Analysis of Hospital Stays in a Nosocomial Infection Control Data

Jessica Hathaway, Matthew Hill, Lilshay Rogers, Heaven Tate
Mentor: Dr. Julian A.D. Allagan (ECSU)
Principal Investigator: Dr. Linda B. Hayden (ECSU)

Abstract

In this report we developed and analyzed several linear regression models to predict hospital stays (or length of Stay) of patients in the U.S using the SENIC project data from CDC-Atlanta. We 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 U.S? In at least one model, we found that Risk of infection, Nurses, Census and Regions influenced the variable “Stay” the most.

Keywords: Data Analysis, Linear regression, Nosocomial

Introduction

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 al4 (see Appendix C, page 1348). The data set contains no missing value although some scaling was found necessary for the purpose of our analysis. As mentioned in the introduction, for our analysis, the original SENIC project data was split into two data: The training data that we called ENIC contains observations 1-70 and the testing data which we 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.

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

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

Methodology

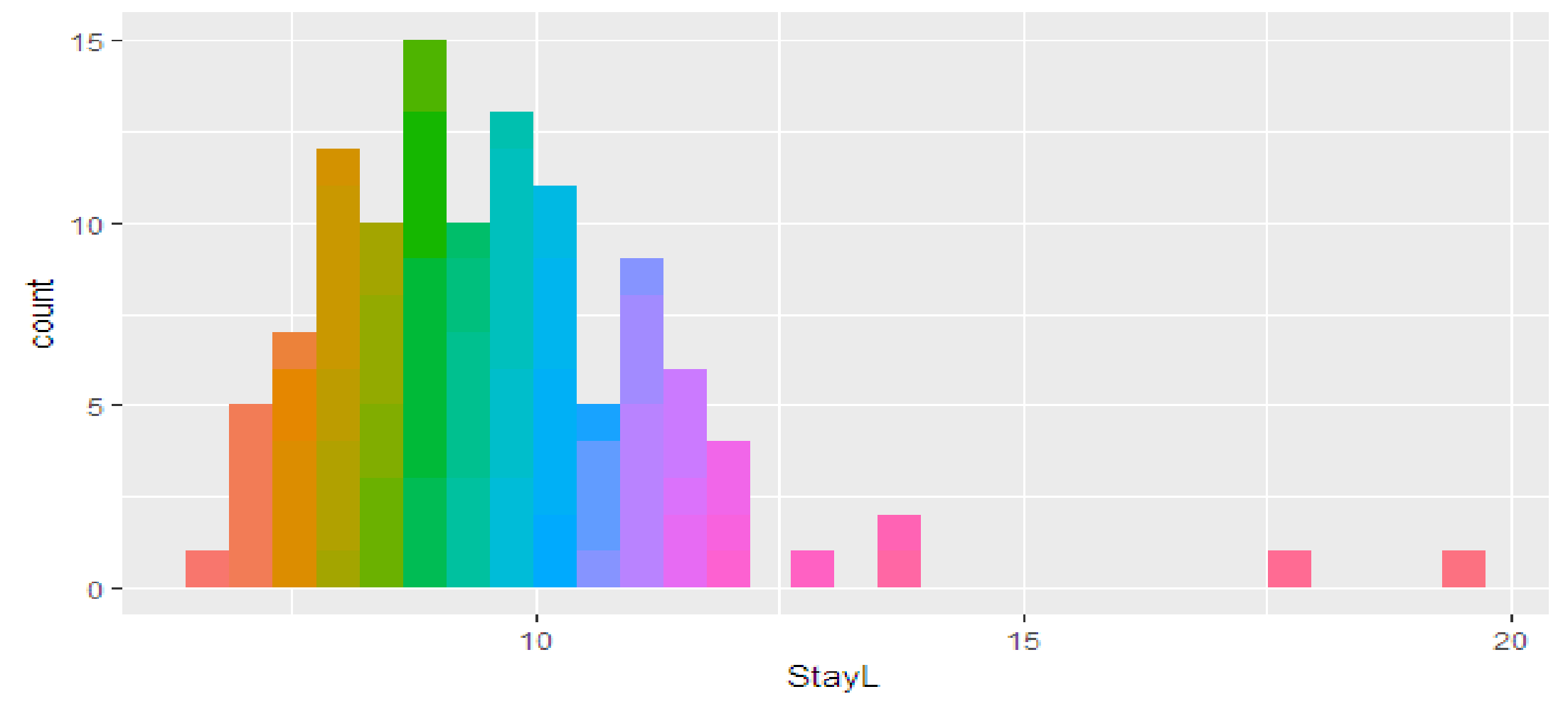
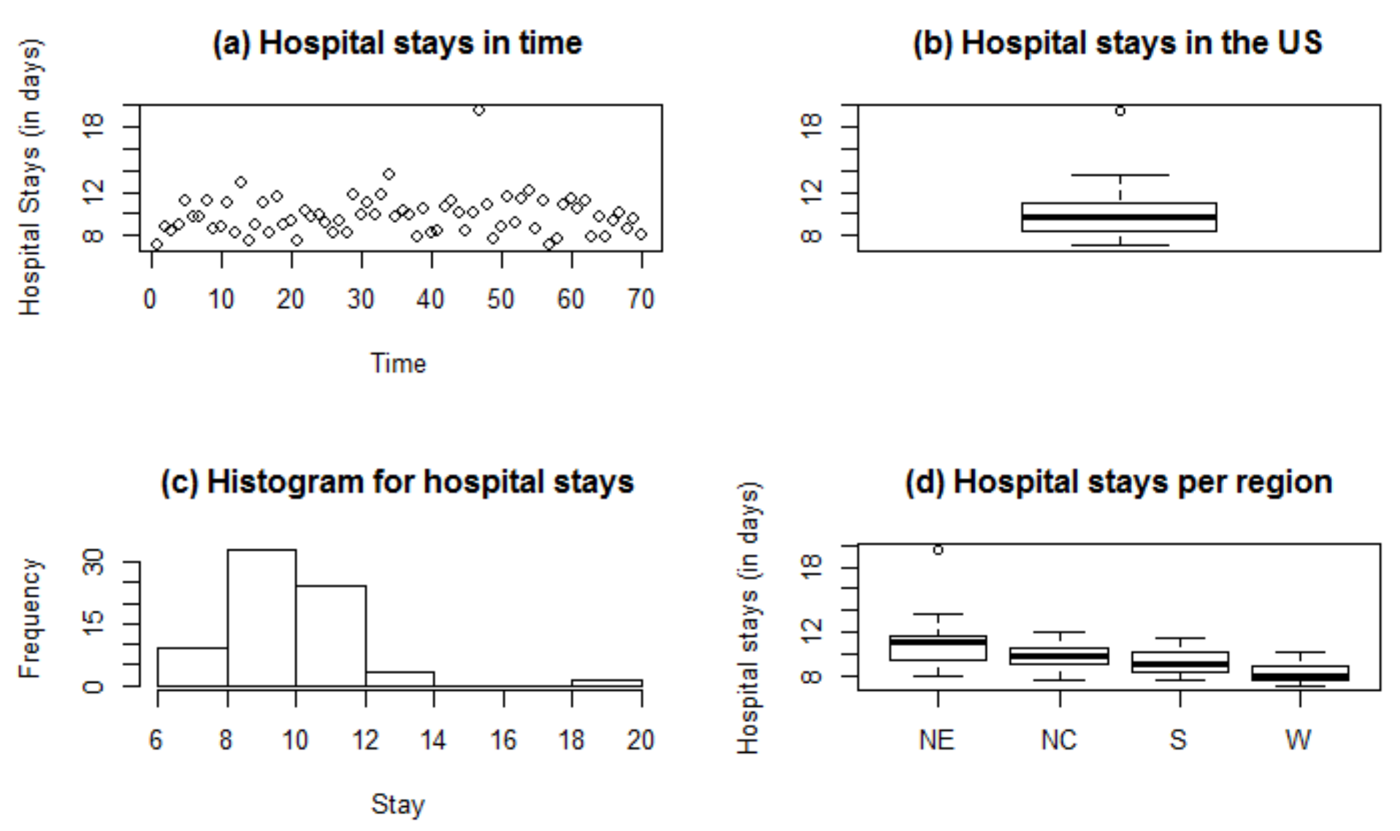
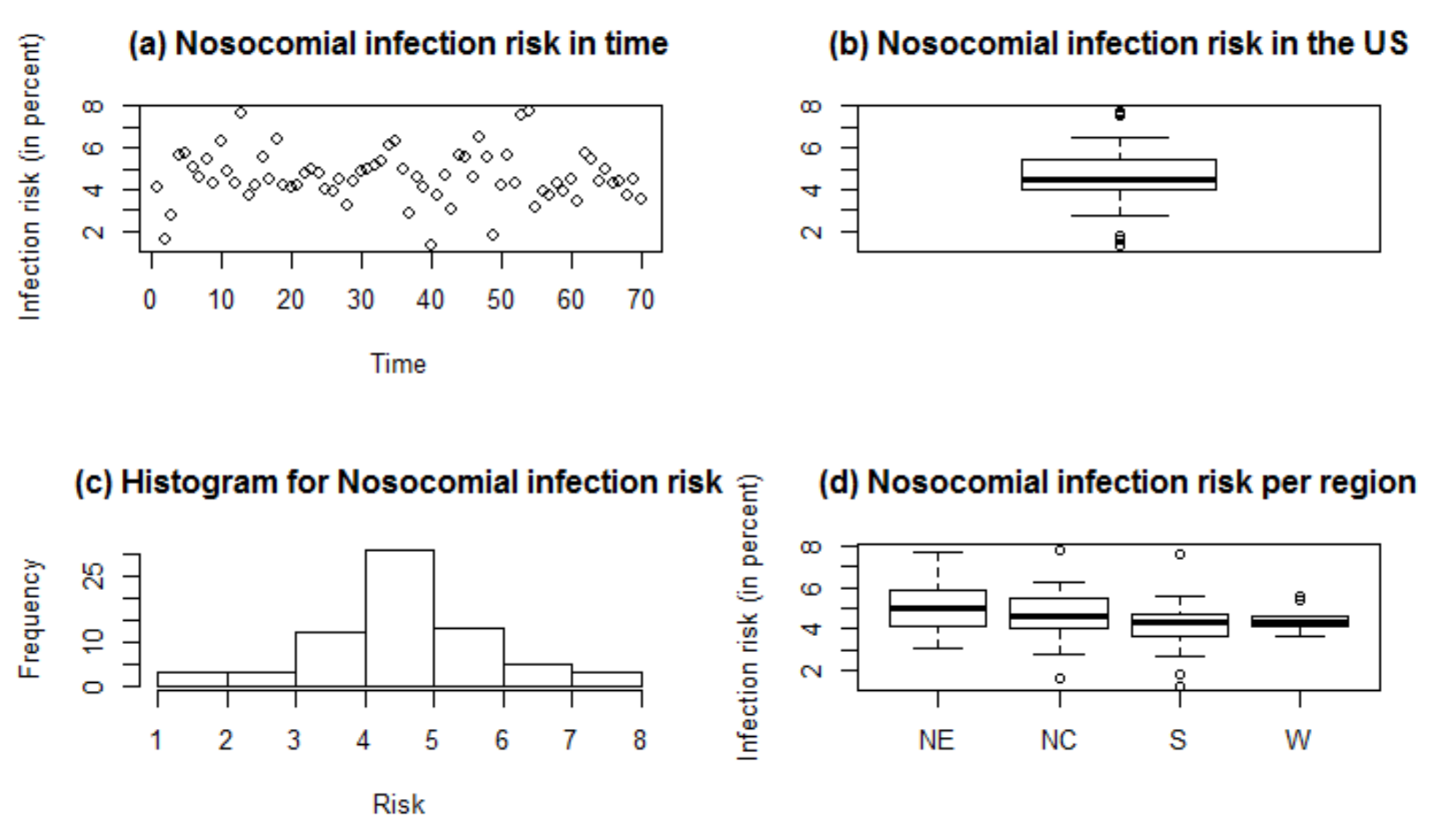


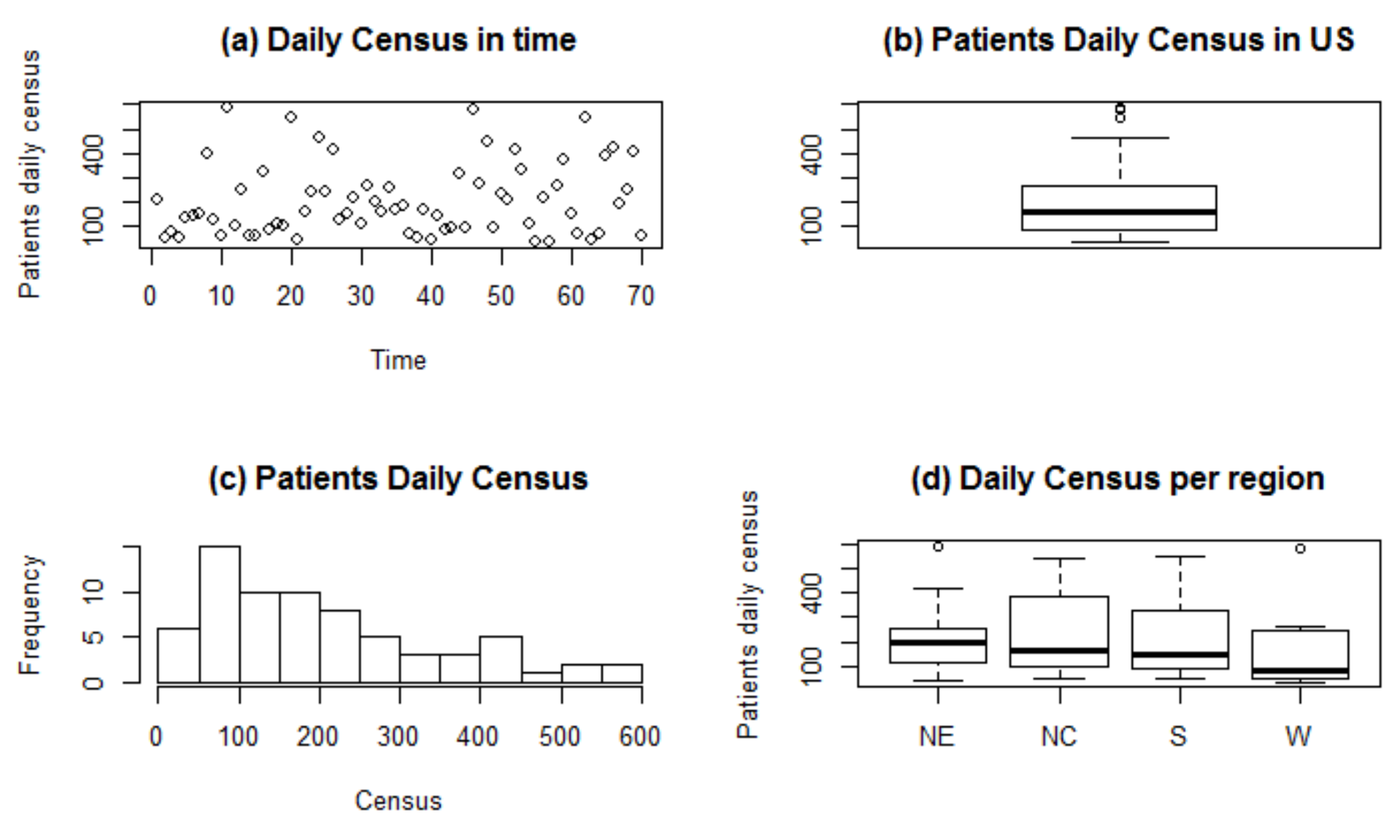
Figure 1: Distribution of Length of the response Stay



Plot 1: Hospital stays data plots from ENIC data



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



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

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

Variables	Mallows’ Cp
Stay, Age, Risk, Xray, Region, Census, Nurses	4.298737
Stay, Age, Risk, Region, Census, Nurses	4.576915
Stay, Risk, Region,Census, Nurses	4.970154

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

Variables for Each Submodel	Adj. R^2
Stay, Age, Risk, Xray, Region, Census, Nurses	0.5115
Stay, Age, Risk, Region, Census, Nurses	0.493
Stay, Risk, Region, Census, Nurses	0.4736

Table 4: Top 3 sub-models listed based on PRESSp-value

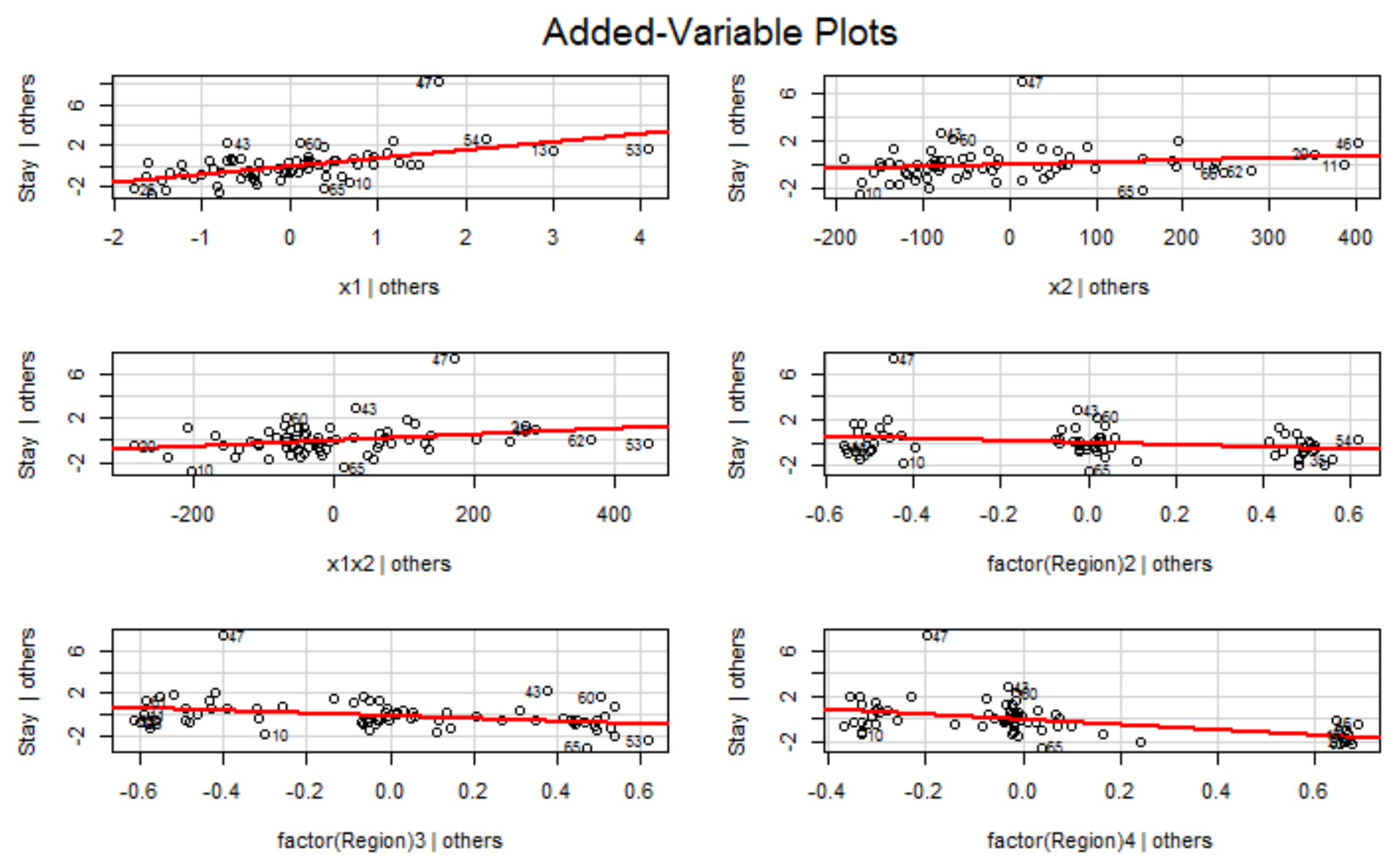
Variables for Each Submodel	PRESS p-value
Stay, Age, Risk, Xray, Region, Census, Nurses	151.0711
Stay, Age, Risk, Region, Census, Nurses	150.7552
Stay, Risk, Region, Census, Nurses	148.574

Table 5: Correlation matrix for chosen model

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

	Estimate	Standard Error	P-value
Intercept	10.620649	0.317679	< 0.001
Risk*	0.781185	0.156062	< 0.001
Census*	0.001769	0.001170	0.135545
Risk*Census*	0.002806	0.001258	0.029308
North Central	-0.959352	0.437287	0.031941
South	-1.244376	0.444660	0.006804
West	-2.223747	0.561232	< 0.001

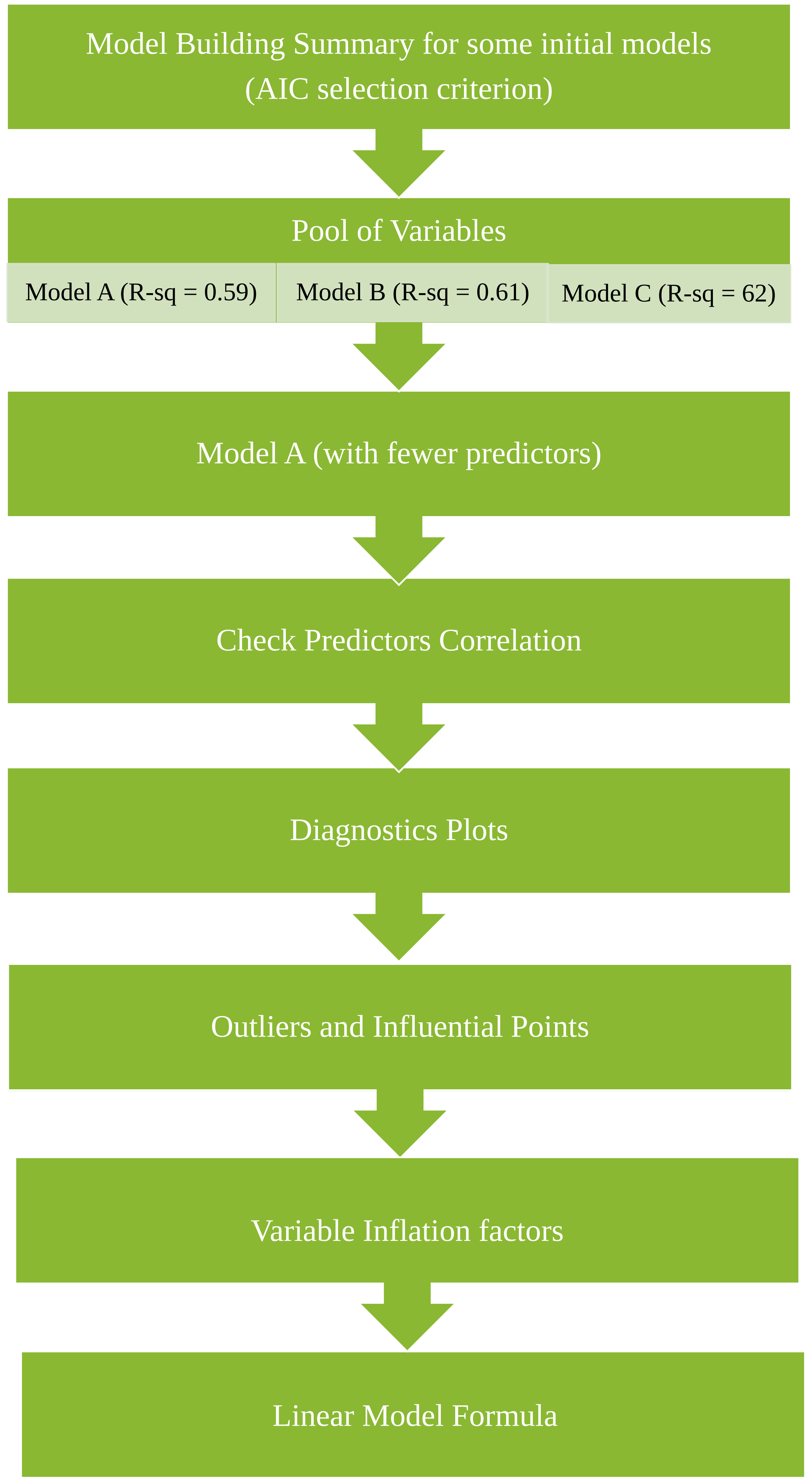


Plot 4: Added-value plots for final model

Final Linear Regression Model Formula

Average Stay ≈
11+0.8Risk+0.002Census+0.003Risk●Census-0.96NC-1.2S-2.2W

Alternative Model Selection Process without breaking up the original data



Model Building Summary for some initial models (AIC selection criterion)

Pool of Variables

Model A (R-sq = 0.59) | Model B (R-sq = 0.61) | Model C (R-sq = 62)

Model A (with fewer predictors)

Check Predictors Correlation

Diagnostics Plots

Outliers and Influential Points

Variable Inflation factors

Linear Model Formula

Conclusion

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 R^2 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, Xray, Beds, Affiliation, Region, Census, Nurses, Services. Our process initially yielded 3 sub-models from which we determined, based on the adjusted R^2 and the PRESS values, one single “best” model. This model contained the response variable Stay and the predictors Risk^*, Census^*, $\text{Risk}^*\text{Census}^*$, and Region where (*) indicate that these variables are “centered”. We 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). We were surprised to see that *Age* was not a significant predictor for *Stay* even though we 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, we actually checked the mean of hospital stays for the SENIC project data and observed a similar value (9.6 days). We 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 we included their record also in our final model analysis. A reasonable argument could be made to delete this hospital record, say, we wanted to limit our final analysis to hospitals for which the Stay is less than 14 days (according to ENIC data). In the end we 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, we chose to compare it to a known model (from a past project) and we 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, we 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). We did not think that adding any new or replacing an existing predictor would improve the overall significance of our final model and yet, we are mindful that there is no “perfect” model. A further analysis with the goal of arriving at an improved linear regression model (compared to our final model) 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 we found that a second-order regression model was not appropriate for our model with the selected predictors, we could not rule out such order if one considers other predictors for the model and perhaps more data is needed to for a better predictive model.

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.
[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.