STAT 100 Notes January 28, 2010

Today:

1. A primer on regression
2. “Predicting the Quality and Prices of Bordeaux Wines”

A primer on regression:

What is it? Given items with measurements X,Y recorded for each one, regression is a method for predicting Y from X. (This is the simplest form).

What is good about it? It uses a simple model, and if the model is correct it produces the best estimates of Y given X. Because the model is simple, it is usually fairly easy to see if the model is reasonably correct.

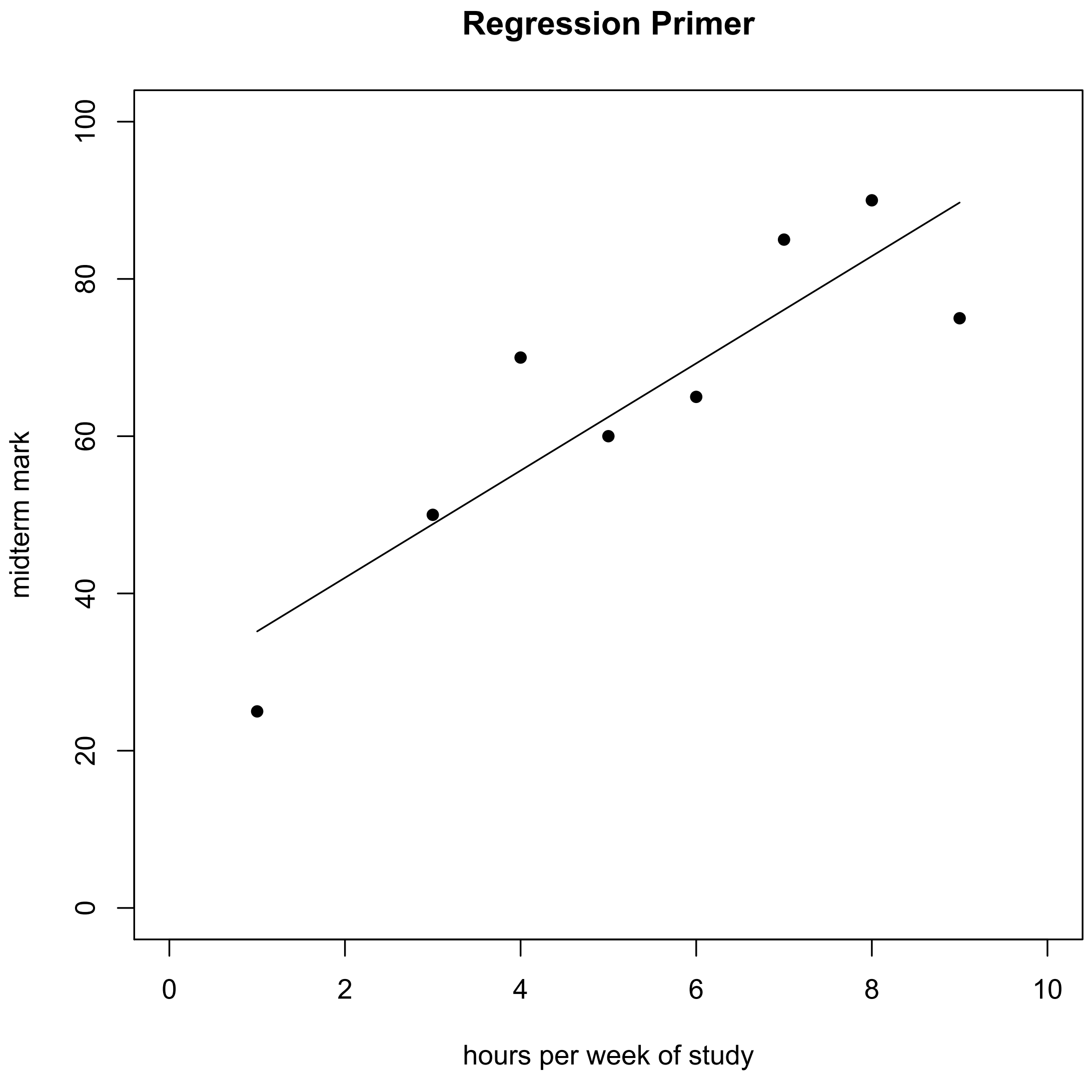
For example? You can predict the price of a Bordeaux wine 20 years after the vintage year by using temperature information in the vintage year. More later …

What is the model? The simplest version is this: Y = a + bX + e where e is normal error with mean 0 and of some unknown SD. That is, the pairs of measurements (X_i ,Y_i), for i = 1,2,3,…,n provides information about the values of a, b and the unknown SD. This is easier to see in a graph: suppose we want to predict midterm percentage mark based on the hours-of-study per week so far in STAT 100. We will collect data from 8 students to try to figure out the relationship (fictional data!).

Here is the fictional data:

hours	midterm%
1	25
3	50
4	70
5	60
6	65
7	85
8	90
9	75

Graphically we have



The line is the regression line. In this case it turned out that the intercept, a=28.4 and the slope, b= 6.8. We will soon see why this particular line was used to fit the data.

But suppose we have this line: what does it tell us about the prediction of Y from, X, i.e. the prediction of midterm mark given the hours of study per week?

If the line is the best fit to the data (among lines) then we have some simple inferences:

a) For each hour of study per week, the midterm mark increases by 6.8 percent
b) For any hours of study per week, x, we predict y as a + b x. For example, 7 hours of study per week is predicted to be associated with a midterm grade of 28.4 + 6.8 x 7 which is 76%.

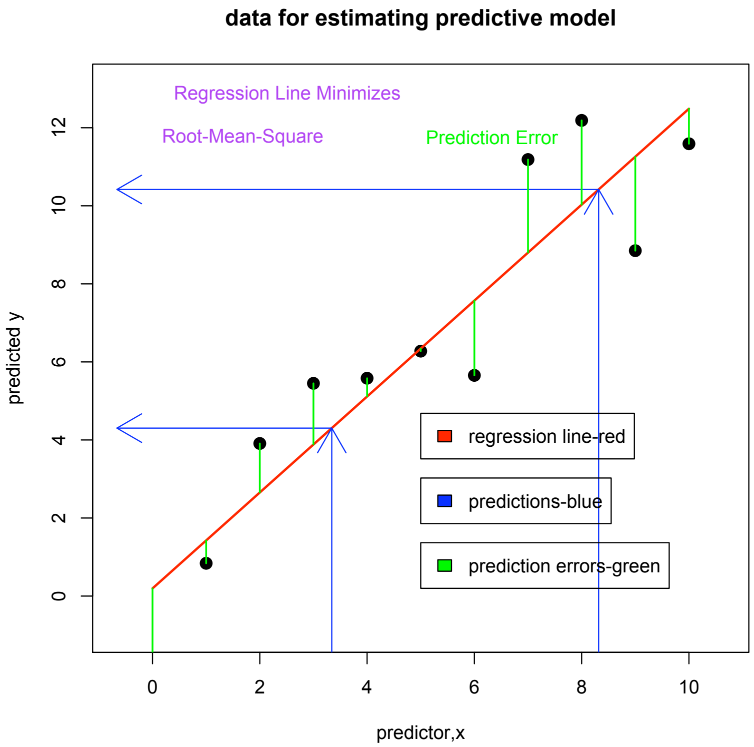
A few “buts”:

1.The relationship suggested in a) is not necessarily causal, because the data is not from an experiment (sure, its fictional, but we imagined it as selected at random from the class, and that would be observational data.) However, in this case the association probably would be causal, don’t you think?

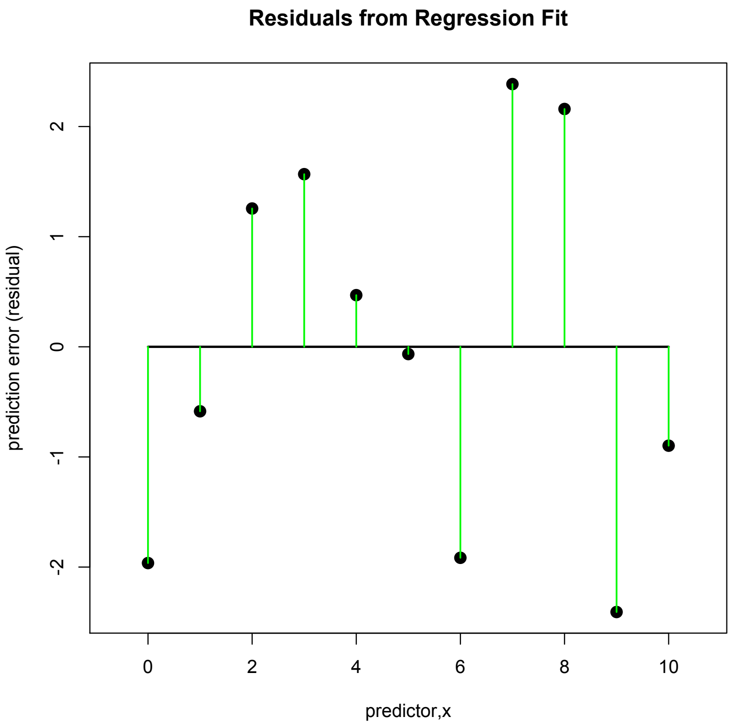
2. The apparent relationship in the graph could be due to random variation. Maybe if we did it for the whole class, the line would be flat, indicating no relationship between X and Y. (although this is unlikely in this case!). The weight of evidence for this apparent relationship to be real (not simply due to random selection) requires a more thorough analysis.

3. The prediction of Y for an X (of midterm mark for a certain number of hours of study per week) is not expected to be perfect for two reasons: i) the prediction line for the whole class is not exactly estimated from the random sample of 8 students, and ii) even if the prediction line were a perfect estimate of the whole-class line, the points are not exactly on the line, so using the line for prediction will involve error.

This last point is important – it turns out that the “best” line is the one that minimizes the prediction errors in some sense. Here is another demo to explore this aspect …



The blue lines show how the line is used for prediction of y from x. Note the small green "error" lines. They represent prediction errors, because they are the error experienced if the red line were used to predict the y-value at the associated x-value. If we focus on these prediction errors, we get the following graph (with the scale much expanded).



These residuals show the part of the prediction of Y that cannot be guessed based on X.

Note that the SD of the residuals could be estimated from this. Recall that the SD of the prediction errors (the SD of the "e" in the model) was one of the unknown *parameters* of the model. a and b are also parameters (the intercept and slope.)

There are lots of details to this. But the main thing is that you use regression to predict Y from X.

Note that we might want to predict Y from two predictors X1 and X2. That is not really much more complicated. The model is

Y = a + b1 X1 + b2 X2 +e

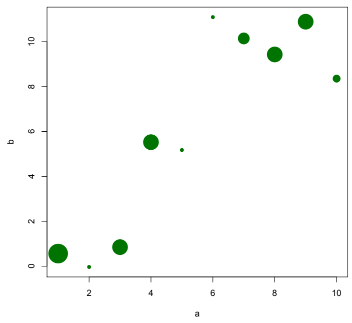
The estimates of a,b1,b2 and the SD of e can be found from formulas or software. This is the kind of thing that is used on p 416 of the "Predicting Quality and Prices of Wines" article.

Predicting Quality and Prices of Wines (pp 407-423)

Red wine from the most famous Bordeaux vineyards is bought by collectors as soon as it is produced - a few months after the harvest of the grapes. The quality of wine produced in the Bordeaux region of France varies greatly from on year to the next, largely dependent on the weather of the year before harvest. Also, this wine tends to improve with age. So there is a business in purchasing wine at the year of vintage and selling it many years later at a price much higher than the original cost. The 1990 prices for wines produced in the 1960s are shown in Table 1 on p 408. The fame of certain vineyards and the year of vintage seem to affect the price. But weather is an even greater influence. While this is well-known by those investing in new-vintage wines, the exact relationship of these influences to the price in later years is not well-known. This is where the regression method comes in handy.

The averages in Table 1 suggest the effect of vintage and chateaux on price. But to see the effect of age (years the wine is in the bottle before selling), we need a data set over a longer period – see Figure 1. Note that the logarithm. Fig 2. seems to straighten out the best-fitting regression line – this is the result of the price increasing by a certain constant percentage each year. This plot allows us to adjust the prices for age. What is left in price differences is the effect of vintage and chateaux, but not age.

The effect of weather can be studied further, since by the time the wine is produced, we know the detail of the weather that was present while the wines grapes were grown. A rough idea of the influence is given by Fig 3. Hot summers and dry harvest period are clearly associated with higher prices. The text also mentions the positive effect of wet winters before the growing season. (Q: how could this be added to the graph?) These three influences are built into the regression equation on p 416.



It seems that the Bordeaux experts were not aware of this regression equation! Look at Table 2 which shows relative prices of each vintage as a function of vintage and age. The prices for 1989 predicted by the regression equation (which is based on growing season weather only) is a pretty good predictor of the 1989 prices (determined by the wine-buyers in 1989). The experts' relative prices tend to decline as the years go by (down the column). The regression equation was a good forecaster of this decline, and would have saved the investor from paying too much.

Comment re: Stock Market Advisers, Wine Critics, and Art Critics

KLW 2010/01/28