NBM2 Task 1: Multiple Regression for Predictive Modeling

RathiPriyanka Srinivasan

Department of Information Technology, Western Governors University

D208: Predictive Modeling

Dr. Daniel Smith

March 4, 2021

A. Research Question

1. Is there an association between total charges and patient income, age, marital status, and gender?
2. Effective interventions to reduce readmission rates share features that involve a multifaceted approach that covers both inpatient and outpatient settings and delivery 'by dedicated transitional care personnel' (Kripalani et al., 2014). The goal of this research question is to help hospitals create community public health programs focusing on specific populations that are associated with higher hospital charges. This goal is reasonable within the scope of the data provided as it includes data on the total charges the patient has incurred (Scenario 1: Medical Readmission, 2021) and various demographic information about each patient that was admitted to the hospital, including: age, marital status, income, and gender (Scenario 1: Medical Readmission, 2021).

B. Method Justification

1. Assumptions associated with using the multiple regression model include assuming that: the variables are normally distributed, there is a linear relationship between the target and predictor variables, the predictor variables were measured without error, the variance of errors for the predictor variable is the same across all values of the target variable, and little or no multicollinearity (Osborne & Waters, 2002).
2. R will be the chosen tool for data preparation, analysis and manipulation. R is flexible to use as there is a large user community that provides packages for various data preparations, data manipulations, and generally has higher time performance for smaller datasets (Brittain et al., 2018). R can also be used for modeling scatterplots and visualizing the distribution of data (Kabacoff, 2017).

3. Multiple regression will be an appropriate analysis as it will give more insight to the influence the predictor has on the target variable. Multiple regression will also help forecast how much the dependent variable will be impacted if the independent variable changes. Additionally, multiple regression will also help the hospital administration predict patient trends and other future values. The data quality is suitable as it was extracted directly from a hospital chain's database (Scenario 1: Medical Readmission, 2021). There are also over 10,000 entries within the CSV file (Scenario 1: Medical Readmission, 2021) which allows for a more accurate R-Squared calculation.

C. Data Preparation

1. The data was scrubbed of all patient medical information. The TotalCharge column was kept as the target variable. The predictor variables include 4 patient demographic data columns including: 'age', 'income', 'marital', and 'gender' (Scenario 1: Medical Readmission, 2021). The categorical 'marital and 'gender variables will be converted to numerical by marking each relevant or 'Yes' entry as '1' and non-relevant or 'No' entry as '0'.

2. The mean is greater than the median for the target and all predictor variables, suggesting that the data is not symmetric and skewed to the right. The observations for the target variable and the predictor variables 'age' and 'income' are spread within approximately two standard deviations on each side of the mean. Observations for the predictor variables 'income' and 'marital' are spread within approximately four standard deviations

	age	income	marital	gender	totalcharge
nbr.val	1.000000e+04	1.000000e+04	1.000000e+04	1.000000e+04	1.000000e+04
nbr.null	0.000000e+00	0.000000e+00	7.977000e+03	5.232000e+03	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	1.800000e+01	1.540800e+02	0.000000e+00	0.000000e+00	1.938312e+03
max	8.900000e+01	2.072491e+05	1.000000e+00	1.000000e+00	9.180728e+03
range	7.100000e+01	2.070950e+05	1.000000e+00	1.000000e+00	7.242416e+03
sum	5.351170e+05	4.049050e+08	2.023000e+03	4.768000e+03	5.312173e+07
median	5.300000e+01	3.376842e+04	0.000000e+00	0.000000e+00	5.213952e+03
mean	5.351170e+01	4.049050e+04	2.023000e-01	4.768000e-01	5.312173e+03
SE.mean	2.063854e-01	2.852115e+02	4.017348e-03	4.994864e-03	2.180394e+01
CI.mean.0.95	4.045569e-01	5.590720e+02	7.874811e-03	9.790940e-03	4.274011e+01
var	4.259493e+02	8.134562e+08	1.613908e-01	2.494867e-01	4.754117e+06
std.dev	2.063854e+01	2.852115e+04	4.017348e-01	4.994864e-01	2.180394e+03
coef.var	3.856827e-01	7.043913e-01	1.985837e+00	1.047581e+00	4.104524e-01

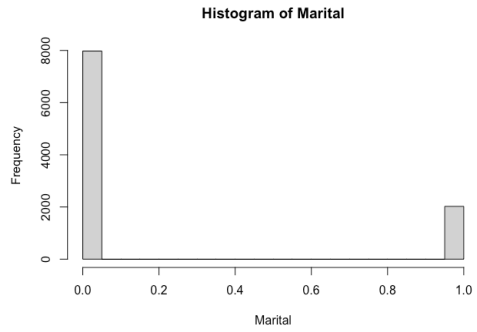
on each side of the mean. A higher standard deviation indicates a greater spread within the data. The minimum, maximum, and range will be relatively the same for the 'marital' and 'gender' variables as the predictor variables were converted from qualitative to quantitative data.

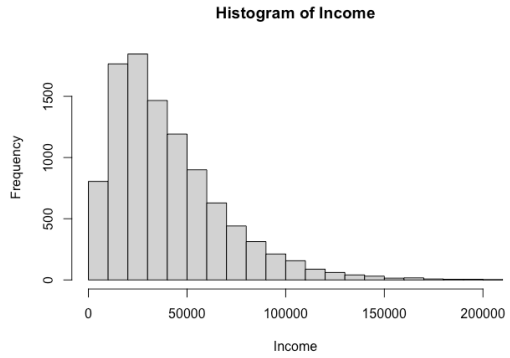
3. The following steps were used to prepare the medical dataset for analysis:
    a. Install relevant R packages including: tidyverse, openxlsx, plotly, zoo and psych
    b. Access the installed packages mentioned above
    c. Load the data from the provided CSV file (Scenario 1: Medical Readmission, 2021)
    d. Remove missing and null values
    e. Review summary statistics to identify which columns have missing or null values
    f. Create a subset of data ('tcharge') that includes columns relevant to the research question (Part A – Research Question, #1)
    g. Change all column names in 'tcharge' to lowercase
    h. In 'tcharge', convert all categorical predictor variables 'gender' and 'marital' to quantitative by converting all character inputs to numerical data

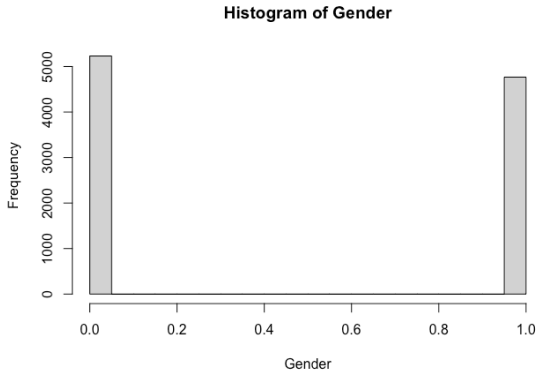
The screenshot of the annotated code in RStudio has been provided:

```
#install packages for cleaning and plotting data
install.packages("tidyverse")
install.packages("openxlsx")
install.packages("plotly")
install.packages("zoo")
install.packages("psych")

#access installed packages for data preparation and analysis
library(tidyverse)
library(openxlsx)
library(plotly)
library(zoo)
library(psych)
library(plyr)
library(e1071)
library(olsrr)
library(caret)

#load data from CSV file - patient medical data extracted from hospital database
medData <- read.csv("/Users/rathipriyankasrinivasan/Desktop/WGU/Term 2/Predictive Modeling/Medical Data/medical_clean.csv")

#remove missing values and nulls
complete.cases(medData)
#review which columns have null values - no NA/null values
summary(medData)

#create subset of columns relevant to research question - includes medical conditions and readmissions
tcharge <- subset(medData, select = -c(1:15, 20:40, 42:50))

#change all column names to lowercase
names(tcharge) <- tolower(names(tcharge))

#convert marital status to 0 if not married, if married convert to 1; then change character type to integer
tcharge$marital[tcharge$marital == "Separated"| tcharge$marital == "Divorced"| tcharge$marital=="Never Married"| tcharge$marital =="Widowed" ] <-0
tcharge$marital[tcharge$marital == "Married"] <- 1
tcharge$marital <- as.integer(tcharge$marital)

#convert gender to 0 if female, if male convert to 1, if nonbinary convert to 2; then change character type to integer
tcharge$gender[tcharge$gender == "Female"] <-0
tcharge$gender[tcharge$gender == "Male"] <-1
tcharge$gender[tcharge$gender == "Nonbinary"] <-0
tcharge$gender <- as.integer(tcharge$gender)
```

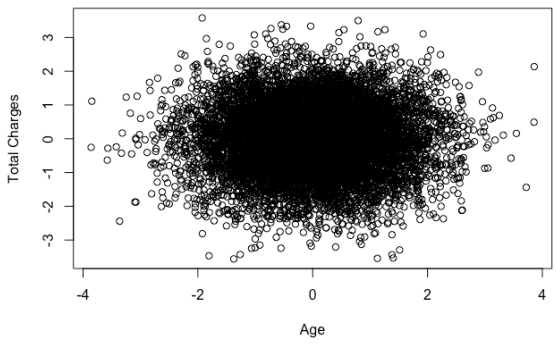
4. Univariate analysis for all predictor variables is provided below:

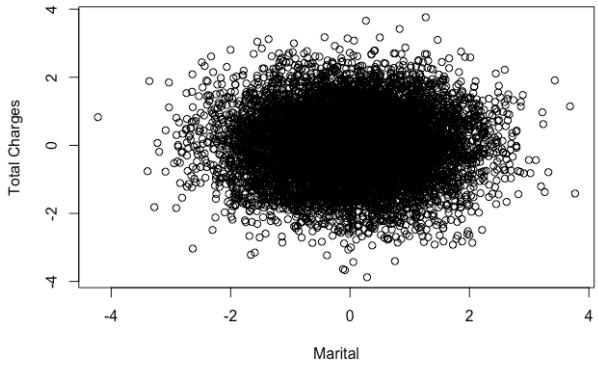


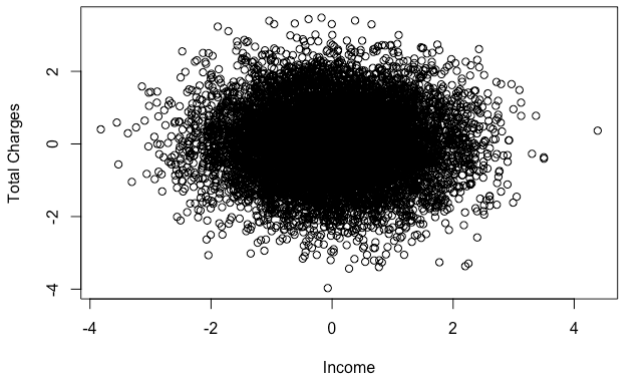


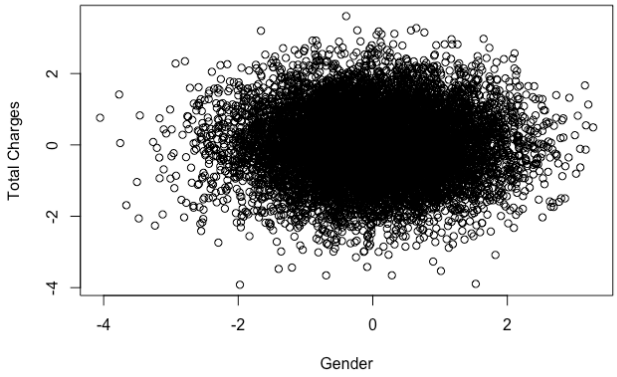


Bivariate analysis with the target variable and each predictor variable is included below:









D. Model Comparison and Analysis

1. Initial multiple regression model (all predictors):

   Total Charges = 1.77*Age – .001080*Income + 14.15*Marital + 21.90*Gender + 5248.

   Other variables that were included with the dataset were excluded based on relevancy, singularities, or a high level of collinearity. Residual error for the initial multiple regression model is 2180.

```
Residuals:
   Min     1Q Median     3Q    Max
 -3404  -2135   -117   2147   3946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.248e+03  7.243e+01  72.453   <2e-16 ***
age          1.770e+00  1.057e+00   1.674   0.0941 .
income      -1.080e-03  7.646e-04  -1.412   0.1579
marital      1.415e+01  5.429e+01   0.261   0.7943
gender       2.190e+01  4.366e+01   0.502   0.6159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2180 on 9995 degrees of freedom
Multiple R-squared:  0.0005171, Adjusted R-squared:  0.0001171
F-statistic: 1.293 on 4 and 9995 DF,  p-value: 0.2702
```

2. Forward step-wise regression will be the variable selection technique used for analyzing patient medical data. Advantages to using forward step-wise regression include processing predictor variables and adjusting the multiple regression model to choose the

best predictor variables from the available options (NCSS Statistical Software, 2021). Forward step-wise regression also allows for step by step analysis of the model by providing information about the quality of predictor variables and how they influence the multiple regression model as they are added in the variable selection technique (NCSS Statistical Software, 2021). Observing which demographic factors have greater influence on the patient total charges will be crucial facilitating community health programs and reducing readmission rates over the long run.

3. Reduced multiple regression model: Total Charges = 1.783*Age + 5216.769. The residual standard error for the reduced multiple regression model is 2180.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3397.1 -2135.5   -92.2  2151.3  3920.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5216.769     60.589  86.100   <2e-16 ***
tcharge$age    1.783      1.056   1.688   0.0915 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2180 on 9998 degrees of freedom
Multiple R-squared:  0.0002848, Adjusted R-squared:  0.0001848
F-statistic: 2.848 on 1 and 9998 DF,  p-value: 0.09151
```

4. Forward stepwise regression starts with no predictor variables in the model, then selects the predictor variable with the highest R-squared. At each step, predictor variables are then selected by observing the greatest influence on R-squared as it increases. The analysis is finished when none of the predictor variables are significant. In this case, significance is measure by the model evaluation metric, p-value = 0.05.

5. Output for forward stepwise regression:

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	age	3e-04	2e-04	1.3234	182126.1383	2180.1924

Confusion Matrix:

```
Overall Statistics

               Accuracy : 1
                 95% CI : (0.9996, 1)
    No Information Rate : 0.0161
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:
```

	Class: 18	Class: 19	Class: 20	Class: 21	Class: 22	Class: 23	Class: 24	Class: 25	Class: 26	Class: 27	Class: 28	Class: 29
Sensitivity	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000
Specificity	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000
Pos Pred Value	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000
Neg Pred Value	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000
Prevalence	0.0133	0.0137	0.012	0.0125	0.0141	0.0137	0.0144	0.013	0.0144	0.0135	0.0136	0.0124
Detection Rate	0.0133	0.0137	0.012	0.0125	0.0141	0.0137	0.0144	0.013	0.0144	0.0135	0.0136	0.0124
Detection Prevalence	0.0133	0.0137	0.012	0.0125	0.0141	0.0137	0.0144	0.013	0.0144	0.0135	0.0136	0.0124
Balanced Accuracy	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000

6. Annotated code for implementation of multiple regression models:

```
#multiple linear regression with ALL predictor variables
AllPred <- lm(tcharge$totalcharge ~., data=tcharge)
summary(AllPred)

#stepwise regression
ols_step_forward_p(AllPred, penter=.05)

#reduced multiple regression model
redPred <- lm(tcharge$totalcharge~tcharge$age, data=tcharge)
summary(redPred)
```

E. Data Summary and Implications

1. The reduced logistic regression equation Total Charges = 1.783*Age + 5216.769 shows that patient total charges is influenced by 1.783 unit of change in the patient's age. Older patients who are seeking medical assistance tend to incur higher total charges. Creating a public health initiative that focuses on older patients may prove to reduce patient total charges and thus, readmissions, over time. These results may lead to a better understanding of patient trends, but they do not completely answer all of the questions related to patient readmissions. One reason for this discrepancy could be that most hospital systems extract data from various EHR systems. A fragmented way of storing data can lead to misinterpretations as the original intent for each source the data was stored in may

differ (Adibuzzaman et. al., 2018). Another issue with analyzing healthcare data is the privacy concerns associated with sensitive patient information (Adibuzzaman et. al., 2018), which can limit the validation and reproducibility process often found within data analysis (Adibuzzaman et. al., 2018).

2. The given analysis suggests that developing public health programs that target older populations that are known for having a higher propensity for medical conditions will help reduce total patient charges and hopefully, readmission rates. For example, working with local nursing and assisted-living homes and advocating for meal programs that offer healthier alternatives could help make an impact on health conditions and reduce medical costs. Another public health initiative is help older community members develop patient efficacy by creating educational programs within local libraries or community centers.

References

Adibuzzaman, M., DeLaurentis, P., Hill, J., & Benneyworth, B. D. (2018). Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2017, 384–392.

Brittain, J., Cendon, M., Nizzi, J., & Pleis, J. (2018). Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance. *SMU Data Science Review*, 1(2).

NCSS Statistical Software. *Stepwise Regression*. (2021). https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

Kabacoff, R. (2017). Scatterplots. *Quick-R*. https://www.statmethods.net/graphs/scatterplot.html

Kripalani, S., Theobald, C. N., Anctil, B., & Vasilevskis, E. E. (2014). Reducing hospital readmission rates: current strategies and future directions. Annual review of medicine, 65, 471–485. https://doi.org/10.1146/annurev-med-022613-090415

Osborne, J. & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2). http://PAREonline.net/getvn.asp?v=8&n=2

Scenario 1: Medical Readmission. (2021). *Medical Data and Dictionary Files*. Western Governors University. https://access.wgu.edu/ASP3/aap/content/g9rke9s0rlc9ejd92md0.html