A Multivariate Regression Analysis of Hospital Stays in a Nosocomial Infection Control Data

Dr. Julian Allagan
Mentor
Elizabeth City State University
adallagan@ecsu.edu

Jessica Hathaway
Researcher
Elizabeth City State University
hathawayjessica18@gmail.com

Matthew Hill
Researcher
Elizabeth City State University
Mghill242@students.ecsu.edu

Heaven Tate
Researcher
Elizabeth City State University
Tate.heaven@gmail.com

Lilshay Rogers
Researcher
Elizabeth City State University
Lilshay1999.lr@gmail.com

Dr. Linda B. Hayden
Principal Investigator
Elizabeth City State University
haydenl@mindspring.com

***Abstract*— In this report the team developed and analyzed several linear regression models to predict hospital stays (or length of Stay) of patients in the US using the SENIC project data from CDC-Atlanta. The team examined several potential exploratory variables and their relations with the response variable "Stay", with the goal of determining what leading factors influenced the length of stay of patients in this Nosocomial (hospital acquired) infection control data. In particular, our report aimed at answering the following: given the data, what leading factors help explain the hospital stays of patients in US? In at least one model, the team found that Age and Regions influenced the variable "Stay" the most.**

Keywords— Data Analysis, Linear regression, Nosocomial

I. INTRODUCTION

Due to budgetary concerns, hospitals, insurance company and other healthcare providers often seek ways to become more efficient. According to a report by Gonzalez [1], in 2011, hospital in-patient expenses accounted for almost one-third of all healthcare expenditures compared to prescription medicine which accounted for about one-fifth of total medical expenses in the United States. Few years later, in a brief statistical report [2], Weiss and Elixhauser noted that, in 2012, there were 36.5 million hospital stays in the US with an average length of stay of 4.5 days and with an average cost of $10,400 per stay. Because some of these costs are often bared by hospitals and insurance providers, many hospitals are now looking for means to optimize their capacity by discharging patients on time and reduce patients' readmission rate. Our objective in this report is to find some statistical relationship between hospital stays and some meaningful hospital-related variables and for this reason, from the research the team chose to explore a Nosocomial (hospital-acquired) Infection Control (SENIC project [3]) data. The team notes that the SENIC project original goal was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. Because of its wide range of indicators or variables (12) its size (113 hospitals) and the reliability of its source (Center for Disease Control, Atlanta, GA [3]), the team deemed the SENIC project data valuable for our analysis. Thus, in this report the team developed and analyzed a statistical regression model to predict hospital stays of patients in the US using the SENIC project data. In particular, our report aimed at answering the following: given the data, what leading factors help explain the hospital stays of patients in US? All data exploration and analysis in our report were solely based on the SENIC data.

II. METHODOLOGY

SENIC PROJECT DATA

This data set consists of a random sample of 113 hospitals. For each hospital, the following 12 variables (See Table 1, below) is provided in the order they appeared in the statistics textbook by Kutner et al. **[4]** (see Appendix C, page 1348). The data set contains no missing value although some scaling was found necessary for the purpose of our analysis.

Variable Name	*Description*
ID number	1-113
Length of stay (Stay)	Average length of stay (in days) of all patients in hospitals
Age (Age)	Average age (in years) of patients
Infection risk (Risk)	Average estimated probability (in percent) of acquiring infection in hospital
Routine culturing ratio (Culturing)	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
Routine chest X-ray ratio (X-ray)	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
Number of beds (Beds)	Average number of beds in hospital during study period
Method school affiliation (Affiliation)	1 = Yes, 2 = No
Region (Region)	Geographic region of the country, where: 1= NE, 2 = NW, 3= S, 4 = W
Average daily census (Census)	Average number of patients in hospital per day during study period
Number of nurses (Nurses)	Average number of full-time equivalent nurses during study period (number full time plus one half the number part time)
Available facilities and services (Services)	Percent of 35 potential facilities and services that are provided by the hospital

Table 1: Variables and their description in the SENIC project data

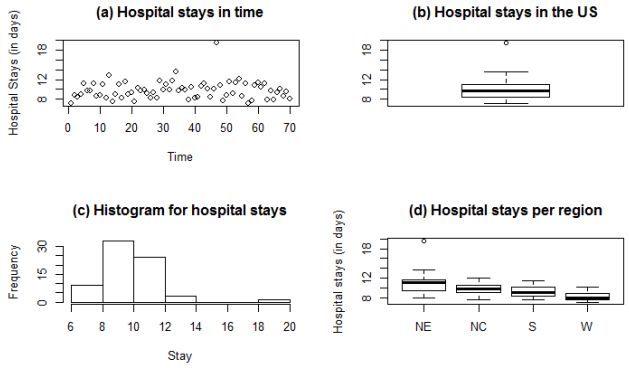
As mentioned in the introduction, for our analysis, the original SENIC project data was split into two data: The training data that the team called ENIC contains observations 1-70 and the testing data which the team called ENIC2 contains the remaining observations (71-113) from SENIC project. Our basic plots, model selection, and diagnostics were done based on ENIC while ENIC2 was used to help validate our final proposed model. The plots and the statistics were generated using the free statistical software R and the codes used are made available in the Appendix.

EXPLOITATIONS OF SOME USEFUL VARIABLES

As previously mentioned, our data exploration was based on the ENIC data, which contains the first 70 observations of the SENIC project data. Here, the team explore some basic information about some of the variables that were later selected in section 2.3 to help build our model.

"RESPONSE" STAY

Plot 1 summarizes some basic information about Stay. In plot 1(a), the team noticed that the hospital stays of patients is independent of time, suggesting that neither time nor a particular event (such as a disease outbreak) influenced the data record. Plot 1(b) indicates that the median of hospital stays is about 10 days. The minimum number of hospital stays is about 7 days and the maximum number of hospital stays is about 14 days. About half of the recorded number of hospital stays is between 9 and 11 days. Plot 1(c) shows that the distribution of the hospital stays is relatively normal with an average close to 10 days and the highest frequency on the record is from the group with 8-10 days of hospital stays. In Plot 1(d), the team see that hospital stays differ from region to region with the highest recorded being in Northeast (NE) and the lowest recorded being in the West (W). Also, the median hospital stays in Northeast (NE), North central (NC), South (S) and West (W) is about 11, 10, 9 and 8, respectively. Finally, each plot shows one hospital stay being an outlier (about 20 days) which comes from a hospital record in NE (See Plot 1(d)).

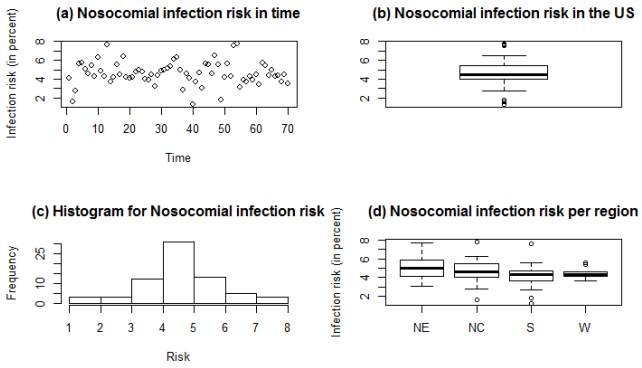


Plot 1: Hospital stays data plots from ENIC data

"PREDICTOR" RISK

Likewise, as discussed in the previous section, the team examined Plot 2 for Risk. Plot 2(a) suggests that time was not a factor in the risk of infection record. Plot 1(b) indicates that the median risk of hospital-acquired infection in the US is about 4.5%. The minimum percentage of Risk is about 3% and a maximum of about 7%. Several outliers were spotted at around 1% and around 8%. Plot 1(c) shows that the distribution of Risk is normal with an average of 4.6% in US. In Plot 1(d), the team see that Risk varies with region with the

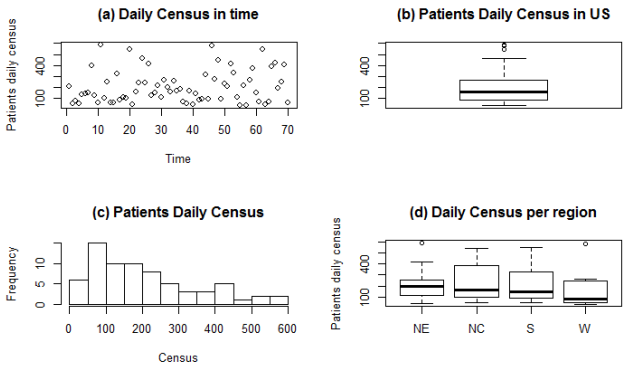
highest records coming from the Northeast (NE) and the lowest records coming from the south (S). The south and the west have the lowest median Risk with west showing the smallest range in Risk. Some outliers can be spotted for all regions except for the northeast.



Plot 2: Hospital-acquired Infection Risk data plots from ENIC data

"PREDICTOR" CENSUS

Plot 3(a) also indicates that the average daily census record is independent of time. The median daily census is about 160 patients in Plot 3(b) with a couple of potential outliers being reported, one from NE and the other from W as shown in Plot 3(d). The minimum daily census is in the low 40's while the maximum daily census is about 500 patients. Further, Plot 3(c) shows that the distribution of Census is right skewed and the Census average is about 205 patients (from computation). The most recorded Census is between 50-100 patients. In Plot 3(d), the team see that Census varies from region to region with the highest median being in Northeast (NE) and the lowest median being in the West (W).



Plot 3: Average patients' daily census data plots from ENIC data

MODEL SELECTION PROCESS

Due to the large amount of candidate variables, the team relied primarily on the Mallows' Cp selection criterion as a model-building technique to arrive at a list of "good" models. This selection method also allows to test whether or not there is a potentially "good" model that includes the variable Stay. Based on their Cp values (in general, the smaller, the better the model), the team selected the top three subsets of "good" variables and recorded them in Table 2. Clearly, each subset includes "Stay", the variable the team wish to be for our model "response" variable.

Variables	***Cp***
Stay, Age, Risk, X-ray, Region, Census, Nurses	4.298737
Stay, Age, Risk, Region, Census, Nurses	4.576915
Stay, Risk, Region, Census, Nurses	4.970154

Table 2: Top 3 best subset of "predictors" and their Cp values

For each subset of variables shown in Table 2, the team used Stay as "response" and the remaining variables as "predictors" and fit a linear regression model for each case. The team recorded the adjusted R2 values (in general, the higher, the better) for each model as shown in Table 3.

Variables for Each Sub model	***R^2a***
Stay, Age, Risk, X-ray, Region, Census, Nurses	0.5115
Stay, Age, Risk, Region, Census, Nurses	0.493
Stay, Risk, Region, Census, Nurses	0.4736

Table 3: Top 3 sub-models and their R2a value

Because the adjusted R2 values for the models are fairly close to one another, the team decided to look at their prediction sum of squares (PRESS) statistic as illustrated in Table 4, below.

Variables for Each Sub model	***PRESS***
Stay, Age, Risk, X-ray, Region, Census, Nurses	151.0711
Stay, Age, Risk, Region, Census, Nurses	150.7552
Stay, Risk, Region, Census, Nurses	148.574

Table 4: Top 3 sub-models listed based on PRESS

Since the team had no objective reason to go with a larger model, the team selected the model with the least PRESS value (in general, the smaller, the better) and the least amount of predictors (for the simplicity of the model) and called it Model 2. The team note here that PRESS is used not only for regression model comparison but also to assess a model's predictive capability, which the team later discussed in section 4.

The team then decided to check the correlation factor for each pair of predictors in Model 2. Below is the correlation matrix with listed correlation coefficient (r) between each pair.

	Stay	Risk	Region	Census	Nurses
Stay	1	0.53	-0.48	0.27	0.20
Risk		1	-0.20	0.25	0.25
Region			1	-0.035	-0.038
Census				1	0.93
Nurses					1

Table 5: Correlation matrix for model 2

The team note that the result in the Table 5 is also supported by the scatterplot from our R-output (see R-1 in Appendix). From this table, the team can observe that Stay is fairly related to Risk and Region. Also, Census appeared to be strongly related (r =0.93) to Nurses and the correlations among the remaining pairs of predictors is much weaker. For instance, Region is hardly related Nurses (r =-0.038) and Census (r =-0.035) while, Risk is barely related to any other predictors (r <0.3).

Due to the strong relation between Census and Nurses, the team decided to look at the model regression output to determine the significance of each variable. The regression output (See R-2 in Appendix) indicated that Census contributed significantly to the model (p-value= 0.03) while Nurses barely (p-value=.09), if both predictors were kept in the model. For this reason, the team dropped the predictor Nurses from Model 2 and denoted the resulting model, Model 3, whose variables are Stay, Risk, Region, and Census.

After dropping Nurses from Model 3, the variable Census became less significant (p-value=0.12) for our model (See R-3 in Code, for the regression output). In other words, despite the fact the variable Risk and Region, each, achieved a statistical significance for the model, the variable Census did not, at alpha =0.05. Due to a strong relation between Nurses and Census from our earlier observation, the team had reason to believe that the number of nurses available in a hospital certainly affected not only the census registration but more importantly the length of stay of patients in the hospital; the less staff or nurses a hospital has the less likely they will keep patients longer in the hospital. So, the team decided to replace Census in the model by Nurses, and found that the predictor Nurses contribution to the model was significantly smaller (SSR=1.4) compared to the contribution of the variable Census (SSR=4.9). See ANOVA in R-3, for details. The team made a final decision (albeit with some doubt) to keep Census while being mindful that it "reflects" Nurses.

CHECKING FOR ADDITIONAL POTENTIAL VARIABLES FOR MODEL 3

The team decided to add some two-way interaction variable to Model 3 and check their significance to the model. Recall that interaction variables are often constructed by multiplying together the corresponding relevant variables. The team anticipated some structural multicollinearity issues to occur as a result of adding this new predictor to our model. As a remedy to this issue for our data, Kutner et al3 suggested "centering" the predictors, which is done by simply subtracting the mean of the predictor values in the data set from each predictor value. The team proceeded to "center" the predictors Risk, Census and denoted their corresponding centered variables Risk* and Census*, respectively, and the predictor Risk*Census* which depicts the interaction between "centered" Risk and "centered" Census. The team chose to add asterisks (*) to each variable to make a distinction between the new variable and its raw or original form.

The regression output (See R-4 in Appendix, or Table 6) of our new or updated Model 3 showed that there was a strong evidence (p-value= 0.029) for two-way interactions between Risk* and Census*. Due to the hierarchy principle2, from now on, the team will no longer be concerned with the significance of the variable Census* in our model since the interaction variable Risk*Census* is significant. Moreover, the team tested other two-ways interaction variables from the predictors listed in Model3—(See R-5 in Appendix) It was clear from these results that no other two-way interaction variable would be significant for the model, whether or not they are included as a single interaction term or along with other interaction terms. In fact, the team noticed with the interaction terms, some of the regions appeared significant while others were not.

Also, the team checked whether or not a quadratic term for either Risk or Census would be significant. The regression outputs (See R-6 in Appendix) indicated that neither quadratic term was significant (p-value>.1) for our model (with or without the interaction term). Our final Model 3, includes therefore the following predictors, Risk*, Census*, Risk*Census*, Region and a summary of its regression output is shown in Table 6 and more details can be obtained with R-4 in Appendix.

	Estimate	*Standard Error*	*p-value*
Intercept	10.620649	0.317679	< 0.001
Risk*(x1)	0.781185	0.156062	< 0.001
Census*(x2)	0.001769	0.001170	0.135545
Risk*Census*(x1x2)	0.002806	0.001258	0.029308
Factor (Region) 2	-0.959352	0.437287	0.031941
Factor (Region) 3	-1.244376	0.444660	0.006804
Factor (Region) 4	-2.223747	0.561232	< 0.001

Table 6: Summary of Model 3 Regression result where *variable=(variable-mean (variable))

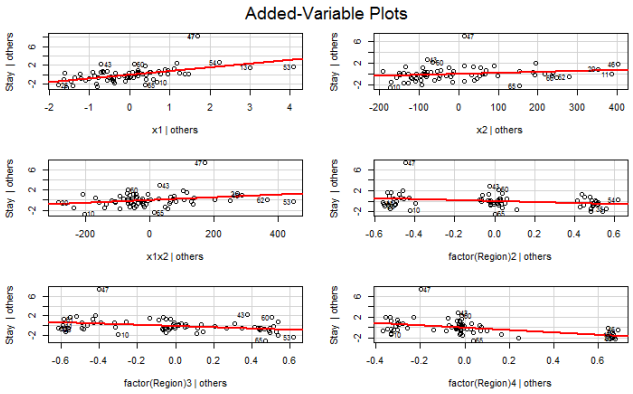
As the team settled on Model 3 (with the interaction term), the team decided to check whether or not it meets basic linear

regression model assumptions.

III. DIAGNOSTICS

SOME USEFUL PLOTS

From the regression output, the variables Risk* (or x1), Census* (or x2), Risk*Census* (or x1x2), Region explained about 50% of the reduction in variation of the average length of stay of the patients in the hospital. The model appeared to be significant (F-value=10.48, p-value<.001). The finding from the added-value plot (See Plot 4, below) also supported the parameters estimates (See Table 6, for instance) of our linear regression model and the significance of each predictor for the model. For instance, from Table 6, the estimated regression coefficient for Risk* (x1) was positive (.78) and was certainly the largest. From Plot 4, looking at the first of the six plots (Top left), Stay increases as x1 increases given the effect of other predictors being held constant and the slope of line illustrating their relationship was the steepest.



Plot 4: Added-value plots for Model 3

VARIANCE INFLATION

The variance inflation factor for each predictor of model, as illustrate in Table 7, is quite small, hence the team considered it satisfactory for the model.

	GVIF	*Df*	*GVIF^(1/(2*Df))*
Risk*(x1)	1.367873	1	1.169561
Census*(x2)	1.085810	1	1.042022
Risk*Census*(x1x2)	1.266721	1	1.125487
factor(Region)	1.118933	3	1.018906

Table 7: Variance Inflation Factor for Model 3

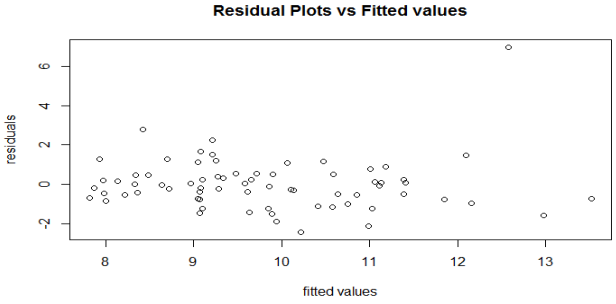
RESIDUALS

The residual vs fit plot (See Plot 5) shows that the residuals appeared randomly scattered (in a "horizontal band") around the 0 line. This suggests that the relationship between Stay and the indicated variables in the model is in fact linear and the variances of the error terms are constant. Furthermore, the normal probability plot (See Plot 6) suggests that the error terms are normally distributed. Thus, the team concluded that Model3 met the required assumptions for a linear regression model and its fitted regression function is given by **Stay=11+0.8Risk*+0.002Census*+0.003Risk*Census*-0.96NC-1.2S-2.2W**, where

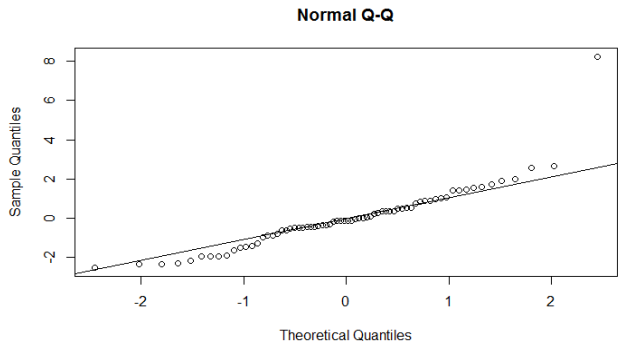
Stay is the average hospital stays, Risk* is the centered infection risk, Census* is the centered Census, and NC, S, and W, represent the north central, the south and the west regions, respectively.

SOME BASIC INTERPRETATIONS

Recall that from ENIC, the average risk of infection is 4.6% and the average daily census is 205 patients. Our model shows that if either the infection risk is at 4.6% or the daily census is about 205 patients, then the team can expect the hospital stays to be about 10.6 days, 9.6 days, 9.4 days, and 8.4 days, for the northeast, north central, south and west regions, respectively. Moreover, when considering, say, the effect of infection risk (Risk) on the hospital stays (Stay), the team can expect that for each percent increase in the infection risk beyond the average 4.6%, the hospital stays increase/decrease by 0.78 +0.003Census*, -0.96 +0.003Census*, -1.2 +0.003Census*, -2.2+0.003Census*, for the northeast, the north central, the south, and for the west regions, respectively. It is clear that due to the interaction between Risk and Census, the effect of Risk on Stay depends on Census and likewise the effect of Census on Stay depends on Risk which can easily be verified.



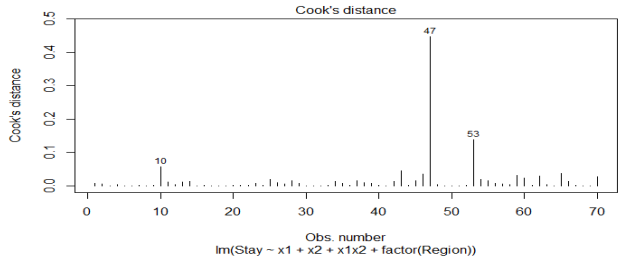
Plot 5: Residual vs Fitted



Plot 6: Normal Q-Q

OUTLIERS AND INFLUENTIAL POINTS

As it can be observed in the previous two plots, there is some indication that ENIC contains at least one outlier and previous plots (See Plots 1-3, for instance) also support this finding. Further analysis (See R-7 in Appendix or Plot 7, for instance) also suggested that case #47 is a potential high leverage point. To see whether or not it had any significant influence on our model, the team proceeded to drop case #47 from ENIC. The new regression output (See R-8 in Appendix) indicated that there was no significant change in the parameter estimates, not even in the R2 value. Further, the team had no objective reason to believe that this observation was recorded in error nor do the team think it was not representative of the SENIC project original data. For these aforementioned reasons, the team decided to keep this hospital record with our final model, Model 3 and proceeded to test its predictive ability.



Plot 7: Cook's distance measure for influential observations in the ENIC data

IV. MODEL VALIDATION

INTERNAL AND EXTERNAL VALIDATION

Here the team check the predictive capability of our final model, Model 3 (or Training) whose variables are Stay, Risk*, Census*, Risk*Census*, and Region. As stated in the introduction, for testing, the team used ENIC2 data, whose entries are observations 71-113 from SENIC. For this reason, the team calculated the prediction errors which are the differences between the actual response values in ENIC2 and the predictions by Model 3, and summarized the predictive ability of Model 3 by the mean squared prediction error (MSPR). The team found (see R-9, in Appendix) that MSPR= 3.057, giving an indication of how well Model 3 will predict in the future. When compared it its Mean Square Error (MSE=1.889), the team found no significant difference between these two values. For a further test on the predictive capability of Model 3, the team decided to fit the regression model identified by our model selection process to ENIC2, our testing dataset. The team compared the estimated regression coefficients and the estimated standard errors of both models. Below (Table 8) is a summary.

	Training		***Testing***	
	Estimate	*Standard Error*	*Estimate*	*Standard Error*
Intercept	10.620649	0.317679	10.7060034	0.4008130
Risk*(x1)	0.781185	0.156062	0.7070530	0.1500185
Census*(x2)	0.001769	0.001170	0.0021519	0.0013864
Risk*Census*(x1x2)	0.002806	0.001258	0.0025843	0.0009835
factor(Region)2	-0.959352	0.437287	-1.3894892	0.4970425
factor(Region)3	-1.244376	0.444660	-1.7047812	0.4766940
factor(Region)4	-2.223747	0.561232	-2.6082614	0.5811453

Table 8: Comparison between Model 3 and Testing data: Estimated regression coefficients

These estimates appear to be reasonably similar across all estimated parameters. This is good evidence that Model 3 can be applied to data beyond ENIC. The team also observed that, the standard error of model Training (1.37) was fairly closed to that of Model Testing (1.11). For further details of these values, see R-10 in Appendix. In addition, the team decided to compare Model 3 to a formerly proposed model in the next section.

A KNOWN OR PROPOSED MODEL

The SENIC project data had been assigned to some members of this group project at first as an exercise. Several models the were proposed for predicting the average length of stay of patients in the SENIC data. The work involved some basic data plots, model comparison and some analysis of the statistical significance of the predictors of each proposed model. One final model was recommended which, for the purpose of this report, the team called Model B. First, here is some basic information about Model B and how it was

obtained.

Model B had the following predictors: Stay (as response variable), Age, Risk, Census. This model was based on the SENIC project data (with observations 1-113) and it was built as follow: The team started with two sub-models and for convenience the team called them Model B1 and Model B2. Model B1 had predictors Age, Risk and Services for the response Stay, and Model B2 had predictors Beds, Risk and Services for the response Stay. The coefficient of multiple determinations (R2) for these models turned out to be the same (about 0.32) and it was decided to consider a starting model which the team called Model B0. Model B0 had Age and Risk as predictors for the response Stay. Then, for each of the following predictors, Routine Culturing, Routing Chest X-ray, Census and Nurses, the team computed and compared their marginal contribution to Model B3. Variable Census contributed the most (about 13%) to Model B0. The final model B had Age, Risk and Census as predictors for the response Stay

MODEL COMPARISON

The team note here that, after getting our final Model 3, the team went back and randomly selected seven other potential model that were initially identified through our original Mallows' Cp selection process. The team did not see any evidence that they were better; their R2 values were significantly lower with higher MSE and, in many cases, one or more predictors from each model were statistically insignificant at $\alpha = 0.05$. For any further interest or details on these fitted regression models, please see R-11 in Appendix. The team also noted that, based on its coefficient of multiple determinations, Model B also appeared to be better compared to the seven models the team previously referred to. The regression output for Model B and its ANOVA result can be obtained from R-12 in Appendix.

So the team decided to compare Model 3 to Model B on the basis of their overall predictive capability. The results are presented in Table 9. These values, particularly PRESS and Predicted R2 (R2pred =1-PRESS/SSTO) indicated that Model 3 is a better predictive model of Stay, even though Model 3 was built on fewer observations (70 observations) compared to Model B (113 observations). Observe that for both models, R2 is not quite high relative to R2pred, an indication that neither model is "overfitting", i.e., using more predictors (whether useful or not) that they actually needed.

	Model 3	***Model B***
R^2 *(the higher the better)*	.50	.49
Adjusted R^2 *(the higher the better)*	.452	.396
Standard Error *(the smaller the better)*	1.37	1.48
PRESS *(the smaller the better)*	147.7434	269.6774

Table 9: Some indicators of the predictive ability of Model 3 vs Model B

V. CONCLUSION AND DISCUSSION

In order to find what predictors help explain patients hospital stays (or Stay) in the SENIC project data, Mallows' cp selection criterion along with adjusted R2 and their PRESS information were used as the main model-building techniques. The SENIC data was divided into two groups, ENIC (for training) and ENIC2 (for testing). Our selection (on ENIC) process began with the following pool of variables Stay, Age, Risk, Culturing, X-ray, Beds, Affiliation, Region, Census, Nurses, Services. Our process initially yielded 3 sub-models from which the team determined based on the adjusted R2 and the PRESS values, one single "best" model. This model contained the response variable Stay and the predictors Risk*, Census*, Risk*Census*, and Region where (*) indicate that these variables are "centered". The team tested and found that there was a significant linear regression relationship between the response variable Stay and these predictors. Our analysis of the model suggested that about 50% of the variation in hospital stays (Stay) could be explained by the infection risk (Risk*), the average number of daily census (Census*) and the geographic region (Region). The team were surprised to see that Age was not a significant predictor for Stay even though the team would naturally think that older patients are more likely to have longer hospital stays than younger ones. Another surprise from our data exploration of Stay indicated that the average hospital stays in the US is almost 10 days, which is unusually high (more than twice the reported average in a recent report [2] as mentioned in the introduction). In fact, the team actually checked the mean of hospital stays for the SENIC project data and observed a similar value (9.6 days). The team had no knowledge of the cause of this significant difference in this mean value (compared to an earlier report [2]) and certainly our model did not intend to determine the cause of hospital stays. One unusually high hospital stays case was reported by one hospital (ID 47) in the northeast and yet it had no unduly influence on our model so the team included their record also in our final model analysis. A reasonable argument could be made to delete this hospital record, say, the team wanted to limit our final analysis to hospitals for which the Stay is less than 14 days (according to ENIC data). In the end the team did not think such limit on the training data ENIC was needed and could increase our prediction error, particularly when the validation data (ENIC2) had several records of 15 days or more of hospitals stays. To further determine the predictive capability of our model, the team chose to compare it to a known model (from a past project) and the team found some evidence that our final model not only was better, but also shows some signs that it could be applied to data beyond ENIC.

Throughout our model selection process, the team

had strived not to exclude any important predictor (to avoid an "underfitted" model) while keeping the model simple with the least possible amount of predictors (to avoid an "overfitted" model). The team did not think that adding any new or replacing an existing predictor would improve the overall significance of our final model and yet, the team are mindful that there is no "perfect" model. A further analysis with the goal of arriving at an improved linear regression model (compared to Model 3) would perhaps be to test other two-way interactive variables (using other predictors) to see whether or not there is a possible reduction in the overall prediction error of the model. Also, although the team found that a second-order regression model was not appropriate for our model with the selected predictors, the team could not rule out such order if one considers other predictors for the model and perhaps more data is needed to create a better predictive model.

ACKNOWLEDGMENTS

The 2018 Mathematics team would like to thank the CERSER principal investigator Dr. Linda B. Hayden and our mentor Dr. Julian Allagan.

REFERENCES

[1] **Gonzalez JM.** National Health Care Expenses in the U.S. Civilian Noninstitutionalized Population, 2011. MEPS Statistical Brief No. 425. Rockville, MD: Agency for Healthcare Research and Quality, 2013. http://meps.ahrq.gov/data_files/publications/st425/stat425.pdf

[2] **Weiss AJ** (Truven Health Analytics), Elixhauser A (AHRQ). Overview of Hospital Stays in the United States, 2012. HCUP Statistical Brief #180. October 2014. Agency for Healthcare Research and Quality, Rockville, MD. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb180-Hospitalizations-United-States- 2012.pdf."NSF - National Science Foundation". *Nsf.gov*. N.p., 2017. Web. 22 Mar. 2017.

[3] **Special issue, The SENIC Project**," American Journal of Epidemiology 111 (1980), pp. 465-653. Data obtained from Robert W. Haley, M.D. Hospital Infections Program, Center for Infectious Disease, Center for Disease Control, Atlanta, Georgia 30333.

[4] **Kutner, Nachtsheim, Neter and Li**, Applied Linear Statistical Methods 5ed., McGraw-Hill, 2004.

Appendix

R-codes for the Analysis of Hospital Stays in a Nosocomial Infection Control data

```
#reading a txt table with header
ENICall<-read.table("hospital-all.txt",header=TRUE)
#attach(ENICall)
#splitting data
ENIC = ENICall[1:70,]
ENIC #verifying the content of the data
#creating variables
Stay<-ENIC$Stay
Age<-ENIC$Age
Risk<-ENIC$Risk
Culturing<-ENIC$Culturing
Xray<-ENIC$Xray
Beds<-ENIC$Beds
Affiliation<-ENIC$Affiliation
Region<-ENIC$Region
Census<-ENIC$Census
Nurses<-ENIC$Nurses
Services<-ENIC$Services
######################plots of Stay
par(mfrow=c(2,2))
plot(Stay, main="(a) Hospital stays in time", xlab="Time",ylab="Hospital Stays (in days)")
boxplot(Stay, main="(b) Hospital stays in the US")
hist(Stay, main="(c) Histogram for hospital stays")
boxplot(Stay[Region==1], Stay[Region==2],Stay[Region==3],Stay[Region==4],main="(d) Hospital stays per region",
names=c("NE","NC","S","W"),ylab="Hospital stays (in days)")
mean(Stay) #9.8
summary(Stay)
IQR(Stay)
######################plots of Risk
par(mfrow=c(2,2))
plot(Risk, main="(a) Nosocomial infection risk in time", xlab="Time",ylab="Infection risk (in percent)")
boxplot(Risk, main="(b) Nosocomial infection risk in the US")
hist(Risk, main="(c) Histogram for Nosocomial infection risk")
boxplot(Risk[Region==1], Risk[Region==2],Risk[Region==3],Risk[Region==4],main="(d) Nosocomial infection risk per
region", names=c("NE","NC","S","W"),ylab="Infection risk (in percent)")
mean(Risk) #4.6%
summary(Risk)
IQR(Risk)
######################plots of Census
par(mfrow=c(2,2))
plot(Census, main="(a) Daily Census in time", xlab="Time",ylab="Patients daily census")
boxplot(Census, main="(b) Patients Daily Census in US")
hist(Census, main="(c) Patients Daily Census")
boxplot(Census[Region==1], Census[Region==2],Census[Region==3],Census[Region==4],main="(d) Daily Census per region",
names=c("NE","NC","S","W"),ylab="Patients daily census")
mean(Census) #205
summary(Census)
IQR(Census)
#verifying the content of the table
#names(ENIC)<-c("ID","Stay","Age", "Risk", "Culturing", "Xray", "Beds", "Affiliation", "Region", "Census", "Nurses",
"Services")
pairs(Stay~Age +Risk+Culturing+ Xray +Beds+ Affiliation+Region+Census+Nurses+Services,main="Scatterplot matrix for
SENIC project data")
#Stay Risk    Census Nurses
```

```
#model selection
library(leaps)
X<-ENIC[,3:12]
Y<-ENIC[,2]
#-Mallows Cp  is finding a single best model
bestmodel=leaps(X,Y, names=names(ENIC)[3:12], method="Cp")
bestmodel$which[ order( bestmodel$Cp ), ]
#To print Cp criterion in increasing order
sort( bestmodel$Cp ) #smallest cp values 4.298737  4.576915  4.970154
#   Age  Risk Culturing  Xray  Beds Affiliation Region Census Nurses Services
#6   TRUE  TRUE     FALSE  TRUE FALSE       FALSE   TRUE   TRUE   TRUE    FALSE
#5   TRUE  TRUE     FALSE FALSE FALSE       FALSE   TRUE   TRUE   TRUE    FALSE
#4  FALSE  TRUE     FALSE FALSE FALSE       FALSE   TRUE   TRUE   TRUE    FALSE

#candidate variables: Age  Risk
#modelA: Age+Risk+X-ray+Region+Census+Nurses
#modelB: Age+Risk+Region+Census+Nurses   modelA minus X-ray
#modelC: Risk+Region+Census+Nurses      modelA minus X-ray minus Age

##############################################################################################
#some test models
modelA<-lm(Stay~Age+Risk+Xray+factor(Region)+Census+Nurses)
modelB<-lm(Stay~Age+Risk+factor(Region)+Census+Nurses)#modelA minus X-ray
modelC<-lm(Stay~Risk+factor(Region)+Census+Nurses)#modelA minus X-ray minus Age
summary(modelA) #Multiple R-squared:  0.5115
summary(modelB) #Multiple R-squared:  0.493
summary(modelC) #Multiple R-squared:  0.4736

##############################################################################################
#picking the better submodel
sum((modelA$residuals/(1-hatvalues(modelA)))^2) #PRESSpA 151.0711
sum((modelB$residuals/(1-hatvalues(modelB)))^2) #PRESSpB 150.7552
sum((modelC$residuals/(1-hatvalues(modelC)))^2) #PRESSpB 148.574
#####################################################################
#R-1-checking correlation
pairs(Stay~Risk+Region+Census+Nurses,main="Scatterplot matrix for Model 2")
# Census appears to be strongly correlated with Nurses
#upon checking the marginal contribution of Census and Nurses, the team decided to drop Nurses and keep
c1 <- data.frame(Stay, Risk, Region,Census, Nurses)  #correlation coefficients
cor(c1)
#R-2#################################
modelC<-lm(Stay~Risk+Census+Nurses+factor(Region))#modelA minus X-ray minus Age
summary(modelC) #Adj Rsq 0.4412  Rsq 0.4736,
#####MODEL2#############################################################################################
#R-3##############
model2_c<-lm(Stay~Risk+factor(Region)+Census)
summary(model2_c) # Rsq Adj: .4179  Multiple R-squared:  0.460
anova(model2_c) #Census SSR 4.97
model2_n<-lm(Stay~Risk+factor(Region)+Nurses)
summary(model2_n) #Rsq Adj .4 and Rsq .44
anova(model2_n) # Nurses SSR 1.46
############MODEL3#############################################################
#centering and renaming variables
x1<-(Risk-mean(Risk))
x2<-(Census-mean(Census))
x1x2<-x1*x2
x3<-factor(Region)
```

```
x1x3<-x1*x3
x2x3<-x2*x3
#R-4 ##########Model 3 with interactive term: FINAL MODEL
model3<-lm(Stay~x1+x2+x1x2+factor(Region))
summary(model3)
#R5##############checking other interactive terms
model4<-lm(Stay~x1+x2+x1x2+factor(Region)+x1*factor(Region)) #Risk and Region interaction term
summary(model4)
model5<-lm(Stay~x1+x2+factor(Region)+x1x2+x2*factor(Region)) #Census and Region interaction term
summary(model5)
model6<-lm(Stay~x1+x2+factor(Region)+x1x2+x2*factor(Region)+x1*factor(Region)) #three interaction terms
summary(model6)
model7<-lm(Stay~x1+x2+factor(Region)+x1*factor(Region)) #Risk and Region interaction term only
summary(model7)
model8<-lm(Stay~x1+x2+factor(Region)+x2*factor(Region)) #Census and Region interaction term only
summary(model8)
#R6#####testing quadratic terms for model 3
x1sq<-x1^2  #Risk sq
x2sq<-x2^2   #Census sq
model3A<-lm(Stay~x1+x2+x1x2+factor(Region)+x1sq)
summary(model3A)
model3B<-lm(Stay~x1+x2+x1x2+factor(Region)+x2sq)
summary(model3B)
model3C<-lm(Stay~x1+x2+factor(Region)+x1sq)
summary(model3C)
model3D<-lm(Stay~x1+x2+factor(Region)+x2sq)
summary(model3D)
########MODEL ASSUMPTIONS ##############################################################################
par(mfrow=c(1,1))
plot(model3$fitted,model3$residuals, main="Residual Plots vs Fitted values", xlab="fitted values", ylab="residuals")
plot(model3$residuals, type='l',main="Residual Plots vs Fitted values", xlab="fitted values", ylab="residuals")
qqnorm(model3$residuals, main="Normal Q-Q")
qqline(model3$residuals)
########FITTED MODEL PARAMETERS#########################################################
#pairs(Stay~Risk+Region+Census+X1X3,main="Scatterplot Matrix for Model 2")
summary(model3)$coefficients[,1] #extracting coefficients
#Avearage Stay=10.620648626 +0.781184591Risk 0.001769013Census -0.959352302(Region)2 -1.244375962(Region)3 -
2.223746724(Region)4 -0.011152462Census+ 0.002805531Risk*Census
####################################added variable plot
avPlots(model3, id.n=5, id.cex=.8)
coef(model3)
#the plots support the linear model. From the estimated parameters of the regression output,
#Y decreases as X1 ( similarly X2, X3) increases, given other variables are being held constant.
#This is also true for each added value plot
+-########################################## variance inflation factors
  #check VIFs ARE GOOD
  vif(model3)
  summary(model3) #Multiple R-squared:  0.4996,      Adjusted R-squared:  0.452
anova(model3) #MSE 1.9  Residual standard error: 1.373
#F-statistic: 10.48 on 6 and 63 DF,  p-value: 4.877e-08

#R-7Influential points#####################
summary(influence.measures(model3)) #possible influential points are 47, 53
round(dffits(model3),2) #confirms 47 with dffits=2.69
## Assessing Outliers
boxplot(Stay, main="Average length of stay of the patients in the hospital", ylab="Length of stay")
outlierTest(model3) # Bonferonni p-value for most extreme obs  #47 is an outlier
```

```
qqPlot(model3, main="QQ Plot") #qq plot for studentized resid
leveragePlots(model3) # leverage plots
plot(model3, which=4, cook.levels=cutoff)  #confirms #47 as influential
#R-8 model 3 with deleted row 47 ########################################MODEL 3 with deleted row 7
summary(model3) #Multiple R-squared:  0.503,         Adjusted R-squared:  0.4565
anova(model3) #MSE 1.9  Residual standard error: 1.384
#F-statistic: 10.66 on 6 and 63 DF,  p-value: 3.82e-08
##VALIDATION MODEL#############################################################################
ENIC2<-ENICall[1:70,]
#testData<-ENICall[1:70,]
#newtestData<-data.frame(Stay, Risk,Region,Census, data=testData)
#####EXTENDING COLUMNS#########################################################
testData<-ENIC2
#attach(testData)
Y1<-testData$Stay
X1<-testData$Risk
X2<-testData$Census
testDatA<-(round(within(testData,centered_Risk<-(X1-mean(X1))),1))#including centered risk in testData
testDatB<-(round(within(testDatA,centered_Census<-(X2-mean(X2))),1)) #including centered Census in testData
testDatB
testData3<-(round(within(testDatB,Risk_Census<-(centered_Census)*(centered_Risk)),1)) #including centered Census in
testData
####CREATING INDICATORS#########################################################
#creating NE indicator variable
testData3$NE<-0 #assigning 0 to NE
testData3$NE[testData3$Region==1]<-1 #NE indicator
#creating NW indicator variable
testData3$NW<-0 #assigning 0 to NE
testData3$NW[testData3$Region==2]<-1 #NW indicator
#creating S indicator variable
testData3$S<-0 #assigning 0 to NE
testData3$S[testData3$Region==3]<-1 #S indicator
#creating W indicator variable
testData3$W<-0 #assigning 0 to W
testData3$W[testData3$Region==1]<-1 #W indicator
testData_update<-testData3
testData_update
#verifying the categorical data entries
v<-testData_update$Stay
v1<-testData_update$centered_Risk
v2<-testData_update$centered_Census
v1v2<-testData_update$Risk_Census
v3<-testData_update$Region
r1<-testData_update$NE
r2<-testData_update$NW
r3<-testData_update$S
r4<-testData_update$W
#R9########## computing MSPR
X.1v <- cbind(rep(1,length(v)),v1,v2,v1v2,r2,r3,r4)
X.1v
b.1t<-coef(model3) #beta hats
(yhat.1v <- X.1v %*% b.1t) #fitted values
(MSPR.1v <- sum((v-yhat.1v)^2)/length(v)) #MSPR 3.057 MSE 1.889  is generalizable
training<-model3
testing<-lm(v~v1+v2+v1v2+factor(v3))
#R-10#################################
anova(training) #MSE:1.886  PRESS=147.7434  SSTO=237.475  R-sq pred=1-PRESS/SSTO=.378
```

```
#compared to for Model B#PRESS= 269.6774 SSTO=409.211 R-sq pred=1-PRESS/SSTO =.341
anova(testing)  #MSE: 1.241
summary(training) # RSE#1.37 R-sq .5
summary(testing) # RSE# 1.11 R-sq: .7
coef(training)
#(Intercept)            x1            x2          x1x2 factor(Region)2 factor(Region)3
#10.620648626     0.781184591     0.001769013     0.002805531    -0.959352302    -1.244375962
#factor(Region)4
#-2.223746724
coef(testing)
# (Intercept)          v1          v2        v1v2  factor(v3)2  factor(v3)3  factor(v3)4
#10.706003392  0.707053034  0.002151941  0.002584325 -1.389489215 -1.704781173 -2.608261439
############################some extra analysis
hist(Stay)
#R-11##########################testing other submodels
trainingData<-ENICall[1:70,]
Training1<-lm(Stay~x1+factor(Region))
summary(Training1)
anova(Training1)#0.4392, Adjusted R-squared:  0.4047 MSE 2.049
Training2<-lm(Stay~x1+Culturing+factor(Region))
summary(Training2) #culturing not significant
anova(Training2)#0.4392, Adjusted R-squared:  0.4047 MSE 2.049
Training3<-lm(Stay~x1+Xray+factor(Region))
summary(Training3) #Xray not significant
anova(Training3)#0.44,     Adjusted R-squared:  0.4047 MSE 2.049
Training4<-lm(Stay~x1+Beds+factor(Region))
summary(Training4) #significant at .18
anova(Training4)#0.45,     Adjusted R-squared:  0.4047 MSE 2.049
Training5<-lm(Stay~x1+Census+factor(Region))
summary(Training5) #significant at .12
anova(Training5)#0.46,     Adjusted R-squared:  0.4047 MSE 2.049
Training6<-lm(Stay~x1+Nurses+factor(Region))
summary(Training6) #NOT significant
anova(Training6)#0.46,
Training7<-lm(Stay~x1+Services+factor(Region))
summary(Training7) #NOT significant
anova(Training7)#0.46
#R-12#### Exercise model summary
modelB<-lm(Stay~Age+Risk+Census)
anova(modelB) #PRESS= 269.6774 SSTO=409.211 R-sq pred=1-PRESS/SSTO =.341
summary(modelB)
anova(model3) #Risk SSR 67.24  SSTO: 237
sum((testing$residuals/(1-hatvalues(testing)))^2) #PRESS for testing
```