SENIC STUDY

Arden Miller

Executive Summary

A regression model was identified that related the estimated probability of acquiring nosocomial infection in a hospital to the levels of five numeric variables (stay, culture, xray, nurses and services) and to the region in which the hospital was located. The data indicates that for fixed levels of stay, culture, xray, and services the mean value of risk is slightly higher for region 4 (West) than for the other 3 regions. It was found that nurses was not required in a model that contained the other explanatory variables and thus provides no additional information about risk. The variables stay, culture, xray and nurses were all found to be related in a positive manner with risk.

1 The Data

This data was collected from hospitals in the United States to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection. The data represents a random sample of 113 hospitals. The following variables were recorded for each hospital.

stay: the average length of stay of all patients in the hospital (in days)

risk: the average estimated probability of acquiring infection in the hospital (in percent)

culture: the ratio of number of cultures performed to number of patients without symptoms of hospital-acquired infections, times 100

xray: the ratio of number of X-rays performed to number of patients without symptoms of pneumonia, times 100

region: a factor indicating geographic region: 1 = Northeast, 2 = North-central, 3 = South, and 4 = West

nurses: the average number of full-time nurses employed at the hospital during the study period

services: the percent of 35 potential facilities and services that are provided by the hospital

	risk	stay	culture	xray	nurses	services
minimum	1.3	6.7	1.6	39.6	14	5.7
lower quartile	3.7	8.3	8.4	69.5	66	31.4
median	4.4	9.4	14.1	82.3	132	42.9
upper quartile	5.2	10.5	20.3	94.1	218	54.3
maximum	7.8	19.5	60.5	133.5	656	80

Table 1: Summaries of the variables in the dataset

Table 1 gives the five number summary for each of the numerical variables in the data set. The values for risk range from 1.3% to 7.8% and are fairly symmetrically distributed around a median value of 4.4%.

Figure ?? contains pairwise scatter plots of the data. These plots indicate that there is a positive relationship between risk and each of the numeric explanatory variables (stay, culture, xray, nurses and services). There appears to be a strong curvilinear relationship between nurses and services. Somewhat weaker relationships exist between stay, culture, and xray. Two hospitals (both from region 1) have unusually large values for stay.

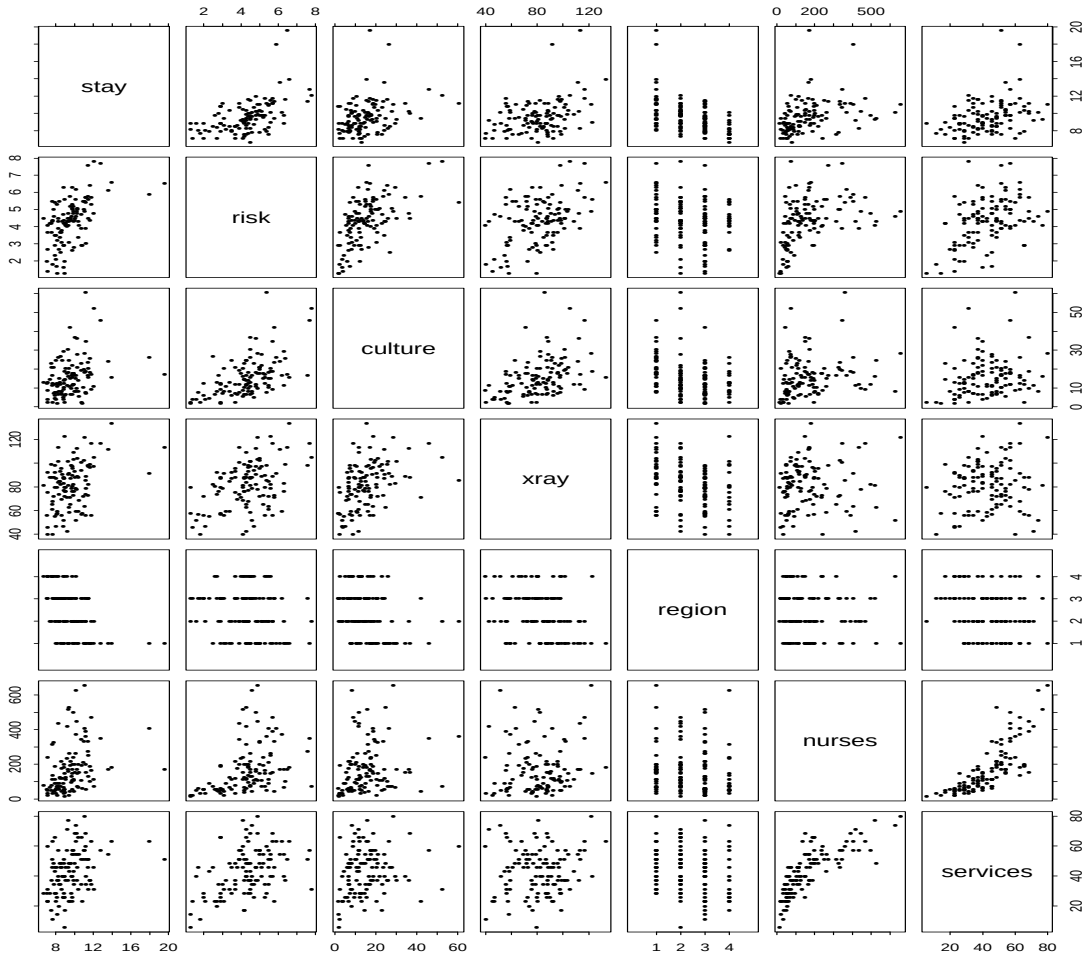

Figure 1: Pairwise Scatter Plots of the Variables in the SENIC Data

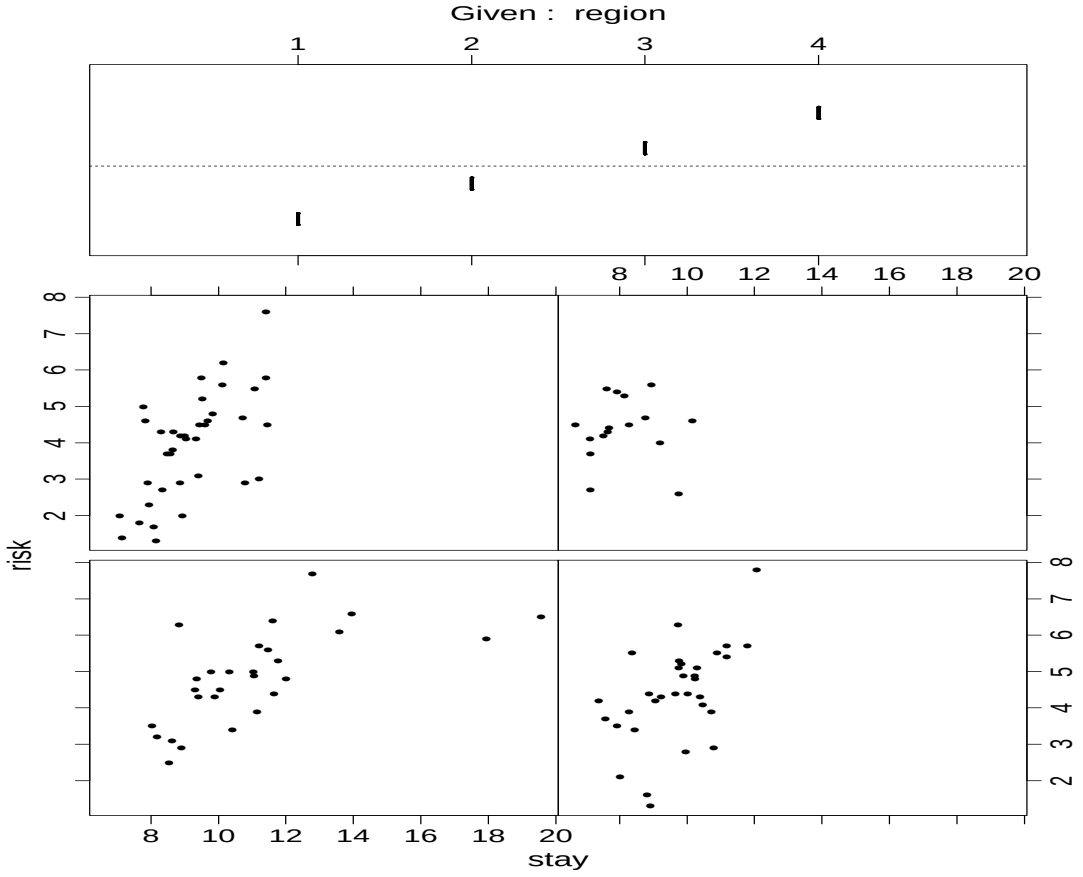

Figure 2: Risk versus Stay for each Region

2 Variables that are Related to Nosocomial Infection

A regression model was used to model risk as a function of the six possible explanatory variables included in the data. It was found that the number of nurses employed did not supply any additional information to the other explanatory variables so it was dropped from the model. Given the strong relationship seen in the pairs plot between nurses and services, it seems likely that nurses is supplying much the same information about risk of infection as services.

For the full data set there was some weak evidence that the way stay affected risk was different for the different regions. Figure **??** contains plots of risk versus stay for each of the four regions. For region 1 (Northeast) the two hospitals that have unusually large values of stay have much smaller values of risk than would be expected by extrapolating from the other hospitals in region 1. These two unusual observations will have a large influence on the estimated relationship between risk and stay for region 1. If we remove these two points there is a strong, positive, linear relationship between risk and stay for region 1. For regions 2 and 3 the relationship between risk and stay appears to be similar to that for region 1 although not as strong. Region 4 has a smaller range of values for stay than the other 3 regions and is therefore much less informative about the relationship between risk and stay. From Figure **??** it would seem that if the 2 hospitals with unusually large values of stay are ignored, there is no evidence that stay has different effects on risk for the different regions. It may be worth investigating these 2 hospitals since their lower than expected values for risk may indicate they have better procedures for controlling nosocomial infection. Given these hospitals are so unusual in terms of their values

		Estimated coefficient	Standard error
Intercept:	Region 1	−1.903	0.774
	Region 2	-1.636	0.719
	Region 3	-1.579	0.677
	Region 4	-0.743	0.645
	stay	0.366	0.077
	xray	0.011	0.005
	culture	0.047	0.010
	services	0.018	0.006

Table 2: Fitted Coefficients for the Regression Model

of stay, the fitted coefficients reported below were obtained ignoring these 2 observations. As a result, the fitted model will not be valid for hospitals with large (greater than 14) values for stay.

Table **??** contains the estimated coefficients for the fitted regression model. The estimated intercept for each region is negative. This is not of concern as it simply represents the value for the fitted regression model when all the numeric regressors are set to 0. For any set of reasonable values for the numeric regressors, the fitted model produces a positive estimate of risk. The differences between the intercepts for the 4 regions indicate the differences in the predicted risk rates for hospitals from different regions that have the same values for all the numeric explanatory variables. Therefore if all other explanatory variables are equal, hospitals in region 4 (West) are predicted to have on average a value for risk that is 0.8 higher than region 3 (South), 0.9 higher than region 2 (North-central), and 1.2 higher than region 1 (North-east). The standard errors for the intercepts suggest that it is quite possible that the intercepts for regions 1, 2 and 3 are really all the same. There is some evidence that the mean risk for region 4 is higher than for the other 3 regions. The coefficients for the numeric explanatory variables indicate the increase in risk that is expected to occur with each unit increase in that variable provided all other explanatory variables are held constant. For example, the model indicates that for each increase of 1 day in the average length of stay of all patients in the hospital, the estimated risk of infection increases by 0.366. A range of reasonable values can be obtained by taking the estimated coefficient and adding and subtracting 2 standard errors. For stay this gives from 0.21 to 0.52 as the set of plausible values for the average increase in risk per unit increase in stay. The estimated coefficients for culture and stay change significantly if observation 8 is removed: culture increases from 0.047 to 0.059 and xray decreases from 0.108 to 0.084.

It should be noted that observational data was used to produce the fitted model for risk. Therefore, we should not infer a causal relationship between the regressors in this model and risk. For example, a hospital that decides to reduce the number of X-rays performed based on this model will not necessarily experience a decrease in risk. What we can say is that the regressors identified are related to risk. The relationship between each regressor and risk could be causal or could be due to some other variable that causes changes in both that regressor and risk.

The fitted model for this data explained slightly less than 60% of the variability observed in risk and therefore approximately 40% cannot be explained using these explanatory variables. It may be worthwhile to investigate other variables that may be linked to risk.

Statistical Appendix

The main goal of this assignment was to identify the numerical variables that effect risk (the probability of nosocomial infection) and to determine if there are differences between regions. I started by fitting the model that contained all the numerical regressors, region, and all the 2-way interactions involving region. These interactions allowed me to assess whether the effects of the numerical regressors were different for different regions. Most of the 2-way interactions were not significant and therefore I tried simplifying the model by eliminating these sequentially from the model. In the end I found that only the stay:region interaction was significant and it was marginal. I also found that nurses was not needed in the model and so it was dropped. The following ANOVA table was obtained for this model:

```
Terms added sequentially (first to last)
             Df Sum of Sq  Mean Sq  F Value      Pr(F)
stay          1  57.30511 57.30511 70.38647 0.00000000
region        3   8.71443  2.90481  3.56791 0.01672565
xray          1  13.28700 13.28700 16.32009 0.00010383
services      1  11.54060 11.54060 14.17503 0.00027867
culture       1  21.72096 21.72096 26.67932 0.00000119
stay:region   3   5.76849  1.92283  2.36176 0.07571207
Residuals   102  83.04325  0.81415
```

The diagnostic plots for this model do not indicate any obvious problems. There are a number of observations that are flagged as being influential by the "influence.measures" command. Two of these (47 and 112) correspond to the points that had usually large values of stay. The trellis plot of risk vs stay conditioned on regions (see Figure **??**) indicated that the apparent interaction between stay and region may be largely due to these 2 points. If these points are removed the p-value for stay:region increases to 0.125. Based on this I decided to remove these points, fit a model with no stay:region interaction, and indicate in my report that the fitted model is not valid for hospitals with large values of stay.

The diagnostic plots for this model are given in Figure **??**. These indicate no obvious problems with the model assumptions.

Given that points (47 and 112) are deleted, then observation 8 shows up as being highly influential (a large Cook's distance). Observation 8 has the largest value for culture of any hospital and if it is deleted the fitted coefficient for culture increases from 0.047 to 0.059 and the fitted coefficient for xray decreases from 0.108 to 0.084. None of the other fitted coefficients change appreciably.

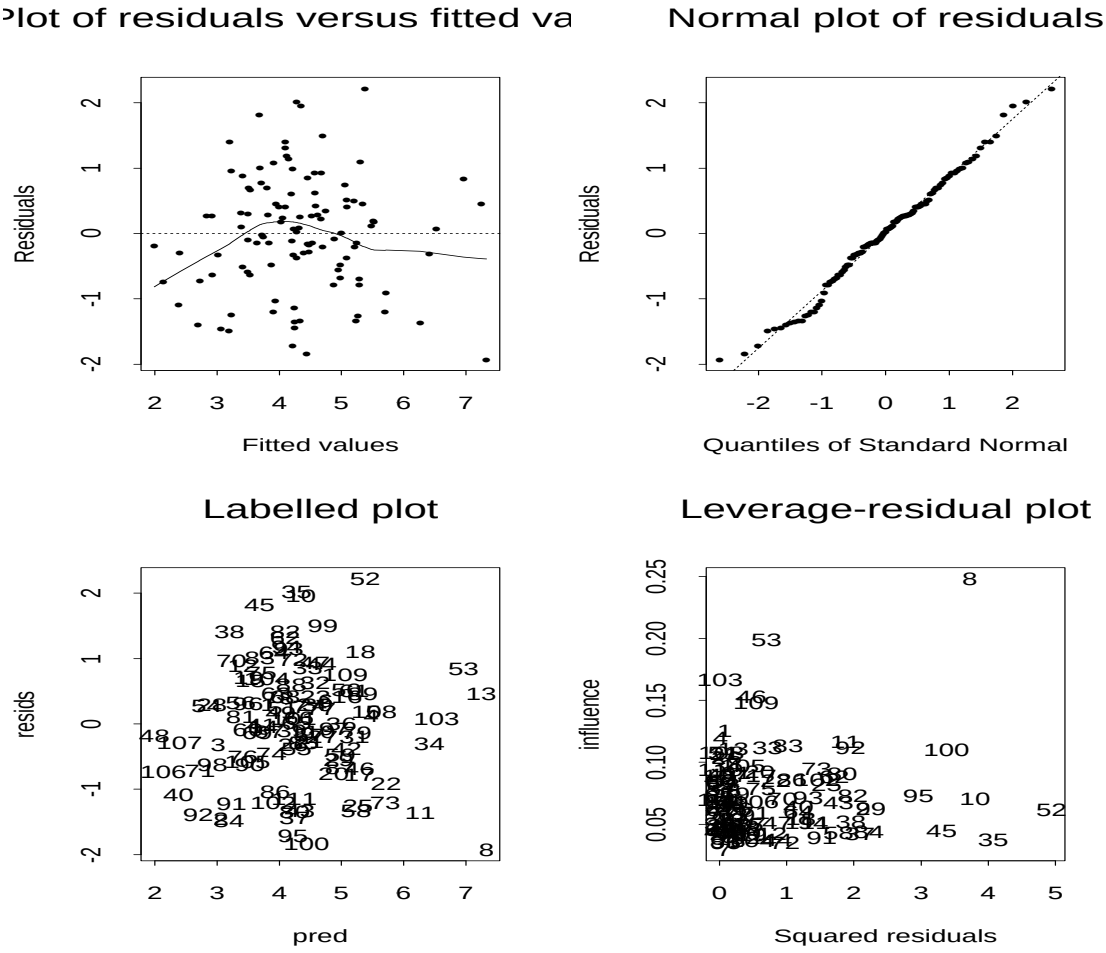

Figure 3: Diagnostic plots for final model

475.330 Assignment 3: Marking Guide

This assignment asks the students to analyze data collected to study the risk of nosocomial infection at hospitals in the United States. They need to identify which of the candidate regressors affect risk (the estimated probability of nosocomial infection), explain what is the impact on risk of these regressors, and discuss whether there are differences between regions.

Report	14 marks
Statistical Appendix	6 marks
Total	20 marks

Report for person investigating nosocomial infection (14 Marks)

This part of the assignment should describe their findings in terms that a non-statistician can understand. They should identify those regressors that they feel impact on risk and discuss the nature of the way these variables affect risk.

- Presentation - 5 marks for a generally well laid out, coherent report. The reader should not have to search for the important parts among a lot of details. Look for good use of graphs and clear explanations of the way the numerical regressors affect risk and a clear discussion of differences between regions. At some point in their report (i.e. an Executive Summary) they should summarize their main findings in a short paragraph. Give: (1) 5 marks for a clear, precise report, that is easy to follow (2) 3 marks if the report is difficult to follow or contains a lot of unnecessary detail, or would be difficult for a non-statistician to understand, (3) 1 mark if it would be very difficult for anyone to understand.
- Content - 9 marks for a report that accomplishes the following
  - Identifies those regressors that affect risk.
  - Explains the impact that these regressors have on risk in non-technical terms.
  - Discusses differences between regions in a sensible manner.
  - Discusses relationships between the regressors. They should identify a strong relationship between nurses and services and comment on it. They may also comment on weaker relationships that are evident between some of the other regressors.
  - They should identify two observations that have unusually large values of *stay*. These points turn out to have a big impact on a possible stay:region interaction (see model answers).
  - I believe any (sensible) model will have a R^2 value of around . They should comment that this indicates that a significant amount of the variability in risk is not explained by the given set of regressors.
  - Indicate that causal relationships cannot be inferred based on this data.

– Give extra credit if they find other interesting aspects of this data that I haven't mentioned.

Note that it is essential that their report covers the first 3 points in this list. An assignment that gets full marks for content **must** accomplish these 3 and **should** accomplish most/some (but not necessarily all) of the others.

Statistical Appendix (6 Marks)

This appendix should outline the reasons that they came to the conclusions they presented in the first part of the analysis. They are not required to give a detailed account of everything they did but they should present a coherent account of their analysis. They should do the following:

1. Identify a suitable model for this data.
   - I intended that they only consider the linear effects for the numerical variables, region, and the interactions involving region and each of the numerical regressors. If they do this their analysis should be very similar to that in the model answer.
   - They may have entertained interactions between the numerical regressors. If they did this they may find 1 or 2 that are significant but these should only have minor effects on the fitted model.
   - They may have tried to transform some of the explanatory variables. Given the amount of scatter in the partial plots I don't think this was necessary but they may get a small improvement by transforming some of the regressors. This is fine as long as their model does not become too complicated. Transforming the regressors may have some impact on the fitted model they choose - they may include log(nurses) instead of services. However, it should have minimal impact on the conclusions reached. Just make sure they interpret the fitted model properly relative to the transformed regressors.
2. Do diagnostic plots for their chosen model
3. Identify influential points and access what effect deleting these points has on the fitted model.

Note:

- Include short comments indicating why a student has lost marks.
- Include a break up of marks into presentation, content, and statistical appendix.