M5_A2

Sofia Cantu

2024-09-17

13. Regresión Múltiple

En la base de datos, Al corte, se describe un experimento realizado para evaluar el impacto de las variables: fuerza, potencia, temperatura y tiempo sobre la resistencia al corte. Indica cuál es la mejor relación entre estas variables que describen la resistencia al corte.

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

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors
```

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

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

```
library(zoo)
```

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

```
##    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
## 7      30       90         225     15        48.8
## 8      40       90         225     15        37.8
## 9      30       60         175     25        26.6
## 10     40       60         175     25        23.4
## 11     30       90         175     25        38.6
## 12     40       90         175     25        52.1
## 13     30       60         225     25        39.5
## 14     40       60         225     25        32.3
## 15     30       90         225     25        43.0
## 16     40       90         225     25        56.0
## 17     25       75         200     20        35.2
## 18     45       75         200     20        46.9
## 19     35       45         200     20        22.7
## 20     35      105         200     20        58.7
## 21     35       75         150     20        34.5
## 22     35       75         250     20        44.0
## 23     35       75         200     10        35.7
## 24     35       75         200     30        41.8
## 25     35       75         200     20        36.5
## 26     35       75         200     20        37.6
## 27     35       75         200     20        40.3
## 28     35       75         200     20        46.0
## 29     35       75         200     20        27.8
## 30     35       75         200     20        40.3
```

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

```r
# Resumen estadístico
resumen <- summary(M)
print(resumen)
```

```
##      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
```

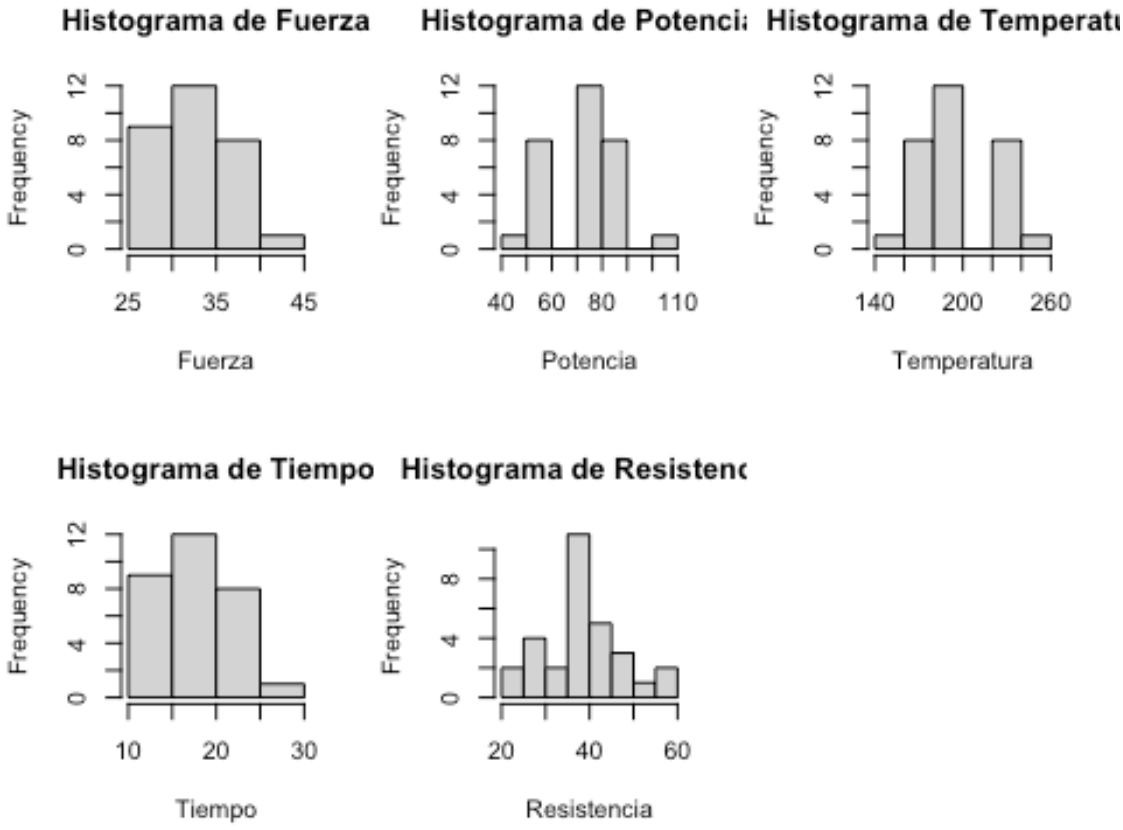
```r
# Desviación estándar
desviacion_estandar <- sapply(M, sd)
print(desviacion_estandar)
```

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

```r
# Gráficos

# Histogramas
par(mfrow=c(2,3))
for(col in names(M)) {
  hist(M[[col]], main=paste("Histograma de", col), xlab=col)
}

# Boxplots
par(mfrow=c(2,3))
```




```r
for(col in names(M)) {
  boxplot(M[[col]], main=paste("Boxplot de", col), ylab=col)
}

# Gráfico de dispersión de Resistencia vs otras variables
```

```r
ggplot(M, aes(x = Fuerza, y = Resistencia)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Resistencia vs Fuerza")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```r
ggplot(M, aes(x = Potencia, y = Resistencia)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Resistencia vs Potencia")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```r
ggplot(M, aes(x = Temperatura, y = Resistencia)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Resistencia vs Temperatura")
```

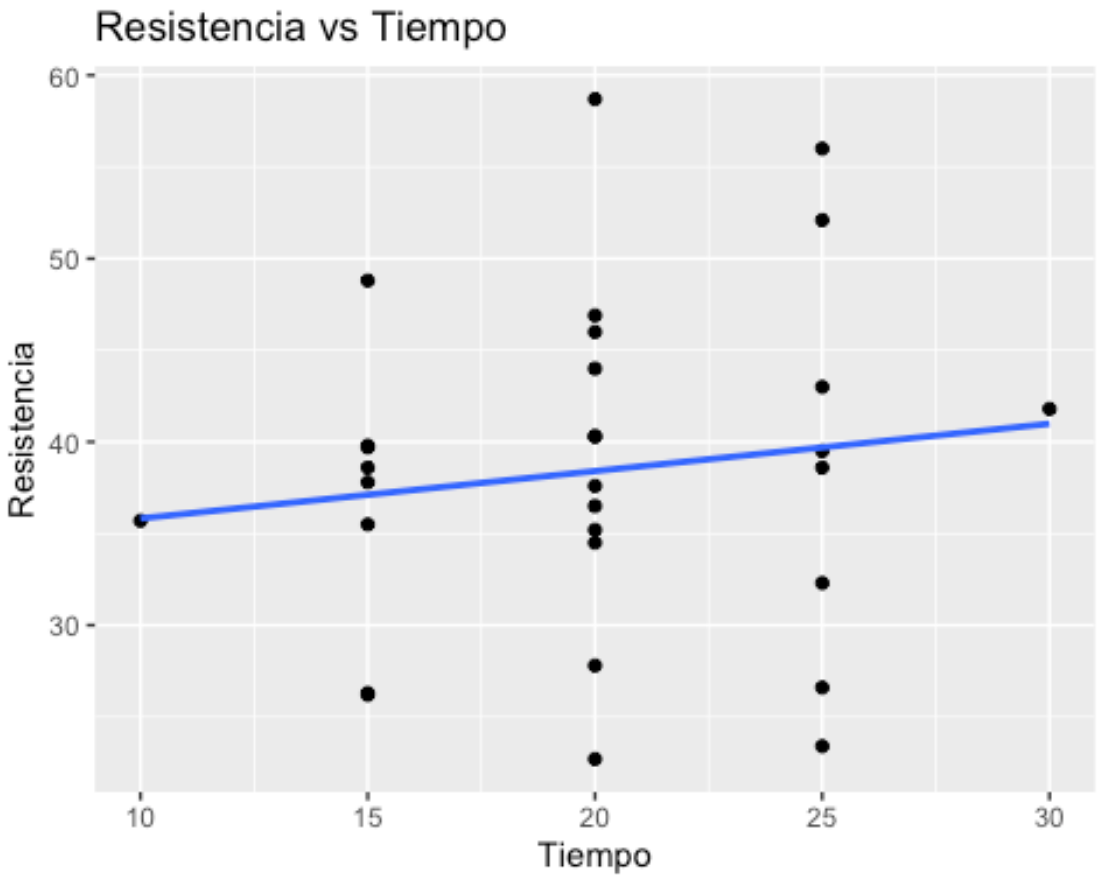
```
## `geom_smooth()` using formula = 'y ~ x'
```

```r
ggplot(M, aes(x = Tiempo, y = Resistencia)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Resistencia vs Tiempo")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```r
# Matriz de correlación
cor_matrix <- cor(M)
print(cor_matrix)
```

```
##                Fuerza  Potencia Temperatura    Tiempo Resistencia
## Fuerza      1.0000000 0.0000000   0.0000000 0.0000000   0.1075208
## Potencia    0.0000000 1.0000000   0.0000000 0.0000000   0.7594185
## Temperatura 0.0000000 0.0000000   1.0000000 0.0000000   0.3293353
## Tiempo      0.0000000 0.0000000   0.0000000 1.0000000   0.1312262
## Resistencia 0.1075208 0.7594185   0.3293353 0.1312262   1.0000000
```



2. Encuentra el mejor modelo de regresión que explique la variable Resistencia.

Analiza el modelo basándote en la significancia del modelo.

Economía de las variables

Un modelo más simple con menos variables, que explique una cantidad comparable de la variabilidad en la variable dependiente, se considera más económico.

```
# Modelo completo con todas las variables
modelo_completo <- lm(Resistencia ~ Fuerza + Potencia + Temperatura + Tiempo,
data = M)

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

# Modelo 3: Excluir las variables Tiempo y Potencia
modelo_3 <- lm(Resistencia ~ Fuerza + Temperatura, data = M)
```

Significación global (Prueba para el modelo)

```
# Significación global del modelo completo
summary(modelo_completo)$fstatistic

##    value    numdf    dendf
## 15.60004  4.00000 25.00000

# Significación global del modelo 2
summary(modelo_2)$fstatistic

##    value    numdf    dendf
## 19.91157  3.00000 26.00000

# Significación global del modelo 3
summary(modelo_3)$fstatistic

##     value     numdf     dendf
##  1.841301  2.000000 27.000000
```

Significación individual (Prueba para cada βi)

```
# Significación individual del modelo completo
summary(modelo_completo)$coefficients

##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -37.4766667 13.09964183 -2.860892 8.412416e-03
## Fuerza        0.2116667  0.21057361  1.005191 3.244356e-01
## Potencia      0.4983333  0.07019120  7.099655 1.928265e-07
## Temperatura   0.1296667  0.04211472  3.078892 4.991622e-03
## Tiempo        0.2583333  0.21057361  1.226808 2.313237e-01

# Significación individual del modelo 2
summary(modelo_2)$coefficients

##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) -32.3100000 12.52409564 -2.5798270 1.588861e-02
## Fuerza        0.2116667  0.21260900  0.9955678 3.286360e-01
## Potencia      0.4983333  0.07086967  7.0316874 1.818400e-07
## Temperatura   0.1296667  0.04252180  3.0494163 5.218346e-03

# Significación individual del modelo 3
summary(modelo_3)$coefficients

##              Estimate  Std. Error   t value   Pr(>|t|)
## (Intercept) 5.0650000 18.95636449 0.2671926 0.79135124
## Fuerza      0.2116667  0.35539713 0.5955779 0.55641713
## Temperatura 0.1296667  0.07107943 1.8242503 0.07920071
```

Variación explicada por el modelo

```
# R-cuadrado ajustado del modelo completo
summary(modelo_completo)$adj.r.squared
```

```
## [1] 0.6681928
```

```
# R-cuadrado ajustado del modelo 2
summary(modelo_2)$adj.r.squared
```

```
## [1] 0.6617473
```

```
# R-cuadrado ajustado del modelo 3
summary(modelo_3)$adj.r.squared
```

```
## [1] 0.05483897
```

Conclusiones de esta sección

Economía de las variables: Modelo completo: Incluye las cuatro variables (Fuerza, Potencia, Temperatura, Tiempo). Modelo 2: Excluye Tiempo, manteniendo Fuerza, Potencia y Temperatura. Modelo 3: Solo incluye Fuerza y Temperatura.

Significación global (Prueba para el modelo) Modelo completo: F-statistic = 15.60004 Modelo 2: F-statistic = 19.91157 [el más significativo globalmente (mejor ajuste)] Modelo 3: F-statistic = 1.841301 [no es un buen modelo]

Significación individual (Prueba para cada βi) Modelo completo: - Potencia y Temperatura son variables significativas ($p < 0.05$). - Fuerza y Tiempo no son significativos. Modelo 2: - Potencia y Temperatura son significativas ($p < 0.05$). - Fuerza no es significativa. Modelo 3: - Ninguna de las variables (Fuerza, Temperatura) es significativa ($p > 0.05$).

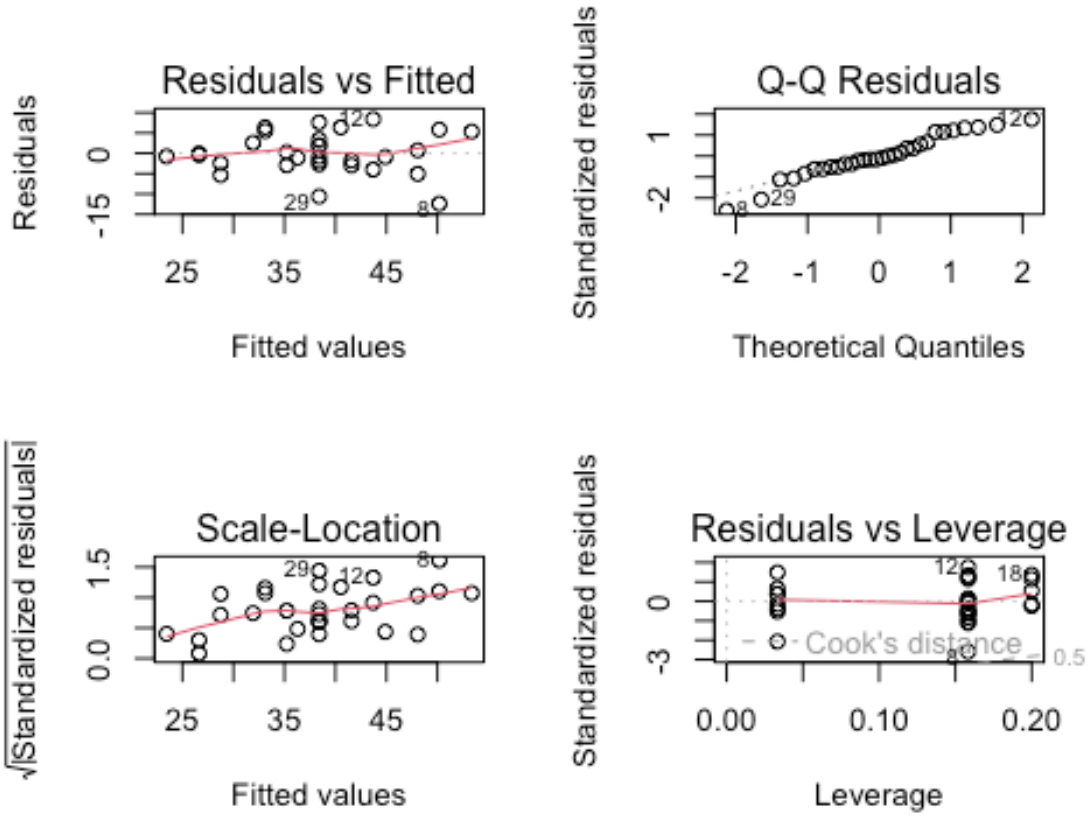
4. Variación explicada por el modelo (R^2 ajustado) Modelo completo: R^2 ajustado = 0.6681928 Modelo 2: R^2 ajustado = 0.6617473 Modelo 3: R^2 ajustado = 0.05483897

El Modelo 2 es una buena opción porque es más económico que el Modelo completo (menos variables), tiene la mejor significación global (mayor F-statistic), y apenas pierde un poco en R^2 ajustado en comparación con el Modelo completo.

3. Analiza la validez del modelo encontrado

Análisis de residuos (homocedasticidad, independencia, etc)

```
# Graficar los residuos del modelo
par(mfrow = c(2, 2))  # Para mostrar varias gráficas en una ventana
plot(modelo_2)
```



No multicolinealidad de Xi

```
# Instalar e importar la librería lmtest si no la tienes
# install.packages("lmtest")
library(lmtest)

# Prueba de Breusch-Pagan
bptest(modelo_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_2
## BP = 5.8003, df = 3, p-value = 0.1217
```

4. Conclusiones sobre el modelo final encontrado

Interpreta en el contexto del problema el efecto de las variables predictoras en la variable respuesta.

4.1 Análisis de residuos (gráficos) - Residuos vs Ajustados: No se observa un patrón claro en los residuos, lo cual es bueno porque indica que no hay problemas evidentes de heterocedasticidad. Los puntos parecen estar distribuidos aleatoriamente en torno a la

línea cero, aunque hay algunos puntos ligeramente alejados. - Normal Q-Q: La mayoría de los puntos siguen la línea diagonal, lo que sugiere que los residuos están distribuidos de manera aproximadamente normal. Sin embargo, hay algunos puntos al final que se desvían, lo que podría indicar leves problemas de normalidad en los extremos. - Scale-Location: Los residuos están distribuidos de manera relativamente homogénea a lo largo de los valores ajustados, lo que apoya la suposición de homocedasticidad. La línea roja es relativamente plana, lo cual es un buen indicador. - Residuos vs Leverage: No se observan puntos con un alto "leverage" o influyentes que puedan tener un impacto excesivo en el modelo. Los puntos están distribuidos de forma homogénea.

4.2. Prueba de Breusch-Pagan (Homocedasticidad) - La prueba de Breusch-Pagan tiene un p-valor de 0.1217, que es mayor a 0.05. Esto indica que no hay evidencia suficiente para rechazar la hipótesis nula de homocedasticidad. En otras palabras, no parece haber un problema de heterocedasticidad en el modelo, y los errores parecen tener varianza constante.

4.3 Interpretación de las variables predictoras (Modelo 2) - Potencia: Es una de las variables más significativas (p-valor muy bajo), lo que indica que tiene un impacto fuerte y positivo en la resistencia al corte. A medida que aumenta la potencia, también aumenta la resistencia al corte. Esto puede interpretarse como que el aumento de la potencia durante el experimento tiene un efecto directo en mejorar la resistencia del material al corte. - Temperatura: También es significativa (p-valor bajo). Su coeficiente positivo indica que un aumento en la temperatura aumenta la resistencia al corte, lo que sugiere que a mayor temperatura, el material podría volverse más resistente al corte en las condiciones del experimento. - Fuerza: No es una variable significativa en este modelo, lo que significa que su influencia sobre la resistencia al corte no es significativa en el contexto de este experimento.

4.4 Conclusiones finales El Modelo 2 se presenta como la opción más apropiada para explicar la Resistencia al corte, ofreciendo un equilibrio entre simplicidad y precisión. Este modelo identifica la Potencia y la Temperatura como los factores más influyentes, con la Potencia ejerciendo un impacto más significativo. La ausencia de heterocedasticidad y multicolinealidad fortalece la fiabilidad del modelo. Su capacidad para predecir cómo los cambios en estos parámetros afectan la resistencia al corte lo convierte en una herramienta valiosa para la optimización de procesos industriales que buscan mejorar esta propiedad del material.

5. Consulta los apoyos sobre regresión para revisar códigos (opcional)