Predicting the Risk of Infection

Sangeeta Mondal

Objective:

The goal of this study was to see if specific demographic variables helped to predicted the risk of obtaining infection in the hospital. In particular, we hypothesize that there was a higher risk of acquiring an infection in hospitals with higher bed occupancy rate controlling for region, age, census, length, and culture.

Exploring Distributions

Univariate Distributions:

Upon observing the univariate distributions, it was difficult to see what kind of transformations could be applied. However, it was noted that both length and census had heavily right skewed distributions (not pictured) suggesting that these variables might require transformations in the model. Below, a summary of sample characteristics of the hospitals is listed.

Characteristics	**Hospitals (N= 113)**
Region	
Northcentral	32 (28.3)
Northeast	28 (24.7)
West	16 (14.2)
South	37 (32.7)
Age, years, median (IQR)	53.2 (50.9-56.2)
Length, days, median (IQR)	9.42 (8.34-10.47)
Culture, cultures/patients without signs or symptoms of infections * 100, median (IQR)	14.1 (8.4-20.3)
Census, patients/day, median (IQR)	143 (68-252)
Bed_rate, patients/day/number of beds, median (IQR)	.756 (.663-.818)
Beds, beds, median (IQR)	186 (106-312)
Risk, estimated probability of acquiring infection in hospital $\times$ 100, median (IQR)	4.4 (3.7-5.2)

Table 1 : Summary of Sample Characteristics

Bivariate Distributions:

Pearson Correlation Coefficients, N = 113

	id	length	age	risk	culture	xray	beds	msch	region	census	nurses	svcs	bed_rate
id	1.00000	-0.02239	0.03726	-0.21132	-0.26742	-0.16602	-0.03565	0.00607	0.10151	-0.02706	-0.13353	-0.09786	-0.18779
length	-0.02239	1.00000	0.18891	0.53344	0.32668	0.38248	0.40927	-0.29695	-0.49213	0.47389	0.34037	0.35554	0.42841
age	0.03726	0.18891	1.00000	0.00109	-0.22585	-0.01885	-0.05882	0.14513	-0.02043	-0.05477	-0.08294	-0.04045	-0.10961
risk	-0.21132	0.53344	0.00109	1.00000	0.55916	0.45339	0.35977	-0.23303	-0.19228	0.38141	0.39398	0.41260	0.28730
culture	-0.26742	0.32668	-0.22585	0.55916	1.00000	0.42496	0.13972	-0.24274	-0.30828	0.14295	0.19890	0.18513	0.12496
xray	-0.16602	0.38248	-0.01885	0.45339	0.42496	1.00000	0.04582	-0.08670	-0.29634	0.06291	0.07738	0.11193	0.10963
beds	-0.03565	0.40927	-0.05882	0.35977	0.13972	0.04582	1.00000	-0.59118	-0.10563	0.98100	0.91550	0.79452	0.28016
msch	0.00607	-0.29695	0.14513	-0.23303	-0.24274	-0.08670	-0.59118	1.00000	0.10267	-0.61476	-0.58824	-0.52439	-0.23901
region	0.10151	-0.49213	-0.02043	-0.19228	-0.30828	-0.29634	-0.10563	0.10267	1.00000	-0.15274	-0.11268	-0.21153	-0.30379
census	-0.02706	0.47389	-0.05477	0.38141	0.14295	0.06291	0.98100	-0.61476	-0.15274	1.00000	0.90790	0.77806	0.41511
nurses	-0.13353	0.34037	-0.08294	0.39398	0.19890	0.07738	0.91550	-0.58824	-0.11268	0.90790	1.00000	0.78351	0.32695
svcs	-0.09786	0.35554	-0.04045	0.41260	0.18513	0.11193	0.79452	-0.52439	-0.21153	0.77806	0.78351	1.00000	0.28477
bed_rate	-0.18779	0.42841	-0.10961	0.28730	0.12496	0.10963	0.28016	-0.23901	-0.30379	0.41511	0.32695	0.28477	1.00000

Table 2 : Correlation Matrix of All Variables

A correlation matrix was utilized to observe linear relationships amongst the different predictor variables and identify other potentially significant relationships. The most highly correlated relationships are: census and beds, nurses and beds, svcs and beds, census and nurses, svcs and census, nurses and svcs. However, since only census is considered in the model, the correlations mentioned do not make significant contributions to the model. Specific to the model, the correlation matrix suggests that risk is most highly correlated with length and culture. There exists some correlation between risk to xray, beds, census, nurses, svcs, and bed_rate. There is very little correlation between risk and age. Bed_rate is moderately correlated to all variables in the initial model (age, census, etc.).

Simple Linear Regressions Between Risk and Each Predictor:

In order to visualize marginal relationships, risk was plotted against every predictor variable (except for region). Such visualizations were particularly created with a linear regression line and loess smoother to check for nonlinearity. Risk and bed_rate seem to have a linear relationship (not pictured). There are 3 nonlinear relationships: risk and length, risk and culture, risk and census. All seem to require power transformations (figures to the left). Furthermore, there does not seem to be any sort of relationship between risk and age as suggested by both the regression and smoother (figure to the very right).

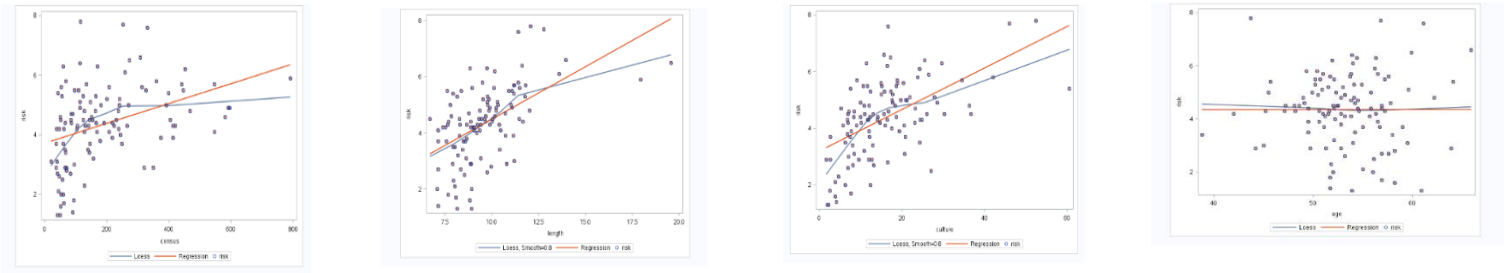

Figure 1: SLR of Risk on Predictors that require Transformations

	Intercept (B0)	Regression Coefficient (B1)	R^2
Age	4.33	0.0003	0
Census	3.72	0.003	0.14
Culture	3.2	0.073	0.31
Length	0.74	0.37	0.28
Bed_Rate	1.89	3.35	0.08
Region			
Region NC	4.33	0.05	0.0003
Region W	4.35	0.03	0.0001
Region S	4.56	-0.63	0.05

Table 3: Results of Regressing Risk on Each Predictor (before transformations)

Consideration of Nonlinear Relationships between Risk and Predictor:

Initial transformations were made based on the marginal regression plots above. Since risk versus length was a simple, monotonic function that resembled a bulge, appropriate logarithmic transformations were made (based on the bulge rule). Since the risk versus culture and risk versus census regression plots had very right skewed tails, appropriate logarithmic transformations were made. Visualizations for risk vs transformations of predictors for marginal relationships are not shown because they are similar to those in residual component analysis shown below. In addition, it can be shown from the table below that the R^2 value increases for select transformations. Furthermore, transformations were not made to risk versus age plot since there was no relationship; no transformation was made to risk versus bed_rate since a good fit had already existed.

	Intercept (B0)	Regression Coefficient (B1)	R^2
Census	0.57	0.76	0.213
Culture	1.24	1.22	0.4
Length	-4.96	4.14	0.3

Table 4: Results of Regressing Risk on Each Predictor (after transformations)

Consideration of Transformations:

To see whether transformations above were appropriate for the MLR model, residual component analysis was carried out. As mentioned above, transformations were based on comparing the loess curves to the regression line. Indeed, the initial transformations (based on SLR models) had also linearized the non-linear relationships in the MLR model (illustrated below). These graphs of the partial regressions were quite similar to those of the marginal regressions.

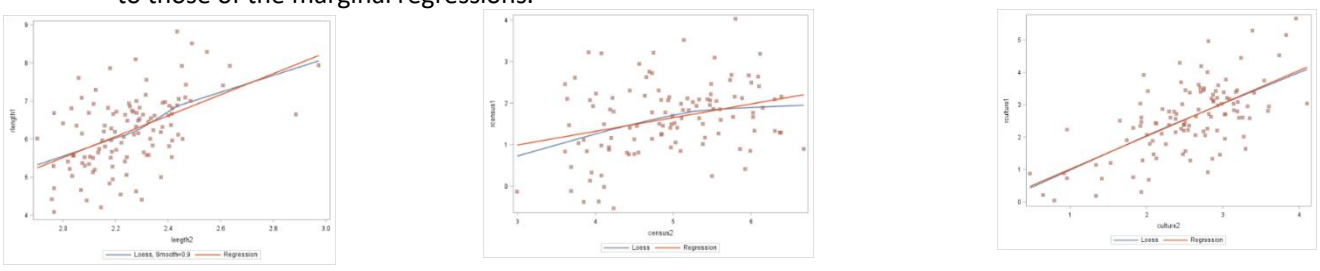

Figure 2: Residual Component Analysis Using Transformations from SLR

Investigation of Possible Interactions:

Interaction terms were investigated based on using varying model selection criterion- Mallow's CP, AIC, and BIC. Terms were created from all possible 2-way combinations of predictor terms. All interaction and lower order terms were included in the MLR model. Forward, backward, and stepwise procedures were carried out for each criteria. In addition, lower order terms were ensured to always be kept in the model. Results from backward procedures were not given as much weight since the resulting models were all very unparsomonious. Interaction terms which appeared most frequently from forward and stepwise procedures based on criterion were included in the model. In this case, bed_rate*regw(i4) and regs*census2(i17) were included in every forward and stepwise procedures so it was included in the model. No output is shown since they all gave very similar results (and was very repetitive).

Check Collinearity:

In the final regression model, VIF values for the interaction and dummy variables were quite inflated (however, the VIF values for dummy variables were ignored). All other predictors in the model had low VIF values (less than 10; not shown here). Interaction terms naturally have higher VIFs due to the fact that they are correlated to other predictors in the model. Hence, the non mean centered model was kept.

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|	VIF
regs	1	-2.18	1.09	-1.99	0.04	44.51
regw	1	3.67	1.20	3.04	0.003	29.76
i4	1	-3.88	1.79	-2.17	0.03	27.27
i17	1	0.48	0.21	2.24	0.02	44.35

Table 5: Model with Selected VIFs

Checking Model Assumptions:

In order to check model assumptions, a residual plot and a QQ plot were used to verify that model assumptions were met. Residuals exhibit random scatter in the residual vs predicted values plot. QQ plot show no heavy tails. However, since there were some outliers, it suggess a not entirely normal distribution. Regardless, for the most part, linear model assumptions were not violated.

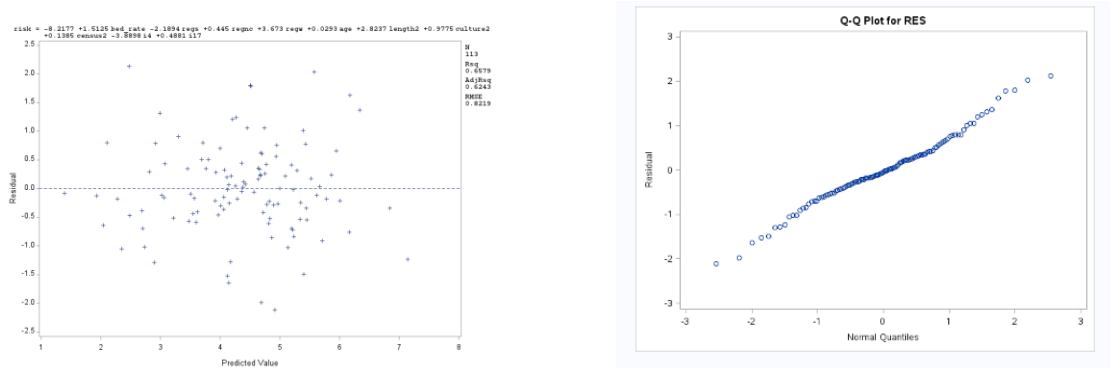

Figure 3: Residual Analysis- Residual vs Predicted Value (Left); QQ Plot (Right)

Outliers and Influential Assumptions:

We see here that observation 101 is an outlier because it has very high leverage, residual, and Cook's Distance values. Although there were other observations with high Cook's D values, they were not investigated as they did not have high residual and leverage values as observation 101 did. A sensitivity analysis was carried out to see if observation 101 was influential on the predictor of interest, bed_rate(occupancy). Since bed_occupancy estimates, standard errors, and p-values remained fairly similar when running the model with and without influential observations (1.51 vs 1.49, .960 vs .967, .12 vs .11 respectively), influential observations were kept in the model.

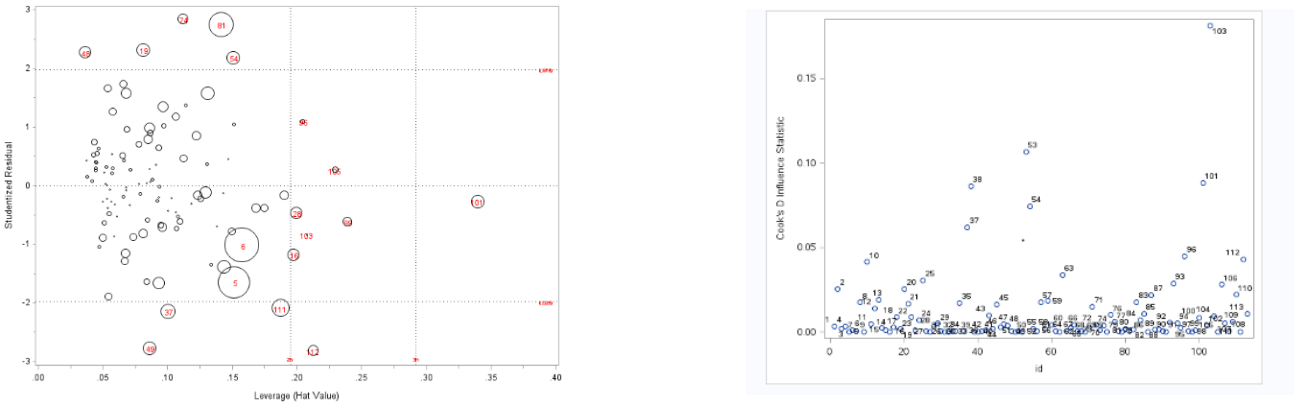

Figure 4: Influence- Bubble Plot of Residual vs Leverage (left); Cook's D vs ID (right)

The Final Model:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-8.21	1.54	-5.31	<.0001
bed_rate	1	1.51	0.96	1.58	0.11
regs	1	-2.18	1.09	-1.99	0.04
regnc	1	0.44	0.23	1.91	0.05
regw	1	3.67	1.20	3.04	0.003
age	1	0.02	0.01	1.49	0.14
length2	1	2.82	0.64	4.35	<.0001
culture2	1	0.97	0.12	7.53	<.0001
census2	1	0.13	0.15	0.88	0.38
i4	1	-3.88	1.79	-2.17	0.03
i17	1	0.48	0.21	2.24	0.02

F-value: <.001 Root MSE: .82 R-Sq: .65 Adj R-Sq: .62

Table 6: Final Regression Model

Below I interpret the final model:

-Intercept-When all predictor variables are set to 0, there is a -8.21% mean infection risk for the northeastern region. (This does not make sense because there is no such thing as a negative infection risk).
-Bed_rate- Controlling for all other variables, a one unit increase in patients/day/number of beds (occupancy ratio) is associated with a 1.51% increase in mean infection risk.
-Regs- When there are 0 patients and all other variables are controlled for, the mean infection risk for the southern region is 2.18 percentage points lower than for the northeastern region. (This may be unrealistic since it is rare for a hospital to be empty).
-Regnc- Controlling for all other variables, the mean infection risk for the northcentral region is .44 percentage points higher than for the northeastern region.
-Regw-When the bed occupancy ratio is 0 and all other variables are controlled for, the mean infection risk is 3.67 percentage points higher for the western region than for than for the northeastern region. (This may also be unrealistic since hospitals would have to be empty for bed occupancy to be 0).
-Age- Controlling for all other variables, a one year increase in age is associated with a .02% increase in mean infection risk.
-Length2 (Log (Length))- Controlling for all other variables, a 1% increase in day length of stay is associated with a .028% increase in mean infection risk.
-Culture2 (Log(Culture))- Controlling for all other variables, a 1% increase in culturing ratio (number of cultures/number of patients without sign or symptoms of infection) is associated with a .0097% increase in mean infection risk.
-Census2 (Log(Census))- Controlling for all the other variables, a 1% increase in average number of patients is associated with a .0013% increase in mean infection risk.
-I4 (bed_rate*regw)- Controlling for all other variables, a one unit increase in bed occupancy ratio is associated with a 3.88% decrease in mean infection risk when comparing the western region to the northeastern region.
-I17 (regs*census2)- Controlling for all other variables, a one unit increase in bed occupancy ratio is associated with a .48% decrease in mean infection risk when comparing the southern region to the northeastern region.

Conclusion

The final model includes the 6 original predictors and 2 interaction terms: bed_rate(bed occupancy ratio), length2 (log length), census2 (log census), culture2 (log culture), region (regs- southern region; regnc- north central region; regw- western region), age, i4 (bed_rate*regw), and i17(regs*census2). Based on the final model, we conclude that a higher bed occupancy rate is associated with a higher mean infection risk. The .65 R^2 value and high F value for the model (P-value = <.001) also substantiates the correlation between bed_rate and infection risk. However, since the p-value for the Bed_rate (occupancy ratio) t-test is not statistically significant (p-val = .12), how much weight should be given to this association should be questioned.

Limitations

Limitations include: 1) the sample size was somewhat small (only 113 observations) 2) outliers were included in model development so whether it may apply to other data is questionable 3) the methods of data collection are unknown so we cannot verify whether there were any errors in data collection.