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

James Tsai – Spring 2015, 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. What are the admission times again?

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 **This is crossing the four square plot on page 9 … since the data are from a single hospital, any inference can only generalized to that hospital. It is the nature of the study (observational) that prevents causal inference.** 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. Ahhh …. Here it is! It is the actual time of admission.
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.

The UNIVARIATE Procedure
Variable: Avg_Admit_Time

Moments			
N	67	Sum Weights	67
Mean	0.45002643	Sum Observations	30.1517706
Std Deviation	0.15822723	Variance	0.02503586
Skewness	0.78507661	Kurtosis	1.09106192
Uncorrected SS	15.2214602	Corrected SS	1.65236661
Coeff Variation	35.1595428	Std Error Mean	0.01933053

Figure 1. Descriptive Statistics for Average Admit Time.

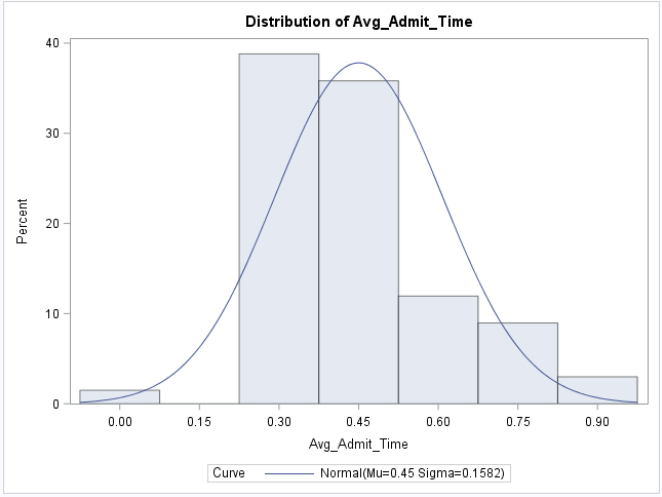

Figure 2. Histogram for Average Admit Time.

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.

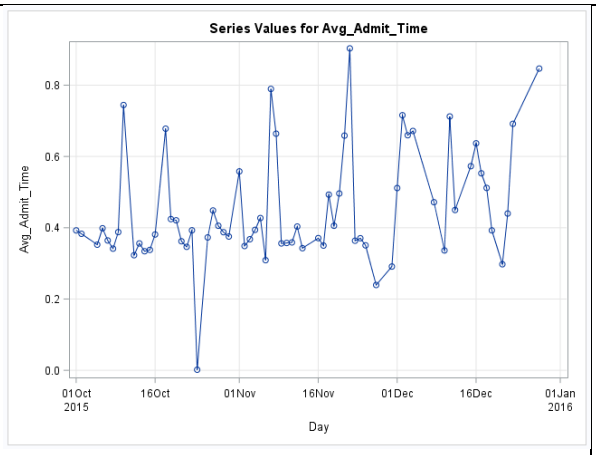

Figure 3. Time Series for Average Admit Time.

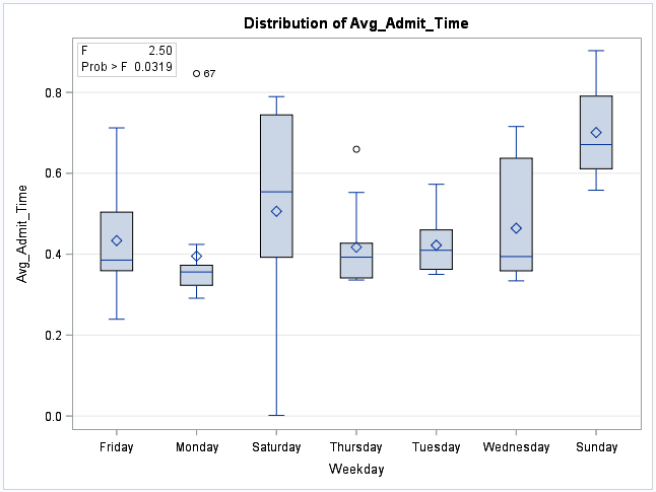

Figure 4. Box Plot of Average Admit Time by Weekday.

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.

Pearson Correlation Coefficients, N = 67
Prob > |r| under H0: Rho=0

	Avg_Admit_Time	Avg_Age	Patient_Count	Day	Cat0	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8	Cat9
Avg_Admit_Time	1.00000	-0.38615 0.0012	-0.52125 <.0001	0.35863 0.0029	.	-0.13551 0.2742	-0.08269 0.5059	-0.36755 0.0022	-0.05292 0.6706	-0.52416 <.0001	0.02123 0.8646	-0.40647 0.0006	-0.18623 0.1313	-0.05188 0.6767
Avg_Age	-0.38615 0.0012	1.00000	0.29044 0.0171	-0.29955 0.0138	.	0.19207 0.1195	0.12327 0.3203	0.18266 0.1390	0.03353 0.7877	0.31867 0.0086	-0.05146 0.6792	0.11157 0.3687	0.00883 0.9435	0.08478 0.4951
Patient_Count	-0.52125 <.0001	0.29044 0.0171	1.00000	-0.53015 <.0001	.	0.30580 0.0118	0.22112 0.0721	0.82842 <.0001	0.21254 0.0842	0.93099 <.0001	0.17524 0.1561	0.52330 <.0001	0.36141 0.0027	-0.02275 0.8550
Day	0.35863 0.0029	-0.29955 0.0138	-0.53015 <.0001	1.00000	.	-0.26877 0.0279	-0.07435 0.5499	-0.37192 0.0019	-0.06529 0.5996	-0.55304 <.0001	0.09386 0.4500	-0.19709 0.1099	-0.21302 0.0835	-0.22608 0.0658
Cat0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Cat1	-0.13551 0.2742	0.19207 0.1195	0.30580 0.0118	-0.26877 0.0279	.	1.00000	0.35869 0.0029	0.25770 0.0353	0.20702 0.0928	0.24850 0.0426	-0.10346 0.4047	0.03452 0.7816	-0.10733 0.3873	-0.07046 0.5710
Cat2	-0.08269 0.5059	0.12327 0.3203	0.22112 0.0721	-0.07435 0.5499	.	0.35869 0.0029	1.00000	0.22475 0.0675	-0.04420 0.7225	0.14802 0.2319	-0.03422 0.7834	0.17454 0.1578	-0.08297 0.5044	-0.03077 0.8048
Cat3	-0.36755 0.0022	0.18266 0.1390	0.82842 <.0001	-0.37192 0.0019	.	0.25770 0.0353	0.22475 0.0675	1.00000	0.29752 0.0145	0.59817 <.0001	0.20245 0.1004	0.41547 0.0005	0.21209 0.0849	-0.05737 0.6447
Cat4	-0.05292 0.6706	0.03353 0.7877	0.21254 0.0842	-0.06529 0.5996	.	0.20702 0.0928	-0.04420 0.7225	0.29752 0.0145	1.00000	0.10409 0.4019	0.03751 0.7631	-0.06425 0.6055	0.02341 0.8508	-0.04420 0.7225
Cat5	-0.52416 <.0001	0.31867 0.0086	0.93099 <.0001	-0.55304 <.0001	.	0.24850 0.0426	0.14802 0.2319	0.59817 <.0001	0.10409 0.4019	1.00000	0.00429 0.9725	0.46229 <.0001	0.31183 0.0102	-0.02199 0.8598
Cat6	0.02123 0.8646	-0.05146 0.6792	0.17524 0.1561	0.09386 0.4500	.	-0.10346 0.4047	-0.03422 0.7834	0.20245 0.1004	0.03751 0.7631	0.00429 0.9725	1.00000	-0.07968 0.5216	0.27957 0.0219	-0.15487 0.2108
Cat7	-0.40647 0.0006	0.11157 0.3687	0.52330 <.0001	-0.19709 0.1099	.	0.03452 0.7816	0.17454 0.1578	0.41547 0.0005	-0.06425 0.6055	0.46229 <.0001	-0.07968 0.5216	1.00000	0.02017 0.8713	0.02836 0.8198
Cat8	-0.18623 0.1313	0.00883 0.9435	0.36141 0.0027	-0.21302 0.0835	.	-0.10733 0.3873	-0.08297 0.5044	0.21209 0.0849	0.02341 0.8508	0.31183 0.0102	0.27957 0.0219	0.02017 0.8713	1.00000	0.11557 0.3517
Cat9	-0.05188 0.6767	0.08478 0.4951	-0.02275 0.8550	-0.22608 0.0658	.	-0.07046 0.5710	-0.03077 0.8048	-0.05737 0.6447	-0.04420 0.7225	-0.02199 0.8598	-0.15487 0.2108	0.02836 0.8198	0.11557 0.3517	1.00000

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. I am not surprised to see the high correlation between the "Cat" variables and Patient_count as they are describing the same idea. I would maybe consider: 1) In the following analysis, try a model with all Cat variables and without the patient_count variables. This will allow you to keep the category information in the model. 2) Let LASSO and or Stepwise selection have a crack at it and see what it comes up with.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.78467	0.10060	7.80	<.0001	0
Patient_Count	1	-0.00500	0.01709	-0.29	0.7710	59.13103
Avg_Age	1	-0.00371	0.00157	-2.35	0.0218	1.11518
Cat3	1	0.00337	0.02199	0.15	0.8787	11.70673
Cat5	1	-0.00571	0.01817	-0.31	0.7546	27.86778
Cat7	1	-0.05305	0.03202	-1.66	0.1027	1.44497

Figure 6. Confirmation of high VIF due to collinearity between patient count and categories.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.78730	0.09934	7.93	<.0001	0
Patient_Count	1	-0.00975	0.00233	-4.18	<.0001	1.09213
Avg_Age	1	-0.00375	0.00156	-2.40	0.0193	1.09213

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:

$$\mu_{Avg\_Admit\_Time} = \beta_0 + \beta_1(Patient\_Count) + \beta_2(Avg\_Age)$$

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 R^2 = 0.3319. Interesting.

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.

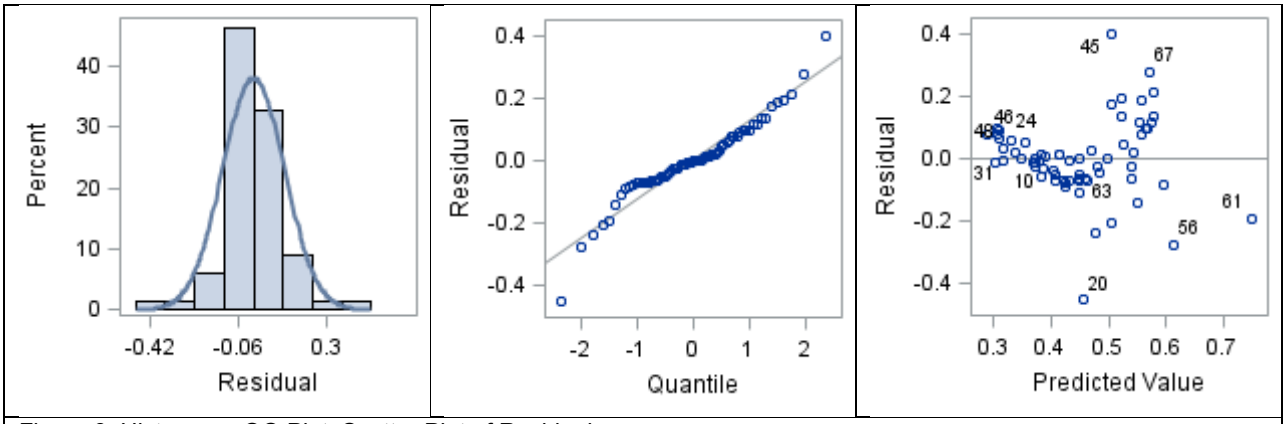

Figure 8. Histogram, QQ Plot, Scatter Plot of Residuals.

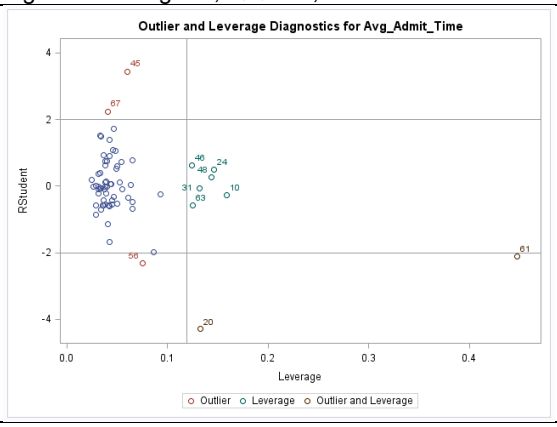

Figure 9. Studentized Residuals.

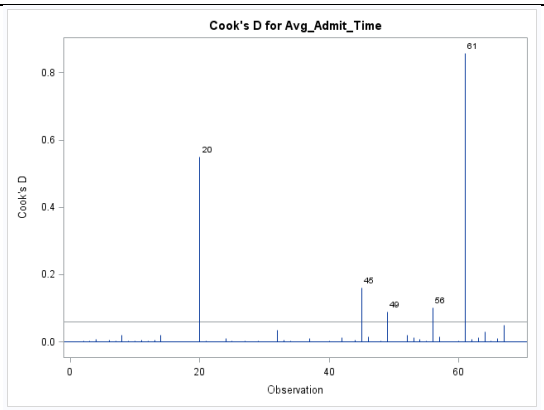

Figure 10. Cook's D Distances

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 $\alpha = 0.05$ 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. Need to have confidence intervals here … but I love the interpretation. 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. (Will tend to make the estimated ***average*** admission time earlier.) There seems to be a lurking variable here. Maybe hospitals with low patient counts specials in procedures that tend to make people come in later. This is why I think it might be useful to drop patient_count and keep the cat variables that hold the same information just broken down by category. It very will might be that the actual variables that is driving the change in admit time hasn't been recorded in this data set … but we should still keep looking.

Also, external or at least internal crossvalidation would be nice to see as well.

$$\boldsymbol{Avg\_Admit\_Time = 0.80971 + (-0.01146 * Patient\_Count) + (-0.00369 * Avg\_Age)}$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.59072	0.29536	21.59	<.0001
Error	62	0.84816	0.01368		
Corrected Total	64	1.43888			

Root MSE	0.11696	R-Square	0.4105
Dependent Mean	0.45535	Adj R-Sq	0.3915
Coeff Var	25.68625		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.80971	0.12570	6.44	<.0001	0
Patient_Count	1	-0.01146	0.00215	-5.33	<.0001	1.12197
Avg_Age	1	-0.00369	0.00198	-1.87	0.0664	1.12197

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. I think you are ok here … time itself is a continuous measurement and you have recorded it on 67 different days. 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. You have a point with the violation of #3. **This is an interesting problem with missing data. FYI A more sophisticated model would impute values for these missing data based on a separate regression possibly based on patient number, day of the week, age, etc.**

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.

Hi James,

This was an interesting problem and you obviously spent a lot of time on the project. I think you did a good job with the fundamentals of building a model, looking at residuals, and interpreting the statistics. The following are few things that I think were either omitted or could have been added to make a stronger analysis. Given the current analysis you have an 84% but can get up to a 92% by doing the following:

1. **Fit a different model dropping patient_number and adding some or all of the categorical patient number variables. Compare this to the model you did fit with AIC and CVPRESS.**
2. **Provide confidence intervals for the coefficient estimates in your final model.**
3. **Break the weekday variable into a two level variable for M-Th and call it 'weekday' and F-Sat and call it 'weekend' and see if that categorical variable is significant in predicting admission time (see if it is significant in the model.)**
4. **Not necessary for the grade but would be interesting … perform an external cross validation of your final model.**

1) I went ahead and dropped the patient_number parameter and added back cat3, cat5, and cat7 variables. These 3 variables had the highest correlation to the average admit time in the Pearson Correlation Matrix. After running the LASSO selection, avg_age, cat5, and cat7 were retained. The prediction model is statistically significant at the $\alpha = 0.05$ level, with n=64, F=17.02, and p-value <0.0001. The R^2 value is 0.4557, and is higher than the original R^2 value of 0.4105 where patient_number parameter was used in lieu of cat5 and cat7 in the model.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.78078	0.09898	7.89	<.0001	0
Avg_Age	1	-0.00370	0.00156	-2.37	0.0210	1.11505
Cat3	1	-0.00260	0.00812	-0.32	0.7499	1.61906
Cat5	1	-0.01084	0.00463	-2.34	0.0223	1.83340
Cat7	1	-0.05575	0.03043	-1.83	0.0718	1.32461

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	AIC	CV PRESS
0	Intercept		1	-177.0664	1.7197
1	Cat5		2	-187.6569	1.3478
2	Avg_Age		3	-186.8633	1.3017
3	Cat7		4	-200.2564*	1.2242*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	Cat3	1.2334	>	1.2242

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.79613	0.12361	6.44	<.0001	0
Avg_Age	1	-0.00355	0.00195	-1.82	0.0733	1.16248
Cat5	1	-0.01423	0.00366	-3.89	0.0003	1.44144
Cat7	1	-0.06010	0.02609	-2.30	0.0247	1.26081

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.65568	0.21856	17.02	<.0001
Error	61	0.78321	0.01284		
Corrected Total	64	1.43888			

Root MSE	0.11331	R-Square	0.4557
Dependent Mean	0.45535	Adj R-Sq	0.4289
Coeff Var	24.88455		

2) Provided below are the 95% confidence intervals for the coefficients, which includes average age, cat5, and cat7.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.79613	0.12361	6.44	<.0001	0.54896	1.04330
Avg_Age	1	-0.00355	0.00195	-1.82	0.0733	-0.00745	0.00034547
Cat5	1	-0.01423	0.00366	-3.89	0.0003	-0.02156	-0.00691
Cat7	1	-0.06010	0.02609	-2.30	0.0247	-0.11227	-0.00793

3) I ended up creating a variable called "DAY1" for weekdays (Mon-Fri) and weekends (Sat-Sun). The reference value 0 was set for weekdays. Interestingly, not only was incorporating this parameter significant in the model (p-value = 0.0013), the R^2 increased to 0.5429 after adding this parameter to model. There is significant evidence to confirm that the average admission times on weekends differ than that of weekdays. We should also keep this variable in the final model because it intuitively makes sense that the hospital operates on a different schedule during the weekends. Our new regression model is:

$$Avg\_Admit\_Time = 0.83493 + (-0.00498 * Avg\_Age) + (-0.00857 * Cat5) + (-0.05431 * Cat7) + (0.14789 * DAY1)$$

The result of the regression shows that on average we would expect a decrease of 0.00498 (equivalent of -7.1712 minutes – CI (-12.4992, -1.8432)) in the average admit time for each unit increase of average age, a decrease of 0.00857 (equivalent of -12.34 minutes – CI (-23.2128, -1.4544)) for each unit increase of cat5 patient count, a decrease of 0.05431 (equivalent of 78.2064 minutes – CI (-147.816, -8.5968)) for each unit increase of cat7 patient count, and a increase of 0.14789 (equivalent of 212.96 minutes - CI (87.0768, 338.8464)) if it's on a weekend (Saturday/Sunday).

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.78119	0.19530	17.82	<.0001
Error	60	0.65769	0.01096		
Corrected Total	64	1.43888			

Root MSE	0.10470	R-Square	0.5429
Dependent Mean	0.45535	Adj R-Sq	0.5124
Coeff Var	22.99280		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.83493	0.11479	7.27	<.0001	0.60533	1.06454
Avg_Age	1	-0.00498	0.00185	-2.69	0.0091	-0.00868	-0.00128
Cat5	1	-0.00857	0.00378	-2.27	0.0269	-0.01612	-0.00101
Cat7	1	-0.05431	0.02417	-2.25	0.0283	-0.10265	-0.00597
DAY1	1	0.14789	0.04370	3.38	0.0013	0.06047	0.23531

4) I will need to get permission to collect the data for this year, but it would definitely be interesting to see how the model holds up as we collect at least a full year's of data.

SAS Code Used For Corrections

```
/* Q1 - Updated Model */
PROC REG DATA=WORK.PROJECT1 outest=PROJECT1RESULT plots(label) = (rstudentbyleverage cooksd);
   MODEL AVG_ADMIT_TIME=AVG_AGE Cat3 Cat5 Cat7 / 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=AVG_AGE Cat3 Cat5 Cat7 / selection=LASSO(choose=AIC stop=CV) CVdetails  ;
RUN;
QUIT;
ODS GRAPHICS OFF;

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=AVG_AGE Cat5 Cat7 / AIC VIF CLI; *CORRB INFLUENCE CLB;
   RUN;
QUIT;


/* Q2 Confidence Intervals for Coefficients */
PROC REG DATA=WORK.PROJECT1_2;
   MODEL AVG_ADMIT_TIME=AVG_AGE Cat5 Cat7 / CLB;
   RUN;
QUIT;

/* Q3 Weekday/Weekend significance */
DATA WORK.PROJECT1_3;
   SET WORK.PROJECT1_2;
   IF Weekday = "Monday" or Weekday = "Tuesday" or Weekday = "Wednesday" or Weekday = "Thursday"
or Weekday = "Friday" then DAY1=0; ELSE DAY1=1;
RUN;

PROC PRINT DATA=WORK.PROJECT1_3; RUN;

PROC REG DATA=WORK.PROJECT1_3;
   MODEL AVG_ADMIT_TIME=AVG_AGE Cat5 Cat7 DAY1/CLB;
   RUN;
QUIT;
```

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;
```