Regression Analysis: Hospital-Acquired Infections

Sarah Brockman Anonymous Anonymous Anonymous

1 May 2019

1 Introduction

1.1 General Background

This project will focus on *The Study on the Efficacy of Nosocomial Infection Control (SENIC)*, which was an endeavor to determine whether infection surveillance and control programs in the United States have lowered the rates of nosocomial (hospital-acquired) infections between 1970 and 1976. The main purpose of this study is to determine if certain hospital characteristics have a relationship with the probability of contracting a hospital-acquired infection (HAI). A few of these hospital characteristics include average length of stay, number of nurses, and affiliation with a medical school, among others. Some of these characteristics may facilitate an increase in HAIs and may unnecessarily put more patients at risk. It is important to determine which factors indicate a higher risk, and then inform hospitals so the administration can implement practices that preclude these highly-preventable infections.

1.2 Main Goals

The goal of our analysis is to determine if there is a strong relationship between a number of hospital characteristics and the probability of contracting a hospital-acquired infection at that hospital. The full list of the 9 characteristics is as follows: average length of stay, average age of patients, routine culturing ratio, routine chest x-ray ratio, number of beds, medical school affiliation, average daily census, number of nurses, and available hospital facilities. These variables will all be used to predict the infection risk, or the probability of contracting an HAI. Of course, it is likely that not all of these variables will be useful in predicting the infection risk, but our goal is to find the ones with the strongest predictive capacity. We will use least-squares multiple regression with various statistical procedures such as forward stepwise regression to determine which variables share the strongest regression relationship with the response variable, which in our case is probability of infection. The variables that have a strong relationship can be warning flags for potential infection, and are those to which hospitals should pay close attention.

We will present the results from our exploratory data analysis in Section 2. In Section 3, we will fit an initial regression model. In Section 4, we will transform some of our predictors and refit our model. In Section 5, we will add interaction and polynomial terms to our model and perform certain model selection techniques to find a final model, and we present our conclusions and findings in Section 6. Our SAS code for the project can be found in Appendix A.

1.3 Dataset Description

The data from the SENIC study includes nine hospital characteristics (plus the dependent variable, infection risk) from 113 different hospitals, sampled from the original 338 hospitals surveyed. The nine hospital characteristics include average length of stay (LOS), average patient age (Age), routine culturing ratio (RCR), routine chest x-ray ratio (RCX), average number of beds (Beds), average daily census (Census), average number of nurses (Nurses), percent of available facilities and services (Facilities), and affiliation with a medical school (Univ). Of these variables, Univ is the only binary predictor, while all the others are quantitative (i.e., they are values in $\mathbb{R}$). We may refer to the variables by their names in parentheses in the remainder of this report. In Section 2, we will summarize our findings from our exploratory data analysis, and in Section 3 we will begin using these predictors to fit a multiple regression model.

2 Exploratory Data Analysis

To begin our data analysis, we studied the basic descriptive statistics (e.g. mean, standard deviation, etc.) of our predictors to get a general feel of the dataset. These descriptive statistics can be found below in Figure (1).

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
ID	identification number (1-113)	113	57.0000000	32.7643099	1.0000000	113.0000000
LOS	average length of stay (days)	113	9.6483186	1.9114560	6.7000000	19.5600000
Age	average age (years)	113	53.2318584	4.4616074	38.8000000	65.9000000
RCR	routine culturing ratio (see descrip.)	113	15.7929204	10.2347074	1.6000000	60.5000000
RCX	routine chest x-ray ratio (see descrip.)	113	81.6283186	19.3638261	39.6000000	133.5000000
Beds	average number of beds	113	252.1681416	192.8426868	29.0000000	835.0000000
Univ	medical school affil: binary, 1=Yes, 0=No	113	0.1504425	0.3590971	0	1.0000000
Census	average daily census	113	191.3716814	153.7595639	20.0000000	791.0000000
Nurses	average number of nurses	113	173.2477876	139.2653897	14.0000000	656.0000000
Facilities	percent of available facil. and serv.	113	43.1592920	15.2008613	5.7000000	80.0000000
y	average infection risk (percent)	113	4.3548673	1.3409080	1.3000000	7.8000000

Figure 1: Basic Descriptive Statistics for All Predictors

Next, we studied the Pearson Correlation Coefficients between each of our quantitative predictor variables. We observed that the correlation coefficients between average number of beds and the following predictor variables were very strong (i.e., correlation coefficient ≥ 0.80): percentage of available facilities, average number of nurses, and average daily census. This makes sense because intuitively, if there are on average more beds, then there are more facilities for more patients to go to, and as a result, more nurses are needed to care for those patients. The only other correlation coefficient that was very strong (correlation coefficient ≥ 0.80) was that between the average number of nurses and the average daily census. This also intuitively reasonable because if there are, on average, more patients in each hospital per day during the study period, then there would need to be, on average, more fully registered and licensed practical nurses during the study period to treat those patients. In Section 3, we analyze the Variance Inflation Factors (VIFs) of all the predictors to see which variables need to be removed due to high multicollinearity.

Following the analysis of the correlations, we studied each of the predictors in detail. We analyzed their relationship with the dependent variable to see if it was apparently linear. We also checked the distributions of each variable to see if they were approximately normally distributed. The variables LOS, Age, Census, RCX, and Facilities all showed an approximate linear relationship with the

dependent variable. They also exhibited approximately normal distributions, however LOS and Census were skewed to the right. Census and Facilities had a few outliers, but we decided these would be negligible in our future analysis. None of these predictors present a compelling need for transformations. One interesting observation was that we have no data for patients under the age of 38, which means our analysis cannot extend to children or adolescents.

The variables RCR, Beds, and Nurses were all heavily skewed to the right, so they could all benefit from transformations such as $\sqrt{X}$ or $\log_{10}X$, which we explore in more detail in Section 4. RCR appears to have a linear relationship with infection risk, but Beds and Nurses do not.

All the previously mentioned predictors are quantitative variables. Our remaining variable, Univ, is our only binary predictor. Univ indicates whether or not a hospital has a medical school affiliation, 1 being 'yes' and 0 being 'no'. Univ shows only two outliers, which are not a problem for our analysis. Additionally, there are a lot more hospitals without a medical school affiliation than with an affiliation (96 vs. 17), so the data distribution is slightly uneven.

3 Initial Regression Model

3.1 Regression Model Specification

3.1.1 Model with All Predictors

To begin the model fitting portion of our analysis, we fit a simple linear regression model to all of our available predictor variables. This model contains only first-order terms (no polynomial or interaction terms), and no predictor variables are excluded or transformed at this time. This serves as our initial baseline model, and we will alter it in subsequent analysis outlined in the remainder of this report. The initial regression model is below, in Equation (1):

$$\begin{aligned}\hat{Y} = &-0.6776 + 0.1734\ \text{LOS} + 0.0133\ \text{Age} + 0.0495\ \text{RCR} + 0.0119\ \text{RCX} - 0.0020\ \text{Beds} \\ &- 0.5021\ \text{Univ} + 0.0019\ \text{Census} + 0.0021\ \text{Nurses} + 0.0176\ \text{Facilities}.\end{aligned} \tag{1}$$

Our ultimate goal is to take this initial regression function and alter its variables so that it is minimal yet still provides maximal predictive power. We will achieve this by transforming our variables in Section 4 and adding interaction and polynomial terms in Section 5. We will then use selection techniques to determine the best model for this dataset.

3.1.2 Multicollinearity: Variance Inflation Factor

We ideally want all of our predictors to be completely independent of all other predictors. This is unlikely to happen, however; the variables will have some degree of multicollinearity. We can check how related each predictor is to the other predictors using a measure called the Variance Inflation Factor. The goal is to remove predictors with high VIF values, i.e., predictors that are strongly related with the other predictors. This removes some redundancy from our model. The VIF values (along with the regression coefficients) of each predictor can be found in Figure (2). It can be seen that the predictors Beds and Census both have VIF values greater than 10 (≈ 33.80 and 36.30 respectively), which is a typical cutoff point. Therefore, we need to remove one (and potentially both) from the model. We chose to remove the variable Census, because after removing each variable separately, the model without Census had a slightly higher R^2 value. Nurses has a VIF of nearly 10 (≈ 7.28), but this is not quite high enough to warrant removal from the model, so we leave it in.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.67757	1.20093	-0.56	0.5738	0
LOS	1	0.17344	0.06866	2.53	0.0131	2.12451
Age	1	0.01329	0.02220	0.60	0.5506	1.20979
RCR	1	0.04950	0.01081	4.58	<.0001	1.50913
RCX	1	0.01187	0.00545	2.18	0.0317	1.37470
Beds	1	-0.00197	0.00271	-0.72	0.4705	33.80318
Univ	1	-0.50208	0.32990	-1.52	0.1311	1.73083
Census	1	0.00186	0.00353	0.53	0.5999	36.30499
Nurses	1	0.00205	0.00174	1.18	0.2421	7.28229
Facilities	1	0.01763	0.01016	1.73	0.0858	2.94396

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.71571	1.19457	-0.60	0.5504	0
LOS	1	0.19002	0.06079	3.13	0.0023	1.67696
Age	1	0.01195	0.02197	0.54	0.5878	1.19375
RCR	1	0.04848	0.01059	4.58	<.0001	1.46027
RCX	1	0.01170	0.00542	2.16	0.0333	1.36974
Beds	1	-0.00071434	0.00130	-0.55	0.5844	7.82687
Univ	1	-0.46561	0.32141	-1.45	0.1504	1.65440
Nurses	1	0.00225	0.00170	1.32	0.1895	6.96274
Facilities	1	0.01718	0.01009	1.70	0.0916	2.92340

Figure 2: Variance Inflation Factors: Initial Model (Left) and Model without Census (Right)

After removing Census, we refit the model and obtained the following regression line, as specified in Equation (2):

$$\begin{aligned}\hat{Y} = &-0.7157 + 0.1900\ \text{LOS} + 0.0119\ \text{Age} + 0.0485\ \text{RCR} + 0.0117\ \text{RCX} \\ &- 0.0007\ \text{Beds} - 0.4656\ \text{Univ} + 0.0022\ \text{Nurses} + 0.0172\ \text{Facilities}.\end{aligned} \tag{2}$$

It is important to note that no regression coefficients changed a large amount after removing one of the predictors with a high VIF value. However, we still check to see if the new model with Census removed has more predictors with high VIF values. We confirmed that there are no remaining predictors with high VIF values that need to be removed, as can be seen in Figure (2).

3.2 Coefficient Estimates

The coefficient estimates (the $\hat{\beta}$'s) in our initial model (with Census removed) are found in Table (1). These are our preliminary slopes for our regression line, and will change once we add more complex terms to our model.

Predictor	Intercept	LOS	Age	RCR	RCX	Beds	Univ	Nurses	Facil.
$\hat{\boldsymbol{\beta}}$	-0.7157	0.1900	0.0119	0.0485	0.0117	-0.0007	-0.4656	0.0022	0.0172

Table 1: Coefficient Estimates for Initial Model

3.3 Residual Diagnostics

3.3.1 Residuals versus Fitted Values

Figure (3) shows the residuals of our initial model plotted against the fitted values. There is no apparent pattern in the residuals, which is in keeping with the assumption that our model is linear. There also seems to be constant variance and even distribution around zero. Therefore, our current model's residuals support our assumptions for linear regression error terms.

3.3.2 Normal Probability Plot of Residuals

Figure (3) shows the Normal Probability Plot of the residuals of the initial model. We want this plot to be as close to linear as possible, and it is very close. This means the error terms are approximately normally distributed, which indicates our model supports the assumptions for the error terms. The R^2 value for this initial model is $R^2 = 0.5343$. This is decently high already, but we can still improve our model by removing any redundancy and raising its capacity.

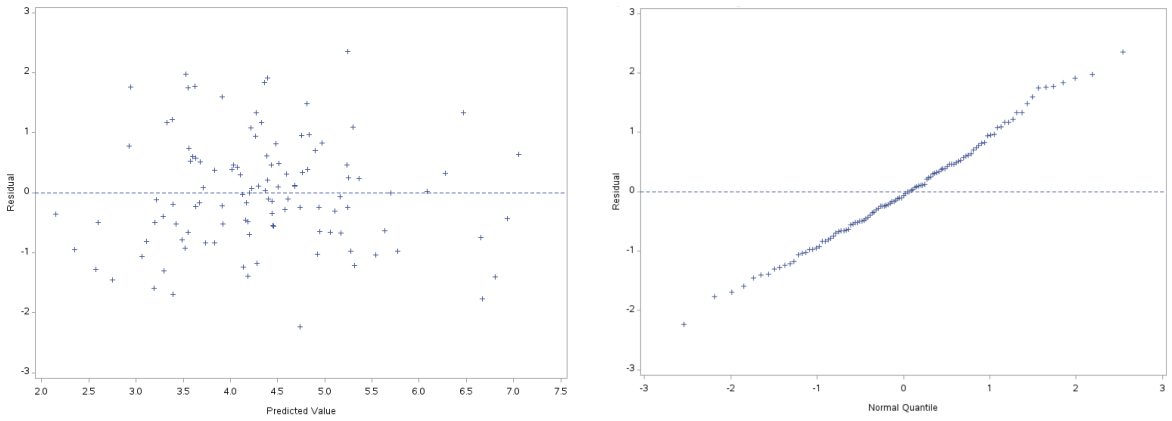

Figure 3: Residuals vs. Fitted Values (Left) and Normal Probability Plot of Residuals (Right)

4 Transformations of Covariates and Response Variables

During our data exploration (see Section 2), we observed that some of our predictors had some outlying data points, and thus these predictors could benefit from a transformation of some type. The variables that we thought could use transformations were RCR, Nurses, and Beds. We decided to transform RCR and Beds using square root, and Nurses using log base 10. We decided that a transformation of the dependent variable Y was not necessary since our initial model upheld the assumptions for linear regression error terms. As we add more complex terms to the model, it may become apparent that a transformation of Y is needed.

4.1 Regression Model Specification

4.1.1 Regression Model Specification with Transformed Variables

After refitting our model to take the transformations into account, we obtained the following regression function, as seen in Equation (3):

$$\begin{aligned}\hat{Y} = &-4.5168 + 0.1658\ \text{LOS} + 0.02778\ \text{Age} + 0.4091\ \text{sqrtRCR} + 0.0082\ \text{RCX} \\ &- 0.2494\ \text{Univ} + 2.1817\ \text{logNurses} - 0.0416\ \text{sqrtBeds} - 0.0081\ \text{Facilities}.\end{aligned} \tag{3}$$

The coefficients of the non-transformed predictors did not change dramatically, but the coefficients of the transformed predictors did change by approximately 1-3 orders of magnitude.

4.1.2 Multicollinearity: Variance Inflation Factor

The new model with transformed variables has no predictors with VIF values greater than 10, so no more variables need to be dropped from the model at this time. The VIF values for this revised version of the model can be found in Figure (4). The VIF values for the transformed predictors are decently low, and their presence did not have any large effect on the VIF values of the other variables. Thus, we can continue to add more terms to the model, such as interaction terms and polynomial terms, as described in Section 5.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-4.51683	1.35323	-3.34	0.0012	0
LOS	1	0.16576	0.05606	2.96	0.0038	1.67907
Age	1	0.02777	0.02059	1.35	0.1803	1.23365
RCRsqrt	1	0.40908	0.08539	4.79	<.0001	1.59697
RCX	1	0.00820	0.00513	1.60	0.1125	1.44012
Univ	1	-0.24942	0.28898	-0.86	0.3901	1.57455
logNurses	1	2.18167	0.60186	3.62	0.0004	6.93943
sqrtBeds	1	-0.04158	0.03841	-1.08	0.2815	6.70613
Facilities	1	-0.00811	0.01108	-0.73	0.4655	4.14547

Figure 4: Variance Inflation Factors for Predictors in Transformed Model

4.2 Coefficient Estimates

The coefficient estimates (the $\hat{\beta}$'s) for our model with the transformations $\log_{10}(\text{Nurses})$, $\sqrt{\text{RCR}}$, and $\sqrt{\text{Beds}}$ can be found in Table (2). As stated in Section 4.1.1, the coefficients did not change dramatically. The signs of all the slopes remained the same except for Facilities, which changed from positive to negative. The slopes for the transformed variables all got higher as well, and did not decrease.

Pred.	Intercept	LOS	Age	$\sqrt{\text{RCR}}$	RCX	$\sqrt{\text{Beds}}$	Univ	$\log_{10}(\text{Nurses})$	Facil.
$\hat{\boldsymbol{\beta}}$	-4.5168	0.1658	0.0278	0.4091	0.0082	-0.0416	-0.2494	2.1817	-0.0081

Table 2: Coefficient Estimates for Transformed Model

4.3 Residual Diagnostics

4.3.1 Residuals versus Fitted Values

The plot of residuals versus fitted values for the model with transformations on RCR, Beds, and Nurses can be found in Figure (5). The residuals in this plot are relatively evenly distributed around zero; The negative residuals seem to be slightly closer to zero than the positive residuals, but this is not a cause for concern at the moment. The residuals also have constant variance as the predicted values increase, so the error terms of the transformed model are still supporting our assumptions. There is also no pattern in the residuals, so a linear model is fitting the data well.

4.3.2 Normal Probability Plot of Residuals

The normal probability plot of the residuals of the transformed model can be found in Figure (5). The plot is still roughly linear, so our transformed model's error terms are still approximately normally distributed, in keeping with our error term assumptions. Since the plot is linear, we still do not need any transformations for our dependent variable, Y. This line does look slightly more wavy than the normal probability plot of the original fitted model (see Figure (3)), but our R^2 value for the new model is $R^2 = 0.6044$. This is an increase of about 0.0701 from the initial model. R^2 is guaranteed to increase as more terms are added to the model, but here, we are not adding new terms to the model, only transforming a subset of the current terms. Thus, the transformed model is doing a slightly better job at explaining the relationship between the predictors and the dependent variable than the initial model.

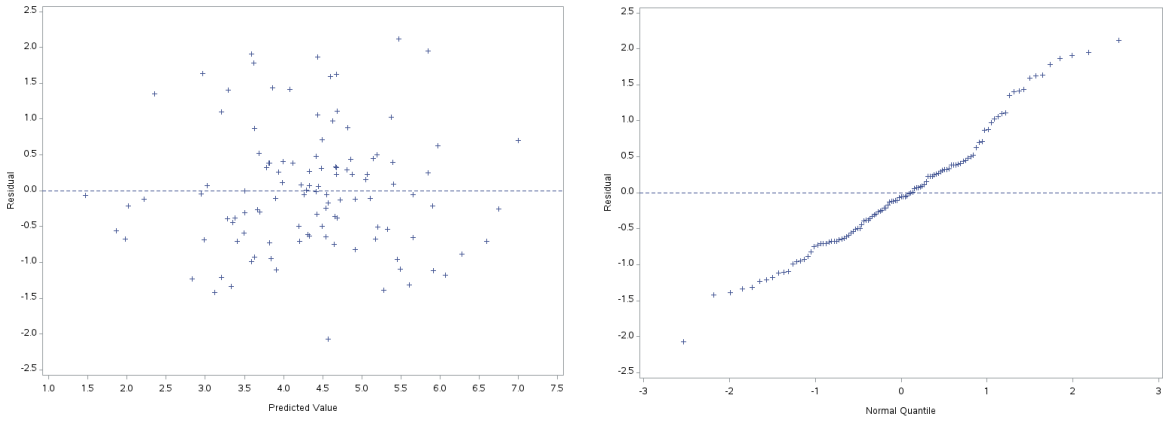

Figure 5: Residuals vs. Fitted Values for Transformed Model (Left) and Normal Probability Plot of Residuals for Transformed Model (Right)

5 Addition of Interaction Effects and Polynomial Terms into Model

In this section, we explore the effects of interaction and polynomial terms in our model. The interaction and polynomial terms and their names we use to reference them are as follows:

- RCRsquare = RCR × RCR
- NurseRCR = Nurses × RCR
- RCRFac = RCR × Facilities
- NursesFac = Nurses × Facilities
- RCRRCX = RCR × RCX
- LOSNurses = LOS × Nurses
- NursesSquare = Nurses × Nurses
- RCRcube = RCR × RCR × RCR

These terms were chosen primarily based on intuition, as we thought these predictors in conjunction could have greater predictive power. For example, a hospital with more available facilities will most likely have a larger staff of nurses to supervise and care for patients in these facilities. We also observed after some trial-and-error testing that polynomial terms of RCR seemed to increase the adjusted R^2 value a bit, so we added quadratic and cubic terms of RCR. After adding these new

terms to the model, we perform some selection techniques in the following sections to choose the best model out of the models created from different subsets of these interaction terms and polynomial terms together with our base predictors.

5.1 Regression Model Specification

5.1.1 Model Selection

To find the best model, we reference the following five criteria: R_p^2, adjusted R^2 ($R_{a,p}^2$), Mallow's C_p criterion, Akaike's Information Criterion (AIC_p), and Schwarz's Bayesian Criterion (SBC_p). We want the model with the highest R_p^2 and $R_{a,p}^2$, and the lowest C_p, AIC_p, and SBC_p. We can use SAS to generate these criteria for a large number of models and see which one best satisfies our requirements. After analyzing the output of the selection procedure using these criteria, we observed that the best model used around 6-8 predictors. One such model uses the predictors LOS, Age, RCRsqrt, RCX, logNurses, Facilities, RCRsquare, and NurseRCR. The corresponding values for the criteria for the model with these predictors can be found in Table (3).

Criterion	R_p^2	$R_{a,p}^2$	C_p	AIC_p	SBC_p
Value	0.6421	0.6146	4.6293	-32.8104	-8.26394

Table 3: Values for the Five Model Selection Criteria

Plots for the five criteria versus the number of predictors in the model can be seen in Figure (6). It can be seen that the optimal number of predictors $p-1$ is around 6-8.

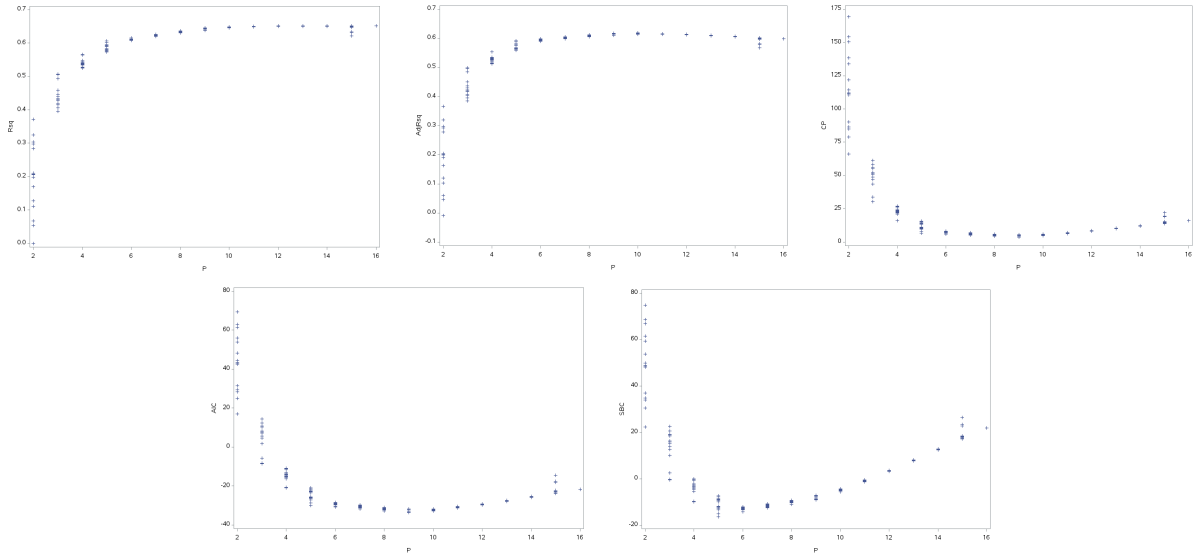

Figure 6: Five Criteria for Model Selection: R_p^2 (Top Left), $R_{a,p}^2$ (Top Middle), C_p (Top Right), AIC_p (Bottom Left), and SBC_p (Bottom Right)

Although the above criteria give us a pretty good idea what a good model is, we can do even better. We can use a stepwise selection procedure to continuously add and remove variables from the model and measure their effects. We can do this procedure automatically using SAS. For

stepwise selection, we consider all the base predictors as well as all the interaction and polynomial terms listed in the beginning of Section 5. Then, variables are added to the model if their p-value is below a threshold of $\alpha = 0.05$. Variables are removed from the model if p-value is greater than a threshold of $\alpha = 0.10$. This process repeats iteratively until no variables can be added or removed. After running this procedure on all of our predictors, we obtained the following regression model, as described in Equation (4):

$$\hat{Y} = -3.28427 + 0.20036 \text{ LOS} + 0.62552 \text{ RCRsqrt} + 1.77334 \text{ logNurses} - 0.000128 \text{ NurseRCR}. \quad (4)$$

5.1.2 Multicollinearity: Variance Inflation Factor

The Variance Inflation Factors for the final model can be found in Figure (7). Upon reviewing the VIFs, we determined that these predictors should all be included in our final model as the VIF for each respective predictor is below three. Of course, since these values are all very low, none of these predictors need to be removed. The p-values for each predictor is also included in Figure (7). Each p-value is below 0.0001, with the exception of NurseRCR. These values are a lot lower than the p-values of the variables in previous versions of the model (see Figures (2) and (4)). This indicates that we have found a model that is more explanatory of the relationship between the predictors and infection risk, albeit with less redundancy.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-3.28427	0.74174	-4.43	<.0001	0
LOS	1	0.20036	0.04770	4.20	<.0001	1.26843
RCRsqrt	1	0.62552	0.08691	7.20	<.0001	1.72664
logNurses	1	1.77334	0.31939	5.55	<.0001	2.03950
NurseRCR	1	-0.00012776	0.00003821	-3.34	0.0011	2.83830

Figure 7: Variance Inflation Factors for Final Model found by Stepwise Selection

5.2 Coefficient Estimates

After performing stepwise selection, we determined that our final model should include the following predictors: LOS (length of stay), RCRsqrt (square root of routine culturing ratio), logNurses (log base 10 of Nurses) and NurseRCR (Nurses multiplied by routine culturing ratio). The coefficient estimates for the model found by stepwise selection can be found below in Table (4). The R^2 value for this model is 0.6064, 0.0020 more than the R^2 value of our transformed model, with fewer predictors involved.

Predictor	Intercept	LOS	RCRsqrt	logNurses	NurseRCR
$\hat{\beta}$	-3.28427	0.20036	0.62552	1.77334	-0.000128

Table 4: Coefficient Estimates for Stepwise Selection Model

5.3 Residual Diagnostics

5.3.1 Residuals versus Fitted Values

The plot for residuals versus fitted values for our model obtained from stepwise selection can be found in Figure (8). There is no pattern in the plot, so our linear model is still appropriate. The residuals are also evenly distributed around zero and have constant variance as the predicted value increases. This is in keeping with our assumptions for the error terms, so this model is a good fit.

5.3.2 Normal Probability Plot of Residuals

The normal probability plot for the model from stepwise selection can be found in Figure (8). We want this plot to be as close to linear as possible, and it is quite close, except for a slight curve in the upper end. The close approximation to linearity tells us our residuals are normally distributed, which tells us our regression model from stepwise selection is a good fit for the data.

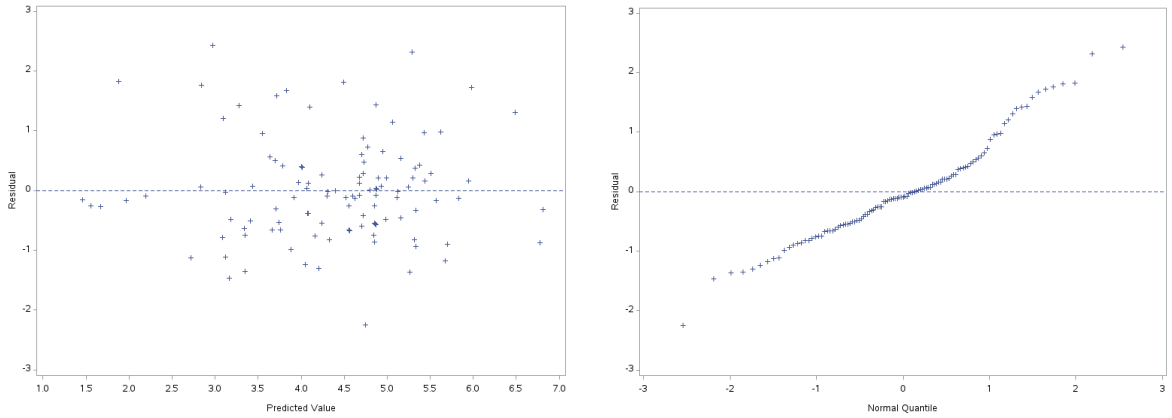

Figure 8: Residuals vs. Fitted Values for Stepwise Selection Model (Left) and Normal Probability Plot of Residuals for Stepwise Selection Model (Right)

6 Final Model and Conclusions

6.1 Final Model Specification

Our final model is repeated below, in Equation (5):

$$\hat{Y} = -3.28427 + 0.20036 \text{ LOS} + 0.62552 \text{ RCRsqrt} + 1.77334 \text{ logNurses} - 0.000128 \text{ NurseRCR}. \quad (5)$$

This is the model found by forward stepwise regression. After generating multiple different regression models and transforming and eliminating predictor variables based upon VIF, correlation coefficients, and the five model selection criteria, we have decided that this is one of the best models we can fit given the data we currently have available. We made sure this model was a good fit for the data by analyzing the plot of the residuals vs. fitted values as well as the normal probability plot. These plots can be found in Figure (8), and are discussed in more detail in Section 5.3. Based on the plots, our linear model is a good fit for the data.

To further guarantee our model is a good fit for the data, we analyzed the R^2 and adjusted R^2, along with the other three model selection criteria, to make sure the values are to be expected. The model selection criteria for our final model can be found below in Table (5).

Criterion	R_p^2	$R_{a,p}^2$	C_p	AIC_p	SBC_p
Value	0.6064	0.5918	5.0000	-30.0673	-16.4303

Table 5: Values for the Five Model Selection Criteria

These values are about what we expected to see from our final model. Other models with a similar subset of predictors performs roughly the same, with negligible difference in performance. Our final R^2 value is 0.6064, which means about 60% of the variation in the data is explained by the model, which is pretty high given we only have 113 data points. The adjusted R^2 is also very close to R^2, potentially indicating that variables kept in the model are not padding the R^2 value with their presence (as R^2 always increases with more variables in the model). The p-values for all the final predictors can be found above in Figure (7). They are all very low, indicating all final predictors provide a significant contribution to the model.

6.2 Conclusions

In our initial model, we had nine predictor variables: LOS (length of Stay), Age, RCR (routine culturing ratio), RCX (routine chest x-ray ratio), Beds, Census, Nurses, Univ (medical school affiliation), and Facilities. At first, our group believed that LOS, Age, and Nurses would be the most significant predictors of infection risk. After running our regression analysis and having to remove all but one of our original predictors (LOS), it became clear that our original hypothesis of the most significant factors was somewhat accurate. Our final model includes only one original predictor, two transformed predictors, and one interaction term. Our model suggests that as the predictor variables increase, the chance of a hospital acquired infection also increases with the only exception being NurseRCR (where there was a very slight decrease). After taking a step back, this result perhaps seems intuitive. The longer your length of stay (LOS) or perhaps the more interaction you have with a nurse (logNurses), the greater your chance of acquiring an infection during your hospital stay.

Through analyzing this data collected by SENIC from 1970 to 1976, we were able to identify the greatest potential risks to patients seeking healthcare treatment in hospitals, which involve length of stay, routine culturing ratio, and number of nurses. This data would have allowed the researchers to share their results with the respective hospitals in the hopes of effectively reducing the rates of nosocomial infections by, for example, trying to lower the total amount of time patients spend in the hospital. This information is not only helpful for hospitals but also for patients, so they know exactly what to watch out for when they visit a hospital.

Since this data is quite outdated, an updated study with a greater number of hospitals would be invaluable in determining the greatest infection risks today. Medical practices and protocols have changed a lot in the past 30 years, so this exact model may not be able to reliably predict infection risk in today's hospitals. A future avenue for the study could include surveying more hospitals in a wide variety of regions, and including more predictors such as number of non-invasive surgeries, and maybe the original cause of the patient's visit to the hospital. These could be great predictors of the risk of acquiring a nosocomial infection.

A SAS Code