Multiple Linear Regression of Average Patient Admit Times for Diseases of the Circulatory System

James Tsai – Spring 2016, MSDS6372 – Experimental Statistics II

# Introduction

Per medical research, the diseases and disorders of the circulatory system account for the highest mortality rate as compared to other ailments. The causes are numerous, from hereditary and genetic factors to diet and lifestyle. For this study, we will examine the admission times of patients who were admitted to the hospital with circulatory diseases with at least a one night stay. Better understanding of the admission times potentially could improve patient care by more efficient utilization of hospital resources.

# Problem Statement

Develop a model based on the explanatory variables that can be used to predict admission times of patients with diseases related to the circulatory system.

# Constraints and Limitations

The analysis was performed on data acquired from a single hospital and therefore no causal inferences can be made between the response and explanatory variables. Furthermore, this analysis was performed on a limited set of data taken over a 3 months period between October 1, 2015 and December 31, 2015 and thus maybe missing data, which may have significant affects on, patient admit times. For example, seasonal and temperature influences potentially could change the conclusions significantly. To ensure HIPAA compliance, data that can uniquely identify a patient such as DOB or patient ID has also been removed. Finally, the variables in the data set may not have any important predictors and it is possible that there are much better explanatory variables that describe patient admission times.

# Data Set Description

The following table describes the variables used in the final analysis. The 67 rows of data represent 67 days aggregated from a larger set of data using a python script. The height and weight data was ultimately removed from this final data set as only about 20% of the original data set contained height and weight data of the patient.

|  |  |
| --- | --- |
| **Element** | **Description** |
| Average Admit Time | The average patient admission time normalized between 0 and 1 in a 24-hour day. For example a value of “0” would represent 12:00 AM and a value of “0.5” would represent 12:00 PM. Derived from the Admit Time field. |
| Average Age | The average age of the patients admitted during that day. Derived from the Age field. |
| Day | The date represented as a number from January 1, 1960. SAS defines day 1 as January 1, 1960. Thus, in order to properly represent the date, we have to count the number days from January 1, 1960. For example, 20362 would represent October 1, 2015. This information is necessary for the time series procedure to work properly and also to keep proper count since there were no activities on certain days. Excel function “DATEDIF” was used to assist in converting the date format. |
| Weekday | The weekday of the week such as “Thursday”. Derived from the Admit Date field |
| Patient Count | The total patient counts for that day. |
| Category 0 Count | The total patient counts that fall in category 0. Derived from ICD-10 Code field. |
| Category 1 Count | The total patient counts that fall in category 1. |
| Category 2 Count | The total patient counts that fall in category 2. |
| Category 3 Count | The total patient counts that fall in category 3. |
| Category 4 Count | The total patient counts that fall in category 4. |
| Category 5 Count | The total patient counts that fall in category 5. |
| Category 6 Count | The total patient counts that fall in category 6. |
| Category 7 Count | The total patient counts that fall in category 7. |
| Category 8 Count | The total patient counts that fall in category 8. |
| Category 9 Count | The total patient counts that fall in category 9. |

The following table describes the original data extracted from HL7 A01 message archives using a python script. The A01 message type contains information about inpatients that were admitted to the hospital system for one or more days. The extract contains 632 rows of data and is delimited by commas.

|  |  |
| --- | --- |
| **Element** | **Description** |
| Admit Time | The time of admission at the hospital in HH24: MM:SS format. |
| Admit Date | The date of admission at the hospital in MM/YY/DDDD format. |
| Height | The height of the patient in meters. |
| Weight | The weight of the patient in kilograms. |
| Age | The age of the patient at the time of admission. |
| ICD-10 Code | The primary medical code diagnosed at the time the patient was admitted to the hospital (i.e. I35.9) |

The following table describes how the ICD-10 codes are logically grouped for diseases of the circulatory system.

|  |  |  |
| --- | --- | --- |
| **Category Assignment for Data Analysis** | **Code Range** | **Description** |
| Category 0 | I00-I02 | Acute rheumatic fever |
| Category 1 | I05-I09 | Chronic rheumatic heart diseases |
| Category 2 | I10-I15 | Hypertensive diseases |
| Category 3 | I20-I25 | Ischemic heart diseases |
| Category 4 | I26-I28 | Pulmonary heart disease and diseases of pulmonary circulation |
| Category 5 | I30-I52 | Other forms of heart disease |
| Category 6 | I60-I69 | Cerebrovascular diseases |
| Category 7 | I70-I79 | Diseases of arteries, arterioles, and capillaries |
| Category 8 | I80-I89 | Diseases of veins, lymphatic vessels, and lymph nodes, not elsewhere classified |
| Category 9 | I95-I99 | Other and unspecified disorders of the circulatory system |

# Response Variable Analysis

The descriptive statistics for the Average Admit Time is shown on Figure 1. The mean of 0.45 for the average patient admission time translates to a time of 10:48 AM. From Figure 2, we can see that the histogram shows that the majority of Average Admit Times occur in the morning. The bar with the value of 0.30 corresponds to 7:20 AM. This is a reasonable result, as the hospital tends to be more active towards the earlier portion of the day.

|  |  |
| --- | --- |
|  |  |
| Figure 1. Descriptive Statistics for Average Admit Time. | Figure 2. Histogram for Average Admit Time. |

An examination of the time series plot of the average admit time over the 3-months period by day shows potential evidence of negative serial correlation as evidenced by the sharp peaks and low valleys in Figure 3. Again this is expected as not only are schedules on Saturday and Sundays at the hospital are not as structured, there are also usually fewer patients being admitted leading to less predictable averages. The box plots in Figure 4 shows evidence of this where Saturday has a skewed distribution of average admit times, and Sunday has a significantly higher mean admit time of 0.7 which translates to 4:48 PM. Furthermore, we observe in the time series plots there are more days where no patients were admitted for circulatory diseases in December as compared to October and November.

|  |  |
| --- | --- |
|  |  |
| Figure 3. Time Series for Average Admit Time. | Figure 4. Box Plot of Average Admit Time by Weekday. |

We will continue the analysis assuming that there is no serial correlation and the observations are independent. We will address the serial correlation in the second part of the analysis.

# Explanatory Variables Analysis and Screening

After examining the Pearson Correlation Coefficients, it is apparent that there is a high-degree of correlation between the patient count and specific diagnosis categories. This is to be expected, as the sum of all the categories equals the total patient count. We should exclude all categories if patient count is included in the model as there is a high-degree of collinearity between the patient count and categories. Average patient age also has some correlation with the average admission time. We will ignore the Day variable for now as it potentially involves serial correlation.

|  |
| --- |
|  |
| Figure 5. Pearson Correlation Matrix. |

We first include the initial explanatory variables of patient count, average age, cat3, cat5, and cat7. Figure 6 confirms that there are VIF (Variation Inflation Factors). After removing the cat3, cat5, and cat7, the VIF are at acceptable levels. Note that including just the patient count is a simpler model than including specific categories of the diseases. It also must be noted that 59% of the patients in this study fall under the category 5 of diseases, which explains the high correlation between patient count and category 5.

|  |  |
| --- | --- |
|  |  |
| Figure 6. Confirmation of high VIF due to collinearity between patient count and categories. | Figure 7. Removal of cat3, cat5, and cat7 yields VIF at acceptable levels below 10. |

# Model Selection

We can move forward with the following simplified model:

While this is already a simplified model, we will use LASSO regression analysis method to see if we can pare down the model even more. This step should be performed since a simpler model is more desirable and easier to explain. The LASSO regression analysis revealed no improvement in existing model and both yielded a R2 = 0.3319.

We need to address the assumptions for regression modeling at this point. In Figure 8, we see no evidence against normality in the residual histogram and QQ plot. However, the scatter plot of the residuals raises some concern for assumption of constant variance.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | |  |
| Figure 8. Histogram, QQ Plot, Scatter Plot of Residuals. | | | |
|  | |  | |
| Figure 9. Studentized Residuals. | | Figure 10. Cook’s D Distances | |

The studentized residuals and Cook’s D Distances in Figure 9 and Figure 10 identified observations 20 and 61, which are both outliers and high leverage. Upon further investigation, we notice that both these observations only contain one patient for that day. We note that observation 20 corresponds to the low point of average admit time in the time series plot of Figure 3. Observation 61 also has an unusually low average age due to the fact there was only one patient for that day. We will remove these two points from the data set, but keep in mind that these are valid data entries and further investigation should be performed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Observation** | **Avg\_Admit\_Time** | **Avg\_Age** | **Patient\_Count** | **Weekday** |
| 20 | 0.001563 | 87 | 1 | Saturday |
| 61 | 0.552604 | 8 | 1 | Thursday |

With the assumptions of normality and constant variance met, and outliers removed, we move forward with the final prediction model shown below and it is statistically significant at the level, with n=64, F=21.59, and p-value <0.0001 (See Figure 11). The result of the regression shows that on average we would expect a decrease of -0.01146 (equivalent of -16.5024 minutes) in the average admit time for each unit increase of patient count and a decrease of -0.00369 (equivalent of -5.3136 minutes) for each unit increase of average age. The model predicts that the more patients for a given day and the higher the average age will tend to make the admission times earlier.

|  |  |
| --- | --- |
|  |  |
| Figure 11. Final regression model and parameter estimates. | |

# Serial Correlation

In examining for serial correlation, we need to first address the assumptions that our data meets the criteria for a time series. From Wikipedia, a time series is a sequence of data points that:

1. Consists of successive measurements made over a time interval
2. The time interval is continuous
3. The distance in this time interval between any two consecutive data point is the same
4. Each time unit in the time interval has at most one data point

In this data set, the time interval cannot be considered continuous and thus the #2 assumption is violated. There are a total of 88 days between October 1, 2015 and December 28, 2015. However, only 67 days are accounted for in the data set since there are days where no patients were admitted with diseases of the circulatory system. We can confirm in Figure 3 and in the data set that there are days where no patients were admitted. For this reason alone, we cannot perform a serial correlation analysis, as our unit of measure of one day is not consistent.

# Conclusions

Although analyzing the data set of patients resulted in a regression model for predicting admission times, we must keep in mind that this model only explains around 41% of the correlation in admission times. It could very well be that there are other variables that predict the response variable more accurately. Furthermore, the data that was collected only covers a 3-month period, which may bias the results. For example, some researchers have found that heart attacks increase during the winter holiday season (ref: <http://www.webmd.com/heart/features/the-truth-behind-more-holiday-heart-attacks>). Some of the following steps maybe taken in the future could potentially increase our chances of making a better model:

* Collect a much larger data set, preferably at least one year, in order to account for any seasonal influences. Account for seasonality in the model.
* Better understand the nuances behind the admission of each type of circulatory disease category. For instance, certain diagnosis leads to a planned admission time (heart palpitations) vs. non-planned admission time (heart failure). Perhaps add planned and unplanned as another categorical variable.
* Further research the mechanics of how the hospital staff operates and model accordingly. Perhaps weekends should follow a different model, as the hospital is less busy and may not follow the same admission scheduling as during the weekdays.

This is a first attempt at creating a regression model for average admission times and perhaps in the future as more data becomes available, an enhanced model can be created which helps in improving the quality of services rendered to patients.

# APPENDIX

**/\* First Data Load \*/**

PROC IMPORT OUT=WORK.PROJECT1

DATAFILE="C:\Users\james\Documents\My SAS Files\9.4\MSDS6372\Project1\project1\_final.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

PROC PRINT DATA=WORK.PROJECT1; RUN;

**/\* Figure 1 - Descriptive Statistics for Average Admit Time \*/**

**/\* Figure 2 - Histogram for Average Admit Time \*/**

PROC UNIVARIATE DATA=WORK.PROJECT1;

VAR AVG\_ADMIT\_TIME;

HISTOGRAM/NORMAL(MU=EST SIGMA=EST);

RUN;

**/\* Figure 3 - Time Series for Average Admit Time \*/**

ODS GRAPHICS ON;

PROC TIMESERIES DATA=WORK.PROJECT1 PLOT=SERIES;

ID DAY INTERVAL=DAY;

VAR AVG\_ADMIT\_TIME;

RUN;

**/\* Figure 4 - Box Plot for Average Admit Time by Weekday \*/**

PROC GLM DATA=WORK.PROJECT1;

CLASS WEEKDAY;

MODEL AVG\_ADMIT\_TIME=WEEKDAY;

RUN;

**/\* Figure 5 - Pearson Correlation Matrix \*/**

PROC CORR DATA=WORK.PROJECT1; RUN;

**/\* Figure 6 - VIF Patient\_Count, Avg\_Age, Cat3, Cat5, Cat7 \*/**

PROC REG DATA=WORK.PROJECT1 outest=PROJECT1RESULT plots(label) = (rstudentbyleverage cooksd);

MODEL AVG\_ADMIT\_TIME=PATIENT\_COUNT AVG\_AGE Cat3 Cat5 Cat7 / AIC VIF CLI; \*CORRB INFLUENCE CLB;

RUN;

QUIT;

**/\* Figure 7 - VIF Patient\_Count, Avg\_Age \*/**

**/\* Figure 8 - (Histogram, QQ Plot, Scatter Plot) of Residuals \*/**

**/\* Figure 9 - Studentized Residuals \*/**

**/\* Figure 10 - Cook's D Distances \*/**

**/\* R-Squared = 0.3915 \*/**

PROC REG DATA=WORK.PROJECT1 outest=PROJECT1RESULT plots(label) = (rstudentbyleverage cooksd);

MODEL AVG\_ADMIT\_TIME=PATIENT\_COUNT AVG\_AGE / AIC VIF CLI; \*CORRB INFLUENCE CLB;

RUN;

QUIT;

ODS GRAPHICS ON;

PROC GLMSELECT DATA=WORK.PROJECT1

SEED=1 plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);

MODEL AVG\_ADMIT\_TIME=PATIENT\_COUNT AVG\_AGE/ selection=LASSO(choose=AIC stop=CV) CVdetails ;

RUN;

QUIT;

ODS GRAPHICS OFF;

**/\* Figure 11 - Final regression model and parameter estimates \*/**

**/\* R-Squared = 0.4105 \*/**

**/\* Second Data Load with Observations 20 and 61 Removed \*/**

PROC IMPORT OUT=WORK.PROJECT1\_2

DATAFILE="C:\Users\james\Documents\My SAS Files\9.4\MSDS6372\Project1\project1\_final\_adjusted.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

PROC REG DATA=WORK.PROJECT1\_2 outest=PROJECT1RESULT plots(label) = (rstudentbyleverage cooksd);

MODEL AVG\_ADMIT\_TIME=PATIENT\_COUNT AVG\_AGE/ AIC VIF CLI; \*CORRB INFLUENCE CLB;

RUN;

QUIT;