Estadística II - Taller 01 **Semestre: 2023-01**

Profesores: Francisco Javier Rodríguez Cortés, Julieth Verónica Guarín Escudero

Monitor: Matheo Muñoz Betancur

Se presenta una base de datos que recopila información de diferentes aplicaciones de la Play Store de Google. Dicha información presentada en variables que muestran atributos de las aplicaciones como el número de veces que esta ha sido instalada, el total de calificaciones que ha recibido y su distribución en las categorías que van desde una hasta cinco estrellas, etcétera.

Tabla 1: Vista previa de algunas variables

Titulo	Total de calificaciones	Total de calificaciones cinco estrellas
Call of Duty®: Mobile - Season 4: Spurned & Burned	13572148	10501443
Mystic Messenger	419193	356875
Blockudoku® - Block Puzzle Game	414430	272911
City Driving 3D	272721	138933
Slugterra: Slug it Out 2	481615	412995

Considere la cantidad total de calificaciones como la covariable (X) y a la cantidad de calificaciones en la categoría de cinco estrellas como la variable respuesta (Y).

Su tarea como analista es realizar las siguientes tareas usando el software estadístico R.

1. Realice la lectura de la base de datos, posteriormente filtre para solo quedarse con aquellas observaciones que tengan menos de 4121627 calificaciones totales y pertenezcan a la categoría de juegos de acción, seleccione solo la covariable y la variable respuesta. Finalmente guarde dichas observaciones en una nueva base de datos.

2. Elabore un gráfico de dispersión de los datos, luego de esto analícelo.

3. Escriba la ecuación del modelo de regresión, junto con sus supuestos. Ajuste un modelo de regresión lineal simple y añada la recta de regresión a la gráfica generada anteriormente.

4. Realice una interpretación del parámetro β_0, ¿qué unidades tiene? Determine si este es significativo usando $\alpha = 0.05$ y si tiene sentido en el contexto de los datos.

5. Repita el proceso anteriormente enunciado con el parámetro β_1.

6. Calcule un intervalo de confianza del 95 % para ambos parámetros del modelo. Antes de calcularlo responda, ¿dichos intervalos deberían contener al cero?

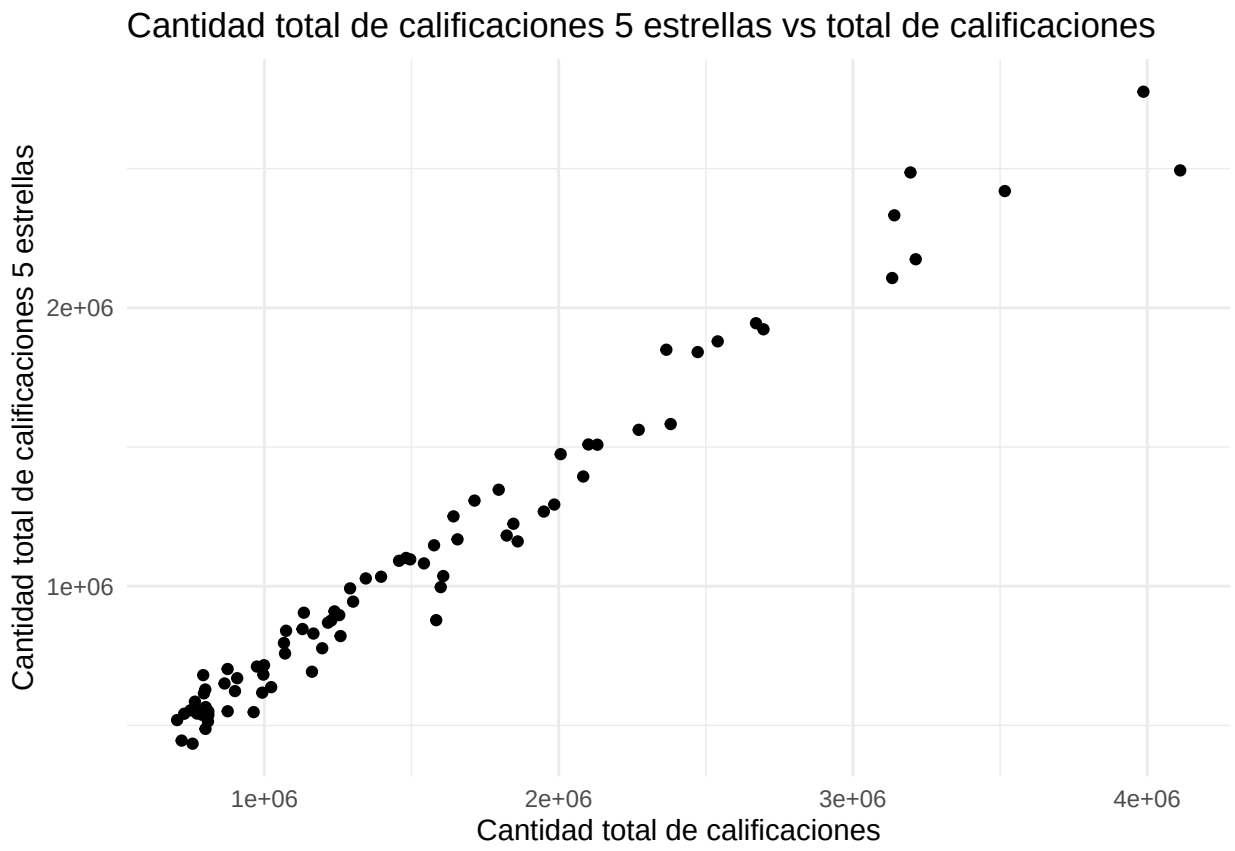
Solución

Ejercicio 1

```
datos.modelo <- read.csv("android-games.csv") %>%
  filter(total_ratings <  4121627, category == "GAME ACTION") %>% #filtra
  select(total_ratings, five_star_rating) #selecciona las columnas que necesito
```

Ejercicio 2

```
p <- ggplot(data = datos.modelo, aes(x=total_ratings, y=five_star_rating)) + geom_point(
p
```



Como se puede observar, existe una correlación lineal entre ambas variables, por lo que es posible plantear un modelo de regresión lineal simple. También, se espera que β_1 sea positivo pues existe una relación positiva entre ambas variables, a mayor número de la cantidad total de calificaciones, mayor es la cantidad total de calificaciones 5 estrellas.

Ejercicio 3

El modelo de regresión lineal es pues

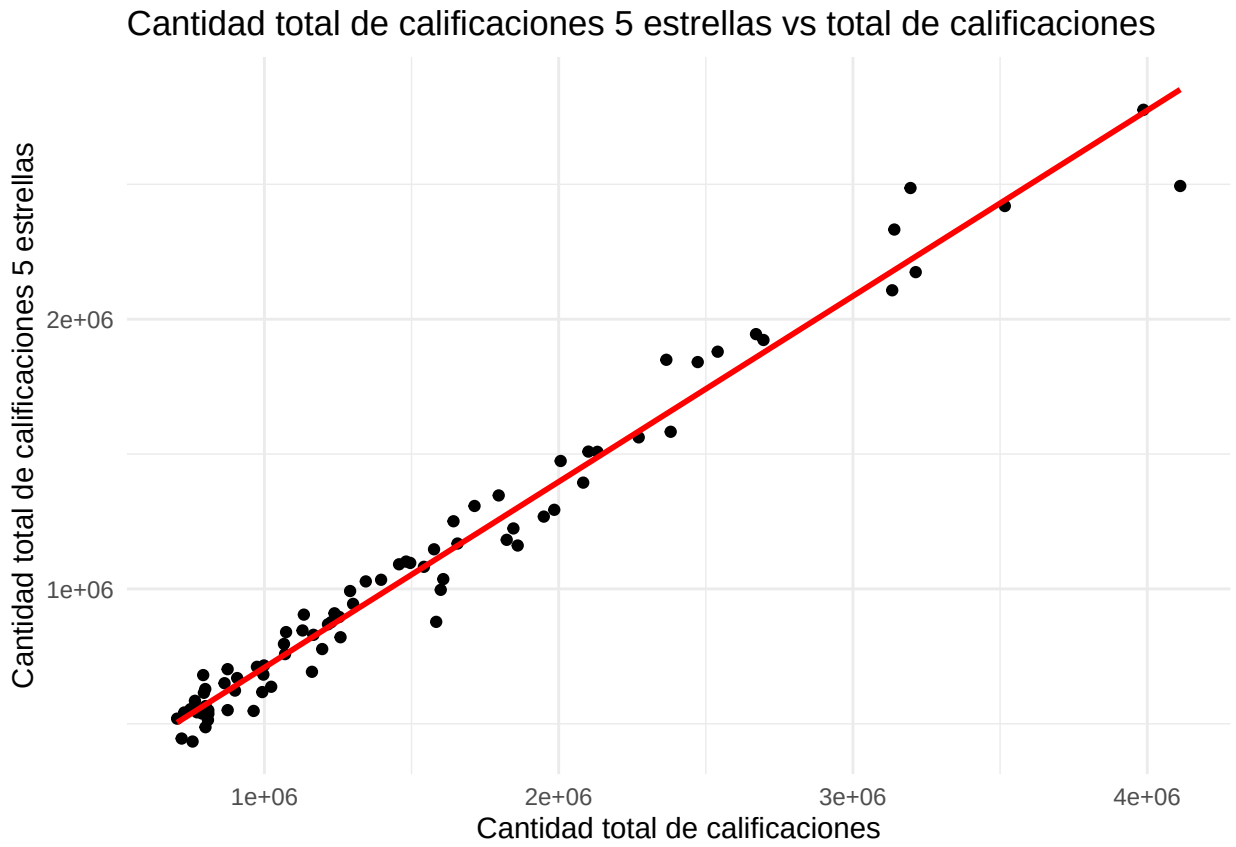
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Cuyo supuesto es

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Y donde $1 \leq i \leq 80$. Ahora, ajustando el modelo y agregando a la gráfica de dispersión el ajuste:

```
mod <- lm(five_star_rating ~ total_ratings, data=datos.modelo)
resumen <- summary(mod) #resumen del modelo
beta0 <- coef(mod)[1] #intercepto
beta1 <- coef(mod)[2] #pendiente
p + geom_smooth(color="red", method= "lm", formula = "y~x", se=F) #grafica con ajuste
```



Ejercicios 4 y 5

Tabla 2: Resúmen de coeficientes

	Estimación	Error estándar	t_0	Valor-P
β_0	20303.7800	22173.2195	0.9157	0.3627
β_1	0.6885	0.0128	53.7437	0.0000

- $\hat{\beta}_0$ no tiene interpretación y dado que su valor-P es mayor a un nivel de significancia $\alpha = 0.05$, no es significativo, además tiene unidades de calificaciones de 5 estrellas
- $\hat{\beta}_1$ es la tasa de cambio de la cantidad de calificaciones de 5 estrellas por la cantidad de calificaciones totales y el parámetro tiene un valor-P mucho menor al nivel de significancia, por lo que es significativamente distinto de 0 e interpretable.

Ejercicio 6

Se espera que el 0 esté contenido en el intervalo de confianza del intercepto y que no lo esté en la pendiente, pues si lo contiene el parámetro no es significativo. Así pues:

```
#función confint permite hallar los intervalos de confianza
inter.beta0 <- confint(mod, "(Intercept)", level = 0.95)
inter.beta1 <- confint(mod, "total_ratings", level = 0.95)

#para crear la tabla:
inter.conf <- data.frame(c(inter.beta0[1], inter.beta1[1]), c(inter.beta0[2], inter.beta
inter.conf %>%
   kable(col.names = c("Límite inferior", "Límite superior"),
        booktab = T, caption = "Intervalos de confianza",
        row.names = T, escape=  F) %>%
  kable_styling(latex_options = c("hold_position")) %>%
  column_spec(2:3, latex_valign = "m") %>%
  row_spec(0:5, align = "c")
```

Tabla 3: Intervalos de confianza

	Límite inferior	Límite superior
β_0	-2.383971e+04	6.444727e+04
β_1	6.629525e-01	7.139579e-01

```
sum()
```

```
[1] 0
```