Coeficiente de correlacion

Iván L. Hernández Buda

2023-08-29

Carga de los datos

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

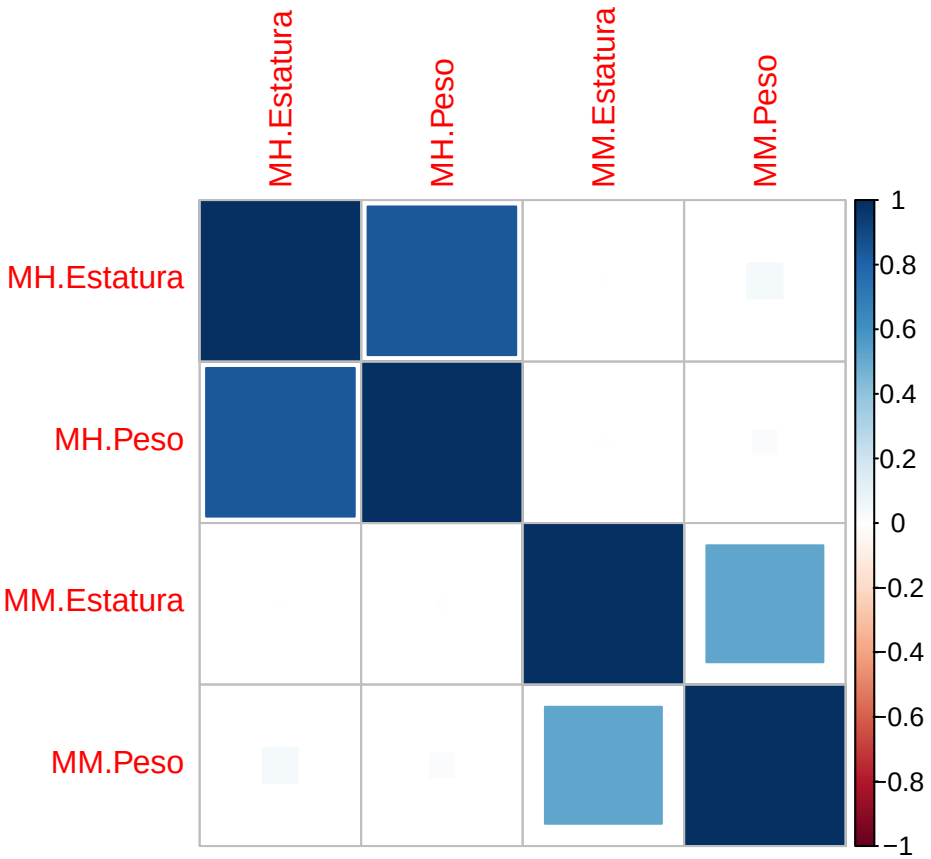
```
library(ggplot2)

M = read.csv("Estatura-peso_HyM.csv")

MM = subset(M,M$Sexo=="M") # Mujer
MH = subset(M,M$Sexo=="H") # Hombre
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
```

Matriz de correlación

A través de la matriz de correlación, podemos observar que la estatura de los hombres influye en su peso, y a su vez, la estatura de las mujeres influye en su peso, y viceversa.

```
corr_matrix <- cor(M1)
corrplot(corr_matrix, method = "square")
```



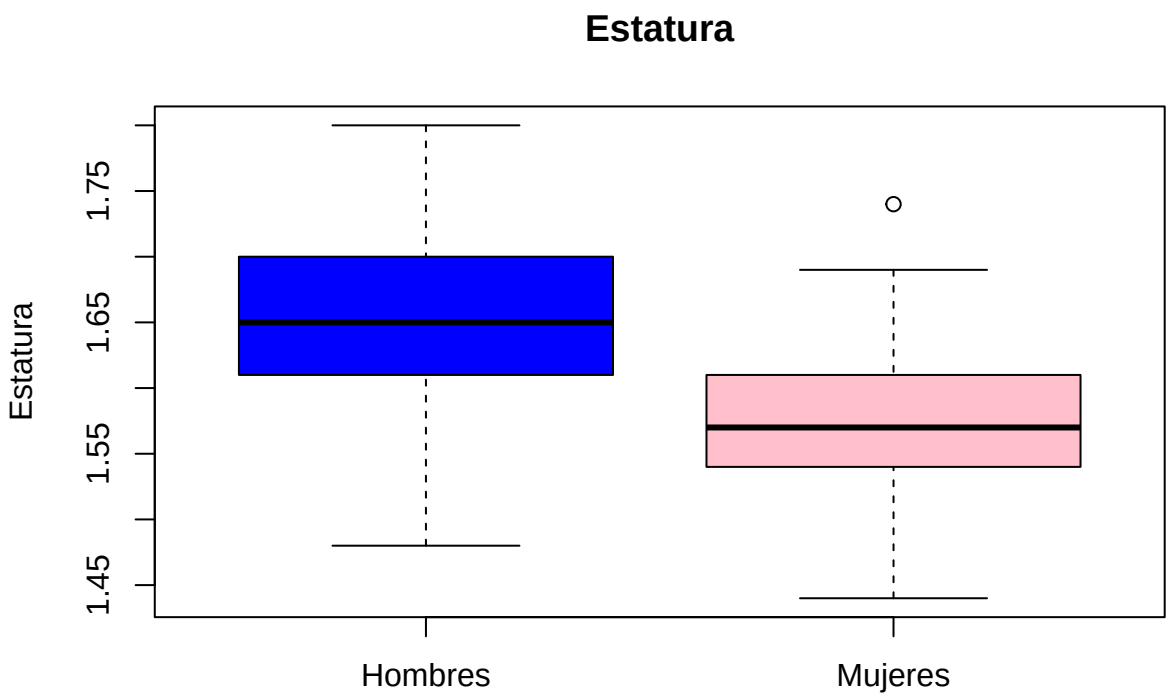
Breve análisis estadístico

Algo que cabe destacar, es que los datos de estatura para hombres y mujeres tiene muy poca varianza, pero por el otro lado, la varianza del peso de hombres y mujeres si tiene una varianza alta.

```
n = 4
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)

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"),
        main="Estatura")
```




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

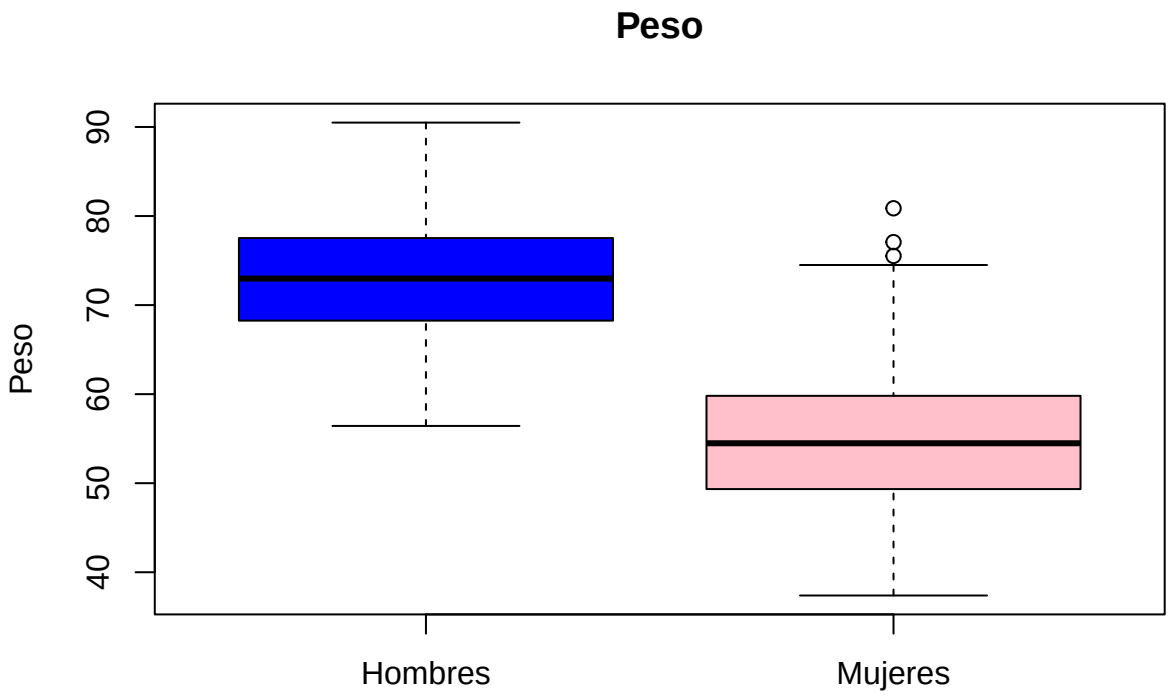
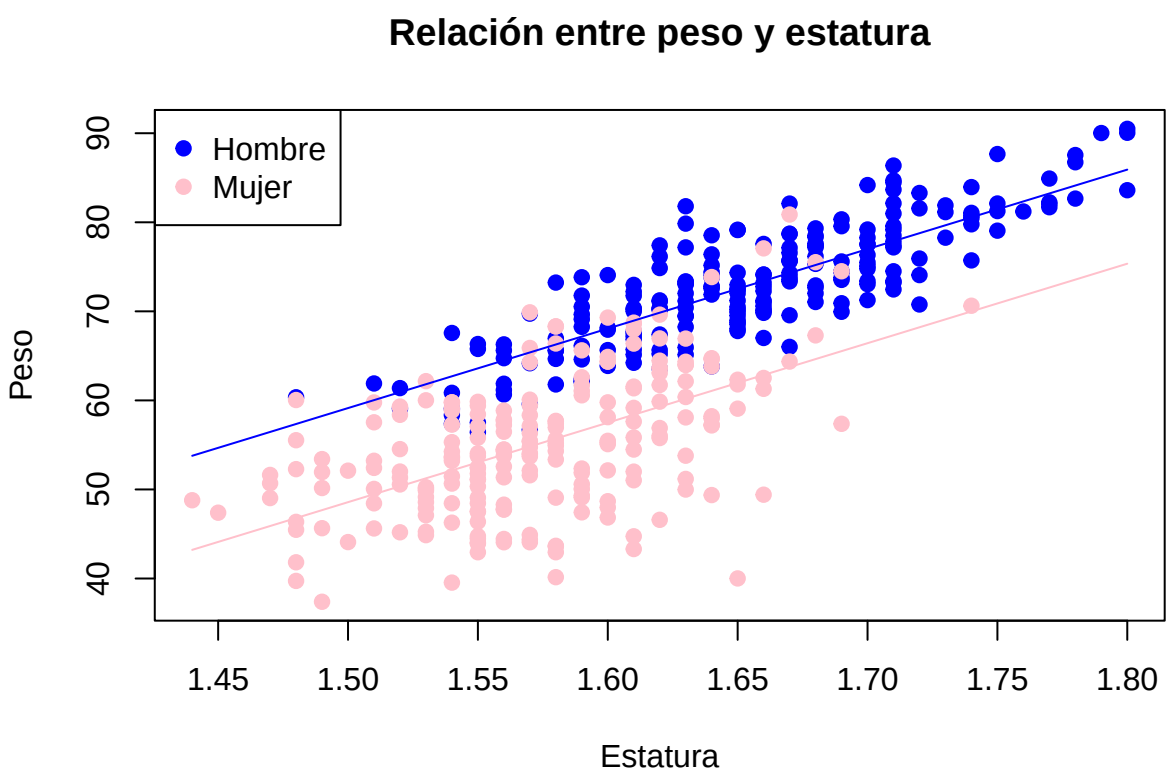


Diagrama de dispersión y recta de ajuste

En la gráfica de dispersión, se puede observar como los datos de las mujeres, en cuanto a peso y estatura, es encuentran menores que los de hombres. También, se puede apreciar que la línea de regresión suele ajustarse en la mayoría de los datos a la de los hombres.

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

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

Ym = function(x){b0 + b2 + b1*x}
Yh = function(x){b0 + b1*x}
color = c("blue","pink")
plot(M$Estatura, M$Peso, col=color[factor(M$Sexo)], pch=19,
     ylab="Peso", xlab='Estatura', main="Relación entre peso y estatura")
x = seq(min(M$Estatura), max(M$Estatura),0.01)
lines(x, Ym(x), col="pink")
lines(x, Yh(x), col="blue")
legend("topleft", legend=c("Hombre","Mujer"), pch=19, col=color)
```



Modelo con interacción

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

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

Ecuación de regresión que mejor se ajusta

Modelo sin interacción

```
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
```

Interpretación de coeficientes

Intercept

El intercept nos indica que cuando la estatura es 0 y el sexo es 0, el peso será de -83.7, lo cual no tiene sentido en este contexto.

Estatura

Esto indica que, manteniendo constante el valor de Sexo, por cada unidad adicional en la estatura de una persona, se espera que el peso aumente en 94.660 unidades. Este coeficiente muestra un valor 'p' muy cercano a 0, lo que denota alta significancia.

SexoM

En este caso, no tiene sentido en el contexto el valor del coeficiente, pero, también, al observar su valor 'p', podemos concluir que no es significante en lo absoluto.

Estatura×SexoM

Este coeficiente de interacción indica el efecto que tiene la estatura en el peso de ambos géneros. También presenta un valor 'p' muy alto, por lo que se concluye que no es significativo. En otras palabras, no hay evidencia sólida de una interacción significativa entre la estatura y el género en relación con el peso.

Valor estadístico R^2

El valor del R^2 ajustado es 0.7832, lo que indica que aproximadamente el 78.32% de la variabilidad en el peso se explica por la variabilidad en las variables independientes incluidas en el modelo, mientras, el porcentaje

restante se puede atribuir a la aleatoriedad.

Variable Dummy:

Mujer, Sexo = 0 Hombre, Sexo = 1

Significancia global e individual

$\alpha = 0.03$

Hipotesis

- Sobre el efecto del modelo (significación global): H_0: El modelo no tiene efecto significativo sobre la variable respuesta, es decir, $\beta_1 = beta_2 = 0$ H_1: El modelo tiene efecto significativo sobre la variable respuesta, es decir, Al menos un $\beta_i \neq 0$
- Sobre las (significación individual): Para β_1: H_0: El coeficiente β_1 no tiene efecto significativo sobre la variable respuesta, es decir, $\beta_1 = 0$ H_1: El coeficiente β_1 tiene efecto significativo sobre la variable respuesta, es decir, $\beta_1 \neq 0$ Para β_2: H_0: El coeficiente β_2 no tiene efecto significativo sobre la variable respuesta, es decir, $\beta_1 = 0$ H_1: El coeficiente β_2 tiene efecto significativo sobre la variable respuesta, es decir, $\beta_1 \neq 0$

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

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

cat("Ecuación de regresión de mejor ajuste:", "\n")
```

```
## Ecuación de regresión de mejor ajuste:
```

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

```
## Peso -74.7546 + 89.26035 * Estatura -10.56447 * Sexo
```

Donde, Estatura: es la variable independiente que representa la estatura de una persona. Sexo M: es una variable indicadora que es 1 si es masculino y 0 si es femenino.

Verificación de significancia de $\vec{\beta_1}$

```
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
```

Interpretación

```
cat("Regresión lineal para mujeres:", "\n")
```

```
## Regresión lineal para mujeres:
```

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

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

```
cat("Regresión lineal para hombres:", "\n") #sexo = 0
```

```
## Regresión lineal para hombres:
```

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

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

##¿Qué información β_0 proporciona sobre la relación entre la estatura y el peso de hombres y mujeres?

β_0 representa la intersección del modelo, es decir, es el¿valor base del peso cuando las variables Estatura y Sexo son iguales a 0. Pero en el contexto del problema no proporciona información directa sobre la relación entre la estatura y el peso de hombres y mujeres en este análisis específico

##¿Cómo interpretas β_1 en la relación entre la estatura y el peso de hombres y mujeres?

β_1 describe el cambio promedio en el peso debido a un cambio en la estatura, teniendo en cuenta las diferencias entre hombres y mujeres en el conjunto de datos. La magnitud de β_1 proporcionan información sobre cómo la estatura influye en el peso de hombres y mujeres en el análisis de regresión.

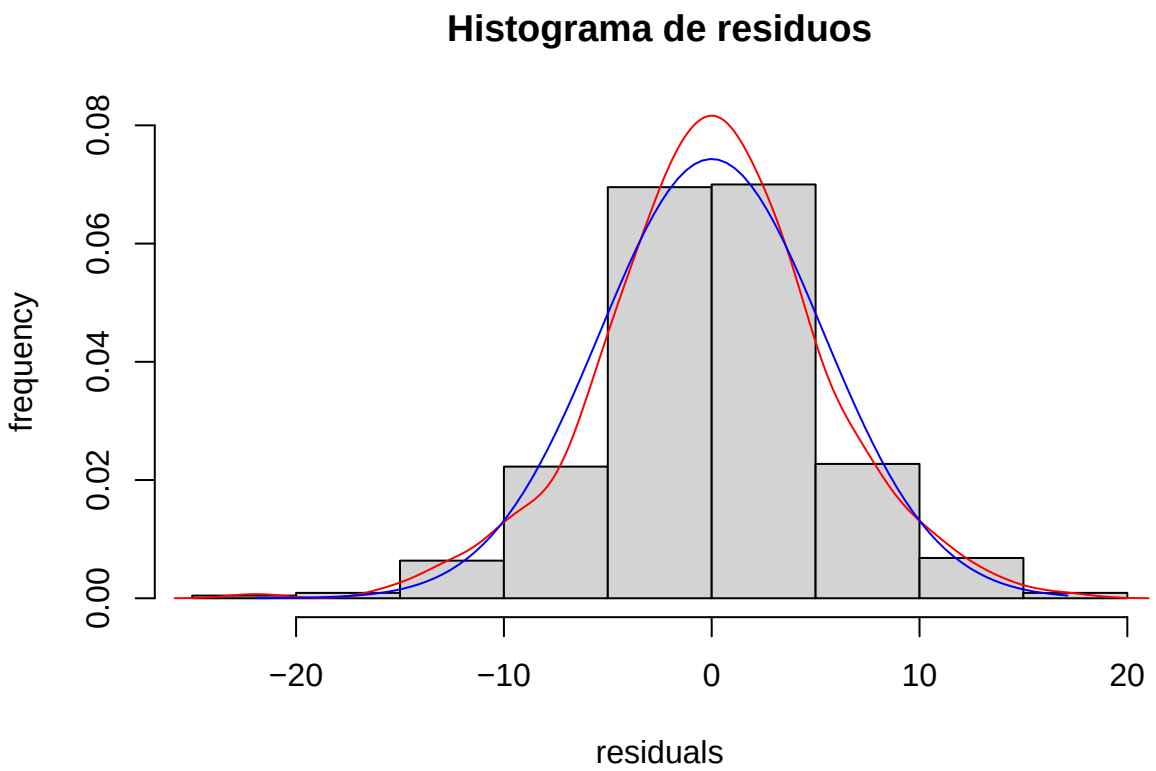
Análisis de validez

Normalidad y media cero

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

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

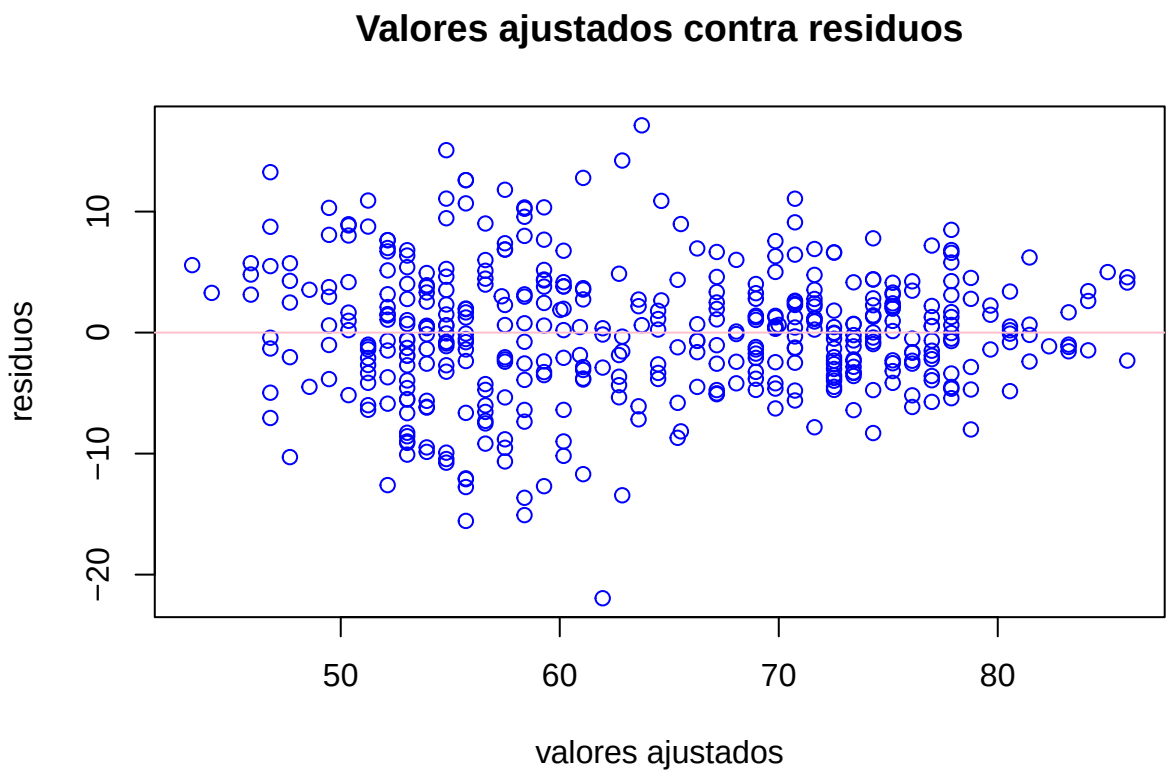
```
hist(A$residuals, freq=FALSE, ylim=c(0,0.08), ylab='frequency', xlab='residuals', main='Histograma de r
lines(density(A$residual), col="red")
curve(dnorm(x, mean=mean(A$residuals), sd=sd(A$residuals)),
      from=min(A$residuals), to=max(A$residuals), add=TRUE, col="blue")
```



Podemos observar en el histograma que, efectivamente, la media de la distribución de los residuos se encuentra en cero, pero, la normalidad se puede observar con el valor 'p' es algo mayor a $\alpha = 0.03$, por lo que rechazamps la hipótesis nula y decimos que la distribución de los residuos no es normal.

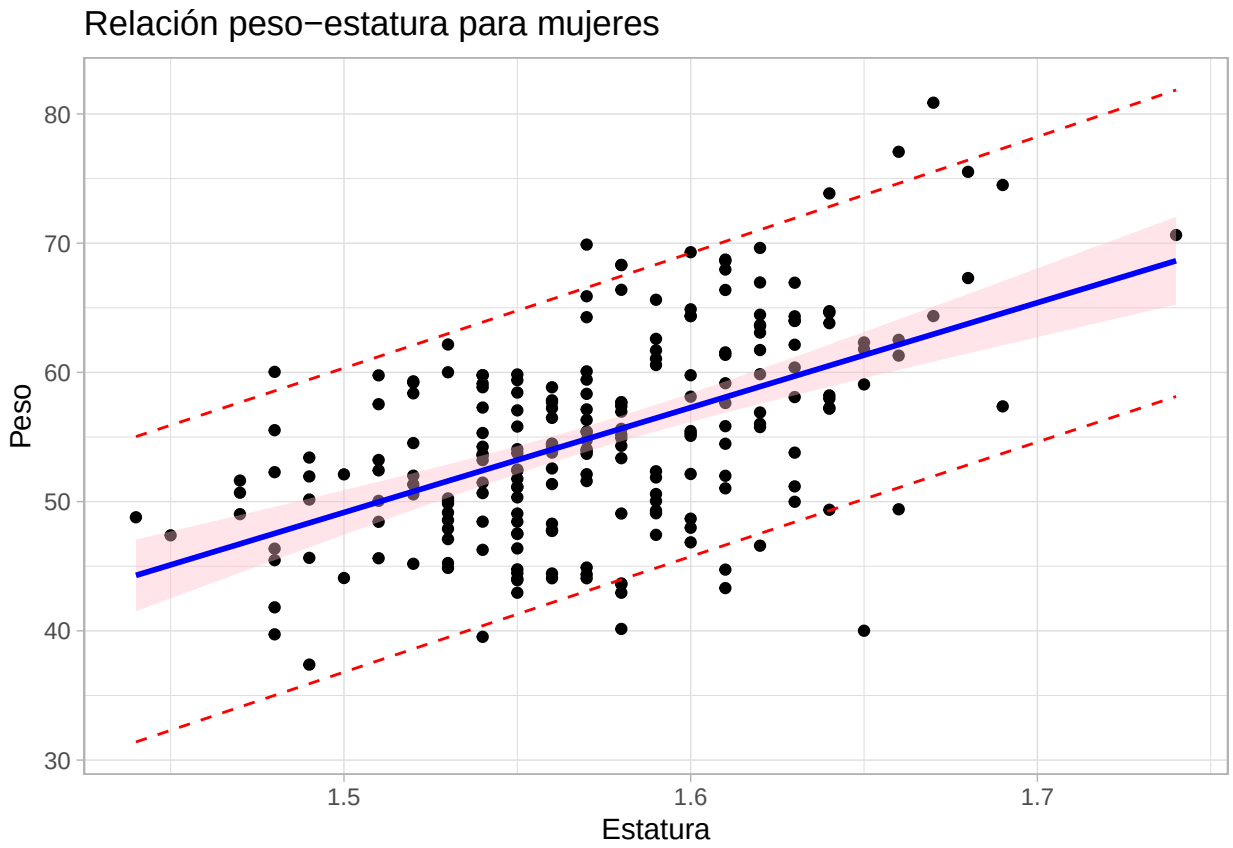
Homocedasticidad e independencia

```
plot(A$fitted.values, A$residuals, col="blue",
     xlab='valores ajustados', ylab='residuos', main='Valores ajustados contra residuos')
abline(h=0, col='pink')
```

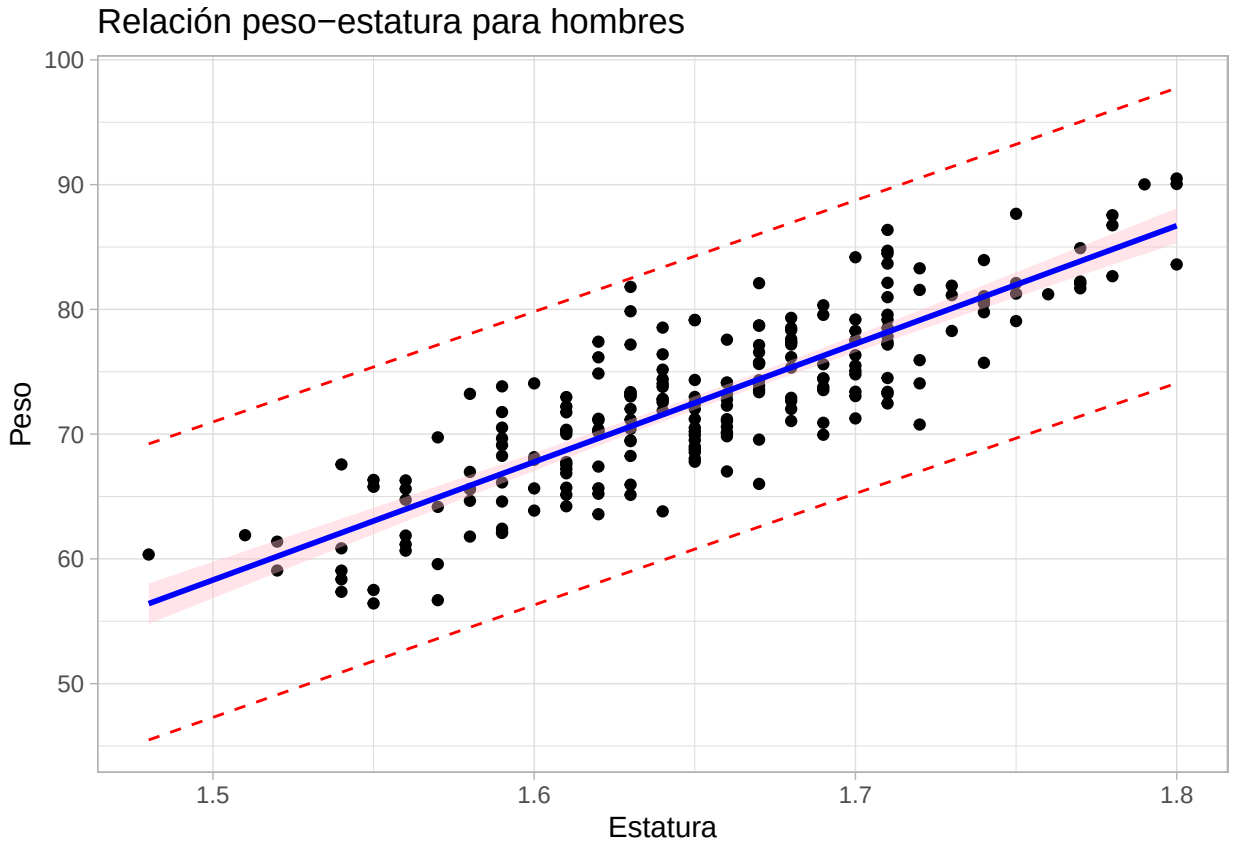


Intervalos de confianza

```
Ip = predict(object=A,interval="prediction",level=0.97)
datos = cbind(M, Ip)
datosM = subset(datos, Sexo=='M')
datosH = subset(datos, Sexo=='H')

ggplot(datosM,aes(x=Estatura,y=Peso))+
ggtitle("Relación peso-estatura para 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='pink') +
theme_light()
```




```
ggplot(datosH,aes(x=Estatura,y=Peso))+
ggtitle("Relación peso-estatura para 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='pink') +
theme_light()
```



##¿Cuáles son las diferencias y similitudes de estos gráficos con respecto a los que ya habías analizado?. Estos gráficos, ¿cambian en algo las conclusiones que ya habías obtenido?

Entre las similitudes visuales se encuentran que las rectas parecen tener la misma pendiente y el mismo bias. Pero, entre las diferencias, matemáticamente las ecuaciones de las rectas son diferentes con respecto al modelo con interacción, debido a ese parámetro extra que se agrega en las ecuaciones de las rectas y, que este modelo actual, no considera.