hw7

Hyun Suk (Max) Ryoo (hr2ee)

10/24/2021

Set Up

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ------------------------------------------------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts --------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(datasets)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
data <- swiss
head(data)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
```

```
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

Question 1

1. For this first question, you will continue to use the dataset swiss which you also used in the last homework. Load the data. For more information about the data set, type ?swiss. The goal of the data set was to assess how fertility rates in the Swiss (French-speaking) provinces relate to a number of demographic variables.
   A) In the previous homework, you fit a model with the fertility measure as the response variable and used all the other variables as predictors. Now, consider a simpler model, using only the last three variables as predictors: Education, Catholic, and Infant.Mortality. Carry out an appropriate hypothesis test to assess which of these two models should be used. State the null and alternative hypotheses, find the relevant test statistic, p-value, and state a conclusion in context. (For practice, try to calculate the test statistic by hand.)

```
full_model <- lm(Fertility~., data=data)
summary(full_model)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

```
simple_model <- lm(Fertility ~ Education + Catholic + Infant.Mortality, data=data)
summary(simple_model)
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality,
##     data = data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4781  -5.4403  -0.5143   4.1568  15.1187
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      48.67707    7.91908   6.147 2.24e-07 ***
## Education        -0.75925    0.11680  -6.501 6.83e-08 ***
## Catholic          0.09607    0.02722   3.530  0.00101 **
## Infant.Mortality  1.29615    0.38699   3.349  0.00169 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.505 on 43 degrees of freedom
## Multiple R-squared:  0.6625, Adjusted R-squared:  0.639
## F-statistic: 28.14 on 3 and 43 DF,  p-value: 3.15e-10
```

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ H_A : not all$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$are zero The results of the partial F tests are shown below.

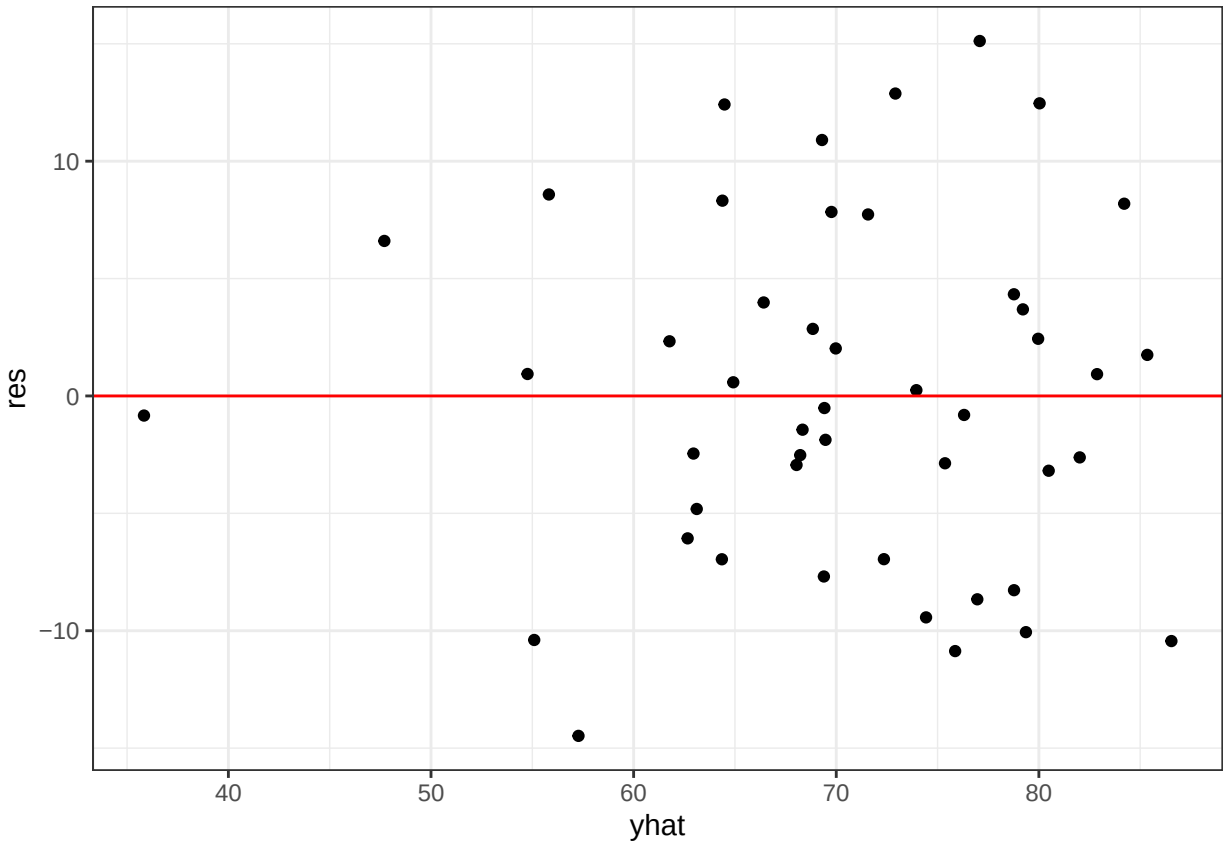
```
anova(simple_model, full_model)
```

```
## 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 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the partial F test, we can observe a F-statistic of 3.0891 and a P-value of 0.05628. Since this p-value is above 0.05 we fail to reject the null hypothesis given our cut off value is 0.05. Therefore we should utilze the reduced model with only the three predictors.

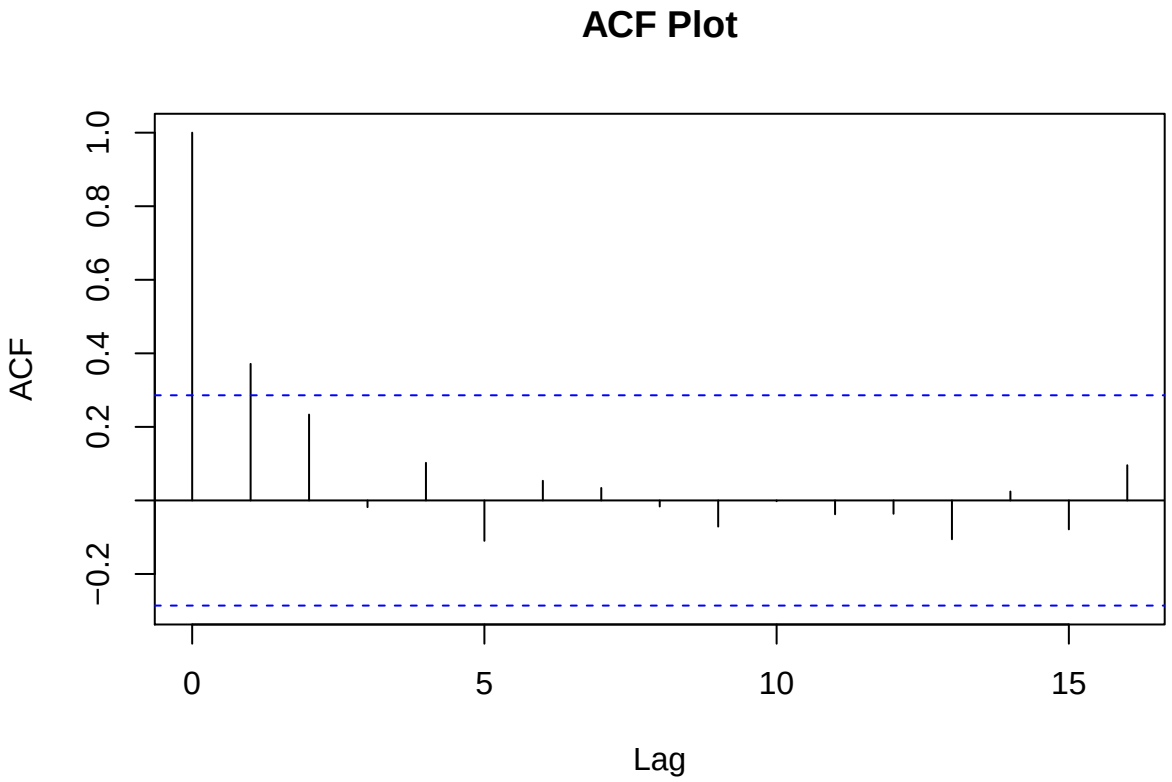
B) For the model you decide to use from part 1a, assess if the regression assumptions are met.

```
yhat = simple_model$fitted.values
res = simple_model$residuals
data %>%
  ggplot(aes(yhat, res)) + geom_point() + theme_bw() + geom_hline(yintercept=0, color="red")
```



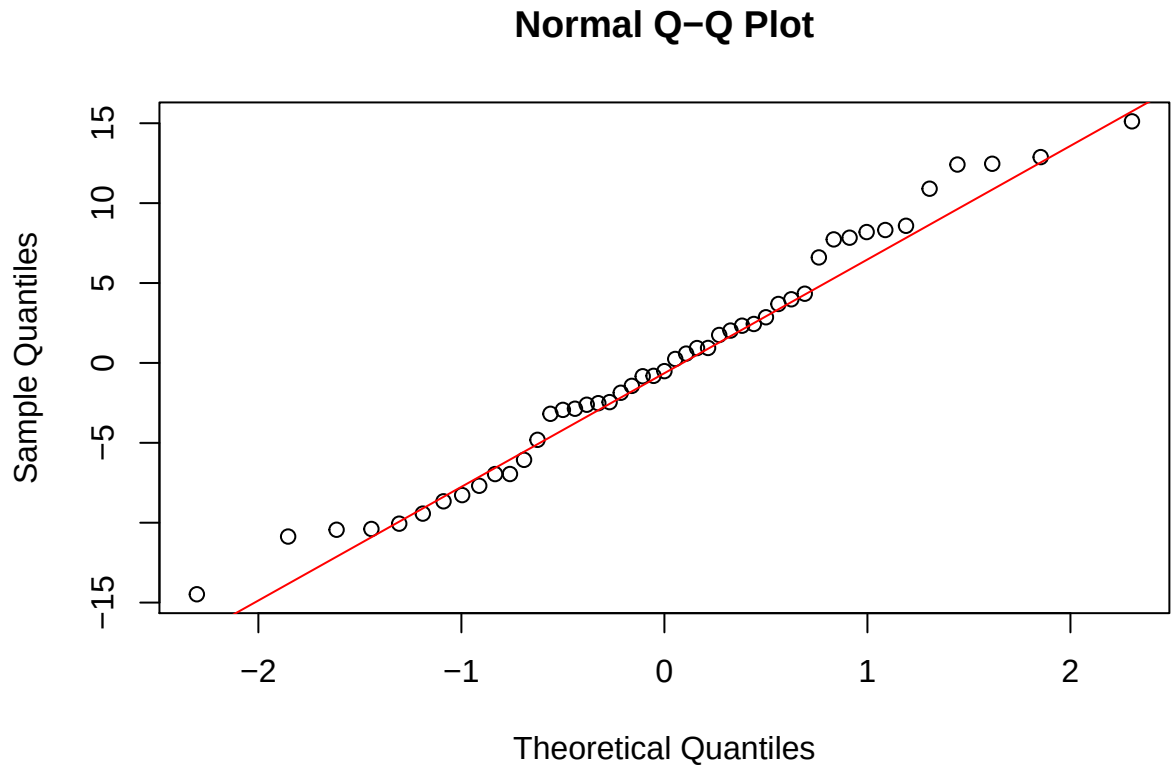
From the above residual plot we can seee that the residuals are even scattered around 0, which is one asumption we need to check.

```
acf(simple_model$residuals, main="ACF Plot")
```



From the ACF plot we can observe the residual at lag one has some correlation, which may be a concern in some context.

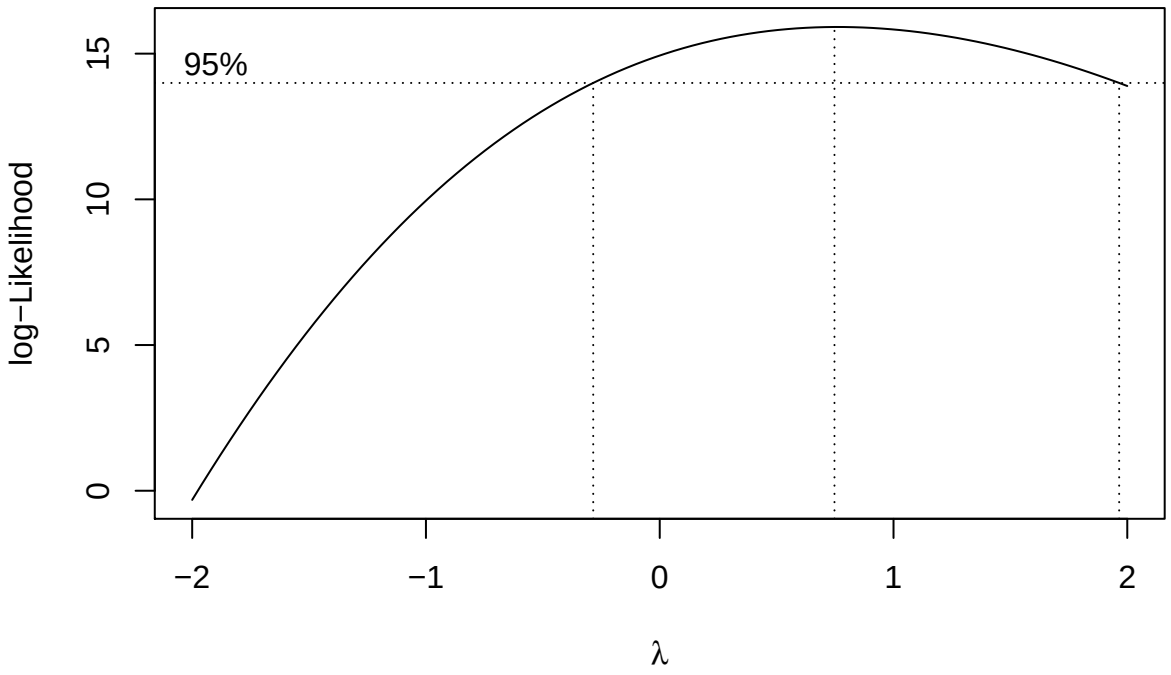
```
{
qqnorm(simple_model$residuals)
qqline(simple_model$residuals, col="red")
}
```



The qqplot shows that the residuals fall close to the red line, which means that the normality assumption can be assumed.

The final step is to check whether we need to transform the response variable. We can check this by observing if 1 falls between the boundaries of the boxcox plot, which by the figure below is true.

```
boxcox(simple_model)
```



Question 2 No R required (InfctRsk)

2.A. Based on the t statistics, which predictors appear to be insignificant? From the table provided in the homework output, we can say that the Age, Cenus, and Beds variables appear to be insignificant

2.B. Based on your answer in part 2a, carry out the appropriate hypothesis test to see if those predictors can be dropped from the multiple regression model. Show all steps, including your null and alternative hypotheses, the corresponding test statistic, p-value, critical value, and your conclusion in context.

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ H_A : not all$\beta_3, \beta_4, \beta_5$are zero

Since this question does not require R, we will find the F-Statisitic for the partial F test by hand.

$$F - Statistic = \frac{\frac{0.136+5.101+0.028}{3}}{\frac{105.413}{107}}$$
$$F - Statistic = \frac{1.755}{0.9851682}$$
$$F - Statistic = 1.781422$$

The p - value can be found using a simple computation of $1 - pf(1.781422, 3, 107)$, which resulted in 0.1550925.

```
1 - pf(1.781422, 3, 107)
```

```
## [1] 0.1550925
```

This p-value is greater than 0.05. We can also check by finding the cricial value, which can be computed using $qf(0.95, 3, 107)$, which is 2.68949.

```
qf(0.95,3,107)
```

```
## [1] 2.68949
```

Since this value is greater than our f-statistic computed we fail to reject the null hytpohesis. Therefore, we can conclude that for the data provided we can utilize the simpler model.

2.C. Suppose we want to decide between two potential models:

- Model 1: using x1, x2, x3, x4 as the predictors for InfctRsk
- Model 2: using x1, x2 as the predictors for InfctRsk Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

$H_0 = \beta_3 = \beta_4 = 0$ $H_A =$ at least one of the beta values is non zero

Since this question does not require R, we will find the F-Statisitic for the partial F test by hand.

$$F - Statistic = \frac{\frac{0.136+5.101}{2}}{\frac{105.413+0.028}{113-5}}$$
$$F - Statistic = \frac{2.6185}{0.9763056}$$
$$F - Statistic = 2.68205$$

The p - value can be found using a simple computation of $ 1 - pf(2.68205, 2, 108)$, which resulted in 0.07297992

```
1 - pf(2.68205, 2, 108)
```

```
## [1] 0.07297992
```

This p-value is greater than 0.05. We can also check by finding the cricial value, which can be computed using $qf(0.95, 3, 107)$, which is 3.080387

```r
qf(0.95,2,108)
```

```
## [1] 3.080387
```

Since our f-statistic value is less than the critical value and the p-value is greater than 0.05 we fail to reject the null hypothesis. THerfore, we can utilize the simpler model (model 2).

Question 3 No R required (Left Arm, Left Foot, Rt Foot)

3.A. Explain how this output indicates the presence of multicollinearity in this regression model.

The output indicates the presence of multicollinearity in this regression model because although the F test states our model is significant in predicting the response varible, none of the individual predictors are significant. Testing each variable one by one is needed and the possibility of multicolinearity needs to be considered.