# Lecture 15

February 16, 2005

Last time we went over the assumptions needed to do statistical inference with the regression model.

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.

Which, put more succinctly looks like $Y = X\beta + E$ where $E$ is a random vector consisting of the random numbers $\epsilon_i$, and $\epsilon_i \sim N(0, \sigma^2)$ and they are independent of each other. With this in place, we can compute t-statistics for the slopes as :

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

and they look nearly identical to the form as for simple linear regression. There is a new test, however, we should talk about, called the "test for overall regression."

The null hypothesis is "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.

The F-test appears in teh ANOVA table printout and also at the very bottom of the summary of your regression. (Use the "summary" command.")

There's another type of test we'll find useful. Suppose this is our model:

$$y_i = \beta_0 + \beta_1 x_i 1 + \beta_2 x_i 2 + \ldots + \beta_p x_i p + \epsilon_i$$

and we want to test the hypothesis that we don't need $\beta_1$ AND $\beta_2$ AND $\beta_3$.

Null hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$. Put differently:

Does adding these three variables significantly contribute towards predicting $Y$ given that the other terms are already in the model?

So what we probably want to look at is the sums of squares due to regression from the big model divided by the sums from the small model. The big model is usually called the "full" model and the other the "reduced".

In fact what we look at is how the sums of squares due to regression has changed as a percent of the total sums of squares, with a slight modification:

$$\frac{\text{SSRegression(full)} - \text{SSRegression(reduced)}/k}{\text{SSE(full)}/\text{df}}$$

where $df = n-$ number of parameters estimated in full model.

The sampling distribution of this statistic is, if the assumptions hold, called and F distribution with k and (n- number of parameters) degrees of freedom. The p-value is the probability that a randomly selected F will be bigger than the observed.

Some examples will make this clear.

These are data collected from several sites in california regarding mussels.

Warning: beware of missing values. Then things might go differently. Next: added variable plots, diagnostics (hat matrix, cook's distance, residuals)