STAT 100 Review Notes April 7, 2010

3. (Time Series - 67 choices) Several topics fit here: random walks, the stock market, smoothing, simulation, forecasting time series, and a few others.

4. (Regression - 64 choices) Regression and regression lines, residual plots, correlation, covariates, data mining, and weather forecasting.

Time Series

To discuss Time Series, we need to say a bit more about "Independence" and its relationship to "random sampling".

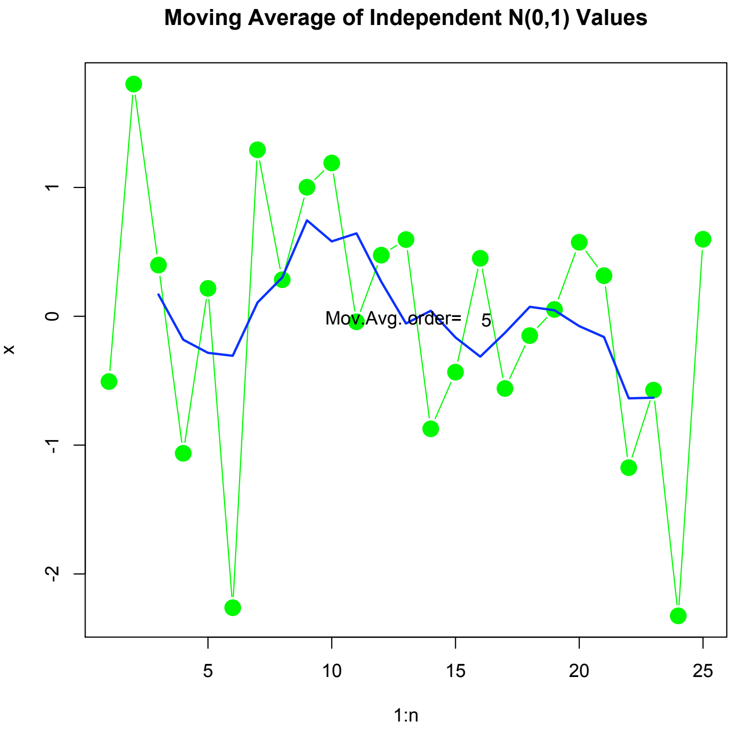
Last day we described the operation of taking a random sample from an infinite population that had an equal proportion of the values {-1, -0.5, 0, 3}. This was in connection with our simulation of the portfolio of risky companies. It did not matter whether we were sampling with replacement or not, since the population is infinite, the items removed from the population at one time do not change the relative frequencies of future samples from the same population. (At times we used the equivalent description that the population only had four elements in it, not infinite, and we were sampling with replacement.) In random sampling from an infinite population (or sampling with replacement), the resulting values selected are "independent" in the sense that the last item sampled does not affect the selection of future values, that is, does not change the distribution for future selections.

One consequence of this is that, in a fair coin tossing scenario, if we observe 10 heads in a row, the chance of a tail on the next toss is still 50%. Or if, in random sampling from the population of registered SFU students, if we select 10 females in a row, the chance of a male in the next selection is still almost the same as it was for each of the first 10 selections ("almost" because the list of registered students is so large that for a small sample, it acts like an infinite list: removing the 10 females will not change the proportions much.) Random samples selected with replacement have the characteristic that adjacent selections are independent.

For a **time series**, adjacent values are usually **not independent (="dependent").**
Consider a couple of examples:
Ex1: the weather tomorrow is dependent on the weather today
Ex2: the fuel consumption series I showed you had waves up and down over the seasons. Clearly knowing one value at one time would help to predict the value at the next time. (dependence).
Ex3: a random walk has successive values that are dependent.
Ex4: the stock market index has successive values that are dependent.

This last example needs a bit more explanation. While the change in price from one day to the next is not predictable, the actual price itself tomorrow is partly predictable from today's price, because daily changes are usually small relative to longer run changes. Intel's stock price (p 361) on Aug 30 in 2002 was close to what it was on Aug 29, 2002, and much less than on April 30, 2002, and so the Aug 29 price did help to predict the Aug 30 price. But the Aug 29 price, or any prior prices, did not help to predict whether the Aug 30 price was more or less than the Aug 29 price. So the successive values of the time series are dependent, even though this does not help in the prediction of the direction of future 1-step price changes.

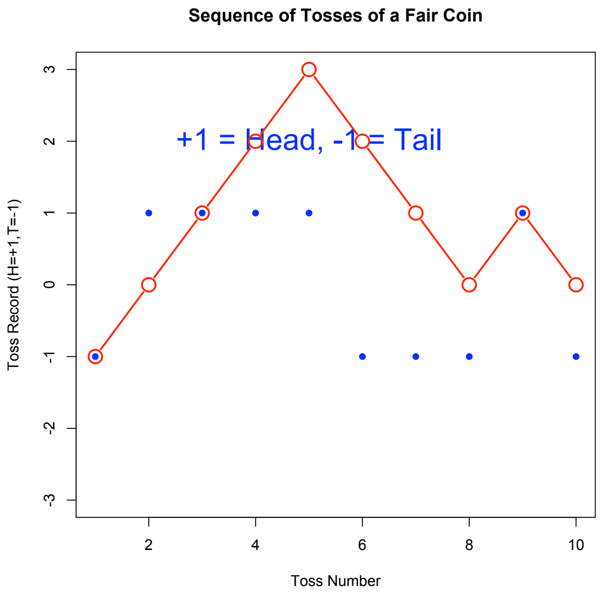
Ex5: Suppose we simulate a random sample of 25 N(0,1) values. Then we treat this as a time series and compute a moving average of order 5. Are the successive values of the original 25 values independent? (Yes). Are the successive values of the moving average independent? (No). Is the original series predictable? (No). Is the moving average predictable? (Yes). Is the 1-step direction of change of the moving average predictable? (No).



OK. So the above discussion should clarify what is meant by the typical dependence of time series data. Many time series have this "slow wave" nature, revealing the dependence of successive values.

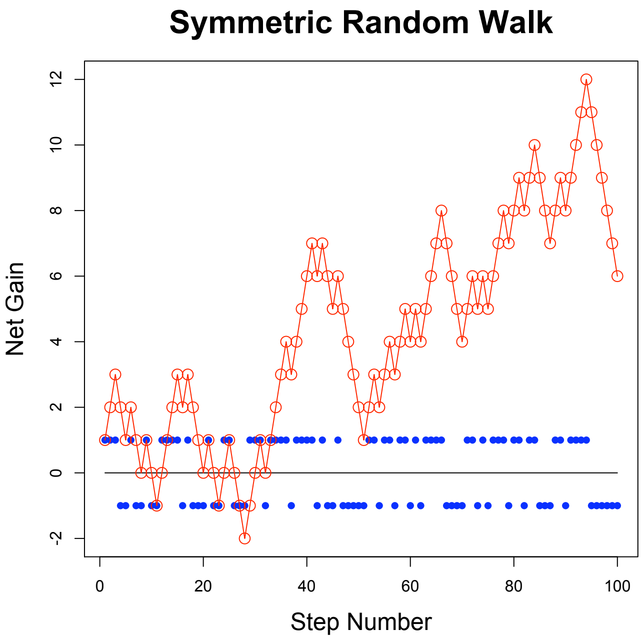
The "slow wave" nature just described is responsible for an illusion of predictability. For example, in the above graph there seems to be a downward trend, but we know from the way this was simulated that this is an illusion. The average of future values is 0 since all the green dots are from a N(0,1) distribution.

Another demonstration of this illusion was done with a random walk. Recall a random walk is a cumulative sequence of 0s and 1s in which the 0s represent "tail" outcomes of a fair coin and 1s represent "head" outcomes. See graph next page. (simple.walk())



Note that the red line is the random walk – the blue dots are just to show how the red line was simulated.

The interesting illusion of the random walk requires a longer simulation. (rw())



The apparent tendency in this particular simulation for the “net gain” to be increasing is an “illusion” in the sense that this tendency is not useful for prediction of the future trend. The chance of an upward trend in future values is the same as the chance for a downward trend.

Stock Market

One reason for looking at random walks is that stock market price indexes (and even stock prices themselves) have been shown to be well modeled by random walks (p 366). The symmetric random walk we study is not the best model for the stock market since

i) it has been the case for about 50 years that the stock market index increases about 10 percent per year (in Canada), so when we look at daily changes in the index, it apparently has a very small but upward trend.

ii) The steps of the Toronto Stock Exchange Index are not ±1, but rather they vary from about -3% to + 3% with changes less than 1% being typical. Since the index is currently at about 14,000, changes of ±100 are more usual that ±1.

Nevertheless, the nature of the time series of the TSE Index (now called the S&P/TSX Index) has a similar nature to the simple random walk that we simulate. (Moreover, it is quite easy to extend our analysis to be more realistic, but the extra complication does not produce any new phenomena of relevance to STAT 100).

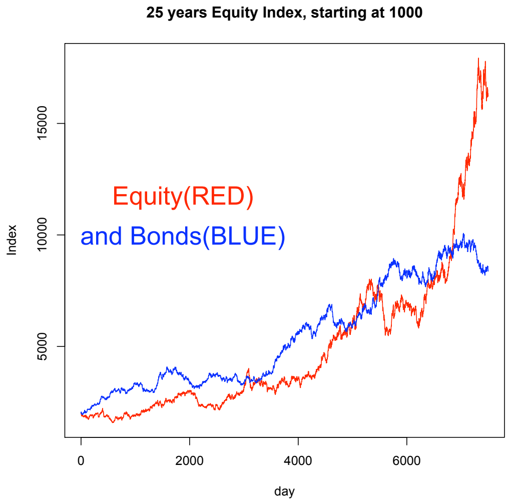
The basic result of our random walk simulation is what we have already described: apparent trends are useless for prediction of future trends.

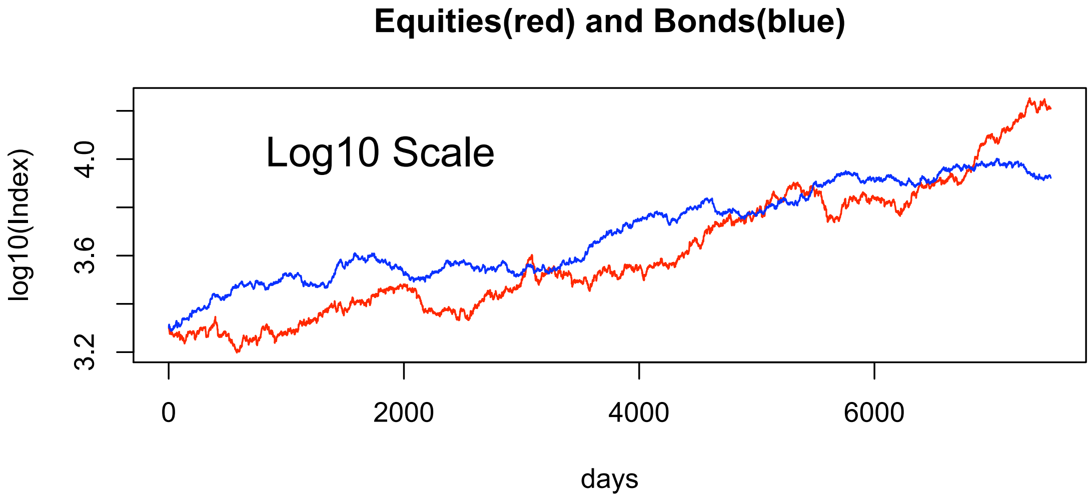
However, the article by Cleary and Sharpe (pp 359-372) provide valuable real world information. The most useful of this is the failure of investment experts to provide advantageous information to the investor (p 369). Our study of random walks shows why the experts are having such a hard time with this.

Another piece of real world information from the Cleary & Sharpe article is their study of forecasting methods, summarized on p 364. They show that the only methods that work at all are the very short term ones that say, tomorrow is pretty much like today! (See the MAPE column and discussion p 364).

Note that when I was discussing the stock market index above, I used percentages to describe the typical daily changes. As the index moves from 14000 (as it was before the recent meltdown) to 7000 (as it was after the meltdown) and back up to 12000 (as it is currently), a 1 percent change changes from 140 points to 70 points and back up to 120 points. Actual point changes day-to-day were smaller when the index was around 7000 than when the index is around 12000 or 14000. But the typical percentage changes stayed in roughly the same range of values over the entire period.

The graph on the next page shows the result of a simulation which is based on a realistic daily distribution of index percentage changes, and this distribution stays the same over the entire time period.





The second graph shows the same time series but with the index converted to a logarithmic scale. Because percentage changes have a distribution that applies to the entire period, there is not a "blow-up" for the more recent values, wheras the original graph did show this. It is much easier to assess the rate of increase from the logarithmic graph than from the raw data graph. These two graphs show what is typical of real data on stock prices, and in fact on many time series that show growth. See for example the graph in the Cleary and Sharpe article on p 368: larger levels of the stock price are associated with larger variability day-to-day.

[Unnecessary to know the following fact:
So what is a logarithm anyway? Logarithms are computed relative to a "base", so you have to know which base you are using before you compute a logarithm.
The base b logarithm of a number is the power to which b must be raised to equal the number. Often we use b=10. $\text{Log}_{10}(a) = c$ means that $10^c = a$.]

The thing to remember is that logarithmic changes preserve percentage changes.

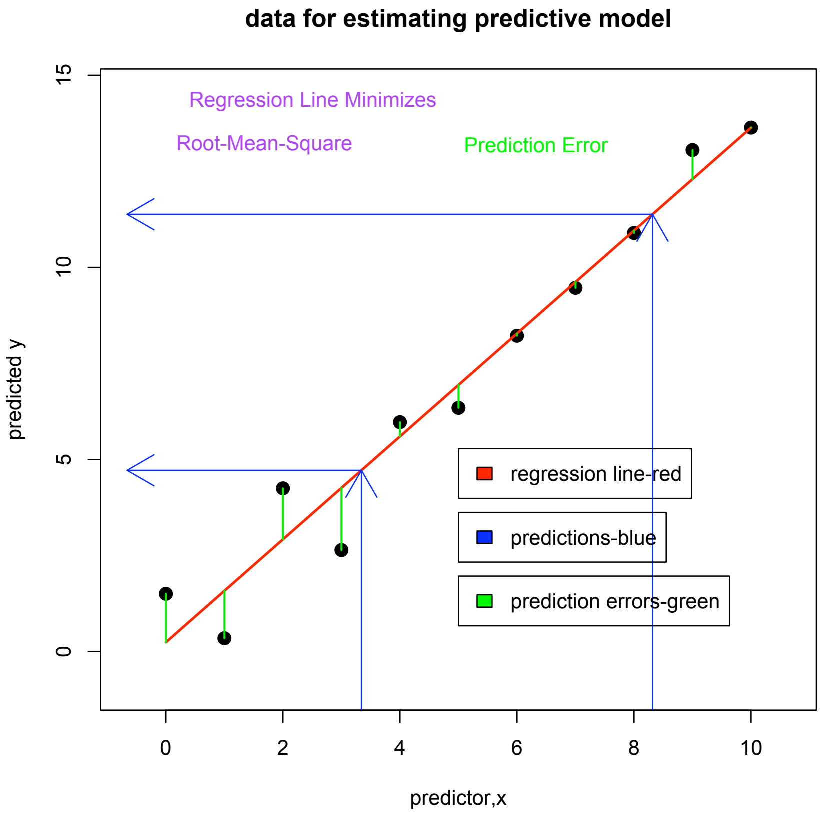
Population Growth

One other topic that could come under the heading of time series is population growth. This was discussed in the article on the Africanized Bee Invasion. A curve called a logistic curve was used to graph the typical pattern of population growth. See p 129. The curve is an "S" curve: increasing growth rate then decreasing growth rate. After a while a stable population can be reached. The reason for this is explained in another graph on p 130 (and was discussed in an assignment).

The next big topic that students requested review of is Regression & Correlation: Although we discussed these techniques separately, they are closely related.

Regression & Correlation:

First let me remind you of the way in which regression aids in the prediction of one variable from another:



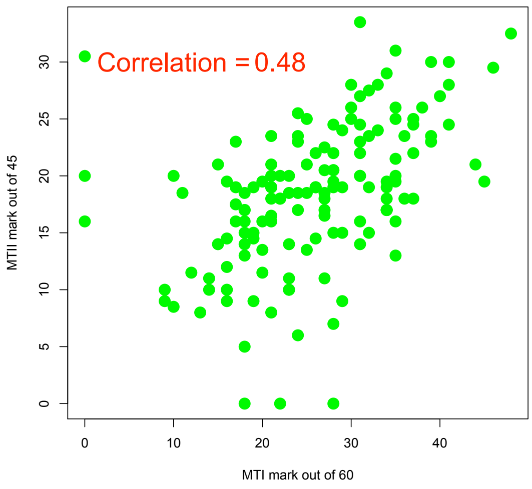
I assume most students understand this graph. I want to spend some time relating regression to correlation. Many data sets are of the following form:

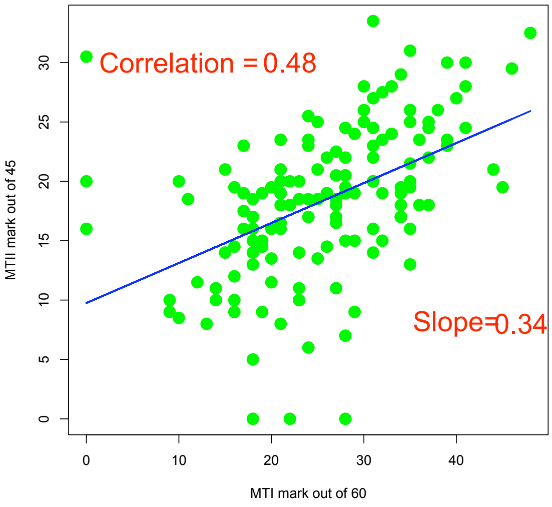
	Var 1	Var 2	Var 2
Case 1			
Case 2			
Case 3			
Case 4			
Case 5			

The cells would contain the data. For example, the "Cases" could be 5 students, and the "Variables" could be MTI, MTII, and Final Exam. In this example, it would not be surprising if the students with the higher marks on MTI also had the higher marks on MTII and on the Final Exam. Of course, there would not be a perfect fit. We say the correlation would not be perfect. The correlation between two variables is defined in the way I defined it in the April 6 lecture, but the idea is that it measures the extent to which the two variables are large and small together. The correlation coefficient lies between -1 and +1, and a correlation of say .9 between MTI and MTII would suggest that the marks on MTI were very closely related to the marks on MTII

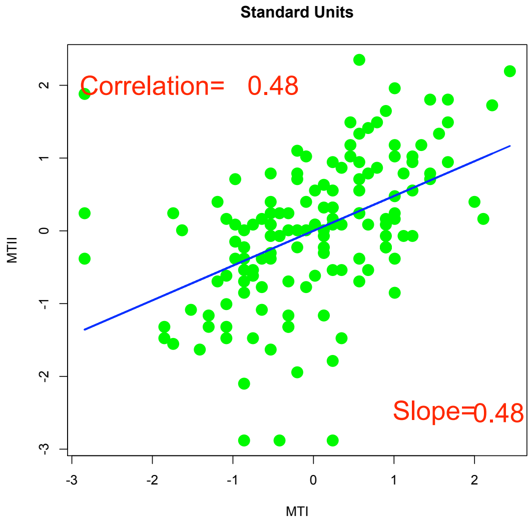
– the highest and lowest marks on MTI would likely be for the same people that the highest and lowest marks were obtained on MTII.

Actually, MTI and MTII marks are NOT highly correlated. The Correlation Coefficient between MTI and MTII in STAT 100 is 0.48. It is positive, suggesting that there is some consistency of performance, but it is not a very close correlation. But even with this modest correlation, there is some ability of the MTI score to predict the MTII score. It is the regression line that does this prediction. Traditionally we put the predictor (MTI) on the x-axis and the predicted variable on the y-axis.





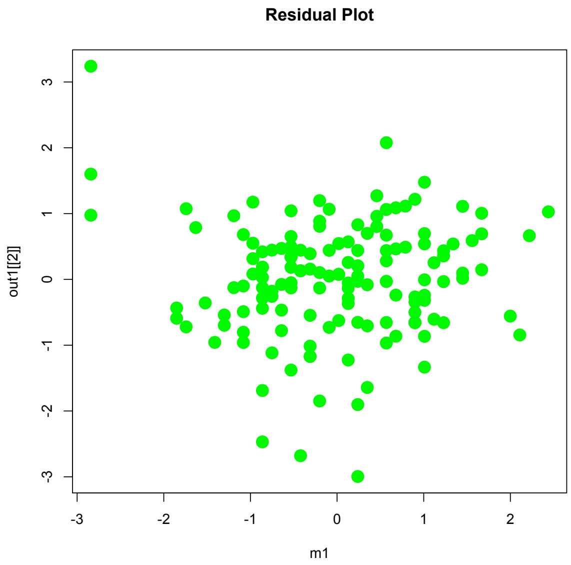
Note that the slope and the correlation are different. However, if we first convert the MT marks to standard units, we get:



The picture looks the same, but the scales are the new standard units.
When we multiply the coordinates, in standard units, and average them, we get 0.48.

When we do a regression, we should always check the residual plot: it is just the same points but with the height of the regression line at each x-value subtracted out from each y-value.

What we look for in a residual plot is some systematic pattern – if there is one, we could improve our predictive fit by adding this pattern. For example, it does look like there is some curvature(bowl-shape) in the residual plot, so maybe our prediction equation should have been a curve instead of a straight line.



Now we move on to a topic related to regression that arose in the articles on turkey mail (p 373ff) and on wine (p407).

Covariates:

Regression is for prediction of a y-variable from one or more x-variables. In turkey mail some of the predictors were not of interest even though they would affect the response (the y-variable): Age, Title, Registration Date. In the wine article, Age was not of interest since it was the affect of weather that was the real determinant of quality, but it certainly affects quality. These variables that are "nuisance" variables are called covariates. We always need to adjust our predictions for the covariates in order focus our analysis on the effects of interest.

Data Mining:

The term just refers to searching for information in a large data set. However, modern uses of the term usually refer to data sets that were collected for one purpose but are now used for a different purpose. For example, records of credit card transactions are collected for accounting purposes, but also provide information about the buying preferences of the credit card holders. Another example is web site search engines, which collect data on searches to optimize future searches, but also provide information about the searcher that can be exploited for commercial purposes. These huge data sets provide a challenge to the statistical analyst since the form of the data was not originally designed for the alternate use.

The example in the readings had to do with the fundraising activities of the Paralyzed Veterans for America(pp 307-322). In this case the data set was generated by the mailout of funding requests in the previous year. Each request cost money and the new use of the data in the current year was to try to reduce the mailing list to the portion that were most likely to contribute, thus saving money by omitting the cost of the ones removed. In the previous year, everyone on the original unrefined mailing list received a request. Even though a variety of demographic data had been collected for this original mailing list, these data had not been used. The large number of background variables and the variable quality of the data made the task of improving the mailing list a challenging task. The competition described in the article had different teams use different methods of analysis, but most used regression since prediction of response was the goal.

Weather Forecasting (pp 171-181)

In the lecture on weather forecasting, regression is used to base future forecasts on current atmospheric data. It is admitted that this does not work very well, since a small change in the current specifications can lead to a large change in the predicted weather (This is what Lorenz showed, p172). But the equations are still used as follows: the current atmospheric conditions are entered as a probability distribution rather than a set of numbers, and the output of the regression equation thus produces a distribution forecast, called by the meteorologists a “dynamical forecast”. This provides an imperfect but still useful forecast of the weather 6 hours hence and for the next few days. Be cause the data on current conditions is complicated to collect, assemble and analyze, forecasts for shorter than 6 hours are not feasible. More conventional regression methods work better in the very short term (p 178).

(KLW 2010/04/08)