D208 TASK 1: LINEAR REGRESSION MODELING

Cali Ciccarelli

07/04/2024

Instructor: Dr Eric Straw

**PART 1:**

**Research Question:**

What causes a change in the monthly charge for customers?

**Goal of Analysis:**

The goal of this analysis is to find independent variables that have a strong correlation with the dependent variable, “MonthlyCharge”. Through the analysis, I plan on analyzing each variable to understand the distribution of all independent/dependent variables. I then hope to create a model that will show statistically significant correlations and let me know which independent variables I can focus on for the regressed model.

**PART 2:**

**Multiple Linear Regression Assumptions:**

* First, the dependent variable is expected to be a continuous, numerical variable. The monthly charge for services is a great example of a continuous numerical variable.
* Another assumption is there should be homoscedasticity which means that the data points surrounding the best line of fit should be within similar distances to the line (Fein, 2022) .
* The relationship between the dependent and independent variables should be linear. We can check this by creating scatterplots (Fein, 2022).
* There should be no multicollinearity. This means there should not be a strongly correlated relationship between the independent variables in the regression.
* “The residuals (errors) should be approximately normally distributed,” (Fein, 2022). A normal distribution is signified by a bell-shaped curve.

**Benefits of Using Python:**

Data analysis with Python is much more efficient. It allows for data to be cleaned on a wide scale by using a code applicable to the data set. It also allows for quick and accurate data analysis using the python packages that allow for statistical calculations and graph/model creations. The code can also be set to automate and perform analysis as updated data comes in.

**Technique:**

The purpose of multiple linear regression is to identify strongly correlated explanatory variables for a dependent variable. This makes the method a great choice to see what appears to influence the monthly charge customers are given. The regression will help to find which independent variables are the most effective at explaining the increase or decrease in the monthly charge. Specifically, the model should help predict how much each variable will increase or decrease the monthly charge for a customer. This will be shown in the coefficients of the model. The model will also explain how likely these predictions are to being correct by providing an R squared value. With this information, stakeholders will be able to make decisions to increase monthly charges and bring in more revenue.

**PART 3:**

1. **Data Cleaning:**

First, I will check for duplicated data. I will run the command, df\_clean.duplicated().any() and look for a return of “false”. This will confirm that there are no duplicated records. Next, I will review the quality of the data and look for null values and outliers for numerical data. For the null values, I will use the command “.isna” and “.sum” to locate the null values in the dataset. I then will fill N/As as appropriately needed. For the data, only InternetService had n/a values, however it is likely that n/a means internet service is not provided for the customer. For these n/a values, I filled it with a category named, “None”. For outliers, I will use the “stats.zscore” function. I will also use the “.abs” value command. Then I will look for values that had a z\_score greater than 3. I also changed the z score to 4 and then 5 so I could review the highest/lowest values for each column an ensure they were within reason. I will also check the categorical variables and ensure there are no typos/errors with the data by reviewing the unique values for each. I will do this by using the “.unique” command. Finally, I plan on dropping the columns that are not needed for my research question. I will drop 'Caseorder' ,'Customer\_id' ,'Interaction','UID' ,'City' ,'State' ,'County' ,'Zip' ,'Lat' ,'Lng' ,'Population' ,'Area' ,'TimeZone' ,'Job' ,'PaymentMethod’, and Items 1-8.

CSV file of the cleaned/transformed data is attached in a separate file.

1. **Summary Statistics:**

Below you will find the summary statistics output within my python code. A big key factor to note is ‘MonthyCharge’ has a mean of 172.62. We will be looking into how each of these variables appear to affect this mean. For categorical variables it lists the count of unique variables and the top variable within the category. For numerical variables, we get the mean, standard deviation, minimum value, maximum value, median and each quartile value.

Copy of output for statistics (The variable name appears below the stats):

count 10000.0000

mean 2.0877

std 2.1472

min 0.0000

25% 0.0000

50% 1.0000

75% 3.0000

max 10.0000

Name: Children, dtype: float64

count 10000.000000

mean 53.078400

std 20.698882

min 18.000000

25% 35.000000

50% 53.000000

75% 71.000000

max 89.000000

Name: Age, dtype: float64

count 10000

unique 5

top Divorced

freq 2092

Name: Marital, dtype: object

count 10000

unique 3

top Female

freq 5025

Name: Gender, dtype: object

count 10000

unique 2

top No

freq 7350

Name: Churn, dtype: object

count 10000.000000

mean 10.001848

std 2.976019

min 0.099747

25% 8.018214

50% 10.018560

75% 11.969485

max 21.207230

Name: Outage\_sec\_perweek, dtype: float64

count 10000.000000

mean 12.016000

std 3.025898

min 1.000000

25% 10.000000

50% 12.000000

75% 14.000000

max 23.000000

Name: Email, dtype: float64

count 10000.000000

mean 0.994200

std 0.988466

min 0.000000

25% 0.000000

50% 1.000000

75% 2.000000

max 7.000000

Name: Contacts, dtype: float64

count 10000.000000

mean 0.398000

std 0.635953

min 0.000000

25% 0.000000

50% 0.000000

75% 1.000000

max 6.000000

Name: Yearly\_equip\_failure, dtype: float64

count 10000

unique 3

top Month-to-month

freq 5456

Name: Contract, dtype: object

count 10000

unique 2

top No

freq 5166

Name: Port\_modem, dtype: object

count 10000

unique 2

top No

freq 7009

Name: Tablet, dtype: object

count 10000

unique 3

top Fiber Optic

freq 4408

Name: InternetService, dtype: object

count 10000

unique 2

top Yes

freq 9067

Name: Phone, dtype: object

count 10000

unique 2

top No

freq 5392

Name: Multiple, dtype: object

count 10000

unique 2

top No

freq 6424

Name: OnlineSecurity, dtype: object

count 10000

unique 2

top No

freq 5494

Name: OnlineBackup, dtype: object

count 10000

unique 2

top No

freq 5614

Name: DeviceProtection, dtype: object

count 10000

unique 2

top No

freq 6250

Name: TechSupport, dtype: object

count 10000

unique 2

top No

freq 5071

Name: StreamingTV, dtype: object

count 10000

unique 2

top No

freq 5110

Name: StreamingMovies, dtype: object

count 10000.000000

mean 34.526188

std 26.443063

min 1.000259

25% 7.917694

50% 35.430507

75% 61.479795

max 71.999280

Name: Tenure, dtype: float64

count 10000.000000

mean 3392.341550

std 2185.294852

min 155.506715

25% 1236.470827

50% 3279.536903

75% 5586.141370

max 7158.981530

Name: Bandwidth\_GB\_Year, dtype: float64

count 10000.000000

mean 172.624816

std 42.943094

min 79.978860

25% 139.979239

50% 167.484700

75% 200.734725

max 290.160419

Name: MonthlyCharge, dtype: float64

1. **Univariate and bivariate visualizations of the distributions of the dependent and independent variables:**

**Univariate numerical:**

**A group of blue and white graphs

Description automatically generated**

**A graph of a bar

Description automatically generated with medium confidence**A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a computer screen

Description automatically generated**A screenshot of a graph

Description automatically generated**

A screenshot of a computer screen

Description automatically generated**A screenshot of a graph

Description automatically generatedBivariate Visuals:**

**A graph of data on a white background

Description automatically generatedA screenshot of a computer program

Description automatically generatedA screenshot of a computer screen

Description automatically generated**

**Converting Data:**

In order to complete my analysis for my research question, I will need to convert all categorical data into measurable, numerical data. For the yes/no columns, I plan on replacing the no with 0 and yes with 1 to represent the lack of or presence of the variable. For the categorical variables that cannot be represented ordinally, I will be using one hot encoding. I will create dummy variables using the pd.get\_dummies() function in python (Bobbit, 2021). This will create binary measuring for each categorical variable. I will drop the first variable to avoid multicollinearity.

**Part 4:**

***Initial Linear Regression Model:***

A document with numbers and text

Description automatically generated

***Heat Maps to Assist in Model Reduction Determination:***

A screenshot of a graph

Description automatically generatedA graph with a red line

Description automatically generatedI decided to use heat maps to look for multicollinearity and to find the independent variables that were correlated the highest with the dependent variable, ‘MonthlyCharge’. Heat maps are helpful because they use color differences that are user friendly in determining where correlations are found. I first eliminated variables that appeared to have no effect on the ‘MonthlyCharge’. Then I reviewed the new heat map with the correlation values listed. From here I was able to determine the variable, ‘Churn’, needed to be removed since it was highly correlated with other variables. I decided to only include the variables that were correlated by over .30 in my new reduced model.

A screenshot of a graph

Description automatically generated

***Reduced Linear Regression Model:***

A screenshot of a computer

Description automatically generated

**Analyze and Compare the Models:**

The initial regression model is showing to technically fit the data set better than the regression model, however this is misleading. The R squared value is higher in the initial model compared to the reduced model because it is incorrectly inflated in the initial model due to high multicollinearity (Datatab, 2021). This means that some of the variables are highly correlated with each other causing an unstable model. This is something I had to look for when creating my heat map analysis. I ensured to rid of multicollinearity by removing the churn variable. This variable was shown in my heat maps to be highly correlated to other variables. The reduced model is now stable without multicollinearity and still has a high R squared value at .754.

***Residual plots:***

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated A graph of a number of graphs

Description automatically generated with medium confidence

(Bobbit, 2021)

* The reduced model’s residual standard error was about .43 for each independent variable.
* Executable Code is attached to the project in a separate file.

**Part 5**

**Regression Equation:**

Y = 111.15 + 32.87\*Multiple + 41.88\*StreamingTV + 52.52\*StreamingMovies

**Interpretation of the coefficients of the reduced model:**

* When it is true that a customer streams movies, the monthly charge increases by around 52.53.
* When it is true that a customer streams TV, the monthly charge is expected to increase by around 41.88.
* When it is true that a customer has multiple devices, the monthly charge is expected to increase by 32.87.

**Statistical and practical significance of the reduced model**

The reduced model is shown to be both practical and significant. ‘MonthlyCharge’ is highly dependent on the results of ‘StreamingTV’, ‘StreamingMovies’, and ‘Multiple’. This information can be used to predict the ‘MonthlyCharge’ with a high probability.

**Limitations of the data analysis:**

Limitations of my linear regression model include being unable to understand the effect ‘Churn’ has on ‘MonthlyCharge’ due to high multicollinearity. If I had left it in the model, I would have an unstable and unreliable model. Although without it, I am unable to predict the monthly charge based on the churn variable. If it was left in the model, the prediction would have been unreliable, since the model cannot calculate what change was caused by what variable.

**Recommend a course of action based on your results.**

My recommendation would be to advertise more for streaming movies and tv if the stakeholders are wanting overall higher monthly charges. They could offer more of a selection or even partner up with movie/tv streaming companies. They can also advertise for family plans for multiple lines since multiple lines is known to increase the monthly charge. I would say that the ideas are up to the stakeholders, however they should understand that when a customer streams tv/movies or has multiple lines, they will be receiving more money for the company.

**References**:

Bobbitt, Z. (2021, May 31). *How to create a residual plot in Python*. Statology. <https://www.statology.org/residual-plot-python/>

Bobbitt, Z. (2021, May 31). *How to Use Pandas Get Dummies – pd.get\_dummies*. Statology. https://www.statology.org/pandas-get-dummies/

DATAtab. (2021, February 10). *Multicollinearity (in Regression Analysis)* [Video]. YouTube. <https://www.youtube.com/watch?v=G1WX5GiFSWQ>

Fein, E. C., Gilmour, J., Machin, T., & Hendry, L. (2022, June 16). *Section 5.3: Multiple Regression Explanation, Assumptions, Interpretation, and write up*. Pressbooks. https://usq.pressbooks.pub/statisticsforresearchstudents/chapter/multiple-regression-assumptions/