Lecture 14

February 14, 2005

Last time, we talked about breaking up the sums of squares in a regression:

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

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 $y = \bar{y}$.

And together these give us an old friend:

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

Today we're going to talk about the ANOVA table, which is another type of output we can examine for regression that is not terribly useful for simple regression, but more useful for multiple. The columns are

1. Source
2. df
3. mean ss
4. F statistic
5. p-value

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.

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.

And one more "coincidence": take the square-root of the "F value" associated with length: the square root of 1875.2 is 43.30 which (coincidence!) is equal to the t-value given in the summary output for length.

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.

Also notice (from the summary) that the coefficient for length is different. In fact, it has changed sign!

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.

Now most software packages don't break the table up like this, they just give the sums of squares due to regression, which in this case we get by adding the sums of squares due to length + diameter. But here's what's important: the values of the sums of squares depends on which order you enter the variables! This means that rings = a + b length + c diameter will produce different regression sums of squares than rings = a + b diameter + c length. The reason is that the sums of squares are calculated like this

1. first fit a model with the first predictor, and calculate the sums of squares due to regression. This is the amount of variation explained by the first predictor.
2. then add the second predictor to the model and see how much additional variation is explained given that you already had the first one in the model
3. note that the sums of squares of residuals will be the same.

So when we put in length first, we get a sums of squares of 13454.5 for length and 1059 for diameter. This tells us a story. We started out with our variation as 13545.5 + 1059 + 28897.1 = 43501.6. Now if you tell me the the length of the shell, I can explaain 13454.5 of that variation. If you tell then give me more information – that is, assuming I know the length, I also get to know the diameter, then I now can explain an additional 1059 of the variation. (note that this isn't all that much.

If, on the other hand, you first give me diameter, then of the initial 43501.6 I can explain 14335.7. KNowing the length then explains an additional 177.9. Hardly any more.

Think of it as buying variation. Each parameter buys you a bit of the total variation, and you want to own as much as possible. Obviously in this case, you're better off buying diameter before you buy length, but we'll talk about that more later.

inference 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.

We still compute t-statistics the same way as before:

$$t = \frac{\text{estimator} - 0}{\text{SE}}$$

but the interpretation is slightly different. The null hypothesis is now "there is no relationship between length and rings *assuming that diameter is included in the model*." In other words, there is no "additional" information. And the same is true of all parameters. The interpreation of the parameters is that they are the change in mean value of y holding all other parameters constant.

Note that there is something called the "F-statistic" at the bottom of the summary. We'll talk about the exact form of this in a bit, but it is testing the null hypothesis: "all of the slopes are 0" against the alternative "at least one of the slopes is not 0." This is sometimes called the test for "overall regression", which doesn't seem to mean much, but essentially tests whether you're wasting your time with these variables (assuming linearity, normality, etc.)

The formula is

$$F = \frac{\text{SSReg}/(p-1)}{\text{SSE}/(n-p-1)} = \frac{\text{MSReg}}{\text{MSE}}$$

If the sums of squares due to errors is big (remember that the sums of squares due to regression is equal to the total sums of squares minus the sums of square due to error) then this term is 0. So if the term is big, that must mean that our errors are small and we explained some variance. And then F will be big.

This is a way of thinking we'll use again and again.

Next time: why do slopes change depending on what other variables are present?