Assignment 4
MY 459 / MY 559

Benjamin Lauderdale
Methodology Department
London School of Economics

Class/Homework Assignment

We will need the mgcv library and our rmse function for this week's assignment.

```
library(mgcv)

## Loading required package: nlme
## This is mgcv 1.8-4. For overview type 'help("mgcv-package")'.

rmse <- function(fitted,observed) sqrt(mean((observed-fitted)^2))
```

To fit a generalized additive model, you can use the gam command, which works almost exactly like the glm command. For example, the hypothetical command

```
gam.fit <- gam(yvar ~ s(xvar1) + xvar2 + as.factor(xvar3)}
```

would fit an additive model for the outcome variable yvar as a function of a smooth spline of xvar1, a linear term for xvar2 and dummy variables for xvar3. Just as with the glm command, if you do not specify a family in the command, normal errors will be assumed.

While it is not required, the smooth spline term of the additive model formula should be given a maximum degrees of freedom to consider. While the gam command will use generalized cross-validation (GCV) to determine the smoothness of the spline function for a particular variable, for computational efficiency reasons it will only check a range of degrees of freedom up to a particular limit.[1] The default for this is 9 (k=10), which is effectively arbitrary. It is a good idea to specify the maximum degrees of freedom explicitly.

```
gam.fit <- gam(yvar ~ s(xvar1,k=10) + xvar2 + as.factor(xvar3)}
```

The maximum degrees of freedom specified for the xvar1 spline by the above command is $k-1=9$.

[1] This is slightly inaccurate, see the documentation for further explanation.

Data

When governments report employment data, they often perform seasonal adjustments. Seasonal adjustment is useful because certain times of year are consistently better/worse for employment levels. If we care about the long-run trends, we need to get rid of this seasonal noise so we can figure out whether the economy is doing well or not. But we also might be interested in the patterns of seasonal noise themselves, so we do not necessarily want to just smooth them away.

The data set we will consider for this assignment is the US, non-farm, employment level from 1992 to 2012 (`USPayrollData.csv`). Non-farm employment is considered for much the same reason that we want to do the seasonal adjustment: farm employment is highly seasonal. There are four variables:

- Year - Integer
- Month - Integer
- PAYNSA - Total nonfarm employees in thousands of persons, not seasonally adjusted.
- PAYEMS - Total nonfarm employees in thousands of persons, seasonally adjusted.

The data set includes both the figures reported by the US government without seasonal adjustment (PAYNSA), as well as the figures with the official seasonal adjustment (PAYEMS). We will focus on the variable PAYNSA, using PAYEMS to see how close we can come to reproducing the US government's method for performing the seasonal adjustments. We are going to create our own seasonal adjustment by building and interpreting an additive model for PAYNSA.

1. PAYNSA is reported in thousands of jobs. We could log-transform this variable before analysing it using an additive model. How will the interpretation of our regression function change if we log-transform PAYNSA? Explain the connection between the additivity assumption in the GAM and this decision about whether to log-transform PAYNSA.

2. Using logPAYNSA as your dependent variable, calculate the root mean square error with respect to the seasonally adjusted data (i.e. comparing logPAYNSA to logPAYEMS. Give a one sentence description of what this RMSE tells us.

3. Create a continuous time variable for time which is equal to:

$$\text{Time} = \text{Year} + \frac{\text{Month} - 1/2}{12}$$

   Plot both the nonadjusted and seasonally adjusted figures against Time.

4. Fit an additive model for employment with dummy variables for each month of the year and a smooth spline of the variable Time using the default number of degrees of freedom.

   (a) Create a table with the dummy variable coefficients for each month. Describe the trend across months: which are the lowest and which are the highest employment months? Which month is closest to the average of the months?

   (b) Create a plot of the smooth spline for Time (just call the plot command on the saved gam object).

   (c) Create a plot of the model residuals as a function of Time. Is there evidence of oversmoothing in the residuals? If there is, adjust the degrees of freedom for the spline until you are satisfied that there is no longer a problem.

5. How can you use the additive model you have just fit to calculate a measure of seasonally adjusted employment? (Hint: think carefully about the three additive components of the model: the smooth spline for Time, the monthly dummy variables, and the residuals.) Explain how you are going to construct your seasonally adjusted employment estimates.

6. Calculate the RMSE of your seasonally adjusted employment measure by comparison to the seasonally adjusted employment measure provided with the data. Compare this to the results from Question 2, and explain what this comparison tells you about how well you have done at replicating the official seasonal adjustments.