Universidad Simón Bolívar
Departamento de Cómputo Científico y Estadística
CO-3321 Estadística Para Ingenieros
Prof. Pedro Ovalles

Proyecto Final

Autores
José Barrera **15-10123** | Carlos Sivira **15-11377** | Miguel Colorado **14-10237**

Sartenejas, noviembre de 2019

Resumen

En el presente proyecto se trabajan los datos proporcionados por el departamento de ventas de una empresa, referentes al número de productos vendidos, las ganancias en millones generadas en 4 regiones y el presupuesto en millones asignado a distintos medios de publicidad (facebook, periodico, instagram, tv y ebay). Los datos para este estudio se encuentran en un archivo de texto llamado "datosproy.txt". La información se encuentra en una tabla con 200 entradas o filas y 7 columnas representando la información antes descrita.

De forma adicional, se proporciona una tabla de presupuestos para la realización del inciso 6, correspondiente a estimar el valor de las ventas para una serie de presupuestos en publicidad con respecto al mejor modelo encontrado en el inciso 3.

Pregunta	Resultados
1	La mayor inversión en publicidad, se realiza mediante Ebay. Las variables facebook e instagram poseen las mayores correlaciones con la variable ventas.
2	Región 1: ***12.26927 - 15.23277***. Región 2: ***11.73804 - 14.94596***. Región 3: ***13.99225 - 17.02407***. Región 4: ***12.62179 - 15.19862***.
3	El mejor modelo es el que se genera utilizando como variable independiente la columna facebook. Se obtuvo: ***ventas = 7.072897 + 0.047335(facebook)***
4	El mejor modelo múltiple encontrado contiene las variables facebook e instagram sin datos atípicos. Se obtuvo : ***ventas = 3.570642 + 0.04288(facebook) + 0.196206(instagram)***
5	***No es posible rechazar la hipótesis nula***, no existe suficiente evidencia para afirmar que las ventas en la región 1 son superiores a 150.
6	Las ventas estimadas son 21.3, 22.2, 23.1, 23.6, 26.0 para los presupuestos 300, 320, 338, 350 y 400 respectivamente. El modelo se ajusta en buena medida.
7	No hay suficiente evidencia para concluir que las ventas medias de las variables de estudio difieren con respecto a las regiones.
8	Tanto los resultados del inciso 2 como 7 indican que no se puede afirmar que las medias varían por región.

Planteamiento del problema

El objetivo general, es estudiar la relación del número de ventas (la variable dependiente) con los presupuestos asignados a cada medio y las ingresos por región. Con el fin de poder asesorar el departamento de ventas de la empresa. Para ello se realizan los análisis:

1. Realice un análisis descriptivo y exploratorio de los datos.

2. Calcular el intervalo de confianza del 95% para las medias de ventas por región.

3. Encuentre el modelo de regresión simple que mejor se ajuste a los datos.

4. Consiga el modelo múltiple más apropiado. Considere un nivel del 5%.

5. Estudios previos indican que las ventas en la región 1 muestran un precio de 150 (millones), aunque estudios suponen que dicha cantidad es superior a la mostrada por este análisis.

6. Para el modelo obtenido en el inciso 3, realice la predicción correspondiente a 5 ventas.

7. ¿Existe suficiente evidencia que permita concluir que las ventas media de las variables de estudio difieren con respecto a las regiones?

8. ¿Cómo compararía los resultados del intervalo de confianza del inciso 2 con el ANOVA?

Metodologia

1. Para el analizar los datos se emplean herramientas gráficas como: diagramas de caja e histogramas, e indicadores numéricos: media, cuartiles, mediana, desviación estándar y coeficiente de variación para cada variable. Además de un estudio de la matriz de correlación.

2. Para calcular los intervalos de confianza del 95%, se procede a ejecutar la función t.test(ventas_por_region, conf.level = 0.95)$conf.int sobre las ventas en cada región.

3. Se estiman los parámetros desconocidos mediante el método de mínimos cuadrados para encontrar los modelos lineales entre la variable dependiente *ventas* con las variables dependientes, *Región* y demás medios de publicidad. Se establece el modelo que mejor se ajusta mediante la comparación de: su error, ajuste, y en base a un análisis de residuales en el que se busca principalmente normalidad y homocedasticidad.

4. Para estimar el mejor modelo múltiple, se utiliza el método de regresión hacia atrás paso a paso partiendo del modelo que contiene a todas las variables. Se tiene especial cuidado en el estudio de las normalidad de los residuales, en la significancia de las variables sobre el resultado y en los valores del ajuste y error al calcular cada modelo.

5. Se realiza una prueba de hipótesis sobre la igualdad de media para muestras grandes. Se calcula el estadístico Z y su p-valor asociado. Adicionalmente, se corrobora el resultado usando la función t.test().

6. Establecido el mejor modelo lineal en el inciso 3, se generan los valores predictivos de ventas para 5 nuevos presupuestos en publicidad. También se generan puntos equidistantes para calcular las bandas del intervalo de predicción y confianza para dicho modelo, empleando los comandos, *sequence* y *predict* de R.

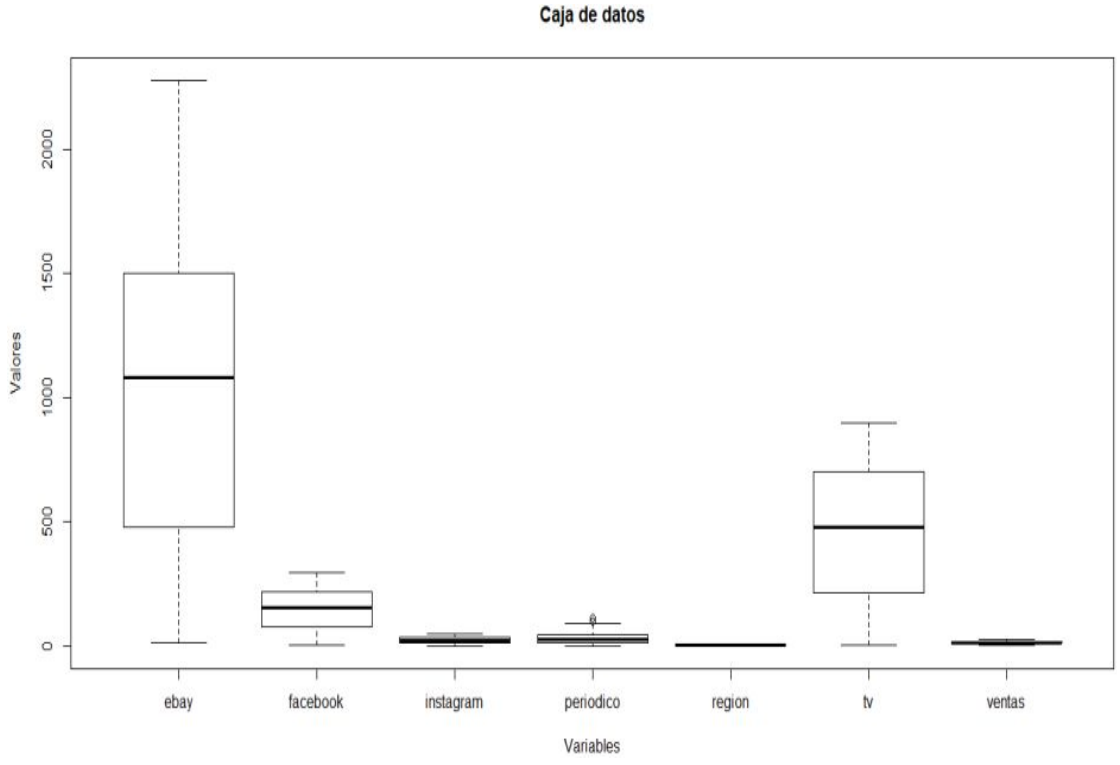
7. Se emplea el procedimiento de análisis de varianza para el diseño de un cofactor. Como H0: no existe diferencia de medias con respecto al país de procedencia y Ha: que si existe alguna diferente. Para esta prueba se utiliza el comando *anova*.

8. Se comparan los resultados del inciso 2 y 7 dado que los intervalos de confianza y las pruebas de hipótesis teóricamente arrojan resultados consonantes.

Desarrollo

1. Mediante el cálculo de las estadísticas descriptivas, pudimos observar que las ventas tienen una distribución normal

El presupuesto destinado a la publicidad en Ebay es un 50% de las veces mayor que cualquier presupuesto dirigido a otro medio publicitario. Análogo, para la Tv con el resto de medios restantes. E igualmente para Facebook. Por lo tanto las mayores inversiones en publicidad se dirigieron a Ebay con diferencia, por el contrario Instagram recibió la menor inversión total. Sin embargo, el rango de presupuesto asignado a cada medio no es estrictamente superior a ningún otro.



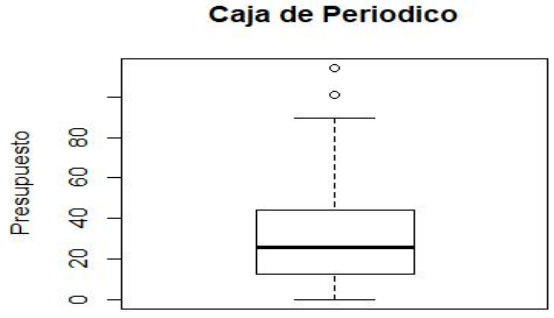
En general el valor promedio de inversión de cada medio publicitario se encuentra cerca muy cercano al valor ubicado en la mitad de los datos. Además, se observa que los cuartiles correspondientes presentan simetría respecto a la mediana, exceptuando el caso de la inversión en periódicos.

Finalmente, se aprecia que la diferencia en el número de ventas por región no es significativo.

Tabla 1. Resúmen Estadístico para las variables

Resúmen Estadístico								
Variable	**Mínimo**	**1er Cuartil**	**Mediana**	**3er Cuartil**	**Máximo**	**Media**	**Desviacion estandar**	**Coeficiente de Variación**
Ventas Totales	1.6	10.5	12.9	17.4	27	14.12	5.18	0.36
Ventas en Facebook	0.7	75.3	151.5	219.8	296.4	149	85.06	0.57
Ventas en periódicos	0.3	12.8	25.6	44.3	114	30.5	21.76	0.71
Ventas en Instagram	0	9.9	23.3	36.6	49.6	23.3	14.94	0.64
Ventas en Televisión	1	215	479	703	900	467.9	273.62	0.58
Ventas en Ebay	10.1	479.9	1083.4	1499.7	2277.6	1060.5	617.84	0.58
Número de Ventas Por Region	1	2	2	3	4	2.497	1.11	0.44

En el caso de la inversión publicitaria en periódicos, se evidencia que hay inversiones inusualmente altas, lo cual concuerda con el estudio de su coeficiente de variación, el cual es relativamente alto (0.71).



Calculada la matriz de correlación se observa que los medios Facebook e Instagram son los que poseen mayor relación con Ventas, y dichas relaciones son las más significativas. Esta información permite tener una idea de las variables que se ajustan mejor para la elaboración de un modelo lineal que aproxima el resultado de las ventas por región.

Tabla 2. Matriz de Correlación de las variables

Matriz De Correlación							
Variables	**Ventas**	**Facebook**	**Periodico**	**Instagram**	**Tv**	**Ebay**	**Region**
Ventas	1	0.77	0.24	0.57	-0.06	0.24	0.05
Facebook	0.77	1	0.06	0.05	-0.04	0.27	0.03
Periodico	0.24	0.06	1	0.36	-0.03	0.06	0.02
Instagram	0.57	0.05	0.36	1	-0.05	0.07	0.1
Tv	-0.06	-0.04	-0.03	-0.05	1	-0.11	0
Ebay	0.24	0.27	0.06	0.07	-0.11	1	-0.1
Region	0.05	0.03	0.02	0.1	0	-0.1	1

2. El estudio de los intervalos de confianza del 95% para el promedio de ventas del productos por región arroja los siguientes resultados:

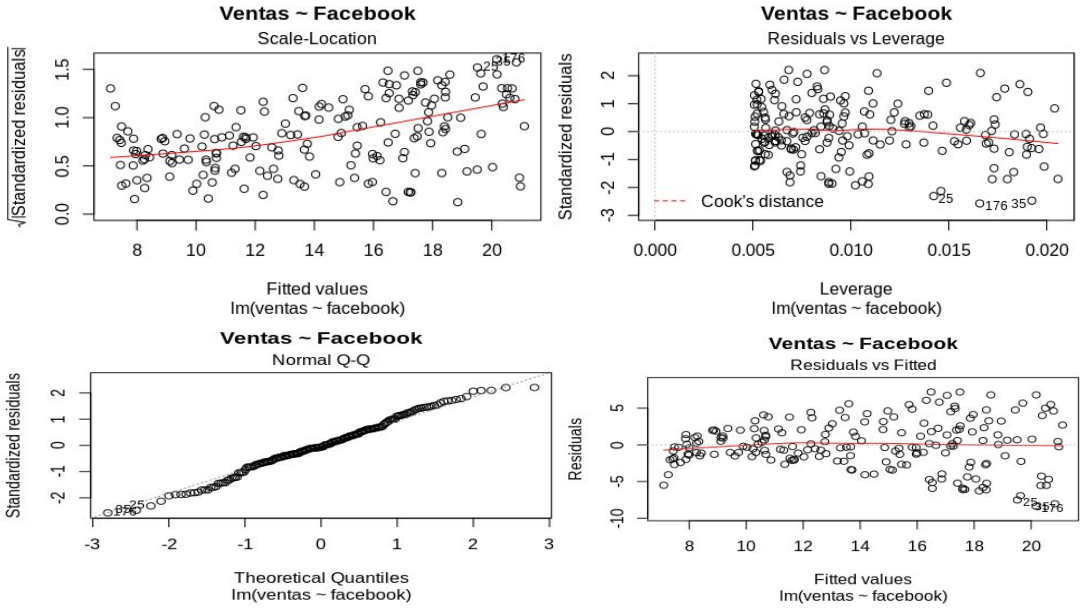
Tabla 3. Intervalos de confianza para la media de ventas por región

Intervalos de confianza para la media de ventas por región				
Región	**1**	**2**	**3**	**4**
Intervalo	12.26927 - 15.23277	11.73804 - 14.94596	13.99225 - 17.02407	12.62179 - 15.19862

Se evidencia que el valor medio de las ventas es similar entre cada una de las regiones. En el caso de las regiones 1 y 4, el intervalo difiere en un máximo de 0.4 décimas, lo cual indica que el promedio de ventas en estas regiones está igualado. Así mismo, en la región 2 se aprecia el intervalo que presenta el menor rango de valores, siendo el límite superior muy cercano a la media de ventas total (14.12), en la región 3 se aprecia el mayor rango de valores para la media de las ventas. Esto se complementa con el estudio previo del histograma de las ventas por región donde la frecuencia de ventas es similar.

3. Para obtener el modelo que más se ajuste a los datos, se calculan modelos lineales simples con ventas como variable dependiente para cada una de los medios de publicidad (variables independientes). Para ello se emplea el comando *lm* de R. Luego, para cada uno de los modelos se analizó el comportamiento de sus residuales, significancia, su normalidad, el error y su ajuste. Se obtuvo que en general los modelos eran bastante normales y sus variables poseían la mayor significancia (excepto por región y tv que no poseen significancia a ningún nivel, por lo cual fueron descartados como posibles mejores modelos).

También se descartó el modelo respecto a instagram por presentar una fuerte asimetría en sus valores mínimos y máximos para los residuales respecto a la mediana. Entre los modelos restantes, aquel con el presupuesto asignado a Facebook como variable independiente es el que presenta el menor error (3.279) y el mayor ajuste de todos (0.6005), a la vez que sus datos son independientes de los residuales pues, en la gráfica Residual vs Fitted, no se aprecia ningún patrón, que estos tienen forma normal, y además son homocedásticos dado que, en la gráfica Residuals vs Leverage, no se aprecia ningun patron. Por lo tanto es considerado como el mejor modelo lineal simple.



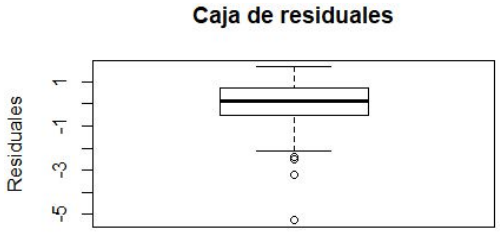
4. Para el ajuste del modelo múltiple, se utiliza el método de regresión hacia atrás paso a paso partiendo del modelo que contiene a todas las variables. Luego, para cada modelo obtenido, se elimina la variable cuyo valor de t, en valor absoluto, es menor. Esto es, se elimina la variable que tiene la menor significancia sobre el modelo. Este proceso se repite hasta obtener un modelo que cumpla con que todas las variables tengan una significancia igual o menor a 0.05.

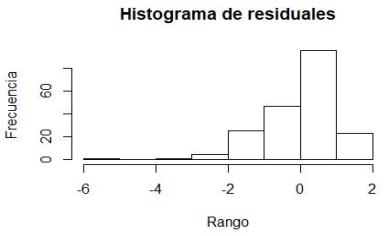
Tabla 4. Muestra del método regresión hacia atrás paso a paso.

Modelos múltiples								
Variable	**Mínimo**	**1er Cuartil**	**Mediana**	**3er Cuartil**	**Máximo**	**Ajuste**	**Error**	**T mínimo**
Ventas ~ facebook + periodico + instagram + tv + ebay + region	-8.6264	-0.8568	0.3028	1.1658	2.7218	0.8923	1.703	0.076
Ventas ~ facebook + instagram + tv + ebay + region	-8.6134	-0.8553	0.2854	1.1696	2.7148	0.8928	1.698	0.1
Ventas ~ facebook + instagram + tv + region	-8.6196	-0.8595	0.2744	1.1774	2.7183	0.8934	1.694	0.127
Ventas ~ facebook + instagram + region	-8.6306	-0.8414	0.2913	1.1594	2.7272	0.8939	1.689	1.214
Ventas ~ facebook + instagram	-8.8196	-0.8501	0.2429	1.2016	2.8138	0.8937	1.691	-
Ventas ~ facebook + instagram con datos atípicos eliminados	-2.8224	-0.6602	0.0229	0.8263	2.3964	0.9486	1.158	-

La tabla anterior muestra los resultados del método de regresión hacia atrás paso a paso. La primera variable en ser eliminada es periódico por tener el menor valor de t en valor absoluto. Sucesivamente se eliminan ebay, tv y región por la misma razón. Finalmente, se obtiene el modelo múltiple que cumple con la significancia solicitada usando las variables facebook e instagram.

Este resultado tiene sentido, ya que, las variables facebook e instagram son las que presentan los mejores modelos simples, en cuanto a error y ajuste, y tienen la mejor correlación con la variable ventas.

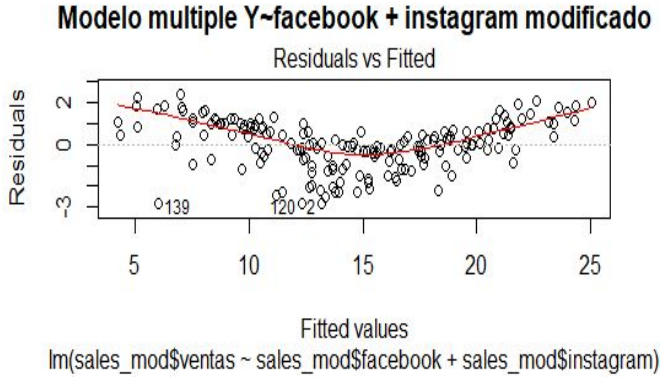


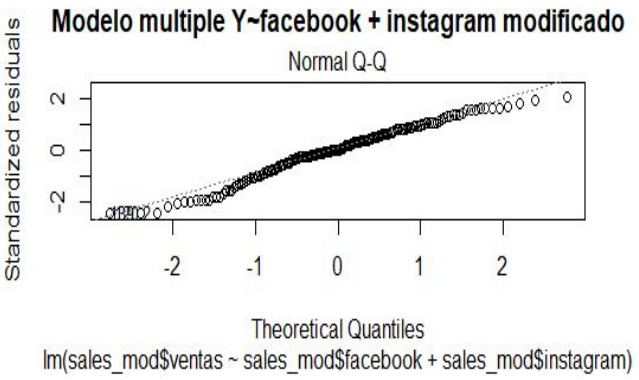


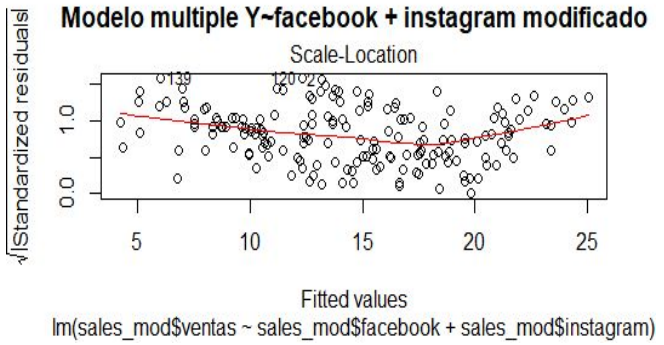
El modelo obtenido presenta tres problemas fundamentales. El primero es que los residuales carecen de simetría del máximo, mínimo y cuartiles respecto de la mediana. Segundo, al estudiar la normalidad de los residuales se observan valores atípicos que afectan la efectividad del método. Por esta razón, se procede a realizar correcciones en los datos de entrada para ajustar la normalidad de los residuales.

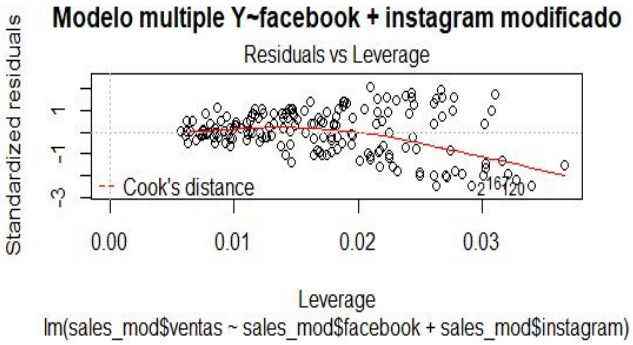
Las correcciones corresponden con retirar del conjunto de datos los siguientes índices: 128, 124, 56, 6, 77, 3, 74, 25, 186, 176, 171, 167, 164, 156, 129, 130, 101, 35, 80, 163, 133.

De esta manera, se consigue un modelo que cumple con el nivel de significancia solicitado con especial cuidado en la distribución normal y simétrica de los residuales.



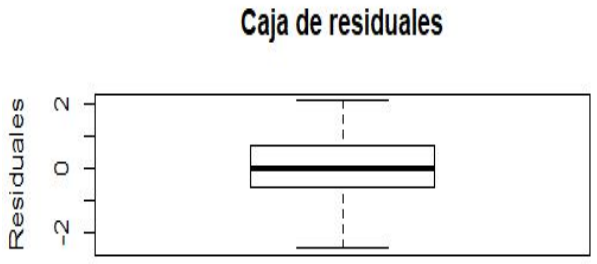


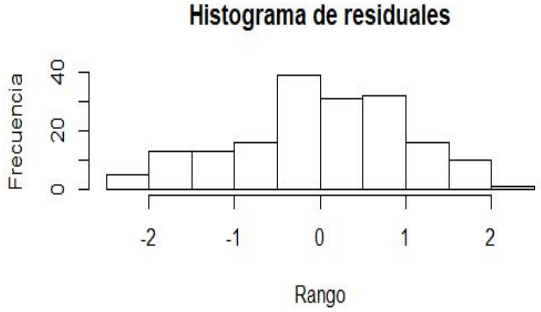




Se puede observar que para este modelo los datos son independientes de los residuales pues, en la gráfica Residual vs Fitted, no se aprecia ningún patrón, los residuales tienen forma normal, y además estos son homocedásticos, ya que en la gráfica Residuals vs Leverage no se aprecia ningun patron. Así mismo, existe una cierta simetría entre los residuales que supera a la del modelo anterior obtenido.

Este nuevo modelo reduce significativamente el error y ajusta de mejor manera la normalidad de los residuales. Por lo tanto, este es el modelo múltiple más apropiado que cumple con el nivel de significancia del 95%.





5. El resultado del análisis del precio de los productos en la región 1 se obtiene mediante la realización de una prueba de hipótesis. En particular, una prueba sobre la media.

Se plantea como hipótesis nula Ho: $\mu=150$

Se plantea como hipótesis alternativa Ha : $\mu>150$.

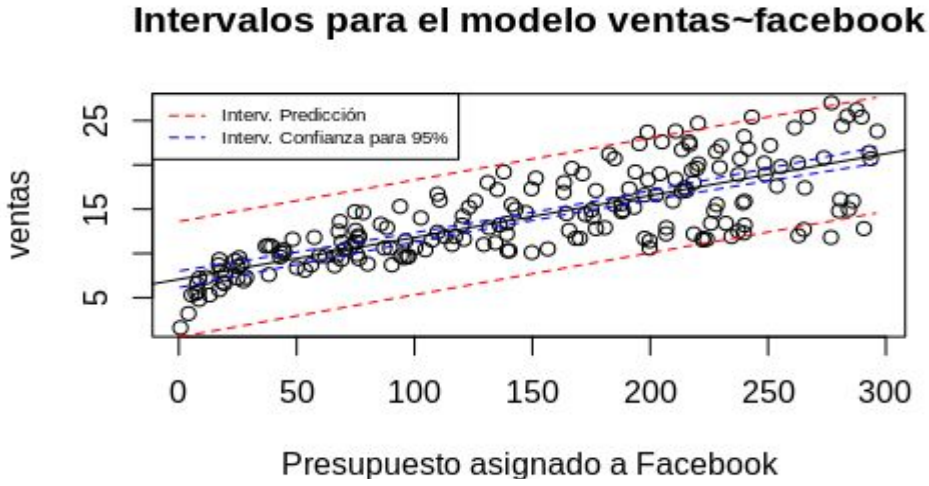
Una vez establecidas las hipótesis se realiza la prueba, usando un nivel de confianza de 99% ($\alpha = 0.01$), y ya que se trata de muestras grandes, se realiza una prueba mediante el cálculo del estadístico Z = (media muestral - $\mu 0$) / (desviación estándar / raíz del número de

datos). De este cálculo, se obtiene como resultado Z = -184.88 con un p-valor asociado igual a 1. Dado que para todo nivel de α, no es posible rechazar la hipótesis nula, se tiene que no existe suficiente evidencia para afirmar que las ventas en la región 1 son superiores a 150.

Observación: Al tomar como hipótesis alternativa Ha : μ<150, y realizando la prueba anterior calculando calculando el p-valor para cola inferior. Se observa que el resultado del p-valor (0.0483), la hipótesis nula debe ser rechazada para un nivel de significancia del 90% y 95%. Pero para un nivel de significancia del 99%, esta hipótesis no es rechazada, lo cual concuerda con el resultado obtenido en la prueba original. Esto es interesante, ya que, la media de ventas en la región 1 es 13.75102. Un valor muy lejano a 150.

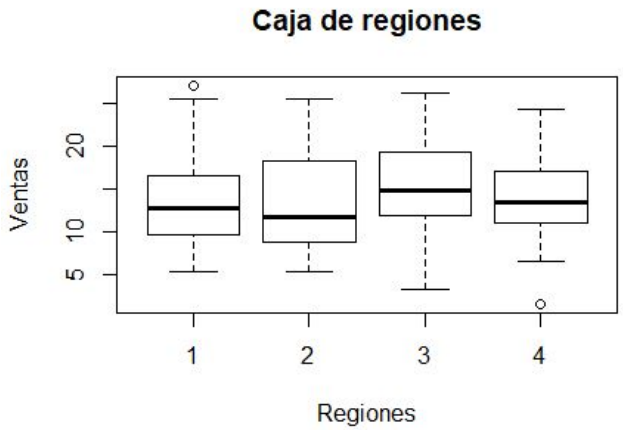
6. Se emplea el modelo ventas ~ facebook, considerado como el mejor modelo lineal simple, según lo argumentado en el inciso 3, para generar predicciones para las ventas dados los 5 nuevos presupuestos. Este cálculo fue realizado con el comando *predict* de R, que sustituye los valores en el modelo previamente definido. Se obtiene que las ventas estimadas son 21.27333, 22.22003, 23.07206, 23.64007, 26.00681 para los presupuestos 300, 320, 338, 350 y 400 respectivamente.

Luego, se genera una secuencia de puntos entre el mínimo y el máximo, de uno en uno, para los valores que se quieren estimar para generar las bandas de confianza y predicción para el modelo. Finalmente se grafican las bandas previamente mencionadas junto con los puntos iniciales para analizar el modelo.



Se observa que a pesar de que prácticamente todos los puntos se encuentran dentro de las bandas del intervalo de predicción, un porcentaje significativamente menor se encuentra entre las bandas del intervalo de confianza del 95%. Esto indica que aun cuando el modelo lineal no tiene un ajuste tan alto como se quisiera, dado que la mayoría de los datos están fuera de la banda del intervalo de confianza, sigue siendo un modelo que se ajusta en buena medida ya que casi todos los datos siguen estando en la banda de predicción.

7. Se realiza un análisis de varianza para un diseño de un factor, donde H0: Las medias de ventas entre las regiones son iguales y Ha: alguna difiere. Se observa que los valores de Q2 son cercanos entre sí, de manera que las medias de cada región se encuentran dentro del rango intercuartil resto de las regiones, los cuales son similares entre sí.



Como p-valor > 0.03 (nivel de significancia de la prueba) no se rechaza la hipótesis nula. Entonces no hay suficiente evidencia para concluir que las ventas medias de las variables de estudio difieren con respecto a las regiones.

8. En el anova del inciso 7 se obtuvo que no había suficiente evidencia para concluir que las ventas medias de las variables de estudio difieren con respecto a las regiones con una significancia de 0.03. Esto puede compararse con los resultados obtenidos en el inciso 2 para los intervalos de confianza de las medias debido a que, a pesar de que el nivel de confianza era del 95%, se tenían intervalos similares donde la intersección entre estos intervalos no da vacío, así que en ambos casos se puede concluir que las medias son iguales.

De hecho, observando el p-valor del anova, el cual dió como resultado 0.1747, no se podría rechazar la hipótesis de que las medias son iguales con un nivel de significancia del 5% el cual corresponde con una confianza del 95%.

Conclusiones y Recomendaciones

Del estudio inicial de los datos, se observa que la mayor inversión en publicidad, por mucho, se realiza mediante Ebay. Luego del estudio de la matriz de correlación se observa que esta variable (ebay) tiene una correlación muy baja con la variable ventas, por lo que se desaconseja asignar grandes cantidades de presupuesto a este medio, si lo que se desea es obtener un impacto proporcional sobre las ventas. Por otro lado, las variables facebook e instagram son las que poseen las mayores correlaciones con la variable ventas.

El mejor modelo lineal simple es el conseguido con la variable facebook, a este le sigue el modelo conseguido con la variable instagram. Sin embargo, el modelo que mejor predice el resultado de las ventas es el modelo múltiple que depende de ambas variables. Por lo tanto, se recomienda priorizar la asignación de fondos en publicidad a los medios de publicitarios Facebook e Instagram, por lo antes mencionado.

Del estudio de las ventas, se observa que la diferencia de estas por región es poco significativa. Por lo que la ganancia obtenida por la venta de un producto no depende de la región donde este fue vendido. Esto se corresponde con los resultados del análisis de varianza, donde no hubo suficiente evidencia para concluir que las ventas medias difieren con respecto a las regiones. También se estudió la media de las ventas para el caso de la región 1. No obstante, en vista de lo anterior podemos extrapolar este caso para el resto de regiones y afirmar que su media para cada región definitivamente no es mayor a 150 millones.

Se debe tomar especial cuidado al momento de seleccionar los datos. Se evidenció la existencia de valores negativos, que por obvias razones, carecen de sentido. Estos valores errados fueron eliminados del conjunto de datos antes de realizar algún estudio.

Por otra parte, se debe cuidar el formato de los datos a utilizar. Se corrigieron los valores que tenían comas en vez de puntos para indicar decimales, ya que, al ser cargados en el programa, no se reconocían los valores de la fila que los contiene.

Bibliografia

- R Documentation and Manuals. Disponible en: https://www.rdocumentation.org/
- Ovalles, Villalta y Martínez. Intervalos de Confianza, parte 1 Muestras grandes.
- Ovalles, Villalta y Martínez. Modelos lineales, parte 1.
- Ovalles, Villalta y Martínez. Modelos lineales, parte 2.
- Ovalles, Villalta y Martínez. Modelos lineales, parte 3.
- Ovalles, Villalta y Martínez. Análisis de Varianza, parte 1.

Anexo Código En R