2020_02_27

Fernando Anorve

3/3/2020

Regresion

Ejemplo 1 - Anscombe dataset

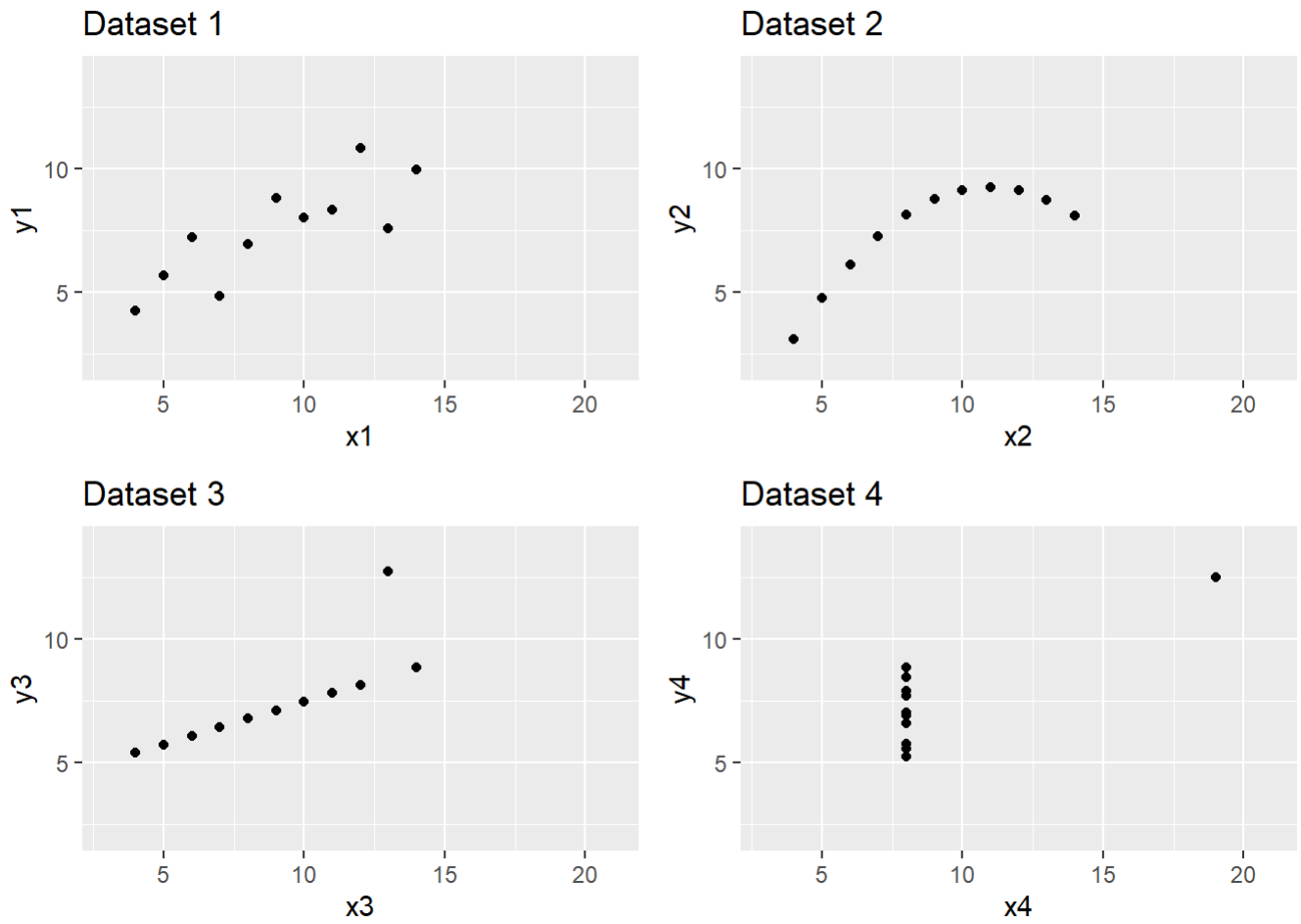
Por que no hay que confiar ciegamente en un coeficiente? No siempre R^2 tiene la razón absoluta.

Para verlo usaremos un conjunto de datos bastante conocido “anscombe”, construido por el estadístico Francis Anscombe en 1975 para demostrar la importancia de graficar datos y buscar comportamientos atípicos antes de construir modelos. Esta colección consta de cuatro conjuntos con los que usualmente se calculan modelos de regresión lineal simple.

```
data("anscombe")
anscombe
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

```
p1 <- qplot(x1, y1, data = anscombe) + ggtitle("Dataset 1") + ylim(c(2,14)) + xlim(c(3,21))
p2 <- qplot(x2, y2, data = anscombe) + ggtitle("Dataset 2") + ylim(c(2,14)) + xlim(c(3,21))
p3 <- qplot(x3, y3, data = anscombe) + ggtitle("Dataset 3") + ylim(c(2,14)) + xlim(c(3,21))
p4 <- qplot(x4, y4, data = anscombe) + ggtitle("Dataset 4") + ylim(c(2,14)) + xlim(c(3,21))

grid.arrange(p1, p2, p3 , p4, nrow = 2)
```



¡Hagan sus apuestas! (no realmente)

- ¿Cual creen que tenga “mejor valor” de R^2?
- ¿Cuál línea creen que tenga una pendiente más pronunciada?

Pero más importante:

- ¿Cuál(es) creen que se ajuste mejor a un modelo lineal?

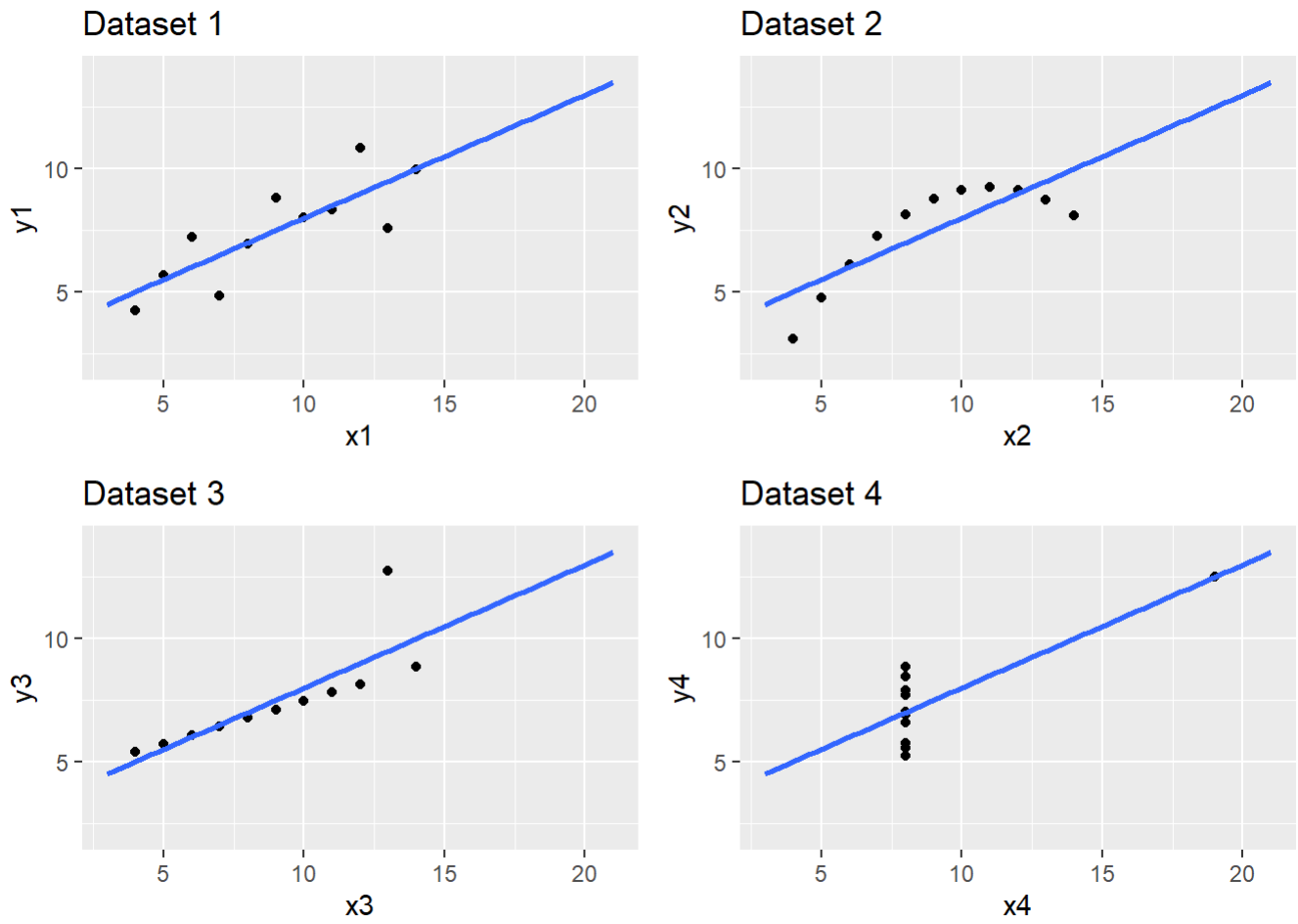
```
lm1 <- lm(y1 ~ x1, data = anscombe)
lm2 <- lm(y2 ~ x2, data = anscombe)
lm3 <- lm(y3 ~ x3, data = anscombe)
lm4 <- lm(y4 ~ x4, data = anscombe)


info <- rbind(lm1$coefficients,lm2$coefficients,lm3$coefficients,lm4$coefficients)
info <- cbind(info , c(summary(lm1)$r.squared , summary(lm2)$r.squared , summary(lm3)$r.squared
, summary(lm4)$r.squared ))
colnames(info)[2:3] = c("(slope)", "r.squared")

round(info, digits = 2)
```

```
##      (Intercept) (slope) r.squared
## [1,]           3     0.5      0.67
## [2,]           3     0.5      0.67
## [3,]           3     0.5      0.67
## [4,]           3     0.5      0.67
```

¡En realidad tienen los mismos coeficientes y el mismo R^2 !

Ya con la linea de ajuste, vemos que no todos quedan igual

```
p1 <- p1 + geom_smooth(method='lm',se=F, fullrange = T)
p2 <- p2 + stat_smooth(method='lm',se=F, fullrange = T)
p3 <- p3 + stat_smooth(method='lm',se=F, fullrange = T)
p4 <- p4 + stat_smooth(method='lm',se=F, fullrange = T)

grid.arrange(p1, p2, p3 , p4, nrow = 2)
```



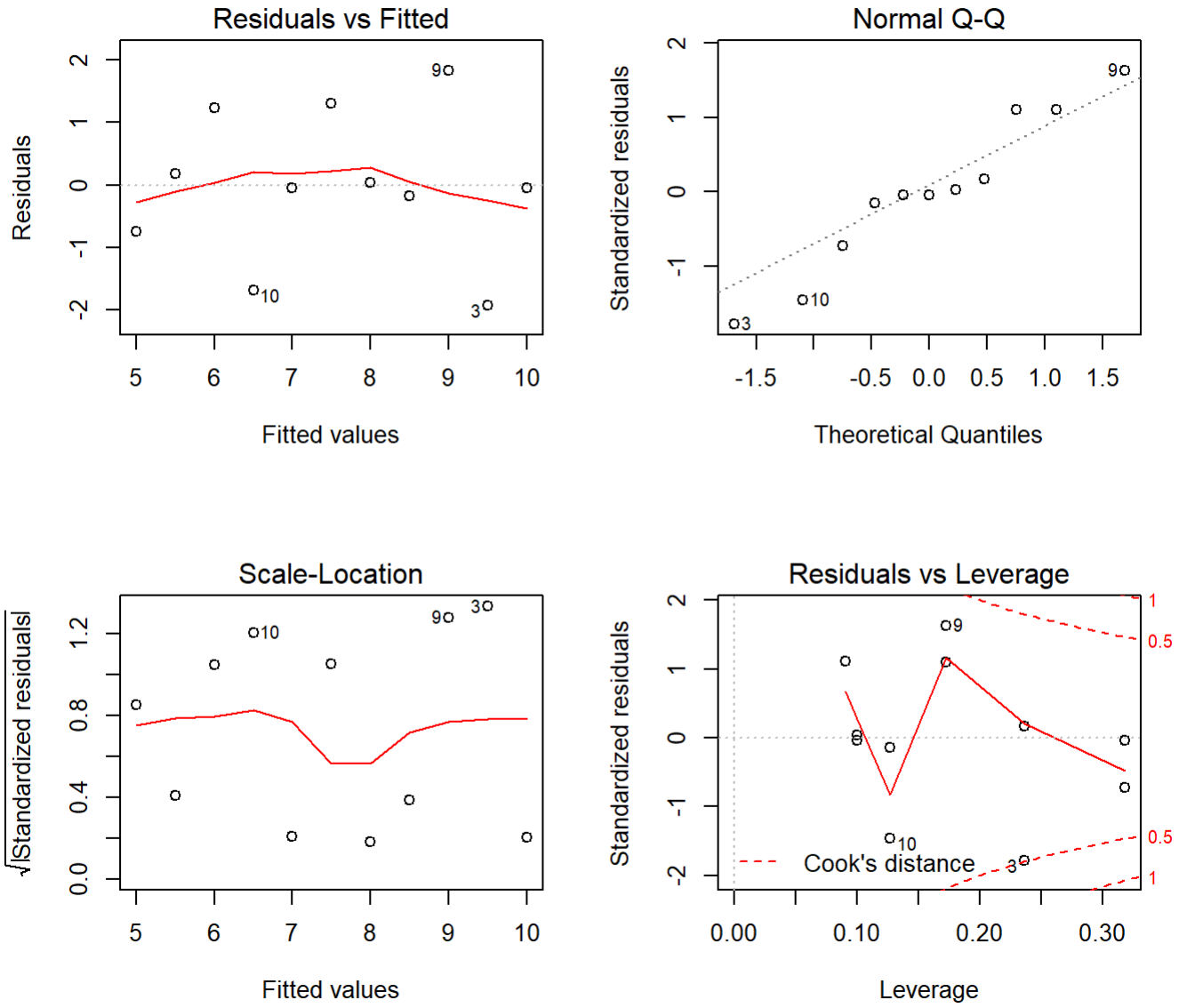
Veamos por ejemplo qué pasa con los primeros tres:

Los plots contienen:

1. Residuals vs Fitted, para observar si los residuales parecen tener patrones lineales (buscamos residuos distribuidos de forma uniforme)
2. Normal Q-Q, para revisar si los residuos parecen seguir una distribución normal (buscamos una tendencia lineal)

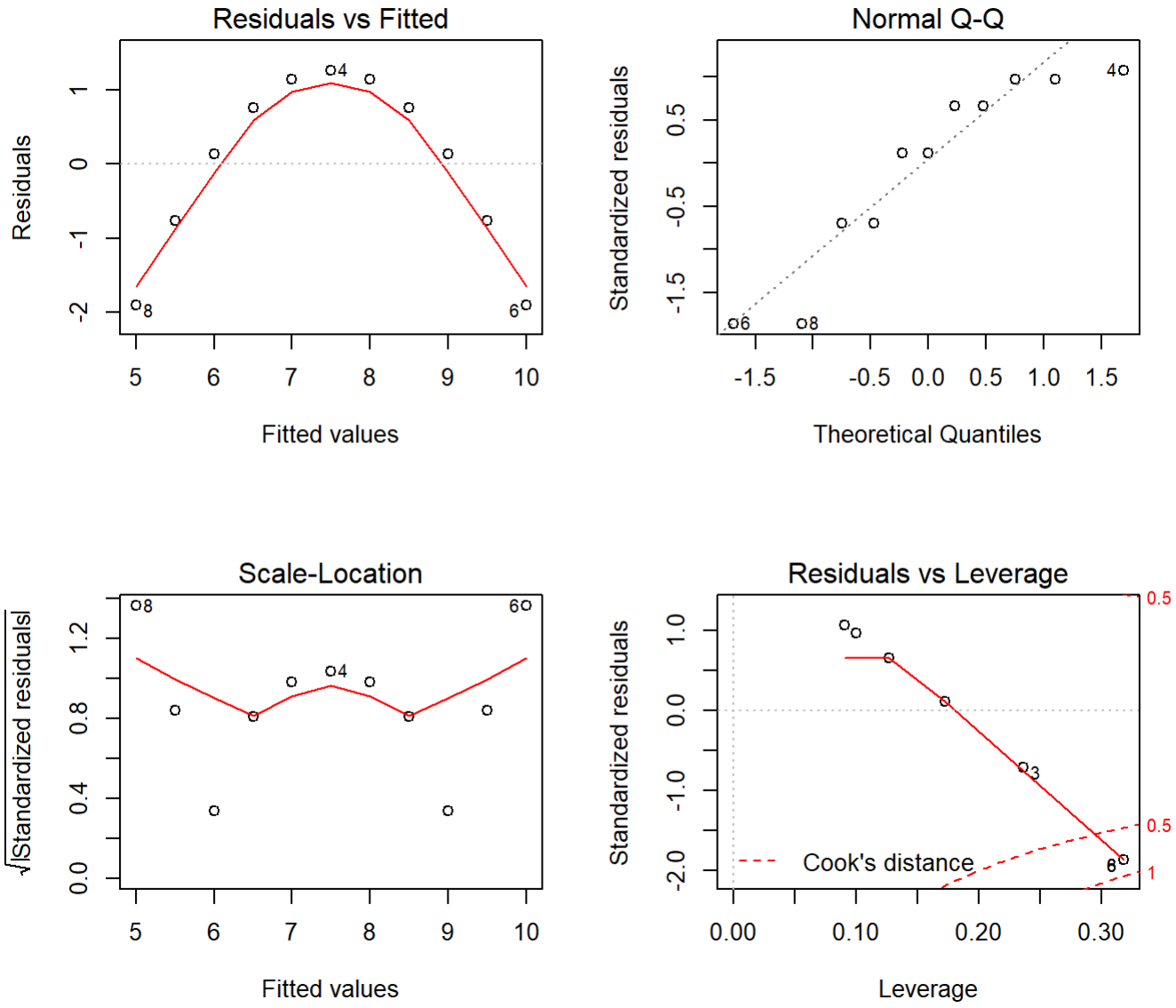
3. Scale-Location, para estudiar la homocedasticidad de las varianzas (buscamos una residuos distribuidos de forma uniforme)
4. Residuals vs Leverage, para buscar posibles observaciones influyentes

```
par(mfrow = c(2,2))
plot(lm1)
```



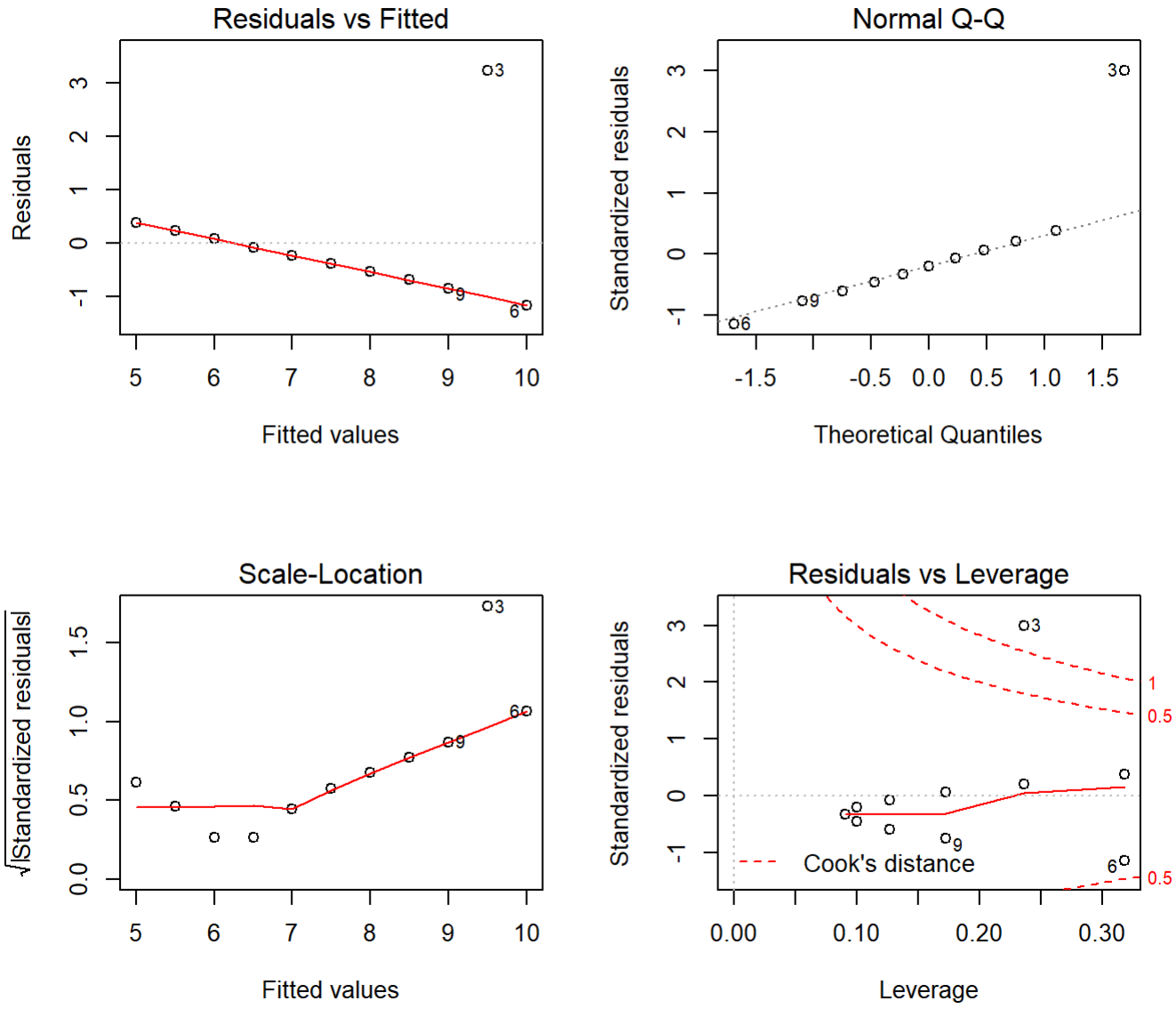
Todo parece estar en orden respecto a lm1

```
par(mfrow = c(2,2))
plot(lm2)
```



La primera gráfica muestra un comportamiento no lineal, por lo que nuestro modelo probablemente no se de utilidad respecto a los datos

```
par(mfrow = c(2,2))
plot(lm3)
```



La tercera observación se sale de los rangos en la última gráfica. Evidencia de valor erróneo o atípico, habría que revisar el contexto.

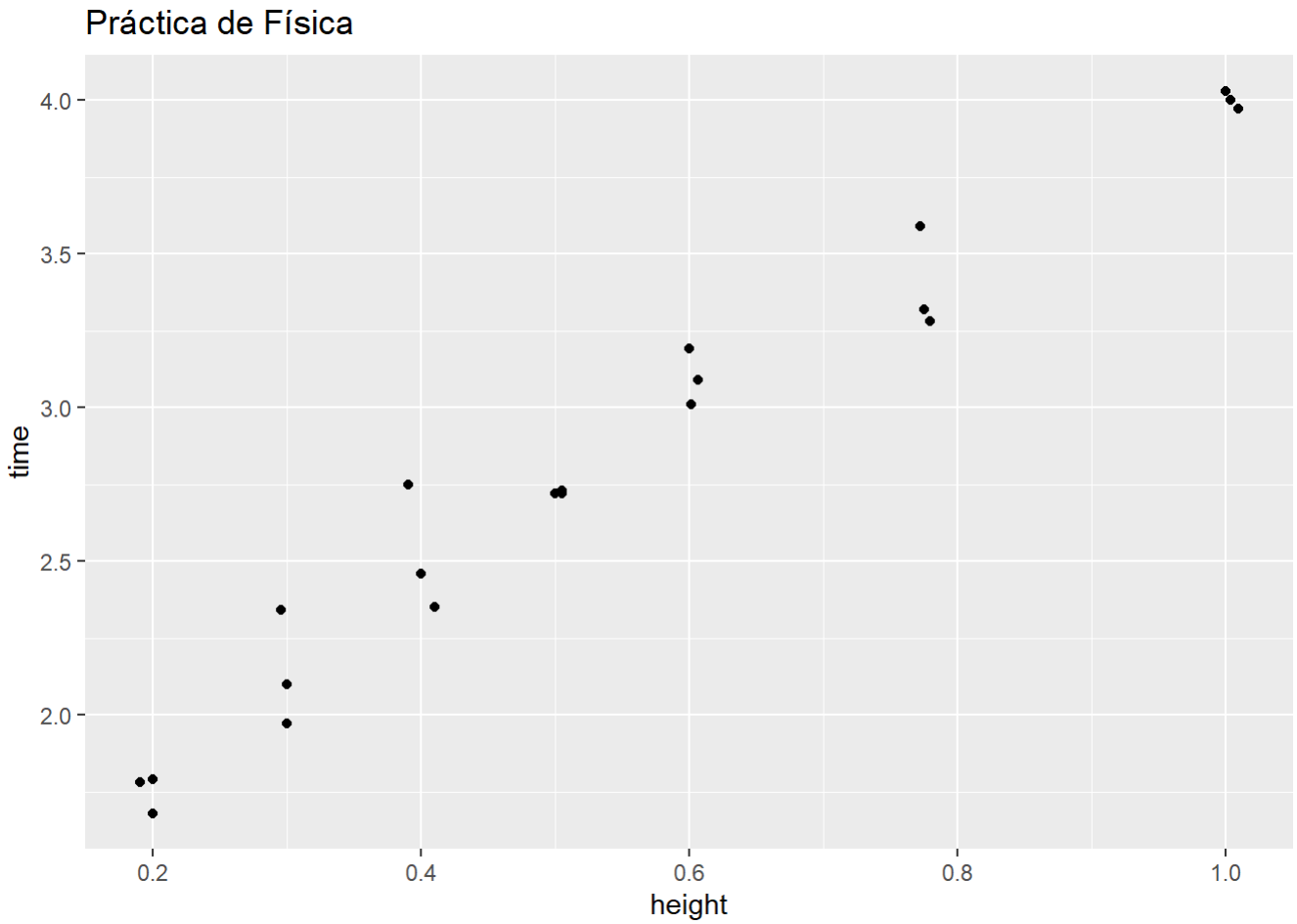
Ejemplo 2 - práctica de física

El archivo *investigation01_en.pdf* contiene un ejemplo práctica de física que sirve como referencia para profesores de preparatoria.

El propósito de esta práctica es tratar de hallar la relación entre el tiempo que necesita una pelota para rebotar seis veces dependiendo de la altura desde la que cae.

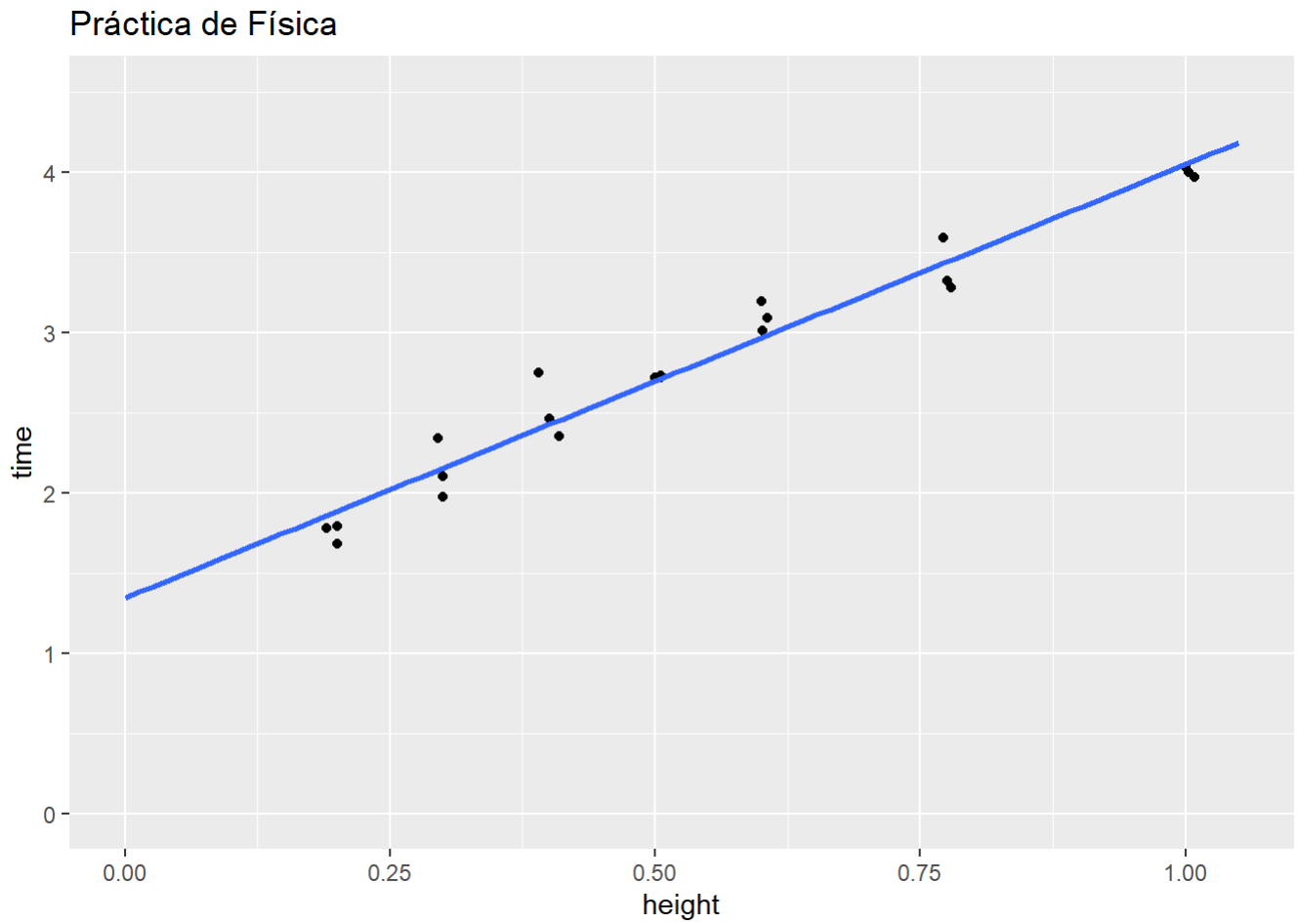
Se obtienen los siguientes resutados

```
results  <- read.csv("physics_lab.txt")

fisica = ggplot(results, aes(x = height , y = time)) + geom_point() + ggtitle("Práctica de Física")

fisica
```



Un modelo lineal luce bastante conveniente. De hecho, el valor de R^2 es 0.9614

```
fisica + geom_smooth(method='lm',se=F, fullrange = T)+ ylim(c(0,4.5)) + xlim(c(0,1.05))
```



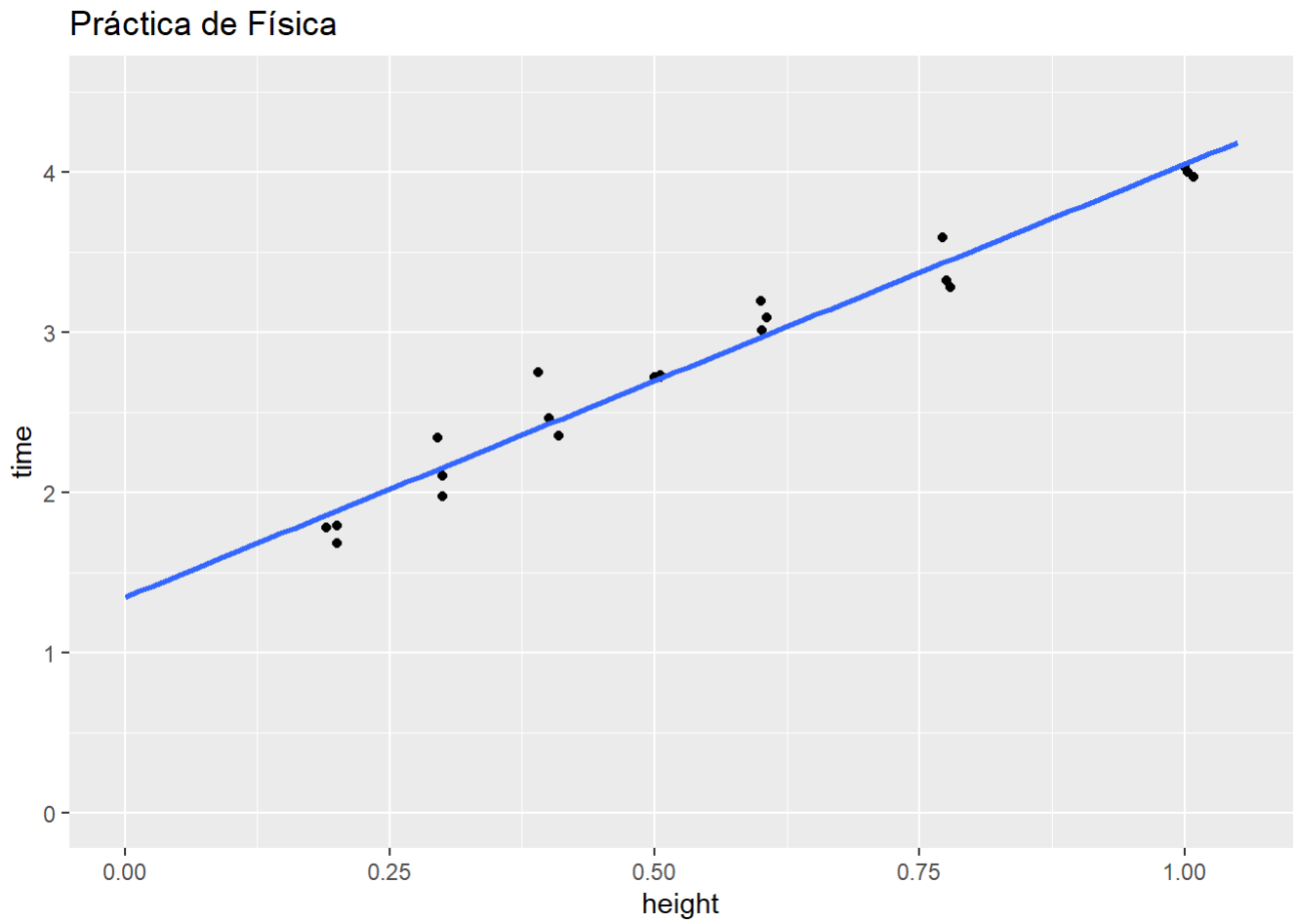

```
modelo_fisica = lm(time ~ height, data = results)

summary(modelo_fisica)
```

```
##
## Call:
## lm(formula = time ~ height, data = results)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20382 -0.10172 -0.01738  0.04170  0.35233
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.34293    0.07448   18.03 2.08e-13 ***
## height       2.70445    0.12430   21.76 6.85e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1479 on 19 degrees of freedom
## Multiple R-squared:  0.9614, Adjusted R-squared:  0.9594
## F-statistic: 473.4 on 1 and 19 DF,  p-value: 6.849e-15
```

Pero… ¿tiene sentido en términos físicos?

Si se lanza desde una altura cercana a cero, ¿cuánto tardaría en rebotar seis veces?

```
fisica + ylim(c(0,4.5)) + xlim(c(0,1.05)) + geom_smooth(method='lm',se=F, fullrange = T)
```




```
modelo_fisica$coefficients
```

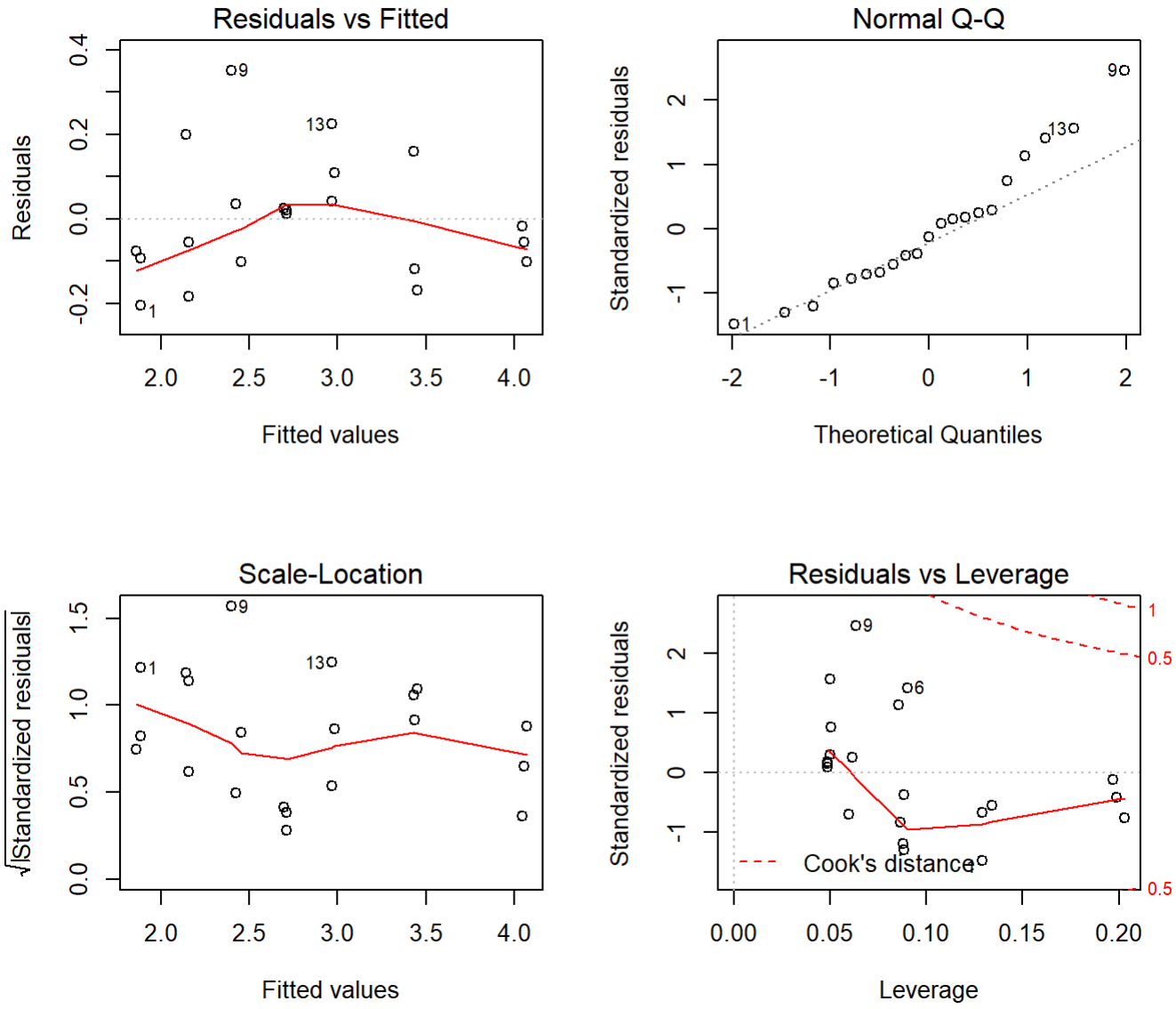
```
## (Intercept)      height
##    1.342931    2.704450
```

¡Poco menos de un segundo y medio!

Eso… no tiene mucho sentido

¿Qué dicen las gráficas?

```
par(mfrow = c(2,2))
plot(modelo_fisica)
```



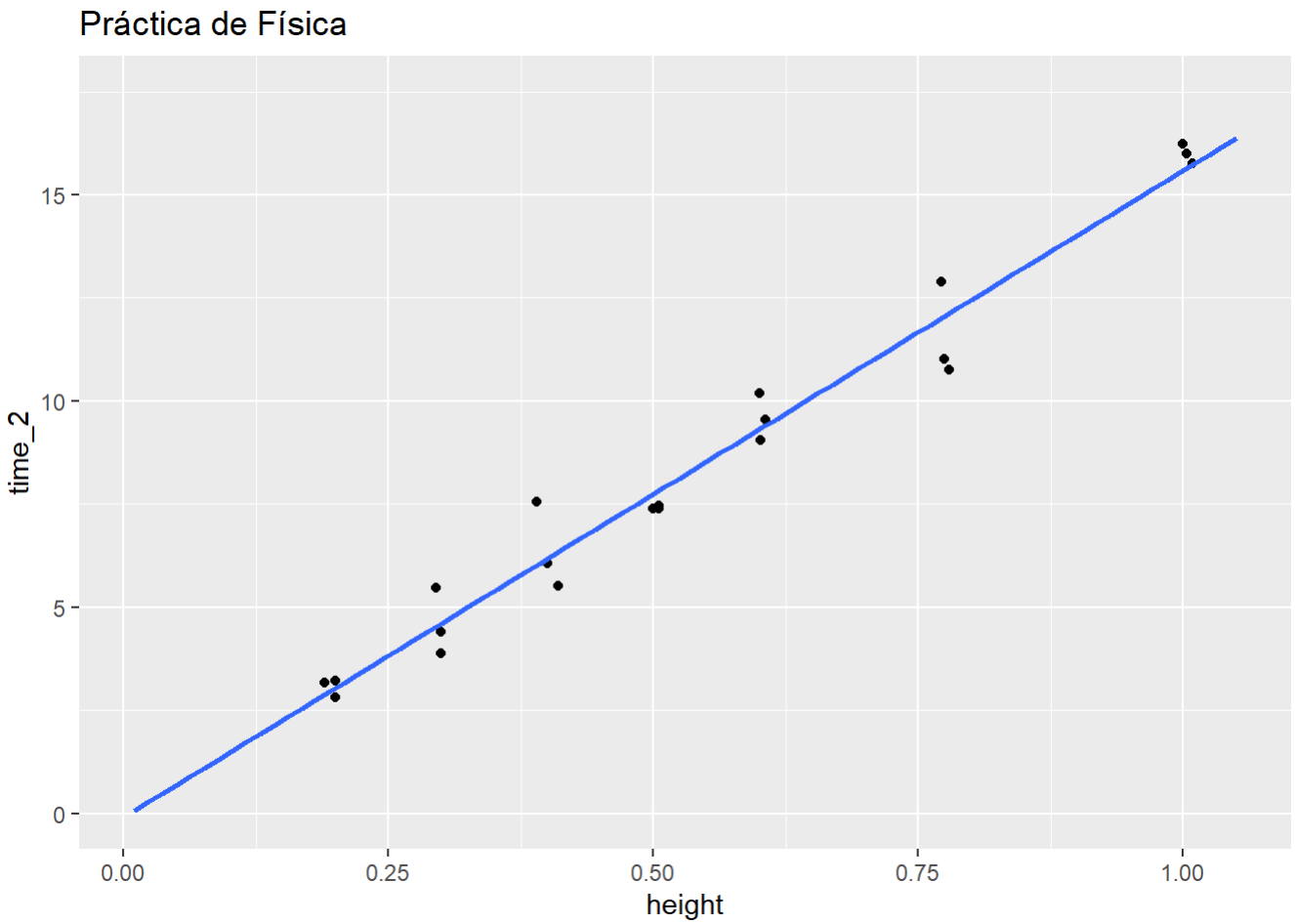
No parecen muy convincentes: la primera gráfica podría sugerir un comportamiento no lineal

La teoría dice que para una sola caída (con tiempo t_1):

$$H = \frac{1}{2} g \cdot t_1^2$$

¿Qué pasa si reescalamos T a T^2?

```
results$time_2 = results$time^2
fisica_2 = ggplot(results, aes(x = height , y = time_2)) + geom_point() + ggtitle("Práctica de F
ísica") +
  geom_smooth(method='lm',se=F, fullrange = T)+ ylim(c(0,17.5)) + xlim(c(0.01,1.05))
fisica_2
```




```
modelo_2 <- lm(time_2 ~ height, data = results)
summary(modelo_2)
```

```
##
## Call:
## lm(formula = time_2 ~ height, data = results)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3627 -0.3696 -0.1236  0.3646  1.5442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1003     0.3691  -0.272    0.789
## height       15.6886     0.6160  25.469 3.78e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7329 on 19 degrees of freedom
## Multiple R-squared:  0.9715, Adjusted R-squared:   0.97
## F-statistic: 648.7 on 1 and 19 DF,  p-value: 3.78e-16
```

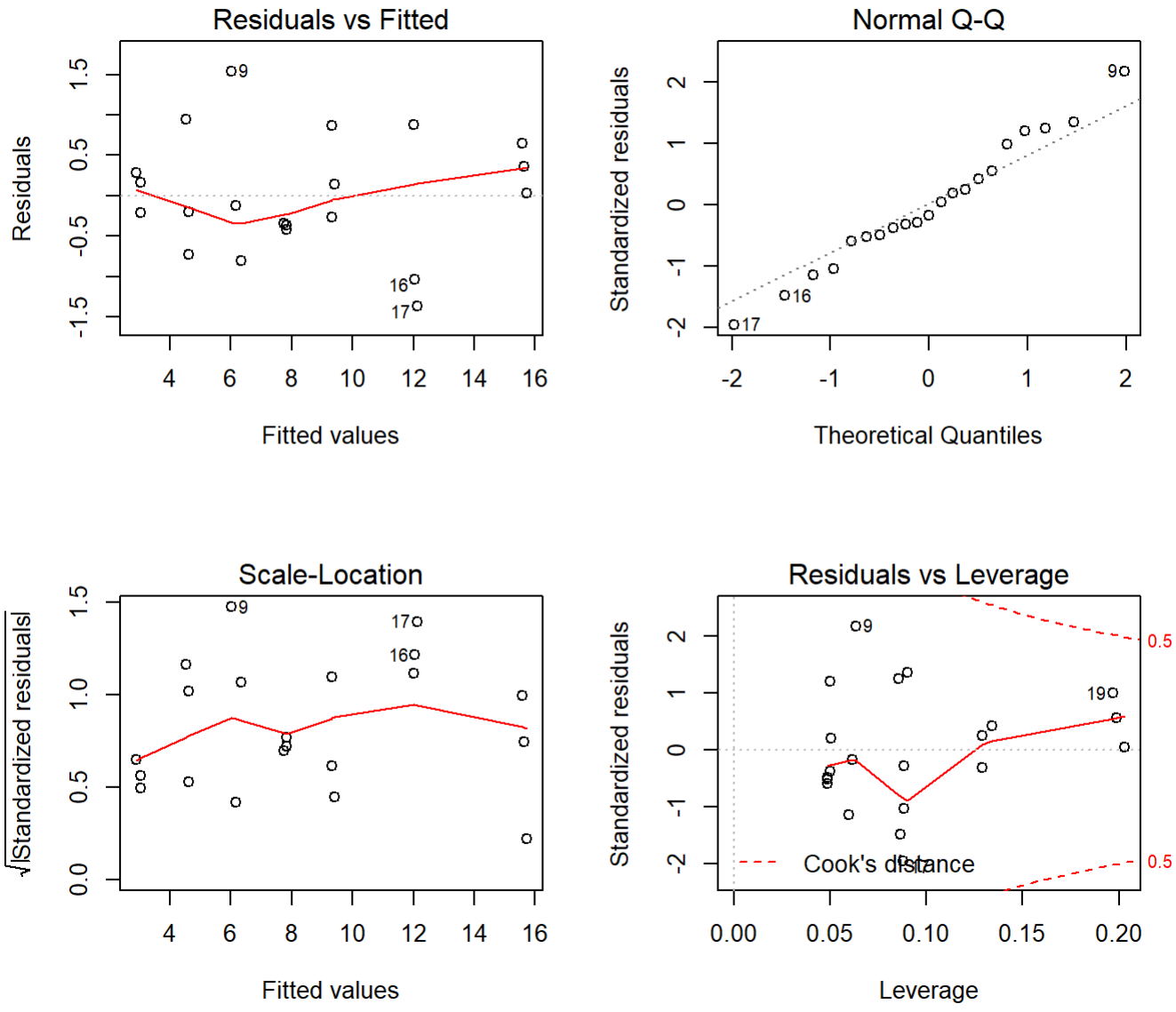
```
modelo_2$coefficients
```

```
## (Intercept)      height
##  -0.1002614  15.6885529
```

Eso… tiene más sentido

¿Qué dicen las gráficas?

```
par(mfrow = c(2,2))
plot(modelo_2)
```



¡Luce mucho mejor!

Ejemplo 3: Galapagos

El siguiente ejemplo concierne al numero de especies de tortugas en las islas galapagos Hay treinta casos (islas) y siete variables en el conjunto de datos.

El primer paso es leer en R y examinar los datos de gala.txt

```
gala <- read.table("gala.txt", header=T) # read the data into R
gala
```

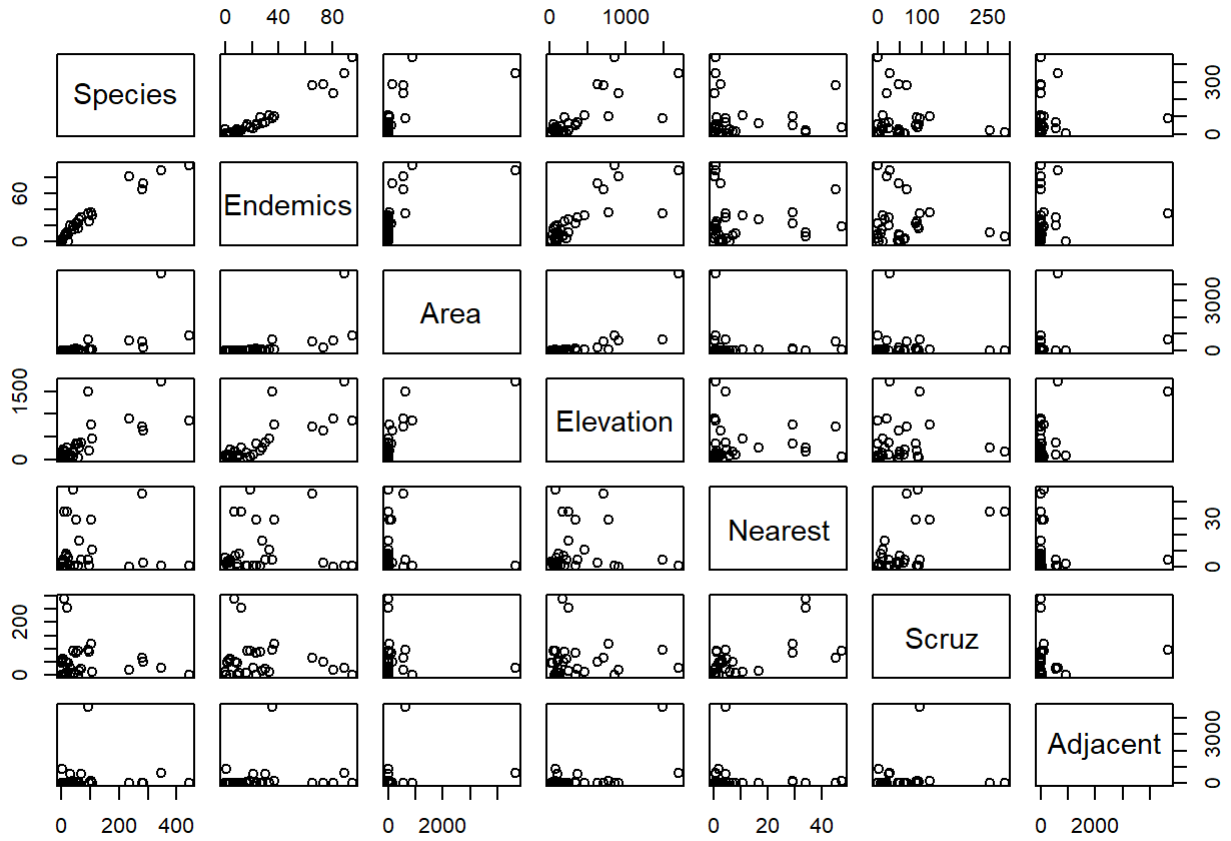
```
##              Species Endemics    Area Elevation Nearest Scruz Adjacent
## Baltra            58       23   25.09       346     0.6   0.6     1.84
## Bartolome         31       21    1.24       109     0.6  26.3   572.33
## Caldwell           3        3    0.21       114     2.8  58.7     0.78
## Champion          25        9    0.10        46     1.9  47.4     0.18
## Coamano            2        1    0.05        77     1.9   1.9   903.82
## Daphne.Major      18       11    0.34       119     8.0   8.0     1.84
## Daphne.Minor      24        0    0.08        93     6.0  12.0     0.34
## Darwin            10        7    2.33       168    34.1 290.2     2.85
## Eden               8        4    0.03        71     0.4   0.4    17.95
## Enderby            2        2    0.18       112     2.6  50.2     0.10
## Espanola          97       26   58.27       198     1.1  88.3     0.57
## Fernandina        93       35  634.49      1494     4.3  95.3  4669.32
## Gardner1          58       17    0.57        49     1.1  93.1    58.27
## Gardner2           5        4    0.78       227     4.6  62.2     0.21
## Genovesa          40       19   17.35        76    47.4  92.2   129.49
## Isabela          347       89 4669.32      1707     0.7  28.1   634.49
## Marchena          51       23  129.49       343    29.1  85.9    59.56
## Onslow             2        2    0.01        25     3.3  45.9     0.10
## Pinta            104       37   59.56       777    29.1 119.6   129.49
## Pinzon           108       33   17.95       458    10.7  10.7     0.03
## Las.Plazas        12        9    0.23        94     0.5   0.6    25.09
## Rabida            70       30    4.89       367     4.4  24.4   572.33
## SanCristobal     280       65  551.62       716    45.2  66.6     0.57
## SanSalvador      237       81  572.33       906     0.2  19.8     4.89
## SantaCruz        444       95  903.82       864     0.6   0.0     0.52
## SantaFe           62       28   24.08       259    16.5  16.5     0.52
## SantaMaria       285       73  170.92       640     2.6  49.2     0.10
## Seymour           44       16    1.84       147     0.6   9.6    25.09
## Tortuga           16        8    1.24       186     6.8  50.9    17.95
## Wolf              21       12    2.85       253    34.1 254.7     2.33
```

Las variables son:

- Especies - El numero de especies de tortugas encontradas en la isla
- Endemics - El numero de especies endemicas
- Area - El area de la isla (km2)
- Elevation - La elevacion mas alta de la isla (m)
- Nearest - La distancia a la isla mas cercana (km)
- Scruz - La distancia desde la isla de Santa Cruz (km)
- Adjacent - El area de la isla adyacente mas cercana (km2)

Los datos fueron tomados de Johnson and Raven (1973) ,Weisberg (1985). Algunos datos fueron sustituidos por simplicidad

```
pairs(gala)
```



Potencialmente hay una observacion atipicamente con gran Area. Cual es?

```
which(gala$Area>2000)
```

```
## [1] 16
```

```
gala[16,]
```

```
##         Species Endemics    Area Elevation Nearest Scruz Adjacent
## Isabela     347       89 4669.32      1707     0.7  28.1   634.49
```

Cooks distance.

```
cooks.distance(lm(Species~.,dat=gala))
```

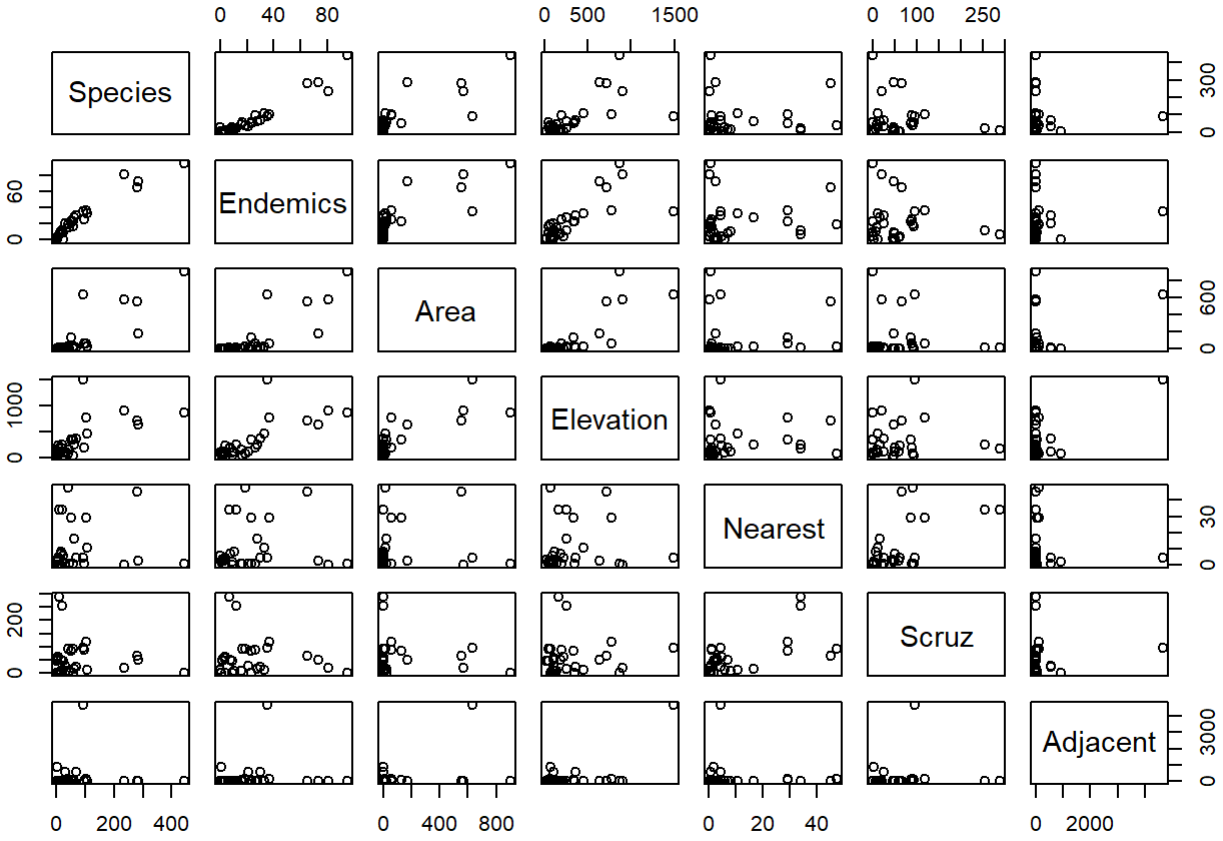
```
##       Baltra    Bartolome     Caldwell     Champion      Coamano
## 3.173296e-03 8.243862e-02 1.771526e-03 1.143940e-04 1.016584e-02
## Daphne.Major Daphne.Minor       Darwin         Eden      Enderby
## 1.112359e-03 4.648268e-02 3.662081e-03 1.601596e-03 3.262790e-03
##     Espanola   Fernandina     Gardner1     Gardner2     Genovesa
## 1.042326e-03 4.891519e+00 1.805953e-06 5.285715e-03 1.358381e-01
##      Isabela     Marchena       Onslow        Pinta       Pinzon
## 4.965186e+01 7.577071e-03 1.410599e-03 5.692596e-03 2.138641e-05
##   Las.Plazas       Rabida SanCristobal  SanSalvador    SantaCruz
## 9.508673e-04 1.072623e-02 2.935472e-01 4.237064e-01 7.381213e-01
##      SantaFe   SantaMaria      Seymour      Tortuga         Wolf
## 1.827796e-02 4.809711e-03 2.016842e-04 3.654999e-04 1.185074e-03
```

```
cooks.distance(lm(Species~.,dat=gala))[16] #Yikes!
```

```
##  Isabela
## 49.65186
```

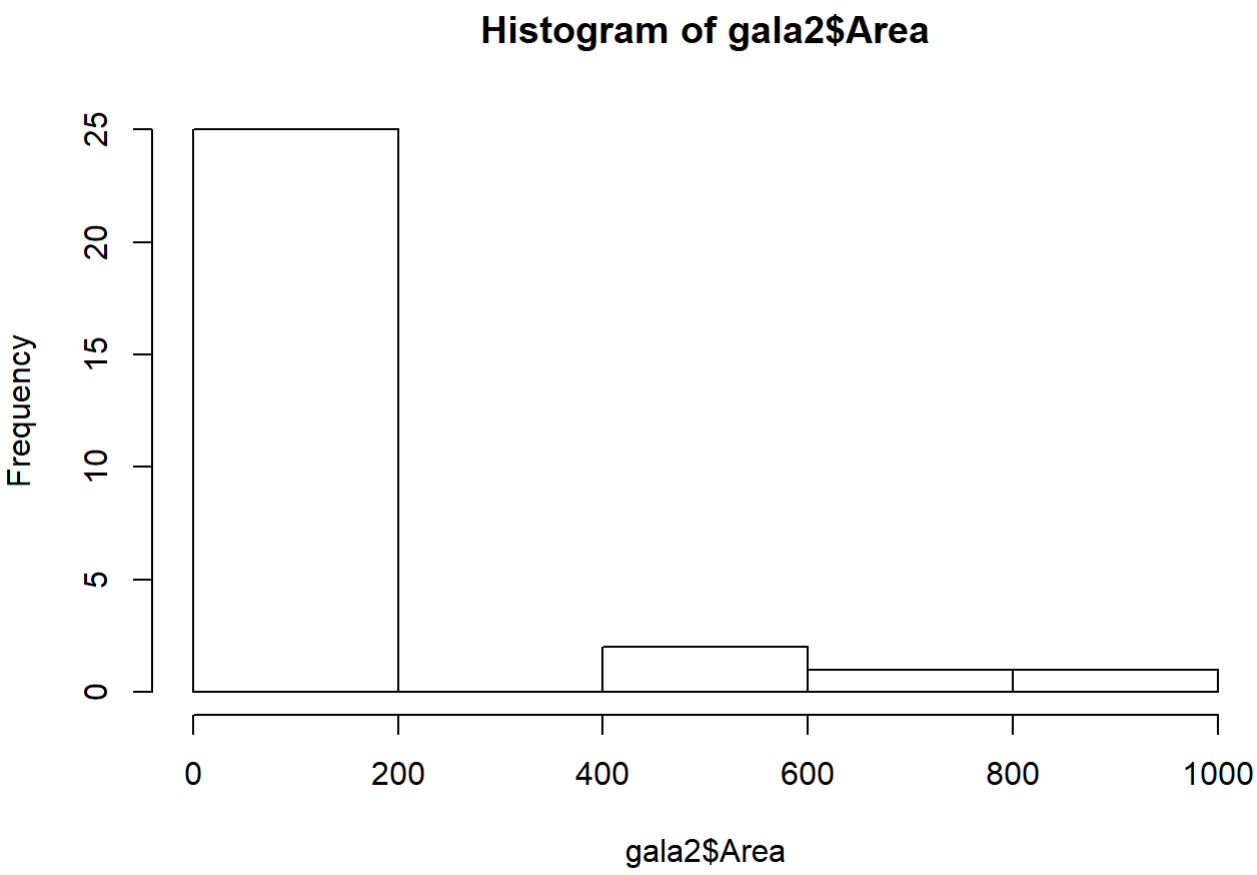
La quitamos?

```
gala2=gala[-16,]
pairs(gala2)
```



Ayuda, pero la variable Area sigue teniendo valores inusualmente grandes Sera conveniente transformar?

```
hist(gala2$Area)
```




```
hist(log(gala2$Area))
```

Se ve mejor! La funcion powerTransform de la libreria car nos puede sugerir una transformacion exponencial para el resultado

```
library(car)
```

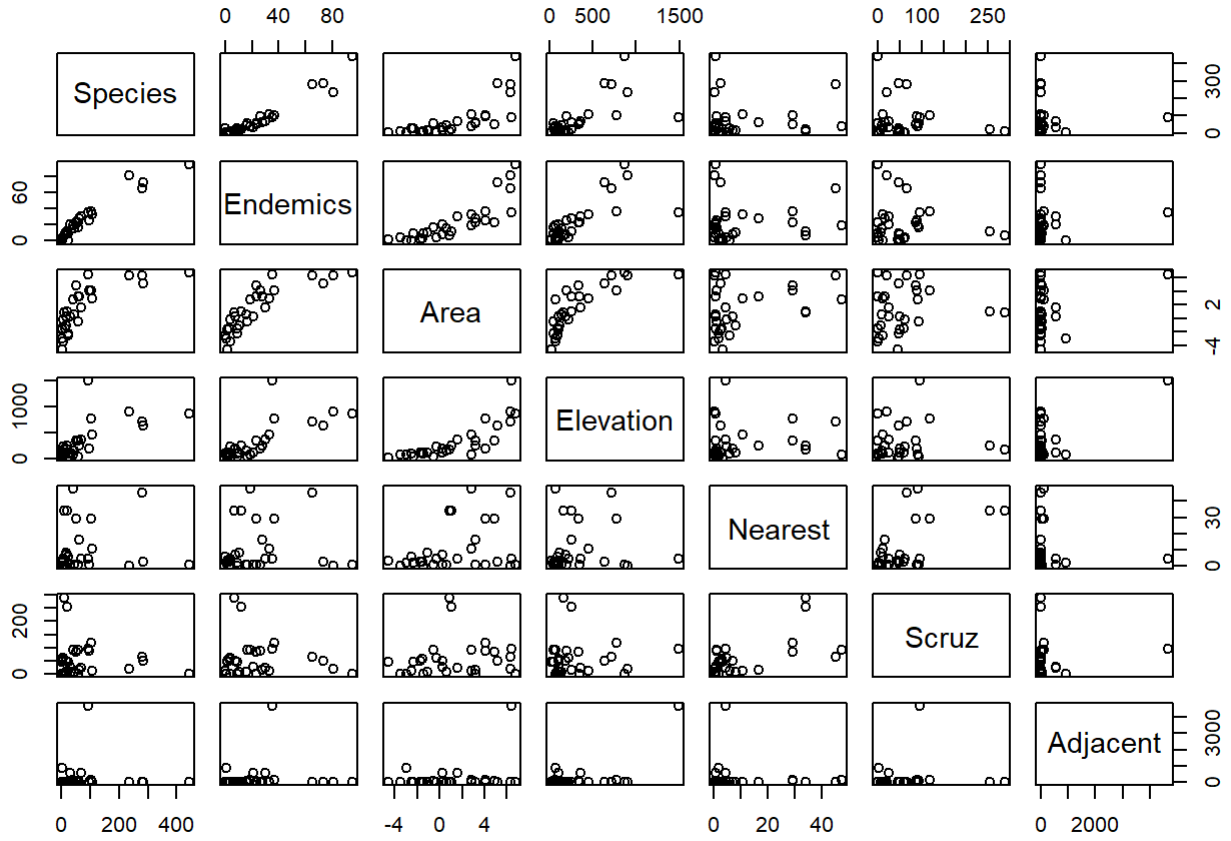
```
## Loading required package: carData
```

```
summary(powerTransform(gala2$Area~1))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.0202           0      -0.1407       0.1004
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df    pval
## LR test, lambda = (0) 0.1071203  1 0.74345
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 171.2435  1 < 2.22e-16
```

Nos quedamos con la transformacion.

```
gala3=gala2
gala3$Area=log(gala3$Area)
pairs(gala3)
```



En Ajacent y en EleVation hay valores sospechosos que no se parecen al resto de los datos Cual es el Adjacent inusual?

```
which(gala3$Adjacent>2000)
```

```
## [1] 12
```

```
gala3[12,]
```

```
##            Species Endemics     Area Elevation Nearest Scruz Adjacent
## Fernandina      93       35 6.452822      1494     4.3  95.3  4669.32
```

```
cooks.distance(lm(Species~.,dat=gala3))
```

```
##       Baltra    Bartolome     Caldwell     Champion      Coamano
## 1.547819e-04 7.539983e-02 1.496224e-03 1.353185e-04 5.817249e-03
## Daphne.Major Daphne.Minor       Darwin         Eden      Enderby
## 2.217566e-03 4.403956e-02 2.779313e-03 1.773447e-04 2.783621e-03
##     Espanola   Fernandina     Gardner1     Gardner2     Genovesa
## 1.368854e-01 5.030427e+00 3.281663e-04 7.870793e-03 5.236820e-02
##     Marchena       Onslow        Pinta       Pinzon   Las.Plazas
## 2.664783e-04 3.560567e-03 6.121615e-02 5.824198e-06 1.306707e-03
##       Rabida SanCristobal  SanSalvador    SantaCruz      SantaFe
## 1.224619e-02 2.461352e-01 6.023689e-01 8.085411e-01 1.627952e-02
##   SantaMaria      Seymour      Tortuga         Wolf
## 1.184794e-05 1.637333e-04 1.480477e-03 8.263478e-03
```

```
cooks.distance(lm(Species~.,dat=gala3))[12]
```

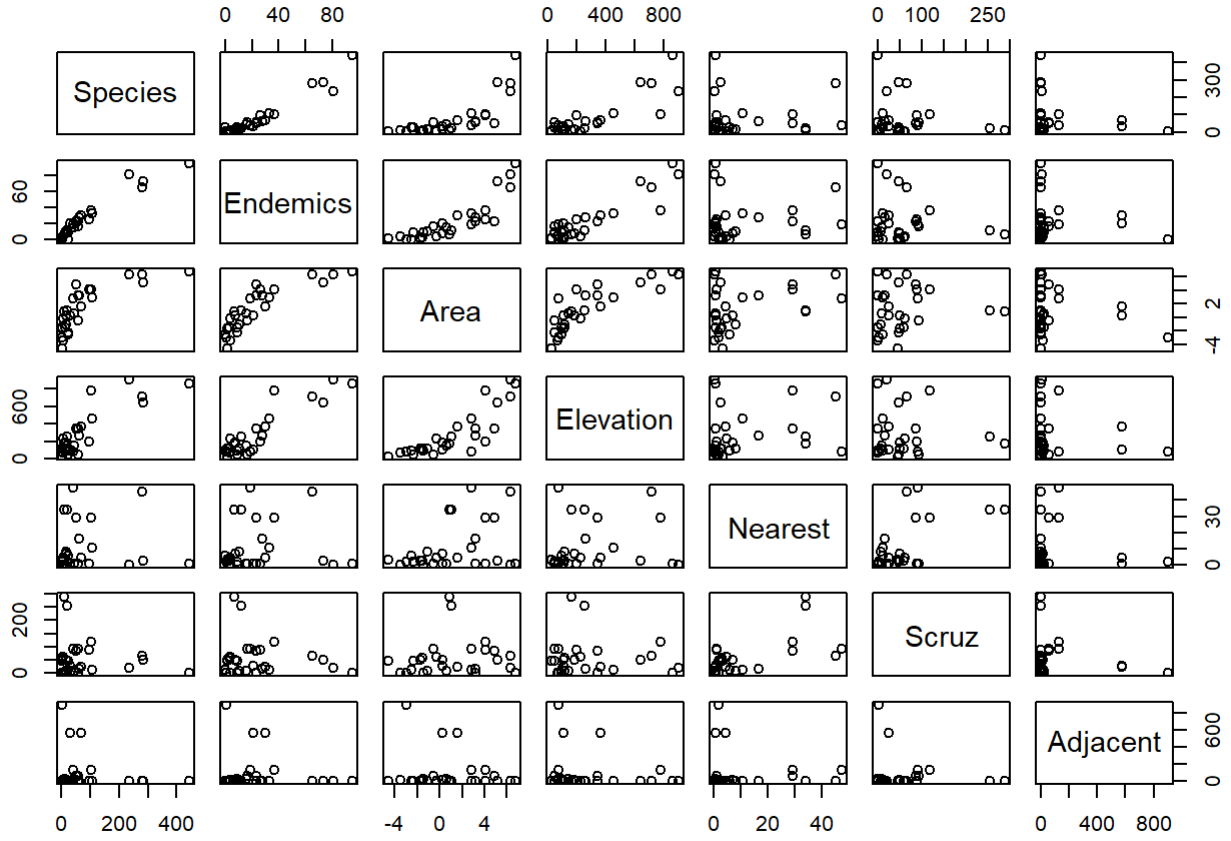
```
## Fernandina
##   5.030427
```

```
qf(0.5,ncol(gala3)+1,nrow(gala3)-ncol(gala3)-1)
```

```
## [1] 0.94805
```

La misma observacion tiene alta Adjacent y Elevation. Hay que quitar

```
gala4=gala3[-12,]
pairs(gala4)
```



Mucho mejor, de nuevo!

Species relacionado con Endemics, Area y Elevation. La relacion entre Nearest, Scruz y Adjacent no parece ser obvia, si es que existe Edemics, Area y Elevation tambien se relacionan, nada sorprendente

transformacion de MV Yeo-Johnson? sugerira transformar Adjacent?

```
summary(powerTransform(cbind(Endemics,Area,Elevation,Nearest, Scruz,Adjacent)~1,gala4,family="yj
Power"))
```

```
## yjPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Endemics     0.3018         0.5       0.0748       0.5289
## Area         0.9932         1.0       0.7997       1.1867
## Elevation    0.0887         0.0      -0.2509       0.4283
## Nearest     -0.2538         0.0      -0.6300       0.1225
## Scruz        0.2202         0.0      -0.0023       0.4427
## Adjacent    -0.2850        -0.5      -0.5174      -0.0525
##
##  Likelihood ratio test that all transformation parameters are equal to 0
##                                    LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 90.2666  6 < 2.22e-16
```

Hay evidencia de quer algunas transformaciones podrian ser utiles, aunque las podemos ignorar por ahora (por cuestión de tiempo)

El investigador esta interesado en dos cosas: 1. Una forma buena en promedio de predecir el numero de especies 2. Un modelo relacionando Species con las caracteristicas geograficas de la isla (es decir, no le interesa tanto incorporar “Endemics” al modelo)

Por que suena logico el punto 2 ?

Respuesta: Es más fácil buscar una imagen satelital sobre características geográficas de una isla y tratar de predecir a partir de ello el número de especies de tortugas… Que enviar a un ejército de becarios a que cuenten, por ejemplo, el número de especies endémicas… ¡Cuando bien pudieron haber contado el número de especies de tortugas!

```
fit1=lm(Species~Endemics+Area+Elevation+Nearest+Scruz+Adjacent,data=gala4)
summary(fit1)
```

```
##
## Call:
## lm(formula = Species ~ Endemics + Area + Elevation + Nearest +
##     Scruz + Adjacent, data = gala4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.112 -13.217  -1.016   9.355  64.014
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.605948  13.063527  -2.113   0.0467 *
## Endemics      5.098972   0.625684   8.149 6.11e-08 ***
## Area         -7.146263   4.014268  -1.780   0.0895 .
## Elevation    -0.032813   0.052240  -0.628   0.5367
## Nearest       0.327837   0.546231   0.600   0.5548
## Scruz        -0.005821   0.104674  -0.056   0.9562
## Adjacent     -0.029412   0.025721  -1.144   0.2657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.26 on 21 degrees of freedom
## Multiple R-squared:  0.9459, Adjusted R-squared:  0.9304
## F-statistic: 61.15 on 6 and 21 DF,  p-value: 3.266e-12
```

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: Species
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Endemics   1 288892  288892 361.6168 1.026e-14 ***
## Area       1   2544    2544   3.1842   0.08881 .
## Elevation  1    229     229   0.2868   0.59793
## Nearest    1    406     406   0.5076   0.48402
## Scruz      1     13      13   0.0157   0.90134
## Adjacent   1   1045    1045   1.3076   0.26570
## Residuals 21  16777     799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Endemics esta relacionado fuertemente y Area tambien hasta cierto punto

Endemics, Area y Elevation se relacionan, por lo que los valores p pueden ser enganiosos Este modelo parece funcionar

Y si ponemos primero elevacion?

```
fit1b=lm(Species~Area+Elevation+Endemics+Nearest+Scruz+Adjacent,data=gala4)
anova(fit1b)
```

```
## Analysis of Variance Table
##
## Response: Species
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Area       1 184695  184695 231.1889 8.280e-13 ***
## Elevation  1  45023   45023  56.3567 2.248e-07 ***
## Endemics   1  61948   61948  77.5422 1.710e-08 ***
## Nearest    1    406     406   0.5076    0.4840
## Scruz      1     13      13   0.0157    0.9013
## Adjacent   1   1045    1045   1.3076    0.2657
## Residuals 21  16777     799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora parecen cambiar las variables significativas, y hemos obtenido otras distintas a Endemics, que nos puede ayudar a responder la segunda pregunta

```
fit1c=lm(Species~Area+Elevation+Nearest+Adjacent+Scruz+Endemics,data=gala4)
anova(fit1c)
```

```
## Analysis of Variance Table
##
## Response: Species
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Area       1 184695  184695 231.1889 8.280e-13 ***
## Elevation  1  45023   45023  56.3567 2.248e-07 ***
## Nearest    1   8081    8081  10.1154  0.004503 **
## Adjacent   1    544     544   0.6808  0.418591
## Scruz      1   1728    1728   2.1635  0.156147
## Endemics   1  53057   53057  66.4134 6.113e-08 ***
## Residuals 21  16777     799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hay que ser cuidadosos al interpretar una tabla ANOVA. En este caso particular, los predictores significativos cambian según el orden en el que