.193740   0.048012   4.035 7.74e-05 ***
MMAX         0.314586   0.046797   6.722 1.79e-10 ***
CACH         0.186278   0.024540   7.591 1.15e-12 ***
CHMIN        0.196694   0.059239   3.320  0.00107 **
CHMAX        0.128308   0.043763   2.932  0.00376 **
```

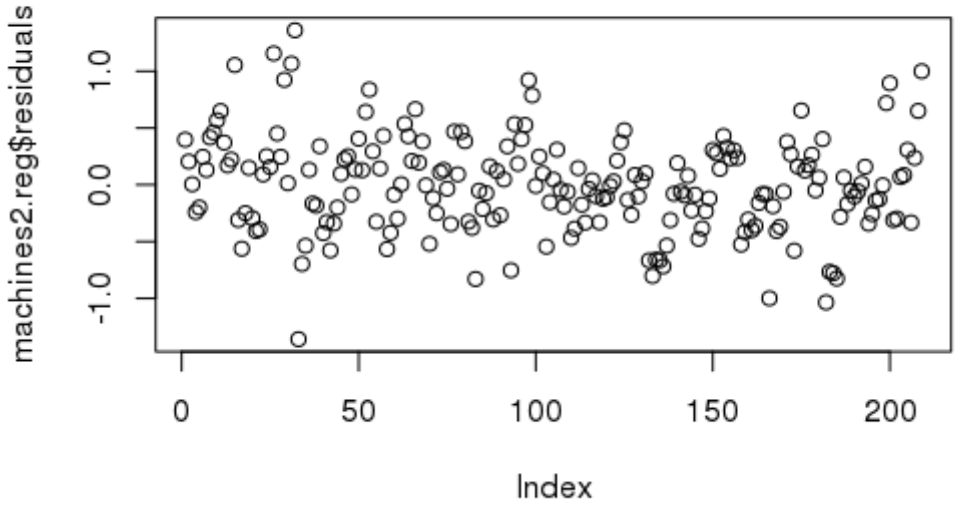
se observa que todas las variables menos MYCT pasa la prueba, demostrando que no hay datos suficiente para indicar que es cero.

```
Residual standard error: 0.4311 on 202 degrees of freedom
Multiple R-squared:  0.8289,    Adjusted R-squared:  0.8238
F-statistic: 163.1 on 6 and 202 DF,  p-value: < 2.2e-16
```

Tambien se puede observar un alto R^2 y R^2 ajustado, y ambos con mucha similitud. Dando confianza al modelo creado.

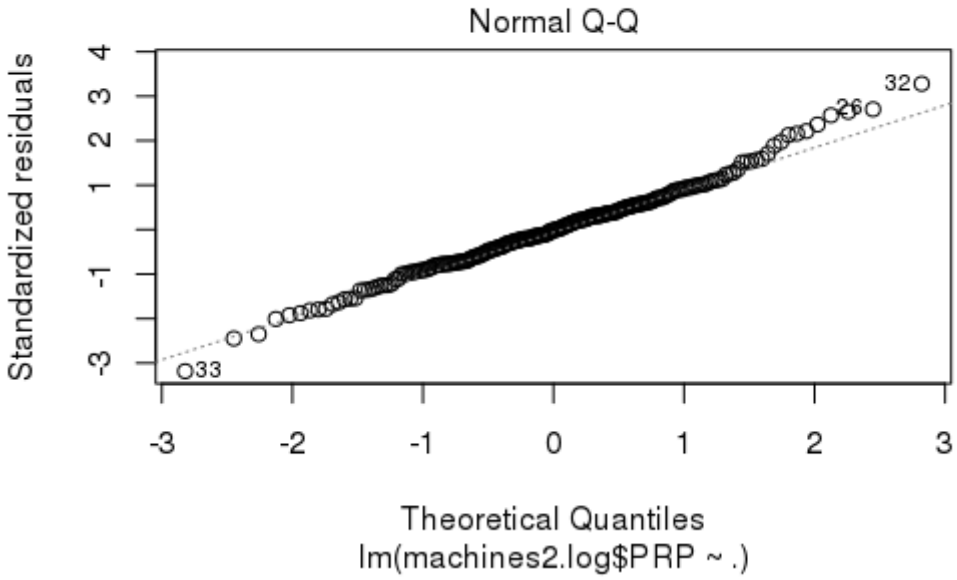
Comprobación de los demas supuestos:

- INDEPENDENCIA:
  - `data:  machines2.reg`
  - `DW = 1.1608, p-value = 1.337e-10`
  - `SE comprueba el supuesto`
- HOMOCEDASTICIDAD:
  - En la imagen se demuestra que se cumple con este supuesto ya que no se observa ninguna relacion entre los distintos puntos.

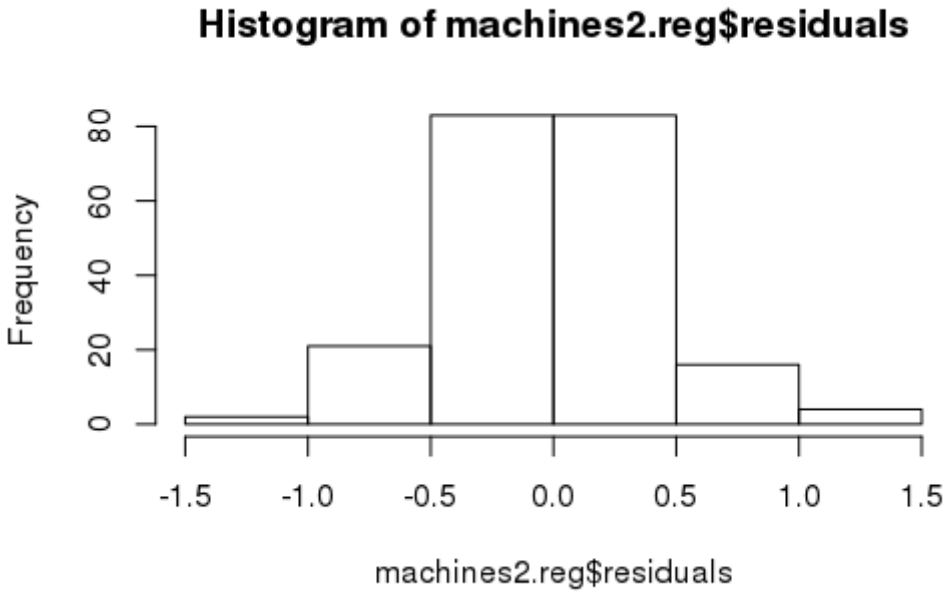


  -

- NORMALIDAD:
  - se observa graficamente al ver que la linea de puntos sigue la grafica Qqplot



- Tambien al ver una forma de campana en el histograma



- NO-COLINEALIDAD:

	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
MYCT	1.0000000	-0.3356422	-0.3785606	-0.3209998	-0.3010897	-0.2505023	-0.3070994
MMIN	-0.3356422	1.0000000	0.7581573	0.5347291	0.5171892	0.2669074	0.7949313
MMAX	-0.3785606	0.7581573	1.0000000	0.5379898	0.5605134	0.5272462	0.8630041
CACH	-0.3209998	0.5347291	0.5379898	1.0000000	0.5822455	0.4878458	0.6626414
CHMIN	-0.3010897	0.5171892	0.5605134	0.5822455	1.0000000	0.5482812	0.6089033
CHMAX	-0.2505023	0.2669074	0.5272462	0.4878458	0.5482812	1.0000000	0.6052093
PRP	-0.3070994	0.7949313	0.8630041	0.6626414	0.6089033	0.6052093	1.0000000

Los atributos MMAX y MMIN parecerian estar relacionados, por lo que el supuesto no se cumpliria

e. Concluya acerca de la bondad del error de estimación de la recta ajustada. ¿Cuál es la diferencia con respecto a la regresión lineal simple?

f. Utilizando la función de regresión resultante, realice la predicción para las siguientes instancias del nuevo modelo bdm-2016 y concluya respecto de esas predicciones:

Adviser	bdm-2016	80	2048	16000	256	4	32
Amdahl	bdm-2016	95	4096	32000	32	8	128
Bti	bdm-2016	188	1024	6000	0	16	32

los valores de las predicciones fueron:
1) 5482.269
2) 10883.938
3) 2093.638

estos estan MUY por encima de los datos originales, por lo que pude haber tenido un error al clasificar que variable es de que tipo y por eso el error.

4. Ahora, explore el dataset crimen.arff en Weka y realice el procedimiento para el análisis de regresión.

- **a. En caso de ser posible, estime los parámetros utilizando una de las herramientas y concluya al respecto.**
- **b. Realice las estimaciones para las instancias que aparecen en crimen-prediccion.csv. Documente los resultados.**

En la carpeta ej4 en el archivo “modelo.txt” se encuentra el modelo lineal creado por weka.

No puedo lograr hacer las predicciones, como resultado de la misma dice error.