Lecture 8

January 28, 2005

Influential Points

Regressions can be affected by exceptional points. It is possible to determine whether an outlier is influential simply by removing it and refitting the model and seeing for yourself. But there are other ways. One useful method is Cook's distance. We'll be able to explore this more when we study multiple regression, but for now, we'll have a simple version.

Let's define $\hat{y}_j$ to be the predicted value at x_j using all of the data. Now suppose we remove one of the observations, say the ith one. We can recompute the line, and figure out what the predicted value is at x_j. We'll call this $\hat{y}_{(i),j}$. Then for a given point that we want to remove, we can compute the distance between where we'd predict the line to be with it and without it for our data-set. Cook's distance is a weighted version of this:

$$D_i = \frac{1}{(RSS/(n-2)} \sum_{i}^{n} (\hat{y}_{(i),j} - \hat{y}_j)^2$$

We can compute this for every point in our data set. If D_i is big for a given point, this means it has a big influence on the line.

R does this fairly easily. Type plot(lmobject) and it gives you a series of 4 diagnostic plots. The first we've seen already (almost): a plot of the residuals against the fitted values (which will have the same shape as the residuals against the x), the next two ignore, and the last is a plot of Cook's distances.

Let's try this with our tree data.

Notice the 31st point is influential. Let's take it out and see what we get.

Note there is still influential points. There will always be. The purpose of this is not to remove them, but to be aware that not all points are equal – some are more equal than others. In this case it looks like the bigger diameter trees have the greatest influence on the regression, which might signal a slight difference in the type of tree being measured.

But even this pattern is not unusual. You'll see quite often that the influential points are off to the extrems (the smallest or biggest).

How big is too big? Well, the cook's distances should all add up to 1 (with simple regression), and that gives you some gauge about whether an influential point is "too big".

Inference

We turn now to the problem of trying to make inferences about a population. The context here is that we think of the data as a subset of a larger population, and we try to understand how our estimates mgiht vary so that we can understand how close they are to the population values.

And there are many estimates we're concerned about now:

1. the intercept
2. the slope
3. the mean for any given x
4. a y for any given x
5. standard deviation of "errors"
6. correlation

Also, in many science and policy contexts, you want to know about how the response would change if you were to change x.

Another approach and paradigm is discussed in the homework. Sometimes you don't care so much about the coefficients of the model. You care only about making predictions. And so you only want to know if your regression model is useful or not for predicting y values. In this context, a good question to evaluate your model might be "how good a job will I do on a new data set?" And this question can be addressed by many techniques, one of which is called cross-validation. In the HW you will do a very simple version of cross-validation that illustrates the method.

We've made a big deal so far about the mean function ; this is the function that describes how the mean y values varies for different x values. In a linear model, $E(Y|x) = \beta_0 + \beta_1 x$. But we need to also describe how individual observations are determined.

Conceptually, our model is this: response is *mean plus random "error"*. The mean is given by the mean function. The random error is a normally distributed random number with mean 0 and some standard deviation that we'll call σ. Hence, for the ith observation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma)$ and the ϵ are indpt of each other.

Some interesting features to note.

1. The X's are known values and are known perfectly (there is no measurement error).
2. σ measures how much variation there is about the line. Because the points are normally distributed about the line, we should expect that about 68% of the points are within 1 σ of the line, 95% within 2, etc.
3. The standard deviation σ is the same for all values of x.
4. the errors are indpt of each other, which means that knowledge about how far one observation is from the line should not give us any information about how far any other is.
5. the y's are also normal random variables.

It is a challenge to a scientist who plans to use this model to design an experiment that satisfies these conditions. It is a challenge to the data analyst to understand what this model says about how we must think about the data, and whether this leads to meaningful interpretations.

For example, with the tree data, do you really believe that we measured the circumference of these trees perfectly? If someone else were to measure them again, would they get exactly the same thing? Probalby not, and this, as we'll see, is a serious problem which we'll discuss later.

So we have to think of these data now in this context: we have a few trees of a particular circumference. Each tree "wants" to be a particular volume, but for various reasons, some trees are a bit bigger and some are a big smaller, and so they vary about the central value. And this variation does NOT depend on the circumference or size of the tree. So small trees vary about their mean the same way that tall trees do.

More Diagnostics

Before interpreting a model, then, we have to check these assumptions. So added to the assumption of linearity, which we've already discussed, we now have to check for independence and normality. The best way to do check normality is to make a qqnorm plot of the residuals. You can also make a histogram for a second opinion.

Checking independence is harder, and sometimes impossible. Sometimes this requires interviewing the person who caught the data to make sure that measurements were independent. Sometimes you get lucky and can see a violation of this. For example, if you know the order the data were collected, plot the residuals against this order. There should be no pattern. If you see all of the positives on one end and the negatives on the other, for instance. Or you see perfect oscilations. Any of these things might be signs of a failure of independence.

Another important thing to look for now is that the variance is constant for all values of x. The same residual plot of residuals against x will help: there should NOT be a "fan" shape. the residuals should have roughly the same spread no matter which slice you look at.

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 ± 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.