William Andrés Gómez Roa

Pontifícia Universidad Javeriana

Regresión Lineal

Profesor Mario Saavedra

REGRESIÓN LINEAL EN R

DATOS: FORMAS Y TAMAÑOS DE 3 VARIEDADES DE TRIGO MEDIDAS CON RAYOS-X

Los DATOS aquí examinados son clases de granos de Trígo pertenecientes a tres variedades diferentes: Kama, Rosa y Canadian, con 70 elementos cada una, seleccionadas al azar para el experimento. Se utilizaron "rayos X suaves" para tener visualización de alta calidad de la estructura interna del núcleo.Las imágenes se registraron en placas KODAK de rayos X de 13x18 cm. Los estudios se llevaron a cabo utilizando granos de trigo cosechados con cosechadora provenientes de campos experimentales, explorados en el Instituto de Agrofísica de la Academia de Ciencias de Polonia en Lublin.

Estos datos fueron donados el 29 de septiembre de 2012 al repositorio de Machine Learning "UCI"

Instalamos Librerias

```r
install.packages('GGally')
install.packages('dplyr')
install.packages('statsr')
install.packages('ggfortify')
install.packages('tidyverse')
install.packages('olsrr')
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

Cargamos los datos

```
df<-read.csv("seeds_dataset.csv", header = FALSE, sep = ";", dec =
".")
colnames(df) <- c("Area", "Perimeter", "Compactness",
"Length_Kernel","Width_Kernel","Asymmetry_Coeff","Length_Groove","Vari
ety_Wheat")
df <- subset(df, select =
c(Area,Perimeter,Compactness,Length_Kernel,Width_Kernel,Asymmetry_Coef
f,Length_Groove,Variety_Wheat))
df
```

```
    Area  Perimeter Compactness Length_Kernel Width_Kernel
Asymmetry_Coeff
1   15.26 14.84     0.8710      5.763         3.312         2.2210

2   14.88 14.57     0.8811      5.554         3.333         1.0180

3   14.29 14.09     0.9050      5.291         3.337         2.6990

4   13.84 13.94     0.8955      5.324         3.379         2.2590

5   16.14 14.99     0.9034      5.658         3.562         1.3550

6   14.38 14.21     0.8951      5.386         3.312         2.4620

7   14.69 14.49     0.8799      5.563         3.259         3.5860

8   14.11 14.10     0.8911      5.420         3.302         2.7000

9   16.63 15.46     0.8747      6.053         3.465         2.0400

10  16.44 15.25     0.8880      5.884         3.505         1.9690

11  15.26 14.85     0.8696      5.714         3.242         4.5430

12  14.03 14.16     0.8796      5.438         3.201         1.7170

13  13.89 14.02     0.8880      5.439         3.199         3.9860

14  13.78 14.06     0.8759      5.479         3.156         3.1360

15  13.74 14.05     0.8744      5.482         3.114         2.9320
```

16	14.59	14.28	0.8993	5.351	3.333	4.1850
17	13.99	13.83	0.9183	5.119	3.383	5.2340
18	15.69	14.75	0.9058	5.527	3.514	1.5990
19	14.70	14.21	0.9153	5.205	3.466	1.7670
20	12.72	13.57	0.8686	5.226	3.049	4.1020
21	14.16	14.40	0.8584	5.658	3.129	3.0720
22	14.11	14.26	0.8722	5.520	3.168	2.6880
23	15.88	14.90	0.8988	5.618	3.507	0.7651
24	12.08	13.23	0.8664	5.099	2.936	1.4150
25	15.01	14.76	0.8657	5.789	3.245	1.7910
26	16.19	15.16	0.8849	5.833	3.421	0.9030
27	13.02	13.76	0.8641	5.395	3.026	3.3730
28	12.74	13.67	0.8564	5.395	2.956	2.5040
29	14.11	14.18	0.8820	5.541	3.221	2.7540
30	13.45	14.02	0.8604	5.516	3.065	3.5310
⋮	⋮	⋮	⋮	⋮	⋮	⋮
181	11.41	12.95	0.8560	5.090	2.775	4.957
182	12.46	13.41	0.8706	5.236	3.017	4.987
183	12.19	13.36	0.8579	5.240	2.909	4.857
184	11.65	13.07	0.8575	5.108	2.850	5.209
185	12.89	13.77	0.8541	5.495	3.026	6.185
186	11.56	13.31	0.8198	5.363	2.683	4.062
187	11.81	13.45	0.8198	5.413	2.716	4.898
188	10.91	12.80	0.8372	5.088	2.675	4.179

189	11.23	12.82	0.8594	5.089	2.821	7.524
190	10.59	12.41	0.8648	4.899	2.787	4.975
191	10.93	12.80	0.8390	5.046	2.717	5.398
192	11.27	12.86	0.8563	5.091	2.804	3.985
193	11.87	13.02	0.8795	5.132	2.953	3.597
194	10.82	12.83	0.8256	5.180	2.630	4.853
195	12.11	13.27	0.8639	5.236	2.975	4.132
196	12.80	13.47	0.8860	5.160	3.126	4.873
197	12.79	13.53	0.8786	5.224	3.054	5.483
198	13.37	13.78	0.8849	5.320	3.128	4.670
199	12.62	13.67	0.8481	5.410	2.911	3.306
200	12.76	13.38	0.8964	5.073	3.155	2.828
201	12.38	13.44	0.8609	5.219	2.989	5.472
202	12.67	13.32	0.8977	4.984	3.135	2.300
203	11.18	12.72	0.8680	5.009	2.810	4.051
204	12.70	13.41	0.8874	5.183	3.091	8.456
205	12.37	13.47	0.8567	5.204	2.960	3.919
206	12.19	13.20	0.8783	5.137	2.981	3.631
207	11.23	12.88	0.8511	5.140	2.795	4.325
208	13.20	13.66	0.8883	5.236	3.232	8.315
209	11.84	13.21	0.8521	5.175	2.836	3.598
210	12.30	13.34	0.8684	5.243	2.974	5.637

	Length_Groove	Variety_Wheat
1	5.220	1
2	4.956	1

3	4.825	1
4	4.805	1
5	5.175	1
6	4.956	1
7	5.219	1
8	5.000	1
9	5.877	1
10	5.533	1
11	5.314	1
12	5.001	1
13	4.738	1
14	4.872	1
15	4.825	1
16	4.781	1
17	4.781	1
18	5.046	1
19	4.649	1
20	4.914	1
21	5.176	1
22	5.219	1
23	5.091	1
24	4.961	1
25	5.001	1
26	5.307	1
27	4.825	1
28	4.869	1
29	5.038	1
30	5.097	1
⋮	⋮	⋮
181	4.825	3
182	5.147	3
183	5.158	3
184	5.135	3
185	5.316	3
186	5.182	3
187	5.352	3
188	4.956	3
189	4.957	3
190	4.794	3
191	5.045	3
192	5.001	3
193	5.132	3
194	5.089	3
195	5.012	3
196	4.914	3
197	4.958	3
198	5.091	3
199	5.231	3
200	4.830	3
201	5.045	3

```
202 4.745         3
203 4.828         3
204 5.000         3
205 5.001         3
206 4.870         3
207 5.003         3
208 5.056         3
209 5.044         3
210 5.063         3
```

Análisis Exploratorio de los Datos

```
head(df)
```

```
  Area  Perimeter Compactness Length_Kernel Width_Kernel
Asymmetry_Coeff
1 15.26 14.84     0.8710      5.763         3.312         2.221

2 14.88 14.57     0.8811      5.554         3.333         1.018

3 14.29 14.09     0.9050      5.291         3.337         2.699

4 13.84 13.94     0.8955      5.324         3.379         2.259

5 16.14 14.99     0.9034      5.658         3.562         1.355

6 14.38 14.21     0.8951      5.386         3.312         2.462

  Length_Groove Variety_Wheat
1 5.220         1
2 4.956         1
3 4.825         1
4 4.805         1
5 5.175         1
6 4.956         1
```

```
summary(df)
```

```
      Area         Perimeter      Compactness      Length_Kernel
 Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
 1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262
 Median :14.36   Median :14.32   Median :0.8734   Median :5.524
 Mean   :14.85   Mean   :14.56   Mean   :0.8710   Mean   :5.629
 3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980
 Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675
  Width_Kernel    Asymmetry_Coeff  Length_Groove    Variety_Wheat
 Min.   :2.630   Min.   :0.7651   Min.   :4.519   Min.   :1
 1st Qu.:2.944   1st Qu.:2.5615   1st Qu.:5.045   1st Qu.:1
 Median :3.237   Median :3.5990   Median :5.223   Median :2
 Mean   :3.259   Mean   :3.7002   Mean   :5.408   Mean   :2
```

```
 3rd Qu.:3.562   3rd Qu.:4.7687   3rd Qu.:5.877   3rd Qu.:3  
 Max.   :4.033   Max.   :8.4560   Max.   :6.550   Max.   :3  
```

```
dim(df)
```

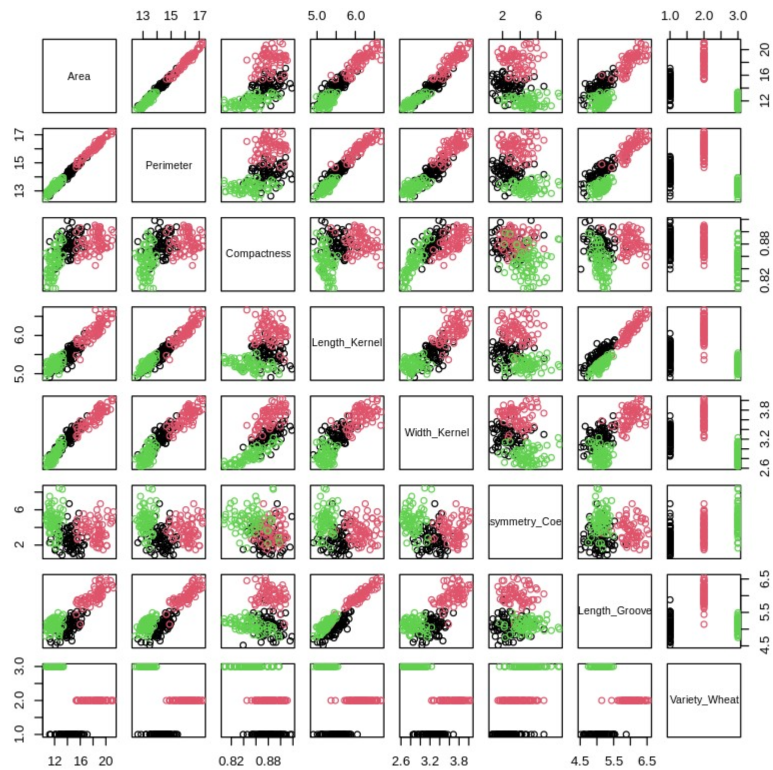
```
[1] 210   8
```

```
names(df)
```

```
[1] "Area"            "Perimeter"       "Compactness"    
"Length_Kernel"  
[5] "Width_Kernel"    "Asymmetry_Coeff" "Length_Groove"  
"Variety_Wheat"
```

VISUALIZACION

```
plot(df, col=df$Variety_Wheat)
```

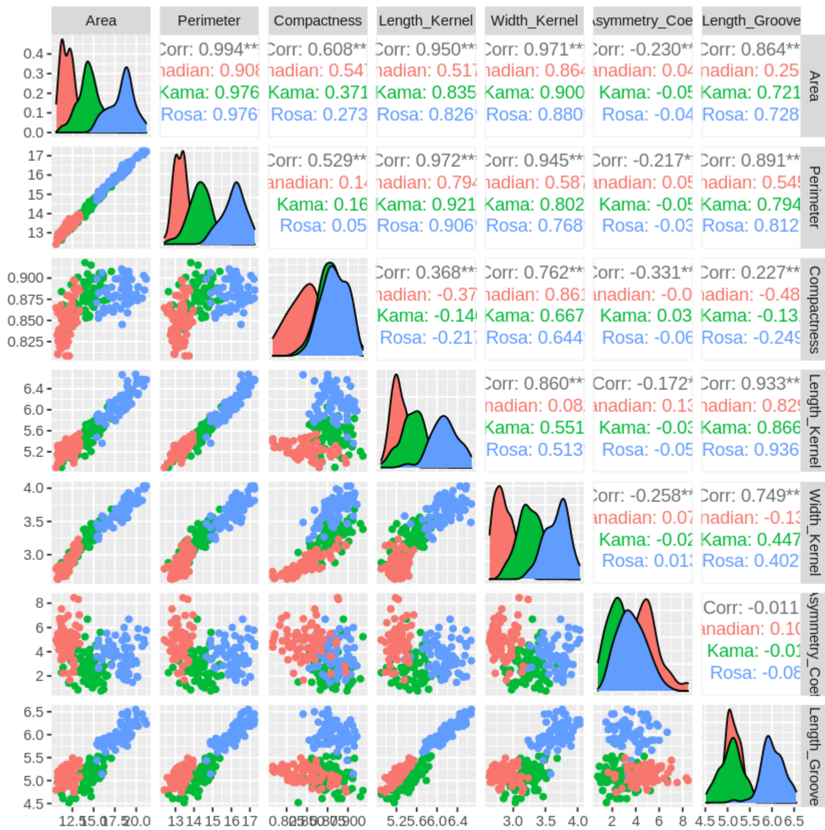



```
library('GGally')
ggcorr(df, method=c("everything", "pearson"))
```


```
library('dplyr')
df<- df %>% mutate(Nombre =
                     case_when(Variety_Wheat ==1 ~ "Kama",
                               Variety_Wheat ==2~ "Rosa",
                               Variety_Wheat ==3~ "Canadian")
)
```

```
Attaching package: ‘dplyr’


The following object is masked from ‘package:MASS’:

    select


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

```
ggpairs(df, columns=1:7, ggplot2::aes(colour=Nombre))
```



Del anterior Análisis Exploratorio podemos decir 2 cosas:

Por un lado, el dataset contiene variables que pueden ser en la practica más sencillas que medir que otras, por este motivo es de gran interes poder predecir el valor de esas variables que son más extrañas, más dificiles de medir o más costosoas de conseguir. Por lo tanto optaremos por escoger un modelo de Regresión LIneal que pueda explicar alguna de estas variables (variable Y). Estas variables son: 'Assymetry Coefficient', 'Compactness' y 'Length of Groove Kernel'.

Por otro lado, Visualizando las graficas de Dispersion y distribucion de los datos, junto con la matriz de correlacion podemos ver A SIMPLE VISTA vemos varios modelos lineales interesnates:

- Width Kernel vs Compactnnes
- Length Kernel vs Compactnnes

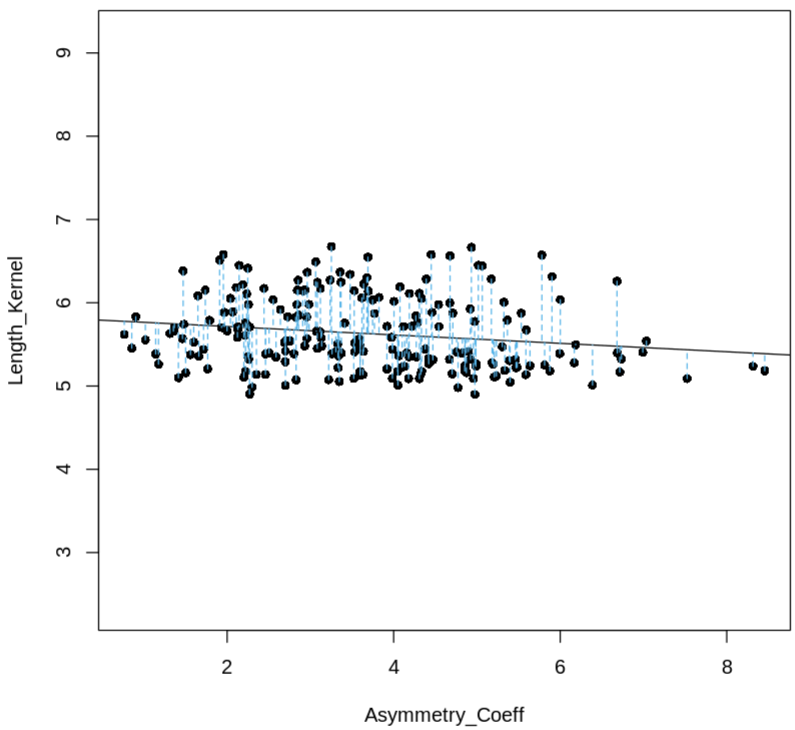
Sin emabrgo es de gran interes poder predecir el "Coeficiente de SImetria". Por tanto, tomando el mayor valor de la matriz de correlación tenemos !?

- Length of Kernel vs Assymetry Coefficient

Regresión LIneal para predecir el coeficiente de simetría

```
library(statsr)
plot_ss(x =Asymmetry_Coeff, y = Length_Kernel, data = df)
```

```
Click two points to make a line.
Call:
lm(formula = y ~ x, data = pts)

Coefficients:
(Intercept)            x
    5.81560     -0.05056

Sum of Squares:  39.82
```



Modelo 1

```
lm1<- lm(Asymmetry_Coeff~Length_Kernel, data=df)
summary(lm1)
```

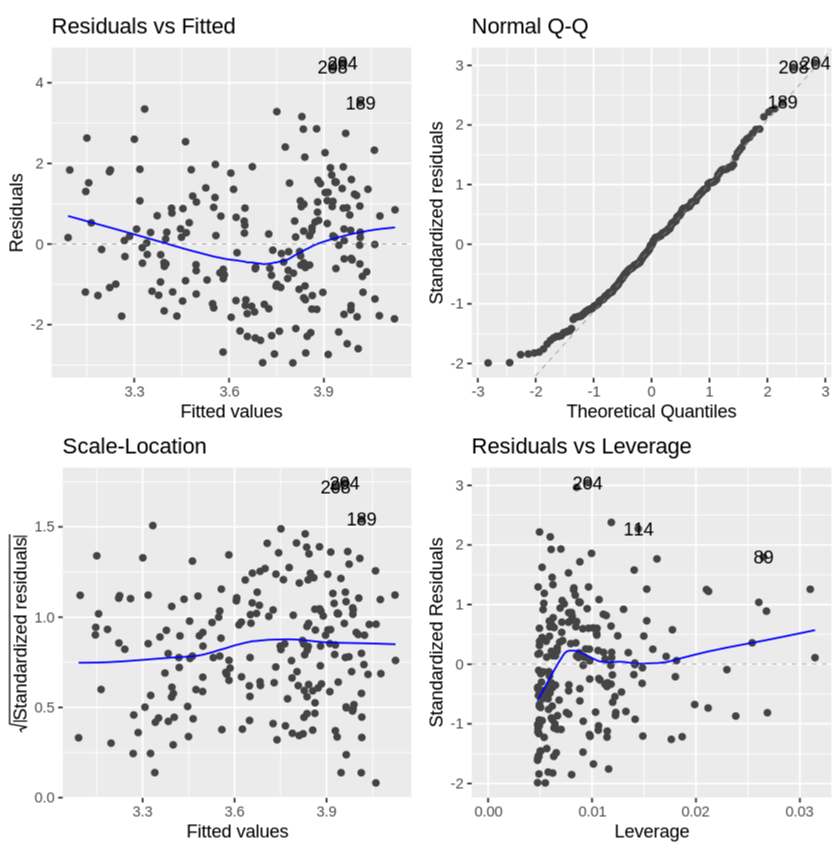
```
Call:
lm(formula = Asymmetry_Coeff ~ Length_Kernel, data = df)

Residuals:
```

```
   Min     1Q Median     3Q    Max 
-2.947 -1.157 -0.019  0.977  4.496 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     6.9772     1.3088   5.331 2.53e-07 ***
Length_Kernel  -0.5822     0.2318  -2.512   0.0128 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.485 on 208 degrees of freedom
Multiple R-squared:  0.02943,	Adjusted R-squared:  0.02477 
F-statistic: 6.308 on 1 and 208 DF,  p-value: 0.01278
```

```
confint(lm1)
```

```
                  2.5 %     97.5 %
(Intercept)    4.396997  9.5573285
Length_Kernel -1.039206 -0.1252041
```

```
library(ggfortify)
autoplot(lm1)
```




```
library(MASS)
AIC(lm1)
```

```
[1] 765.9681
```

Viendo Otros Modelos

```
lm2 <- lm(Asymmetry_Coeff~Area, data=df)

lm3 <- lm(Asymmetry_Coeff~Compactness, data=df)

lm4 <- lm(Asymmetry_Coeff~Compactness*Length_Kernel, data=df)

lm5 <- lm(Asymmetry_Coeff~Compactness*Length_Kernel*Area, data=df)

summary(lm2)
```

```
Call:
lm(formula = Asymmetry_Coeff ~ Area, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0453 -1.0670 -0.0326  0.9476  4.5010 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.46155    0.52757  10.352  < 2e-16 ***
Area        -0.11863    0.03487  -3.402 0.000803 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.467 on 208 degrees of freedom
Multiple R-squared:  0.0527,	Adjusted R-squared:  0.04815 
F-statistic: 11.57 on 1 and 208 DF,  p-value: 0.0008028
```

```
summary(lm3)
```

```
Call:
lm(formula = Asymmetry_Coeff ~ Compactness, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9463 -0.9196 -0.0655  0.8393  5.1017 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22.071      3.627   6.085 5.51e-09 ***
Compactness  -21.092      4.163  -5.067 8.90e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.422 on 208 degrees of freedom
Multiple R-squared:  0.1099,	Adjusted R-squared:  0.1056 
F-statistic: 25.67 on 1 and 208 DF,  p-value: 8.903e-07
```

```
summary(lm4)
```

```
Call:
lm(formula = Asymmetry_Coeff ~ Compactness * Length_Kernel, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1292 -0.9251 -0.1255  0.8454  5.1483 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)  
(Intercept)                 145.09      60.55   2.396   0.0175 *
Compactness                -160.31      69.17  -2.318   0.0214 *
Length_Kernel               -22.75      11.08  -2.054   0.0413 *
Compactness:Length_Kernel    25.74      12.64   2.036   0.0430 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.412 on 206 degrees of freedom
Multiple R-squared:  0.1302,	Adjusted R-squared:  0.1176 
F-statistic: 10.28 on 3 and 206 DF,  p-value: 2.454e-06
```

```
summary(lm5)
```

```
Call:
lm(formula = Asymmetry_Coeff ~ Compactness * Length_Kernel * 
    Area, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9037 -0.9179 -0.1585  0.8510  4.8715 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -458.273    423.963  -1.081    0.281
Compactness                     559.235    486.218   1.150    0.251
Length_Kernel                   100.619     79.143   1.271    0.205
Area                             20.826     26.133   0.797    0.426
Compactness:Length_Kernel      -120.791     90.769  -1.331    0.185
Compactness:Area                -25.530     29.848  -0.855    0.393
Length_Kernel:Area               -4.731      4.651  -1.017    0.310
Compactness:Length_Kernel:Area    5.723      5.318   1.076    0.283

Residual standard error: 1.388 on 202 degrees of freedom
```

```
Multiple R-squared:  0.1767,	Adjusted R-squared:  0.1481
F-statistic: 6.192 on 7 and 202 DF,  p-value: 1.41e-06
```

```
AIC(lm1)
```

```
[1] 765.9681
```

```
AIC(lm2)
```

```
[1] 760.8719
```

```
AIC(lm3)
```

```
[1] 747.7998
```

```
AIC(lm4)
```

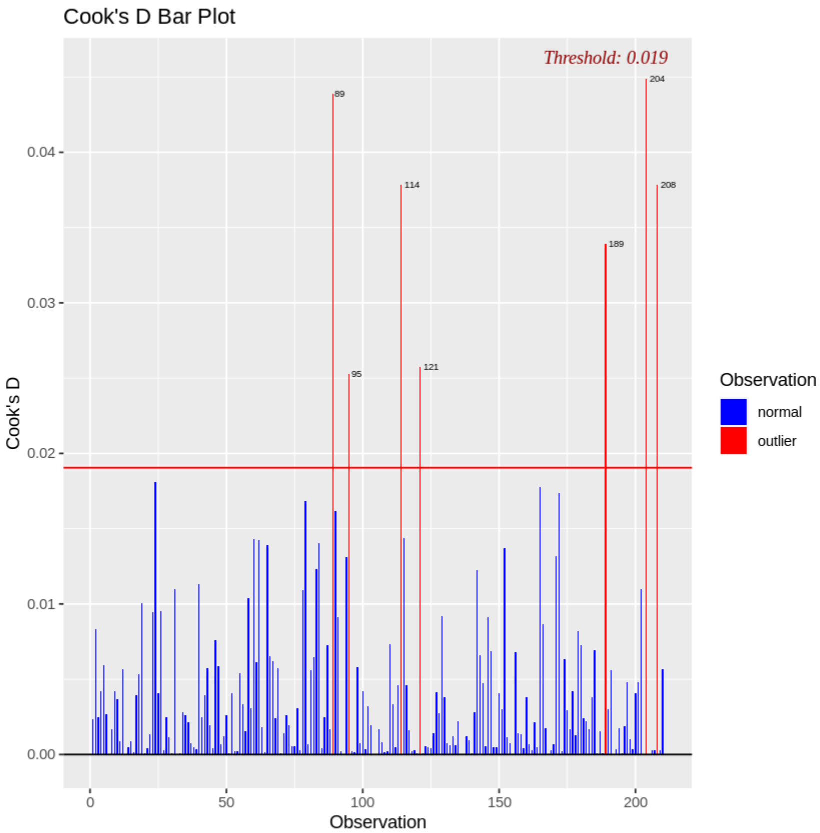
```
[1] 746.9416
```

```
AIC(lm5)
```

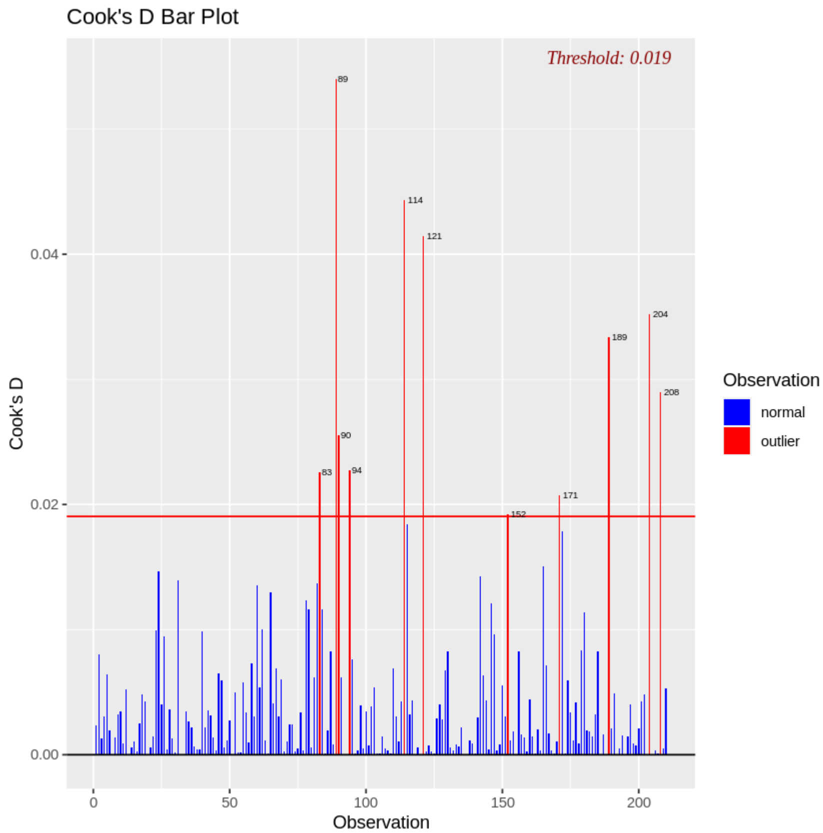
```
[1] 743.4211
```

Podemos ver que el AIC es coherente con el coeficiente de correlación de pearson y dicen que el mejor modelo sera el 1. Pero si vemos los resultados estadisticos de las pruebas de hipotesis es mucho mejor el modelo 3.

```
library(olsrr)
ols_plot_cooksd_bar(lm1)
```

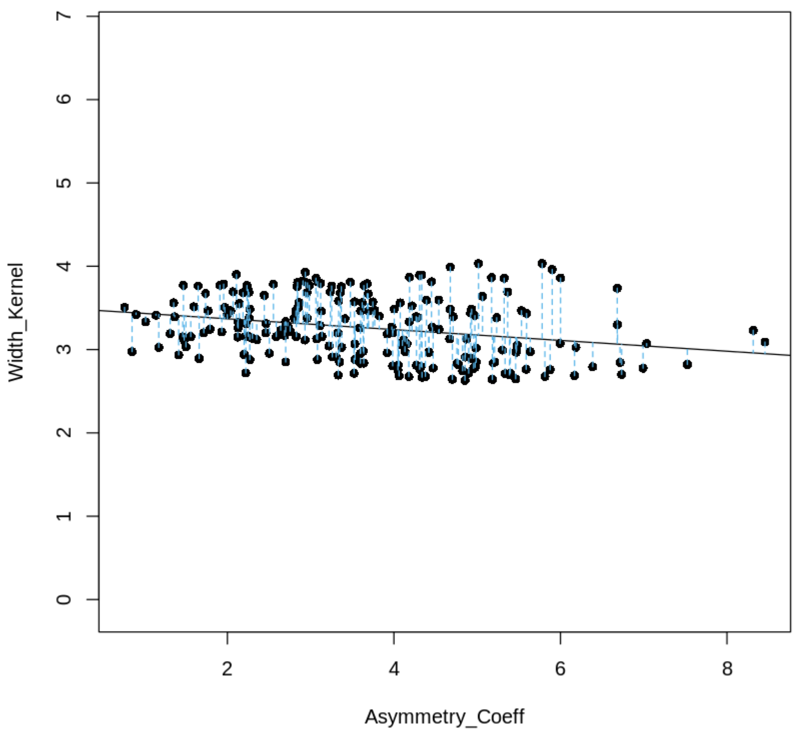



```
ols_plot_cooksd_bar(lm2)
```



VIendo los resultados de una regresion Lineal más "ideal" : Width_Kernel vs Compactness

```
plot_ss(x =Asymmetry_Coeff, y = Width_Kernel, data = df)

Click two points to make a line.
Call:
lm(formula = y ~ x, data = pts)

Coefficients:
(Intercept)            x
    3.49846     -0.06482

Sum of Squares:  27.832
```



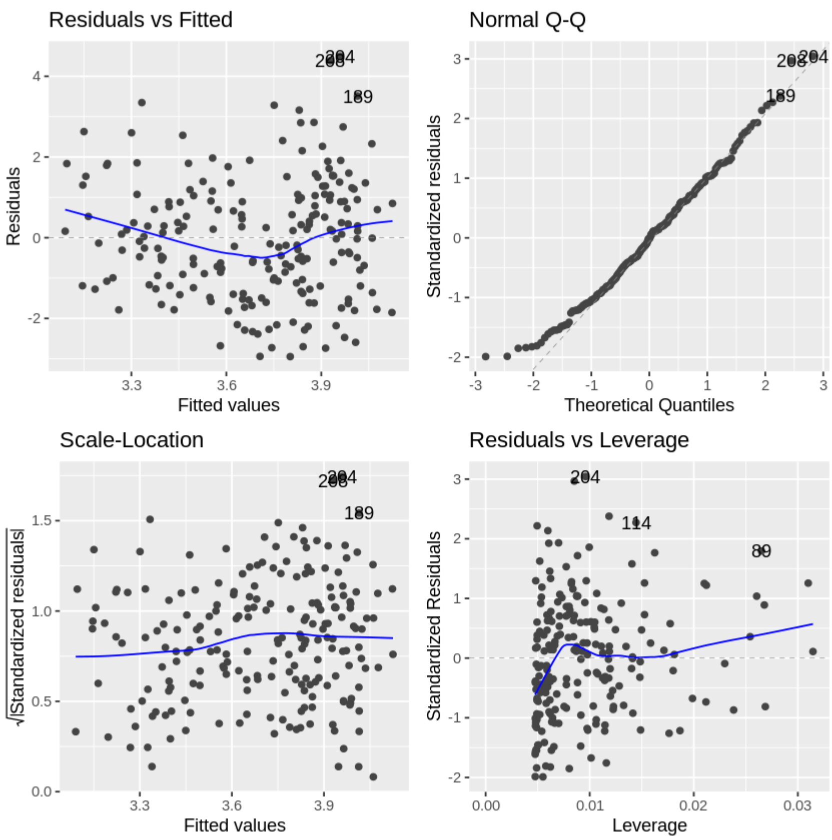

```
lmZ<- lm(Asymmetry_Coeff~Length_Kernel, data=df)
summary(lmZ)
```

```
Call:
lm(formula = Asymmetry_Coeff ~ Length_Kernel, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-2.947 -1.157 -0.019  0.977  4.496 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     6.9772     1.3088   5.331 2.53e-07 ***
Length_Kernel  -0.5822     0.2318  -2.512   0.0128 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.485 on 208 degrees of freedom
Multiple R-squared:  0.02943,	Adjusted R-squared:  0.02477 
F-statistic: 6.308 on 1 and 208 DF,  p-value: 0.01278
```

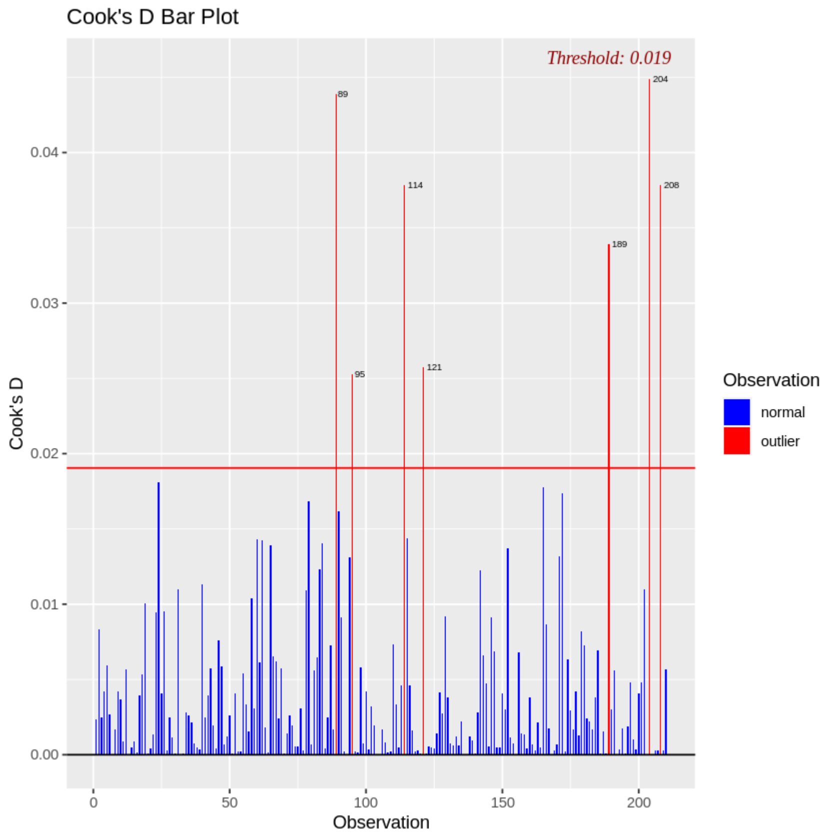
```
confint(lmZ)
```

```
                2.5 %      97.5 %
(Intercept)     4.396997  9.5573285
Length_Kernel -1.039206 -0.1252041
```

```
autoplot(lmZ)
```




```
ols_plot_cooksd_bar(lmZ)
```




```
AIC(lmZ)
```

```
[1] 765.9681
```

Conclusion

Podemos decir que se ajusto un modelo lineal a unos datos que no tenian estructura lineal evidente,pero que era de gran valor predecir la variable Y; aquí obtuvimos 1 modelo que tenía mejor cCoeficientes de Pearson y AIC, en contraste encontré un modelop que era mejor con las pruebas de hipotesis en general excepto en los residuales, en general se escogería el modelo con mejor AIC.

Ademas ajustamos un modelo a unos datos con grafico de dispersión lineal y coeficiente de person muy alto, le idea era tener una idea general del comportamiento de las pruebas de hipotesis, graficas, AIC, entre ambos casos descritos, pero no se encontro ningun patron evidente.

William Andrés Gómez Roa