**Performance Assessment**

**NBM3 TASK 1: LINEAR REGRESSION MODELING**

Bader Ale

Department of Information Technology, Western Governors University

D208 Exploratory Data Analysis

Professor William Sewell

July 2023

Table of Contents

[Part I: Research Question 3](#_Toc139710099)

[Part II: Method Justification 3](#_Toc139710100)

[Part III: Data Preparation 4](#_Toc139710101)

[Works Cited 5](#_Toc139710102)

# Part I: Research Question

For this performance assessment, my research question is: **Does either age, income, doctor visits, and days spent in the hospital, which caused a greater increase in the total charge to the patient**. This will help determine which of these factors are the biggest influence of a patient’s total charge during their stay. The goal is to help patients determine what has the biggest influence on their hospital charges.

# Part II: Method Justification

For this performance assessment, we will be using multivariate linear regression to determine the relationship between our explanatory variables and our target variable. In order to do this, we must first take into account certain assumptions. First of all, there has to exist a linear relationship between each of the explanatory variables (independent variables, X) and the target variable (dependent variable, Y). Secondly, we must determine if no multicollinearity exists between the explanatory variables. This is because the linear model will not be great if more than one variable affects the independent variable the same way. Thirdly, all explanatory variables must be independent of each other. Lastly, the residuals of the model should have a constant variance, or deviation from the true value, at every point in the model (Zach, 2021).

Python was chosen over R for two reasons. The first reason is the existing familiarity with Python; the only thing to learn was which libraries needed to be imported and the parameters. Even though R was writes specifically for statistical analysis and Python as more of a general programming language, many libraries have been created for Python to perform like R. The second reason is its flexibility and ease of reading/structuring.

Multiple linear regression (MLR) is the appropriate technique for analyzing the research question because we effectively want to predict a total charged to the patient depending on several other, independent factors. Multiple linear regression, or multivariate linear regression was used in lieu of regular, univariate linear regression because of the multiple explanatory variables considered.

# Part III: Data Preparation

# Works Cited

Zach. (2021, November 16). *The Five Assumptions of Multiple Linear Regression*. Retrieved July 3, 2023, from Statology: https://www.statology.org/multiple-linear-regression-assumptions/