**Performance Assessment: Task 1**

**A1. Research Question**

My research question for this performance assessment is, “What factors have an effect on the number of days a patient stays in the hospital during an initial visit?”

**A2. Goals**

The goal of this analysis is to determine whether there are variables within the provided dataset that directly correlate to the number of initial days.

**B1. Summary of Assumptions**

There are four assumptions of a linear regression model. The first assumption is linearity, meaning there is a linear relationship between the target and predictor variables. The second assumption is homoscedasticity, meaning there is equal variance of residuals. The third assumption is that there is no multicollinearity, meaning no strong relationship between the predictor variables. The fourth assumption is normality, meaning the residuals are normally distributed.

**B2. Tool Benefits**

For this performance assessment I chose to use Python. One benefit of using Python is that it is easy to read and build upon. Another benefit of Python is that the regression summary function includes more model evaluation metrics which makes it easier to evaluate the model.

**B3. Appropriate Technique**

Multiple linear regression is an appropriate technique for analysis because the target variable is continuous. A linear regression model allows me to analyze the relationship between initial days and my chosen predictor variables.

**C1. Data Cleaning**

Upon loading the data, I included ‘keep\_default\_na=False’ to keep values of NONE from getting converted to nulls. I renamed the 8 survey response variables using the .rename() function to better align the names with the data and to make my life easier. The first goal of my data cleaning process was to assess for duplicates, which I accomplished using the .duplicated() function. Once I determined there were no duplicates, I assessed for nulls using the .isnull() function. After confirmation of no null values, I created boxplots of all quantitative values to assess for outliers using the seaborn .boxplot() function. There were 3 variables that contained outliers. To remove the outliers, I replaced the outliers with nulls and imputed the nulls with the median value.

**C2. Summary Statistics**

The dependent variable for my model is initial days. The twelve independent variables for my model are: children, age, income, vitamin D levels, doctor visits, full meals eaten, total charges, additional charges, timely treatment, active listening, complication risk, and asthma. I performed summary statistics on the data set for the quantitative variables using the .describe() function in order to quickly get the characteristics of each variable, including the full range and average value. The function cannot be used to assess categorical variables, so I also used .value\_counts() function to obtain the summary statistics for complication risk and asthma.

A table with numbers and text

Description automatically generated

A table with numbers and text

Description automatically generated

A number of numbers and symbols

Description automatically generated with medium confidenceA black text on a white background

Description automatically generated

**C3. Visualizations**

A graph of a number of days

Description automatically generated

A graph with blue squares

Description automatically generatedA graph of age and age of patients

Description automatically generatedA blue graph with white text

Description automatically generatedA chart with a blue line

Description automatically generated with medium confidenceA graph of a doctor visits

Description automatically generatedA graph of a number of meals

Description automatically generatedA graph of blue and white lines

Description automatically generated with medium confidenceA graph of a number of patients

Description automatically generated with medium confidenceA graph of a survey response

Description automatically generatedA graph of blue rectangular objects

Description automatically generated with medium confidence

A graph of a number of blue bars

Description automatically generatedA graph with blue rectangles

Description automatically generated

A graph of numbers and lines

Description automatically generated with medium confidenceA blue and white chart with numbers

Description automatically generated

A blue and white chart

Description automatically generatedA chart of blue dots

Description automatically generated

A graph of a doctor visits

Description automatically generatedA graph of food items

Description automatically generated with medium confidence

A graph showing a number of patients

Description automatically generatedA blue and white chart

Description automatically generated

A graph of a survey response

Description automatically generatedA graph of blue dots

Description automatically generated

A diagram of a box plot

Description automatically generatedA graph of a box plot

Description automatically generated

**C4. Data Transformation**

According to the data dictionary provided with the data set, the survey responses were rated “on a scale of 1 to 8 (1 = most important, 8 = least important),” which felt counterintuitive. To make interpreting results of my analysis more straight forward, I chose to reverse the variables so that 1 = least important and 8 = most important using the .replace() function. I completed this prior to my univariate and bivariate analysis. I re-expressed my categorical variables as numeric. Since both variables contained ordinal data, I used the .replace() function again.

A screenshot of a computer program

Description automatically generated

**A close-up of a computer screen

Description automatically generatedC5. Prepared Data Set**

See “D208\_PA1\_MV\_clean” csv file for prepared data set.

**D1. Initial Model**

A screenshot of a computer

Description automatically generated**D2. Justification of Model Reduction**

After observing the regression results of my initial model, I tested for multicollinearity between the predictor variables by obtaining the variance inflation factor (VIF) for each. I removed the variable with the highest VIF value greater than 10, then reran the VIF values. I continued removing variables one by one until all VIF values were less than 10. From there, I used the backward stepwise elimination method to remove the least significant figures one by one until all the remaining predictor variables had p-values less than 0.05. This method ensured that I was left with a refined final model with statistically significant features.

**D3. Reduced Linear Regression Model**

A screenshot of a computer

Description automatically generated**E1. Model Comparison**

The probability f-statistic value for both my initial and final model is 0.00, implying that both regression models are meaningful. The adjusted r-squared value for my initial and final model are also very similar (0.982 vs. 0.981) which implies that the variables removed during model reduction did not add to the power of my regression model.

**E2. Output and Calculations**

**A graph with blue dots

Description automatically generated**A blue line with a red line

Description automatically generated

**A number with black lines

Description automatically generated with medium confidence**

**E3. Code**

See “D208 PA1 MV ipynb file for code.

**F1. Results**

The regression equation for my final model is:

This equation can be interpreted as follows: Keeping all things constant, one unit increase in total charges is associated with ~0.01 unit *increase* in initial days, one unit increase in additional charges is associated with ~0.00 unit *decrease* in initial days, and one unit increase in complication risk is associated with ~2.70 *decrease* in initial days.

The p-values for all my variables are below 0.05. The probability f-statistic for my model is also below 0.05. The adjusted r-squared value is also acceptable and close to 1.00. Due to these results, my model qualifies as statistically significant. My model is also practically significant, as hospital will want to know what factors are related to initial days to reduce initial days and decrease hospital spend.

**F2. Recommendations**

Based on my regression results and analysis, I would recommend building a new initial model that includes all variables in the data set. This model was a good start and showed that correlation does exist in the data. I would expand on this analysis to hopefully find more variables that can be included in the model. In the follow up analysis, I would lower the allowed VIF and see if that lowers the condition number of the final model.

**G. Panopto Demonstration**

Link to video included with submission.

**H. Sources of Third-Party Code**

**I. Sources**