Actividad 12: Regresión Lineal - Análisis de los errores

Daniela Jiménez Téllez

2024-09-04

Importación de datos

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

En este caso se utilizará el modelo con interacción:

```
# Hombres

hombres <- subset(datos, Sexo == "H")
modelo_hombres <- lm(Peso ~ Estatura, data = hombres)

# Mujeres

mujeres <- subset(datos, Sexo == "M")
modelo_mujeres <- lm(Peso ~ Estatura, data = mujeres)

# Sin interacción

modelo_sin_interaccion <- lm(Peso ~ Estatura + Sexo, data = datos)

# Con interacción

modelo_con_interaccion <- lm(Peso ~ Estatura * Sexo, data = datos)
```

La validez del modelo

1. Analiza si el (los) modelo(s) obtenidos anteriormente son apropiados para el conjunto de datos. Realiza el análisis de los residuos:

- **Normalidad de los residuos**

La hipótesis estadística es:

H_0 : Los datos se comportan como una distribución normal.

H_1 : Los datos no se comportan como una distribución normal.

Con un $\alpha = 0.03$

```
# Hombres

normH <- shapiro.test(resid(modelo_hombres))
cat("Los resultados de la normalidad para el modelo de hombres usando el test de Shapiro-Wilk so
n: \n")
```

```
## Los resultados de la normalidad para el modelo de hombres usando el test de Shapiro-Wilk son:
```

```
print(normH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_hombres)
## W = 0.99356, p-value = 0.4597
```

```
# Mujeres

normM <- shapiro.test(resid(modelo_mujeres))
cat("Los resultados de la normalidad para el modelo de mujeres usando el test de Shapiro-Wilk so
n: \n")
```

```
## Los resultados de la normalidad para el modelo de mujeres usando el test de Shapiro-Wilk son:
```

```
print(normM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_mujeres)
## W = 0.99659, p-value = 0.9144
```

```
# Sin interacción

normSI <- shapiro.test(resid(modelo_sin_interaccion))
cat("Los resultados de la normalidad para el modelo con interacción usando el test de Shapiro-Wi
lk son: \n")
```

```
## Los resultados de la normalidad para el modelo con interacción usando el test de Shapiro-Wilk
son:
```

```
print(normSI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_sin_interaccion)
## W = 0.99337, p-value = 0.0501
```

```
# Con interacción

normCI <- shapiro.test(resid(modelo_con_interaccion))
cat("Los resultados de la normalidad para el modelo con interacción usando el test de Shapiro-Wi
lk son: \n")
```

```
## Los resultados de la normalidad para el modelo con interacción usando el test de Shapiro-Wilk
son:
```

```
print(normCI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_con_interaccion)
## W = 0.99356, p-value = 0.05772
```

En el caso del modelo de **hombres**: Dado que el p-value > 0.03, se acepta H_0, por lo tanto los datos se comportan como una normal.

En el caso del modelo de **mujeres**: Dado que el p-value > 0.03, se acepta H_0, por lo tanto los datos se comportan como una normal.

En el caso del modelo sin **interacción**: Dado que el p-value > 0.03, se acepta H_0, por lo tanto los datos se comportan como una normal.

En el caso del modelo con **interacción**: Dado que el p-value > 0.03, se acepta H_0, por lo tanto los datos se comportan como una normal.

- **Verificación de media cero**

La hipótesis estadística es:

$H_0 : \mu_e = 0$

$H_1 : \mu_e \neq 0$

```
# Hombres

medH <- t.test(resid(modelo_hombres))
cat("El resultado del test de media cero para el modelo de hombres es: \n")
```

```
## El resultado del test de media cero para el modelo de hombres es:
```

```
print(medH)
```

```
##
##  One Sample t-test
##
## data:  resid(modelo_hombres)
## 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
```

```
# Mujeres

medM <- t.test(resid(modelo_mujeres))
cat("El resultado del test de media cero para el modelo de mujeres es: \n")
```

```
## El resultado del test de media cero para el modelo de mujeres es:
```

```
print(medM)
```

```
##
##  One Sample t-test
##
## data:  resid(modelo_mujeres)
## 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
```

```
# Sin interacción

medSI <- t.test(resid(modelo_sin_interaccion))
cat("El resultado del test de media cero para el modelo sin interacción es: \n")
```

```
## El resultado del test de media cero para el modelo sin interacción es:
```

```
print(medSI)
```

```
##
##  One Sample t-test
##
## data:  resid(modelo_sin_interaccion)
## t = 2.4085e-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
## 6.163788e-17
```

```
# Con interacción

medCI <- t.test(resid(modelo_con_interaccion))
cat("El resultado del test de media cero para el modelo con interacción es: \n")
```

```
## El resultado del test de media cero para el modelo con interacción es:
```

```
print(medSI)
```

```
##
##  One Sample t-test
##
## data:  resid(modelo_sin_interaccion)
## t = 2.4085e-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
## 6.163788e-17
```

En el caso del modelo de **hombres**: Se rechaza H_0 ya que la media es diferente de 0.

En el caso del modelo de **mujeres**: Se rechaza H_0 ya que la media es diferente de 0.

En el caso del modelo sin **interacción**: Se rechaza H_0 ya que la media es diferente de 0.

En el caso del modelo con **interacción**: Se rechaza H_0 ya que la media es diferente de 0.

A pesar de que en todos los casos se rechazó H_0 ya que la media no es 0, esta se acerca mucho.

- **Homocedasticidad e independencia**

Para homocedasticidad:

H_0 : La varianza de los errores es constante (homocedasticidad).

H_1 La varianza de los errores no es constante (heterocedasticidad).

Para independencia:

H_0 Los errores no están correlacionados.

H_1 Los errores están correlacionados.

```
library(lmtest)
```

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

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

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

```
### HOMOCEDASTICIDAD

# Hombres

hom1H <- bptest(modelo_hombres)
hom2H <- gqtest(modelo_hombres)

cat("Los resultados de los tests de homocedasticidad para el modelo de hombres son: \n")
```

```
## Los resultados de los tests de homocedasticidad para el modelo de hombres son:
```

```
print(hom1H)
```

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

```
print(hom2H)
```

```
##
##  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
```

```
# Mujeres

hom1M <- bptest(modelo_mujeres)
hom2M <- gqtest(modelo_mujeres)

cat("Los resultados de los tests de homocedasticidad para el modelo de mujeres son: \n")
```

```
## Los resultados de los tests de homocedasticidad para el modelo de mujeres son:
```

```
print(hom1M)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_mujeres
## BP = 8.4976, df = 1, p-value = 0.003556
```

```
print(hom2M)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo_mujeres
## GQ = 1.4265, df1 = 108, df2 = 108, p-value = 0.03313
## alternative hypothesis: variance increases from segment 1 to 2
```

```
# Sin interacción

hom1SI <- bptest(modelo_sin_interaccion)
hom2SI <- gqtest(modelo_sin_interaccion)

cat("Los resultados de los tests de homocedasticidad para el modelo sin interacción son: \n")
```

```
## Los resultados de los tests de homocedasticidad para el modelo sin interacción son:
```

```
print(hom1SI)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_sin_interaccion
## BP = 48.202, df = 2, p-value = 3.413e-11
```

```
print(hom2SI)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo_sin_interaccion
## GQ = 3.2684, df1 = 217, df2 = 217, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

```
# Con interacción

hom1CI <- bptest(modelo_con_interaccion)
hom2CI <- gqtest(modelo_con_interaccion)

cat("Los resultados de los tests de homocedasticidad para el modelo con interacción son: \n")
```

```
## Los resultados de los tests de homocedasticidad para el modelo con interacción son:
```

```
print(hom1CI)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_con_interaccion
## BP = 59.211, df = 3, p-value = 8.667e-13
```

```
print(hom2CI)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo_con_interaccion
## GQ = 3.2684, df1 = 216, df2 = 216, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

```
### INDEPENDENCIA

# Hombres

ind1H <- dwtest(modelo_hombres)
ind2H <- bgtest(modelo_hombres)

cat("Los resultados de los tests de independencia para el modelo de hombres son: \n")
```

```
## Los resultados de los tests de independencia para el modelo de hombres son:
```

```
print(ind1H)
```

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

```
print(ind2H)
```

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

```
# Mujeres

ind1M <- dwtest(modelo_mujeres)
ind2M <- bgtest(modelo_mujeres)

cat("Los resultados de los tests de independencia para el modelo de mujeres son: \n")
```

```
## Los resultados de los tests de independencia para el modelo de mujeres son:
```

```
print(ind1M)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_mujeres
## DW = 1.8062, p-value = 0.07532
## alternative hypothesis: true autocorrelation is greater than 0
```

```
print(ind2M)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_mujeres
## LM test = 1.4655, df = 1, p-value = 0.2261
```

```
# Sin interacción

ind1SI <- dwtest(modelo_sin_interaccion)
ind2SI <- bgtest(modelo_sin_interaccion)

cat("Los resultados de los tests de independencia para el modelo sin interacción son: \n")
```

```
## Los resultados de los tests de independencia para el modelo sin interacción son:
```

```
print(ind1SI)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_sin_interaccion
## DW = 1.8663, p-value = 0.07325
## alternative hypothesis: true autocorrelation is greater than 0
```

```
print(ind2SI)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_sin_interaccion
## LM test = 1.3595, df = 1, p-value = 0.2436
```

```
# Con interacción

ind1CI <- dwtest(modelo_con_interaccion)
ind2CI <- bgtest(modelo_con_interaccion)

cat("Los resultados de los tests de independencia para el modelo con interacción son: \n")
```

```
## Los resultados de los tests de independencia para el modelo con interacción son:
```

```
print(ind1CI)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_con_interaccion
## DW = 1.8646, p-value = 0.07113
## alternative hypothesis: true autocorrelation is greater than 0
```

```
print(ind2CI)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_con_interaccion
## LM test = 1.3453, df = 1, p-value = 0.2461
```

2. No te olvides de incluir las hipótesis en la pruebas de hipótesis que realices.

Se incluyen en la parte anterior.

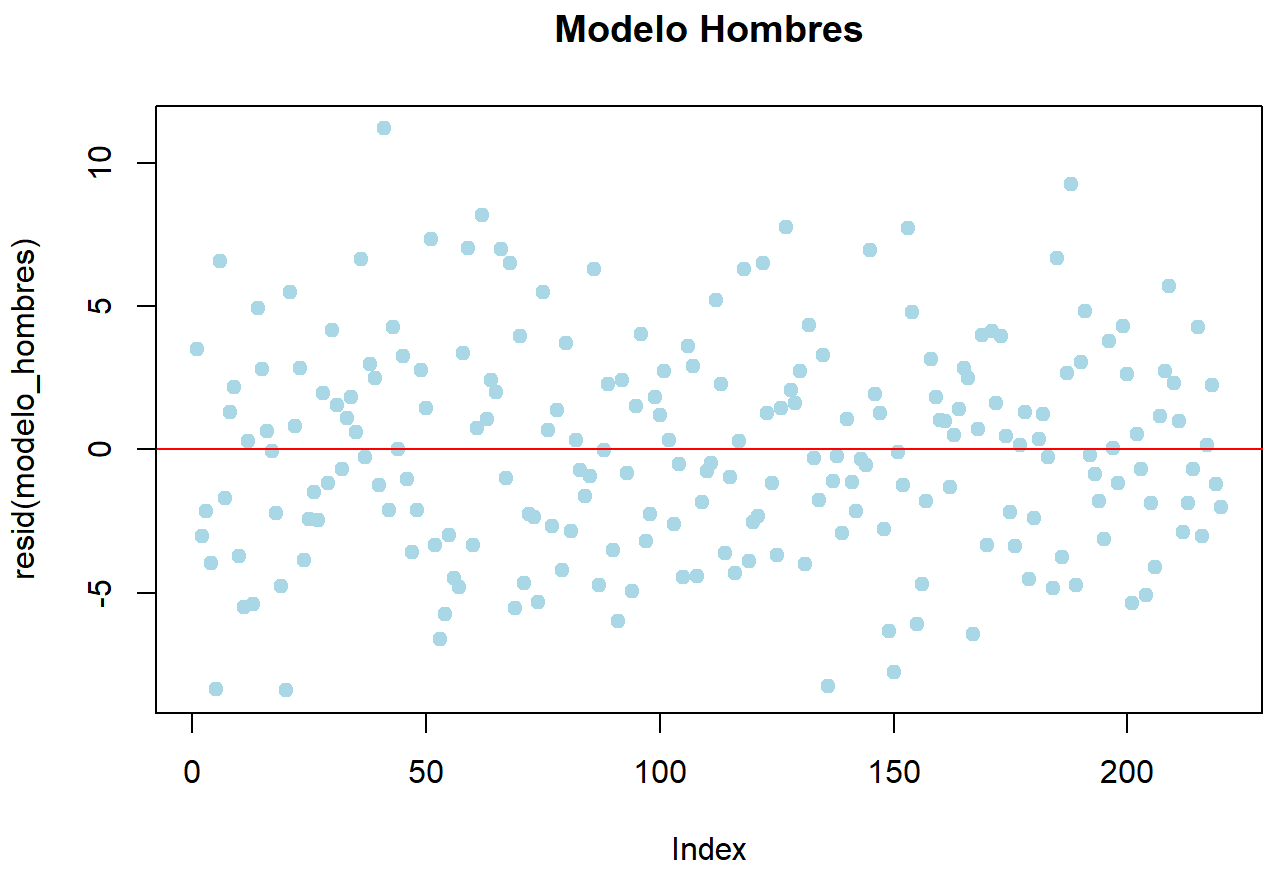
3. Interpreta en el contexto del problema cada uno de los análisis que hiciste.

Para concluir, se pueded decir que los modelos de hombres y mujeres por separado pasan todas las pruebas. Por otro lado, los modelos con y sin interacción, no pasaron todos los parámetros. Primeramente, estos apenas pasaron la prueba de normalidad. Finalmente, se puede decir que los errores están correlacionados, y hay demás fallas en los tests de homocedasticidad e indedpendencia, por lo que no son buenos modelos.

4. Utiliza el comando: plot(modelo). Observa las gráficas obtenidas y contesta:

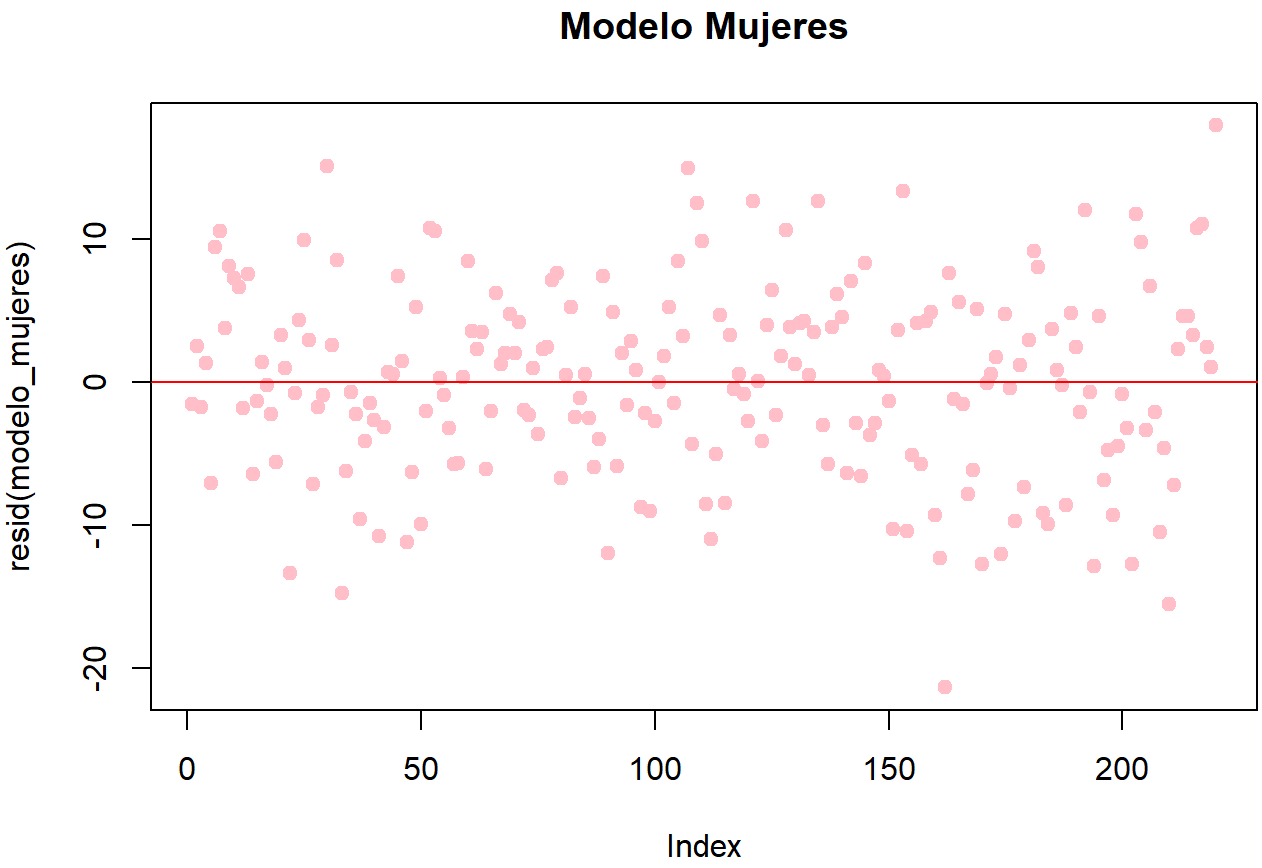
```
# Hombres

plot(resid(modelo_hombres), col = "lightblue", pch = 19, main = "Modelo Hombres")
abline(h = 0, col = "red")
```



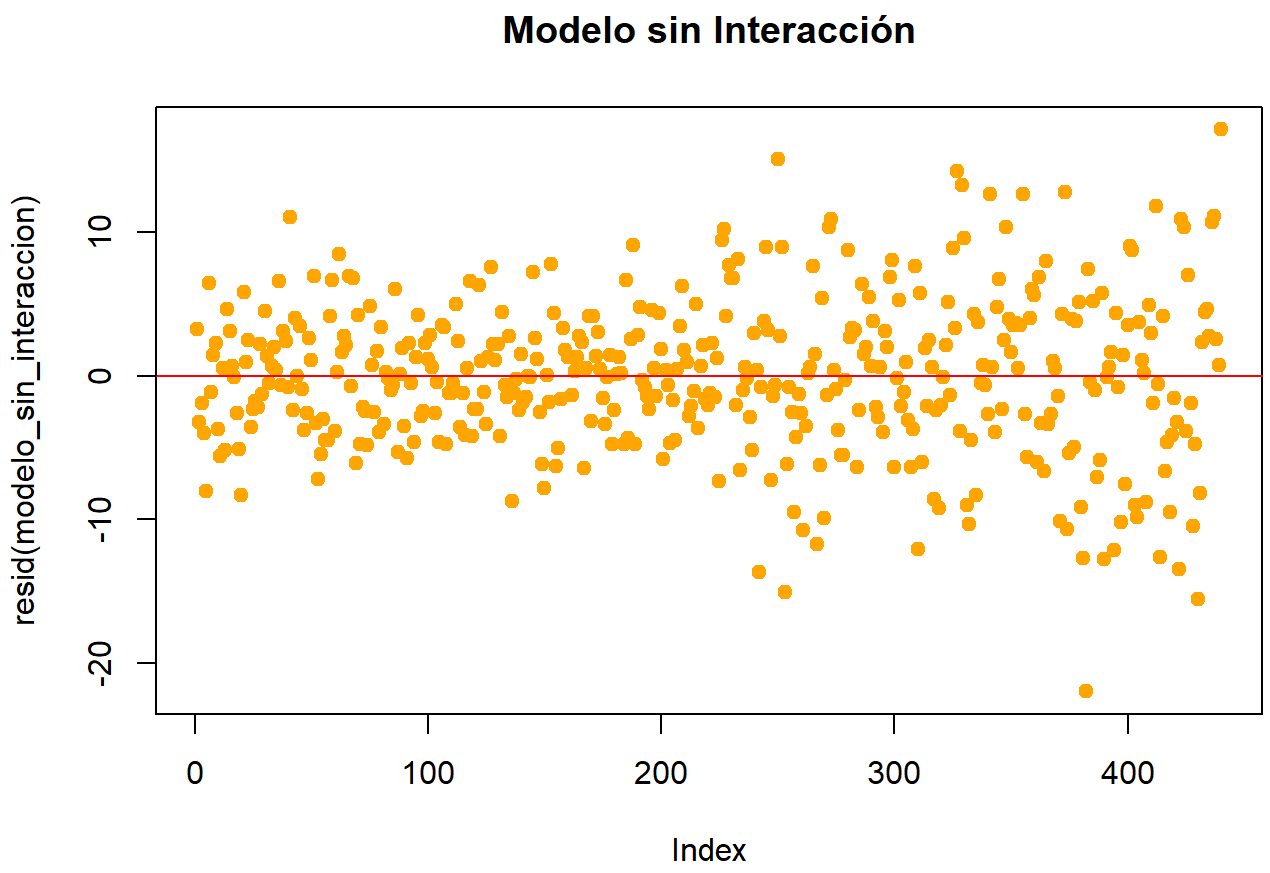
```
# Mujeres

plot(resid(modelo_mujeres), col = "pink", pch = 19, main = "Modelo Mujeres")
abline(h = 0, col = "red")
```



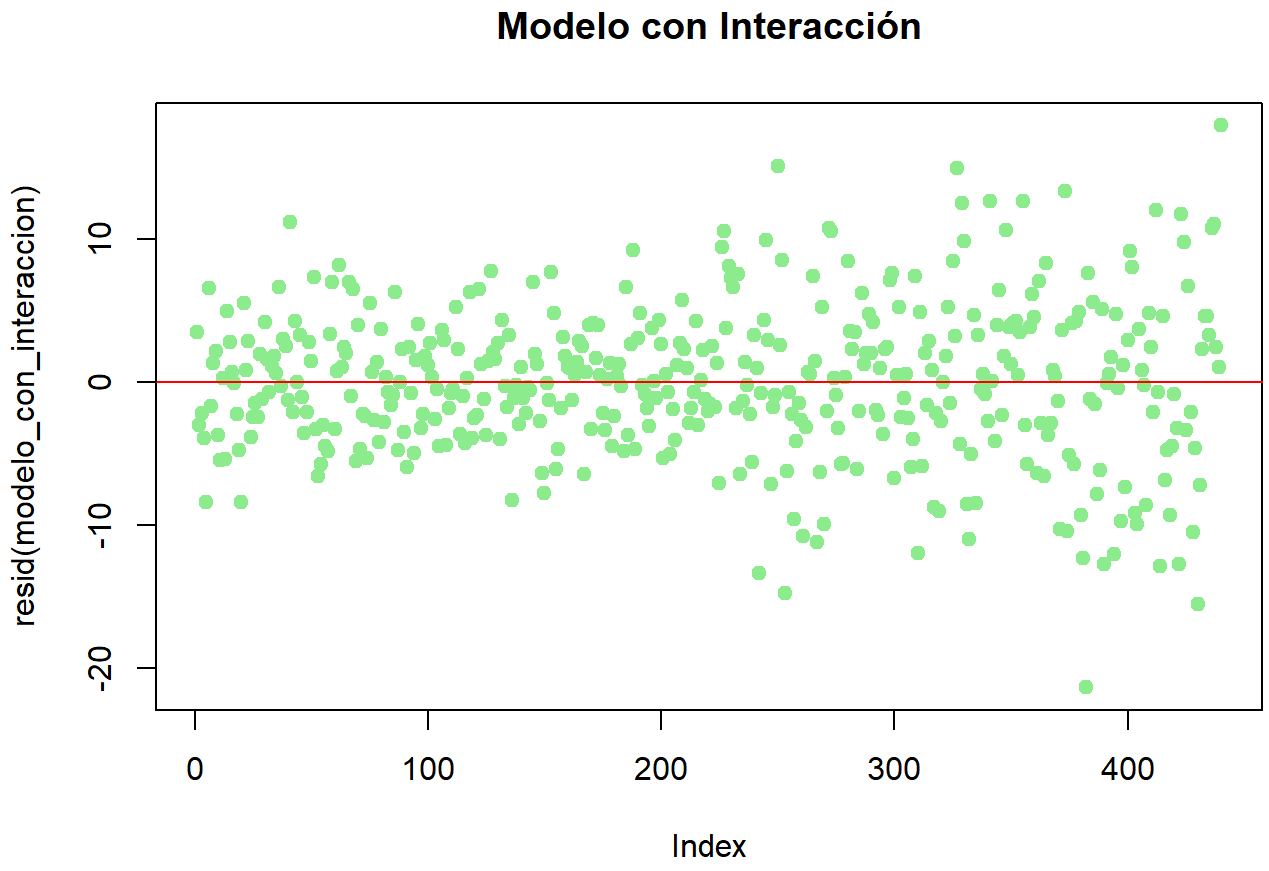

```
# Sin interacción

plot(resid(modelo_sin_interaccion), col = "orange", pch = 19, main = "Modelo sin Interacción")
abline(h = 0, col = "red")
```




```
# Con interacción

plot(resid(modelo_con_interaccion), col = "lightgreen", pch = 19, main = "Modelo con Interacció
n")
abline(h = 0, col = "red")
```



- **¿Cuáles son las diferencias y similitudes de estos gráficos con respecto a los que ya habías analizado?**

Me parece que en este caso, tanto la recta como los datos se muestran más estables (?) (menos inclinados.)

- **Estos gráficos, ¿cambian en algo las conclusiones que ya habías obtenido?**

Sí. Yo creo que al principio no tenía mucha idea sobre qué es el buen desempeño de un modelo, y después de todas estas pruebas que se hicieron, y al poder compararlos, me di cuenta que el mejor modelo es el de hombres ya que no hay tanta dispersión en los datos y mostró un desempeño decente. Igualmente, el de mujeres pasó las pruebas, pero no fue el mejor.

5. Emite una conclusión final sobre el mejor modelo de regresión lineal que conjunte lo que hiciste en las tres partes de esta actividad.

Finalmente, la conclusión final es que el mejor modelo en este caso es el de los hombres, y después el de las mujeres. Por otro lado, los modelos con y sin interacción no fueron muy eficientes.

Intervalos de confianza

1. Con los datos de las estaturas y pesos de los hombres y las mujeres construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción de Y para el mejor modelo seleccionado.

```
library(ggplot2)
```

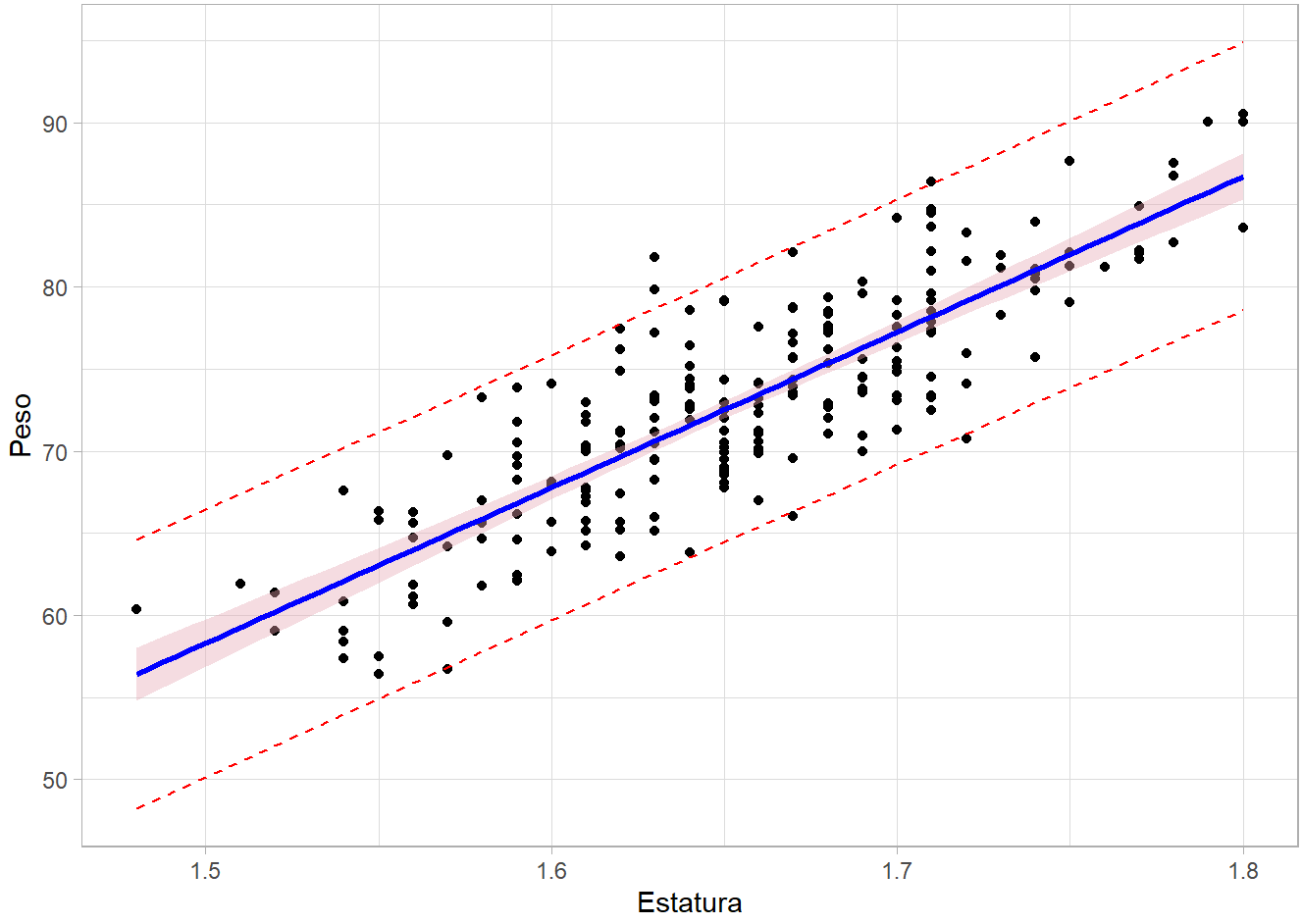
```
# Hombres

predH = predict(modelo_hombres, interval = "prediction",level = 0.97)
```

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

```
datosH = cbind(hombres, predH)

ggplot(datosH, aes(x = Estatura, y = Peso)) + geom_point() + geom_line(aes(y = lwr), color = "re
d", linetype = "dashed") + geom_line(aes(y = upr), color = "red", linetype = "dashed") + geom_sm
ooth(method = lm, formula = y ~ x, se = TRUE, level = 0.97, col = "blue", fill = "pink2") + them
e_light()
```



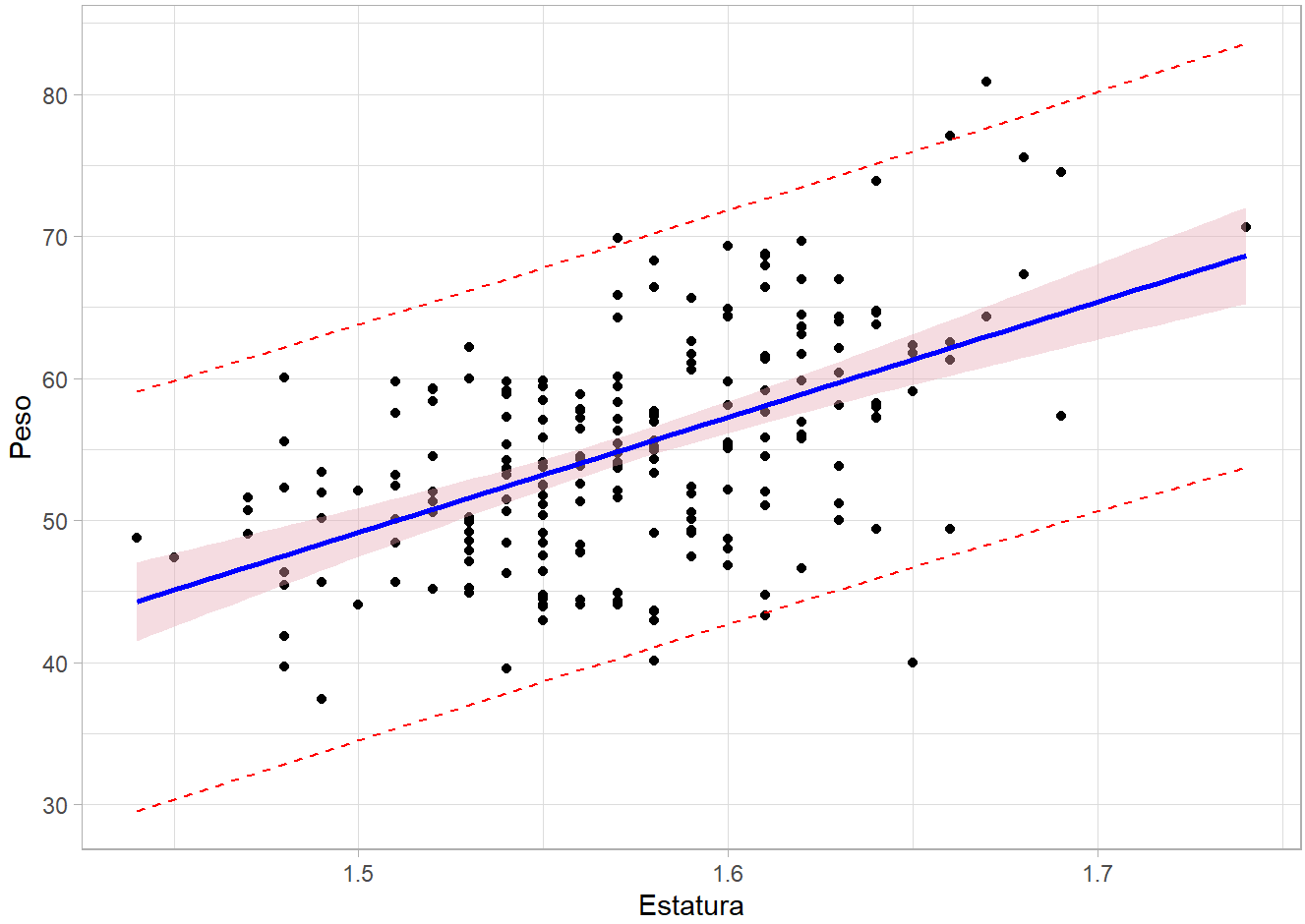

```
# Mujeres

predM = predict(modelo_mujeres, interval = "prediction",level = 0.97)
```

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

```
datosM = cbind(mujeres, predM)

ggplot(datosM, 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 = "pink
2") +
  theme_light()
```



2. Interpreta y comenta los resultados obtenidos.

Se puede concluir que en la gráfica del modelo de los hombres hay mejor dispersión en los datos, así como que los intervalos de predicción son más angostos que en el de las mujeres. Por otro lado, en el caso de los intervalos de confianza, igual modelos observar cómo son más cercanos a la recta, que en el caso de las mujeres.