Aziz Saries and Tyler Le
STA 108
Dr. Amy T. Kim

I. Introduction

This project aims to predict the average length of stay of patients in a hospital using a multiple linear regression model. The data we have chosen is the "SENIC.csv" file, which contains data from 113 hospitals. We will use age and infection risk as predictor variables for our base model and test to see which additional predictor would be most helpful in this base model. The additional predictors we are comparing will be the ratio of the number of cultures performed to patients without signs or symptoms of hospital-acquired infections, the average number of beds in a hospital, the average number of full-time equivalent registered and licensed practical nurses, and the percent of 35 potential facilities and services that the hospital provides.

II. Summary

As denoted in Figure 2, all explanatory variables are positively correlated with average length of stay. Additionally, there is some multicollinearity between some predictor variables, as seen in Figure 1. However, we will disregard this going forward.

III. Variable Selection

Now, we will compare the four additional predictors to determine which is the best to add to the base model. The criteria we will be using to determine which of these additional predictors is best is the partial R^2 value. The first additional predictor is ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infections, times 100. When this predictor is added to the base model we obtained a partial R^2 value of 0.0117. This means that 1.17% of the error for the base model is reduced when we add our ratio of cultures performed variable. The next additional predictor is the average number of beds in the hospital during the study period. When this predictor was added to the base model, the partial R^2 value was 0.0886. This means that 8.86% of the error for the base model is reduced when we add our variable of average number of beds. The remaining partial R^2 values for average number of full time nurses and percent of facilities provided were 0.0374 and 0.0363, respectively. This means that 3.74% of the error in our base model is reduced when we add our average number of nurses variable, and 3.63% of the error in our base model is reduced when we add our percentage of potential facilities variable. Due to the average number of beds in the hospital having the highest partial R^2 value of 0.0886, this will be the best variable to add to our base model. This conclusion is also supported on Page 5, where we calculated multiple F-tests and concluded that our beds variable had a statistically significant effect on average length of stay and had the greatest F* value among all four potential additional variables. This model that includes our best additional variable will be referred to as proposed model 1.

IV. Model Comparison and Fit

We will now compare proposed model 1 to another model that uses average number of beds, infection risk, and available facilities and services as predictors. This model will be referred to as proposed model 2. We will compare these two models using adjusted R^2. Proposed model 1 provides an adjusted R^2 of 0.3632. Proposed model 2 provided an adjusted R^2 of 0.3226. Therefore, we conclude that proposed model 1 fits better in terms of adjusted R^2 and will be our final model.

V. Model Diagnostics

When checking the assumption of constant variance, we can see in Figure 4 that roughly ¾ of our data has constant variance but the last ¼ of the data has a much greater variability. Examining further, we can see in and Figure 7 we can see that our Fligner-Killeen test allowed us to conclude that the variance is non-constant. Our normality assumption is satisfied and we can see this looking at Figure 5, because a large portion of our data does not deviate significantly from the line. We see that there are a total of 6 values that have a Cook's Distance greater than our cutoff (4/113) = 0.035. The 6 values are n = [20, 43, 47, 63, 81, 112]. The significant Cook's Distances were [0.04, 0.07, 0.38, 0.08, 0.05, 0.39]. Therefore, by Cook's Distance, these six values are outliers. We found numerous observations with high leverage, an unusual result in average length of stay with usual predictor values. These high leverage values are at n = [2, 8, 10, 11, 13, 20, 21, 24, 40, 43, 46, 47, 48, 52, 53, 54, 63, 65, 75, 78, 81, 85, 93, 104, 106, 107 112], all observations can be found in Figure 11. Using the studentized residuals method, observations 47 and 112 were flagged as outliers. Therefore, we will remove observations 47 and 112 from the model, as they are the points that were flagged by all three tests.

VI Interpretation

We will now interpret this final model's β values and the corresponding confidence intervals, along with the R^2 value. All confidence intervals were conducted using an $\alpha = 0.05$. β0 represents that if there is a hospital with an average age of 0, average risk of infection of 0%, and an average number of beds 0, we expect the average length of stay of the patients to be 4.0245 days. This parameter had a confidence interval of (1.195, 6.854) which can be interpreted as: we are 95% confident that the average length of stay for a hospital with an average age of 0, average risk of infection of 0%, and an average number of beds of 0, will be between (1.195, 6.854) days. Additionally, β1, which represents the value associated with our explanatory variable "age" this value can be explained as: the average change in length of stay when the average age of patients goes up by 1 will be 0.0515, holding risk of infection and number of beds constant. This parameter had a corresponding confidence interval of (0.00069, 0.1023), which can be interpreted as: we are 95% confident that when the average age of a hospital goes up by 1 year, the average increase in length of stay for the hospital will be between (0.00069, 0.1023) days, holding all other predictors constant. β2, which represents the value associated with our explanatory variable "risk" can be explained as: the average change in length of stay when the average risk of infection goes up by 1 will be 0.528, holding all other predictors constant. The corresponding confidence interval for this parameter was (0.347,0.708) which can be interpreted as : we are 95% confident that when the average risk of infection in a hospital goes up by 1%, the average increase in length of stay for the hospital will be between (0.347,0.708), holding all other predictors constant. Finally, the last explanatory variable β3 that corresponds to our "beds" variable which represents: the average change in length of stay when the average number of beds goes up by 1 will be 0.0018, holding all other predictors constant. This parameter's confidence interval of (0.0005, 0.0031) can be interpreted as : we are 95% confident that when the average number of beds in a hospital goes up by 1, the average increase in length of stay for the hospital will be between (0.0005, 0.0031), holding all other predictors constant. This final model generated an R^2 value of 0.3663. This value means that 36.63% of the variation in the average length of stay for a hospital can be attributed to the average age, average risk of infection and average number of beds.

VII. Prediction

Given an average age of 50, risk of obtaining an infection in hospital of 6%, and average number of beds in the hospital being 250, we expect that the average length of stay in the hospital will be 10.2145 days.

VII Conclusion

Something interesting was that infection had the highest correlation, this could be because hospitals with high infection rates would have to keep patients for longer as they treat the infection. One limitation of our model is that it assumes linearity. There could be much stronger connections between our predictor variables and our explanatory variables, but we could be unable to observe them due to the relationship being nonlinear.

Plots, Graphs, and Hypothesis tests

Figure 1

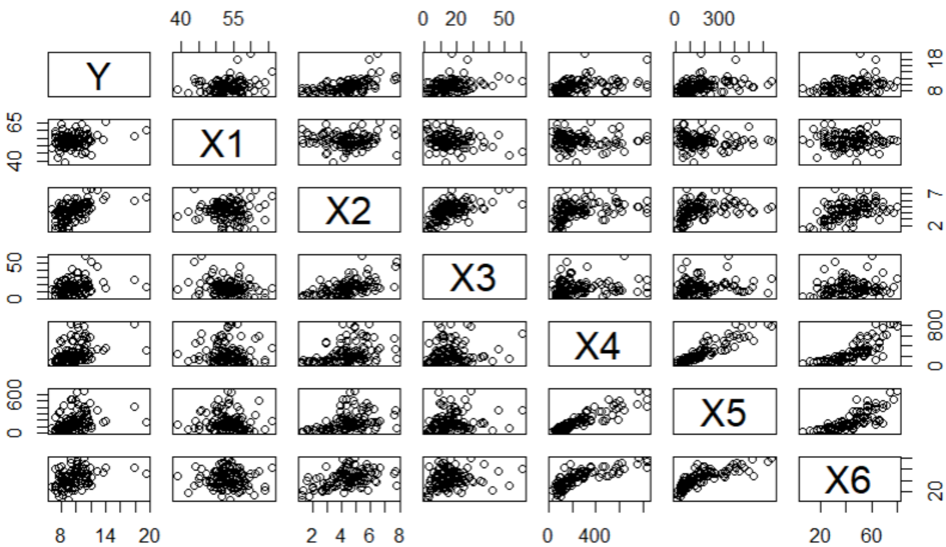

Figure 2

Correlation Matrix:

	Y	X1	X2	X3	X4	X5	X6
Y	1.00	0.1889140	0.5334438	0.3266838	0.4092652	0.3403671	0.3555379
X1	0.1889140	1.00	0.001093166	-0.2258468	-0.05882316	-0.08294462	-0.04045138
X2	0.5334438	0.001093166	1.00	0.5591589	0.35977000	0.39398134	0.41260068
X3	0.3266838	-0.2258468	0.5591589	1.00	0.13972495	0.19889983	0.18513114
X4	0.4092652	-0.05882316	0.35977000	0.13972495	1.00	0.91550415	0.79452438
X5	0.3403671	-0.08294462	0.39398134	0.19889983	0.91550415	1.00	0.78350550
X6	0.3555379	-0.04045138	0.41260068	0.18513114	0.79452438	0.78350550	1.00

Figure 3

F-Tests Tests for β3 = 0 α = 0.05

All Hypothesis Tests : H_0 : β3 = 0 (No relationship between X3 and Y) vs. H_a : β3 ≠ 0 (There is a relationship between Y and X3)

Culturing : F* = 1.287 < F(0.95, 1, 109) = 3.928. Therefore we fail to reject H_0 , and we conclude that the ratio of cultures performed on patients without signs or symptoms of hospital-acquired infections has no significant effect on average length of stay in the hospital at significance α = 0.05.

Bed : F* = 10.59094 > F(0.95, 1, 109) = 3.928. Therefore we reject H_0, and we conclude that the average number of beds in hospital during study period has significant effects on the average length of stay in the hospital at significance α = 0.05.

Nurse : F* = 4.231 > F(0.95, 1, 109) = 3.928. Therefore we reject H_0, and we conclude that the average number of full-time equivalent registered and licensed practical nurses has significant effects on the average length of stay in the hospital at significance α = 0.05.

Facility : F* = 4.116 > F(0.95, 1, 109) = 3.928. Therefore we reject H_0, and we conclude that the percent of 35 potential facilities and services provided by the hospital has significant effects on the average length of stay in the hospital at significance α = 0.05.

Figure 4

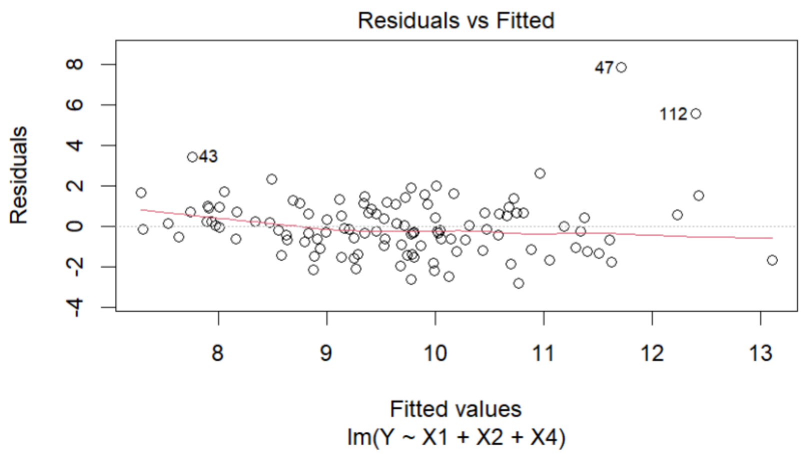

Figure 5

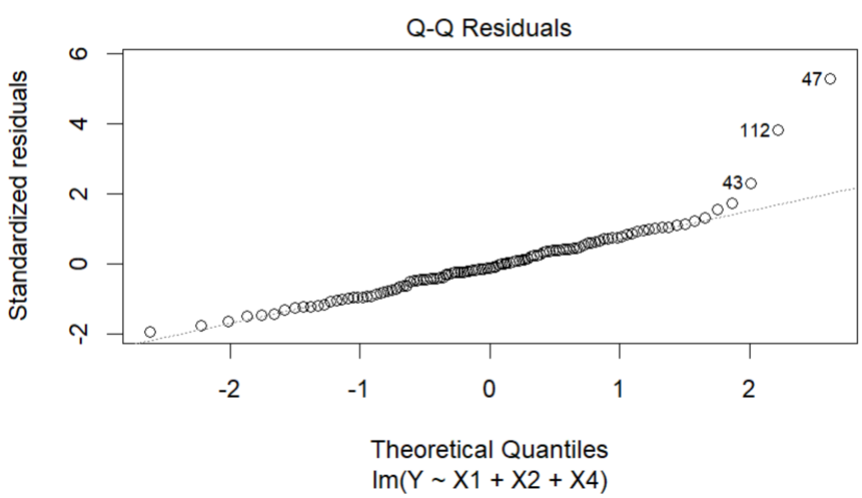

Figure 6

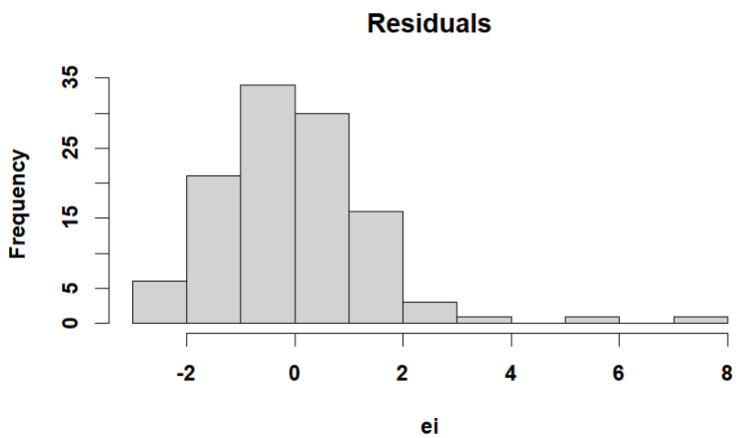

Figure 7

Fligner-Killeen test

Fligner-Killeen test of homogeneity of variances

Fligner-Killeen:med chi-squared = 1.0478, df = 1, p-value = 0.306

Figure 8

Cook's Distance outputs:

	Y <dbl>	X1 <dbl>	X2 <dbl>	X3 <dbl>	X4 <int>	X5 <int>	X6 <dbl>	Group <fctr>
20	9.35	53.8	4.1	15.9	833	519	77.1	Lower
43	11.20	45.0	3.0	7.0	130	56	34.3	Upper
47	19.56	59.9	6.5	17.2	306	172	51.4	Upper
63	7.93	64.1	5.4	7.5	68	49	28.6	Lower
81	10.79	44.2	2.9	2.6	461	196	65.7	Upper
112	17.94	56.2	5.9	26.4	835	407	62.9	Upper

6 rows

Figure 9

Visual for Cook's Distance

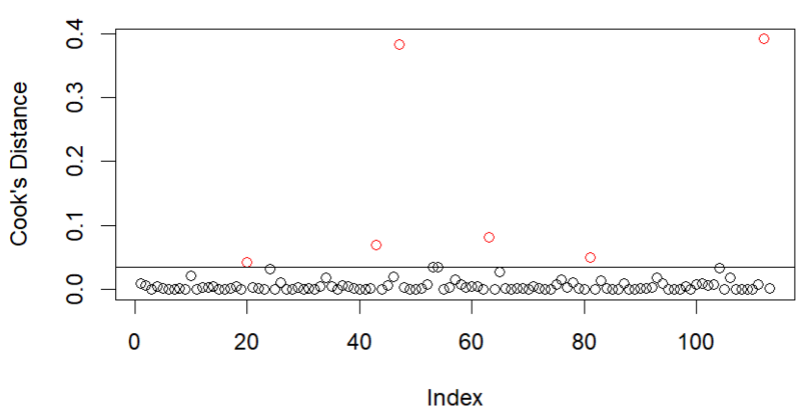

Figure 10

High-Leverage Points

	Y	X1	X2	X3	X4	X5	X6	Group
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<fctr>
2	8.82	58.2	1.6	3.8	80	52	40.0	Lower
8	11.18	45.7	5.4	60.5	640	360	60.0	Upper
10	8.84	56.3	6.3	29.6	85	66	40.0	Lower
11	11.07	53.2	4.9	28.5	768	656	80.0	Upper
13	12.78	56.8	7.7	46.0	322	349	57.1	Upper
20	9.35	53.8	4.1	15.9	833	519	77.1	Lower
21	7.53	42.0	4.2	23.1	95	49	17.1	Lower
24	9.84	62.2	4.8	12.0	600	497	57.1	Upper
40	8.16	60.9	1.3	1.9	73	21	14.3	Lower
43	11.20	45.0	3.0	7.0	130	56	34.3	Upper

	Y	X1	X2	X3	X4	X5	X6	Group
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<fctr>
46	10.16	54.2	4.6	8.4	831	629	74.3	Upper
47	19.56	59.9	6.5	17.2	306	172	51.4	Upper
48	10.90	57.2	5.5	10.6	593	211	51.4	Upper
52	9.23	51.6	4.3	11.6	620	420	71.4	Lower
53	11.41	61.1	7.6	16.6	535	273	51.4	Upper
54	12.07	43.7	7.8	52.4	157	76	31.4	Upper
63	7.93	64.1	5.4	7.5	68	49	28.6	Lower
65	7.78	45.5	5.0	20.9	489	329	48.6	Lower
75	8.45	38.8	3.4	12.9	235	124	48.6	Lower
78	10.23	53.2	4.9	9.9	752	446	68.6	Upper

	Y	X1	X2	X3	X4	X5	X6	Group
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<fctr>
81	10.79	44.2	2.9	2.6	461	196	65.7	Upper
85	8.09	56.9	1.7	7.6	92	61	45.7	Lower
93	8.92	53.9	1.3	2.2	56	14	5.7	Lower
104	13.95	65.9	6.6	15.6	356	182	62.9	Upper
106	10.80	63.9	2.9	1.6	130	62	22.9	Upper
107	7.14	51.7	1.4	4.1	115	19	22.9	Lower
112	17.94	56.2	5.9	26.4	835	407	62.9	Upper

Figure 11

Outliers(Determined by Studentized-Residuals) :

	Y	X1	X2	X3	X4	X5	X6	Group
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<fctr>
47	19.56	59.9	6.5	17.2	306	172	51.4	Upper
112	17.94	56.2	5.9	26.4	835	407	62.9	Upper

2 rows

Figure 12

	2.5%	97.5%

Intercept	1.1950405969	6.853895593
X1	0.0006946865	0.102256525
X2	0.3473343496	0.708012831
X3	0.0005103288	0.003091285

Prediction:

age = 50, risk = 6, bed = 250	Average length of stay = 10.21449

Summaries for Models

Model using X1, X2, X3

Residuals :

Min | 1Q | Median | 3Q | Max
-3.2413 -0.7188 -0.1385 0.7706 7.8325

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.52455	1.91664	0.795	0.4281
Age	0.09152	0.03497	2.617	0.0101
Risk	0.67105	0.13672	4.908	3.24e-06
Culturing	0.02086	0.01839	1.135	0.2590

Residual std. error: 1.588 on 109 degrees of freedom
Multiple R-squared: 0.328, Adjusted R-squared: 0.3095
F-statistic: 17.73 on 3 and 109 DF, p-value: 1.915e-09

ANOVA Table

	df	SS	MS	Fs	Pr(>F)
Age	1	14.604	14.604	5.7885	0.01781
Risk	1	116.356	116.356	46.1188	6.1e-10
Culturing	1	3.248	3.248	1.2874	0.25902

Residuals	109	275.002	2.523		

<u>Model using X1, X2, X4</u>

Residuals :

Min | 1Q | Median | 3Q | Max
-2.8400 -0.9489 -0.1688 0.6872 7.8479

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.6171132	1.7922815	0.902	0.36891
Age	0.0873712	0.0323702	2.699	0.00806
Risk	0.6249491	0.1152349	5.423	3.56e-07
Bed	0.0026122	0.0008027	3.254	0.00151

Residual std. error: 1.525 on 109 degrees of freedom
Multiple R-squared: 0.3802, Adjusted R-squared: 0.3632
F-statistic: 22.29 on 3 and 109 DF, p-value: 2.487e-11

ANOVA Table

	df	SS	MS	Fs	Pr(>F)
Age	1	14.604	14.604	6.2768	0.013712
Risk	1	116.356	116.356	50.0093	1.54e-10
Bed	1	24.642	24.642	10.5909	0.001514
Residuals	109	253.609	2.327		

<u>Model using X1, X2, X5</u>

Residuals :

Min | 1Q | Median | 3Q | Max
-3.0580 -0.8241 -0.1065 0.6724 7.9144

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)

(Intercept)	1.723830	1.843554	0.935	0.3518
Age	0.086906	0.033337	2.607	0.0104
Risk	0.662303	0.120268	5.507	2.46e-07
Nurse	0.002390	0.001162	2.057	0.0421

Residual std. error: 1.568 on 109 degrees of freedom
Multiple R-squared: 0.3454,   Adjusted R-squared: 0.3274
F-statistic: 19.17 on 3 and 109 DF, p-value: 4.67e-10

ANOVA Table

	df	SS	MS	Fs	Pr(>F)
Age	1	14.604	14.604	5.943	0.01639
Risk	1	116.356	116.356	47.350	3.931e-10
Nurse	1	10.397	10.397	4.231	0.04208
Residuals	109	267.853	2.457		

Model using X1, X2, X6

Residuals :

Min | 1Q | Median | 3Q | Max
-2.9998 -0.8594 -0.1823 0.7109 7.7619

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.38646	1.86622	0.743	0.4591
Age	0.08371	0.03325	2.518	0.0133
Risk	0.65845	0.12135	5.426	3.52e-07
Facility	0.02174	0.01071	2.029	0.0449

Residual std. error: 1.568 on 109 degrees of freedom
Multiple R-squared: 0.3448, Adjusted R-squared: 0.3267
F-statistic: 19.12 on 3 and 109 DF, p-value: 4.931e-10

ANOVA Table

	df	SS	MS	Fs	Pr(>F)
Age	1	14.604	14.604	5.9370	0.01645
Risk	1	116.356	116.356	47.3017	3.998e-10
Facility	1	10.125	10.125	4.1162	0.04491
Residuals	109	268.125	2.460		

Proposed Model 2

Residuals :
Min | 1Q | Median | 3Q | Max
-2.9504 -0.9707 -0.1877 0.7496 8.4363

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	6.467380	0.615157	10.513	< 2e-16
Risk	0.647707	0.121908	5.313	5.76e-07
Bed	0.003018	0.001272	2.373	0.0194
Facility	-0.009285	0.016524	-0.562	0.5753

Residual std. error: 1.573 on 109 degrees of freedom
Multiple R-squared: 0.3407, Adjusted R-squared: 0.3226
F-statistic: 18.78 on 3 and 109 DF, p-value: 6.854e-10

Best Model with Outliers Removed

Residuals :
Min | 1Q | Median | 3Q | Max
-2.4276 -0.8017 -0.1076 0.8103 3.0420

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)

(Intercept)	4.024468	1.427286	2.820	0.00573
Age	0.051476	0.025616	2.010	0.04700
Risk	0.527674	0.090971	5.800	6.8e-08
Bed	0.001801	0.000651	2.766	0.00668

Residual std. error: 1.19 on 107 degrees of freedom
Multiple R-squared: 0.3663, Adjusted R-squared: 0.3485
F-statistic: 20.61 on 3 and 107 DF, p-value: 1.291e-10