Efficacy of Nosocomial Infection Control

Authors: Mark Makris, Sully Fagbemi

The main focus of this study was to show how infection surveillance and control programs contribute to the overall nosocomial infection is U.S. hospitals. The data collected and analyzed was provided by Dr. Junyong Park and is believed to be from 110 hospitals in the U.S out of 338 hospitals surveyed. We have broken down the key variables that were analyzed in the experiment as follows:

KEY

F2 - Length of Stay
F3 - Age
F4 - Infection Risk
F5 - Routing Culturing Risk
F6 - Routine Chest X-Ray Ratio
F7 - Number of Beds
F8 - Medical School affiliation
F9 - Region
F10 - Average daily census
F11 - Number of nurses
F12 - Available facilities and services
F13 - log10(length of stay/F2)

Other Terms used in analysis of our data set includes:

C_p - Mallows's Candidate Predictor; Used to evaluate the fit of a regression model by using the OLS method. Low values typically indicate better precision of the model being examined.
R^2 - Coefficient of Determination; this is the degree of variance in the response variable that is determined from the explanatory variables.

Statistical Model Chosen & Interpretation In Terms of The Model

In trying to control the experimentwise error rate, we have chosen a 95% confidence interval. Forward selection with a significance value of .05 for variable additions to the regression model

Rationale Behind The Model, Assumptions Made, & Reasoning.

Assumptions: As for any Multiple Regression analysis, we have assumed the following:

1. Expected Values of the Errors is Zero.

2. The Errors all have the same Variance, i.e. $Var(\varepsilon i) = \sigma\hat{}2$ for all i's

3. The errors are independent of each other

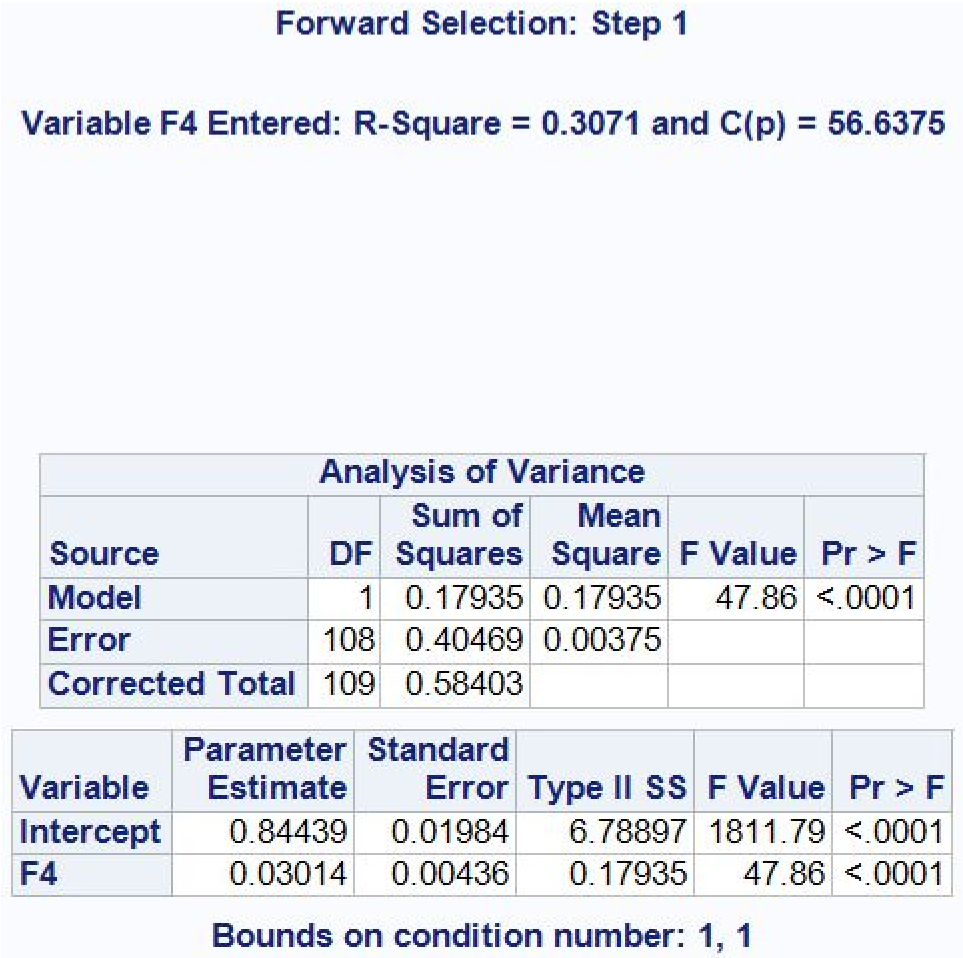
4. The errors are all normally distributed.

We looked at a couple of different ways to create our model including forward, backwards, and stepwise selection. Then we also looked at different significance levels to test and decided on the model with the highest r squared and had a normally distribution on the residuals to make sure we weren't missing any patterns.

Descriptive Summary of Data

We decided to test the nosocomial infection rate by evaluating the average length of hospital stay in the data. Forward Selection was chosen as the preferred analytical procedure in this study and we have shown the evolution in how the most effective model was built. Starting with the response variable (Log10[Length of Stay]), different independent variables were added in succession till we got the best R squared value.

The steps of the process have been shown below:



Forward Selection: Step 1

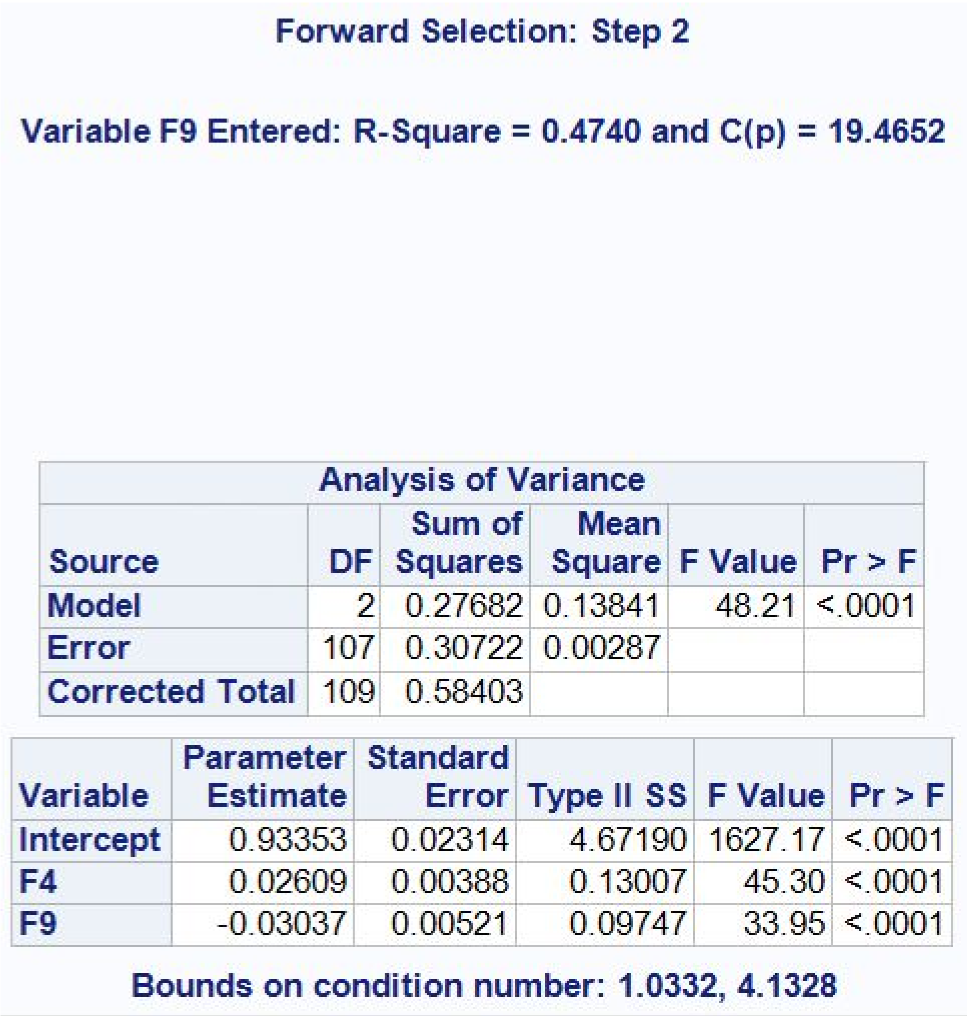
Variable F4 Entered: R-Square = 0.3071 and C(p) = 56.6375

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.17935	0.17935	47.86	<.0001
Error	108	0.40469	0.00375		
Corrected Total	109	0.58403			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.84439	0.01984	6.78897	1811.79	<.0001
F4	0.03014	0.00436	0.17935	47.86	<.0001

Bounds on condition number: 1, 1

After the first step we see the r squared values shows that 30.71% of the residual values or randomness is explained by the first variable(infection risk). That is a low R squared value and we would like it to be higher. Also the C(p) is quite large which means that it is not necessarily a good fit to just have the one variable explaining the length of stay. Variable inserted here is the **Infection Risk rate**.



Forward Selection: Step 2

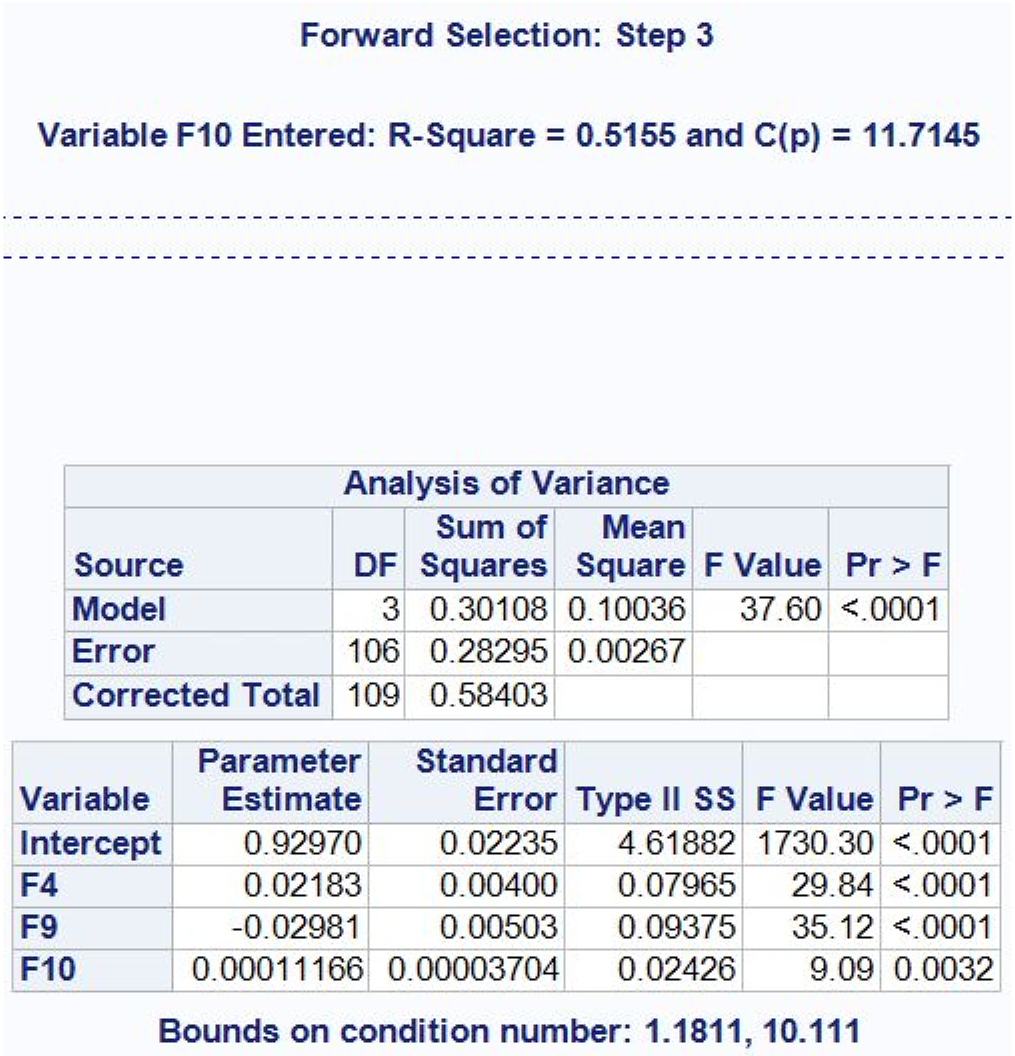
Variable F9 Entered: R-Square = 0.4740 and C(p) = 19.4652

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.27682	0.13841	48.21	<.0001
Error	107	0.30722	0.00287		
Corrected Total	109	0.58403			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.93353	0.02314	4.67190	1627.17	<.0001
F4	0.02609	0.00388	0.13007	45.30	<.0001
F9	-0.03037	0.00521	0.09747	33.95	<.0001

Bounds on condition number: 1.0332, 4.1328

After the second step which involved inclusion of the **Region** variable, we see the r squared values shows that 47.40% of the residual values or randomness is explained by the two variables(infection risk,region) which is 16.69% more than with just the one variable(infection risk). But the C(p) is still large which means that it is not necessarily a good fit to just have the two variables explaining the length of stay. The r squared is also not where we would like it to be, since it does not explain half the randomness.



Forward Selection: Step 3

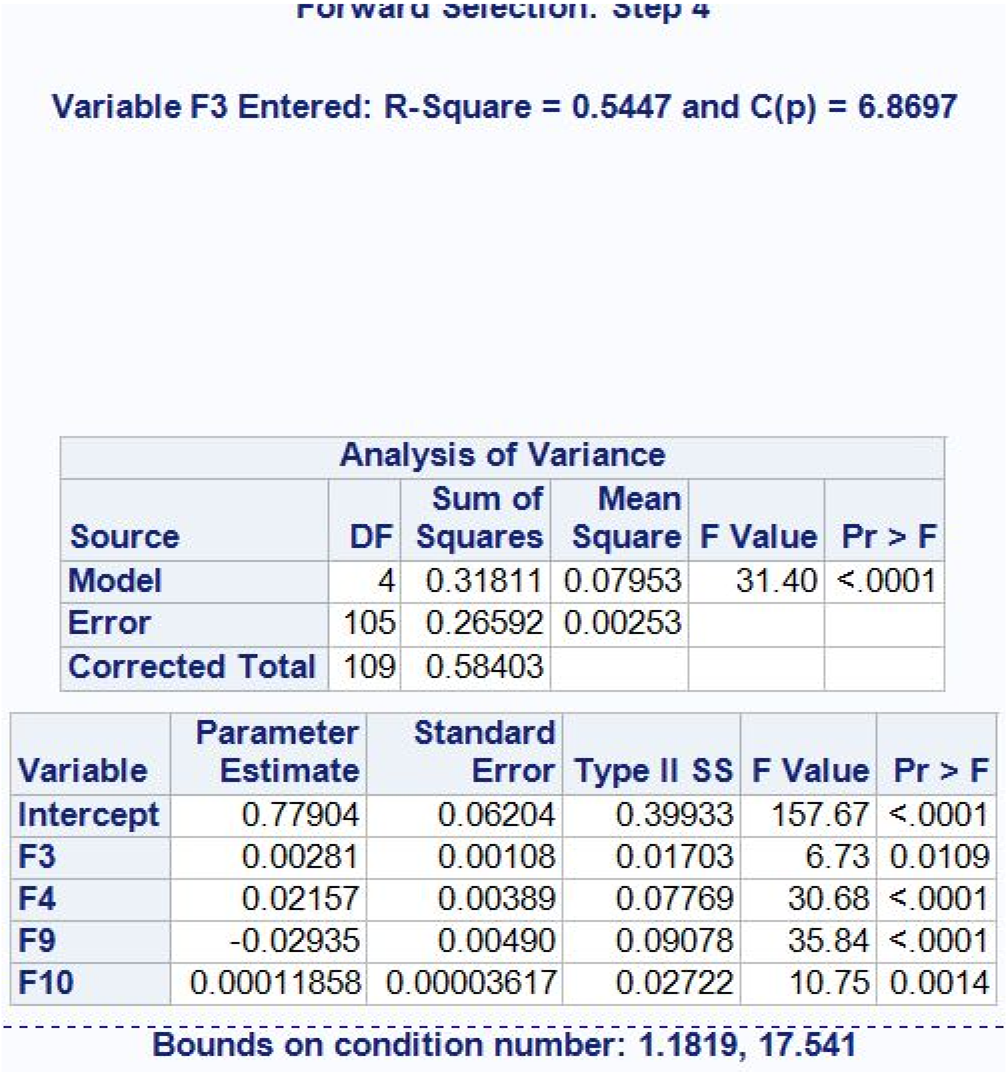
Variable F10 Entered: R-Square = 0.5155 and C(p) = 11.7145

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.30108	0.10036	37.60	<.0001
Error	106	0.28295	0.00267		
Corrected Total	109	0.58403			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.92970	0.02235	4.61882	1730.30	<.0001
F4	0.02183	0.00400	0.07965	29.84	<.0001
F9	-0.02981	0.00503	0.09375	35.12	<.0001
F10	0.00011166	0.00003704	0.02426	9.09	0.0032

Bounds on condition number: 1.1811, 10.111

After the third step, the **Average Daily Census** variable was included, and we see the r squared values shows that 51.55% of the residual values or randomness is explained by the three variables(infection risk,region,average daily census) which is 4.15% more than with just the two variables(infection risk,region). But the C(p) is still large which means that it is not necessarily a good fit to just have the three variables explaining the length of stay. Although now half of the randomness is explained by the model.



Forward Selection: Step 4

Variable F3 Entered: R-Square = 0.5447 and C(p) = 6.8697

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.31811	0.07953	31.40	<.0001
Error	105	0.26592	0.00253		
Corrected Total	109	0.58403			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.77904	0.06204	0.39933	157.67	<.0001
F3	0.00281	0.00108	0.01703	6.73	0.0109
F4	0.02157	0.00389	0.07769	30.68	<.0001
F9	-0.02935	0.00490	0.09078	35.84	<.0001
F10	0.00011858	0.00003617	0.02722	10.75	0.0014

Bounds on condition number: 1.1819, 17.541

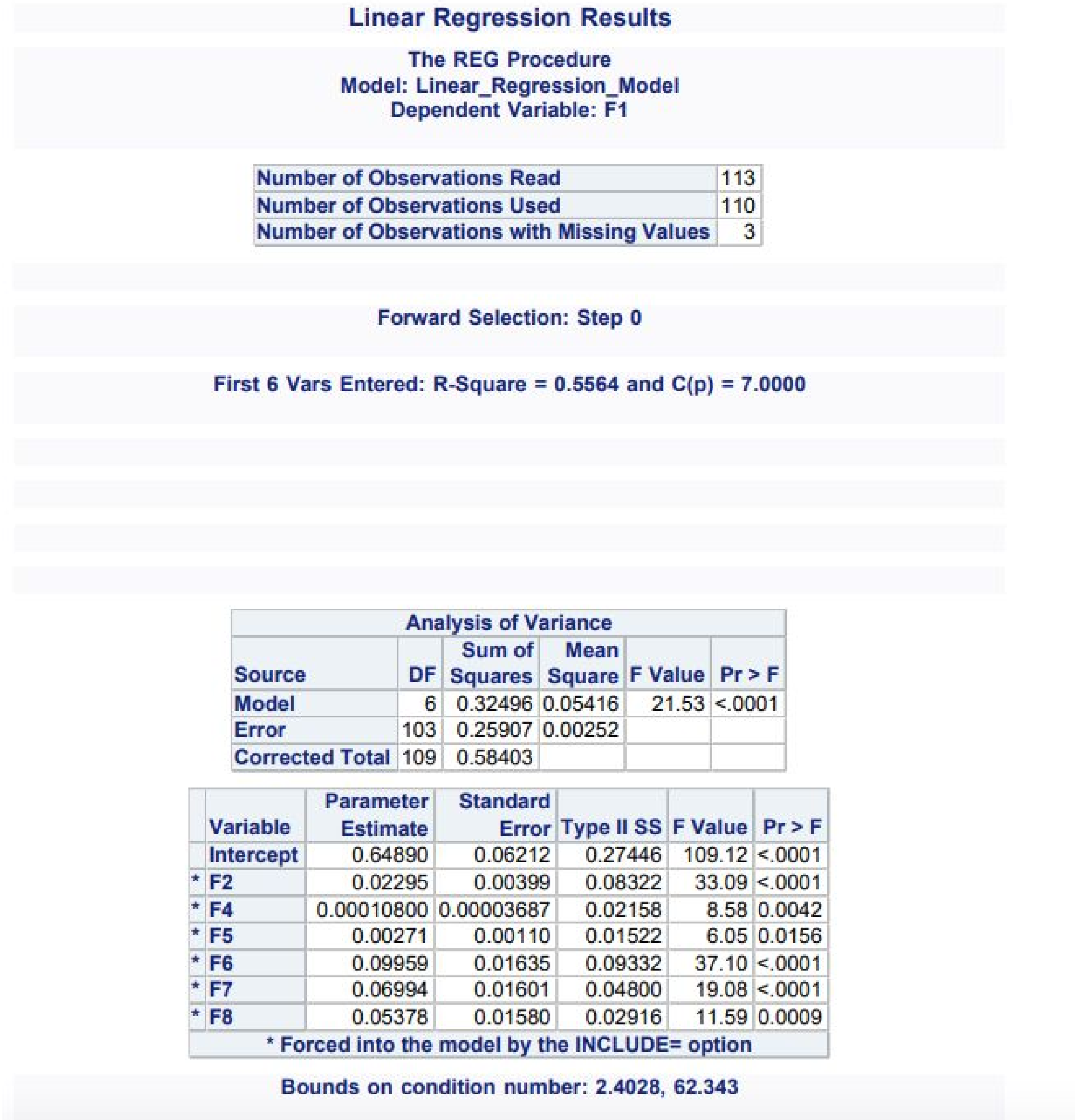
After the fourth step which involved the addition of the **Age** variable, we see the r squared values shows that 54.47% of the residual values or randomness is explained by the four variables(infection risk,region,average daily census,age) which is 2.92% more than with just the three variables(infection risk,region, average daily census). We also see that the new variables are not adding much to the explanation of the length of stay because of the small increase in r squared. The C(p) is actually somewhat small which means that it is a good fit to explain the length of stay.

Regression formula(variables):
F13 = 0.77904 + 0.02157F4 - 0.02935F9 + 0.00011858F10 + 0.00281F3

Regression formula(named variables):
log10(length of stay) = 0.77904 + 0.02157(infection risk) - 0.02935(region) + 0.00011858(average daily census) + 0.00281(age)

Since Region is an explanatory variable in our model, we further expanded our regression function to each individual Region, to evaluate the contribution of each region in the overall length of stay. The results of the regression is displayed below

*The following F values are not the same as the rest of the document for just this image.



Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: F1

Number of Observations Read	113
Number of Observations Used	110
Number of Observations with Missing Values	3

Forward Selection: Step 0

First 6 Vars Entered: R-Square = 0.5564 and C(p) = 7.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	0.32496	0.05416	21.53	<.0001
Error	103	0.25907	0.00252		
Corrected Total	109	0.58403			

	Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
	Intercept	0.64890	0.06212	0.27446	109.12	<.0001
*	F2	0.02295	0.00399	0.08322	33.09	<.0001
*	F4	0.00010800	0.00003687	0.02158	8.58	0.0042
*	F5	0.00271	0.00110	0.01522	6.05	0.0156
*	F6	0.09959	0.01635	0.09332	37.10	<.0001
*	F7	0.06994	0.01601	0.04800	19.08	<.0001
*	F8	0.05378	0.01580	0.02916	11.59	0.0009

* Forced into the model by the INCLUDE= option

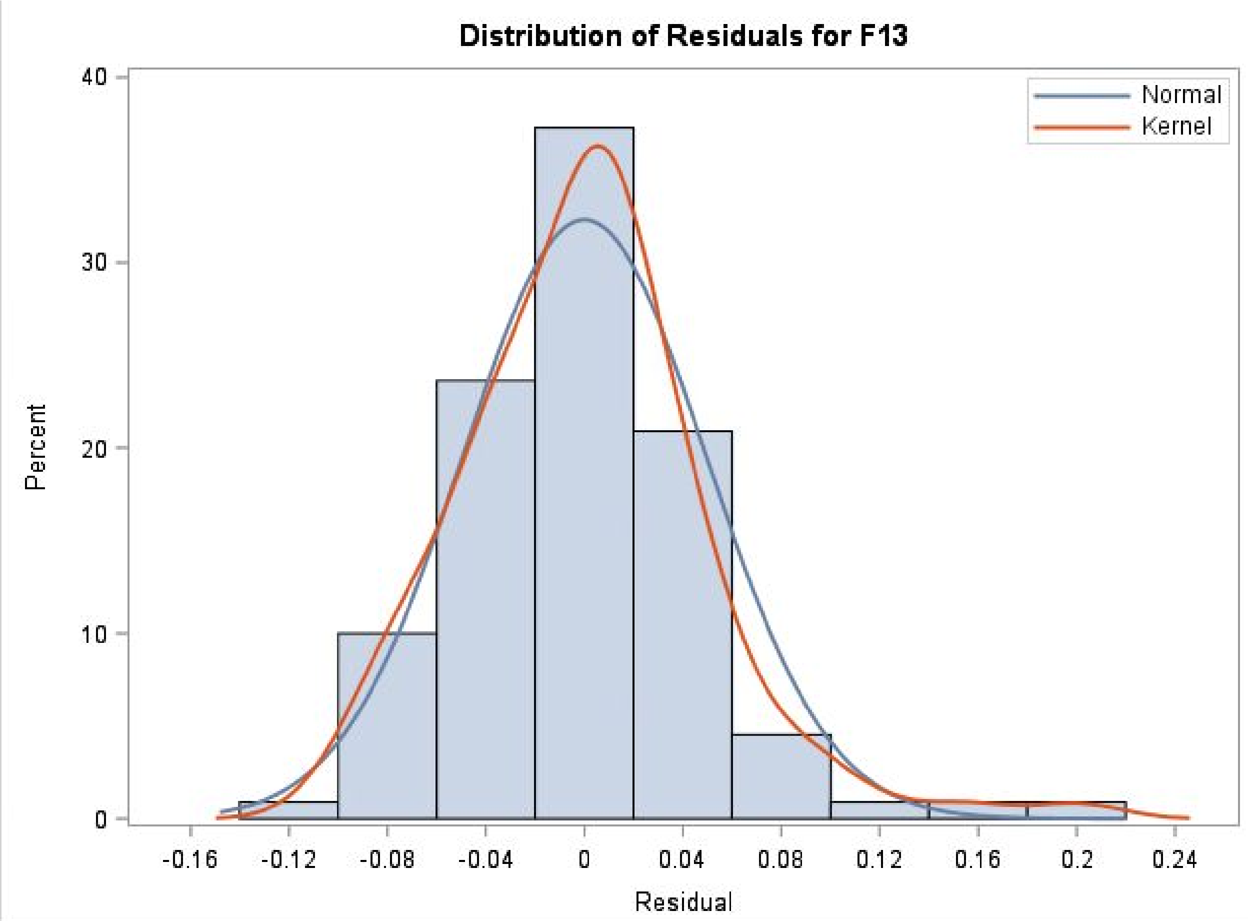
Bounds on condition number: 2.4028, 62.343

Where F2 is now the Infection Risk, F4 is Average Daily Census, F5 is the Age, F6 is NE Region, F7 is NC Region and F8 is S Region, and we have relaxed the W region because when NE, NC and S all equal 0, we assume this represents the W Region. We see a slight change has occurred in the Regression function due to the collinearity in the explanatory values.

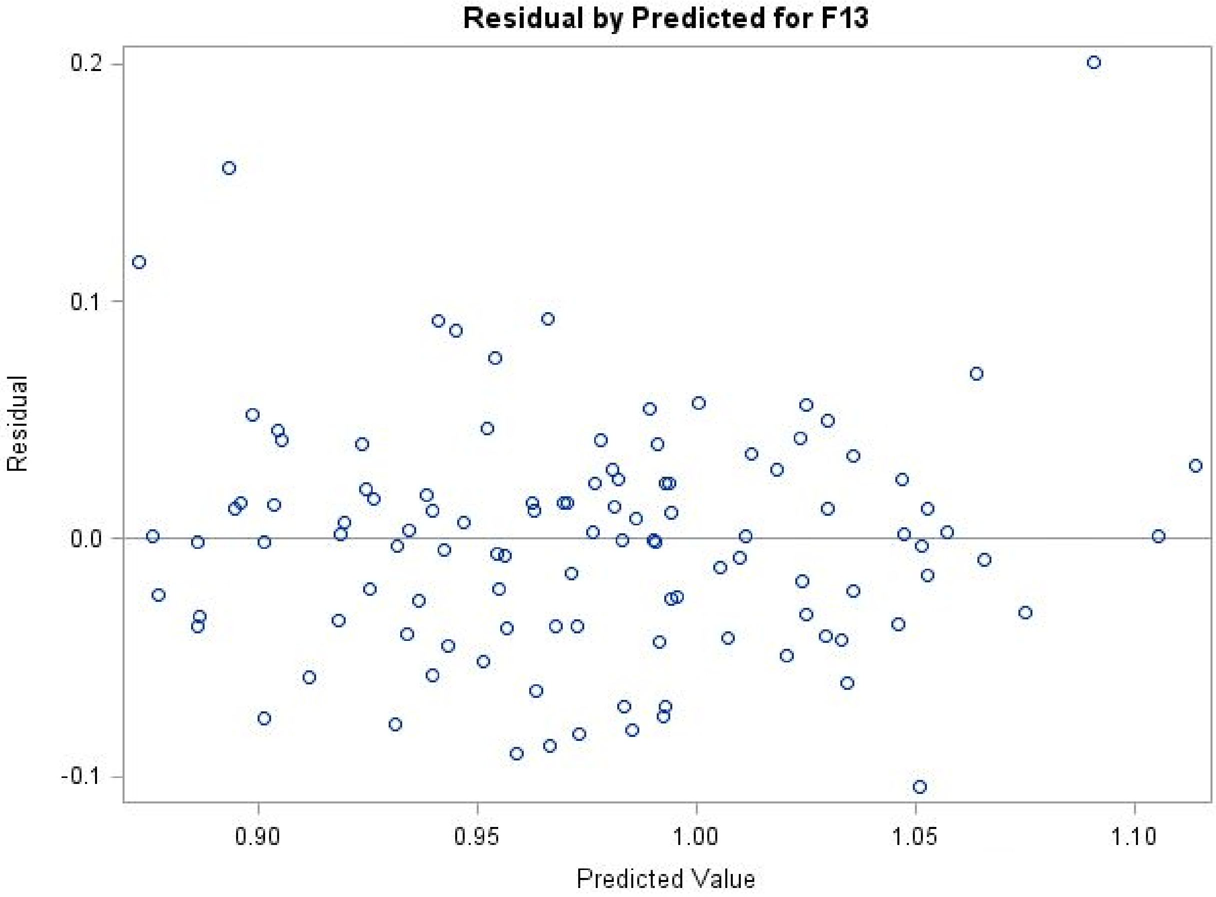
Thus, the expanded regression function is now:

log10(length of stay) = .64890 + 0.02295(infection risk) + 0.00010800(average daily census) + 0.00271(age) + 0.09959(NE) + 0.06994(NC) + 0.05378(S)

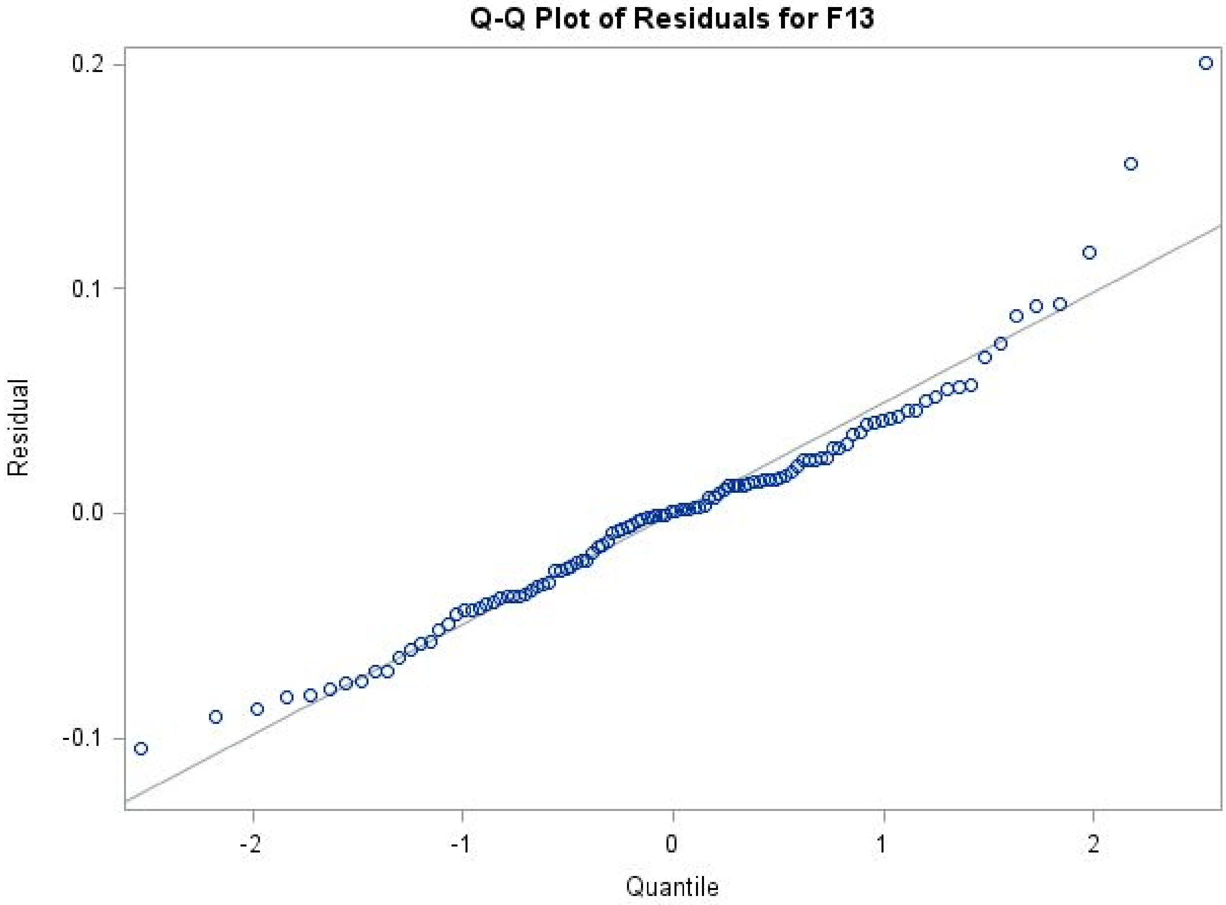
Associate output



This graph shows that the residual are normal with a mean of 0 which is what we are looking for because it means that the overall residuals should end up being 0. Normal distribution is what we want to find in our model because it shows that there is not something shifting the data.



This plot shows that the residuals are evenly spread around 0 and there is no hidden trend that we are missing in our equation. If we saw a pattern of any sort then it would mean that we missed some sort of explanation of the dependant variable with our independent variables.



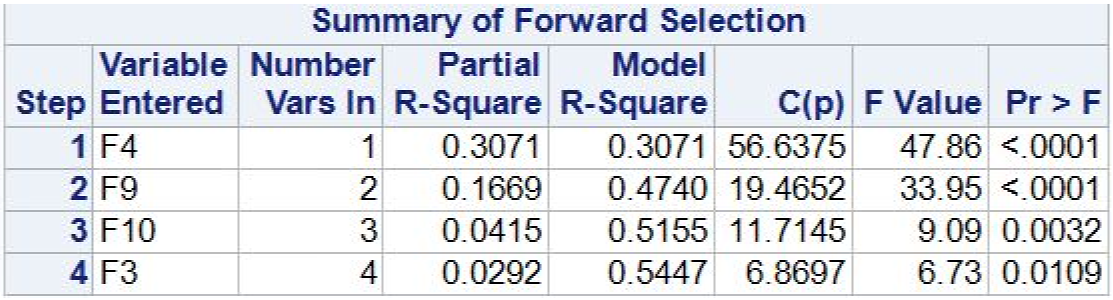
Evaluating the residual plot of the Length of stay, we see it is approximately normally distributed which confirms that the explanatory variables used have enough predictable value on our response variable.

Estimated Values of ‘y’ For Last 3 Rows In The Data

ID = 111: y = .92652, length of day = 2.53
ID = 112: y = 1.12867, length of day = 3.09
ID = 113: y = .92742, length of day = 2.53

Appendix

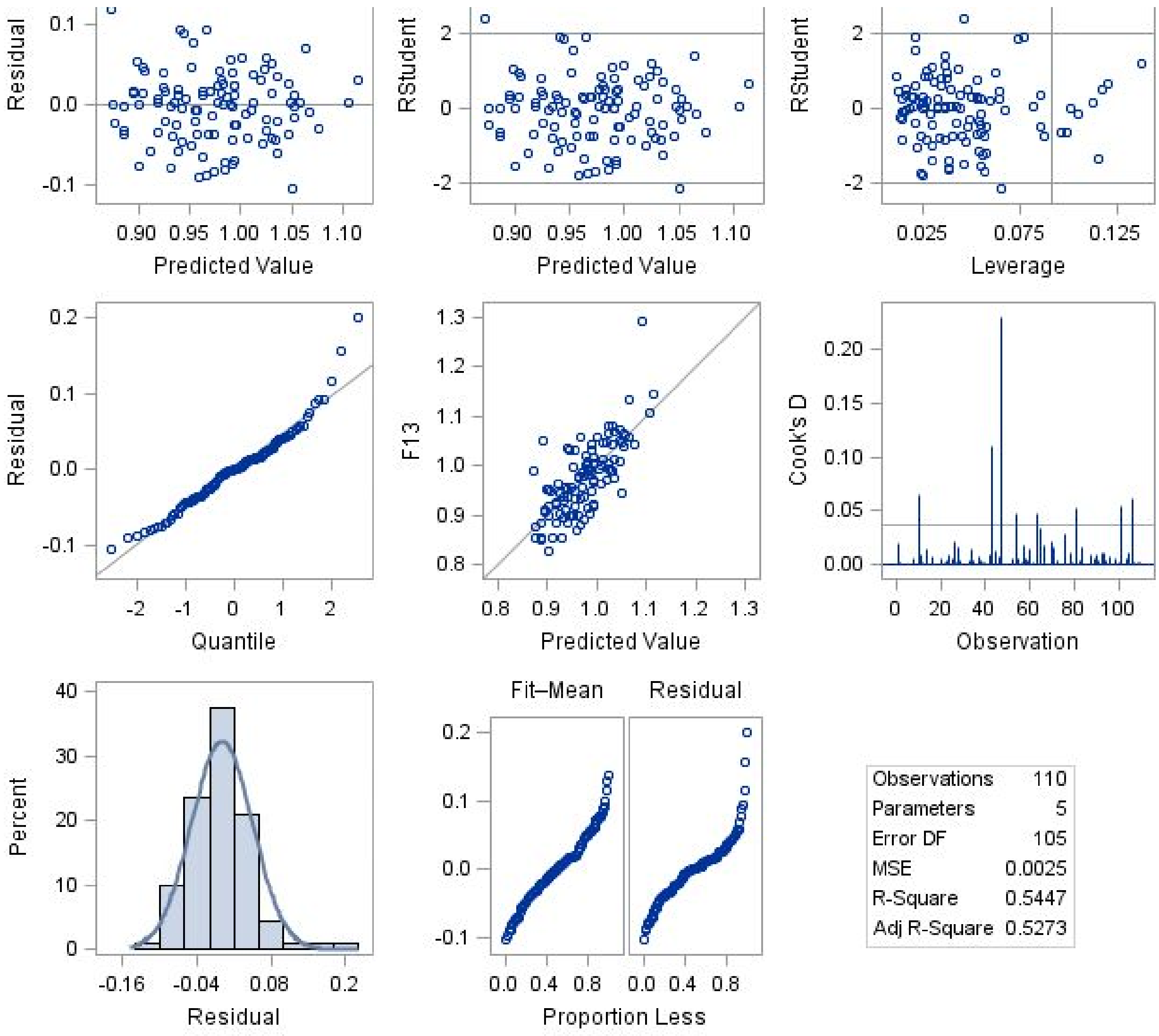
Subsequent data from our full model has been attached below as well as those from the reduced value.



Summary of Forward Selection

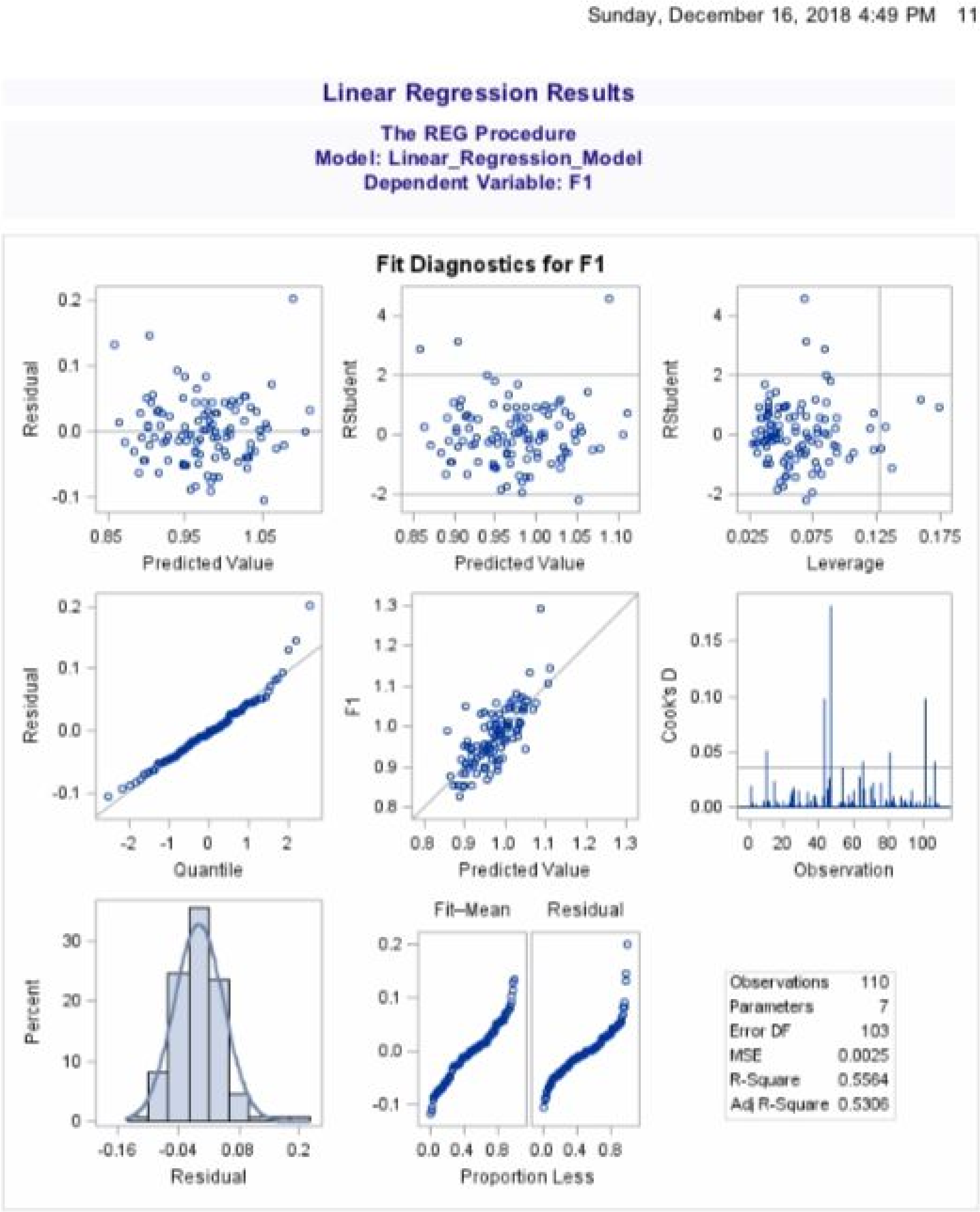
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	F4	1	0.3071	0.3071	56.6375	47.86	<.0001
2	F9	2	0.1669	0.4740	19.4652	33.95	<.0001
3	F10	3	0.0415	0.5155	11.7145	9.09	0.0032
4	F3	4	0.0292	0.5447	6.8697	6.73	0.0109

F4 - Infection Risk
F9 - Region
F10 - Average Daily Census
F3 - Age



Observations	110
Parameters	5
Error DF	105
MSE	0.0025
R-Square	0.5447
Adj R-Square	0.5273

For the expanded regression function involving the individual regions, we have the following summary statistics:



Sunday, December 16, 2018 4:49 PM 11

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: F1

Fit Diagnostics for F1

Observations	110
Parameters	7
Error DF	103
MSE	0.0025
R-Square	0.5564
Adj R-Square	0.5306

In both cases, we see significant evidence to suggests the residuals are approximately normal and the linear regression calculated have some explanatory value on the length of stay in hospital.

Contribution

Mark Makris:

First I had to run multiple tests for selection to find the best model. I had to see which selection method and significance level would result in the best possible model for the data. I also described some of the tables for the step the selection model went through to choose the relevant variables. I Calculated the last 3 rows of data as well as what the actual length of days would be for those values. I also added some of the tables and plots to the document that were relevant for us to discuss and explain the results of.

Sully Fagbemi:

Did some reading on Ch. 13, to find out how to start with the building of our initial regression model and the things to consider in making a good model. Came up with the assumptions for our analysis, did some writing on the descriptive summary of our findings and also ran the descriptive data for our expanded regression function as well as analyzing the SAS output. Also helped in addressing the appendix as well as proofreading the article.