**D208 PA**

Jon Isak

D208: Predictive Modeling

Eric Straw

Contents

[Part I: Research Question 3](#_Toc152037714)

[Part II: Method Justification 3](#_Toc152037715)

[Part III: Data Preparation 4](#_Toc152037716)

[Part IV: Model Comparison and Analysis 6](#_Toc152037717)

[Part V: Data Summary and Implications 11](#_Toc152037718)

[Part VI: Demonstration 12](#_Toc152037719)

**D208 PA**

## Part I: Research Question

A1. I will use multiple linear regression modeling to research the question, what are the factors contributing to higher monthly charges for customers? This question aims to explore the relationship between different factors and monthly charge. Identifying the relationship could allow the company to gain insights into what drives higher costs. These insights can be used for pricing strategies to optimize revenue.

A2. The goal of the data analysis is to uncover and understand the relationships between variables and monthly charges. Another goal is to identify significant predictors to create a model.

## Part II: Method Justification

B1. A multiple linear regression model relies on four conditions. The model’s residuals are close to normal. The residuals have similar variability. The residuals are independent, and each variable has a linear relationship with the outcome (Libretexts, 2023).

B2. I will conduct the analysis using Python within a Jupyter notebook environment. Python serves as an excellent choice due to its rich resources, including extensive documentation, tutorials, and community support, which are useful for each phase of the analysis. Moreover, Python has numerous powerful libraries like Pandas, NumPy, Matplotlib, Seaborn, sklearn, and Statsmodels, enhancing its suitability for a wide spectrum of data analysis tasks. Pandas offers efficient data manipulation capabilities, allowing seamless handling and transformation of datasets. NumPy provides essential numerical operations and efficient array handling, enabling quick and optimized mathematical computations. Additionally, Matplotlib and Seaborn facilitate the creation of insightful visualizations. Sklearn provides a wide array of machine learning algorithms and tools for data mining and analysis. Statsmodels offers statistical models and tests for exploring data. The usefulness of Python libraries makes it an ideal choice for this analysis, offering the necessary tools to perform data processing, statistical modeling, visualization, and result interpretation effectively. (Terra, 2019).

B3. Multiple linear regression is a suitable technique for analyzing the factors influencing monthly charges due to its capability to handle multiple predictors simultaneously. In this context, the research aims to estimate the relationship between the dependent variable, monthly charges, and several independent variables that potentially impact these charges. This method accommodates the inclusion of numerous predictors, allowing the assessment of how each independent variable contributes to the variability observed in the monthly charges. Furthermore, it enables the identification of which factors, among the multiple variables considered, hold stronger associations with the dependent variable. By leveraging this technique, a comprehensive understanding of how various factors collectively affect monthly charges can be attained, facilitating more informed decision-making processes within the studied domain.

## Part III: Data Preparation

The provided churn dataset is assumed to have been cleaned according to the file description. An assessment will be conducted to verify its cleanliness by checking for missing values and data types. The goals of this verification process are to detect any missing entries and ensure consistency with the data dictionary. No missing values were found, and the data types across columns were consistent with the expected values outlined in the data dictionary. This indicates that the dataset does not require additional cleaning and is suitable for subsequent analysis. The verification code is included with the file.

C2. The dependent variable, 'MonthlyCharge', had a mean of $172.62 with a standard deviation of $42.94. The minimum was $79.98, and the maximum was $290.16. The median was $167.48. Regarding the independent variables, 'Age' spans from 18 to 89 years, with an average age of approximately 53.08 years and a standard deviation of 20.70. 'Children' shows a mean of around 2.09 children per household, with a maximum of 10 children. 'Income' had a mean income of $39,806 and ranged from $348.67 to $258,900.70. The 'Tenure' variable had a range from 1 to 72 months, with an average of 34.53 months. 'OnlineSecurity' ranged from 0 to 1 for no and yes. A mean of .3576 indicates that 35.76% of customers signed up for online security. For ‘InternetService’, 4408 had fiber, 3463 had DSL, and 2129 had none.

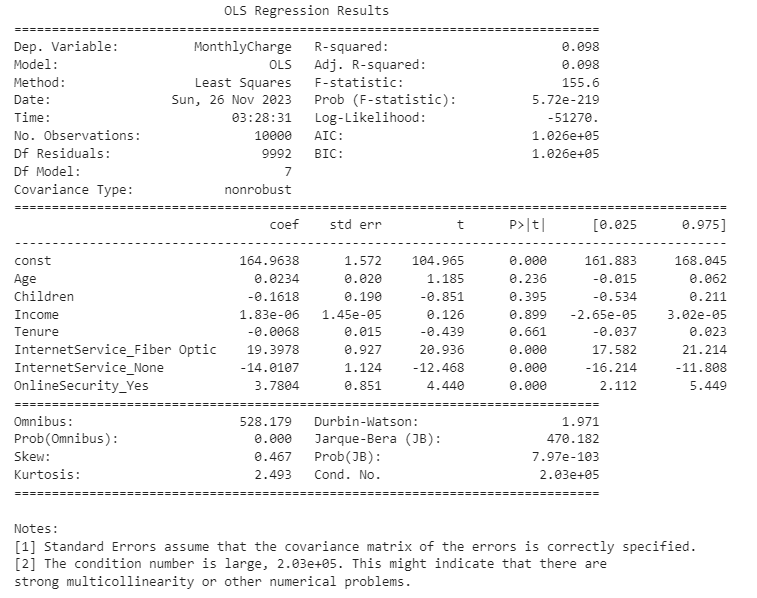
C3. The bivariate visualizations and summaries reveal insightful patterns in the independent variables concerning monthly charges. Regarding the histograms, a histogram of monthly charges shows a bell curve with a left skew. A distribution of age shows a uniform distribution. Children and income show a left skew. Tenure shows a bimodal distribution. A distinction of internet service indicates the frequency order as fiber optic, DSL, then none. A distribution of monthly security indicates most customers answered no.

C4. The data needs to be transformed to create a suitable dataset for analysis. The objectives of these transformations include standardizing numeric variables and encoding categorical variables for effective interpretation within the regression model. The first goal involves standardization of the numeric variables, 'MonthlyCharge', 'Age', 'Children', 'Income', and 'Tenure'. This aims to ensure these variables share a comparable scale, allowing for meaningful comparisons and interpretations in the regression model. The second goal pertains to encoding categorical variables, ‘InternetService’ and ‘OnlineSecurity,’ into a numerical format suitable for regression analysis. For example, by converting ‘OnlineSecurity’ into numeric values like 0 for no and 1 for yes using label encoding, the model can effectively interpret this categorical data. These transformations will prepare the data for the comprehensive analysis (Upadhyay, 2020)

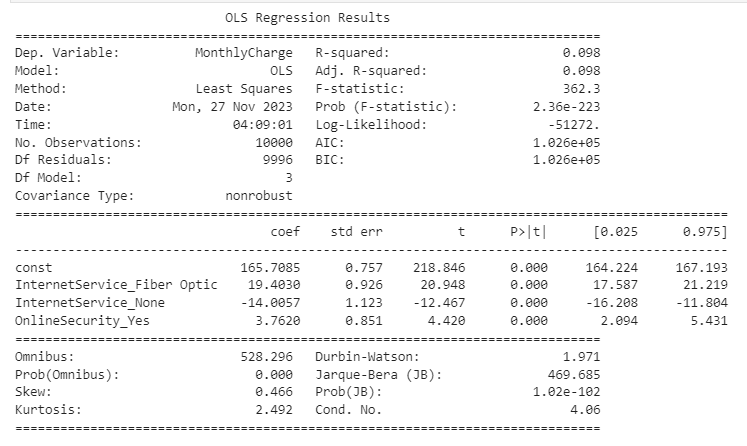
C5. The file is attached.

## Part IV: Model Comparison and Analysis

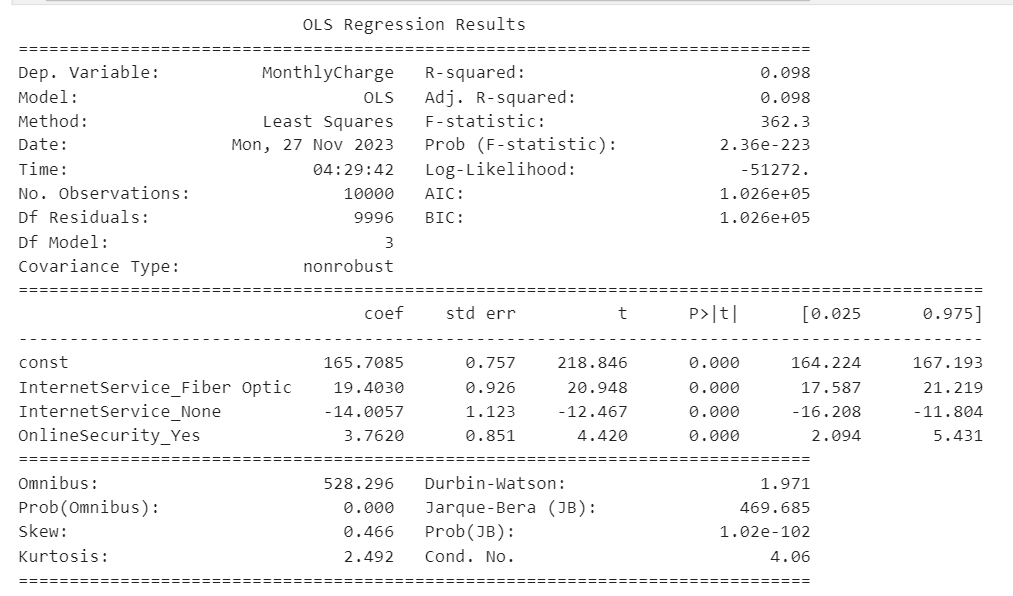
D1. The OLS Regression results showcase the statistics and coefficients of a multiple linear regression model aimed at understanding factors impacting customer monthly charge. The model accounts for approximately 9.8% of the variance in MonthlyCharge, signifying its low ability to explain variations. Adjusted R-squared is similar at .098. Age, Children, Income, and Tenure do not appear to be statistically significant predictors of MonthlyCharge as indicated by their p-values above the .05 level of significance. However, InternetService\_Fiber Optic, InternetService\_None, and OnlineSecurity\_Yes are all statistically significant predictors of Monthly. For the significant variables, the coefficient for fiber optic is 19.40, -14.01 for no internet, and 3.78 for online security,



D2. I will use backward elimination to remove insignificant variables based on p-values. This method removes a variable with a p-value higher than the significance level of .05. This removes variables that have no significant impact on monthly charges. The elimination process showed InternetService\_Fiber Optic, InternetService\_None, and OnlineSecurity\_Yes were the only significant variables. R-squared and Adjusted R-squared remain the same as the initial model, indicating that the model's ability to explain the variance in MonthlyCharge has not changed with the reduced number of predictors.

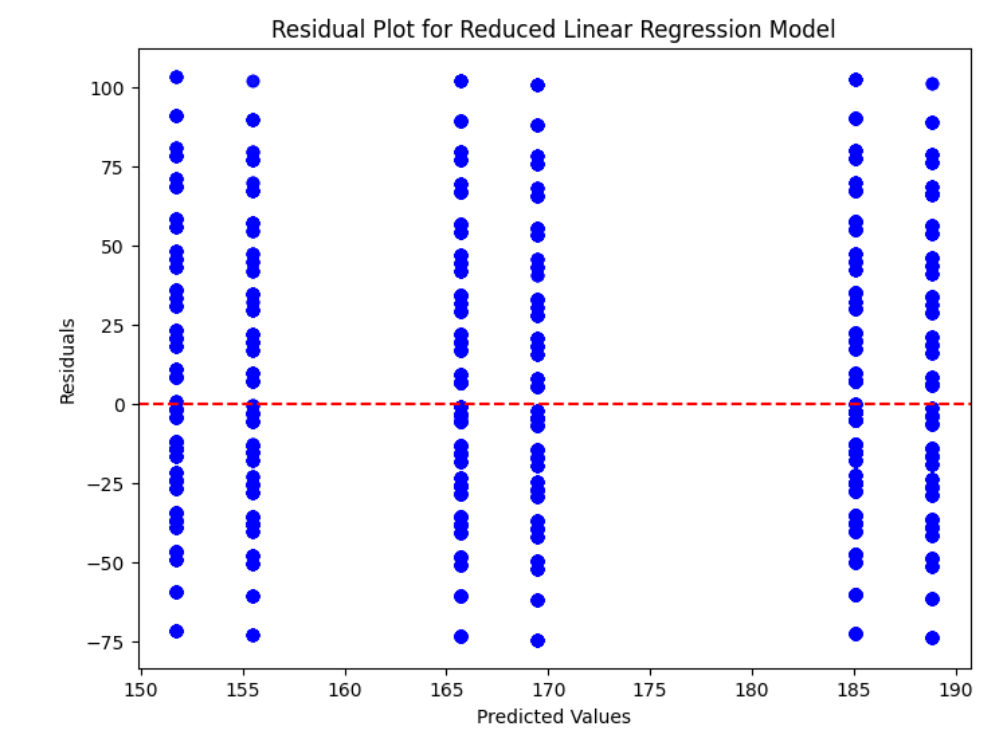


D3. The reduced model has an R-squared value of 0.098, similar to the initial model. This suggests that the reduced model has the same explanatory power as the initial model. The F-statistic of 362.3 and p-value of 2.36e-223 indicate that the reduced model is significant. 'nternetService\_Fiber Optic' has a coefficient 19.4030. This indicates that customers with fiber tend to have higher monthly charges. 'OnlineSecurity\_Yes' has a positive coefficient of 3.7620. This indicates that customers with online security have higher charges. ‘InternetService\_None’ has a coefficient of -14.0057, indicating that customers without internet have lower charges.



E1. The initial model’s aim was to identify factors influencing monthly charge, and it consisted of multiple independent variables. However, after scrutinizing the model's p-values, InternetService\_Fiber Optic, InternetService\_None, and OnlineSecurity\_Yes were the only statistically significant predictor for monthly charge. This prompted the need for a reduced model focusing solely on these variables. The model evaluation metric, specifically the R-squared value, served as a key benchmark in assessing both models. The initial multiple linear regression model exhibited an R-squared of 0.098, suggesting that approximately 9.8% of the variance in monthly charge was accounted for by the variables included in the model. In contrast, the reduced linear regression model, with significant variables, retained the same R-squared value of 0.098. While having three values maintained a significant association with monthly charge in the reduced model, the consistent R-squared value across both models implies that these three values alone explain a similar proportion of variance in churn as the multiple variables in the initial model. This suggests that the inclusion of additional factors in the initial model did not substantially enhance the model's explanatory power beyond the impact captured solely by InternetService\_Fiber Optic, InternetService\_None, and OnlineSecurity\_Yes.

E2. The residual standard error calculated for the reduced linear regression model came out at 40.78. This indicates that, on average, the actual monthly charge values deviate from the predicted values by about 40.78 units. For the residuals, a mean of 4.49e-13 is close to 0 indicating the model is good at predicting the residuals. A mean of -5.16 suggests that the model may underestimate monthly charge. The residual plot shows six lines with the mean of the residuals close to 0, indicating that the model is a good fit. The graph doesn’t show any patterns, indicating the model is a good fit.



E3. The file is attached.

## Part V: Data Summary and Implications

F1. The regression equation is MonthlyCharge = 165.71 + 19.40 \* InternetService\_Fiber Optic - 14.01 \* InternetService\_None + 3.76 \* OnlineSecurity\_Yes for the reduced model. InternetService\_Fiber Optic, InternetService\_None have been selected as predictors because they are significant with a p value below .05. Regarding the coefficients, InternetService\_Fiber Optic suggests an increase of $19.40 in increase in monthly charge. InternetService\_None decreases monthly charge by $14.01. OnlineSecurity\_Yes increases monthly charge by $3.76. While the model is significant, the R-squared value of .098 suggests that the model is limited since it explains 9.8% of the variance in MonthlyCharge. This suggests that there are other factors that influence monthly charge.

F2. Based on the limited explanatory power of selected variables, it's crucial to delve deeper into additional factors influencing monthly charge. Exploring other variables could offer comprehensive insights. The company should use domain knowledge to continue refining the mode by adding other variables that may be relevant. Using the knowledge that customers without internet have lower monthly charges, the company should consider marketing internet services to customers who do not have internet to increase revenue and try to get current customers to upgrade to fiber optic. The company should also consider marketing online security as an add-on to internet customers to further increase revenue.

## Part VI: Demonstration

G. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6ff1bc47-50fd-4f9a-b04d-b0c3005d1417

H.

Scikit. (n.d.). Sklearn.preprocessing.StandardScaler. Scikit-Learn. Retrieved November 14, 2023, from https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

I

Libretexts. (2023, April 27). *Checking Model Assumptions using Graphs*. https://stats.libretexts.org/Courses/Cerritos\_College/Introduction\_to\_Statistics\_with\_R/14%3A\_Multiple\_and\_Logistic\_Regression/14.03%3A\_Checking\_Model\_Assumptions\_using\_Graphs

Terra, J. (2019, March 5). *Python for Data Science and Data Analysis*. Simplilearn.Com. https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article

Upadhyay, A. (2020, July 23). StandardScaler and Normalization with code and graph. *Analytics Vidhya*. https://medium.com/analytics-vidhya/standardscaler-and-normalization-with-code-and-graph-ba220025c054