Stat 6021: Homework Set 7 Solutions

1. (a) Let x_1, x_2 denote the variables we wish to drop: Agriculture and Examination.

$H_0 : \beta_1 = \beta_2 = 0$, H_a : at least one of the coefficients in H_0 is non zero

Note: The reduced model, R, has 3 predictors: Education, Catholic, and Infant Mortality, while the full model, F, has all 5 predictors. There are 2 approaches to find the result of this partial F test.

Approach 1: Use the `anova()` function to compare the reduced and full models.

```
> anova(reduced,result)
Analysis of Variance Table

Model 1: Fertility ~ Education + Catholic + Infant.Mortality
Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     43 2422.2
2     41 2105.0  2     317.2 3.0891 0.05628 .
```

From the partial F test, the p-value is greater than 0.05, so we fail to reject the null hypothesis at 0.05 significance level. This means we should select the reduced model with just the three predictors: Education, Catholic, and Infant Mortality.

Approach 2: use the `anova()` function on the full model to produce the sequential sum of squares.

```
> anova(result2)
Analysis of Variance Table

Response: Fertility
                 Df Sum Sq Mean Sq F value    Pr(>F)
Education         1 3162.7  3162.7 61.6004 1.073e-09 ***
Catholic          1  961.1   961.1 18.7187 9.478e-05 ***
Infant.Mortality  1  631.9   631.9 12.3080  0.001109 **
Agriculture       1  264.2   264.2  5.1454  0.028641 *
Examination       1   53.0    53.0  1.0328  0.315462
Residuals        41 2105.0    51.3
```

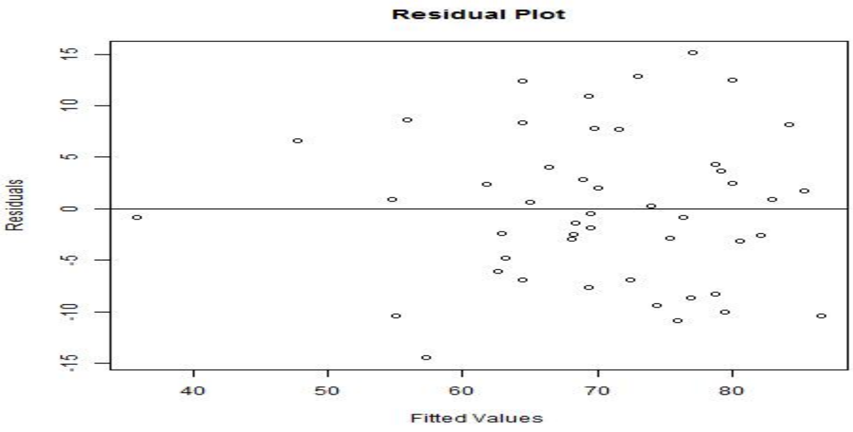
The partial F statistic is

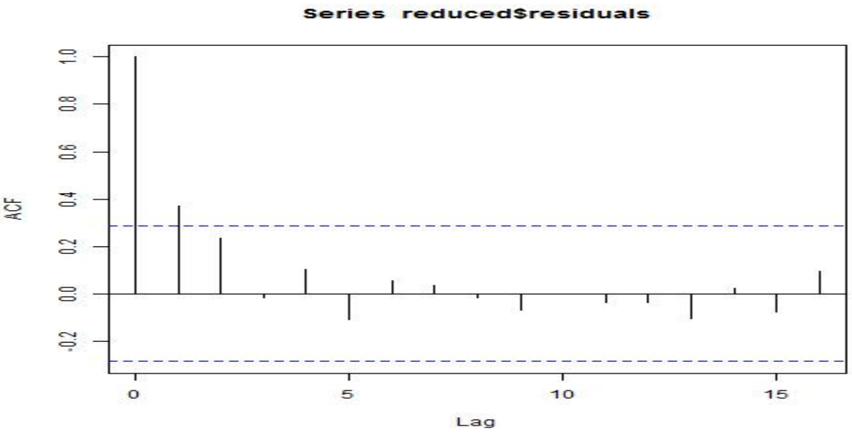
$$\begin{aligned}
F - stat &= \frac{[SS_R(F) - SS_R(R)]/r}{SS_{res}(F)/(n-p)} \\
&= \frac{[264.2 + 53]/2}{2105/41} \\
&= 3.079
\end{aligned}$$

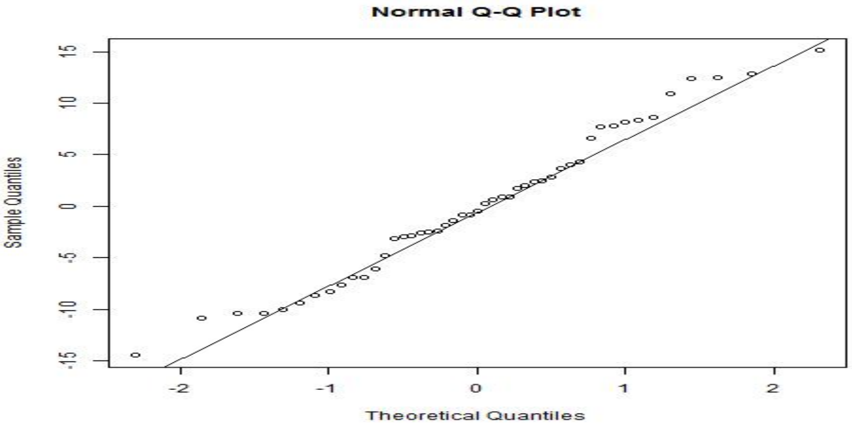
where r denotes the number of predictors to drop, which is 2. $(n - p)$ denotes the DF of the error for the full model, which is given as 41 in the output. We have a p-value of 0.057, using 1-pf(3.079,2,41), or a critical value of 3.226 using qf(0.95,2,41). So we fail to reject the null hypothesis and use the reduced model.

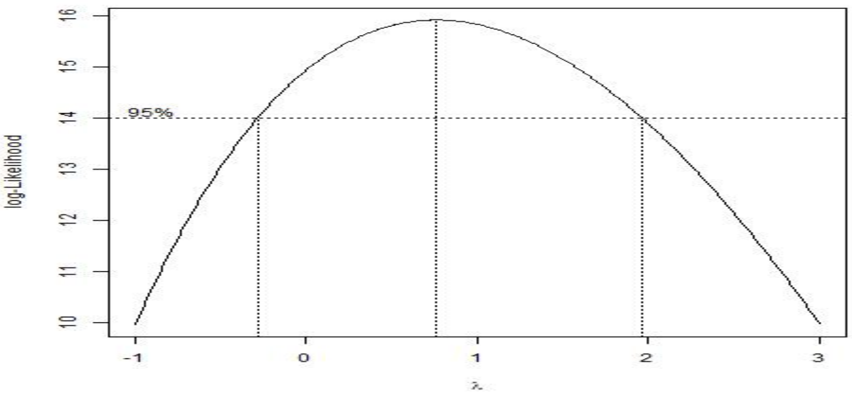
The two approaches should theoretically give the same results, the differences are due to rounding in the output.

(b) The regression assumptions mostly appear to be met. From the residual plot, we note the residuals are evenly scattered around 0 at random, with a constant vertical variance. The ACF plot shows the residuals at lag 1 have a slight correlation, so this is a concern. We need to consider whether the order of the rows matter in the original dataframe, or if they were ordered in some matter. If not, one could randomize the order of the rows and check for the autocorrelation again (probably will not be correlated). From the QQ plot, the normality assumption is reasonably met as the residuals fall close to their theoretical values under normality. The Box Cox plot indicates we do not need to transform the response variable, so the constant variance assumption is met.









2. (a) *Age*, *Census*, Beds have insignificant t tests.

   (b) $H_0 : \beta_3 = \beta_4 = \beta_5 = 0.$
   H_a : not all $\beta_3, \beta_4, \beta_5$ are zero.

   Note: The reduced model has x_1 and x_2 as predictors, the full model has x_1, x_2, x_3, x_4, x_5 as predictors.

   Use given table below to calculate partial F statistic

```
> anova(result)
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305 58.1676 1.044e-11 ***
Cultures    1  33.397  33.397 33.8995 6.154e-08 ***
Age         1   0.136   0.136  0.1376   0.71144
Census      1   5.101   5.101  5.1781   0.02487 *
Beds        1   0.028   0.028  0.0279   0.86759
Residuals 107 105.413   0.985
```

$$\begin{aligned} F - stat &= \frac{[SS_R(F) - SS_R(R)]/r}{SS_{res}(F)/(n-p)} \\ &= \frac{[0.136 + 5.101 + 0.028]/3}{105.413/107} \\ &= 1.781 \end{aligned}$$

The corresponding p-value is $1 - pf(1.781, 3, 107) = 0.1551$, which is greater than 0.05. The critical value is 2.689, found using qf(0.95,3,107), is greater than our partial F statistic. So we fail to reject the null. Data suggests that the coefficients for *Age*, *Census*, and *Beds* can be dropped from the model, and we go with the reduced model.

(c) In this question, model 1 is the Full model,F, while model 2 is the Reduced model, R. So,
F: $E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
R: $E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$H_0 : \beta_3 = \beta_4 = 0, H_a :$ at least one of the coefficients in H_0 is nonzero.

$$\begin{aligned} F - stat &= \frac{(SS_R(x_1, x_2, x_3, x_4) - SS_R(x_1, x_2))/r}{SS_{res}(x_1, x_2, x_3, x_4)/(n-p)} \\ &= \frac{(0.136 + 5.101)/2}{(105.413 + 0.028)/(113 - 5)} \\ &= 2.682 \end{aligned}$$

Because the SS_T is constant as long as we have the same response variable, the SS_R for x_5 goes into the SS_{res} when the full model uses the first 4 predictors. The p-value is 0.0729834, found by using `1-pf(2.682,2,108)`. The critical value is `qf(0.95,2,108)` which is 3.0804. Since our test statistic is less than the critical value, we fail to reject the null. This means we go with the simpler of the two models, model 2.

3. Some indications that we have multicollinearity:

   - Insignificant t tests for predictors that should be useful in predicting the response variable.
   - Significant ANOVA F test indicating our model is useful in predicting the response.

4.

$$\begin{aligned}
\boldsymbol{X}^{*'}\boldsymbol{Y}^{*} &= \begin{bmatrix} x_{11}^{*} & \cdots & x_{n1}^{*} \\ \vdots & & \vdots \\ x_{1k}^{*} & \cdots & x_{nk}^{*} \end{bmatrix} \cdot \begin{bmatrix} y_1^{*} \\ \vdots \\ y_n^{*} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^{*} y_i^{*} \\ \vdots \\ \sum_{i=1}^{n} x_{ik}^{*} y_i^{*} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} \frac{1}{\sqrt{n-1}} \frac{(x_{i1}-\bar{x_1})}{s_1} \frac{1}{\sqrt{n-1}} \frac{(y_i-\bar{y})}{s_y} \\ \vdots \\ \sum_{i=1}^{n} \frac{1}{\sqrt{n-1}} \frac{(x_{ik}-\bar{x_k})}{s_k} \frac{1}{\sqrt{n-1}} \frac{(y_i-\bar{y})}{s_y} \end{bmatrix} \\
&= \begin{bmatrix} r_{Y1} \\ \vdots \\ r_{Yp} \end{bmatrix}
\end{aligned}$$