Statistics 36-315: Statistical Graphics and Visualization

Homework 7
Date: March 20, 2006 Due: via Blackboard, March 29, 2006, 12:30pm

1. (20% of Grade) This question is based on two readings from Wainer (2005) distributed in class.
   
   (a) In Chapter 13, Wainer describes two different time series datasets, one for winning times in the Boston Marathon and the second on faxes from Price-Waterhouse. As time series, what do they convey about the uses of trend lines?
   
   (b) In Chapter 21, Wainer returns to the graph for winning times in the Boston Marathon. What does his analysis there suggest regarding the use of smoothers for capturing the effect of trends in time series? How does the discussion relate to the possible use of loess smoothers?

2. (50% of Grade) In this problem, you will take a seemingly complex time-series and show that it is actually pretty simple. The time series is the yearly incidence rate of melanoma in Connecticut. Download melanoma.csv and the source lab08.r. This problem includes writing R code. Be sure to turn in all of your code, for all parts.

   (a) Graph melanoma versus time using connected dots. Adjust the aspect ratio according to the 45° rule, which should be done for all remaining plots. The upward trend is pretty disturbing (the rate is normalized for population), but there is more to the data than that.

   (b) Decompose the melanoma series into a linear trend and its residual. Start by graphing a trend line on top of a scatterplot. Use span=1.

   (c) Now extend the frame with a column for the residual. Graph the residual series with a trend line of appropriate smoothness. The series has length 37, so choose a span of the form \( k/37 \). The result should be a smooth oscillation, without small bumps. You will probably need to adjust the aspect ratio for this plot.

   (d) Plot the auto-correlation function of the oscillatory component. What is oscillatory component’s period, and how strong is the periodicity? Indicate where on the plot this is shown.

   (e) Make a spiral plot and calendar plot of the oscillatory component, using your estimated period. Make them easily-readable.

   (f) Find the cycle which is most out of sync with respect to the others. (“One of these things is not like the others...”) Identify it on both plots and give a brief explanation of how both plots show this.

   (g) Now extend the frame to include the second fit and its residual. The first fit is the “trend component”, the second fit the “oscillatory component”, and the final residuals are the “residual component”. Graph the residual component using your favorite method. There is a pattern among the large residuals. What is it?
(h) Melanoma is mainly caused by solar radiation. The column **sunspot** is the number of sunspots observed on the sun during that year. Make a superposition of **sunspot** and your oscillatory component. These series have very different scales, so you will need to standardize them first. The function **match.quantiles** will take a matrix whose columns are time series and returns a new matrix where the columns are standardized by a scale and shift in order to match the first column.

(i) The melanoma oscillation is related to the sunspot number. Plot the cross-correlation function between the oscillatory component and the sunspot data. This will tell you the precise lag between the two. The cross-correlation is computed similar to **acf**, as follows:

\[ ccf(ts1,ts2) \]

where **ts1** and **ts2** are vectors, e.g. columns of your frame. What is the lag between sunspots and melanoma? Indicate where on the plot this is shown.

(j) What connection do your analysis and plots reveal between sunspots and melanoma?

3. (30% of grade) In this problem, you will use the rate-of-change transformation to understand population growth in the U.S. A matrix of population data is in **hw7.csv**. Each column is a state and each row is a decade during which a census was taken. The functions you need are in **lab9.r**.

(a) Start by making a line chart where the horizontal axis is time. This orders the states by total population, which obviously favors larger states.

(b) Now change it into a growth chart. Plotting the growth since 1790 is more fair to small states, except Ohio which wasn’t a state until 1803. Still, there isn’t much to get from this plot except a basic ordering.

(c) Now make a new matrix containing the rate-of-change transformation of the population growth curves. This is done using the function **rate**, which takes a matrix as input and returns a new matrix where each number is the instantaneous percent change between rows.

(d) Make a line chart of the rate-of-change for each state over time. Use the option **ylim** to zoom in on the area between -10% and 50% growth rate, and choose a good aspect ratio. If you can’t print in color, you can hand in the plot in black and white, though it obviously won’t be as readable.

(e) Based on the similarity of their rate curves, pair up the states. For each of the three pairs, point out briefly what is the main similarity between the growth rates of the two states, and the main difference. Refer to dates but otherwise avoid using numbers in your description.

(f) After 1940, one state is not like the others. Which is it and how is it different? (Notice that this is much harder to see in the first two plots.)