BIOSTAT 200B: Project 1

Chad Pickering

2/26/2018

Introduction

As part of the SENIC (Study on the Efficacy of Nosocomial Infection Control) project in 1975-1976, a sample of 113 hospitals in the U.S. containing information concerning patients and hospital services were taken. The study only included continental U.S. hospitals regarded as short-term stay, general medical and surgical hospitals not owned by the federal or state government. Hospitals needed a minimum of 50 beds, and exclusions included patients under 18, patients on burn, dialysis, ophthalmology, otolaryngology, oral surgery, newborn, pediatric, rehabilitation, or psychiatry services, and patients hospitalized for less than 24 hours. Infections included were limited to UTIs, surgical wound infections, pneumonias, and bacteria-related maladies. We can use this information to understand whether the risk of acquiring an infection in a hospital is elevated when bed occupancy rates are higher, correcting for a slew of variables, described in the methods section. We will also examine interactions and collinearity between variables, as well as influential points.

Methods

During the study period, researchers measured number of beds, as well as bed occupancy rates, defined as the average number of patients in the hospital during that given period divided by the average number of beds in the hospital (this is made into a percentage for effective analysis). Additionally, we study the ratio of nurses to number of beds, the average age of patients in each hospital, the average length of stay for patients per hospital, the routine culturing ratio (defined as the number of cultures performed compared to the number of patients without signs or symptoms of hospital-acquired infection), and recorded the region of the country of each hospital. Typically, a hospital's average patient is middle-aged (in their 50s) and has a mean length of stay between 8 and 11 days (Table 1). In the sample, a hospital has a few hundred beds on average, a bed occupancy rate hovering between 65% and 80%, and has about 68 nurses per 100 beds on average.

A scatterplot matrix features correlations of each covariate (excluding the categorical region) in the full model that we consider (Figure 1). Average length of stay and routine culturing ratio are both moderately to highly correlated with risk, suggesting that they may be significant predictors of infection risk. Average number of beds, nurse to bed ratio, and bed occupancy rate are mildly correlated with infection risk, too. Length of stay and bed occupancy rate, as well as number of beds and average length of stay are also moderately correlated with each other, which may suggest various collinearity implications, discussed later.

Using residual analysis on each regressor individually (regressing on infection risk), we find that bed occupancy rate and age have thicker tails, but are otherwise normal (and satisfies other conditions). Length of stay has two outliers with high leverage (more on this later) as can be seen in the right tail of the univariate density plot (Figure 1), but is otherwise linear. A transformation here will not be made on all 113 observations to bring a mere two into alignment with the others; we will assess influence of these points to see if down-weighting or removal is appropriate. The residual plot of nurse rate has some slight deviations in the lower fitted values that may suggest a transformation in the independent variable, but all attempts were no better than the original. Both routine culturing ratio and average number of beds have a mild dependency structure in the fitted residuals (concave down behavior), and are corrected with a natural log transformation, which will be implemented in the multiple linear regression model.

Additionally, a single interaction term, average number of beds per region, chosen based on its contribution to the model shown in a sequential sums of squares analysis, is included for completeness and ease of interpretation. No a priori interactions between continuous variables are known, so none of these terms are added to the final model.

As mentioned earlier, length of stay (VIF=2.26) is mildly correlated with three other predictors, average number of beds (r_{ij}=0.446, VIF=7.27), bed occupancy rate (r_{ij}=0.428, VIF=1.65), and routine culturing ratio (r_{ij}=0.313, VIF=1.57). Bed occupancy rate is also mildly correlated with average number of beds (r_{ij}=0.365). The only variance inflation factor (VIF) that is concerning is that of the average number of beds, where the standard error for its beta estimate is inflated by around 2.7. The interaction term in the model does significantly increase the VIFs of its components, region and number of beds. However, this will not significantly impact the inference necessary to answer the research question, so no direct adjustments to the model are made.

There also exist a few influential points in the dataset. Because of its use as an effective summary of both leverage and residual outlyingness, Cook's distance is the final metric used to determine overall influence - a total of nine observations are flagged when using the cut-off of $\frac{4}{100}$, including the two outliers in the design space of length of stay. All insignificant continuous predictors in the full MLR model (Table 2) remain insignificant in the MLR model without the influential points (Table 3); while bed occupancy rate shows signs of greater significance, the p-value does not fall below any reasonable significance threshold. Upon reviewing all nine influential observations, each of them seem to only have one or two outcomes that are "extreme" relative to their respective means or medians. Overall, the full MLR with all 113 observations contains more natural variability and violates very few assumptions, and thus should be analyzed in place of the trimmed version. The two high-leverage points in the length of stay variable, though, should probably be down-weighted using a more advanced method; a log-transformation was attempted to decrease influence, but results showed only mild change in effect (no action was taken to reduce interpretation complexity). It should be noted that in both MLR models shown, the regression diagnostic plots look reasonable albeit evidence of slightly thick tails on the QQ plot.

Conclusion

Whether or not we look at the MLR that excludes influential points, we have evidence that the data does not support the hypothesis; the risk of a hospital-acquired infection is not significantly higher at hospitals with a higher bed occupancy rate. A 1% increase in bed occupancy rate (e.g. 70% to 71%) is associated with a mean increase in infection risk of 0.008% (probability), holding all variables constant, a very small increase which is statistically insignificant (Table 2). However, we can say that a one day increase in the mean length of stay for patients in a hospital increases the mean risk of infection by about 0.28%, adjusting for all other variables, which is significant according to the model with all observations. Additionally, although outside of the scope of the question, an increase in routine culturing ratio is also associated with a significant increase in risk according to Table 2.

A potentially major limitation with our model is that we were not given any information a priori to include any interaction terms motivated by previous work - this may affect the accuracy of inference. Additionally, the hospitals that we can make inference to directly depends on the types of patients accepted or services offered; because of the exclusions made, some of which are mentioned in the introduction, inference has a smaller scope, but it should be sufficient to mention this alongside the report. Lastly, if other research questions were proposed, transformations of certain variables would need to be reconsidered to make interpretation more straightforward.

Figures and Tables

Figure 1. Scatterplot and correlation matrix.

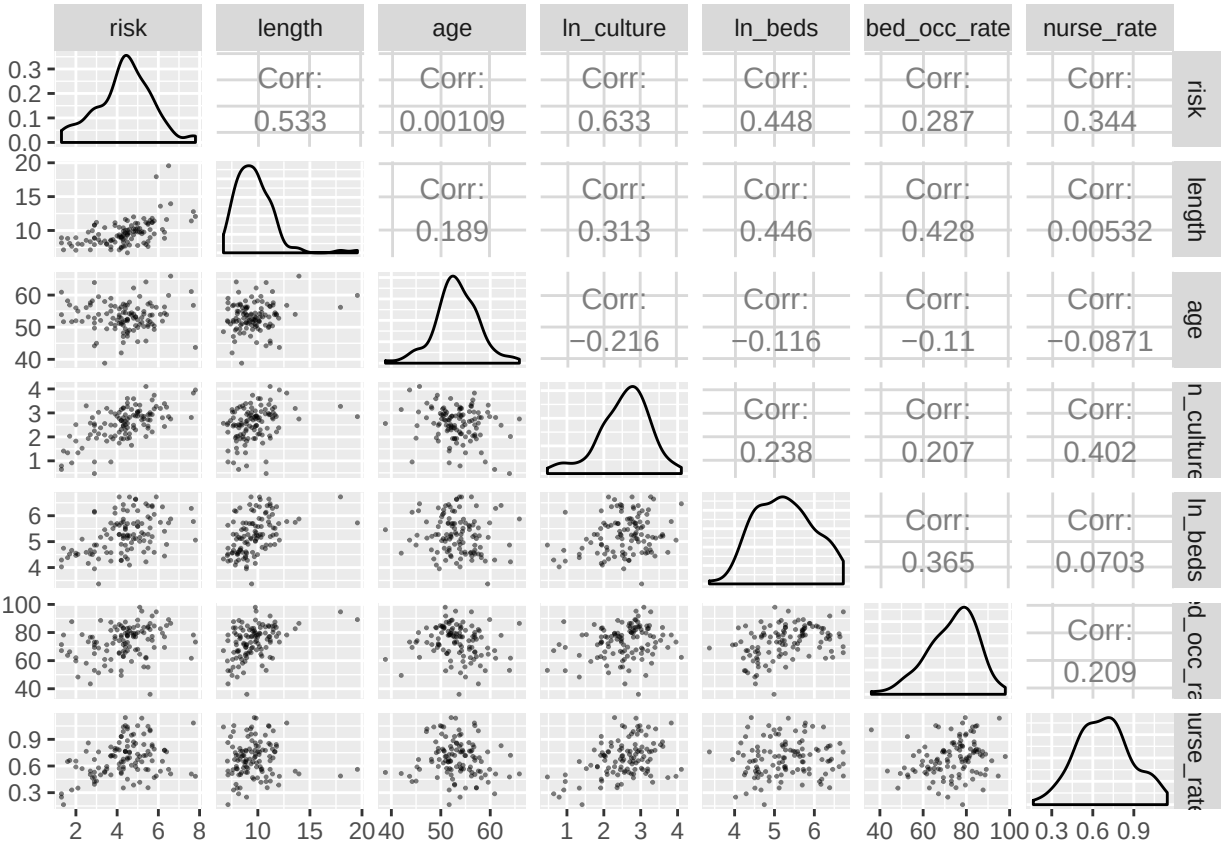

Table 1. Characteristics of the cohort.

Characteristic	Hospitals (n=113)
Infection risk	4.36 ± 1.34
Length of stay (days)	9.42 (8.34-10.47)
Average age of patients	53.23 ± 4.46
Routine culturing ratio	14.1 (8.4-20.3)
Average number of beds	186.0 (106.0-312.0)
Geographic region	
Northeast	28 (24.8)
North central	32 (28.3)
South	37 (32.7)
West	16 (14.2)
Average bed occupancy rate (%)	73.45 ± 11.50
Average nurses per bed	0.68 ± 0.20

Note: All statistics are either mean ± SD, median (IQR), or discrete count (percentage).

Table 2. Multiple linear regression model (n=113).

Term	Unadj. Estimate (95% CI)	P-value	Adj. Estimate (95% CI)	P-value
Intercept			-3.463 (-7.046, 0.118)	0.058
Bed occ. rate	0.034 (0.012, 0.055)	**0.002**	0.008 (-0.010, 0.025)	0.398
Age of patients	0.000 (-0.056, 0.057)	0.991	0.028 (-0.012, 0.069)	0.167
Length of stay	0.374 (0.263, 0.486)	**<0.001**	0.280 (0.155, 0.404)	**<0.001**
Rtn. cult. ratio $*$	1.221 (0.940, 1.502)	**<0.001**	0.919 (0.633, 1.204)	**<0.001**
Number of beds $*$	0.821 (0.513, 1.129)	**<0.001**	-0.027 (-0.610, 0.555)	0.926
Nurses per bed	2.269 (1.103, 3.434)	**<0.001**	0.724 (-0.211, 1.659)	0.128
Region (NE)	Ref	...	...	Ref
Region (NC)	-0.467 (-1.139, 0.206)	0.172	-1.304 (-4.908, 2.230)	0.475
Region (S)	-0.934 (-1.585, -0.283)	**0.005**	-3.367 (-6.860, 0.126)	0.059
Region (W)	-0.480 (-1.294, 0.335)	0.246	1.342 (-2.807, 5.492)	0.523
Beds $*$: Region (NE)			Ref	...
Beds $*$: Region (NC)			0.331 (-0.345, 1.006)	0.334
Beds $*$: Region (S)			0.695 (0.034, 1.355)	**0.040**
Beds $*$: Region (W)			-0.077 (-0.893, 0.738)	0.851

$*$ natural log transformed predictor

Table 3. Multiple linear regression model with influential points removed (n=104).

Term	Adj. Estimate (95% CI)	P-value
Intercept	-5.098 (-8.188, -2.008)	**0.001**
Bed occ. rate	0.012 (-0.003, 0.026)	0.105
Age of patients	0.033 (-0.002, 0.067)	0.062
Length of stay	0.377 (0.245, 0.509)	**<0.001**
Rtn. cult. ratio $*$	0.831 (0.587, 1.075)	**<0.001**
Number of beds $*$	0.057 (-0.455, 0.569)	0.825
Nurses per bed	0.621 (-0.172, 1.413)	0.123
Region (NE)	Ref	...
Region (NC)	-0.114 (-3.299, 3.072)	0.944
Region (S)	-3.063 (-6.139, 0.013)	0.051
Region (W)	3.891 (0.252, 7.530)	**0.036**
Beds $*$: Region (NE)	Ref	...
Beds $*$: Region (NC)	0.107 (-0.485, 0.699)	0.720
Beds $*$: Region (S)	0.617 (0.037, 1.196)	**0.037**
Beds $*$: Region (W)	-0.515 (-1.216, 0.185)	0.148

$*$ natural log transformed predictor
Note: Points removed: 10, 37, 38, 47, 53, 54, 96, 101, 112