STEVE G MWANGI
R LAB 6
03/11/2020

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.

Submitted

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(scenic.df$MedSchAff)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    2.00    1.85    2.00    2.00
```

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.

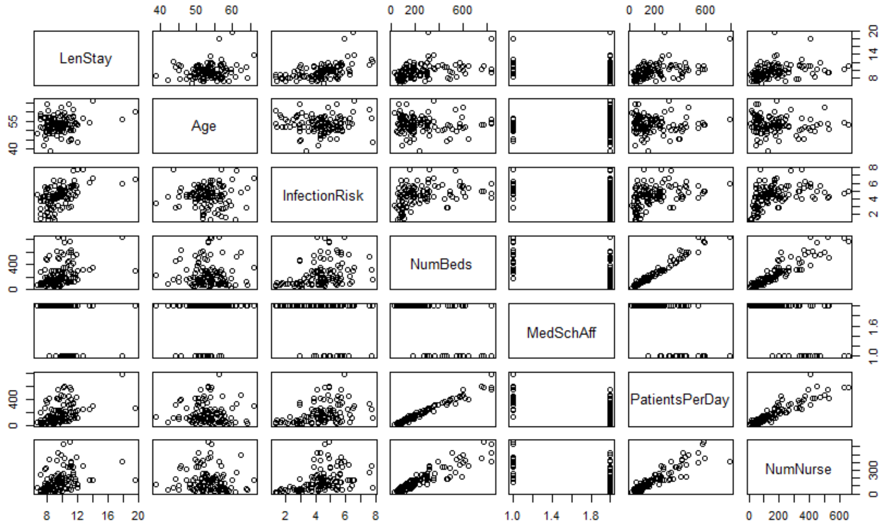

Figure 1 Paired Scatter plot for all quantitative variables

The scatter plots of *MedSchAff* vs. et al are all pointing to the predictor being a factor.

PatientsPerDay correlates positively with NumNurse(which also correlates positively with NumBeds), as well as with NumBeds.

Some of the graphs point to a violation of constant variance, such as infection risk vs each (NumBeds, PatientsPerDay,NumNurse).

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).

Length of Stay	**Age**	**Number of Beds**	**Patients per day**	**Number of Nurse**
1.571815	1.071767	31.777783	33.197632	6.450622

Table 1 VIF for the data in the scenic linear model

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

As can be seen in the linear plots of the predictors with VIF > 5, there exists high multicollinearity among these variables which were pointed out in C3.

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 publicationquality table. Do these values indicate any issues?

Length of Stay	**Age**	**Patients Per Nurse**
1.042785	1.037170	1.006300

Table 2 VIFs after dropping a column and using patients per nurse instead.

These new VIFs do not indicate any underlying issues point to multiple collinearity. They are ideal.

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.

```
> scenicStep.aic=step(scenic2.lm)
Start:  AIC=18.51
InfectionRisk ~ LenStay * MedSchAff + Age * MedSchAff + PatientsPerNurse *
    MedSchAff

                          Df Sum of Sq    RSS    AIC
- LenStay:MedSchAff        1   0.29172 115.83 16.795
- MedSchAff:PatientsPerNurse  1   0.44079 115.98 16.941
- MedSchAff:Age            1   1.66883 117.21 18.131
<none>                                115.54 18.511
```

The first parameter to be removed is the *LenStay:MedSchAff* which R removes due to ho w the interaction between the parameters, once removed, results in a lower AIC = 16.8

```
Step:  AIC=16.8
InfectionRisk ~ LenStay + MedSchAff + Age + PatientsPerNurse +
    MedSchAff:Age + MedSchAff:PatientsPerNurse

                             Df Sum of Sq    RSS    AIC
- MedSchAff:Age               1     1.443 117.27 16.195
- MedSchAff:PatientsPerNurse  1     1.750 117.58 16.490
<none>                                    115.83 16.795
- LenStay                     1    52.352 168.18 56.936
```

The next step parameter removed is *MedSchAff:Age(interaction)* which R removes due to how the interaction between the parameters, once removed, results in a lower AIC = 16.19

```
Step:  AIC=16.19
InfectionRisk ~ LenStay + MedSchAff + Age + PatientsPerNurse + 
    MedSchAff:PatientsPerNurse

                             Df Sum of Sq    RSS    AIC
- Age                         1     1.968 119.24 16.075
<none>                                    117.27 16.195
- MedSchAff:PatientsPerNurse  1     2.554 119.83 16.630
- LenStay                     1    56.846 174.12 58.856
```

The next step parameter removed is *Age(main)* which R removes due to how the parameters, once removed, results in a lower AIC = 16.08

```
Step:  AIC=16.08
InfectionRisk ~ LenStay + MedSchAff + PatientsPerNurse + MedSchAff:PatientsPe
rNurse

                             Df Sum of Sq    RSS    AIC
- MedSchAff:PatientsPerNurse  1     2.095 121.34 16.043
<none>                                    119.24 16.075
- LenStay                     1    55.242 174.48 57.092
```

The next step parameter removed is *MedSchAff:PatientsPerNurse(interaction)* which R removes due to how the interaction between the parameters, once removed, results in a lower AIC = 16.04

```
Step:  AIC=16.04
InfectionRisk ~ LenStay + MedSchAff + PatientsPerNurse

                   Df Sum of Sq    RSS    AIC
- MedSchAff         1     0.726 122.06 14.717
<none>                          121.34 16.043
- PatientsPerNurse  1    21.509 142.84 32.484
- LenStay           1    53.179 174.51 55.112
```

Last step parameter removed is *MedSchAff (main)* which R removes due to how the main effect of the parameter, once removed, results in a lower AIC = 14.717

```
Step:  AIC=14.72
InfectionRisk ~ LenStay + PatientsPerNurse

                   Df Sum of Sq    RSS    AIC
<none>                          122.06 14.717
- PatientsPerNurse  1    22.012 144.07 31.453
- LenStay           1    62.624 184.69 59.513
```

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

$$\boldsymbol{\hat{Y} = 1.5741 + 0.3924\hat{X}_1\beta_1 - 0.8450\hat{X}_2\beta_2}$$

```
> scenicStep.aic

Call:
lm(formula = InfectionRisk ~ LenStay + PatientsPerNurse, data = scenic.df)

Coefficients:
     (Intercept)           LenStay  PatientsPerNurse


> summary(scenicStep.aic)

Call:
lm(formula = InfectionRisk ~ LenStay + PatientsPerNurse, data = scenic.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6062 -0.5384 -0.0393  0.6126  2.7684 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.57407    0.54494   2.889  0.00466
LenStay           0.39240    0.05223   7.512 1.64e-11
PatientsPerNurse -0.84504    0.18973  -4.454 2.03e-05
                    
(Intercept)      ** 
LenStay          ***
PatientsPerNurse ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.053 on 110 degrees of freedom
Multiple R-squared:  0.3939,	Adjusted R-squared:  0.3828 
F-statistic: 35.74 on 2 and 110 DF,  p-value: 1.099e-12
```

Interpretation: We see that for every unit increase in Length of Patient Stay, after effects of Patients Per Nurse are accounted for, the infection risk goes up by a factor of 0.39240 plus a standard 1.57407units. Additionally, after the effects of Length of stay are accounted for, the infection risk decreases by a factor of 0.84504 plus the standard 1.57407, for every unit increase in Patients Per Nurse.

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.

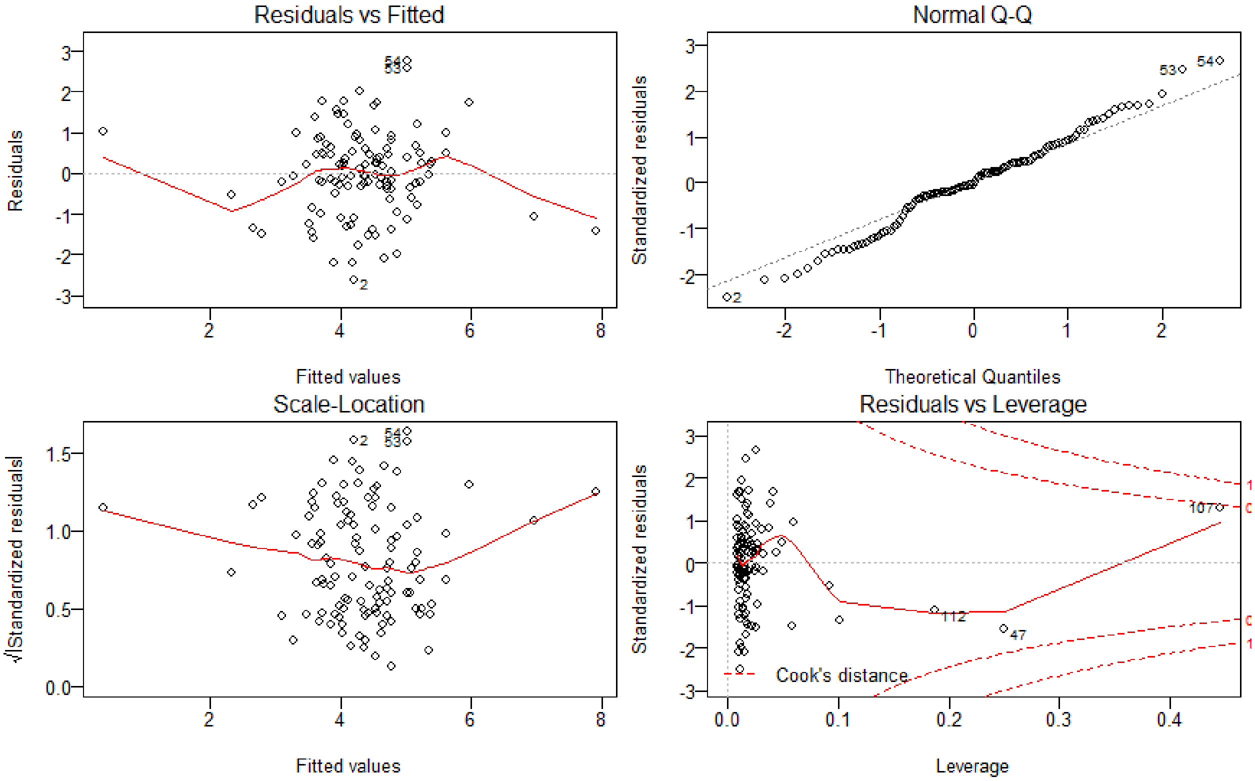

Table 3 Residual diagnosis plots generated using R

The cook's plot as can be seen in the fourth plot, does not highlight points that might be problematic and of great influence on our model. Instead, the other three plots highlight outlier points 53 and 54, which also seem to have high leverage. Outside from these two points, the plots are pointing to a residual distribution that would be decent since constant variance assumption is retained and the residual points are independently and identically distributed.

C10 (4) Interpret 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.

Even though there is a high correlation among some variables, the VIF pointed out multicollinearity in the data set, which was at first hinted to by the initial plots. This confirmation enabled the points that had significant VIFs to be considered and removed or used to estimate a new model with an additional column that was the Patients Per Nurse. The model selection using the AIC criterion facilitated the arrival of a model that was simple and fit the data appropriately, which was limited to just two predictor variables which had the lowest AIC score of 14.72 (without any additional interacting predictors). But in sum, we found that the infection risk could be predicted or explained by taking into account the number of patients per nurse and the length/number of days patient stayed in the hospital; and this infection risk was inversely proportional to the former, but directly proportional to the latter.