TMATH 410 Computer Assignment 6

Lectures 12-13, Chapters 5, 6, 9, 11

Objectives:

1. Practice evaluating collinearity
2. Use R's model selection utility
3. Perform diagnostics on the final model selected

Data for computer assignment

For this assignment we will investigate a data set from a study in 1975-1976 that investigated the risk of hospital-acquired infection in the US. The data were published in the American Journal of Epidemiology in 1980, and obtained from Neter et al. The variables in the data set (for this assignment) are: the average length of stay per patient (LenStay), the average age of patients (Age), mean probability of infection (as a percentage; InfectionRisk), average number of beds in hospital (NumBeds), whether the hospital is affiliated with a medical school (MedSchAff; 1=Yes, 2=No), the average number of patients per day (PatientsPerDay), and the number of nurses in the hospital (NumNurse).

C1. (4) Submit the R script you used for this assignment.

Download the dataset named "SENICForAssgn6.csv" from Canvas and read it into your R session. For this assignment I name the R data frame senic.df. Verify that your data frame has 7 columns and 113 rows. Below are the first 3 lines of the data set.

```
##   LenStay  Age InfectionRisk NumBeds MedSchAff PatientsPerDay NumNurse
## 1    7.13 55.7           4.1     279         2            207      241
## 2    8.82 58.2           1.6      80         2             51       52
## 3    8.34 56.9           2.7     107         2             82       54
```

Also make sure to code the MedSchAff column as a factor.

```
senic.df$MedSchAff=factor(senic.df$MedSchAff)
```

C2 (2) Create a summary table of the MedSchAff column and report the number of hospitals affiliated with a medical school and the number not affiliated.

```
summary(senic.df$MedSchAff)
```

Now we will assess collinearity among the quantitative predictor variables.

C3 (3) Create paired scatterplots for each of the quantitative variables. The easiest way to do this is to simply enter: plot(senic.df). This creates all pairwise scatterplots for all variables in a data frame. Make sure your plotting window is big enough! For this problem it is ok for the graphic to not be of publication-quality. Include the graph in your report and comment on any patterns you see.

C4 (2) Estimate the full model that predicts infection risk as a linear combination of the quantitative explanatory variables. Use the vif function in the car package to calculate variance inflation factors for the predictors in your linear model, and present the results in a publication-quality table. (Note, you can install the car package in RStudio by going to the packages tab, then install, then type car. Then type car::vif(lmObject.lm) where lmObject.lm is the name of your linear model object)

```
senic1.lm=lm(InfectionRisk~LenStay+Age+NumBeds+PatientsPerDay+NumNurse,
             data=senic.df)
```

C5 (2) Comment on the values for VIF, in the context of your graphs in C3.

Based on the results of you analysis above, you decide to eliminate number of beds as a predictor variable, and to combine the mean number of patients and the number of nurses into a single variable that measures the patients per nurse. Create a new column in the senic.df data frame that is is the PatientsPerDay/NumNurse. Call the new column PatientsPerNurse.

C6 (2) Estimate a new linear model that predicts InfectionRisk by a linear combination of LenStay, Age, and PatientsPerNurse, then calculate VIFs for this new model and report the results in a publication-quality table. Do these values indicate any issues?

Let's see what the built-in model selection function in R returns as a "final" model. We will use the step function with AIC as a selection criterion. First we estimate a full linear model, including MedSchAff and interactions between MedSchAff and other quantitative predictors. We will use LenStay, Age, and PatientsPerNurse as our quantitative variables.

```
senic2.lm=lm(InfectionRisk~LenStay*MedSchAff+Age*MedSchAff+
               PatientsPerNurse*MedSchAff,data=senic.df)
```

C7. (5) Use the step function model selection for our full linear model using AIC as the criterion. In your lab document summarize which variable was eliminated each step, and explain why.

```
# AIC is the default
senicStep.aic=step(senic2.lm)
senicStep.aic
```

C8. (3) Produce a summary of the final model selected by the step function. Provide an interpretation of each partial regression coefficient.

C9. (3) Give residual plots of your final model. Comment on any patterns you see, and whether you believe any major regression assumptions are violated. Decide if there are any influential points in your model.

C10 (4) Intepret the results of your model selection procedure in the context of what we have learned about infection risk through this data analysis. Include in your discussion consideration of which variables were included in the final model and which variables were not included.