Lecture 13

February 11, 2005

Last time we talked about the abalone data set, and I'm hoping with that context some of the mathematical ideas will now be more concrete.

First, lets tackle notation.

Our model is

$$\text{nrings} = \beta_0 + \beta_1 \text{length} + \ldots \beta_p \text{shell}$$

We can write this as vectors:

$$\begin{pmatrix} 15 \\ 7 \\ 9 \\ \vdots \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} + \beta_1 \begin{pmatrix} 91 \\ 70 \\ 106 \\ \vdots \end{pmatrix} + \ldots + \beta_p \begin{pmatrix} 30 \\ 14 \\ 42 \\ \vdots \end{pmatrix}$$

$$\begin{pmatrix} 15 \\ 7 \\ 9 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 91 & \ldots & 30 \\ 7 & 70 & \ldots & 14 \\ 9 & 106 & \ldots & 42 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \beta_p \end{pmatrix}$$

More generally, we let y_i be the response from the ith abalone. And x_{ij} is the ith abalone's value for predictor j. For example, x_{i1} represents length. So

$$\begin{pmatrix} y_1 \\ y_2 \\ 9 y_3 cr \vdots \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{18} \\ x_{21} & x_{22} & \ldots & x_{28} \\ x_{31} & x_{32} & \ldots & x_{38} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \beta_p \end{pmatrix}$$

or most succinctly

$$Y = X\beta$$

is the way we represent a hyper-plane in matrix notation.

Our conception is that the LHS contains data on a response variable, and our goal is to find the vector β that minimizes the mean-squared error:

$$(Y - X\beta)'(Y - X\beta)$$

First, why is this MSE? Note that if y is an n by 1 vector

$$y'y = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = y_1^2 + y_2^2 + \dots y_n^2 = \sum y_i^2$$

so you can see that this is just the least squares formulation.

The solution to these equations is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

And I leave it to the homework for you to convince yourself that these are the right equations. Your handout helps motivate this equation quite a bit, too. (Write down what $X'X$ and its inverse are.)

Often these are rewritten. First , let

$$H = (X'X)^{-1}X'$$

And there are nice properties:

1. These estimates are linear combinations of the y values.
2. the correlation between the predicted values and the observed values is the highest possible correlation you can achieve, limiting yourself to linear combinations of the y-values. This corelation is called the multiple correlation coefficient.

Analysis of Variance There is one part of the summary output that we have not yet talked about. I'm going to go back and remind you of some things we talked about when doing simple regression.

This equation:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

shows that the total sum of squares (the deviations of the observations about the mean) consists of two parts. The first is the part that can be attributed to the regression line – how much the regression line varies about the mean. The second is the errors– due to the fact that the points do not lie on the regression line. In fact, you can think of the LHS as the "null model" – what if there were no relationship at all, and we just fit the line $y = \bar{y}$.

We used this fact to find a way of evaluating the fit. This suggests that if the points lie close to the line, then the part of the variation due to the regression line should be a large fraction of the total variation:

$$\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

and this is equal to r^2.

Each sum of squares has something called the degrees of freedom – this tells us the number of independent pieces of information involving the n independent numbers $y_1, \ldots, y_n$ are needed to compute the sum of squares.

For example, the total sum of squares needs only n-1, because if you know the mean and you know the first n-1 y's, then you know the last (since $\bar{y} = \sum_{-}^{n-1} y_i + y_n$ implies that $y_n = \bar{y} - \sum_{i}^{n-1}$.

We can similarly figure out that the degrees of freedom for the ss due to regression is 1 (which is the number of parameters needed in addition to the intercept) and the SS for the errors needs n-2, which is the sample size minus the number of parameters estimated. This is often put into a table

First colum has the sources of variation in this order: SS due to regression, SS do to residual (or error) and then total . These corespond to The three terms above.

Next column has the degrees of freedom, which here are 1, n-2 and n-1. Finally, the third column has the mean squared error, which is the SS divided by the degrees of freedom.

In R, if you type "anova(lmfit)" where lm is a linear model object, you will get this table. For example, if we fit abalone rings on length, the SS due to regression (which is says are SS due to length) are 13454.5 and the mean is the same value because we divided by 1 degree of freedom.

The residuals sum of squares, the error sum of squares, is 299956.1, and we divide this by 4175 = n-2 to get 7.2. The square root of 7.2 is 2.68 which is the same as the residual standard error reported in the summary.

And r-squared must be $13454.5/(13454.5 + 29956.1) = .3099$.

Notice that the total SS is 43410.6.

Let's look what happens to the table when we add a term, say diameter. Notice that the SS due to length is the same, and the total ss is the same. But now the SS due to the residuals has been reduced and a SS due to diameter introduced. It is as if we broke off some of the SS from the errors.

This is one way of thinking about multiple regression. We start with lots of variation about the average. We then add a predictor variable that reduced the variation. Now we add another that reduces it even more.

But one thing we'll examine is that sometimes it doesn't reduce it very much. And if it doesn't reduce it very much, maybe that's because it's not needed – it just doesn't explain any of the variation.

We again assume that the deviations about the plane have a constant variance. These deviations

are $y - \hat{y}$, The residual sum of squares is the squared sum of these:

$$(y - \hat{y})'(y - \hat{y})$$

and we can get an estimate of their standard deviation by dividing by the degrees of freedom: $n - (number of parameters estimated) = n - p - 1$

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}$$

To make inferences we need assumptions. Here's a list although we don't always need all of the assumptions.

1. Y is a random variable with some probability distribution with a finite mean and variance.
2. The observations on y are statistically independent. (We've also stated this same assumption in terms of the errors, but they amount to the same thing.
3. The mean value of y is a **linear** function of the predictors $X_1, \ldots X_p$.
4. Conditioned on any fixed subset of the predictors, the variance of y is a constant value.
5. Conditioned on the prectors, Y is normally distributed or, put differently, the errors are normally distributed.

One consequence is that the estimators are unbiased, and also normally distributed.