10Es_A01571214_Lautaro_Coteja

A01571214 - Lautaro Coteja

2024-09-06

R Markdown

Estadistica - Actividad / Tarea 10: Regresion Lineal

Parte 1

Hipotesis: * $H_0: \beta_1 = 0$ * $H_1: \beta_1 \neq 0$

```
M = read.csv("C:/Users/lauta/Downloads/Estatura-peso_HyM.csv")

# Separar los datos por sexo
MM = subset(M, M$Sexo == "M")
MH = subset(M, M$Sexo == "H")

# Crear un dataframe con las variables por sexo
M1 = data.frame(MH$Estatura, MH$Peso, MM$Estatura, MM$Peso)

head(M1)
```

```
##   MH.Estatura MH.Peso MM.Estatura MM.Peso
## 1        1.61   72.21        1.53   50.07
## 2        1.61   65.71        1.60   59.78
## 3        1.70   75.08        1.54   50.66
## 4        1.65   68.55        1.58   56.96
## 5        1.72   70.77        1.61   51.03
## 6        1.63   77.18        1.57   64.27
```

```
n = 4
d = matrix(NA, ncol = 7, nrow = n)

# Calcular las medidas descriptivas para cada variable
for (i in 1:n) {
  d[i, ] = c(as.numeric(summary(M1[, i])), sd(M1[, i]))
}

m = as.data.frame(d)

row.names(m) = c("H-Estatura", "H-Peso", "M-Estatura", "M-Peso")
names(m) = c("Minimo", "Q1", "Mediana", "Media", "Q3", "Maximo", "Desv Est")
```

```r
# Mostrar el dataframe con las medidas descriptivas
m
```

```
##            Minimo      Q1 Mediana     Media      Q3 Maximo   Desv Est
## H-Estatura   1.48  1.6100   1.650  1.653727  1.7000   1.80 0.06173088
## H-Peso      56.43 68.2575  72.975 72.857682 77.5225  90.49 6.90035408
## M-Estatura   1.44  1.5400   1.570  1.572955  1.6100   1.74 0.05036758
## M-Peso      37.39 49.3550  54.485 55.083409 59.7950  80.87 7.79278074
```

```r
# Calcular la matriz de correlacion para hombres, mujeres y combinado
correlacion_hombres = cor(MH[, c("Estatura", "Peso")])
correlacion_mujeres = cor(MM[, c("Estatura", "Peso")])
correlacion_combinado = cor(M[, c("Estatura", "Peso")])

correlacion_hombres
```

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

```r
correlacion_mujeres
```

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

```r
correlacion_combinado
```

```
##           Estatura      Peso
## Estatura 1.0000000 0.8032449
## Peso     0.8032449 1.0000000
```

```r
# Regresion hombres
A_hombres = lm(MH$Peso ~ MH$Estatura)

# Regresion mujeres
A_mujeres = lm(MM$Peso ~ MM$Estatura)

# Regresion combinada
A_combinado = lm(Peso ~ Estatura, data = M)

# Validacion del modelo para hombres
summary(A_hombres)
```

```
##
## Call:
## lm(formula = MH$Peso ~ MH$Estatura)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.3881 -2.6073 -0.0665  2.4421 11.1883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -83.685      6.663  -12.56   <2e-16 ***
## MH$Estatura   94.660      4.027   23.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 218 degrees of freedom
## Multiple R-squared:  0.7171, Adjusted R-squared:  0.7158
## F-statistic: 552.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

```
# Validacion del modelo para mujeres
summary(A_mujeres)
```

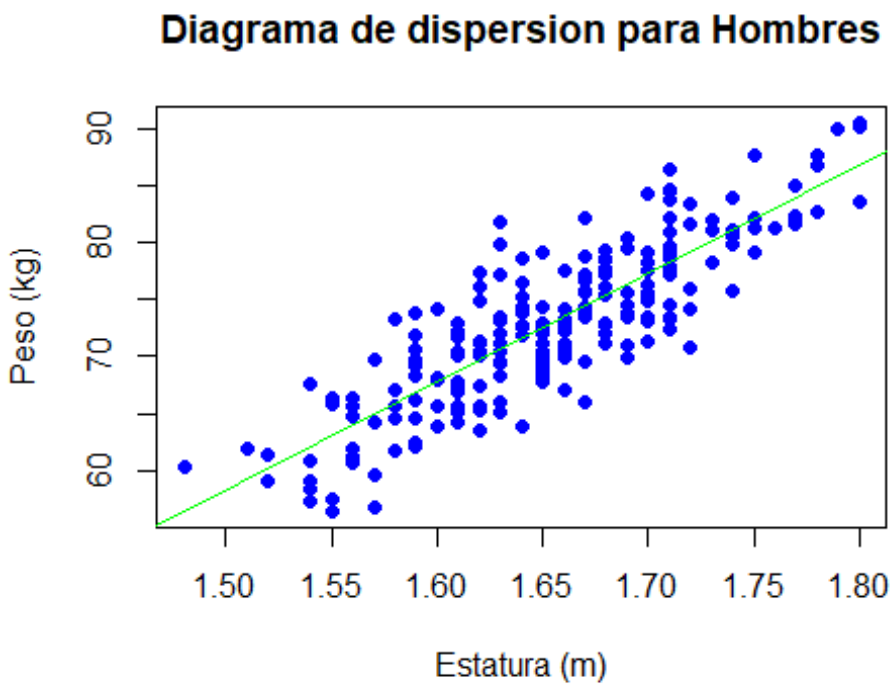
```
##
## Call:
## lm(formula = MM$Peso ~ MM$Estatura)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -4.1942   0.4004   4.2724  17.9114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.560     14.041  -5.168 5.34e-07 ***
## MM$Estatura   81.149      8.922   9.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.65 on 218 degrees of freedom
## Multiple R-squared:  0.2751, Adjusted R-squared:  0.2718
## F-statistic: 82.73 on 1 and 218 DF,  p-value: < 2.2e-16
```

```
# Validacion del modelo combinado
summary(A_combinado)
```

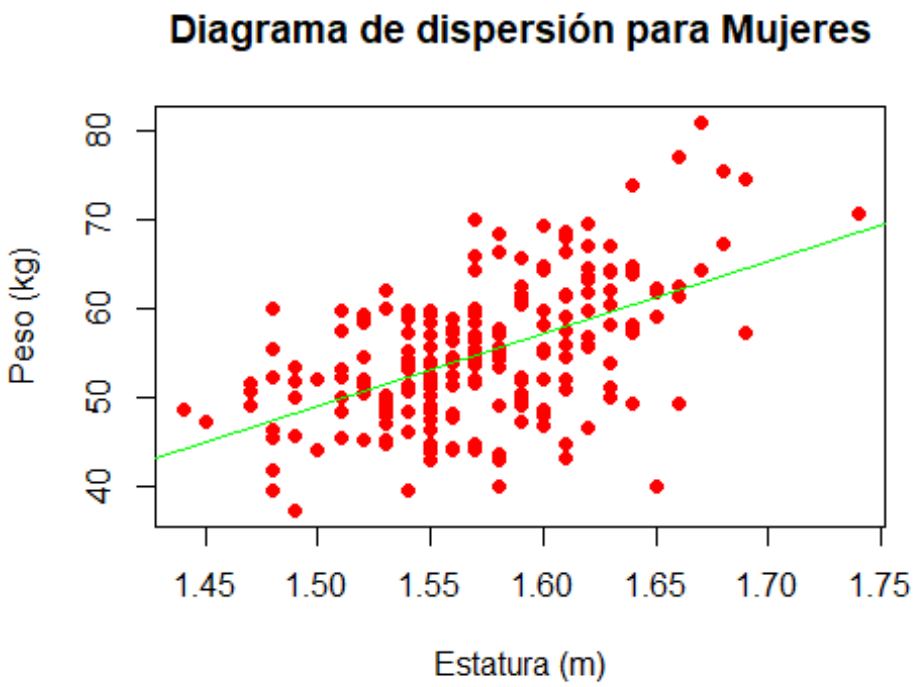
```
##
## Call:
## lm(formula = Peso ~ Estatura, data = M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.8653  -3.7654   0.6706   5.0142  15.6006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -151.883      7.655  -19.84   <2e-16 ***
## Estatura     133.793      4.741   28.22   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.883 on 438 degrees of freedom
## Multiple R-squared:  0.6452, Adjusted R-squared:  0.6444
## F-statistic: 796.5 on 1 and 438 DF,  p-value: < 2.2e-16
```

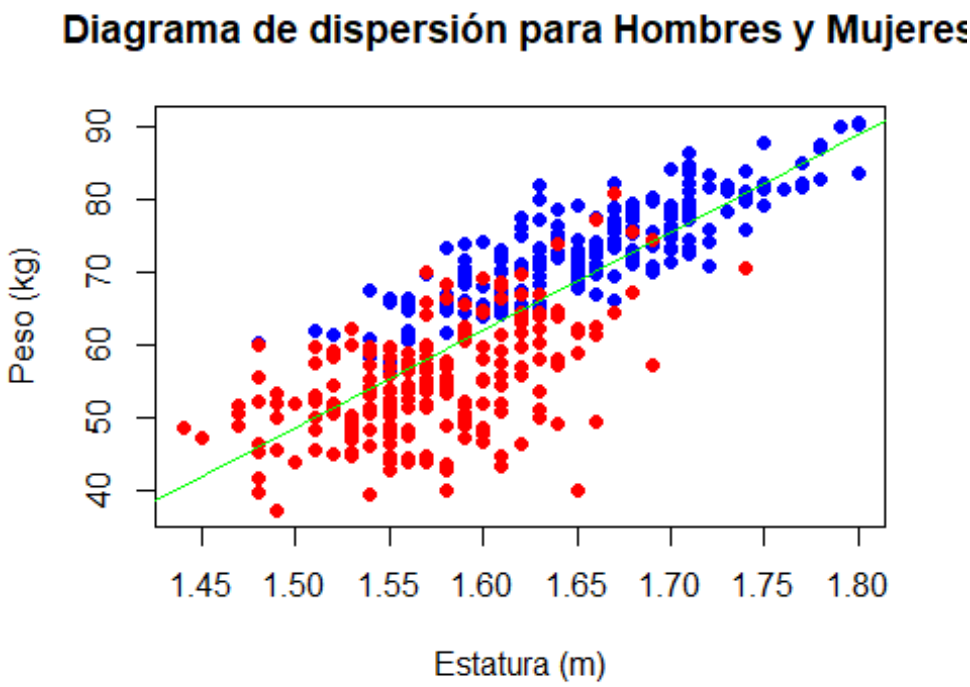
```r
# Diagrama de dispersion con recta de mejor ajuste para hombres
plot(MH$Estatura, MH$Peso, main="Diagrama de dispersion para Hombres",
xlab="Estatura (m)", ylab="Peso (kg)", pch=19, col="blue")
abline(A_hombres, col="green")
```




```r
# Diagrama de dispersion con recta de mejor ajuste para mujeres
plot(MM$Estatura, MM$Peso, main="Diagrama de dispersión para Mujeres",
xlab="Estatura (m)", ylab="Peso (kg)", pch=19, col="red")
abline(A_mujeres, col="green")
```

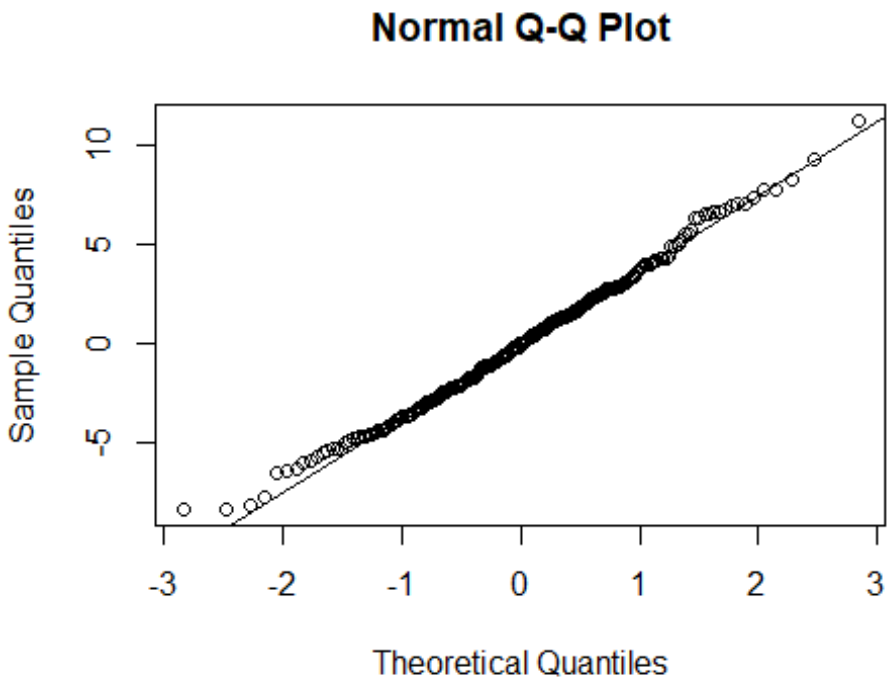



```r
# Diagrama de dispersion con recta de mejor ajuste para ambos sexos
plot(M$Estatura, M$Peso, main="Diagrama de dispersión para Hombres y
Mujeres", xlab="Estatura (m)", ylab="Peso (kg)", pch=19, col=ifelse(M$Sexo ==
"H", "blue", "red"))
abline(A_combinado, col="green")
```

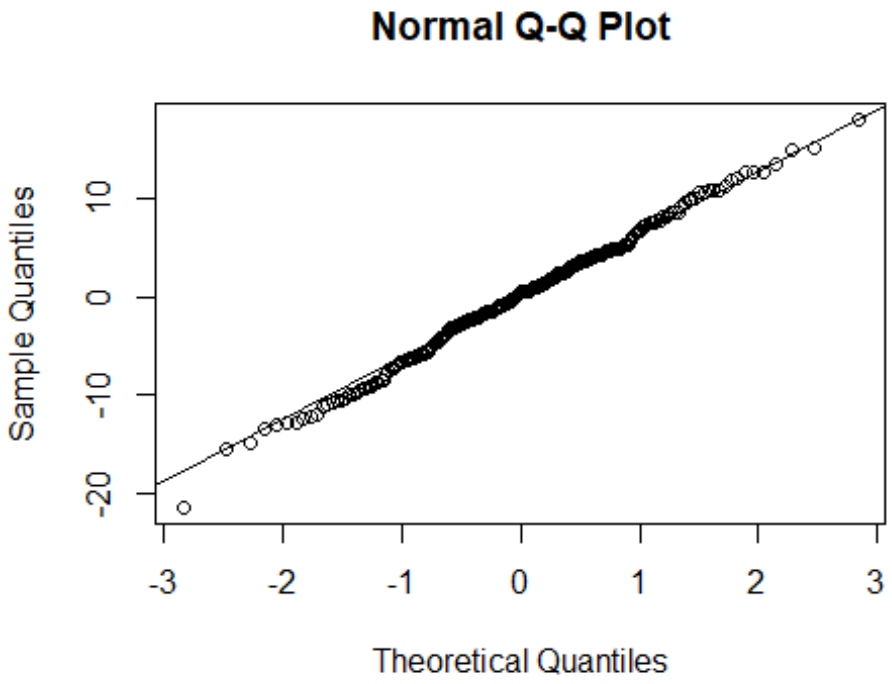



```r
# Prueba de normalidad para los residuos de hombres
shapiro.test(A_hombres$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  A_hombres$residuals
## W = 0.99356, p-value = 0.4597

# Prueba de normalidad para los residuos de mujeres
shapiro.test(A_mujeres$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  A_mujeres$residuals
## W = 0.99659, p-value = 0.9144

# Prueba de normalidad para los residuos del modelo combinado
shapiro.test(A_combinado$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  A_combinado$residuals
## W = 0.97683, p-value = 1.803e-06
```

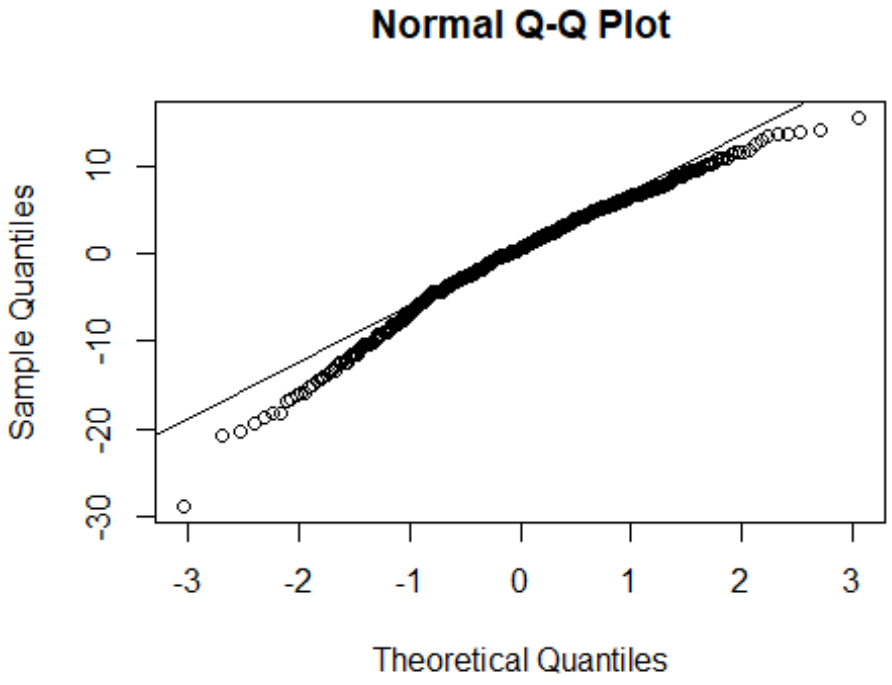
```r
# QQ plot para los residuos de hombres
qqnorm(A_hombres$residuals)
qqline(A_hombres$residuals)
```



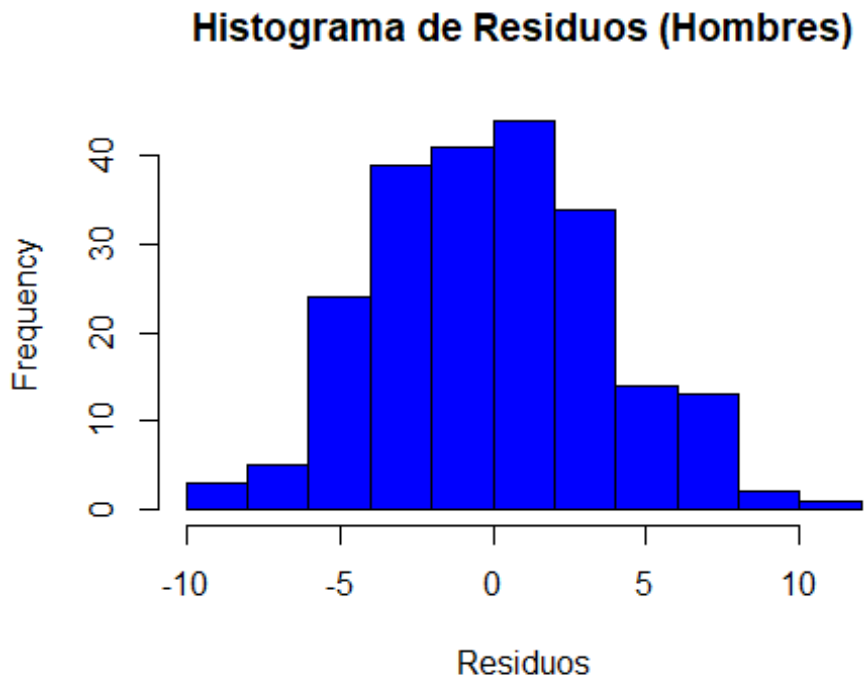

```r
# QQ plot para los residuos de mujeres
qqnorm(A_mujeres$residuals)
qqline(A_mujeres$residuals)
```



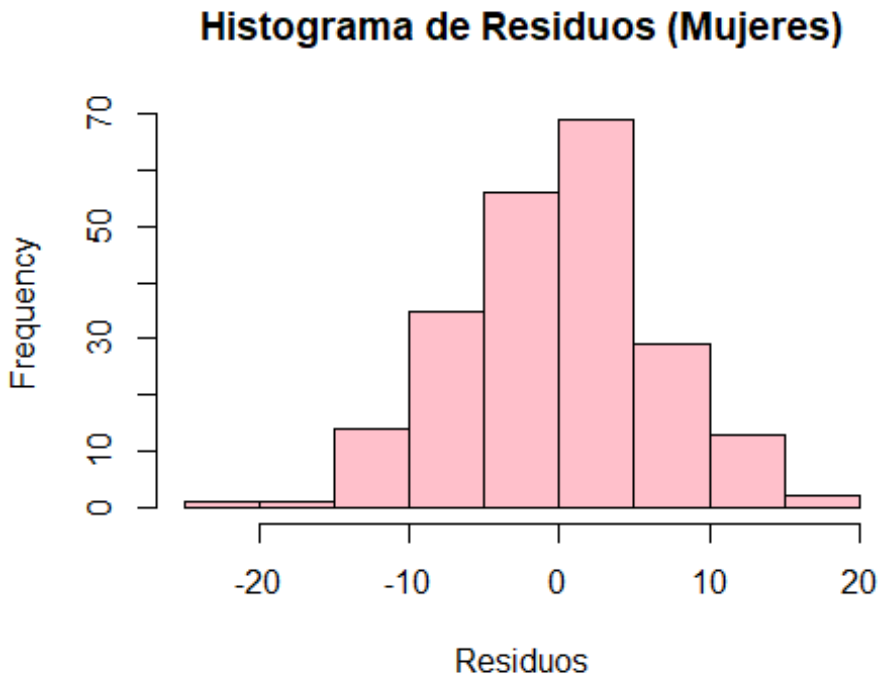

```r
# QQ plot para los residuos del modelo combinado
qqnorm(A_combinado$residuals)
qqline(A_combinado$residuals)
```



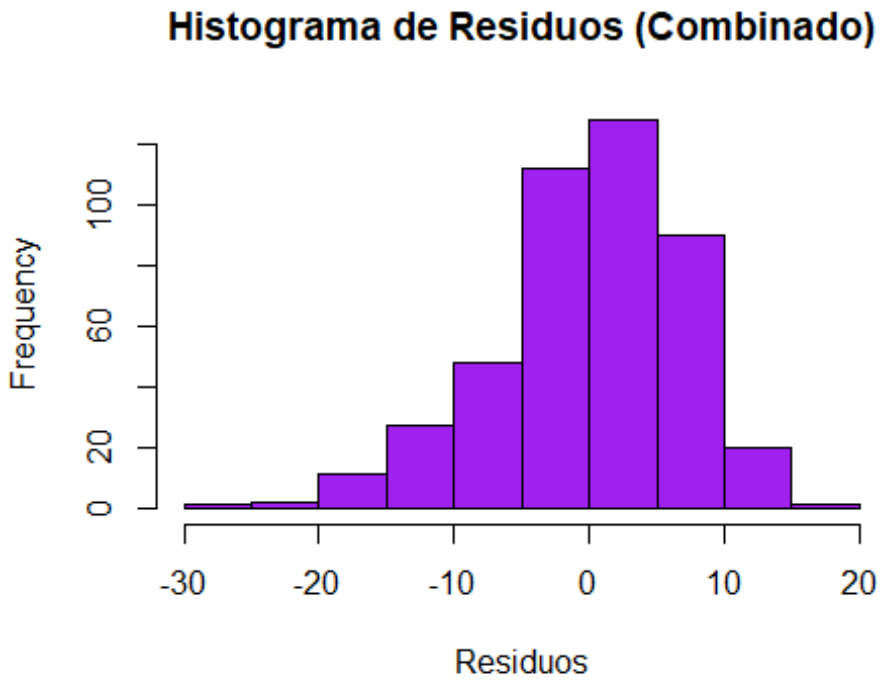
```r
# Histograma para los residuos de hombres
hist(A_hombres$residuals, main = "Histograma de Residuos (Hombres)", xlab =
"Residuos", col = "blue")
```




```r
# Histograma para los residuos de mujeres
hist(A_mujeres$residuals, main = "Histograma de Residuos (Mujeres)", xlab =
"Residuos", col = "pink")
```




```r
# Histograma para los residuos del modelo combinado
hist(A_combinado$residuals, main = "Histograma de Residuos (Combinado)", xlab
= "Residuos", col = "purple")
```



```r
# Prueba t para verificar si la media de los residuos es cero (hombres)
t.test(A_hombres$residuals)
```

```
##
##  One Sample t-test
##
## data:  A_hombres$residuals
## t = 4.5495e-16, df = 219, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.4876507  0.4876507
## sample estimates:
##    mean of x
## 1.125698e-16
```

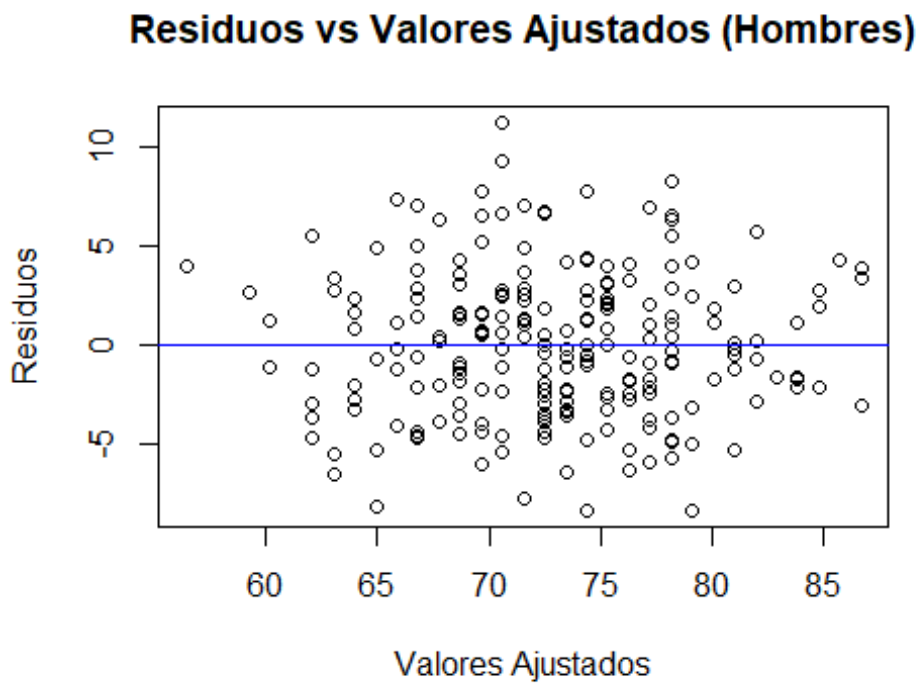
```r
# Prueba t para verificar si la media de los residuos es cero (mujeres)
t.test(A_mujeres$residuals)
```

```
##
##  One Sample t-test
##
## data:  A_mujeres$residuals
## t = -3.9979e-16, df = 219, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.881609  0.881609
## sample estimates:
##     mean of x
## -1.788342e-16
```

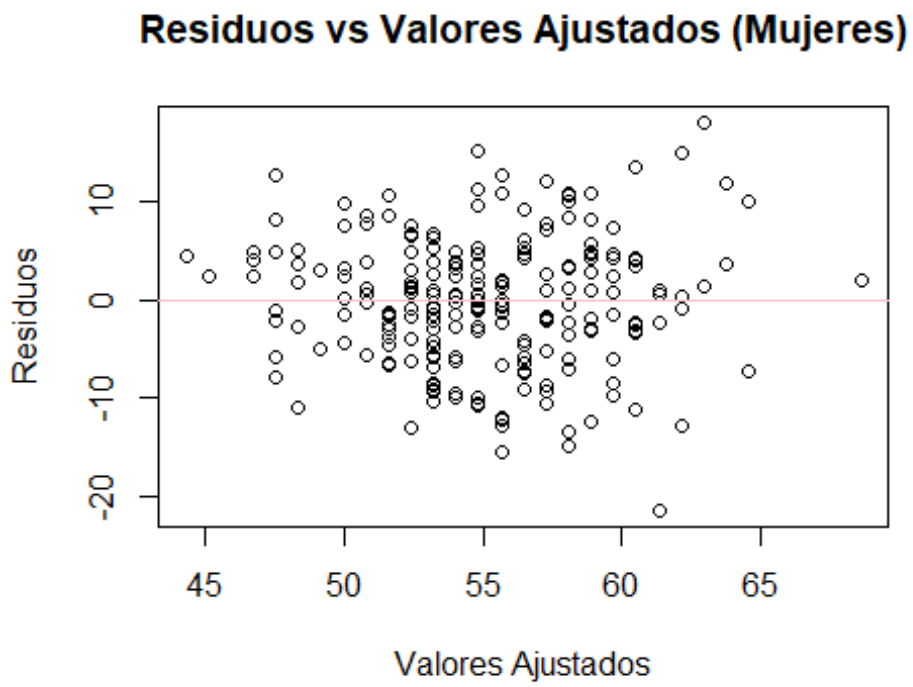
```r
# Prueba t para verificar si la media de los residuos es cero (combinado)
t.test(A_combinado$residuals)
```

```
##
##  One Sample t-test
##
## data:  A_combinado$residuals
## t = 2.7844e-15, df = 439, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.6441362  0.6441362
## sample estimates:
##   mean of x
## 9.12569e-16
```

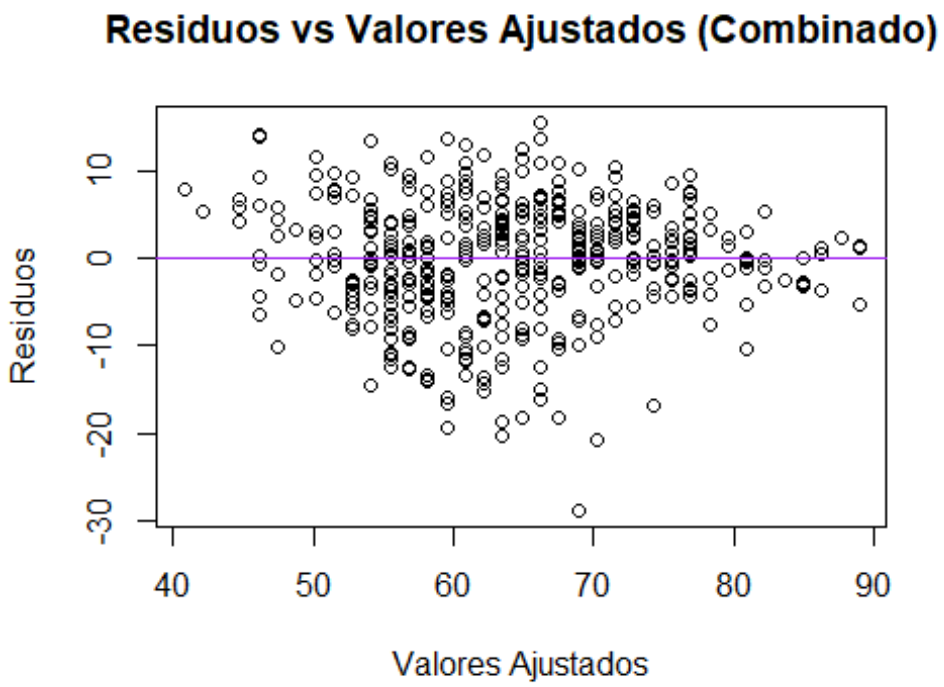
```r
# Grafico (hombres)
plot(A_hombres$fitted.values, A_hombres$residuals, main = "Residuos vs
Valores Ajustados (Hombres)", xlab = "Valores Ajustados", ylab = "Residuos")
abline(h = 0, col = "blue")
```




```r
# Grafico (mujeres)
plot(A_mujeres$fitted.values, A_mujeres$residuals, main = "Residuos vs
Valores Ajustados (Mujeres)", xlab = "Valores Ajustados", ylab = "Residuos")
abline(h = 0, col = "pink")
```




```r
# Grafico (combinado)
plot(A_combinado$fitted.values, A_combinado$residuals, main = "Residuos vs
Valores Ajustados (Combinado)", xlab = "Valores Ajustados", ylab =
"Residuos")
abline(h = 0, col = "purple")
```



Conclusion En base a el analisis hecho hasta ahora, se concluye que la estatura es un predictor importante del peso tanto en hombres como en mujeres. Por los modelos, se nota que un modelo especifico para cada sexo es mejor que un modelo combinado.

Parte 2 CON INTERACCION

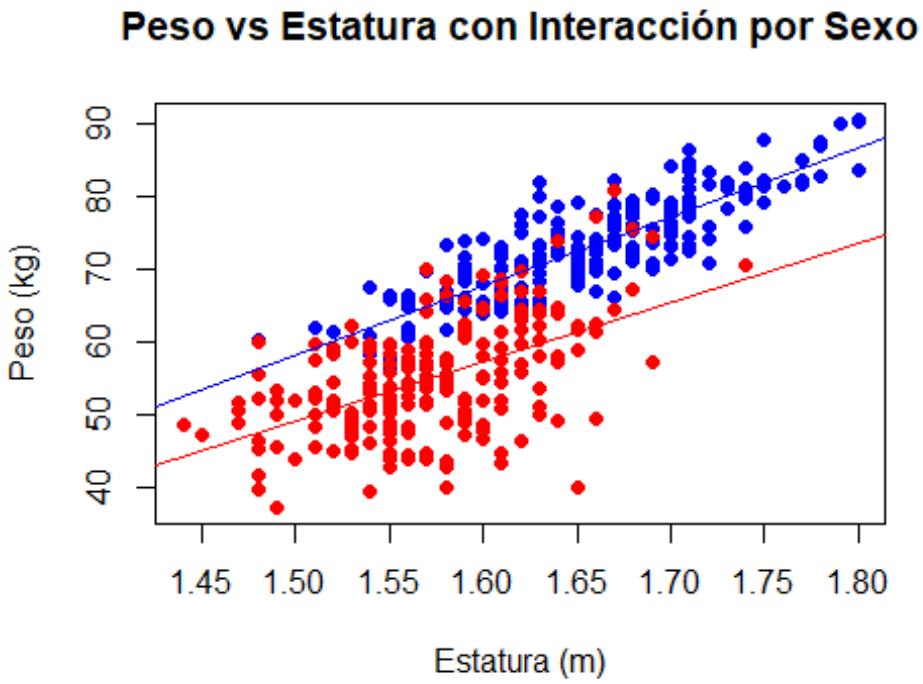
```
modelo_interaccion = lm(Peso ~ Estatura * Sexo, data = M)

# Hipotesis (Modelo)
# H0: El modelo no es significativo, todos los coeficientes son igual a 0.
# H1: Al menos uno de los coeficientes no es igual a 0.

summary(modelo_interaccion)

##
## Call:
## lm(formula = Peso ~ Estatura * Sexo, data = M)
##
## 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 ***
## Estatura         94.660      5.882  16.092   <2e-16 ***
```

```
## SexoM            11.124     14.950   0.744    0.457
## Estatura: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
```

```
# Interpretacion variables dummy
# El modelo incluira variables dummy para el sexo, estas son codificadas de manera que uno de los niveles actua como referencia y el otro como variable dummy.

# R²
summary(modelo_interaccion)$r.squared
```

```
## [1] 0.7847011
```

```
# Explicacion
# Nos indica el porcentaje de variacion en el peso que es explicado por la estatura y la interaccion con el sexo. En base a los resultados, nos dice que la variabilidad en el peso se explica por la estatura, el sexo y la interaccion entre estatura y sexo. Tiene mayor R^2 que los otros modelos, lo que dice que este tiene el mejor ajuste general, aunque la mejora no vale tanto debido a que el termino de interaccion no es significativo.

# Hipotesis (Significancia)
# H0: βî = 0 (el coeficiente no es significativo)
# H1: βî ≠ 0 (el coeficiente es significativo)

# Diagrama de dispersio8n con rectas de mejor ajuste diferenciadas por sexo
plot(M$Estatura, M$Peso, col=ifelse(M$Sexo == "H", "blue", "red"), pch=19,
xlab="Estatura (m)", ylab="Peso (kg)", main="Peso vs Estatura con Interacción por Sexo")
abline(lm(Peso ~ Estatura, data=M[M$Sexo == "H", ]), col="blue")
abline(lm(Peso ~ Estatura, data=M[M$Sexo == "M", ]), col="red")
```



Intercepto

En el modelo para hombres da un valor negativo lo cual no es una situacion realista, En el para mujeres es lo mismo que el de los hombres, y en el combinado nos da un valor que no es tan preciso y es mas generalizado, igual que en el modelo con interaccion, pero dice que la relacion entre peso y estatura es diferente en base al sexo.

Coeficientes

En el modelo para hombres dice que por cada metro adicional de estatura, el peso aumenta para el hombre, en el caso del modelo para mujeres la relacion es al reves, y en el combinado indica que en promedio, por cada metro adicional de estatura, el peso aumenta tambien, y en el modelo con interaccion dice que el efecto de la estatura en el peso es diferente para las mujeres.

Conclusion

En base a los analisis realizados, dado que el termino de interaccion no es significativo, el modelo combinado sin interaccion puede ser suficiente para describir la relacion entre estatura y peso. Como extra es un modelo mas sencillo y con buena R^2. Pero si se desea un analisis mas detallado para cada sexo, los modelos separados son los mas convenientes, especialmente porque el R^2 es mas alta en los hombres. Asi que se termina diciendo que

la estatura es un predictor significativo del peso en ambos sexos, pero la fuerza de esta relacion varia entre hombres y mujeres.

Parte 3 Analisis de loS errores

```
library(nortest)
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

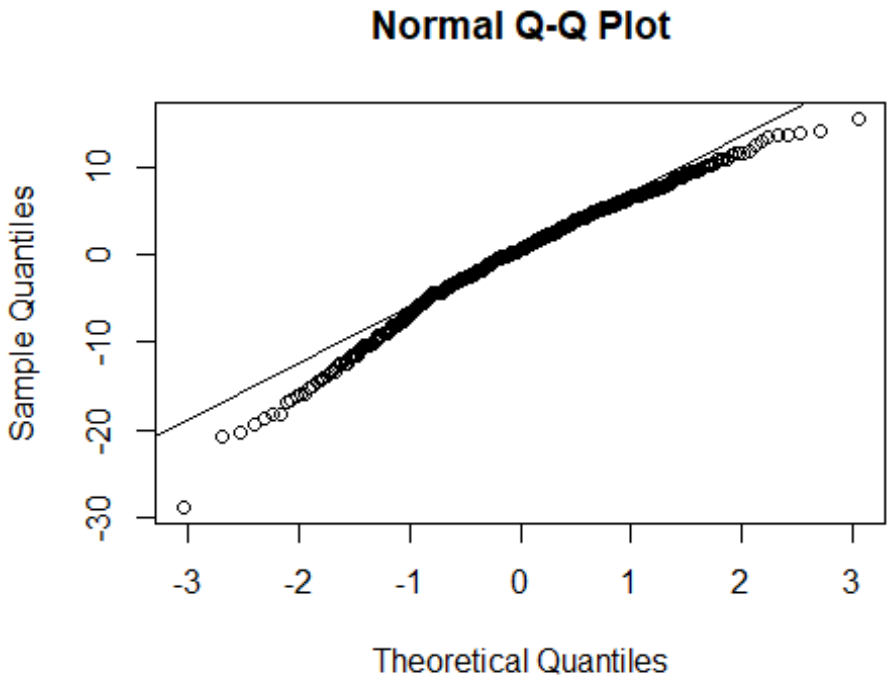
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)

A_combinado = lm(Peso ~ Estatura, data = M)

summary(A_combinado)
```

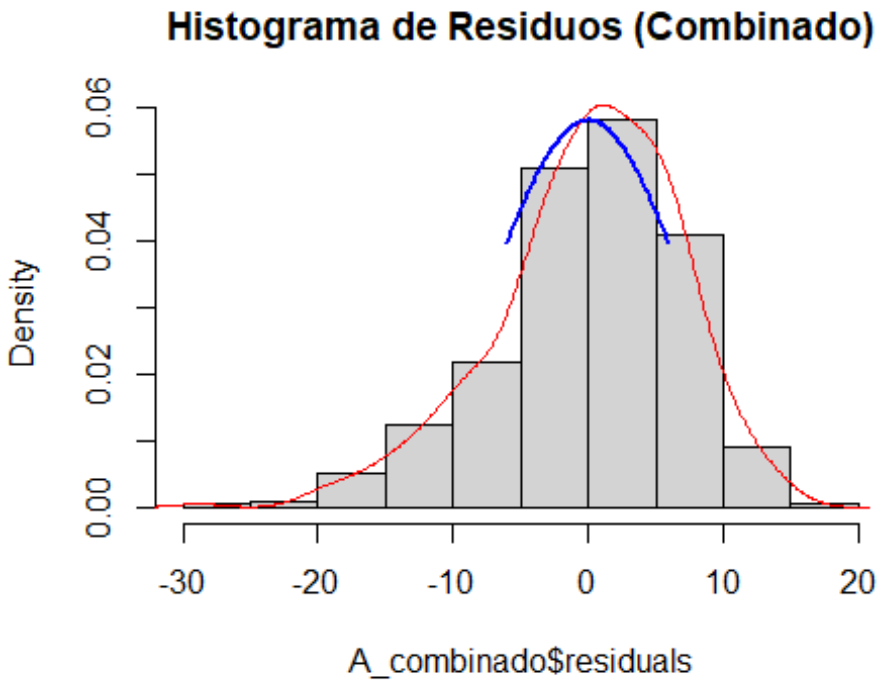
```
##
## Call:
## lm(formula = Peso ~ Estatura, data = M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.8653  -3.7654   0.6706   5.0142  15.6006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -151.883      7.655  -19.84   <2e-16 ***
## Estatura     133.793      4.741   28.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.883 on 438 degrees of freedom
## Multiple R-squared:  0.6452, Adjusted R-squared:  0.6444
## F-statistic: 796.5 on 1 and 438 DF,  p-value: < 2.2e-16
```

```
# 1. Normalidad de los residuos
# Hipotesis:
# H0: Los residuos siguen una distribucion normal.
# H1: Los residuos no siguen una distribucion normal.

# Prueba de Normalidad de Anderson-Darling
ad.test(A_combinado$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  A_combinado$residuals
## A = 2.4766, p-value = 2.888e-06

# Graficos de normalidad
qqnorm(A_combinado$residuals)
qqline(A_combinado$residuals)
```




```
hist(A_combinado$residuals, freq = FALSE, main = "Histograma de Residuos
(Combinado)")
lines(density(A_combinado$residuals), col = "red")
curve(dnorm(x, mean = mean(A_combinado$residuals), sd =
sd(A_combinado$residuals)),
      from = -6, to = 6, add = TRUE, col = "blue", lwd = 2)
```

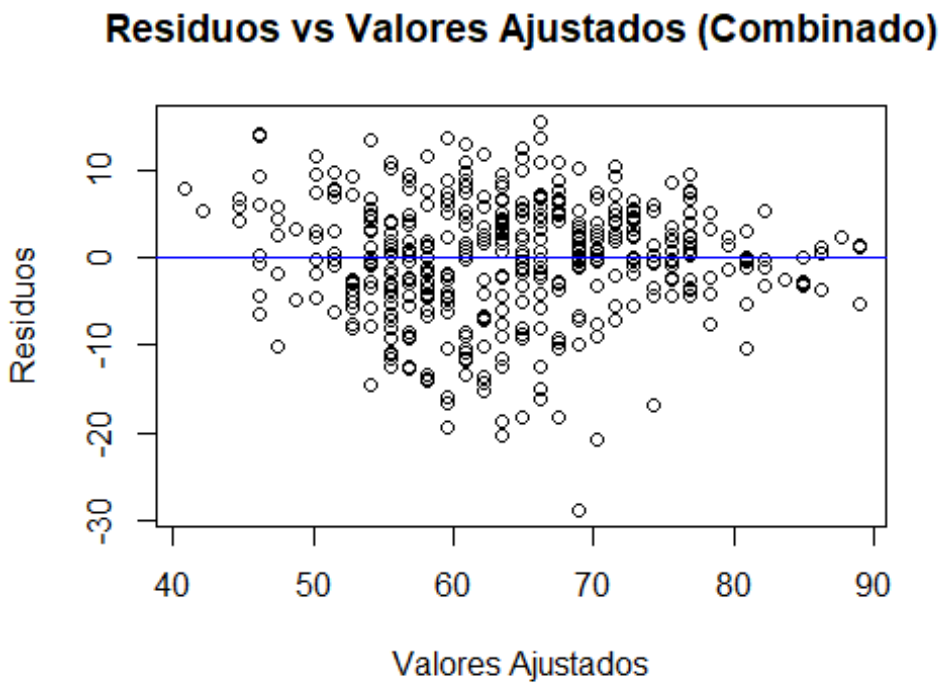



```r
# 2. Verificacion de media cero
# Hipotesis:
# H0: La media de los residuos es igual a cero.
# H1: La media de los residuos no es igual a cero.

t.test(A_combinado$residuals)
```

```
##
##  One Sample t-test
##
## data:  A_combinado$residuals
## t = 2.7844e-15, df = 439, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.6441362  0.6441362
## sample estimates:
##   mean of x
## 9.12569e-16
```

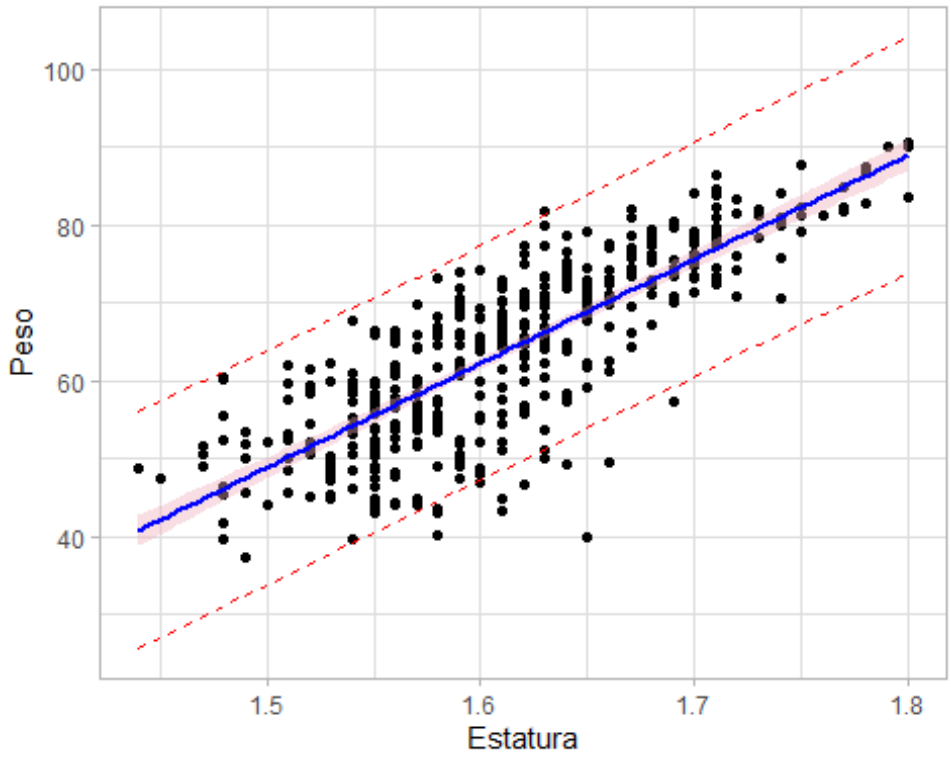
```r
# 3. Homocedasticidad e independencia
# Hipotesis de homocedasticidad:
# H0: La varianza de los errores es constante (homocedasticidad).
# H1: La varianza de los errores no es constante (heterocedasticidad).

# Grafico de residuos vs valores ajustados
plot(A_combinado$fitted.values, A_combinado$residuals, main = "Residuos vs
Valores Ajustados (Combinado)",
```

```
     xlab = "Valores Ajustados", ylab = "Residuos")
abline(h = 0, col = "blue")
```




```
# Prueba de homocedasticidad: Breusch-Pagan
bptest(A_combinado)

##
##  studentized Breusch-Pagan test
##
## data:  A_combinado
## BP = 5.7194, df = 1, p-value = 0.01678

# Prueba de independencia: Durbin-Watson
dwtest(A_combinado)

##
##  Durbin-Watson test
##
## data:  A_combinado
## DW = 1.4034, p-value = 1.586e-10
## alternative hypothesis: true autocorrelation is greater than 0

# 4. Intervalos de prediccion y confianza
# Construir la grafica de los intervalos de confianza y prediccion
Ip = predict(A_combinado, interval = "prediction", level = 0.97)

## Warning in predict.lm(A_combinado, interval = "prediction", level = 0.97):
predictions on current data refer to _future_ responses
```

```r
datos1 = cbind(M, Ip)

ggplot(datos1, aes(x = Estatura, y = Peso)) +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, level = 0.97, col =
"blue", fill = "pink2") +
  theme_light()
```



Conclusiones

De acuerdo a la normalidad de los residuos con Anderson-Darling, el resultado sugiere que no siguen una distribucion completamente normal, lo cual puede afectar a la validez. En este contexto, la falta de normalidad de los residuos, sugiere que el modelo combinado podria no ser tan recomendable. La verificacion de media cero dio que no podemos rechazar la hipotesis nula, por lo que la media de los residuos es 0. Esto indica que no hay un sesgo sistematico en el modelo, lo cual es bueno. La homocedasticidad e independencia indicaron que se rechaza la hipotesis nula por lo que existe heterocedasticidad, y ademas los residuos no son independientes. Finalmente, los intervalos de prediccion y confianza muestra como los valores de peso se ajustan en funcion de la estatura, tambien los intervalos de prediccion son mas amplios que los de confianza. Estos intervalos permiten estimar el rango probable de pesos para una estatura dada, pero debido a los problemas de normalidad y heterocedasticidad, pueden no ser 100% precisos.