A3-Regresión Múltiple-Detección datos atípicos

Eliezer Cavazos

2024-09-17

```
oCorte = read.csv("C:\\Users\\eliez\\OneDrive\\Desktop\\Clases\\AlCorte.csv")
#leer la base de datos
```

1. Haz un análisis descriptivo de los datos: medidas principales y gráficos

```
oModelo = lm(Resistencia~., data=oCorte)
oPasos = step(oModelo, direction="both", trace=1)
```

```
## Start:  AIC=102.96
## Resistencia ~ Fuerza + Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Fuerza       1     26.88  692.00 102.15
## - Tiempo       1     40.04  705.16 102.72
## <none>                      665.12 102.96
## - Temperatura  1    252.20  917.32 110.61
## - Potencia     1   1341.01 2006.13 134.08
##
## Step:  AIC=102.15
## Resistencia ~ Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Tiempo       1     40.04  732.04 101.84
## <none>                      692.00 102.15
## + Fuerza       1     26.88  665.12 102.96
## - Temperatura  1    252.20  944.20 109.47
## - Potencia     1   1341.02 2033.02 132.48
##
## Step:  AIC=101.84
## Resistencia ~ Potencia + Temperatura
##
##               Df Sum of Sq     RSS    AIC
## <none>                      732.04 101.84
## + Tiempo       1     40.04  692.00 102.15
## + Fuerza       1     26.88  705.16 102.72
## - Temperatura  1    252.20  984.24 108.72
## - Potencia     1   1341.01 2073.06 131.07
```

2. Encuentra el mejor modelo de regresión que explique la variable Resistencia. Analiza el modelo basándote en:

Significancia del modelo: Economía de las variables Significación global (Prueba para el modelo) Significación individual (Prueba para cada βi) Variación explicada por el modelo

```
oModeloNulo = lm(Resistencia~1, data=oCorte)
oPasos2 = step(oModeloNulo, scope=list(lower=oModeloNulo, upper = oModelo),
direction="forward")
```

```
## Start:  AIC=132.51
## Resistencia ~ 1
##
##               Df Sum of Sq     RSS    AIC
## + Potencia     1   1341.01  984.24 108.72
## + Temperatura  1    252.20 2073.06 131.07
## <none>                     2325.26 132.51
## + Tiempo       1     40.04 2285.22 133.99
## + Fuerza       1     26.88 2298.38 134.16
##
## Step:  AIC=108.72
## Resistencia ~ Potencia
##
##               Df Sum of Sq    RSS    AIC
## + Temperatura  1   252.202 732.04 101.84
## <none>                     984.24 108.72
## + Tiempo       1    40.042 944.20 109.47
## + Fuerza       1    26.882 957.36 109.89
##
## Step:  AIC=101.84
## Resistencia ~ Potencia + Temperatura
##
##          Df Sum of Sq    RSS    AIC
## <none>                 732.04 101.84
## + Tiempo  1    40.042 692.00 102.15
## + Fuerza  1    26.882 705.16 102.72
```

```
#oModelo2 = lm(Resistencia ~ Potencia + Temperatura, data=oModelo)
oModelo2 = summary(oPasos2)
```

```
oN = length(oCorte$Resistencia)
oPasos3=step(oModelo,direction="both",k=log(oN))
```

```
## Start:  AIC=109.97
## Resistencia ~ Fuerza + Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Fuerza       1     26.88  692.00 107.76
## - Tiempo       1     40.04  705.16 108.32
## <none>                      665.12 109.97
```

```
## - Temperatura  1    252.20  917.32 116.21
## - Potencia     1   1341.01 2006.13 139.69
##
## Step:  AIC=107.76
## Resistencia ~ Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Tiempo       1     40.04  732.04 106.04
## <none>                      692.00 107.76
## + Fuerza       1     26.88  665.12 109.97
## - Temperatura  1    252.20  944.20 113.68
## - Potencia     1   1341.02 2033.02 136.69
##
## Step:  AIC=106.04
## Resistencia ~ Potencia + Temperatura
##
##               Df Sum of Sq     RSS    AIC
## <none>                      732.04 106.04
## + Tiempo       1     40.04  692.00 107.76
## + Fuerza       1     26.88  705.16 108.32
## - Temperatura  1    252.20  984.24 111.52
## - Potencia     1   1341.01 2073.06 133.87
```

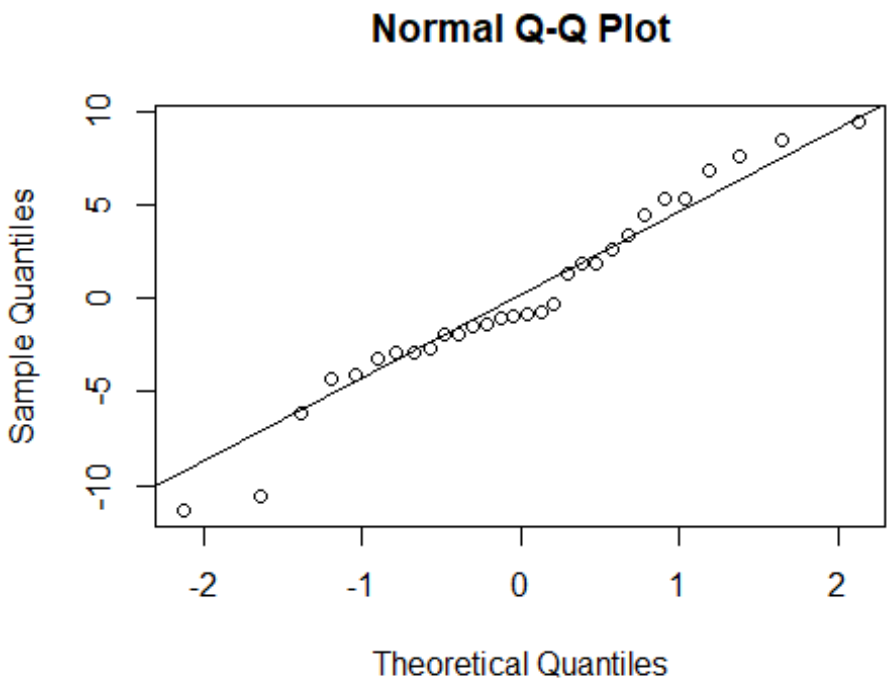
```
summary(oPasos3)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = oCorte)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3233  -2.8067  -0.8483   3.1892   9.4600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.90167   10.07207  -2.472  0.02001 *
## Potencia      0.49833    0.07086   7.033 1.47e-07 ***
## Temperatura   0.12967    0.04251   3.050  0.00508 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.207 on 27 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6619
## F-statistic: 29.38 on 2 and 27 DF,  p-value: 1.674e-07
```

3. Analiza la validez del modelo encontrado:

Análisis de residuos (homocedasticidad, independencia, etc)

```r
qqnorm(oPasos3$residuals)
qqline(oPasos3$residuals)
```




```r
t.test(oPasos3$residuals)
```

```
##
##  One Sample t-test
##
## data:  oPasos3$residuals
## t = 8.8667e-17, df = 29, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.876076  1.876076
## sample estimates:
##    mean of x
## 8.133323e-17
```

Homocedasticidad

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(oPasos3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  oPasos3
## BP = 4.0043, df = 2, p-value = 0.135
```

Linealidad

```
dwtest(oPasos3)
```

```
##
##  Durbin-Watson test
##
## data:  oPasos3
## DW = 2.3511, p-value = 0.8267
## alternative hypothesis: true autocorrelation is greater than 0
```

No multicolinealidad de Xi

```
library(car)
```

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

```
vif(oPasos3)
```

```
##    Potencia Temperatura
##           1           1
```

4. Analisis Datos Atipicos

Estandarizacion extrema residuos

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

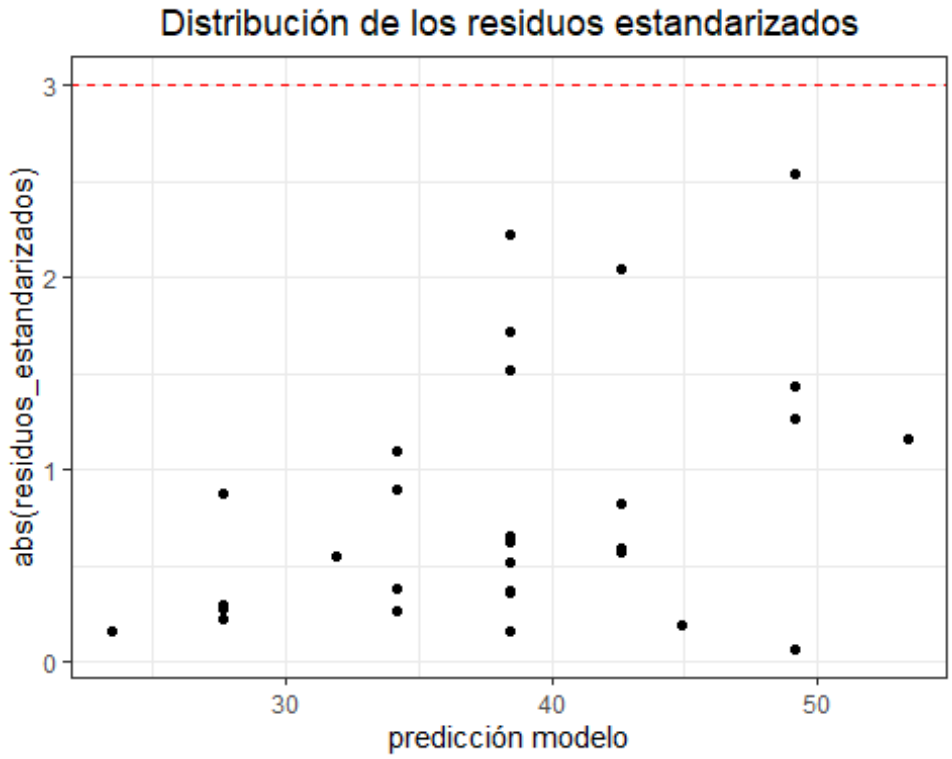
```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
oCorte$residuos_estandarizados <- rstudent(oPasos3)


ggplot(data = oCorte, aes(x = predict(oPasos3), y =
abs(residuos_estandarizados))) +
geom_hline(yintercept = 3, color = "red", linetype = "dashed") +
# se identifican en rojo observaciones con residuos estandarizados absolutos
> 3
geom_point(aes(color = ifelse(abs(residuos_estandarizados) > 3, 'red',
'black'))) +
scale_color_identity() +
labs(title = "Distribución de los residuos estandarizados", x = "predicción
modelo") +
theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



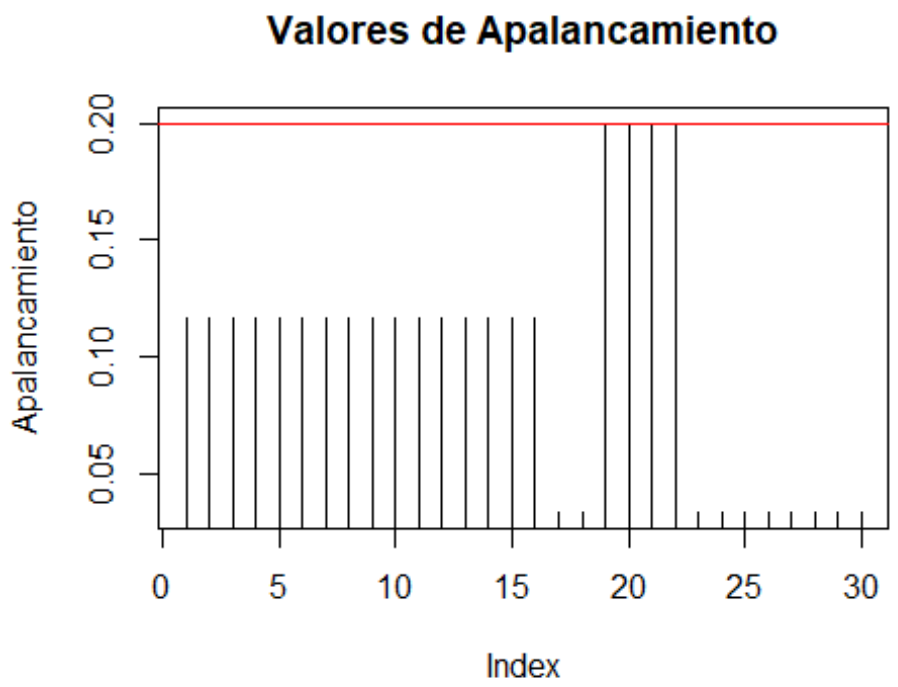

```
Atipicos = which(abs(oCorte$residuos_estandarizados)>2)
oCorte[Atipicos, ]

##    Fuerza Potencia Temperatura Tiempo Resistencia residuos_estandarizados
## 8      40       90         225     15        37.8               -2.535832
## 12     40       90         175     25        52.1                2.043589
## 29     35       75         200     20        27.8               -2.216952
```

Distancia de Leverange

```
leverage = hatvalues(oPasos3)

plot(leverage, type="h", main="Valores de Apalancamiento",
ylab="Apalancamiento")
abline(h = 2*mean(leverage), col="red") # Límite comúnmente usado
```



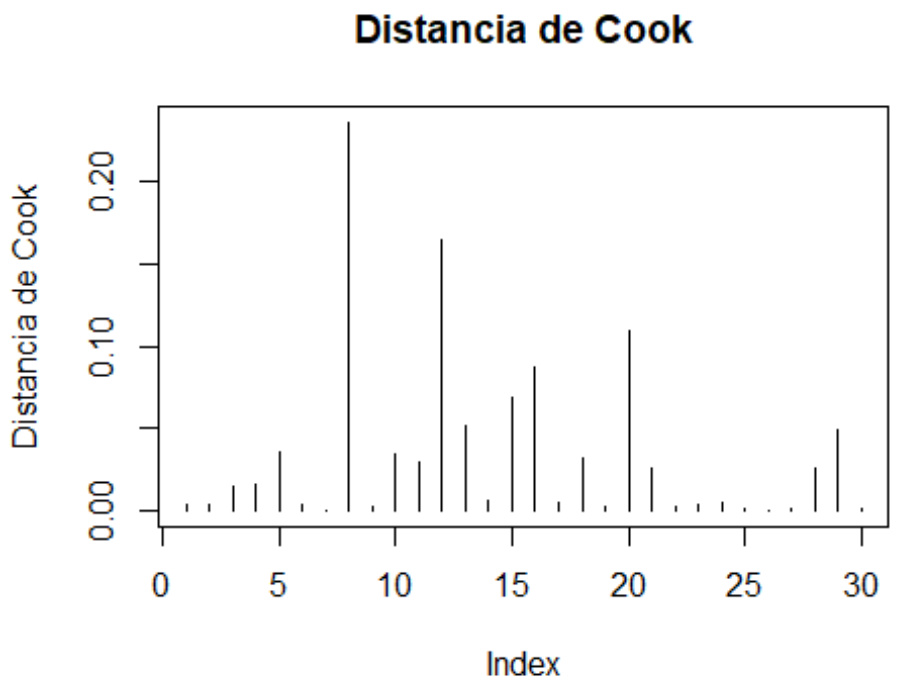

```
high_leverage_points = which(leverage > 2*mean(leverage))
oCorte[high_leverage_points, ]

##    Fuerza Potencia Temperatura Tiempo Resistencia residuos_estandarizados
## 19     35       45         200     20        22.7                -0.159511
## 20     35      105         200     20        58.7                 1.154355
```

Distancia de Cook

```
cooksd <- cooks.distance(oPasos3)

plot(cooksd, type="h", main="Distancia de Cook", ylab="Distancia de Cook")
abline(h = 1, col="red") # Límite comúnmente usado
```



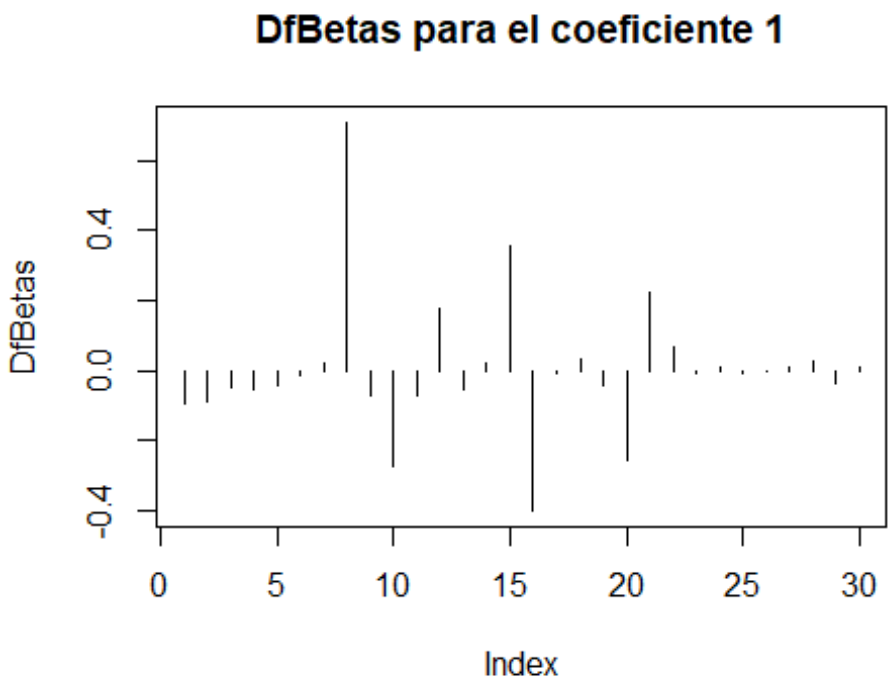

```
puntos_influyentes = which(cooksd > .15)
oCorte[puntos_influyentes, ]
```

```
##    Fuerza Potencia Temperatura Tiempo Resistencia residuos_estandarizados
## 8      40       90         225     15        37.8                -2.535832
## 12     40       90         175     25        52.1                 2.043589
```

DFBetas

Coeficiente 1

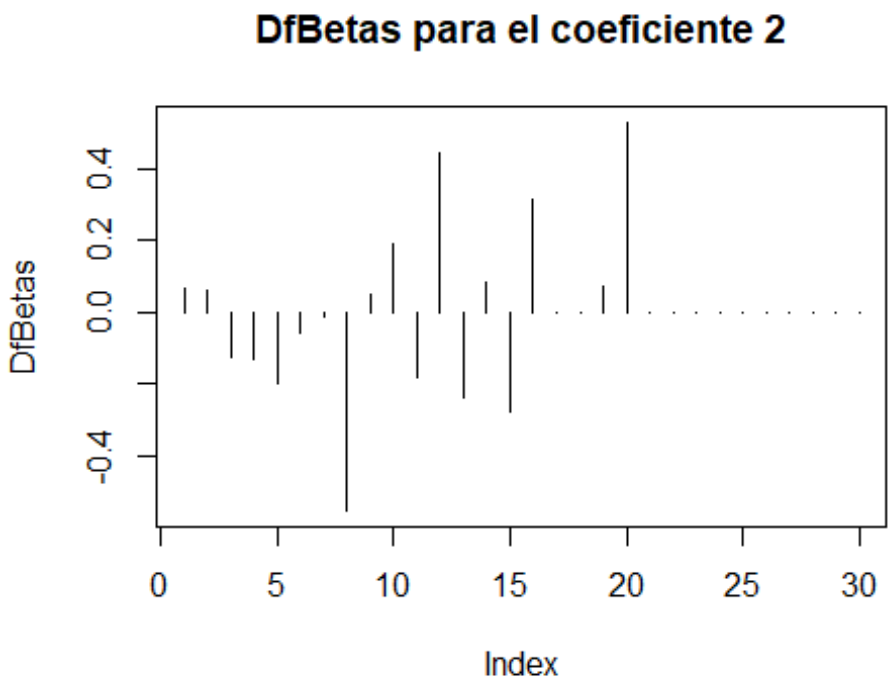
```
dfbetas_values = dfbetas(oPasos3)
plot(dfbetas_values[, 1], type="h", main="DfBetas para el coeficiente 1",
    ylab="DfBetas")
abline(h = c(-1, 1), col="red") # Límites comunes
```




```
puntos_influyentes = which(abs(dfbetas_values[, 1]) > 1)
```

Coeficiente 2

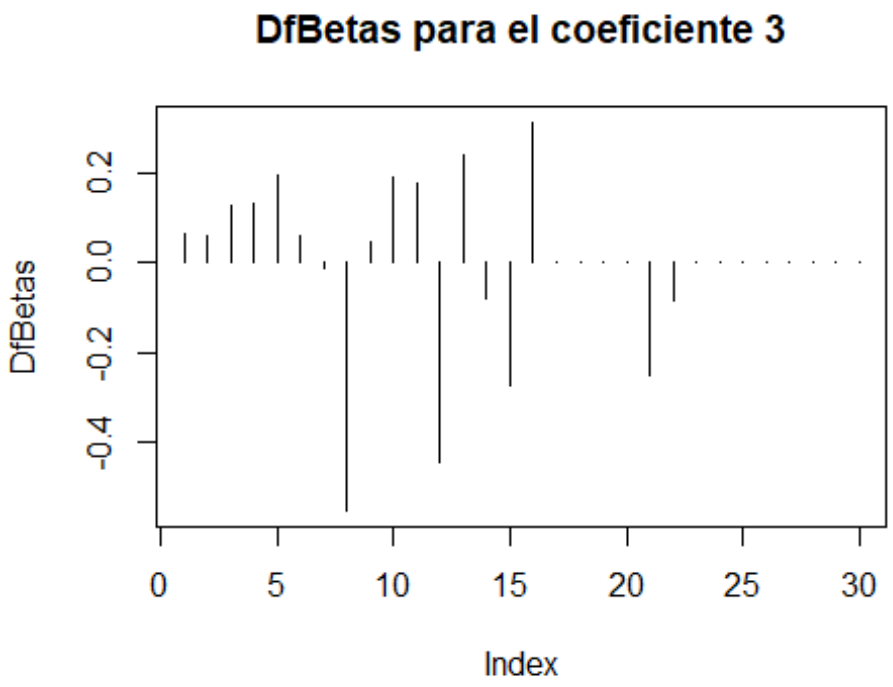
```
dfbetas_values = dfbetas(oPasos3)
plot(dfbetas_values[, 2], type="h", main="DfBetas para el coeficiente 2",
    ylab="DfBetas")
abline(h = c(-1, 1), col="red") # Límites comunes
```




```
puntos_influyentes = which(abs(dfbetas_values[, 2]) > 1)
```

Coeficiente 3

```
dfbetas_values = dfbetas(oPasos3)
plot(dfbetas_values[, 3], type="h", main="DfBetas para el coeficiente 3",
    ylab="DfBetas")
abline(h = c(-1, 1), col="red") # Límites comunes
```



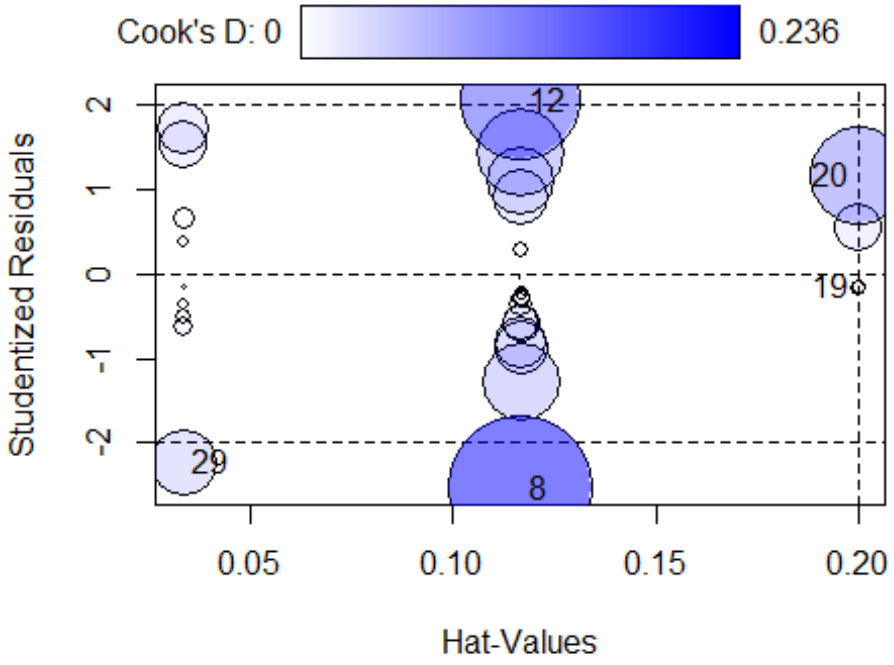

```
puntos_influyentes = which(abs(dfbetas_values[, 3]) > 1)
```

Datos Influyentes

```
influencia = influence.measures(oPasos3)
summary(influencia)

## Potentially influential observations of
##   lm(formula = Resistencia ~ Potencia + Temperatura, data = oCorte) :
##
##    dfb.1_ dfb.Ptnc dfb.Tmpr dffit cov.r   cook.d hat
## 8   0.71  -0.55    -0.55    -0.92  0.65_*  0.24   0.12
## 19 -0.04   0.07     0.00    -0.08  1.40_*  0.00   0.20
## 21  0.22   0.00    -0.25     0.27  1.35_*  0.03   0.20
## 22  0.07   0.00    -0.09    -0.09  1.39_*  0.00   0.20

library(car)
influencePlot(oPasos3)
```




```
##       StudRes        Hat       CookD
## 8  -2.535832 0.11666667 0.235696235
## 12  2.043589 0.11666667 0.164507739
## 19 -0.159511 0.20000000 0.002199712
## 20  1.154355 0.20000000 0.109693544
## 29 -2.216952 0.03333333 0.049338917
```

Conclusion

Se puede identificar que los datos atipicos que influyen serian la distancia de Cook y al tener un valor alto se identifica que tiene un efecto considerable en la varianza de los coeficientes del modelo