act10_A01742161

Rogelio Lizárraga

2024-08-30

La recta de mejor ajuste (Primera entrega)

```
install.packages("ggplot2")
```

```
## Installing package into '/home/gojo/R/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
```

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

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

```
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
```

Observamos que la estatura y el peso están fuertemente correlacionados para los hombres, pero no tanto para las mujeres. Además, vemos que la media y la mediana están muy cercanas en la estatura de los hombres y mujeres, y en el peso de los hombres, pero estas se encuentran más alejadas entre sí para las mujeres. También, observamos que la desviación estándar es mayor para el peso de las mujeres que para el peso de los hombres.

3. Encuentra la ecuación de regresión de mejor ajuste:

Probaremos primero el modelo de regresión lineal para la base de datos de hombres: ## a. Realiza la regresión entre las variables involucradas

```
Modelo1H  = lm(Peso~Estatura, MH)
Modelo1H
```

```
##
## Call:
## lm(formula = Peso ~ Estatura, data = MH)
##
## Coefficients:
## (Intercept)     Estatura
##      -83.68        94.66
```

Para esta tenemos la ecuación

$$Peso = -83.68 + 94.66 Estatura$$

b. Verifica el modelo:

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

```
summary(Modelo1H)
```

```
##
## 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
```

Verifica la significancia del modelo con un alfa de 0.03. Verifica la significancia de $\hat{\beta}_i$ con un alfa de 0.03. Verifica el porcentaje de variación explicada por el modelo

Observamos que el valor p es < 0.03 para β_0 y β_1, por lo que se rechaza H_0 y llegamos a la conclusión que los coeficientes del intercepto y la estatura sí son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Finalmente, tenemos un coeficiente de determinación del 0.7171, por lo que el modelo (la estatura de los hombres) explica el 71.71% de la varianza en el peso de los hombres.

Ahora, probaremos el modelo de regresión lineal para la base de datos de mujeres: ## a. Realiza la regresión entre las variables involucradas

```
Modelo1M  = lm(Peso~Estatura, MM)
Modelo1M
```

```
##
## Call:
## lm(formula = Peso ~ Estatura, data = MM)
##
## Coefficients:
## (Intercept)     Estatura
##      -72.56        81.15
```

Para esta tenemos la ecuación

$$Peso = -72.56 + 81.15 Estatura$$

b. Verifica el modelo:

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

```
summary(Modelo1M)
```

```
##
## 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 ***
```

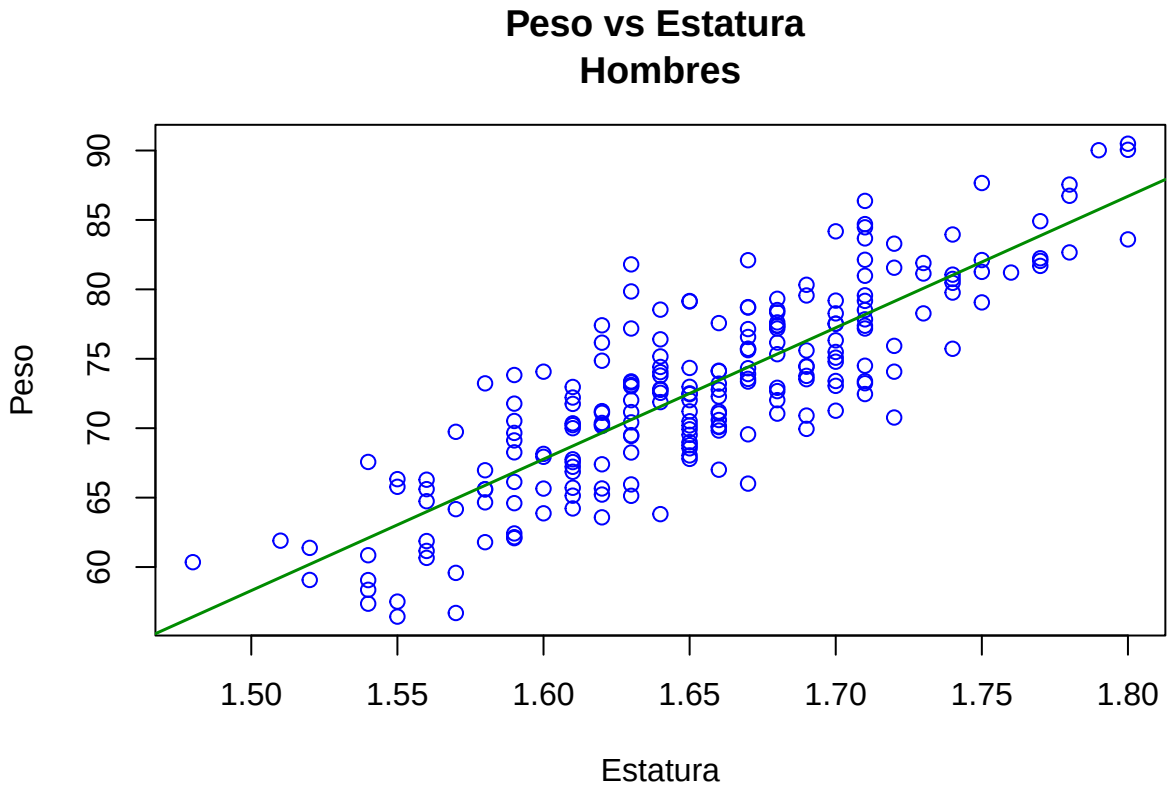
```
## Estatura      81.149      8.922   9.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.65 on 218 degrees of freedom
## Multiple R-squared:  0.2751, Adjusted R-squared:  0.2718
## F-statistic: 82.73 on 1 and 218 DF,  p-value: < 2.2e-16
```

Verifica la significancia del modelo con un alfa de 0.03. Verifica la significancia de $\hat{\beta}_i$ con un alfa de 0.03. Verifica el porcentaje de variación explicada por el modelo

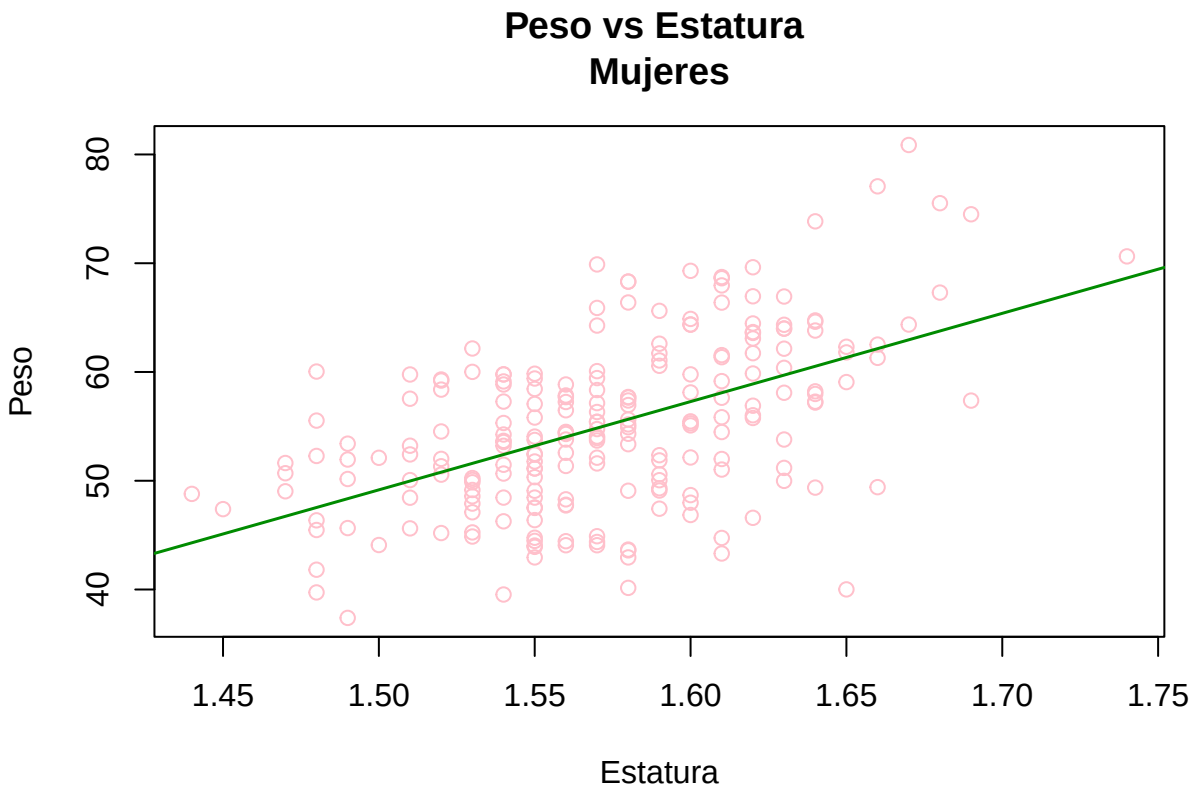
Observamos que el valor p es < 0.03 para β_0 y β_1, por lo que se rechaza H_0 y llegamos a la conclusión que los coeficientes del intercepto y la estatura sí son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Es decir, sí hay una asociación entre las variable dependiente e independiente (peso y estatura). Finalmente, tenemos un coeficiente de determinación del 0.2718, por lo que el modelo (la estatura de las mujeres) explica el 27.18% de la varianza en el peso de las mujeres, lo cual es una variación explicada muy baja.

4. Dibuja el diagrama de dispersión de los datos y las rectas.

```
plot(MH$Estatura, MH$Peso, col ='blue', main = 'Peso vs Estatura \n Hombres ', ylab = 'Peso', xlab = 'Es
abline(Modelo1H, col = 'green4', lwd = 1.5, pch = 19)
```



```r
plot(MM$Estatura, MM$Peso, col ='pink', main = 'Peso vs Estatura \n Mujeres ', ylab = 'Peso', xlab = 'Es
abline(Modelo1M, col = 'green4', lwd = 1.5, pch = 19)
```



Finalmente, haremos el modelo del peso y estatura de hombres y mujeres en conjunto.

```r
Modelo2 = lm(Peso~ Estatura + Sexo, M)
Modelo2
```

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

En este modelo tenemos la ecuación

$$Peso = -74.75 + 89.26 Estatura - 10.56 SexoM$$

, donde $SexoM$ es una variable dummy que indica que: si es mujer, $SexoM = 1$ y $SexoM = 0$ si es otro caso.

```r
summary(Modelo2)
```

```
##
## 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
```

Observamos que el valor p es < 0.03 para β_0, β_1 y $beta_2$, por lo que se rechaza H_0 y llegamos a la conclusión que los coeficientes del intercepto, la estatura y el sexo sí son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Es decir, sí hay una asociación entre las variable dependientes e independientes (peso, contra estatura y sexo). Finalmente, tenemos un coeficiente de determinación del 0.7837, por lo que el modelo (estatura + sexo) explica el 78.37% de la varianza en el peso de las personas, lo cual es una explicación buena.

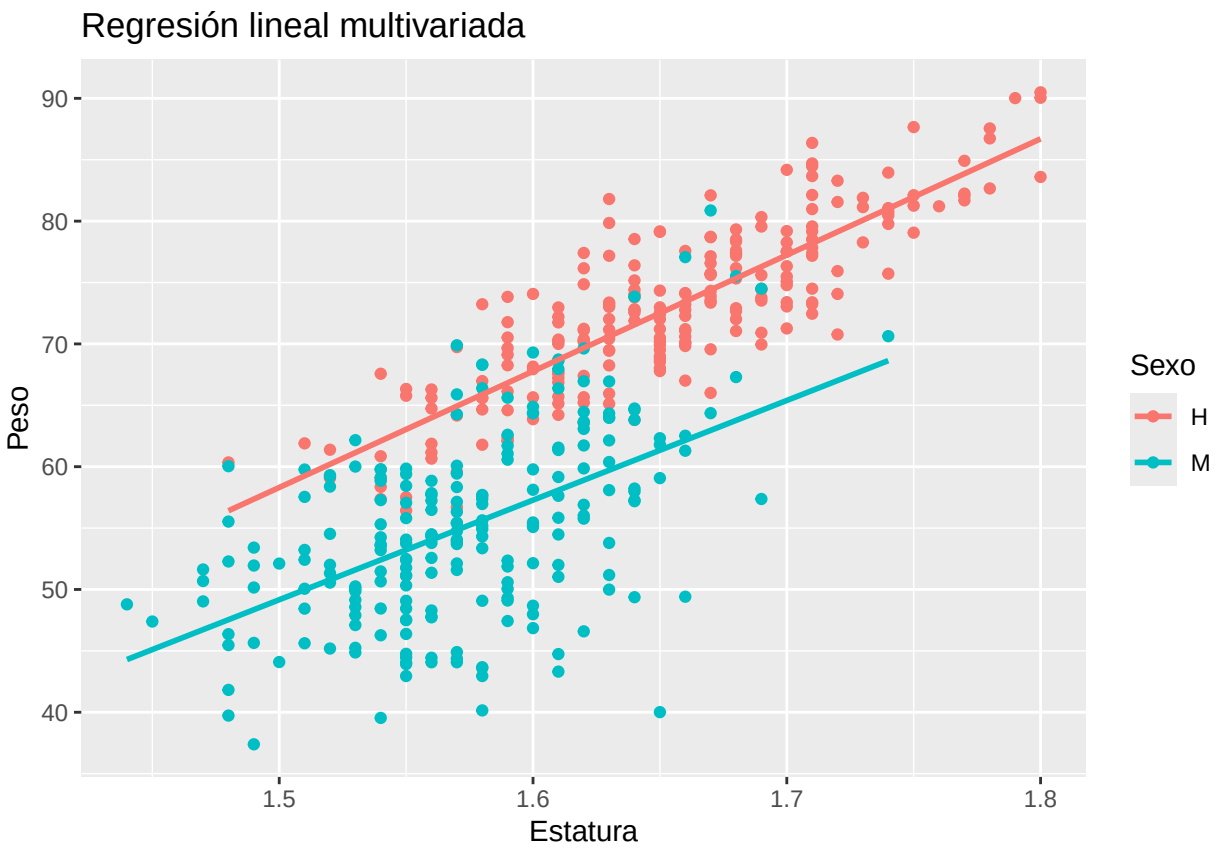
Conclusión mejor modelo

De esta manera, observamos que el mejor modelo es el de estatura + sexo de todo el conjunto de datos, pues es significativo y tenemos el mayor coeficiente de determinación con un valor del 78.37%, por lo que este explica el 78.37% de la variación del peso de las personas.

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

```
ggplot(M, aes(x = Estatura, y = Peso, colour = Sexo)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Regresión lineal multivariada",
       x = "Estatura",
       y = "Peso",
       color = "Sexo")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Este modelo de regresión lineal multivariada logra el mejor ajuste del modelo, separando por sexo y logrando un 78.37% de explicación de la variación del peso de las personas.

5 y 6. Interpreta en el contexto del problema cada uno de los análisis que hiciste.

El valor p representa en cada modelo si es significativo o no este. Es decir, que si hay una asociación real entre las variables y no debida al azar o por casualidad. Por otro lado, $SexoM$ representa si el sexo de la persona es mujer o no. El intercepto de esta variable representa el valor inicial que tomaría el modelo si es mujer (-10.56) y si no lo es (0). Indica que existe una relación negativa entre el sexo y el peso: Si es mujer, menor peso esperado.

a. ¿Qué información proporciona $\hat{\beta}_0$ sobre la relación entre la estatura y el peso de hombres y mujeres?

β_0 es el intercepto y representa el peso cuando la estatura es cero. Es decir, el valor incicial que tiene el modelo y forma parte de la ecuación de regresión.

b. ¿Cómo interpretas $\hat{\beta}_1$ en la relación entre la estatura y el peso de hombres y mujeres?

β_1 representa el cambio en el peso por cada unidad de cambio en la estatura. Es decir, +1 m de estatura = + 35 kg en el peso. ∴ existe una relación positiva y directa entre la estatura y el peso: mayor estatura,

mayor peso.