Detecting_Flu_Epidemics_via_Search_Engine_Query_Dat

Suchitra

```
fluTrain = read.csv("FluTrain.csv")
str(fluTrain)
```

```
## 'data.frame':    417 obs. of  3 variables:
##  $ Week   : Factor w/ 417 levels "2004-01-04 - 2004-01-10",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ ILI    : num  2.42 1.81 1.71 1.54 1.44 ...
##  $ Queries: num  0.238 0.22 0.226 0.238 0.224 ...
```

```
#View(head(fluTrain))
which.max(fluTrain$ILI)
```

```
## [1] 303
```

```
#highest ILI in which week
fluTrain[303,]
```

```
##                        Week      ILI Queries
## 303 2009-10-18 - 2009-10-24 7.618892       1
```

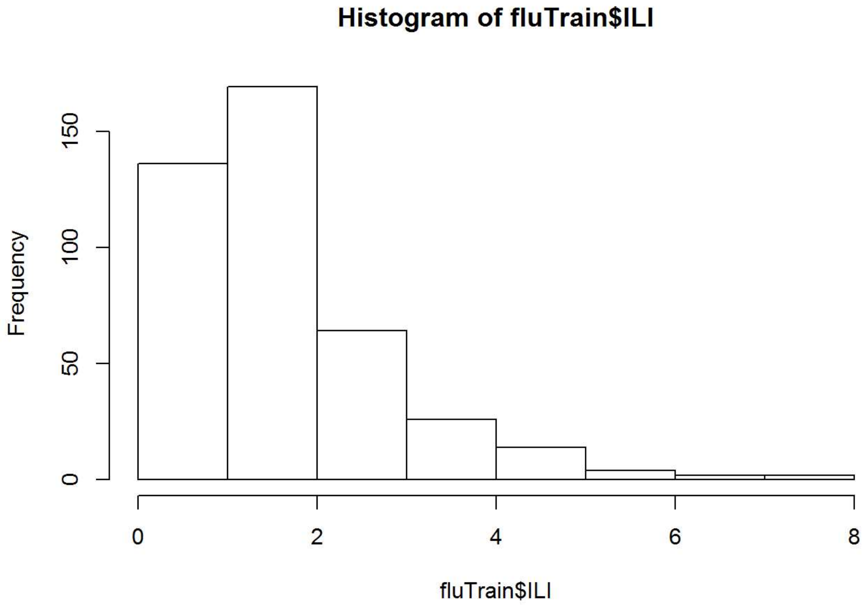
```
which.max(fluTrain$Queries)
```

```
## [1] 303
```

```
fluTrain[303,]
```

```
##                        Week      ILI Queries
## 303 2009-10-18 - 2009-10-24 7.618892       1
```

```
hist(fluTrain$ILI)
```



```
plot( fluTrain$Queries,log(fluTrain$ILI))
# can use a linear model as the relationship is linear
flureg= lm(log(ILI)~ Queries, data = fluTrain)
summary(flureg)
```

```
## 
## Call:
## lm(formula = log(ILI) ~ Queries, data = fluTrain)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.76003 -0.19696 -0.01657  0.18685  1.06450 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.49934    0.03041  -16.42   <2e-16 ***
## Queries      2.96129    0.09312   31.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2995 on 415 degrees of freedom
## Multiple R-squared:  0.709,  Adjusted R-squared:  0.7083 
## F-statistic:  1011 on 1 and 415 DF,  p-value: < 2.2e-16
```

```
#adjusted r^2 =  0.7083
#R-squared = exp(-0.5*Correlation)

fluTest = read.csv("FluTest.csv")
PredTest1 = exp(predict(flureg, newdata=fluTest))

which(fluTest$Week == "2012-03-11 - 2012-03-17")
```

```
## [1] 11
```

```
PredTest1[11]
```

```
##       11 
## 2.187378
```

```
#(Observed ILI - Estimated ILI)/Observed ILI

SSE = sum((PredTest1-fluTest$ILI)^2)
RMSE = sqrt(SSE / nrow(fluTest))
RMSE
```

```
## [1] 0.7490645
```

```
#training a time series model
# we take the lag as 2 weeks

#install.packages("zoo")
install.packages("zoo", repos = "http://cran.us.r-project.org")
```

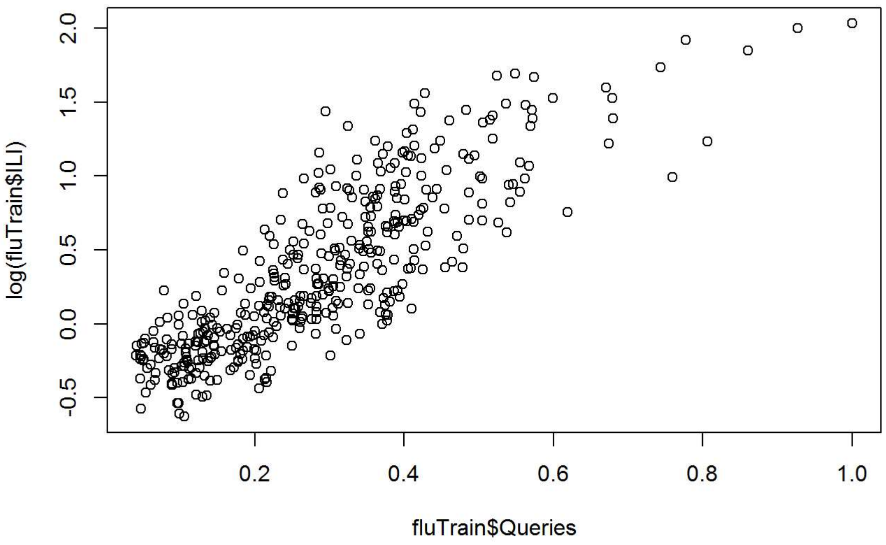
```
## Installing package into 'C:/Users/Suchitra/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)
```

```
## package 'zoo' successfully unpacked and MD5 sums checked
## 
## The downloaded binary packages are in
##  C:\Users\Suchitra\AppData\Local\Temp\RtmpcRxfFW\downloaded_packages
```

```
library(zoo)
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```



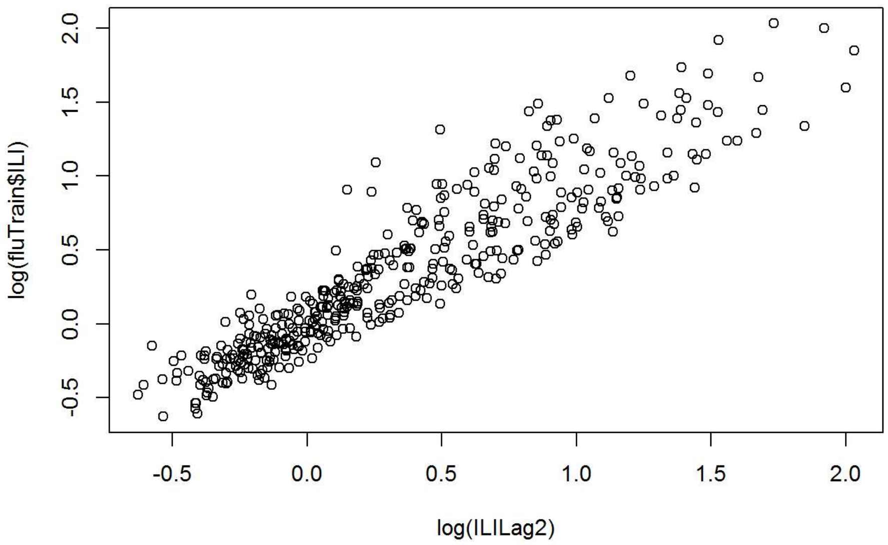
```
ILILag2 = lag(zoo(fluTrain$ILI), -2, na.pad=TRUE)

fluTrain$ILILag2 = coredata(ILILag2)
summary(fluTrain)
```

```
##                         Week          ILI           Queries       
##  2004-01-04 - 2004-01-10:  1   Min.   :0.5341   Min.   :0.04117  
##  2004-01-11 - 2004-01-17:  1   1st Qu.:0.9025   1st Qu.:0.15671  
##  2004-01-18 - 2004-01-24:  1   Median :1.2526   Median :0.28154  
##  2004-01-25 - 2004-01-31:  1   Mean   :1.6769   Mean   :0.28603  
##  2004-02-01 - 2004-02-07:  1   3rd Qu.:2.0587   3rd Qu.:0.37849  
##  2004-02-08 - 2004-02-14:  1   Max.   :7.6189   Max.   :1.00000  
##  (Other)                :411                                     
##     ILILag2      
##  Min.   :0.5341  
##  1st Qu.:0.9010  
##  Median :1.2519  
##  Mean   :1.6754  
##  3rd Qu.:2.0580  
##  Max.   :7.6189  
##  NA's   :2
```

```
#training a time series model

plot(log(ILILag2), log(fluTrain$ILI))
```



```
flureg2 = lm(log(ILI)~ Queries + log(ILILag2), data = fluTrain)
summary(flureg2)
```

```
## 
## Call:
## lm(formula = log(ILI) ~ Queries + log(ILILag2), data = fluTrain)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.52209 -0.11082 -0.01819  0.08143  0.76785 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.24064    0.01953  -12.32   <2e-16 ***
## Queries       1.25578    0.07910   15.88   <2e-16 ***
## log(ILILag2)  0.65569    0.02251   29.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1703 on 412 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9063, Adjusted R-squared:  0.9059 
## F-statistic:  1993 on 2 and 412 DF,  p-value: < 2.2e-16
```

```
#adjusted r^2 = 0.9059

#evaluating the time series model

ILILag2_test = lag(zoo(fluTest$ILI), -2, na.pad=TRUE)

fluTest$ILILag2 = coredata(ILILag2_test)
summary(fluTest)
```

```
##                        Week         ILI           Queries      
##  2012-01-01 - 2012-01-07: 1   Min.   :0.9018   Min.   :0.2390  
##  2012-01-08 - 2012-01-14: 1   1st Qu.:1.1535   1st Qu.:0.2772  
##  2012-01-15 - 2012-01-21: 1   Median :1.3592   Median :0.3924  
##  2012-01-22 - 2012-01-28: 1   Mean   :1.6638   Mean   :0.4094  
##  2012-01-29 - 2012-02-04: 1   3rd Qu.:1.8637   3rd Qu.:0.4874  
##  2012-02-05 - 2012-02-11: 1   Max.   :6.0336   Max.   :0.8054  
##  (Other)                :46                                   
##     ILILag2      
##  Min.   :0.9018  
##  1st Qu.:1.1359  
##  Median :1.3409  
##  Mean   :1.5188  
##  3rd Qu.:1.7606  
##  Max.   :3.6002  
##  NA's   :2
```

```
fluTest$ILILag2[1]= fluTrain$ILI[416]
nrow(fluTrain)
```

```
## [1] 417
```

```
fluTest$ILILag2[2]= fluTrain$ILI[417]
summary(fluTest)
```

```
##                        Week         ILI           Queries      
##  2012-01-01 - 2012-01-07: 1   Min.   :0.9018   Min.   :0.2390  
##  2012-01-08 - 2012-01-14: 1   1st Qu.:1.1535   1st Qu.:0.2772  
##  2012-01-15 - 2012-01-21: 1   Median :1.3592   Median :0.3924  
##  2012-01-22 - 2012-01-28: 1   Mean   :1.6638   Mean   :0.4094  
##  2012-01-29 - 2012-02-04: 1   3rd Qu.:1.8637   3rd Qu.:0.4874  
##  2012-02-05 - 2012-02-11: 1   Max.   :6.0336   Max.   :0.8054  
##  (Other)                :46                                   
##     ILILag2      
##  Min.   :0.9018  
##  1st Qu.:1.1535  
##  Median :1.3592  
##  Mean   :1.5368  
##  3rd Qu.:1.8554  
##  Max.   :3.6002  
## 
```

```
fluTest$ILILag2[1]
```

```
## [1] 1.852736
```

```
fluTest$ILILag2[2]
```

```
## [1] 2.12413
```

```
PredTest2 = exp(predict(flureg2, newdata=fluTest))
SSE = sum((PredTest2-fluTest$ILI)^2)
RMSE = sqrt(SSE / nrow(fluTest))
RMSE
```

```
## [1] 0.2942029
```

```
#RMSE = 0.2942029
#less the rmse, better the model
```