# A2 - Regresión Múltiple

Juan Bernal

2024-09-17

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.

```
data = read.csv('AlCorte.csv')
head(data)
```

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

# *Significancia del modelo:

## a. Economía de las variables

La elección de las variables que formaran el modelo de regresión lineal múltiple serán elegidas con base en los procesos de elección de variables hacia delante, atrás y mixto. Además, se utilizarán el criterio de información de Akaike y el criterio de Schwartz para determinar cuál es el mejor modelo (mientras menor sea el criterio, mejor es el modelo).

```
modelo_completo = lm(data$Resistencia ~ data$Potencia+data$Temperatura+data$Tiempo+data$Fuerza) # Modelo de regresión para explicar la resistencia con base en todas las otras variables
modelo_nulo = lm(Resistencia~1, data = data) # Modelo de regresión para explicar la resistencia
```

*Elección de variables hacia DELANTE*

```
Paso_for_aic = step(modelo_nulo, scope = list(lower = modelo_nulo, upper =
modelo_completo), direction = "forward") # Criterio AIC
```

```
## Start:  AIC=132.51
## Resistencia ~ 1
##
##                     Df Sum of Sq     RSS    AIC
## + data$Potencia      1   1341.02  984.24 108.72
## + data$Temperatura   1    252.20 2073.06 131.07
## <none>                            2325.26 132.51
## + data$Tiempo        1     40.04 2285.22 133.99
## + data$Fuerza        1     26.88 2298.38 134.16
##
## Step:  AIC=108.72
## Resistencia ~ data$Potencia
##
##                     Df Sum of Sq    RSS    AIC
## + data$Temperatura   1   252.202 732.04 101.84
## <none>                            984.24 108.72
## + data$Tiempo        1    40.042 944.20 109.47
## + data$Fuerza        1    26.882 957.36 109.89
##
## Step:  AIC=101.84
## Resistencia ~ data$Potencia + data$Temperatura
##
##                Df Sum of Sq    RSS    AIC
## <none>                       732.04 101.84
## + data$Tiempo  1    40.042 692.00 102.15
## + data$Fuerza  1    26.882 705.16 102.72
```

```
n = length(data$Resistencia)
Paso_for_bic = step(modelo_nulo, scope = list(lower = modelo_nulo, upper =
modelo_completo), direction = "forward", k = log(n)) # Criterio BIC
```

```
## Start:  AIC=133.91
## Resistencia ~ 1
##
##                     Df Sum of Sq     RSS    AIC
## + data$Potencia      1   1341.02  984.24 111.52
## + data$Temperatura   1    252.20 2073.06 133.87
## <none>                            2325.26 133.91
## + data$Tiempo        1     40.04 2285.22 136.79
## + data$Fuerza        1     26.88 2298.38 136.97
##
## Step:  AIC=111.52
## Resistencia ~ data$Potencia
##
##                     Df Sum of Sq    RSS    AIC
## + data$Temperatura   1   252.202 732.04 106.04
## <none>                            984.24 111.52
## + data$Tiempo        1    40.042 944.20 113.68
## + data$Fuerza        1    26.882 957.36 114.09
##
## Step:  AIC=106.04
## Resistencia ~ data$Potencia + data$Temperatura
##
##                Df Sum of Sq    RSS    AIC
## <none>                       732.04 106.04
## + data$Tiempo   1    40.042 692.00 107.76
## + data$Fuerza   1    26.882 705.16 108.32
```

La elección de variables hacia delante con ambos criterios AIC y BIC indican que el modelo que mejor explica la Resistencia es aquel que toma como variables independientes Potencia y Temperatura. El menor AIC obtenido fue 101.84 y el BIC fue 106.04.

*Elección de variables hacia ATRÁS*

```
Paso_back_aic = step(modelo_completo, direction = "backward") # Criterio AIC
```

```
## Start:  AIC=102.96
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo +
##     data$Fuerza
##
##                     Df Sum of Sq     RSS    AIC
## - data$Fuerza        1     26.88  692.00 102.15
## - data$Tiempo        1     40.04  705.16 102.72
## <none>                           665.12 102.96
## - data$Temperatura   1    252.20  917.32 110.61
## - data$Potencia      1   1341.02 2006.13 134.08
##
## Step:  AIC=102.15
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo
##
##                     Df Sum of Sq     RSS    AIC
## - data$Tiempo        1     40.04  732.04 101.84
## <none>                           692.00 102.15
## - data$Temperatura   1    252.20  944.20 109.47
## - data$Potencia      1   1341.02 2033.02 132.48
##
## Step:  AIC=101.84
## data$Resistencia ~ data$Potencia + data$Temperatura
##
##                     Df Sum of Sq     RSS    AIC
## <none>                           732.04 101.84
## - data$Temperatura   1     252.2  984.24 108.72
## - data$Potencia      1    1341.0 2073.06 131.07
```

```
Paso_back_bic = step(modelo_completo, direction = "backward", k = log(n)) # Criterio BIC
```

```
## Start:  AIC=109.97
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo +
##     data$Fuerza
##
##                     Df Sum of Sq     RSS    AIC
## - data$Fuerza        1     26.88  692.00 107.76
## - data$Tiempo        1     40.04  705.16 108.32
## <none>                           665.12 109.97
## - data$Temperatura   1    252.20  917.32 116.21
## - data$Potencia      1   1341.02 2006.13 139.69
##
## Step:  AIC=107.76
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo
##
##                     Df Sum of Sq     RSS    AIC
## - data$Tiempo        1     40.04  732.04 106.04
## <none>                           692.00 107.76
## - data$Temperatura   1    252.20  944.20 113.68
## - data$Potencia      1   1341.02 2033.02 136.69
##
## Step:  AIC=106.04
## data$Resistencia ~ data$Potencia + data$Temperatura
##
##                     Df Sum of Sq     RSS    AIC
## <none>                           732.04 106.04
## - data$Temperatura   1     252.2  984.24 111.52
## - data$Potencia      1    1341.0 2073.06 133.87
```

La elección de variables hacia atrás con ambos criterios, AIC y BIC, indican que el modelo que mejor explica la Resistencia es aquel que toma como variables independientes Potencia y Temperatura. El menor AIC obtenido fue 101.84 y el BIC fue 106.04.

*Elección de variables MIXTO*

```
Paso_both_aic = step(modelo_completo, direction="both", trace=1) # Criterio AIC
```

```
## Start:  AIC=102.96
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo +
##     data$Fuerza
##
##                     Df Sum of Sq     RSS    AIC
## - data$Fuerza        1     26.88  692.00 102.15
## - data$Tiempo        1     40.04  705.16 102.72
## <none>                           665.12 102.96
## - data$Temperatura   1    252.20  917.32 110.61
## - data$Potencia      1   1341.02 2006.13 134.08
##
## Step:  AIC=102.15
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo
##
##                     Df Sum of Sq     RSS    AIC
## - data$Tiempo        1     40.04  732.04 101.84
## <none>                           692.00 102.15
## + data$Fuerza        1     26.88  665.12 102.96
## - data$Temperatura   1    252.20  944.20 109.47
## - data$Potencia      1   1341.02 2033.02 132.48
##
## Step:  AIC=101.84
## data$Resistencia ~ data$Potencia + data$Temperatura
##
##                     Df Sum of Sq     RSS    AIC
## <none>                           732.04 101.84
## + data$Tiempo        1     40.04  692.00 102.15
## + data$Fuerza        1     26.88  705.16 102.72
## - data$Temperatura   1    252.20  984.24 108.72
## - data$Potencia      1   1341.02 2073.06 131.07
```

```
Paso_both_bic = step(modelo_completo, direction="both", trace=1, k = log(n)) # Criterio BIC
```

```
## Start:  AIC=109.97
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo +
##     data$Fuerza
##
##                    Df Sum of Sq     RSS    AIC
## - data$Fuerza       1     26.88  692.00 107.76
## - data$Tiempo       1     40.04  705.16 108.32
## <none>                          665.12 109.97
## - data$Temperatura  1    252.20  917.32 116.21
## - data$Potencia     1   1341.02 2006.13 139.69
##
## Step:  AIC=107.76
## data$Resistencia ~ data$Potencia + data$Temperatura + data$Tiempo
##
##                    Df Sum of Sq     RSS    AIC
## - data$Tiempo       1     40.04  732.04 106.04
## <none>                          692.00 107.76
## + data$Fuerza       1     26.88  665.12 109.97
## - data$Temperatura  1    252.20  944.20 113.68
## - data$Potencia     1   1341.02 2033.02 136.69
##
## Step:  AIC=106.04
## data$Resistencia ~ data$Potencia + data$Temperatura
##
##                    Df Sum of Sq     RSS    AIC
## <none>                          732.04 106.04
## + data$Tiempo       1     40.04  692.00 107.76
## + data$Fuerza       1     26.88  705.16 108.32
## - data$Temperatura  1    252.20  984.24 111.52
## - data$Potencia     1   1341.02 2073.06 133.87
```

La elección de variables mixta con ambos criterios, AIC y BIC, indican que el modelo que mejor explica la Resistencia es aquel que toma como variables independientes Potencia y Temperatura. El menor AIC obtenido fue 101.84 y el BIC fue 106.04.

*Modelo final*

Dados los resultados anteriores, llegamos a la conclusión de que el modelo de regresión lineal múltiple que mejor explica la Resistencia es aquel que toma como variables indenpendientes Potencia y Temperatura.

```
rl = lm(Resistencia~Potencia+Temperatura, data = data)
summary(rl)
```

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

## b. Significación global (Prueba para el modelo)

Hipótesis del modelo:

- $H_0 : \beta = 0$ El modelo no es significante
- $H_1 : \beta \neq 0$ El modelo es significante

Dado un valor de significancia estándar de $\alpha = 0.05$ y que el p-value del modelo es 1.67e-07, contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que el modelo es significante.

## c. Significación individual (Prueba para cada $\beta i$)

Hipótesis de variables:

- $H_0 : \beta_0 = \beta_1 = 0$
- $H_1 : \exists \beta_i \neq 0$

Dado un valor de significancia estándar de $\alpha = 0.05$ y los p-values de los coeficientes del modelo, contamos con suficiente evidencia para rechazar las hipótesis iniciales, es decir, los 3 coeficientes son significativos.

## d. Variación explicada por el modelo

La varianza de la Resistencia es explicada en un 68.52% por el modelo de regresión lineal múltiple que explica la Resistencia con base en la Potencia y Temperatura.

# 2. Analiza la validez del modelo encontrado:

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

1. Normalidad de los residuos

Prueba de hipótesis:

- $H_0$ : Los datos provienen de una población normal
- $H_1$ : Los datos no provienen de una población normal

Regla de decisión: $p - value < \alpha$ se rechaza $H_0$

```
library(nortest)
ad.test(rl$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  rl$residuals
## A = 0.41149, p-value = 0.3204
```

Dado el valor p de la prueba de normalidad, sabemos con un 95% de confianza que los datos del modelo provienen de una población normal. Es decir, no hay suficiente evidencia para rechazar la hipótesis inicial.

## 2. Verificación de media cero

Prueba de hipótesis:

- $H_0 : \mu = 0$
- $H_1 : \mu \neq 0$

Regla de decisión: $p - value < \alpha$ se rechaza $H_0$

```
t.test(rl$residuals)
```

```
##
##  One Sample t-test
##
## data:  rl$residuals
## t = 4.2338e-17, df = 29, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.876076  1.876076
## sample estimates:
##    mean of x
## 3.883612e-17
```

Dado el valor p de la prueba de media cero, sabemos con un 95% de confianza que la media de los residuos es cero. Es decir, no hay suficiente evidencia para rechazar la hipótesis inicial.

## 3. Homocedasticidad

Prueba de hipótesis para homocedasticidad:

- $H_0$ : La varianza de los errores es constante (homocedasticidad)
- $H_1$ : La varianza de los errores no es constante (heterocedasticidad)

Regla de decisión: $p - value < \alpha$ se rechaza $H_0$

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
bptest(rl) # Test de Breusch-Pagan para Homocedasticidad
```

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

Dado un nivel de significancia estándar de 0.05, no contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que la varianza de los errores es constante, es decir, hay homocedasticidad.

## 4. Independencia

Prueba de hipótesis para independencia:

- $H_0$ : Los errores no están correlacionados
- $H_1$ : Los errores están correlacionados

Regla de decisión: $p - value < \alpha$ se rechaza $H_0$

```
dwtest(rl) # Test de Durbin-Watson para Independencia
```

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

Dado un nivel de significancia estándar de 0.05, no contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que los errores no están correlacionados, es decir, hay independencia.

## 5. Linealidad

Prueba de hipótesis para linealidad:

- $H_0$ : No hay términos omitidos que indican linealidad
- $H_1$ : Hay una especificación errónea en el modelo que indica no linealidad

Regla de decisión: $p - value < \alpha$ se rechaza $H_0$

```
resettest(rl)
```

```
##
##  RESET test
##
## data:  rl
## RESET = 0.79035, df1 = 2, df2 = 25, p-value = 0.4647
```

Dado un nivel de significancia estándar de 0.05, no contamos con suficiente evidencia para rechazar la hipótesis inicial, por lo que no hay términos omitidos que indican linealidad, es decir, hay linealidad

# *No multicolinealidad de Xi

```
cor(data)
```

```
##                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
```

Notemos en la tabla de correlaciones que no existe relación alguna entre la Potencia y Temperatura.

```
library(car)
```

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

```
vif(rl)
```

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

Además, los bajos valores del VIF indican que, efectivamente, no hay multicolinealidad entre las variables independientes del modelo.

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

El mejor modelo encontrado es el que predice la resistencia al corte de acuerdo a la potencia y temperatura, pues explica un 68% de la variación de los datos. Argumentando que es el mejor dado que usa menos variables y sigue dando una buena explicación de la variación de la resistencia al corte. Además, el modelo cumple con todos los supuestos de validez.

4. Consulta los apoyos sobre regresión para revisar códigos:

*Verificación Significancia del Modelo

*Validez del Modelo

*Regresión lineal Múltiple