**D208 Linear Regression Model**

**Nhi Le**

**Western Governors University**

**August 16, 2023**

**Part I: Research Question**

**A.1. Research question:**

How is the amount of data usage of customers, in GB, affected by some explanatory variables?

**A.2. Goals:**

From this analysis, the stakeholders can understand what factors affect the data used by customers. They will be able to get insights into the predictable amount of data usage of customers and create a data limit strategy to increase customer experience with our services. Moreover, the stakeholders can think of future marketing or sales strategies related to data usage.

**Part II: Method Justification**

**B.1. Summary of assumptions:**

The four assumptions of a multiple linear regression model (MLR):

* There is a linear relationship between an explanatory variable and a response variable.
* Among explanatory variables, they are not too highly correlated with one another.
* yi observations are selected independently and randomly from the population.
* Residuals normally need to be distributed with a mean of 0 (Hayes, 2023).

**B.2. Tool benefits:**

In this analysis, I will be using Python. Its benefits are: Firstly, Python has a lot of libraries and packages that can help me to create my predictive model with simple syntax. I can use Python from cleaning data using NumPy and Pandas libraries, to visualizing data using Matplotlib and Seaborn, to machine learning with Scikit-learn (Pruciak, n.d.). Secondly, Python has been around for a long time and has a global IT community. I can find support and solutions quickly for my codes if I need it, which helps me to complete this analysis effectively and accurately.

**B.3. Appropriate technique:**

Multiple linear regression (MLR) is an appropriate technique because the research question is for data usage by customers, in GB. Bandwidth\_in\_GB is a continuous variable, which can be used in MLR. There are many explanatory variables in the dataset that can help to predict the data usage by customers, for example: children, age, and income. By using MLR, we can find the relationship between these explanatory variables and bandwidth\_in\_GB variable by adding or removing them from the regression equation. MLR is convenient in this predictive model using multiple independent variables, instead of just a single linear regression. In summary, using MLR will help to answer the research question and help the executive leaders to decide what areas affect the data usage by customers and focus on them, having strategies to improve our services.

**Part III: Data Preparation**

**C.1. Data cleaning:**

My data cleaning goals are to detect missing data, duplicate data, and outliers, then decide to treat them with appropriate methods:

* Import dataset churn\_clean.csv into Jupyter Notebook.
* Get information (column names, data types), and statistical details (count, min, max, mean, std, percentile) of the dataset.
* Find missing data and impute missing data with meaningful measures of central tendency (mean, median, or mode).
* Find outliers and treat them by removing them, retaining them, excluding them, or imputing them with the median.
* Rename the Item1 – Item8 columns to easily recognized names (For example: ‘Item1’ renamed to ‘TimelyResponse’)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

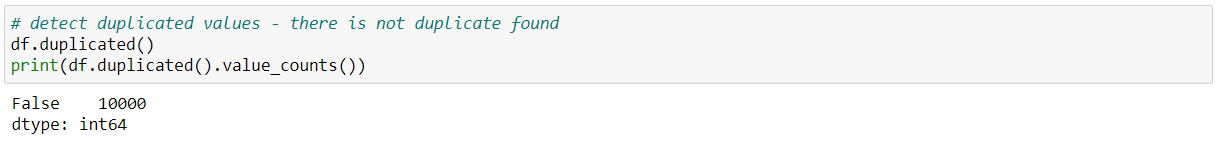
Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



A white background with black text

Description automatically generated

A white background with black dots

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

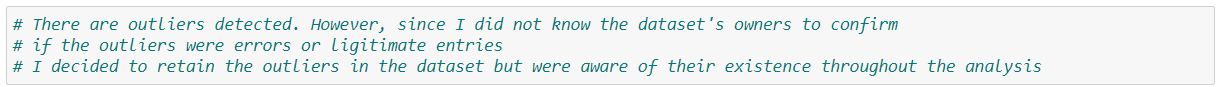
Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated



A white screen with blue text

Description automatically generated

**C.2. Summary statistics:**

The dependent variable used to answer the research question was ‘Bandwidth\_GB\_Year’ (the average yearly amount of data used, in GB, per customer). This is a continuous variable. This variable has 10,000 values and the mean of average of data usage of a year was 3392.34 GB.

The dependent variables used to answer the research question all have 10,000 records:

* The 9 continuous variables including:
* ‘Children’, ‘Age’, ‘Income’ are demographic variables on billing statement for each customer. The average number of children that customers had was 2. The age range of customers was 18 to 89. The average income was 39806.93.
* ‘Email’ is numeral variable to record the number of marketing or correspondence emails sent. On average, customers received 12 emails, and the maximum emails customers got was 23. ‘Contacts’ is numeral variable for how many times customer contacted technical support. Customers contacted support 1 time on average.
* ‘Outage\_sec\_perweek’ shows system outages in the customer’s neighborhood’s average of seconds per week. The longest time that the customers experienced system outages was 21 seconds per week.
* ‘Tenure’ is numerical variable to record how many months the customer has been with the provider. The customers have used our services from 1 to 72 months. ‘MonthlyCharge’ is the monthly charge for the customer. Customers paid us 172.62 per month. ‘Yearly\_equip\_failure’ is numeral variable to show the number of time customer’s equipment failed and needed to reset or replaced last year. This did not happen a lot, but some customers had to experience this up to 6 times.
* The 24 categorical variables including:
* 8 categorical variables reflect customer’s satisfaction ratings on a scale of 1 to 8 (1 = most important, 8 = least important): ‘item1’ – Timely response, ‘item2’ – Timely fixes, ‘item3’ – Timely replacements, ‘item4’ – Reliability, ‘item5’ – Options, ‘item6’ – Respectful response, ‘item7’ – Courteous exchange, ‘Item8’ – Evidence of active listening. The values are from 1 to 8. The mean of each variable is around 3.50.
* ‘Gender’ is categorical variable to reflect the gender of customer. There are 3 unique values for gender.
* ‘Churn’ is categorical variable as the values are ‘Yes’ or ‘No’. It recorded whether the customers canceled service last month. There were 7350 records for ‘No’. ‘Techie’ has Yes/No value. This categorical variable reflects if the customer thinks that they are good at technology. ‘No’ was the top value.
* ‘Contract’ is categorical variable on what kind of contract customer has ‘Month-to-month’, ‘One Year’, or ‘Two Year’. 5456 customers had Month-to-month contracts.
* ‘Tablet’ is categorical variable answering if the customer has a tablet. ‘Port\_modem’ is categorical variable answering if the customer has a portable modem. The values are ‘Yes’ or ‘No’. The top values for these 2 variables were ‘No’.
* ‘InternetService’ shows customer’s internet service provider. There were 3 unique values.
* ‘Phone’, ‘Multiple’, ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘TechSupport’, ‘StreamingTV’, ‘StreamingMovies’, ‘PaperlessBilling’ are services that the company provides. The values of these variables are ‘Yes’ or ‘No’ to reflect if the customer signed up for. Top values for ‘Phone’ and ‘PaperlessBilling’ were ‘Yes’, the rest had the top values as ‘No’.

Here are the screenshots of the summary statistics output of the variables (including: count, mean, std, min, 25%, 50%, 75%, max):

A white screen with blue text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A table with numbers and text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a screen

Description automatically generated

**C.3. Visualizations:**

Univariate visualizations:

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph with lines and dots

Description automatically generated

A screenshot of a graph

Description automatically generated

Bivariate visualizations:

A screen shot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

C.4. Data transformation:

My data transformation goals are to make sure I can use all the independent variables for the analysis. To make that happen, I will need to encode categorical values into numerical values since I can only use numerical values in multiple linear regression.

* Drop variables that will not be needed for the analysis.
* Create dummy variables for categorical variables.
* Encode categorical values to numerical values: For those variables with Yes/No values, the dummy value is 1 for Yes and 0 for No. For the Gender variable, it has Male, Female, and Nonbinary. The DummyFemale is 1 when Gender is Female and else it is 0. Contract has 3 values: Month-to-month, One Year, and Two year. DummyMonthtoMonth is 1 when Contract is Month-to-month, else it is 0. InternetService has 3 values: Fiber Optic, DSL, and None. DummyFiberOptic is 1 when InternetsService is Fiber Optic, else it is 0.
* Spot-check the statistical details of the dataset to make sure categorical values are encoded correctly.
* Extract the prepared dataset as CSV file named ‘churn\_prepared’.

A screenshot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated



**C.5. Prepared data set:**

The prepared data set will be submitted as ‘churn\_prepared.csv’ along with this doc file.

**Part IV: Model Comparison and Analysis**

**D.1. Initial model:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bandwidth\_GB\_Year = -284.95 + 30.39\*Children + -3.32\*Age + 0.00\*Income + -0.35\*Outage\_sec\_perweek + 0.07\*Email + 2.03\*Contacts + 1.17\*Yearly\_equip\_failure + 82.11\*Tenure + 9.28\*MonthlyCharge + -0.12\*TimelyResponse + 0.49\*TimelyFixes + -2.35\*TimelyReplacements + -0.54\*Reliability + 0.85\*Options + 1.23\*Respectfulness + 0.21\*Courteous + 2.75\*Listening + -63.79\*DummyFemale + 12.99\*DummyChurn + -1.81\*DummyTechie + 0.76\*DummyPort\_modem + 0.50\*DummyTablet + 0.72\*DummyPhone + -227.31\*DummyMultiple + 54.23\*DummyOnlineSecurity + -116.52\*DummyOnlineBackup + -32.57\*DummyDeviceProtection + -110.47\*DummyTechSupport + -166.63\*DummyStreamingTV + -2.58\*DummyPaperlessBilling + -2.06\*DummyMonthtoMonth + -485.50\*DummyFiberOptic + -278.15\*DummyStreamingMovies

**D.1. Justification of model reduction:**

The feature selection method I used to create the reduced model was backward stepwise elimination. I removed the least significant independent variables that had a p-value greater than 0.05. Then, I needed to verify if my model met all the 4 assumptions. Firstly, I used heat maps to see the relationships between the independent variables and the dependent variables. I only kept those independent variables that had correlations with the dependent variable more than 0.05 or less than -0.05. Secondly, I checked multicollinearity by calculating VIF (Variance Inflation Factor) and kept those variables having VIF less than 5. Thirdly, I created residual plots to verify independence of observations, homoscedasticity, and normality of residuals.

**D.3. Reduced linear regression model:**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Bandwidth\_GB\_Year = -15.56 + 15.05\*DummyChurn + 82.05\*Tenure + 4.06\*MonthlyCharge + 53.68\*DummyStreamingTV + -355.62\*DummyFiberOptic

**E.1. Model comparison:**

The model evaluation metric I used to compare the initial model and reduced model was R-squared/ Adjusted R-squared. The R-squared/ Adjusted R-squared for the initial model is 0.997. However, the condition number is large which might suggest strong multicollinearity. The R-squared/ Adjusted R-squared for the reduced model is 0.994. The difference between the two models is insignificant. Although the R-squared/ Adjusted R-squared shows that the initial model is better than the reduced model, the reduced model met all 4 assumptions of multiple linear regression. By removing all those other 28 predictor variables, our model still explains 99.40% of the variance, which is very impressive. Therefore, the reduced model is a good fit to explain the dependent variable.

**E.2. Output and calculations:**

Through the residual plots, the points are plotted randomly spread around the 0 line. The observations are independent and random. The spreads of residuals are not based on one side so the reduced model does not violate heteroscedasticity. The model’s residual standard error was calculated. In the real world, it is rare to acquire a perfect model. Therefore, the result is acceptable.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A white rectangular sign with black text

Description automatically generated

**E.3. Code:**

I attached a Jupiter Notebook file for my code as ‘D208(1).ipynb’

**Part V: Data Summary and Implications**

**F.1. Results:**

The final multiple linear regression equation including 5 independent variables:

Bandwidth\_GB\_Year = -15.56 + 15.05\*DummyChurn + 82.05\*Tenure + 4.06\*MonthlyCharge + 53.68\*DummyStreamingTV + -355.62\*DummyFiberOptic

The coefficients suggest that for every 1 unit of:

* DummyChurn (Customers that left our services): Bandwidth\_GB\_Year will increase 15.05 units.
* Tenure (The amounts of months that the customers stay with our services): Bandwidth\_GB\_Year will increase 82.05 units.
* MonthlyCharge: Bandwidth\_GB\_Year will increase 4.06 units.
* DummyStreamingTV (Customers had Streaming TV): Bandwidth\_GB\_Year will increase 53.68 units.
* DummyFiberOptic (Customers had Internet Service with Fiber Optic): Bandwidth\_GB\_Year will decrease 355.62 units.

P-values for Tenure, MonthlyCharge, DummyStreamingTV, and DummyFiberOptic are at 0.000. The P-value for DummyChurn is 0.003. The p-values are very low, which proves that the reduced model is statistically significant. The model is also practically significant. Through the heat map, I know that there is a linear relationship between each of these 5 independent variables and the dependent variable. Through the reduced model, I can learn how each of these independent variables impact the amount of bandwidth usage per year. Then I can improve our services and reduce churn.

There are 2 limitations of the analysis: Firstly, the linear regression model can be the best fit for the dataset that has linear relationships between the independent variables and dependent variable when the underlying conditions are met. However, we might have missed some independent variables that do not have linear relationships with the dependent but have other meaningful impacts to the dependent variable. Secondly, this data set is small and only contains data from the last year. If we had the data in the time series for at least 3 years, we could track and find more accurate trends.

**F.2. Recommendations:**

From the analysis, I see that churn, monthly charge, tenure, Streaming TV service, and Fiber Optic Internet Service affect the customer’s average data usage, in GB, in a year. I see that when the customers stay with us longer, they will have a higher amount of data usage. To use Streaming TV, customers need to use a lot of data. In summary, to reduce retention, we need to increase our data providing capacity to make sure our loyal customers have great experiences with us. We can use the reduced model to predict the bandwidth customers need and provide them with data limits equal or higher than the predicted amounts. Sales and marketing can assure customers that the longer they stay with us, the higher the data limit we will give them. For customers with Streaming TV, we can give them higher data limits since the service needs to use more data.

**Part VI: Demonstration**

**G. Panopto demonstration:**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=be57ff38-d370-4b3a-98ac-b06001502b49>

**H. Sources of third-party code:**

Data to fish. Data to Fish. (n.d.). https://datatofish.com/multiple-linear-regression-python/

Zach. (2022, October 12). How to calculate VIF in python. Statology. https://www.statology.org/how-to-calculate-vif-in-python/

GeeksforGeeks. (2022, February 21). How to create a residual plot in Python. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/

1. **Sources:**

Hayes, A. (2023, May 9). *Multiple linear regression (MLR) definition, formula, and example*. Investopedia. https://www.investopedia.com/terms/m/mlr.asp

Pruciak, M. (n.d.). *Why is python a great choice for data analysis?*. Agile Software Development Agency in Europe. https://www.ideamotive.co/blog/python-for-data-analysis