Question 5

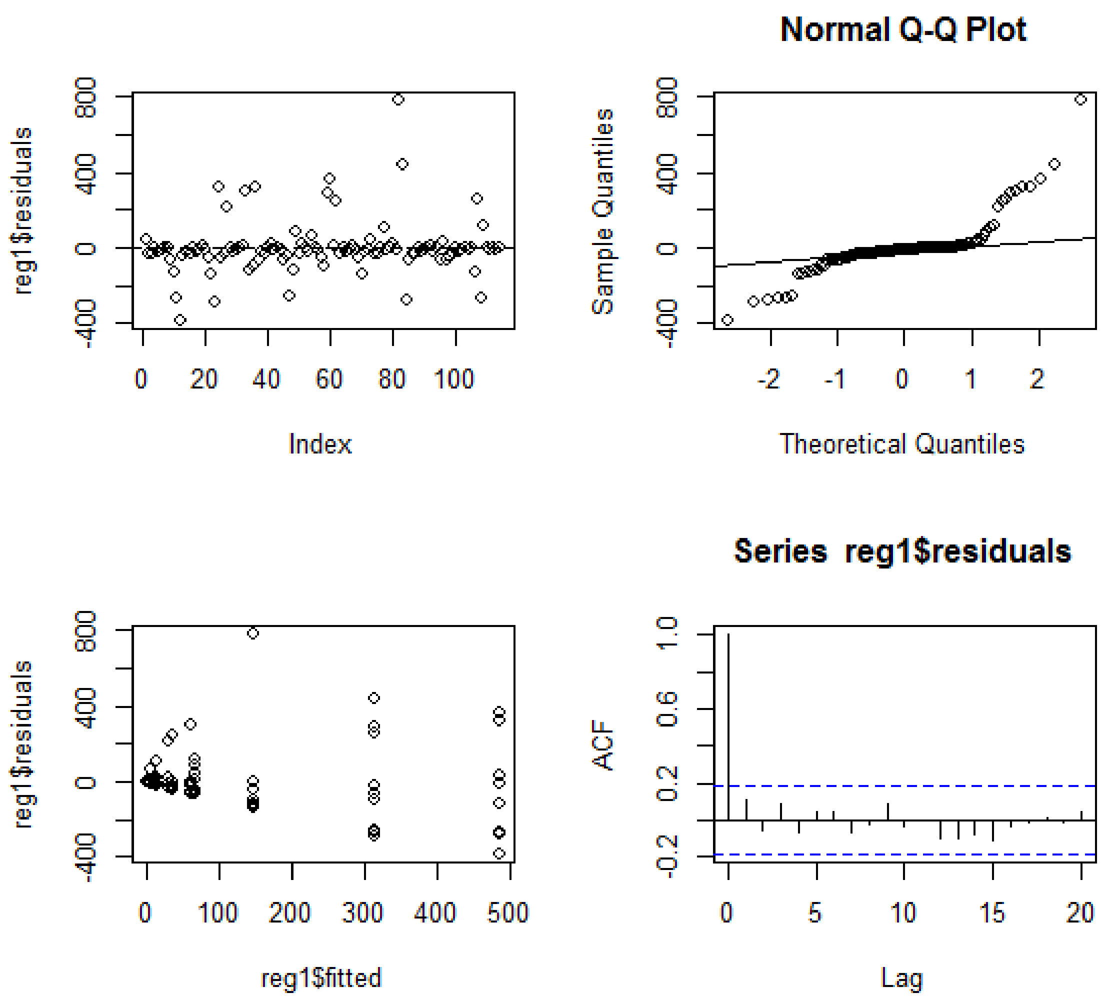
a) Fit a seasonal regression model and provide full residual diagnosis

```
lightning <- c(109, 5, 0, 0, 0, 0, 4, 12, 0, 18, 51, 103, 23, 19, 0, 0,
0, 0, 18, 3, 12, 12, 24, 812, 15, 1, 248, 0, 0, 0, 0, 21, 365, 29, 217,
806, 0, 12, 0, 0, 38, 0, 0, 0, 0, 108, 56, 369, 155, 11, 51, 0, 0, 75,
0, 4, 14, 47, 609, 856, 80, 285, 0, 4, 0, 0, 12, 1, 10, 9, 298, 481,
108, 1, 0, 0, 121, 0, 3, 38, 50, 927, 757, 209, 6, 5, 0, 1, 0, 0, 0,
25, 39, 153, 249, 519, 1, 0, 0, 1, 0, 0, 0, 1, 66, 22, 573, 224, 191,
38, 16, 0, 1, 3, 2, 33, 119, 487, 103, 48, 145) # read in the data as a
vector
light.all <- ts(lightning, start = c(2001,9), frequency = 12)
light.train <- ts(lightning[1:114], start = c(2001,9), frequency = 12)
light.test <- ts(lightning[115:121], start = c(2011,3), frequency = 12)
month = factor(cycle(light.train)) # as the month is used every year
without specific year
reg1 = lm(light.train~month)
summary(reg1)

##
## Call:
## lm(formula = light.train ~ month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -383.56   -35.70    -8.23     0.40   779.78
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.000     46.815   0.342   0.7332
## month2          -8.200     66.206  -0.124   0.9017
## month3         -11.889     68.020  -0.175   0.8616
## month4          -4.333     68.020  -0.064   0.9493
## month5          45.778     68.020   0.673   0.5025
## month6         131.222     68.020   1.929   0.0565 .
## month7         298.889     68.020   4.394 2.73e-05 ***
## month8         470.556     68.020   6.918 4.12e-10 ***
## month9          52.800     66.206   0.798   0.4270
## month10         21.700     66.206   0.328   0.7438
## month11         15.500     66.206   0.234   0.8154
## month12        -15.400     66.206  -0.233   0.8165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 148 on 102 degrees of freedom
## Multiple R-squared:  0.5065, Adjusted R-squared:  0.4532
## F-statistic: 9.515 on 11 and 102 DF,  p-value: 1.365e-11
```

The fitted model is: $light.train = 16.000 - 8.2000month2 - 11.889 - 4.333month4 + 45.778month5 + 131.222month6 + 298.889month7 + 470.556month8 + 52.800month9 + 21.700month10 + 15.500month11 - 15.400month12

```r
par(mfcol=c(2,2))
plot(reg1$residuals)
abline(h=0,lty=2)
plot(reg1$fitted, reg1$residuals)
qqnorm(reg1$residuals)
qqline(reg1$residuals)
acf(reg1$residuals)
```



Comments: The residual plot has many outliers, and fitted vs. residual plot also looks like a linear trend. The Q-Q plot has light tails. The ACF plot looks fine with no significant value after h=0, but we might want to consider refit the model to reduce the large outliers in the plot.

b) Predict the number of lightning strikes in the last 7 months

```
PI<-
predict.lm(reg1,newdata=data.frame(month=factor(cycle(light.test))),
interval="prediction")
PI

##          fit         lwr      upr
## 1   4.111111 -305.412230 313.6345
## 2  11.666667 -297.856675 321.1900
## 3  61.777778 -247.745564 371.3011
## 4 147.222222 -162.301119 456.7456
## 5 314.888889    5.365548 624.4122
## 6 486.555556  177.032214 796.0789
## 7  68.800000 -239.171836 376.7718

(light.test>PI[,"lwr"] & light.test<PI[,"upr"]) # see if light.test is
in PI

##        Mar   Apr   May   Jun   Jul   Aug   Sep
## 2011  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

Comments: as the result reveals, we have two test data not in the prediction interval. With the 95% perdiction level we might want to consider refitting the model.

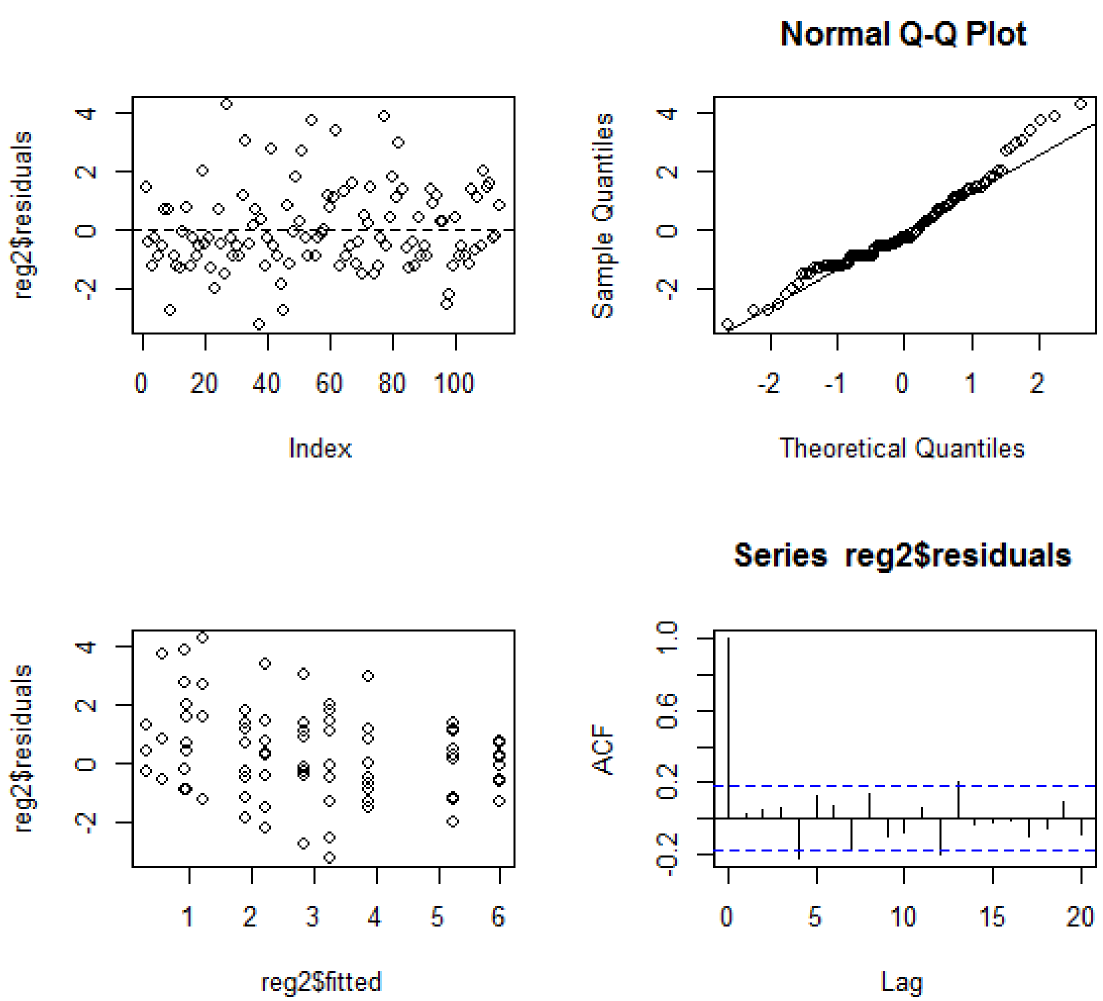
c) Refit the model with log transformation

```
reg2<-lm(log(light.train+1)~month)
summary(reg2)

##
## Call:
## lm(formula = log(light.train + 1) ~ month)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2683 -0.9388 -0.2996  0.8237  4.2873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.91607    0.47680   1.921 0.057486 .
## month2      -0.34437    0.67430  -0.511 0.610657
## month3       0.02894    0.69277   0.042 0.966760
## month4       0.96834    0.69277   1.398 0.165215
## month5       1.90603    0.69277   2.751 0.007025 **
## month6       2.94850    0.69277   4.256 4.63e-05 ***
## month7       4.33019    0.69277   6.251 9.61e-09 ***
## month8       5.07286    0.69277   7.323 5.80e-11 ***
## month9       2.35226    0.67430   3.488 0.000718 ***
## month10      1.31742    0.67430   1.954 0.053467 .
```

```
## month11       0.31412    0.67430   0.466 0.642319
## month12      -0.61650    0.67430  -0.914 0.362722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.508 on 102 degrees of freedom
## Multiple R-squared:  0.603,  Adjusted R-squared:  0.5602
## F-statistic: 14.09 on 11 and 102 DF,  p-value: 4.378e-16
```

```
par(mfcol=c(2,2))
plot(reg2$residuals)
abline(h=0,lty=2)
plot(reg2$fitted, reg2$residuals)
qqnorm(reg2$residuals)
qqline(reg2$residuals)
acf(reg2$residuals)
```




```
logPI2<-
predict.lm(reg2,newdata=data.frame(month=factor(cycle(light.test))),
```

```
interval="prediction")
PI2 = exp(logPI2)-1
light.test>PI2[,"lwr"] & light.test<PI2[,"upr"]
```

```
##       Mar  Apr  May  Jun  Jul  Aug  Sep
## 2011 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Comment: the model looks MUCH BETTER. Residul plots look random with no significant trend as we suppose, and the Q-Q plot looks more like a normal distribution. And the ACF value still shows no relation to the lag h. Most importantly, now we can fit all the test value in the prediction interval.

d)Compare the fit and performance of the two models. Which, if any, satisfies the fundamental assumptions of a regression model?

The second model definitely more satisfy the fundamental assumptions of a regression model. It has zero mean errors with normal distribution, and it fits the test value well.