7. Regresión lineal

Diego Rodríguez

2023-08-29

Carga de los datos

```
df = read.csv("Estatura-peso_HyM.csv")
summary(df)
```

```
##     Estatura          Peso           Sexo
##  Min.   :1.440   Min.   :37.39   Length:440
##  1st Qu.:1.560   1st Qu.:54.49   Class :character
##  Median :1.610   Median :64.53   Mode  :character
##  Mean   :1.613   Mean   :63.97
##  3rd Qu.:1.660   3rd Qu.:73.22
##  Max.   :1.800   Max.   :90.49
```

La base de datos contiene 3 variables, dos de ellas numericas (la estatura y el peso) y una categórica (Sexo). El rango de valores de la estatura abarca de 1.44 a 1.8 unidades, teniendo una media de 1.613 unidades. El peso va de 37.39 unidades como mínimo y un máximo de 90.49 unidades, con una media de 63.97.

```
table(df['Sexo'])
```

```
## Sexo
##   H   M
## 220 220
```

Por último la variable Sexo se encuentra balanceada en 220 H y 220 M.

```
sum(is.na(df))
```

```
## [1] 0
```

No hay valores nulos

La recta de mejor ajuste

```
dfH = subset(df, df$Sexo == 'H')
dfM = subset(df, df$Sexo == 'M')

matriz_correlacion_H <- cor(dfH[c('Estatura', 'Peso')])

cat("Matriz de correlacion para Hombres:")
```

```
## Matriz de correlacion para Hombres:
```

```
print(matriz_correlacion_H)
```

```
##           Estatura      Peso
## Estatura 1.0000000 0.8468348
```

```
## Peso     0.8468348 1.0000000
```

```
matriz_correlacion_M <- cor(dfM[c('Estatura', 'Peso')])

cat("\n Matriz de correlacion para Mujeres:")
```

```
##
##  Matriz de correlacion para Mujeres:
```

```
print(matriz_correlacion_M)
```

```
##           Estatura      Peso
## Estatura 1.0000000 0.5244962
## Peso     0.5244962 1.0000000
```

Matriz de correlación para Hombres: - La correlación entre "Estatura" y "Peso" para hombres es 0.8468348. Esto indica una correlación positiva bastante fuerte entre la estatura y el peso para los hombres. En otras palabras, a medida que la estatura de los hombres aumenta, tiende a haber un aumento en el peso, y viceversa.

Matriz de correlación para Mujeres: - La correlación entre "Estatura" y "Peso" para mujeres es 0.5244962. Esto indica una correlación positiva más débil entre la estatura y el peso para las mujeres en comparación con los hombres. Aunque todavía existe una relación positiva, no es tan fuerte como en el caso de los hombres.

En el caso de los hombres, la relación es más fuerte, lo que sugiere que la estatura y el peso están más vinculados en este grupo. Para las mujeres, la relación es más débil, lo que indica que la estatura y el peso no están tan estrechamente relacionados como en el caso de los hombres.

```
media_estatura_H <- mean(dfH$Estatura)
desviacion_estandar_estatura_H <- sd(dfH$Estatura)

media_peso_H <- mean(dfH$Peso)
desviacion_estandar_peso_H <- sd(dfH$Peso)

cat("Estatura hombres:\n")
```

```
## Estatura hombres:
```

```
cat("Media:", media_estatura_H, "\n")
```

```
## Media: 1.653727
```

```
cat("Desviación Estándar:", desviacion_estandar_estatura_H, "\n\n")
```

```
## Desviación Estándar: 0.06173088
```

```
cat("Peso hombres:\n")
```

```
## Peso hombres:
```

```
cat("Media:", media_peso_H, "\n")
```

```
## Media: 72.85768
```

```
cat("Desviación Estándar:", desviacion_estandar_peso_H, "\n")
```

```
## Desviación Estándar: 6.900354
```

```
media_estatura_M <- mean(dfM$Estatura)
desviacion_estandar_estatura_M <- sd(dfM$Estatura)
```

```r
media_peso_M <- mean(dfM$Peso)
desviacion_estandar_peso_M <- sd(dfM$Peso)

cat("\n \n Estatura mujeres:\n")
```

```
##
##
##  Estatura mujeres:
```

```r
cat("Media:", media_estatura_M, "\n")
```

```
## Media: 1.572955
```

```r
cat("Desviación Estándar:", desviacion_estandar_estatura_M, "\n\n")
```

```
## Desviación Estándar: 0.05036758
```

```r
cat("Peso mujeres:\n")
```

```
## Peso mujeres:
```

```r
cat("Media:", media_peso_M, "\n")
```

```
## Media: 55.08341
```

```r
cat("Desviación Estándar:", desviacion_estandar_peso_M, "\n")
```

```
## Desviación Estándar: 7.792781
```

Para hombres: - Estatura hombres: - Media: La media de la estatura de los hombres es aproximadamente 1.653727. - Desviación Estándar: La desviación estándar de la estatura de los hombres es aproximadamente 0.06173088. Una desviación estándar relativamente baja sugiere que las estaturas de los hombres tienden a agruparse cerca de la media.

- Peso hombres:
  - Media: La media del peso de los hombres es aproximadamente 72.85768.
  - Desviación Estándar: La desviación estándar del peso de los hombres es aproximadamente 6.900354. Una desviación estándar moderada sugiere que los pesos de los hombres varían, pero no de manera extrema.

Para mujeres: - Estatura mujeres: - Media: La media de la estatura de las mujeres es aproximadamente 1.572955. - Desviación Estándar: La desviación estándar de la estatura de las mujeres es aproximadamente 0.05036758. Una desviación estándar relativamente baja indica que las estaturas de las mujeres tienden a agruparse cerca de la media.

- Peso mujeres:
  - Media: La media del peso de las mujeres es aproximadamente 55.08341.
  - Desviación Estándar: La desviación estándar del peso de las mujeres es aproximadamente 7.792781. Una desviación estándar moderada sugiere que los pesos de las mujeres varían, pero no de manera extrema.

En comparativa, los hombres tienden a pesar hasta 8 gramos más en promedio y una estatura hasta 17 cm más en promedio.

Modelo con sexo

```r
A = lm(df$Peso ~ df$Estatura + df$Sexo)
A
```

```
##
```

```
## Call:
## lm(formula = df$Peso ~ df$Estatura + df$Sexo)
##
## Coefficients:
## (Intercept)  df$Estatura     df$SexoM
##      -74.75        89.26       -10.56
```

Verificacion del modelo

```
summary(A)
```

```
##
## Call:
## lm(formula = df$Peso ~ df$Estatura + df$Sexo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.9505  -3.2491   0.0489   3.2880  17.1243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -74.7546     7.5555  -9.894   <2e-16 ***
## df$Estatura  89.2604     4.5635  19.560   <2e-16 ***
## df$SexoM    -10.5645     0.6317 -16.724   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.381 on 437 degrees of freedom
## Multiple R-squared:  0.7837, Adjusted R-squared:  0.7827
## F-statistic: 791.5 on 2 and 437 DF,  p-value: < 2.2e-16
```

Significancia global: F-statistic = 791.5 un valor muy alto lo que significa que el modelo es muy representativo. Significancia individual: t-values de variables son muy grandes lo que significa que el estadístico de cada variable se encuentra muy alejado de la media, es decir, la zona de rechazo está muy alejada de la media. Por lo tanto, los valores P de las variables también son extremadamente pequeños, por lo que se rechaza la hipotesis nula.

```
b0 = A$coefficients[1]
b1 = A$coefficients[2]
b2 = A$coefficients[3]
```

###Ecuacion general

```
cat("Peso= ", b0, "+", b1, "Estatura", b2, "SexoM")
```

```
## Peso=  -74.7546 + 89.26035 Estatura -10.56447 SexoM
```

###Ecuacion del modelo

```
#Para mujeres
cat("Peso =", b0 + b2, "+", b1, "Estatura", "\n")
```

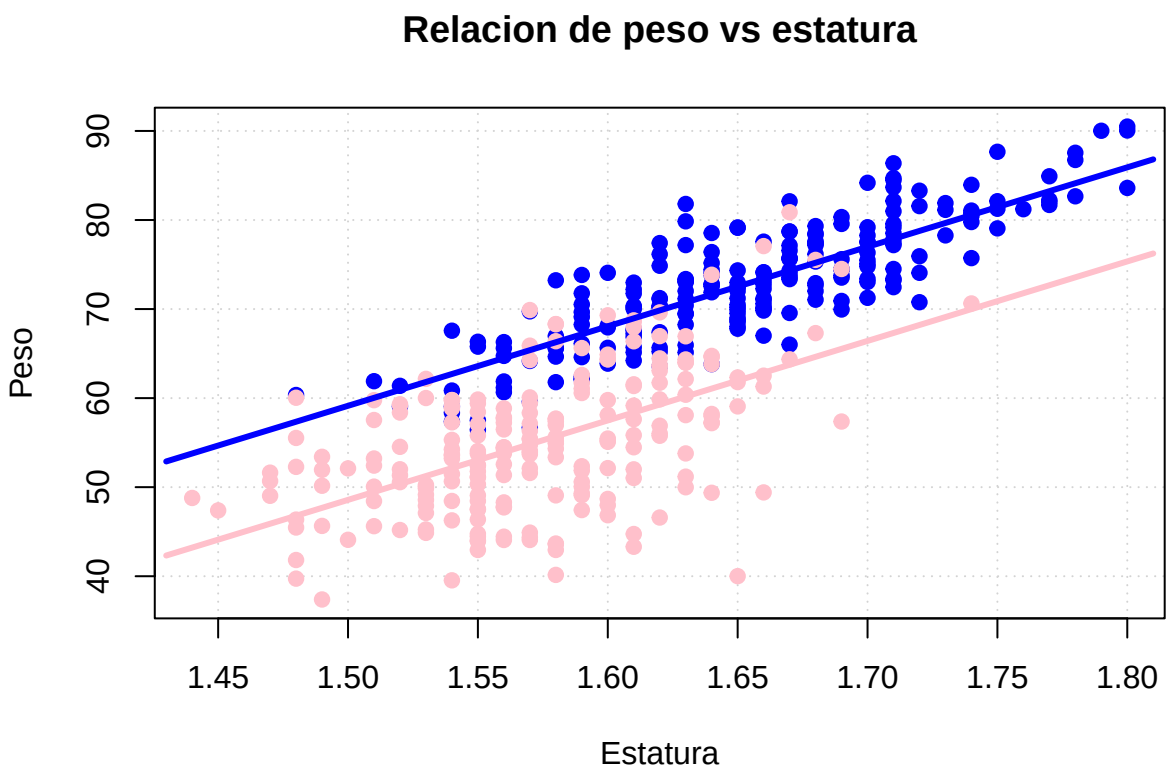
```
## Peso = -85.31907 + 89.26035 Estatura
```

```
#Para hombres
cat("Peso=", b0, "+", b1, "Estatura")
```

```
## Peso= -74.7546 + 89.26035 Estatura
```

Grafica

```
Ym = function(x){b0+b2+b1*x}
Yh = function(x){b0+b1*x}
colores = c("blue", "pink")
plot(df$Estatura, df$Peso, col=colores[factor(df$Sexo)], pch=19,
     ylab = "Peso", xlab = "Estatura", main = "Relacion de peso vs estatura",
     grid())
x = seq(1.43, 1.81, 0.01)
lines(x,Ym(x), col = "pink", lwd = 3)
lines(x, Yh(x), col = "blue", lwd = 3)
```



##

Modelo 2

Buscando interaccion entre las variables.

```
B = lm(df$Peso ~ df$Estatura*df$Sexo)
B
```

```
##
## Call:
## lm(formula = df$Peso ~ df$Estatura * df$Sexo)
##
## Coefficients:
##          (Intercept)           df$Estatura              df$SexoM
##               -83.68                 94.66                 11.12
## df$Estatura:df$SexoM
##               -13.51
```

```
summary(B)
```

```
##
## Call:
```

```
## lm(formula = df$Peso ~ df$Estatura * df$Sexo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -3.1107   0.0204   3.2691  17.9114
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -83.685      9.735  -8.597   <2e-16 ***
## df$Estatura            94.660      5.882  16.092   <2e-16 ***
## df$SexoM               11.124     14.950   0.744    0.457
## df$Estatura:df$SexoM  -13.511      9.305  -1.452    0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.374 on 436 degrees of freedom
## Multiple R-squared:  0.7847, Adjusted R-squared:  0.7832
## F-statistic: 529.7 on 3 and 436 DF,  p-value: < 2.2e-16
```

p value de las variables sexo y estatura muy grande, lo que confirma que no hay efecto de interaccion entre las variables. Nos quedamos con el modelo A.

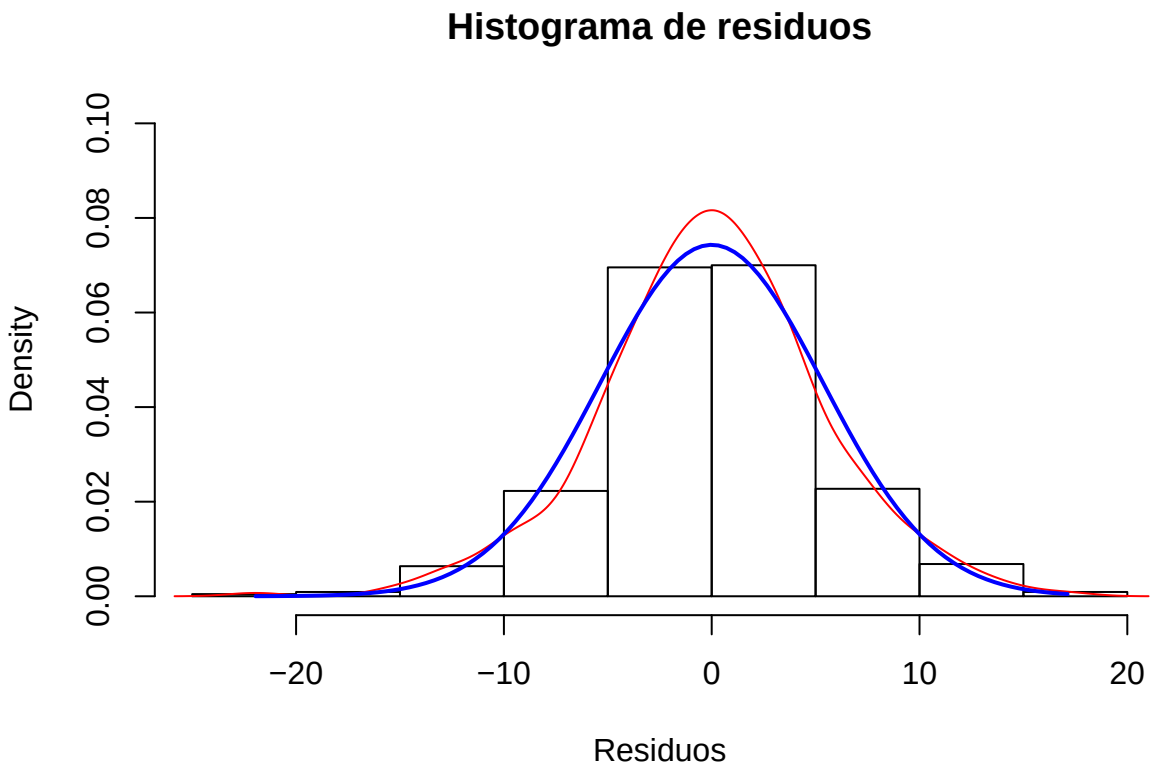
Analisis de residuos

```
shapiro.test(A$residuals)
```

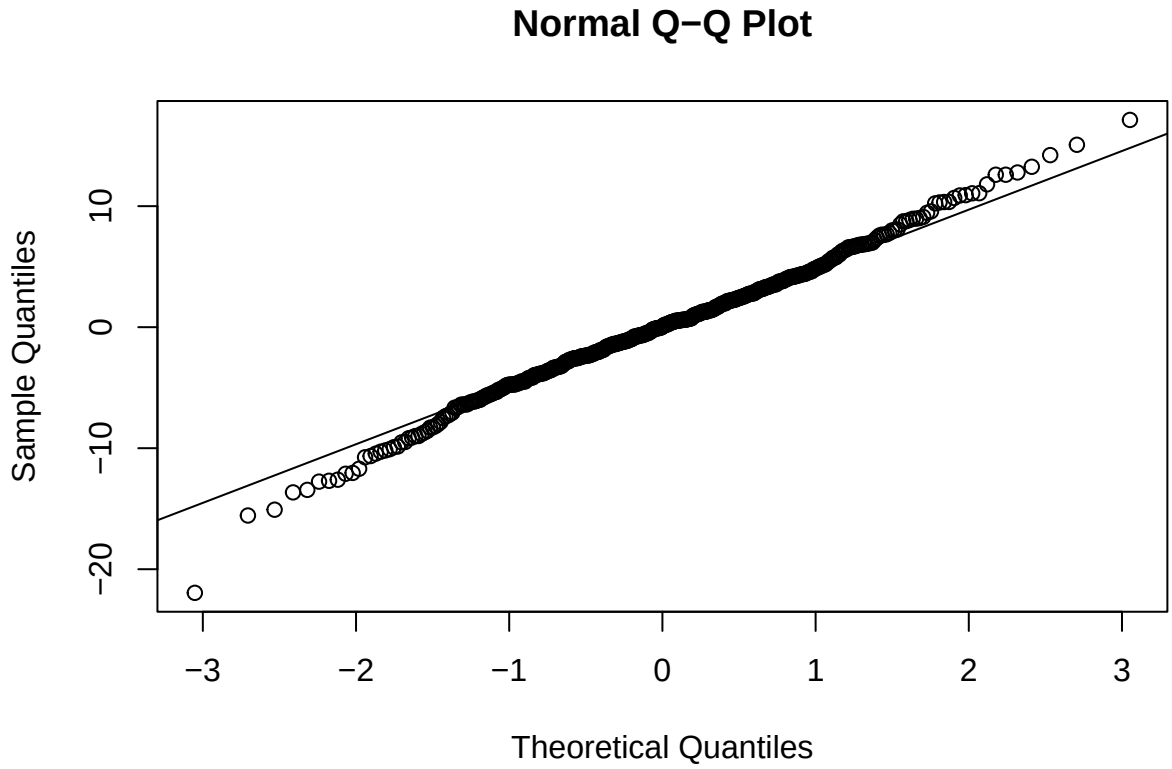
```
##
##  Shapiro-Wilk normality test
##
## data:  A$residuals
## W = 0.99337, p-value = 0.0501
```

p value mayor a 0.03 entonces no rechazamos la hipotesis nula, los residuos siguen una distribución normal, la cual se puede apreciar en las siguientes gráficas, histograma de densidad y qqplot.

```
hist(A$residuals,  freq = FALSE, ylim = c(0,0.1), xlab = "Residuos", col = 0, main = "Histograma de resi
lines(density(A$residuals), col = "red")
curve(dnorm(x, mean = mean(A$residuals), sd = sd(A$residuals)),
      from = min(A$residuals), to = max(A$residuals), add = TRUE,
      col = "blue", lwd =2)
```




```
qqnorm(A$residuals)
qqline(A$residuals)
```



Se confirma normalidad en los residuos.

Media Cero

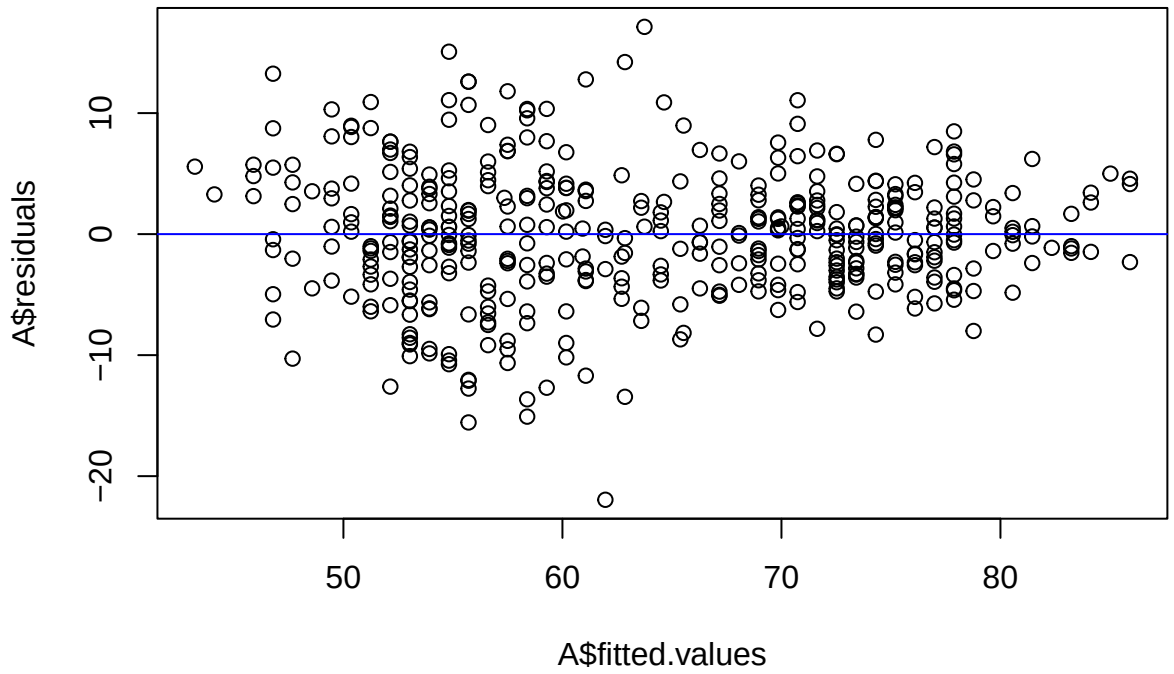
```
t.test(A$residuals)
```

```
##
##  One Sample t-test
##
## data:  A$residuals
## t = 6.8638e-16, df = 439, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5029859  0.5029859
## sample estimates:
##    mean of x
## 1.756605e-16
```

Los resultados de esta prueba indican que no hay evidencia estadística para afirmar que la media de la muestra sea significativamente diferente de cero. Las diferencias observadas en la muestra son tan pequeñas en comparación con la variabilidad de los datos que no son estadísticamente significativas.

Homocedasticidad

```
plot(A$fitted.values,A$residuals)
abline(h=0, col="blue")
```



La gráfica de residuos demuestra que existe homoestecidad ya que los residuos se distribuyen de manera lineal y sin varianza.