Directions

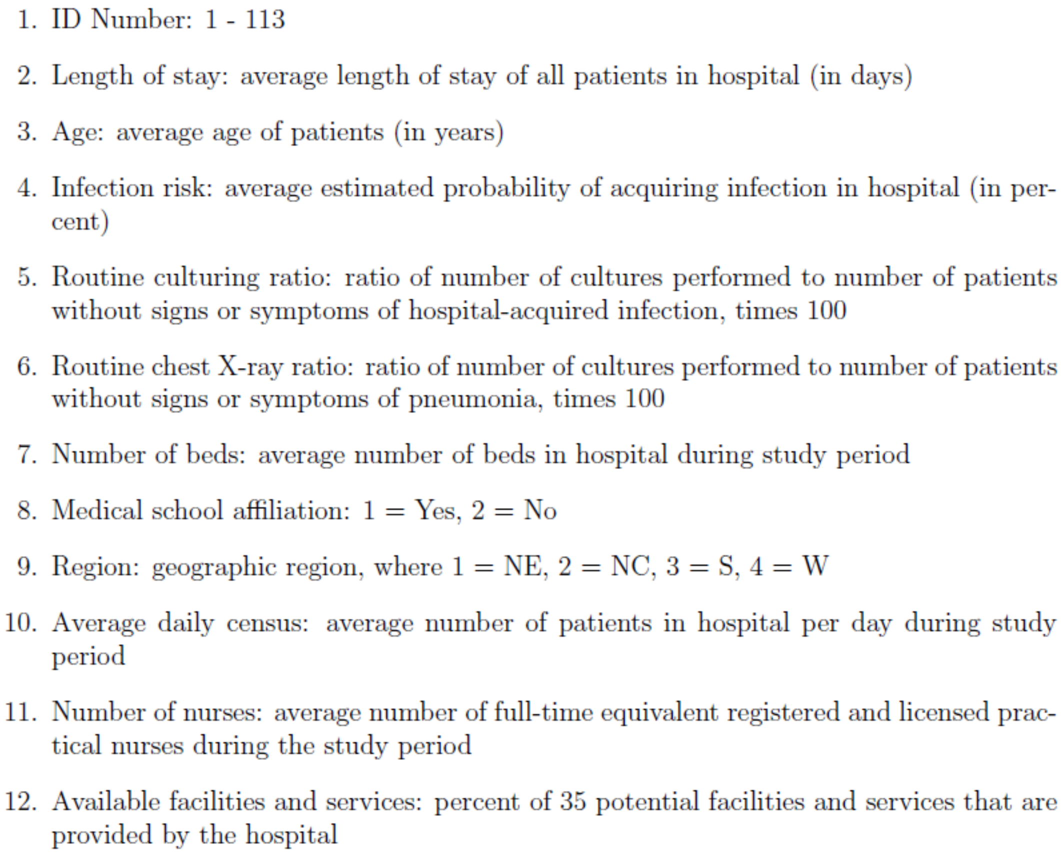
Using `RMarkdown` in `RStudio`, complete the following questions. Launch RStudio and open a new RMarkdown file or use the class RMarkdown template provided and save it on your working directory as a `.Rmd` file. At the end of the activity, save your **pdf** generated from `RMarkdown+Knitr` and submit your homework on the Blackboard.

If you have questions, please post them on the **lesson discussion board**.

All questions are mandatory. Some **R-codes** and **output** from the code have been provided for you.

R codes and output must be clearly shown. Homework submitted after the due date will attract a penalty of 10 points per day after the due date.

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (`SENIC Project`) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. The dataset consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each record in the dataset has an identification number and provides information on 11 other variables for a single hospital. The 12 variables are:



1. ID Number: 1 - 113
2. Length of stay: average length of stay of all patients in hospital (in days)
3. Age: average age of patients (in years)
4. Infection risk: average estimated probability of acquiring infection in hospital (in percent)
5. Routine culturing ratio: ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
6. Routine chest X-ray ratio: ratio of number of cultures performed to number of patients without signs or symptoms of pneumonia, times 100
7. Number of beds: average number of beds in hospital during study period
8. Medical school affiliation: 1 = Yes, 2 = No
9. Region: geographic region, where 1 = NE, 2 = NC, 3 = S, 4 = W
10. Average daily census: average number of patients in hospital per day during study period
11. Number of nurses: average number of full-time equivalent registered and licensed practical nurses during the study period
12. Available facilities and services: percent of 35 potential facilities and services that are provided by the hospital

The `average length of stay in a hospital` (`Y`) is assumed to be related to `infection risk`, `available facilities and services`, and `routine chest X-ray ratio`.

1. Run three separate regression models for each of the three `potential predictors` (i.e., your first model is Y = β_0 +$\beta_1 X_1$ where X_1 = infection risk). Plot the three estimated regression functions over the data in three separate graphs. Does a linear relationship appear to provide a good fit for each of the three predictor variables?
2. Which predictor leads to the smallest `MSE` (a.k.a., `unexplained (error) variation`)? Which predictor variable has the highest R^2? So, which of the three accounts for the largest reduction in variability of the average length of stay?

Predictor	**MSE**	R^2
Infection risk		
Facilities and Services		
Chest X-ray ratio		

3. Obtain model residuals and use them for model diagnostics. Do you identify any issues with model assumptions? Refer to the last lesson for a review of the approach.

Testing linearity and constant variance by plotting fitted value against against residuals.

4. Delete cases 47 ($X = 6.5$, $Y = 19.56$) and 112 ($X = 5.9$, $Y = 17.94$) and refit the model for length of stay and infection risk. From this fitted model, obtain `prediction intervals` for new Y observations at X = 6.5 and X = 5.9. Does what was observed (i.e., $Y = 19.56, 17.94$) fall into the bounds of the respective prediction intervals? Discuss the significance of this.

5. Build the "best" regression model for Y. Begin first with variable selection using the regsubsets function. Justify your final choice of model using criterion-based methods such as BIC and adjusted R^2, all of which can be extracted from the regsubsets model object using the summary function.

6. Once you have identified your final model, check for and comment on any issues with:
   - Multicollinearity between predictors
   - Outliers and influential points
   - Appropriateness of predictors (i.e., is any transformation of predictors necessary?)
   - Normality of residuals
   - Constant variance of residuals.

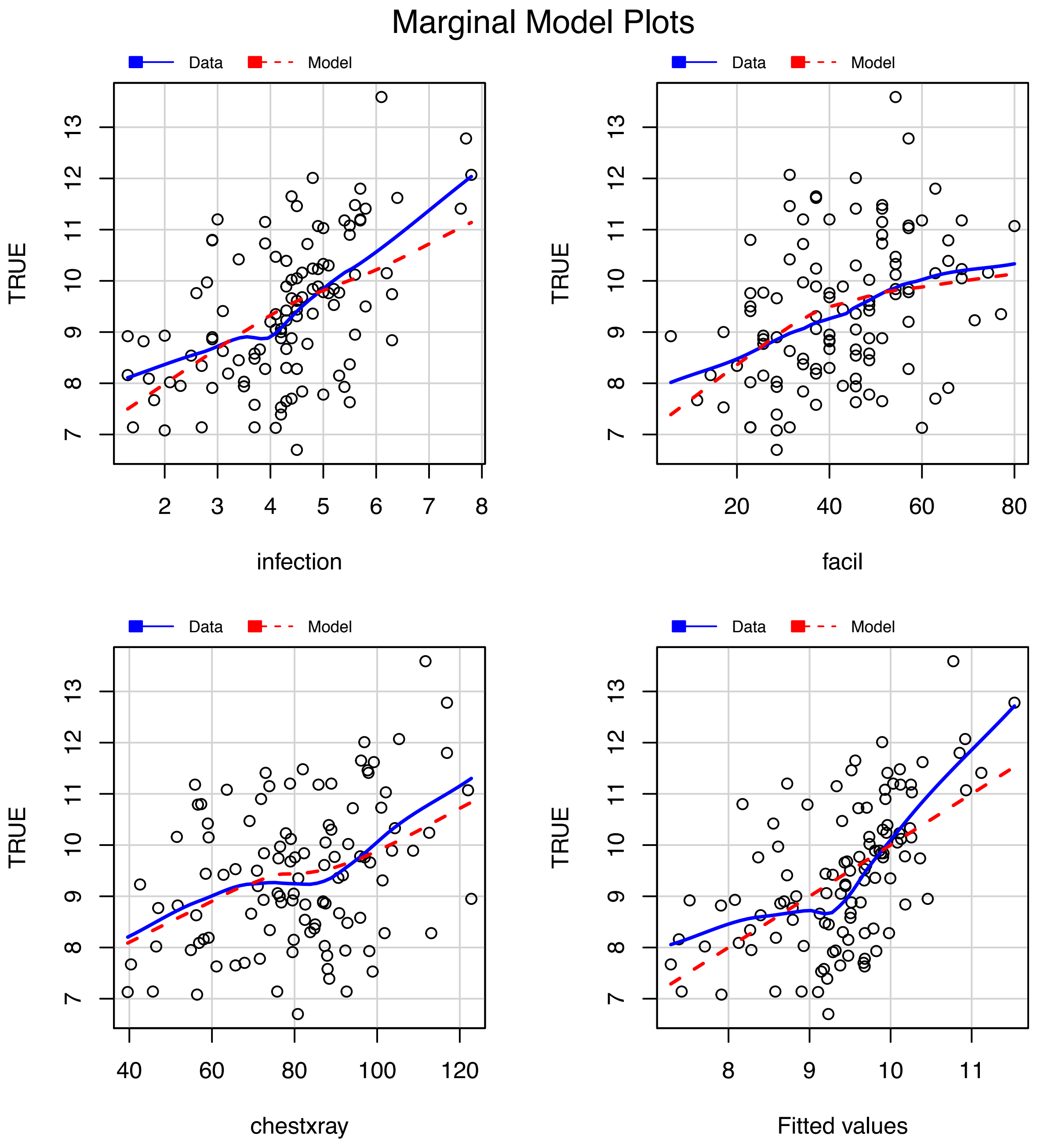
```r
# Normality of residuals

par(mfrow = c(1, 2))
shapiro.qqnorm(residuals(lmBM), cex = 2)
shapiro.qqnorm(residuals(lmBMsub), cex = 2)
```

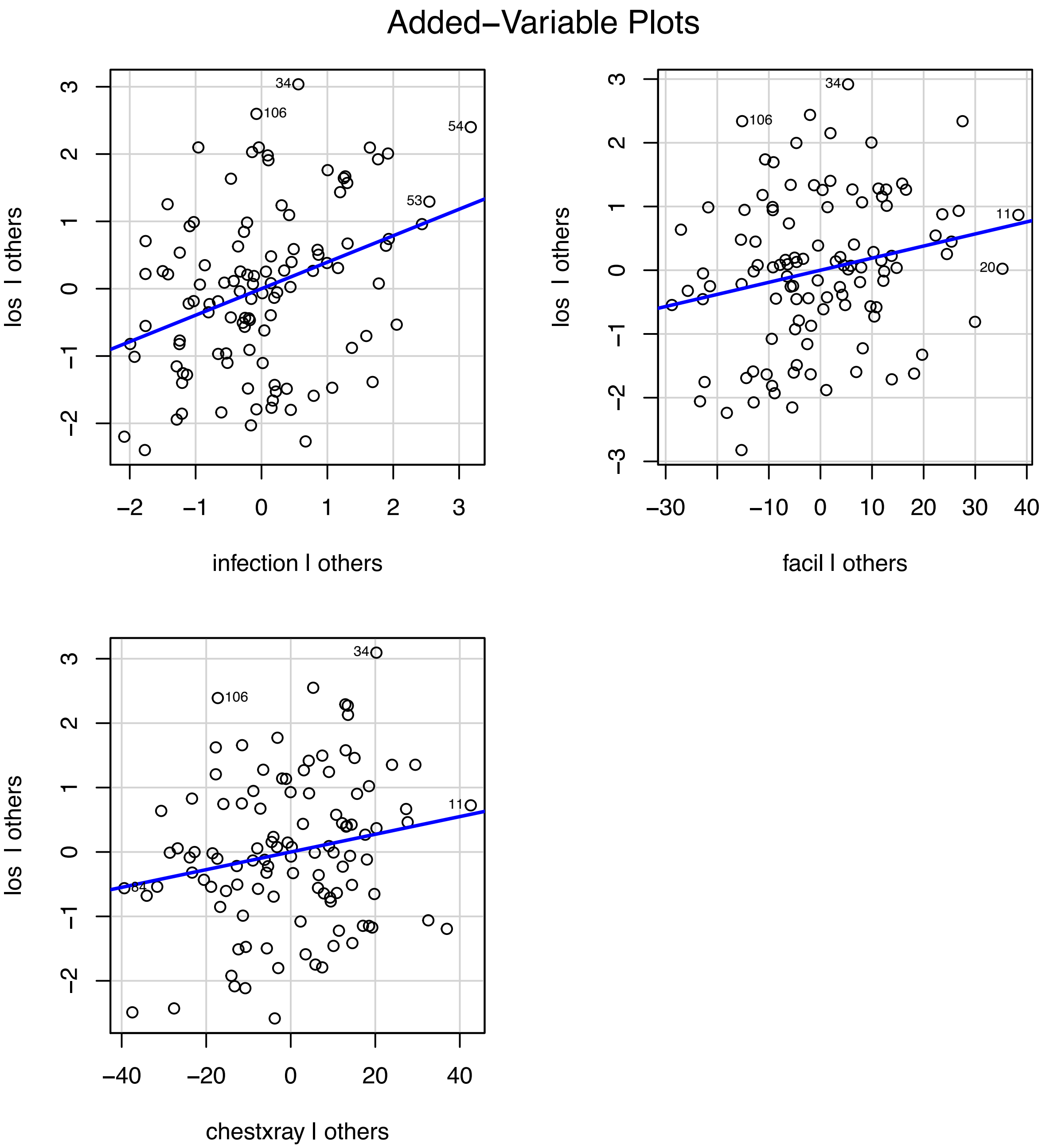
7. Provide an intuitive interpretation of your final model. In other words, explain your findings to me as if I have a minimal working knowledge of statistics.

Extra Plots (Not Required)

```r
library(car)
marginalModelPlots(lmBMsub)
```



```r
avPlots(lmBMsub, id = list(n = 2, cex = 0.6))
```


Added−Variable Plots
los | others
infection | others
−2
−1
0
1
2
3
34
106
54
53
los | others
facil | others
−30
−10
0
10
20
30
40
34
106
11
20
los | others
chestxray | others
−40
−20
0
20
40
34
106
11
84