A3-Regresión Múltiple-Detección datos atípicos

Ozner Leyva

2024-09-24

```
library(car)

## Loading required package: carData

library(ggplot2)

M = read.csv("C:/Users/ozner/Downloads/AlCorte.csv")
```

Parte 1: Haz un análisis descriptivo de los datos: medidas principales y gráficos (ya se hizo en la actividad A2 y la profesora dijo que no se tenía que subir nada en esta sección)

Parte 2: Encuentra el mejor modelo de regresión que explique la variable Resistencia (ya se hizo en la actividad A2 y la profesora dijo que no se tenía que subir nada en esta sección)

Parte 3: Analiza la validez del modelo encontrado (ya se hizo en la actividad A2 y la profesora dijo que no se tenía que subir nada en esta sección)

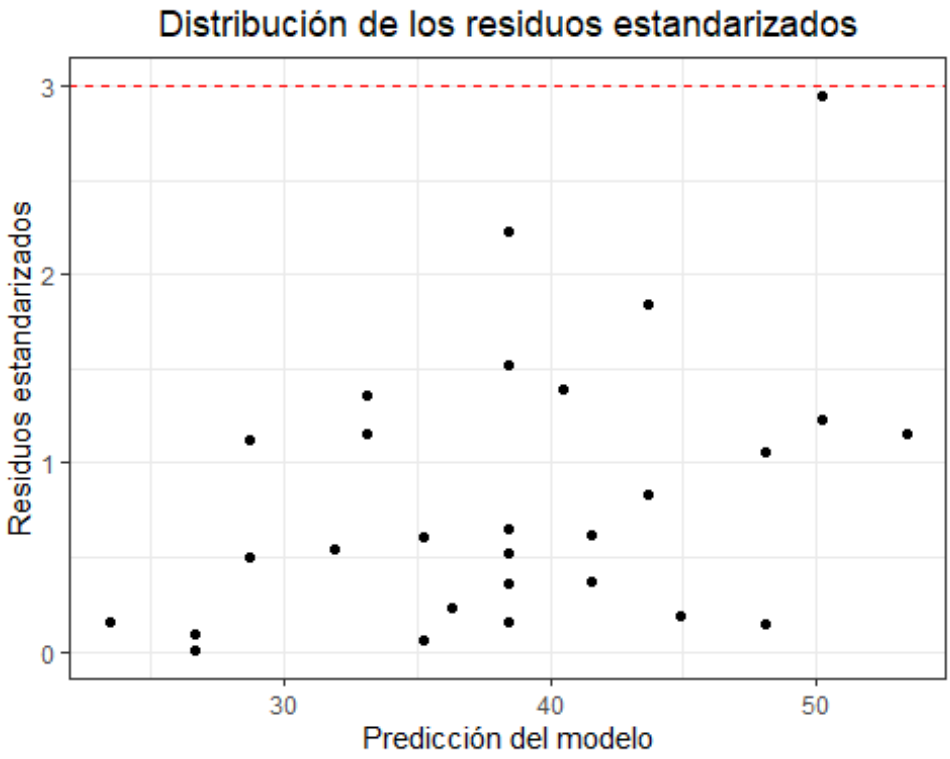
Parte 4: Haz el análisis de datos atípicos e incluyentes del mejor modelo encontrado

```
# Modelo 2: Excluir la variable Tiempo
modelo_2 <- lm(Resistencia ~ ï..Fuerza + Potencia + Temperatura, data =
M)
```

4.1 Detección de Datos Atípicos

```
# Calcular residuos estandarizados
M$residuos_estandarizados <- rstudent(modelo_2)

# Gráfico para visualizar los residuos estandarizados
ggplot(data = M, aes(x = predict(modelo_2), 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))
```



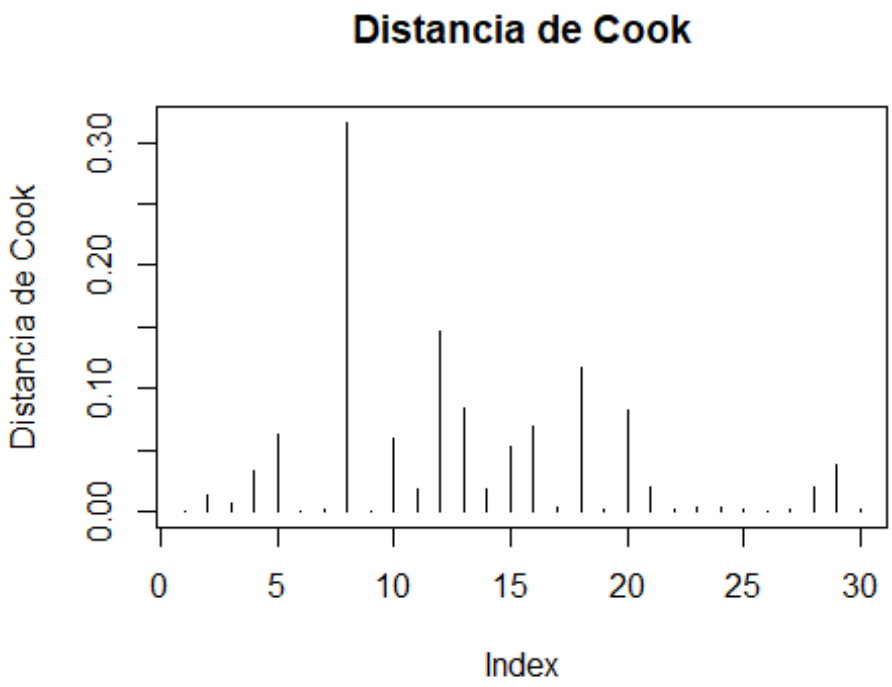

```
# Identificar observaciones con residuos estandarizados mayores a 3
Atipicos <- which(abs(M$residuos_estandarizados) > 3)
M[Atipicos, ]
```

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

4.2 Detección de Datos Influyentes

```
# Calcular la distancia de Cook
cooksd <- cooks.distance(modelo_2)

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



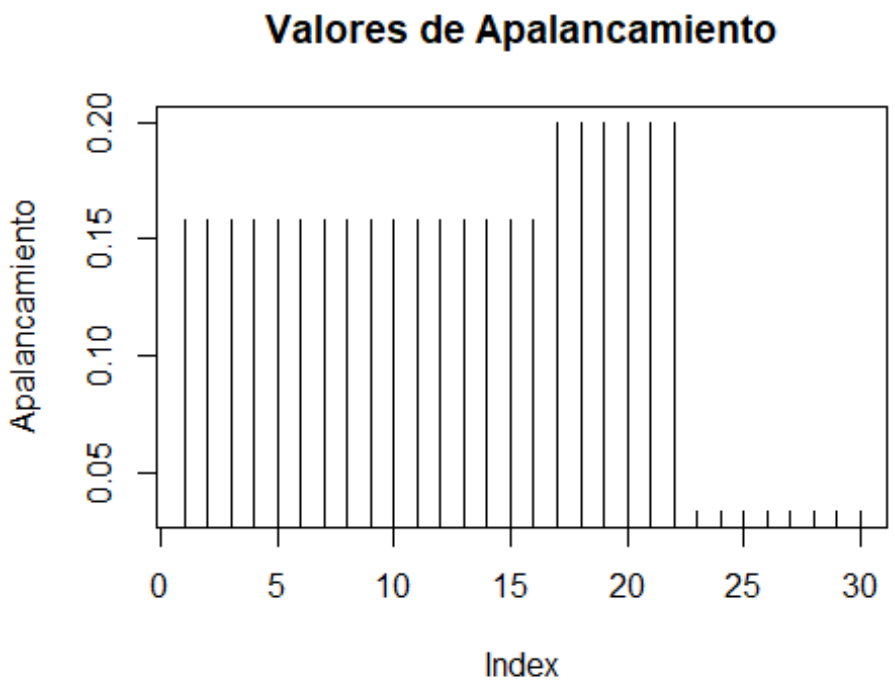

```
# Identificar puntos influyentes (distancia de Cook > 1)
puntos_influyentes <- which(cooksd > 1)
M[puntos_influyentes, ]
```

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

4.3 Análisis del Leverage

```
# Calcular leverage
leverage <- hatvalues(modelo_2)

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




```
# Identificar puntos con leverage alto
high_leverage_points <- which(leverage > 2 * mean(leverage))
M[high_leverage_points, ]
```

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

4.4 Resumen de Medidas de Influencia

```
# Resumen de medidas de influencia
influencia <- influence.measures(modelo_2)
```

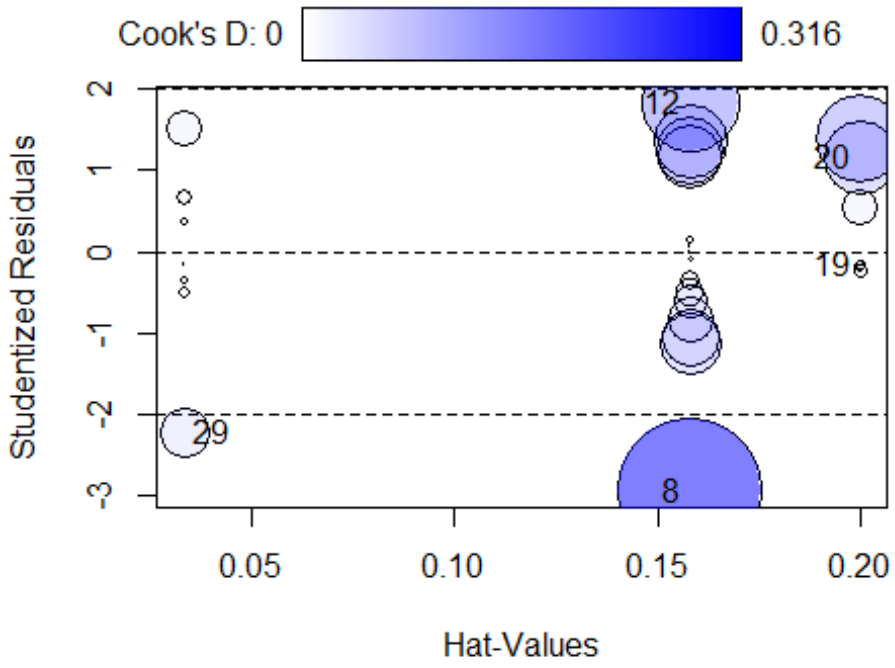
```
## Warning in abbreviate(vn): abreviatura utilizada con caracteres no
ASCII
```

```
summary(influencia)
```

```
## Potentially influential observations of
##   lm(formula = Resistencia ~ ï..Fuerza + Potencia + Temperatura,
data = M) :
##
##   dfb.1_  dfb.ïFrz dfb.Ptnc dfb.Tmpr dffit   cov.r   cook.d hat
## 8  1.07_* -0.66    -0.66    -0.66    -1.28_*  0.42_*  0.32   0.16
```

```
# Gráfico combinado de leverage, distancia de Cook y residuos
estandarizados
```

```r
library(car)
influencePlot(modelo_2)
```




```
##       StudRes        Hat       CookD
## 8  -2.9506791 0.15833333 0.315846303
## 12  1.8370333 0.15833333 0.145428149
## 19 -0.1593699 0.20000000 0.001649243
## 20  1.1544398 0.20000000 0.082243207
## 29 -2.2229729 0.03333333 0.036992066
```

Interpretación de esta actividad

1. Distribución de los residuos estandarizados

Al analizar el gráfico de la distribución de los residuos estandarizados, se puede ver que ninguno de los residuos excede el umbral de 3 desviaciones estándar. Esto nos sugiere que no hay datos atípicos extremos en términos de los valores residuales. Sin embargo, los residuos que se acercan al valor de 3 desviaciones estándar, como es el caso de la observación número 8, podrían considerarse potencialmente atípicos y merecen un examen más detallado.

En resumen, aunque no hay residuos extremos, algunos valores residuales cercanos al límite de 3 desviaciones estándar podrían indicar observaciones que merecen una mayor revisión.

2. Distancia de Cook

El gráfico de la distancia de Cook nos muestra que la mayoría de las observaciones tienen valores bajos, lo cual es esperado en un modelo ajustado adecuadamente. Sin embargo, la observación número 8 se destaca por tener un valor de Cook más alto, cercano a 0.32, aunque todavía por debajo del umbral común de 1. Esto nos sugiere que, si bien la observación número 8 tiene cierta influencia en el modelo, su impacto no es lo suficientemente significativo como para requerir una intervención inmediata.

3. Valores de Apalancamiento

El gráfico de valores de apalancamiento nos muestra que las observaciones con índices más bajos tienen un mayor leverage en comparación con las demás. En otras palabras, estas observaciones iniciales ejercen una mayor influencia en el ajuste del modelo.

Además, las observaciones con alto leverage, como las cercanas a los índices 12, 19 y 20, también pueden tener un impacto considerable en el ajuste del modelo, especialmente si presentan residuos elevados. Esto significa que estos datos con alto leverage podrían estar influyendo de manera significativa en los resultados del modelo.

4. Resumen de Medidas de Influencia

El resumen de las medidas de influencia, que incluye los valores de residuos estudentizados, apalancamiento y distancia de Cook, destaca varias observaciones, en particular las números 8, 12, 19, 20 y 29.

De estas, la observación número 8 es la más notable, ya que presenta valores destacados en cuanto a residuos estandarizados, leverage y distancia de Cook. Esto nos sugiere que la observación número 8 podría estar ejerciendo una influencia significativa en el ajuste del modelo.

5. Gráfico combinado de Residuos vs. Apalancamiento (Influence Plot)

En el gráfico combinado de residuos estandarizados versus leverage (influence plot), la observación número 8 se destaca por presentar un residuo estandarizado negativo cercano a -3, combinado con un valor de leverage moderado y una distancia de Cook considerable.

También, las observaciones 12 y 20 también presentan valores importantes de leverage, aunque sus residuos estandarizados no son tan extremos como en el caso de la observación 8.

Conclusión general

La observación número 8 es la que más destaca como un posible punto atípico e influyente. Si bien no excede los umbrales críticos establecidos, sería prudente examinar esta observación más a fondo o considerar su impacto en el modelo.

Además, las observaciones 12, 19 y 20 presentan valores de leverage elevados. Esto indica que dichas observaciones están relativamente alejadas del centro de los datos en el espacio de los predictores, y podrían afectar el ajuste del modelo si también presentan residuos grandes.

En resumen, aunque el modelo parece ser relativamente estable, sería recomendable hacer pruebas adicionales o considerar modelos alternativos que puedan manejar mejor estas observaciones que parecen ser atípicas.