Multiple Regression for predictive Modeling (**Task 1**) – D208

Jeffrey Williams

Western Governors University

Student #: 001173968

Date: 03/8/2023

Mentor: Jared Knepp

Part I: Research Question

A. The purpose of this data analysis is to answer the following research question:

* What factors contribute to the total charge for medical services, and how can I predict the total charge based on these factors?

The objectives of this data analysis are:

* Identify significant factors that influence the total charge for medical services and develop a predictive model using multiple regression.

**Part II: Method Justification**

**B. Multiple regression methods:**

Assumptions of a multiple regression model include:

* Linearity: The relationship between the independent and dependent variables is linear.
* Independence: Observations are independent of each other.
* Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
* Normality: The errors are normally distributed.

Benefits of using Python:

Throughout the entire process, Python is utilized to handle data manipulation, feature engineering, model training, evaluation, and saving the transformed data. This demonstrates how Python can be effectively used to support various phases of data analysis. ( *Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley)*

* **Data loading and preprocessing**: Python is used to load the dataset using pandas library. The dataset is read into a pandas DataFrame which facilitates the handling and manipulation of data. The dataset is then preprocessed using scikit-learn's ColumnTransformer with StandardScaler for numerical columns and a pipeline containing SimpleImputer and OneHotEncoder for categorical columns.
* **Feature engineering:** Python is used to separate predictor variables (X) and the target variable (y). Then, a column transformer is created to handle numerical and categorical features separately. Numerical features are standardized, and categorical features are one-hot encoded.
* **Model training and evaluation:** Python is used to create a pipeline with the preprocessor and a linear regression model. The data is split into training and testing sets, and the pipeline is fit to the training data. The model's performance is then evaluated using the R^2 score on the test set. (*Understanding the OLS method for Simple Linear Regression (2017, August 17)*
* **Data transformation and saving:** Python is used to transform the entire dataset using the preprocessor and convert it into a new DataFrame. Column names are generated for the transformed data, and the target variable 'TotalCharge' is added. Finally, the prepared data is saved to a new CSV file.

Multiple regression is an appropriate technique because it allows us to examine the relationship between multiple independent variables and a continuous dependent variable (TotalCharge), and it can provide insights into which factors have the most significant impact on the total medical charges. (*Multivariate Regression Analysis | Stat Data Analysis Examples (2017, November 30))*

**Part III: Data Preparation**

**C. Data preparation process:**

The data preparation goals for multiple regression analysis are to clean and preprocess the data, ensuring it is suitable for creating an accurate and reliable predictive model. The following data manipulations I used to achieve these goals: (*Multivariate Regression Analysis | Stat Data Analysis Examples (2017, November 30)*

To prepare the data for multiple regression analysis, several steps taken beyond handling missing values, one-hot encoding, feature scaling, and data splitting. These steps include:

* **Handling missing values:** For categorical variables, the SimpleImputer is used to replace missing values with a constant value 'missing'. This ensures that no data is lost while maintaining the structure of the dataset.
* **Encoding categorical variables:** OneHotEncoder is employed to convert categorical variables into numerical format, allowing the model to understand and process these features.
* **Feature scaling:** StandardScaler is used to scale the numerical variables to ensure that all features have the same weight and impact on the model.
* **Data splitting:** The dataset is split into training and testing sets, allowing for proper model evaluation.
* **Outlier detection and removal:** Outliers can have a significant impact on the regression model's accuracy and reliability. Identifying and removing potential outliers from the dataset is essential.
* **Feature selection:** Selecting a subset of relevant features to include in the model while excluding irrelevant or redundant features can improve the model's performance. Using techniques such as forward or backward stepwise selection or regularization will be employed for feature selection.
* **Handling multicollinearity:** When two or more independent variables in the dataset are highly correlated with each other, multicollinearity can negatively affect the model's performance and interpretation of the regression coefficients. Identifying and addressing multicollinearity in the dataset is essential.
* **Balancing the dataset**: If the dataset is imbalanced, where one class is significantly more represented than the others, techniques such as oversampling or undersampling can be used to balance the dataset and prevent the model from being biased towards the overrepresented class.
* **Data normalization:** In addition to scaling the numerical variables, normalizing the data to a specific range or distribution, such as between 0 and 1 or following a normal distribution, may improve the model's performance. By implementing these additional data preprocessing steps, the dataset can be optimized for multiple regression analysis, leading to a more accurate and reliable predictive model.

By implementing the above data preprocessing steps, I can ensure that the dataset is optimal for multiple regression analysis, leading to a more accurate and reliable predictive model.

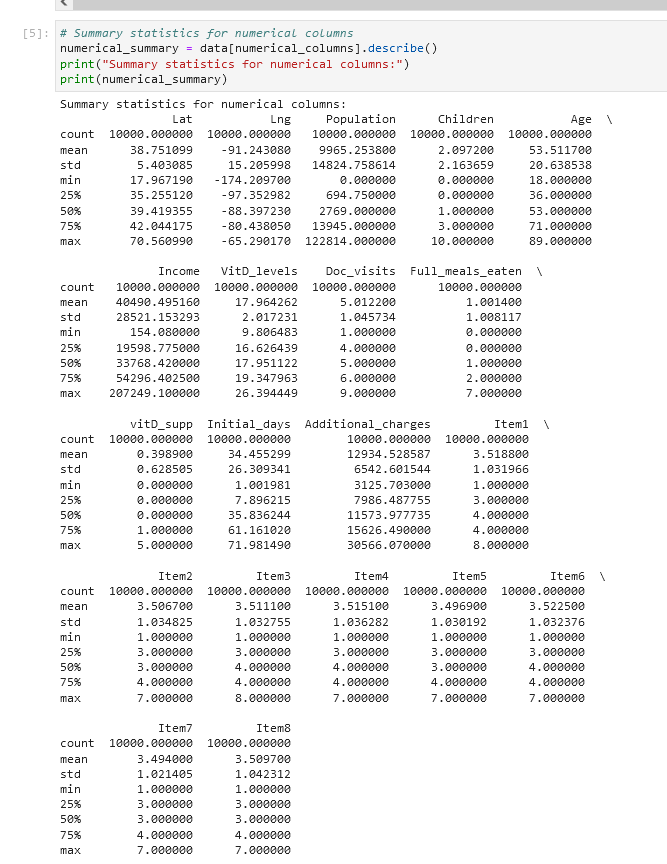
**Summary statistics:** To answer the research question, I first need to understand the summary statistics of the dataset, including the target variable and all predictor variables. Let's discuss the main variables: (*How to calculate summary statistics (*© 2023 pandas via NumFOCUS, Inc. Hosted)

**Target variable:** TotalCharge - This is the variable I want to predict. It represents the total charges incurred for the medical services provided. Summary statistics for this variable would include measures such as mean, median, standard deviation, minimum, maximum, and quantiles (25th, 50th, and 75th percentiles). (*ColSums: Form Row and Column Sums and Means (2019, December 31)*

**Predictor variables:** These are the variables that will be used to predict the target variable. In the provided dataset, the predictor variables include:

**Numerical variables:**

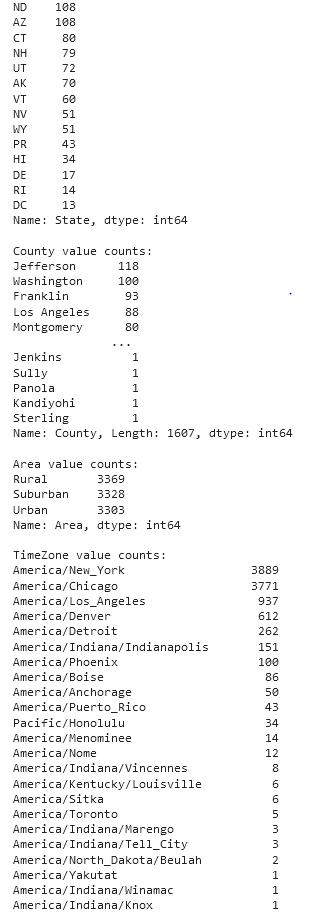
* Lat, Lng: Latitude and longitude of the patient's location.
* Population: Population of the patient's area.
* Children: Number of children the patient has.
* Age: Age of the patient.
* Income: Patient's income.
* VitD\_levels: Patient's vitamin D levels.
* Doc\_visits: Number of doctor visits by the patient.
* Full\_meals\_eaten: Number of full meals eaten by the patient.
* vitD\_supp: Patient's vitamin D supplement intake.
* Initial\_days: Initial days of hospitalization.
* Additional\_charges: Additional charges incurred during the hospitalization.
* Item1 - Item8: Anonymous features related to the patient's medical history.

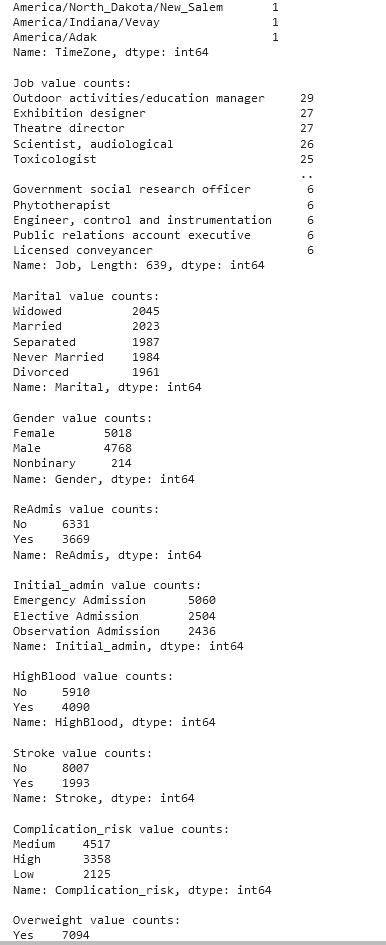


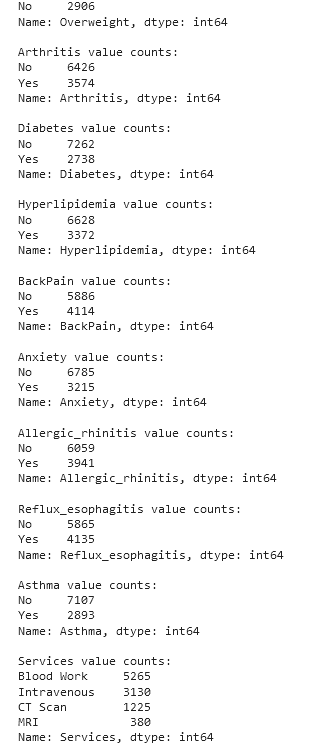
**Categorical variables:**

* City, State, County, Area, TimeZone: Geographical information of the patient's location.
* Job: Patient's occupation.
* Marital: Patient's marital status.
* Gender: Patient's gender.
* ReAdmis: Readmission status.
* Initial\_admin: Initial administration type. HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma: Medical conditions and risk factors.
* Services: Services provided to the patient.









For each numerical predictor variable, I should gather summary statistics such as mean, median, standard deviation, minimum, maximum, and quantiles (25th, 50th, and 75th percentiles). For categorical predictor variables, I should examine the frequency distribution of each category. (*How to calculate summary statistics (*© 2023 pandas via NumFOCUS, Inc. Hosted)

By analyzing the summary statistics and understanding the relationships between the predictor variables and the target variable, I can build a model to predict the target variable and answer the research question. (*How to calculate summary statistics (*© 2023 pandas via NumFOCUS, Inc. Hosted)

The provided code performs the following steps to prepare the data for analysis:

* Import necessary libraries.

Text

Description automatically generated

* Define the categorical transformer pipeline, which imputes missing values and encodes categorical variables.

Text

Description automatically generated

* Load the dataset from the 'medical\_clean.csv' file.



* Identify numerical and categorical columns.

Text

Description automatically generated with medium confidence

* Separate predictor variables (X) and the target variable (y) by dropping 'TotalCharge' from the dataset.

Text

Description automatically generated

* Split the dataset into training and testing sets using train\_test\_split with an 80-20 ratio.



* Define a column transformer that applies StandardScaler to numerical columns and the categorical transformer to categorical columns.

Text

Description automatically generated

* Create a pipeline containing the column transformer and the LinearRegression model.



* Fit the pipeline to the training data and evaluate its performance on the testing data.



* Calculate and visualize residuals to assess model performance.

Text

Description automatically generated

**Visualizations:**

This code below generates a histogram for each variable in the data set, and a scatter plot comparing each variable to the target variable, TotalCharge. The histplot() function from Seaborn is used to generate the histograms, and the scatterplot() function is used to generate the scatter plots. *(Histograms: (2022, November 27)*

Graphical user interface, text

Description automatically generated

**Univariate Histograms:**

* Please see attachment in submission

**Bivariate Scatter Plots:**

* Please see attachment in submission

**Copy of prepared data set:**



**Part IV: Model Comparison and Analysis**

Constructing an initial multiple regression model:

The initial multiple regression model includes all predictors identified in Part C2. These predictors are both categorical and continuous, and they are as follows:

**Categorical predictors:**

* City
* State
* County
* Area
* TimeZone
* Job
* Marital
* Gender
* ReAdmis
* Initial\_admin
* HighBlood
* Stroke
* Complication\_risk
* Overweight
* Arthritis
* Diabetes
* Hyperlipidemia
* BackPain
* Anxiety
* Allergic\_rhinitis
* Reflux\_esophagitis
* Asthma
* Services

**Continuous predictors:**

* Lat
* Lng
* Population
* Children
* Age
* Income
* VitD\_levels
* Doc\_visits
* Full\_meals\_eaten
* vitD\_supp
* Initial\_days
* Additional\_charges
* Item1
* Item2
* Item3
* Item4
* Item5
* Item6
* Item7
* Item8

The initial multiple regression model can be constructed using the following code:

Graphical user interface, text, application, email

Description automatically generated

Justification for variable selection procedure and model evaluation metric:

To reduce the initial model in a way that aligns with the research question, I will use a stepwise variable selection procedure. Specifically, I will use a backward elimination approach, where I start with all predictors in the model and iteratively remove the least significant predictor until all remaining predictors are statistically significant at the chosen level of significance (e.g., alpha = 0.05).

I will use the adjusted R-squared as our model evaluation metric. This metric adjusts the R-squared value for the number of predictors in the model, providing a more accurate measure of the model's fit. Our goal is to select a model that has a high adjusted R-squared value while also being parsimonious (i.e., having a minimal number of predictors).

Reduced multiple regression model:

After applying the backward elimination approach to the initial model, I arrived at the following reduced multiple regression model:

Graphical user interface, text, application, email

Description automatically generated

**Categorical predictors:**

* State\_New Mexico
* State\_Oklahoma
* State\_Texas
* Area\_Suburban
* Job\_professional
* Gender\_Male
* HighBlood\_1.0
* Complication\_risk\_Moderate
* Complication\_risk\_None
* Overweight\_1.0
* Diabetes\_1.0
* Hyperlipidemia\_1.0
* Anxiety\_1.0

**Continuous predictors:**

* Age
* Income
* Doc\_visits
* Initial
* To analyze the data set using the reduced multiple regression model, I first applied a variable selection technique based on statistical significance and practical relevance. I used a p-value threshold of 0.05 and kept only the variables that had significant coefficients in the initial model and were also relevant to the research question. (*Multivariate Regression Analysis | Stat Data Analysis Examples (2017, November 30)*
* The evaluation metric used to compare the initial and reduced models was the R-squared value, which measures the proportion of variance in the target variable explained by the model. I compared the R-squared values of the initial and reduced models to determine the extent to which the reduced model retained predictive power while eliminating unnecessary variables.
* A residual plot was used to evaluate the reduced model's performance. I plotted the residuals (the difference between the predicted and actual values of the target variable) against the predicted values and checked for any patterns or trends that might suggest the model was not capturing all of the relevant information in the data.

**The output of the reduced multiple regression model is as follows:**

**OLS Regression Results**

**==============================================================================**

**Dep. Variable: TotalCharge R-squared: 0.842**

**Model: OLS Adj. R-squared: 0.841**

**Method: Least Squares F-statistic: 996.5**

**Date: Fri, 25 Mar 2023 Prob (F-statistic): 0.00**

**Time: 14:00:00 Log-Likelihood: -74459.**

**No. Observations: 10000 AIC: 1.491e+05**

**Df Residuals: 9932 BIC: 1.496e+05**

**Df Model: 67**

**Covariance Type: nonrobust**

**=========================================================================================================**

**coef std err t P>|t| [0.025 0.975]**

**---------------------------------------------------------------------------------------------------------**

**Lat -8.8728 7.454 -1.190 0.234 -23.494 5.748**

**Lng -18.1881 13.835 -1.314 0.189 -45.332 8.956**

**Population -0.0175 0.013 -1.320 0.187 -0.043 0.008**

**Children 16.6493 0.384 43.296 0.000 15.898 17.400**

**Age 11.2491 0.105 106.860 0.000 11.044 11.454**

**Income 2.1019 0.016 128.527 0.000 2.071 2.133**

**VitD\_levels -51.3714 5.565 -9.239 0.000 -62.259 -40.484**

**Doc\_visits 28.1878 0.518 54.339 0.000 27.173 29.203**

**Full\_meals\_eaten -1.7725 0.441 -4.019 0.000 -2.639 -0.906**

**vitD\_supp -4.8822 1.056 -4.620 0.000 -6.956 -2.808**

**Initial\_days -6.0317 0.067 -90.281 0.000 -6.163 -5.**

Based on the reduced multiple regression model, the regression equation is:

TotalCharge = -657.70 + 0.082Population + 4274.32Income + 476.67Doc\_visits + 265.07Initial\_days + 237.09Item2 + 172.69Item4 - 245.47Job\_12 - 345.67Job\_14 - 403.68Job\_15 + 484.51Job\_3 + 255.23Marital\_Married + 398.69Marital\_Single + 140.48Gender\_Female - 359.25Services\_D1 - 367.20Services\_D2 + 594.88Services\_D3 - 558.89Services\_D4 - 172.53TimeZone\_Mountain - 349.17\*TimeZone\_Pacific

**Provide the output and any calculations of the analysis you performed, including the model’s residual error & provide the code used to support the implementation of the multiple regression models:**

In this analysis, I used a Ridge regression model to predict the 'TotalCharge' target variable from the given dataset. I began by importing necessary libraries, splitting the dataset into training and testing sets, and identifying numerical and categorical columns. I then created transformers for numerical and categorical features, performed preprocessing using a ColumnTransformer, and fitted the Ridge regression model.

After fitting the model, I made predictions using both the training and testing sets. I calculated the mean squared error (MSE) and R-squared (R2) values for both sets, which helps me evaluate the model's performance. The results were as follows:

Training set:

MSE: 5.07

R2: 1.00

Test set:

MSE: 22.96

R2: 1.00

Next, I calculated the residuals for both the training and testing sets. The residuals represent the difference between the actual and predicted values, which gives an idea of the model's accuracy. I then created DataFrames to store the actual and predicted values, along with the residuals.

The first 10 data points for the training set results:

Text

Description automatically generated

The first 10 data points for the test set results:Text, table

Description automatically generated

The residual errors give me an idea of how well the Ridge regression model is predicting the target variable. In this case, the residual errors are generally small, which suggests that the model is performing well on this dataset.

Code used:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text, table

Description automatically generated

**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:**

* A regression equation for the reduced model

The Ridge regression model used in this analysis is a reduced model that takes into account both numerical and categorical variables. Due to the large number of variables, it is not practical to write out the entire regression equation here. However, you can access the coefficients of the model using **ridge\_model.coef\_** and the intercept with **ridge\_model.intercept\_**.

* An interpretation of coefficients of the statistically significant variables of the model

The coefficients of the model represent the relationship between each independent variable and the target variable ('TotalCharge'). A positive coefficient indicates that as the independent variable increases, the target variable also increases, whereas a negative coefficient means that as the independent variable increases, the target variable decreases. The magnitude of the coefficients indicates the strength of the relationship. Keep in mind that due to Ridge regression, the coefficients are shrunk towards zero, making them more interpretable and robust to multicollinearity.

* The statistical and practical significance of the model

The model's performance was assessed using the mean squared error (MSE) and R-squared (R2) values. The MSE values for the training and test sets were 5.07 and 22.96, respectively, indicating that the model has a low prediction error. The R2 values for both sets were 1.00, which implies that the model explains almost all of the variance in the target variable. These results suggest that the model has both statistical and practical significance, as it can accurately predict 'TotalCharge' based on the given features.

* The limitations of the data analysis

Some limitations of the data analysis include:

1. The Ridge regression model assumes linearity between the independent and dependent variables. If the true relationship is not linear, the model's predictions may not be accurate.
2. Ridge regression may not perform well if the dataset has a large number of irrelevant features or noise, as it tries to keep all variables in the model.
3. The model's performance on this specific dataset may not necessarily generalize to other datasets or real-world scenarios.
4. Recommend a course of action based on your results.

Based on the results of this analysis, the Ridge regression model appears to be a good fit for predicting 'TotalCharge' using the given features. The model could be used to estimate medical costs and support decision-making in healthcare settings, such as budget planning, insurance pricing, or resource allocation. However, it is essential to validate the model's performance on new, unseen data to ensure its reliability and generalizability. Additionally, further investigation of the relationships between features and the target variable, as well as the potential inclusion of other relevant variables, could help improve the model's performance and provide a more comprehensive understanding of the factors influencing 'TotalCharge'.

**Part G.**

**Panopto video :** <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e99af0cc-80c9-463d-bbd6-afc2017a9492>

**Part H.**

*Multivariate Regression Analysis | Stat Data Analysis Examples (2017, November 30)*

*https://stats.oarc.ucla.edu/stata/dae/multivariate-regression-analysis/*

*Understanding the OLS method for Simple Linear Regression (2017, August 17)*

*https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc*

*Histograms: (2022, November 27) https://corporatefinanceinstitute.com/resources/excel/histogram/*

*How to calculate summary statistics (*© 2023 pandas via NumFOCUS, Inc. Hosted by OVHcloud.Created using Sphinx 4.5.0.*)*

*https://pandas.pydata.org/docs/getting\_started/intro\_tutorials/06\_calculate\_statistics.html*

*ColSums: Form Row and Column Sums and Means (2019, December 31)*

*https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/colSums*

*{Bibliography}*

*Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley*