Estadística II - Taller 02 **Semestre: 2023-02**

Monitor: Matheo Muñoz Betancur

Se presenta una base de datos que recopila información de diferentes clientes de un Ecommerce, a groso modo su tarea consiste en explorar dicha base de datos para ajustar un modelo de regresión adecuado donde la variable respuesta (Y) es la cantidad anual gastada por cliente.

Tabla 1: Vista previa de la base de datos

Email	Avatar	Cantidad gastada al año por cliente
mstephenson@fernandez.com	Violet	587.9511
hduke@hotmail.com	DarkGreen	392.2049
pallen@yahoo.com	Bisque	487.5475
riverarebecca@gmail.com	SaddleBrown	581.8523
mstephens@davidson-herman.com	MediumAquaMarine	599.4061

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

1. Realice la lectura de la base de datos, seleccione únicamente las variables numéricas.
2. Elabore un gráfico de dispersión de las variables para encontrar aquella que presente una mejor relación lineal con respecto a la variable respuesta.
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. **Nota:** seleccione aleatoriamente el 80 % de los datos para ajustar el modelo.
4. Realice la prueba de significancia para la pendiente, luego realice la prueba de significancia de la regresión usando análisis de varianza. ¿Ambos enfoques permiten llegar a la misma conclusión? ¿Qué relación existe entre una prueba y la otra?
5. Dé una interpretación de los parámetros β_0 y β_1 del modelo, en caso que sea posible hacerlo.
6. Calcule el R^2 usando sumas de cuadrados, y realice una interpretación de este.
7. Use el modelo para predecir la cantidad anual total gastada por cliente en el 20 % de los datos que no usó para ajustar el modelo. Calcule los respectivos intervalos de confianza y de predicción. ¿Cuáles intervalos son más anchos? ¿Por qué cree usted que esto sucede?

Solución

Ejercicio 1

Primero, se realiza la lectura de la base de datos con

```
datos <- read.csv("Ecommerce_Customers.csv")
```

Para seleccionar las variables numéricas, se observa primero la estructura de la variable datos, así:

```
str(datos)
```

```
'data.frame':   500 obs. of  8 variables:
 $ Email               : chr  "mstephenson@fernandez.com" "hduke@hotmail.com" "pallen@ya
 $ Avg..Session.Length : num  34.5 31.9 33 34.3 33.3 ...
 $ Time.on.App         : num  12.7 11.1 11.3 13.7 12.8 ...
 $ Time.on.Website     : num  39.6 37.3 37.1 36.7 37.5 ...
 $ Length.of.Membership: num  4.08 2.66 4.1 3.12 4.45 ...
 $ Address             : chr  "835 Frank Tunnel\nWrightmouth, MI 82180-9605" "4547 Arche
 $ Yearly.Amount.Spent : num  588 392 488 582 599 ...
 $ Avatar              : chr  "Violet" "DarkGreen" "Bisque" "SaddleBrown" ...
```

Se puede observar que hay 5 variables númericas, así, se hace la selección de estas.

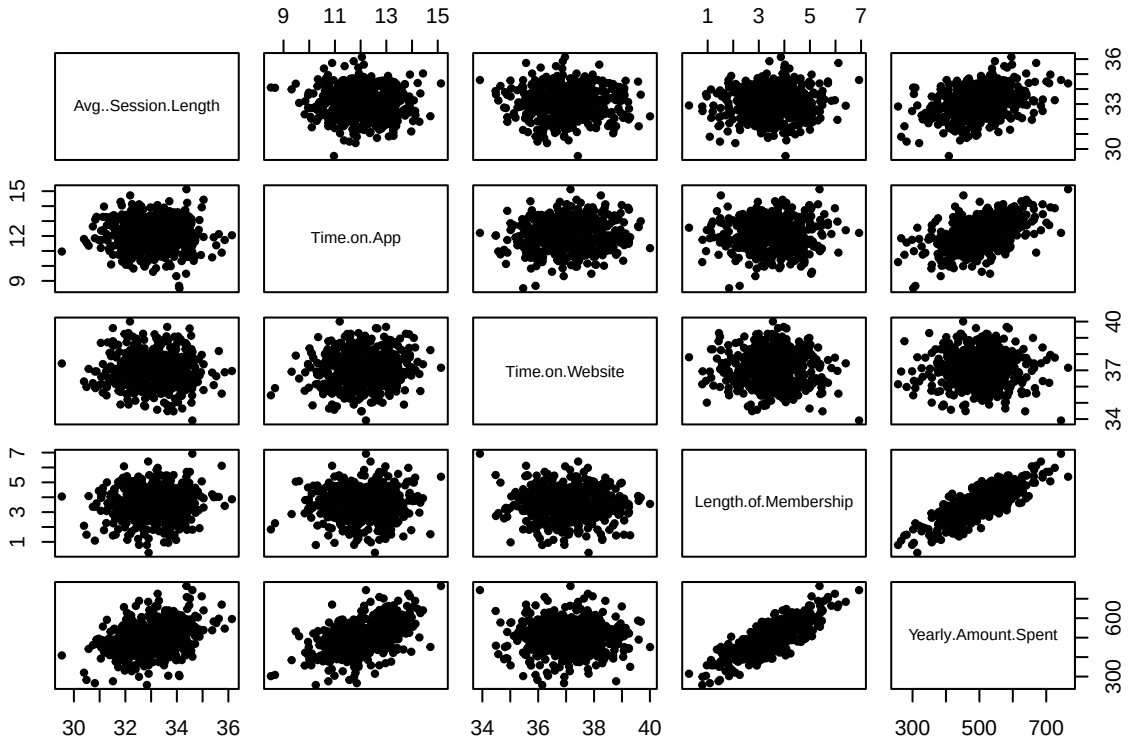
```
datos.filtrados <- datos %>%
  select(Avg..Session.Length:Length.of.Membership, Yearly.Amount.Spent)
```

Note que la variable *Yearly.Amount.Spent* será la variable respuesta Y.

Ejercicio 2

El gráfico de dispersión que relacion todas las variables entre sí para observar entre cuáles puede existir linealidad basta con hacer lo siguiente a estos datos:

```
plot(datos.filtrados, pch=20)
```



Así, es claro que la covariable X que mejor relación lineal presenta con la variable cantidad anual gastada por cliente, la cual es la variable respuesta Y es la duración de su membresía (*Length.of.Membership*). Por esto, se crea la nueva variable en **R** con los datos a utilizar:

```
datos.modelo <- datos.filtrados %>%
  select(Yearly.Amount.Spent, Length.of.Membership)
```

El número de datos para el ajuste son 80 % del número de datos.

```
n <- 0.8 * nrow(datos.modelo)
n
```

```
[1] 400
```

Ejercicio 3

Considerando la duración de la membresía como la covariable X y la cantidad anual gastada como la respuesta Y, se plantea el siguiente modelo de regresión:

$$Y_i = \beta_0 + \beta_1 X_i, \ \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Donde $1 \leq i \leq n$ con $n = 400$. Seleccionando aleatoriamente 400 datos de la variable datos.modelo y ajustando el modelo de regresión:

```
n <- 500 #Para recuperar el valor total de n
set.seed(9012913) #Para permitir reproducibilidad de la aleatoriedad
filas <- sample(1:n, 0.8*n)
datos.ajuste <- datos.modelo[filas, ]
datos.prediccion <- datos.modelo[-filas,  ]
mod <- lm(Yearly.Amount.Spent ~ Length.of.Membership, data=datos.ajuste)
summary(mod)
```

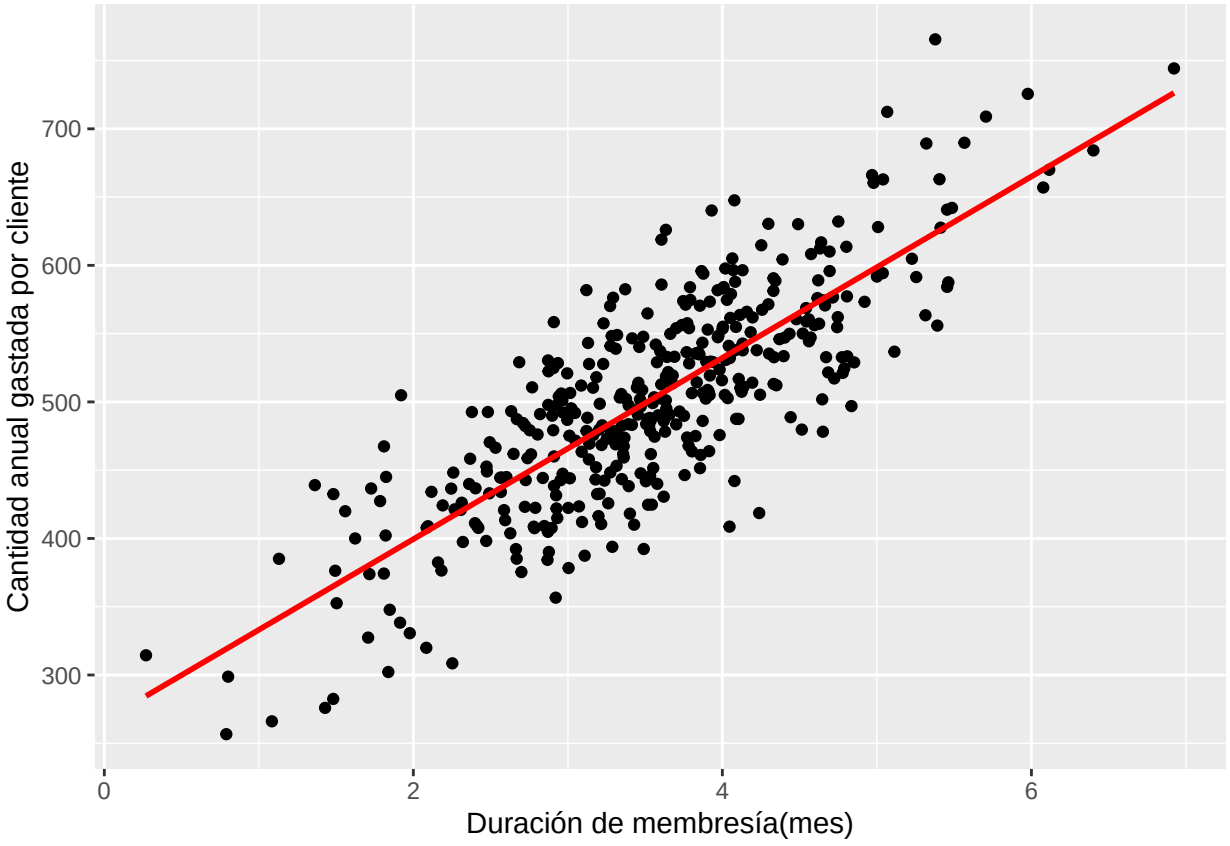
```
Call:
lm(formula = Yearly.Amount.Spent ~ Length.of.Membership, data = datos.ajuste)

Residuals:
     Min       1Q   Median       3Q      Max
-129.475  -29.460   -1.944   32.280  141.800

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          266.684      8.802   30.30   <2e-16 ***
Length.of.Membership  66.393      2.398   27.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.57 on 398 degrees of freedom
Multiple R-squared:  0.6583,    Adjusted R-squared:  0.6574
F-statistic: 766.8 on 1 and 398 DF,  p-value: < 2.2e-16
```

Aquí se puede observar que:

- $\hat{\beta}_0 = 266.684$, cuyo error estándar es 8.802, su estadístico de prueba para significancia individual es $T = 30.30$ y su valor-P es del orden de 0.
- $\hat{\beta}_1 = 66.393$, cuyo error estándar es 2.398, su estadístico de prueba para significancia individual es $T = 27.69$ y su valor-P es del orden de 0.

Ahora, se presenta la gráfica de Y vs X con el ajuste:

```
p <- ggplot(datos.ajuste, aes(x=Length.of.Membership, y= Yearly.Amount.Spent)) + geom_po
p
```



Ejercicio 4

En el ejercicio anterior se presentó el valor-P para la prueba de significancia de la pendiente que plantea

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

El cual arrojó un valor cercano a 0, mucho menor a cualquier valor de significancia, por lo que se rechaza la hipótesis nula y la pendiente es significativa. Para usar análisis de varianza para la prueba de significancia de la regresión, se presenta a continuación la tabla anova:

```
anova(mod)
```

```
Analysis of Variance Table

Response: Yearly.Amount.Spent
                     Df  Sum Sq Mean Sq F value    Pr(>F)
Length.of.Membership   1 1663089 1663089  766.79 < 2.2e-16 ***
Residuals            398  863224    2169
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De aquí es posible afirmar que la regresión es significativa, lo cual no contradice el resultado dado por la hipótesis individual. También, cuando se usa análisis de varianza en un modelo lineal con sólo un parámetro que hace significativa la regresión, el estadístico T y F de las pruebas están relacionados pues $F_0 = T_0^2$, esto se puede ver en R

```
F.subcero <- 766.79
T.subcero <- 27.69
T.subcero^2
```

```
[1] 766.7361
```

Ejercicio 5

- $\hat{\beta}_0$: Como el valor de $X = 0$ no hace parte de los datos, entonces no es interpretable
- $\hat{\beta}_1$: Cuando una persona aumenta su tiempo de suscripción en un mes, se estima que su gasto anual medio incrementa 66.393USD

Ejercicio 6

Tomando en cuenta los valores que pueden ser extraídos de la tabla anova para la suma cuadrática de regresión y la suma cuadrática de error, se tiene:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}} = \frac{1663089}{1663089 + 863224} = 0.6583068$$

El R^2 es la proporción de la variabilidad de la respuesta explicada por la regresión. Un $R^2 = 0.6583068$ significa que la regresión explica aproximadamente 65.83 % de la variabilidad de la respuesta.

Ejercicio 7

El intervalo de confianza para una observación futura es más ancho, lo cual puede ser observado en su fórmula y esto es debido a que su predicción se hace considerando un error de predicción, el cual al hallar su varianza hace que se infle su valor respecto al valor de la respuesta media. Una función en **R** que permite obtener el intervalo de confianza para la respuesta media es la que sigue(**Nota:** para hacer esta inferencia, es necesario hacer uso sólo de los datos que se utilizaron en la regresión):

```r
int.conf.media <- predict(mod, newdata = datos.ajuste, interval = "confidence", level =

int.conf.media[1:5, ] #5 primeras entradas de la tabla
```

```
         fit      lwr      upr
37  389.3367 380.1370 398.5364
26  447.3884 441.4014 453.3754
218 588.8789 581.1825 596.5752
85  558.5910 552.4877 564.6943
153 624.6270 614.7741 634.4798
```

Para los intervalos de predicción similarmente en **R**:

```r
int.conf.futuro <- predict(mod, newdata = datos.prediccion, interval = "prediction", lev

int.conf.futuro[1:5, ] #5 primeras entradas de la tabla
```

```
        fit      lwr      upr
5  561.8879 470.1172 653.6585
6  631.4145 539.2821 723.5469
10 479.3281 387.6430 571.0131
12 513.2152 421.5404 604.8901
19 367.3742 275.2080 459.5405
```

Note acá que los datos para predicción no pueden ser datos usados en la predicción y deben estar entre el mínimo y máximo de los datos de la misma, es decir, deben ser de interpolación.