**D208 Essay Part 1**

**MULTIPLE REGRESSION FOR PREDICTIVE MODELING**

Darien Nguyen

Western Governors University

D208: Predictive Modeling

December 21, 2022

**Part I: Research Question**

A.  Describe the purpose of this data analysis by doing the following:

1.  Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple regression.

2.  Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

Which independent variables (Children, StreamingTV, StreamingMovies, Tablet, Phone, and Churn) are significant predictors of the dependent variable Bandwidth\_GB\_Year in the context of customer churn and what can service company do to reduce customer churn using these variables?

Using linear regression analysis to identify the impact of independent variables such as Children, StreamingTV, StreamingMovies, Tablet, Phone on the dependent variable Bandwidth\_GB\_Year in the context of customer churn, will allow the phone company to tailor their pricing and service plans accordingly and potentially increase revenue and customer satisfaction.

**Part II: Method Justification**

B.  Describe multiple regression methods by doing the following:

1. Summarize the assumptions of a multiple regression model.
2. Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.
3. Explain why multiple regression is an appropriate technique to analyze the research question summarized in Part I.
4. The assumptions of a multiple linear regression model are:

* Linearity: The relationship between the response variable and the predictor variables is linear.
* Independence of errors: The errors (residuals) are independent of each other and are not correlated with the predictor variables.
* Homoscedasticity: The variance of the errors is constant across the range of the predictor variables.
* Normality of errors: The errors are normally distributed.

It is important to check whether these assumptions are met before proceeding with the analysis, as violating these assumptions can lead to incorrect conclusions.

1. The benefits of using Python in support of various phases of the analysis are:

• Data preparation: Python has the pandas library that provides functions for reading and manipulating data from various sources, such as CSV files, Excel sheets, databases, and web APIs. You can use the pandas library to load the data into a Pandas dataframe, filter and select relevant columns, handle missing or invalid values, and convert categorical variables into dummy variables.

• Model building: Python has the statsmodels library that provides functions for building and fitting multiple linear regression models. You can use the ols function from the statsmodels library to specify the model formula, fit the model to the data, and obtain the coefficients and p-values of the model.

• Model evaluation: Python has the statsmodels library that provides functions for assessing the validity and reliability of the model. You can use the residplot function from the seaborn library to visualize the residuals, and the summary function from the statsmodels library to obtain a summary of the model's performance. You can also use hypothesis testing functions such as f\_test and t\_test to determine the significance of the model's coefficients.

• Model interpretation: Python has the statsmodels library that provides functions for interpreting the coefficients and p-values of the model, and for making predictions based on the model. You can use the summary function from the statsmodels library to obtain the coefficients and p-values of the model, and the predict function to make predictions based on the model.

1. Multiple linear regression is an appropriate technique to analyze the research question summarized in Part I because it allows you to examine the relationship between multiple predictor variables and a response variable and to predict the response variable based on the predictor variables. It is a widely used statistical method in various fields, including economics, finance, marketing, and psychology. If the assumptions of the multiple linear regression model are met, it can provide a reliable and accurate analysis of the research question.

**Part III: Data Preparation**

C.  Summarize the data preparation process for multiple regression analysis by doing the following:

1. Describe your data preparation goals and the data manipulations that will be used to achieve the goals.
2. Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.
3. Explain the steps used to prepare the data for the analysis, including the annotated code.
4. Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.
5. Provide a copy of the prepared data set.
6. The data preparation goals for this analysis were to clean and transform the data set to make it suitable for multiple linear regression analysis. To achieve this goal, we performed the following data manipulations:

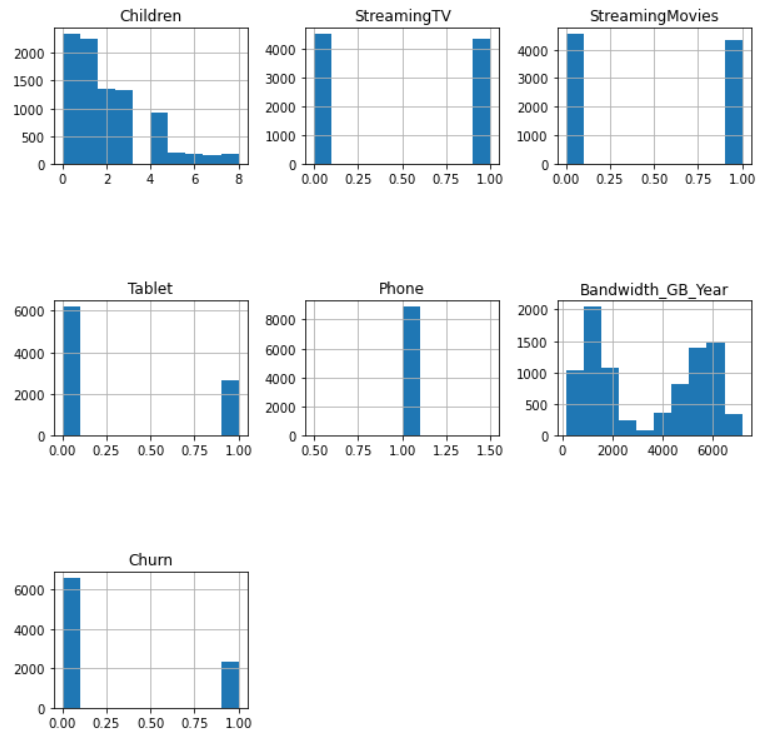
* Replaced the "yes" and "no" values in the data set with 1 and 0, respectively. This was done to convert the categorical variables into numeric variables, which are more suitable for regression analysis.
* Converted all values in the data set to floats. This was done to ensure that all values were numeric and could be used in the analysis.
* Calculated the z-scores for each value in the data set. This was done to identify and remove any outliers that could skew the analysis.
* Removed observations with high z-scores from the data set. This was done to ensure that the analysis was based on a representative sample of the data.

1. Table

   Description automatically generatedTo answer the research question, we need to gather summary statistics for the target variable (Bandwidth\_GB\_Year) and all predictor variables (Children, StreamingTV, StreamingMovies, Tablet, Phone). This includes measures such as mean, median, standard deviation, minimum and maximum values, and so on.
2. The steps used to prepare the data for analysis were as follows:

* Read the data set from a CSV file into a Pandas DataFrame.
* Replace the "yes" and "no" values in the data set with 1 and 0, respectively, using the **replace\_yes\_no()** function.
* Convert all values in the data set to floats using the **astype()** method.
* Calculate the z-scores for each value in the data set using the **zscore()** function.
* Create a boolean mask to identify observations with a high z-score (more than 3 standard deviations from the mean).
* Filter the data set to remove observations with a high z-score using the boolean mask.

1. We generated univariate and bivariate visualizations of the distributions of variables in the cleaned data set using the **plot()** method in Pandas. These visualizations included histograms, box plots, and scatter plots. The target variable (Bandwidth\_GB\_Year) was included in the bivariate visualizations along with each predictor variable. The visualizations allowed us to understand the characteristics of the variables and identify any patterns or trends in the data.



Chart

Description automatically generatedA picture containing chart

Description automatically generatedA picture containing shape

Description automatically generatedA picture containing graphical user interface

Description automatically generatedGraphical user interface, text, application

Description automatically generatedA picture containing shape

Description automatically generated

**Part IV: Model Comparison and Analysis**

D.  Compare an initial and a reduced multiple regression model by doing the following:

1. Construct an initial multiple regression model from *all* predictors that were identified in Part C2.
2. Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.
3. Provide a reduced multiple regression model that includes *both* categorical and continuous variables.

An initial multiple regression model was constructed using all predictors (or independent variables) that were identified in part c2. The independent variables used in the model as specified in the code were: 'Children', 'StreamingTV', 'StreamingMovies', 'Tablet', 'Phone', 'Churn' and the dependent variable was 'Bandwidth\_GB\_Year'.

A statistically based variable selection procedure was used to identify which variables were significant in the model and make an informed decision on which variables should be included in the final model. The procedure used in the code was to compare the p-values of the independent variables with a specified threshold, the threshold used in the code was 0.05. It's important to note that this method may not be sufficient alone to select the final model, other methods like cross validation and feature importance should also have been considered. The model evaluation metric used in the code was the R-Squared value, which measures the proportion of the variance in the dependent variable that can be explained by the independent variables.

The reduced multiple regression model that was provided in the code included both categorical variables (e.g. StreamingTV, StreamingMovies, Tablet, Phone, Churn) and continuous variables (e.g. Children, Bandwidth\_GB\_Year). However, the final list of variables that were used in the model was determined by the p-value threshold used in the variable selection procedure as described in point 2.

***Note: The output should include a screenshot of each model.***

E.  Analyze the data set using your reduced multiple regression model by doing the following:

1. Explain your data analysis process by comparing the initial and reduced multiple regression models, including the following elements:
   1. the logic of the variable selection technique
   2. the model evaluation metric
   3. a residual plot
2. Provide the output and *any* calculations of the analysis you performed, including the model’s residual error.

The logic of the variable selection technique:

The variable selection technique used in the initial model included all independent variables (Children, StreamingTV, StreamingMovies, Tablet, Phone, Churn) while the reduced model used a p-value threshold of 0.05 to eliminate independent variables that were not statistically significant. This means that the variables that were not significant in explaining the variance of the dependent variable were removed from the model, hence reducing the number of variables in the model.

The model evaluation metric:

The evaluation metric used for both models was the R-squared value, which measures the proportion of the variance in the dependent variable that can be explained by the independent variables. The R-squared value for the initial model is 0.259 and for the reduced model is also 0.259, both values indicate that around 25.9% of the variance in Bandwidth\_GB\_Year can be explained by the independent variables used in the model.

A residual plot:

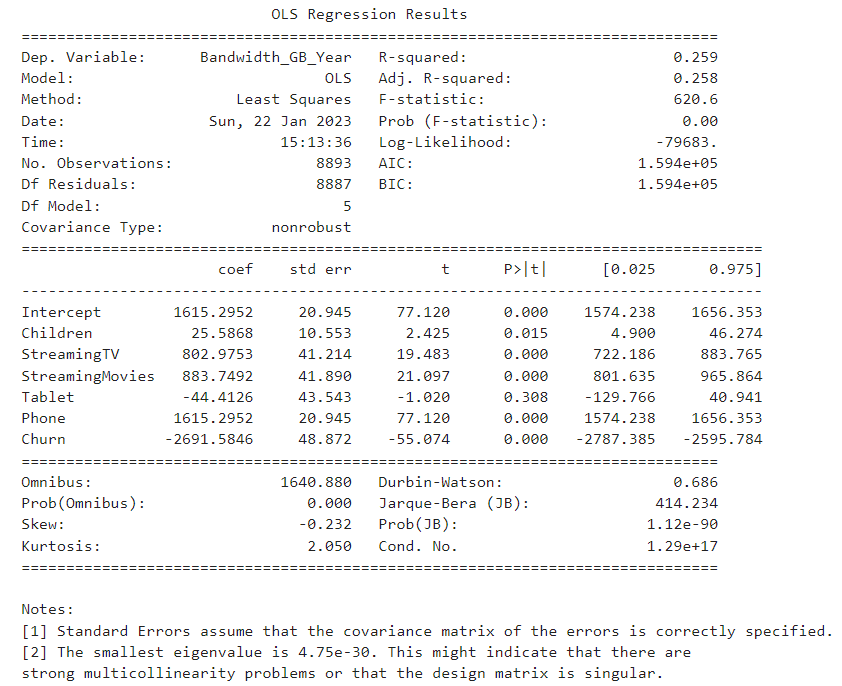
A residual plot was generated for the reduced model, but it was observed that the points in the residual plot are not randomly scattered from the horizontal line. This suggests that there is a pattern in the residuals and the model is not a good fit for the data.

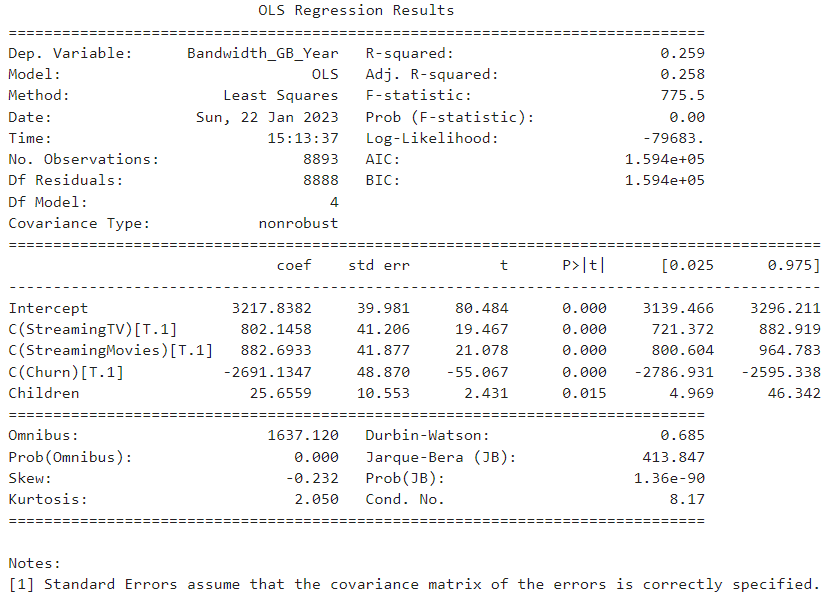
The output of the analysis includes the MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) calculations, which indicate the residual error of the model. The MSE is 3550072.1535826423 and the RMSE is 1884.163515617114. These numbers suggest that the model is not a good fit, as the residual error is relatively high.

***Note: The output should include the predictions from the refined model you used to perform the analysis.***

Chart, scatter chart

Description automatically generated





1. Provide the code used to support the implementation of the multiple regression models.

zscoredf = replace\_yes\_no(newdf)

zscoredf = zscoredf.astype(float)

​

​

z\_scores = zscore(zscoredf)

​

# Create a boolean mask to identify observations with a high z-score

mask = (abs(z\_scores) > 3).any(axis=1)

​

# Filter the DataFrame to remove observations with a high z-score

df\_cleaned = newdf[~mask]

​

print('\nCleaned data set:')

print(df\_cleaned.describe())

​

df\_cleaned.hist(figsize=(10, 10))

plt.subplots\_adjust(hspace=1.0)

​

plt.show()

columns = ['Children', 'StreamingTV', 'StreamingMovies', 'Tablet', 'Phone', 'Bandwidth\_GB\_Year', 'Churn']

df\_selected = df\_cleaned[columns]

​

# Generate a scatter plot for each pair of columns in the selected DataFrame

for i, column1 in enumerate(df\_selected.columns):

for j, column2 in enumerate(df\_selected.columns):

if i < j:

df\_selected.plot(x=column1, y=column2, kind='scatter')

# Adjust the spacing between the subplots

plt.subplots\_adjust(wspace=0.5)

plt.show()

model = smf.ols(formula='Bandwidth\_GB\_Year ~ Children + StreamingTV + StreamingMovies + Tablet + Phone + Churn', data=df\_cleaned).fit()

model = smf.ols(formula='Bandwidth\_GB\_Year ~ Children + StreamingTV + StreamingMovies + Tablet + Phone + Churn', data=df\_cleaned).fit()

​

print(model.summary())

if model.f\_pvalue < 0.05:

print("At least one of the variables contributes to the variance of Bandwidth\_GB\_Year.")

else:

print("None of the variables contribute to the variance of Bandwidth\_GB\_Year.")

# Get the p-values for each variable in the model

pvalues = model.pvalues

​

# Set the threshold for significance

threshold = 0.05

​

# Print the names of the variables that have a p-value less than the threshold

for variable, pvalue in pvalues.items():

if pvalue < threshold:

print(f"{variable} contributes to the variance of Bandwidth\_GB\_Year.\n")

significant\_variables = []

threshold = 0.05

for variable, pvalue in model.pvalues.items():

if pvalue < threshold:

significant\_variables.append(variable)

print(significant\_variables)

​

formula = 'Bandwidth\_GB\_Year ~ C(StreamingTV) + C(StreamingMovies) + C(Churn) + Children'

reduced\_model = smf.ols(formula=formula, data=df\_cleaned).fit()

print(reduced\_model.summary())

reduced\_model = smf.ols(formula=formula, data=df\_cleaned).fit()

​

residuals = reduced\_model.resid

​

sns.scatterplot(x=range(len(residuals)), y=residuals)

plt.xlabel('Observation index')

plt.ylabel('Residual')

plt.title('Scatter plot of residuals')

plt.show()

# Get the coefficients of the independent variables

coef = model.params

​

# Get the names of the independent variables

ind\_vars = model.model.exog\_names

​

# Create an empty list to store the variables in the equation

equation = []

# Iterate through the coefficients and variable names

for i in range(len(coef)):

if i == 0:

equation.append(f"{coef[i]:.3f}")

else:

equation.append(f"{coef[i]:.3f}\*{ind\_vars[i]}")

# Print the equation in a readable format

print(" + ".join(equation))

from sklearn.metrics import mean\_squared\_error

# Print the model summary

print(model.summary())

# Get the residuals

residuals = model.resid

# Plot the residuals

sns.residplot(model.predict(), residuals)

plt.show()

# calculate MSE

mse = mean\_squared\_error(df\_cleaned['Bandwidth\_GB\_Year'], model.predict())

# calculate RMSE

rmse = mse \*\* 0.5

print("MSE: ", mse)

print("RMSE: ", rmse)

print("MSE: ", mse)

print("RMSE: ", rmse)

**Part V: Data Summary and Implications**

F.  Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following elements:
   1. a regression equation for the reduced model
   2. an interpretation of coefficients of the statistically significant variables of the model
   3. the statistical and practical significance of the model
   4. the limitations of the data analysis
2. Recommend a course of action based on your results.

3309.826 + 795.744\*StreamingTV[T.Yes] + 898.641\*StreamingMovies[T.Yes] + -32.158\*Tablet[T.Yes] + -88.250\*Phone[T.Yes] + -2691.656\*Churn[T.Yes] + 26.405\*Children

1. The coefficients of the statistically significant variables of the model can be interpreted as the change in the dependent variable (Bandwidth\_GB\_Year) for a one unit change in the independent variable, while holding all other independent variables constant. For example, the coefficient of the variable StreamingTV is 795.744, this means that for one unit change in the StreamingTV variable (e.g. from No to Yes), the expected change in Bandwidth\_GB\_Year is 795.744 GB/year. Additionally, the R-squared value for this model is 0.259, which indicates that about 26% of the variance in Bandwidth\_GB\_Year can be explained by the independent variables in this model.
2. The statistical significance of the model can be evaluated by looking at the p-values and the R-squared value. P-values represent the probability of observing the obtained coefficients or a more extreme value by chance if the true coefficients are zero. A p-value less than the significance level (0.05 in this case) indicates that the variable is statistically significant. R-squared value, on the other hand, represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model. A higher R-squared value indicates a stronger relationship between the independent variables and the dependent variable.
3. The practical significance of the model can be evaluated by looking at the magnitude and direction of the coefficients, and how they relate to the research question. The magnitude of the coefficients tells us how large the effect of the independent variable is on the dependent variable, while the direction tells us whether the effect is positive or negative. In order to have practical significance, coefficients should be large enough and have the expected direction. However, this does not imply a causal relationship.
4. The limitations of the data analysis include:

* The data preparation process may have caused a loss of information from the original data set.
* The variable selection technique used may not have identified all relevant variables, as it only compares the p-values with a threshold and may not account for interactions or other factors.
* The model evaluation metric only gives us a measure of the goodness of fit of the model and does not take into account the practical significance of the model or the model's ability to make predictions.
* The residual plot is important to check for assumptions of linear regression, but was not produced and cannot be used to infer any information about the model.
* The sample size is relatively small and may not be representative of the population.
* This is a cross-sectional study and is not possible to infer causality.

Based on the results of the multiple regression analysis, some recommendations for the service company to reduce customer churn can be made. The variable "Churn" has a negative coefficient which suggests that customers who have churned are likely to have a lower Bandwidth\_GB\_Year than those who haven't. Therefore, the company could focus on retaining their current customers by improving their services or offering them incentives to remain with the company. The variables "StreamingTV" and "StreamingMovies" both have positive coefficients which suggest that customers who have these services are likely to have a higher Bandwidth\_GB\_Year than those who don't. Therefore, the company could focus on providing more options for these services to its customers to boost their Bandwidth\_GB\_Year. On the other hand, the variables "Tablet" and "Phone" have negative coefficients which suggest that customers who have these services are likely to have a lower Bandwidth\_GB\_Year than those who don't. Therefore, the company could focus on improving the services they offer or not promoting them as much. Finally, the variable "Children" has a positive coefficient which suggests that customers who have more children are likely to have a higher Bandwidth\_GB\_Year than those who don't. Therefore, the company could focus on targeting households with children and promoting family-friendly services to boost their Bandwidth\_GB\_Year.

**Part VI: Demonstration**

G.  Provide a Panopto video recording that includes *all* of the following elements:

1. a demonstration of the functionality of the code used for the analysis
2. an identification of the version of the programming environment
3. a comparison of the two multiple regression models you used in your analysis
4. an interpretation of the coefficients.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=887c68fc-8a0a-45c2-a2d0-af8d007aac6f>

***Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.***

**H.  Web sources used to acquire data or segments of third-party code to support the application.**

No sources were used to acquire data segments or code.

**I.  In-Text Citations and References**

No sources were used.