Estimating Substantive Effects in Binary Outcome Panel Models: A Comparison

Casey Crisman-Cox, Texas A&M University

Dummy variable maximum likelihood estimation for binary response panel models struggles to estimate coefficients or substantive quantities in either short or rare event panels. The standard response is a conditional maximum likelihood that consistently estimates the coefficients on time-varying covariates but makes substantive effects impossible to compute. In light of this problem, multiple suggestions have appeared for computing these effects, but there is little to no guidance as to when one solution may be preferred to another. I address this question by comparing one of these approaches, a correlated random effects estimator, to the maximum likelihood dummy variable estimator. I find that when the number of within-group observations is small or events are rare, the correlated random effects estimator is preferred, but as panels get longer the difference between approaches fades.

Across all subfields of political science, researchers find themselves considering the task of estimating and interpreting panel models with a binary dependent variable.[1] So long as panels are not too short and events are not too rare, there is no problem; a maximum-likelihood dummy variable (MLDV) estimator (i.e., ordinary logit with a dummy variable for each unit) can produce reasonably good estimates of key parameters (Coupé 2005; Greene 2004; Katz 2001). But in situations in which either condition fails (either short panels, rare events, or both), the MLDV is problematic. Given the prevalence of rare events in political science data (e.g., civil or interstate conflict onset) and the frequent use of surveys with a limited number of waves, it is important to ask how we can produce substantively meaningful estimates in these situations.

A recent look through top journals suggests that the main alternative to the MLDV (regardless of sample size or rareness) is a conditional maximum likelihood (CML) estimator developed by Chamberlain (1980).[2] The advantages of this approach are clear: the estimator is fast and readily available, and it can produce good coefficient estimates for time-varying covariates in almost all settings. The drawback of this approach is equally clear: the estimator works by eliminating all group-specific constants from the estimation problem. Removing these constants means, as King (2001) puts it, that "first differences, and indeed every quantity of interest but one, are impossible to compute correctly from estimates of the fixed-effects [logit] model" (499). This refrain is sometimes cited directly by scholars as they justify an alternative approach (e.g., Hafner-Burton, Hyde, and Jablonski 2013, n. 90). Such a trend is alarming, as it frequently leads scholars to embrace less robust estimators that may not fully account for the unobserved heterogeneity common to panel data (i.e., a pooled logit or random effects), in order to produce substantive results of interest.[3]

In response to these limitations, various ad hoc alternatives have appeared to sidestep the interpretation concern. For example, some scholars (e.g., Escribà-Folch, Meseguer, and Wright 2018; Pardos-Prado and Xena 2019) have fit the model using the CML and then produced substantive effects

Casey Crisman-Cox (c.crisman-cox@tamu.edu) is an assistant professor of political science at Texas A&M University, College Station, TX 77843.

Data and supporting materials necessary to reproduce the numerical results in the article are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). An online appendix with supplementary material is available at https://doi.org/10.1086/709839.

1. Some cursory Google Scholar searches find that, since 2015, about 15% of articles in the top three general-interest journals (*American Political Science Review* [*APSR*], *American Journal of Political Science* [*AJPS*], and *Journal of Politics* [*JOP*]) contain the terms "fixed," "effects," and "logit."

2. Looking at the replication files for *APSR*, *AJPS*, and *JOP* articles in 2018 in which the authors mention using a fixed-effects logit, I find that more than half the time the authors mean CML, rather than MLDV, even when they do not use the term "conditional" or cite Chamberlain.

3. This is not to say that random effects is never a good strategy. There are lots of cases in which analysts will have strong theoretical reasons for preferring random effects or another strategy, such as generalized estimating equations (GEE). However, the ability to compute substantive effects should not be one of them.

The Journal of Politics, volume 83, number 2. Published online March 18, 2021. https://doi.org/10.1086/709839

by assuming that the unestimated unit-specific constants are all zero (the margins command in Stata uses this assumption when applied to the CML). While this strategy provides a path toward producing substantive effects, it can lead to exaggerated effect sizes to the extent that the true constants may be far from zero. The propagation of such methods, and their recent use in top journals, suggests that there is a need to consider and clarify the available tools for producing substantive effects.[4]

In this article, I provide this clarity by considering the use of an econometric approach known as the CRE.[5] I compare the CRE to both the CML and MLDV and provide new advice to scholars about how and when these tools are useful for producing substantive effects. In doing so, I contribute to the long-running discussion of how to interpret panel data models with a binary outcome by offering new guidance on where and when to use these tools in applied research, while also providing new guidelines for producing estimates of key quantities of interest.

Both the CRE and MLDV work well in ideal conditions (long, non-rare-events panels), which suggests that the MLDV should be preferred, as it makes fewer functional form assumptions. In contrast, when either T is small or events are rare, the CRE is almost always preferred to the MLDV when it comes to estimating coefficients, unit-specific constants, predicted probabilities, and marginal effects. However, in these two situations in which the CRE is preferred, analysts are often faced with large numbers of all-zero or all-one units. These homogeneous-outcome units introduce an additional nuance in comparing theses methods: the MLDV excludes them for numeric reasons, while the CRE does not.[6] Ideally, analysts have theoretical expectations about which units are and are not of interest to their research question (e.g., politically relevant dyads in international relations). To the extent that some of these theoretically interesting units may be all-zero/one, the CRE can be used to find marginal effects, while the MLDV still faces numeric issues. In the absence of strong theory about which all-zero units are relevant, analysts may be interested in estimating effects that are conditional on units having experienced an event, which can be estimated by either the CRE or the MLDV (indeed the MLDV can only produce these conditional effects in the presence of homogeneous units). Overall, analysts should carefully consider which units are of interest to them, employ multiple methods and samples that allow them to estimate the quantities of interest to them, and discuss any differences that appear. Of the approaches considered here, only the CRE can estimate effects that include interesting homogeneous units, which enables these comparisons.[7]

Before proceeding, it is worth reiterating a point: with long, nonrare panels, the MLDV is a fine tool for producing full-sample substantive effects. As Coupé (2005) and Greene (2004) both demonstrate, with as few as 10 observations and reasonably common data (about 30% and 50% in Greene and Coupé, respectively), the MLDV is not bad. These longer panels, typically referred to as time series cross-sectional (TSCS) data, are thought to be the main format of panel data in political science (e.g., Greene 2003). However, this focus on TSCS overlooks the use of survey panels in which it is often easier (or cheaper) to have fewer waves with lots of individuals. Additionally, as both the simulations below and analysis by Cook et al. (2020) demonstrate, it takes increasingly longer panels for the MLDV to produce good estimates as events get rarer. As a result, we should be aware of how well the CRE performs in situations in which we do not have ideal data for the MLDV.

THE PROBLEMS WITH PANELS

As mentioned, the MLDV is a reasonable choice so long as the number of within-unit observations is not too small. With very short panels, the issues are well known and covered in most econometric textbooks. In the most extreme case of only two observations per unit, the MLDV estimates of the time-varying coefficients converge to twice their true value (Hsiao 1986, 161).[8] Within political science, short panels are

4. Of course the linear probability model (LPM) with fixed effects is also an option, particularly in more complicated situations such as when using instrumental variables or modeling complex dynamics. When considering the LPM in the simulations below, I find that it is a very good choice if the analyst is only interested in the average marginal effect (AME), but it is notably worse than correlated random effects (CRE) at estimating other substantive quantities (e.g., predicted probabilities). Additionally, the LPM estimator is known to be inconsistent if its predicted probabilities fall outside the unit interval (Horrace and Oazaca 2006), which is not uncommon in rare events situations. In these cases, it would be a good idea to compare the effect sizes to the methods used here in order to assess the extent of the problem.

5. While the CRE has appeared only sporadically in political science, most notably by Bell and Jones (2015) who promote its use in linear panels, the CRE has some history in economics, dating back to Mundlak (1978). Several other alternatives to the CRE, MLDV, and CML are considered in the appendixes (apps. A–J are available online).

6. The CML also removes these units, but the CML does not produce marginal effects estimates, making this issue less concerning.

7. Another approach that allows for this is a penalized maximum likelihood (PML) estimator proposed by Cook, Hays, and Franzese (2020). However, in almost every simulation considered here, the CRE performs better (see apps. B and H).

8. Note that while the coefficients will converge to twice their true value, this does not mean that the marginal effects are also inflated by a factor of two. Greene (2004) finds that while marginal effects are still biased in short-T panels, they tend to be less biased than coefficient estimates.

relatively common in survey-based research. For example, a review of top journals (*APSR*, *AJPS*, and *JOP*) since 2016 finds no shortage of short-panel surveys with a binary outcome, including Goldman (2018; $T = 2$), Goren and Chapp (2017; $T = 2$), Hale and Colton (2017; $T = 2$), Pardos-Prado and Xena (2019; $T = 4$), and Simonovits and Kézdi (2016; $T = 4$), among others.[9] While this type of data is certainly less common than TSCS, it is clearly present, and analysts working with these data still face the very open question of which tools do and do not work well in their cases.

The length issue becomes more nuanced if we consider rare events as well. In the simulations below, the number of within-unit observations needed to overcome the MLDV's issues increases as events become rarer. Standard rules of thumb suggest that the MLDV's problems disappear with 8, 10, or 20 within-group observations (e.g., Coupé 2005; Katz 2001), but these guidelines may be overly optimistic, depending on the rareness of the data-generating process. Analysis by Cook et al. (2020) confirms this point, as they find that the MLDV struggles at estimating quantities of interest even when $T = 50$ if events are rare. Their findings suggest that as events get rarer, the MLDV's issues persist with common TSCS dimensions.

A related concern with rare events data sets is that there are often many units that experience no within-unit variation in the outcome variable. The MLDV removes all of these units from the data. But what does it mean for a unit to experience no variation, and does it matter? Regarding what it means to be an all-zero unit, there are two reasons why a unit might fall into this category: it is incapable of experiencing the event or it has characteristics that make it unlikely to experience the event (i.e., these all-zero units are an artifact of only observing them for a finite number of periods).[10] For example, does the United States not experience a civil war in a particular data set because it is truly incapable (probability zero)? Or is it just because the event is unlikely (probability small but greater than zero)?

If the former is true, then it may be reasonable to remove these cases, but how do we know? This general question has motivated years of discussion among international relations scholars about the idea of "political relevance," which provides some theoretic guidance for determining which homogeneous units should be included or excluded. Theoretical guidelines like this are perhaps more satisfying than dropping all-zero units for purely technical reasons, as the MLDV does, particularly when they make up a large percentage of the data. The CRE's main value added is its ability to produce marginal effects for interesting all-zero units, and as such, it can provide estimates for either the full sample or any given theoretically relevant subset (that may or may not include some all-zero units).

Regarding whether the inclusion of all-zero units matters, the answer depends on the quantity of interest. On the one hand, these units provide no additional information on the coefficients for time-varying covariates when analysts are using the CML or MLDV.[11] On the other hand, they provide information on marginal effects. To see this, consider that the population AME of a specific variable is a function of all the members in the population, including those that never experience the event within the sample. For example, it may require a very large decrease in US military capability until a civil war becomes likely, but this is the nature of nonlinear models: some units will be in the flatter ends of the logistic S curve. The variables of interest will have smaller effects among these units, but including them is part of building an average effect estimate that applies to the full population. As a result, it is perhaps often frustrating that, while we can obtain good coefficient estimates from the MLDV, we cannot always translate them into good substantive effects that apply to the full population (or even a subpopulation of interest) when there are large numbers of all-zero units in the sample of interest. Additionally, when the number of within-unit observations is finite, any given all-zero unit may not even be from the flatter ends of the S curve. Even units that exist in the middle of the S curve (where marginal effects can be sizable) might be all zero by chance, while still being relevant to the research question. The exclusion of these units can make conditional-on-having-experienced-an-event estimates less interesting to researchers who want to make an inference about either the full population or a theoretically interesting group rather than relying on the dependent variable to determine a unit's relevance.

However, including large numbers of all-zero units without a theoretical basis may not always be the right approach either. After all, it could be the case that a large number of these units are actually irrelevant to the research question, and including them leads to attenuated effect estimates. Without a theory to determine a homogeneous unit's relevance, inferences that are conditional on a unit having experienced the

9. For unbalanced panels, T is listed as the median number of within-unit observations.

10. I frequently refer to these no-variation units as all zero, but the same discussions apply to all-one cases. The reason for this terminology is that the all-zero situation appears to be far more common in political science data, particularly with rare events.

11. The CRE's random effects framework allows for more pooling across units than the alternatives, and so it is potentially more sensitive to the inclusion or exclusion of these units.

event might be a good way to sidestep concerns about averaging in units that are irrelevant in the sense that they cannot (or are very unlikely to) experience the event. Overall, analysts may want to estimate multiple effects from multiple samples (i.e., the full sample, the MLDV sample, and one or more theoretically motivated samples), discuss any differences, and choose the quantities that best suit their research question. The CRE framework allows for estimating and comparing these quantities, while the MLDV is limited to just the conditional effect.

ESTIMATORS

Consider the binary outcome panel model

$$\begin{aligned} y_{it} &= \mathbb{I}(x_{it}'\beta + z_i'\alpha + \varepsilon_{it} > 0), \\ i &= 1, \ldots, N; \quad t = 1, \ldots, T, \end{aligned} \tag{1}$$

where ε_{it} is distributed logistic with mean zero and scale one, x_{it} is a column vector of time-varying predictors, z_i is a column vector of time-invariant predictors, and $\mathbb{I}(\cdot)$ is the indicator function. Additionally, z_i is unobserved and possibly correlated with x_{it}. Because z_i is unobserved, we replace it with an individual-level constant $c_i = z_i'\alpha$. From here on, I consider fitting and interpreting models of this form.

Perhaps the most obvious approach is to use the MLDV and estimate c_i directly. However, there are two main issues with the dummy variable approach. First, the MLDV struggles with bias when T is small. Second, if the outcomes for group i are all zero (all one), then the maximum likelihood estimate of c_i is negative (positive) infinity (Hsiao 1986, 160–61). If c_i is finite, this results in separation bias. These units end up being dropped to avoid the numeric problems of separation, but doing so limits the MLDV to only being able to consider marginal effects that are conditional on a unit having experienced the event.[12]

Conditional maximum likelihood

Chamberlain's (1980) CML estimator avoids some of the potential pitfalls of dummy variables by working around the problem. Chamberlain's insight is that the within-unit sum $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for c_i. Using the properties of sufficient statistics, Chamberlain factors each c_i out of the joint distribution of y_i, which removes it from the estimation problem.

There are three advantages of this approach: it is consistent, it is implemented in standard statistical software,[13] and it allows for arbitrary correlation between the unobserved heterogeneity and the observed independent variables. The primary drawback is that without estimates of the unit constants, it is impossible to estimate the substantive quantities that are of interest to many applied researchers.

Correlated random effects

The CRE model is another alternative. While it has a relatively long history in econometrics, it has only recently received notice in political science (Bell and Jones 2015). While a few different approaches fall under this heading, I focus on the most easily implemented and widely accessible version: Mundlak's (1978) CRE estimator.

The CRE builds on the ordinary random effects model, which assumes that the c_i's are distributed such that $\mathrm{E}[c_i|x_{i1}, \ldots, x_{iT}] = \mathrm{E}[c_i]$. In other words, this assumption states that there is no relationship between the unobserved heterogeneity and the observed covariates. When this assumption is violated, the random effects estimates of β suffer from omitted variable bias.

With the CRE, however, the independence assumption is relaxed by imposing a functional form on each c_i to explicitly relate it to the observables. But relaxing independence requires the analyst to model the relationship between x_{it} and c_i. Mundlak's CRE imposes the functional form

$$c_i = \psi_0 + \bar{x}_i'\psi + u_i, \tag{2}$$

for each i, where $\bar{x}_i = (1/T)\sum_{t=1}^{T} x_{it}$ and $u_i \sim N(0, \sigma_u^2)$. The main advantage of this approach over alternative CRE specifications is its simplicity. Mundlak's model uses a random effects logit of y_{it} on x_{it} and $\bar{x}_i$.[14] Once the random effects model is fit, the group-specific intercepts are constructed using equation (2). Note that when $\psi = 0$ we are left with the ordinary random intercepts model.

The CRE is an attractive compromise between random and fixed effects. It allows for (modeled) correlation between c_i and the observables while still estimating c_i. With these estimates, analysts can compute substantive effects. The CRE also allows for the inclusion of time-invariant characteristics in the specification of c_i. Among the estimators considered here, only the CRE can accommodate time-invariant traits.

Furthermore, by staying within the random effects framework, the model allows for more pooling of information than the MLDV, which may be helpful in a rare-events context. To see this, consider two units with similar covariate profiles

12. Whether this restriction is consequential depends on whether the homogeneous units being dropped are relevant to the data-generating process or are otherwise interesting to the analyst.

13. Stata's xtlogit with the fe option and R's survival::clogit functions.

14. As such, analysts can use prepackaged maximum likelihood routines such as Stata's xtlogit with the re option or R's lme4::glmer functions, or the model can be fit with prepackaged Bayesian routines using R's rstanarm::stan_glmer (see app. I).

in which one of them never experiences the event. With the MLDV, this unit's estimated constant would be pushed toward negative infinity, and so it cannot be included. In contrast, the CRE's functional form imposes some built-in limits on how much c_i can inflate, by tying it to the observables (and as such to units that do experience the event). This pooling is desirable when an all-zero unit is relevant to the data-generating process (i.e., the true c_i is finite). However, when truly irrelevant units dominate the data, their inclusion may attenuate the estimated effects. I return to this point below. Another drawback of the CRE is that the analyst is required to specify a functional form on the correlation, although this can be done in very general ways (as in Núñez 2017).[15]

SIMULATIONS

Consider the data-generating process:

$$y_{it} = \mathbb{I}((x_{it} + z_i)\beta + z_i\alpha + \varepsilon_{it} > 0),$$

where $x_{it} \sim N(0, 1)$, $z_i \sim N(-4, 1)$, $\beta = -2$, and ε_{it} is distributed standard logistic. Additionally, z_i is unobserved, and $x^{\star}_{it} = x_{it} + z_i$ is the only observable. This setup produces an average correlation of about 0.7 between $x^{\star}$ and z. When $\alpha = 3.25$, rare-events panel data are produced (typically less than 5% of within-group observations are one) but not when $\alpha = 2.25$ (about 35% of within-group observations are one). Beyond allowing α to vary, all N, T combinations with $T \in \{3, 5, 10, 25\}$ and $N \in \{20, 40, 60, 80, 100\}$ are considered. For each combination of α, N, and T, I generate 1,000 data sets and fit the model to the data using the CRE, MLDV, and CML.[16] Throughout, I compare the approaches using root-mean-squared error (RMSE). Additional results and simulations can be found in the appendix.

Figure 1 presents the RMSE in estimating β, while figure 2 considers the RMSE in estimating c. The top row of each figure contains the results for rare events. The first thing to note here is that the MLDV tends to do worse than all of the other estimators at estimating β in every setting, but the gap is especially pronounced in short panels and with rare events. Again, this highlights the fact that the value of T that makes the MLDV good enough is not as clear as it is in Coupé (2005) or Katz (2001) because the rareness of events affects this decision.

The next thing to note in figure 1 is the relative performance of the CRE and CML. Both do very well at estimating β, with the CRE doing slightly better than this gold-standard estimator in many settings. As N increases beyond what is presented here, these tiny differences will disappear as the CML estimate will eventually converge to the true value of β as N goes to infinity. All the estimators perform well with long-enough panels, but the CRE and CML are clearly the best choices when length is limited.

Turning to figure 2, we see the same gap between the CRE and the MLDV. This figure is a little trickier to interpret, as the MLDV only estimates c_i for heterogeneous units, and so it is only judged within the units it considers. In contrast, the CRE has the more difficult task of pinning down c_i for the homogeneous units as well. As a result, the comparisons are slightly less than straightforward. However, appendix D compares only the parameters estimated by the MLDV, and the conclusions are unchanged. Despite not having to consider the homogeneous units, the MLDV struggles with estimating c_i. Even in the ideal situation ($T = 25$ and nonrare events), it is worse at producing these estimates than the CRE.

Obtaining good point estimates is important, but our main objective is to obtain good estimates of substantive effects, such as predicted probabilities ($\hat{p}_{it}$) and marginal effects. Figure 3 shows the RMSE in estimating predicted probabilities. Again, the MLDV's predicted probabilities are only for those units it considers, while the CRE considers all units. Given the above results, it is unsurprising that the CRE dominates here. In comparing the MLDV to the CRE, we see that even in ideal conditions (large T and nonrare events) the MLDV is still notably worse at producing predicted values. As with estimating the unit-specific constant c, the CRE's performance extends to the more direct comparison that only considers the heterogeneous units considered by the MLDV (see app. D).

Perhaps the most important quantity to any applied researcher is the marginal effect that an individual covariate has on the probability that $y = 1$. One common approach is to consider the AME, given as

$$\widehat{\text{AME}} = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{T}\sum_{t=1}^{T}\Lambda(x_{it}\hat{\beta} + \hat{c}_i)(1 - \Lambda(x_{it}\hat{\beta} + \hat{c}_i))\hat{\beta},$$

where Λ is the logistic cumulative distribution function and N_k refers to the number of units considered by the model, such that $N_{\text{MLDV}} \leq N_{\text{CRE}} = N$ as the MLDV only considers

15. Furthermore, as demonstrated in app. F, this simple version of the CRE can still perform quite well when the relationship between x and z is nonlinear.

16. Intuitively, increasing T tends to reduce the proportion of all-zero units. However, many political science data sets have a large T and a high proportion of all-zero units. To gain insight into how these estimators perform in those cases, I conduct an additional Monte Carlo simulation in app. E, which has a longer panel $T = 25$ and much rarer events than considered in these simulations. The basic results presented here are unchanged: the CRE tends to be the best choice for rare events.

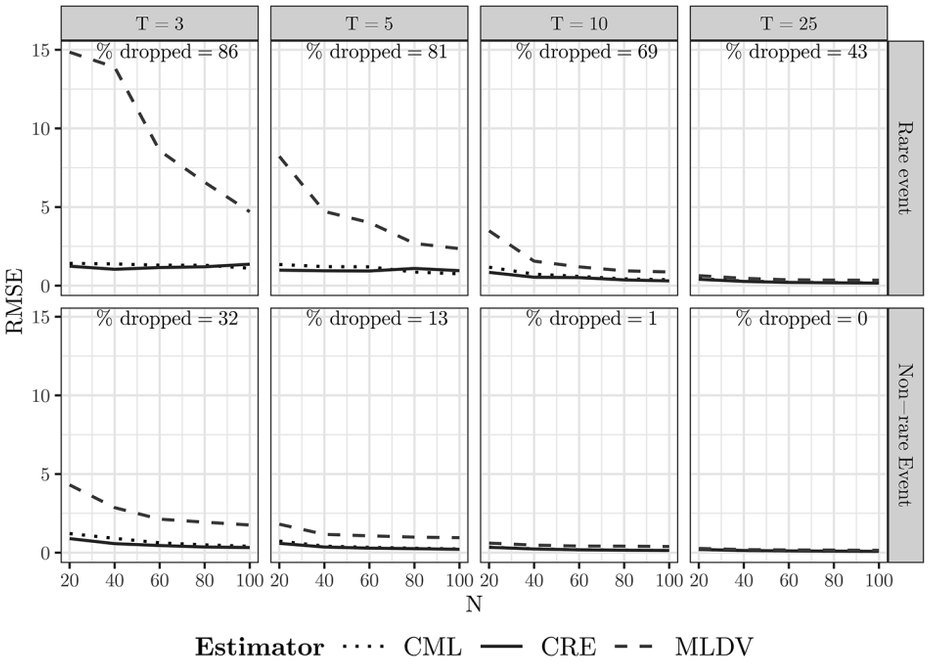

Figure 1. RMSE in estimating β. Percentage dropped refers to the average percentage of units that are dropped by the MLDV and CML (i.e., the percentage of all-zero/all-one units) across the simulated data sets within each panel.

heterogeneous units. The RMSE in these estimates is presented in figure 4.[17]

Note that the MLDV is disadvantaged here, as the AME it produces only considers the heterogeneous units, which means that it is actually a conditional average marginal effect (cAME). Nevertheless, the cAME is the only version of the AME that the MLDV produces, so we want to know how well it approximates the true AME.[18] Specifically, we are comparing these two estimates to the true AME that includes the homogeneous units. In most of the panels, the CRE is preferred for estimating this quantity. In the large T and non-rare-events settings, the approaches are effectively tied. This is where the MLDV should, and does, perform well. However, both estimators do a good job under these conditions, making the MLDV's good performance here underwhelming.

Overall, the choice between estimators appears to depend on the panel structure. When dealing with short panels, rare events, or both, the CRE tends to be the best choice (in terms of both bias and RMSE), particularly when marginal effects or predicted probabilities are of interest. As panels get longer, both estimators produce better results, but as events get rarer, panels need to be longer for the MLDV to catch up to the CRE. However, in the cases in which the estimators are roughly equal (T is big and events are not too rare), analysts may prefer the MLDV for its weaker assumption on c_i.

Inference on heterogeneous units

In the simulations above, the all-zero units are all relevant in the sense that all observations are generated with a nonzero (although sometimes still small) probability that $y_{it} = 1$. However, there are reasonable concerns that real-world data may contain some (or many) irrelevant cases. For example, in the Green, Kim, and Yoon (2001) data, considered below, there are thousands of country-dyads that never experience a dispute, and while it is not obvious that they should all be excluded from analysis, it is also not obvious that they are all of interest for computing marginal effects. In these cases, analysts may be interested in estimating either a cAME that only considers the heterogeneous units or the marginal effect for a theoretically interesting subset (e.g., politically relevant dyads) to either compliment or substitute an unconditional AME that averages over all the units. Analysts can then determine which of these quantities best answers their specific research questions.

17. Considering the marginal effect at the mean does not change any conclusions.

18. An alternative approach might be to assume that the marginal effect for the all-zero units is effectively zero and average additional zeros with the cAME to get a better estimate of the population AME. However, in many cases, particularly when T is small, this may not be a reasonable assumption. Doing so in the above Monte Carlos tends to improve on the MLDV's AME estimate, most notably when events are rare and the assumption is more likely satisfied, but this improved performance is still almost always worse than using the CRE.

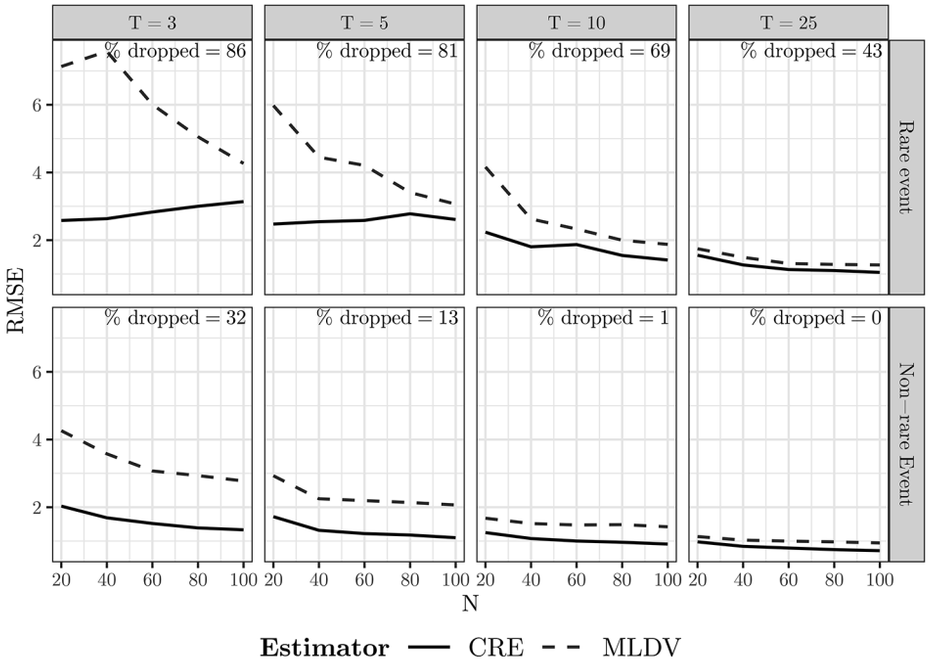

Figure 2. RMSE in estimating c (averaged over c_i). Percentage dropped refers to the average percentage of units that are dropped by the MLDV (i.e., the percentage of all-zero/all-one units) across the simulated data sets within each panel.

Estimating the cAME is straightforward for the MLDV, as the cAME is the AME for the subsample used by the MLDV. For the CRE, however, there are two different ways to produce subgroup estimates like the cAME. One approach is to fit the CRE to the full sample and then only produce effects for the heterogeneous units.[19] Another approach is to fit the CRE using only the heterogeneous units and produce the cAME.[20] While the differences in these approaches may seem minor, these changes to the underlying sample can notably affect results, as the CRE's random effects framework allows for some pooling of information.

Overall, the choice between the full-sample CRE or the restricted CRE (rCRE) depends on an untestable assumption about the relative balance of (ir)relevant units.[21] To the extent to which the omitted units are relevant to the data-generating process, the CRE provides the better estimates of the cAME than either the rCRE or the MLDV (as shown in app. D), which is the result of the CRE's ability to pool some information from relevant units to improve the estimates for the units of interest. However, the CRE's estimates of the cAME get worse as irrelevant units dominate the data. To explore the extent of this problem, I consider two additional and extreme case in the appendixes to test the limits of the full-sample CRE. In appendix E, the CRE substantially outperforms both the MLDV and the rCRE at estimating the cAME even when most of the data are practically irrelevant, which is to say most units are extremely unlikely to ever experience the event (on average, $\Pr(y_{it} = 1) < 0.0004$ for 90% of observations within each simulated data set).[22] Likewise in appendix D, I consider a Monte Carlo where 90% of the data are totally irrelevant ($c_i = -\infty$). As expected, the full-sample CRE performs relatively worse here compared to other situations, but it is still either the best choice or only slightly worse than the alternatives at finding the cAME. As such, using the CRE to estimate the cAME appears to be a safe choice in most situations. Overall, analysts should consider which quantities of interest provide the best approach for their data and research questions and select the estimator(s) accordingly. When events are relatively common or the cAME is of interest, the MLDV is frequently a good choice, so long as T is not too small, but to the extent that some or all of the all-zero units are of interest, the CRE is better.

19. This approach is easily implemented in R by using the data option in the margins::margins command (glmer models are supported as of version 0.3.22).

20. Like the MLDV, the cAME is the AME for this approach as only the MLDV subsample is used.

21. From here on, the term CRE will be used to refer to the full-sample CRE, while the rCRE will be used to refer to using the heterogeneous subsample. Any other subsample approaches will be denoted separately.

22. This particular experiment has the added benefit of providing some insight into how the estimators perform in data that closely resemble the Green et al. (2001) data considered below in terms of the overall rareness of events.

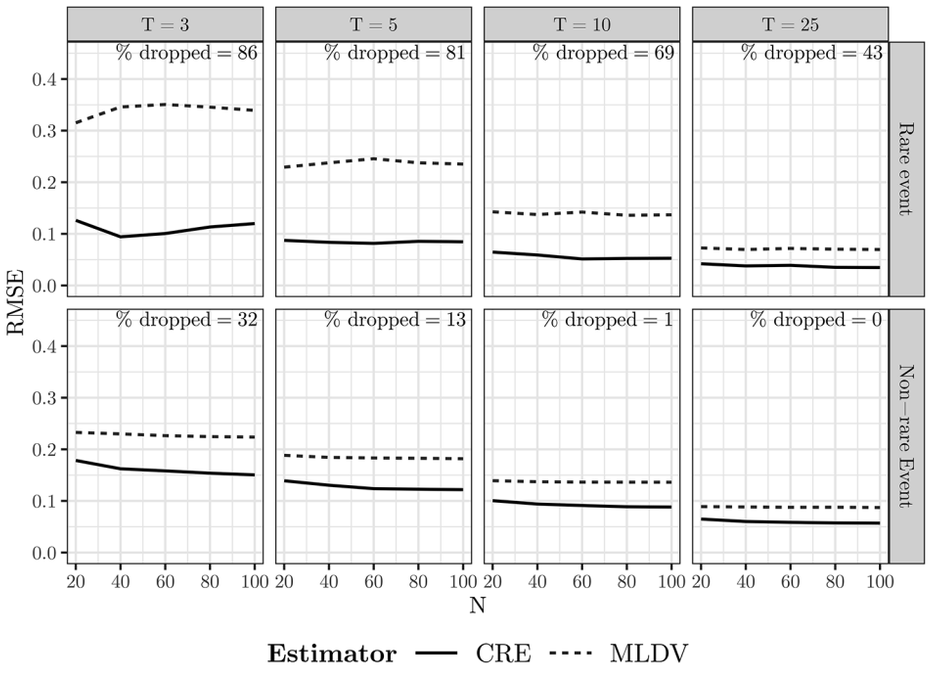

Figure 3. RMSE in estimating predicted probabilities. Percentage dropped refers to the average percentage of units that are dropped by the MLDV (i.e., the percentage of all-zero/all-one units) across the simulated data sets within each panel.

Before considering the empirical applications, I report the following conclusions.

1. When either T is small or events are rare, the CRE outperforms the MLDV across a wide range of settings.
2. When T is large and events are sufficiently common, the MLDV is a good choice for estimating all quantities of interest.
3. With lots of all-zero units, analysts face a theoretic choice. To the extent that they believe some of these

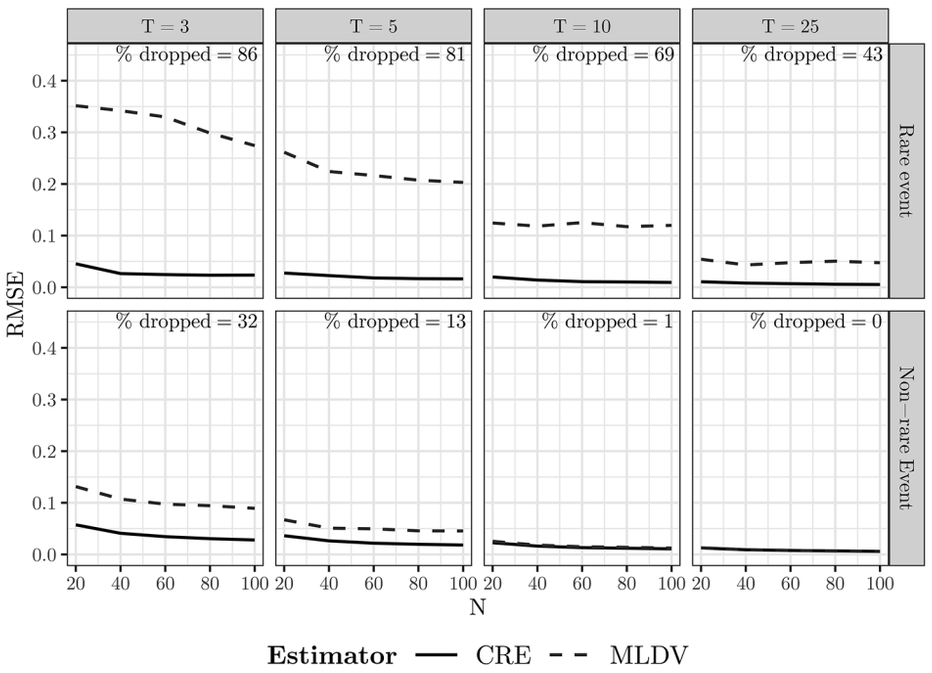

Figure 4. RMSE in estimating the AME. Percentage dropped refers to the average percentage of units that are dropped by the MLDV (i.e., the percentage of all-zero/all-one units) across the simulated data sets within each panel.

units are relevant to their study, the full-sample CRE provides very good estimates of the AME, cAME, and other theoretically interesting subsample effects. Restricted sample estimators (MLDV or rCRE) may be preferred if analysts believe that the omitted units are overwhelmingly irrelevant.

Additional simulations

In the appendixes, I consider several additional simulations that provide insight into the performance of these estimators. In appendix A, I present the bias results for the above simulations and consider alternative estimators: Cook et al.'s (2020) PML estimator; Beck's (2015) two-step; ordinary random effects; GEE; and the LPM. The CRE is still the best choice (or very close to the best) for most quantities of interest. The LPM is worth discussing given its advantages over the various nonlinear estimators considered here when it comes to more complicated data problems such as instrumental variables and dynamic models. As mentioned above, the LPM does a fine job at identifying the AME, but it is an order of magnitude worse at producing predicted probabilities than the CRE. Overall, while the CRE and LPM both do a very good job at finding the AME, the CRE does better at other important quantities.

Two other potential concerns regarding the CRE are addressed in appendixes F and G. In the former, I change the data-generating process such that x_{it} and z_i are nonlinearly related. This experiment is designed to explore the sensitivity of the CRE's functional form assumption. As expected, this change worsens the CRE's overall performance, but it still tends to be a better choice than the MLDV for estimating substantive quantities when the data are rare or T is small. In the latter, I explore the issue of rarely changing covariates and the effect this common problem has on the CRE, by restricting the within-unit standard deviation of x_{it} to be 0.5. The above recommendations are unchanged by introducing these various difficulties.

APPLICATIONS

I consider three applications to illustrate the strengths and weakness of the above approaches. The first is a non-rare-events panel with a large number of within-group observations on the relationship between remittances and protest behavior, by Escribà-Folch et al. (2018). I include this study to illustrate that these tools will agree with each other under ideal conditions but also differ from other solutions in the literature. In this case, the MLDV might be weakly preferred when we consider the similarity in results and the flexibility in functional form.

The next example considers short ($T = 2$) survey data from Goldman (2018). Here, both the rCRE and the MLDV produce estimated effects that are larger than the cAME or AME produced by the full-sample CRE. Overall, the full-sample CRE may be preferred to both the rCRE and MLDV in this example for two reasons. First, the rCRE and MLDV are nearly identical in both point estimates and effects, which, given the shortness of the panel, casts suspicion on the rCRE results. Second, given that the MLDV is known to be problematic with short panels, it is unlikely that its estimated cAME is a quantity of interest. In contrast, the full-sample CRE performed well in the short-panel simulations and can estimate either the cAME, if the analyst is only interested in heterogeneous units, or the AME, which applies to the entire population of interest.

The final example is a rare-events panel on democratic peace by Green et al. (2001). In the rare-events example, the MLDV produces cAME estimates that are extraordinarily large compared to the CRE's estimates of the full-sample AMEs, which matches both the above Monte Carlo evidence and results from Cook et al. (2020). As result, this is a case in which decisions about the all-zero units result in large substantive differences. With so many all-zero units, its not clear that either the cAME or the full-sample AME will be the most interesting substantive quantity. Instead, international relations scholars are often interested in politically relevant dyads, which is a subset of the full sample that contains some, but not all, of the all-zero units. The CRE, but not the MLDV, provides a useful framework for considering the marginal effects on this theoretically motivated subset. As a result, this example demonstrates the CRE's range of uses for applied researchers who may want to consider all, some, or none of the all-zero units. Overall, this example demonstrates the CRE's ability to consider multiple quantities of interest. Analysts can then choose a strategy that makes the most sense for their research question and data.

Together, all three applications reinforce three major points from the Monte Carlos. First, the MLDV is a good choice when conditions are ideal. Second, the CRE tends to do very well under a variety of conditions and is typically the best choice in both short panels and rare-events settings, particularly when there are multiple quantities of interest. Third, it is important to compare estimates across models and consider why any differences emerge; the CRE allows for these comparisons.

Long panel from Escribà-Folch, Meseguer, and Wright

I begin with Escribà-Folch et al. (2018), who consider the effect that remittances have on protest behavior. Here, I focus on their individual-level analysis in sub-Saharan Africa. In this panel, there are 614 districts (N), with a median of 15 individuals within each district (T), and within each district an average of 12.5% of individuals have protested. As a result we have a

longer panel with a relatively common event. In this situation, we expect that the CRE, rCRE, MLDV, and CML will all produce similar estimates. This case is included to show two things: it is important to consider a method that estimates c when producing marginal effects, and the different estimators work as expected in ideal conditions.

The dependent variable is from the 2008 Afrobarometer and records whether an individual attended a demonstration or march. The first hypothesis of interest is, Does receiving remittances lead to an increased probability of attending a demonstration or march? Remittances are measured on a 0–5 scale that records how often an individual received a remittance. Escribà-Folch et al. (2018) are also interested in whether this effect is attenuated by how progovernment a specific district is. They measure this by combining three Afrobarometer criteria: trust in the president, trust in the ruling party, and presidential performance. The empirical model of interest is

$$\Pr(\text{protest}_{it} = 1) = \Lambda(\text{remit}_{it}\beta_1 + (\text{remit}_{it} \times \text{progovernment}_i)\beta_2 + x_{it}'\gamma + c_i),$$

where x_{it} is a vector of controls, i indexes geographical districts, and t indexes individuals.

The control variables include whether individuals use a cell phone regularly, as well as their age (logged), education, wealth, sex, and employment status. Escribà-Folch et al. (2018) fit the model using the CML. Table 1 reports their CML results, as well as results from the CRE, rCRE, and MLDV. All four models overwhelmingly agree in their estimates of the coefficients and standard errors.

Turning to marginal effects, I follow Escribà-Folch et al. (2018) and estimate the AME of raising remittances from 0 to 5.[23] This is given by

$$\widehat{\text{AME}}_{\text{remit}} = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{T_i}\sum_{t=1}^{T_i}(\Lambda(\text{remit}_{it}\hat{\beta}_1 + (\text{remit}_{it} \times \text{progovernment}_i)\hat{\beta}_2 + x_{it}'\hat{\gamma} + \hat{c}_i) \times (1 - \Lambda(\text{remit}_{it}\hat{\beta}_1 + (\text{remit}_{it} \times \text{progovernment}_i)\hat{\beta}_2 + x_{it}'\hat{\gamma} + \hat{c}_i)) \times (\hat{\beta}_1 + \text{progovernment}_i\hat{\beta}_2)) \times 5,$$

where N_k again refers to the number of units considered by the model.

23. It is worth pointing out that Escribà-Folch et al. (2018) do their best to overcome the shortcomings of the CML and produce an AME estimate by assuming $c_i = 0$ for all i. Their effort speaks to the importance of exploring and promoting tools like the CRE, as it allows an analyst to estimate effects without such a strong assumption. Following their code, I find cAMEs of about 0.14 and −0.01 for the two cases considered in fig. 5, the former is about twice as large as any of the other methods. For more information, see app. J.

Table 1. Replicating Escribà-Folch, Meseguer, and Wright (2018): Coefficient Estimates

	CML	MLDV	rCRE	CRE
Remittances	.16	.17	.16	.16
	(.07)	(.07)	(.07)	(.07)
Remit × progovernment	−.22	−.22	−.21	−.22
	(.13)	(.14)	(.13)	(.13)
Cell phone	.31	.33	.30	.30
	(.09)	(.09)	(.09)	(.09)
Age (log)	−.17	−.19	−.18	−.18
	(.10)	(.10)	(.10)	(.10)
Education	.09	.09	.09	.09
	(.02)	(.02)	(.02)	(.02)
Wealth	.29	.31	.30	.30
	(.13)	(.13)	(.13)	(.13)
Male	.20	.21	.20	.20
	(.07)	(.07)	(.07)	(.07)
Employment	.04	.05	.04	.04
	(.05)	(.05)	(.05)	(.05)
Travel	.04	.05	.05	.05
	(.07)	(.07)	(.07)	(.07)
Groups	469	469	469	614
Observations	8,626	8,626	8,626	10,295

Note. Dependent variable = protest. Standard errors in parentheses. Coefficients on the group means (CRE) and unit constants (MLDV) are omitted.

There are three things to note from figure 5. First, in this long T, nonrare example, the MLDV's cAME is roughly similar (if larger) than the AME estimated by the CRE, which matches the Monte Carlo evidence for best case data. Second, the assumption used by Escribà-Folch et al. (2018) inflates the AME. The CRE and MLDV roughly agree with each other as to the magnitude and strength of the effect (0.06 and 0.07, respectively), which is less than half of the effect estimated by Escribà-Folch et al.'s approach. This decrease in magnitude speaks to the importance of including c_i in estimating effects. Third, the CRE's cAME estimate is a little smaller than either the MLDV or rCRE. This difference can be because either the excluded units are truly irrelevant (CRE's cAME is likely worse) or the excluded units are relevant (the rCRE and MLDV are likely worse). As mentioned above, this is an untestable assumption that requires theoretical reasons for either including or excluding these units. However, the differences across the cAME estimates are definitely minor (less than a 10% difference), and the CRE, rCRE, and MLDV all appear to perform well in this example. If the full sample is of theoretic interest, then the CRE may be preferred, but if analysts are worried about the CRE's functional form assumption, then the MLDV may be preferred,

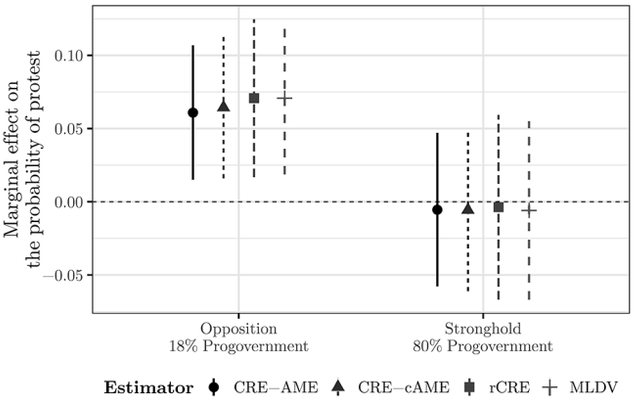

Figure 5. Average marginal effects and 95% confidence intervals of remittances on the probability of protesting by progovernment sentiment.

with the cAME offering a good approximation of the AME in this best case scenario.

Short panel from Goldman

Having satisfied ourselves that everything works as expected in the Escribà-Folch et al. (2018) case (relatively common events with a large number of within-group observations), we now look at a case with a short ($T = 2$) panel. Here, we are examining the role of gender bias in the 2008 Democratic presidential primary. The full survey includes five waves, but only waves 1 and 3 address the question of interest. In this case, the rCRE and the full-sample CRE produce notably different estimates for both the coefficients and cAMEs, and the cAMEs are notably different from the estimated full-sample AMEs. Interestingly, the rCRE is nearly identical to the known-to-be-biased MLDV. As a result, the full-sample CRE is perhaps the preferred modeling choice in this example, even if we are interested in the cAME. These differences highlight why it is important to consider multiple estimators and samples when possible.

As Goldman (2018) notes, the issue of gender bias in politics is contentious, with plenty of work finding evidence both for and against double standards for women. He takes on this question by considering whether (perceived) gender favoritism in a candidate affects a voter's willingness to support a female candidate. He argues that voters who fear gender favoritism will expect women politicians to favor so-called women's issues and to implement policies biased against men. In this framework, gender bias is about the expected policies rather than perceptions about how women should act in the political sphere.

He tests his theory that a fear of gender favoritism leads to gender discrimination at the polls using the National Annenberg Election Study internet panel survey that was conducted in five waves from fall 2007 to winter 2009. The fear of gender favoritism is measured using the four questions about perceived gender favoritism in candidates. Specifically, the survey asks the extent to which the individual believes female officeholders are biased toward women over men on issues of (1) favoring women applicants over men when providing government jobs, (2) promoting educational programs that benefit girls at the expense of boys, (3) supporting spending that favors women, and (4) focusing on so-called women's issues. Respondents were asked the extent to which they agreed with these four statements on a five-point scale, and Goldman averages their responses to create his measure of perceived favoritism.

This gender favoritism measure is the main variable of interest. Additional controls include a similar measure for perceived racial favoritism by black officeholders, the strength of their partisan identification, whether they were contacted by a campaign, their political interest, how viable they think Clinton is as a candidate, whether they rate their ideology as closer to Clinton or a competitor (either John Edwards or Barack Obama), and finally which candidate (Clinton or Edwards/Obama) is most in agreement with (or acceptable to) them on six different issues.[24] This last value is a combination of various survey questions and is coded 1 if the individual is measured as closer to Clinton, 0 if the individual is closer to Obama or Edwards, and 0.5 if there is a tie. Our model of interest focuses on what explains each individual's top choice over the course of the 2008 Democratic primary, while controlling for the unobserved and fixed differences among individuals. Table 2 reports the results for the CML, MLDV, rCRE, and CRE.

It is clear from looking at table 2 that the MLDV and the rCRE are identical to each other but different from the other two approaches. While the CRE results are relatively close to the CML estimates (estimates are mostly within 1 standard error), the MLDV and rCRE coefficient estimates are substantially further away and uniformly larger in magnitude. In many cases, we see the expected MLDV result where the regression coefficients are nearly twice what the CML produces. That the rCRE exhibits the same trend is worrisome in this example and suggests that this may be a case in which there are relevant units being excluded by restricting ourselves to the MLDV sample. The CRE's ability to use information from these units can help it find estimates that are closer in magnitude to the consistent estimates produced by the CML.

Turning our attention to the marginal effects, we continue to see differences. In figure 6, we see that the MLDV and rCRE continue to be identical, while also producing the largest results. For some variables the cAMEs are roughly similar across

24. The issues are the economy, health care, Iraq, immigration, trade, and homeland security.

Table 2. Replicating Goldman: Coefficient Estimates

	CML	MLDV	rCRE	CRE
Perception of gender favoritism	−.84	−1.68	−1.68	−1.24
	(.55)	(.78)	(.78)	(.52)
Perception of racial favoritism	.80	1.59	1.59	1.43
	(.52)	(.74)	(.74)	(.46)
Strength of party ID	1.21	2.43	2.43	1.31
	(1.04)	(1.47)	(1.47)	(1.02)
Contacted by campaign	.09	.19	.19	.36
	(.55)	(.77)	(.77)	(.48)
Political interest	−.16	−.31	−.31	−.06
	(.51)	(.72)	(.72)	(.47)
Viable candidate	.97	1.95	1.95	1.01
	(.15)	(.22)	(.22)	(.14)
Issue agreement	1.69	3.39	3.39	1.56
	(.25)	(.36)	(.36)	(.21)
Ideological agreement	5.99	11.98	11.98	6.90
	(1.31)	(1.85)	(1.85)	(1.10)
Observations	806	806	806	2,990
Individuals	403	403	403	1,495

Note. Dependent variable = support for Clinton. Standard errors in parentheses. Only coefficients common to all models are included. Coefficients on the group means (CRE) and unit constants (MLDV) are omitted.

the three approaches to estimating this quantity, while in others, most notably the self-perceptions of candidate viability, issue agreement, and ideological agreement, the differences are very noticeable (on average the MLDV/rCRE produce cAME estimates that are 1.65 times as large as the CRE's cAME). Moving to the main variable of interest (perceptions of gender favoritism), both the MLDV and the rCRE suggest that conditional on having changed his vote, increasing a male Democrat's belief that female politicians are gender biased against men leads him to be about 27 percentage points less likely to support Hillary Clinton, on average. This is a remarkably strong effect, but it appears to be driven by the modeling choice. Changing our focus to the CRE, we still see a negative cAME, but it is a smaller decrease of about 23 percentage points (or about 11% smaller than the MLDV/rCRE estimate).

Which estimator is preferred if the cAME is of interest? As mentioned, above, the answer depends on an untestable assumption about the relevance of the homogeneous units, but so long as there are not too many totally irrelevant people who change their vote with probability zero (i.e., intercepts of positive or negative infinity), the full-sample CRE benefits from being able to use information in the homogeneous units. In this particular case, the equivalence of the MLDV and rCRE casts additional suspicion on the CRE, as the MLDV is known to be problematic when $T = 2$.

However, it is not obvious that the cAME is the main quantity of interest for this specific research question, which is whether perceptions of gender favoritism affect whether male Democrats vote for Clinton, not just male Democrats who happened to change their votes. In small-T settings like this, it is likely that many of the homogeneous individuals are relevant and interesting to the researcher. Indeed, among the "never Clinton" voters the average predicted probability of voting for Clinton is about 0.24, suggesting that many of these voters might be relevant to the study and many might become heterogeneous units if observed for more periods. When considering the full sample of male Democrats, we find that the marginal effect of increasing gender bias is now, on average, a much smaller 14 percentage point decrease in the likelihood of supporting Clinton. The ability to estimate the full-sample AME when it is of interest is the main value added of the CRE in this example. To the extent that we are interested in making an inference about all male Democrats rather than only male Democrats who changed their votes, analysts have to choose between using the MLDV's cAME as an approximation of this quantity of interest or using the CRE to estimate it directly. In contrast to the Escribà-Folch et al. (2018) example, the MLDV's cAME estimate does not

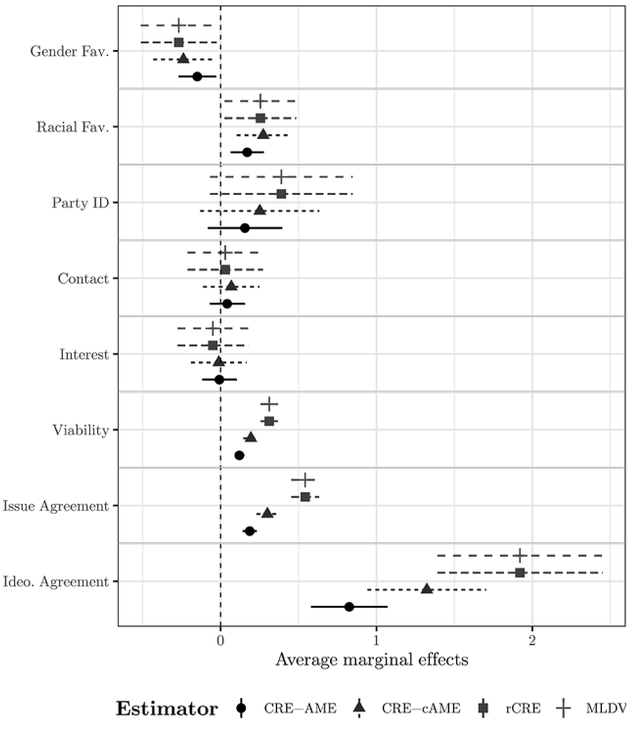

Figure 6. Average marginal effect on support for Clinton

roughly match the AME estimated by the CRE, making this choice is consequential.

Rare events panel from Green, Kim, and Yoon

I now consider a rare-events example by reproducing the interstate dispute analysis from Green et al. (2001), where N is 3,075 country dyads. The dependent variable is the onset of a militarized interstate dispute, and the variables of interest are joint democracy (measured as the lowest polity score in dyad i in year t), a dummy for contiguous states,[25] the logged capability ratio (higher to lower), an indicator for whether the dyad contains allies, the three-year average growth rate in gross domestic product (GDP) per capita (dyadic minimum), and the bilateral trade-to-GDP ratio (dyadic minimum).

The data are organized into an unbalanced panel of dyad-years with 3,075 dyads from 1951 to 1992. On average, dyads are observed for $T = 30$ years, with a range of 20–42 years, with a total of 93,755 dyad-year observations. Only 198 dyads ever engage in an interstate dispute, so the CML, MLDV, and rCRE only consider 6,353 observations, or about 7% of the total data, while the CRE is fit to the full data set. However, international relations has long been interested in a theoretically interesting subsample known as "politically relevant dyads" (Lemke and Reed 2001; Maoz and Russett 1993). Political relevance is usually defined as any dyad that contains at least one major power or a pair of contiguous countries (Lemke and Reed 2001, 127). This is typically thought of as the set of dyads where interaction is plausible, and as such, it helps define a set of relevant all-zero units. Many studies in international conflict present results for both all dyads and politically relevant dyads (e.g., Chatagnier and Kavakli 2017), and I follow this here, by including a politically relevant model (pCRE) that considers all the heterogeneous units plus the set of politically relevant all-zero units.

The first data column in table 3 contains the CML results, which exactly match those reported by Green et al. (2001). The remaining columns contain the MLDV, rCRE, pCRE, and CRE estimates, for which almost nothing changes in the estimates of β. However, major differences emerge when we consider marginal effects. In this case, I consider three quantities of interest: the full-sample AME (CRE), politically relevant AME (CRE and pCRE), and the cAME (CRE, rCRE, and MLDV).

Figure 7 shows these potential quantities of interest. As in Cook et al. (2020), we see that the MLDV produces effects that are notably larger than an AME estimated from the full sample, which means that choices about the all-zero units are consequential. Similar to the above examples, the MLDV estimates of the cAME are an average of 21% larger than the CRE estimates but only about 3% larger, on average, than the rCRE estimates.[26] Unlike in the Goldman example, none of the estimated cAMEs differ by an order of magnitude.

Whether the cAME is of interest here depends on an analyst's specific research question, and the best way to estimate it will depend on how prevalent truly irrelevant units are in the data.[27] With just over 90% of the data being all-zero units, there are real questions about the (ir)relevance of these homogeneous units. However, to the extent that the MLDV subsample likely omits some relevant units, the cAME not only gets worse as an approximation of whatever the true marginal effect of interest is, but the MLDV and the rCRE tend to get worse at estimating it.

Table 3. Replicating Green, Kim, and Yoon's (2001) Democratic Peace Analysis: Coefficient Estimates

	CML	MLDV	rCRE	pCRE	CRE
Min democracy	−.003	−.003	−.003	−.003	−.003
	(.01)	(.01)	(.01)	(.01)	(.01)
Min trade	−.07	−.07	−.07	−.06	−.06
	(.19)	(.19)	(.19)	(.20)	(.18)
Contiguity	1.90	1.99	1.78	1.91	2.34
	(.34)	(.35)	(.31)	(.33)	(.37)
Capability ratio	.39	.40	.39	.40	.40
	(.14)	(.14)	(.14)	(.14)	(.14)
Min growth	−.06	−.06	−.06	−.06	−.06
	(.01)	(.01)	(.01)	(.01)	(.01)
Ally	−1.07	−1.11	−.92	−.97	−1.18
	(.43)	(.44)	(.41)	(.41)	(.44)
Groups	198	198	198	643	3,075
Observations	6,353	6,353	6,353	20,563	93,755

Note. Dependent variable = dispute onset. Standard errors in parentheses. Only coefficients common to all models are included. Coefficients on the group means (CRE) and unit constants (MLDV) are omitted.

25. In 25 dyads, contiguity varies.

26. The large differences between the MLDV and CRE estimates of the cAME is mostly driven by the estimated effect of democracy. However, if we exclude this outlier, then the MLDV effects are still about 15% larger, on average.

27. Of particular interest here is the experiment in app. E, which focuses on extremely rare events with a large number of all-zero units and the task of estimating the cAME. In that experiment, the full-sample CRE has, on average, a very slight bias toward zero when estimating the constants, while the MLDV and rCRE are more biased in the same direction. As a result, all three estimates of the cAME trend away from zero, with the CRE being much less affected than the other two even when most units are practically irrelevant (large, negative, but finite unit constants). These trends match the results in this example and provide some evidence that the full-sample CRE may be a safe choice for these quantities of interest.

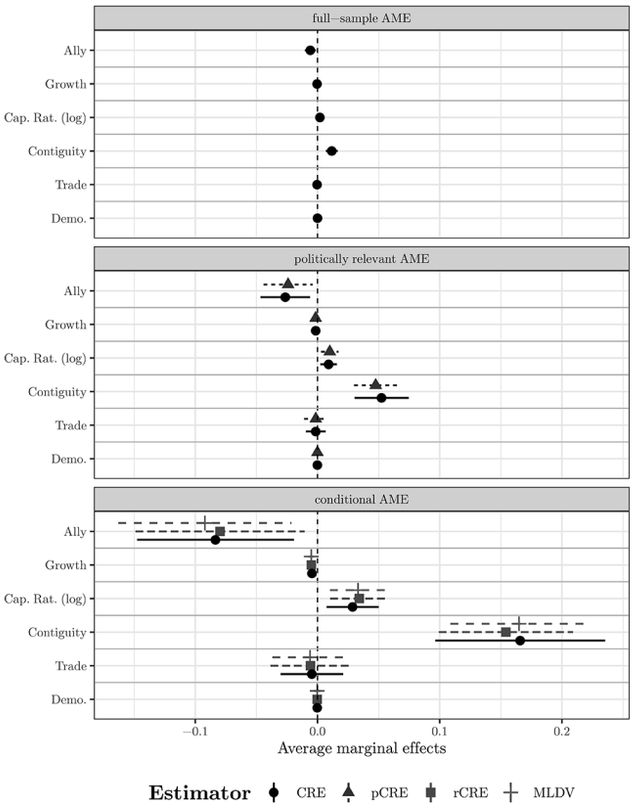

Figure 7. Average marginal effects for Green, Kim, and Yoon's (2001) dispute data.

Similarly, it may also be unlikely that the full-sample AME will always be of interest in cases like this. After all, if there are a thousands of all-zero units, many are probably irrelevant in the sense that they will almost certainly never go to war within any time frame of interest. This returns us to the concept of politically relevant dyads. As mentioned, many conflict researchers compare full-sample models to politically relevant models, and the CRE framework allows for this in ways that the MLDV does not.

In looking at these two sections of figure 7, three things stand out. First, both the full-sample CRE and the pCRE produce similar estimates of the politically relevant AME (differences are less than 10%).[28] Second, the politically relevant AME tends to be between the full-sample and conditional AMEs. Given that its construction, in this example, focused on which all-zero units to include, this middle ground is expected. Third, the CRE framework provides an easy-to-use tool for considering many different quantities of interest, and they can all be found within the context of a single fitted model (the full-sample CRE).

The final choice among the various methods and samples comes down to two factors: the quantity of interest and expectations about how many/which units are (ir)relevant. When only the heterogeneous units are of interest to the analyst, the cAME can be estimated in multiple ways, with the choice among the MLDV, rCRE, and CRE depending on the analyst's beliefs about how many of the omitted units are truly zero probability. But when there are theoretically important all-zero units, then the CRE framework (either a full-sample or restricted model depending on the theory) should be the main tool for analysts interested in substantive effects, as the MLDV's technical limitations preclude the inclusions of these units. Overall, analysts should always look for any differences between the approaches, consider why these differences appear (e.g., short T, rare events, or too many irrelevant units), and choose an approach accordingly.

CONCLUSION

In this article, I contribute to the long-running debate within political methodology regarding the tools for estimating substantive effects from binary outcome panel data. Specifically, I compare a CRE approach to the traditional methods (CML and MLDV) and show that it offers real promise over these and other existing strategies. When faced with the task of analyzing binary outcomes, I offer the following guidance:

1. With small-T or rare-events panels, CRE tends to be the best choice for estimating β, predicted probabilities, and a full-sample AME.
2. For large-T panels, the MLDV works well enough that it should be preferred to the CRE because of the MLDV's weaker functional form assumptions (so long as events are not too rare or there are not a lot of potentially relevant or theoretically interesting homogeneous units).
3. When considering data with a larger number of homogeneous units, analysts may be concerned about including potentially irrelevant (probability zero) units. To the extent that there are some or many relevant homogeneous (nonzero probability) units, the full-sample CRE tends to be a safe choice for estimating the full-sample AME, the cAME, or any other quantity. However, if the homogeneous units are believed to be overwhelmingly irrelevant, then a restricted-sample estimator like the rCRE may be a preferred approach to finding the cAME. Additionally, the CRE framework can allow for other theoretical definitions

28. As before, the choice between the full-sample CRE and the restricted pCRE depends on the analyst's beliefs about how many of the "politically irrelevant" all-zero dyads are truly irrelevant.

of relevance (e.g., political relevance) that retain some theoretically relevant all-zero units. This flexibility in considering all, some, or none of the homogeneous units is the CRE's main advantage over the MLDV.

4. Overall, analysts should compute their substantive effects of interest using multiple samples when possible, discuss any differences, and choose an approach that best suits their research question and their beliefs about the relevance of the homogeneous units in their data. As demonstrated most directly in the Goldman example, the partial pooling that occurs within the CRE can make sampling choices consequential.

The analysis above is clearly not the last word in this discussion. As mentioned, Mundlak's CRE is not the only CRE model. Other approaches offer a more general functional form than Mundlak's but at the cost of model complexity and computational feasibility. For example, Núñez (2017) offers a very flexible model that relies on basis expansion and penalized regression. Such an approach minimizes the CRE's functional form assumption, but it burdens analysts with increased computation and interpretation complexities. Furthermore, Mundlak's CRE seems to work well in practice and is easy to implement. Future analysis comparing these more advanced CREs may inform the comparison further but should explicitly consider the added computational expense.

Additionally, while the penalized likelihood methods proposed by Cook et al. (2020) tend to perform worse than the CRE (as shown in the appendixes), they may still offer a path forward in the debate about estimating substantive effects. For example, Rainey (2016) recommends simulating multiple partial prior distributions to ensure reasonable results. Future work should consider how adjusting the prior penalty can improve the penalized likelihood approach, particularly with rare events. Such improvements in the penalty may also lead the method to outperform the CRE.

ACKNOWLEDGMENTS

Thanks to Scott Cook, Michael Gibilisco, Jacob Montgomery, Ryan Kennedy, Yannis Vassiliadis, the editor Tom Clark, and four anonymous referees for comments and suggestions. This article also benefits from participants at the 2019 Texas Methods Conference and the European Political Science Association's annual meeting. Any errors are my own.

REFERENCES

Beck, Nathaniel. 2015. "Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effects: What Are the Issues?" Unpublished manuscript.

Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3 (1): 133–53.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225–38.

Chatagnier, J. Tyson, and Kerim Can Kavakli. 2017. "From Economic Competition to Military Combat: Export Similarity and Economic Conflict." *Journal of Conflict Resolution* 61 (7): 1510–36.

Cook, Scott J., Jude C. Hays, and Robert J. Franzese. 2020. "Fixed Effects in Rare Events Data: A Penalized Maximum Likelihood Solution." *Political Science Research and Methods* 8 (1): 92–105.

Coupé, Tom. 2005. "Bias in Conditional and Unconditional Fixed Effects Logit Estimation: A Correction." *Political Analysis* 13:292–95.

Escribà-Folch, Abel, Covadonga Meseguer, and Joseph Wright. 2018. "Remittances and Protest in Dictatorships." *American Journal of Political Science* 62 (4): 889–904.

Goldman, Seth K. 2018. "Fear of Gender Favoritism and Vote Choice during the 2008 Presidential Primaries." *Journal of Politics* 80 (3): 786–99.

Goren, Paul, and Christopher Chapp. 2017. "Moral Power: How Public Opinion on Culture War Issues Shapes Partisan Predispositions and Religious Orientations." *American Political Science Review* 111 (1): 110–28.

Green, Donald P., Soo Yeon Kim, and David H. Yoon. 2001. "Dirty Pool." *International Organization* 55 (2): 441–68.

Greene, William. 2004. "The Behavior of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *Econometrics Journal* 7:98–119.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Pearson Prentice-Hall.

Hafner-Burton, Emilie M., Susan D. Hyde, and Ryan S. Jablonski. 2013. "When Do Governments Resort to Election Violence?" *British Journal of Political Science* 44:149–79.

Hale, Henry E., and Timothy J. Colton. 2017. "Who Defects? Unpacking a Defection Cascade from Russia's Dominant Party, 2008–2012." *American Political Science Review* 111 (2): 322–37.

Horrace, William C., and Ronald L. Oazaca. 2006. "Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model." *Economic Letters* 90 (3): 321–27.

Hsiao, Cheng. 1986. *Analysis of Panel Data*. 1st ed. Cambridge: Cambridge University Press.

Katz, Ethan. 2001. "Bias in Conditional and Unconditional Fixed Effects Estimation." *Political Analysis* 9 (4): 379–84.

King, Gary. 2001. "Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data." *International Organization* 55 (2): 497–507.

Lemke, Douglas, and William Reed. 2001. "The Relevance of Politically Relevant Dyads." *Journal of Conflict Resolution* 45 (1): 126–44.

Maoz, Zeev, and Bruce M. Russett. 1993. "Normative and Structural Causes of Democratic Peace, 1946–1986." *American Political Science Review* 87 (3): 624–38.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46 (1): 69–85.

Núñez, Lucas. 2017. "Partial Effects for Binary Outcome Models with Unobserved Heterogeneity." Unpublished manuscript.

Pardos-Prado, Sergi, and Carla Xena. 2019. "Skill Specificity and Attitudes toward Immigration." *American Journal of Political Science* 63 (2): 286–304.

Rainey, Carlisle. 2016. "Dealing with Separation in Logistic Regression Models." *Political Analysis* 24:339–55.

Simonovits, Gábor, and Gábor Kézdi. 2016. "Economic Hardship Triggers Identification with Disadvantaged Minorities." *Journal of Politics* 78 (3): 882–92.