Breaking down the pipeline

Ben Klemens

30 March 2009

Let me get back to entry #002, where I pointed out the value of distinguishing between inferential and descriptive techniques.

I believe my first few tries at understanding statistics failed because I took classes that didn't make this distinction. Consider good ol' ordinary least squares (OLS), which is often all the stats an undergrad will learn. Here are the steps:

- Clean your data, producing an input matrix X and a dependent vector Y. This is via various computer-code matrix manipulations and substitutions for missing data.
- Find the line of best fit, with coefficients $\beta = (X'X)^{-1}X'Y$, using pure linear algebra.
- Test the elements of β, using methods most folks recognize as statistical, regarding comparing a statistic against a distribution.

The point here is that each of these steps is a different world from the others: computer trickery, linear algebra, and t-distributions basically have nothing in common. Like most undergrad courses, I'll pass on the first step, and assume a perfect data set. Then we're back to the distinction from prior episodes: the second step is purely descriptive, and the third step is purely inferential.

As usual, there is little or no benefit to confounding the descriptive and inferential step. They evolved separately; talking distributions provides nothing for the understanding of linear algebra; talking linear projections does nothing to forward an understanding of the Normal, t, or F distribution.

But you'd think that linear projections and t-tests are joined at the hip from the many stats classes and textbooks that present the steps above as an unbreakable pipeline.

Fun fact: errors do not need to be Normally distributed for the OLS projection to be the line of best fit (i.e., to minimize squared error). I've met a number of extremely intelligent people who thought otherwise. This even appeared in a draft textbook I was peer-reviewing last week.

This fun fact is from the Gauss-Markov theorem, which is a linear algebra and minimization exercise that has no need for math regarding distributions. But at this point, you can see where that confusion came from: when these people learned OLS, they simultaneously learned the part about projection (the Gauss-Markov theorem) and the part about hypothesis testing (based on Normally- or t-distributed errors).

I also think that there is a truly prevalent perception that that the *purpose* of OLS is the hypothesis test at the end of the pipeline, that that's all that OLS does: it tests whether one column of the X matrix affects or does not affect a column of the Y matrix. The other numbers that the software spits out—because the software also merges the descriptive and inferential steps—are just irrelevant.

The OLS pipeline goes from inputs to projection coefficients, pauses for air, then goes from projection coefficients and variances to confidence levels. Those who believe that there's no middle step, and it goes straight from inputs to confidence levels, are prone to miss out on a number of points:

- Missing the inappropriateness of OLS when Y can't be expressed as a linear combination of the elements of X. OLS is appropriate if Y is a linear function of X_1^2, but a student looking for the final confidence interval may not have an eye out for squaring or other such such transformations.
- Failing to realize that OLS is one of an infinite number of alternate models that one could use to test a hypothesis, where those tests also conclude with a t-test or F-test.
- Ignoring practical significance, such as when a coefficient is statistically significant but implies a minimal change in outputs given a reasonable change in inputs.

I could think of a few more, mostly bad habits about grubbing for p-values. Such bad habits are from a failure to balance the descriptive and the inferential.

Policy implications At this point, my recommendations should be obvious: when teaching a standard pipeline like OLS, be clear as to which steps are inferential steps, and which are descriptive. Teach them separately, as a set of pipe joints each of which has value by itself, and at the end mention that what you'd taught to that point can be welded together to form a smooth pipeline.

There are a few common ways by which parts get merged and OLS sold as an inferential technique only. I have several examples of their use on my bookshelf, and recommend that they be avoided.

The most common method of confounding is to open the section on OLS with a list of assumptions required for both description and inference. This has minor benefits over introducing each assumption as it is needed—it makes it easy to memorize the list for the test—but has the major disadvantage of not giving context as to why these assumptions to be memorized are necessary. I'm pretty sure that it's these all-assumptions-first textbooks that make it so easy for me to find people who think Normally-distributed errors somehow fit into the Gauss-Markov theorem.

Some textbooks on my bookshelf literally class regression in the inferential statistics part of the book, and heavily focus on the interpretation of those darn p-values. See above about encouraging significance-grubbing.

As you can imagine, my own writing takes pains to make the distinction, and *Modeling with Data* covers most methods twice: once purely descriptively and with no

inference, once a chapter or two later covering only inference. I don't think any continuity is lost because students have to flip between p 274 and p 307.

I used OLS as an example here, but one could apply it to may parts of what's covered in a probability and statistics class. For example, all the distributions are almost always covered in one long list, even though some are for description of natural data and others are for inference using constructed statistics. As Kmenta [1986] explains, "There are no noted parent populations whose distributions could be described by the chi-squared distribution." [I like Kmenta's textbook because it does a wonderful job of telling the reader exactly whether he's talking about description or inference at any given point. Too bad it's out of print.] From a purist's point of view, all distributions are just functions, but from the student's point of view, the crucial question is what s/he will do with each distribution. If their use is different, then that needs to be made clear, meaning that methods for inference and those for description need to be clearly distinguished.

References

Jan Kmenta. *Elements of Econometrics*. Macmillan Publishing Company, 2nd edition, 1986.