STA 437/1005, Fall 2010 — Assignment #1

Due at start of lecture on November 15. Please hand it in on 8 1/2 by 11 inch paper, stapled in the upper-left corner, without any folder or other packaging around it. If really necessary, you can submit it by email (to radford@stat.utoronto.ca*), but please do this only if you can't easily hand in a paper copy.*

This assignment is worth 10% of the course grade. It is to be done by each student individually. You may discuss this assignment in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion of this assignment with any written notes or other recordings, nor receive any written or other material from anyone else by other means such as email.

In this assignment, you will analyse two datasets taken from the scientific literature, exploring them graphically, and applying principal components analysis. You should do your analyses using R.

For both datasets, you should produce a report explaining your findings, and justifying them with suitable plots or textual output from R. You should hand in only a reasonable amount of output, as needed for your analysis, not every conceivable plot. You *must* hand in a listing of the R commands you used for the analysis, including the commands used to produce the plots you hand in. You do not need to repeat in your report what I say below about the datasets.

The datasets, along with some relevant links, are available from the course web page, at

`http://www.utstat.utoronto.ca/~radford/sta437/`

I will also post some hints on using R for this assignment. You should check the web page regularly in case there are any corrections made to the assignment (any corrections will be highlighted in bold at the top of the web page).

The two parts are worth equal marks.

Part I: This dataset contains the daily average wind speeds at 12 locations in the Republic of Ireland, for every Monday from 1961 to 1978, and every preceding Sunday, extracted from a larger dataset containing average wind speeds for all days from 1961 to 1978.

The data is discussed in the following paper:

> Haslett, J. and Raftery, A. E. (1989) "Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource" (with discussion), *Applied Statistics*, vol. 38, pp. 1–50.

These authors' purpose was prediction of wind power resources at possible new locations, but for this assignment, we will just consider predicting the average wind speed on Mondays from the wind speeds on the preceding Sundays. When looking at pairs of days a week apart, the Sunday/Monday pairs are almost independent, so we can ignore the time series aspect of the full dataset. (Of course, a Sunday and the following Monday are not independent; that dependence is what we will be trying to model.)

Two data files, one for Sundays and one for Mondays, are on the webpage. Each has a header line giving the names of the variables, which are abbreviations for the 12 locations.

The order of the 12 locations is roughly from South to North. The observations have labels composed of year/month/day (two digits each). This format is suitable for reading with the `read.table` function using the `head=TRUE` option.

We will sometimes look at average wind speeds over all 12 locations. You can compute these averages using the `rowMeans` function.

The paper by Haslett and Raftery says that taking the square roots of the wind speeds improves normality of the data, and that there is an effect of season on wind speed. You should examine whether these claims are correct, looking at the average wind speed over all 12 locations on Sundays. You should consider three possibilities — no transformation, taking the log of the average wind speed, and taking the square root of the average wind speed. For each possibility, you should fit a linear model of average wind speed or its transformation (using the `lm` function) with sin and cos of the "time of year angle" as covariates. The time of year angle can be computed as 2*pi*((week/52.18)%%1), where "week" is a vector of week numbers, and 52.18 is the number of weeks in a year. Here, "%%1" takes the fractional part of a number. The week numbers start at 1 for the first Sunday, and increase by 1 for each observation; such a sequence can be obtained with something like `1:nrow(data)`.

After fitting this linear model for the seasonal effect, you should find the residuals, and check whether there are still any seasonal effects in the residuals, and whether the residuals appear to be normally distributed. Report your conclusions. Checking for seasonal effects can be done by plotting observations against the year angle. Checking for normality can be done using histograms and QQ plots.

For the rest of the analysis, you should look at the the square roots of the wind speeds minus the fitted value for that day, from the linear model fit to the square root of the average wind speed on Sunday. You can ignore the difference in time of year between a Sunday and a following Monday. (Note that this procedure is not necessarily the best — it assumes that all 12 locations have the same seasonal effects, and that the order of taking square roots and averages isn't important — but it's what we'll do for this assignment.)

You should start by looking at pair-wise scatterplots of this (transformed) Sunday data from the 12 locations. Report any outliers that you find, and comment on the correlations between locations. You can also look at the correlation matrix.

Next, using the transformed data for Sundays, you should find the principal components, using the `prcomp` function, without scaling (ie, use the covariance matrix rather than the correlation matrix). You should look at the components found (the rotation and standard deviations) and comment on what meaning (if any) can be assigned to them.

Finally, you should look at predicting the (transformed) average wind speed on Monday using data from the previous Sunday. You should consider linear models with the following sets of covariates:

- Just the (transformed) average wind speed on the previous Sunday.
- All the (transformed) wind speeds on the previous Sunday.
- The first 1, 2, or 3 principal components from the previous Sunday.

Comment on the performance of each of these methods, using "Adjusted R-Squared" (output by `summary(lm(...))`) as the criterion for how good each linear model is. Can the 12 measurements be condensed to fewer numbers with little or no loss of predictive performance?

Part II: This data set contains the levels of expression of genes in samples of lung cancer cells from 39 patients. The aim is to use the expression levels of these genes to predict whether the lung cancer will recur in a patient after surgery.

The data is from the following study:

> Wigle, D., Jurisica, I., N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, S. Keshavjee, G. Darling, T. Winton, B. Breitkreutz, P. Jorgenson, M. Tyers, F. A. Shepherd, M.S. Tsao. (2002) "Molecular profiling of non-small cell lung cancer and correlation with disease-free survival", *Cancer Research*, vol. 62, pp. 3005–3008.

The measurement for a particular gene and patient is the ratio of the expression level of that gene in the patient's cancer cells divided by the expression level of that gene is a reference sample from other tissues. There may have been some normalization done as well (I can't tell what exactly was done from the paper and data file). All the measurements are positive.

The study above started with about 19,000 genes, and then looked a subset of 2899 genes. I reduced the number further to the 1644 genes for which the expression measurements were missing in at most two patients. I filled in any missing measurements using the median of the non-missing measurements over all 1644 genes and 39 patients. (This is simple, but rather crude; a better method would be desirable in a real application.)

Two data files are provided. One contains the expression levels of the 1644 genes (with missing values filled in) for the 39 patients, with a header line giving information identifying the gene. The observations are labelled with identifiers for the patients. The other data file contains information on the patients (ordered the same as in the gene expression data file), with a header line giving the names of the variables ("id", "type", "stage", and "recur"). Only the "recur" variable is used in this assignment. It is 0 if no recurrence of cancer was observed, and 1 if cancer did recur. Both data files are suitable for reading with the `read.table` function using the `head=TRUE` option.

In this assignment, we will see how well the "recur" variable can be predicted from the gene expression data, by fitting a linear model. A linear model is not really appropriate for a response variable like this, which takes values 0 or 1 — later, we will cover logistic regression models, which would be more appropriate. Nevertheless, it is meaningful to look at how well a linear model does, as a way of seeing how informative the gene expression data is about cancer recurrence.

It is not possible to do a standard least squares fit of a linear model for recurrence using all 1644 gene expression measurements as covariates, with only 39 patients. (Least squares requires that the number of observations be at least as large as the number of covariates.) Accordingly, we will look at predicting recurrence using some small number of principal components found from the 1644 gene expression measurements.

Several issues arise in doing this:

- Should the gene expression measurements be transformed before finding principal components?
- Are there any outliers that should be removed or adjusted before finding principal components?

- Should the principal components be found from the covariance matrix or the correlation matrix?
- How many principal components should be used in a linear model for recurrence?

You should investigate these issues for this assignment, and report your results. For transformations, you should consider at least the possibilities of no transformation and a log transformation. You should assess how well a set of covariates does at predicting the "recur" variable using the "Adjusted R-Squared" output from `summary(lm(...))`. You may also want to look at scatterplots of pairs of principal components (or the original gene expression measurements) in which patients with and without recurrence of cancer are identified by different colours or different plot symbols.