11. Regresión Lineal con Interacción

Juan Bernal

2024-08-30

La recta de mejor ajuste (Primera entrega)

Analiza la base de datos de estatura y peso 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")
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
```

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

```
cor(M1,method = "pearson")
```

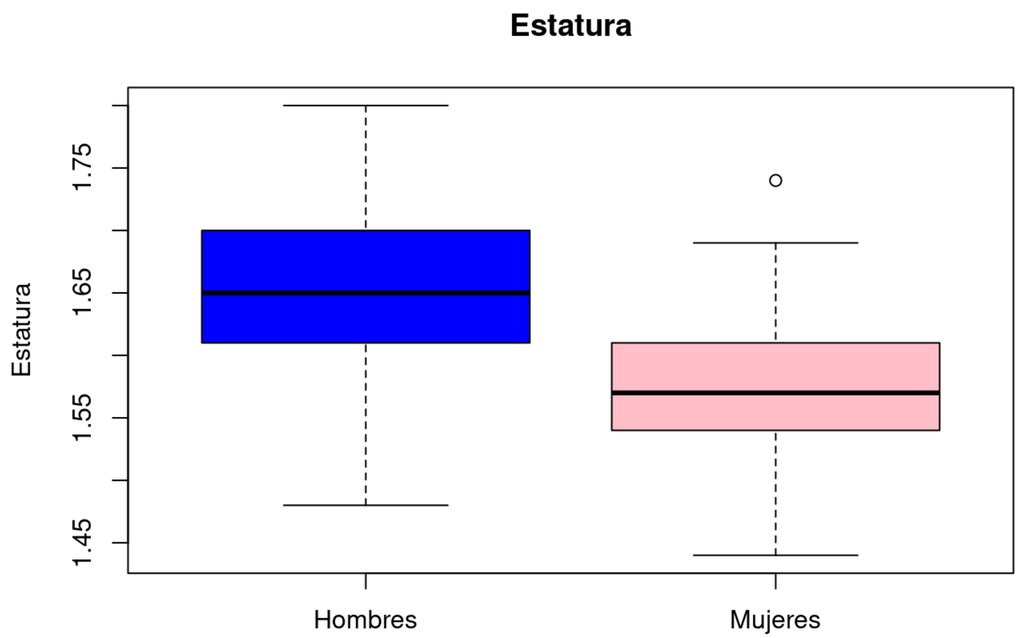
```
##              MH.Estatura     MH.Peso  MM.Estatura    MM.Peso
## MH.Estatura 1.0000000000 0.846834792 0.0005540612 0.04724872
## MH.Peso     0.8468347920 1.000000000 0.0035132246 0.02154907
## MM.Estatura 0.0005540612 0.003513225 1.0000000000 0.52449621
## MM.Peso     0.0472487231 0.021549075 0.5244962115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

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

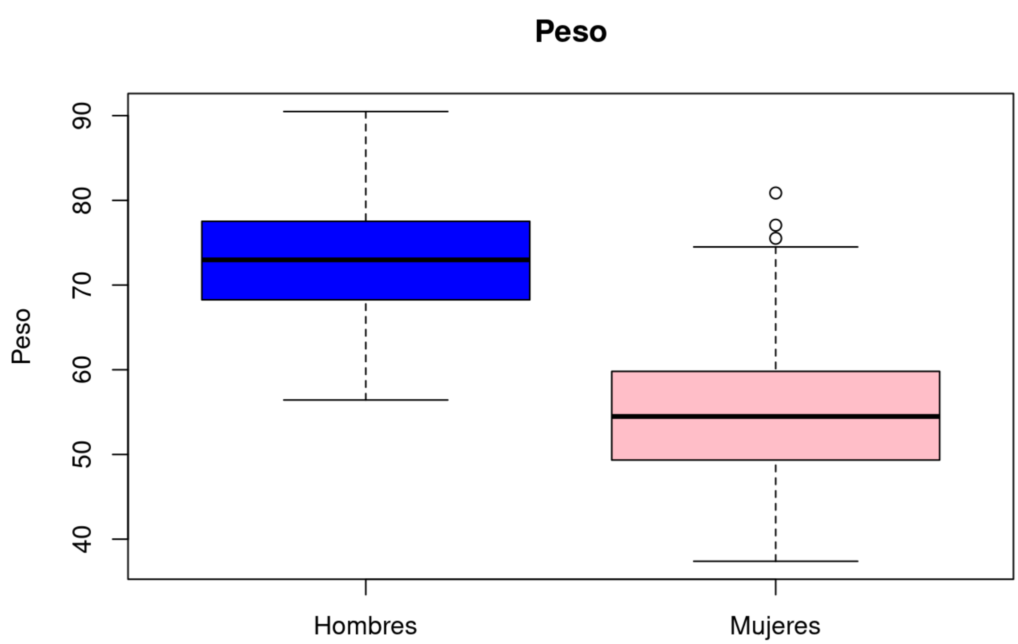
```
n=4 #número de variables
d=matrix(NA,ncol=7,nrow=n)
for(i in 1:n){
  d[i,]<-c(as.numeric(summary(M1[,i])),sd(M1[ ,i]))
}
m=as.data.frame(d)

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")
```



3. Encuentra la ecuación de regresión de mejor ajuste:

A. Realiza la regresión entre las variables involucradas

B. Verifica el modelo:

a. Verifica la significancia del modelo con un alfa de 0.03.

b. Verifica la significancia de β_i con un alfa de 0.03.

c. Verifica el porcentaje de variación explicada por el modelo.

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

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

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

Hipótesis:

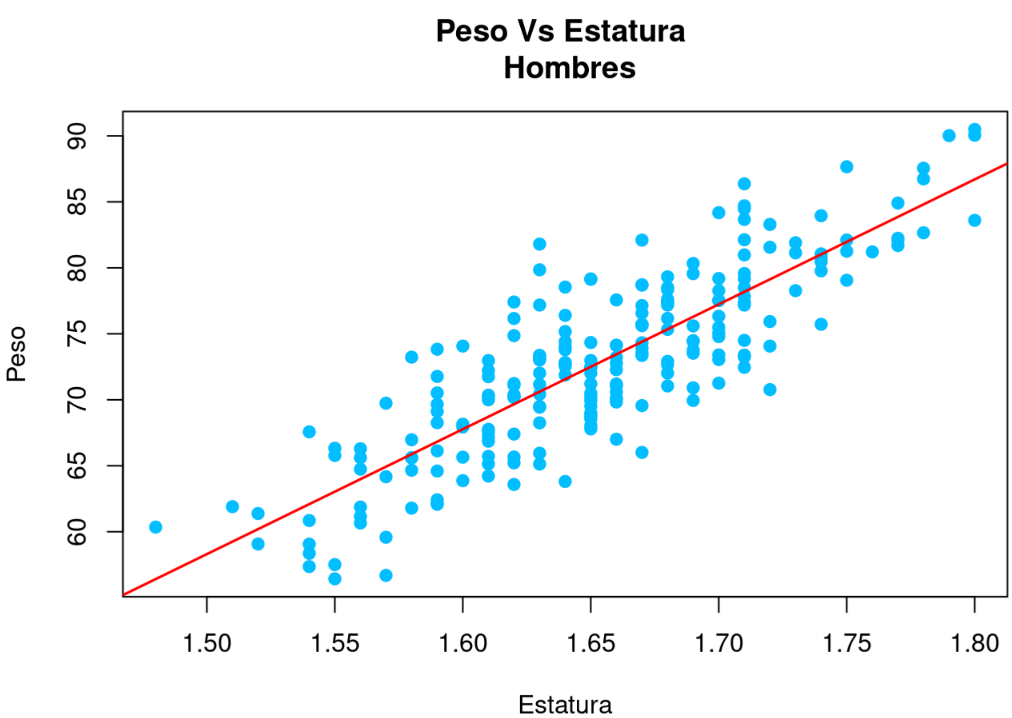
- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Este modelo con solo hombres, demuestra que la estatura es significante para el cálculo del peso, pues es menor a un nivel de 0.03 de confianza.

Modelo Hombres:

- $E_H = -83.685 + 94.66 Peso$

```
plot(MH$Estatura, MH$Peso, col = "#00BFFF", main = "Peso Vs Estatura \n Hombres", ylab = "Peso", xlab = "Estatura", pch = 1
9)
abline(rl_h, col = "red",lwd=1.6)
```



```r
rl_m = lm(Peso~Estatura, MM) # Regresión lineal de la Estatura con base en el Peso de Hombres
summary(rl_m)
```

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

Hipótesis:

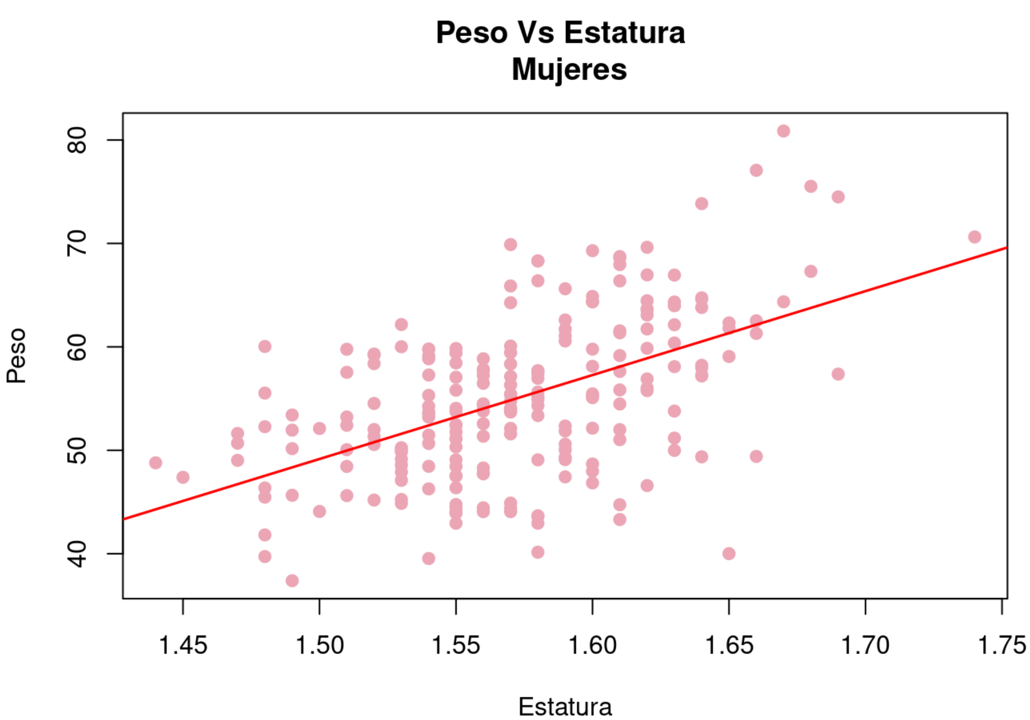
- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Este modelo con solo mujeres, demuestra que la estatura es significante para el cálculo del peso, pues es menor a un nivel de 0.03 de confianza.

Modelo Mujeres:

- $E_M = -72.56 + 81.149 Estatura$

```r
plot(MM$Estatura, MM$Peso, col = "pink2", main = "Peso Vs Estatura \n Mujeres", ylab = "Peso", xlab = "Estatura", pch = 19)
abline(rl_m, col = "red",lwd=1.6)
```



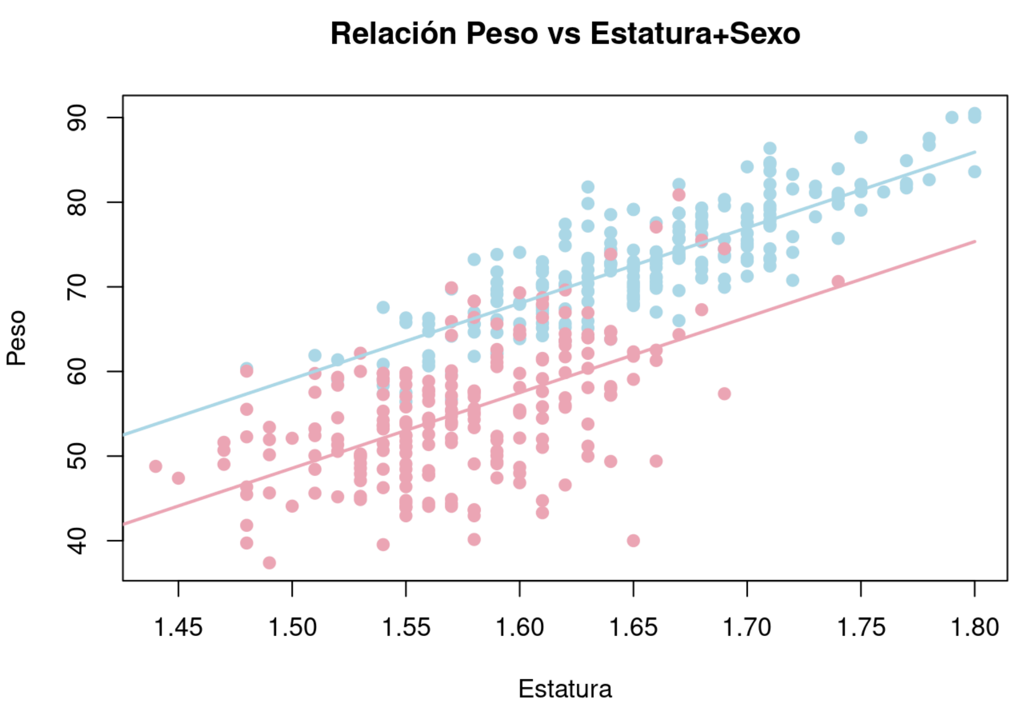

```r
rl = lm(Peso~Estatura+Sexo, M)
summary(rl)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura + Sexo, data = M)
##
## 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 ***
## Estatura     89.2604     4.5635  19.560   <2e-16 ***
## 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
```

El modelo que considera la estatura y sexo para predecir el peso de la persona es significativo, pues su p-value es menor a un 0.03 de significancia.

- E = -74.7546 + 89.2604 Estatura - 10.5645 SexoM

```
b0 = rl$coefficients[1] # Beta 0
b1 = rl$coefficients[2] # Beta 1
b2 = rl$coefficients[3] # Beta 2

Ym = function(x){b0 + b1*x + b2} # Lleva Beta 2, porque las mujeres se representan con un 1
Yh = function(x){b0 + b1*x} # No lleva Beta 2, porque los hombres se representan con un 0

colores = c('lightblue', 'pink2')
plot(M$Estatura, M$Peso, col = colores[factor(M$Sexo)], pch = 19, ylab = "Peso", xlab = "Estatura", main = "Relación Peso vs
Estatura+Sexo")
x = seq(1.40, 1.80, 0.01)
lines(x, Ym(x), col = 'pink2', lwd=2)
lines(x, Yh(x), col = 'lightblue', lwd=2)
```



El modelo con solo hojmbres explica un 71.71% de la variación, el de solo mujeres explica un 27.51% y el modelo que considera ambos sexos explica un 78.37% de la variación.

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

Los análisis anteriores demuestran la importancia de la estatura y sexo de la persona para poder predecir su peso. Además, se hace un análisis por separado de hombres y mujeres para demostrar que existe una diferencia significativa entre el modelo que define la estatura de un hombre y la de una mujer.

6. Interpreta en el contexto del problema:

A. ¿Qué información proporciona β̂0 sobre la relación entre la estatura y el peso de hombres y mujeres?

El coeficiente $\hat{\beta}_0$ es el intercepto del modelo. Representa el valor esperado del peso cuando la estatura y la variable de sexo (SexoM) son iguales a cero.

B. ¿Cómo interpretas β̂1 en la relación entre la estatura y el peso de hombres y mujeres?

Indica que la estatura tiene un efecto positivo significativo en el peso, con un incremento en la estatura resultando en un incremento en el peso, independientemente del sexo.

Otro Modelo (Segunda Entrega)

7. Propón un nuevo modelo. Esta vez toma en cuenta la interacción de la Estatura con el Sexo y realiza los mismos pasos que hiciste con los modelos anteriores:

a. Obtén el modelo e interpreta las variables Dummy

```
rl_es = lm(Peso~Estatura*Sexo, M) # Regresión lineal de la Estatura con base en el Peso de Hombres
summary(rl_es)
```

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

Hipótesis:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$
- E = -83.685 + 94.66 Estatura + 11.124 SexoM - 13.511 Estatura:SexoM

b. Significancia del modelo:

- Valida la significancia del modelo con un alfa de 0.03 (incluye las hipótesis que pruebas)

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

El modelo que considera la estatura, sexo y la interacción entre ambas variables para predecir el peso de la persona es significativo, pues su p-value es menor a un 0.03 de significancia.

- Validala significancia de βi con un alfa de 0.03 (incluye las hipótesis que pruebas)

- $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$
- $H_1 : \exists \beta_i \neq 0$

El p_value de β_0 es menor a 2e-16, así como el p_value de β_1. Los p_value de β_2 y β_3 son mayores a 0.0. De este análisis de los p-values obtenemos que β_0 y β_1 son significativos con un alfa de 0.03 para el modelo, y β_2 y β_3 no son significativos con un alfa de 0.03, ni siquiera con un alfa de 0.01.

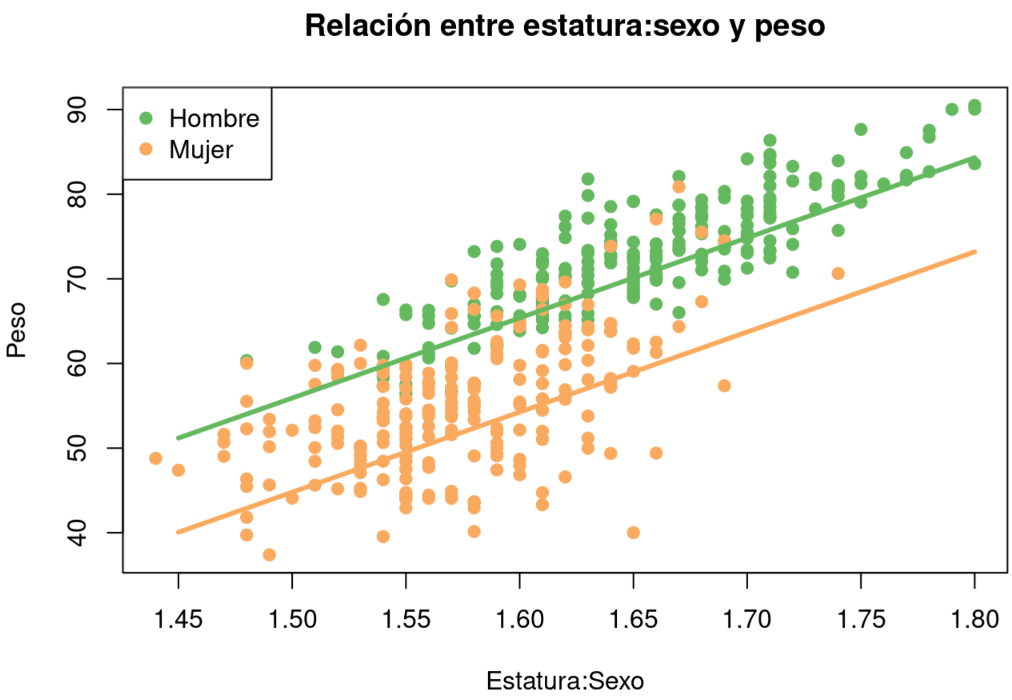
Esto nos dice que el intercepto del modelo y la estatura son importantes para explicar el peso de una persona, pero el sexo de la persona y la interacción entre la estatura y el sexo no son importantes para dicha explicación.

- Indica cuál es el porcentaje de variación explicada por el modelo.

El modelo que considera la estatura, sexo e interacción entre estatura y sexo explica un 78.47% de la variación.

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

```
b0_A=rl_es$coefficients[1]
b1_A=rl_es$coefficients[2]
b2_A=rl_es$coefficients[3]
b3_A = rl_es$coefficients[4]

Ym=function(x){b0_A+b2_A+b1_A*x+b3_A}
Yh=function(x){b0_A+b1_A*x+b3_A}

colores=c("#66BD63","#FDAE61")
plot(M$Estatura,M$Peso,col=colores[factor(M$Sexo)],pch=19,ylab="Peso",xlab="Estatura:Sexo",main="Relación entre estatura:sex
o y peso")
x=seq(1.45,1.80,0.01)
lines(x,Ym(x),col="#66BD63",lwd=3)
lines(x,Yh(x),col="#FDAE61",lwd=3)

legend("topleft", legend=c("Hombre","Mujer"), pch=19, col=c("#66BD63","#FDAE61"))
```



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

Los análisis anteriores demuestran la importancia de la estatura de la persona para poder predecir su peso. Además, se hace un análisis de interacción entre las variables “Estatura” y “Sexo” para demostrar que el sexo de la persona no es significativo para el modelo que define la estatura de una persona, y que no existe interacción entre estas variables Estatura:Sexo.

8. Interpreta en el contexto del problema:

a. ¿Qué información proporciona β_0 sobre la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores.

El coeficiente $\hat{\beta}_0$ es el intercepto del modelo. Representa el valor esperado del peso cuando la estatura, el sexo (SexoM) y la interacción “Estatura:SexoM” son iguales a cero. En el caso del modelo con interacción, $\hat{\beta}_0$ es menor que en los otros 3 modelos sin interacción, esto es para balancear la nueva variable “Estatura:SexoM”.

b. ¿Cómo interpretas β_i en la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores.

$\hat{\beta}_1$ Indica que la estatura tiene un efecto positivo significativo en el peso, con un incremento en la estatura resultando en un incremento en el peso, independientemente del sexo. Es mayor que en los 3 modelos anteriores, debido a las nuevas consideraciones de “Interacción” del modelo.

$\hat{\beta}_2$ Indica que el sexo tiene un efecto positivo en el peso, donde dependiendo del sexo de la persona resultará en un incremento en el peso. Es mayor que en los 3 modelos anteriores, debido a las nuevas consideraciones de “Interacción” del modelo, así como el incremento del intercepto y la estatura.

$\hat{\beta}_3$ Indica que la interacción “Estatura:SexoM” tiene un efecto negativo en el peso, donde dependiendo de dicha interacción resultará en un decremento en el peso. Esta nueva variable es la causa de los incrementos y decrementos en los demás $\hat{\beta}_i$, pues aporta más valor, pero no es significante.

c. Indica cuál(es) de los modelos probados para la relación entre peso y estatura entre hombres y mujeres consideras que es más apropiado y explica por qué.

El modelo sin interacción resulta mejor, pues todas las betas son significativas a un alfa de 0.03. Además de que contamos con un efecto menos (la interacción) y obtenemos un modelo que explica tan bien la varianza como el modelo con interacción, pues existe solo una diferencia de 0.1, explicando el modelo sin interacción un 78.37% de los datos y el modelo con interacción un 78.47%.