Stats 531
Winter, 2016
Midterm Exam

Name: ______________________________ **UMID #:** ______________________________

There are 3 sections (A, B and C) containing a total of 20 points. Points will be awarded for clearly explained and accurate answers.

Only pens and/or pencils should be out of your bag for the duration of the exam. You may not use access any electronic device, paper notes, or books during the exam.

Section	Points	Score
A	5	
B	11	
C	4	
Total	20	

We consider Google flu trends as a proxy for nationwide epidemiological reporting data on flu. Google flu trends (GFT) is a time series that was published by Google from 2008 to 2015. GFT uses search query data to try to reproduce the Centers for Disease Control time series of influenza-like illness (ILI). ILI is measured as the percentage of all hospital visits in the USA that are caused by flu-like symptoms (high fever with a cough). So far as GFT is a reliable proxy for ILI, it has the advantage that it is instantaneously available. It takes a few weeks for the ILI data to be assembled.

The two time series are shown in Figure 1. Both ILI and GFT are published each week.

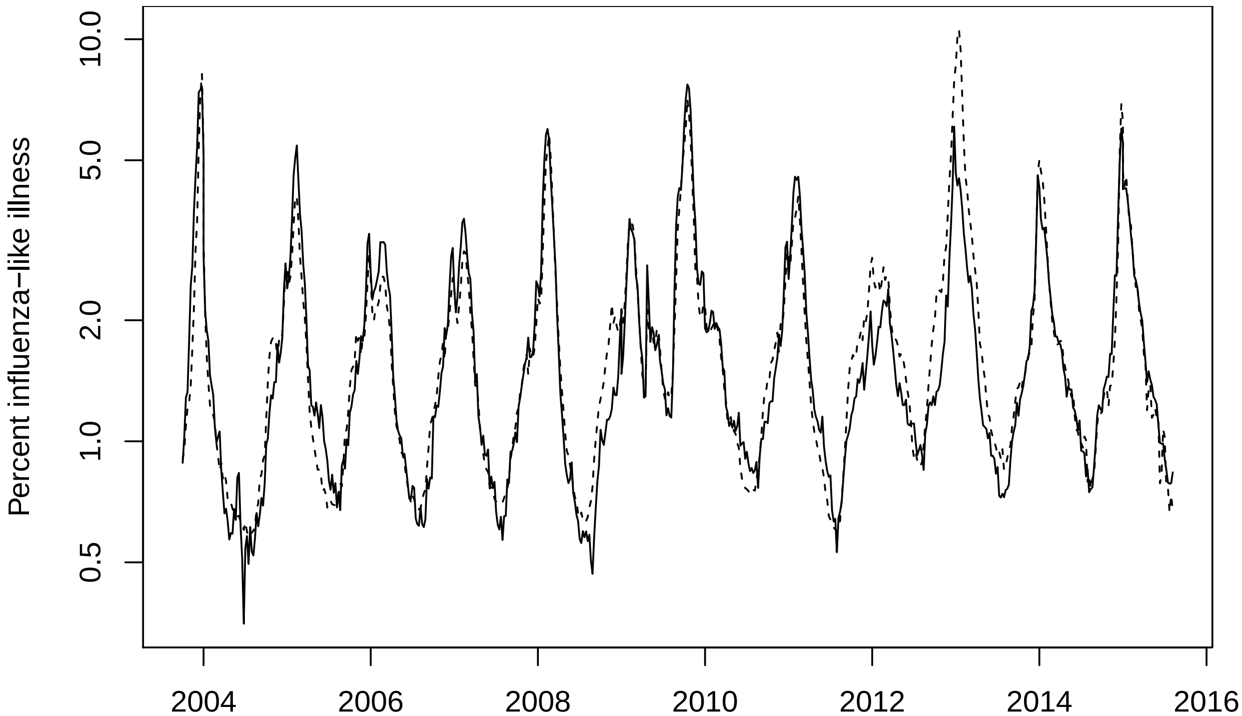

Figure 1: ILI (solid line) and GFT (dashed line) from September 2003 to June 2015, plotted on a log scale.

Section A. Exploratory data analysis.

A1. [3 points]. Look at Figures 1 and 2. Interpret these figures to describe strengths and weaknesses of GFT as a proxy for ILI.

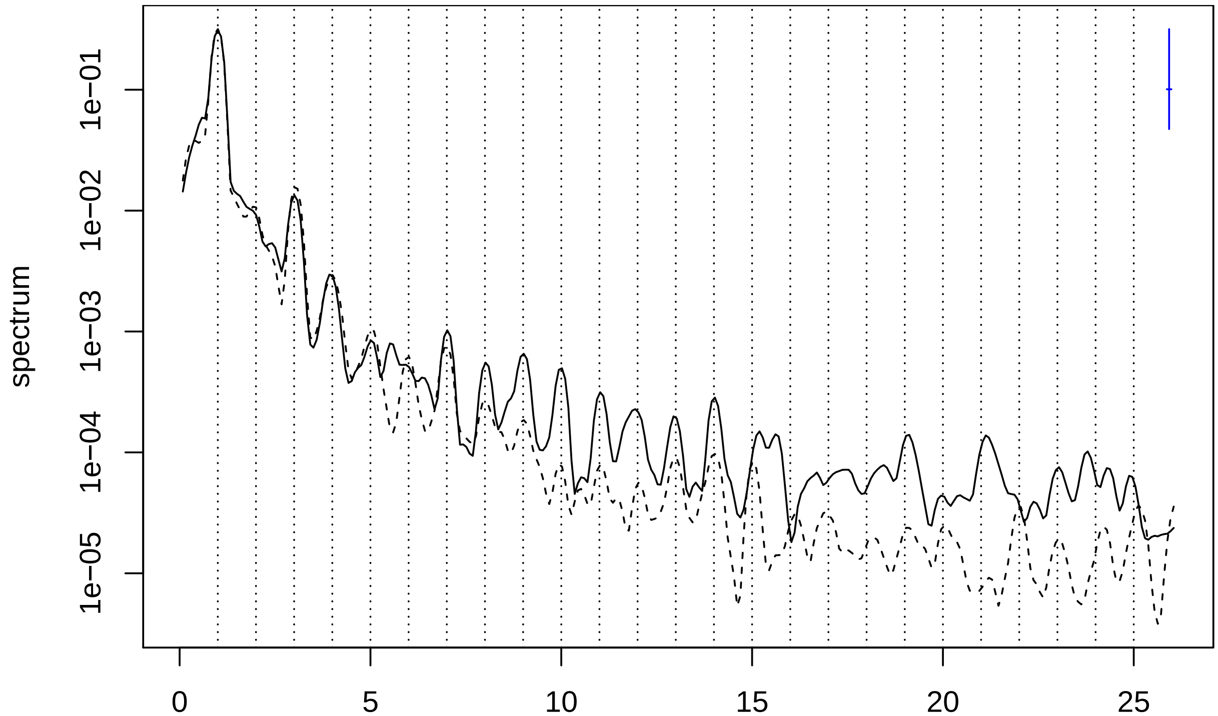

Figure 2: Smoothed periodogram for log(ILI) (solid line) and log(GFT) (dashed line).

A2. [2 points]. What are the units of frequency in Fig. 2? Explain how you reach your answer.

Section B. Fitting a model.

Can we do better than GFT? A simple way to do that would be to model the error arising from GFT, together with considering a linear transformation of GFT. This can be done by fitting a regression with ARMA errors model, as follows.

```
##
## Call:
## arima(x = log(ILI), order = c(1, 0, 1), xreg = log(GFT))
##
## Coefficients:
##          ar1      ma1  intercept  log(GFT)
##       0.9163  -0.1607     0.0375    0.8372
## s.e.  0.0183   0.0477     0.0402    0.0301
##
## sigma^2 estimated as 0.009379:  log likelihood = 566.96,  aic = -1123.93
```

B1. [5 points]. Write in full detail the model for which the above computation gives a maximum likelihood estimate.

Now we consider a table of AIC values for different ARMA(p,q) error specifications:

	MA0	MA1	MA2	MA3	MA4
AR0	-248.10	-678.77	-862.10	-952.50	-984.17
AR1	-1114.80	-1123.93	-1122.60	-1120.73	-1118.77
AR2	-1122.89	-1122.65	-1122.55	-1123.36	-1122.74
AR3	-1122.22	-1124.96	-1123.29	-1119.51	-1120.18
AR4	-1120.79	-1118.88	-1123.20	-1121.66	-1119.34

B2. [2 points]. What do the results in this table suggest about the suitability of the ARMA(1,1) choice made above for the regression error model.

B3. [2 points]. Explain the evidence in this AIC table for or against numerical difficulties in maximization and/or evaluation of the likelihood.

B4. [2 points]. The two panels in Figure 3 show a smoothed periodogram and a sample autocorrelation function for the residuals of the above regression with ARMA errors. Interpret these figures to help assess this model specification and suggest possible improvements.

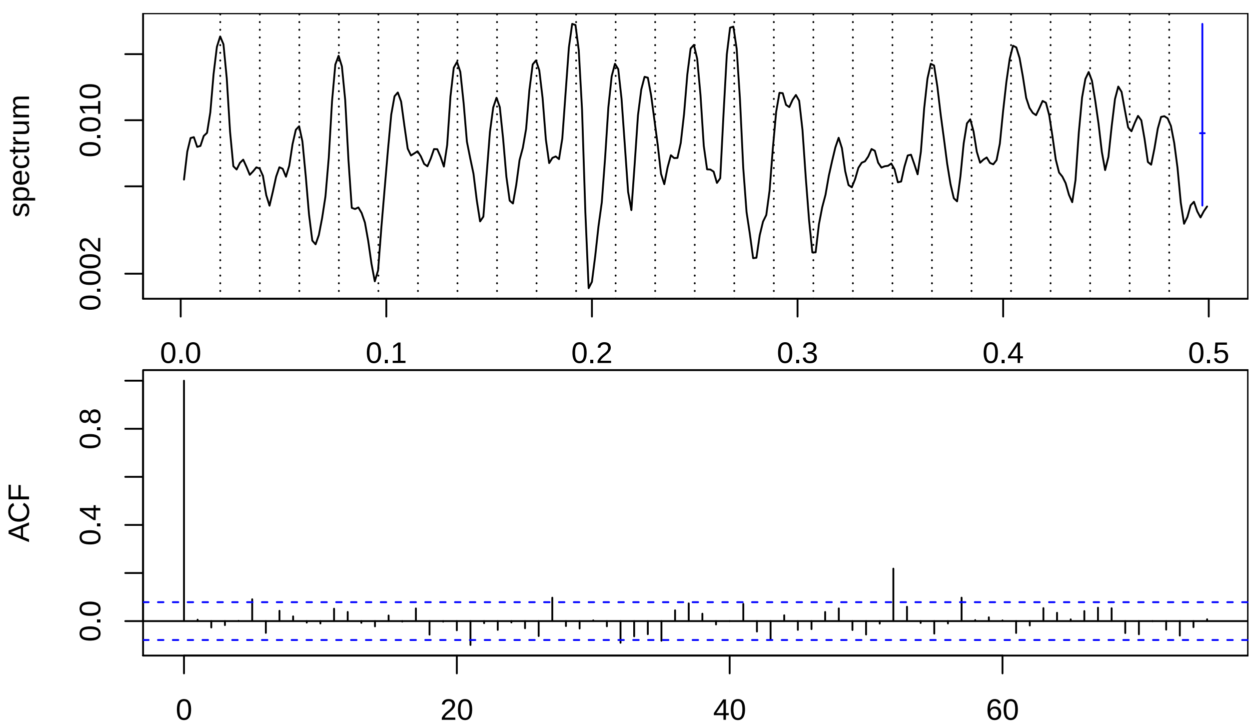

Figure 3: Spectrum and sample autocorrelation function for the residuals of the regression with ARMA errors fitted above

Section C. Consideration of the logarithmic transformation.

C1. [4 points]. What issues would you consider when deciding whether to analyzing ILI and GFT on a logarithmic scale, as we have done above, or on an untransformed scale? As part of your answer, you may consider the analysis below.

Fitted regression with ARMA errors on an untransformed scale:

```
##
## Call:
## arima(x = ILI, order = c(1, 0, 1), xreg = GFT)
##
## Coefficients:
##          ar1     ma1  intercept     GFT
##       0.8870  0.2096     0.4526  0.7202
## s.e.  0.0199  0.0410     0.1072  0.0270
##
## sigma^2 estimated as 0.05038:  log likelihood = 45.63,  aic = -81.26
```

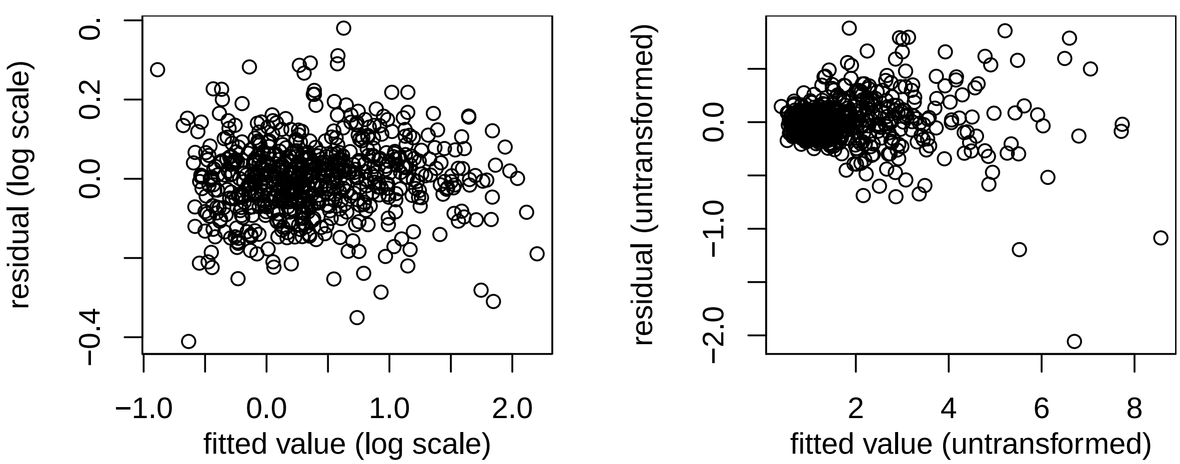

Figure 4: Residual vs fitted value plots for the regression on the log scale (left hand side) and natural, untransformed scale (right hand side).