Actividad 7. Regresión Lineal

Luis Ángel Guzmán Iribe - A01741757

2023-08-29

```
library(readr)
M = read.csv("Estatura-peso_HyM-2.csv")
```

La recta de mejor ajuste.

```
MM = subset(M,M$Sexo=="M")
MH = subset(M,M$Sexo=="H")
M1=data.frame(MH$Estatura,MH$Peso,MM$Estatura,MM$Peso)

n=4 #número de variables
d=matrix(NA,ncol=7,nrow=n)
for(i in 1:n){
  d[i,]<-c(as.numeric(summary(M1[,i])),sd(M1[ ,i]))
}
m=as.data.frame(d)
```

1. Obtén la matriz de correlación de los datos que se te proporcionan. Interpreta.

```
cor(M1)
```

```
##              MH.Estatura     MH.Peso  MM.Estatura    MM.Peso
## MH.Estatura 1.0000000000 0.846834792 0.0005540612 0.04724872
## MH.Peso     0.8468347920 1.000000000 0.0035132246 0.02154907
## MM.Estatura 0.0005540612 0.003513225 1.0000000000 0.52449621
## MM.Peso     0.0472487231 0.021549075 0.5244962115 1.00000000
```

```
cor(M$Estatura, M$Peso)
```

```
## [1] 0.8032449
```

Con la matriz de correlación podemos apreciar realmente 2 relaciones importantes, estatura y peso para hombres, y estatura y peso para mujeres. Esto sugiere que únicamente existe una correlación aislada entre estas variables, podemos apreciar el mismo fenomeno cuando se calcula la correlación entre, la estatura y el peso de la población en general, obteniendo un coeficiente de correlación de 0.803, lo cual me lleva a pensar que la correlación más importante para un modelo en la población general, es esta.

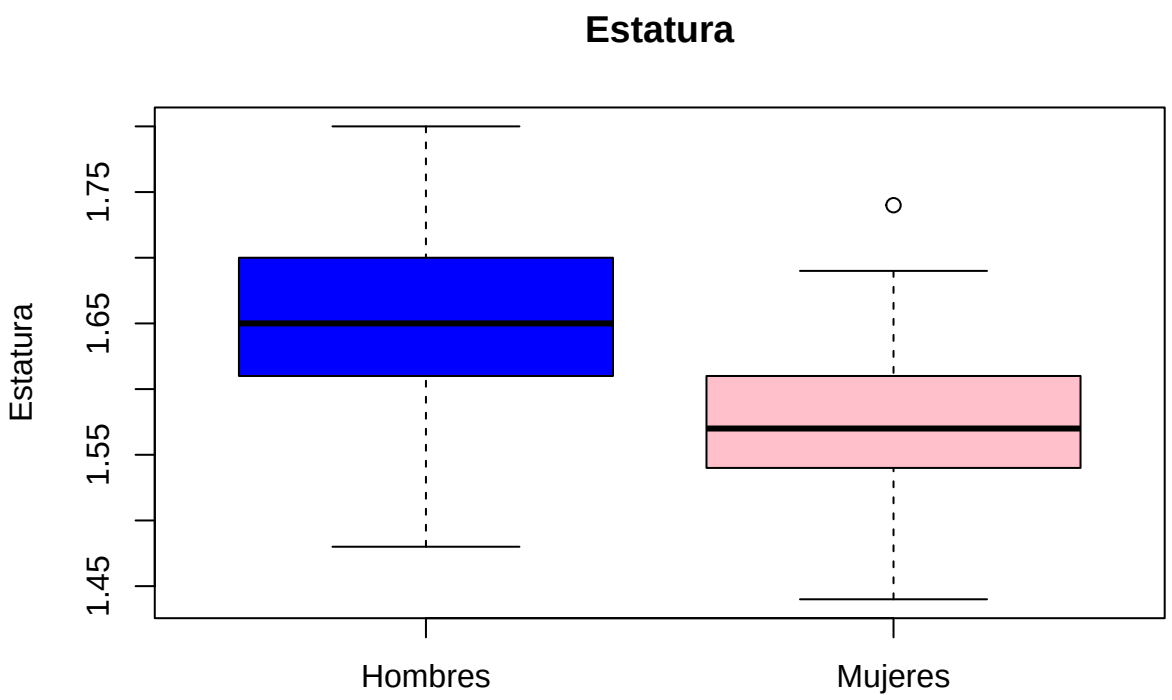
2. Obtén medidas (media, desviación estándar, etc) que te ayuden a analizar los datos.

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

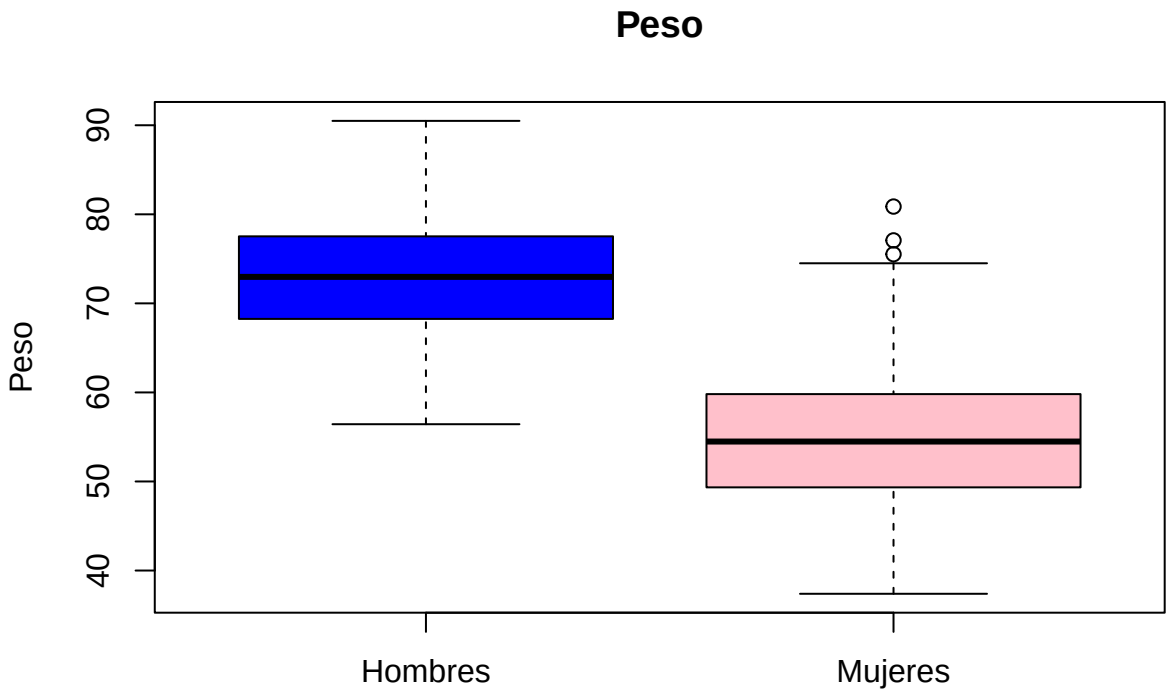
```
##            Minimo      Q1 Mediana     Media      Q3 Máximo   Desv Est
```

```
## 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
```

```
boxplot(M$Estatura~M$Sexo, ylab="Estatura", xlab="", col=c("blue","pink"), names=c("Hombres", "Mujeres")
```




```
boxplot(M$Peso~M$Sexo, ylab="Peso",xlab="", names=c("Hombres", "Mujeres"), col=c("blue","pink" ), main=
```



###

3. Encuentra la ecuación de regresión de mejor ajuste.

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

3.1 Realiza la regresión entre las variables involucradas.

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

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

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

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

3.2 Verifica el modelo. Modelo sin interacción.

```r
summary(A)
```

```
##
## Call:
## lm(formula = M$Peso ~ M$Estatura + M$Sexo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.9505  -3.2491   0.0489   3.2880  17.1243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -74.7546     7.5555  -9.894   <2e-16 ***
## M$Estatura   89.2604     4.5635  19.560   <2e-16 ***
## M$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
```

Modelo con interacción.

```r
B = lm(M$Peso~M$Estatura * M$Sexo)
summary(B)
```

```
##
## Call:
## lm(formula = M$Peso ~ M$Estatura * M$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 ***
## M$Estatura            94.660      5.882  16.092   <2e-16 ***
## M$SexoM               11.124     14.950   0.744    0.457
## M$Estatura:M$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
```

3.2.1 Verifica la significancia del modelo con un alfa de 0.03. Modelo sin interacción.

Dado que el valor p de nuestro modelo es 2.2e-16, o en otras palabras, funcionalmente 0, podemos afirmar que el modelo es significativo, rechazando la hipotesis nula de $\beta = 0$, sugiriendo que al menos uno de estos coeficientes es útil para explicar Y.

Modelo con interacción.

Nuestro valor p para el modelo que toma en cuenta la interacción de generos es también funcionalmente 0, por lo que podemos rechazar la hipotesis nula, y también afirmar que al menos uno de los coeficientes es significativo para explicar el modelo.

3.2.2 Verifica la significancia de $\hat{B}_i$ con un alfa de 0.03. Modelo sin interacción.

Podemos apreciar que para todas los coeficientes $\hat{B}_i$ el valor Pr(>|t|) es representado como <2e-16, de nuevo, funcionalmente es 0, definitivamente por debajo de nuestro margen alpha = 0.03, lo que nos permite indicar que los 3 coeficientes fueron encontrados como significativos.

Modelo con interacción.

En este caso, este modelo difiere del anterior, en este caso, el atributo de sexo no es significativo para explicar el comportamiento de los datos, con un valor p de 0.457, muy por arriba de nuestro margen de 0.03, lo mismo sucede con la relación entre sexo y estatura.

3.2.1 Verifica el porcentaje de variación explicada por el modelo. Modelo sin interacción.

El porcentaje de variación explicada por el modelo es dado por el valor R^2, en este caso 0.7827.

Modelo sin interacción

El valor R^2 es de 0.7832.

4.Dibuja el diagrama de dispersión de los datos y la recta de mejor ajuste.

Modelo

```
# Para mujeres SexoM = 1
cat("Para mujeres", "\n")
```

```
## Para mujeres
```

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

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

```r
# Para mujeres SexoM = 2
cat("Para hombres", "\n")
```

```
## Para hombres
```

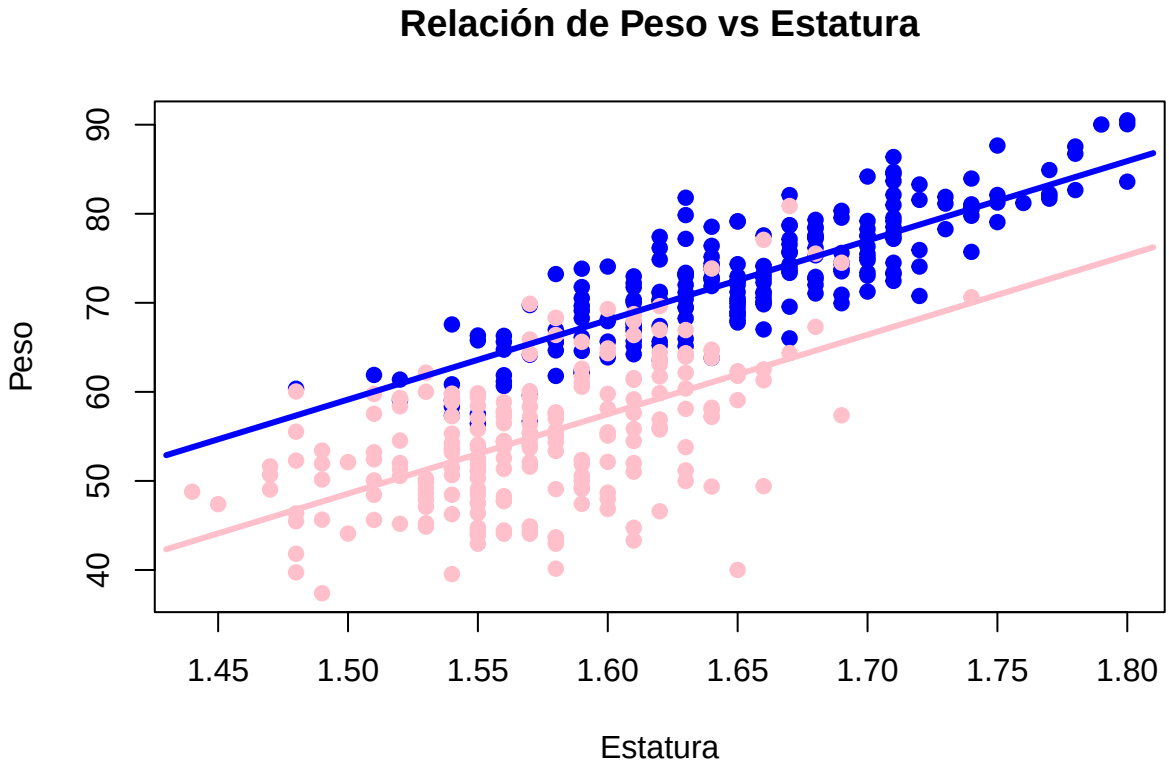
```r
cat("Peso = ", b0, "+", b1, "Estatura\n")
```

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

Gráfica

```r
Ym = function(x){b0+b2+b1*x}
Yh = function(x){b0+b1*x}

colores = c("blue","pink")
plot(M$Estatura, M$Peso, col=colores[factor(M$Sexo)], pch=19, ylab="Peso", xlab="Estatura", main="Relac
x = seq(1.43, 1.81, 0.01)
lines(x, Ym(x), col="pink", lwd=3)
lines(x, Yh(x), col="blue", lwd=3)
```



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

Basados en el analisis, podemos concluir que el modelo sin interacción es el más apto para describir el conjunto de datos con el que se trabaja. Considero que esto es así porque en este modelo todas las variables resultaron ser significativas para la predicción de los datos. Este modelo, en el que no se considera interacción, resulta en 2 modelos que se obtienen al considerar la naturaleza binaria de la variable sexo, lo cual lleva a 2 diferentes intersecciones para los modelos. Se hablará de esto con más detalle en el siguiente punto.

6. Interpreta en el contexto del problema:

6.1 ¿Qué información proporciona B_0 sobre la relación entre la estatura y el peso de hombres y mujeres? B_0 representa la intersección del modelo, o el punto en el que este toca 0, en este caso no tiene

sentido, dado que necesitaríamos un peso negativo y no tiene sentido que exista una persona con estatura 0, pero de lo que si habla, es que la relación entre el peso y estatura de los hombres es la misma, solo que los hombres tienen un mayor peso para su estatura de manera natural, los hombres tienen casi 10 kilos más que una mujer con la misma estatur.

6.2 ¿Cómo interpretas B_1 en la relación entre la estatura y el peso de hombres y mujeres?
B_1 representa la pendiente del modelo, o la proporción con la que crece el peso en función de la estatura. En el modelo, podemos apreciar que este valor es identico para hombres y mujeres, es decir, por cada centimetro extra de estatura, los hombres y mujeres ganan el mismo peso, solo sucede que los hombres empiezan con más peso desde el inicio.

Validación del Modelo

1. Retoma el notebook en el que realizaste el análisis de regresión que encontraste 'La recta de mejor ajuste'

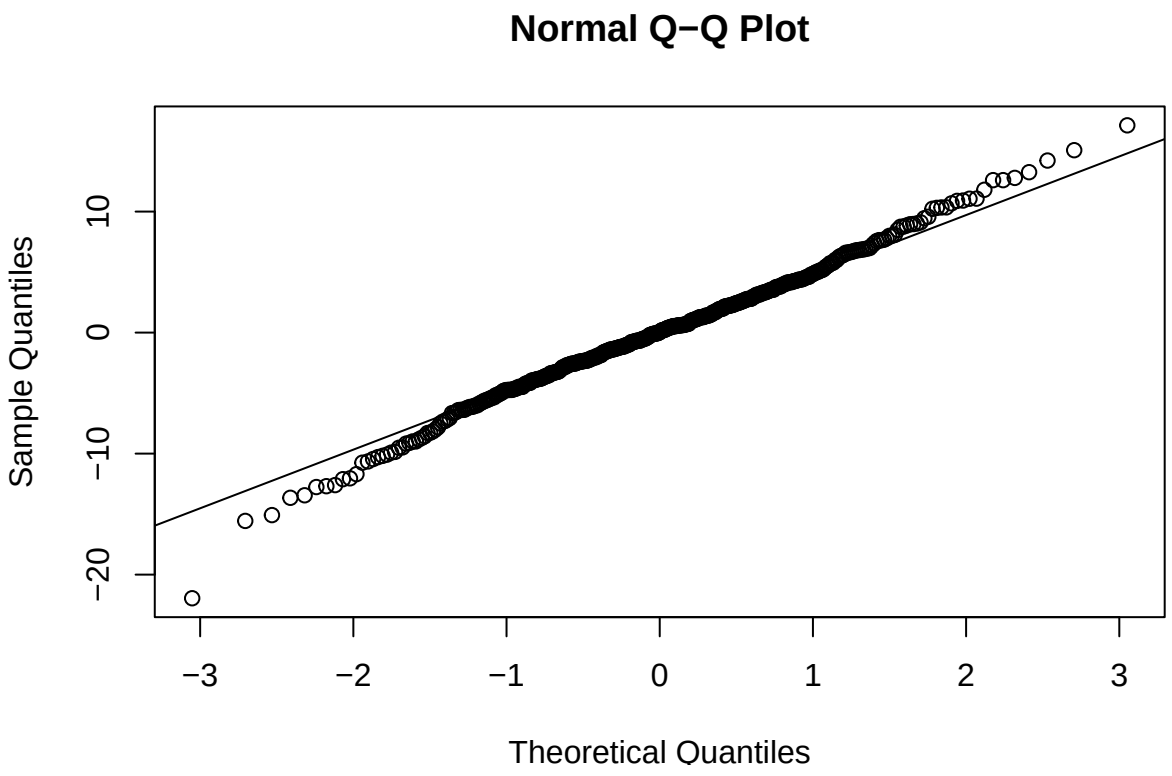
2. Analiza si el (los) modelo(s) obtenidos son apropiados para el conjunto de datos. Realiza el análisis de los residuos:

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

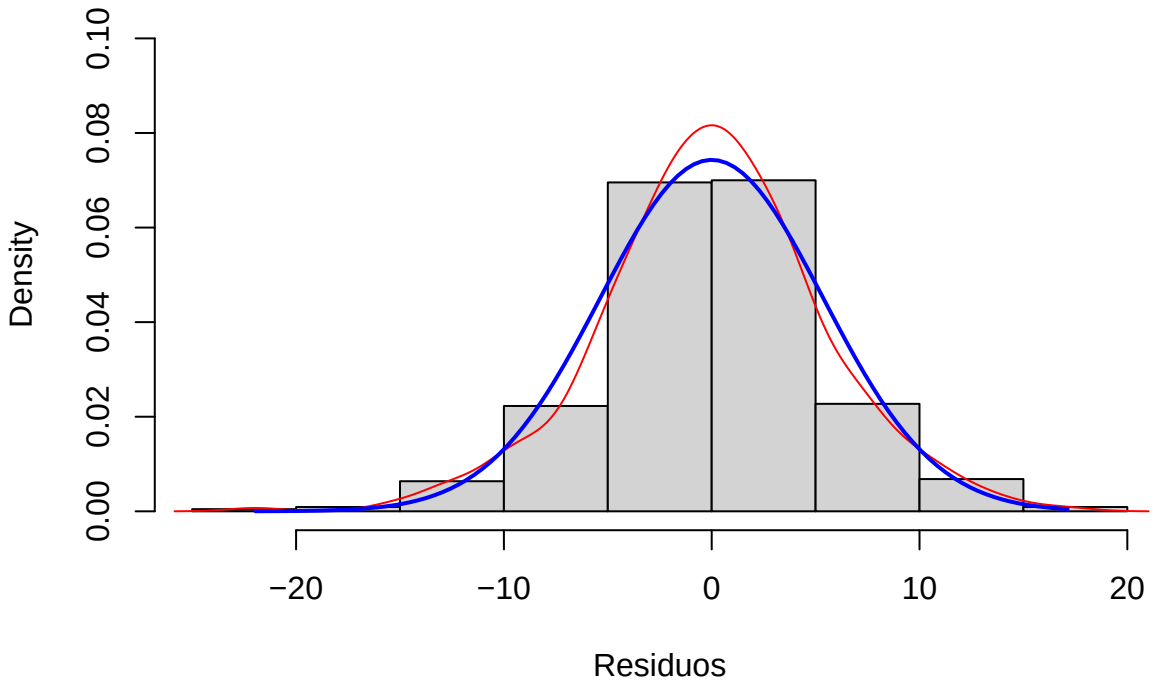
2.1 Normalidad de los residuos

```
##
##  Anderson-Darling normality test
##
## data:  A$residuals
## A = 0.79651, p-value = 0.03879
```

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




```
hist(A$residuals,freq=FALSE, ylim = c(0, 0.1), xlab = "Residuos")
lines(density(A$residual),col="red")
curve(dnorm(x,mean=mean(A$residuals),sd=sd(A$residuals)), from=min(A$residuals), to=max(A$residuals), ad
```



Si tomamos a alpha como 0.05, otro nivel común de significancia diferente al previamente usado 0.03, el conjunto de datos pasa la prueba de normalidad con un valor p de 0.03879. También podemos ver como en la qqplot

se ajusta satisfactoriamente a una distribución normal, aunque cabe decir que con más peso en las colas (una distribución platicúrtica)

2.2 Verificación de media cero

- Paso 1. Definir la hipotesis

  $H_0 : \mu = 0$

  $H_1 : \mu \neq 0$

- Paso 2. Regla de decisión

  Nivel de confianza $= 0.95$

  $\alpha = 0.05$

- Paso 3. Análisis del resultado

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

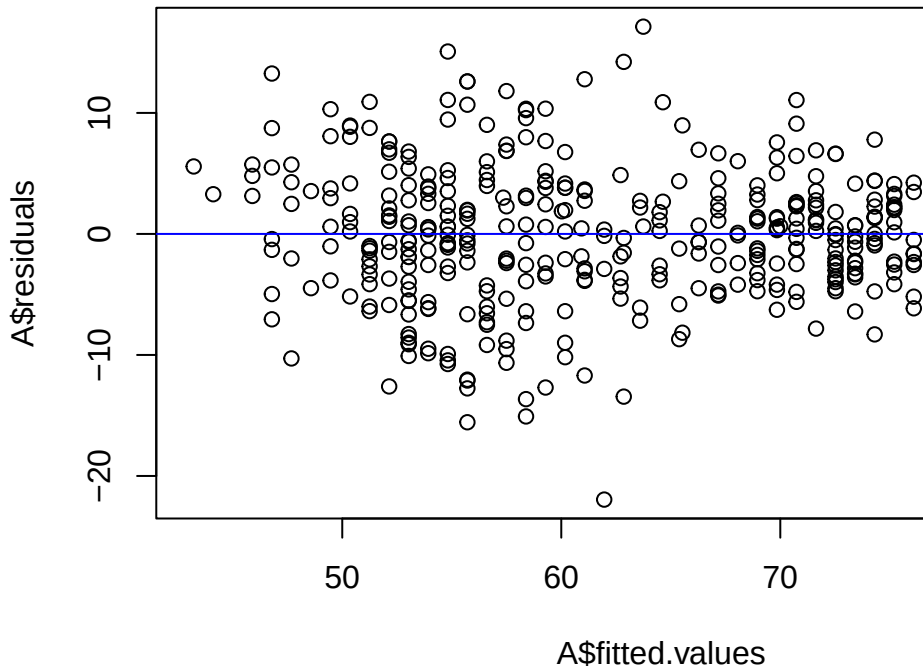
```
##
##  One Sample t-test
##
## data:  A$residuals
## t = 6.941e-16, df = 439, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5029859  0.5029859
## sample estimates:
##    mean of x
## 1.776357e-16
```

$$\alpha < p$$

- Paso 4. Conclusion

Como mi valor p es mayor que alfa ($1 > 0.05$) no puedo rechazar la hipotesis nula, lo que me permite afirmar con un alto grado de seguridad, que la hipotesis nula es correcta, y la media de los residuos es 0.

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



2.3 Homocedasticidad e independencia

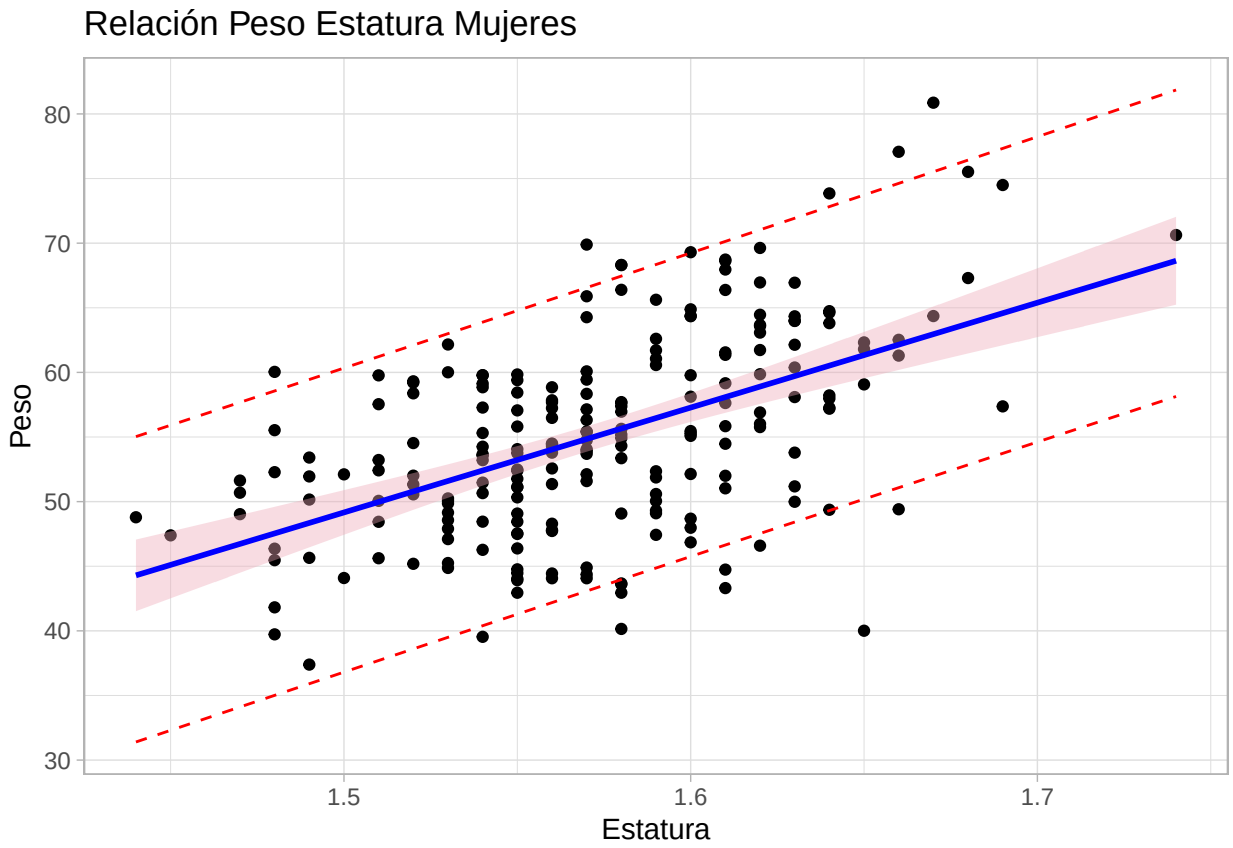
El gráfico muestra un conjunto de datos con simetría y homocedasticidad, los residuos se reparten de manera equitativa a lo largo de la gráfica, lo que lleva a pensar que existe independencia entre los residuos del modelo.

3. Conclusión final.

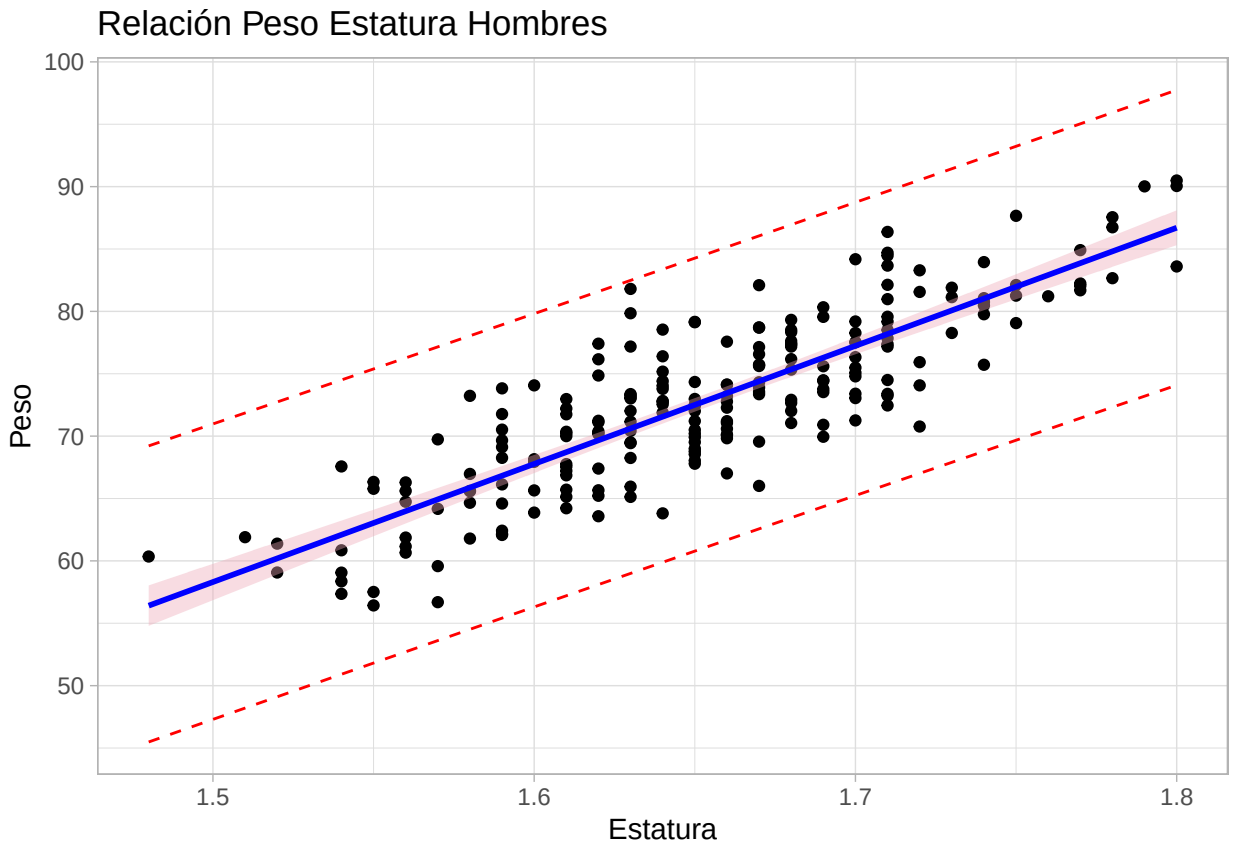
En conclusión, el modelo se ajusta de manera satisfactoria a los datos y a los resudios, en cuanto a estos ultimos provó un alto nivel de normalidad, una media igual a 0 con un alto nivel de exactitud, y una distribución Homocedastica e independiente de los datos.

Intervalos de confianza

```
Ip=predict(object=A,interval="prediction",level=0.97)
M2=cbind(M,Ip)
M2m = subset(M2, Sexo=="M")
M2h = subset(M2, Sexo=="H")
library(ggplot2)
ggplot(M2m,aes(x=Estatura,y=Peso))+
  ggtitle("Relación Peso Estatura Mujeres") +
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()
```




```
ggplot(M2h,aes(x=Estatura,y=Peso))+
  ggtitle("Relación Peso Estatura Hombres") +
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()
```



Los intervalos de confianza nos permiten el rango con el que podemos afirmar con un 97% de confianza que se encuentra la media de los datos en la población general. En este caso, podemos observar que el intervalo para las mujeres es ligeramente mayor que el de los hombres, lo que indicaría que existe una mayor variabilidad para los datos en el caso de la muestra de las mujeres que de los hombres.