Lecture 10

February 2, 2005

Load the tree.txt data into R. We're going to talk about confidence intervals and hypothesis tests. These commands are covered in R Notes to Lecture 8 and the theory is covered in the "summary" handout from Monday.

Fit a model to predict tree Volume based on Height. What's the model?

All CIs are of the form estimate ± constant times SE.

CI for parameters For slope:

$$\hat{\beta}_1 \pm t_{df}(\alpha/2)\hat{\sigma}\sqrt{1/\text{SXX}}$$

For intercept

$$\hat{\beta}_0 \pm t_{df}(\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}}$$

where $\hat{\sigma} = \sqrt{\text{RSS}/n-2}$.

R doesn't output these explicitly, but they are easily calculated, since it gives the standard errors in the output summary.

Type summary(tree.lm). What are the standard errors?

So you can now make these, once you know how to find the constant.

To find the constant, you need to find a value that has area $\alpha/2$ **above** it in a t-distribution with degrees of freedom equal to n-2. . The command qt(p, df) will give you the number in a t-distribution that has area p below it. So we want to type either qt(1-p, df) or -qt(p,df). Here, for p we substitute $\alpha/2$, and we determine *alpha* based on our confidence level: Conf level $= 1-\alpha$. For a 95% conf. level, $\alpha = .05$.

Hypothesis tests for parameters These are computed easily by R. The t statistic it reports is simply

est/SE

and is used to test the null hypothesis that the true value is 0 against the alternative that it is not. This statistic follows a t-distribution with n-2 degrees of freedom if the null hypothesis is true. The p-value is $P(T > t) + P(T < -t)$ where t is the reported test statistic, and T follows a t-distribution wiht n-2 degrees of freedom. This is calculating how unusual our observed value was. Small p-value means very unusual – which makes us doubt the null hypothesis.

CI for means and predictions 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.