Multiple Regression, Sample Size, and Power

Cyrus Samii

Consider a completely randomized experiment with a randomized binary treatment, $D_i = 0, 1$, where 0 is control and 1 is treatment. We perform the experiment on a sample of N units, of which the proportion p are assigned to treatment and the rest to control. The variances for potential outcomes under treatment and control are S_1^2 and S_0^2, respectively. We estimate the treatment effect by regressing (OLS) the observed outcomes, Y, on the treatment. The coefficient on D, $\hat{\beta}_R$, is our treatment effect estimator. (The R subscript indicates that the data are created via randomization.)

We know from before that the sampling and randomization variance for $\hat{\beta}_R$ is (rescaling by N)

$$NV_{\hat{\beta}_R} = \left(\frac{S_1^2}{p} + \frac{S_0^2}{1-p} \right).$$

Now, suppose we were to run a regression on some set of covariates, X, where the regression specification is such that we include D, $X - \bar{X}$, and then the interactions between D and $X - \bar{X}$. The coefficient on D from this regression, $\beta_{R,cov,interact}$, is the "centered interaction estimator". It has lots of nice properties—see Lin (2013). If X is discrete, it is equivalent to the post-stratification estimator (Miratrix et al., 2013). As Lin shows, $\beta_{R,cov,interact}$ computes the difference in covariate-adjusted treatment and control means. It is consistent for the average treatment effect, and its asymptotic variance is,

$$NV_{\hat{\beta}_{R,cov,interact}} \xrightarrow{a} \left(\frac{S_1^2(1-R_1^2)}{p} + \frac{S_0^2(1-R_0^2)}{1-p} \right)$$

where R_1^2 and R_0^2 are the usual terms of the "proportion of variance explained" among the treatment group and control group outcomes, respectively. When $S_1^2 = S_0^2 = S^2$, this reduces to,

$$NV_{\hat{\beta}_R,hom} = \frac{S^2(1-R_Y^2)}{p(1-p)} = \frac{\text{Var}(Y)(1-R_Y^2)}{\text{Var}(D)}.$$

where R_Y^2 is the variance explained in Y by the full set of covariates and interactions. (If $S_1^2 \neq S_0^2$ but we redefine $S^2(1-R_Y^2)$ as the maximum of $S_1^2(1-R_1^2)$ or $S_0^2(1-R_0^2)$, then $NV_{\hat{\beta}_R,hom}$ is a conservative approximation for the limit of $NV_{\hat{\beta}_{R,cov,interact}}$.)

Now, if this is not a randomized experiment, things are a little more complicated. The reason is that in non-randomized studies, we are likely to have correlation between D and X, and this actually has consequences for power. Let's call the centered interaction estimator in this case $\hat{\beta}_{NR,cov,interact}$ (NR stands for "non-randomized"). Remember that by the Frisch-Waugh-Lovell theorem, the coefficient on D, after controlling for X, converges to,

$$\hat{\beta}_{NR,cov,interact} \xrightarrow{a} \frac{\text{Cov}(\tilde{Y}, \tilde{D})}{\text{Var}(\tilde{D})},$$

where $\tilde{}$ means "residualized with respect to everything else in the regression." That is, $\hat{\beta}_{NR,cov,interact}$ is equal to the coefficient of the regression of $\tilde{Y}$ on $\tilde{D}$. Then, by analogy to the expression for $NV_{\hat{\beta}_R,hom}$, we can define

$$NV_{\hat{\beta}_{NR},hom} = \frac{\text{Var}(Y)(1-R_Y^2)}{\text{Var}(D)(1-R_D^2)},$$

where R_D^2 is the variance explained D by the full set of covariates and interactions. Therefore, the consequences of going from a randomized to a non-randomized design is to inflate the variance of the estimator of the coefficient on D by,

$$D^2(\hat{\beta}_{NR,cov,interact}) = \frac{\frac{\mathrm{Var}(Y)(1-R_Y^2)}{\mathrm{Var}(D)}}{\frac{\mathrm{Var}(Y)(1-R_Y^2)}{\mathrm{Var}(D)(1-R_D^2)}} = \frac{1}{1-R_D^2}.$$

This is the so-called "collinearity penalty." It implies that the *effective sample size* in a non-randomized design goes down as the *degree of confounding* increases. The reason, heuristically, is that as the degree of confounding increases, then comparisons between treatment and control units become increasingly *imbalanced* as we move over values of X. Fixing sample size, unbalanced comparisons are more variable than balanced comparisons, as the variance is driven by the smallest group that is included in the comparison.

References

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics* (In press).

Miratrix, L. W., J. S. Sekhon, and B. Yu (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of Royal Statistical Society Series B* (forthcoming).