A3-Regresión Múltiple-Detección datos atípicos_fer

Fernanda Pérez

2024-09-17

```
datos <- read.csv("D:/Downloads/AlCorte.csv")

head(datos)
```

```
##   Fuerza Potencia Temperatura Tiempo Resistencia
## 1     30       60         175     15        26.2
## 2     40       60         175     15        26.3
## 3     30       90         175     15        39.8
## 4     40       90         175     15        39.7
## 5     30       60         225     15        38.6
## 6     40       60         225     15        35.5
```

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

```
summary(datos)
```

```
##      Fuerza      Potencia    Temperatura      Tiempo    Resistencia
##  Min.   :25   Min.   : 45   Min.   :150   Min.   :10   Min.   :22.70
##  1st Qu.:30   1st Qu.: 60   1st Qu.:175   1st Qu.:15   1st Qu.:34.67
##  Median :35   Median : 75   Median :200   Median :20   Median :38.60
##  Mean   :35   Mean   : 75   Mean   :200   Mean   :20   Mean   :38.41
##  3rd Qu.:40   3rd Qu.: 90   3rd Qu.:225   3rd Qu.:25   3rd Qu.:42.70
##  Max.   :45   Max.   :105   Max.   :250   Max.   :30   Max.   :58.70
```

```
desviacion_estandar <- apply(datos[, -1], 2, sd)
varianza <- apply(datos[, -1], 2, var)
desviacion_estandar
```

```
##    Potencia Temperatura      Tiempo Resistencia
##   13.645765   22.742941    4.548588    8.954403
```

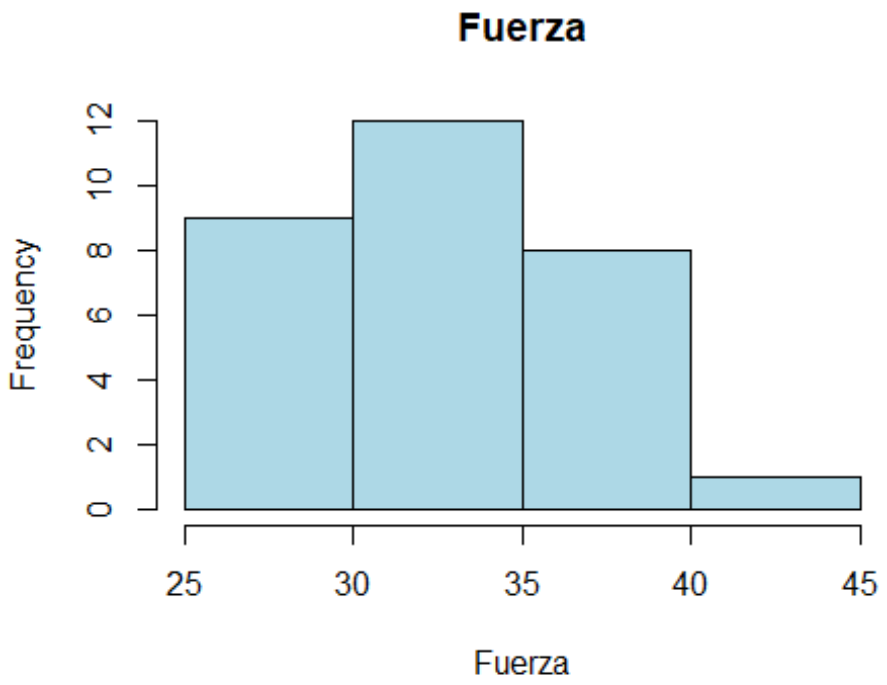
```
varianza
```

```
##    Potencia Temperatura      Tiempo Resistencia
##   186.20690   517.24138    20.68966    80.18133
```

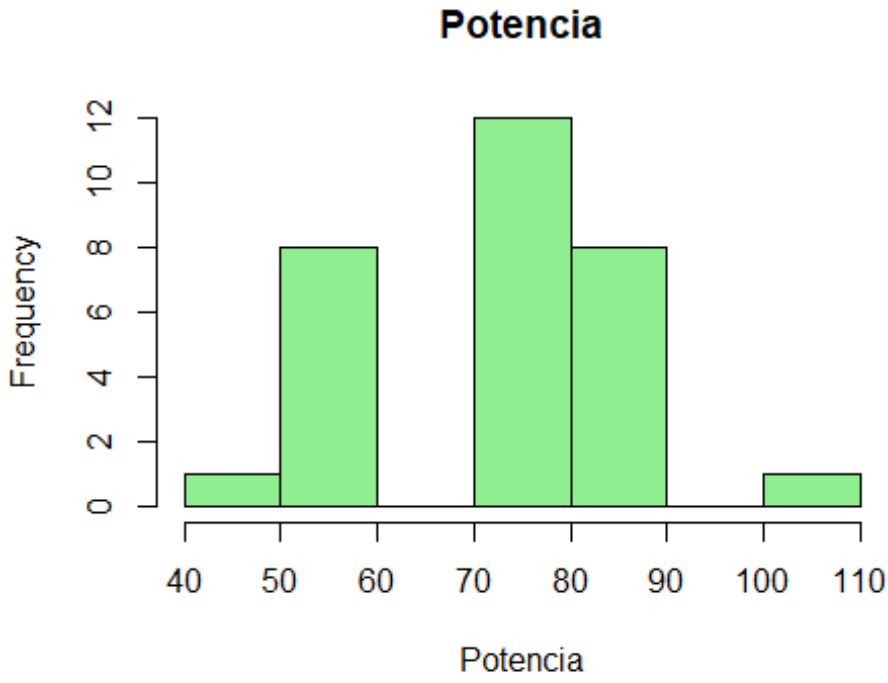
Se observa la desviación estandar de nuestras variables numericas y se calcula la varianza de las variables numericas

Histogramas de las variables

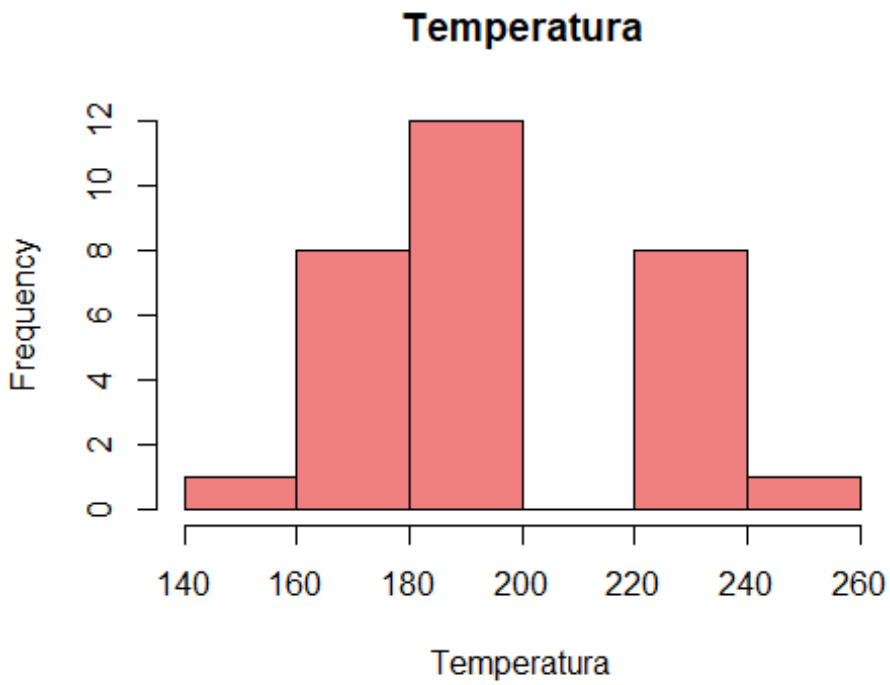
```
hist(datos$Fuerza, main="Fuerza", xlab="Fuerza", col="lightblue")
```



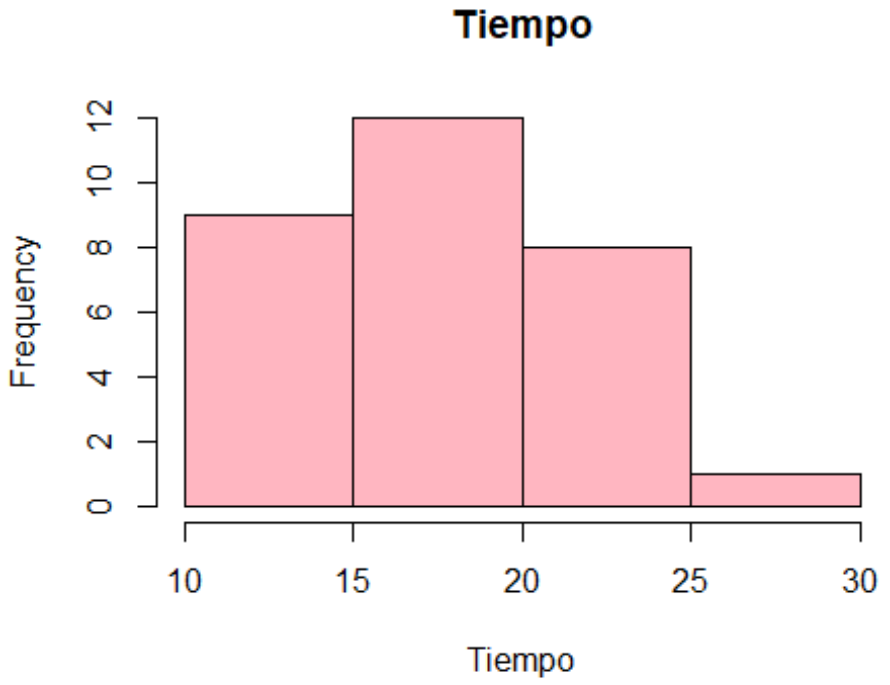

```r
hist(datos$Potencia, main="Potencia", xlab="Potencia", col="lightgreen")
```



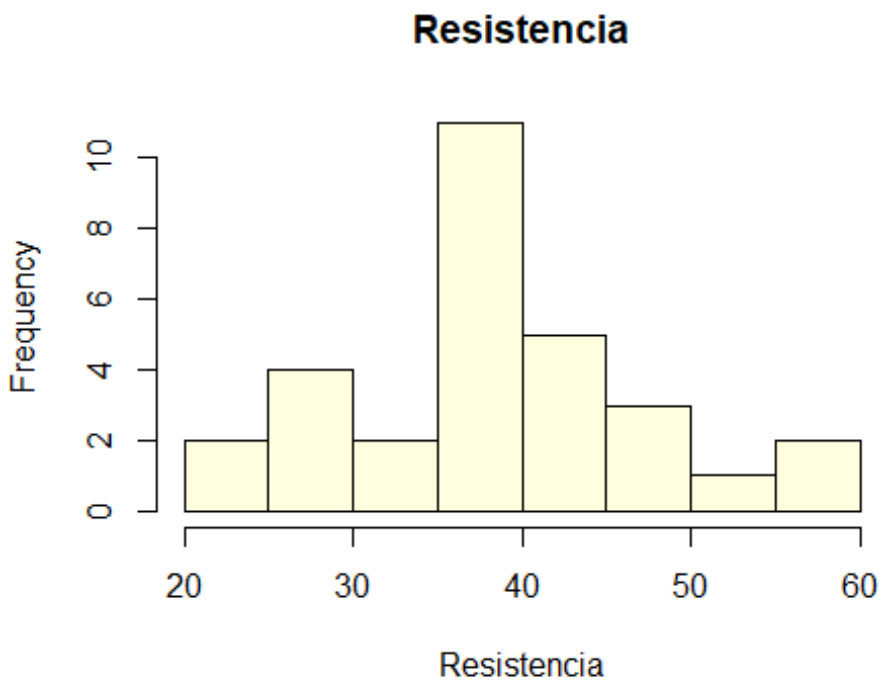

```r
hist(datos$Temperatura, main="Temperatura", xlab="Temperatura",
col="lightcoral")
```




```r
hist(datos$Tiempo, main="Tiempo", xlab="Tiempo", col="lightpink")
```

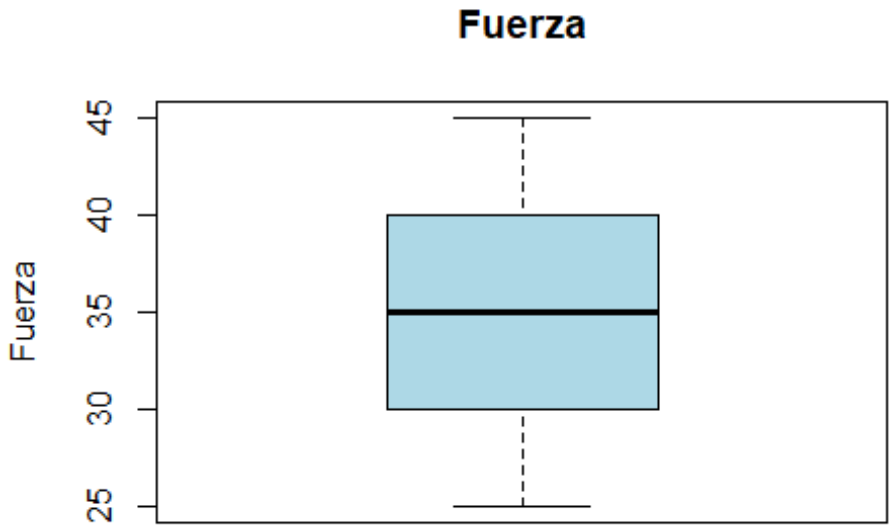



```r
hist(datos$Resistencia, main="Resistencia", xlab="Resistencia",
col="lightyellow")
```

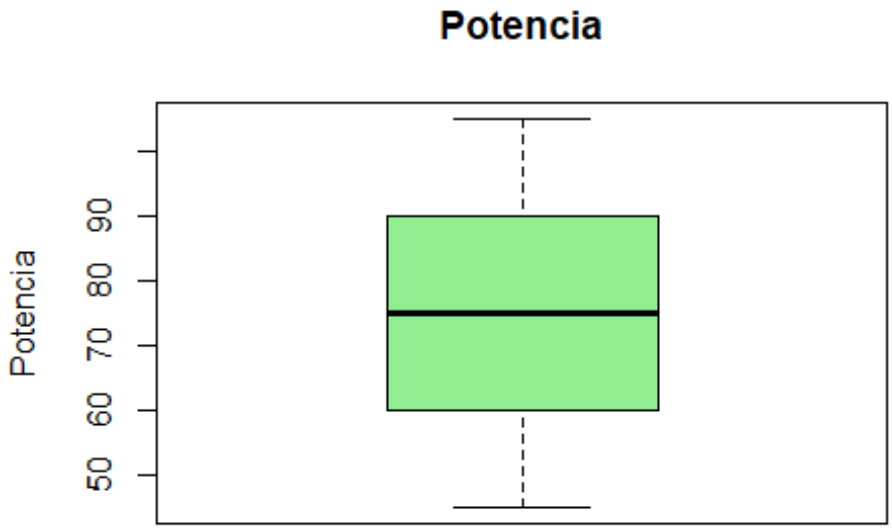


Boxplot de las variables

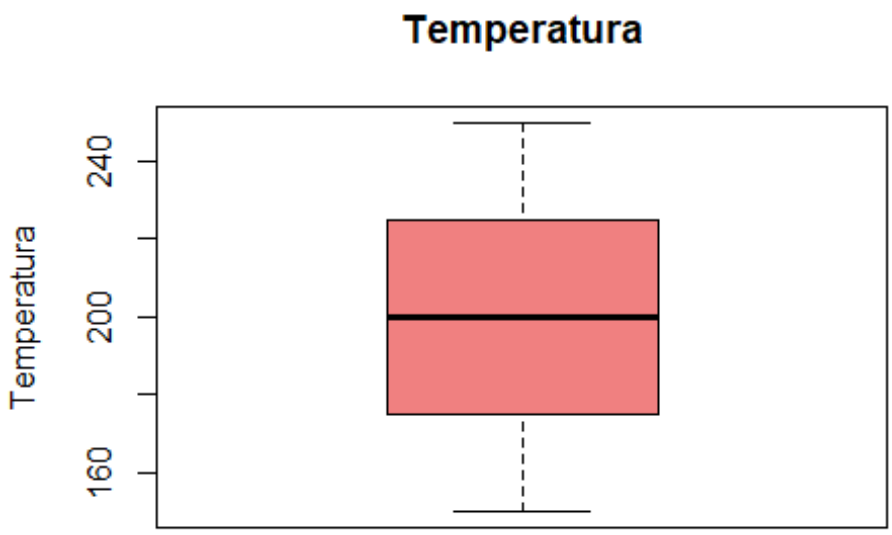
```
boxplot(datos$Fuerza, main="Fuerza", ylab="Fuerza", col="lightblue")
```



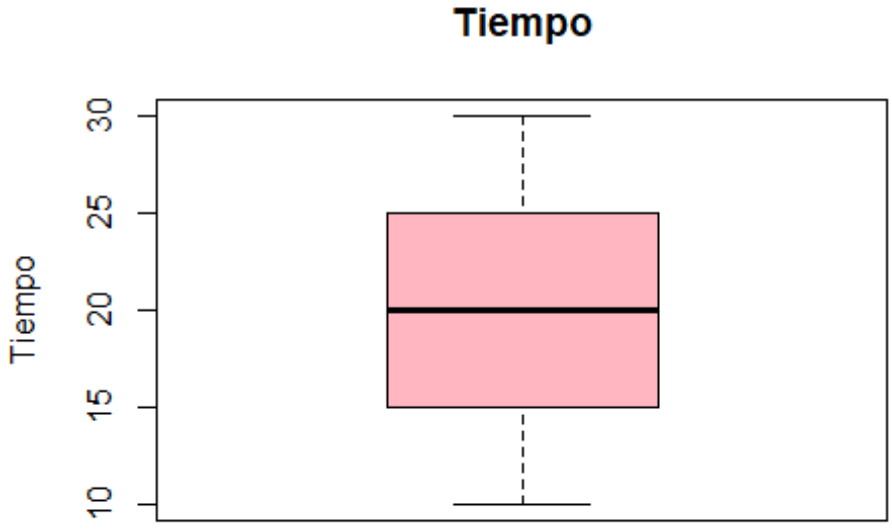
```r
boxplot(datos$Potencia, main="Potencia", ylab="Potencia",
col="lightgreen")
```



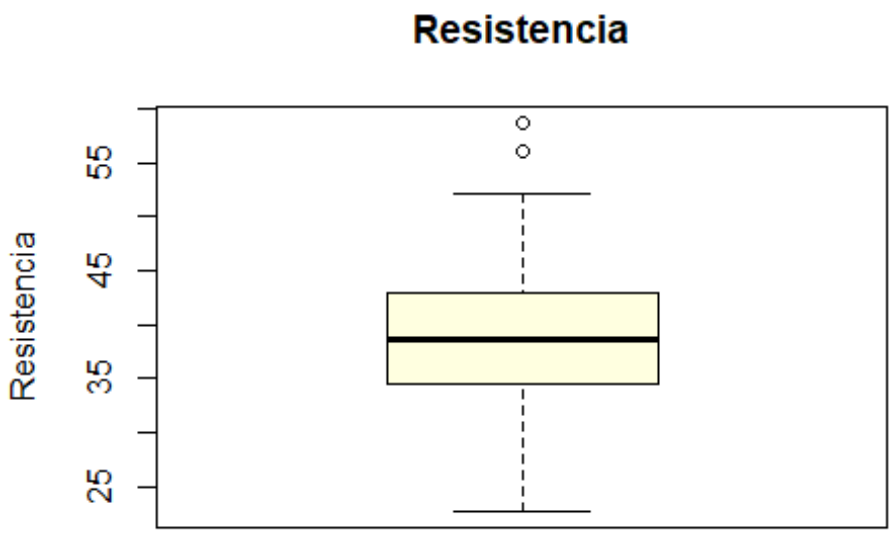

```r
boxplot(datos$Temperatura, main="Temperatura", ylab="Temperatura",
col="lightcoral")
```




```r
boxplot(datos$Tiempo, main="Tiempo", ylab="Tiempo", col="lightpink")
```




```r
boxplot(datos$Resistencia, main="Resistencia", ylab="Resistencia",
col="lightyellow")
```



Con estos gráficos vemos la distribución de las variables y la relación con la resistencia al corte. Para darnos una idea general de los datos con los que estamos rabajando.

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

Vamos a tener 3 métodos: mixto , forward y backward; para el modelo completo(cuenta con todas las variables predictoras) y para el modelo nulo (no cuenta con predictores)

```
modeloCompleto <- lm(Resistencia ~ ., data = datos)

modeloNulo <- lm(Resistencia ~ 1, data = datos)
```

Modelo mixto: se añaden o eliminan variables de acuerdo el criterio de mejor ajuste:

```
pasosMixto <- step(modeloCompleto, 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
```

```
summary(pasosMixto)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = datos)
##
## 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
```

Modelo forward: este metodo se comienza con el modelo nulo y se van añadiendo las variables más significativas

```
pasosForward <- step(modeloNulo, scope = list(lower = modeloNulo, upper =
modeloCompleto), 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
```

```
summary(pasosForward)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = datos)
##
## 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
```

modelo backward: este metodo se comienza con el metodo completo y se van eliminando las variables significativas

```
pasosBackward <- step(modeloCompleto, direction = "backward")
```

```
## 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
## - 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
## - Temperatura  1     252.2  984.24 108.72
## - Potencia     1    1341.0 2073.06 131.07
```

```
summary(pasosBackward)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = datos)
##
## 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
```

De los 3 métodos que tenemos mixto, forward y backward vemos que el modelo final que es más adecuado incluye potencia y temperatura como sus variables predictoras significativas para lograr explicar la resistencia.

Vemos que potencia y temperatura son altamente signifcativa en todos los modelos ya que en ambos casos p<0.001.

Significancia del modelo:

Economía de las variables

Significación global (Prueba para el modelo)

Significación individual (Prueba para cada βi)

Variación explicada por el modelo

```
anova(pasosMixto)
```

```
## Analysis of Variance Table
##
## Response: Resistencia
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Potencia     1 1341.02 1341.02  49.461 1.465e-07 ***
## Temperatura  1  252.20  252.20   9.302  0.005082 **
## Residuals   27  732.04   27.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(pasosForward)
```

```
## Analysis of Variance Table
##
## Response: Resistencia
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Potencia     1 1341.02 1341.02  49.461 1.465e-07 ***
## Temperatura  1  252.20  252.20   9.302  0.005082 **
## Residuals   27  732.04   27.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(pasosBackward)
```

```
## Analysis of Variance Table
##
## Response: Resistencia
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Potencia     1 1341.02 1341.02  49.461 1.465e-07 ***
## Temperatura  1  252.20  252.20   9.302  0.005082 **
```

```
## Residuals   27  732.04   27.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que en estos resultados de anova nos aseguramos de demostrar que potencia y temperatura son predictores significativos y robustos para explicar la variabilidad en la Resistencia.

```
summary(pasosMixto)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -24.9016667 10.07206836 -2.472349 2.001412e-02
## Potencia      0.4983333  0.07085806  7.032839 1.465430e-07
## Temperatura   0.1296667  0.04251483  3.049916 5.082118e-03
```

```
summary(pasosForward)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -24.9016667 10.07206836 -2.472349 2.001412e-02
## Potencia      0.4983333  0.07085806  7.032839 1.465430e-07
## Temperatura   0.1296667  0.04251483  3.049916 5.082118e-03
```

```
summary(pasosBackward)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -24.9016667 10.07206836 -2.472349 2.001412e-02
## Potencia      0.4983333  0.07085806  7.032839 1.465430e-07
## Temperatura   0.1296667  0.04251483  3.049916 5.082118e-03
```

Vemos que los coeficientes de los modelos son consistentes y por los interceptos la línea de regresión no pasa por el origen

```
summary(pasosMixto)$r.squared
```

```
## [1] 0.6851783
```

```
summary(pasosMixto)$adj.r.squared
```

```
## [1] 0.6618581
```

```
summary(pasosForward)$r.squared
```

```
## [1] 0.6851783
```

```
summary(pasosForward)$adj.r.squared
```

```
## [1] 0.6618581
```

```
summary(pasosBackward)$r.squared
```

```
## [1] 0.6851783
```

```
summary(pasosBackward)$adj.r.squared
```

```
## [1] 0.6618581
```

Vemos que tenemos R ^2 y R^2 ajsutados con valores similares y robustos

```
length(coef(pasosMixto)) - 1
```

```
## [1] 2
```

```
length(coef(pasosForward)) - 1
```

```
## [1] 2
```

```
length(coef(pasosBackward)) - 1
```

```
## [1] 2
```

Vemos que independientememte de cual de los 3 modelos elijamos en todos al final se seleccionan 2 variables predictoras: potencia y temperatura.

Vemos que cada modelo acaba con las mismas 2 variables , lo cual nos refleja la importancia de estas 2 variables.

Analiza la validez del modelo encontrado:

Con los 3 modelos se encontraron resultados muy similares ya que al final todos los modelos terminaron con las mismas 2 variables significativas: potencia y temperatura. El R^2 ajustado tambien fueron constantemente altos en todos los métodos. Y como al final todos los metodos seleccionaron las msima variables tienen una economia similar ya sin las variables innecesarias.

Vemos que los metodos tienen un resultado extremadamente similar , es decir son validos y efectivos, sin embargo el método forward podria ser preferible por el hecho de que por su naturaleza incremenat es potencialmente más eficiente en identificar el modelo óptimo haciendo menos calculos en comparación con el mixto o el backward.

Análisis de residuos (homocedasticidad, independencia, etc)

```
if (!require(lmtest)) {
    install.packages("lmtest")
    library(lmtest)
}
```

```
## Cargando paquete requerido: lmtest
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

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

```r
if (!require(car)) {
    install.packages("car")
    library(car)
}
```

```
## Cargando paquete requerido: car
```

```
## Cargando paquete requerido: carData
```

```r
modeloForward <- lm(Resistencia ~ Potencia + Temperatura, data = datos)

bptest(modeloForward)
```

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

```r
dwtest(modeloForward)
```

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

Con estos resultados vemos que cumple adecuadamente con las suposiciones de homocedasticidad e independencia de los residuos. ### No multicolinealidad de Xi

```r
vif(modeloForward)
```

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

Por los valores de VIF vemos que no hay multicolinealidad entre las 2 variablese predictoras de nuestro modelo. Un VIF igual a 1 es un valor ideal, y tambien nos indica que cada predictor es completamente independiente de los otros o sea que esta libre de cualquier problema de multicolinealidad.

Emite conclusiones sobre el modelo final encontrado e interpreta en el contexto del problema el efecto de las variables predictoras en la variable respuesta

En cada paso se fue realizando un análisis por cada resultado encontrado, despues de análisar estos resultados cuidadosamente vemos que le modelo de forward ademas de ser un muy buen modelo y cumplir con todos los requisitos cu,ple con las suposiciones básicas de homocedasticidad e independencia de los residuos y está libre de problemas de multicolinealidad , o sea que el modelo forward es estadisticamente

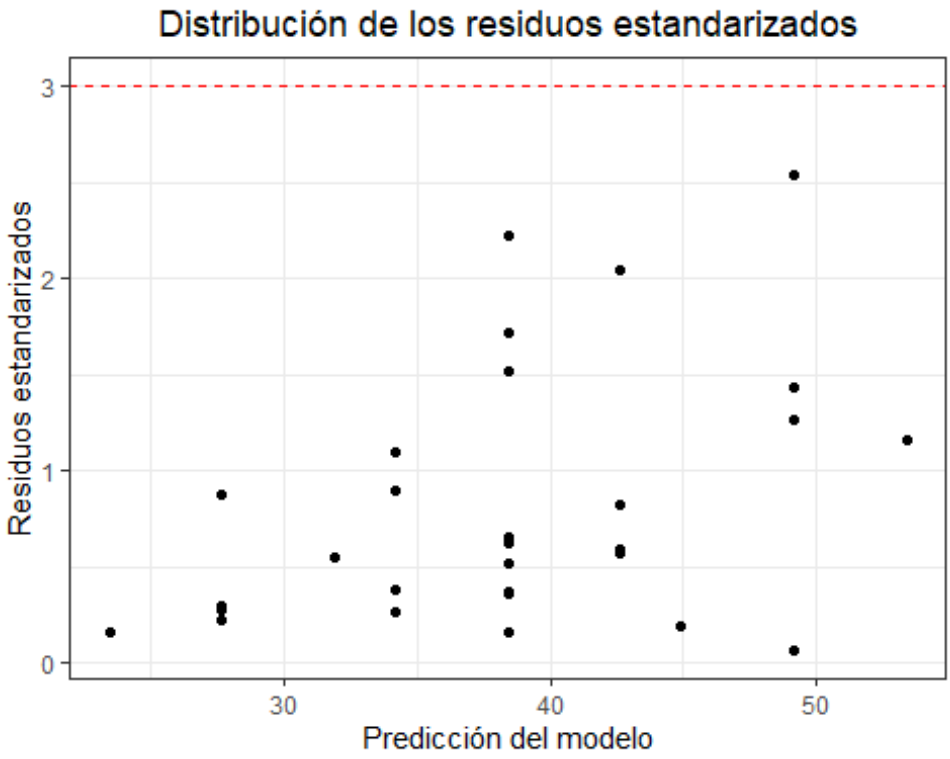
sólido y confiable para la predicción de resistencia al corte basada en las variables de Potencia y Temperatura.

Haz el análisis de datos atípicos e incluyentes del mejor modelo encontrado

Residuos Estandarizados y Observaciones Atípicas

Se tratan de encontrarlos residuos que se enceuntran a más de 3 desviaciones estándar de los valores ajustados para lograr identificar observaciones atípicas.

```
if (!require(dplyr)) {
    install.packages("dplyr")
    library(dplyr)
}

## Cargando paquete requerido: dplyr

##
## Adjuntando el paquete: '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

datos$residuos_estandarizados <- rstudent(modeloForward)

library(ggplot2)
ggplot(data = datos, aes(x = predict(modeloForward), y =
abs(residuos_estandarizados))) +
  geom_hline(yintercept = 3, color = "red", linetype = "dashed") +
  geom_point(aes(color = ifelse(abs(residuos_estandarizados) > 3, 'red',
'black'))) +
  scale_color_identity() +
  labs(title = "Distribución de los residuos estandarizados", x =
"Predicción del modelo", y = "Residuos estandarizados") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
atipicos <- which(abs(datos$residuos_estandarizados) > 3)
datos[atipicos, ]
```

```
## [1] Fuerza                  Potencia                Temperatura
## [4] Tiempo                  Resistencia
residuos_estandarizados
## <0 rows> (o 0- extensión row.names)
```

En el grafico de los residuos estandarizados vemos que no hay observaciones con residuos que excedan el umbral de 3 desviaciones estándar, o sea que no se pasan de la linea roja, lo cual nos dice que no hay datos atípicos severos en términos de residuos estandarizados.

Y tambien los residuos están distribuidos de manera uniforme y no vemos patrones evidentes.

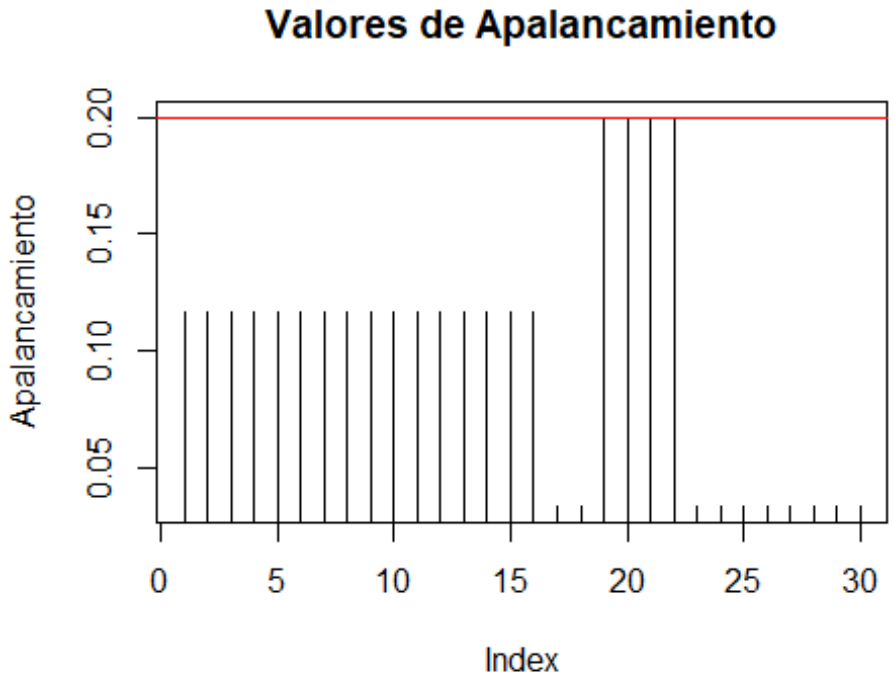
Leverage

Aquí se calculan los valores de apalancamiento y se visualizan cuáles observaciones tienen un alto apalancamiento o sea , influencia en el ajuste del modelo.

“El apalancamiento (leverage) mide la distancia de un punto de datos con respecto a la media de los predictores. Un punto de datos con alto apalancamiento está alejado del”centro” de los valores predictores.”

```
leverage <- hatvalues(modeloForward)
```

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




```r
high_leverage_points <- which(leverage > 2 * mean(leverage))
datos[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
```

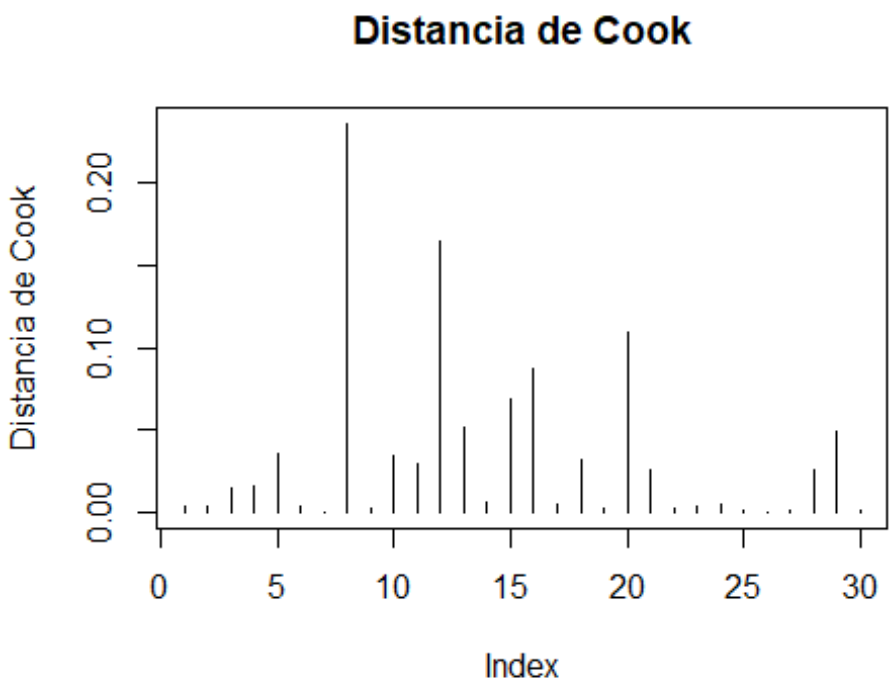
vemos que a pesar de que las observaciones 19 y 20 tienen un apalancamiento alto, no parecen ser problemáticas en términos de su influencia sobre el modelo.

Distancia de Cook

Con la distancia de cook medimos la influencia que tiene cada observación en todos los coeficientes del modelo. Las observacione que tengan un valor de Cook mayor a 1 se consideradan influyentes.

```r
cooksd <- cooks.distance(modeloForward)

plot(cooksd, type = "h", main = "Distancia de Cook", ylab = "Distancia de
```

```
Cook")
abline(h = 1, col = "red")
```



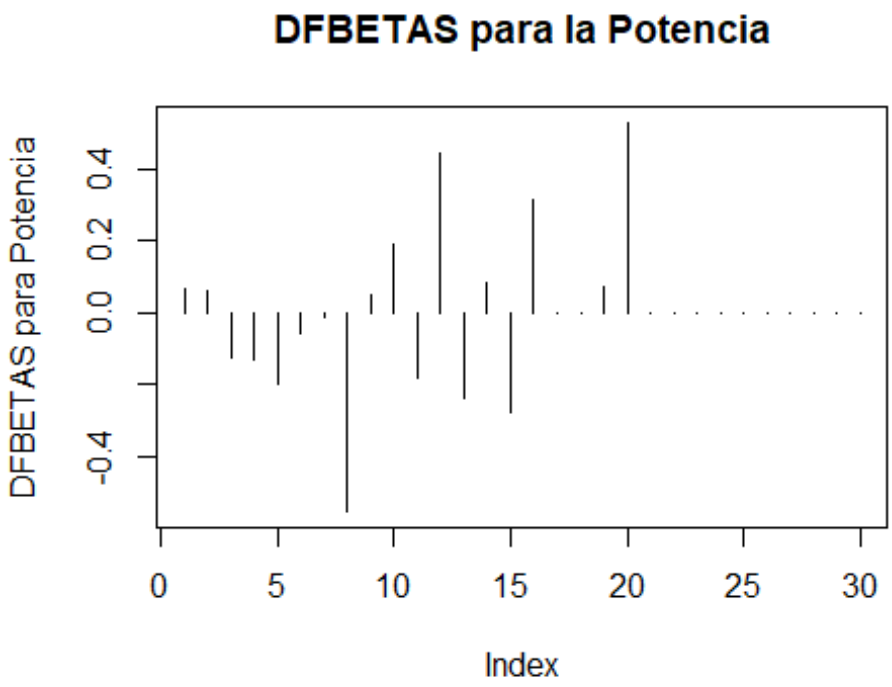

```
puntos_influyentes <- which(cooksd > 1)
datos[puntos_influyentes, ]

## [1] Fuerza                  Potencia                Temperatura
## [4] Tiempo                  Resistencia
residuos_estandarizados
## <0 rows> (o 0- extensión row.names)
```

vemos que ningún punto supera el valor de 1. No hay observaciones con una distancia de Cook que sea lo suficientemente alta como para ser consideradas influyentes en el modelo, o sea que ningún punto tiene un impacto excesivo sobre los coeficientes del modelo de regresión.

DFBETAS

Los valores de DFBETAS nos van a indicar el cambio en los coeficientes de regresión en el caso de que una observación específica es eliminada. Los valores de DFBETAS que sean mayores a 1 (en términos absolutos) son preocupantes.

```
dfbetas_values <- dfbetas(modeloForward)

plot(dfbetas_values[, 2], type = "h", main = "DFBETAS para la Potencia",
ylab = "DFBETAS para Potencia")
abline(h = c(-1, 1), col = "red")
```




```
puntos_influyentes_dfbetas <- which(abs(dfbetas_values[, 2]) > 1)
datos[puntos_influyentes_dfbetas, ]

## [1] Fuerza                  Potencia               Temperatura
## [4] Tiempo                  Resistencia
residuos_estandarizados
## <0 rows> (o 0- extensión row.names)
```

Ninguno de los valores de DFBETAS supera el umbral común de 1 absoluto , lo que nos indica que ninguna observación tiene una influencia significativa en la estimación del coeficiente de Potencia.

Medidas de Influencia General

por ultimo se hace un resumen general de las medidas de influencia

```
influencia <- influence.measures(modeloForward)
summary(influencia)

## Potentially influential observations of
##   lm(formula = Resistencia ~ Potencia + Temperatura, data = datos) :
##
##    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
```

Vemos que ninguna observación parece que es excesivamente influyente en el modelo en términos de Cook's Distance. Por otro lado , las observaciones 19, 21, y 22 podrían estar afectando la estabilidad del modelo.

##Conclusión

El análisis de datos atípicos e influyentes realizado sobre el modelo de regresión que predice la resistencia a partir de las variables potencia y temperatura nos revela lo siguiente:

En el análisis no se detectaron observaciones altamente influyentes o atípicas que comprometan la validez general del modelo. Aunque algunas observaciones, en particular las 19, 21 y 22 nos muestran ciertos indicios de influencia sobre la estructura del modelo. En conclusión el modelo es robusto y se comporta bien con los datos actuales, y no tenemos la necesidad de tener que eliminar o ajustar observaciones. específicas.