Regression: Why?

February 5, 2018

Three justifications for estimating a linear regression

- Best least squares fit in population: $\min E[(Y - X\beta)^2]$.
- Mean of $Y \mid X$ is linear, we want to find it: $E[Y \mid X] = X\beta$.
- $Y \mid X$ has a known (or parametrically estimable) distribution with $X\beta$ as a location parameter.

Population LS fit

- Useful for predicting Y from new draws from same population of X's.
- Does not require any assumptions on the distribution, other than finite variances.
- In particular, not necessarily true that $E[Y \mid X] = X\beta$.
- OLS achieves the semi-parametric efficiency bound.

Semi-parametric efficiency bound?

- But don't we know that GLS is more efficient than OLS?
- SEB: Any other estimator that gives improved performance under some distributional assumption, must give worse performance under some other distribution that satisfies the same minimal identifying moment condition. ($E[Y'X]) = X'X\beta$).
- Since we don't try to estimate a nonlinear $E[Y \mid X]$ or a fancy distribution for the residuals $Y - X\beta$, we don't have to rely on a lot of complicated assumptions and keep the number of parameters to consider small.
- This makes it plausible that asymptotic distribution theory applies well in a moderate sized sample. See Angrist/Pischke.

So should we, e.g., try to model the distribution of errors in a linear regression?

- SEB seems to say this is pointless.
- But we know that, e.g., correctly modeling a t distribution for the errors delivers efficiency gains.
- How can it not give us an advantage if we model the errors as t at the start, test against alternatives in a sieve scheme or a Bayesian ∞-dimensional prior that in some sense covers the whole space?

Maybe we should, if we have a good prior.

- There are two possibilities: One is that our sieve scheme does not in a relevant sense cover the whole space, so that there are distributions of the residual for which our scheme does worse than OLS (or GLS).

- The other is that it is general enough, in which case our prior probability on more complex models always keeps us far enough away from certainty about the actual error distribution that we get no asymptotic efficiency gain.

Maybe we should, if we have a good prior.

- There are two possibilities: One is that our sieve scheme does not in a relevant sense cover the whole space, so that there are distributions of the residual for which our scheme does worse than OLS (or GLS).

- The other is that it is general enough, in which case our prior probability on more complex models always keeps us far enough away from certainty about the actual error distribution that we get no asymptotic efficiency gain.

- But this is only asymptotics! If we have a good idea that with high probability the distribution of errors is t, we will do better than OLS in modest sized samples. Only in very large samples, where estimation accuracy has in any case become very high, do we start doing only as well as OLS.