Estadística II - Práctica 02 **Semestre: 2021-02**

Profesores: Carlos M. Lopera-Gómez y Raúl Alberto Pérez

Monitor: Simon Pedro Galeano

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 es la cantidad anual gastada por cliente.

Cuadro 1: Vista previa de la base de datos

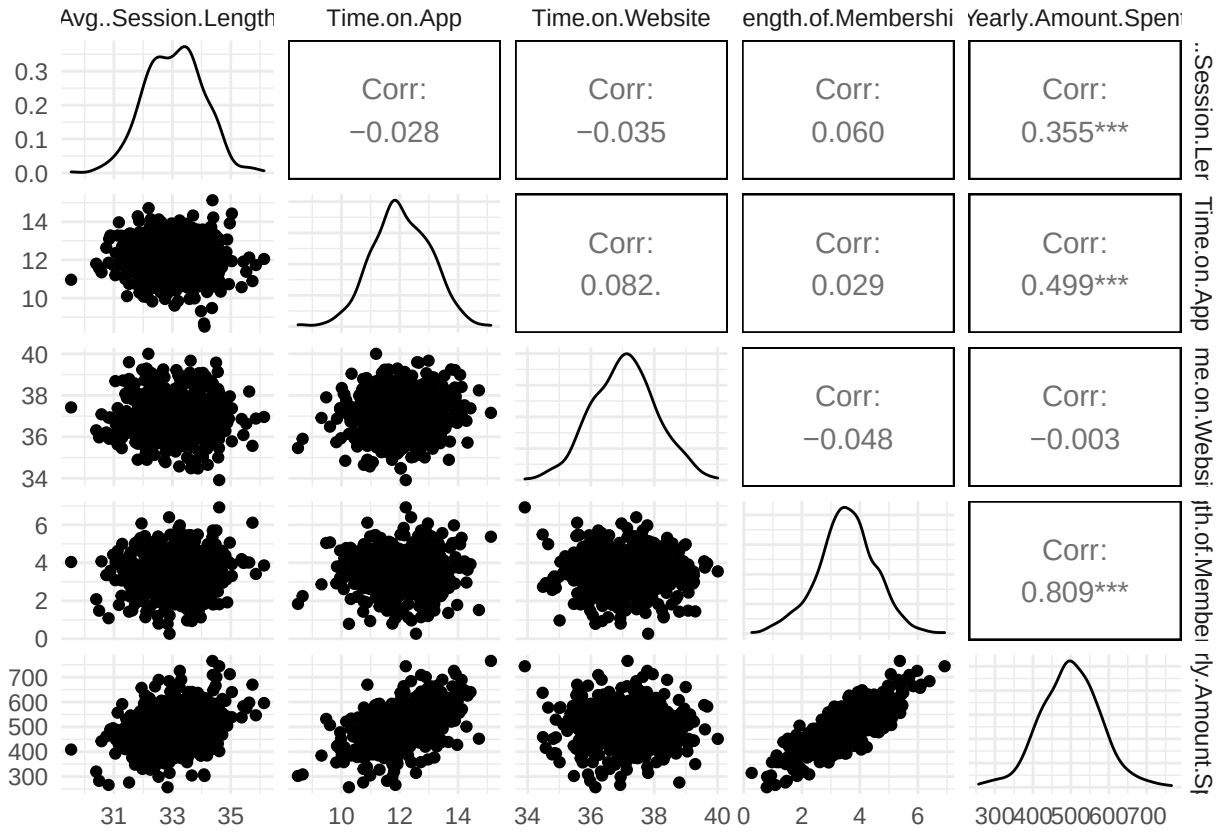
Email	Avatar	Cantidad gastada al año por cliente
mstephenson@fernandez.com	Violet	587.9511
hduke@hotmail.com	DarkGreen	392.2049
pallen@yahoo.com	Bisque	487.5475
riverarebecca@gmail.com	SaddleBrown	581.8523
mstephens@davidson-herman.com	MediumAquaMarine	599.4061

Su tarea 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. De una interpretación de los parámetros β_0 y β_1 del modelo, claro está, si es posible hacerlo.
6. Calcule el R^2 usando el coeficiente de correlación y usando sumas de cuadrados, compare estos entre sí y compárelos con las salidas de R. Realice una interpretación de este.
7. Use el modelo para predecir las 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 2



Del gráfico anterior se puede notar que la variable que presenta una mejor relación líneal (al menos de manera gráfica) con la variable respuesta es la duración de la membresía.

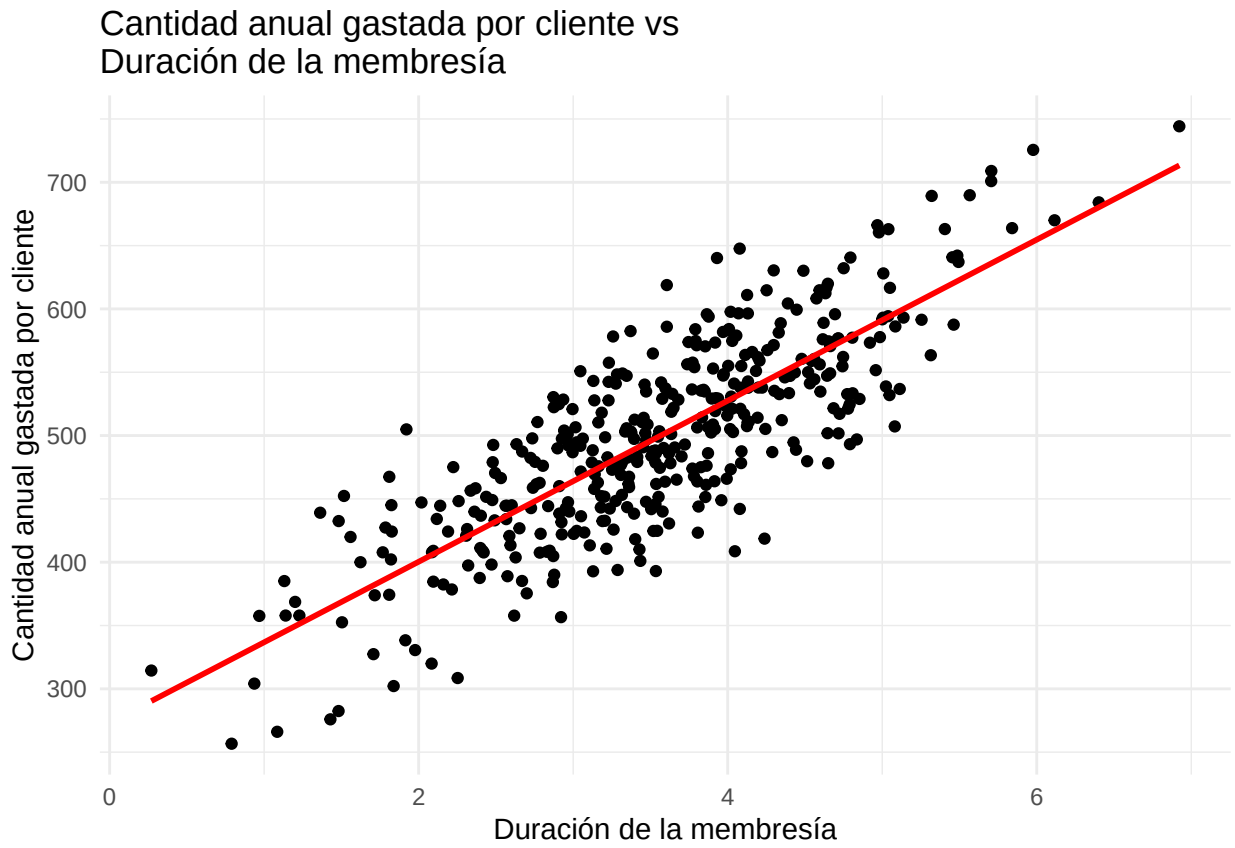
Por tanto se escoge a la duración de la membresía como covariable.

Ejercicio 3

Se realiza la especificación del modelo de regresión, teniendo en cuenta que y y x hacen referencia a la variable respuesta y a la covariable respectivamente.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ 1 \leq i \leq 400$$

Posteriormente se realiza presenta el diagrama de dispersión de los datos con la recta de regresión ajustada.



Ejercicio 4

Cuadro 2: Resumen de las pruebas

	Significancia de la pendiente	Significancia de la regresión
Estadístico de prueba	27.95	781.2
P-valor	<2e-16	<2.2e-16

Ejercicio 5

Para el parámetro β_0 no se posee interpretación útil, mientras que para el parámetro β_1 se **estima** que el gasto anual por cliente aumenta en **promedio** 63.584 dólares por cada año de suscripción adicional.

Ejercicio 6

Cuadro 3: Comparación de R^2

	R^2
Salida de R	0.6625
Correlación	0.66247
Suma de cuadrados	0.66247

Se puede observar que independiente de como se calcule, el coeficiente de determinación es el mismo para los métodos de computo de este considerados anteriormente.

Ejercicio 7

Cuadro 4: Predicciones, intervalos de confianza y de predicción

Estimación	Límite inferior intervalo de confianza	Límite superior intervalo de confianza	Límite inferior intervalo de predicción	Límite superior intervalo de predicción
532.7766	527.6097	537.9435	441.4604	624.0929
442.5773	436.5702	448.5843	351.2096	533.9449
534.1706	528.9568	539.3843	442.8517	625.4895
471.5807	466.6528	476.5087	380.2777	562.8838
481.3253	476.6149	486.0357	390.0338	572.6168

Cabe resaltar que los intervalos de predicción son más anchos que los intervalos de confianza para la respuesta media.