**Performance Assessment: D208 – Multiple Regression**

**A. Research Question**

**1.** For this assessment, the research question is as follows: using multiple linear regression, can we predict a patients TotalCharge based on their age, gender, the initial days, the initial admission, the services, doc visits, and complication risk?

**2.** The goal of this data analysis is to determine whether we can predict the total charges a patient paid based on these variables: age, gender, initial\_days, services, doc\_visits, initial\_admin, and complication\_risk. These variables were chosen by my own personal preference as they provide good insight into what could lead to the TotalCharge. For example, an older patient with a high number of doc visits and an emergency admission should, theoretically, pay more in charges to the hospital.

**B. Method Justification**

**1.** “Five main assumptions underlying multiple regression models must be satisfied: (1) linearity, (2) homoskedasticity, (3) independence of errors, (4) normality, and (5) independence of independent variable” (“Basics of Multiple Regression and Underlying Assumptions”, n.d.).

* Linearity: this assumption states that there is a linear relationship between the target variable and all of the predictor variables
* Homoskedasticity: this assumption states that “the residuals have constant variance at every point in the linear model” (Zach, 2021).
* Independence of errors: this assumption states that “none of the predictor variables are highly correlated with each other” (Zach, 2021).
* Normality: this assumption states that each of the residuals are all normally distributed.
* Independence of independent variable: this assumption states that each observation is independent of each other.

**2.** One benefit of using R is that R is a free and open-sourced platform that anyone can download and use. Another benefit of using R is that R has numerous free packages to install that allows the user to do more with the data than the original code allows. There are also great visualization tools, and it is easy to pick up and learn and apply numerous statistical methods to your data set.

**3.** Since the target variable, in this case TotalCharge, is a continuous variable, multiple linear regression is an appropriate way to analyze this research question.

**C. Data Preparation**

**1.** In order to properly clean the data to prepare it for analysis, I will first remove unwanted variables from the data set, and then check for any missing data. If there is any missing data, I will remove them from the data set. Next, I will remove all duplicate entries from the data set. The next step to clean the data is to check for outliers. Categorical data cannot have outliers so I will use boxplots to check for outliers in the continuous data and then remove them if necessary. The code has been attached as an RScript file alongside this written assessment.

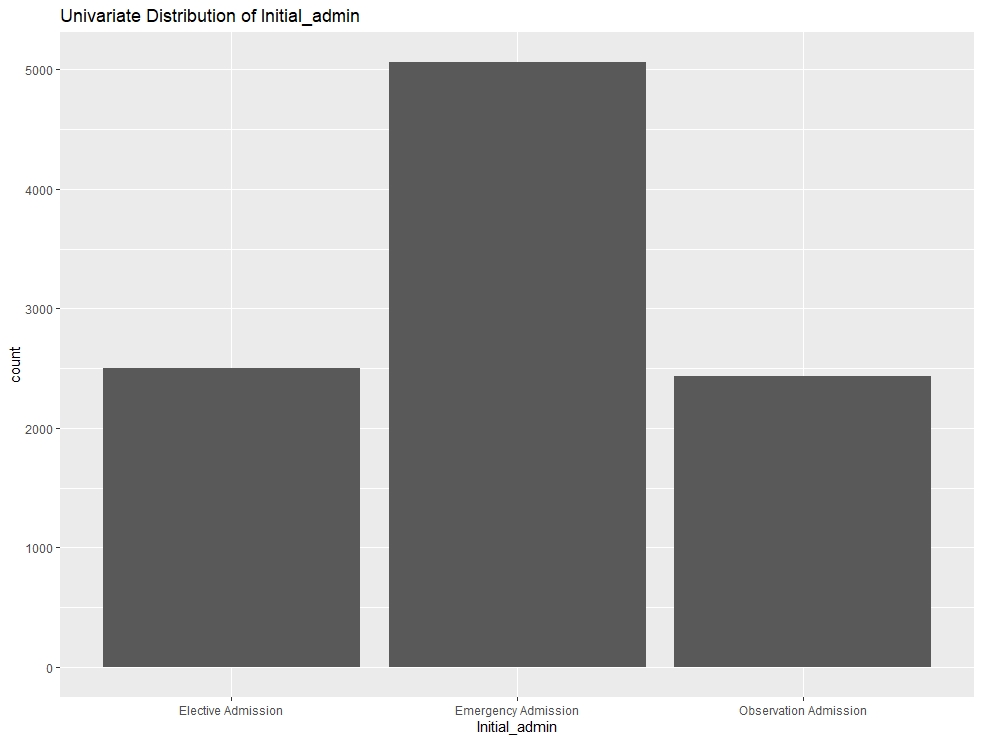
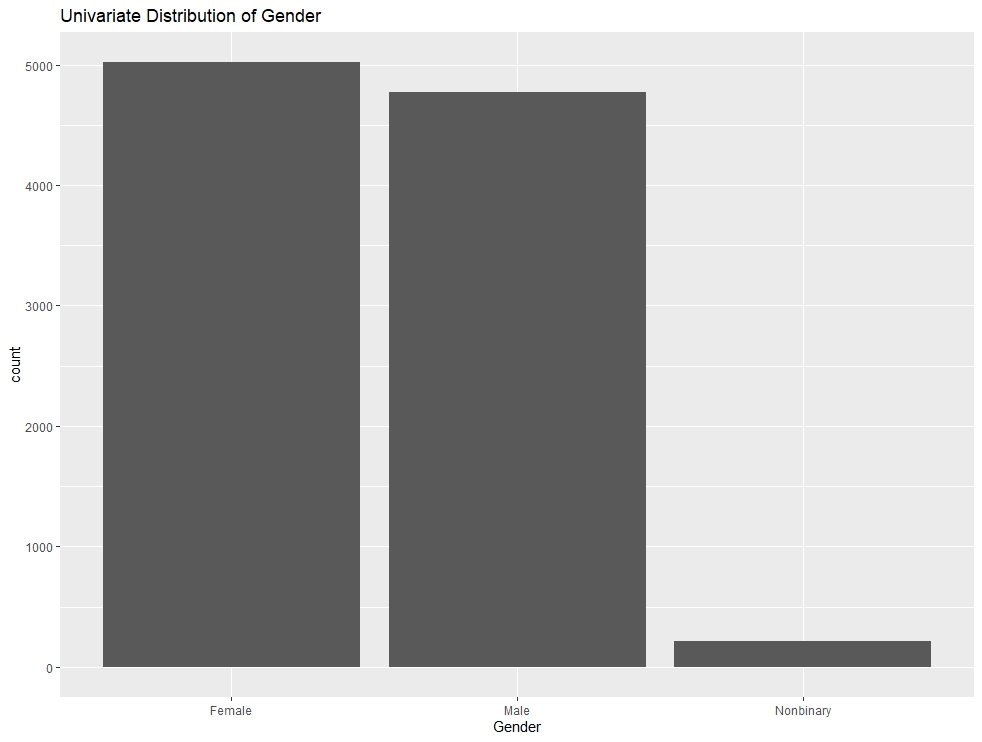
**2.** The following is a screenshot of the summary statistics for all variables:

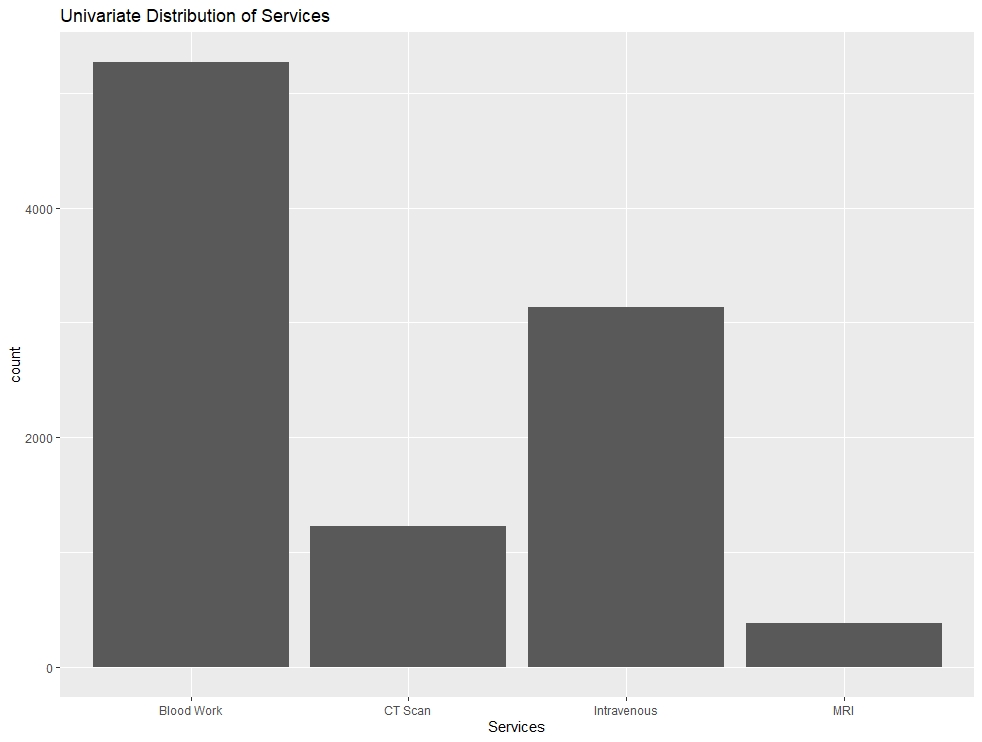
A screenshot of a computer

Description automatically generated

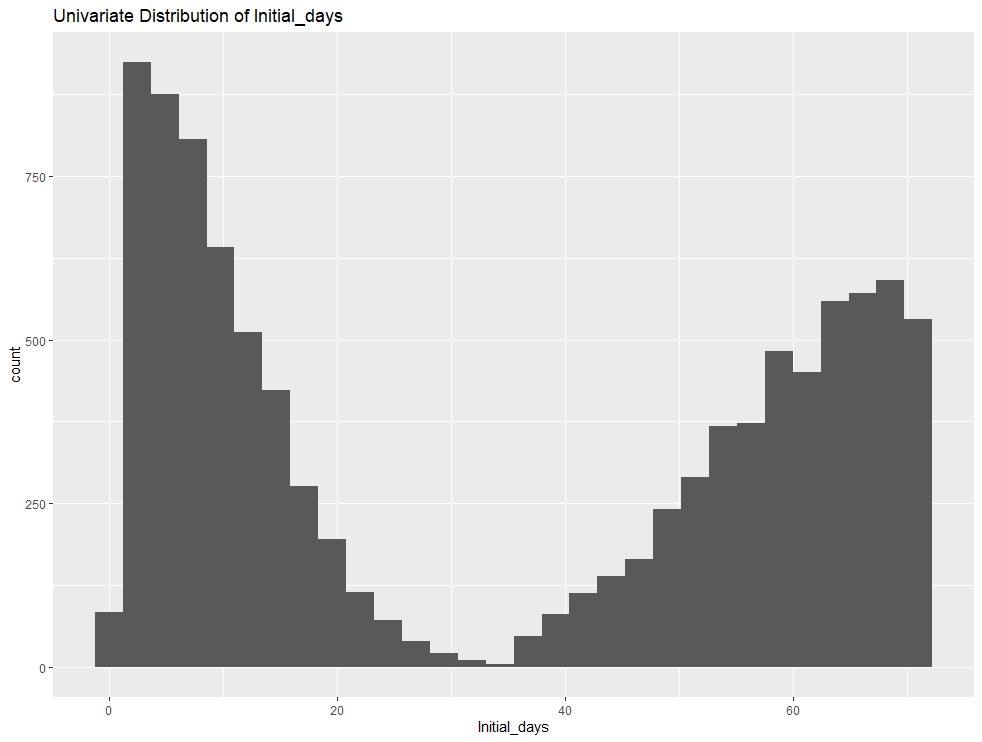
* Age: the average age is 53.51 years old with the lowest age of 18 and an oldest of 89
* Gender: 5018 Female, 4768 Male, and 214 nonbinary
* Doc\_visits: the average number of doctor visits is 5.012 with the fewest number of visits being 1 and the most being 9
* Initial\_admin: 2504 elective admission, 5060 emergency, and 2436 observation
* Complication\_risk: 3358 high, 2125 low, and 4517 medium
* Services: 5265 blood work, 1225 CT scan, 3130 intravenous, and 380 MRI
* Initial\_days: the average number of initial days is 34.455 with the most being 71.981 and the least being 1.002
* TotalCharge (target variable): the average total charge is $5312 with the lowest being $1938 and the highest being $9181

**3.** Univariate distributions of each variable as well as bivariate distributions using the target TotalCharge variable have been generated in R. The RScript file showing these distributions has been attached alongside this written assessment. The following screenshots also showcase each of the visualizations that were created:

A graph showing a number of squares

Description automatically generatedA graph showing a number of bars

Description automatically generatedA graph of a number of bars

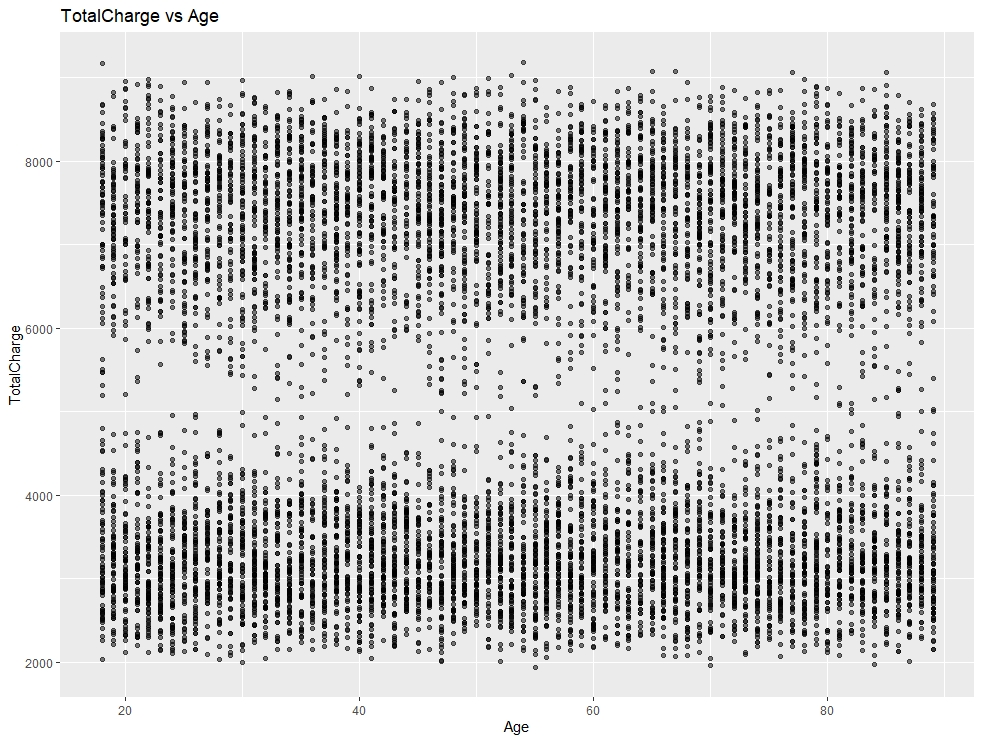
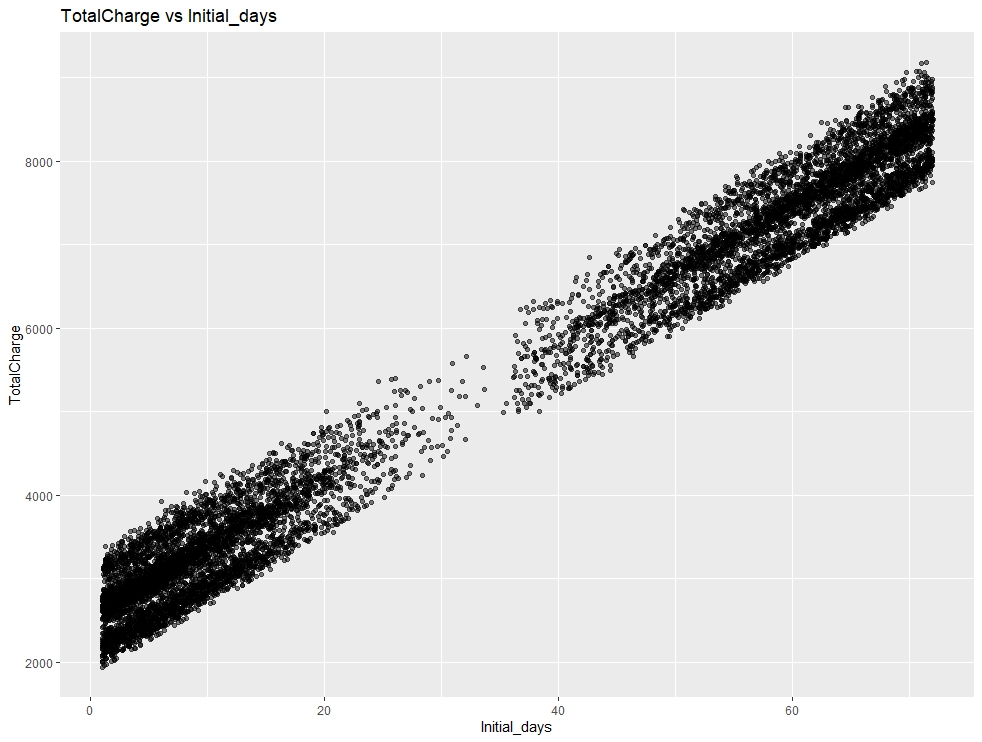
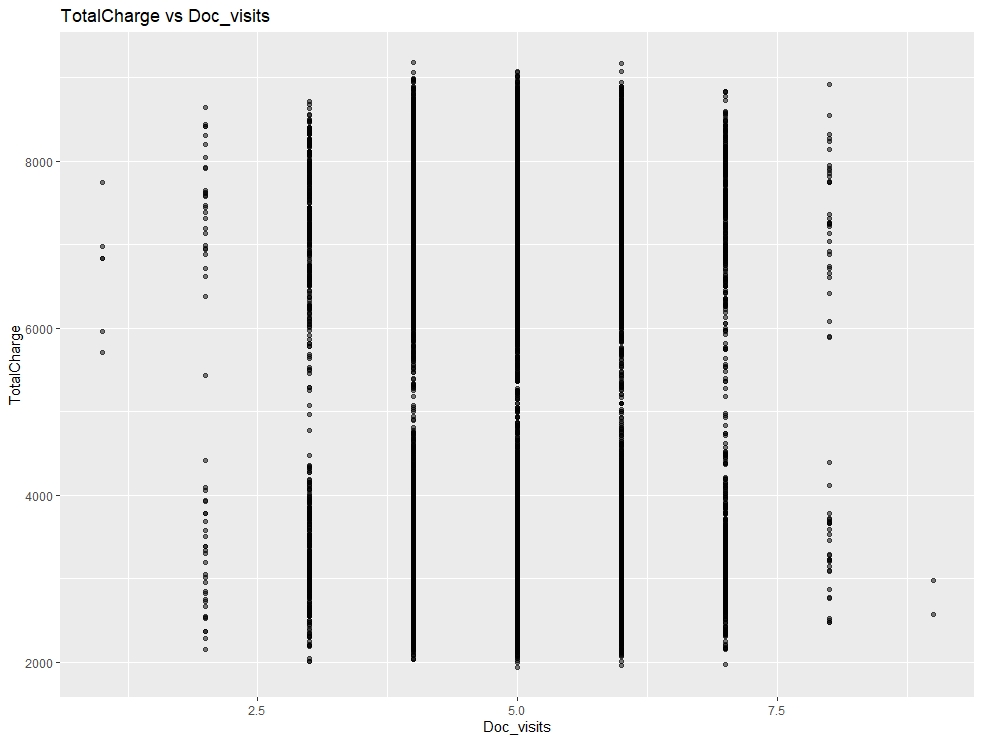
Description automatically generatedA graph showing a number of levels with Willis Tower in the background

Description automatically generatedA graph showing different colored squares

Description automatically generatedA graph showing different colored squares

Description automatically generatedA graph showing different colored squares

Description automatically generatedA graph showing different colored rectangles

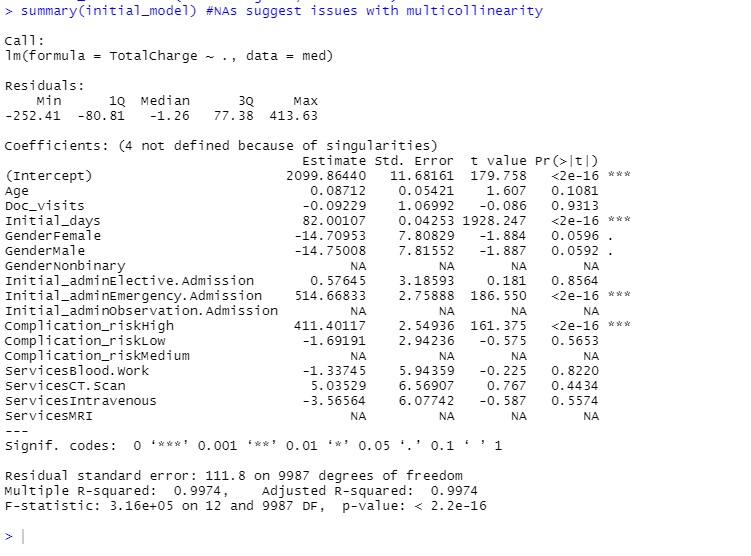
Description automatically generated 

**4.** In order to transform the data to prepare it for analysis, I only need to encode the categorical data by creating dummy variables in order to run the linear regression model. This process is called one-hot encoding. I will not be scaling the continuous variables as that will affect the interpretability of the model. The code has been attached as an RScript file alongside this written assessment.

**5.** A copy of the cleaned data set has been attached alongside this written assessment.

**D. Model Comparison**

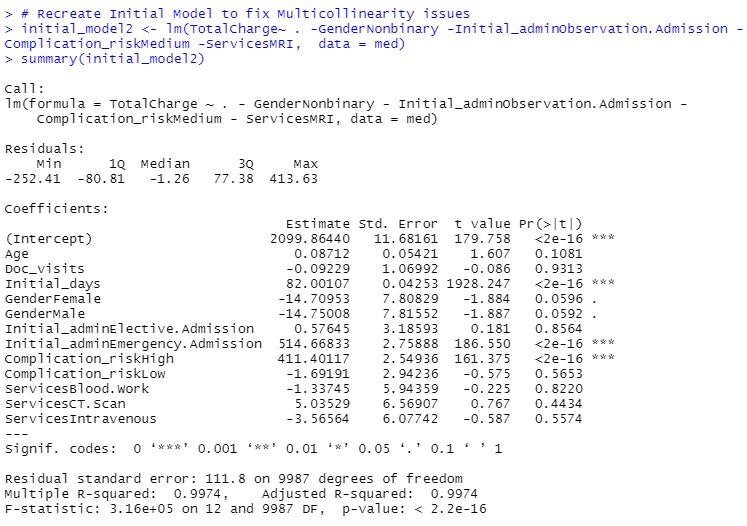
**1.** In initial model was created in R using all of the predictor variables. The following screenshot shows the summary of that model:



Right away, the first thing you will notice is a handful of variables are returning as NA. This means that those variables have multicollinearity and are dependent on each other or upon other variables. The next screenshot shows the aliases of the model:



Here you can see that GenderNonbinary, Initial\_adminObservation.Admission, Complication\_riskMedium, and ServicesMRI are dependent on other variables. In order to properly create the linear model, these variables should be removed as well to fix the issues with multicollinearity. After omitting these variables in the model, the new initial model summary looks as follows:



As you can see, there are no more NAs, and the issues with multicollinearity as been resolved.

**2.** After creating this second version of the initial model without the multicollinearity issues, stepwise reduction was chosen as the model feature selection procedure in order to reduce the number of predictors in the model. This method of feature selection uses backwards elimination and forward selection as a way to determine the appropriate number of predictors to include in the model while retaining the highest variance of the data as possible (in other words, the adjusted R-squared value). Stepwise starts at both the beginning and the end of the predictors set and automatically adds and removes predictors in each direction based on the AIC, or Akaike information criterion. The algorithm continues this process until all of the predictors are significantly significant within this reduced model. This generates the best overall model possible and reduces the overall number of predictors.

**3.** Using stepwise regression in the “both” direction, a reduced model was created in R. The following screenshot showcases the summary of the reduced model:

A screenshot of a computer

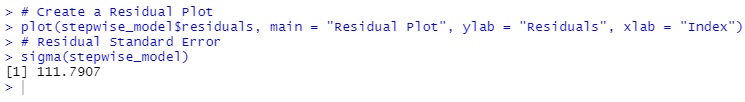
Description automatically generated

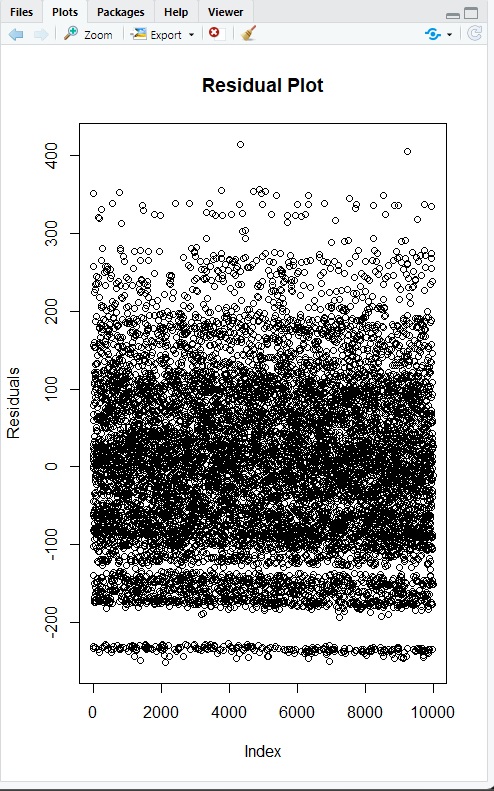
Using this method, the model was reduced to include just the following predictors: age, initial\_days, GenderFemale, GenderMale, Initial\_adminEmergency.Admission, Complication\_riskHigh, and ServicesCT.Scan.

**E. Analysis**

**1.** As stated above, the initial model was reduced to include these predictors: age, initial\_days, GenderFemale, GenderMale, Initial\_adminEmergency.Admission, Complication\_riskHigh, and ServicesCT.Scan. To restate, these predictor variables are included in the final model after applying stepwise reduction to the initial model to generate the model with the least number of predictors with the highest variance. The stepwise algorithm was applied to the initial model which determined that the variables mentioned above are the most statistically significant. We can now compare and evaluate the two models by observing the value for adjusted R-squared. This model evaluation metric determines how good of a fit the model is by explaining the variability while adjusting based on the number of predictor variables. In this case both models have an adjusted R-squared value of 0.9974. This means that the reduced model, despite having fewer variables, maintained the same variability as the initial model, which indicates the reduced model is a VERY strong fit as it contains 99.74% of the variability against the TotalCharge variable.

**2.** The following screenshots showcase the residuals plot and the residual standard error, as well as the code to create them:





Upon observing the residuals plot, the residuals seem to be uniformly distributed around the horizontal axis. There are no clear patterns within the residuals, but there are noticeable outliers the further from 0 you get in each Y direction. The residual standard error, as seen in the code snippet, is 111.7907. This indicates that you can expect the predicted TotalCharge to be off by +- $111.79.

**3.** A copy of the code as an RScript file has been attached alongside this written assessment.

**F. Data Summary and Implications**

**1.** The following list is a discussion of the results of the data analysis:

* From the intercept and coefficients from the reduced model summary, we can come up with the following regression equation:

TotalCharge = 2097.13424 + (0.08690 \* Age) + (82.00128 \* Initial\_days) + (-14.77567 \* GenderFemale) + (-14.78232 \* GenderMale) + (514.37051 \* Initial\_adminEmergency.Admission) + (411.92997 \* Complication\_riskHigh) + (7.14155 \* ServicesCT.Scan).

* The following is an interpretation of the coefficients:
  + 2097.13424: this is the intercept. This is the estimated TotalCharge when all the other predictors are at 0, which is not practical in this case since someone who has been in the hospital for 0 days should not be paying anything, but this is a predicted estimate.
  + 0.08690: this coefficient is stating that for every additional year older the patient is, that the total charge is predicted to increase by $0.09. In other words, a 50-year-old patient (50\*0.08690=4.345) is predicted to pay an additional $4.35.
  + 82.00128: this coefficient is stating that for every additional day, the total charge is predicted to increase by $82.
  + -14.775677: this coefficient is predicting the total charge to decrease by $14.78 if the patient is female.
  + -14.78232: this coefficient is predicting the total charge to decrease by $14.78 if the patient is male.
  + 514.37051: this coefficient states that if the initial reason for admission is an emergency, then the total charge is expected to increase by $514.37.
  + 411.92997: this coefficient is stating that if the patient has a high complication risk, then the total charge is expected to increase by $411.93.
  + 7.14155: this coefficient is stating that if the patient received a CT scan, then the total charge is predicted to increase by $7.14.
* All of the predictors in the reduced model have p-values at or lower than 0.05 which indicates that they are all statistically significant in predicting the total charge. The except to this is the “age” predictor, which has a p-value of 0.1088. The reason why the algorithm might have chosen to keep this variable despite a slightly less significant p-value is that the predictor still could have created a better fit within the reduced model. It is also possible that the data might be dependent on age and thus it was kept in. From a practical perspective, this regression equation is stating the patients who spend more time in the hospital, were admitted for an emergency, and/or have a high complication risk, are being charged significantly more money by the hospital.
* The data analysis does have some limitations. One such limitation is that this does not take a patients insurance into account which could be affecting just how much the patient is paying, but it is also possible that the total charge is before insurance even comes into play. Another limitation is that the predictors could still be affected by multicollinearity which can affect how the coefficients are interpreted. One other limitation is that we chose the variables that are statistically significant for the reduced model according to the stepwise algorithm, but other variables could potentially have a practical effect or still be significant.

**2.** Based upon the results of the data analysis, I recommend a focus on efficiency. The number of days initially in the hospital contribute greatly to the total charge, so any strengths to how efficient doctors can be to get patients discharged as soon as possible will greatly affect how much a patient is paying. We should especially optimize efficiency within the emergency department since emergency admittance and patients with high complication risks also significantly affects the patients total charge.

**G. Sources**

“Basics of Multiple Regression and Underlying Assumptions.” CFA Institute, www.cfainstitute.org/en/membership/professional-development/refresher-readings/multiple-regression#:~:text=Five%20main%20assumptions%20underlying%20multiple,whether%20these%20assumptions%20are%20satisfied. Accessed 9 June 2023.

Zach. “The Five Assumptions of Multiple Linear Regression.” Statology, 16 Nov. 2021, www.statology.org/multiple-linear-regression-assumptions/.