En el siguiente estudio, realizamos el estudio estadístico de la empresa HBAT en busca de la influencia de ciertas conductas de la empresa sobre la probabilidad de que los clientes recomienden comprar en HBAT.

Estudio HBAT

Probabilidad de recomendación de la firma

Mathias Badilla
Universidad de Concepción

Resumen.

En el siguiente análisis realizamos un estudio a la empresa HBAT industries, empresa dedicada a la fabricación y venta de productos del papel. Indagamos como las distintas variables que ellos consideraron en su estudio a 2000 clientes, influyen en la probabilidad de que un cliente recomiende la marca. Para llevar a cabo todo esto usamos el método de mínimos cuadrados ordinarios mediante el programa Gretl y luego entregaremos la regresión lineal con sus respectivos factores y su correcta interpretación.

Abstract.

In the following analysis, we carried out a study of HBAT industries, a company dedicated to the manufacture and sale of paper products. We investigated how the different variables that they considered in their study to 2000 customers, influence the probability that a customer recommends the brand. To carry out all of this, we use the Ordinary least squares method through the Gretl program and then we deliver the linear regression with its respective factors and its correct interpretation.

1.-Introduccion

HBAT industries, empresa dedicada a la fabricación y venta de productos de papel, está realizando un estudio con 2000 clientes, los cuales se dividen en dos segmentos del mercado, el primero son los diarios y el segundo son las revistas.

HBAT esta intereseda en analizar como es que diversas variables de estudios que ellos definieron, inlfuyen en la probabilidad de que sus clientes recomienden la marca.

Para ello consideraremos los resultados de su estudio, extrayendo una muestra de 200 clientes y realizaremos el modelo de mínimos cuadrados ordinarios con el programa Gretl.

Posteriormente procederemos a analizar los supuestos de la regresión lineal aplicadas a nuestra regresión, dando valores que aprueben o rechacen las hipótesis, acompañando en los casos correspondientes con tablas y gráficas, esto con el fin para dar correctas interpretaciones sobre los resultados que nos entregue el modelo sobre los factores que acompañen a cada variable.

2. Metodología.

2.1 Elección de la muestra

La forma en que se eligieron las 200 observaciones aleatorias del total de 2000 observaciones sobre las percepciones de los clientes de HBAT Industries, se explica a continuación.

El proceso fue el siguiente, al costado de las 2000 observaciones se impuso un numero aleatorio entre 0 y 1 de 8 decimales, los cuales fueron ordenados de menor a mayor esto afectando el orden de las observaciones. Fue así que se seleccionaron las primeras 200 observaciones.

Esta muestra fue exportada a otro archivo Excel el cual fue ingresado al programa Gretl donde se procedió a seleccionar el modelo de "Mínimos cuadrados ordinarios" (Desde ahora MCO) para realizar las primeras pruebas estadísticas de la muestra y así empezar su estudio.

2.2 Primeras pruebas

Para que una regresión lineal sea válida, hay 5 supuestos básicos que debe cumplir y son en los que nos preocuparemos para que nuestra regresión lineal sea válida y se pueda dar una correcta interpretación.[3]

-**Linealidad**. La forma en que trabajaremos con este supuesto será mediante el contraste Reset de Ramsey.

-**Independencia**. Esto lo probaremos mediante la prueba de Durbin-Watson y de manera gráfica.

-**Homocedasticidad**. Este supuesto lo trabajaremos mediante el test de White.

-**Normalidad**. Se trabajará con la prueba de normalidad ejecutada por Gretl.

-**Colinealidad**. Será estudiando mediante el factor de inflación de varianza.

Une vez teniendo nuestra muestra, procedemos a realizarles algunos análisis para ver su calidad e ir comparando con los resultados que vayamos obteniendo más adelante e identificar posibles problemas que pudiese tener con los supuestos que debiese cumplir la regresión lineal.

Los primeros datos que obtenemos al realizar el método de MCO son:

R-cuadrado=0.61401

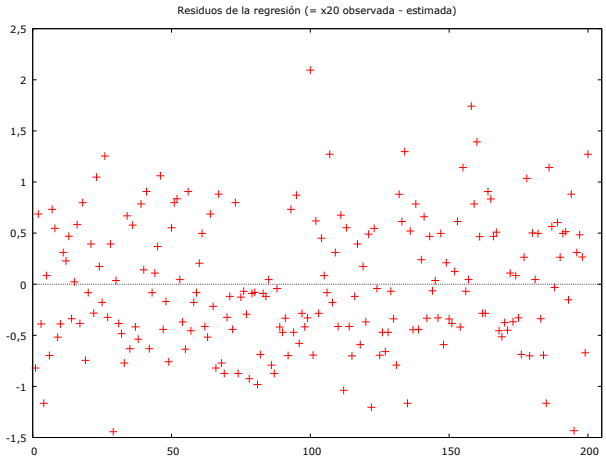
R-cuadrado ajustado=0.573267

Dado que estas son variables que nos ayudan a entender que tan efectivas son las variables independientes explicando la variable dependiente, podemos notar que ambos valores son superiores al 0.5 lo cual nos da una buena señal en cuanto a la explicación que están dando las variables.

Partiendo por el primer supuesto mencionado, haremos análisis de linealidad, como ya fue mencionado usaremos el contraste Reset de Ramsey, el cual nos entrega

Valor-p=0.382995

Por lo que aceptamos la hipótesis de que no hay problemas de linealidad en nuestra muestra, cumpliendo el supuesto de linealidad.



Y observando la gráfica de residuos/número de observación, generado por Gretl, podemos notar que su distribución no tiene una tendencia de dependencia con los valores anteriores.

Considerando ambos test, podemos concluir que **hay pruebas suficientes para afirmar que se cumple el supuesto de Linealidad**

Mejores estimadores lineales insesgados (Desde ahora MELI) ni son tampoco eficientes.

El siguiente análisis que se hizo a la muestra, fue analizar la Autocorrelación mediante la prueba de Durbin-Watson y así comprobar el supuesto de independencia. El valor d fue calculando exportando los residuos calculados por Gretl a Excel e incluir la formula, dándonos el siguiente valor,

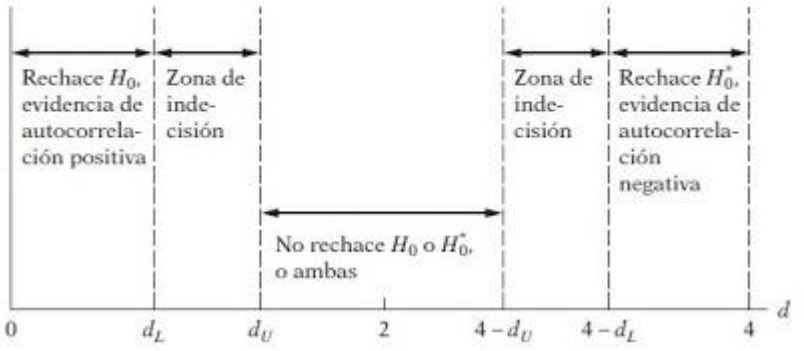
d=2.6824

Luego del programa Gretl se obtuvieron los valores dL y dU que nos permitirán establecer las zonas de rechazo y de incertidumbre para nuestra muestra, obteniéndose los valores siguientes, y siguientes intervalos

dL=1.5653

dU=1.9787

Estadístico *d* de Durbin-Watson



2

Rechazo: [0 ; 1.5653] U [2.4347 ; 4] Indecisión: [1.5653 ; 1.9787] U [2.0213 ; 2.4347]

Como podemos ver, nuestro valor "d" se encuentra dentro de la zona de Indecisión y por tanto es no concluyente, es mejor rechazar H_0 de la prueba de Durbin-Watson y **se concluye que sí hay problemas de linealidad dentro de nuestra muestra.**

El siguiente supuesto que debe cumplirse es el de Homocedasticidad, en este punto accedimos a la prueba de White, desde el programa Gretl, de donde obtuvimos un valor-p,

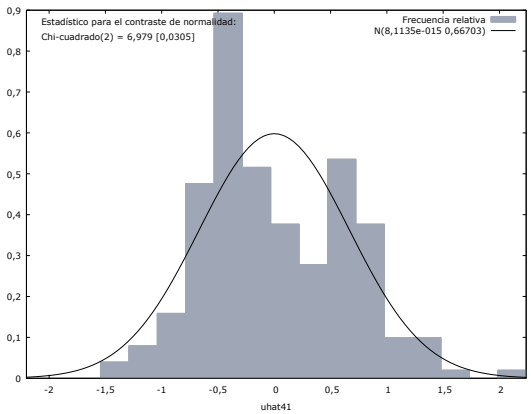
Valor-p=0.0397334

Considerando el valor-p entregado, procedemos a rechazar la hipótesis nula de no haber heterocedasticidad, por lo que se puede concluir que **tenemos pruebas suficientes para afirmar que en nuestra muestra sí hay Heterocedasiticidad**, esto trayendo de implicancia que nuestros estimadores ya no son los Mejores estimadores lineales insesgados (Desde ahora MELI) ni son tampoco eficientes.

Continuando con el análisis de normalidad, se realiza el test de normalidad, arrojando los siguientes resultados,

Valor-p=0.03502

Y de manera anexa, obteniéndose la siguiente gráfica,



Por tanto, en consideración de ambos análisis**, puede concluirse que poseemos pruebas suficientes para afirmar que la muestra no posee una distribución normal**.

Nos queda analizar el ultimo supuesto, el de colinealidad. Para realizar esta prueba se procede a calcular el Factor de inflación de la varianza (Desde ahora VIF, por sus siglas en ingles de "Variance inflation factor"). Usando Gretl para calcular el VIF, obtenemos los siguientes resultados,

Variable	VIF	Variable	VIF
X6	3.326	X13	2.113
X7	4.258	X14	5.100
X8	4.434	X15	1.165
X9	4.941	X16	2.994
X10	1.859	X17	72.856
X11	83.896	X18	92.961
X12	5.711	----------	----------

Se considera que hay problemas de multicolinealidad cuando el VIF alcanza valores superiores o iguales a 10, en nuestra muestra, el VIF alcanzo valores sobre 10 en 3 variables, por lo **que tenemos pruebas suficientes para afirmar problemas de colinealidad.**

2.3 Corrigiendo los problemas a los supuestos de regresión lineal

Nos damos cuenta de que hay varios problemas que deberíamos solucionar para poder hacer cumplir los supuestos de una regresión lineal, partiremos por la heterocedasticidad ya que problemas con esta puede producir que variables que sí son significativas aparezcan como no significativas, lo cual podría influir en tomar malas decisiones sobre las variables.

Las primeras tres medidas tomadas, sin éxito ninguna ya que los valores p arrojados en las pruebas se aproximaban a 0.0023, fueron las siguientes,

-Transformación de la variable dependiente, aplicando logaritmo natural.

-Transformación de la variable dependiente, aplicando el cuadrado de la variable.

-Transformación de la variable dependiente, dividiéndola por su desviación estándar.

Luego se procedió a realizar una transformación de todas las variables, excluyendo las dicotómicas, aplicándoles logaritmo natural, sin tampoco tener una real influencia en nuestro problema de heterocedasticidad.

Finalmente se optó por realizar una transformación de todas las variables independientes, exceptuando las dicotómica a cada una se le dividió por la desviación estándar de la muestra y luego se volvió a aplicar el test de White, dándonos el siguiente valor,

Valor-p=0.055833 => Rechazar hipótesis de heterocedasticidad.

Ya teniendo ese valor-p podemos decir que **no hay razones para creer que nuestra muestra posee heterocedasticidad, si no más bien homocedasticidad**.

Ahora, luego de haber hecho esta transformación en las variables las interpretaciones sobre los cambios en las variables independientes no se mantienen, hay que considerar el n aumento en desviaciones estándar de Y, por el aumento de una desviación estándar en X.

Como ya veíamos, tenemos 3 valores con un VIF sobre 10, haremos las 3 pruebas, eliminando uno por uno, se presentan las siguientes condiciones respecto al VIF de cada variable

Eliminada	Mantenida	x11	x17	x18
X11		-	<10	<10
X17		<10	-	<10
X18		<10	<10	-

Dado que no hay distinción respecto al cual borrar, procederemos a buscar aquella que tanga mayor valor-p, buscando así la que menos influencia tenga.

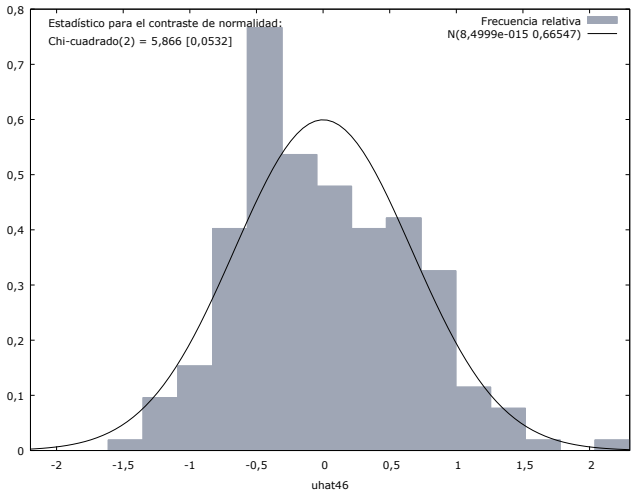
Variable	Valor-p
x11	0.5175
x17	0.5007
x18	0.5019

Por lo que se procede a eliminar la variable x11, asociada a “Linea de productos”, dejando esta resolución un efecto **de corrección sobre el problema de colinealidad.**

Sobre el supuesto de normalidad, notamos lo siguiente, al omitir las variables ficticias asociadas a la variable x1, nuestro modelo se comportaba de manera normal, obteniéndose los siguientes resultados.

Valor-p=0,0532489

Y la siguiente grafica,



Podemos notar como omitiendo esta variable notamos una corrección en la distribución normal que tienen los residuos, pero esto a su vez nos provoca un cambio en la heterocedasticidad, dejándonos el siguiente resultado la prueba de White,

Valor-p=0,00117329

Considerando lo acontecido con las variables, y luego de varios intentos realizando inclusive segundas transformaciones, se decidió no extrae la variable ficticia x1 **manteniendo los problemas de normalidad.**

Analizando el ultimo supuesto, el de independencia es que luego de ya todas las correcciones realizadas se procede a recalcular el valor “d” con los nuevos residuos generados, obteniéndose lo siguiente,

d=2.7101

Donde con la actualización de las variables dL, dU tenemos,

dL=1.5878

dU=1.9547

Podemos notar como ahora nuestra prueba nos arroja directamente a la zona de rechazo, por lo que las transformaciones realizadas a nuestro modelo no son suficientes para corregir los problemas de independencia. Al igual que con la normalidad, se intentaron segundas transformaciones de las variables independientes, pero ninguna fue influyente en corregir los problemas, por lo que podemos aseverar **mantener problemas de Independencia**

3.- Resultados.

3.1 Para los supuestos.

Ya hemos analizado los 5 supuestos aplicados a nuestra regresión, donde al iniciar el estudio presentábamos problemas con Heterocedasticidad, Colinealidad , normalidad e Independencia, pero mediante transformaciones a las variables independientes, y eliminación de alguna de las variables, logramos corregir los problemas asociados a la heterocedasticidad y la colinealidad, mantener los problemas en normalidad e independencia nos traerá problemas sobre no tener estimadores MELI, por lo que las interpretaciones de nuestra formula de regresión lineal no nos dará una correcta interpretación.

3.1 Regresión lineal

$$pX_{20} = 7.88 - 0.194X_{1.1}.0.0106X_{1.2} - 0.135X_2 - 0.823X_3 + 0.0760X_4 - 0.687X_5 + 0.368X_6 + 0.141X_7 + 0.339X_8 + 0.0638X_9 - 0.139X_{10} + 0.118X_{12} + 0.0358X_{13} - 0.301X_{14} + 0.149X_{15} + 0.124X_{16} - 0.0170X_{17} + 0.0239X_{18}$$

Podemos ver ya nuestra regresión lineal terminada, cabe recordar que la interpretación de los valores no es la forma directa por la transformación de las variables independientes antes mencionada (Interpretación en el punto 2.2), ya con la regresión terminada podemos ver la influencia que cada variable seleccionada a estudio por HBAT industries tiene sobre la variable dependiente “Probabilidad de recomendación de la firma”

4.- Conclusiones y discusiones

4.1 Efectos relacionados.

Pudimos notar en reiteradas oportunidades como es que la busca de la corrección algunos de los supuestos, impactaba en otro, buscando arreglar la normalidad mediante el aumento de la muestra, la normalidad se vio corregida, pero aumento nuevamente la heterocedasticidad. Finalmente llegar a modelo optimo es juego en ciclo con las distintas variables y como las aplicamos o transformamos para poder tener una regresión lineal adecuada.

4.2 Variables muy influyentes.

La variable ficticia x1, fue una variable que trajo problemas con el supuesto de normalidad, quitarla implicaba arreglar la normalidad pero generar problemas de heterocedasticidad, aquí notamos como es que mismas variables pueden influir mucho en el desarrollo de un modelo, no se le aplicaron transformaciones por ser ficticia, pero deben haber otros métodos no utilizados en esta investigación que nos permitiese hacer un trabajo optimo con tales variables.

4.3 Residuo como fuente de estudio.

Ya comentábamos como la variable x1 generaba problemas en la normalidad, esto se descubrió analizando la grafica de normalidad y dándonos cuenta que los principales problemas estaban en los residuos negativos, ahí se decidió estudiar aquellas variables que su factor que la acompañe fuese negativo, así es que llegamos a la variable x1 como fuente del problema.

5.-Bibliografía.

1. Emelina Lopez Gonzales (1998)
   Tratamiento de colinealidad en regresión múltiple
   Psicothema, 1998. Vol.10,n2, pp. 491-507

2.- Guillermo Atilano (2019)

   Econometría 1

   http://sgpwe.izt.uam.mx/Curso/31576.Econometria-I/Tema/16317.Autocorrelacion.html

3.- Raul Martin Martin

   Supuestos del modelo de regresión lineal

   Escuela superior de informatica

Anexo

Durante la realización de la tarea siento como ciertos conceptos se fueron haciendo reales, si bien no logré llegar a la solución óptima del problema, el tiempo invertido, el cual no pensé honestamente seria tanto, me ayudo a entender de buena manera todos los conceptos visto en clases. Encontré, si bien un poco estresante la tarea, a la vez muy divertida, se generaron instancia de discusión con los compañeros las cuales son muy buenas a la hora de aprender y más considerando el contexto actual.

Fue un poco difícil lidiar con los problemas que se iban arreglando y generando otros errores, de todas formas, quedara pendiente estudiar las otras formas de solucionar ese tipo de situaciones.

"Declaro que este informe ha sido desarrollado por el autor en forma estrictamente individual sin participación de terceros, y que el contenido es totalmente original y no incluye partes de otros trabajos similares. Entiendo que, si se comprueba lo contrario, el profesor se reserva el derecho a modificar la nota, ya sea que esta comprobación ocurra durante este proceso de evaluación o incluso después de terminado el semestre".

Mathias Badilla Salamanca

Estudiante de pregrado

Ingeniería civil industrial

Universidad de Concepción