Analyzing Multiple Surveys: Results from Monte Carlo Experiments

Eduardo L. Leoni
Department of Political Science, Columbia University
7th Floor, International Affairs Bldg.
420 W. 118th Street
New York, NY 10027
email: ell2002@columbia.edu

Abstract

In the Political Analysis "Special Issue on Multilevel Modeling of Large Clusters" (Autumn 2005) several papers discuss the adoption of two step models to analyze data from multiple surveys. Complementing that issue, in this paper we perform Monte Carlo experiments comparing the two syep logit model's performance to maximum likelihood, bayesian and pooled estimation using clustered standard errors. We find that at low levels of residual variation across groups, pooling the data is the most efficient method, but even clustered standard errors provide incorrect coverage. We show that jackknife estimation can overcome this limitation. When the level of residual variation is higher the random effects estimators (two-step, ML or Bayesian) outperform pooled estimation both in efficiency and quality of the inferences. We also find that pooled estimation is less robust to violations of the random effects assumptions. The two step approach is somewhat less efficient than alternate approaches but has important advantages: it is computationally simpler and easier to extend to new models. We conclude by replicating two studies with data from multiple surveys, which illustrate the importance of allowing coefficients to vary across groups and the flexibility of the two step approach.

1 *Introduction*

The increasing availability of cross-country and repeated cross-sections survey datasets has sparked interest among political scientists on what are the appropriate strategies to analyze them. We are particularly interested in the following situation. The researcher has at hand a dataset with several hundreds of individual observations in each group/survey, and is mainly interested in the relationship between survey level covariates, or "context", and individual level behavior. These relationships might be either direct (that is, additive) or indirect (through an interaction with an individual level variable). There are at least two characteristics of this problem that demand careful attention. The first is the correlation in the errors introduced by the grouped nature of the data. In this paper, we discuss at length the assumptions that have to be made about the error term in these models, showing why is it important to take into account their group level components. A second feature is the usually small number of observed groups (e.g. countries) but several hundreds of individual level observations within each group. A third characteristic is common to most survey data. Namely, the dependent variable is very often qualitative in nature.

In a special issue of *Political Analysis* dedicated to the multilevel analysis of large clusters survey data, most authors advocated variants of a two step strategy. In the first step, they estimate separate regression models for each survey including only variables that vary within survey as regressors. In the second step, using a variety of weighting schemes, they estimate regression models where the dependent variable are the estimated parameters in the first step. At this stage, variables that vary across surveys are used as independent variables.

However, as noted by Franzese and Beck on the same issue, it remains unclear if there is any advantage in applying the two step strategy over "one step" estimation methods such as pooled regression with cluster robust standard errors or maximum likelihood random effects regression. Our objective in this paper is to partly fill this gap.

We first discuss the mixed model (also known as random coefficient or multilevel model) assumptions for the linear regression and the logit cases. We then perform a set of Monte Carlo experiments of logit models that suggest the following: a) when the true level of residual variation at the group level is small, a pooled strategy is the most efficient; b) unfortunately, the regular standard errors, and even the cluster robust standard errors, are too small, leading to rejection rates much lower than the nominal level we set (95%) . As an alternative, we suggest a jackknife estimator of the standard error, which performs quite well under these circumstances and is computationally fast and simple to implement. We also investigate maximum likelihood and Bayesian multilevel model procedures. As expected, they turn out to be the most efficient estimators of those considered. Even when the residual level of variation at the group level is nil, they are just as efficient as the pooled estimator. We next consider the two step estimator proposed by Hanushek (1974). The procedure is the least efficient when the group level residual variation is close to zero, but the penalty is modest. We find that this procedure, alone or combined with (hc3) robust standard errors also produces reliable inferences.

In the next set of experiments we violate the random effects assumptions by introducing correlation between the group level error terms and the individual level variable. We show that this violation affects much more the pooled than the other estimators. Not only is the pooled estimator much less efficient, but the inferences made are incorrect (too liberal) even with the jackknife standard errors.

The overall message is clear. The pooled estimator is uniformly worse than the other estimators as long as there is some group level residual variation. The inefficiency and reliability of inferences are particularly troublesome if there is correlation between individual level variable and the group level error term. The two step estimator is slightly less efficient than the Bayesian or ML estimators and some differences in the performance at inference persist. However, there are more similarities than differences, so the choice of estimator can be left to be more a matter of (computational) convenience and personal preferences than anything else.

The last section applies some of these methods to actual datasets. We first replicate the analysis of (Baker, 2005), showing that the specific choice of which random effects to use did not matter much in this application, but cluster standard errors perform poorly. We next focus our attention on the research on happiness from the economics perspective, showing that the two step approach easily extends to the ordered logit case. We discuss Blanchflower and Oswald (2004), who completely ignore the clustered nature of the data, leading to severely optimistic standard errors. This example also shows that including dummy variables for each survey is insufficient to deal with the heterogeneity in this case, and one should also allow the other coefficients to vary across surveys.

2 *The linear regression case*

In order to build some intuition we start with a linear regression model. Let s index the groups (e.g. states, or countries) and i index individual observations (e.g. survey respondents) within groups. For simplicity sake, let's assume the model has one independent variable x at the individual level, one (z) at the group level and their interaction ($x \cdot z$). It has the following form[1]:

$$y_{is} = \gamma_{00} + \gamma_{10} x_{is} + \gamma_{01} z_s + \gamma_{11} x_{is} \cdot z_s + v_{is} \tag{1}$$

To simplify notation, we also assume that there are T individual observations in each of the S groups (i.e., the panels are balanced.) The models we consider make different assumptions about the unobserved effect v_{is}. Do we allow it to be correlated with x? With z? Do we want to make assumptions about the form of this correlation or not? Lastly, do we think the error term is systematically related to the groups themselves or is most of the variation contained at the individual level?

In such general terms, it is hard or impossible to answer any of those questions. Therefore, it is useful to be more specific about what questions we are trying to answer when pooling

[1] The nonstandard subscripts will be justified below

cross-sectional survey data.

The defining characteristic of pooled cross-sectional data is that it is feasible (and practical) to do a group by group analysis. Therefore, if we are interested in how x_{is} affects y_{is} in a particular group s, one possibility is to do the analysis using solely observations from that group. On the other hand, note that such parameter does not appear by itself in Equation 1. We can rewrite Equation 1 in hierarchical form so that it does.

$$y_{is} = \beta_{0s} + \beta_{1s} x_{is} + \varepsilon_{is} \tag{2}$$

$$\beta_{0s} = \gamma_{00} + \gamma_{01} z_s + u_{0s} \tag{3}$$

$$\beta_{1s} = \gamma_{10} + \gamma_{11} z_s + u_{1s} \tag{4}$$

Note that we transformed the error term v_{is} in the original equation to an error components format. It is now defined as $\varepsilon_{is} + u_{0s} + u_{1s} \cdot x_{is}$. The models and estimation methods we consider impose different assumptions about u_0, u_1 and ε.[2]

2.1 *Complete pooling*

The first, and perhaps most obvious, option is to estimate Equation 1 by pooling all the data and performing a linear regression. If x and z are uncorrelated with the composite error term v_{is}, while $V(u_0) = 0$ and $V(u_{1s}) = 0$, the OLS estimator will be best linear unbiased. If either $V(u_0) \neq 0$ or $V(u_{1s}) \neq 0$, OLS will be consistent, but its standard errors will be wrong. It is easy to construct a standard error that is robust to arbitrary heterokedasticity while also taking into

[2] We find it best to avoid philosophical distinctions between random and fixed effects models and concentrate on the properties of the unobserved effects. We do not want to imply that the philosophical distinctions are unimportant. (For an illuminating discussion, see Western (1998).) We only assume the researcher thinks that explaining variation across the groups s is a valid exercise, and that the number of observations in each panel is large enough to allow for dummy variables estimation.

account the group level correlation introduced by u_{0s} and u_{1s} (Wooldridge, 2002):

$$\frac{N-1}{N-k}\frac{S}{S-1}(Q'Q)^{-1}\left(\sum_{j=1}^{S}\left[\left(\sum_{i\in \text{groups}s}(e_i q_i)\right)'\left(\sum_{i\in \text{groups}s}(e_i q_i)\right)\right]\right)(Q'Q)^{-1} \tag{5}$$

Where N denotes the total number of observations, S the number of groups, e is the residual from the OLS regression and Q is a matrix formed by combining x, z and $x \cdot z$ with a column of ones. Finally, q_i is the row vector corresponding to individual i in matrix Q.

This is the approach favored by Franzese (2005). Its main benefit is that of consistency in the presence of heteroskedasticity of unknown form, since some of the methods described below rely on a quite restrictive heteroskedastic process. There are two main problems with pooled estimation. First, cluster robust standard errors have known properties only when the number of groups is large relative to the number of observations within groups (Wooldridge, 2003), which of course is the opposite situation we are interested in. Although there are small sample corrections for the simple linear regression case(Long and Ervin, 2000; Mackinnon and White, 1985), we are unaware of any such adjustments for the clustered case. The second issue concerns efficiency. If the level of heteroskedasticity is high and has a known form, complete pooled estimation can be quite inefficient compared to the random effects models described below.

2.2 Dummy Variables

Let's define the dummy variables model to be the case where the unobserved effects u_0 and u_1 are estimated without assuming a distribution across groups. Assume $V(u_{1s}) = 0$ and $E(\varepsilon_{is}|x_{is}) = 0$. If we make the within transformation, or more simply include a dummy variable for each group, we get what is known in econometrics as the fixed effects estimator. We can recover estimates of γ_{11} and γ_{00} but not of γ_{01} (since z_s is perfectly collinear with the set of dummy variables.) The error component u_{0s} can follow any distribution. Most importantly, it can be correlated with x_{is} which is the dummy variables model major advantage. The main disadvantage is that it does not allow us to estimate the direct effect of z_s. The model is consistent as $S \to \infty$ with T fixed.

It is not the most efficient approach, however, if the orthogonality condition between v_{is} and the data (x_{is} and z_s) is indeed satisfied.

If we do not wish to assume that $V(u_{1s}) = 0$ we can estimate a dummy variables model where different intercept and slope vectors for every group are estimated. It consists of a full set of interactions between x_{is} and the group level dummies. This varying intercepts, varying slopes model is equivalent to the group by group analysis, and is consistent as T increases. Unfortunately it does not allow us to estimate directly any of the parameters in Equation 1.

2.3 Random Effects

If we are willing to assume that the unobserved effects u_0 and u_1 are uncorrelated with the regressors we can estimate a random effects model. We do it either directly in the form of Equation 1 (the "one step" approach) or by estimating in a first step Equation 2 separately for each group s and then using these estimates as the dependent variables in a second step (the "two step" approach.)[3]

If, in addition to the orthogonality assumption between the disturbances and the explanatory variables, we also assume that $Var(u_{1s}) = 0$ we can estimate the model by what is known in the econometrics literature as the random effects model, although it could more accurately be referred to as a random intercepts model. In the linear case we can use a quasi-demeaning transformation of the dependent and independent variables to estimate the model. The transformation has a simple form:

$$\lambda = 1 - \sqrt{\frac{V(\varepsilon)}{V(\varepsilon) + TV(u_0)}} \tag{6}$$

Assume for the moment that λ is known. We can then transform Equation 1 as:

[3]The one vs. two step terminology is a bit misleading. First because under some conditions, the two procedures are equivalent (Jusko and Shively, 2005; Amemiya, 1978). Secondly because what is called "one step" might involve many "steps" if iterative procedures are used, or in any event more than one step if we use some form of feasible generalized least squares.

$$y_{is} - \lambda \bar{y}_{is} = \tag{7}$$

$$\gamma_{00}(1-\lambda) + \gamma_{10}(x_{is} - \lambda \bar{x}_{is}) + \gamma_{01}(z_s - \lambda \bar{z}_s) + \gamma_{11}(x_{is} - \lambda \bar{x}_{is})(z_s - \lambda \bar{z}_s) + v_{is} - \lambda \bar{v}_{is} \tag{8}$$

Notice that as T increases, $\lambda \rightarrow 1$ and the estimates tend to those of the dummy variables model. This also implies that as T gets larger, the correlation between u_0 and x_{is} will matter less and less, and the particular advantage of the dummy variables model also becomes less important. The problem of correlation between u_0 and x, therefore, should not impact the random effects estimator by much in data from multiple large surveys, with its hundreds, sometimes thousands of individual level observations per group.

What if $V(u_1) \neq 0$? One option is to fully exploit the covariance structure and the assumed strict exogeneity of the disturbance terms ε_{is}, u_{0s} and u_{1s}. By assuming the disturbances are independent normal variates, we can construct a log-likelihood function and maximize it by the usual maximum likelihood methods, as long as we begin with good starting values. One can use the *lme4* package by Bates and Debroy (2003) which uses the ECME algorithm to provide such starting values and then computes the (restricted) maximum likelihood estimates. The *xtmixed* feature in *Stata 9* can also estimate these modules, although with less flexibility than the *lmer* package. Finally, one can use specialized software for multilevel analyses (e.g. HLM and MLWin) to estimate them.

2.3.1 Two-Step Estimators

The fact that we have large T, and therefore are able to get consistent estimates of β_{0s} and β_{1s} immediately suggests a two step estimator. The first step involves estimating a macro-group by macro-group regression model. That is, we estimate the model in Equation 2 separately for each s grouping. This produces S values each for the slope and intercept parameters that then become

dependent variables in the regressions given by Equations 3 and 4. The group specific intercepts β_{0s} and coefficients β_{1s} can thus be seen as a reduced-form parameters, with restrictions set by the contextual level model.

In the second step we can do a separate analysis for each of the individual level coefficients estimated in the first stage. Since heteroskedasticity is introduced by the fact that the dependent variable is estimated, most likely with unequal error variances across groups, weighting can bring some efficiency gains. Saxonhouse (1976) and Wooldridge (2003) suggest weighting by the inverse of the standard error. The logic is simple: we should downweight the observations that have more imprecise estimates. Hanushek (1974), Borjas (1982) and Lewis and Linzer (2005), on the other hand, argue that there are multiple sources of variation implied by the model: some related to the e_{is} term (the sampling variance), while others are related to the macro-level disturbances (in our case, u_{0s} and u_{1s}). Weighting by the inverse of the standard errors neglects the existence of macro-level disturbances. To appropriately correct for heteroskedasticity in regression explaining the intercepts $\hat{b}_{0s}$ one should weight each observation by

$$w_s = \sqrt{V(u_0) + V(\hat{\beta}_{0s})} \tag{9}$$

(A similar calculation could be done for each individual level coefficient.)

We already have an estimate of the sampling variance of each s, which is $V(\hat{\beta}_{0s})$, since we have the standard error of $\hat{\beta}_{0s}$. The remaining issue is to derive an estimate of $V(u_0)$. We gain some information about it by using the estimated σ^2 from an (unweighted) ordinary least squares regression of $\hat{\beta}_{0s}$ on z_s. However, we should adjust it to take into account the fact that it reflects both sources of error.

Following Hanushek (1974), Lewis and Linzer (2005) show that $V(u_0)$ can be estimated by:

$$\frac{\sum_s r_{0s}^2 - \sum_s V(\hat{\beta}_{0s}) + tr((Z'Z)^{-1}z'GZ)}{C - l} \tag{10}$$

where r_{0s} are the residuals of the OLS (unweighted) second level regression, $V(\hat{\beta}_{0s})$ is the estimated variance of the intercept in survey s, Z is the matrix of group level independent

variables, G is a diagonal matrix with $V(\beta_{0s})$ as the diagonal elements and l is the number of independent contextual variables.

With a small number of groups Lewis and Linzer (2005) provide Monte Carlo experiments that show the FGLS estimator to be 10% more efficient than the OLS counterpart.[4] The standard errors, however, are about 10% too small, so there is the usual trade-off between robust inference and efficiency.

3 Binary Dependent Variables

The linear model has limited utility in our application of interest, since the overwhelming majority of dependent variables derived from survey data is qualitative in nature. Does any of the intuition for linear models travel to more complicated models such as logit and probit? In this section we outline the logit model, following closely the discussion of the linear case. The presentation is brief, since much of the terminology and issues were outlined in the linear model section.

Let y^* be the (unobservable) utility differential between the two alternatives. We observe y instead, which equals 1 if y^* is positive and 0 if it is negative. We want to explain y^* using an explanatory variable (x) that varies within groups and another (z) that varies only across groups:

$$y^*_{is} = \gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s + v_{is} \tag{11}$$

$$v_{is} = \varepsilon_{is} + u_{0s} + u_{1s} \cdot x_{is} \tag{12}$$

3.1 Complete pooling

[4] There are 30 groups in their experiments.

If ε_{is} follows a logistic distribution we will have:

$$P(y_{is} = 1) = logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s) \tag{13}$$

We can estimate this model by maximum likelihood. The obvious problem is that we are assuming that observations are independent within groups, i.e. that both $\mathbf{u_0}$ and $\mathbf{u_1}$ are zero. This may lead to serious underestimation of the standard errors(Moulton, 1990).

3.1.1 Logit with cluster standard errors

One possibility is to calculate standard errors that are robust to heteroskedasticity. One can calculate robust standard errors when the data is drawn from clusters (which we have been referring to as groups) in the following way(Wooldridge, 2003). Let

$$\mathbf{e_{is}} = 1 - logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s)(x_{is}, z_s) \tag{14}$$

if $y_{is} = 1$ and

$$\mathbf{e_{is}} = -logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s)(x_{is}, z_s) \tag{15}$$

if $y_{is} = 0$. The robust variance matrix is given by (Wooldridge, 2003)

$$\frac{S}{S-1}\hat{\mathbf{V}}\left[\sum_{c=1}^{S}\left(\sum_{i \in \texttt{group}c} \mathbf{e_{is}}' \sum_{i \in \texttt{group}c} \mathbf{e_{is}}\right)\right]\hat{\mathbf{V}} \tag{16}$$

with S denoting the total number of groups. Cluster standard errors, however, were designed to handle data structures with many groups and a relatively small number of observations in each group. Although small S corrections have been proposed for the generalized equations framework(Murray, Varnell and Blitstein, 2004), we are not aware of such corrections in the *vanilla* logit/probit case. On the other hand, the logit model just described is inconsistent in the presence of heteroskedasticity, omitted variables (even if uncorrelated with the regressors) and other forms of model specification. The robust variance estimation in (quasi-) maximum

likelihood models can in some circumstances provide an appropriate asymptotic covariance matrix for the (biased) estimator (p.834 Greene, 2001). We will investigate some of these claims in our Monte Carlo experiments.

3.2 Dummy Variables logit

Recall that we defined a dummy variables model to be the case when the unobserved effects u_0 and u_1 can be (consistently) estimated. In the panel data case (small T, large S), directly estimating the unobserved effects leads to inconsistency as $S \to \infty$ due to the incidental parameters problem. That is, with each additional group s we also have to estimate a new u_0 (and, if we allow the β_1 coefficient to vary, u_1). On the other hand, Monte Carlo experiments by Heckman (1981) and Katz (2001) show that this is not an acute problem with T larger than 20.

With $N/S \to \infty$ (i.e. as the group sizes tend to infinity), standard asymptotic results of maximum likelihood estimators can be applied separately for each group. As in the linear case, however, we don't get direct estimates of both parameters of interest γ_{01} and γ_{11}.

$$P(y_{is} = 1) = logit^{-1}(\beta_{0s} + \beta_{1s}x_{is}) \tag{17}$$

3.3 Random effects logit

If we assume that the group level unobserved effects u_0 and u_1 follow a multivariate normal distribution $N(0, \sigma^2)$ independent of v_{is} we can estimate a random effects logit model. The tricky part in estimating this likelihood is how to calculate an integral that appears because of the group level error terms. For logit models with random intercepts (i.e. "random effects logit") in Stata, this is done through Gauss-Hermite quadrature, which is a bad approximation if the number of observations in each group is large and/or $\frac{\sigma^2}{1+\sigma^2}$ is large. From the manual (StataCorp, 2003, p.139) "as a rule of thumb, you should use this quadrature approach only for

small to moderate panel sizes . . . 50 is a reasonably safe upper bound."[5] A second option is to rely on approximations such as *Penalized Quasi-Likelihood* Breslow and Clayton (1993). Although PQL yields biased estimates when cluster sizes are small, Monte Carlo evidence indicate that the approximation is quite good with cluster sizes of 50 or more (Breslow, 2003).

It is important to highlight, however, that this problem is one of implementation/computation. We can sidestep this issue by implementing different integration technique. In particular, by assigning a prior distribution, one can estimate a Bayesian hierarchical logit model using some combination of the Gibbs and Metropolis-Hastings samplers (Gelman, 2004*b*). The Bayesian model can be set up as follows.

$$
\begin{aligned}
P(y_{is} = 1) &= logit^{-1}(\beta_{0s} + \beta_{1s} x_{is}) && (18)\\
\theta_{1s} &= \gamma_{00} + \gamma_{01} z_s && (19)\\
\theta_{2j} &= \gamma_{10} + \gamma_{11} z_s && (20)\\
\beta &\sim MVN(\theta, \Sigma) && (21)\\
\gamma_{00} &\sim N(\mu_{00}, \sigma^2_{00}) && (22)\\
\gamma_{01} &\sim N(\mu_{01}, \sigma^2_{01}) && (23)\\
\Sigma &\sim Inv - Wishart_{\nu_0}(\Lambda_0^{-1}) && (24)
\end{aligned}
$$

β_{0s} and β_{1s} are the group specific coefficients, which we assume to have a multivariate normal distribution with variance-covariance matrix Σ and mean vector composed by θ_{1s} and θ_{2j}. Normal prior distributions for γ and inverse-Wishart prior distribution for Σ complete the model.

[5] Stata 9 introduced an adaptive Hermite quadrature that might perform better, although it is slower. In addition, Stata releases until January 12th 2007 had a numerical problem known as "underflow" *Additions to Stata since release 9.0* (2007). The problem caused particularly large clusters, with as few as three hundred observations, to be dropped from the analysis. For the causes and solutions to common numerical problems in statistical computing see Altman, Gill and McDonald (2003).

Basic results from Bayesian statistics tell us that with T large the priors for the individual level model will matter less (than with small T), so we have the estimates of β_{0s} and β_{1s} to approach the group by group analysis with T large.

3.3.1 Two step estimation

Borjas and Sueyoshi (1994) derive the conditions under which we can estimate the structural parameters γ in a two step fashion analogous to the linear case from the previous section. The authors are interested in the case where $V(u_1) = 0$, so the first step involves estimating a logit model with dummy variables indicating membership to each group. In the second step one regresses these dummy variables estimates on the set of group level explanatory variables using a feasible generalized least squares very similar to Hanushek's from section 2.3.1.[6]

The case for consistency and asymptotic normality of γ as $N/S \to \infty$ and $S \to \infty$ when $V(u_1) \neq 0$ can be briefly outlined. The use of standard $\sqrt{N/S}$ asymptotics guarantees consistency and asymptotic normality of the first stage group by group estimates. One difficulty is guaranteeing that we can ignore the first stage asymptotics when analyzing the second stage, which follows as long as "S grows slowly enough" (p.171 Borjas and Sueyoshi, 1994). Achen (2005) emphasizes this point through a clever example in which this condition is not satisfied. The practical recommendation is that for consistency in the second stage estimates we need T in each group to be "large enough". With hundreds of observations in each group, this condition is likely to be satisfied with data coming from multiple surveys.

One major payoff of the two-stage estimation is well summarized by the authors:

> (the) two-stage approach does not require a distributional choice for u, only the

[6] Borjas (1982) provide the feasible generalized least squares procedure that Huber, Kernell and Leoni (2005) (following Borjas and Sueyoshi (1994)) use in their analysis. It estimates $V(u_0)$ by:

$$\frac{\sum_s r_{0s}^2 - \sum_s V(\hat{\beta}_{0s})}{S - l} \tag{25}$$

> orthogonality conditions. In contrast to the alternative approaches which all require correct specification of the distributions of both the individual effects e and the group effects u (and generally require normality of u for computational tractability), our procedure requires only that the distribution of (v_{is}) be correctly specified and that (β) be a consistent estimator. (p.171)

The most restrictive assumption says the individual level error terms are correctly specified. We are imposing a common variance across groups (i.e. no panel specific heteroskedasticity.) The main problem is that if such heteroskedasticity exists, it will bias all first level coefficients. This oft forgotten issue should not been taken lightly, although we will follow those before us by ignoring it henceforth. Another important payoff is that two-step estimation is a much less demanding computational problem. The implementation of the random effects statistical procedures were designed for problems with a much smaller number of observations per panel and, as we will see, run into numerical difficulties when estimating models with data structures we are interested in.

4 *Design of the Experiments*

In order to analyze the performance of some of the estimators discussed thus far we designed a set of Monte Carlo experiments of logit models with large cluster sizes. We generate the dependent variable according to the following scheme:

$$y^*_{is} = \gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s + \sigma_0 u_{0s} + \sigma_1 u_{1s} \cdot x_{is} \tag{26}$$

$$y_{is} \sim 1\left[Uniform(0,1) < logit^{-1}(y^*_{is})\right] \tag{27}$$

The independent variable z, along with error terms u_0 and u_1, are generated as standard normal variates of length S and then replicated T times so they are constant within groups.

Condition	(1)	(2)
S	$(10, 20, 30)$	$(10, 20, 30)$
T	250	250
$Corr(x, u_0)$	0	0.5
$Corr(x, u_1)$	0	0.5
σ_0	$(0, 0.4, 0.8)$	$(0, 0.4, 0.8)$
σ_1	$(0, 0.4, 0.8)$	$(0, 0.4, 0.8)$

Table 1: Experimental Conditions

The independent variable x and the uniform distribution is drawn at the individual level. In all experimental conditions $\gamma_{00} = -0.5$, $\gamma_{01} = 0.5$, $\gamma_{10} = 1.5$, $\gamma_{11} = 2$, $x_{is} \sim N(0, 0.1^2)$, $z_s \sim N(0, 0.3^2)$, $Corr(z, u_0) = 0$, $Corr(z, u_1) = 0$ and $Corr(x, z) = 0.5$. The conditions that do change are listed in Table 4.

We focus on just two aspects of the simulation results: efficiency and inference. The first set of experimental conditions concentrate on the effect of the residual level of variation and the number of groups on the small sample properties of the estimators. We expect the pooled estimator to perform best in terms of efficiency if the residual level of variation across groups (σ_0 and σ_1) is low. The question is how much we gain, given the large sample sizes within groups. At high level of residual variation, the random effects model should be more efficient than the pooled estimator.

Regarding inference, recall that the standard analyses of the random effects and cluster standard errors rely on fixed T and large S asymptotics. Much less is known about the case with small S and large T. At least until analytical results are available, Monte Carlo experiments can provide some guidance to practitioners.

The second set of experiments violates one of the random effects assumptions. We specify the individual level variable x and both group level disturbance terms u_0 and u_1 to be correlated at the .5 level. We expect the random effects estimators to be more robust than the pooled estimator in this context. The reasoning is as follows. First, note that this kind of violation of the strict exogeneity assumption is exactly the one we are guarded against when using fixed

effects estimators. Secondly, recall that the random effects estimators approximate the fixed effects counterparts as T gets large. It follows that with sufficiently large T the random effects estimators won't be affected much by violations of this kind.

For each simulated data-set we estimate the following models: maximum likelihood and Bayesian estimation (both using the *lme4* (Bates and Debroy, 2003) package for *R*), pooled logit (with cluster and jackknife standard errors), and two step Hanushek (with regular standard errors and with hc3 robust standard errors.

5 *Results*

We present the results from our Monte Carlo experiments in Figures 1 and 2, which are structured as follows. The left column displays the efficiency of each estimator while the right column displays their performance regarding inference. The top row in each figure plots the results for γ_{01} – the additive effect – and the bottom row plots γ_{11}, the interactive effect, of the group level variable z.

Efficiency is measured using the root mean squared error criterion. For γ_{j1}, the root mean squared error is calculated as $\sqrt{\sum_{m=1}^{M}(\gamma_{j1m} - \hat{\gamma}_{01})^2/M}$, where M is the number of simulations for each experimental condition and m indexes the simulations. There are separate lines for each of the four estimators (Pooled, Hanushek, ML and Bayes), and a separate panel for each level of residual variation.

Rejection rates are calculated as follows. In each simulation we compute the 95% confidence intervals for each point estimate and associated standard error around the estimated vale of the parameter using a *quasi* t-test. The rejection rate is the proportion of simulations in which the true parameter is included in this region. Thus, this proportion should approximate 0.95 for correct inference. We should note that for the Bayesian estimator we use the simulations from the posterior distribution to construct Bayesian 95% intervals around the true parameter. As

shown in the right columns of the figures, each estimator of the standard error is presented in its own panel, while the different lines now display the different levels of residual variation.

In all plots we have the number of groups in the x-axis , which can be 10, 20 or 30.

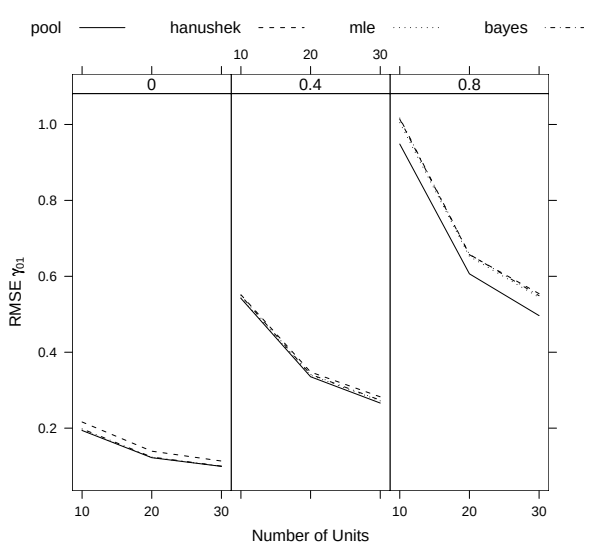
5.1 When the random effects assumptions are valid

The plots in Figure 1 present the results for the case in which the random effects assumptions are valid. As can be seen in the left column plots, the two step estimator is the least efficient at low levels of residual variation but the difference all but disappears at higher levels. The pooled estimator is the most efficient when the degree of residual variation is low and is slightly more efficient overall for parameter γ_{01} (the main effect.) However, this is more than offset by the very large inefficiency attained by γ_{11}.

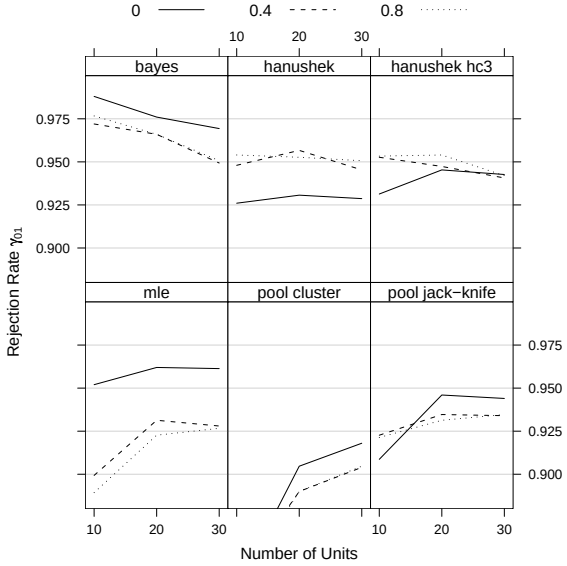
We observe in the plots what will constitute a recurring feature: the good performance of the pooled estimator when the residual level of variation is small. To perform inference, however, we need not only point estimates but also standard errors with appropriate size. Recall that we have reasons to suspect that the cluster standard errors do not have good properties when the number of groups is small and the number of observations in each group is large.

We show the rejection rates in the right panels in Figure 1. As we suspected, cluster standard errors do not provide appropriate coverage. When the number of groups is ten, the rejection rate is less than 90%. Even for S as large as 30 the over-confidence of cluster standard errors can be substantial. It is important to notice that this occurs even when the residual level of variation is zero! This can be very unfortunate if we find ourselves in a situation where the residual level of variation seems to be low, since the pooled estimator has attractive properties in terms of efficiency and unbiasedness in this scenario.

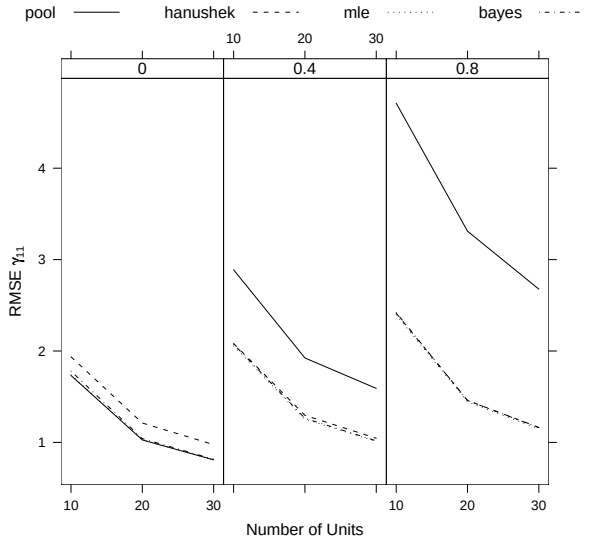
It is interesting to note that a similar problem occurs in the linear regression case when one uses the "regular" robust standard errors, i.e., heteroskedasticity consistent standard errors with no correction for small samples. In the linear regression case Long and Ervin (2000) recommends



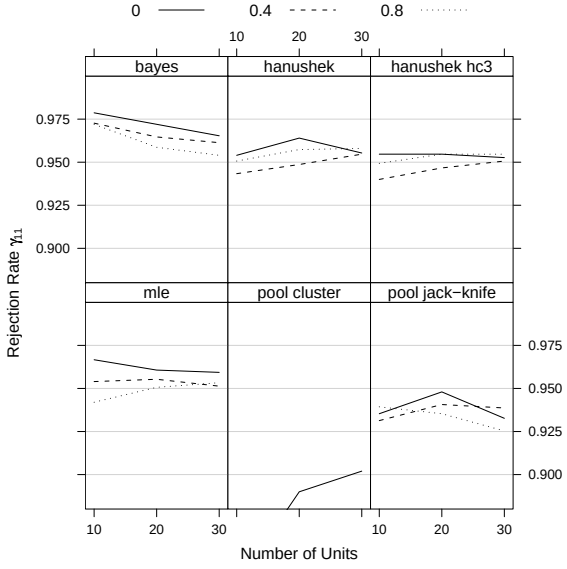
(a) RMSE for γ_{01}



(b) Rejection rates for γ_{01}



(c) RMSE for γ_{11}



(d) Rejection rates for γ_{11}

Figure 1: Monte Carlo Experiments of logit models. These panels reflect the case when the random effects assumptions are valid.

using a different robust standard error formula, known as *hc3*, whenever the sample size is below 500. The *hc3* estimator was created as an approximation to a jackknife estimator analyzed by Efron (Mackinnon and White, 1985), but unfortunately no such approximations exist for the cluster standard error case.

The obvious alternative is to estimate the jackknife numerically. In the cluster case one estimates the model with all groups and then perform S replications deleting one cluster (group) each time. In each iteration, we calculate:

$$jack_{01}^{c} = S \cdot \hat{\gamma}_{01} - (S-1) \cdot \tilde{\gamma}_{01}^{s} \tag{28}$$

where $\hat{\gamma}_{01}$ denotes the estimate from the full model, and $\tilde{\gamma}_{01}^{s}$ denotes the estimate of the model with group s deleted. Then,

$$\frac{V(\mathbf{jack_{01}})}{S} \tag{29}$$

serves as the jackknife estimator of the standard error of $\hat{\gamma}_{01}$. As always, a parallel calculation is done for γ_{11} (or any other parameter of interest). The panels labeled "pooled - jack–knife" demonstrate that the procedure works quite well. The jackknife estimator line is always quite close to the 95% nominal rate at every level of residual variation we looked at.

The Hanushek estimator performs well in terms of inference. We also produced *hc3* robust standard errors. Even though the Hanushek estimator is designed to take into account the heteroskedasticity generated by the fact that the dependent variable is being estimated, it is a good idea to make the analysis more robust if feasible. We see that the costs of doing so seem to be small even when the heteroskedastic process is correctly modeled.

Finally, the Bayesian and the Maximum Likelihood estimators perform quite similarly. They are more efficient than the two step estimators and produce inferences close to the nominal rate. There is a tendency to be too conservative when the number of groups is small, but this probably is more of a feature than a problem.

5.2 Violating the random effects assumptions

We now proceed to analyze the second set of experiments, wherein we let the group level unobserved effects u_0 and u_1 be correlated with the individual level variable x (column 2 in Table 4). Recall that this kind of endogeneity is the prime reason to adopt the fixed effects formulation in panel methods. We argued, however, that this is much less worrisome when the number of observations in each group is large. The following results support this intuition.

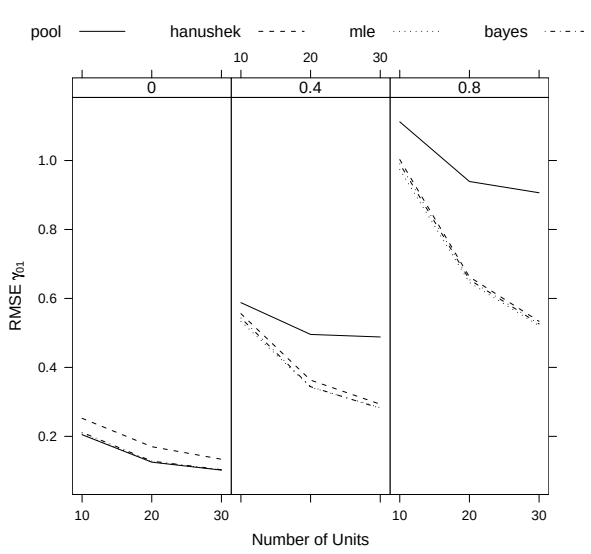
The plots in Figure 2 show that the pooled estimator is quite inefficient even at modest levels of residual variation. The panels also show that the two step Hanushek estimator relative inefficiency is accentuated at low levels of residual variation. The ML and Bayesian estimators continue to be the most efficient across the board.

Notice that the correlation between x and the group level error terms creates problems for the some of the estimators. The Bayesian estimator is quite conservative when the number of groups is small. More seriously, for γ_{01} we see that the pooled estimator is extremely liberal at even moderate levels of residual variation, irrespective of the number of groups or the specific (jackknife vs. cluster) estimator of the standard error. The MLE rejection rates for the same parameter is sometimes too liberal, other times too conservative, with no clear pattern. Finally, the Hanushek estimator seems to provide correct inferences.

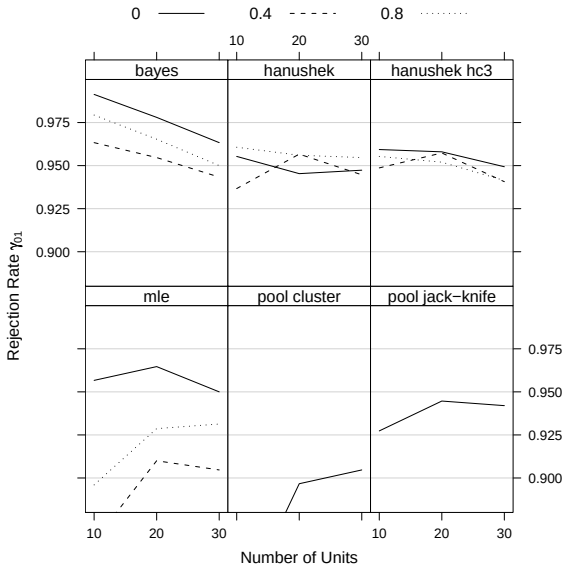
5.3 Summary

We cannot make very general statements using solely Monte Carlo experiments. However, it is clear that the results matched our expectations. We summarize the main findings as follows:

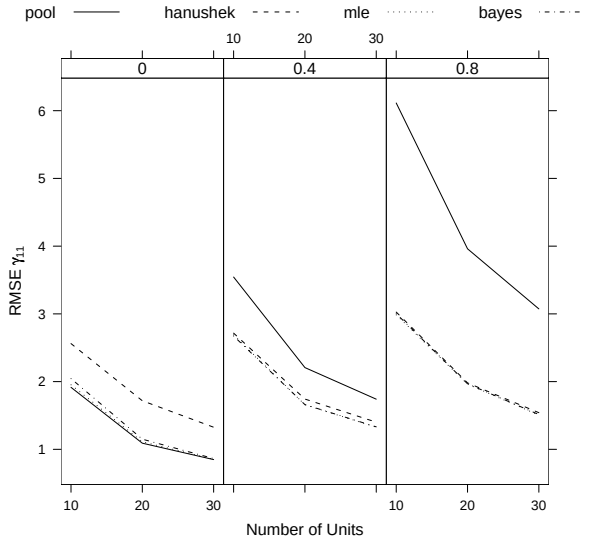
1. At low levels of residual variance, the best choice is to estimate the pooled model. The jackknife estimate of the standard error provides appropriate rejection rates as long as the random effects assumptions hold.
2. However, researchers are seldom in the position of explaining most of the variation across



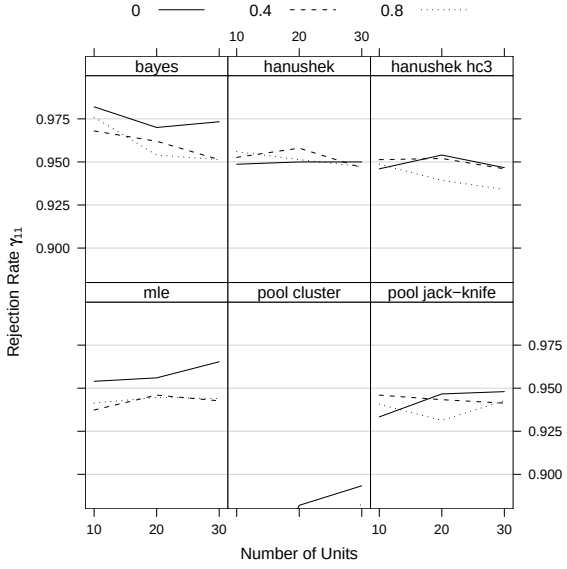
(a) RMSE for γ_{01}



(b) Rejection rates for γ_{01}



(c) RMSE for γ_{11}



(d) Rejection rates for γ_{11}

Figure 2: Monte Carlo Experiments of logit models. These panels reflect the case when the survey level disturbances are correlated with the individual level variable (i.e. the random effects assumptions are violated.)

surveys, and cannot be sure that the random effects assumptions really apply. Therefore, we strongly suggest the estimation of random effects models. They are more efficient and also provide better inferences at a wider array of circumstances.

3. The particular random effects estimator to use seem to be a second order matter, at least in these experiments. Two-step methods are, as expected, less efficient, but only slightly. On the other hand, they seem to be more stable, and computationally cheap. It can also be straightforwardly extended to other non-linear models, such as conditional logit (Glazerman, 1998) or ordered logit (later in this paper).

4. The Bayesian estimator was just as efficient as the MLE, and slightly more conservative. Having the posterior distribution of the parameters, however, makes prediction and model checking much easier (Gelman, 2004*a*, chapter 6), while also providing uncertainty estimates for the variance parameters. And it can be very fast using the *mcmcsamp* function in the *lme4* package.

5. The ML estimator frequently produced warnings of non-convergence, even after much fiddling with the optimization options. Although ignoring the warnings still produced adequate performance in our experiments, it would be irresponsible to ignore them in actual research. We also experimented with other approximations to the log-likelihood function with dismal results. Adaptive-quadrature in some circumstances only converged in about 15% of the simulations. Primo, Jacobsmeier and Milyo (2007) report similar problems when estimating random effects models with large clusters using Stata.

6 *Application I: Trade preferences across individuals and nations*

Why are some individuals staunch supporters of free trade while others remain adamantly opposed to the opening of the economy? Political scientists and economists have tried to answer

this question by using the expectations of the economic theories of trade, typically some variant of the Hecksher-Ohlin theory. This theory assumes the economy is divided into factors (usually labor and capital) and predicts that countries will export goods that are intensive in the most abundant factor and import goods that are intensive in the factor which is more scarce. For public opinion research purposes, a theory that divides the economy into labor and capital is not terribly useful. Thus, researchers have typically used a version of the theory that takes as basic factors of production skilled versus unskilled labor. The specific prediction in this version is that "liberalization raises the relative wages of skilled workers in skill-abundant countries while lowering the relative wages of skilled workers in skill-scarce countries." (Baker, 2005, p.926).

The implications for predicting support for trade across individuals should be clear: in skill rich countries, individual skill should be positively correlated with support for free trade, while in skill-scarce countries the opposite relationship should hold. However, while the expected relationship has been confirmed by studies performed in developed countries, researchers have so far not found the negative correlation between skill and preferences for free trade in less developed economies. It is this particular puzzle that Andy Baker tries to solve in his paper. While staying in the realms of the Hecksher-Ohlin theories of trade, Baker looks at the opposite side of the economic activity. That is, while trade theories are usually concerned with the economic activity of individuals in the *production* of goods, Baker argues that they should also look at the activity of individuals as consumers, with particular attention to how the bundles of goods consumed vary across individuals, since some tend to consume a higher portion of their income with exportable goods than others. The prediction of a model that takes into account this variation in consumption patterns is the following: “the propensity to consume skill-intensive goods should be negatively correlated with support for free trade in skill-abundant countries and positively associated with pro-trade in skill-scarce countries.”

The basic hypotheses can be summarized as follows: the standard H-O theory predicts that the coefficient on individual level skill should be high among countries rich in skill and low

among countries in which skill is scarce. In addition, at countries with average skill levels, the effect of individual skill should be approximately zero. The H-O theory when applied to the consumer side of the economy, on the other hand, predicts that individuals that consume a high portion of skill-rich goods should be more averse to free trade in high-skill countries than in low-skill countries. Baker measures this propensity by individual income. An alternative explanation, proposed by Wood (1997), is that the education of some individuals are so low that they are unable to benefit from unskilled labor-intensive exports. He also tests for the the impact of land vs. capital, using town size as a proxy at the individual level and measuring the relative amount of capital in the whole economy using aggregate data. Finally, he introduces in the estimation *political interest* as a control variable, also interacted with country skill level. All the individual level variables so far are centered and standardized in each country (i.e. they are country specific z-scores.) The country level variables are centered at the international median.

The author uses a random coefficients logit model (with the package HLM) to predict trade preferences across a broad cross-section of countries present in the World Values Survey data. We start with some exploratory analysis. Each panel in Figure 3 plots the group-by-group logit coefficients and confidence intervals of the specified independent variables for each of the 41 countries in the dataset. In the x axis we have country skill, while the solid lines are the two step regression predicted values (with country skill as an independent variable.) Below each plot we print the coefficient and standard errors for the corresponding two step regression. [7]

In the first row we have the three individual level covariates that are interacted with country skill in the author's model. They all turn out to be significant at conventional levels. The effect of individual skill is highly correlated with country skill, fitting quite snugly the predictions of the standard H-O model. More worrying are the coefficients plotted on the second row. Gender and, to a lesser extent, nationalism and the country level intercepts, are significantly correlated

[7] Plotting coefficient estimates across values of a survey level covariate with corresponding confidence intervals is what Gelman and Hill (2007) calls the "secret weapon".

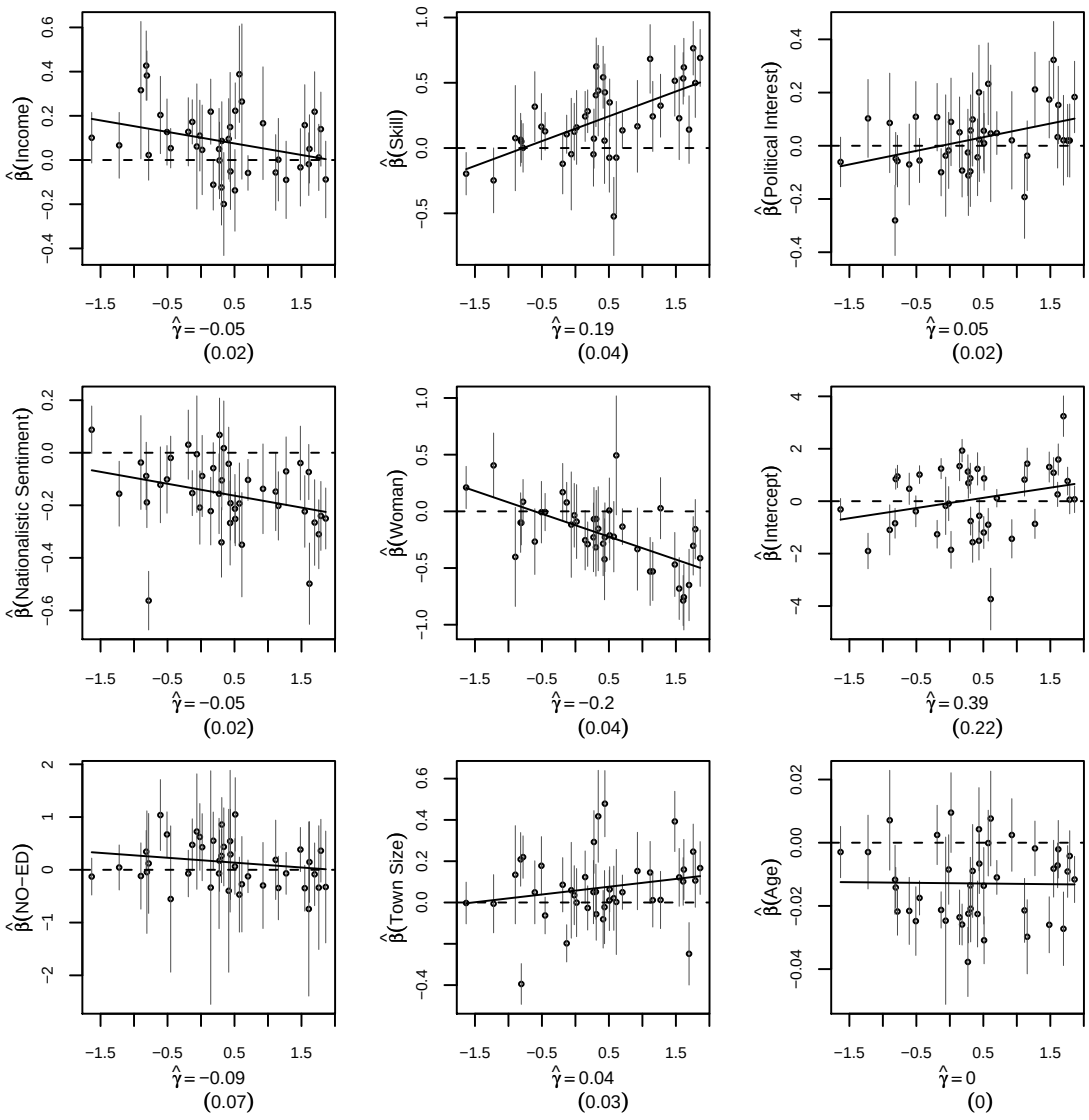


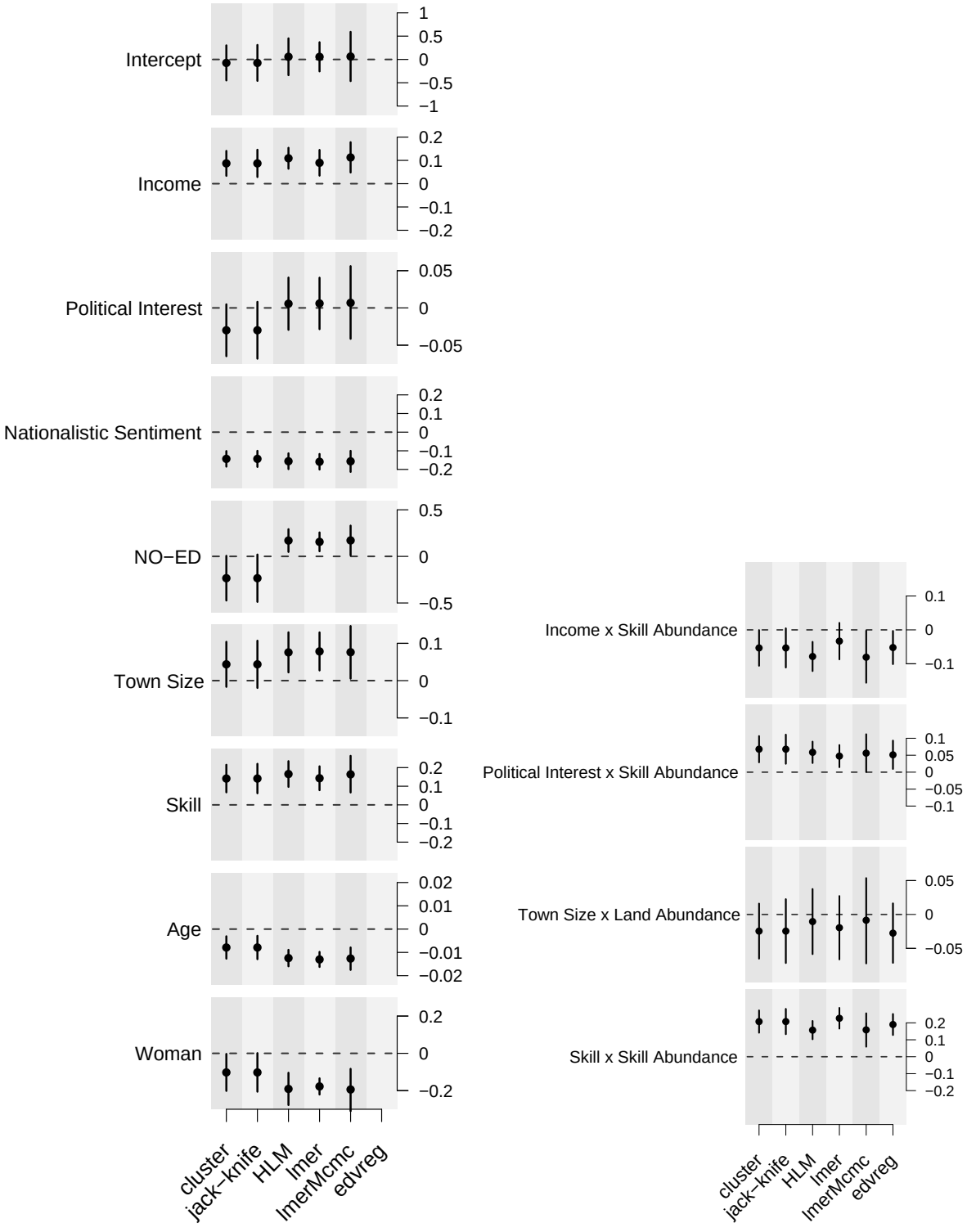
Figure 3: Country-specific estimates of the individual level variables plotted against country skill abundance. The two-step estimates of the relationship is displayed below each panel.

with country skill, but Baker did not include such interactions in his model.

The author provided all that was needed to fully replicate his results. However, since he used multiple imputation and such task is very time consuming in some of the estimation methods we use, we decided to use just the first of the imputed datasets. The results are presented in Table 6 and Figure 4. We focus on the graphical displays, since it facilitates comparisons across models.

In a result consistent with our Monte Carlo experiments, the pooled logit coefficients can be quite different from those of the other models. The most extreme case is that of the *No-Ed* coefficients, measuring the effects of having very low skills on the support for free trade. Pooled logit supports the opposite inference from those of the other models, and all coefficients are statistically significant at the 95% level. Therefore, had the author chosen a different estimation technique, he would find support for Wood's explanation for the lack of support for free trade among lower-middle income countries.

Although the random effects estimators present broadly similar inferences, there are important caveats. First, the *lmer* produced convergence warnings, and we see in table 6 that its estimates were quite different from both HLM and *lmer MCMC*. This suggests the use of the Bayesian estimator for data structures of this kind. The results from *HLM* are close to *lmer MCMC*'s, but the standard errors are much smaller. It ranges from 0.54 to 0.75 of those estimated by the bayesian procedure. Although in our experiments the latter was slightly over-conservative, the magnitude of the differences reported here call into question HLM's standard errors. One possible cause is the use of "robust" standard errors in the HLM package. They are substantively *smaller* than the regular standard errors, in particular for the cross-level interactions. It is 50% smaller for the *Town Size · Land Abundance* coefficient, and 27% smaller for the *Skill · Skill Abundance* coefficient. Even for some individual level variables (*Skill* and *No-Ed*) the robust standard errors are about 10% smaller than the regular standard errors. This is recalls the patterns from our Monte Carlo results, and suggests that one should not use robust standard



(a) Individual Level Predictors

(b) Country Level Interactions

Figure 4: Replication of Baker (2005). The results from the HLM model are slightly different from those reported by the author since we use just the first of his 10 multiple imputation datasets.

	Pooled	Jack-knife	HLM6	Lmer	Lmer MCMC	Hanushek
Intercept	-0.074	-0.074	0.058	0.056	0.063	
	(0.186)	(0.191)	(0.196)	(0.155)	(0.268)	
Income	0.087	0.087	0.109	0.09	0.112	
	(0.027)	(0.029)	(0.022)	(0.027)	(0.033)	
Income x Skill Abundance	-0.053	-0.053	-0.078	-0.033	-0.081	-0.052
	(0.026)	(0.029)	(0.021)	(0.027)	(0.039)	(0.024)
Political Interest	-0.03	-0.03	0.006	0.006	0.007	
	(0.017)	(0.019)	(0.017)	(0.017)	(0.025)	
Political Interest x Skill Abundance	0.068	0.068	0.059	0.047	0.056	0.051
	(0.019)	(0.021)	(0.016)	(0.016)	(0.028)	(0.021)
Nationalistic Sentiment	-0.143	-0.143	-0.156	-0.159	-0.156	
	(0.021)	(0.021)	(0.021)	(0.021)	(0.029)	
NO-ED	-0.234	-0.234	0.17	0.156	0.171	
	(0.119)	(0.126)	(0.061)	(0.050)	(0.081)	
Town Size	0.044	0.044	0.076	0.078	0.076	
	(0.030)	(0.031)	(0.027)	(0.025)	(0.035)	
Town Size x Land Abundance	-0.025	-0.025	-0.011	-0.02	-0.009	-0.028
	(0.020)	(0.023)	(0.024)	(0.023)	(0.032)	(0.022)
Skill	0.141	0.141	0.165	0.143	0.164	
	(0.037)	(0.039)	(0.034)	(0.032)	(0.050)	
Skill x Skill Abundance	0.207	0.207	0.157	0.227	0.159	0.19
	(0.033)	(0.037)	(0.027)	(0.030)	(0.050)	(0.031)
Age	-0.008	-0.008	-0.012	-0.013	-0.013	
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
Woman	-0.102	-0.102	-0.191	-0.178	-0.194	
	(0.049)	(0.051)	(0.043)	(0.022)	(0.058)	

Table 2: Regression results

errors when the number of groups is even of moderate size.

7 Application II: Happiness over time in the United States

One key advantage of the two step method over alternative procedures is how easy it is to adapt it to unique features of the data at hand. In this section we show how to the two step model can easily be extended to ordered multinomial response variables.

The substantive question we address is the following. Does economic growth raise the well-being of the population? Studying the trend in reported well-being over time, Easterlin (1974) argues that happiness is relative. That is, people compare themselves to others when judging their own well-being. Thus, increasing the average income in the population should bring little improvement in average reported well-being. Blanchflower and Oswald (2004) analyze responses to reported well-being questions in both the United States and Great Britain, finding support for the Easterlin hypothesis. However, they go a step further by using individual level data (instead of aggregate responses) and analyzing the trends by subgroups.

Among other interesting results, the authors find that general levels of happiness in the US population have declined over time in the United States from 1972 to 1998. Since income per capita has increased in the period, this is evidence in favor of the Easterlin hypothesis. In addition, by analyzing sub-samples of the data, they find that happiness among women has declined, while happiness among man has remained stable and among blacks it has increased. They measure happiness using the General Social Survey, which asks "taken all together, how would you say things are these days – would you say you are very happy, pretty happy or not too happy?"

7.1 Two-step ordered logit

The ordered structure of the responses are dealt with by using an ordered logit model. By the latent variable formulation we have:

$$y^*_{is} = \gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s + v_{is} \quad (30)$$

$$v_{is} = \varepsilon_{is} + u_{0s} + u_{1s} \cdot x_{is} \quad (31)$$

We observe $y_{is} \in \{1, 2, 3\}$. Assuming ε_{is} follows a logistic distribution and ignoring u_{0s} and u_{1s} for the moment, we have[8]:

$$P(y_{is} = 1) = 1 - logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s) \quad (32)$$

$$P(y_{is} = 2) = logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s) \quad (33)$$

$$-logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s - cut) \quad (34)$$

$$P(y_{is} = 3) = logit^{-1}(\gamma_{00} + \gamma_{10}x_{is} + \gamma_{01}z_s + \gamma_{11}x_{is} \cdot z_s - cut) \quad (35)$$

The cutpoint *cut* is estimated along the other coefficients in the model. The terms u_{0s} and u_{1s} are not in general zero, of course. It is possible to estimate this random effects logit model using Bayesian or maximum likelihood techniques. Two step estimation is also possible, the only difficulty here is that in most situations we would like the cut point to be the same across groups. Fortunately, one can easily constrain the cutpoints by estimating a single ordered logit regression with group level indicators and interactions and use the estimated coefficients as dependent variables in the second step.

That is, we can estimate:

[8]For reasons that will be clear soon, we use the parameterization described in equation 6.12 of Gelman and Hill (2007), where there is a constant and the first cutpoint is constrained to be zero. The parameterization in Stata and Venables and Ripley (2002) estimates an extra cutpoint and constrains the constant to be zero instead.

$$P(y_{is} = 1) = 1 - logit^{-1}(\beta_{0s} + \beta_{1s}x_{is}) \quad (36)$$

$$P(y_{is} = 2) = logit^{-1}(\beta_{0s} + \beta_{1s}x_{is}) - logit^{-1}(\beta_{0s} + \beta_{1s}x_{is} - cut) \quad (37)$$

$$P(y_{is} = 3) = logit^{-1}(\beta_{0s} + \beta_{1s}x_{is} - cut) \quad (38)$$

and then regress β_0 and β_1 on Z in the second step.

7.2 Results

In Table 7.2 we show our replication for the first column of table 2 in Blanchflower and Oswald (2004).[9] The authors completely ignored the clustered nature of the data and simply estimated a pooled ordered logit model. This lead to serious underestimation of the standard errors, particularly for the time trend, which obviously only varies at the survey level. In Figure 5 we compare the size of the regular and cluster standard errors to jackknife standard errors. The regular standard errors for the time trend is less than half of the jackknife standard errors, and even the cluster standard errors should be about 15% larger. The consequence is that the claim made by Blanchflower and Oswald right at the abstract, namely that "(r)eported levels of well-being have declined over the last quarter of a century in the US", is not warranted in such general terms by the data.

In the fourth column we show the results for a two step model in which the first step is an ordered logit with time dummies. These time dummies are then regressed on time in the second step. The standard error for the time trend is as large as the cluster standard errors, but the coefficient itself is smaller. More importantly, note how the standard errors for the other coefficients are more similar to the regular standard errors than for either the jackknife or the cluster. This is expected, since this model, despite taking into account u_{0s}, still assumes $u_{1s} = 0$.

[9] We measure time and age in decades in order to make the table more pleasant to read.

	Regular	Cluster	Jackknife	Hanushek
Intercept	1.74	1.74	1.74	1.9
	(0.08)	(0.08)	(0.08)	(0.05)
$y^* \geq$ very happy	-1.03	-1.03	-1.03	-0.87
	(0.08)	(0.06)	(0.07)	(0.05)
Time	-0.03	-0.03	-0.03	-0.02
	(0.01)	(0.02)	(0.03)	(0.02)
Age	0.16	0.16	0.16	0.09
	(0.03)	(0.03)	(0.03)	(0.01)
Age^2	0	0	0	-0.01
	(0.00)	(0.00)	(0.00)	(0.00)
Male	-0.05	-0.05	-0.05	-0.05
	(0.02)	(0.03)	(0.03)	(0.02)
Black	-0.73	-0.73	-0.73	-0.73
	(0.03)	(0.04)	(0.04)	(0.03)
Other race	-0.14	-0.14	-0.14	-0.15
	(0.06)	(0.08)	(0.08)	(0.06)

Table 3: Regression results. Time measure in decades, 1972=0.

Therefore, including dummies and neglecting possible interactions (i.e. that coefficients other than the intercept might vary across groups) can be worse than simply using cluster standard errors for individual level covariates.

We can relax this assumption by interacting each of the individual level independent variables with the time dummies, which we show graphically in Figure 6.[10] We add for good measure the 2000, 2002, 2004 and 2006 GSS surveys.

The age predictor is centered at 30, thus the intercept reflects the predicted value of y^* for a 30 (instead of zero) years old white woman. We can notice a slight downward trend in the coefficients since 1972, but substantial variation over short periods of time in the coefficients remain. Interestingly enough, the trend for the male coefficients is positive, yielding a flat trend for men over the period (.01). We should keep in mind that the first level coefficients

[10]Blanchflower and Oswald indirectly take into account the varying u_{1s} by analyzing subsamples of women, men, black, whites, etc.

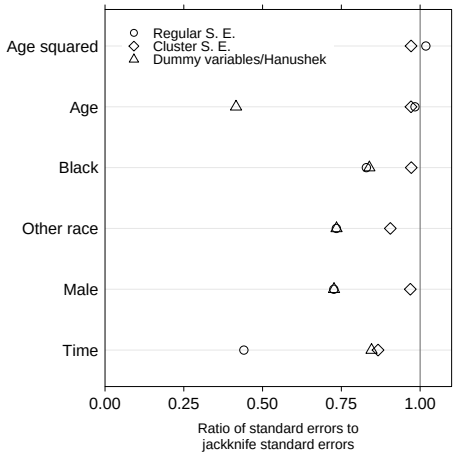


Figure 5: Ratio of standard errors calculated with regular, cluster, and Hanushek (with time dummies) standard errors to those calculated using the jackknife.

are significantly different from zero in only four of the 26 surveys. Furthermore, the estimated happiness gap between men and women is very small throughout the period. In sum, the evidence for an increasing gap between men's and women's happiness is underwhelming,

The difference in happiness between blacks and whites is much more pronounced, as is its downward trend in absolute terms throughout the period. No similar trend is observed in the difference between whites and those in the "other races" category. Finally, we find a downward trend in the linear coefficient of age (in decades). While in the 1970s increasing age was predicted to sizably increase happiness, in the 2000s the predicted increase is basically null. The panel to the right reveals that there is not much evidence for a trend in the quadratic term nor for a quadratic term at all.

This application shows how important interactions can be when modeling multilevel structures. As Western (1998) argued, assuming away the possibility that coefficients vary across clusters can lead to misleading and inefficient estimates. The two step approach proved to be useful and easy to extend in this situation, allowing all coefficients to vary across time.

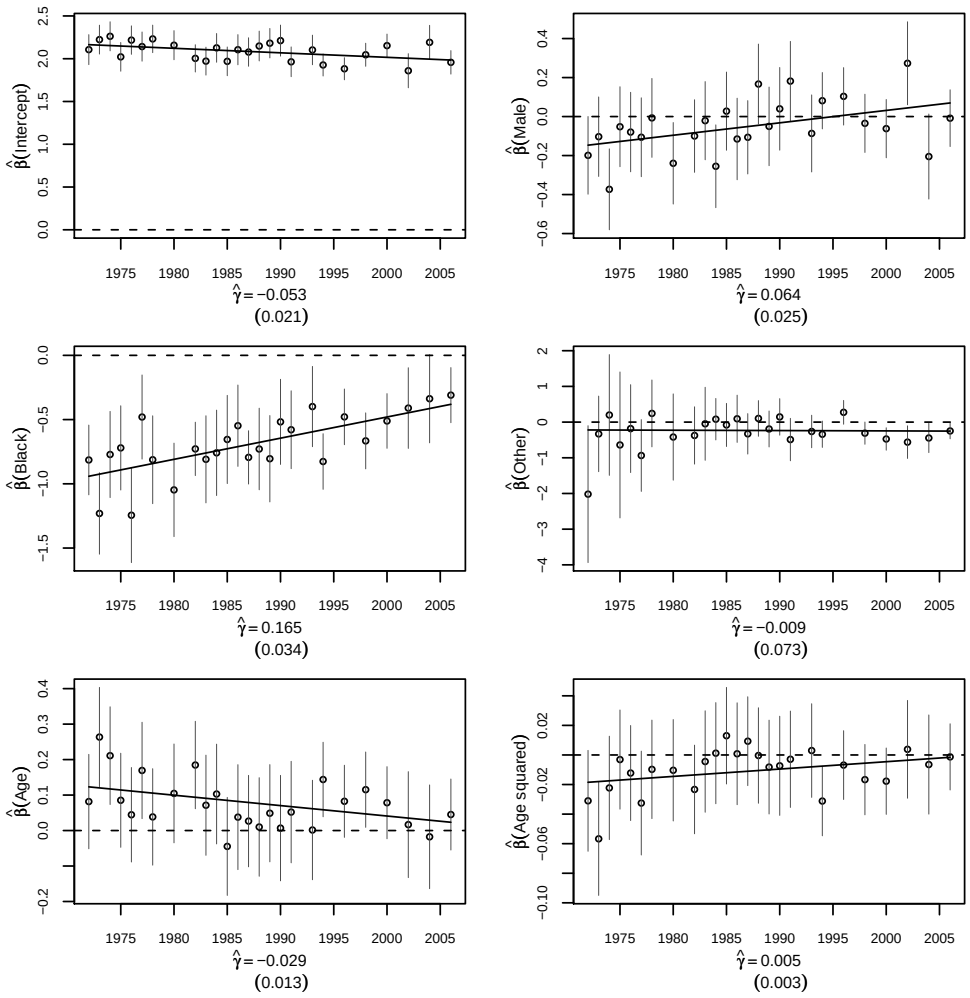


Figure 6: Survey specific estimates of the individual level variables plotted against year of the survey. The two-step estimates of the relationship is displayed below each panel.

8 Conclusion

The typical conclusion of a Monte Carlo study argues that further studies are needed in order to establish generality of the findings, solve the problems encountered and so on. Although a more thorough analysis of specialized software is needed for cases with a small number of clusters and very large cluster sizes, we argue that the evidence presented so far is convincing enough to support some basic contentions in support for the use of two step methods in cross-national survey research.

The first can be stated simply: two-step models *are* "random effects" models. That is, they are just a different method to estimate the same model, as long as the cluster sizes are sufficiently large. Indeed, under the random effects assumptions and in the real data we examined, results from any of the full random effects models and the two step models strongly agree. However, our Monte Carlo results show that some two step model appear to have better numerical stability and robustness to a wider set of conditions than maximum likelihood methods. The two-step approach is also very flexible, as we showed extending it to the ordered logit case.

Our second contention is that cluster (and jackknife) standard errors often perform poorly with large cluster sizes. Researchers are likely to find themselves in situations where some of the variation[11] across surveys can be explained by the survey level regressors but certainly not all of it. When this is the case, random effects models can be much more efficient and no less robust than cluster standard errors. They also outperform cluster standard errors when some of the random effects assumptions are not met. It is noteworthy that in both applied examples we get very different inferences between the cluster standard errors and random effect analyses. Although both random effects and pooled estimation are biased when individual level regressors are correlated with the survey level disturbances, the bias is attenuated in random effects models

[11] With large cluster sizes, checking the residual level of variation can be accomplished through graphical displays, although more formal residual analysis is also possible.

as the number of respondents increase.

The final advantage of two step methods is computational and practical. Breaking down the analysis into two largely separate steps facilitates exploratory analysis and model checking for the researcher, and alleviates convergence problems and computational costs. As an added bonus, the two step methods we discussed are implementable with not much work in any modern statistical software.

The statistical and computational apparatus of random effects modeling center on problems largely irrelevant when cluster sizes are large. So, when group-by-group analysis is both feasible and practical, the Monte Carlo evidence presented here largely supports the use of random effects estimation in two steps.

References

Achen, Christopher H. 2005. "Two-Step Hierarchical Estimation: Beyond Regression Analysis." *Political Analysis* 13(4):447–456.

Additions to Stata since release 9.0. 2007. January 12th 2007 update.

Altman, M., J. Gill and M.P. McDonald. 2003. *Numerical Issues in Statistical Computing for the Social Scientist*. First ed. New York: John Wiley & Sons.

Amemiya, Takeshi. 1978. "A Note on a Random Coefficients Model." *International Economic Review* 19(3):793–796.

Baker, Andy. 2005. "Who Wants to Globalize? Consumer Tastes and Labor Markets in a Theory of Trade Policy Beliefs." *American Journal of Political Science* 49(4):925–939.

Bates, Douglas and Saikat Debroy. 2003. "Linear mixed models and penalized least squares." *Journal of Multivariate Analysis* .

Blanchflower, David G. and Andrew J. Oswald. 2004. "Well-being over time in Britain and the USA." *Journal of Public Economics* pp. 1359 – 1386.

Borjas, George J. 1982. "On regressing regression coefficients." *Journal of Statistical Planning and Inference* 7(2):131–137.

Borjas, George J. and Glenn T. Sueyoshi. 1994. "A two-stage estimator for probit models with structural group effects." *Journal of Econometrics* 64(1-2):165–182.

Breslow, Norm. 2003. "Whither PQL?" UW Biostatistics Working Paper Series. Working Paper 192.
URL: *http://www.bepress.com/uwbiostat/*

Breslow, Norm and D. Clayton. 1993. "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association* 88:9–25.

Easterlin, R. A. 1974. *Nations and Households in Economic Growth: Essays in Honour of Moses Abramowitz*. Vol. Nations and Households in Economic Growth: Essays in Honour of Moses Abramowitz Academic Press, New York chapter Does economic growth improve the human lot? Some empirical evidence.

Franzese, Robert. 2005. "Empirical Strategies for Various Manifestations of Multilevel Data." *Political Analysis* .

Gelman, Andrew. 2004*a*. *Bayesian data analysis*. Texts in statistical science 2nd ed. Boca Raton, Fla.: Chapman & Hall/CRC.

Gelman, Andrew. 2004*b*. "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis* .

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Glazerman, Steven. 1998. Determinants and Consequences of Parental School Choice PhD thesis Harris Graduate School of Public Policy Studies, University of Chicago.

Greene, William. 2001. *Econometric Analysis*. Prentice Hall.

Hanushek, Eric A. 1974. "Efficient Estimators for Regressing Regression Coefficients." *American Statistician* 28(2):66–67.

Heckman, James J. 1981. The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process. In *Structural Analysis of Discrete Data*, ed. Charles F. Manski and Daniel Mcfadden. Cambridge, MA: MIT Press.

Huber, John D., Georgia Kernell and Eduardo L. Leoni. 2005. "Institutional Context, Cognitive Resources and Party Attachments Across Democracies." *Political Analysis* 13(4):365–386.

Jusko, Karen L. and Phillips W. Shively. 2005. "Applying a Two-Step Strategy to the Analysis of Cross-National Public Opinion Data." *Political Analysis* .

Katz, Ethan. 2001. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation." *Political Analysis* 9(4):379–384.

Lewis, Jeffrey B. and Drew A. Linzer. 2005. "Estimating Regression Models in Which the Dependent Variable Is Based on Estimates." *Political Analysis* .

Long, Scott J. and Laurie H. Ervin. 2000. "Using heterscedasticity consistent standard errors in the linear regression model." *The American Statistician* 54(3).

Mackinnon, James O. and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29:305–325.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *The Review of Economics and Statistics* 72(2):334–338.

Murray, David M., Sherri P. Varnell and Jonathan L. Blitstein. 2004. "Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments." *American Journal of Public Health* 94(3):423–432.

Primo, David M., Matthew L. Jacobsmeier and Jeffrey Milyo. 2007. "Estimating the Impact of State Policies and Institutions with Mixed-Level Data." *State Politics and Policy Quarterly* 7(4):446–459.

Saxonhouse, Gary R. 1976. "Estimated Parameters as Dependent Variables." *American Economic Review* 66(1):178–183.

StataCorp. 2003. *Stata Cross-Sectional Time-Series Reference Manual Release 8*. College Station, TX: Stata Press.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42(4):1233–1259.

Wood, Adrian. 1997. "Openness and Wage Inequality in Developing Countries: The Latin American Challenge to East Asian Conventional Wisdom." *Work Bank Economic Review* 11(1):33–57.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, Mass.: MIT Press.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* pp. 133–138.