Report on *Economics* Manuscript 1247 "An estimation of worker and firm effects with censored data"

The authors develop and implement a method to address censored earnings data when estimating linear regression models with two high-dimensional effects. Their method is applied to the workhorse log-linear decomposition of earnings into worker- and firm-specific effects introduced by Abowd, Kramarz, and Margolis (1999). The proposed method, "fill-in iterated least squares" (FILS) works by imputing censored observations given an estimate of the model parameters and then re-estimating the model parameters by application of least-squares to the actual and imputed data. The authors derive properties of their method in a simplified setting, then document its properties using simulation studies. They conclude with an application to Spanish matched employer-employee data. Their results indicate that failure to account for censoring biases the parameters on observed characteristics, attributes too much earnings variation to firms, and biases the estimated correlation between worker and firm effects downward.

The topic of this paper is very relevant to the large and growing group of scholars working with matched data. Earnings data are often censored in administrative records from which employer-employee matched data are drawn, but the problem is often ignored. The proposed FILS method is novel, but I think the authors need to be more clear about the precise nature of the contribution. I take their method as a heuristic approach based on an analytic derivation for a simpler case. I think the overall pattern of results described in the paper is informative, and my comments focus on technical considerations that I suspect are not likely to change the paper's qualitative implications.

Comments

1. The derivation and proofs related to the FILS estimator are somewhat ad hoc. The included proofs have to do with a rather restricted case with no regressors and no unobserved heterogeneity. The authors need to be clearer why the derivations are general. Alternatively, it may be fine to simply present the algorithm as a heuristic based on the rigorous derivation of the simpler result. The authors just need to be clearer about what they can prove and what they cannot.

2. The multicollinearity described in footnote 11 seems to indicate the model has not been implemented properly. The solution (adding noise to dummies) is non-standard. This raises the additional question of precisely how the worker and firm effects are separately identified, making it unclear how the results should be interpreted.

3. The authors should use a different method for computing the AKM decomposition in the simulations and in the data application. The authors rely on the statement on page 24 that "Abowd et al. (2004) ... noted that ... sweeping out the worker heterogeneity ... gives exactly the same solution as the Least Squares Dummy Variable estimator". I

am not sure which Abowd et al. (2004) paper from the reference list the citation refers to, but it is important to be clear what is true and what is not. Absorbing worker effects and including firm dummies will give algebraically equivalent solutions for the time-varying observables, but not for the firm and worker effects. Corneliessen (2008) argues clearly that for this method to recover the firm effects the firm dummies must also be de-meaned within worker. For this reason, the results relating to the bias on observed characteristics are accurate, but I am not sure how to interpret the results relating to worker and firm heterogeneity. For the simulations, it should be trivial to use the exact solution method proposed by Abowd, Creecy, and Kramarz (2002) or any other method that properly recovers the coefficients on the fixed effects. Alternatively, they could just focus on how failing to deal with censoring biases estimated parameters associated with the covariates.

4. In follow-up work, the authors could appeal to the statistical literature on iterative methods for dealing with censored data. The proposed method is a sibling to the data augmentation algorithm and associated MCMC methods. I think the current paper can be published without actually implementing such a method, but the possibility should be discussed, since it would benefit future work that builds on this paper. The paper by Chib (1992) would be a useful entry point for this literature, as would Tanner (1997).

References

Chib, S. (1992). Bayes inference in the tobit censored regression model, *Journal of Econometrics* **51**(1): 79–99.

Tanner, M. (1997). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Series in Statistics, Springer New York.