MA684
Homework 7

1. Let's explore some model diagnostic simulations. Consider the following models:

$$Model\ 1: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$Model\ 2: y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$
$$Model\ 3: y_i = \beta_0 + \beta_1 x_i + \sqrt{x_i}\varepsilon_i$$
$$Model\ 4: y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sqrt{x_i}\varepsilon_i$$
$$Model\ 5: y_i = \beta_0 + \beta_1 x_i + \eta_i$$

where $\beta_0 = 0.1, \beta_1 = 0.5, \beta_2 = 2.0, x_i \sim Uniform(0,2), \varepsilon_i \sim N(0,0.2), \eta_i \sim t_4$ distributions. Note that t_4 is the t-distribution with 4 degrees of freedom.

a. Generate n=100 samples from each model. Obtain the scatter plot of the data.
b. For each data, fit the linear regression model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
c. Generate the residual plot and normal QQ plot. In each case, white a short description of the two plots. What features of the plots suggest that your fitted model may not be adequate?

```
#generate samples for ε
eps=rnorm(100,0,0.2)
#generate x values
x=runif(100,0,2)
#generate y values from the model
y=0.1+0.5*x+eps
#create a data frame for the data so that we are not confused with so many datasets
data_mod1=data.frame(x,y)
#Let's have a look at the scatter plot of the data
plot(x,y, main = "Scatterplot for model 1")
#Fit linear model
fit1=lm(y~x,data=data_mod1)
#create a place holder for the two plots
par(mfrow=c(1,2))
#plot residual
plot(fit1, which=1, col="blue")
#QQ plot
plot(fit1,which=2, col="red3")
```

2. We will be using the "surgical" data in the "olsrr" package for this homework. The data is related to patients' survival times who had undergone a type of liver operation. The response variable is "survival," and for now, we consider only the continuous predictors in the dataframe.

a. Create a scatter plot matrix of the data. Describe how the predictors are related to the response variable.

b. Fit a multiple regression model.   Is the model statistically significant? Explain your answer
c. Which of the predictors are statistically significant given the other variables are already in the model? Explain your answer.
d. Carry out the residual analysis. Is the model adequate? Explain your answer.
e. Carry out a hypothesis test for normality and equal variance.   What are your conclusions?
f. Now, we want to fit a new model with the response variable "survival" replaced by "log(survival)" and all the predictors the same as before.   Repeat parts (a)-(e), and explain how the new model is different from the old.