Actividad: Aprendizaje Automático y Regresión Lineal Simple

María Isabel Moreno Carreño

Índice


```
#Si da error generando el pdf, usa estos comandos:
#install.packages("tinytex")
#tinytex::tlmgr_install("xetex")
#install.packages("skimr") # Descomenta si no lo tienes
#install.packages("lmtest") # Descomenta si no lo tienes
```

```r
# Cargar las librerías necesarias
library(readr)
library(ggplot2)
library(scales)
library(tidyverse)
library(dplyr)
library(car)
library(corrplot)
library(ggpubr)
library(skimr)
library(lmtest)
```

Introducción.

Este trabajo pretende utilizar el ejercicio de *Creación de un modelo de aprendizaje automático* para hacer una análisis estadístico en R, un modelo de regresión lineal simple.

Pasos del análisis de regresión lineal simple

1. **Exploración inicial y análisis de correlación**
   - Visualizar la relación entre `harness_size` y `boot_size` con un gráfico de dispersión.
   - Calcular el coeficiente de correlación de Pearson para determinar la fuerza y dirección de la relación.
2. **Ajuste inicial del modelo de regresión lineal simple**
   - Estimar el modelo con todas las observaciones para obtener una primera aproximación de la relación. Comprobar la linealidad
3. **Evaluación de los supuestos del modelo**
   - Comprobar la normalidad de los residuos (test de Shapiro-Wilk, QQ plot).
   - Evaluar la homocedasticidad (test de Breusch-Pagan).
   - Verificar la independencia de los residuos (test de Durbin-Watson).
   - Comprobar la linealidad con el gráfico de residuos vs. valores ajustados.
4. **Partición del conjunto de datos**
   - Dividir el dataset en conjunto de entrenamiento (80%) y conjunto de prueba (20%) de forma aleatoria y reproducible.
5. **Entrenamiento del modelo**
   - Ajustar el modelo de regresión lineal simple utilizando solo el conjunto de entrenamiento.
6. **Predicción sobre el conjunto de prueba**
   - Predecir los valores de `boot_size` en el conjunto de prueba.
   - Comparar las predicciones con los valores reales.
7. **Validación de los supuestos en el modelo entrenado**

- Revisar que los residuos del modelo ajustado en el conjunto de entrenamiento cumplen los supuestos estadísticos.

8. **Evaluación cuantitativa y visual del modelo**
   - Calcular el Error Cuadrático Medio (MSE) y la Raíz del Error Cuadrático Medio (RMSE).
   - Visualizar las predicciones frente a los valores reales.
   - Interpretar el coeficiente de determinación (R^2) y los coeficientes del modelo para entender el impacto del predictor.

Objetivos.

Determinar si el tamaño del arnés de nieve para perros se puede utilizar para predecir el número de bota para el perro.

1. Hipótesis.

Hipótesis nula: No hay correlación entre el tamaño del arnés y el tamaño de la bota.
Hipótesis alternativa: Hay correlación entre el tamaño del arnés y el tamaño de la bota.

2. Carga de datos.

Creación de variables con los datos a tratar. - Formatos / Tipos de datos: Datos Númericos. - Que escala de medida tienen los datos.

```
# Crear el data frame directamente a partir de los vectores
data <- data.frame(
    boot_size = c(39, 38, 37, 39, 38, 35, 37, 36, 35, 40,
                  40, 36, 38, 39, 42, 42, 36, 36, 35, 41,
                  42, 38, 37, 35, 40, 36, 35, 39, 41, 37,
                  35, 41, 39, 41, 42, 42, 36, 37, 37, 39,
                  42, 35, 36, 41, 41, 41, 39, 39, 35, 39),

    harness_size = c(58, 58, 52, 58, 57, 52, 55, 53, 49, 54,
                     59, 56, 53, 58, 57, 58, 56, 51, 50, 59,
                     59, 59, 55, 50, 55, 52, 53, 54, 61, 56,
                     55, 60, 57, 56, 61, 58, 53, 57, 57, 55,
                     60, 51, 52, 56, 55, 57, 58, 57, 51, 59)
)

# Tipo de datos de cada columna.
glimpse(data)
```

```
## Rows: 50
## Columns: 2
## $ boot_size    <dbl> 39, 38, 37, 39, 38, 35, 37, 36, 35, 40, 40, 36, 38, 39, 4~
## $ harness_size <dbl> 58, 58, 52, 58, 57, 52, 55, 53, 49, 54, 59, 56, 53, 58, 5~
```

```r
class(data$boot_size) # Tipo de dato de la columna boot_size
```

```
## [1] "numeric"
```

```r
class(data$harness_size) # Tipo de dato de la columna harness_size
```

```
## [1] "numeric"
```

```r
str(data)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ boot_size   : num  39 38 37 39 38 35 37 36 35 40 ...
##  $ harness_size: num  58 58 52 58 57 52 55 53 49 54 ...
```

```r
summary(data)
```

```
##    boot_size      harness_size
##  Min.   :35.00   Min.   :49.00
##  1st Qu.:36.00   1st Qu.:53.00
##  Median :38.50   Median :56.00
##  Mean   :38.32   Mean   :55.64
##  3rd Qu.:40.75   3rd Qu.:58.00
##  Max.   :42.00   Max.   :61.00
```

En este caso, no se realiza preparación de los datos, ya que son datos que se han generado directamente en el código, y no se han cargado desde un fichero.

3. Exploración inicial y análisis de correlación.

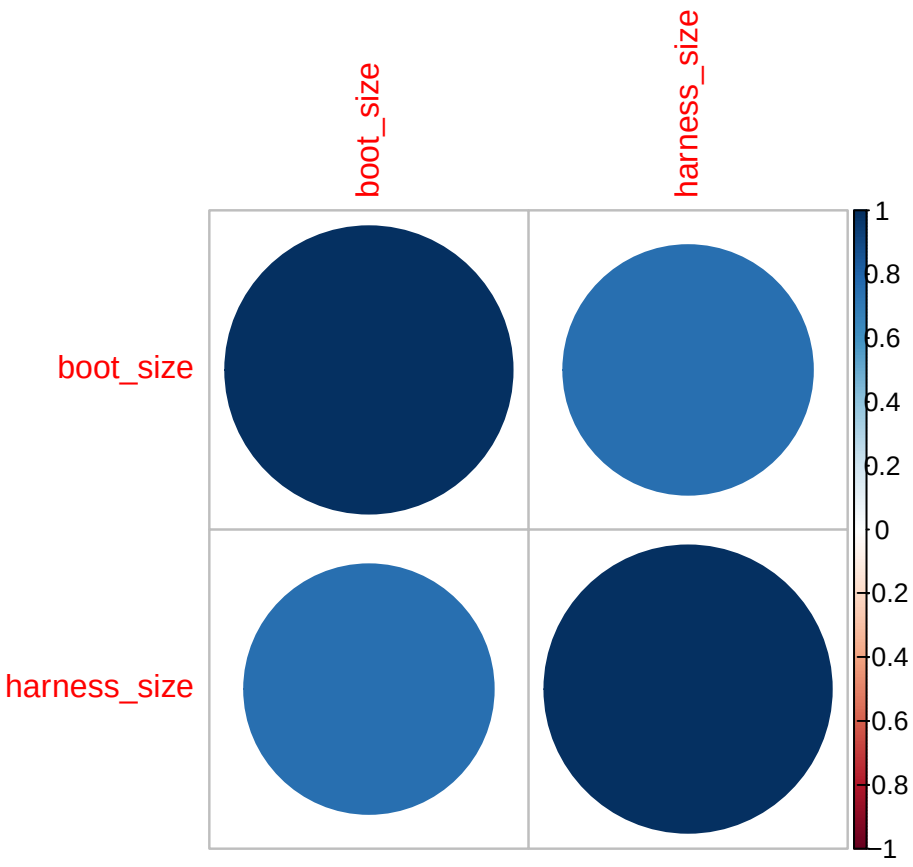
Las variables son cuantitativas, se puede utilizar cor. Se calcula la matriz de correlación lineal.

Lo más importante al principio es calcular la correlación entre las variables, si no hubiera se acepta la hipóteis nula, el tamaño del arnés no puede predecir el tamaño de la bota.

```r
correlation_matrix<- cor(data[,1:2]) # matriz de correlación)
correlation_df <- as.data.frame(as.table(correlation_matrix)) # Convertir a data frame
correlation_matrix
```

```
##              boot_size harness_size
## boot_size    1.0000000    0.7543121
## harness_size 0.7543121    1.0000000
```

```r
corrplot(correlation_matrix)
```




```
# Filtrado quitando diagonal y duplicados - Sentido cuando son más variables.
correlation_df_filtered <- correlation_df %>%
filter(Var1 != Var2) %>%
rowwise() %>%
mutate(pair = paste(sort(c(Var1, Var2)), collapse = "_")) %>%
ungroup() %>%
distinct(pair, .keep_all = TRUE) %>%
arrange(desc(Freq))
# Ver resultados
print(correlation_df_filtered)
```

```
## # A tibble: 1 x 4
##   Var1         Var2       Freq pair
##   <fct>        <fct>     <dbl> <chr>
## 1 harness_size boot_size 0.754 boot_size_harness_size
```

Aplicando el siguiente cuadro de la Fuerza de la Correlación: r> 0.7: Correlación fuerte. 0.3 < r <= 0.7: Correlación moderada. r<= 0.3: Correlación débil.

La correlación entre las variables es fuerte. Aplicar las pruebas estadísticas correspondientes para comprobar si la correlación es significativa.

4. Ajuste inicial del modelo de regresión lineal simple.

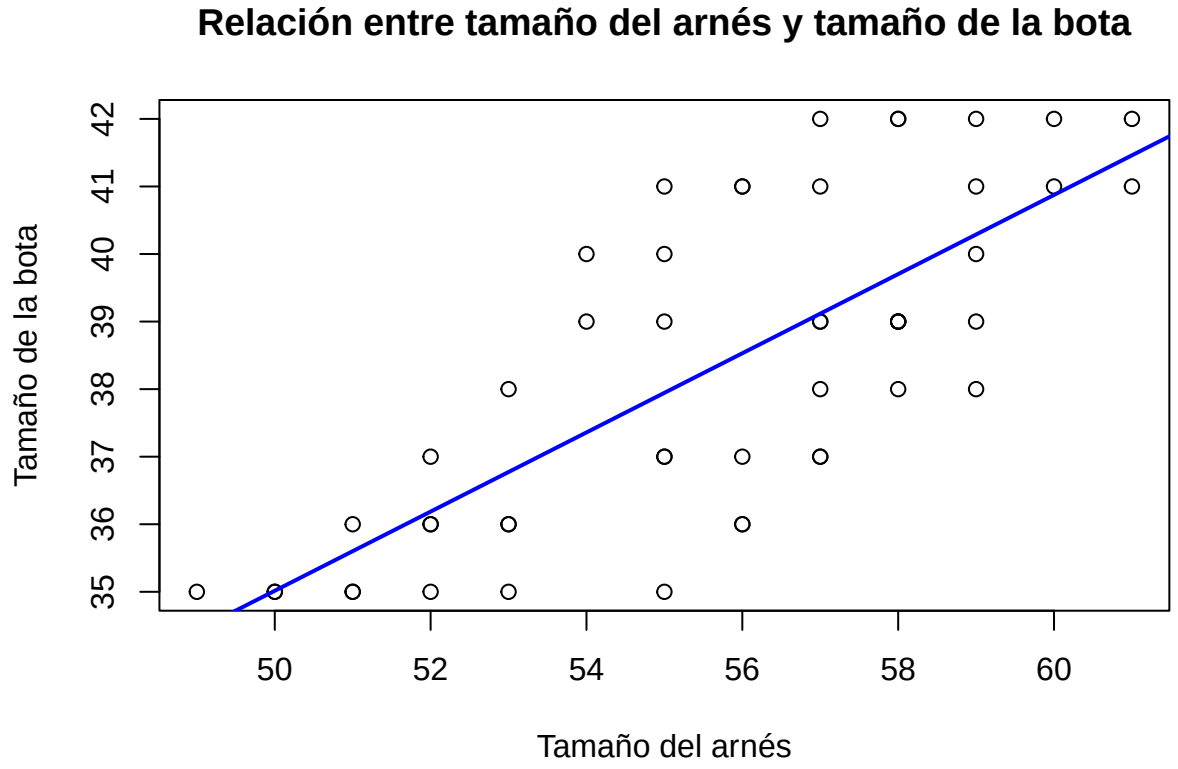
Primero se comprueba si el modelo se ajusta a la linealidad.

```r
modelo <- lm(boot_size ~ harness_size, data = data)



# Comprobar la linealidad del modelo
plot(data$harness_size, data$boot_size,
     main = "Relación entre tamaño del arnés y tamaño de la bota",
     xlab = "Tamaño del arnés",
     ylab = "Tamaño de la bota")

# Añadir la recta de regresión basada en el modelo completo
abline(modelo, col = "blue", lwd = 2)
```




```r
summary(modelo)
```

```
##
## Call:
## lm(formula = boot_size ~ harness_size, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9450 -0.9450 -0.1872  1.1078  3.0550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   5.71911    4.10162   1.394     0.17
## harness_size  0.58593    0.07361   7.960 2.53e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.588 on 48 degrees of freedom
## Multiple R-squared:  0.569,  Adjusted R-squared:   0.56
## F-statistic: 63.37 on 1 and 48 DF,  p-value: 2.529e-10
```

p-valor < 0.001: Rechazamos H (coeficiente $= 0$), y concluimos que harness_size explica significativamente boot_size.

Intercepto - Es la talla estimada de bota cuando el tamaño del arnés es 0 (puede no tener sentido físico, pero es necesario para el modelo matemático).

harness_size 0.586 Por cada unidad que aumenta el tamaño del arnés, la talla de la bota aumenta 0.586 unidades, de forma promedio.

R^2 es 0.569, significa que el modelo explica el 56.9% de la variabilidad en la talla de bota se explica por el tamaño del arnés.

5. Evaluación de los supuestos del modelo.

Para asegurarnos que el modelo es válido, se cumple la normalidad, homocedasticidad e independencia de los residuos.

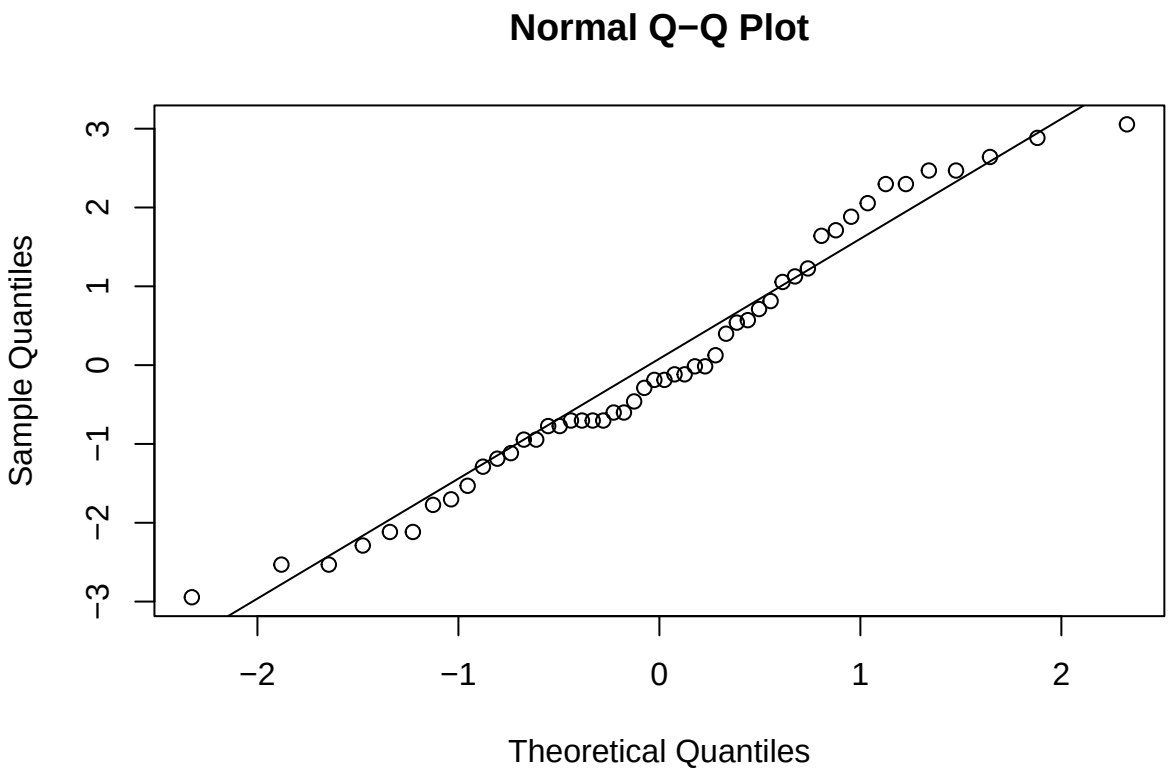
5.1 Normalidad de los residuos.

Comprobar la normalidad de los residuos para la validez del modelo.

```
shapiro.test(residuals(modelo))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo)
## W = 0.96725, p-value = 0.1784
```

```
qqnorm(residuals(modelo)); qqline(residuals(modelo))
```



Como p > 0.05, no se rechaza la hipótesis nula de normalidad de los residuos.

Esto significa que los residuos del modelo pueden considerarse normales, lo cual es un requisito clave para la validez de los intervalos de confianza y pruebas t en regresión

5.2 Homocedasticidad de los residuos.

Estudia si la varianza es constante.

```
bptest(modelo)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 1.0313, df = 1, p-value = 0.3099
```

Como p > 0.05, no se rechaza la hipótesis nula de homocedasticidad.

Esto significa que la varianza de los residuos es constante (no hay heterocedasticidad).

Condición favorable para la validez del modelo de regresión lineal.

5.3 Independencia de los residuos.

```
dwtest(modelo)
```

```
##
##  Durbin-Watson test
##
## data:  modelo
## DW = 1.7172, p-value = 0.1532
## alternative hypothesis: true autocorrelation is greater than 0
```

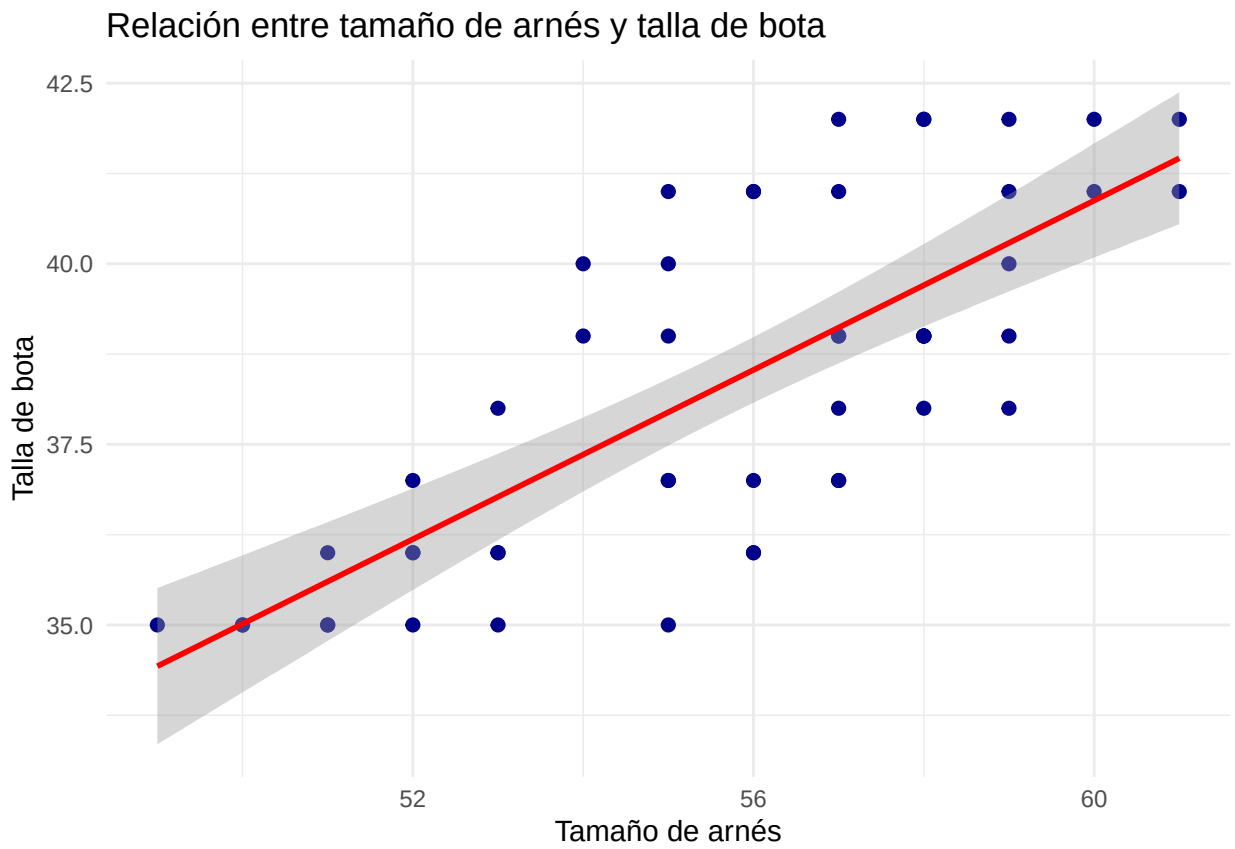
Como p > 0.05, no hay evidencia estadística para rechazar la hipótesis nula de independencia de residuos.

Esto indica que no hay autocorrelación significativa positiva en los residuos.

```
library(ggplot2)

ggplot(data, aes(x = harness_size, y = boot_size)) +
  geom_point(color = "darkblue", size = 2) +             # Puntos observados
  geom_smooth(method = "lm", se = TRUE, color = "red") + # Recta de regresión + IC
  labs(
    title = "Relación entre tamaño de arnés y talla de bota",
    x = "Tamaño de arnés",
    y = "Talla de bota"
  ) +
  theme_minimal()
```

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



6. Resumen de resultados.

El modelo de regresión lineal simple muestra que el tamaño del arnés de nieve para perros puede predecir significativamente el tamaño de las botas.

La correlación entre las variables es fuerte, y el modelo explica el 56.9% de la variabilidad en la talla de bota, con un nivel de significancia $p < 0.001$. El coeficiente de regresión indica que por cada unidad que aumenta el tamaño del arnés, la talla de la bota aumenta en promedio 0.586 unidades.

Los residuos cumplen con los supuestos de normalidad, homocedasticidad e independencia, lo que valida el modelo. El tamaño del arnés es un predictor útil para determinar el tamaño de las botas de nieve para perros.

Por tanto, se concluye que el tamaño del arnés es un predictor fiable para estimar la talla de las botas de nieve para perros.

Se recomienda incluir otras variables relacionadas para aumentar el porcentaje de explicación del modelo, como la raza del perro, el peso o la edad, que pueden aportar mayor precisión al modelo de predicción.

7. Preparación de los datos y Entrenamiento del modelo.

El objetivo es dividir los datos en un conjunto de entrenamiento y otro de prueba, para entrenar el modelo y luego predecir el tamaño de las botas en el conjunto de prueba.

```r
# Dividir los datos en conjunto de entrenamiento y prueba
set.seed(123) # Para reproducibilidad
train_indices <- sample(1:nrow(data), size = 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]


# Entrenar el modelo de regresión lineal
modelo_entrenado <- lm(boot_size ~ harness_size, data = train_data)
# Resumen del modelo entrenado
summary(modelo_entrenado)
```

```
##
## Call:
## lm(formula = boot_size ~ harness_size, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9386 -0.8285 -0.1519  1.0946  3.0614
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.40010    4.35289    1.47     0.15
## harness_size  0.57343    0.07855    7.30 9.69e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.561 on 38 degrees of freedom
## Multiple R-squared:  0.5837, Adjusted R-squared:  0.5728
## F-statistic: 53.29 on 1 and 38 DF,  p-value: 9.694e-09
```

El siguiente paso es predecir los valores con los datos que no han sido usados para el entrenamiento.

```r
# Predecir en el conjunto de prueba
predicciones <- predict(modelo_entrenado, newdata = test_data)

# Comparar las predicciones con los valores reales

resultados <- data.frame(
  Real = test_data$boot_size,
  Predicho = predicciones
)
print(resultados)
```

```
##    Real Predicho
## 1    39 39.65892
## 2    38 39.65892
## 6    35 36.21835
## 16   42 39.65892
## 23   37 37.93864
## 34   41 38.51206
## 35   42 41.37920
```

```
## 38   37 39.08549
## 44   41 38.51206
## 47   39 39.65892
```

7.1 Evaluación de los supuestos del modelo entrenado.

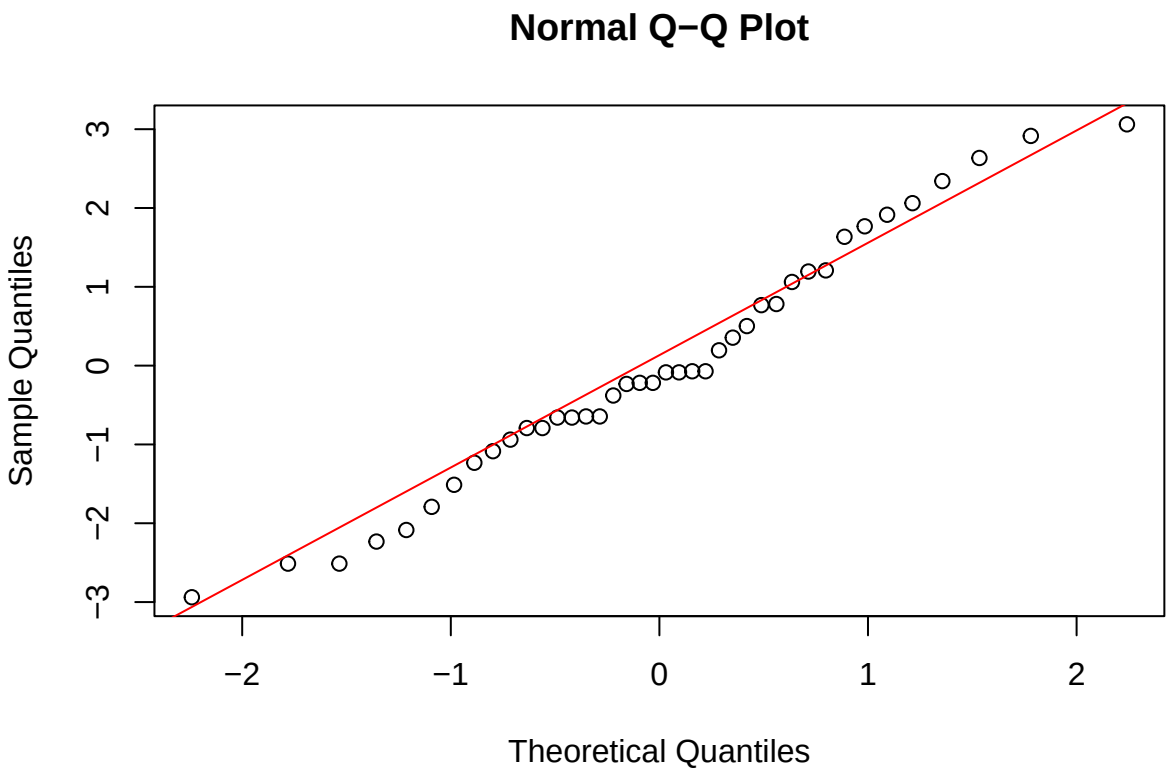
En este paso se comprueba los supuestos del modelo lineal en el modelo entrenado.

- Normalidad de los residuos.
- Homocedasticidad de los residuos.
- Independencia de los residuos.

```
# Comprobar la normalidad de los residuos
shapiro.test(residuals(modelo_entrenado))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo_entrenado)
## W = 0.97659, p-value = 0.5648
```

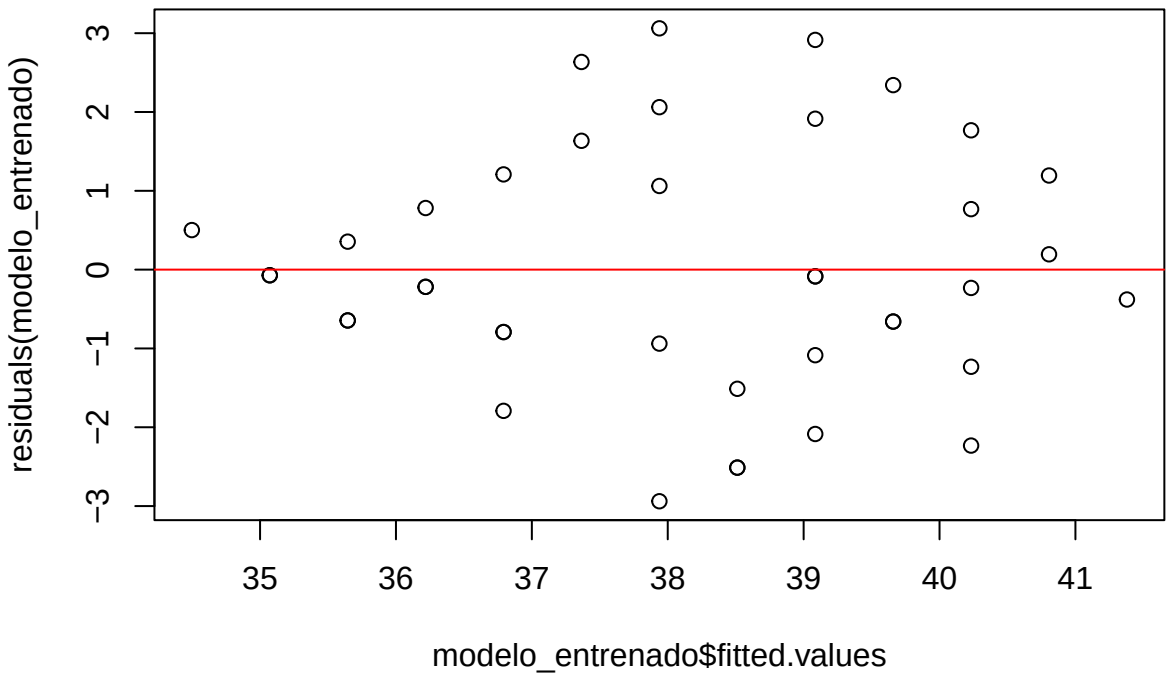
```
# visualización de la normalidad.
qqnorm(residuals(modelo_entrenado))
qqline(residuals(modelo_entrenado), col = "red")
```




```r
# Comprobar la homocedasticidad de los residuos
bptest(modelo_entrenado)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_entrenado
## BP = 1.2055, df = 1, p-value = 0.2722
```

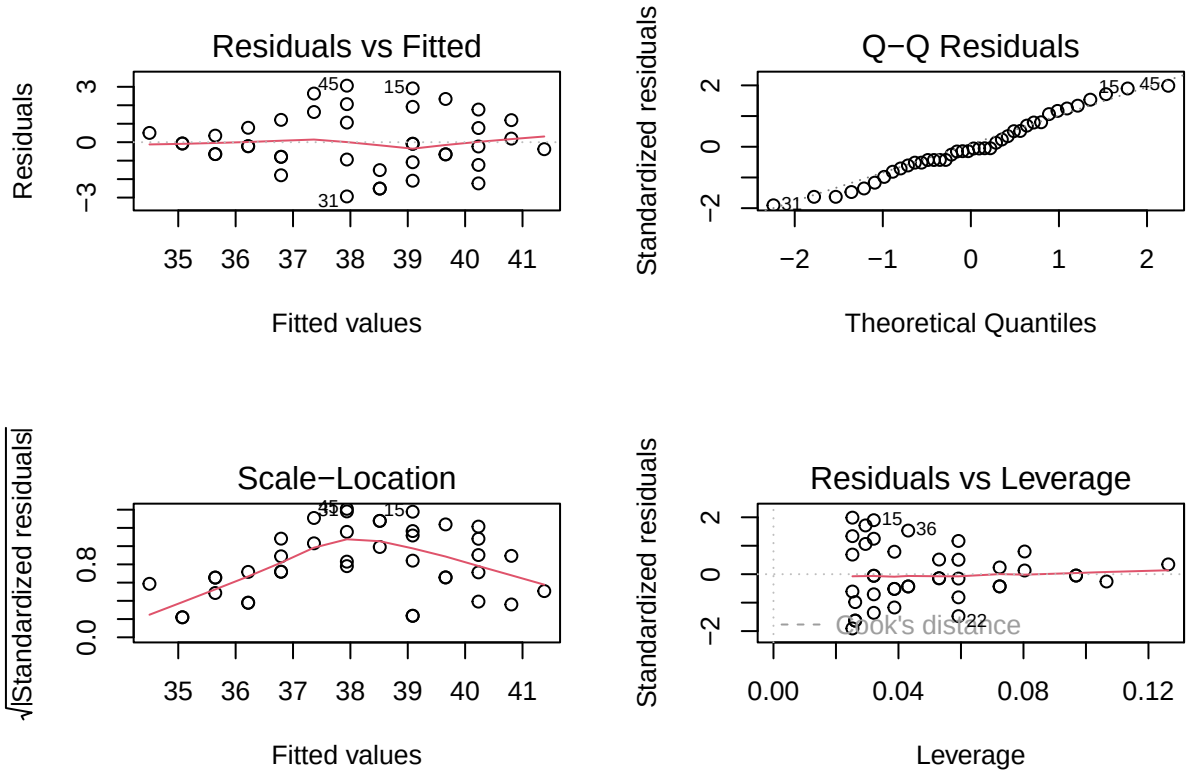
```r
# Visualización
plot(modelo_entrenado$fitted.values, residuals(modelo_entrenado))
abline(h = 0, col = "red")
```




```
# Comprobar la independencia de los residuos
dwtest(modelo_entrenado)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_entrenado
## DW = 2.2338, p-value = 0.7776
## alternative hypothesis: true autocorrelation is greater than 0
```

```
par(mfrow = c(2, 2)) # Para mostrar dos gráficos en una fila
plot(modelo_entrenado)
```



Shapiro - Wilk: Normalidad.

El p-valor > 0.05, no se rechaza la hipótesis nula $\rightarrow$ los residuos pueden considerarse normales.

Breusch-Pagan: Homocedasticidad.

El p-valor > 0.05, no se rechaza la hipótesis nula $\rightarrow$ los residuos tienen varianza constante (no hay heterocedasticidad).

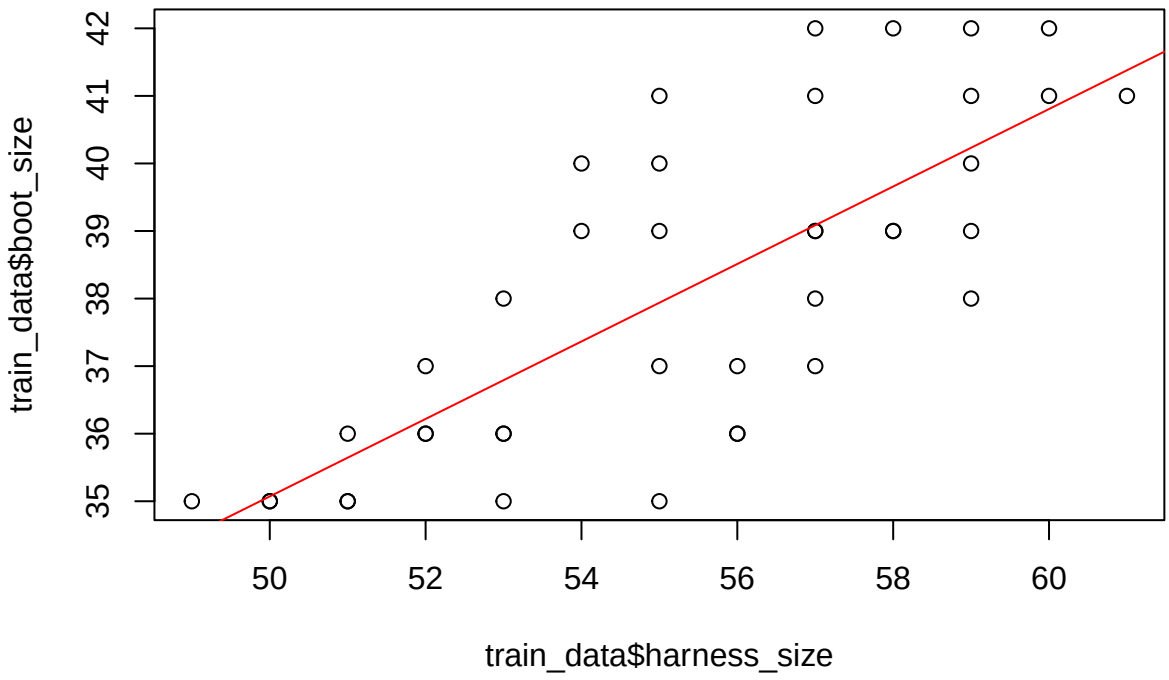
Durbin-Watson: Independencia de los residuos.

El p-valor > 0.05, no hay evidencia estadística para rechazar la hipótesis nula de independencia de residuos $\rightarrow$ no hay autocorrelación significativa positiva en los residuos.

7.2 Linealidad del modelo entrenado

Para comprobar la linealidad de la relación entre las variables, se puede visualizar la relación entre el tamaño del arnés y la talla de bota, junto con la línea de regresión ajustada. Esto nos ayudará a ver si la relación es lineal.

```
plot(train_data$harness_size, train_data$boot_size)
abline(modelo_entrenado, col = "red")
```

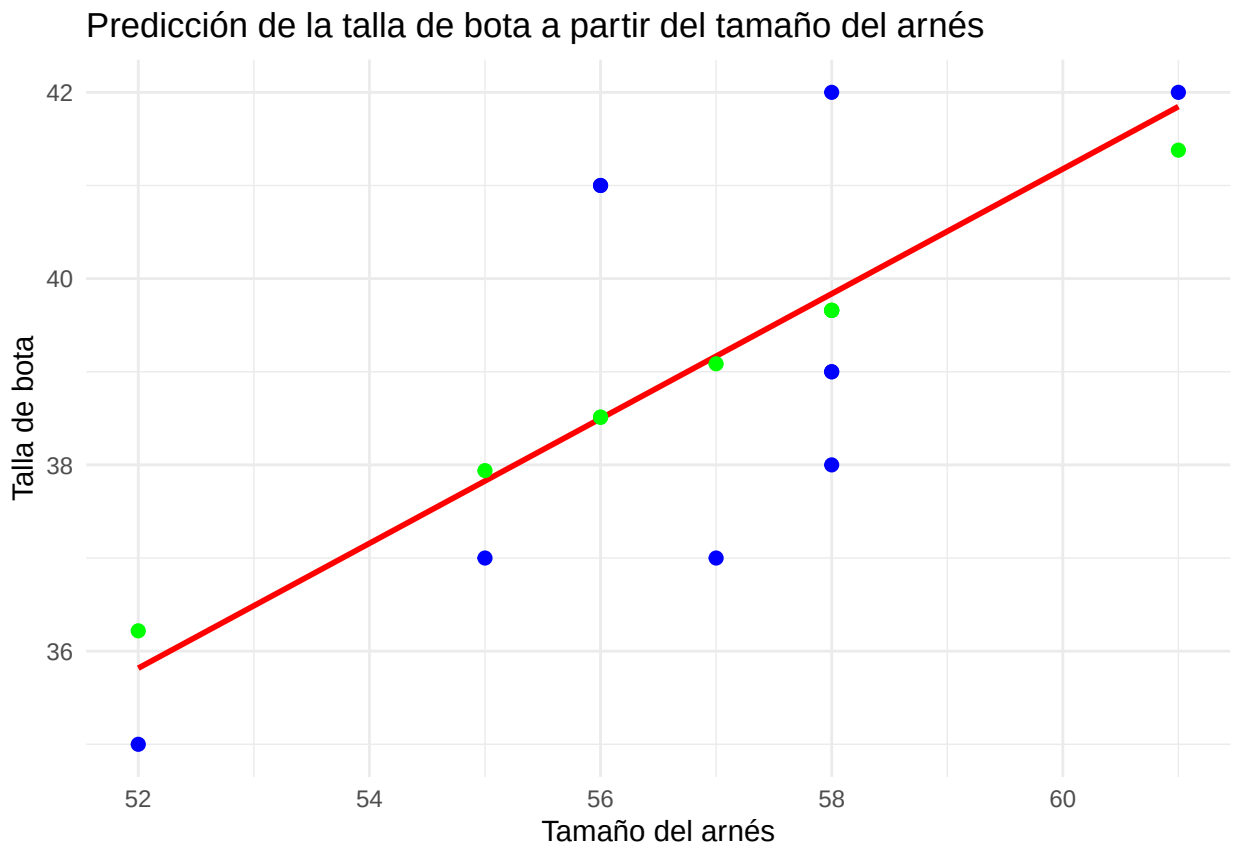


7.3 Visualización predicciones frente a valores reales.

```
library(ggplot2)

ggplot(test_data, aes(x = harness_size, y = boot_size)) +
  geom_point(color = "blue", size = 2) +  # Observaciones reales
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Línea de regresión
  geom_point(aes(y = predicciones), color = "green", size = 2) +  # Predicciones
  labs(
    title = "Predicción de la talla de bota a partir del tamaño del arnés",
    x = "Tamaño del arnés",
    y = "Talla de bota"
  ) +
  theme_minimal()
```

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



7.4 Evaluación cuantitativa del modelo.

El siguiente paso es calcular el error cuadrático medio (MSE), este valor indica la precisión del modelo en las predicciones.

```
predicciones <- predict(modelo_entrenado, newdata = test_data)
residuos_test <- test_data$boot_size - predicciones

mse <- mean(residuos_test^2)
rmse <- sqrt(mse)
print(paste("MSE:", round(mse, 2)))
```

```
## [1] "MSE: 2.86"
```

```
print(paste("RMSE:", round(rmse, 2)))
```

```
## [1] "RMSE: 1.69"
```

El MSE es una medida de la precisión del modelo, cuanto más bajo sea, mejor será el modelo en términos de predicción. Un MSE de 2.86 indica que, en promedio, las predicciones del modelo se desvían de los valores reales en 2.86 unidades al cuadrado. El modelo se desvia en promedio ±1.69 unidades de la talla de bota real.

8. Ejemplo de predicción.

Calcular la talla de bota para un perro con un tamaño de arnés de 52.5

```r
predict(modelo, newdata = data.frame(harness_size = 52.5))
```

```
##        1
## 36.48019
```

9. Conclusiones finales.

El análisis confirma que existe una relación lineal significativa entre el tamaño del arnés y la talla de la bota. El modelo de regresión lineal simple explica el 56.9% de la variabilidad en la talla de bota ($R^2 = 0.569$), y presenta un error cuadrático medio (MSE) de 2.86, lo que se traduce en un error promedio de predicción (RMSE) de aproximadamente 1.69 unidades.

Esto sugiere que, aunque el tamaño del arnés es un buen predictor, las predicciones individuales pueden desviarse en torno a 1.69 unidades respecto al valor real, por lo que sería recomendable incorporar otras variables explicativas si se desea mejorar la precisión del modelo.