Act12. Regresión Lineal - Análisis de errores

Andrés Villarreal González

2024-09-04

Analisis de errores

```
# Cargar datos
datos <- read.csv("Estatura-peso_HyM.csv")

# Convertir la columna 'Sexo' a 1 para Hombres (H) y 0 para Mujeres (M)
datos$Sexo <- ifelse(datos$Sexo == "H", 1, 0)
```

1. Modelo sin Interacción

```
# Modelo sin interacción (solo estatura)
modelo_sin_interaccion <- lm(Peso ~ Estatura, data = datos)
summary(modelo_sin_interaccion)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura, data = datos)
##
## 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
```

2. Modelo con interacción

```
# Crear la interacción entre Estatura y Sexo
datos$Estatura_Sexo <- datos$Estatura * datos$Sexo

# Modelo con interacción entre estatura y sexo
modelo_con_interaccion <- lm(Peso ~ Estatura + Sexo + Estatura_Sexo, data
= datos)
summary(modelo_con_interaccion)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura + Sexo + Estatura_Sexo, data = datos)
##
## 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)    -72.560     11.346  -6.395 4.13e-10 ***
## Estatura        81.149      7.209  11.256  < 2e-16 ***
## Sexo           -11.124     14.950  -0.744    0.457
## Estatura_Sexo   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
```

```
# Prueba ANOVA para verificar la significancia del modelo con interacción
anova(modelo_con_interaccion)
```

```
## Analysis of Variance Table
##
## Response: Peso
##                Df Sum Sq Mean Sq   F value Pr(>F)
## Estatura        1  37731   37731 1306.5938 <2e-16 ***
## Sexo            1   8097    8097  280.3892 <2e-16 ***
## Estatura_Sexo   1     61      61    2.1085 0.1472
## Residuals     436  12590      29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Resumen del modelo con interacción para verificar la significancia de
los coeficientes
summary(modelo_con_interaccion)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura + Sexo + Estatura_Sexo, data = datos)
##
## 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)    -72.560     11.346  -6.395 4.13e-10 ***
## Estatura        81.149      7.209  11.256  < 2e-16 ***
```

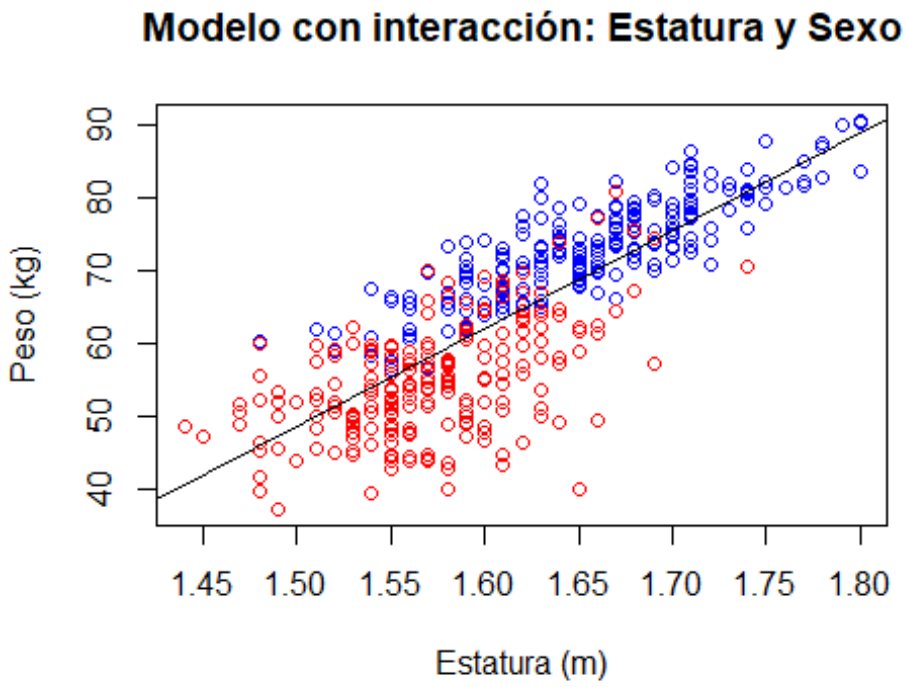
```
## Sexo           -11.124     14.950  -0.744    0.457
## Estatura_Sexo   13.511      9.305   1.452    0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.374 on 436 degrees of freedom
## Multiple R-squared:  0.7847, Adjusted R-squared:  0.7832
## F-statistic: 529.7 on 3 and 436 DF,  p-value: < 2.2e-16
```

```
# R-cuadrado para el modelo con interacción
summary(modelo_con_interaccion)$r.squared
```

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

```
# Diagrama de dispersión con las líneas de regresión para hombres y
mujeres
plot(datos$Estatura, datos$Peso, col = ifelse(datos$Sexo == 1, "blue",
"red"),
     xlab = "Estatura (m)", ylab = "Peso (kg)", main = "Modelo con
interacción: Estatura y Sexo")

# Línea de regresión
abline(lm(Peso ~ Estatura, data = datos), col = "black")
```



3. Modelo solo hombres

```
# Filtrar los datos para hombres
datos_hombres <- subset(datos, Sexo == 1)
```

```r
# Modelo para hombres (solo estatura)
modelo_hombres <- lm(Peso ~ Estatura, data = datos_hombres)
summary(modelo_hombres)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_hombres)
##
## 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 ***
## 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
```

```r
anova(modelo_hombres)
```

```
## Analysis of Variance Table
##
## Response: Peso
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Estatura    1 7478.0  7478.0  552.67 < 2.2e-16 ***
## Residuals 218 2949.7    13.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(modelo_hombres)
```

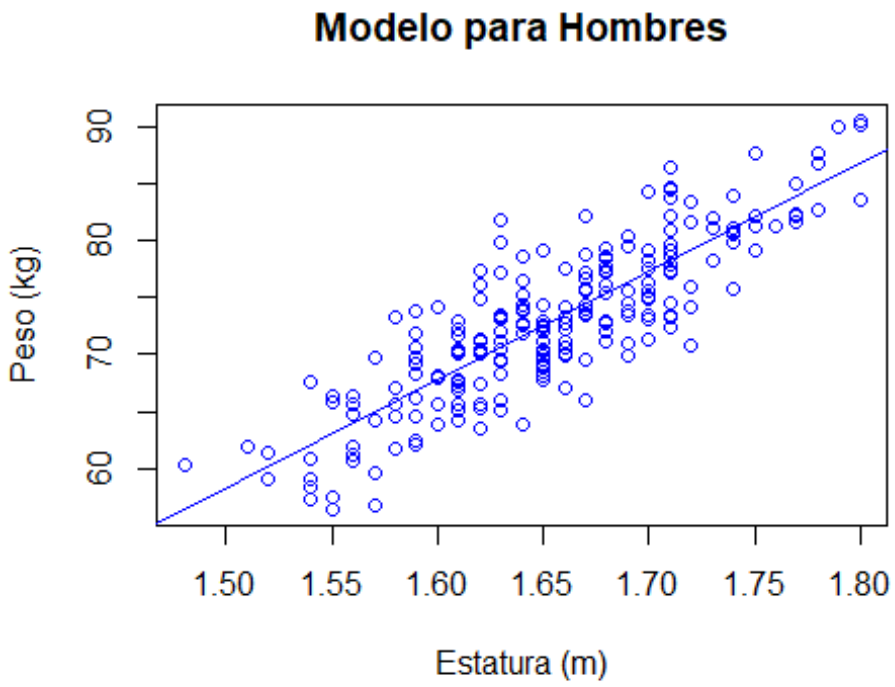
```
##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_hombres)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3881 -2.6073 -0.0665  2.4421 11.1883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -83.685      6.663  -12.56   <2e-16 ***
## 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

summary(modelo_hombres)$r.squared

## [1] 0.7171292

plot(datos_hombres$Estatura, datos_hombres$Peso, xlab = "Estatura (m)",
ylab = "Peso (kg)", main = "Modelo para Hombres", col = "blue")
abline(modelo_hombres, col = "blue")
```



4. Modelo solo mujeres

```
# Filtrar los datos para mujeres
datos_mujeres <- subset(datos, Sexo == 0)

# Modelo para mujeres (solo estatura)
modelo_mujeres <- lm(Peso ~ Estatura, data = datos_mujeres)
summary(modelo_mujeres)

##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_mujeres)
##
## 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 ***
## 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
```

```
anova(modelo_mujeres)
```

```
## Analysis of Variance Table
##
## Response: Peso
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Estatura    1 3658.6  3658.6   82.73 < 2.2e-16 ***
## Residuals 218 9640.7    44.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

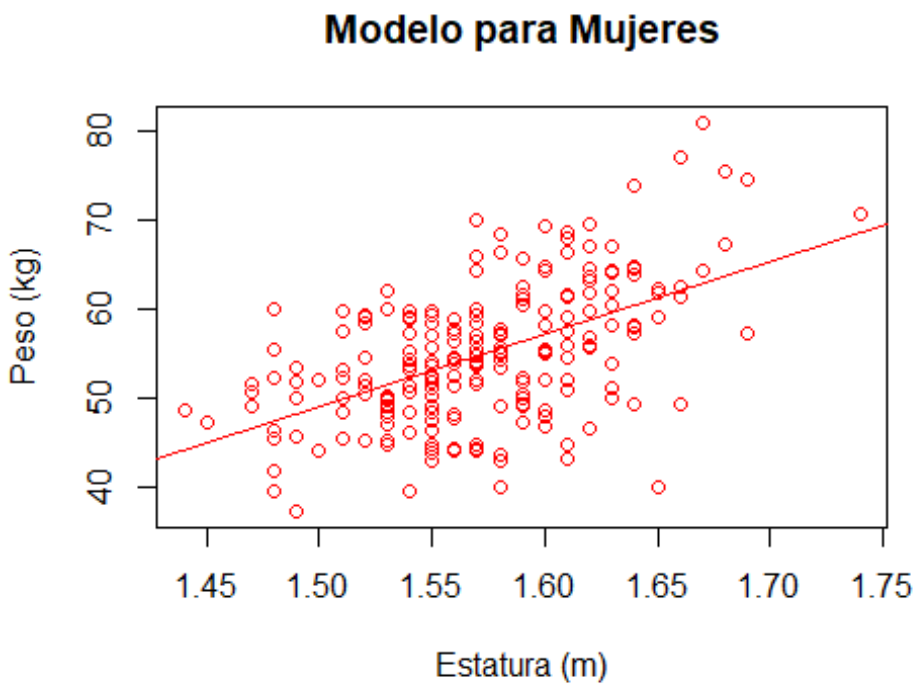
```
summary(modelo_mujeres)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_mujeres)
##
## 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 ***
## 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
```

```
summary(modelo_mujeres)$r.squared
```

```
## [1] 0.2750963
```

```
plot(datos_mujeres$Estatura, datos_mujeres$Peso, xlab = "Estatura (m)",
ylab = "Peso (kg)", main = "Modelo para Mujeres", col = "red")
abline(modelo_mujeres, col = "red")
```



Analisis de Resiudos para modelo_hombres

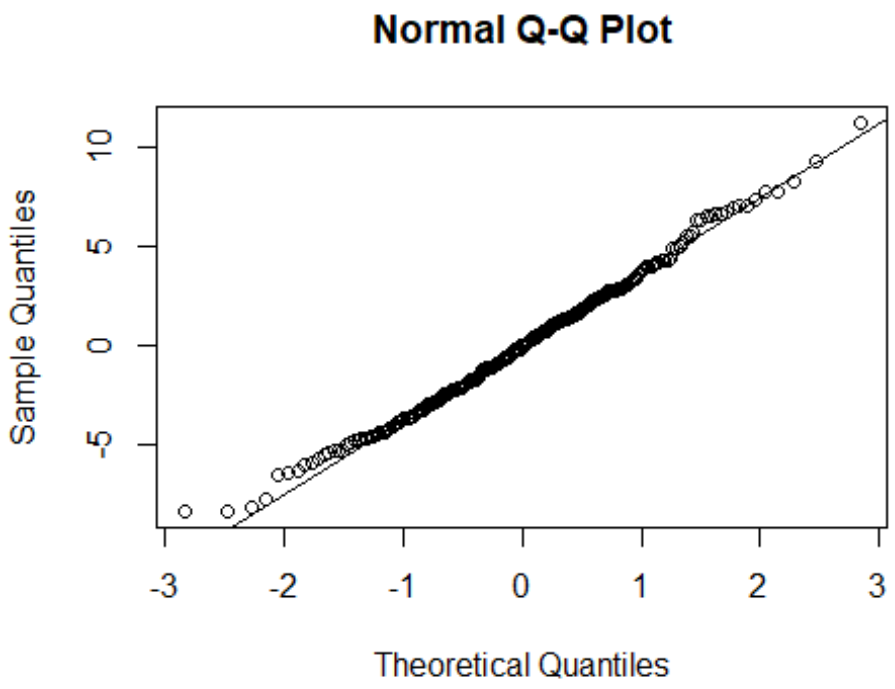
Normalidad de residuos

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

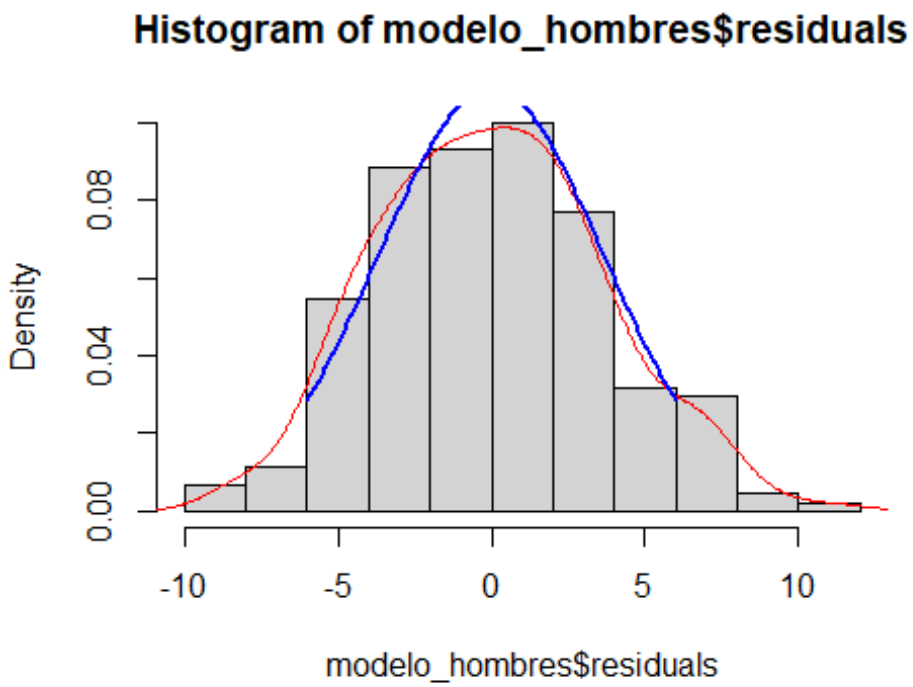
```
library(nortest)
ad.test(modelo_hombres$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  modelo_hombres$residuals
## A = 0.3009, p-value = 0.5771
```

```
qqnorm(modelo_hombres$residuals)
qqline(modelo_hombres$residuals)
```




```
hist(modelo_hombres$residuals,freq=FALSE)
lines(density(modelo_hombres$residual),col="red")
curve(dnorm(x,mean=mean(modelo_hombres$residuals),sd=sd(modelo_hombres$re
siduals)), from=-6, to=6, add=TRUE, col="blue",lwd=2)
```



Verificación de media cero

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

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

```
##
##  One Sample t-test
##
## data:  modelo_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
```

Homocedasticidad e Independencia

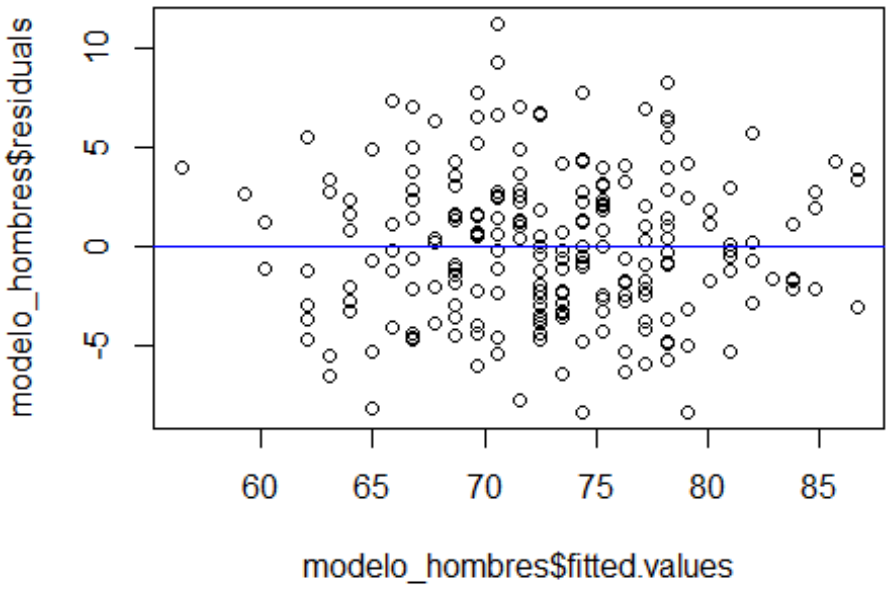
Homocedasticidad

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

Independencia

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

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




```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

dwtest(modelo_hombres)

##
##  Durbin-Watson test
##
## data:  modelo_hombres
## DW = 2.0556, p-value = 0.6599
## alternative hypothesis: true autocorrelation is greater than 0

bgtest(modelo_hombres)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_hombres
## LM test = 0.20778, df = 1, p-value = 0.6485
```

```
library(lmtest)
bptest(modelo_hombres)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_hombres
## BP = 0.93324, df = 1, p-value = 0.334
```

```
gqtest(modelo_hombres)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo_hombres
## GQ = 0.84148, df1 = 108, df2 = 108, p-value = 0.8144
## alternative hypothesis: variance increases from segment 1 to 2
```

Resultados de pruebas para modelo de hombres:

Prueba Normalidad: Debido al valor p mayor a alfa y observando los gráficos no se rechaza H0 por lo que se puede decir que los residuos provienen de una distribución normal.

Prueba media cero: El p-valor es 1, lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula.

Prueba de Homocedasticidad: El p-valor de ambas pruebas es mayor que el nivel de significancia, lo que sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Prueba de Independencia: El p-valor de ambas pruebas es mayor que el nivel de significancia, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula.

```
# Generar los intervalos de predicción con un nivel de confianza del 97%
Ip = predict(object = modelo_hombres, interval = "prediction", level =
0.97)
```

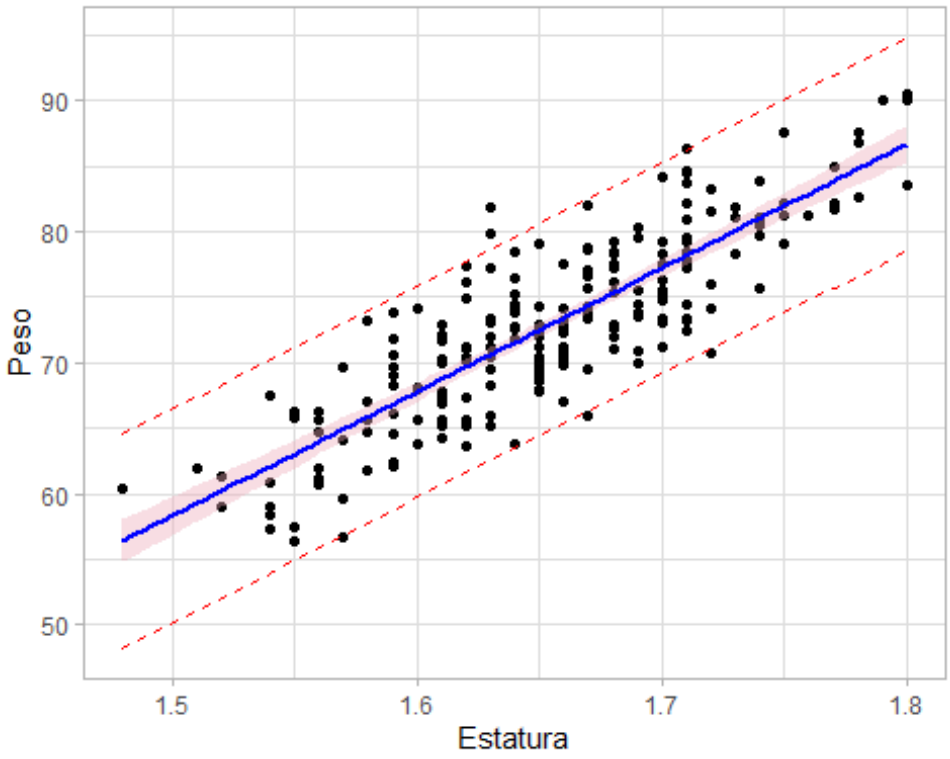
```
## Warning in predict.lm(object = modelo_hombres, interval =
"prediction", : predictions on current data refer to _future_ responses
```

```
# Añadir los intervalos de predicción a los datos originales
datos1 = cbind(datos_hombres, Ip)

# Cargar ggplot2
```

```r
library(ggplot2)

# Crear la gráfica
ggplot(datos1, aes(x = Estatura, y = Peso)) +
  geom_point() +  # Puntos de los datos reales
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +  # Límite inferior del intervalo de predicción
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +  # Límite superior del intervalo de predicción
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.97, col = "blue", fill = "pink2") +  # Línea de regresión con intervalo de confianza
  theme_light()  # Tema de la gráfica
```



El modelo lineal ajustado para hombres muestra una fuerte relación lineal positiva entre la estatura y el peso. El ajuste es razonable, con la mayor parte de los puntos distribuidos alrededor de la línea de regresión. Los intervalos de confianza estrechos sugieren que la media del peso está bien estimada, aunque el intervalo de predicción más amplio refleja la variabilidad que podría existir entre personas con la misma estatura.