Stats 531
Winter, 2016
Midterm Exam

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.

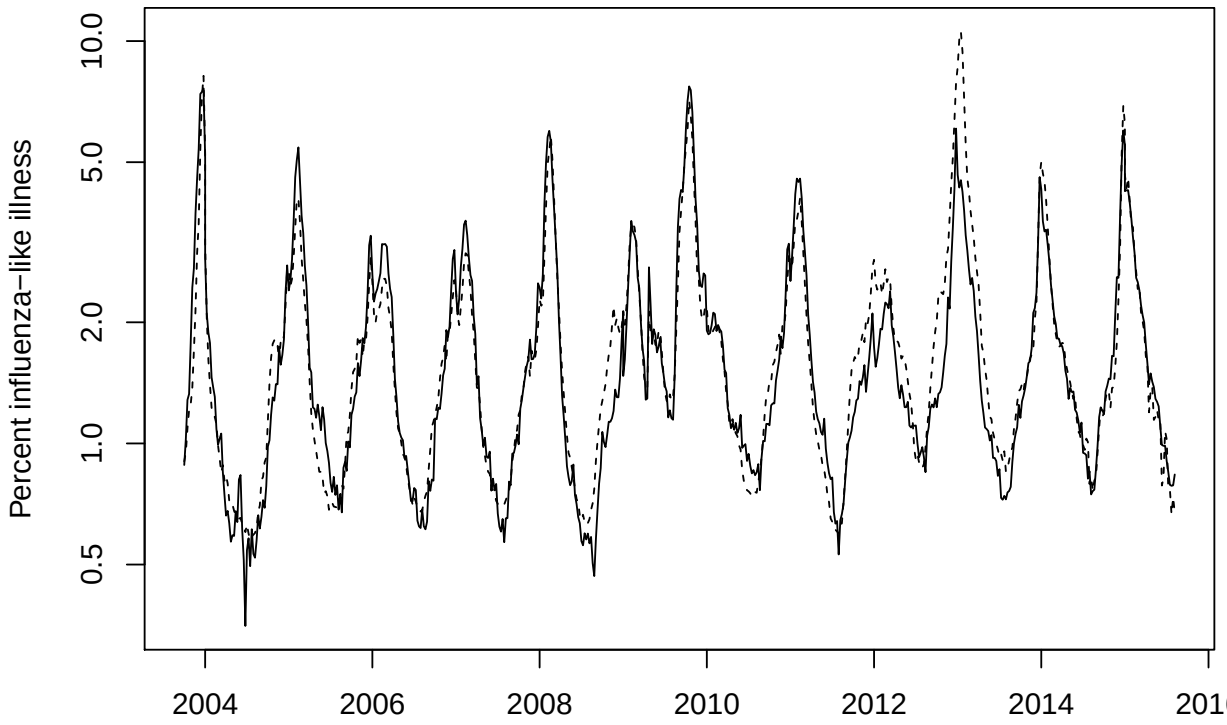

Figure 1: ILI (solid line) and GFT (dashed line) from September 2003 to June 2015, plotted on a log scale.

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.

GFT captures the main features of the ILI data. We can see this from the timeplot, corroborated with the lower power of GFT compared to ILI at high frequencies. It is smoother (has lower power at high frequencies). In particular, ILI has a more complex, less sinusoidal, seasonality than GFT, since ILI has considerably more power at the high seasonal frequencies, with frequencies at an integer number of cycles per year. From Fig. 1, we see that GFT sometimes substantially mis-estimates peaks and troughs in ILI (e.g., the 2013 peak and 2004 trough are over-estimated by a factor of about 2).

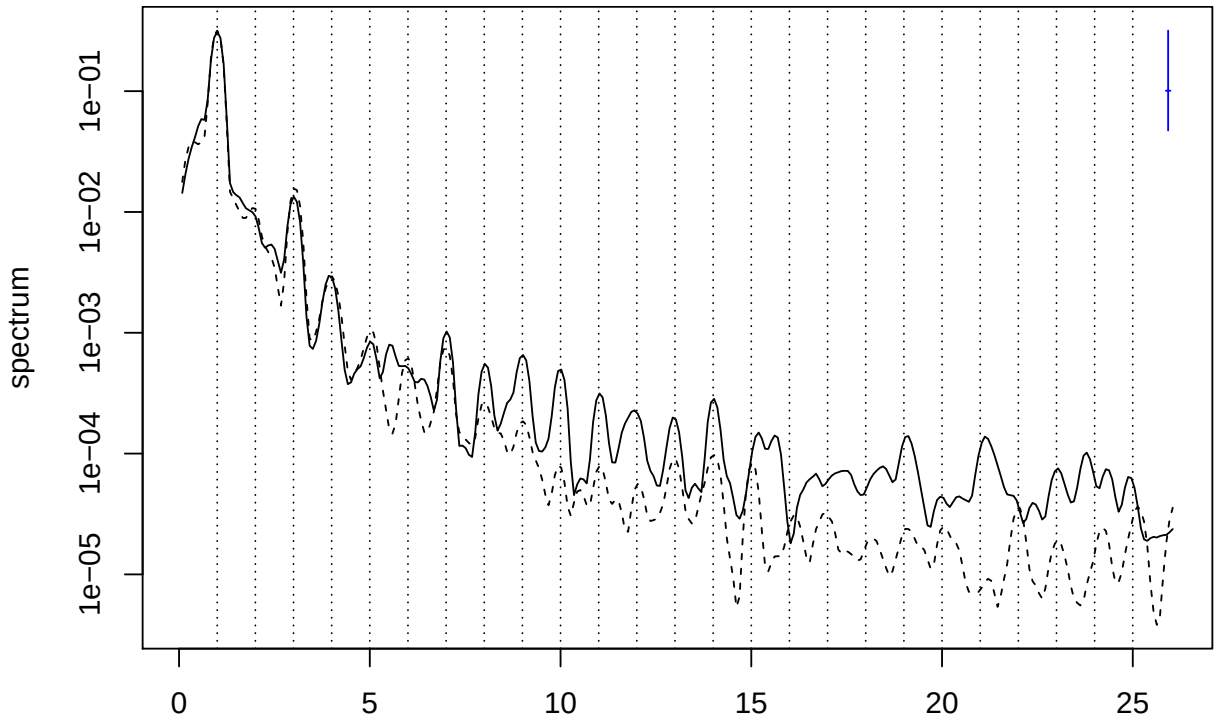

Figure 2: Smoothed periodogram for log(ILI) (solid line) and log(GFT) (dashed line).

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

Cycles per year.

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.

Write $y^*_{1:N}$ for the N values of log(ILI), at times $t_{1:N}$. Write $z_{1:N}$ for the corresponding values of log(GFT). We model $y^*_{1:N}$ conditional on $z_{1:N}$ as a realization of the time

series model $Y_{1:N}$ defined by

$$Y_n = \alpha + \beta z_n + \epsilon_n,$$

for which $\epsilon_{1:N}$ is a stationary, causal, invertible, Gaussian ARMA(1,1) model satisfying a stochastic difference equation,

$$\epsilon_n = \phi\epsilon_{n-1} + \omega_n + \psi\omega_{n-1},$$

where $\{\omega_n\}$ is Gaussian white noise, $\omega_n \sim N[0, \sigma^2]$.

The maximum likelihood estimate computed above corresponds to $\sigma^2 = 0.0094$, $\phi = 0.92$, $\psi = -0.16$, $\alpha = 0.038$ and $\beta = 0.84$.

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.

In this table, only the ARMA(3,1) model has a lower AIC and this difference is small. We prefer to work with a smaller model. Although AIC rewards model simplicity, does so only as far as complexity leads to poor prediction from overfitting. Other considerations are that smaller models reduce problems with parameter identifiability, invertibility, and numerical stability which we know are common when fitting larger ARMA models. There is no compelling reason from this table to choose something other than ARMA(1,1).

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

Adding a parameter in a nested model should not logically be able to increase the AIC by more than 2 units. We can find plenty of situations where that logic is violated.

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.

The estimated spectrum has peaks at many seasonal frequencies (multiples of 1/52 yr^{-1}) and the sample ACF has a substantially nonzero value at the seasonal period (1yr=52week). Apart from this evidence of modest but non-negligible seasonality in the residuals, there is not much other deviation from white noise: the spectrum is otherwise flat, apart from the seasonal peaks, and the sample ACF values at lags other than 52 are small. We could try adding a seasonal component to the model, such as SARMA(1,1)$\times$(1,0)$_{52}$.

Section C. Consideration of the logarithmic transformation.

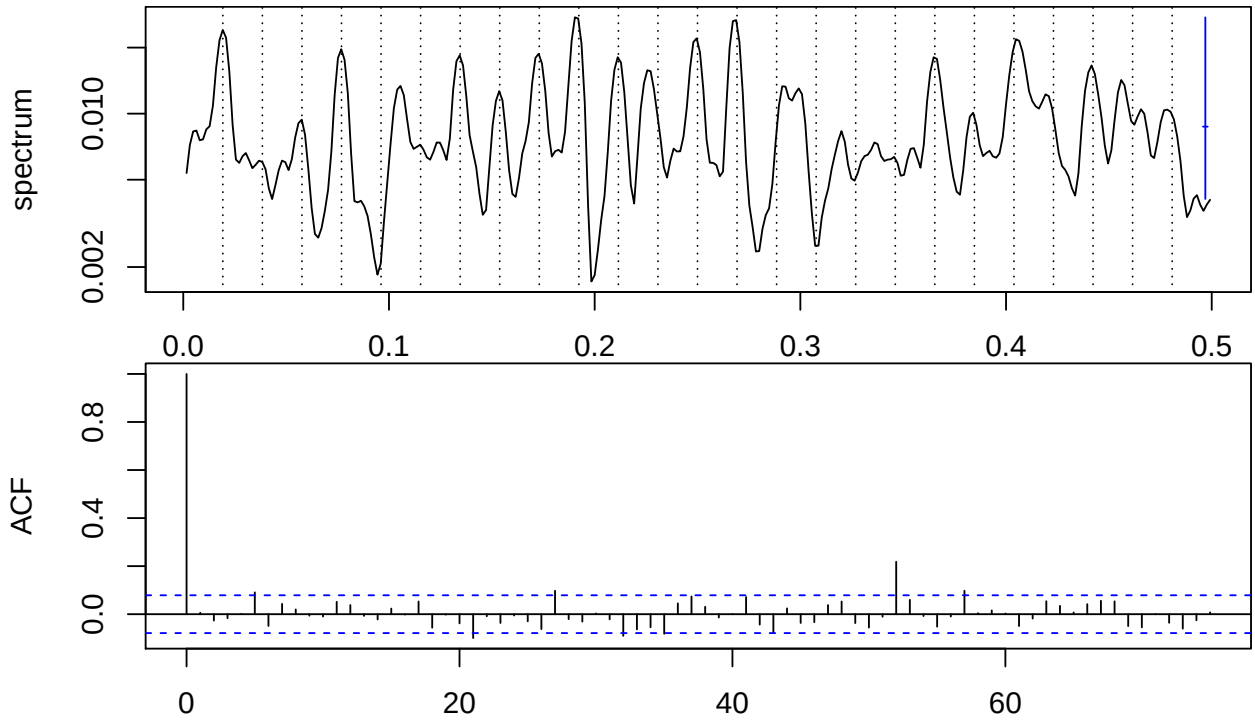

Figure 3: Spectrum and sample autocorrelation function for the residuals of the regression with ARMA errors fitted above

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.

Gaussian white noise is a better model for the residuals on a log scale. To see that, notice the heteroskedasticity in the right hand, untransformed, panel of Fig. 4. Larger values of ILI correspond to larger residuals.

It might be expected that errors in predicting ILI should be larger, in absolute terms, when ILI itself is more prevalent. Fitting on the log scale respects that expectation.

The regression coefficient for GFT is closer to 1 on the log scale, which might be taken to indicate that this is a better scale for approximating ILI with GFT.

AIC values are not directly comparable. However, we can do a Jacobian transformation of the likelihoods, by transforming the likelihood for the log data back to the natural scale. The Jacobian transformation tells us that, if $Z = \log(Y)$, then

$$f_Z(\log(y)) = \frac{1}{y} f_Y(y).$$

Thus, if the data are $z^* = \log(y^*)$, the log likelihood is

$$\log f_Z(z^*) = \log f_Y(y^*) - \log(y^*).$$

Therefore, we should compare the log likelihood of 45.6 (on the untransformed scale)

with

$$567.0 - \sum_{n=1}^{N} \log y_n^* = 325.6.$$

Comparing these log likelihoods shows that the model fits much better on the log scale.

One can ask whether it is more or less scientifically meaningful to model on a log scale. However, this is not too important: One can always transform a fit on a log scale back to the untransformed scale.

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
```

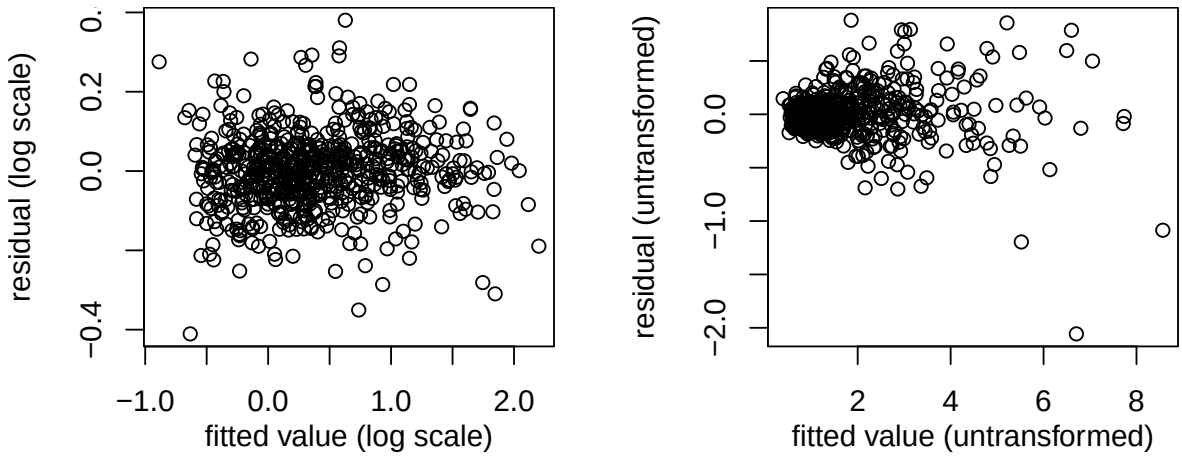

Figure 4: Residual vs fitted value plots for the regression on the log scale (left hand side) and natural, untransformed scale (right hand side).