M5_A3

Sofia Cantu

2024-09-24

13. Regresión Múltiple: Detección datos atípicos

```
 # Librerías
if (!require(car)) install.packages("car")
```

```
## Loading required package: car
```

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

```
library(car)
if (!require(ggplot2)) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
```

```
M = read.csv("~/Downloads/ArchivosCodigos/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

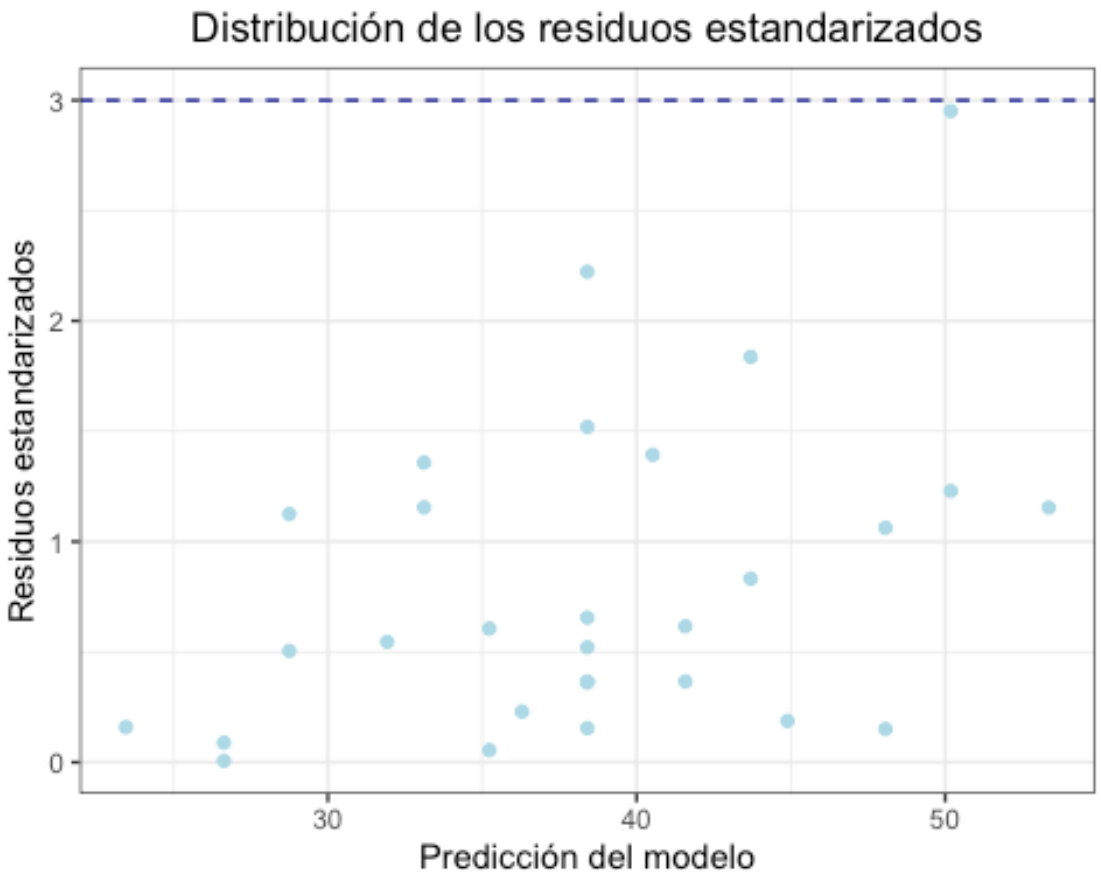
```
# Modelo 2: Excluir la variable Tiempo
modelo_2 <- lm(Resistencia ~ Fuerza + Potencia + Temperatura, data = M)
```

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 = "blue4", linetype = "dashed") +
  geom_point(aes(color = ifelse(abs(residuos_estandarizados) > 3, 'blue4',
'lightblue'))) +
  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))
```



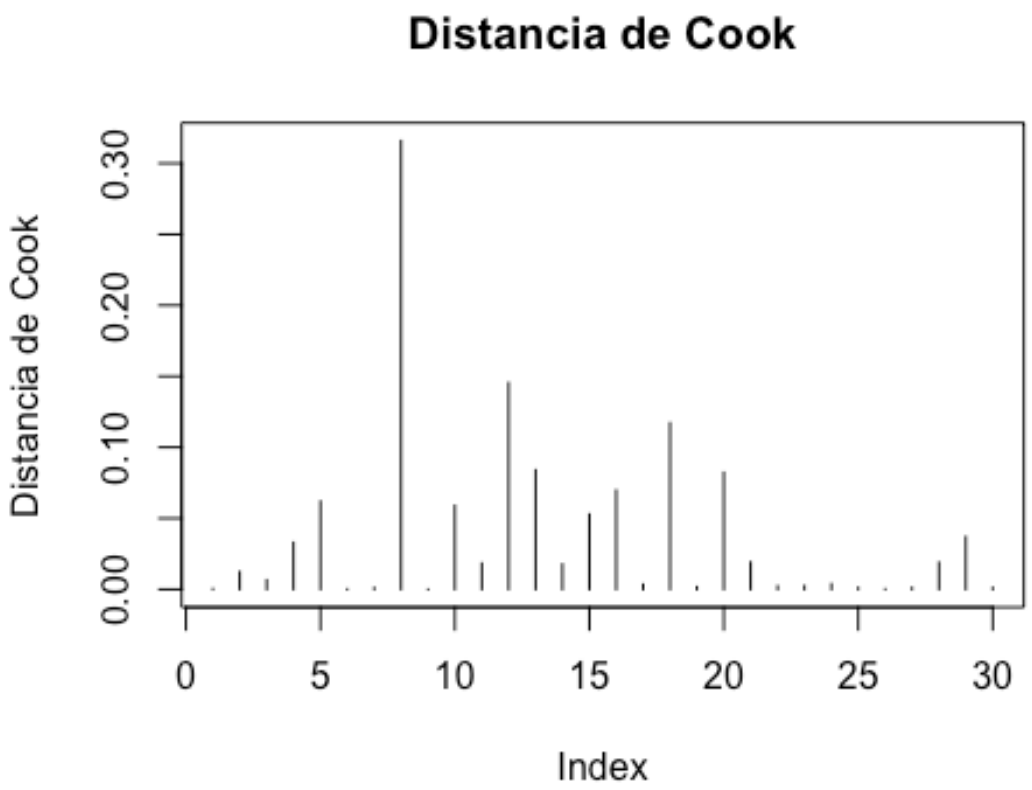

```
# Identificar observaciones con residuos estandarizados mayores a 3
Atipicos <- which(abs(M$residuos_estandarizados) > 3)
M[Atipicos, ]
```

```
## [1] 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)
```

```r
# Gráfico para visualizar la distancia de Cook
plot(cooksd, type="h", main="Distancia de Cook", ylab="Distancia de Cook")
abline(h = 1, col="blue4") # Límite comúnmente usado
```



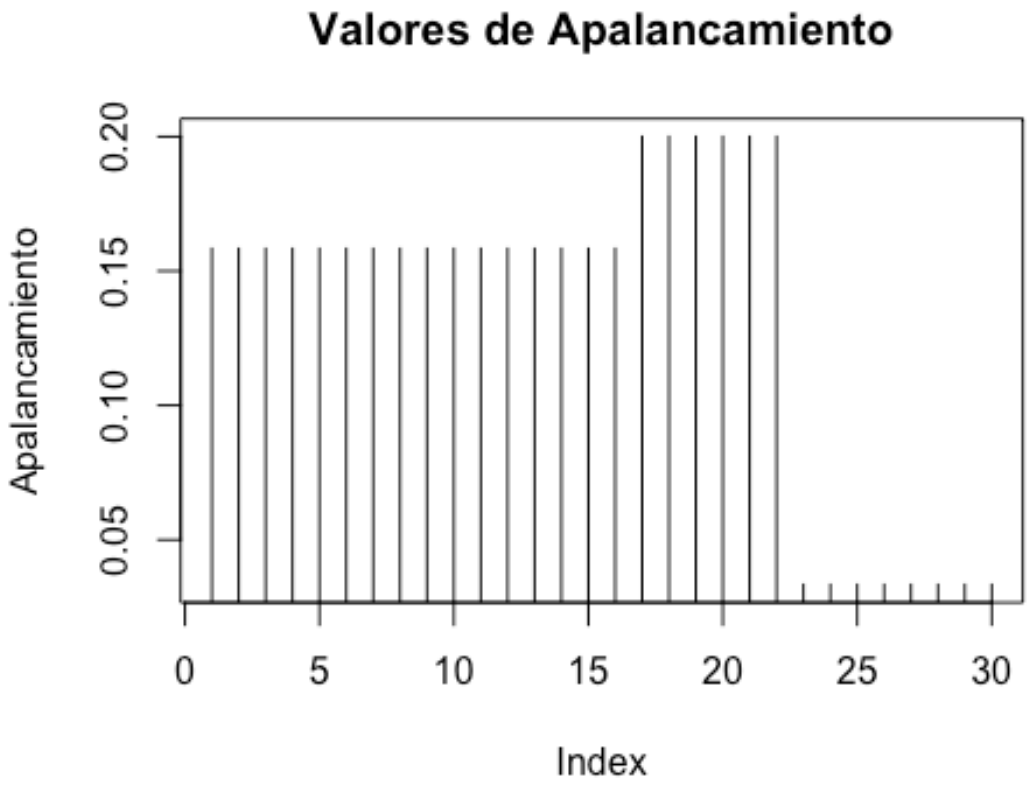

```r
# Identificar puntos influyentes (distancia de Cook > 1)
puntos_influyentes <- which(cooksd > 1)
M[puntos_influyentes, ]
```

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

4.3 Análisis del Leverage

```r
# 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="blue4") # Límite comúnmente usado
```




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

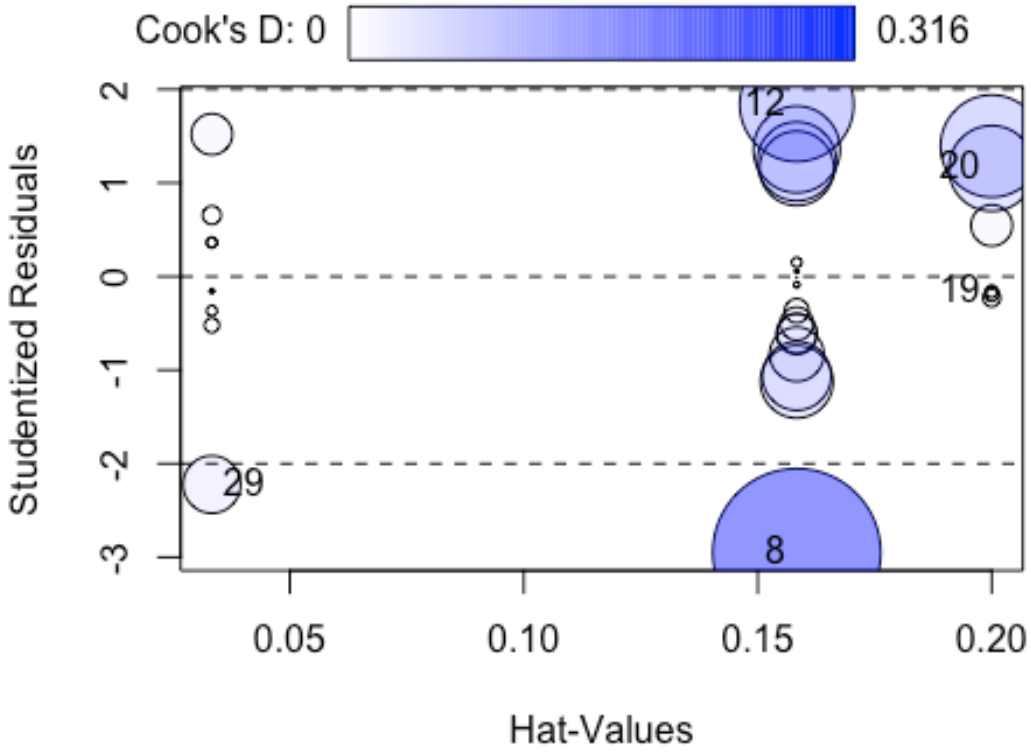
```
## [1] Fuerza                  Potencia                Temperatura
## [4] Tiempo                  Resistencia
residuos_estandarizados
## <0 rows> (or 0-length row.names)
```

4.4 Resumen de Medidas de Influencia

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

```
## Potentially influential observations of
##   lm(formula = Resistencia ~ Fuerza + Potencia + Temperatura, data = M) :
##
##   dfb.1_  dfb.Furz 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
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, observamos que no se presentan residuos que excedan el umbral de 3, lo que indica la ausencia de datos atípicos extremos en términos de sus valores residuales. Sin embargo, los residuos que se aproximan al valor de 3, como es el caso de la observación número 8, podrían considerarse potencialmente atípicos y merecen una revisión más detallada.

2. Distancia de Cook

El gráfico de la distancia de Cook 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 destaca por tener un valor de Cook más elevado, cercano a 0.32,

aunque todavía por debajo del umbral común de 1. Esto sugiere que, si bien la observación número 8 ejerce cierta influencia en el modelo, no es lo suficientemente significativa como para requerir una acción inmediata.

3. Valores de Apalancamiento

El gráfico de valores de apalancamiento muestra que las observaciones con índices más bajos (es decir, las situadas al inicio del conjunto de datos) presentan un mayor leverage en comparación con las demás. Las observaciones con alto leverage (por ejemplo, las cercanas a los índices 12, 19 y 20) pueden tener un impacto considerable en el ajuste del modelo, especialmente si además presentan residuos elevados.

4. Resumen de Medidas de Influencia

El resumen de las medidas de influencia, que incluye StudRes, Hat y CookD, resalta varias observaciones, en particular las números 8, 12, 19, 20 y 29. Entre ellas, la observación número 8 es la más significativa en términos de residuos estandarizados, leverage y distancia de Cook, lo que indica que podría estar influyendo notablemente en el modelo.

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

Finalmente, 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 leverage moderado y una distancia de Cook considerable. Esta observación debe ser revisada cuidadosamente, ya que tiene el potencial de afectar el modelo. Las observaciones 12 y 20 también presentan valores significativos 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 más destacada en términos de ser un punto atípico e influyente. Aunque no excede los umbrales críticos, sería prudente examinarla más a fondo o considerar su impacto en el modelo.

Las observaciones 12, 19 y 20 presentan valores de leverage altos, lo que indica que están relativamente alejadas del centro de los datos en el espacio de los predictores y podrían afectar el ajuste del modelo si combinan leverage con residuos grandes.

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