210A Week 9 Notes

Sampling on the Dependent Variable

Technically, sampling on the dependent variable is when you select cases on the basis of meeting a criteria and then use those cases as evidence for the criteria. Since we're usually more interested in associations than distributions we can broaden this problem to something like "sampling on theory affirmation." This practice is at the center of Karl Popper's positivist approach of "falsifiable hypotheses." Popper complained that Freudians and Marxists just cataloged evidence that supported their theory whereas he argued that real science consists of searching for evidence *against* your theory and failing. He gave the example of the hypothesis "all swans are white" and said the way to test this hypothesis is not to accumulate a vast catalog of white swans, but to search for a black swan. On so doing you would find that there are in fact black swans (in Australia) and so the hypothesis is false. (This is more confusing than it used to be as Nassim Taleb's recent bestseller *The Black Swan* confuses Popper's meaning with the the failure in statistics or finance to model rare events such as catastrophic market crashes). It gets a little more complicated when you deal with the sorts of probabalistic hypotheses sociologists typically use (i.e., we like to say "most" not "all") so in our context good practice is to collect evidence at random, or at least at random with regard to the distribution or association that is of analytical interest.

(Note that because of the way that regression works, "sampling on the independent variable" isn't that big of a deal, at least if you assume that interaction effects are weak or non-existent. In fact we often deliberately use "over-sampling" to ensure adequate sample size to draw inferences about some numerically small but theoretically interesting subpopulation.)

In principle, pretty much everyone agrees with the general principle that we ought to avoid sampling on the dependent variable however this is easier said than done because you can run into problems not just by sampling on the dependent variable but on variables that are highly correlated with the dependent variable. This problem is also known as "sample selection bias." For instance early political polling occasionally misforecast elections because pollsters like to contact people by telephone whereas telephones only saturated the population in the 1970s. Since people who owned phones tended to be wealthier, and wealthier people were more likely to vote Republican, there were several infamous polls that exaggerated the Republican vote. Note that this is a much more subtle error than getting contact numbers from a list of registered Republicans, but the effect is similar.

One common method prone to the sampling on the dependent variable problem is the "strategic site" method, which is especially (but not exclusively) a problem for qualitative work and often is a euphemism for "I got hired to do consulting and figured I could get an article out of it." This comes up when people choose field-sites that are interesting rather than typical or choose field-sites where what is normally a background condition is made salient. For instance, someone interested in race relations might study a town/workplace where race relations are particularly good/bad or might attend some kind of diversity training workshop. There's nothing wrong with choosing such field-sites (or sampling individual respondents from them) but you have to be very careful how you analyze them. There's a big difference between "how race is enacted" and "how race is enacted in an artificially salient context" or "how race is enacted in a particular setting that I found interesting because of the way that race is enacted there." So long as you don't confuse the former with the latter two you're not in trouble, but it's easy to fall into.

Censorship

Very often we have missing data in that cases that should appear in our sample frame can not be observed (or at least they can't be observed as to some variable). Sometimes this is a measurement issue, as when people refuse to answer a survey but in other cases it is a basic fact, as when people are dead and therefore couldn't answer the survey if they wanted to. Likewise often you can measure some things about cases but not others, either because subjects refuse to answer particularly sensitive questions or because the question is not applicable to them.

Censorship is not a problem if you think the cases you can't measure are similar (or would be similar) to the cases you can measure, but if you think that the censorship is itself correlated to the trait then you have a problem that is essentially a weaker form of sampling on the dependent variable. For example, imagine you were measuring the health effects of smoking among the elderly and you found that old smokers weren't in that much worse shape than elderly nonsmokers. The problem is that it's impossible to measure the health of the dead, although it's probably also true that if the dead had managed to survive they would probably be less healthy than people who did in fact live. Since smokers often die in their 50s and 60s you would thus be underestimating the health effects of smoking.

Likewise, imagine that you are trying to measure whether there is wage discrimination against women. You measure it and you find that within occupation and among childless people there are basically no wage differences by gender. Aha, so there is no gender gap *per se*, only a "mommy gap." But there are a few problems with this. First, controlling for occupation is arguably overcontrolling (see below). Second, women have low labor force participation and among women this is probably correlated with earning power. If one of the reasons women choose to stay home is because they would make so little money if they did work (or alternately, are making so much money that they feel they can't afford to stay home) then we face a censored distribution of womens' wages since we might expect that if housewives did get jobs they wouldn't be as good of jobs as those held by those women who do in fact work in the formal economy.

There are some advanced regression techniques for dealing with this like Tobit, propensity score matching, regression discontinuity, and instrumental variables. These techniques are very hard to do well but at least they take the issue seriously. You should especially be on the lookout for studies suffering from censorship bias that don't even try to model the censorship.

Model Uncertainty

A related problem is model uncertainty. This is similar to the issue of publication bias but more complicated and harder to formally model. In classic publication bias, the assumption is that the model is always the same and it is applied to multiple datasets. This is somewhat realistic in fields like psychology where many studies are analyses of original experimental data. However in macro-economics and macro-sociology there is just one world and so to a first approximation what happens is that there is basically just one big dataset that people just keep analyzing over and over. To a lesser extent this is true of micro literatures that rely heavily on secondary analyses of a few standard datasets (e.g., GSS and NES for public opinion; PSID and ADD-health for certain kinds of demography; SPPA for cultural consumption).

What changes between these analyses is the models, most notably assumptions about the basic structure (distribution of dependent variable, error term, etc), the inclusion of control variables, and the inclusion of interaction terms. If there were no measurement error, this wouldn't be a bad thing as it would just involve people groping towards better specifications. However if there is error, then these specifications may just be fitting the error rather than fitting the model. Cristobal Young showed pretty convincingly that this is the case for the religion/development relationship by showing that the analysis is sensitive to the inclusion of data points suspected to be of low quality.[1]

Likewise, Gerber and Malhotra did meta-analyses of the two flagship poli sci journals and two flagship soc journals and find a suspicious number of papers that are just barely significant. and a suspicious dearth of papers that are just barely insignificant [2]Although, they describe the issue as "publication bias," I think the issue is really model uncertainty in that any decent quant can fiddle with transformations, control variables, standard error structure, etc to push a p of .06 to a p of .04.

Regression to the Mean

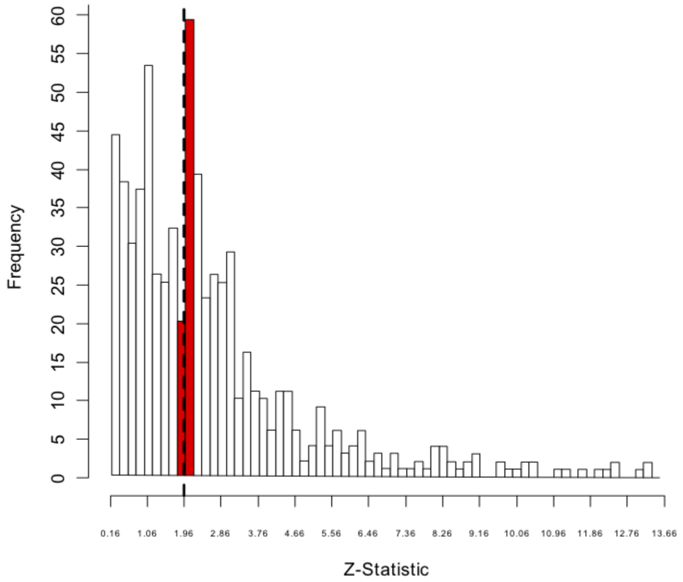
Imagine having all your undergrads write practice essays. You read them all and find the five worst essays, then send these kids to the writing center, or even (martyr that you are) tutor them personally. At the end of the term you see that they were no longer the bottom five but were still in the bottom half.

[1] Young, Cristobal. 2009. "Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth." *American Sociological Review* 74:380-397.

[2] Gerber, Alan S., and Neil Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?." *Sociological Methods and Research* 37:3-30.

Gerber, Alan S., and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3:313-326.

Figure 1a: Histogram of Z-Statistics, APSR & AJPS (Two-Tailed)



Source: Gerber, Alan and Neil Malhotra. 2008. “Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals” *Quarterly Journal of Political Science* 3: 313-326.
(Note, the graph in the *SMR* version looks almost identical.)

Conversely, imagine noticing that most of the faculty brats you know are much smarter than average kids, but not as smart as their parents. In these cases the issue is not necessarily the efficacy of the writing center or the stupefaction of growing up in a college town, but regression to the mean. In the first case it's adverse selection, in the second it's advantageous selection, but the issue is the same.

Regression to the mean occurs whenever you have three conditions:

1. a pre-treatment and post-treatment measure of the key variable (or something similar like two indicators loading on the same latent variable)

2. assignment to the treatment is non-random with respect to the pre-treatment measure

3. the key variable has moderate to low reliability

The reason is that you operationalize effect of the treatment as $(Y_{i1} + e_{i1}) - (Y_{i0} + e_{i0})$. Now it's true that the actual treatment effect would be $Y_{i1} - Y_{i0}$ (assuming that "i" is in the treatment group). But note that e_{i0} is uncorrelated with e_{i1}. Therefore, to the extent that e_{i0} was important to assigning cases to the treatment, a lot of what you think is an effect is really just that the latent value of the cases you selected for treatment weren't as severe as you thought they were. We can demonstrate this with a simulation which shows that the regression to the mean bias is a pretty tight function of the ratio of the signal to noise ratio. If σ_e is anywhere close to σ_y then regression to the mean can be a serious problem. The practical implication is to be *very* skeptical of claims about effects where where the measurement has low reliability and selectivity is built into the system. This is especially an issue with things like evaluating means-tested social policies. Likewise regression to the mean explains much of the placebo effect in medical studies, which is why you can get a placebo effect even with plants and nonsentient patients.