Regresión Lineal

Jacobo Hirsch Rodriguez

2024-08-31

```
datos=read.csv("./estatura_peso.csv") #leer el dataset
```

```
str(datos)
```

```
## 'data.frame':    440 obs. of  3 variables:
##  $ Estatura: num  1.61 1.61 1.7 1.65 1.72 1.63 1.76 1.67 1.67 1.65 ...
##  $ Peso    : num  72.2 65.7 75.1 68.5 70.8 ...
##  $ Sexo    : chr  "H" "H" "H" "H" ...
```

como la variable sexo es un valor tipo char, es necesario hacer una variable dummy que represente dicha columna

```
# Convertir la variable 'sexo' a binaria: 1 para "H" (Hombre), 0 para "M" (Mujer)
datos$sexo_binario <- ifelse(datos$Sexo == "H", 1, 0)
datos$sexo_binario <- as.factor(datos$sexo_binario)
```

```
str(datos)
```

```
## 'data.frame':    440 obs. of  4 variables:
##  $ Estatura    : num  1.61 1.61 1.7 1.65 1.72 1.63 1.76 1.67 1.67 1.65 ...
##  $ Peso        : num  72.2 65.7 75.1 68.5 70.8 ...
##  $ Sexo        : chr  "H" "H" "H" "H" ...
##  $ sexo_binario: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

```
str(datos$sexo_binario)
```

```
##  Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

y vamos a hacer un nuevo dataset que tenga solo las variables numericas

```
# Seleccionar solo las columnas numéricas (incluyendo la nueva variable binaria)
datos_numericos <- datos[sapply(datos, is.numeric)]
```

ahora vamos a obtener la matriz de correlación

```
# Calcular la matriz de correlación
correlacion <- cor(datos_numericos)
print(correlacion)
```

```
##           Estatura      Peso
## Estatura 1.0000000 0.8032449
## Peso     0.8032449 1.0000000
```

podemos considerar que en general los hombres son mas altos que las mujeres, asi como parece ser que mientras más aumenta la estatura también aumenta el peso. ambos casos son resultados que se pueden esperar

Calcular medidas descriptivas

```
medias <- colMeans(datos_numericos)
desviaciones <- apply(datos_numericos, 2, sd)
minimos <- apply(datos_numericos, 2, min)
maximos <- apply(datos_numericos, 2, max)

medidas <- data.frame(Medias = medias, Desviaciones = desviaciones, Minimos = minimos, Maximos = maximos
print(medidas)
```

```
##             Medias Desviaciones Minimos Maximos
## Estatura  1.613341   0.06929171    1.44    1.80
## Peso     63.970545  11.54161456   37.39   90.49
```

Encuentra la ecuación de regresión de mejor ajuste:

```
# Realizar la regresión lineal considerando 'estatura' y 'sexo_binario'
modelo <- lm(Peso ~ Estatura + sexo_binario, data = datos)

# Resumen del modelo para ver la ecuación y los coeficientes
summary(modelo)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura + sexo_binario, data = datos)
##
## 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)   -85.3191     7.1874  -11.87   <2e-16 ***
## Estatura       89.2604     4.5635   19.56   <2e-16 ***
## sexo_binario1  10.5645     0.6317   16.72   <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
```

#Verifica el modelo:

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

```
# Verificar la significancia del modelo completo
modelo_significativo <- summary(modelo)$coefficients[2,4] < 0.03
if (modelo_significativo) {
  print("El modelo es significativo con un nivel de alfa = 0.03")
} else {
  print("El modelo no es significativo con un nivel de alfa = 0.03")
}
```

```
## [1] "El modelo es significativo con un nivel de alfa = 0.03"
```

##Verifica la significancia de bi con un alfa de 0.03.

```
# Verificar la significancia de los coeficientes
coef_significativos <- summary(modelo)$coefficients[,4] < 0.03
print(coef_significativos)
```

```
##   (Intercept)      Estatura sexo_binario1
##          TRUE          TRUE          TRUE
```

Verifica el porcentaje de variación explicada por el modelo

```
# R-cuadrado del modelo
r_cuadrado <- summary(modelo)$r.squared
print(paste("El porcentaje de variación explicada por el modelo es:", r_cuadrado * 100, "%"))
```

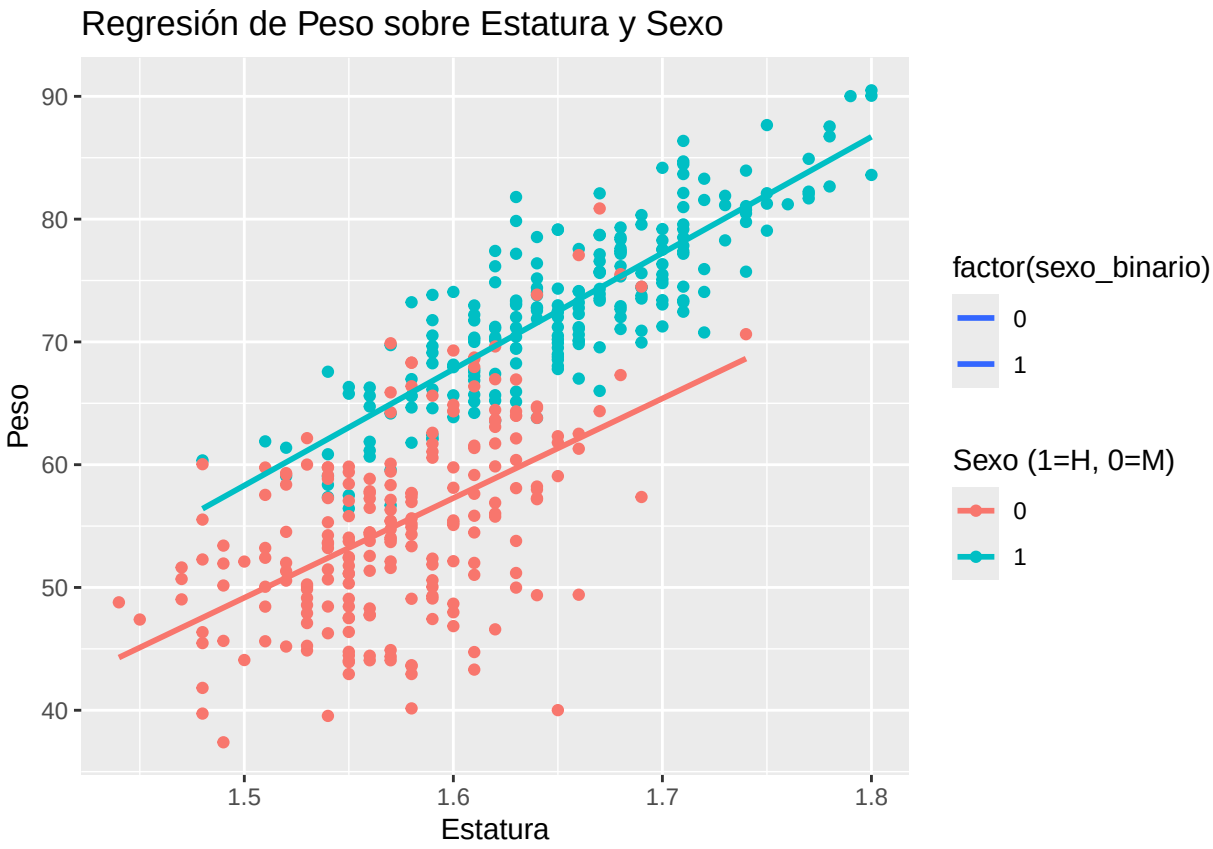
```
## [1] "El porcentaje de variación explicada por el modelo es: 78.3659909400558 %"
```

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

```
# Crear un gráfico con la recta de ajuste para cada grupo
library(ggplot2)

ggplot(datos, aes(x = Estatura, y = Peso, color = factor(sexo_binario))) +
  geom_point() +
  geom_smooth(method = "lm", aes(fill = factor(sexo_binario)), se = FALSE) +
  labs(title = "Regresión de Peso sobre Estatura y Sexo",
       x = "Estatura",
       y = "Peso",
       color = "Sexo (1=H, 0=M)")
```

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



#Conclusiones

el porcentaje de variabilidad es considerado alto al 78%, la ecuación que arroja el modelo se puede escribir como

Peso = -85.3191 + 89.26*Estatura* + *10.56*sexo

donde b0 es el intercepto para las mujeres que es -85.3191, mientras que para los hombres el intercepto es lo mismo mas b2 (el sexo), la ecuación nos dice que por cada unidad (que asumo se encuentra en metros) de estatura, el peso incrementa en 89.26 (asumo que son kilogramos), de la misma forma se puede interpretar que como el sexo es una variable binaria y se sabe que el valor para hombres es 1, esa parte de la ecuacion solo se considera cuando se trata de un hombre por lo que en promedio los hombres pesan 10.56 kilogramos más.

#Regresión lineal con interacción

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:

```
# Realizar la regresión lineal considerando 'estatura' y 'sexo_binario'
new_modelo <- lm(Peso ~ Estatura * sexo_binario, data = datos)

# Resumen del modelo para ver la ecuación y los coeficientes
summary(new_modelo)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura * sexo_binario, data = datos)
```

```
##
## 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)              -72.560     11.346  -6.395 4.13e-10 ***
## Estatura                  81.149      7.209  11.256  < 2e-16 ***
## sexo_binario1            -11.124     14.950  -0.744    0.457
## Estatura:sexo_binario1    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
```

#Obtén el modelo e interpreta las variables Dummy

el p value de el intercepto y de estatura son muy bajos (muy significativos), sumados a que el valor de f es los suficientemente alto para determinar que son explicativos del modelo sin embargo la interacción con el sexo parecen no ser significativos.

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

```r
# Verificar la significancia del modelo completo
modelo_significativo <- summary(new_modelo)$coefficients[2,4] < 0.03
if (modelo_significativo) {
  print("El modelo es significativo con un nivel de alfa = 0.03")
} else {
  print("El modelo no es significativo con un nivel de alfa = 0.03")
}
```

```
## [1] "El modelo es significativo con un nivel de alfa = 0.03"
```

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

```r
# Verificar la significancia de los coeficientes
coef_significativos <- summary(new_modelo)$coefficients[,4] < 0.03
print(coef_significativos)
```

```
##           (Intercept)              Estatura          sexo_binario1
##                  TRUE                  TRUE                  FALSE
## Estatura:sexo_binario1
##                 FALSE
```

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

```r
# R-cuadrado del modelo
r_cuadrado <- summary(new_modelo)$r.squared
print(paste("El porcentaje de variación explicada por el modelo es:", r_cuadrado * 100, "%"))
```

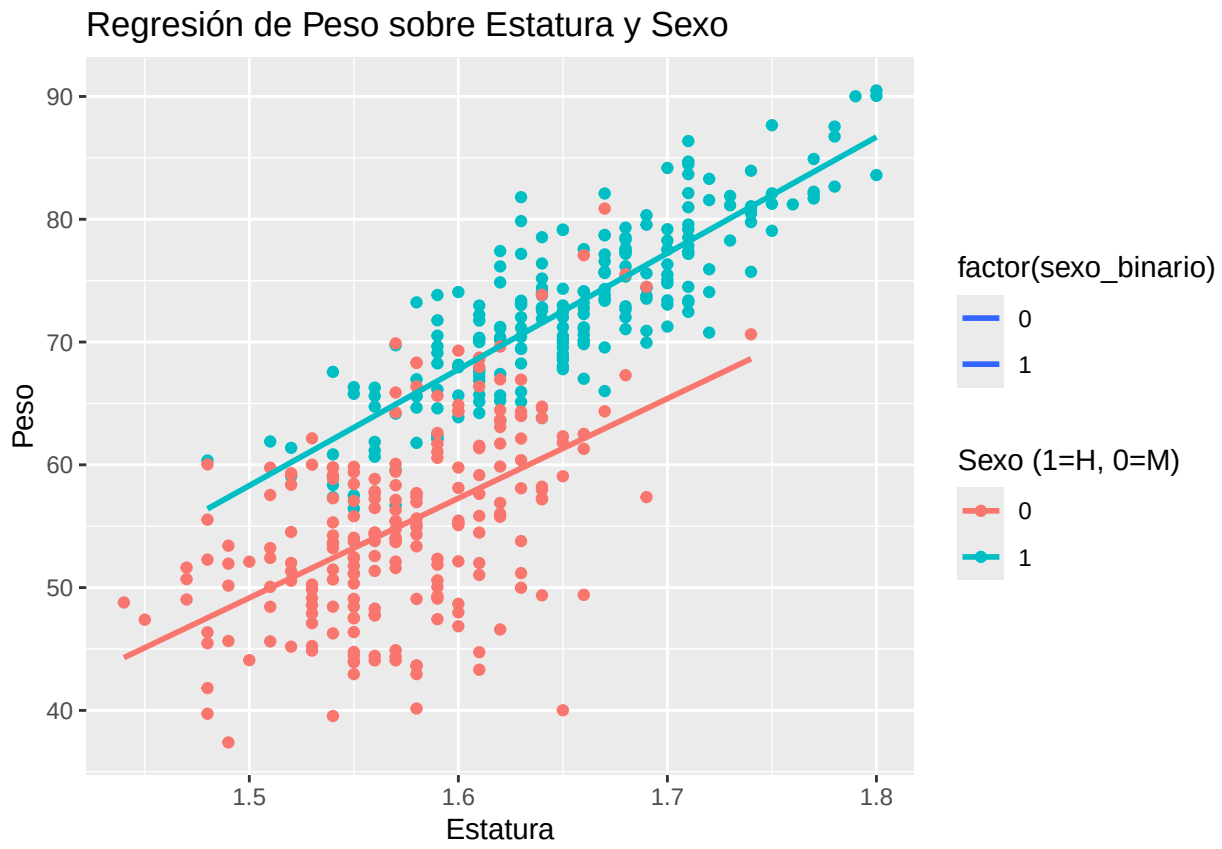
```
## [1] "El porcentaje de variación explicada por el modelo es: 78.4701094355899 %"
```

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

```
# Crear un gráfico con la recta de ajuste para cada grupo
library(ggplot2)

ggplot(datos, aes(x = Estatura, y = Peso, color = factor(sexo_binario))) +
  geom_point() +
  geom_smooth(method = "lm", aes(fill = factor(sexo_binario)), se = FALSE) +
  labs(title = "Regresión de Peso sobre Estatura y Sexo",
       x = "Estatura",
       y = "Peso",
       color = "Sexo (1=H, 0=M)")
```

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



Interpreta en el contexto del problema: ¿Qué información proporciona B0 sobre la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores. ¿Cómo interpretas Bi en la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores. 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.

la ecuacion con interaccion es: Peso = -72.56 + 81.14 * estatura -11.12*sexo* + *13.51*estatura*sexo

se explica la intercepción como el minimo de peso que se tiene en correlacion con el minimo de estatura (ya que un peso ni una estatura negativa tienen sentido)

Aunque el modelo original que no considera la interacción se comporta ligeramente mejor (especificamente 78.47- 78.36 = 0.11 mejor) se descarta el modelo que posea variables de más que se consideran no significativas (como en el modelo de interacción) por que el porcentaje de variacion no considera la aleatoridad de la variable (en cambio la significancia si la considera) por ende se descarta el el modelo con la interacción.