Correlations in regressors and sampling distribution of regression coefficients

Cyrus Samii
Department of Political Science, Columbia University,
and MacMillan Center, Yale University

February 17, 2011

A point that does not receive adequate treatment in the political science literature on grouped data regression is that the consequences of correlation in the errors depends considerably on corresponding correlation in the *regressors* (that is, the X's). The point was elaborated by Moulton (1986) and Angrist and Pischke (2008) explore the consequences in some more detail. To their credit Beck and Katz (1995) and Beck and Katz (1996) raise this point, however they did not elaborate on the consequences for estimating standard errors. To see how this works, consider a simple data generating process with only one covariate taking on values, x_{it}. When the data are assumed to be generated by an iid stochastic process with fixed N and growing T, the central ("meat") term in the formula for the variance of regression coefficients is given by:

$$\text{plim}_T \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t'\mathbf{u}_t\mathbf{u}_t'\mathbf{x}_t = \text{plim}_T \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} x_{it}x_{jt}u_{it}u_{jt} = \mathrm{E}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} x_{it}x_{jt}u_{it}u_{jt}\right).$$

We have not fixed the regressors here, and so by iterated expectations, we can rewrite the above as,

$$\mathrm{E}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left(x_{it}x_{jt}|\mathbf{u}_t\right)u_{it}u_{jt}\right).$$

The interior conditioning on $\mathbf{u}_t$ allows us to check properties of this quantity over different assumptions on the process generating the time-period-specific vectors, $\mathbf{x}_t$.[1] What we see is that if x_{it} are not correlated across-cross-sections ($i = 1, ..., N$)—that is, when there is no contemporaneous correlation in the regressors—then it must be that for $i \neq j$, $\mathrm{E}\left(x_{it}x_{jt}|\mathbf{u}_t\right) = 0$, in which case the terms containing the $u_{it}u_{jt}$ for $i \neq j$ are zero, even if $\mathrm{E}\left(u_{it}u_{jt}\right) \neq 0$. Kezdi (2003) developes this point for an arbitrary number of regressors. In a given sample,

[1] If the $\mathbf{x}_t$ and $\mathbf{u}_t$ are assumed independent of each other we can even drop the conditioning.

a sufficient statistic that allows one to study this possibility is,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} x_{it}x_{jt} \text{ for } i \neq j,$$

which can be put onto a normalized scale by using the "intra-class correlation" for the x_{it}'s. Contemporaneous correlation in the $\mathbf{x}_t$'s is checked computing the intra-class correlation coefficient that groups over time periods. The same point applies to serial correlation, and serial correlation in $\mathbf{x}_i$'s is checked with the intra-class correlation coefficient that groups over cross-sections.

The upshot is that contemporaneous correlation or serial correlation in the errors contributes to the variance of regression coefficient estimates only to the extent that it is matched by corresponding correlations in the regressors. If one suspects, say, contemporaneous correlation in the errors but the contemporaneous correlation in the regressors is very close to zero, then any contemporaneous correlation in the errors will be of little consequence in terms of computing correct standard errors. In such a case, an analyst seeking to minimize parametric assumptions may reasonably prefer estimates that ignore contemporaneous correlation and standard error estimates robust to serial correlation (e.g., ASEs), over a possibly misspecified parametric correction for serial correlation (e.g. Prais-Winsten for AR(1)) with standard error estimates robust to most inconsequential contemporaneous correlation.

The point is illustrated starkly in Figure 1. The top row and bottom row each represent different data generating processes. For both cases, we have $y_{it} = \beta_0 + \beta_1 x_{it} + \epsilon_{it}$, where I've set β_0 to 0 and β_1 to 1 and the x_{it}'s are normal draws. In both cases, the data consist of 50 clusters (indexed by i) of 50 units each (indexed by t), and the ϵ_{it} are normal errors uncorrelated with the x_{it}'s. Within each cluster the ϵ_{it}'s have correlation of 0.9. For the top row, the x_{it}'s had zero correlation within clusters, whereas for the bottom row the correlation is again 0.9 with clusters. We see that when there is no correlation in the x's, then the correlation in the ϵ's doesn't contribute to the variance of regression coefficient over simulations, and so the vanilla OLS standard errors are accurate. The cluster robust standard errors are also accurate although they exhibit a bit more volatility. But when there is correlation in the x's to go along with correlation in the ϵ's, then this clearly affects the sampling variance of the regression coefficient—note how the dashed line is now way out at about 0.14, and the cluster robust standard errors are needed to obtain a nearly unbiased estimate (it still appears to be somewhat attenuated on average; the code I used didn't employ the finite sample correction, so that might explain it). We see that the cluster robust standard errors are quite volatile, but the vanilla OLS standard errors never even come close.

Correlations in the regressors can be checked in sample data, e.g. by checking the intra-class correlations of regressors. The examples here show why it is useful to do so, allowing the analyst to make better-informed decisions about what kinds of standard estimates are likely to be the least biased among a set of imperfect alternatives.

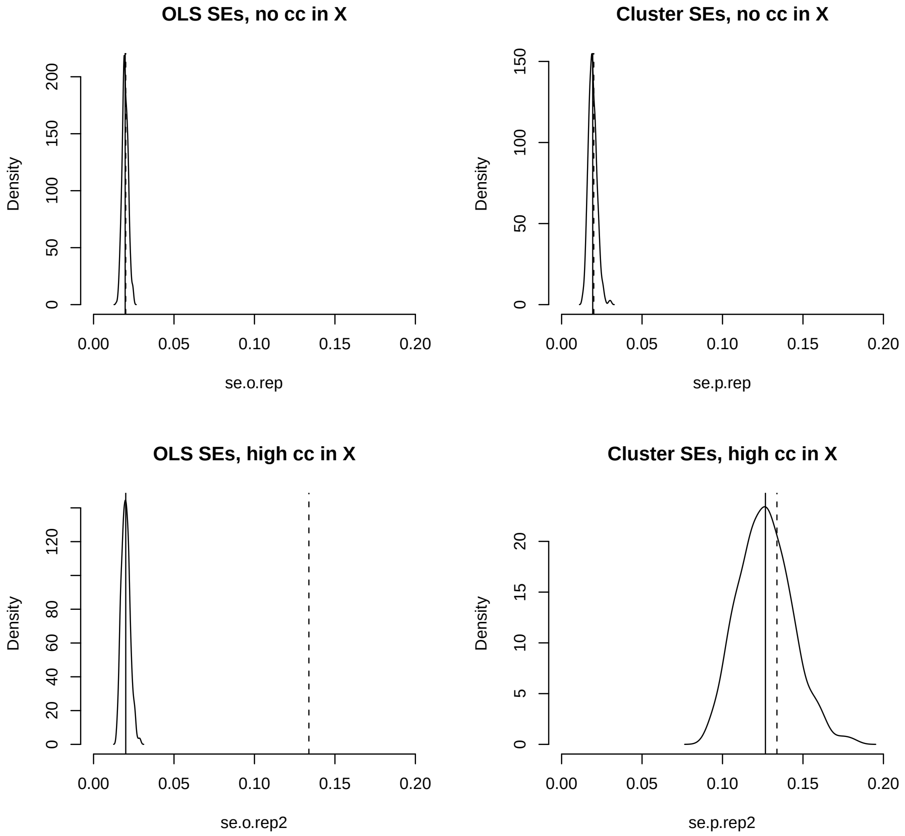

Figure 1: *Kernel density plots showing the distribution of standard error estimates for the coefficient on x_{it} from 500 simulation runs for OLS standard errors and cluster robust standard errors. The dashed line shows the actual standard deviation of the regression coefficient over the 500 runs, and the solid line shows the mean of the standard error estimates. For all four cases, there is substantial intra-cluster correlation in the errors, but only for the bottom two is there any intra-correlation in the x'_{it}s. "cc" in the plot titles refers to "clustered correlation."*

References

Angrist, Joshua D. and Jorn-Steffan Pischke. 2008. *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Beck, Nathaniel and Jonathan Katz. 1995. "What to Do and Not to Do with Time-Series Cross-Section Data." *American Political Science Review* 89(3):634–647.

Beck, Nathaniel and Jonathan N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6(1):1–36.

Kezdi, Gabor. 2003. "Robust Standard Error Estimation in Fixed-Effects Panel Models." Budapest University and Central European University.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3):385–397.