Assessing Sample Selection Bias in Fixed Effects Binary Outcome Models

Lucas Núñez *

Schar School of Policy and Government

George Mason University

May 25, 2023

Abstract

Many estimators that account for unobserved heterogeneity in binary outcome models exclude units without variation in the outcome, a form of sample selection. This raises the specter of sample selection bias. In this article, I discuss the circumstances under which this sample selection leads to sample selection bias and those under which it does not. Additionally, I present a series of simulation studies, with different data generating processes to measure the extent of this sample selection bias when it occurs. The results show that sample selection bias is predominantly a concern when unobserved heterogeneity is systematically correlated observed covariates (and sample selection). The article concludes with recommendations to apply alternative estimators that impose some assumptions but avoid sample selection, and its potential bias.

*Lucas Núñez is an Assistant Professor at the Schar School of Policy and Government, George Mason University. Email: lnunez6@gmu.edu. I thank Jay Goodlife, Lorena Barberia, and other participants at the MPSA Annual Meeting for helpful feedback and suggestions.

1 Introduction

The estimation of binary outcome models involves many challenges that continuous outcome models, usually estimated using linear regression, do not face. An important challenge is dealing with the presence of unobserved heterogeneity; that is, differences across units of analysis that are not measured or observed but influence the outcome nonetheless. These differences can take many forms: the can be random or systematically related to observed covariates; they can affect the levels of the outcome variable directly or through the independent variables.

An important concern with many methods that account for unobserved heterogeneity arise in cases with limited within-unit variation in the outcome variable over time. In binary outcome models (and other discrete outcome models) this lack of variation in the outcome can occur even when the underlying probability of the outcome occurring has changed.[1] This change in the probability of the outcome without changes in the (observed) discrete outcome itself, has important consequences for the estimation of the impact of independent variables on the binary outcome. While these variables potentially impact the probability of the outcome occurring, lack of variation in the binary outcome leads to several estimating techniques to remove, implicitly or explicitly, these units from the estimation procedure. That is, this lack of within-unit variation in the outcome leads to a form of sample selection.

This sample selection has the potential to lead to sample selection bias. Whether this bias materializes depends on the extent to which the units with and without variation in the binary outcome differ in terms of the unobserved heterogeneity and how this unobserved heterogeneity affects the quantities of interest to researchers. When the heterogeneity is mostly random, this sample selection does not generate sample selection bias[2]: the effective estimation sample may be smaller, but is still representative of the overall sample. When the heterogeneity is systematically related to the independent variables, however, sample selection will lead to sample selection bias: the effective estimation sample is no longer representative of the full sample.

[1] Alternatively, we can think of the binary outcome not changing despite the underlying continuous latent construct that generates the outcome does.

[2] There may be other sources of bias, however, depending on the estimator.

In this article I explore the extent of the bias induced by the sample selection that arises from the lack of within-unit variation in a binary outcome. The analysis and discussion focuses on the Conditional Maximum Likelihood Estimator (Chamberlain, 1980; Rasch, 1961) and the Fixed-Effects estimator (both of which induce sample selection) as well as the Correlated Random Effects (Chamberlain, 1980; Mundlak, 1978) and Penalized Correlated Random Effects (Núñez, 2022) estimators (which use the full sample, but impose some additional assumptions).[3] I analyse the extent of this bias along three main dimensions: (1) the type and extent of unit unobserved heterogeneity; (2) the number of time periods in the data; and (3) the extent to which the outcome being studied is rare. This third dimension, the rarity of the events under study, has been the focus of significant research efforts in the context of unobserved heterogeneity. Rare events typically imply limited within-unit variation in the outcome, and thus exacerbates, in some instances severely, this type of sample selection.

2 Unobserved Heterogeneity in Binary Outcome Models

Throughout this article I consider binary outcome data generating processes (DGPs) that follow a generalized linear model with time-invariant unobserved heterogeneity. These DGPs consist of a binary response, y_{it}, and a k-dimensional vector of time-varying characteristics, x_{it}, such that the response for unit i at time t is generated by:

$$y_{it} = \{\alpha_i + \beta_i x_{it} - \varepsilon_{it} > 0\} \tag{1}$$

where $\{A\}$ is an indicator function that takes the value one if A holds and zero otherwise; β_i is a k-dimensional parameter vector that is allowed to vary by unit i; α_i is a unit specific intercept for unit i; and ε_{it} is a unit- and time-specific error. Throughout this article, I assume that

[3] I also discuss other alternative estimators, but do not include them in simulation exercises given that they have been shown to be outperformed by other estimators. In particular, I briefly discuss Penalized Maximum Likelihood Fixed Effects (Cook et al., 2020), and bias corrected Fixed Effects (Dhaene and Jochmans, 2015; Fernandez-Val, 2009; Fernandez-Val and Vella, 2011), and the method proposed in Beck (2018)

the error term from equation 1, ε_{it}, is strictly exogenous. The general notation in equation 1 accommodates a variety of DGPs, depending on whether the parameters are allowed to vary, β_i and α_i, the extent to which are correlated with other parameters or variables in the model when they do (see Table 1 below).

When the error terms are independently and identically distributed according to a known cumulative distribution function $G(\cdot)$, equation 1 can be alternatively written as:

$$Prob(y_{it} = 1|x_{it}, \alpha_i) = G(\alpha_i + \beta x_{it}) \tag{2}$$

Typical choices of $G(\cdot)$ are the normal distribution, which gives the probit model, or the logistic distribution which gives the logit model. Throughout this article I focus on logistic DGPs, for which $G(\cdot) = \Lambda(\cdot)$, the logistic link.

The slope parameters of this DGP are sometimes of interest themselves. When these slopes vary by unit i, researchers may be interested in the average slope instead. However, interest typically lies in estimating partial effects and probabilities. In the presence of unobserved heterogeneity (in the form of varying intercepts or varying slopes), these partial effects are calculated by taking expectations over the unobserved heterogeneity. The partial effects for equation 2 are defined by:

$$PE_j(x) = E\left[\frac{\partial}{\partial x_j} G(\alpha_i + \beta_i x)|x\right], \;\; j = 1, ..., k \tag{3}$$

where x_j denotes the jth element of x. Additionally, researchers may be interested in the average partial effect, defined as:

$$APE_j = E\left[\frac{\partial}{\partial x_j} G(\alpha_i + \beta_i x)\right], \;\; j = 1, .., k \tag{4}$$

where the last expectation is taking with respect to x, α_i and β_i.

There are a variety of DGPs that fit within this simple framework. These are differentiated

by whether β_{1i} and α_i vary by individual, and whether this variation is correlated with the independent variables or not. Here I will focus on only five alternatives, described in Table 1.[4]

Table 1: DGPs with Unobserved Heterogeneity

β_i	α_i	DGP
$\beta_i = \beta \; \forall \; i$	$\alpha_i = \alpha \; \forall \; i$	No Heterogeneity
$\beta_i = \beta \; \forall \; i$	$\alpha_i \perp\!\!\!\perp x$	Random Intercepts
$\beta_i = \beta \; \forall \; i$	$\alpha_i \not\perp\!\!\!\perp x$	Correlated Intercepts
$\beta_i \perp\!\!\!\perp x$	$\alpha_i = \alpha \; \forall \; i$	Random Coefficients
$\beta_i \not\perp\!\!\!\perp x$	$\alpha_i = \alpha \; \forall \; i$	Correlated Coefficients

The first DGP does not contain unobserved heterogeneity of any kind: all units of analysis are assumed to have the same intercept and the same slope. This DGP is included as a baseline that allows for the study of the performance of different estimators in the absence of heterogeneity.

The second and third DGPs consider unobserved heterogeneity in the intercepts. The second one has random intercepts: the intercepts vary by individual, but the variation is independent of the observed covariates in the model. In the third DGP, the intercepts vary by individual and are correlated with the observed covariates (which is the traditional fixed-effects context). In both cases, the slope is assumed to be the same for every unit.

The fourth and fifth DGPs consider unobserved heterogeneity in the slopes. The fourth has random coefficients: the slopes vary by unit, but this variation is independent of the observed covariates. In the fifth DGP, the coefficients not only vary by unit but they are also correlated to the observed covariates variables. In both cases, the intercept is assumed to be the same for every unit.

[4]Other alternatives are combinations of these five. Any issues arising from any of these specifications will also affect other specifications that are combinations of these five.

3 Estimation Strategies

Standard Logit

The most basic estimation strategy for a binary outcome model with logistic errors is the standard logit model. The standard logit model assumes that there is no heterogeneity of any kind and maximizes the following likelihood:

$$\begin{aligned} \log L(\beta, \alpha) &= \sum \Big[y_{it} \log(p_{it}) + (1 - y_{it}) \log(1 - p_{it}) \Big] \\ p_{it} &= \Lambda(\alpha + \beta x_{it}) \end{aligned} \tag{5}$$

If the data generating process follows the No Heterogeneity case, standard logit model will provide the correct estimates for α, β and the average partial effects described in equation 4. Importantly, the logit model does not implicitly nor explicitly discard units without variation in the outcome. Therefore, it does not suffer from sample selection nor sample selection bias. However, if the data generating process follows a data generating process with any unobserved heterogeneity, the estimates will be inconsistent.

Conditional Maximum Likelihood Estimation

The functional form of the logit model permits obtaining a conditional distribution that does not depends on the intercept(s), dubbed Conditional Maximum Likelihood Estimation or CMLE (Rasch, 1961; Andersen, 1970; Chamberlain, 1984). Here I focus on the conditional maximum likelihood for the case $T = 2$. A similar, but more complicated conditional likelihood, can be built to the same effect when $T > 2$ (see, for example Wooldridge, 2010, p. 622)

$$\log L(\beta) = \sum \left[\{y_{i1} - y_{i2} = 1\} \log \Lambda\big((x_{i2} - x_{i1})\beta\big) + \{y_{i1} - y_{i2} = -1\} \log \left(1 - \Lambda\big((x_{i2} - x_{i1})\beta\big) \right) \right] \tag{6}$$

This method estimates β by focusing on those units for which there is variation in the outcome over time. This results in consistent estimates of β when they are the same for all units in the

sample, regardless of whether the intercepts vary by unit. However, if the slope parameters vary in any way, CMLE will produce inconsistent estimates.

The explicit focus on units with variation in the outcome in CMLE, $\{y_{i1} - y_{i2} \neq 0\}$, implies a form of sample selection. If this sample selection is the results of random (independent) variation in the intercepts, it will not produce sample selection bias. If the intercepts vary systematically with the observed covariates but the slopes are the same for every unit, there is no direct bias from sample selection because β is unrelated to the unobserved hetergoeneity. However, it is possible that some *indirect* bias occurs: units without variation in the outcome likely have systematically higher or lower values of the covariates; their exclusion can ultimately lead to bias. In terms of variation in the slopes, if the variation is random, sample selection bias should not occur, although estimates of the average slope will be biased due to misspecification. Finally, cases in which the slopes vary in a way that is correlated with the independent variables have the potential to lead to sample selection bias (in addition to misspecification bias).

While CMLE is attractive because it can produce consistent estimates of β when the heterogeneity is limited to the intercepts, it also has an important shortcoming. By building a conditional likelihood function that does not depend on the intercepts (varying or not), it cannot provide estimates for the Average Partial Effects (nor any other measure that depends on those intercepts). As such, it can be of limited use.

Fixed Effects Logit

Another alternative is to use the Fixed-Effects logit estimator, which treats unit-specific intercepts as parameters to be estimated, usually implemented as dummy variables for each unit under analysis. The log-likelihood function for this estimator is:

$$\begin{aligned} \log L(\beta, \alpha_1, ..., \alpha_N) &= \sum \Big[y_{it} \log(p_{it}) + (1 - y_{it}) \log(1 - p_{it}) \Big] \\ p_{it} &= \Lambda(\alpha + \beta x_{it} + \alpha_i) \end{aligned} \tag{7}$$

This estimator does not explicitly 'remove' units from estimation as CMLE does. However, it does remove them implicitly: units without variation in the outcome can be perfectly explained by setting α_i to be infinitely large (positive or negative), which is not well defined. As a consequence, these observations are removed from estimation nonetheless (see, for example King and Zeng, 2001; Beck, 2018, 2020). The potential for this method to lead to sample selection bias in the estimates of β depends on the same conditions as for CMLE: if slopes are the same for every unit or vary independently of the covariates, there should be no sample selection bias in estimating β. If the slopes vary and are correlated with the covariates, then sample selection bias can occur.

The main benefit of the FE estimator relative to CMLE is that it allows for the estimation of partial effects and probabilities, because it directly produces estimates of the intercepts, at least for units with variation in the outcome. However, directly estimating these unit-intercepts leads to the incidental parameters problem (Neyman and Scott, 1948). The FE estimator's consistency relies on T-asymptotics. This means that for relatively small T-sizes, it can contain a severe bias. Larger T-sizes reduce this problem, but there is no clear-cut threshold: Heckman (1981) suggests the bias due to incidental parameters is negligible for $T = 8$, while Coupe (2005) argues that larger values, like $T = 16$, are necessary. Simulations in this article suggest that even larger T sizes are not enough if the unobserved heterogeneity is particularly pervasive. This bias due to incidental parameters also affects estimates of APEs.[5]

The FE estimator presents additional problems when estimating APEs (and other related quantities), depending on the type of unobserved heterogeneity. While estimates of β do not suffer from sample selection bias when there are varying intercepts (only bias from incidental parameters), APE estimates do. Beyond the fact that unit-specific intercepts are inconsistently estimated, the intercepts for units without variation in the outcome cannot be estimated at all. This sample selection tends to occur for units with small partial effects, as units without variation in the outcome tend to be those with the largest intercepts (positive or negative). As

[5]However, the bias due to incidental parameters is smaller for APEs than it is for the slopes.

a consequence, APEs are overestimated. This will occur in cases in which the intercepts are independent of the covariates as well as when they are not, with the latter having the largest potential for large bias. A simple (although imperfect) solution is to assume that the partial effects for units without variation in the outcome is exactly zero. Heterogeneity in the slopes can also create sample selection bias in cases where this heterogeneity is correlated with the covariates in the model.

Alternative estimators have been proposed to help deal with this issue. Cook et al. (2020) propose a Penalized Maximum Likelihood estimator that imposes Jeffreys prior on the fixed-effects. This constrains them from being infinitely large, thus avoiding discarding observations without variation in the outcome. Crisman Cox (2019), however, note that this method underperforms relative to CRE (see next subsection). Another alternative is to rely on bias-corrected estimators like those proposed by Dhaene and Jochmans (2015), Fernandez-Val (2009) and Fernandez-Val and Vella (2011), which directly attempt to correct for the bias created by this issue, provided the T-dimension of the data is large enough. However, Núñez (2022) argues that the performance of these estimators is relatively poor, unless sample sizes (in N and T) are very large.

Correlated Random Effects & Penalized Flexible Correlated Random Effects

The Correlated Random Effects (CRE) estimator makes explicit assumptions about how unobserved heterogeneity in the intercepts relates to the observed covariates. The most common implementation, drawn from Mundlak (1978), assumes that there are time invariant intercepts that depend on a linear combination of the time-means of the covariates, as described in equation 8.[6] This specification allows for time-invariant intercepts that are correlated with covariates, but only linearly.

[6]Chamberlain (1980) proposes a more general version of Mundlak's specification, modeling the unobserved heterogeneity in the intercepts by projecting the time dimension of the model into a single dimension. This similar to a weighted mean of the covariates across time.

$$
\begin{aligned}
\log L(\beta, \alpha) &= \sum \Big[y_{it} \log(p_{it}) + (1 - y_{it}) \log(1 - p_{it}) \Big] \\
p_{it} &= \Lambda(\alpha + \beta x_{it} + \gamma \bar{x}_i) \\
\bar{x}_i &= \frac{1}{T} \sum_{t=1}^{T} x_{it}
\end{aligned} \tag{8}
$$

The main advantage of CRE is that, by providing an explicit functional form for how the varying intercepts relate to the observed covariates, it allows for the estimation of both the slopes, β, and the Average Partial Effects. Additionally, it does so without implicitly or explicitly 'removing' observations from the data, avoiding issues of sample selection bias due to lack of variation in the outcome.[7] Another important benefit highlighted by the literature, in both linear and non-linear contexts, is that while CRE specifications do impose some restrictions on the unobserved heterogeneity, there can be large gains in precision as more variation is conserved for estimating the parameters of interest (see, for example Clark and Linzer, 2015; Bell and Jones, 2015; Crisman Cox, 2019).

These benefits make CRE a very attractive strategy as it allows for the estimation of partial effects which CMLE cannot, does not suffer from the incidental parameters problem as FE, and uses all the data, including units without variation in the outcome, to produce the estimates. These benefits even apply in contexts with rare events data, in which sample selection is greatest, which is why Crisman Cox (2019) strongly advocates to their use in rare events contexts.

But the CRE estimator does not achieve these benefits for free. Its main shortcoming is that it restricts the way in which time-invariant intercepts can relate to the observed covariates. This leaves the estimator prone to misspecification which can lead to inconsistent estimates of both β and the APEs.[8]

[7] Units without variation in the outcome will typically have less influence on estimates of β, because their behavior will tend to be captured by the $\gamma \bar{x}_i$ term. However, since γ is the same for every unit in the sample, this term cannot fully explain non-varying units as it is not free to become very large. Thus, units without variation in the outcome still contribute to the estimation of β.

[8] Crisman Cox (2019) shows that the CRE estimator can be robust to misspecification of the unobserved heterogeneity. However, simulations results presented below show that when this heterogeneity is pervasive, the

An alternative to the traditional CRE specifications is the Penalized Flexible Correlated Random Effects (PF-CRE) estimator proposed in Núñez (2022). Similar to CRE, this estimator relies on a explicit formulation of the unobserved heterogeneity but relying instead on a flexible polynomial specification that is more robust to misspecification. This flexible specification is coupled with a penalization step that effectively selects polynomial terms to induce efficiency. Thus, this estimator overcomes the misspecification issues present in traditional CRE. At the same time, similar to CRE, it does not suffer from sample selection issues, since the estimation relies on the entirety of the sample, including units without variation in the outcome variable.

3.1 Where Selection Bias Comes In

A form of sample selection will occur whenever an estimation method implicitly or explicitly 'removes' observations in the estimation process. This sample selection can be benign or pernicious, depending on the specifics of the data generating process. As a general rule, sample selection will be pernicious and lead to sample selection bias when the excluded observations differ systematically from the rest; that is, when the units without variation in the outcome exhibit heterogeneous behavior that is in some way correlated with the parameters of interest. When estimating Average Partial Effects, sample selection bias will occur when units have heterogeneous slopes or intercepts (or both) and when the excluded units have systematically different ones compared to the non-excluded ones (in at least one dimension). This implies that random intercepts and random slopes do not lead to sample selection bias; but correlated intercepts and correlated slopes do.

The extent or severity of this sample selection bias, when present, depends on three additional factors. The first is the extent to which the unobserved heterogeneity (in slopes or intercepts) is correlated with the observed components of the model: the stronger the relationship, the larger the sample selection bias. The second factor is the share of units that are excluded due to lack of variation in the outcome; this in turn depends on the extent to which

standard CRE can produce estimates with a large bias due to misspecification.

the event under study is considered rare (or unlikely).[9] The third factor is the number of time-periods in the data: ceteris paribus, more time periods translate into a lower likelihood of units not experiencing variation in the outcome, thus reducing the extent of the sample selection, and consequently the bias caused by it.

It is important to note here that different estimators may suffer from other sources of bias in the presence of unobserved heterogeneity (whether random or systematic), beyond sample selection bias. For this reason, when analyzing the results of the simulations presented in the following section, it is important to consider how any bias in the estimates varies with the extent of the sample selection challenge. This is most easily done by comparing the extent of the bias in simulations with rare or very rare events and those in which the events are not rare.

4 Data Generating Process for Simulations

This section describes specifics of the data generating processes used in the simulations presented below. In all cases, the outcome depends on a latent index, z, with the following functional form:

$$z_{it} = \alpha + \beta_{1i} x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_i \tag{9}$$

where x_{1it}, x_{2it}, and x_{3it} are the values of three independent variables for each unit i in period t; β_{1i} is the coefficient for the first independent variable and, in principle, it is allowed to vary by unit i; β_2 and and β_3 are the fixed coefficients for the other independent variables; α is an intercept; and α_i is a unit-specific additional intercept.

With the addition of an extreme type-I error term, ε, consider the the following latent outcome variable:

$$y^*_{it} = z_{it} - \varepsilon_{it} \tag{10}$$

[9] Please note that while the literature focuses on 'rare' events, very common events create exactly the same problems.

From this, the probability of the outcome occurring is defined as:

$$P(y_{it} = 1) = P(y_{it}^* > 0) = \Lambda(z_{it}) = \frac{1}{1 + e^{-z_{it}}} \tag{11}$$

where $\Lambda(\cdot)$ is the logit link.

In all cases, the three independent variables are normally distributed with means zero and the following covariance matrix:

$$\mathbf{x}_{it} \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{12}$$

$$\Sigma_{i,j} = \begin{cases} \frac{1}{2^{|i-j|}}; & |i-j| \leq 2 \\ 0; & \text{otherwise} \end{cases} \tag{13}$$

The data generating processes of interest are differentiated by the specification for α_i and β_i, which are presented in Table 2.

Table 2: Models of Heterogeneity Considered

Model	β_{1i}	α_i
No Heterogeneity	$\beta_{1i} = 1 \; \forall \; i$	$\alpha_i = 0 \; \forall \; i$
Random Intercepts	$\beta_{1i} = 1 \; \forall \; i$	$\alpha_i \sim \mathcal{N}(0, 1)$
Correlated Intercepts	$\beta_{1i} = 1 \; \forall \; i$	$\alpha_i = \frac{3}{4}(-2x_{1i1} + 0.3x_{2i1}^2) + \frac{1}{4}\epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, 1)$
Random Coefs	$\beta_{1i} \sim \mathcal{N}(1, 1)$	$\alpha_i = 0 \; \forall \; i$
Correlated Coefs	$\beta_{1i} = 1 - x_{1i1} + \frac{1}{2}\epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, 1)$	$\alpha_i = 0 \; \forall \; i$

For the No Heterogeneity case, the coefficient for the first independent variable is set equal to one for every unit and the unit-specific intercept is set to zero. In the Random Intercepts case, the coefficient for the first independent variable is fixed at 1 for all units, and the intercept has a normal distribution centered at zero with variance one. In the Correlated Intercepts case, the coefficient for the first independent variable is also fixed at 1 for all units, but the intercept is correlated with both the first and second independent variables in the model. The specific form of dependence includes a quadratic term for the second independent variable. This quadratic term is included to avoid overstating the degree to which the standard Correlated Random

Effects estimator can recover correct estimates.[10]

The data generating process for the two other models fixes the intercept to zero for all units. In the Random Coefficients case, the coefficient for the first independent variable has a normal distribution with expected value 1 and variance 1. In the Correlated Coefficients case, the coefficient for the first independent variable has a linear correlation with the first independent variable; but it still has expected value equal to 1.

The final three parameters in this data generating process are the slopes of the second and third independent variables, β_2 and β_3, which are set equal to one; and the common intercept, α, which is manipulated to generate different levels of "rare" events, ranging from relatively 'common events' to 'very rare events.'

4.1 Results

The following results use a sample size of $N = 250$ and four different longitudinal sizes: $T = 2, 5, 10, 20$. I present results for the bias in estimates of β_1 (or its expected value) and estimates of the APE. The Appendix includes figures showing the Root Mean Squared Error (RMSE) for the different estimators.

Let us begin with the bias in the estimates of β from the Conditional Maximum Likelihood estimator, described in equation 6, which are presented in the first row of Figure 1. Since this method does not allow for the estimation of Average Partial Effects, only the results for estimated slopes are presented. This estimator has no bias (or a negligible one) for the data generating processes with no heterogeneity, random intercepts, and correlated intercepts. The only exception is for cases in which events are rare and the time-dimension of the data is smallest, $T = 2$.[11] This is expected, as these three data generating processes fit perfectly within the assumptions of CMLE. The rightmost top two panels present the bias of CMLE in estimating the average of β_i. For both data generating processes CMLE produces biased estimates, which

[10] A purely linear dependence between the covariates and the intercept fits exactly in the assumptions of CRE, but may be limiting in real-life applications.

[11] This is not surprising, since a the combination of a rare events and smaller sample size leaves little information in the data to estimate the parameters consistently.

is expected since CMLE does not account for heterogeneity in the slopes of the data generating processes.

Is there a sample selection bias induced by the rarity of the events for CMLE? As the top panels of Figure 1 show, the estimator's bias is about the same, regardless of how rare the events are.[12] For the first three DGPs this is due to the slope of the model, β, being the same for the units with variation in the outcome and those without variation in the outcome. Sample selection occurs, but it is independent of β and thus leads to no bias. In the random coefficients case, the bias is not affected by the rarity of the events because the distribution of β_i is independent of whether an individual experiences variation in the outcome or not. The results for the fifth DGP, correlated coefficients, warrants a more careful description. In this case, the parameter β_i is correlated with the independent variables. Additionally, individuals with more extreme values of the independent variable should be more likely to experience no variation in the outcome. Since the slope for these individuals is different, the rarer the events the larger the bias should be. However, the bias does not change with the rarity of the events, which is somewhat unexpected for this data generating process. Estimates in this case are biased nonetheless, but the rarity of events does not seem to have a noticeable impact.[13]

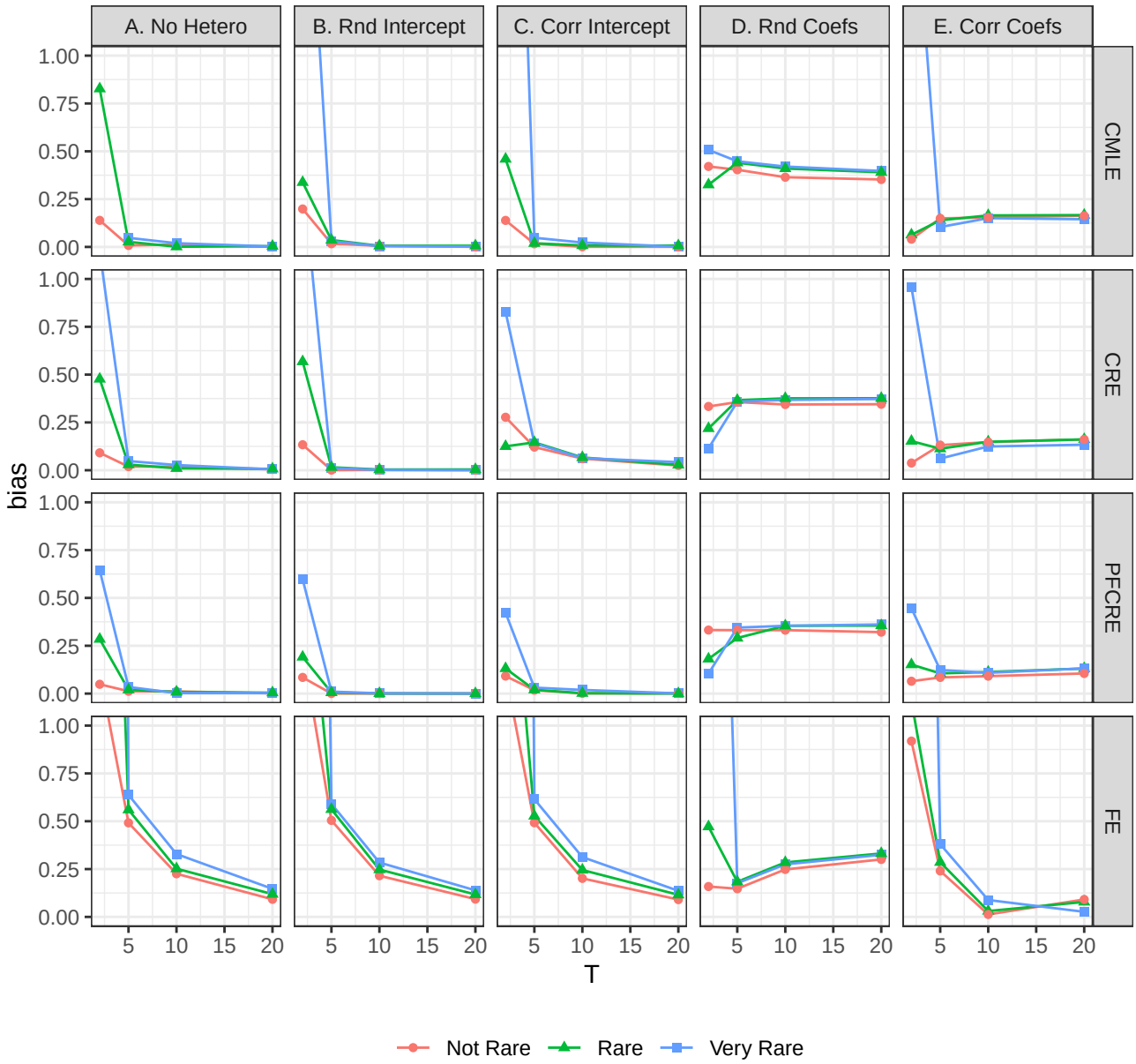
Overall, CMLE estimates β (or average β), do not seem to experience sample selection bias due to the rarity of the event under study. When CMLE is biased, as in the cases with varying coefficients, it is due to model misspecification instead.

The next estimator is the Correlated Random Effects, for which both β and the Average Partial Effect (APE) can be estimated. The bias of CRE for the slope and the APE are presented in the second row of Figures 1 and 2. The GDPs with no heterogeneity and the one with random intercepts fit within the assumptions of the CRE estimator; therefore, we should expect no bias in those cases. This is precisely what the figures show (except for $T = 2$). In terms of correlated intercepts, the traditional CRE estimator is consistent as long as the heterogeneity

[12]The only caveat is for very small T.

[13]It is possible that the heterogeneity in the slopes is not sufficiently pervasive.

Figure 1: Bias in estimated slopes β



All simulations come from a sample size of $N = 250$ units. For Random Coefficients and Correlated Coefficients, bias is calculated relative to the average slope.

in the intercept is linear in the independent variables. The DGP used here, however, violates this assumption, which is the reason why estimates for β and the APE are biased.[14] In the case of random coefficients, the estimates of β are biased due to the assumption (incorrect in this case) of constant slopes, but estimates of the Average Partial Effects have relatively low bias. The reason for this is that the estimates of β factor into the APE in two opposing ways that seem to nearly compensate each other (this also happens in the standard logit model). Finally, CRE produces biased estimates of both β and the APE in the correlated coefficients case.

The rarity of the events under study does not have a noticeable impact in the bias of β estimates from CRE. It does, however, affect the bias of the Average Partial Effect estimates in the Correlated Intercepts and Correlated Coefficients DGPs: the rarer the outcomes, the higher the bias of the APE. However, this increased bias is not due to sample selection as CRE uses all units to produce the estimates. Instead, this bias comes from the fact that CRE is misspecified for both data generating processes; the rarity of the outcome simply exacerbates the misspecification bias, but it is not the cause of it.

The third rows in Figures 1 and 2 present the bias in estimates β and APE from the PF-CRE estimator. The performance of this estimator is similar to that of CRE in the No Heterogeneity and Random Intercept cases, which is expected since CRE is a special case of PF-CRE. For the Correlated Intercepts case, however, PF-CRE does not show any bias (except for $T = 2$ with very rare events), as this estimator is better able to capture non-linearities in the way that the heterogeneous intercepts relate to the observed covariates. Compared to CRE, the improvement of PF-CRE in the Correlated Intercepts case is most notable in APE estimates. For the Random Coefficients and Correlated Coefficients cases, PF-CRE has a similar performance to CRE, as neither of these estimators accommodates for heterogeneity in the slopes.

[14]The bias of β in from CRE in the Correlated Intercepts DGP is not very large. This is because the non-linearity in the heterogeneity comes x_2 rather than x_1 and thus tends to affect the slope of x_1, the focus of Figure 1, less than the other slope. This bias would be larger if the independent variables were more highly correlated to each other. The bias for the APE is larger because the APE depends on directly on all independent variables and coefficients.

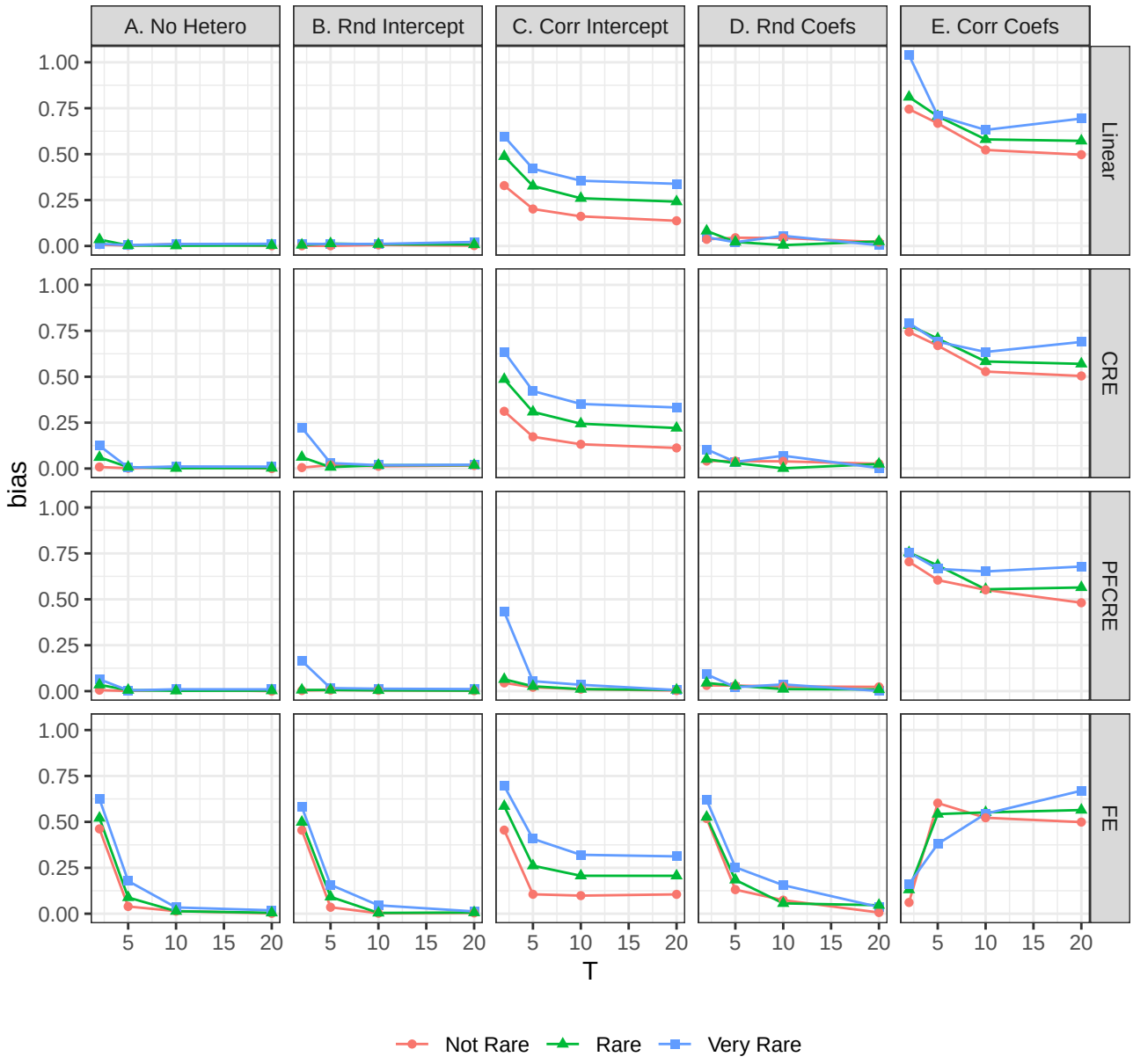
The fourth rows of Figures 1 and 2 shows the bias from the Fixed-Effects estimators. The Fixed-Effects estimator produces biased estimates of β for the No Heterogeneity, Random Intercepts, and Correlated Intercept cases. This bias is predominantly a consequence of the incidental parameters problem and slowly disappears as T increases.[15] The bias in estimates of β tend to be somewhat larger when events are rarer, suggesting that rare events exacerbate the incidental parameters problem. In terms of estimates of Average Partial Effects, the bias disappears quickly for the No Heterogeneity and Random Intercepts cases. For the Correlated Intercept case, however, the bias does not disappear completely as the T-dimension of the data increases and it is much more severe for cases with rare events. This is a clear indication of sample selection bias at play: the units which the FE model implicitly discards in the estimation process have systematically different intercepts that those used in the estimation, which means that Average Partial Effects are calculated on a non-representative subsample of the original data.

For the Random Coefficients case, estimates of β are systematically biased because the FE estimator is misspecified for this case; however, APEs show very little to no bias. In both cases, the bias does not vary with the rarity of events. For the final DGP, Correlated Coefficients, Fixed-Effects estimates of both β and the APEs are bias. Additionally, the bias in APE estimates increases with the rarity of events.

The final estimator considered is the linear probability model with fixed effects. Parameter estimates from this model are approximations of the Average Partial Effect and are thus presented in the first row of Figure 2. In both the No Heterogeneity and Random Intercept cases, the linear probability model shows no bias. In the Correlated Intercepts case, the linear model is biased for a similar reason FE is: units without variation in the outcome do not influence the the estimates of the linear model's slopes, which implicitly means excluding from estimation. This bias increases with the rarity of events and is a consequence of the sample selection

[15]Note that even for $T = 20$, which is relatively large, the FE estimator retains a considerable bias in estimates of β, suggesting that common recommendations of $T = 8$ (Heckman, 1981) or $T = 16$ (Coupe, 2005) as sufficient to overcome this problem may only apply to certain circumstances.

Figure 2: Bias Average Partial Effect



induced by the estimation procedure: the units without variation in the outcome are effectively excluded from estimation (the fixed-effects perfectly explain them) and they are also the units that typically have much smaller partial effects. The bias of the linear model in the Random Coefficients case tends to be relatively small, whereas for the Correlated Coefficients case it tends to be large, and worsen with the rarity of events.

Overall, sample selection bias induced by lack of within-unit variation in the outcome occurs when the quantities of interest exhibit individual variation that correlates with the sample selection. This limits the problem to two data generating process: Correlated Intercepts and Correlated Coefficients.

CMLE estimates of the slopes always experience sample selection. In the Correlated Intercepts case, this does not generate sample selection bias in estimates of β because the slope is the same for every unit. In the Correlated Coefficients case, while CMLE should be affected by sample selection bias, the simulations presented here find no evidence of it in practice.[16]

Similar to CMLE, FE estimates always experience sample selection. In the cases where the slopes are homogeneous, FE estimates of β should not experience bias from this sample selection. Instead, the bias of FE estimates in these cases comes from the incidental parameters problem, which disappears slowly as the number of time periods, T, increases. The sample selection issue does seem to exacerbate this bias, but likely due to sample selection reducing the sample size and making convergence slower. For cases with heterogeneous slopes, the FE estimator shows biased estimates of β due to misspecification, but there is little evidence of sample selection either exacerbating or causing this bias.

In terms of Average Partial Effects, FE estimates in the Correlated Intercepts case show very clear evidence of bias due to sample selection. This is precisely because the units that FE excludes from estimation have systematically different partial effects (very small ones) compared to the rest of the units. For the Correlated Coefficients case, the impact of sample selection on

[16]It is possible that the heterogeneity in the slopes is not large enough to generate sufficient differences in the slopes of units that show no variation in the outcome and units that do.

FE is not clear.

The extent of this sample selection bias generated by limited within-unit variation of course depends on the extent of the unobserved heterogeneity. Appendix Figures B1 and B2 present the bias in estimates of the Average Partial Effects for the Correlated Intercepts and Correlated Coefficients cases, respectively. Both figures include a case in which the heterogeneity is highly correlated with the independent variables and one in which it is only moderately correlated with the independent variables. As both figures show, the bias due to sample selection becomes less problematic the less pervasive the heterogeneity, especially in the case of Correlated Coefficients.

The largest issue with sample selection induced by lack of within unit variation occurs in the Fixed Effect estimates of APEs when the data generating process includes correlated intercepts. A simple correction is to assume that the partial effect for those units without variation is zero. This can typically help reduce the bias in APEs, as Figure 3 shows, although imperfectly. The main issue with this correction is that the partial effects for excluded units are not necessarily exactly zero, but simply likely to be small. As such, this correction can often be excessive.

There are better approaches to resolve the issue of lack of within-unit variation. One solution is to rely on the Penalized Maximum Likelihood Fixed Effects (PML-FE) estimator proposed by Cook et al. (2020). This method imposes Jeffreys prior to the unit-specific dummy variables of the Fixed Effects logit estimator. The imposition of this prior prevents the unit-specific dummies from perfectly predicting the outcome for units with no variation in it. This is accomplished by the penalization directly, as it prevents the dummy variables from going to infinity. As such, it ensures that these units provide some information in estimating β. More importantly, it ensures that they are included in the estimation of the Average Partial Effect. Moreover, units without variation in the outcome will not have exactly zero Partial Effects, as in the simple adjustment presented in Figure 3. Instead, Jeffreys prior bounds the individual partial effects from being exactly zero. there are also bias-correction techniques that have been proposed for the Fixed-Effects estimator (Dhaene and Jochmans, 2015; Fernandez-Val, 2009; Fernandez-Val and Vella, 2011), however, these techniques usually require large sample sizes (in N and T) to

produce good estimates.

An alternative solution is to rely on the PF-CRE method previously discussed and included in the simulations. As shown in Figures 1 and 2 PF-CRE is able to capture unobserved heterogeneity in the intercepts without imposing restrictive assumptions (like CRE does) and without discarding observations without variation in the outcome (like CMLE and FE do). Additionally, Crisman Cox (2019) shows that the traditional CRE approach tends to outperform PML-FE, even when it is misspecified. Given that PF-CRE includes the CRE as a specific case, it should also be expected to outperform PML-FE as well. Additionally, PF-CRE tends to perform well with relatively small sample sizes, whereas bias-correction models for FE tend to require much larger samples.

Overall, sample selection induced by lack of variation in the outcome matters only in a select number of circumstances: when the intercept is correlated with the independent variables and, to a lesser extent, when the slopes are correlated with the independent variables. For the correlated intercept cases, there are multiple solutions available that help reduce or eliminate this bias due to sample selection, while also reasonably avoiding other forms of bias. In the case of correlated coefficients, the problem is a different one: all the methods commonly used for binary outcome models produce biased estimates simply because they are misspecified. Limited variation in the outcome only plays a marginal role in biasing already biased estimates.

5 Conclusion

Binary outcome models with panel data present a variety of challenges, especially in the presence of unobserved heterogeneity. A particular challenge arises in these models when, in addition to (or because of) the unobserved heterogeneity, a non-negligible proportion of the units under analysis do not experience variation in the outcome. Common methods used to estimate these models, like Conditional Maximum Likelihood, the Fixed Effects logits, and linear models with fixed effects, implicitly or explicitly discard or exclude these units from the

estimation procedure, creating a form of sample selection. Depending on the underlying data generating process and the parameters of interest, this bias can be benign, simply leading to a loss of power without any further consequences. But in other circumstances, this bias can be pernicious, not only leading to a loss of power but also creating sample selection bias in the estimates.

In principle, any estimation method that involves a form of sample selection will generate sample selection bias in the estimates if the quantities of interest systematically differ between units with and without variation in the outcome variable. Thus, theoretically, data generating processes with random heterogeneity, in either slopes or intercepts, should not experience any noticeable impact from this type of sample selection. Conversely, data generating processes in which the heterogeneity is in some way correlated with other elements in the model (and thus with selection) are susceptible to sample selection bias.

The simulations presented in this article show that sample selection bias, created by lack of within unit variation in the outcome, does not occur in common estimation methods when the heterogeneity is random, as expected. When the unobserved heterogeneity is not random (and correlated with the independent variables), sample selection bias does occur. Focusing on the estimation Average Partial Effects, which are usually the quantity of interest in much of research, this sample selection bias tends to be a bigger problem in data generating processes with heterogeneous unit-intercepts. While data generating processes with heterogeneous slopes are also influenced by sample selection bias, its impact is more limited. Furthermore, the common estimators discussed in this article are not designed to deal with varying slopes in the first place, meaning that their estimates are biased regardless of sample selection problems.

There is an unavoidable reality, however, to cases in which there is unit unobserved heterogeneity in conjunction with lack of within unit variation in the outcome: some parameters simply are not identified for the entire sample. This is especially the case for data generating processes with unit intercepts, as probabilities and partial effects for units without variation in the outcome are not identified by the data. There are, however, several partial solutions to

this problem. One solution is to use the Correlated Random Effects estimator, which imposes restrictions on the unobserved heterogeneity. If these restrictions are reasonable for the specific application, then quantities of interest can be estimated consistently. Importantly, CRE does not discard observations without variation in the outcome, making it an attractive alternative for rare events data, as advocated by Crisman Cox (2019). Under some circumstances, the validity of these restrictions can be tested, using a Hausman-style test (Hausman, 1978; Núñez, 2022). A related alternative is to rely on the Penalized Flexible Correlated Random Effects (PF-CRE) from Núñez (2022), which reduces the chances of misspecification present in the standard CRE by relying on a flexible polynomial form for the unobserved heterogeneity and ensures a more efficient estimates thanks to a penalization step, thus retaining and improving upon the benefits of traditional CRE.

Overall, however, it is important to recognize than when outcomes are rare and there is little within-unit variation in the outcomes, the data simply does not have enough information to identify and consistently estimate many of the quantities of interest. When this is the case, researchers need resort to extra-data information to aid in the identification and estimation of the quantities of interest. Any method that incorporates extra-data information is as good as the extra information added or the generality of the assumptions underlying it. As such, careful consideration and justification of the validity of this additional information or assumptions is necessary in those cases.

References

Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society, Series B (Methodological)*, 32(2):283–301.

Beck, N. (2018). Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effects: What are the Issues? *http://arxiv.org/abs/1809.06505*.

Beck, N. (2020). Estimating Grouped Data Models with a Binary-Dependent Variable and Fixed Eects via a Logit versus a Linear Probability Model: The Impact of Dropped Units. *Political Analysis*, 28:139–145.

Bell, A. and Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, 3(1):133–153.

Chamberlain, G. (1980). Analysis of Covariance with Qualitative Data. *Review of Economic Studies*, 47:225–238.

Chamberlain, G. (1984). Panel Data. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume II, chapter 22. North-Holland, Amsterdam.

Clark, T. S. and Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods*, 3(2):399–408.

Cook, S. J., Hays, J. C., and Franzese, R. J. (2020). Fixed effects in rare events data: A penalized maximum likelihood solution. *Political Science Research and Methods*, 8:92–105.

Coupe, T. (2005). Bias in Conditional and Unconditional Fixed Effects Logit Estimation: A Correction. *Political Analysis*, 13:292–295.

Crisman Cox, C. (2019). Estimating Substantive Effects in Binary Outcome Panel Models: A Comparison. *Journal of Politics*, Accepted.

Dhaene, G. and Jochmans, K. (2015). Split-Panel Jackknife Estimation of Fixed-Effect Models. *Review of Economic Studies*, 82:991–1030.

Fernandez-Val, I. (2009). Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models. *Journal of Econometrics*, 150:71–85.

Fernandez-Val, I. and Vella, F. (2011). Bias Corrections for Two-Step Fixed Effects Panel Data Estimators. *Journal of Econometrics*, 163:144–162.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6):1251–1271.

Heckman, J. J. (1981). The Incidental Parameters Problem and the Problem of Initial Conditions in Discrete Time-Discrete Data Stochastic Process. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications* . MIT Press, Cambridge.

King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2):137–163.

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

Neyman, J. and Scott, E. L. (1948). Consistent Estimates Based on Partially Constistent Observations. *Econometrica*, 16:1–32.

Núñez, L. (2022). Partial Effects for Binary Outcome Models with Unobserved Heterogeneity. *Journal of Politics*, 84(1):67–85.

Rasch, G. (1961). On General Laws and the Meaning of Measurement in Psychology. *The Danish Institute of Educational Research, Copenhagen.*

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT Press, Cambridge, MA, second edition.

Appendix A Root Mean Squared Error

Figure A1: RMSE β

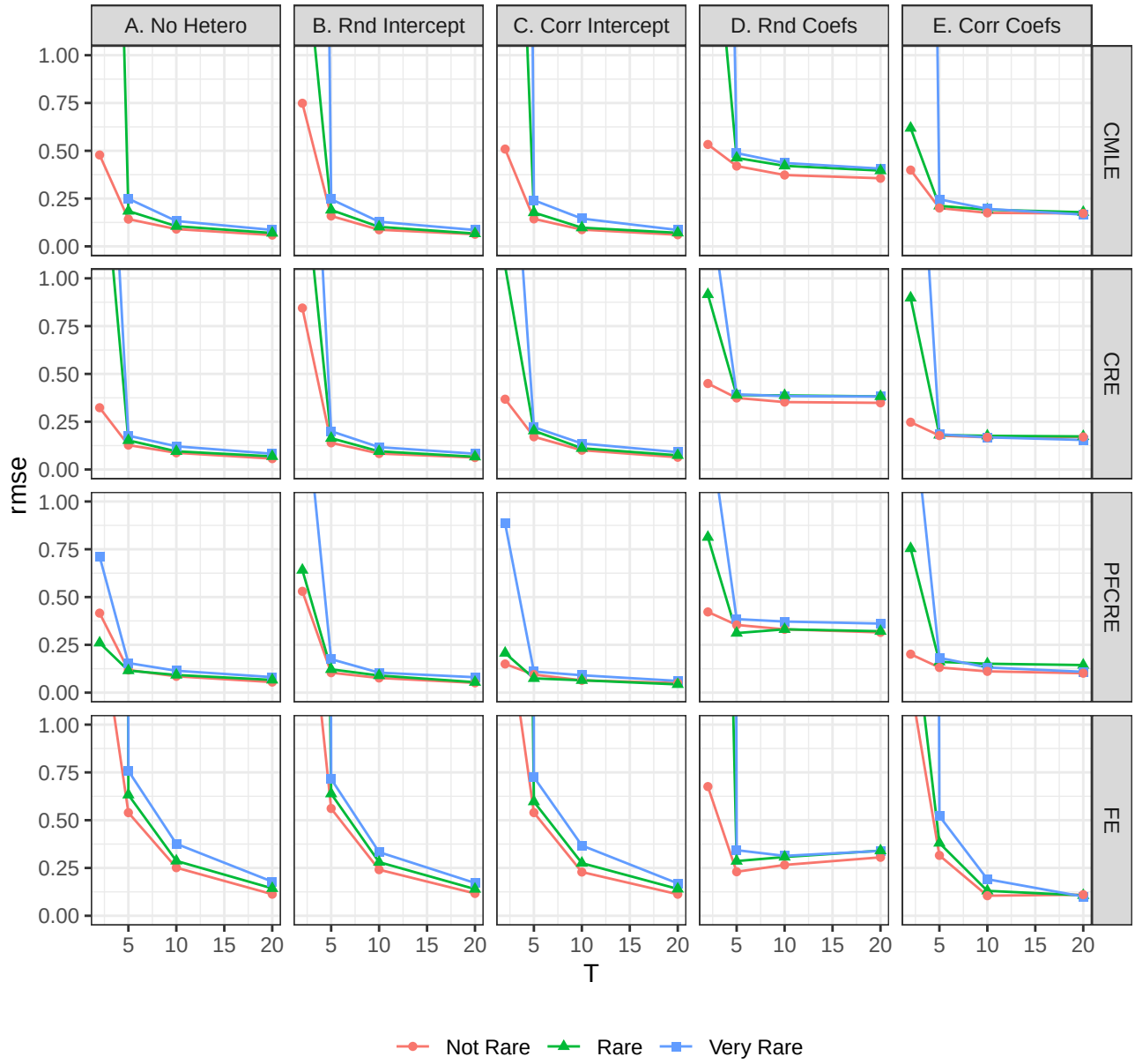
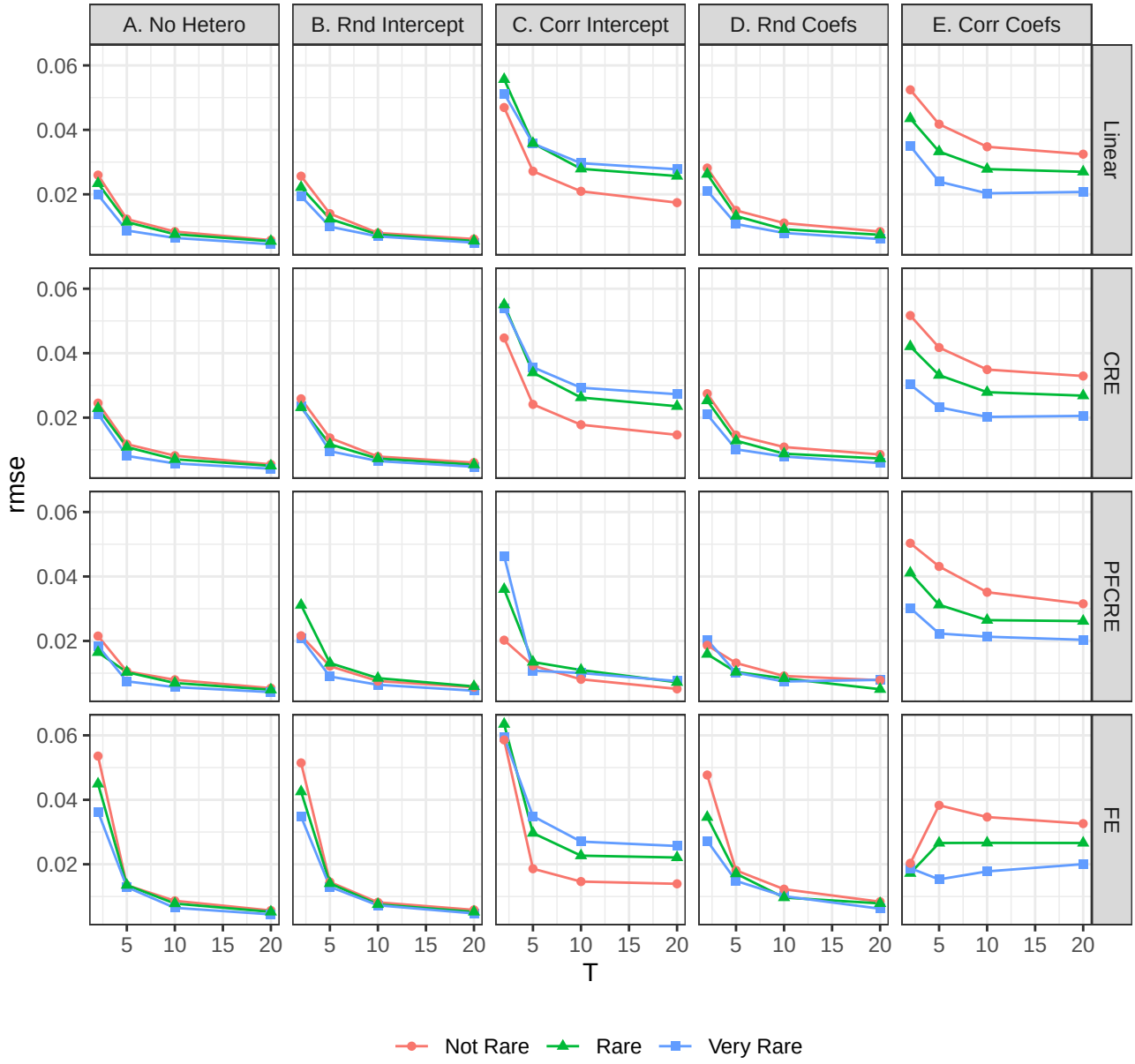


Figure A2: RMSE Average Partial Effect



Appendix B Less Pervasive Heterogeneity

Figure B1: Bias Average Partial Effect – Correlated Intercept

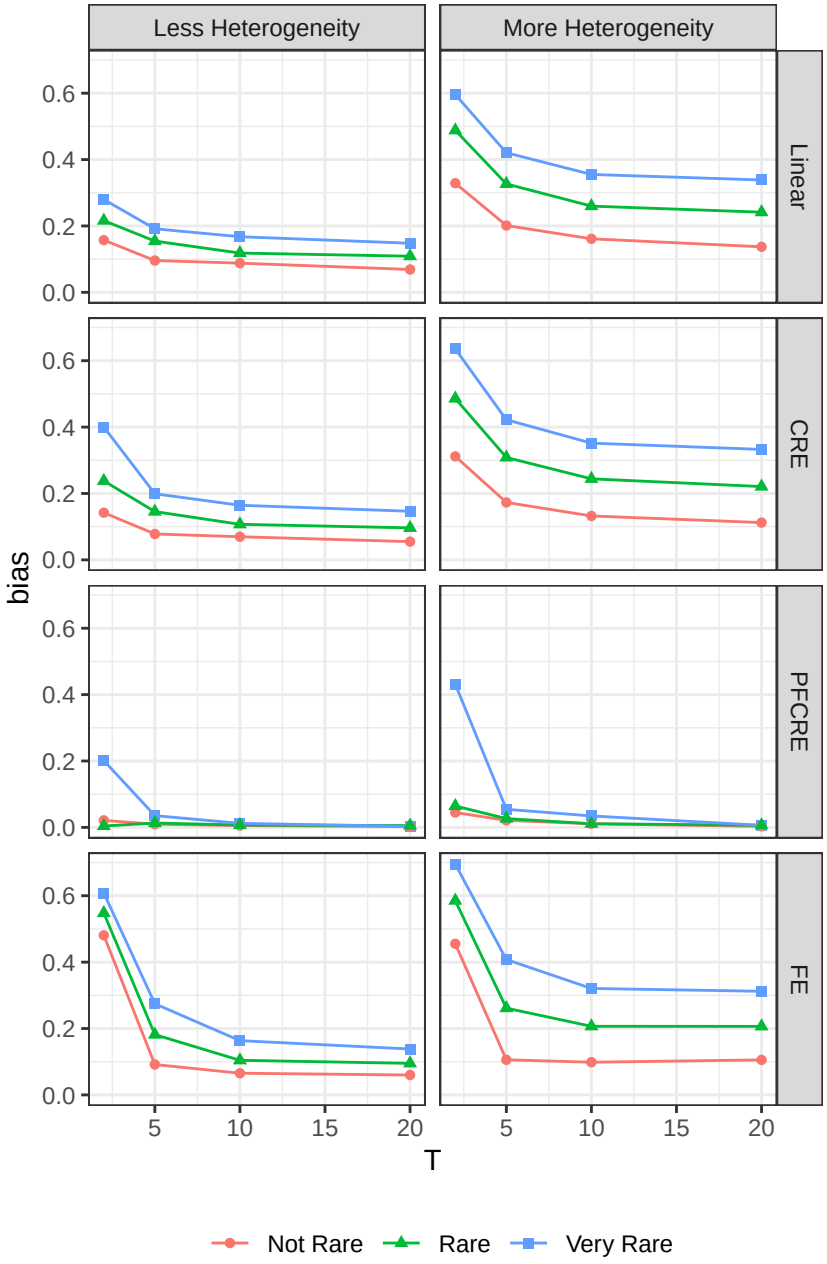


Figure B2: Bias Average Partial Effect – Correlated Coefficients

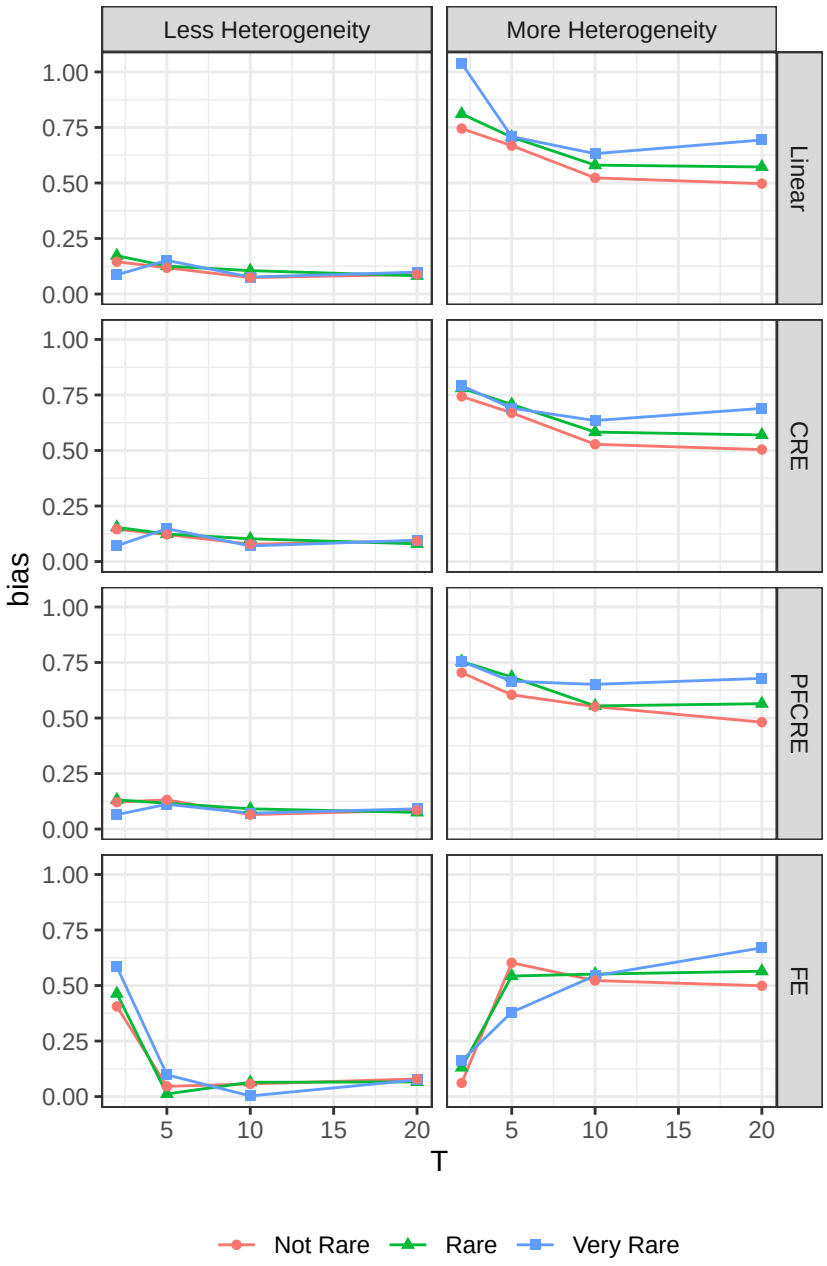
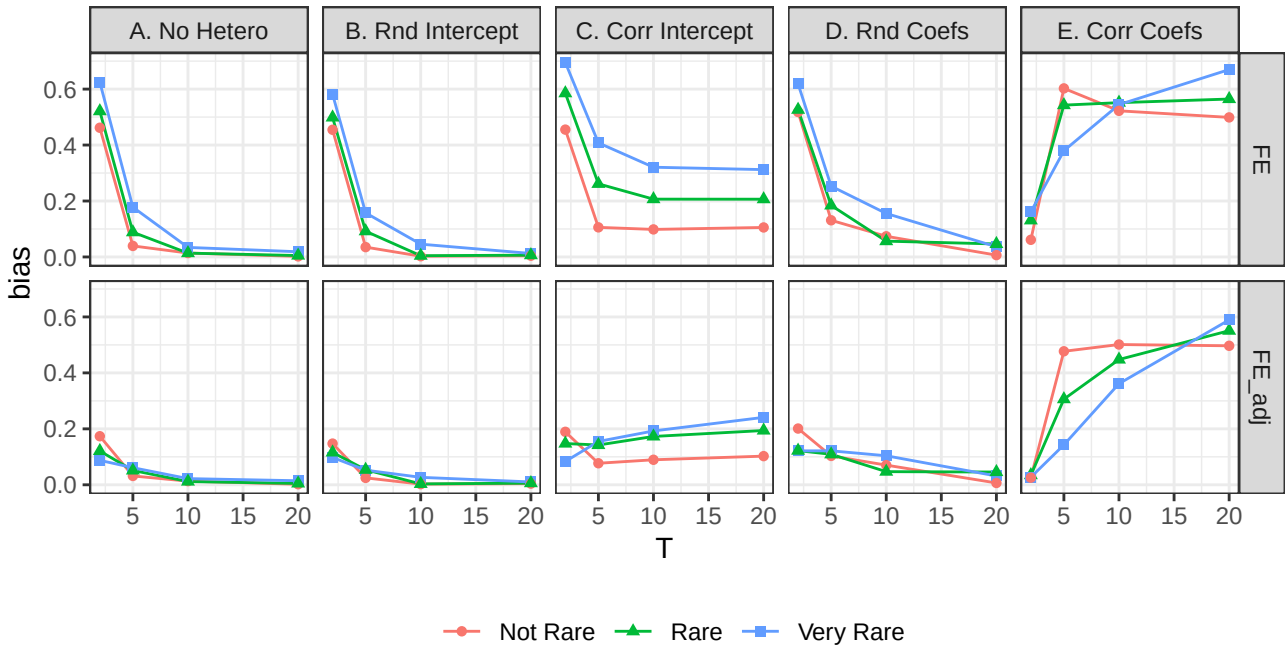


Figure 3: FE & Adjusted FE



The adjusted FE model assumes that the partial effects for units without variation in the outcome is exactly zero.