En el siguiente informe se procederá a presentar un modelo de regresión el cual es **un**[**modelo matemático**](https://economipedia.com/definiciones/modelo-matematico.html)**que busca determinar la relación entre una variable dependiente (Y), con respecto a otras variables, llamadas explicativas o independientes (X).** Asimismo, el modelo busca determinar cuál será el impacto sobre la variable “Y” ante un cambio en las variables explicativas (X). Y asi poder determinar cuáles variables son capaces de explicar un fenómeno determinado, mostrando la predicción de la creciente popularidad de la moda de inspiración coreana que se ha apoderado del comercio minorista electrónico en Filipinas. Los emprendedores se están poniendo al día con esta tendencia creciente al vender sus productos inspirados en la moda coreana en diferentes plataformas en línea.

Introducción Experimentación Recomendaciones

Los modelos de regresión lineal trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variables independientes. En este tipo de modelos es importante testar la heterocedasticidad, la multicolinealidad y la especificación. En este informe se analiza un ejemplo utilizando el Software R, detallando la información entregada por el programa, cuál es su explicación y cómo se interpreta. Se explica cómo introducir variables categóricas en el modelo de regresión lineal múltiple mediante la carga de datos y la creación de variables. Finalmente, se explica la importancia de revisar los supuestos del modelo y asegurar un tamaño de muestra adecuado para que la estimación del modelo sea correcta, para esto se procederá hacer la limpieza de datos, representación gráfica de los datos, representación de correlaciones entre los datos, ejecución de la tarea de regresión, pruebas de significancia, bondad de ajuste del modelo, comprobación de las suposiciones del modelo y la Identificación de outliers y valores influyentes.

En esta sección vamos a analizar el rendimiento de datos reales que contiene nuestro archivo csv, con el fin de diseñar un modelo de regresión, interpretado en el lenguaje de programación R el cual es un software de referencia en el mundo de la Estadística.

La creciente popularidad de la moda de inspiración coreana se ha apoderado del comercio minorista electrónico en Filipinas. Los emprendedores se están poniendo al día con esta tendencia creciente al vender sus productos inspirados en la moda coreana en diferentes plataformas en línea. Por lo cual vamos a desarrollar un modelo de regresión.

¿Porque se usara el modelo de regresión?

Se utilizara regresión, dado que el tipo de variable que utilizaremos son variables cuantitativas continuas que corresponde a problemas de regresión, que en si tienen como resultado un número. Es decir, el resultado de la técnica de machine learning que estemos usando será un valor numérico, dentro de un conjunto infinito de posibles resultados, recordando que las variables cuantitativas continuas admiten cualquier valor, las discretas no.

En nuestro archivo .csv podemos observar diversas variables. Por lo cual de ese grupo hemos tomado las variables; ID de tienda, seguidor, tasa de respuesta, tiempo de respuesta, calificación mala, buena, normal y por calificación por estrella. Quedándonos de es esta forma:

¿Porque no utilizamos las demas variables?

No sé ha utilizado las demas variables, por ser de tipo String y en la mayoría de los casos tienen datos null

Implica almacenarlos de una manera consisistente que nos permita enfocarnos en responder preguntas de los datos en lugar de estar luchando con los datos. Entonces, datos limpios son datos que facilitan las tareas del análisis de datos:

La idea de las bases de datos limpias es hacer el análisis de datos más fácil, permitiendo enfocarnos en entender el problema y no la logística de la data.\\

Una base de datos limpia tiene las siguientes características: \\\\•Cada variable forma una columna.\\

•Cada observación forma una fila.\\

•Cada tipo de unidad observacional forma una tabla.\\

Para poder hacer esto en R primero cargamos nuestra base de datos con las siguientes líneas de programación\\

Una vez cargado los datos, procedemos a limpiarlo

La representación gráfica de datos estadísticos tiene como objetivo ofrecer una visión de conjunto del hecho sometido a investigación, de una manera más directa y perceptible que la mera presentación de los datos numéricos, para esto cargamos nuestras variables a predecir y procedemos a mandar a imprimir nuestras graficas de todas las variables con respecto Y

\begin{figure}[h]

\centering

\includegraphics[width=6cm, height=2cm]{figuras/limp.png}\\

\label{fig:gatos}

\end{figure}

A continuación se observara las gráficas de cada variable

La correlación es un tipo de asociación entre dos variables numéricas, específicamente evalúa la tendencia (creciente o decreciente) en los datos. Dos variables están asociadas cuando una variable nos da información acerca de la otra, para este caso almacenaremos nuestras anteriores variables en una sola matriz y la compararemos con el comando cor.

Primera regresión

Comprobaciones de la suposición del modelo

Podemos ver en la gráfica Q-Q que no cumple con la condición de normalidad por que los valores extremos no se ajustan a la distribución normal dada por la línea

Residuals vs Fitted

Nos permite ver que no existe linealidad, pues la línea de color rojo no tiene una tendencia lineal con respecto al eje horizontal

Scale-location

Presenta heterocedasticidad pues como se ve en la gráfica la línea roja de tendencia no es horizontal

Residual vs Leverage

Se puede observar que hay presencia de valores influyentes como el valor que se encuentra en la posición 27, supera el umbral porque tiene una distancia de cook mayor que 1

Segunda regresión

Grafica Q-Q

Podemos ver que de manera parecida la gráfica tiene una similitud con la primera grafica de Q-Q y de igual manera no cumple con la condición de normalidad por que los valores extremos

#no se ajustan a la distribución normal dada por la líneañ

Residuals vs Fitted

Nos permite ver que no existe linealidad, pues la línea de color rojo no tiene una tendencia lineal con respecto al eje horizontal

Scale-location

Presenta heterocedasticidad pues como se ve en la gráfica la línea roja de tendencia no es horizontal

Residual vs Leverage

Se puede observar que hay presencia de valores influyentes como el valor que se encuentra en la posición 27, supera el umbral porque tiene una distancia de cook mayor que 1

Tercera regresión

comprobaciones dela tercera suposición del modelo

Grafica Q-Q

Podemos ver que de manera parecida la gráfica tiene una similitud con la primera y segunda grafica de Q-Q y de igual manera no cumple con la condicionad de normalidad por que los valores extremos no se ajustan a la distribución normal dada por la línea

Residuals vs Fitted

Nos permite ver que no existe linealidad, pues la línea de color rojo no tiene una tendencia lineal con respecto al eje horizontal

Scale-location

Presenta heterocedasticidad pues como se ve en la gráfica la línea roja de tendencia no es horizontal

Residual vs Leverage

Se puede observar que hay presencia de valores influyentes como el valor que se encuentra en la posición 27, supera el umbral porque tiene una distancia de cook mayor que 1

Residual estandarizado [+2,-2]

posicion 1, 20, 30, 75, 138, 192, 193

hi = influence(reg)$hat

Verificamos si existe un valor que supere el valor de 0.9

Existe un valor influyente 27

cook = cooks.distance(reg)

cook1= which(cook>1)

rs1=which(rs >2,-2)

Segunda regresión

outlier y valores influyentes

rs2 = rstandard(regno)

Residual estandarizado [+2,-2]

Posición 1 20 21 47 72 75 138 193

hi2 = influence(regno)$hat

Verificamos si existe un valor que supere el valor de 0.9

No existen números que superen

cook2 = cooks.distance(regno)

cook3= which(cook2>1)

rs3=which(rs2>2,-2)

Tercera Regresión

outlier y valores influyentes

rs4 = rstandard(regsi)

Residual estandarizado [+2,-2]

Posición 1 20 21 47 75 136 193

hi4 = influence(regsi)$hat

Verificamos si existe un valor que supere el valor de 0.9

No existen que números que superen

cook5 = cooks.distance(regsi)

cook4= which(cook5>1)

rs4=which(rs4 >2,-2)

plot(cook2)