Lecture 9

January 31, 2005

Inference

With these assumptions we get a whole bunch of theory helping us out. We've already shown that the intercept and slope are linear combinations of the y's. And linear combinations of normal random variables are also normal. And the central limit theorem applies to them, so even if the residuals are not perfectly normal, there is still hope for the estimators.

It also turns out that the estimators of slope and intercept are unbiased. In mathematical notation

$$E(\hat{\beta}_i) = \beta_i$$

To prove this you don't need to assume normality, by the way.

What we don't know is how much varying these estimators do. We need to know their standard error.

One can show that

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\mathrm{SXX}}$$

$$\mathrm{Var}(\hat{\beta}_2) = \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}})$$

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\sigma^2 \frac{\bar{x}}{\mathrm{SXX}}$$

the standard errors are the square roots of these.

And our approach, as always, is to replace the unknown with unbiased estimates. So we need to estimate σ^2. What is a reasonable way of doing this? Well, since this term measures the variation about the line, we need to know how far each point is from the line. This information is what the residuals measure. In fact, since $E(y|x) = \hat{y}|x$, you can see that $\frac{\mathrm{RSS}}{(n-2)}$ is a reasonable estimator of this variance. (the n-2 is needed to make the estimator unbiased.) This is called the residual standard error. Substitute this into the equations above, take the square root, and you have the standard errors.

The square-roots of these values are typically provided as output. In R, it prints a table for each coefficient that includes the estimated value of the coefficient and the standard error. R also gives the residual standard error.

Confidence Intervals for coefficients

are found as always: estimator $\pm$ t times SE. It turns out that the estimators, when using estimated versions of the standard errors, follow a t distribution with n-2 degrees of freedom. So t is chosen to be the upper $\alpha/2$ quartile of a t-distribution with n-2 degrees of freedom.

For example, suppose we want a 95% CI of the slope for our tree data. The model we fit (using squared diameter) was:

$$V = -87.12 + 1.54H$$

and the standard error of the intercept is 29.27and of the slope is 0.3839.

So to find the constant multiplier, we find the value from a t distribution with $n - 2$ degrees of freedom that has 2.5% above it. Here, $n = 31$. So the command qt(1-.025, 29) will do the trick. The value is 2.0452. So a CI would be

$$1.54 \pm 2.0452 * .3839$$

or (0.75, 2.32).

Note that this interval does not include 0.

In practice, you can think of CIs as providing a range of plausible values for the population parameter. So, from our data, it is plausible to assume that the true slope is anywhere between 0.75 and 2.32.

Remember the "confidence" in confidence interval is all about the process, and not the result. In repeated collections of data, 95% of our CIs will contain the population value. So there is a 95% chance that our method will be successful, but there is NOT a probability associated with whether or not the population value lies between0.75 and 2.32.

Hypothesis tests

As always, we focus on the slope, since the intercept is often not terribly interesting. (And that is certainly true with the tree data.)

The primary hypothesis is that the population value is 0. This would mean that there was no relationship between x and y. We can form a t-statistic as follows:

$$\frac{\text{estimate} - 0}{\text{se}}$$

This follows a t-distribution with n-2 degrees of freedom because the estimate is normally distributed (or approximately so).

Most software packages will produce the observed t-statistic and its p-value: $\mathrm{P}(T > |t|) + P(T < -|t|)$.

For tree data, $t = 4.021$ and $\mathrm{P}(T > 4.021) + P(T < -4.021) = 0.000378$.

Properties of LS Estimates or "Why Least Squares"? Turns out that thsi method of finding estimators produces some nice properties. (Strictly speaking, we did something called ordinary least squares. more on that later.) Al of these properties assume the relation between mean and x is truly linear and variance is constant.

1. estimates are consistent; roughly, as sample size increases, difference between estimates and population values decrease.
2. invariance to scale. The estimates change in a predictable way if units are changed (say from metric to english)
3. Gauss-markov theorem: these estimates have the smallest variance among all linear, unbiased estimates (linear in y).

We'll leave "ordinary" LS behind when we have non-constant variance.

What about normality?

If residuals are not normal, the CI's and t-tests are still pretty accurate, as long as sample size is large. How to say how large is large, but there are other techniques one can rely on (such as bootstrapping) if this is a concern.

CI for the conditional mean

A common use of regression is to estimate the mean response for a given value of x. For example, for trees with height 80, what's the mean volume? This question is answered by plugging in 80 to the regression equation. But we want to get confidence intervals for this.

What we need to know, then, is how much $\hat{y}$ varies about it's mean. We need the standard error of $\hat{\beta}_0 + \hat{\beta}_1 x$ for some value of x.

$$Var(\hat{\beta}_0 + \hat{\beta}_1 x) = Var(\hat{\beta}_0) + x^2 Var(\hat{\beta}_1) + 2x * Cov(\hat{\beta}_0, \hat{\beta}_1).$$

And we've found each of these up above if we just substitute our estimate of σ^2.

This works out to be

$$\hat{y}|x \pm t(n-2, \alpha/2) * \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

R does this with the predict.lm command.

Note that the confidence interval gets wider as x moves away from $\bar{x}$.

Prediction Intervals

Suppose you want to answer the quesiton: I've found a tree with height 80. What is the volume of this tree? Note that we're no longer asking about the mean of all trees of height 80, just the one.

The answer is the same: plug 80 into the regression equation. Only the confidence intervals change, to reflect the additional uncertainty of predicting a single point.

$$\hat{y}|x \pm t(n-2, \alpha/2) * \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

R does this through the predict.lm command. See the R notes.