

## A2-Regresión Múltiple

Ozner Leyva

2024-09-18

### 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.

```
library(tidyverse)

## -- Attaching core tidyverse packages -----
tidyverse 2.0.0 --
## v dplyr      1.1.2      v readr      2.1.4
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.4.4      v tibble    3.2.1
## v lubridate  1.9.2      v tidyr     1.3.0
## v purrr      1.0.1
## -- Conflicts -----
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors

library(ggplot2)
library(zoo)

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

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

##      i..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

```
resumen <- summary(M)
print(resumen)
```

##	i..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

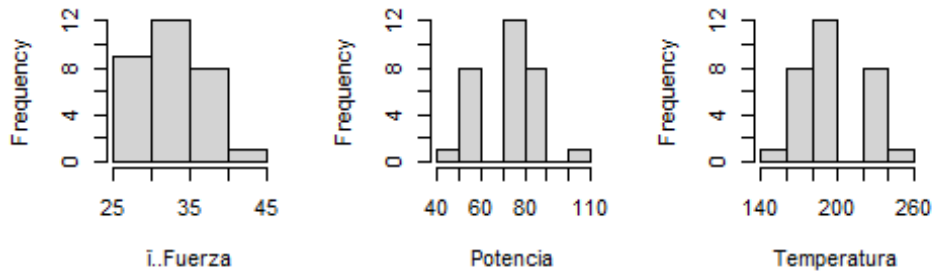
```
desviacion_estandar <- sapply(M, sd)
print(desviacion_estandar)
```

##	i..Fuerza	Potencia	Temperatura	Tiempo	Resistencia
##	4.548588	13.645765	22.742941	4.548588	8.954403

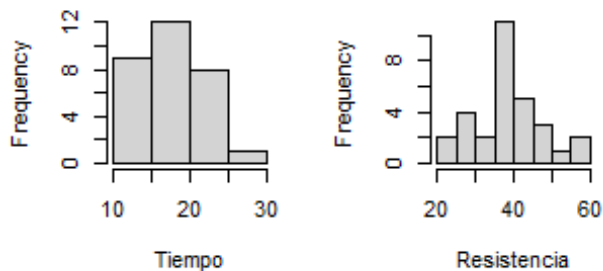
```
par(mfrow=c(2,3))
for(col in names(M)) {
  hist(M[[col]], main=paste("Histograma de", col), xlab=col)
}
```

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

Histograma de i..Fuerza   Histograma de Potenci   Histograma de Temperat



Histograma de Tiempo   Histograma de Resistenci



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

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

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

```
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'
```

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

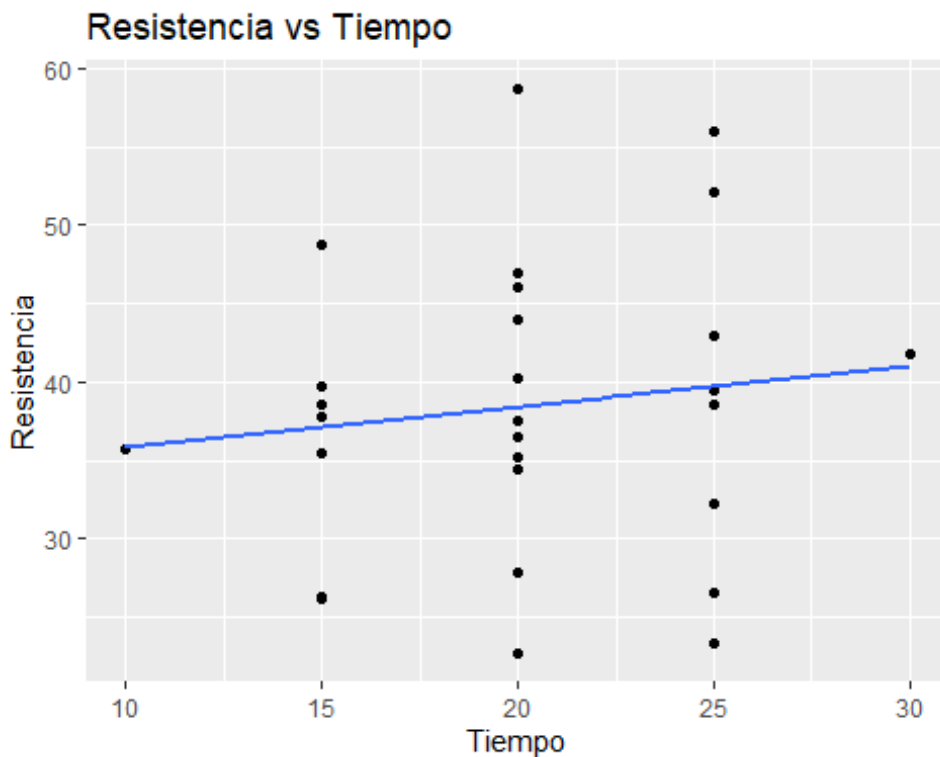
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
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'

cor_matrix <- cor(M)
print(cor_matrix)
```

	ï..Fuerza	Potencia	Temperatura	Tiempo	Resistencia
ï..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 ~ ï..Fuerza + Potencia + Temperatura +
  Tiempo, data = M)
```

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

```
# Modelo 3: Excluir las variables Tiempo y Potencia
modelo_3 <- lm(Resistencia ~ i..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    2.00000 27.00000
```

### Significación individual (Prueba para cada $\beta_i$ )

```
# Significación individual del modelo completo
```

```
summary(modelo_completo)$coefficients
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) -37.476667 13.09964183 -2.860892 8.412416e-03
## i..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
## i..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
```

```
## i..Fuerza    0.2116667    0.35539713  0.5955779  0.55641713
## Temperatura 0.1296667    0.07107943  1.8242503  0.07920071
```

### Variación explicada por el modelo

```
r2_ajustado_completo <- summary(modelo_completo)$adj.r.squared
cat("R-cuadrado ajustado del modelo completo:", r2_ajustado_completo,
"\n\n")
```

```
## R-cuadrado ajustado del modelo completo: 0.6681928
```

*# Modelo 2 (simplificado)*

```
r2_ajustado_modelo2 <- summary(modelo_2)$adj.r.squared
cat("R-cuadrado ajustado del modelo 2:", r2_ajustado_modelo2, "\n\n")
```

```
## R-cuadrado ajustado del modelo 2: 0.6617473
```

*# Modelo 3 (básico)*

```
r2_ajustado_modelo3 <- summary(modelo_3)$adj.r.squared
cat("R-cuadrado ajustado del modelo 3:", r2_ajustado_modelo3, "\n")
```

```
## R-cuadrado ajustado del modelo 3: 0.05483897
```

### Conclusiones de esta sección

Análisis de la eficiencia de los modelos:

Composición de los modelos:

Modelo integral: Incorpora cuatro variables (Fuerza, Potencia, Temperatura y Tiempo). Modelo simplificado: Considera tres variables (Fuerza, Potencia y Temperatura), excluyendo Tiempo. Modelo básico: Se limita a dos variables (Fuerza y Temperatura).

Evaluación de la significación global:

Modelo integral: F-estadístico = 15.60004 Modelo simplificado: F-estadístico = 19.91157 (destaca por su mayor significación global) Modelo básico: F-estadístico = 1.841301 (muestra un ajuste deficiente)

Análisis de la significación individual de las variables:

Modelo integral: • Potencia y Temperatura demuestran significación estadística ( $p < 0.05$ ) • Fuerza y Tiempo carecen de significación  
Modelo simplificado: • Potencia y Temperatura mantienen su significación ( $p < 0.05$ ) • Fuerza no alcanza significación estadística  
Modelo básico: • Ninguna variable (Fuerza, Temperatura) logra significación ( $p > 0.05$ )

Capacidad explicativa de los modelos ( $R^2$  ajustado):

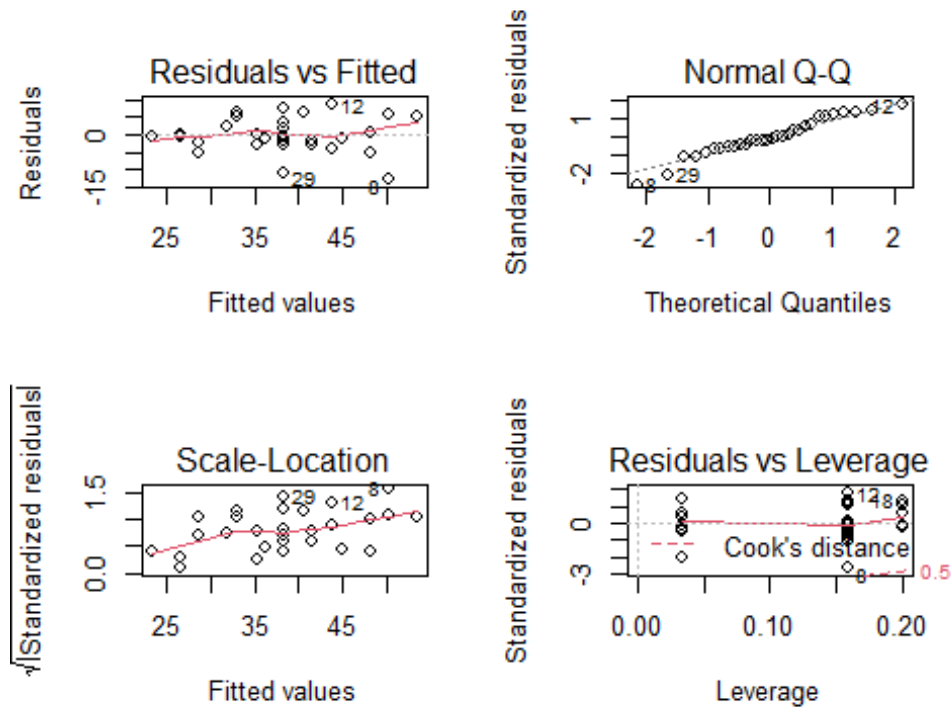
Modelo integral:  $R^2$  ajustado = 0.6681928 Modelo simplificado:  $R^2$  ajustado = 0.6617473 Modelo básico:  $R^2$  ajustado = 0.05483897

Conclusión: El Modelo simplificado es la opción más equilibrada. Nos da una mayor parsimonia en comparación con el Modelo integral al reducir el número de variables, presenta la mejor significación global (F-estadístico más elevado) y solo sacrifica una fracción mínima del  $R^2$  ajustado respecto al Modelo integral. Estos factores lo posiciona como la alternativa más eficiente y robusta para el análisis.

### 3. Analiza la validez del modelo encontrado

**Análisis de residuos (homocedasticidad, independencia, etc)**

```
par(mfrow = c(2, 2))
plot(modelo_2)
```



### No

multicolinealidad de Xi

```
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

Interpretación contextual del modelo predictivo para la resistencia al corte

Evaluación de residuos

a) Residuos vs. Valores ajustados:

Distribución aleatoria de puntos alrededor de la línea cero Ausencia de patrones evidentes, sugiriendo homocedasticidad Presencia de algunos valores atípicos moderados

b) Gráfico Q-Q normal:

Alineación general de puntos con la diagonal, indicando normalidad aproximada Ligeras desviaciones en los extremos, sugiriendo posibles anomalías en la distribución de colas

c) Gráfico de escala-ubicación:

Dispersión homogénea de residuos a lo largo del rango de valores ajustados Tendencia relativamente plana, respaldando la hipótesis de homocedasticidad

d) Residuos vs. Influencia (Leverage):

Ausencia de puntos con alta influencia o leverage Distribución uniforme, sin observaciones que distorsionen significativamente el modelo

Análisis de homocedasticidad (Prueba de Breusch-Pagan)

P-valor: 0.1217 ( $> 0.05$ ) Interpretación: No se rechaza la hipótesis nula de homocedasticidad Conclusión: Los errores presentan varianza constante, validando la suposición de homocedasticidad

Impacto de las variables predictoras (Modelo simplificado)

a) Potencia:

Significancia: Alta (p-valor muy bajo) Efecto: Positivo y sustancial en la resistencia al corte Interpretación: Un incremento en la potencia se asocia directamente con una mayor resistencia al corte del material

b) Temperatura:

Significancia: Considerable (p-valor bajo) Efecto: Positivo en la resistencia al corte Interpretación: Temperaturas más elevadas tienden a aumentar la resistencia al corte del material en las condiciones experimentales

c) Fuerza:

Significancia: No significativa en este modelo Interpretación: La fuerza aplicada no muestra un impacto estadísticamente relevante en la resistencia al corte dentro del contexto experimental

Conclusiones generales

El Modelo simplificado es la opción más adecuada para explicar la variabilidad en la resistencia al corte, ya que nos da un equilibrio óptimo entre parsimonia y capacidad



predictiva. Las variables Potencia y Temperatura se identifican como los factores determinantes, con la Potencia ejerciendo una influencia más pronunciada. La robustez del modelo se puede notar por la ausencia de violaciones significativas en los supuestos de homocedasticidad y multicolinealidad.