12. Análisis de los errores

Juan Bernal

2024-09-04

Analiza la base de datos de estatura y peso de los hombres y mujeres en México y obten el mejor modelo de regresión para esos datos.

```
M = read.csv("Estatura-peso_HyM.csv")
MM = subset(M,M$Sexo=="M")
MH = subset(M,M$Sexo=="H")
```

El 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:

```
rl_h = lm(Peso~Estatura, MH) # Regresión lineal del Peso con base en la Estatura de Hombres
```

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

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

a. Normalidad de los residuos

Prueba de hipótesis:

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

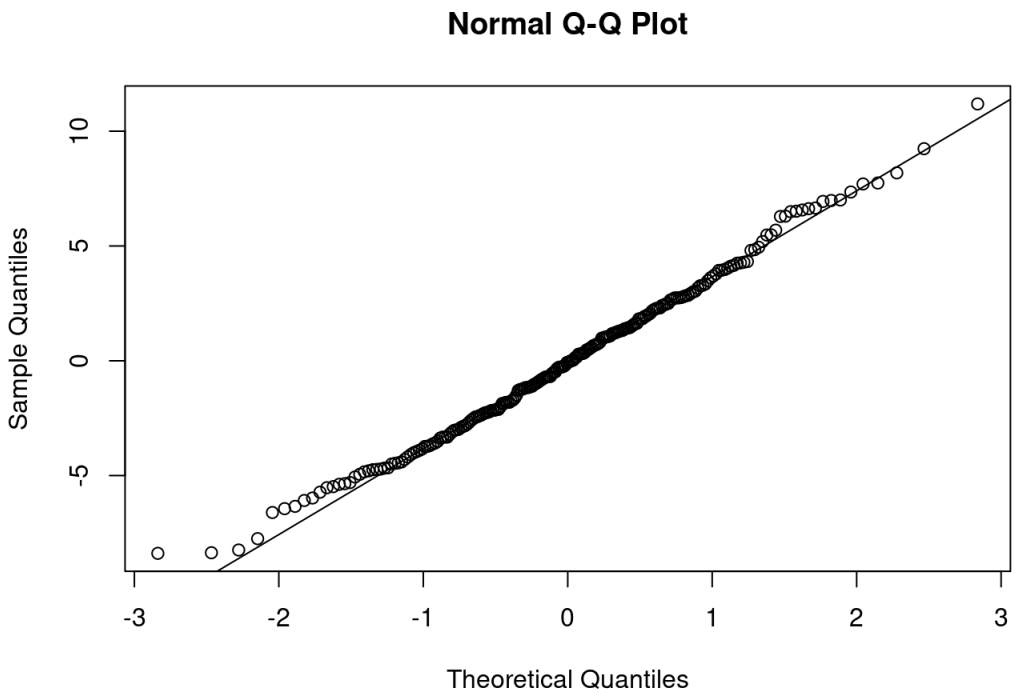
Regla de decisión: $p - value < \alpha$ se rechaza H_0

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

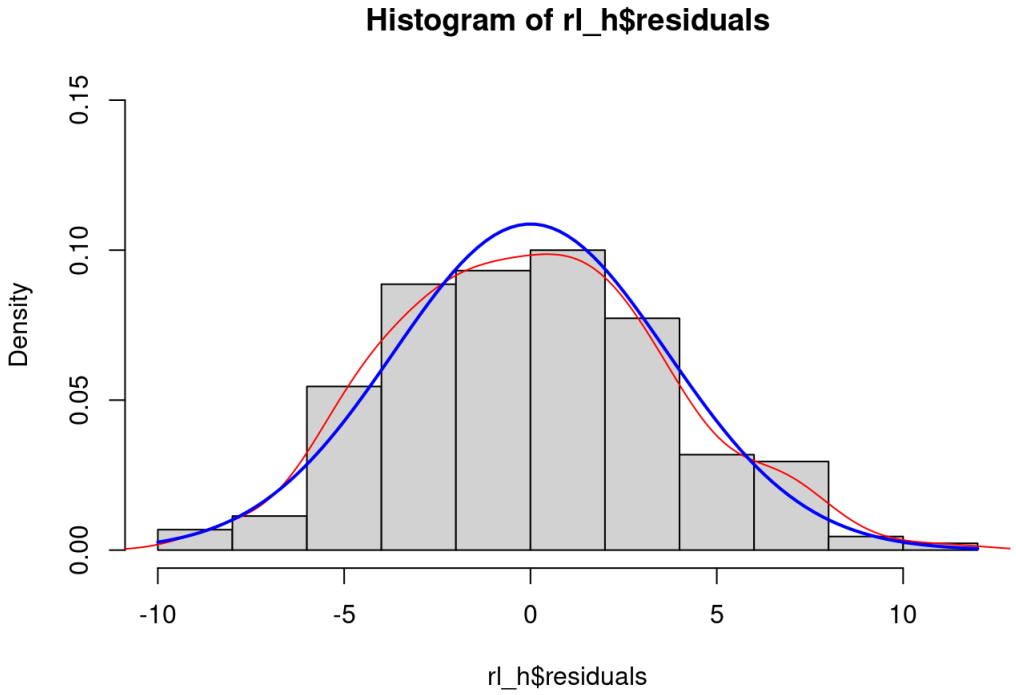
```
##
##  Anderson-Darling normality test
##
## data:  rl_h$residuals
## A = 0.3009, p-value = 0.5771
```

Dado que el p-value es mayor a un α de 0.03, no hay suficiente evidencia para rechazar la hipótesis inicial, por lo que los datos provienen de una población normal.

```
qqnorm(rl_h$residuals)
qqline(rl_h$residuals)
```




```
hist(rl_h$residuals,freq=FALSE, ylim=c(0,0.15))
lines(density(rl_h$residual),col="red")
curve(dnorm(x,mean=mean(rl_h$residuals),sd=sd(rl_h$residuals)), add=TRUE, col="blue",lwd=2)
```



Con las gráficas verificamos que efectivamente los residuos se distribuyen normalmente, pues el histograma se muestra como campana con los dato acumulándose en la media, y el qqplot muestra los datos casi perfectamente sobre la línea.

b. Verificación de media cero

Prueba de hipótesis:

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

Regla de decisión: $p - value < \alpha$ se rechaza H_0

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

```
##
##  One Sample t-test
##
## data:  rl_h$residuals
## t = -7.4842e-17, 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
## -1.851817e-17
```

Dado que el p-value es mayor a un alfa de 0.03, no hay suficiente evidencia para rechazar la hipótesis inicial, por lo que la media de los residuos es 0.

c. Homocedasticidad e independencia

Prueba de hipótesis para homocedasticidad:

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

Regla de decisión: $p - value < \alpha$ se rechaza H_0

Prueba de hipótesis para independencia:

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

Regla de decisión: $p - value < \alpha$ se rechaza H_0

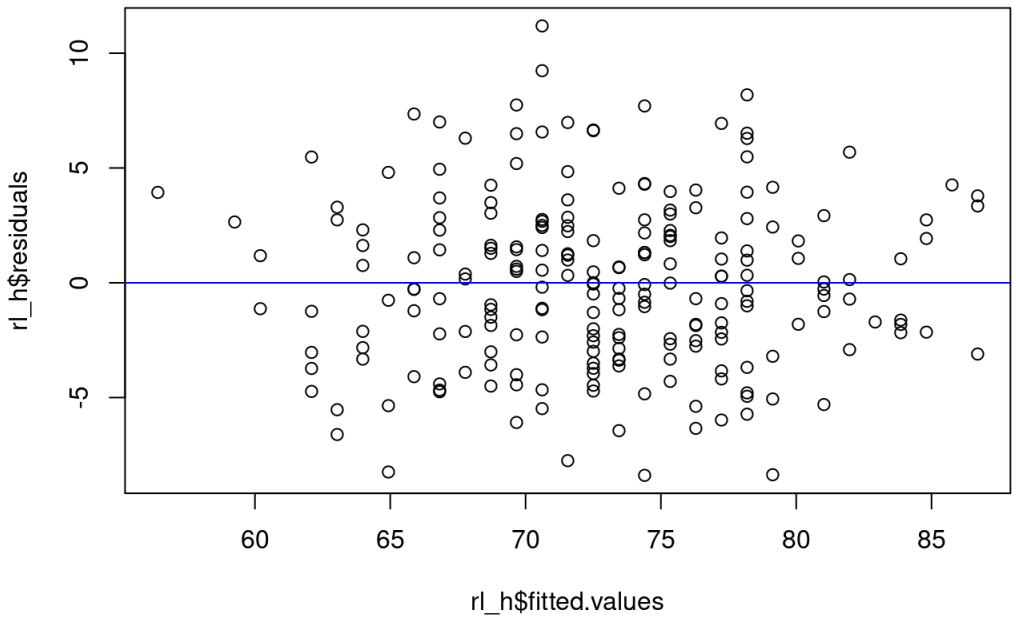
```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

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




```
library(lmtest)
dwtest(rl_h) # Test de Durbin-Watson para Independencia
```

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

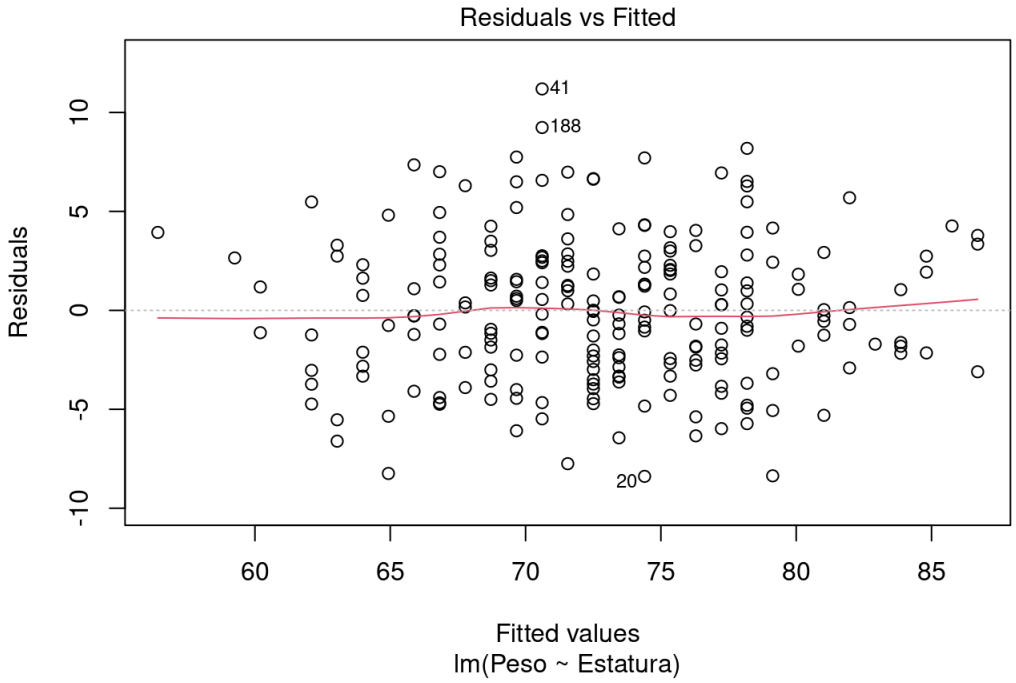
```
bptest(rl_h) # Test de Breusch-Pagan para Homocedasticidad
```

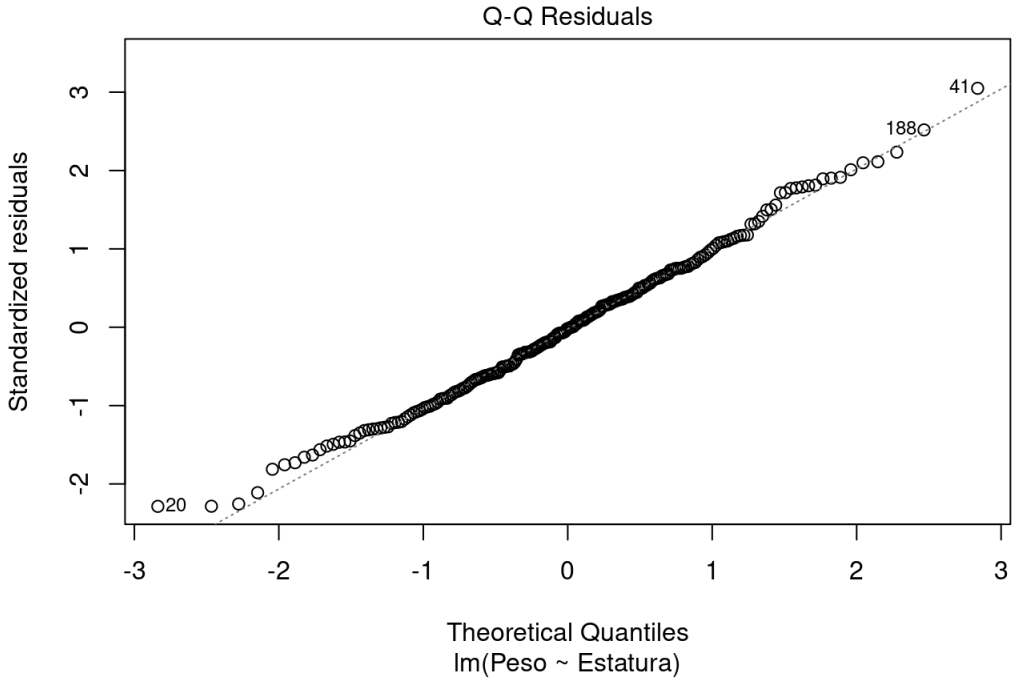
```
##
##  studentized Breusch-Pagan test
##
## data:  rl_h
## BP = 0.93324, df = 1, p-value = 0.334
```

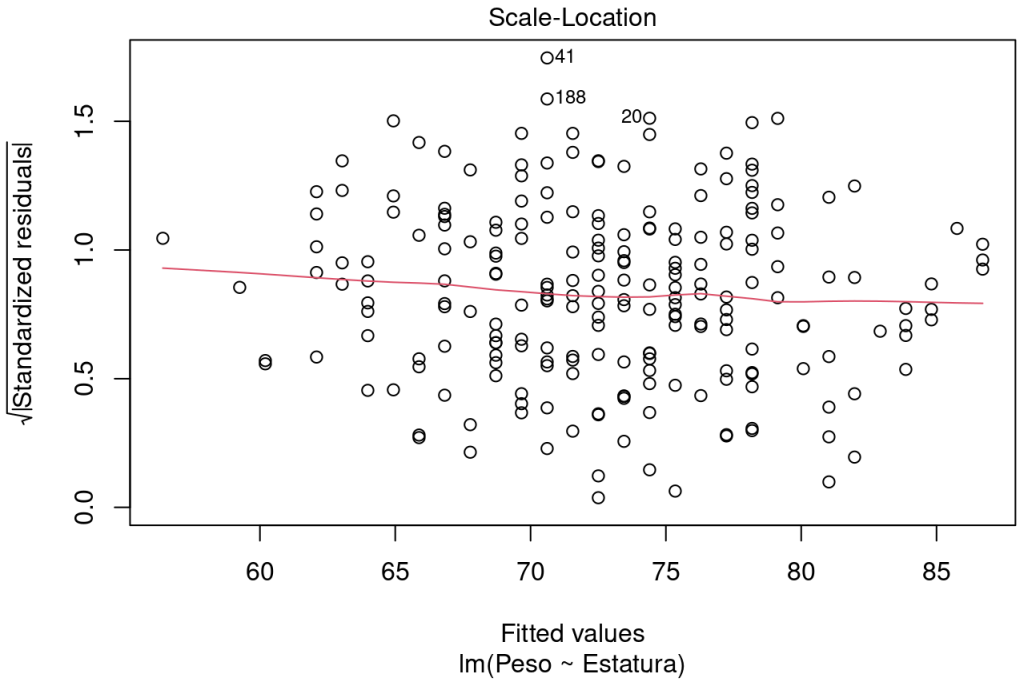
Dado que los p-values son mayores a 0.03, entonces asumimos las hipótesis iniciales, que nos indican que hay homocedasticidad e independencia en los errores.

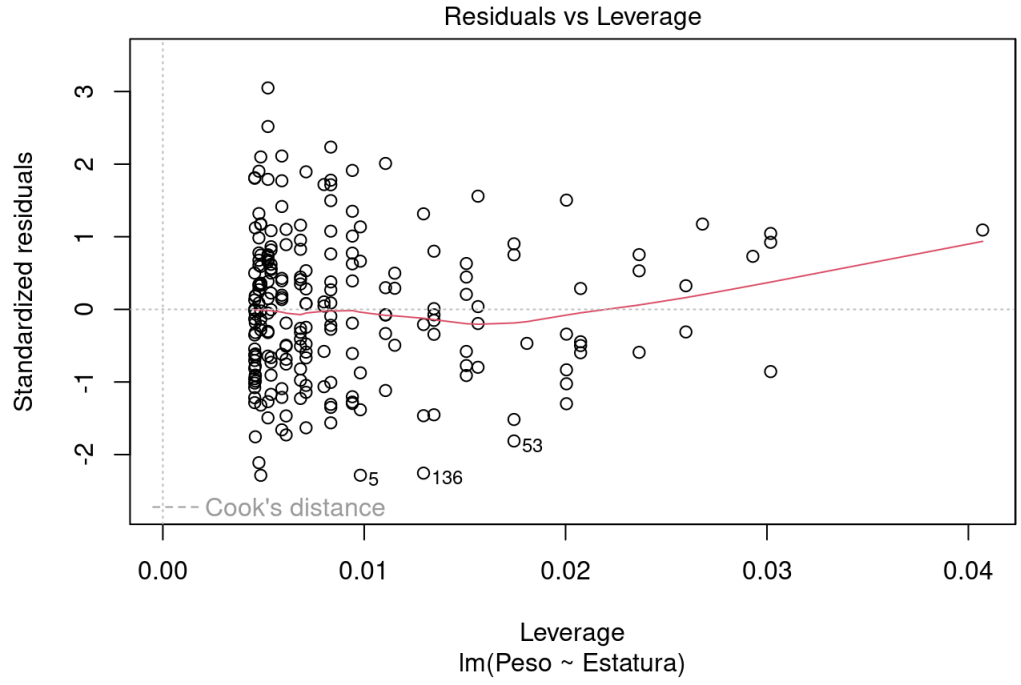
4. Utiliza el comando: plot(modelo). Observa las gráficas obtenidas y contesta:

```
plot(rl_h)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Peso ~ Estatura)
41
188
20
-10
-5
0
5
10
60
65
70
75
80
85


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura)
41
188
20
-2
-1
0
1
2
3





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

La gráficas son parecidas a las obtenidas anteriormente, pues todas demuestran la media 0 obtenida, el comportamiento normal de los datos, y la independencia y homocedasticidad de los mismos.

b. Estos gráficos, ¿cambian en algo las conclusiones que ya habías obtenido?

Los gráficos muestran el comportamiento de la media de los residuos, y podemos notar que la media se mantiene constantemente en 0, lo cual coincide con nuestras pruebas de normalidad.Además, la gráficad de “Residuals vs Leverage” nos muestra que hay datos influyentes en el modelo, es decir, datos que si quitamos podría mejorar el modelo.

Las conclusiones no cambian, si no que ahota contamos con evidencia más sólida.

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

El mejor modelo de regresión lineal para predicción del peso, es el que toma en consideración únicamente la estatura de los hombres. Esto debido al análisis de residuos que llevamos. El modelo con interacción de estatura y sexo tiene mejor R^2, pero el análisis de residuo demostró que no era normal, ni el modelo con estatura y sexo era normal. Por lo que, el mejor modelo con normalidad y R^2 fue el que toma como variable independiente la estatura de los hombres.

Intervalos de Confianza

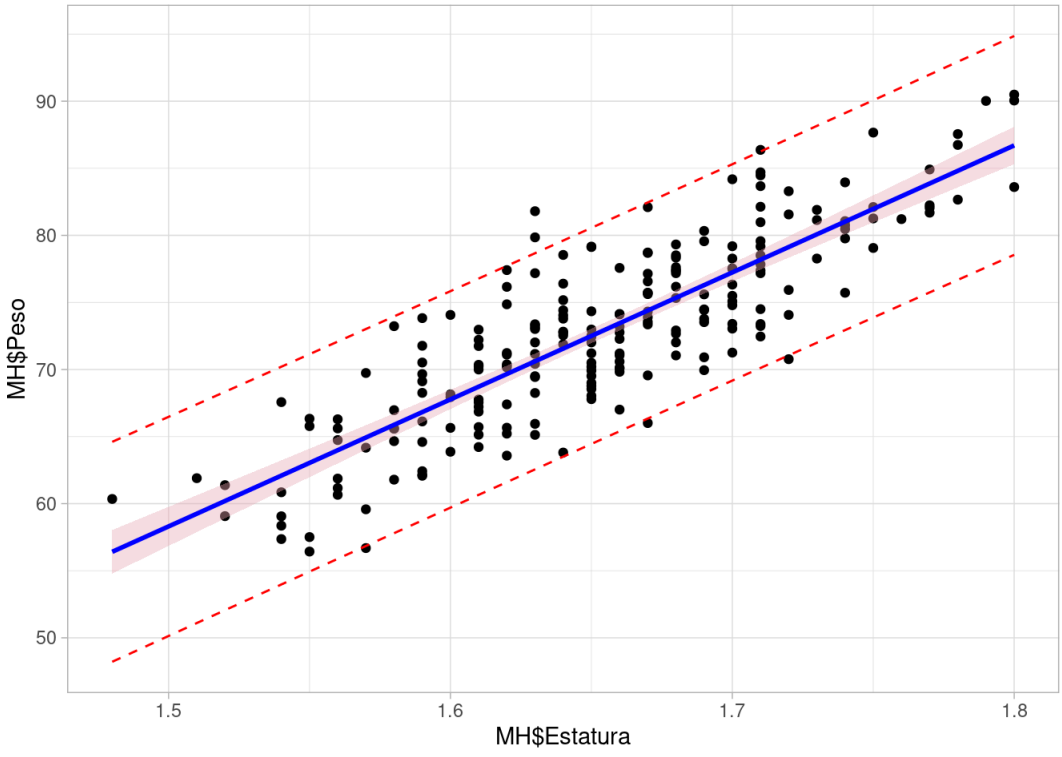
1. Con los datos de las estaturas y pesos de los hombres y las mujeres construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción de Y para el mejor modelo seleccionado.

```
Ip=predict(object=rl_h,interval="prediction",level=0.97)
```

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

```
datos1=cbind(MH,Ip)
```

```
library(ggplot2)
ggplot(datos1,aes(x=MH$Estatura, y=MH$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()
```



2. Interpreta y comenta los resultados obtenidos

En la gráfica se observa la línea azul central que es la línea de regresión lineal que mejor ajusta los datos, la banda rojo claro alrededor de la línea de tendencia representa el intervalo de confianza y las líneas punteadas rojas representan los límites de predicción, que muestran el rango dentro del cual se espera que caigan los valores individuales del peso para una estatura dada.

La gráfica muestra que existe una tendencia de que a mayor estatura corresponde un mayor peso, con cierta variabilidad alrededor de esta tendencia.