# Estimation in the fixed effects ordered logit model

Chris Muris[*]

December 16, 2015

**Abstract**

This paper introduces a new estimator for the fixed effects ordered logit model. The proposed method has two important advantages over existing estimators. First, it estimates the differences in the cut points along with the regression coefficient. The cut point differences provide bounds on a marginal effect, so they can be used to interpret the magnitude of the regression coefficient. Second, the proposed estimator for the regression coefficient is more efficient than any existing estimator. The proposed estimator is based on the observation that the ordered logit model with $J$ outcomes and $T$ observations can be converted to a binary choice logit model in $(J-1)^T$ ways. The conditional maximum likelihood estimator in Chamberlain (1980) can be applied to each of these binary choice models. The proposed procedure optimally combines the information from all these binary choice models. Existing estimators use at most $(J-1)$ of these $(J-1)^T$ transformations. A simulation study reveals that the resulting efficiency gain can be substantial. As an empirical illustration, I examine the income-health gradient for children using data from the Medical Expenditure Panel Survey.

## 1 Introduction

The fixed effects ordered logit model is widely used in empirical research in economics.[1] The model allows a researcher with panel data and an ordinal dependent variable to

[*] Department of Economics, Simon Fraser University. Email: cmuris@sfu.ca. I would like to thank Irene Botosaru, Heng Chen, Bryan Graham, Brian Krauth, Krishna Pendakur, Pedro Portugal, Pedro Raposo, Marcel Voia, and Nathanael Vellekoop for very helpful comments and suggestions. This paper has benefited greatly from the feedback of two anonymous referees, and from seminar participants at the 2014 Seattle-Vancouver Econometrics Conference, Catolica Lisbon, University of Kentucky, and the 2015 Canadian Economic Association meetings.

[1] Examples of empirical research using panel data with ordered responses can be found in many areas of economics. In health economics, Carman (2013) looks at the intergenerational transfers and health. and Frijters et al. (2005) and Khanam et al. (2014) look at the relationship between income and health. In the context of the economics of educations, Fairlie et al. (2014) estimate the effect of same-minority teachers on student outcomes. Allen and Allnutt (2013) are interested in the effect of the "Teach First" program on student achievement. In labor economics, examples of authors using the fixed effects ordered logit model are Das and van Soest (1999), Hamermesh (2001), and Booth and van Ours (2008, 2013). An area of research where fixed effects ordered logit models are heavily used is the empirical research of life satisfaction, see, among many others, Ferrer-i-Carbonell and Frijters (2004), Frijters et al. (2004) and Blanchflower and Oswald (2004). Finally, the fixed effects order logit model is useful for the analysis of (sovereign) credit ratings, see e.g. Amato and Furfine (2003) and Afonso et al. (2013).

control for time-invariant unobserved heterogeneity that is correlated with the observed covariates in an unrestricted way.

In this paper, I propose a new estimator for the fixed effects ordered logit model. This procedure has two advantages over existing methods. First, it simultaneously estimates the differences in the cut points and the regression coefficient. The cut point differences can be used to bound a marginal effect.[2] Existing estimators do not estimate the cut point differences, and can consequently not be used to examine the magnitude of the regression coefficient.[3]

Second, the new estimator for the regression coefficient is more efficient than existing estimators, despite the additional cut point differences that need to be estimated. I show a strict efficiency gain with respect to the most efficient estimator currently around.[4] A simulation study suggests that the efficiency gain can be substantial.

The proposed estimator is based on the observation that the ordered logit model with $J$ outcomes and $T$ observations can be converted to a binary choice logit model in $(J-1)^T$ ways. The conditional maximum likelihood estimator in Chamberlain (1980) can be applied to each of these binary choice models. The proposed procedure optimally combines the information from all these binary choice models. Existing methods do one of the following: (1) they use only $(J-1)$ transformations or (2) they collapse the ordered variable to a binary choice variable, which corresponds to using only one such transformation. These procedures are less efficient for the regression coefficient, and do not provide an estimate of the cut point differences.

To fix ideas, consider the following simple example. Let $y_t \in \{1, 2, 3\}$ be an ordered random variable indexed by time. Choose category 1 a the cutoff category, and consider the transformation $d_{t,1} = 1\{y_t \leq 1\}$. Since the transformed variable $d_{t,1}$ is a binary choice variable, the conditional logit estimator in Chamberlain (1980) can be applied to $d_{t,1}$. The same procedure can be repeated for a transformation based on the other cutoff category $d_{t,2} = 1\{y_t \leq 2\}$. For efficiency, one could then combine the estimators based on $d_{t,1}$ and $d_{t,2}$. The first paper to point this out is Das and van Soest (1999).

I show that there are two additional transformations when we allow the cutoff category to vary over time, namely

$$d_{t,(1,2)} = 1\{y_t \leq t\} = \begin{cases} 1\{y_1 \leq 1\} & \text{if } t = 1, \\ 1\{y_2 \leq 2\} & \text{if } t = 2, \end{cases}$$

and $d_{t,(2,1)} = 1\{y_t \leq 3 - t\}$. These transformations provide additional information about the regression coefficient, and about how far apart the categories are. Therefore, combining the conditional logit estimators based on these transformation allows us

[2] Without the cut point differences, partial effects cannot be computed. This is closely related to an analogous drawback in the context of the fixed effects binary choice logit model, see the discussion in e.g. Chamberlain (1984, p. 1277), Honoré (2002, Section 2), and Wooldridge (2010, section 15.8.3).

[3] See Baetschmann et al. (2015) for a recent contribution, and an overview of the existing procedures.

[4] This estimator was introduced by Das and van Soest (1999). Their estimator, and some variations on it, are discussed in Baetschmann et al. (2015).

to estimate the regression coefficient more efficiently, and simultaneously gives us an estimator for the cut point differences.

**Organization.** The remainder of this paper is organized as follows. Section 2 surveys the related literature. Section 3 introduces the fixed effects ordered logit model and presents the main result concerning the $(J-1)^T$ sufficient statistics. In Section 3.3, I discuss how the cut point differences can be used to bound a certain marginal effect that may be relevant to empirical practice. In Section 4, I introduce the conditional maximum likelihood estimator based on a single transformation, and establish its asymptotic properties. In Section 5, I show how to efficiently combine the information from all such estimators into a single estimator for the regression coefficient and cut point differences. I show that the estimator is at least as efficient as currently available procedures. Section 6 discusses the implementation of the estimator, focusing on its implementation in Stata. I also discuss a composite likelihood procedure designed to overcome finite-sample issues. Section 7 documents the results of a simulation study, and Section 8 contains an empirical illustration on health satisfaction of children as it relates to family income. Finally, Section 9 concludes.

## 2 Related literature

This paper contributes to the literature on estimation in nonlinear, parametric, large-$n$ fixed-$T$ panel data models with fixed effects by providing an estimator for the parameters in the fixed effects ordered logit model.

In such models, estimation is complicated due to the incidental parameters problem, see e.g. Lancaster (2000). For a small class of models, model-specific solutions are available. The binary choice logit model is discussed in detail in the next paragraph. Machado (2004) analyzes the binomial regression model with logistic link function. Truncated and censored regression models are discussed by Honoré (1992), and count data models are discussed in Hausman et al. (1984).

Results for restricted classes of models are sometimes available. Hahn (1997) provides results on efficient estimation in panel data models if the density belongs to the exponential class. Wooldridge (1999) considers two classes of models with multiplicative unobserved heterogeneity. Lancaster (2002) proposes a likelihood-based approach on models that allow an orthogonal reparametrization of the incidental parameter. Bonhomme (2012) proposes a functional differencing approach and shows that it is useful for a class of models including random coefficient models and nonlinear regression models. Excellent reviews for the nonlinear panel data literature are e.g. Arellano and Honoré (2001, Sections 4-7) and Arellano and Bonhomme (2012).

Estimation in the fixed effects ordered logit model is closely related to the literature on fixed effects binary choice logit models. Building on Andersen (1970), Chamberlain (1980) discusses CMLE in the fixed effects binary choice logit model and in an unordered discrete choice logistic model. The logistic case is special. Chamberlain (2010) shows that, in the fixed effects binary choice model with unbounded and strictly exogenous regressors, and i.i.d., unit-variance error terms, $\sqrt{n}$ consistent estimation is possible if and only if the error terms follow a logistic distribution. In a two-period setting, Magnac (2004) relaxes the serial independence assumption on the error terms.

This paper is not the first to consider estimation in the fixed effects ordered logit model. Das and van Soest (1999) discuss how to combine the information from several binary choice models into one estimator. Baetschmann et al. (2015) analyze the estimator in Das and van Soest, discuss different ways of aggregating the information from the binary choice models consider by Das and van Soest. They also introduce a composite likelihood estimator for the regression coefficient which is asymptotically less efficient than the estimator in Das and van Soest, but is shown to be preferable in small samples in an extensive simulation study. They also show that the procedure in Ferrer-i-Carbonell and Frijters (2004) is inconsistent. That procedure uses endogenously determined cut points that vary with $i$ but not with $t$.

Of the 14 empirical papers listed in footnote 1, 9 use methods developed for fixed effects model in ordered response models (Baetschmann et al., 2015, appears three times, Ferrer-i-Carbonell and Frijters's procedure is used by five papers, and the Chamberlain approach and the procedure in Das and van Soest are each used once). Four papers use linear panel data models, ignoring the discrete nature of the dependent variable.

A (correlated) random effects ((C)RE) approach is used in five papers. In a CRE model, the unobserved heterogeneity is modelled as a function of time-invariant characteristics, including time-averaged regressors, with an additive error term that is assumed to be independent of the regressors in the model. That model is more restrictive than a fixed effects model, which does not impose any restrictions on the relationship between the unobserved heterogeneity and the regressors. Honoré (2002, Section 2) points out some further drawbacks of the CRE approach in the context of nonlinear models. The main drawback of the CRE approach is that misspecification of the model for the unobserved heterogeneity will produce an inconsistent estimator.

# 3 Model and main result

Section 3.1 introduces the fixed effects ordered logit model. Section 3.2 shows that the model can be transformed into a fixed effects binary choice logit model by using potentially time-varying cutoff categories. This leads to the main result, which establishes the existence of a sufficient statistic for the unobserved heterogeneity in each of the transformed models. Finally, Section 3.3 discusses how the differences of the cut point parameters in the ordered model can be used bound a marginal effect, thus providing an interpretation for the magnitude of the regression coefficient.

### 3.1 Fixed effects ordered logit model

Consider the ordered logit model with additive unobserved heterogeneity in the latent variable,

$$y_{it}^* = \alpha_i + X_{it}\beta + u_{it},\ t = 1, \cdots, T \geq 2. \tag{1}$$

for an individual $i$ with a vector of regressors $X_{it} \in \mathbb{R}^{1\times K}$, and associated vector of regression coefficients $\beta \in \mathbb{R}^{K\times 1}$. The error terms are assumed to be serially independent conditional on the regressors $X_i = (X_{i1}, \cdots, X_{iT})$ and the unobserved heterogeneity $\alpha_i \in \mathbb{R}$, and to follow a standard logistic distribution

$$(u_{i1}, \cdots, u_{iT}) |\, (\alpha_i, X_i) \sim \text{iid } LOG\,(0, 1)\,. \tag{2}$$

The observed, ordered dependent variable $y_{it} \in \{1, \cdots, J\}$ is linked to the latent variable $y_{it}^*$ through cut points $\gamma_j \in \mathbb{R}$ in the following way:

$$y_{it} = \begin{cases} 1 & \text{if } \quad y_{it}^* < \gamma_1, \\ 2 & \text{if } \gamma_1 \leq y_{it}^* < \gamma_2, \\ \vdots & \vdots \\ J & \text{if } \gamma_{J-1} \leq y_{it}^*. \end{cases} \tag{3}$$

We will refer to the model in equations (1)-(3) as the fixed effects ordered logit model.

A random sample of size $n$ on $(y_i, X_i) = (y_{i1}, \cdots, y_{iT}, X_{i1}, \cdots, X_{iT})$ is available for the estimation of the regression coefficient and the cut points. In our asymptotic analysis, the number of cross-section units diverges to infinity. The number of time periods $T$ can be small, as long as $T \geq 2$.

Conditional on the covariates $X_i$ and the unobserved heterogeneity $\alpha_i$, the probability that the ordered dependent variable $y_{it}$ assumes a particular value $j$ is

$$\begin{aligned} P\left(y_{it} = j |\, X_i, \alpha_i\right) &= P\left(\gamma_{j-1} < \alpha_i + X_{it}\beta + u_{it} < \gamma_j |\, X_{it}, \alpha_i\right) \\ &= \Lambda\left(\gamma_j - \alpha_i - X_{it}\beta\right) - \Lambda\left(\gamma_{j-1} - \alpha_i - X_{it}\beta\right), \end{aligned} \tag{4}$$

where $\Lambda(x) = \exp(x) / (1 + \exp(x))$ is the CDF of the logistic distribution, and we have implicitly set $\gamma_0 = -\infty$ and $\gamma_J = +\infty$. The maximum likelihood estimator is affected by the incidental parameters problem (see Lancaster, 2000), since the number of parameters in the likelihood function,

$$L_n(\beta, \delta, \alpha) = \prod_{i=1}^{n} \prod_{t=1}^{T} \prod_{j=1}^{J} \left[\Lambda\left(\gamma_j - \alpha_i - X_{it}\beta\right) - \Lambda\left(\gamma_{j-1} - \alpha_i - X_{it}\beta\right)\right]^{1\{y_{it}=j\}}, \tag{5}$$

grows with the sample size $n$. The estimator for $\beta$ based on (5) will be inconsistent for $n \to \infty$, if $T$ is fixed.

## 3.2 Transformations

The incidental parameters problem can be avoided in models for which a sufficient statistic for the incidental parameters is available.[5] The main result in this section shows that such a sufficient statistic for the unobserved heterogeneity parameter $\alpha_i$ is available for $(J-1)^T$ different transformations of the fixed effects ordered logit model in (1)-(3).

The ordered dependent variable $y_{it}$ can be transformed into a binary variable by checking whether $y_{it} \leq \pi(t)$ for any time series of cutoff categories $\pi(t)$, $t = 1, \cdots, T$. There are $(J-1)^T$ ways of constructing such a transformation $\pi = (\pi(t))_{t=1}^T$. Denote by $d_{i,\pi}$ the binary time series that results from applying transformation $\pi$ to observation $i$,

$$d_{i,\pi} = \left(d_{i,t,\pi} = 1\left\{y_{it} \leq \pi(t)\right\}, \, t = 1, \cdots, T\right).$$

The transformed observation $d_{i,\pi}$ follows the model:

$$\begin{aligned} y_{it}^* &= \alpha_i + X_{it}\beta + u_{it} && (6) \\ u_i | (\alpha_i, X_i) &\sim \text{iid } LOG(0,1) && (7) \\ d_{i,t,\pi} &= 1\left[\alpha_i + X_{it}\beta + u_{it} < \gamma_{\pi(t)}\right]. && (8) \end{aligned}$$

I will refer to the model in (6)-(8) as the $\pi$-transformed fixed effects binary choice logit model. Denote the number of observations below or at the associated cutoff categories by

$$\bar{d}_{i,\pi} = \sum_{t=1}^{T} d_{i,t,\pi}.$$

Furthermore, denote by $p_{i,\pi}(d)$ the probability distribution of $d_{i,\pi}$ conditional on $\bar{d}_{i,\pi}$, written as a function of the regression coefficient $\beta$ of interest and the cut points $\gamma = (\gamma_1, \cdots, \gamma_{J-1})$,

$$p_{i,\pi}(d | \beta, \gamma) \equiv P\left(d_{i,\pi} = d | \, \bar{d}_{i,\pi} = \bar{d}, X_i, \alpha_i\right).$$

Finally, denote by $F_{\bar{d}}$ the set of all binary $T-$vectors that set exactly $\bar{d}$ elements to 1:

$$F_{\bar{d}} = \left\{ f \in \{0,1\}^T \text{ such that } \bar{f} = \bar{d} \right\}.$$

The following theorem formalizes that $\bar{d}_{i,\pi}$ is a sufficient statistic for $\alpha_i$ in the $\pi$-transformed fixed effects binary choice logit model (6)-(8).

**Theorem 1.** *If the random vector $(y_i, X_i)$ follows the fixed effects ordered logit model in equations (1)-(3), then for any transformation $\pi$, the conditional probability distribution of the $\pi$-transformed dependent variable $d_{i,\pi}$ is given by*

[5] Andersen (1970) derives conditions under which a conditional maximum likelihood estimator (CMLE) may be consistent for the common parameters. I use the sufficient statistic in Chamberlain (1980), who extends the insight in Andersen (1970) to the fixed effects binary choice logit model.

| $t$ | $y_{i,t}$ | $d_{i,t,(1,1)}$ | $d_{i,t,(2,2)}$ | $d_{i,t,(1,2)}$ | $d_{i,t,(2,1)}$ |
|---|---|---|---|---|---|
| *Observation 1* | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 0 | 1 | 1 | 0 |
| | $\bar{d}_{1,\pi}$ | 1 | 2 | 2 | 1 |
| *Observation 2* | | | | | |
| 1 | 2 | 0 | 1 | 0 | 1 |
| 2 | 3 | 0 | 0 | 0 | 0 |
| | $\bar{d}_{2,\pi}$ | 0 | 1 | 0 | 1 |
| *Observation 3* | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 3 | 0 | 0 | 0 | 0 |
| | $\bar{d}_{3,\pi}$ | 1 | 1 | 1 | 1 |

**Table 1:** An illustration of the relationship between the ordered variables $y_{it}$, the transformed variables $d_{i,t,\pi}$ and the sum of the transformed variables $\bar{d}_{i,\pi}$ for the case $J = 3$, $T = 2$, and three different cross-section units.

$$p_{i,\pi}\left(d\right|\beta,\gamma) \quad = \quad \frac{1}{\sum_{f\in F_{\bar{d}}}\exp\left\{\sum_t \left(f_t - d_t\right)\left(\gamma_{\pi(t)} - X_{it}\beta\right)\right\}} \tag{9}$$

*for any $d \in \{0,1\}^T$. This conditional probability distribution does not depend on $\alpha_i$.*

*Proof.* Appendix A.1. □

The remainder of this section discusses the main result. If the cutoff categories are time-invariant, then Theorem 1 simplifies to

$$p_{i,\pi}\left(d\right|\beta) = \frac{1}{\sum_{f\in F_{\bar{d}}}\exp\left\{-\sum_t \left(f_t - d_t\right)X_{it}\beta\right\}},$$

which does not depend on the cut points $\gamma_j$. This expression is identical to the conditional likelihood contribution in the fixed effects binary choice logit model (Chamberlain, 1980).

For a transformation with time-varying cutoff categories, the conditional probability depends on the differences in the cut points. Consider the case $T = 2$ and let $\pi = (j,k)$ with $j \neq k$. The (inverse of the) conditional probabilities are

$$p^{-1}_{i,(j,k)}\left((1,0)\right|\beta,\gamma) \quad = \quad 1 + \exp\left\{(\gamma_k - \gamma_j) - (X_{i2} - X_{i1})\beta\right\}, \tag{10}$$

$$p^{-1}_{i,(j,k)}\left((0,1)\right|\beta,\gamma) \quad = \quad 1 + \exp\left\{(\gamma_j - \gamma_k) - (X_{i1} - X_{i2})\beta\right\}. \tag{11}$$

Table 1 illustrates this case, with $J = 3$. Four transformations are available. The first one is $d_{t,(1,1)} = 1\left\{y_t \leq 1\right\}$, where category 1 is chosen as the cutoff category in both time periods. Consider two observations from Table 1: $y_1 = (1,2)$ and $y_2 = (2,3)$. The transformed data is in column $d_{i,t,(1,1)}$. For observation $y_1$, only the $y_{11}$ is at or below the cutoff, so that $d_{1,1,(1,1)} = 1$ and $d_{1,2,(1,1)} = 0$. For $y_2$, both entries exceed the cutoff category, so that $d_{2,1,(1,1)} = d_{2,2,(1,1)} = 0$. For observation 2, there is no variation in the dependent variable over time, which implies that this transformed observation provides no information on the parameters. The table also presents (column 4) the other time-invariant transformation $\pi = (2,2)$.

Next, consider the time-varying transformation $\pi = (1,2)$, which chooses cutoff categories 1 in period 1, and 2 in period 2. For observation $y_1 = (1,2)$, the ordered variable is at the cutoff category in both periods, so that $d_{1,1,(1,2)} = d_{1,2,(1,2)} = 1$. For observation $y_2 = (2,3)$, the dependent variable exceeds the cutoff category in both periods, so that $d_{2,1,(1,2)} = d_{2,2,(1,2)} = 0$. Neither of those transformed observations have variation, but observation $y_3$ does, since $y_3 = (1,3)$ is transformed into $d_{3,(1,2)} = (1,0)$.

This last transformation, $\pi = (1,2)$, can easily be seen to discard information as it reduces the effective sample size from 3 to 1: only one of the three observations has variation in the transformed dependent variable. Transformation $\pi = (2,1)$ (column 6 in Table 1) is inefficient because all the observations are equivalent after applying that transformation. These examples highlight that the choice of transformation determines the source of variation the researcher uses when estimating the parameters of the model.

Interestingly, for observations with no variation in the ordered dependent variable, there may exist transformations that induce variation in the transformed dependent variable. This happens when the ordered variable is constant, and not equal to 1 or $J$. An example is $y_4 = (2,2)$ and $\pi = (2,1)$. Then $d_{4,(2,2)} = (1,0)$, so that the transformed variable is not constant over time, even though we started from an ordered variable that did not exhibit any variation.

### 3.3 Cut points

In this Section, I show that one can use knowledge of the cut point differences to estimate the minimum required change in the $m$-th regressor to move an arbitrary cross-section unit with $y_{it} = j$ to a category higher than $j$.

An interpretation of the magnitude of the regression coefficient is not available when cut point differences are unknown. This is related to a well-known drawback of the fixed effects binary choice logit model, see e.g. the discussion in Chamberlain (1984, p. 1277) and Wooldridge (2010, p. 622). Consider the marginal effect of a ceteris paribus change in regressor $m$ on the probability that the dependent variable for individual $i$ in period $t$ is at or below $j$:

$$\frac{\partial P\left(y_{it} \leq j \middle| X_{it}, \alpha_i\right)}{\partial X_{it,m}} = \beta_m \Lambda\left(\alpha_i + X_{it}\beta - \gamma_j\right)\left[1 - \Lambda\left(\alpha_i + X_{it}\beta - \gamma_j\right)\right], \qquad (12)$$

where $\beta_m$ is the regression coefficient associated with the $m$-th regressor. This marginal effect depends on the unobserved heterogeneity $\alpha_i$. Although the sign of the regression coefficient determines the sign of (12), the magnitude of the marginal effect unknown in the absence of knowledge on $\alpha_i$.[6] The inability to compute marginal effects for the binary choice model is serious obstacle to its use in empirical applications.

Consider the ordered logit model with $J = 2$ (binary choice). Knowledge of $y_{it}$ tells us whether the latent variable $y_{it}^*$ is to the left or to the right of the cut point $\gamma_1$, but it does not tell us how far away the latent variable is from that cut point. Consequentially, for an arbitrarily large change in $X_{it}$, we can not be sure that the counterfactual latent variable crosses the cut point.

In the ordered logit model with $J > 2$, the same reasoning applies to an observation that falls in one of the extreme categories, $y_{it} \in \{1, J\}$. For such an observation, the associated latent variable $y_{it}^*$ can be arbitrarily far away from the closest cut point ($\gamma_1$ or $\gamma_{J-1}$). In the absence of knowledge on $\alpha_i + u_{it}$, there is no change in $X_{it}\beta$ large enough to guarantee that the latent variable moves across the cut point.

The situation is different when we observe a dependent variable in an intermediate category, $y_{it} = j \in \{2, \cdots, J-1\}$. For such an observation, we know that the latent variable $y_{it}^*$ is in the finite interval $(\gamma_{j-1}, \gamma_j)$. This allows us to bound the marginal effect. For example, a ceteris paribus change in $X_{it}\beta$ that exceeds $\gamma_j - \gamma_{j-1}$ will move $y_{it}^*$ across one of the cut points. By considering the required change in $X_{it}$ to guarantee such a crossing, this bound provides an interpretation of the strength of the relationship between $X_{it}$ and the $y_{it}$. This opportunity is not available for currently available estimators, because they do not yield a consistent estimator of the cut point differences.

To formalize this interpretation, consider an individual $i$ who is in an intermediate category $j$ at time $t$, i.e. $y_{it} = j \in \{2, \cdots, J-1\}$.[7] For such an individual,

$$y_{it}^* = X_{it}\beta + v_{it} \in (\gamma_{j-1}, \gamma_j)\,, \tag{13}$$

where $v_{it}$ is the composite error term $v_{it} = \alpha_i + u_{it}$. A ceteris paribus change in the regressors of $\Delta x$ induces the counterfactual variables

$$\begin{aligned}
\tilde{X}_{it} &= X_{it} + \Delta x, \\
\tilde{y}_{it}^* &= y_{it}^* + (\Delta x)\,\beta \\
&= (X_{it} + \Delta x)\,\beta + v_{it}, \\
\tilde{y}_{it} &= \sum_{j=1}^{J} j \cdot 1\left\{\gamma_{j-1} < \tilde{y}_{it}^* < \gamma_j\right\}.
\end{aligned}$$

The object of interest is the distribution of the counterfactual dependent variable, conditional on the observable random variables $y_{it}$ and $X_{it}$. Let $F_v$ denote the unknown distribution of the composite error term $v_{it} = \alpha_i + u_{it}$, conditional on $X_i$. In Appendix A.2, I obtain the following result for the conditional probability that the counterfactual

[6] The expected marginal effect would require knowledge of the conditional distribution of$\alpha_i$.

[7] The interpretation is only available for a dependent variable in the intermediate categories, and not for $y_{it} \in \{1, J\}$.

dependent variable $\tilde{y}_{it}$ is in a category strictly larger than $j$:[8]

$$P\left(\tilde{y}_{it} > j \middle| y_{it} = j, X_{it}\right) = \begin{cases} 1 & \text{if } (\Delta x)\,\beta > \gamma_j - \gamma_{j-1}, \\ 0 & \text{if } (\Delta x)\,\beta < 0, \\ \frac{F_v(\gamma_j - X_{it}\beta) - F_v(\gamma_j - (X_{it}+\Delta x)\beta)}{F_v(\gamma_j - X_{it}\beta) - F_v(\gamma_{j-1} - X_{it}\beta)} & \text{otherwise.} \end{cases} \tag{14}$$

The first component states that we can be certain that the counterfactual dependent variable is greater than $j$ if the change in the latent variable due to $\Delta x$ is large enough.[9] To turn this into a useful quantity, consider a change of $\Delta x_m$ in the $m$-th regressor $X_{itm}$, with positive regression coefficient $\beta_m > 0$. Introduce the quantity

$$\delta_m^j = \frac{\gamma_j - \gamma_{j-1}}{\beta_m}. \tag{15}$$

The result in (14) implies that

$$\Delta x_m > \delta_m^j \Rightarrow P\left(\tilde{y}_{it} > j \middle| y_{it} = j, X_{it}\right) = 1.$$

As a result, we can interpret $\delta_m^j$ as the minimum required change in $X_{itm}$ to move an arbitrary observation $y_{it} = j$ to a higher category. A small value of $\delta_m^j$ means that the relationship between $X_{it}$ and $y_{it}$ is strong at category $j$. Larger values for $\delta_m^j$ can be due to the distance from category $j$ is far away from $j+1$, or because the $m$-th regressor has little impact on $y_{it}$.[10]

# 4 Conditional maximum likelihood estimation

This section analyzes identification and estimation in one arbitrary $\pi$-transformed fixed effects binary choice model (6)-(8).

[8] The Appendix contains some additional results. For example, an expression similar to the one that follows can be obtained for $P\left(\tilde{y}_{it} = k \middle| y_{it} = j, X_{it} = x\right)$. Here, we focus on the conditional probability $P\left(\tilde{y}_{it} > j \middle| y_{it} = j, X_{it}\right)$ because it leads to an easily interpretable quantity.

[9] The second and third components of this display are not informative. The second component states that if we start from category $j$, and decrease the value of the latent variable, the counterfactual dependent variable will not be greater than $j$. The third component is uninformative because it requires several evaluations of the unknown function $F_v$. For this reason, we will focus on the first component.

[10] An alternative interpretation of the cut point differences is available. Denote the odds ratio for category $j$ by

$$\omega_j\left(X_{it}\right) = \frac{P\left(y_{it} \leq j \middle| X_{it}, \alpha_i\right)}{P\left(y_{it} > j \middle| X_{it}, \alpha_i\right)} = \exp\left(\gamma_j - \alpha_i - X_{it}\beta\right).$$

The ratio of the odds ratios of two categories $j$ and $k$, $j \neq k$ is a function of the cut point difference only:

$$\log\left[\frac{\omega_j\left(X_{it}\right)}{\omega_k\left(X_{it}\right)}\right] = \gamma_j - \gamma_k.$$

This formalizes the idea that, conditional on $X_{it}$, the cut point differences measure the distance between two categories.

The conditional probability (9) in Theorem 1 depends on $\gamma$ only through the differences in the cut points used by transformation $\pi$. To see this, note that the number of non-zero entries for any

$$f \in F_{\bar{d}} = \left\{ f \in \{0,1\}^T \text{ such that } \bar{f} = \bar{d} \right\}.$$

is equal to the number of non-zero entries in $d$. For any $f \in F_{\bar{d}}$, then, we have that $\gamma_{\pi(1)} \sum_t (f_t - d_t) = 0$. This allows us to rewrite the conditional probability in terms of cut point differences $\gamma_{\pi(t)} - \gamma_{\pi(1)}$:

$$\begin{aligned} p_{i,\pi}^{-1}(d|\,\beta,\gamma) &= \sum_{f\in F_{\bar{d}}} \exp\left\{ \sum_t (f_t - d_t)\,\gamma_{\pi(t)} - \sum_t (f_t - d_t)\,X_{it}\beta \right\} \\ &= \sum_{f\in F_{\bar{d}}} \exp\left\{ \sum_t (f_t - d_t)\left(\gamma_{\pi(t)} - \gamma_{\pi(1)}\right) - \sum_t (f_t - d_t)\,X_{it}\beta \right\}. \quad (16) \end{aligned}$$

The estimand in the $\pi$-transformed fixed effects binary choice logit model (6)-(8) therefore consists of the regression coefficient $\beta$ and a subset of the differences of the cut points.

To formally define the estimand, let

$$\tilde{\gamma}_{\pi,\Delta} = \left(\gamma_{\pi(1)} - \gamma_{\pi(1)}, \cdots, \gamma_{\pi(T)} - \gamma_{\pi(1)}\right)$$

be the $T \times 1$ vector of cut point differences that appear in (16). Denote by $n_\pi$ the number of unique, non-zero elements in $\tilde{\gamma}_{\pi,\Delta}$. Collect those elements in an $n_\pi \times 1$ vector $\gamma_{\pi,\Delta}$.[11] The $T \times n_\pi$ selection matrix $S_\pi$ transforms the unique elements in $\gamma_{\pi,\Delta}$ into the time series of cut point differences $\tilde{\gamma}_{\pi,\Delta}$:

$$\tilde{\gamma}_{\pi,\Delta} = S_\pi \gamma_{\pi,\Delta}.$$

Adjoin the selection matrix $S_\pi$ to the stacked regressor matrix to obtain the $T \times (K + n_\pi)$ set of augmented regressors $Z_{i\pi} = \left[-X_i \quad | \quad S_\pi\right]$. The associated $(K + n_\pi) \times 1$ vector of augmented regression coefficients $\theta_\pi = (\beta, \gamma_{\pi,\Delta})$ is the parameter of interest for the $\pi$-transformed fixed effects binary choice model. The conditional probability in Theorem 1 can now be rewritten as a function of the estimand:

$$\begin{aligned} p_{i,\pi}^{-1}(d|\,\theta_\pi) &= \sum_{f\in F_{\bar{d}}} \exp\left\{ \sum_t (f_t - d_t)\left(\gamma_{\pi(t)} - \gamma_{\pi(1)} - X_{it}\beta\right) \right\} \\ &= \sum_{f\in F_{\bar{d}}} \exp\left\{ (f - d)\, Z_{i\pi}\theta_\pi \right\}. \quad (17) \end{aligned}$$

[11] A time-invariant transformation $\pi$ sets $\pi(t) = j$ for all $t$, so that the cut point difference parameter $\gamma_{\pi,\Delta}$ is empty, and $n_\pi = 0$.

Then, the conditional maximum likelihood estimator (CMLE) can be defined as:

$$\hat{\theta}_\pi = \left(\hat{\beta}, \hat{\gamma}_{\pi,\Delta}\right) = \arg\max_{\mathbb{R}^K \times \mathbb{R}^{n_\pi}} \frac{1}{n} \sum_{i=1}^{n} 1\left\{d_i = d\right\} \ln p_{i\pi}\left(d \middle| \theta_\pi\right), \tag{18}$$

as an estimator for $\theta_{\pi,0} = (\beta_0, \gamma_{\pi,\Delta,0})$, the true value of the parameters in the $\pi$-transformed binary choice model implied by the ordered logit model with true parameter values $(\beta_0, \gamma_0)$.

The upcoming proof of consistency of the CMLE relies on the concavity of the criterion function in (18). Let

$$\begin{aligned} s_{i,\pi}\left(d \middle| \theta_\pi\right) &= \frac{\partial \ln p_{i,\pi}\left(d \middle| \theta_\pi\right)}{\partial \theta_\pi}, \\ H_{i,\pi}\left(d \middle| \theta_\pi\right) &= \frac{\partial^2 \ln p_{i,\pi}\left(d \middle| \theta_\pi\right)}{\partial \theta_\pi \partial \theta_\pi'} \end{aligned}$$

be the contribution to the score vector and the Hessian matrix for an individual $i$ with $d_i = d$. In Appendix A.3, I show that

$$s_{i,\pi}\left(d \middle| \theta_\pi\right) = -\frac{\sum_{f \in F_{\bar{d}}} \exp\left(\left(f - d\right) Z_{i\pi} \theta_\pi\right) Z_{i\pi}' \left(f - d\right)'}{\sum_{f \in F_{\bar{d}}} \exp\left(\left(f - d\right) Z_{i\pi} \theta_\pi\right)} \tag{19}$$

and

$$\begin{aligned} & H_{i,\pi}\left(d \middle| \theta_\pi\right) \\ & = -\frac{1}{2} \frac{\sum_{f,g \in F_{\bar{d}}} \exp\left\{\left(f - d\right) Z_{i\pi} \theta_\pi\right\} \exp\left\{\left(g - d\right) Z_{i\pi} \theta_\pi\right\}}{\left(\sum_{f \in F_{\bar{d}}} \exp\left(\left(f - d\right) Z_{i\pi} \theta_\pi\right)\right)^2} Z_{i,\pi}' \left(g - f\right)' \left(g - f\right) Z_{i\pi}. \end{aligned} \tag{20}$$

The second derivative (20) is negative semidefinite: both the denominator and the numerator are positive, and the nonscalar part is an outer product. It follows that the criterion function (18) is globally concave, which helps me to establish consistency and asymptotic normality under fairly weak conditions.

To state a sufficient condition for consistency and asymptotic normality of the CMLE for $\theta_{\pi,0}$, let

$$\tilde{X}_i = \operatorname{vec}\left(X_i'\right) = \begin{bmatrix} X_{i1}' \\ \vdots \\ X_{iT}' \end{bmatrix}$$

be the $KT \times 1$ column vector that stacks the $T$ blocks of $K$ regressors for individual $i$.

**Assumption 2.** *The variance matrix of the regressors, Var $\left(\tilde{X}_i\right)$, exists and is positive definite.*

Assumption 2 guarantees that there is sufficient variation in the regressors for a given individual across time. In particular, it implies that for any two time periods $s \neq t$ , the variance matrix of the difference $X_{it} - X_{is}$ is of full rank. This rules out regressors that are constant across a subset of the sample period, and rules out sets of regressors that perfectly co-vary. This assumption could be relaxed, as identification of the regression coefficient only requires two time periods.

To state the main result in this section, some additional notation is required. First, denote by

$$s_{i,\pi}\left(\theta_\pi\right) = \sum_{d \in \{0,1\}^T} 1\left\{d_i = d\right\} s_{i,\pi}\left(d \middle| \theta_\pi\right)$$

the score contribution of individual $i$, and let $\Sigma_\pi$ be the variance of the score at the true value of the parameters,

$$\Sigma_\pi = E\left(s_{i,\pi}\left(\theta_{\pi,0}\right) s_{i,\pi}\left(\theta_{\pi,0}\right)'\right). \tag{21}$$

Denote the Hessian by

$$H_\pi\left(\theta_\pi\right) = E\left(\sum_{d \in \{0,1\}^T} 1\left\{d_i = d\right\} H_{i,\pi}\left(d \middle| \theta_\pi\right)\right) \tag{22}$$

and write $H_\pi = H_\pi\left(\theta_{\pi,0}\right)$ when the Hessian is evaluated at $\theta_{\pi,0}$.

**Theorem 3.** *Let $\left(\{y_i, X_i\}, i = 1, \cdots, n\right)$ be a random sample from the fixed effects ordered logit model (1)-(3) with true parameter values $(\beta_0, \gamma_0)$, and let $\pi$ be an arbitrary transformation. If Assumption 2 holds, then (i) the CMLE $\hat{\theta}_\pi$ in equation (18) is consistent for $\theta_{\pi,0}$;*

$$\hat{\theta}_\pi \xrightarrow{p} \theta_{\pi,0} \text{ as } n \to \infty;$$

*(ii) a central limit theorem applies to the score,*

$$\frac{1}{\sqrt{n}} \sum_i s_{i,\pi}\left(\theta_{\pi,0}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\pi\right); \tag{23}$$

*(iii) the Hessian $H_\pi\left(\theta_\pi\right)$ exists and is invertible for all $\theta_\pi$; (iv) the CMLE estimator $\hat{\theta}_\pi$ in equation (18) has the following limiting distribution:*

$$\sqrt{n}\left(\hat{\theta}_{\pi,n} - \theta_{\pi,0}\right) \xrightarrow{d} \mathcal{N}\left(0, H_\pi^{-1} \Sigma_\pi H_\pi^{-1}\right) \text{ as } n \to \infty. \tag{24}$$

*Proof.* Part (i) in Appendix A.4, parts (ii)-(iv) in Appendix A.5. □

Results (i) and (iv) of Theorem 3 describe the asymptotic behavior of the CMLE based on a single transformation $\pi$. Parts (ii) and (iii) are intermediate results for a standard proof of asymptotic normality of an extremum estimator. They are stated here because they are essential ingredients for the efficiency result in the next section.

# 5 Efficiency

In this section, I introduce a class of generalized method of moments (GMM) estimators that incorporate the information from all $(J-1)^T$ transformations that can be applied to the fixed effects ordered logit model. The optimal estimator in this class yields an estimator for the regression coefficient $\beta_0$ that is at least as efficient as existing estimators. I also show that it is asymptotically equivalent to the optimal linear combination of all $(J-1)^T$ CMLE estimators, or optimal minimum distance estimator (OMD).

For an arbitrary transformation $\pi$, the estimand targeted by the CMLE is the maximizer

$$\theta_{\pi,0} \quad = \quad \arg\max_{\mathbb{R}^K \times \mathbb{R}^{n_\pi}} E\left[\sum_{d\in\{0,1\}^T} 1\left[d_i = d\right] \ln p_\pi\left(d \middle| \theta_\pi\right)\right]. \tag{25}$$

The representation in terms of the first order conditions is the first order condition:

$$E\left[s_{i\pi}\left(\theta_{\pi,0}\right)\right] = 0. \tag{26}$$

It will be useful to separate the role of the regression coefficient from that of the cut point differences:

$$E\left[s_{i\pi}\left(\theta_\pi\right)\right] = E\begin{bmatrix} s_{i,\pi,\beta}\left(\beta,\gamma_{\pi,\Delta}\right) \\ s_{i,\pi,\gamma}\left(\beta,\gamma_{\pi,\Delta}\right)\end{bmatrix} = E\begin{bmatrix} \partial\ln p_{i\pi}\left(\beta,\gamma_{\pi,\Delta}\right)/\partial\beta \\ \partial\ln p_{i\pi}\left(\beta,\gamma_{\pi,\Delta}\right)/\partial\gamma_{\pi,\Delta}\end{bmatrix}.$$

From this perspective, a transformation $\pi$ provides $K+n_\pi$ first order conditions for $K+n_\pi$ parameters.[12] These first order conditions are moment conditions in the GMM framework. In what follows, we will gather the moment conditions from all transformations.

The set of all $(J-1)^T$ transformations can be separated into a set of $J-1$ time-invariant transformations,[13] and the remaining, time-varying, transformations. We will consider them separately, starting with the time-invariant transformations. For a time-invariant transformation, the number of unique cut point differences is $n_\pi = 0$, so that the parameter of interest consists only of the regression coefficient $\theta_{\pi,0} = \beta_0$. For a single time-invariant transformation, the restriction on the $K\times 1$ vector of scores $s_{i,\pi,\beta}$ exactly identifies $\beta_0$. Taken together, the $(J-1)$ sets of restrictions on the score vectors from the time-invariant transformations overidentify the regression coefficient $\beta_0$ through the $K(J-1)$ moment conditions:

$$E\left[s_{i,1,\beta}\left(\beta_0\right)\right] = E\begin{bmatrix} s_{i,(1,\cdots,1)}\left(\beta_0\right) \\ \vdots \\ s_{i,(J-1,\cdots,J-1)}\left(\beta_0\right)\end{bmatrix} = 0, \tag{27}$$

[12] The derivative $s_{i\pi,\beta}$ gives $K$ moment conditions, and there are $n_\pi$ moment conditions from $s_{i\pi,\gamma}$. There are $K$ elements in the parameter $\beta$, and $n_\pi$ elements in $\gamma_{\pi,\Delta}$.

[13] A time-invariant transformation is one that sets $\pi(t) = j$, $j \in \{1,\cdots,J-1\}$ for all $t$.

if $J > 2$. A GMM estimator based on (27) takes the form

$$\tilde{\beta}_{W_{1,n}} = \arg\min \bar{s}_{1,n}(\beta)' W_{1,n} \bar{s}_{1,n}(\beta), \tag{28}$$

where $\bar{s}_{1,n}(\beta) = \frac{1}{n}\sum_{i=1}^{n} s_{i,1,\beta}(\beta)$, and $W_{1,n}$ is a weight matrix.

Existing consistent estimators for the regression coefficient correspond to different choices for $W_{1,n}$. Setting $W_{1,n} = e_j \otimes I_K$ corresponds to using the time-invariant transformation $\pi(t) = j$. The blow-up-and-cluster (BUC) estimator advocated by Baetschmann et al. (2015) sets $W_{1,n}$ equal to a blockdiagonal matrix with its $j$-th block equal to the the inverse of the Hessian associated with transformation $\pi(t) = j$ (Baetschmann et al., 2015, p. 691). Denote by $\tilde{\beta}^*$ the asymptotically efficient GMM estimator in the class 28, which sets $\text{plim}W_{1,n} = E\left[s_{i,1,\beta}(\beta_0) s_{i,1,\beta}(\beta_0)'\right]^{-1}$. That estimator is asymptotically equivalent to the minimum distance estimator proposed by Das and van Soest (1999), see also Remark 5.

The moment conditions implied by the time-invariant transformations (27) do not exhaust the information in the fixed effects ordered logit model. Each time-varying transformation implies $K + n_\pi$ additional moment conditions of the form (26). These moment conditions involve both $\beta_0$ and $\gamma_{\pi,\Delta,0}$. Gather the cut point difference parameters from all time-varying transformations in $\gamma_\Delta = (\gamma_{\pi,\Delta})_\pi$, a column vector with $n_\gamma \equiv \sum_\pi n_\pi$ elements.[14] Similarly, collect the scores

$$s_{i,2,\gamma}(\beta, \gamma_\Delta) = (s_{i,\pi,\gamma}(\beta, \gamma_{\pi,\Delta}), \pi : n_\pi \geq 1)$$

in a $n_\gamma \times 1$ vector. The scores for the regression coefficients from the time-varying transformations are collected in

$$s_{i,2,\beta}(\beta, \gamma_\Delta) = (s_{i,\pi,\beta}(\beta, \gamma_{\pi,\Delta}), \pi : n_\pi \geq 1).$$

Taking together the restrictions on the scores from all $(J-1)^T$ transformations, we obtain

$$E[s_i(\beta_0, \gamma_{\Delta,0})] = E\begin{bmatrix} s_{i,1,\beta}(\beta_0) \\ s_{i,2,\beta}(\beta_0, \gamma_{\Delta,0}) \\ s_{i,2,\gamma}(\beta_0, \gamma_{\Delta,0}) \end{bmatrix} = 0, \tag{29}$$

which has dimension $K(J-1)^T + n_\gamma$ for the $K + n_\gamma$-dimensional parameter $(\beta, \gamma_\Delta)$. A GMM estimator based on (29) takes the form

$$\left(\hat{\beta}_{W,n}, \hat{\gamma}_{\Delta,W,n}\right) = \arg\min \bar{s}_n(\beta, \gamma_\Delta)' W_n \bar{s}_n(\beta, \gamma_\Delta). \tag{30}$$

where $\bar{s}_n(\beta, \gamma_\Delta) = \frac{1}{n}\sum_{i=1}^{n} s_i(\beta, \gamma_\Delta)$ is the sample analog of the moments in (29). Denote by $\left(\hat{\beta}^*, \hat{\gamma}^*_\Delta\right)$ the asymptotically efficient estimator in the class of estimators of the form (30).

[14] Taking into account the relationship between the components in $\gamma_\Delta$ may lead to a more efficient estimation procedure. For the purpose of this section - showing efficiency increases for the regression coefficient $\beta$ - it is useful to consider each $\gamma_{\pi,\Delta}$ as a separate parameter.

Theorem 4 establishes the asymptotic distribution of $\left(\hat{\beta}^{*}, \hat{\gamma}_{\Delta}^{*}\right)$, and shows that $\hat{\beta}^{*}$ is at least as efficient as $\tilde{\beta}^{*}$, so that adding the moment conditions from the time-varying transformations reduces the asymptotic variance of the regression coefficient estimator. For a formal result, some additional notation is needed. First, denote by

$$\Sigma = E\left[s_i\left(\beta_0, \gamma_{\Delta,0}\right) s_i\left(\beta_0, \gamma_{\Delta,0}\right)^{'}\right] \tag{31}$$

the variance of the entire set of scores (29), and by

$$\Sigma_1 = E\left[s_{i,1,\beta}\left(\beta_0, \gamma_{\Delta,0}\right) s_{i,1,\beta}\left(\beta_0, \gamma_{\Delta,0}\right)^{'}\right]$$

the variance of the restricted set of scores from time-invariant transformations in (27).[15] The diagonal blocks are the transformation-specific variance matrices $\Sigma_\pi$ from (21). Furthermore, denote by $H$ the expected derivative of the scores in (27) with respect to $(\beta, \gamma_\Delta)$, evaluated at the true values of the parameter.[16] Finally, $H_1$ consists of the top left $K\left(J-1\right) \times K$ block of $H$, which stacks the Hessians $H_\pi$ of the time-invariant transformations.

**Theorem 4.** *Let* $\left(\{y_i, X_i\}, i = 1, \cdots, n\right)$ *be a random sample from the fixed effects ordered logit model (1)-(3) with true parameter values* $(\beta_0, \gamma_0)$*, and let Assumption 2 hold. Then, as* $n \to \infty$*,*

$$\begin{aligned} \sqrt{n}\left(\tilde{\beta}^{*} - \beta_0\right) &\overset{d}{\to} \mathcal{N}\left(0, V_1\right), \\ \sqrt{n}\left(\begin{pmatrix} \hat{\beta}^{*} \\ \hat{\gamma}_{\Delta}^{*} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \gamma_{\Delta,0} \end{pmatrix}\right) &\overset{d}{\to} \mathcal{N}\left(0, V\right), \end{aligned}$$

*where*

$$\begin{aligned} V_1 &= \left(H_1^{'}\Sigma_1^{-1}H_1\right)^{-1}, \\ V &= \left(H^{'}\Sigma H\right)^{-1}. \end{aligned}$$

*Furthermore, let* $V_\beta$ *be the top-left* $K \times K$ *block of* $V$*. Then* $V_1 - V_\beta$ *is positive semidefinite.*

*Proof.* The restrictions on $s_{i,2,\gamma}$ exactly identify the cut point differences $\gamma_{\Delta,0}$. Therefore, by Theorem 1 in Ahn and Schmidt (1995), estimation of $\beta_0$ using the restrictions on $s_{i,1,\beta}$ is equivalent to estimation of $(\beta_0, \gamma_{\pi,0})$ using the restrictions on $s_{i,1,\beta}$ and $s_{i,2,\gamma}$. The restrictions on $s_{i,2,\beta}$ therefore provide additional information on $\beta_0$ that is not used for estimation of the additional cut point parameters. A detailed proof is in Appendix A.6. □

[15] The matrix $\Sigma_1$ is the top-left $K\left(J-1\right) \times K\left(J-1\right)$ corner of $\Sigma$.

[16] This matrix has a specific structure, which is omitted for the sake of brevity, but which can be found in Appendix A.7.1.

Theorem 4 states that, for the regression coefficient $\beta_0$, estimation based on all transformations is at least as efficient as estimation based on the time-invariant transformations only. Therefore, the amount of information gained by considering the first order conditions from the time-invariant transformations is greater than or equal to the amount of information required for estimating the additional cut point parameters. Note that this procedure also yields an estimator $\hat{\gamma}^*_\Delta$ for the cut point differences. Finally, note that the estimation procedure can likely be improved by taking into account the relationship between the components of $\gamma_{\Delta,0}$.

*Remark* 5. In Appendix A.7, I show that the optimal minimum distance (MD) estimator based on all $(J-1)^T$ CMLE estimators is asymptotically equivalent to the optimal GMM estimator $\left(\hat{\beta}^*, \hat{\gamma}^*_\Delta\right)$ in Theorem 4. This provides an alternative way to combine the information in all the transformations, which will prove useful for implementation, see Section 6.[17]

*Remark* 6. By inspecting the expression for the optimal weights in Appendix A.7, it can be seen that the weight given to $\hat{\beta}_\pi$ is

$$W_\pi \equiv \left(H^{'}\Sigma^{-1}H\right)^{-1} H^{'}\Sigma^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ H^{\beta\gamma}_{\pi_J} \\ H^{\gamma}_{\pi_J} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This weight will not be equal to zero since $H$ and $\Sigma$ have full rank because of the logistic errors. Since the OMD estimator assigns non-zero weights to the CMLE for the regression coefficient from time-varying transformations. This means that it is strictly more efficient than the OMD procedure in Das and van Soest (1999), which implicitly assigns zero weight to the estimators based on time-varying transformations. By Remark 5, the efficient GMM and OMD estimators based on all transformations are strictly more efficient than the GMM and OMD estimators based on the time-invariant transformations.

# 6 Implementation

This section describes the implementation of the estimators described in Sections 4 and 5. First, I show how the CMLE estimator based on a single, potentially time-

[17] This extends the insight in Baetschmann et al. (2015, p. 691), who only consider time-invariant transformations. They show that the MD estimator in Das and van Soest (1999) is equivalent to the GMM estimator $\tilde{\beta}^*$ in Theorem 4.

varying transformation from Section 4 can be implemented when a computer program for the fixed effects binary choice logit model is available. Second, I discuss how the asymptotically efficient minimum distance estimator in Section 5, Remark 5, can be implemented using standard methods for minimum distance estimators. Additionally, I discuss a composite likelihood estimator (CLE) that is not asymptotically efficient, but addresses a drawback of the optimal minimum distance (OMD) estimator. This CLE extends the blow-up-and-cluster (BUC) estimator in Baetschmann et al. (2015).

All three estimators are easy to implement in Stata (StataCorp, 2015). The CMLE estimator can be obtained using the *clogit* command.[18] The OMD estimator can be obtained using the *suest* command after several calls to *clogit*. The composite likelihood estimator can be obtained by running *clogit* after duplicating the data set using *expand*. This extends the insight by Baetschmann et al. (2015) to time-varying transformations.

## 6.1 CMLE and OMD

If a binary dependent variable $d_i$ and regressors $X_i$ follow a fixed effects binary choice logit model, the conditional probability that forms the basis for the CMLE in Chamberlain (1980), and implemented in Stata's *clogit* is given by

$$p_i^{-1}\left(d|\,\beta\right) = 1 + \exp\left\{-\sum_t \left(f_t - d_t\right) X_{it}\beta\right\}. \tag{32}$$

Now, consider an observation $(y_i, X_i)$ from the fixed effects ordered logit model. The conditional probability associated with the $\pi$-transformed model is given by Theorem 1, equation (9):

$$p_{i,\pi}^{-1}\left(d|\,\beta,\gamma\right) = \sum_{f\in F_{\bar{d}}} \exp\left\{\sum_t \left(f_t - d_t\right)\left(\gamma_{\pi(t)} - X_{it}\beta\right)\right\}. \tag{33}$$

For a time-invariant transformation $\pi$, the cut points $\gamma_{\pi(t)}$ drop out, and the conditional probability (33) is identical to that of the fixed effects binary choice logit model in (32). In that case, the CMLE for the regression coefficient using the $\pi$-transformed model is identical to the CMLE for the binary choice model applied to $d_{i\pi}$.

For a time-varying transformation $\pi$, the conditional probability in equation (33) features the additional term $\sum_t \left(f_t - d_t\right)\gamma_{\pi(t)}$. In Section 4, equation (17), I showed that the conditional probability can be rewritten as

$$p_{i,\pi}^{-1}\left(d|\,\theta_\pi\right) \;=\; \sum_{f\in F_{\bar{d}}} \exp\left\{\left(f-d\right) Z_{i\pi}\theta_\pi\right\},$$

[18] The *clogit* command comes with an option that allows the empirical researcher to cluster standard errors. This clustering will carry over to the ordered logit estimators discussed in this paper. However, care should be exercised when choosing this option. Please refer to Section VII.C.2 in Cameron and Miller (2015) for a discussion of cluster-robust inference in nonlinear fixed effects models. Additionally, note that serial correlation is explicitly ruled out by the fixed effects ordered logit model discussed here.

where $Z_{i\pi}$ is the set of augmented regressors $Z_{i\pi} = \begin{bmatrix} -X_i & | & S_\pi \end{bmatrix}$. As a result, the conditional probability for the $\pi$-transformed model is equivalent to the CMLE for a binary choice fixed effects logit model with the columns in $S_\pi$ as additional regressors. These additional regressors indicate which cut points are used to $\pi$-transform the dependent variable. The coefficient estimates for the elements of $S_\pi$ are estimates for the cut point differences $\gamma_{\pi(t)} - \gamma_{\pi(1)}$.

The OMD estimator in Section 5, Remark 5, is the optimal linear combination of all CMLE estimators. The optimal weights depend on the variance matrix of vector of all CMLE estimators. Stata's *suest* command can be used to estimate that variance matrix. Therefore one can produce the OMD estimator using a simple program that uses *suest* after $(J-1)^T$ calls to clogit. A more detailed description can be found in Appendix A.7.

## 6.2 Composite likelihood estimator

The number of CMLE estimators $(J-1)^T$ can be very large, even for moderate values of $J$ and $T$. The OMD estimator requires an estimate of the variance matrix of these estimators. When the number of entries in that variance matrix approaches or exceeds the number of observations $n$, we expect poor finite-sample performance of the OMD estimator.

Recent contributions to the literature on estimation with many moments, and on the combination of many estimators include Han and Phillips (2006) and Chen et al. (2014). The OMD procedure discussed in Chen et al. (2014) is similar to the estimator introduced in Section 5, the implementation of which is described in the previous subsection. Chen et al. (2014) discuss conditions on the rate of growth of the number of estimators with the sample size under which standard inference applies (see for example their Theorem 3). In this paper, I assume that the number of categories $J$ and the number of time periods $T$ are fixed, so that the conditions in their paper are trivially satisfied.

The composite likelihood estimator (CLE) introduced in this section is an alternative procedure that also incorporates the information from all transformations, but that avoids the estimation of the large variance matrix. The CLE has the added advantage that it imposes the relationship between the elements of $\gamma_\Delta$.[19] The drawback of the CLE is that it sacrifices asymptotic efficiency.

The CLE is defined as the maximizer of the sum of the criterion functions for all the CMLEs, i.e.

$$\hat{\theta}_{cle} = \arg \max_{\mathbb{R}^K \times \mathbb{R}^{n_\pi}} - \sum_\pi \sum_{i=1}^n 1_{\{d_i = d\}} \ln \sum_{f \in F_{\bar{d}}} \exp \left\{ \sum_t (f_t - d_t) \left( \gamma_{\pi(t)} - \gamma_{\pi(1)} - X_{it}\beta \right) \right\}. \tag{34}$$

[19] For example, if $J=3$, $T=2$, then $\gamma_{(1,2),\Delta} = \gamma_2 - \gamma_1$ and $\gamma_{(2,1),\Delta} = \gamma_1 - \gamma_2$ so that $\gamma_{(1,2),\Delta} = \gamma_{(2,1),\Delta}$. The CLE discussed in this section automatically imposes these restrictions.

By imposing the normalization restriction $\gamma_1 = 0$, we can interpret $\gamma_j$ as the difference of the $j$-th cutpoint with respect to $\gamma_1$. This estimator

$$\hat{\theta}_{cle} = \left(\hat{\beta}_{cle}, \hat{\gamma}_{2,cle}, \cdots, \hat{\gamma}_{J-1,cle}\right)$$

estimates the regression coefficient and all the cut point differences.

The CLE is an extension of the "blow up and cluster estimator" (BUC) estimator described in Baetschmann et al. (2015). The difference between the CLE and the BUC estimator is that the CLE estimator takes into account all transformations, whereas the BUC estimator uses time-invariant transformations only. Baetschmann et al. (2015, p. 690) discuss the properties of the BUC estimator. By examining the first order conditions of the objective function, they show that the BUC estimator is a consistent and asymptotically normal GMM estimator with non-optimal weight matrix. As a result, the CLE is not asymptotically efficient. However, it is easy to implement, avoids the estimation of the Hessians or variance matrix for use in a second step, and has excellent finite-sample properties. These insights extend to the CLE estimator.

In Stata, the proposed estimator can be implemented by (i) duplicating each observation $(J-1)^T$ times using *expand*; (ii) generating binary choice variables by applying a different transformation $\pi$ to each duplicate of the original data; (iii) applying *clogit* on these binary choice variables and the augmented regressors discussed in Section 6.1. Baetschmann et al. (2015) provide more details.

In Section 7, I document the usefulness of the CLE approach using a simulation study. I show that the CLE avoids the finite sample bias in situations where $(J-1)^T$ is large relative to $n$. In those cases, the finite sample bias for the OMD estimator is large.

# 7 Simulation study

This simulation section consists of three parts. First, I document the efficiency gains from using the optimal minimum distance (OMD) estimator over existing estimators. Second, I show that the finite sample performance of the OMD estimator deteriorates as $T$ grows, and that the CLE is a good alternativewhen the number of transformations is large relative to the sample size. Finally, I document that the OMD and CLE are robust against misspecification of the error term distribution.[20]

## 7.1 Efficiency gain

The first set of results are for simulation designs with $T = 2$ and $n = 5000$. The reported results are based on $S = 1000$ simulations with data generated from the fixed

[20] Stata code for the simulation study is available from my website, `www.sfu.ca/~cmuris`. Since the CLE performs well regardless of the simulation design, we will use the it in our empirical illustration in Section 8.

effects ordered logit model described in Section 3. In particular,

$$y_{it}^{*} = \alpha_i + \frac{1}{K}\sum_k X_{it,k} + u_{it}.$$

The regressors $X_{it,k}$ are generated from a Normal distribution with a time-varying mean,

$$\begin{pmatrix} X_{i1,k} \\ X_{i2,k} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

The unobserved heterogeneity is generated as $\alpha_i = \frac{1}{2}\left(X_{i1,1} + X_{i2,1}\right)$. The ordered variable $y_{it} \in \{1, \cdots, J\}$ is computed according to

$$y_{it} = \begin{cases} 1 & \text{if} \quad\quad y_{it}^{*} < -1, \\ 2 & \text{if } -1 \leq y_{it}^{*} < 1, \\ j & \text{if } j-2 \leq y_{it}^{*} < j-1, \text{ if } j \in \{3, \cdots, J-1\} \\ J & \text{if } J-1 \leq y_{it}^{*}. \end{cases}$$

I evaluate the following estimators: (i) “Oracle” is the infeasible MLE, based on known $\alpha_i$, that maximizes the unconditional likelihood 5;[21] (ii) “CSLogit” is the MLE based on the unconditional likelihood 5, with the unobserved heterogeneity parameters fixed at 0, i.e. it ignores the unobserved heterogeneity; (iii)-(vi) the CMLEs based on the transformations $(1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$; (vii) “DvS” is the optimal minimum distance estimator based on all time-invariant transformations, proposed by Das and van Soest (1999); (viii) “OMD” is the optimal minimum distance estimator based on all transformations, proposed in Section 5, Remark 5. This is the estimator proposed advocated by this paper.

Table 2 displays the results for the simplest case, $J = 3$ and $K = 1$. Of the feasible estimators, the OMD estimator has the lowest simulated standard error. In particular, it is more than 10% more efficient than any of the currently available estimators. Unsurprisingly, estimators based on a single transformation are outperformed by estimators that efficiently combine several of them (DvS and OMD). The OMD has the largest bias of the feasible estimators, but that bias is small at 0.8. In terms of the cut point parameters, the OMD estimator has the lowest bias at 0.13%, and its simulated standard error is almost 50 lower than that of the most competitive CMLE. The large bias of the CSLogit emphasizes that unobserved heterogeneity is important in this simulation design. The relatively strong performance of the CMLE based on the transformation $(2,1)$ versus $(1,2)$ is due to the increase in the mean of $X_{itk}$ from -1 to 1 when we move from period 1 to period 2.

[21] This estimator does not suffer from the incidental parameters problem, and uses all cross-section units (including those with time-invariant $y_i$). It is asymptotically at least as efficient as the minimum-variance estimator for $\beta$ in the presence of nuisance parameters $\alpha_i$. As such, it provides an upper bound on the performance of the estimators considered in this simulation exercise.

|  | $\beta$ |  | $\gamma_2 - \gamma_1$ |  |
|---|---|---|---|---|
| Estimator | %Bias | RelSD | %Bias | RelSD |
| Oracle | 0.0 | 1.00 | 0.03 | 1.00 |
| CSLogit | 14.2 | 0.93 | 6.29 | 0.95 |
| $\pi = (1, 1)$ | 0.0 | 1.89 | - | - |
| $\pi = (2, 2)$ | 0.2 | 2.28 | - | - |
| $\pi = (1, 2)$ | 0.3 | 4.09 | 0.18 | 3.30 |
| $\pi = (2, 1)$ | 0.2 | 1.90 | 0.28 | 3.70 |
| DvS | 0.6 | 1.52 | - | - |
| **OMD** | **0.8** | **1.35** | **0.13** | **1.51** |

**Table 2:** Simulation results based on 1000 simulations, for $J = 3$, $K = 1$, $T = 2$, $N = 5000$ and $T = 2$. The OMD estimator proposed in this paper is in the bolded last row. The columns "%Bias" list the absolute value of the simulated bias, divided by the true value of the parameter. The columns "RelSD" report the simulated standard deviation divided by the simulated standard deviation of the Oracle estimator.

To investigate the robustness of these findings, Table 3 presents results for designs with different values for $J$ and $K$. In terms of the regression coefficient, notice that the simulated relative standard error of the OMD estimator is always substantially lower than that of the other feasible estimators (only the most competitive ones are displayed). Compared to the Oracle estimator, there is some loss of relative efficiency when $J$ is increased. Increasing the number of regressors from increases relative efficiency. The efficiency gain with respect to the DvS estimator is 11.2% in the benchmark case, and increases to 14.5% and 14.7% when $K$ is increased to 3 and 5. Increasing $J$ reduces the efficiency gain to 8.8%. This is likely due to the finite sample bias associated with estimating the variances and covariancies of $(J-1)^2$ estimators, which is investigated in more detail below. In terms of the cut point differences, the relative standard errors increase when the number of regressors is increased. The number of categories does not seem to have a large effect on the relative standard error. The bias of the OMD estimator increases, likely due to finite sample issues associated with estimating the optimal weights for the additional estimators.

## 7.2 Small sample performance

Section 6.2 discusses a potential drawback of the OMD estimator when the number of estimators is large relative to the sample size. The OMD relies on an estimate of the variance of the CMLE estimators, which may lead to poor finite sample performance when considering a large number of estimators. The CLE was introduced to avoid this issue, at the expense of asymptotic efficiency. I now investigate the relative finite sample performance of the CLE estimator and the OMD estimator.

For that purpose, I decreasing the sample size to $n = 350$, and manipulate the number of time periods $T$. The remaining design parameters are unchanged. For $t > 2$,

|  | $J=3,\ K=3$ |  | $J=3,\ K=5$ |  | $J=5,\ K=5$ |  |
|---|---|---|---|---|---|---|
| Estimator | %Bias | RelSD | %Bias | RelSD | %Bias | RelSD |
| *Coefficient* $\beta$ |  |  |  |  |  |  |
| $\pi=(1,1)$ | 0.17 | 1.78 | 0.90 | 1.70 | 0.90 | 1.80 |
| $\pi=(1,2)$ | 0.30 | 1.49 | 0.80 | 1.33 | 0.80 | 1.40 |
| DvS | 0.03 | 1.43 | 0.61 | 1.36 | 0.41 | 1.37 |
| **OMD** | **0.24** | **1.22** | **0.01** | **1.16** | **1.69** | **1.28** |
| *Cut point* $\gamma_2-\gamma_1$ |  |  |  |  |  |  |
| $\pi=(1,2)$ | 0.47 | 4.20 | 0.57 | 5.02 | 0.57 | 5.03 |
| $\pi=(2,1)$ | 0.69 | 11.01 | 2.02 | 14.17 | 2.02 | 14.95 |
| **OMD** | **0.21** | **1.42** | **0.50** | **1.40** | **1.76** | **1.40** |

**Table 3:** Robustness of simulation results. We change the values of $(J,K)$ away from their respective values of $(3,1)$ in the benchmark simulation design in Table 2. Results are based on $S=1000$ simulations. The columns "%Bias" list the absolute value of the simulated bias, divided by the true value of the parameter. The columns "RelSD" report the simulated standard deviation divided by the simulated standard deviation of the Oracle estimator.

the regressors are generated from a standard normal distribution.

The results for the regression coefficient, based on $S=1000$ simulations, are presented in Table 4. The results for the cut points are qualitatively similar to those for the regression coefficient. We report the results for (i) the Oracle estimator; (ii) for the CMLE estimator based on $\pi=(1,1)$, which has the best performance among the CMLEs for this design; (iii) "BUC", the composite likelihood estimator based on time-invariant transformations, described in Baetschmann et al. (2015); (iv) "DvS", the optimal minimum distance estimator based on time-invariant transformations proposed by Das and van Soest (1999); (v) "CLE", the composite likelihood based on all transformations introduced in Section 6.2; (vi) "OMD", the optimal minimum distance estimator based on all transformations.

The CLE has the lowest relative standard error in all designs. The relative performance of the OMD estimator versus the CLE deteriorates when $T$ increases. In particular, when $T=6$ and $T=8$, the simulated bias for the OMD estimator is very large at 10% and 18%. For these cases, the CLE is an excellent alternative, as it can be seen to dominate the existing estimators in terms of standard error, and also delivers the differences in the cut points. The poor performance of the OMD estimator in these designs is not surprising. The optimal combination of the $(J-1)^T$ CMLEs requires an estimate of a large number of variance and covariance matrices.[22]

[22] For $T=4$, the number of blocks is 121. For $T=6$, this number grows to 2017.

| | $T=2$ | | $T=4$ | | $T=6$ | | $T=8$ | |
|---|---|---|---|---|---|---|---|---|
| Estimator | %Bias | RelSD | %Bias | RelSD | %Bias | RelSD | %Bias | RelSD |
| Oracle | 0.72 | 1.00 | 1.29 | 1.00 | 0.38 | 1.00 | 1.17 | 1.00 |
| $\pi=(1,1)$ | 1.01 | 1.78 | 2.45 | 1.57 | 0.59 | 1.48 | 1.70 | 1.46 |
| BUC | 1.44 | 1.53 | 1.64 | 1.18 | 0.16 | 1.15 | 1.14 | 1.10 |
| DvS | 0.30 | 1.50 | 1.30 | 1.20 | 0.00 | 1.13 | 0.99 | 1.09 |
| **CLE** | **1.83** | **1.40** | **1.87** | **1.15** | **0.20** | **1.10** | **1.05** | **1.07** |
| **OMD** | **0.21** | **1.43** | **1.90** | **1.20** | **10.70** | **1.18** | **18.85** | **1.21** |

**Table 4:** Simulation results comparing the optimal minimum distance estimators to the composite likelihood estimators. Sample size is $n = 350$, and results are based on $S = 100$ simulations. The columns "%Bias" list the absolute value of the simulated bias, divided by the true value of the parameter. The columns "RelSD" report the simulated standard deviation divided by the simulated standard deviation of the Oracle estimator.

### 7.3 Robustness to misspecification

In this subsection, I investigate the sensitivity of the OMD estimator and the CLE to misspecification of the error term distribution. To that end, I generate data from the benchmark specification used for Table 2, with the exception that the error terms $u_{it}$ are generated from a number of alternative distributions. Baetschmann et al. (2015, Section 3.3) document that the DvS and BUC estimator perform well under alternative specification when the focus is on the ratio $\beta_1/\beta_2$. In this subsection, I focus on the ratio $\beta_1/\left(\gamma_2-\gamma_1\right)$. I do not report results for estimators that do not estimate the cut point difference: such estimators are not informative for a model with only one regressor, because $\beta_1$ can only be estimated up to scale.

The results, using $S = 1000$ simulations, and using a Normal, $\chi^2$, and Poisson distribution are presented in Table 5. The results are favorable for the estimators proposed in this paper. The Oracle estimator has substantially more bias than in the correctly specified case, and is now outperformed by the CLE and OMD estimator in terms of bias for two of the designs. The Oracle estimator still has the lowest standard error. The CMLEs for time-varying transformations yield mixed results. The performance for $\pi = (1, 2)$ is reasonable, although not as good as the CLE and OMD. However, the CMLE based on $\pi = (2, 1)$ performs very poorly. The CLE and OMD do not inherit the poor performance of that estimator, and have generally excellent performance under misspecification.

## 8 Empirical illustration

In this section, I investigate the relationship between reported (subjective) children's health status and total household income using the conditional likelihood estimator that uses all $(J-1)^T$ transformations. The analysis in this section follows that in Murasko

| | Poi(2) | | $\chi^2_3$ | | $\mathcal{N}(0,1)$ | |
|---|---|---|---|---|---|---|
| Estimator | %Bias | RelSD | %Bias | RelSD | %Bias | RelSD |
| Oracle | 0.73 | 1.00 | 0.46 | 1.00 | 0.43 | 1.00 |
| $\pi = (1,2)$ | 0.08 | 1.28 | 0.14 | 1.28 | 0.06 | 1.33 |
| $\pi = (2,1)$ | 7.29 | 7.02 | 6.51 | 5.70 | 6.31 | 8.59 |
| **CLE** | **0.05** | **1.22** | **0.14** | **1.17** | **0.04** | **1.30** |
| **OMD** | **0.06** | **1.21** | **1.06** | **1.19** | **0.03** | **1.29** |

**Table 5:** Simulation results for the misspecification analysis. Results are for $\beta_1/(\gamma_2-\gamma_1)$, based on $S = 1000$ simulations. Results are based on $S = 1000$ simulations. The columns "%Bias" list the absolute value of the simulated bias, divided by the true value of the parameter. The columns "RelSD" report the simulated standard deviation divided by the simulated standard deviation of the Oracle estimator.

(2008) and Khanam et al. (2014).[23]

In an influential paper, Case et al. (2002) use pooled cross-section data from the United States to document that reported children's health is positively related to household income. They also find that this relationship is stronger for older children. Currie and Stabile (2003) replicate these findings using the panel data from Canada. They use the availability of panel data to investigate the effect of past health shocks, but do not control for fixed effects. Further empirical evidence for the findings in Case et al. (2002) comes from Murasko (2008). He uses the Medical Expenditure Panel Survey (MEPS) from the United States. He documents that the income-health relationship seems to be weaker in MEPS than it is in the cross-sectional NHIS data used by Case et al. (2002). Murasko confirms that the income-health relationship is stronger for older children. He does not use fixed effects to control for unobserved heterogeneity..

There are two studies that cast doubt on the finding that the health-income relationship becomes stronger with age. First, using pooled cross-section data from England, Currie et al. (2007) confirm the positive relationship between income and health. However, they find that magnitude of the relationship is much smaller than that reported in Case et al. (2002), and they do not find evidence that the strength of the relationship increases with the age of the child. Khanam et al. (2014) use Australian data. This study appears to be the first that that controls for time-invariant unobserved heterogeneity in this literature. They use the BUC estimator in Baetschmann et al. (2015), and do not find evidence that the strength of the health-income relationship increases with age.

My results are based on Panel 16 of the Medical Expenditure Panel Survey (MEPS). The MEPS is a rotating panel collected by the Agency for Healthcare Research Quality since 1996. It gathers information on demographic and socioeconomic variables, and on health and healthcare usage from a nationally representative sample of households. Data from the household's medical provider and employer-based health insurance sup-

[23] Stata code for this example can be downloaded from my website, `www.sfu.ca/~cmuris`.

| | RE | CRE | $(3,3)$ | BUC | CLE |
|---|---|---|---|---|---|
| $\log(\text{Income})_{it}$ | $-0.38$ | $-0.10$ | 0.33 | $-0.09$ | $-0.06$ |
| | (0.03) | (0.06) | (0.26) | (0.07) | (0.08) |
| Age$\times \log(\text{Income})_{it}$ | $-0.014$ | 0.017 | 0.083 | 0.021 | 0.035 |
| | (0.007) | (0.013) | (0.049) | (0.015) | (0.016) |
| Family size | 0.09 | $-0.19$ | $-0.46$ | $-0.20$ | $-0.23$ |
| | (0.04) | (0.14) | (0.60) | (0.15) | (0.16) |
| $\gamma_2 - \gamma_1$ | 2.01 | 2.02 | - | - | 1.87 |
| | (0.05) | (0.05) | | | (0.05) |
| $\gamma_3 - \gamma_2$ | 2.96 | 2.97 | - | - | 2.92 |
| | (0.09) | (0.09) | | | (0.12) |
| $\gamma_4 - \gamma_3$ | 2.73 | 2.74 | - | - | 2.61 |
| | (0.23) | (0.23) | | | (0.29) |

**Table 6:** Estimated health-income relationship for children using MEPS Panel 16, which consists of 4131 children, each observed in 2 years. Age, log(Income) and Family size were normalized to have zero sample mean. The first two columns report the results from random effects (RE) and correlated random effects (CRE) estimator. For the correlated random effects estimator (CRE) model, we model $\alpha_i$ as a linear function of the average of the time-varying regressors, gender, age, and dummies for race and region. Column $(3,3)$ uses Chamberlain's conditional logit estimator using "Good" as the cut off category. Column BUC reports the results from a composite likelihood procedure using the time-invariant transformations. Column CLE uses the composite likelihood procedure proposed by this paper.

plement the household data. The household data is obtained through questionnaires. Importantly, all household members are interviewed regarding their health status.

Panel 16 of the MEPS contains data on 4131 children gathered across 2 years (2011 and 2012). A child is any interviewee who does not reach age 18 by the end of the interview period. I remove observations with nonpositive values for household income, age, and those with family size less than two. I also remove the richest 5% of families. The dependent variable is self-reported health status ($RTHLTH$) reported in rounds 2 and 4.[24] Subjective health is reported on a scale of 1-5, where "1" corresponds to "Poor", "2" to Fair, "3" to "Good", "4" to "Very good", and "5" to "Excellent". The explanatory variable of interest is the log of total household income, and the interaction between age and income. The interaction term between age category and family income allows us to determine whether the relationship between income and health changes with age. Following Murasko, I also control for family size (capped at 5). The reported specification does not include a year dummy: including it does not change the results. The results are reported in Table 6.

[24] Reported health is available for all five interview rounds. However, we only have annual income data. We choose rounds 2 and 4 based on the timing of the interview rounds: the uneven interview rounds are conducted over a period that span multiple years.

I report results for five estimators. The first two columns report the results from random effects (RE) and correlated random effects (CRE) estimator. For the correlated random effects estimator (CRE) model, I model $\alpha_i$ as a linear function of the average of the time-varying regressors, gender, age, and dummies for race and region. Column $(3, 3)$ uses Chamberlain's conditional logit estimator using "Good" as the cut off category. Column BUC reports the results from a composite likelihood procedure using the time-invariant transformations. Column CLE uses the composite likelihood procedure proposed by this paper. Age, log(Income), and Family size have been normalized to have zero sample mean.

There does not seem to be enough evidence in the data to find a positive effect of income on health at the average sample age (coefficient on log(Income)). The CRE estimator and the FE estimators (columns 2-4) do not document a statistically significant relationship. The random effects estimator does find a statistically significant negative relationship, but this is likely due to omitted variable bias. The data, however, do support the finding that the income-health effect increases with age. The CLE estimator proposed in this paper is the only estimator that is efficient enough to pick this effect up. Although the other estimators find the same, positive, sign, the results for those estimators are not statistically significantly different from zero at standard significance levels. None of the estimators find a statistically significant effect of family size, although all point estimates - with the exception of the estimate from the random effects model - produce negative point estimate. It is likely that the this is due to omitted variable bias, which is avoided by the correlated random effects and fixed effects procedures.

In this context, the CLE is clearly to be preferred over the BUC and CMLE estimators: it is the only estimator that detects that the relationship between income and health changes with age. It is also the only fixed effects estimator that produces the cut point differences. The CLE is also to be preferred over the RE estimator, which seems to be biased due to the omitted variable bias that is captured by our fixed effects approach.

It is also interesting to compare the CLE results to those from the CRE. The coefficient estimates from the CRE estimator are very similar to those from the CLE, which suggests that the model for the unobserved heterogeneity used by the CRE is close to correctly specified. However, this model involves so many covariates that the additional restrictions imposed on the unobserved heterogeneity do not reduce the standard errors much in comparison to the CLE. The standard errors are only slightly smaller.

In Section 3.3, I discussed an interpretation for the magnitude of the regression coefficient based on the cut point differences. The quantity $\delta_m^j$ in equation (15) can be interpreted as the the minimum required change in $X_{itm}$ to move an arbitrary observation $y_{it} = j$ to a higher category. For the CLE results in this model, consider the amount of family income required to move a 15-year old that is currently reported to have "Fair" health to have at least "Good" health. The appropriate regression coefficient for a 15-year old is $7.60 * 0.035 - 0.064 \approx 0.20$. As a result, an income increase of more than 900 is required to achieve this. This either suggests that the quantity is a loose bound, or that the relationship between income and health is weak. To get some idea, we can compare the results from the CRE model. If we look at the CRE's predictions for children of age 15 that currently have "Fair" health, we see that a 100% income increase changes their predicted probability of being in category "Good" or above from 0.1415 to 0.1447. This suggests that the relationship between income and health at age 15, although statistically significant, is not very strong..

# 9 Conclusion

I propose a new estimator for the regression coefficient and the difference in the cut points in the fixed effects ordered logit model. The estimator uses $(J-1)^T$ transformations of a time series of ordered discrete outcomes into time series of binary outcomes. Taking into account all these transformations allows me to construct an estimator for the regression coefficient that is more efficient than currently available estimators. Furthermore, the difference in the cut points is estimated, which provides a bound on a marginal effect that is useful in empirical practice. Simulation results suggest substantial efficiency gains, and document robustness to alternative distributional assumptions on the error terms.

It may be possible to extend the main result in this paper to a more general setting, namely the fixed effects ordered choice model that relaxes the logistic assumption and the serial independence assumption. For example, consider the fixed effects ordered logit model in equations (1)-(3), but replace the conditional logit assumption by the assumption that the conditional distribution of the error terms is identically distribution in each error term. Then the transformations described in Section 3.2 would turn the ordered model into $(J-1)^T$ instances of the semiparametric binary choice model studied by Manski (1987). An alternative approach would be to use a pairwise differencing approach, see e.g. Ahn et al. (2015).

# References

[1] Afonso, A., P. Gomes, and P. Rother (2013), "Short- and Long-run Determinants of Sovereign Debt Credit Ratings," *International Journal of Finance and Economics*, 16(1), 1-15.

[2] Ahn, S. C., and Peter Schmidt (1995), "A Separability Result for GMM estimation, with Applications to GLS prediction and Conditional Moment Tests." Econometric Reviews 14 (1): 19-34.

[3] Ahn, H., H. Ichimura, J.L. Powell, and P.A. Ruud (2015), "Simple Estimator for Invertible Index Models," Working paper.

[4] Allen, R. and J. Allnutt (2013), "Matched Panel Data Estimates of the Impact of Teach First on School and Departmental Performance," DoQSS Working Papers 13-11, Institute of Education, University of London.

[5] Amato, J.D., and C.H. Furfine (2003), "Are Credit Ratings Procyclical?," *Journal of Banking and Finance*, 28(11), 2641-2677.

[6] Andersen, E.B. (1970), "Asymptotic Properties of Conditional Maximum-likelihood Estimators," *Journal of the Royal Statistical Society*, Series B, 32, 283-301.

[7] Arellano, M., and S. Bonhomme (2012), "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395-424.

[8] Arellano, M., and B. Honoré (2001), "Panel Data Models: Some Recent Developments," in J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 5*, pp. 3229-3296. Amsterdam: Elsevier.

[9] Baetschmann, G., Staub, K. E., & Winkelmann, R. (2015). "Consistent estimation of the fixed effects ordered logit model." Journal of the Royal Statistical Society: Series A, 178(3), 685-703.

[10] Blanchflower, D.G., and A.J. Oswald (2004), "Well-being Over Time in Britain and the USA," *Journal of Public Economics*, 88, 1359-1386.

[11] Bonhomme, S. (2012), "Functional Differencing," *Econometrica*, 80(4), 1337-1385.

[12] Booth, A.L., and J.C. van Ours (2008), "Job Satisfaction and Family Happiness: The Part-Time Work Puzzle," *Economic Journal*, 118, F77-F99.

[13] Booth, A.L., and J.C. van Ours (2013), "Part-time Jobs: What Women Want?," *Journal of Population Economics*, 26(1), 263-283.

[14] Cameron and Miller (2015), "A Practitioner's Guide to Cluster-Robust Inference*", Journal of Human Resources*, 50(2), 317-373.

[15] Carman, K. (2013), "Inheritances, Intergenerational Transfers, and the Accumulation of Health," *American Economic Review,* Papers and Proceedings, 23(2), 223-237.

[16] Case, A., D. Lubotsky, and C. Paxson (2002), "Economic Status and Health in Childhood: The Origins of the Gradient," *The American Economic Review*, 92(5), 1308-1334.

[17] Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225-238.

[18] Chamberlain, G. (1984), "Panel Data," in Z. Griliches and M.D. Intriligator (Ed.), *Handbook of Econometrics, Volume 2*, pp. 1247-1318. Amsterdam: Elsevier.

[19] Chamberlain, G. (2010), "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78, 159-168.

[20] Chen, X., D.T. Jacho-Chávez, and O. Linton (2014), "Averaging of an Increasing Number of Moment Condition Estimators," *Econometric Theory*, 1-41.

[21] Currie, J. and M. Stabile (2003), "Socioeconomic Status and Child Health: Why Is the Relationship Stronger for Older Children?," *American Economic Review*, 93(5), 1813-1823.

[22] Currie, J., M.A. Shields, and S. Wheatley Price (2007), "The Child Health/Family Income Gradient: Evidence from England," *Journal of Health Economics*, 26, 213-232.

[23] Das, M. and A. van Soest (1999), "A Panel Data Model for Subjective Information on Household Income Growth," *Journal of Economic Behavior and Organization*, 40, 409-426.

[24] Fairlie, R.W., F. Hoffmann, and P. Oreopoulos (2014), "A Community College Instructor Like Me: Race and Ethnicity Interactions in the Classroom," *American Economic Review*, 104(8), 2567-2591.

[25] Ferrer-i-Carbonell, A., and P. Frijters (2004), "How Important is Methodology for the estimates of the determinants of Happiness?," *Economic Journal*, 114, 641-659.

[26] Frijters, P., J.P. Haisken-DeNew, and M.A. Shields (2004), "Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunication," *American Economic Review*, 94, 730-740.

[27] Frijters, P., J.P. Haisken-DeNew, and M.A. Shields (2005), "The Causal Effect of Income on Health: Evidence from German Reunication," *Journal of Health Economics*, 24, 997-1017.

[28] Hahn, J. (1997), "A Note on the Efficient Semiparametric Estimation of Some Exponential Panel Models," *Econometric Theory*, 13(4), 583-588.

[29] Hall, Alastair R. (2005), "Generalized method of moments." Oxford: Oxford University Press.

[30] Hamermesh, D. (2001), "The Changing Distribution Of Job Satisfaction," *Journal of Human Resources*, 36(1), 1-30.

[31] Han, C., and P.C.B. Phillips (2006), "GMM with Many Moment Conditions," *Econometrica,* 74(1), 147-192.

[32] Hausman, J.A., B.H. Hall, and Z. Griliches (1984), “Econometric Models for Count Data With an Application to the Patents-R&D Relationship,” *Econometrica*, 52(4), 909-938.

[33] Honoré, B. (1992). “Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects,” *Econometrica*, 60 (3), 533-565.

[34] Honoré, B. and E. Tamer (2006). “Bounds on Parameters in Panel Dynamic Discrete Choice Models,” *Econometrica*, 74 (3), 611-629.

[35] Honoré, B. (2002), “Nonlinear Models with Panel Data,” *Portuguese Economic Journal*, 1, 163-179.

[36] Khanam R., H.S. Nghiem, and L.B. Connelly (2014), “What Roles Do Contemporaneous and Cumulative Incomes Play in the Income–Child Health Gradient for Young Children? Evidence From an Australian Panel,” *Health Economics*, 23, 879-893,

[37] Lancaster, T. (2000), “The Incidental Parameter Problem Sine 1948,” *Journal of Econometrics*, 95, 391-413.

[38] Lancaster, T. (2002), “Orthogonal Parameters and Panel Data,” *Review of Economic Studies*, 69, 647-666.

[39] Machado, M.P. (2004), “A Consistent Estimator for the Binomial Distribution in the Presence of 'Incidental Parameters': An Application to Patent Data,” *Journal of Econometrics*, 119 (1), 73-98.

[40] Magnac, T. (2004), “Panel Binary Variables and Sufficiency: Generaling Conditional Logit,” *Econometrica*, 72(6), 1859-1876.

[41] Manski, C.F. (1987), “Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data,” *Econometrica*, 55, 357-362.

[42] Murasko, J.E. (2008), “An Evaluation of the Age-profile in the Relationship Between Household Income and the Health of Children in the United States,” *Journal of Health Economics*, 27, 1489-1502.

[43] Newey, W.K., and D.L. McFadden (1994), “Large Sample Estimation and Hypothesis Testing,” in R.F. Engle and D.L. McFadden, *Handbook of Econometrics, Volume 4*, pp. 2111-2245. Amsterdam: Elsevier.

[44] StataCorp (2015), *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

[45] Winkelmann, R., and L. Winkelmann (1998), “Why Are the Unemployed So Unhappy? Evidence From Panel Data,” *Economica*, 65, 1-15.

[46] Wooldridge, J.M. (1999), "Distribution-free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics*, 90, 77-97.

[47] Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data, Second Edition.* Cambridge, Massachusetts: The MIT Press.

# A Proofs and derivations

## A.1 Proof of Theorem 1

*Proof.* Proof of Theorem 1. Consider an observation $(y_i, X_i)$ from the fixed effects ordered logit model (1)-(3). Assign a category $\pi(t) \in \{1, \cdots, J-1\}$ to each time period. Denote by $d_{i,\pi}$ the dichotimization of $y_i$ along $\pi$,

$$d_{i,\pi} = \left(d_{i,t,\pi} = 1\left\{y_{it} \leq \pi(t)\right\}, t = 1, \cdots, T\right).$$

Denote the number of observations below or at the associated cut points by $\bar{d}_{i,\pi} = \sum_{t=1}^{T} d_{i,t,\pi}$. The logistic probability distribution is positive on the real line, so that $P\left(d_{it} = f \middle| X_i, \alpha_i\right) > 0$ for any $f$, and for any value of $\alpha_i$ and $X_i$. This guarantees that all denominators in the following derivations are bounded away from zero. Letting $F_{\bar{d}} = \left\{f \in \{0,1\}^T : \bar{f} = \bar{d}\right\}$, we have

$$\begin{aligned} P\left(d_{i,\pi} = d \middle| \bar{d}_{i,\pi} = \bar{d}, X_i, \alpha_i\right) &= \frac{P\left(d_{i,\pi} = d, \bar{d}_{i,\pi} = \bar{d} \middle| X_i, \alpha_i\right)}{P\left(\bar{d}_{i,\pi} = \bar{d} \middle| X_i, \alpha_i\right)} \\ &= \frac{P\left(d_{i,\pi} = d \middle| X_i, \alpha_i\right)}{\sum_{f \in F_{\bar{d}}} P\left(d_{i,\pi} = f \middle| X_i, \alpha_i\right)} \\ &= \left[\sum_{f \in F_{\bar{d}}} \frac{P\left(\bar{d}_{i,\pi} = f \middle| X_i, \alpha_i\right)}{P\left(d_{i,\pi} = d \middle| X_i, \alpha_i\right)}\right]^{-1} \qquad (35) \end{aligned}$$

Because of conditional serial independence, any arbitrary term in that sum simplifies:

$$\begin{aligned}
&\frac{P\left(d_{i,\pi}=f\middle|\,X_i,\alpha_i\right)}{P\left(d_{i,\pi}=d\middle|\,X_i,\alpha_i\right)} \\
&= \prod_t\left[\frac{P\left(d_{i,t,\pi(t)}=f_t\middle|\,X_i,\alpha_i\right)}{P\left(d_{i,t,\pi(t)}=d_t\middle|\,X_i,\alpha_i\right)}\right] && (36)\\
&= \prod_{t:f_t=1,d_t=0}\left[\frac{P\left(d_{i,t,\pi(t)}=1\middle|\,X_i,\alpha_i\right)}{P\left(d_{i,t,\pi(t)}=0\middle|\,X_i,\alpha_i\right)}\right]\prod_{t:f_t=0,d_t=1}\left[\frac{P\left(d_{i,t,\pi(t)}=0\middle|\,X_i,\alpha_i\right)}{P\left(d_{i,t,\pi(t)}=1\middle|\,X_i,\alpha_i\right)}\right] && (37)\\
&= \prod_{t:f_t=1,d_t=0}\left[\frac{\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}/\left[1+\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}\right]}{1/\left[1+\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}\right]}\right] && (38)\\
&\qquad\times\prod_{t:f_t=0,d_t=1}\left[\frac{1/\left[1+\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}\right]}{\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}/\left[1+\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}\right]}\right]\\
&= \prod_{t:f_t=1,d_t=0}\left[\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}\right]\prod_{t:f_t=0,d_t=1}\left[\frac{1}{\exp\left\{\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right\}}\right].
\end{aligned}$$

For the equality leading to (37), note that the terms for $d_t = f_t$ have identical denominator and numerator. To obtain (38), substitute the expressions for the conditional probability in the fixed effects ordered logit model.

Because $f$ and $d$ have the same number $\left(\bar{d}\right)$ of ones, the sets $\{t : f_t = 1, d_t = 0\}$ and $\{t : f_t = 0, d_t = 1\}$ have the same number of elements. Therefore, we have

$$\begin{aligned}
\frac{P\left(d_{i,\pi}=f\middle|\,X_i,\alpha_i\right)}{P\left(d_{i,\pi}=d\middle|\,X_i,\alpha_i\right)} &= \prod_{t:f_t\neq d_t}\left[\frac{\exp\left\{f_t\left(\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right)\right\}}{\exp\left\{d_t\left(\gamma_{\pi(t)}-X_{it}\beta-\alpha_i\right)\right\}}\right]\\
&= \exp\left\{\sum_{t:f_t\neq d_t}\left(f_t-d_t\right)\left(\gamma_{\pi(t)}-X_{it}\beta\right)\right\}. && (39)
\end{aligned}$$

By plugging (39) back into equation (35), we obtain the equality (9) in Theorem 1. □

## A.2 Bounds on the counterfactual distribution

Consider an individual $i$ who is in an intermediate category $j$ at time $t$, i.e. $y_{it} = j \in \{2, \cdots, J-1\}$. For such an individual,

$$y^*_{it} = X_{it}\beta + v_{it} \in \left(\gamma_{j-1}, \gamma_j\right), \tag{40}$$

where $v_{it}$ is the composite error term $v_{it} = \alpha_i + u_{it}$. Consider a ceteris paribus change in the regressors by $\Delta x$, inducing the counterfactual variables:

$$\begin{aligned} \tilde{X}_{it} &= X_{it} + \Delta x, \\ \tilde{y}^*_{it} &= y^*_{it} + (\Delta x)\,\beta \\ &= (X_{it} + \Delta x)\,\beta + v_{it}, \\ \tilde{y}_{it} &= \sum_{j=1}^{J} j \cdot 1\left\{\gamma_{j-1} < \tilde{y}^*_{it} < \gamma_j\right\}. \end{aligned}$$

We are interested in the distribution of the counterfactual dependent variable, conditional on the observable random variables $y_{it}$ and $X_{it}$, i.e.

$$P\left(\tilde{y}_{it} \middle| y_{it} = j, X_{it} = x\right). \tag{41}$$

This conditional probability depends on $\Delta x$ and on the true values of the model parameters. From (40), we have that, conditional on $X_{it} = x$ and for a fixed $\Delta x$,

$$y_{it} = j \quad \Leftrightarrow \quad v_{it} \in (\gamma_{j-1,0} - x\beta_0, \gamma_{j,0} - x\beta_0) \equiv (l_1, u_1)\,, \tag{42}$$

$$\tilde{y}_{it} = k \quad \Leftrightarrow \quad v_{it} \in (\gamma_{k-1,0} - (x + \Delta x)\,\beta_0, \gamma_{k,0} - (x + \Delta x)\,\beta_0) \equiv (l_2, u_2)\,. \tag{43}$$

The relative position of $(l_1, u_1)$ and $(l_2, u_2)$ are important in what follows. To that end, let $l = \max\{l_1, l_2\}$ and $u = \min\{u_1, u_2\}$. Also, denote by $F_v$ the conditional-on-$X_i$ distribution function of $v_{it}$. Then

$$\begin{aligned} P\left(\tilde{y}_{it} = k \middle| y_{it} = j, X_{it}\right) &= \frac{P\left(\tilde{y}_{it} = k, y_{it} = j \middle| X_{it}\right)}{P\left(y_{it} = j \middle| X_{it}\right)} && (44) \\ &= \frac{P\left(v_{it} \in (l_1, u_1) \cap (l_2, u_2) \middle| X_{it}\right)}{P\left(v_{it} \in (l_1, u_1) \middle| X_{it}\right)} && (45) \\ &= \begin{cases} 0 \text{ if } \ (l_1, u_1) \cap (l_2, u_2) = \emptyset \\ 1 \text{ if } \ (l_1, u_1) \cap (l_2, u_2) = (l_1, u_1) \\ \frac{F_v(u) - F_v(l)}{F_v(u_1) - F_v(l_1)} \text{ otherwise.} \end{cases} && (46) \end{aligned}$$

Similarly, we can consider the event $\tilde{y}_{it} > j$. Conditioning on $X_{it} = x$, we have

$$\tilde{y}_{it} > j \Leftrightarrow v_{it} > \gamma_{j,0} - (x + \Delta x)\,\beta_0 \equiv (l_2, \infty)\,. \tag{47}$$

Then

$$\begin{aligned} P\left(\tilde{y}_{it} > j \middle| y_{it} = j, X_{it}\right) &= \frac{P\left(\tilde{y}_{it} > j, y_{it} = j \middle| X_{it}\right)}{P\left(y_{it} = j \middle| X_{it}\right)} \\ &= \frac{P\left(v_{it} \in (l_1, u_1) \cap (l_2, \infty) \middle| X_{it}\right)}{P\left(v_{it} \in (l_1, u_1) \middle| X_{it}\right)} \\ &= \begin{cases} 1 \text{ if } \ (l_1, u_1) \cap (l_2, u_2 = \infty) = (l_1, u_1) \ \text{ or } l_2 < l_1 \\ 0 \text{ if } \ (l_1, u_1) \cap (l_2, u_2 = \infty) = \emptyset \text{ or } u_1 < l_2, \\ \frac{F_v(u = u_1) - F_v(l = l_2)}{F_v(u_1) - F_v(l_1)} \text{ otherwise.} \end{cases} \end{aligned}$$

Note that the final component of the display has the term $\frac{F_v(u=u_1)-F_v(l=l_2)}{F_v(u_1)-F_v(l_1)}$, which is unknown because $F_v$ is unknown, since the conditional distribution of $\alpha_i$ is unrestricted. The interesting cases are therefore those with $u_1 < l_2$ and $l_2 < l_1$.

We can use the definitions of $l_1$, $u_1$, and $l_2$ to back out the results used in the main text. Those definitions are in (42) and in (47). First, note that

$$u_1 \quad = \quad \gamma_{j,0} - x\beta_0 < \gamma_{j,0} - (x + \Delta x)\,\beta_0 = l_2$$

if and only if $\Delta x \beta_0 < 0$. This result is sensible, but not informative. The interesting case is

$$l_2 = \gamma_{j,0} - (x + \Delta x)\,\beta_0 < \gamma_{j-1,0} - x\beta_0 = l_1$$

which holds if and only if

$$(\Delta x)\,\beta_0 > \gamma_{j,0} - \gamma_{j-1,0}.$$

Therefore, the result of interest can be restated as.

$$P\left(\tilde{y}_{it} > j \middle| \, y_{it} = j, X_{it}\right) \quad = \quad \begin{cases} 1 \text{ if } \ (\Delta x)\,\beta_0 > \gamma_{j,0} - \gamma_{j-1,0}, \\ 0 \text{ if } \Delta x \beta_0 < 0 \\ \frac{F_v(\gamma_{j,0} - X_{it}\beta_0) - F_v(\gamma_{j,0} - (X_{it} + \Delta x)\beta_0)}{F_v(\gamma_{j,0} - X_{it}\beta_0) - F_v(\gamma_{j-1,0} - X_{it}\beta_0)} \text{ otherwise.} \end{cases}$$

In the main text, this result is used to turn the cut point differences into interpretable quantities.

For the sake of completeness, I also work out $P\left(\tilde{y}_{it} = k \middle| \, y_{it} = j, X_{it}\right)$ by simplifying the inequalities in (46) using the definitions of $l_1$, $u_1$, $l_2$, and $u_2$ in (42) and (43). The relevant cases are (i) $(l_1, u_1) \cap (l_2, u_2) = \emptyset$ and (ii) $(l_1, u_1) \cap (l_2, u_2) = (l_1, u_1)$. For case (i), note that

$$\begin{aligned} & (l_1, u_1) \cap (l_2, u_2) = \emptyset \\ \Leftrightarrow \quad & u_1 < l_2 \text{ or } u_2 < l_1 \\ \Leftrightarrow \quad & \gamma_{j,0} - x\beta_0 < \gamma_{k-1,0} - (x + \Delta x)\,\beta_0 \text{ or } \gamma_{k,0} - (x + \Delta x)\,\beta_0 < \gamma_{j-1,0} - x\beta_0 \\ \Leftrightarrow \quad & (\Delta x)\,\beta_0 < \gamma_{k-1,0} - \gamma_{j,0} \text{ or } \ (\Delta x)\,\beta_0 > \gamma_{k,0} - \gamma_{j-1,0}. \end{aligned}$$

For case (ii), note that

$$\begin{aligned} & (l_1, u_1) \cap (l_2, u_2) = (l_1, u_1) \\ \Leftrightarrow \quad & l_2 \leq l_1 \text{ and } u_2 \geq u_1 \\ \Leftrightarrow \quad & \gamma_{k-1,0} - (x + \Delta x)\,\beta_0 \leq \gamma_{j-1,0} - x\beta_0 \text{ and } \gamma_{k,0} - (x + \Delta x)\,\beta_0 \geq \gamma_{j,0} - x\beta_0 \\ \Leftrightarrow \quad & (\Delta x)\,\beta_0 \geq \gamma_{k-1,0} - \gamma_{j-1,0} \text{ and } \ (\Delta x)\,\beta_0 \leq \gamma_{k,0} - \gamma_{j,0} \end{aligned}$$

so that (46) is equal to

$$P\left(\tilde{y}_{it} = k \middle| \, y_{it} = j, X_{it}\right) = \begin{cases} 0 \text{ if } \ (\Delta x)\,\beta_0 < \gamma_{k-1,0} - \gamma_{j,0}, \\ 0 \text{ if } \ (\Delta x)\,\beta_0 > \gamma_{k,0} - \gamma_{j-1,0}, \\ 1 \text{ if } \gamma_{k-1,0} - \gamma_{j-1,0} \leq (\Delta x)\,\beta_0 \leq \gamma_{k,0} - \gamma_{j,0}, \\ \frac{F_v(u) - F_v(l)}{F_v(u_1) - F_v(l_1)} \text{ otherwise.} \end{cases}$$

## A.3 Score and Hessian calculations

We abbreviate $\sum_{f\in F_{\bar{d}}}$ to $\sum_f$ where this does not cause confusion. The following two derivatives will be useful in this section:

$$\begin{aligned}
\frac{\partial p_{i,\pi}^{-1}\left(d|\,\theta_\pi\right)}{\partial\theta_\pi} &= \frac{\partial\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)}{\partial\theta_\pi} \\
&= \sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\left(f-d\right)Z_i \\
\frac{\partial^2 p_{i,\pi}^{-1}\left(d|\,\theta_\pi\right)}{\partial\theta_\pi\partial\theta_\pi'} &= \sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)Z_i'\left(f-d\right)'\left(f-d\right)Z_i.
\end{aligned}$$

Note that $\ln p_{i,\pi}\left(d|\,\theta_\pi\right) = -\ln p_{i,\pi}^{-1}\left(d|\,\theta_\pi\right)$ is of the form $g\left(h\left(x\right)\right)$, with $g:\mathbb{R}\to\mathbb{R}$ is $g(u)=-\ln u$ and $h:\mathbb{R}^p\to\mathbb{R}$, is $h=p^{-1}$. The score in (19) follows immediately. The Hessian in 20 can be computed using the appropriate chain rule,

$$\frac{\partial^2 g\circ h\left(x\right)}{\partial x\partial x'} = \left.\frac{\partial^2 g(u)}{\left(\partial u\right)^2}\right|_{u=h(x)}\frac{\partial h\left(x\right)}{\partial x}'\frac{\partial h\left(x\right)}{\partial x} + \left.\frac{\partial g(u)}{\partial u}\right|_{u=h(x)}\frac{\partial^2 h\left(x\right)}{\partial x\partial x'},$$

which, for our case, evaluates as

$$\begin{aligned}
&-\frac{\partial^2 - \ln p_{i,\pi}^{-1}\left(d|\,\theta_\pi\right)}{\partial\theta_\pi\partial\theta_\pi'} \\
= \;& p_{i,\pi}^2\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\left(f-d\right)Z_i\right]'\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\left(f-d\right)Z_i\right] \qquad (48) \\
& -p_{i,\pi}\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)Z_i'\left(f-d\right)'\left(f-d\right)Z_i\right] \\
= \;& -p_{i,\pi}^2\left(\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\right]\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)Z_i'\left(f-d\right)'\left(f-d\right)Z_i\right] \qquad (49) \right.\\
& \left.\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\left(f-d\right)Z_i\right]'\left[\sum_f \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\left(f-d\right)Z_i\right]\right).
\end{aligned}$$

The difference between two double sums can be manipulated as in Machado (2003, Theorem 1), who uses a similar derivation in the context of a binomial logit model with fixed effects. For an arbitrary choice of $f,g$, the double term features the two expressions

$$\exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\exp\left(\left(g-d\right)Z_{i,\pi}\theta_\pi\right)\left[Z_i'\left(g-d\right)'\left(g-d\right)Z_i - Z_i'\left(f-d\right)'\left(g-d\right)Z_i\right]$$

and

$$\exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\exp\left(\left(g-d\right)Z_{i,\pi}\theta_\pi\right)\left[Z_i'\left(f-d\right)'\left(f-d\right)Z_i - Z_i'\left(g-d\right)'\left(f-d\right)Z_i\right],$$

which sum to

$$\exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\exp\left(\left(g-d\right)Z_{i,\pi}\theta_\pi\right)\left[Z_i^{'}\left(g-f\right)^{'}\left(g-f\right)Z_i\right],$$

so that the second derivative in (49) simplifies to

$$\begin{aligned}&\frac{\partial^2 \ln p_{i,\pi}\left(d\right|\theta_\pi)}{\partial\theta_\pi\partial\theta_\pi^{'}}\\ =\quad&-\frac{1}{2}p_{i,\pi}^2\sum_f\sum_g \exp\left(\left(f-d\right)Z_{i,\pi}\theta_\pi\right)\exp\left(\left(g-d\right)Z_{i,\pi}\theta_\pi\right)\left(Z_{i,\pi}^{'}\left(g-f\right)^{'}\left(g-f\right)Z_{i,\pi}\right).\end{aligned}$$

## A.4 Proof of Theorem 3, part (i)

*Proof.* [Proof of Theorem 3, part (i)] The CMLE is an extremum estimator with sample criterion function

$$Q_n\left(\theta_\pi\right) \quad = \quad \frac{1}{n}\sum_i\sum_{d\in\{0,1\}^T} 1\left[d_i=d\right]\ln p_\pi\left(d\right|\theta_\pi).$$

It follows from equation (20) in Section 4 (derivation in Appendix A.3) that the sample criterion function is concave. The current proof proceeds by verifying the conditions (i, identification) and (iii, pointwise convergence) for Theorem 2.7 in Newey and McFadden (1994), which is a proof of consistency for extremum estimators with concave sample criterion functions. First, I establish pointwise convergence of the criterion function. Second, I establish identification. The result then follows from the aforementioned result.

**Pointwise convergence (boundedness of criterion function).** Boundedness of

$$Q_0=E\left(Q_n\right)=\sum_{d\in\{0,1\}^T}1\left[d_i=d\right]E\left[\ln p_\pi\left(d\right|\theta_\pi)\right]$$

implies pointwise convergence of $Q_n$ to $Q_0$ by the law of large numbers. To bound $E\left[\ln p_\pi\left(d\right|\theta_\pi)\right]$, use a mean-value expansion around $\theta_\pi=0$, i.e. there exist some $\tilde{\theta}_\pi$ in between 0 and $\theta_\pi$ (row-wise) such that

$$\ln p_{i\pi}\left(d\right|\theta_\pi)-\ln p_{i\pi}\left(d\right|0)=s_{i,\pi}\left(d\middle|\,\tilde{\theta}_\pi\right)\theta_\pi.$$

Note that

$$\ln p_\pi\left(d_i\right|0)=-\ln\sum_{f\in F_{\bar{d}}}\exp\left\{\left(f-d\right)Z_{i\pi}0\right\}=-\ln\#\left\{F_{\bar{d}}\right\}<\infty.$$

Therefore, bounding the expectation of the derivative delivers the desired pointwise convergence. I simplify notation for

$$\begin{aligned}
s_{i,\pi}\left(d|\,\theta_{\pi}\right) &= -\frac{\sum_{f}\exp\left(\left(f-d\right)Z_{i\pi}\theta_{\pi}\right)Z_{i\pi}^{\prime}\left(f-d\right)^{\prime}}{\sum_{f}\exp\left(\left(f-d\right)Z_{i\pi}\theta_{\pi}\right)} \\
&\equiv -\frac{\sum_{f}a_{fd}Z_{i\pi}^{\prime}\left(f-d\right)^{\prime}}{a_{d}}
\end{aligned}$$

by setting $a_{fd}=\exp\left(\left(f-d\right)Z_{i\pi}\theta_{\pi}\right)$ and $a_{d}=\sum_{f}a_{fd}$. Note that $a_{fd}/a_{d}\in(0,1)$. Using Jensen's inequality and submultiplicativity,

$$\begin{aligned}
\left\|E\left(s_{i,\pi}\left(d|\,\theta_{\pi}\right)\right)\right\| &\leq E\left\|-Z_{i\pi}^{\prime}\times\sum_{f}\frac{a_{fd}}{a_{d}}\left(f-d\right)^{\prime}\right\| \\
&\leq \left\|\sum_{f}\frac{a_{fd}}{a_{d}}\left(f-d\right)\right\|\times E\left(\left\|Z_{i\pi}\right\|\right)
\end{aligned}$$

which is bounded because of our Assumption 2, which guarantees the existence of second moments of the regressors. Pointwise convergence of $Q_n$ to $Q_0$ follows.

**Identification.** We first make sure that $P\left(d_{i\pi}\in\left\{\left(0,\cdots0\right),\left(1,\cdots,1\right)\right\}\right)\neq1$. If this does not hold, then $Q_0\left(\theta_{\pi}\right)=1$ does not depend on the parameters. This is the population analogue of "time-invariant observations are not informative." This condition is satisfied because for any realization of the real-valued unobserved heterogeneity, the error term will still have some mass in the closets tail. To be more precise, note that

$$\begin{aligned}
P\left(d_{i\pi t}=0\right) &= P\left(u_{it}>\gamma_{\pi(t)}-\alpha_{i}\right) \\
&= \int\Lambda\left(\alpha_{i}-\gamma_{\pi(t)}\right)f\left(\alpha_{i}\right)d\alpha_{i} \\
&= \int_{-\infty}^{\bar{a}}\Lambda\left(\alpha_{i}-\gamma_{\pi(t)}\right)f\left(\alpha_{i}\right)d\alpha_{i}+\int_{\bar{a}}^{\infty}\Lambda\left(\alpha_{i}-\gamma_{\pi(t)}\right)f\left(\alpha_{i}\right)d\alpha_{i} \\
&< \int_{-\infty}^{\bar{a}}1f\left(\alpha_{i}\right)d\alpha_{i}+\sup\nolimits_{\alpha\in(\bar{a},\infty)}\Lambda\left(a-\gamma_{\pi(t)}\right)\int_{\bar{a}}^{\infty}f\left(\alpha_{i}\right)d\alpha_{i} \\
&= \int_{-\infty}^{\bar{a}}f\left(\alpha_{i}\right)d\alpha_{i}+\int_{\bar{a}}^{\infty}f\left(\alpha_{i}\right)d\alpha_{i}=1.
\end{aligned}$$

where I have used the symmetry of $\Lambda(u)=1-\Lambda(u)$, the fact that $\Lambda(u)$ is strictly increasing, and that $\lim_{u\to+\infty}\Lambda(u)=1$. The argument above can be conditioned on $X_i$. Since observations are serially independent after conditioning on $X_i$ and $\alpha_i$ , the above argument implies $P\left(d_{i\pi}=0_T\right)<1$. A two-sided version of the above argument then shows that

$$P\left(d_{i\pi}\in\left\{\left(0,\cdots0\right),\left(1,\cdots,1\right)\right\}\right)\neq1.$$

We have now established that transformed dependent variables are time-varying with positive probability.

The proof that follows is based on a reduction of the $T$ period problem to the 2-period using time periods $(1,t)$, where $t$ is arbitrary.[25] Consider an arbitrary time period $t>1$ and discard information from remaining time periods. The two-period data $(Y_{i1}, Y_{i,t}, X_{i1}, X_{it})$ follows a fixed effects ordered logit model. For any transformation $\pi$, the conditional probabilities in (10) and (11) apply:

$$\begin{aligned}
p^{-1}_{i,(\pi(1),\pi(t))}((1,0)) &= 1+\exp\left\{\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)-\left(X_{it}-X_{i1}\right)\beta\right\},\\
p^{-1}_{i,(\pi(1),\pi(t))}((0,1)) &= 1+\exp\left\{-\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)+\left(X_{it}-X_{i1}\right)\beta\right\}.
\end{aligned}$$

It will be useful to consider separately two types of transformation $\pi$: (i) the cut off categories in the two periods are the same, $\pi(t)=\pi(1)$, or (ii) the cut off category in period $t$ is different from that in period 1, $\pi(t)\neq\pi(1)$.

In case (i), the conditional probabilities in (10) and (11) simplify to

$$1+\exp\left\{-\left(X_{it}-X_{i1}\right)\beta\right\} \text{ and } 1+\exp\left\{\left(X_{it}-X_{i1}\right)\beta\right\}.$$

The assumption on the regressor variance matrix implies that $E\left[\left(X_{it}-X_{i1}\right)'\left(X_{it}-X_{i1}\right)\right]$ is positive definite, so that for any $\beta\neq\beta_0$,

$$\left(\beta-\beta_0\right)' E\left[\left(X_{it}-X_{i1}\right)'\left(X_{it}-X_{i1}\right)\right]\left(\beta-\beta_0\right)>0$$

which implies $\left(X_{it}-X_{i1}\right)\beta\neq\left(X_{it}-X_{i1}\right)\beta_0$. By monotonicity of $p$ in $\left(X_{it}-X_{i1}\right)\beta$, this ensures identification of $\beta_0$. This reasoning is similar to the standard identification proof for logit and probit, see e.g. Example 1.2 in Newey and McFadden (1994).

For case (ii), consider two different values of the difference in the cut point parameter,

$$\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)\neq\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)_0$$

and two values of the regression coefficient $\beta$, $\beta_0$. Let

$$0\neq\delta\equiv\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)-\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)_0$$

Then, it must hold that $\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)-\left(X_{it}-X_{i1}\right)\beta\neq\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)_0-\left(X_{it}-X_{i1}\right)\beta_0$. To see why, assume $\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)-\left(X_{it}-X_{i1}\right)\beta=\left(\gamma_{\pi(t)}-\gamma_{\pi(1)}\right)_0-\left(X_{it}-X_{i1}\right)\beta_0$. Then $\delta\neq 0$ implies $\left(X_{it}-X_{i1}\right)\left(\beta-\beta_0\right)=\delta\neq 0$. This would require some linear combination of $X_{it}$ to be equal to a non-zero constant. However, the variance of any non-zero linear combination of $X_{it}-X_{i1}$ is non-zero, by assumption. We conclude that

[25] Alternatively, a sufficient condition for identification is strict negative definiteness of the Hessian of $Q_0$. In the proof of asymptotic normality in Section A.5, part (iii), I show that this condition holds. The drawback of the alternative approach is that it relies more heavily on Assumption 2. The proof here suggests that Assumption 2 can be relaxed. I do not pursue this relaxation in this paper: it is notationally cumbersome, and does not provide additional insight.

$\gamma_{\pi,\Delta} \neq \gamma_{\pi,\Delta,0}$ implies $(f-d)\, Z_{i\pi}\theta_{\pi} \neq (f-d)\, Z_{i\pi}\theta_{\pi,0}$. Identification follows from strict monotonicity of $1+\exp(u)$.

Since the choice of $t>1$ was arbitrary, we can repeat this argument for all time periods $t>1$ and identify $\gamma_{\pi(t)} - \gamma_{\pi(1)}$, $t>1$, giving us $\gamma_{\Delta,\pi,0}$. The regression coefficient was identified using only two time periods. □

## A.5 Proof of Theorem 3, parts (ii)-(iv)

*Proof.* [Proof of Theorem 3, parts (ii)-(iv)] This proof is structured as follows. (1) First, I will show that $\Sigma_{\pi}$ exists. Part (ii) of Theorem 3 then follows by the random sampling assumption and a standard central limit theorem. (2) Second, I will show part (iii): that the Hessian $H_{\pi}\left(\theta_{\pi}\right)$ exists and is invertible for all $\theta_{\pi}$.

Part (iv) then follows from Theorem 3.1 in Newey and McFadden. To see that the conditions of their Theorem 3.1 is satisfied, note that the CMLE was shown to be consistent in part (i). Their condition (i, true value is in the interior) is satisfied because the parameter space is $\mathbb{R}^{K+n_{\pi}}$. Their condition (ii, objective function is twice differentiable) and the first part of their condition (iv, second derivative is continuous) is satisfied because of the results on the score and hessian in section 4, equations (19) and (20), as derived in A.3. Their condition (iii) is part (ii) of the current proof. The remainder of condition (iv), and condition (v), follows from part (iii) in the current proof (existence and invertibility Hessian). In particular, because I will show that the Hessian exists for all $\theta_{\pi}$, so that a ULLN will apply for any compact set around $\theta_{\pi 0}$, which is all we need for condition (iv) in Newey and McFadden's Theorem 3.1. In what follows, sums are over elements in the set $F_{\bar{d}} = \left\{ f \in \{0,1\}^{T} \,:\, \bar{f} = \bar{d} \right\}$ unless mentioned otherwise.

**Part (ii): Existence of $\Sigma_{\pi}$.** I will show that $\Sigma_{d,\pi}\left(\theta_{\pi}\right) = E\left( s_{i,\pi}\left(d \middle| \theta_{\pi}\right) s_{i,\pi}\left(d \middle| \theta_{\pi}\right)' \right)$ exists for an arbitrary $d$ and an arbitrary $\theta_{\pi}$. Existence of $\Sigma_{\pi}$ then follows because

$$\Sigma_{\pi}\left(\theta_{\pi}\right) = \sum_{d \in \{0,1\}^{T}} 1\left\{d_i = d\right\} \Sigma_{d,\pi}\left(\theta_{\pi}\right),$$

and is then evaluated at $\theta_{\pi} = \theta_{\pi 0}$. To simplify notation, let $a_{fd} = \exp\left(\left(f-d\right) Z_{i\pi}\theta_{\pi}\right)$

and $a_d = \sum_f a_{fd}$. Note that $a_{fd}/a_d = a_f \in (0,1)$. Then

$$\begin{aligned}
E\left(s_{i,\pi}\left(d|\,\theta_\pi\right) s_{i,\pi}\left(d|\,\theta_\pi\right)'\right) &= E\left[\frac{\sum_f \exp\left(\left(f-d\right) Z_{i\pi}\theta_\pi\right) Z_{i\pi}'\left(f-d\right)'}{\sum_f \exp\left(\left(f-d\right) Z_{i\pi}\theta_\pi\right)} \times \right. \\
&\qquad \left. \frac{\sum_f \exp\left(\left(f-d\right) Z_{i\pi}\theta_\pi\right)\left(f-d\right) Z_{i\pi}}{\sum_f \exp\left(\left(f-d\right) Z_{i\pi}\theta_\pi\right)}\right] \\
&= E\left[\frac{\sum_f \sum_g a_{fd}a_{gd}Z_{i\pi}'\left(f-d\right)'\left(g-d\right) Z_{i\pi}}{a_d^2}\right] \\
&< E\left[\sum_f \sum_g Z_{i\pi}'\left(f-d\right)'\left(g-d\right) Z_{i\pi}\right]
\end{aligned}$$

which exists, because because $Z_i$ is a combination of $\tilde{X}_i$ and the selection matrix $S_\pi$. and we know that the second moments of $X_i$ exist by the assumption (2).

**Part (iii): Existence and invertibility of $H_\pi\left(\theta_\pi\right)$.** In Sections 4 and A.3, it was established that the log-likelihood contributions are concave. It follows that the expected log-likelihood is concave. I now establish that concavity of the population criterion function is strict, by showing that the second derivative,

$$H_\pi\left(\theta\right) = E\left(\sum_{d\in\{0,1\}^T} 1\left[d_i = d\right] H_{i,\pi}\left(d|\,\theta_\pi\right)\right)$$

is negative definite.[26] This will be true if, letting $a_{fd}$ and $a_d$ be defined as in the proof of part (ii),

$$\begin{aligned}
E\left[H_{i,\pi}\left(d|\,\theta_\pi\right)\right] &= -\frac{1}{2}E\left[\left[p_{i,\pi}\left(d|\,\theta_\pi\right)\right]^2 \sum_{f,g} a_{fd}a_{gd}Z_{i,\pi}'\left(g-f\right)'\left(g-f\right) Z_{i,\pi}\right] \\
&= -\frac{1}{2}E\left[Z_{i,\pi}'\left(\sum_{f,g}\left[p_{i,\pi}\left(d|\,\theta_\pi\right)\right]^2 a_{fd}a_{gd}\left(g-f\right)'\left(g-f\right)\right) Z_{i,\pi}\right] \qquad (50)
\end{aligned}$$

is negative definite for at least one value $d$. Take an arbitrary value $d$, and , so that $p_{i,\pi}\left(d|\,\theta_\pi\right) = 1/a_d$. Note that $a_{fd} \geq 0$ for all $f, d$. The double sum inside (50) simplifies considerably in notation:

$$\begin{aligned}
\left[p_{i,\pi}\left(d|\,\theta_\pi\right)\right]^2 \sum_{f,g} a_{fd}a_{gd}\left(g-f\right)'\left(g-f\right) &= \frac{1}{a_d^2}\sum_{f,g} a_{fd}a_{gd}\left(g-f\right)'\left(g-f\right) \\
&= \sum_{f,g}\frac{a_{fd}a_{gd}\left(g-f\right)'\left(g-f\right)}{\left(\sum_{h_1} a_{h_1d}\right)\left(\sum_{h_2} a_{h_2d}\right)} \qquad (51)
\end{aligned}$$

[26] This could serve as an alternative proof of identification (see the proof of Theorem 3, part (i)).

For all $f = g$, this term is equal to the zero matrix, since $(f-g)^{'} = 0_T$. Furthermore, the denominator in (51) can be expanded

$$\left(\sum_{h_1} a_{h_1 d}\right)\left(\sum_{h_2} a_{h_2 d}\right) = \sum_{h_1} a_{h_1 d}^2 + \sum_{h_1 \neq h_2} a_{h_1 d} a_{h_2 d}.$$

Since

$$\frac{a_{fd}}{a_{gd}} = \frac{\exp\left\{(f-d) Z_{i,\pi}\theta_\pi\right\}}{\exp\left\{(g-d) Z_{i,\pi}\theta_\pi\right\}} = \exp\left\{(f-g) Z_{i,\pi}\theta_\pi\right\} = a_{fg},$$

any term under the double sum in in (51) can be bounded as follows:

$$\begin{aligned} \frac{a_{fd} a_{gd} (g-f)^{'} (g-f)}{\sum_{h_1} a_{h_1 d}^2 + \sum_{h_1 \neq h_2} a_{h_1 d} a_{h_2 d}} &= \frac{(g-f)^{'}(g-f)}{\sum_{h_1} a_{h_1 f} a_{h_1 g} + \sum_{h_1 \neq h_2} a_{h_1 f} a_{h_2 g}} && (52) \\ &= \frac{(g-f)^{'}(g-f)}{2 + \sum_{h_1} a_{h_1 f} a_{h_1 g} + \sum_{h_1 \neq h_2} a_{h_1 f} a_{h_2 g}} && (53) \\ &\geq \frac{1}{2} (g-f)^{'} (g-f). && (54) \end{aligned}$$

Plugging this lower bound back into (51) gives

$$\sum_{f,g} \frac{a_{fd} a_{gd} (g-f)^{'} (g-f)}{\left(\sum_{h_1} a_{h_1 d}\right)\left(\sum_{h_2} a_{h_2 d}\right)} \geq \frac{1}{2} \sum_{f,g} (g-f)^{'} (g-f),$$

and, finally,

$$\begin{aligned} E\left[H_{i,\pi}\left(d \mid \theta_\pi\right)\right] &\leq -\frac{1}{2} E\left[Z_{i,\pi}^{'} \left(\sum_{f,g} (g-f)^{'}(g-f)\right) Z_{i,\pi}\right]. \\ &= -\frac{1}{2} \sum_{f,g} E\left[Z_{i,\pi}^{'} (g-f)^{'} (g-f) Z_{i,\pi}\right] && (55) \end{aligned}$$

Every term in the finite sum of $E\left[Z_{i,\pi}^{'} (g-f)^{'} (g-f) Z_{i,\pi}\right]$ in equation (55) is finite because of assumption (ii). This establishes the existence of the Hessian for all $\theta_{\pi 0}$.

To see that it is positive definite, take arbitrary $f \neq g$, and note that

$$(g-f) Z_{i,\pi} = \sum_t (g_t - f_t)\left(\gamma_{\pi(t)} - \gamma_{\pi(1)} - X_{it}\right) = (g-f) S_\pi \gamma_{\pi,\Delta} + \Delta_{g,f} X_i,$$

where $\Delta_{g,f} X_i = \sum_t (g_t - f_t) X_{it}$ is a $1 \times K$ vector of time-differenced regressors . The assumption that Var $\left(\tilde{X}_i\right)$, exists and is positive definite implies that

$$\begin{aligned} Var\left(\Delta_{g,f} X_i\right) &= Var\left(Vec\left((g-f) X_i\right)\right) \\ &= Var\left(I_K \otimes (g-f) \tilde{X}_i\right) \\ &= \left(I_K \otimes (g-f)\right) Var\left(\tilde{X}_i\right)\left(I_K \otimes (g-f)^{'}\right) && (56) \end{aligned}$$

is positive definite, since the expression in (56) amounts to selecting a submatrix of a positive definite matrix. Finally, since $Var\left(\Delta_{f,d}X_i\right)$ is positive definite, it follows that $H_\pi\left(\theta\right)$ is negative definite, by equation (55). □

## A.6 Proof of Theorem 4

*Proof.* [Proof of Theorem 4] This proof consists of two parts. The main result is the comparison of the variance of the different estimators. That result requires asymptotic normality of the estimators. I start with the variance comparison (Part 1), and then show that the required asymptotic normality is satisfied (Part 2).

**Part 1: Comparing the asymptotic variances.** On top of the estimators $\tilde{\beta}^*$, and $\hat{\beta}^*$ defined in the main text, define $\check{\beta}^*$ as the efficient estimator based on $s_{i,1,\beta}$ and $s_{i,2,\gamma}$. In Part 2 (below), I show that $\tilde{\beta}^*$, $\check{\beta}^*$ and $\hat{\beta}^*$ are asymptotically normal, i.e.

$$
\begin{aligned}
\sqrt{n}\left(\tilde{\beta}^* - \beta\right) &\overset{d}{\to} \mathcal{N}\left(0, V_1\right)\\
\sqrt{n}\left(\check{\beta}^* - \beta\right) &\overset{d}{\to} \mathcal{N}\left(0, V_2\right)\\
\sqrt{n}\left(\hat{\beta}^* - \beta\right) &\overset{d}{\to} \mathcal{N}\left(0, V_3\right)
\end{aligned}
$$

This part proceeds in two steps. The first step establishes that $\check{\beta}^*$ is equivalent to $\tilde{\beta}^*$, using Theorem 1 in Ahn and Schmidt (1995). Second, let $\check{\theta}^* = \left(\check{\beta}^*, \check{\gamma}^*_\Delta\right)$ and $\tilde{\theta}^* = \left(\tilde{\beta}^*, \tilde{\gamma}^*_\Delta\right)$. Since the latter estimator targets the same parameters, but uses more moment conditions, it follows from e.g. Theorem 6.1 in Hall (2005) that $\tilde{\theta}^*$ is at least as efficient as $\check{\theta}^*$.

**Step 1.** To show that $\check{\beta}^*$ and $\tilde{\beta}^*$ are equivalent, note that our problem can be written in the form of Ahn and Schmidt (1995, Section 2). In their notation $m_1\left(\theta_1\right) = s_{1,\beta}\left(\beta\right)$ and $m_2\left(\theta_1, \theta_2\right) = s_{2,\gamma}\left(\beta, \gamma_\Delta\right)$. Their Theorem 1 then requires that (i) the dimension of $s_{1,\beta}$ is equal to a greater than $\beta$, which holds because $\left(J-1\right)K \geq K$; (ii) the dimension of $s_{2,\gamma}$ is equal to that of $\gamma_\Delta$, which holds by construction; and that (iii) $E\left[\partial s_{1,\beta}/\partial\beta|_{\beta=\beta_0}\right]$ has full rank, and that $E\left[\partial s_{2,\gamma}/\partial\gamma_\Delta|_{\beta=\beta_0,\gamma_\Delta=\gamma_{\Delta,0}}\right]$ is invertible, which holds [...]. Their Theorem 1 therefore applies, so that $\check{\beta}^* = \tilde{\beta}^*$, which implies $V_2 = V_1$.

**Step 2.** To show that $\hat{\beta}^*$ is at least as efficient as $\check{\beta}^*$, note that $\check{\theta}^* = \left(\check{\beta}^*, \check{\gamma}^*_\Delta\right)$ and $\tilde{\theta}^* = \left(\tilde{\beta}^*, \tilde{\gamma}^*_\Delta\right)$ are both estimators for $\left(\beta, \gamma_\Delta\right)$. The former is based on a subset $\left(s_{1,\beta}, s_{2,\gamma}\right)$. The latter is based on the full set $\left(s_{1,\beta}, s_{2,\gamma}, s_{2,\beta}\right)$. The regularity conditions in Theorem 6.1 in Hall are satisfied, and the proof of positive semi-definiteness there applies, which means that $Avar\left(\check{\theta}^*\right) - Avar\left(\hat{\theta}^*\right)$ is psd. This implies that $V_2 - V_3$ is psd.

Taking together the results from Steps 1 and 2, we can conclude that $V_1 = V_2 \geq V_3$.

**Part 2: Asymptotic normality.** To establish asymptotic normality of the estimators, I check the conditions of Theorem 3.4 in Newey and McFadden (1994), which is a result for asymptotic normality of a general GMM estimator. Condition (i, interior solution) is satisfied because the parameter space is open. Condition (ii, continuously differentiable moment function) follows from the properties of the second derivative of the criterion function for each transformation $\pi$, described in detail in Section A.3. Condition (iii, moment function are zero-mean, finite-variance) is satisfied because the moment functions are score, and because each score is finite by part (ii) of Theorem 3. Condition (iv, bounded envelope) is implied by the existence of the Hessian for each transformation, for every value of the parameters, see part (iii) of Theorem 3.

What remains to be shown show is condition (v, invertibility of efficiency bound). For $\tilde{\beta}^{*}$, this amounts to verifying positive definiteness of $H_1^{'}\Sigma_1^{-1}H_1$, where $H_1$ is the $(J-1)\,K \times K$ matrix that stacks the $K \times K$ Hessians from the time-invariantly transformed models on top of each other:

$$H_1 = \begin{bmatrix} H_{(1,\cdots,1)} \\ \vdots \\ H_{(J-1,\cdots,J-1)} \end{bmatrix}$$

and $\Sigma_1 = E\left[s_{i,1,\beta}\left(\beta_0\right) s_{i,1,\beta}\left(\beta_0\right)^{'}\right]$. Since each block in $H_1$ has full rank, the matrix $H_1$ has full column rank, so that $H_1^{'}\Sigma_1^{-1}H_1$ is positive definite if and only if $\Sigma_1$ is positive definite.

Similarly, for $\hat{\beta}^{*}$, this amount to verifying that $H^{'}\Sigma^{-1}H$ is invertible, where $H$ stacks the $(K+n_\gamma)\times(K+n_\gamma)$ Hessians from all the transformations.[27] It is easy to see that $H$ has full column rank, so that positive definiteness of $\Sigma = E\left[s_i\left(\beta,\gamma_\Delta\right)s_i\left(\beta,\gamma_\Delta\right)^{'}\right]$ is necessary and sufficient for invertibility of $H\Sigma^{'-1}H$. If $\Sigma$ is pd, then $\Sigma_1$ is pd, because it is a principal submatrix of $\Sigma$.

Consider the simple case that there are only two transformations $\pi$ and $\pi^{'}$, and that $S_\pi = S_{\pi^{'}}$.[28] Letting $a_{fd} = \exp\left\{(f-d)\,Z_{i\pi}\theta_{\pi 0}\right\}$ and

$$A_d = \frac{\sum_{f,g} a_{fd}a_{gd}Z_{i\pi}^{'}\left(f-d\right)^{'}\left(g-d\right)Z_{i\pi}}{\left(\sum_f a_{fd}\right)^2}$$

obtains that

$$\begin{aligned} \Sigma &= E\begin{bmatrix} s_{i\pi}s_{i\pi}^{'} & s_{i\pi}s_{i\pi^{'}}^{'} \\ s_{i\pi^{'}}s_{i\pi}^{'} & s_{i\pi^{'}}s_{i\pi^{'}}^{'} \end{bmatrix} \\ &= \sum_{d\in\{0,1\}} E\left\{\begin{bmatrix} 1\left\{d_{i\pi}=d\right\} & 1\left\{d_{i\pi}=d\right\}1\left\{d_{i\pi^{'}}=d\right\} \\ 1\left\{d_{i\pi}=d\right\}1\left\{d_{i\pi^{'}}=d\right\} & 1\left\{d_{i\pi^{'}}=d\right\} \end{bmatrix}\otimes A_d\right\} \qquad (57) \end{aligned}$$

using that $S_\pi = S_{\pi^{'}}$so that $Z_{i\pi} = Z_{i\pi^{'}}$, and abbreviating $\sum_f = \sum_{f\in F_{\bar{d}}}$. Consider

$$Q \equiv E\left[\left[\begin{matrix} 1\left\{d_{i\pi}=d\right\} & 1\left\{d_{i\pi}=d\right\}1\left\{d_{i\pi^{'}}=d\right\} \\ 1\left\{d_{i\pi}=d\right\}1\left\{d_{i\pi^{'}}=d\right\} & 1\left\{d_{i\pi^{'}}=d\right\} \end{matrix}\right]\middle| X_i,\alpha_i\right] \qquad (58)$$

[27] This Hessian is described in detail in Appendix A.7.1.

[28] The general case is tedious, and does not provide additional insight, and is therefore omitted.

which can be written as

$$Q = \left[ \begin{array}{cc} P\left(d_{i\pi} = d \middle| X_i, \alpha_i\right) & P\left(d_{i\pi} = d_{i\pi'} = d \middle| X_i, \alpha_i\right) \\ P\left(d_{i\pi} = d_{i\pi'} = d \middle| X_i, \alpha_i\right) & P\left(d_{i\pi'} = d \middle| X_i, \alpha_i\right) \end{array} \right]$$

The matrix $Q$ will be of full rank if $P\left(d_{i\pi} = d_{i\pi'} = d| X_i, \alpha_i\right) < P\left(d_{i\pi} = d| X_i, \alpha_i\right)$ or $P\left(d_{i\pi} = d_{i\pi'} = d| X_i, \alpha_i\right) < P\left(d_{i\pi'} = d| X_i, \alpha_i\right)$. To see that this holds, consider a time period $t$ for which $\pi\left(t\right) \neq \pi^{'}\left(t\right)$. Without loss of generality, let $\pi\left(t\right) < \pi^{'}\left(t\right)$. Then

$$\begin{aligned} P\left(d_{i\pi(t)} \neq d_{i\pi'(t)} \middle| X_i, \alpha_i\right) &= P\left(\gamma_{\pi(t)} < y_{it}^{*} < \gamma_{\pi'(t)}\right) \\ &= \Lambda\left(\gamma_{\pi'(t)} - \alpha_i - X_{it}\beta\right) - \Lambda\left(\gamma_{\pi(t)} - \alpha_i - X_{it}\beta\right) \\ &> 0, \end{aligned}$$

which implies that the unconditional probability $P\left(d_{i\pi} \neq d_{i\pi'}\right) > 0$, which implies that the matrix $Q$ in (58) is invertible. Then, from (57), $\Sigma$ is invertible. □

## A.7 GMM and OMD estimation

Subsection A.7.1 presents the structure of the Hessian involved in the efficient GMM estimator. Subsection A.7.2 contains the derivations underlying Remark 5, which claims that a certain optimal minimum distance estimator is equivalent to the optimal GMM estimator in Theorem 4. Subsection A.7.3 contains some details about the implementation of the estimator in Stata.

### A.7.1 GMM: Structure of $H$

To describe the structure of $H$, I order the transformations by letting $\pi_1, \cdots, \pi_{J-1}$ be the $J-1$ time-invariant transformations, and letting $\pi_J, \cdots, \pi_{(J-1)^T}$ be the remaining transformations. For each transformation, the Hessians of the CMLE, $H_\pi$ in (22), can be dividided into blocks corresponding to the regression coefficient and cut point difference,

$$H_\pi \equiv \begin{bmatrix} H_\pi^\beta & H_\pi^{\beta\gamma} \\ H_\pi^{\gamma\beta} & H_\pi^\gamma \end{bmatrix}.$$

The matrix $H$ has the following structure

$$H = \begin{bmatrix} H^{\beta}_{\pi_1} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ H^{\beta}_{\pi_{J-1}} & 0 & 0 & \cdots & 0 \\ H^{\beta}_{\pi_J} & H^{\beta\gamma}_{\pi_J} & 0 & \cdots & 0 \\ H^{\gamma\beta}_{\pi_J} & H^{\gamma}_{\pi_J} & 0 & \cdots & 0 \\ H^{\beta}_{\pi_{J+1}} & 0 & H^{\beta\gamma}_{\pi_{J+1}} & \cdots & 0 \\ H^{\gamma\beta}_{\pi_{J+1}} & 0 & H^{\gamma}_{\pi_{J+1}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H^{\beta}_{\pi_{(J-1)^T}} & 0 & 0 & \cdots & H^{\beta\gamma}_{\pi_{(J-1)^T}} \\ H^{\gamma\beta}_{\pi_{(J-1)^T}} & 0 & 0 & \cdots & H^{\gamma}_{\pi_{(J-1)^T}} \end{bmatrix}. \tag{59}$$

### A.7.2 OMD: Derivations

Consider the set of CMLE estimators $\hat{\theta}_\pi$ from (18), for all transformations $\pi$. Consider that these CMLE estimators target the parameters $(\beta_\pi, \gamma_{\pi,\Delta})$ without imposing the assumption that $\beta_\pi = \beta_0$ for all $\pi$. Label the transformations $\pi_1, \cdots, \pi_{(J-1)^T}$ , letting $\pi_1 \cdots, \pi_{J-1}$ be the time-invariant transformations.[29] Then denote by $\hat{\tilde{\theta}}$ the estimator that stacks all the CMLE estimators, i.e.

$$\hat{\tilde{\theta}} = \begin{bmatrix} \hat{\beta}_{\pi_1} \\ \vdots \\ \hat{\beta}_{\pi_{J-1}} \\ \hat{\beta}_{\pi_J} \\ \hat{\gamma}_{\pi_J,\Delta} \\ \vdots \\ \hat{\beta}_{\pi_{(J-1)^T}} \\ \hat{\gamma}_{\pi_{(J-1)^T},\Delta} \end{bmatrix}.$$

and denote by $\tilde{\theta}_0$ the true value of the stacked target parameter.

The stacked estimator $\hat{\tilde{\theta}}$ corresponds to the GMM estimator based on the full set of moment conditions in 29, without taking into account that $\beta_\pi = \beta_0$ for all $\pi$, i.e. it is

[29] Note that we have used the ordering and numbering of the CMLEs used in Section A.7.1, introduced before equation (59).

the GMM estimated based on

$$E\begin{bmatrix} s_{i,\pi_1}\left(\beta_{\pi_1}\right) \\ \vdots \\ s_{i,\pi_{J-1}}\left(\beta_{\pi_{J-1}}\right) \\ s_{i,\pi_J}\left(\beta_{\pi_J},\gamma_{\pi_J,\Delta}\right) \\ \vdots \\ s_{i,\pi_J}\left(\beta_{\pi_{(J-1)^T}},\gamma_{\pi_{(J-1)^T},\Delta}\right) \end{bmatrix} = 0. \tag{60}$$

Therefore, the estimator $\hat{\tilde{\theta}}$ can be seen as a GMM estimator that uses the same moment conditions as those used by the optimal GMM estimator in Theorem 4. Consequentially, the variance matrix associated with the moment conditions of this estimator is the same as that for the GMM estimators, namely $\Sigma$ in equation (31). However, the Hessian associated with (60) is

$$H_{MD} = \begin{bmatrix} H^{\beta}_{\pi_1} & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & 0 & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & H^{\beta}_{\pi_{J-1}} & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & H^{\beta}_{\pi_J} & 0 & \cdots & 0 & H^{\beta,\gamma}_{\pi_J} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & H^{\beta,\gamma}_{\pi_J} & 0 & \cdots & 0 & H^{\gamma}_{\pi_J} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & H^{\beta}_{\pi_{J+1}} & \cdots & 0 & 0 & H^{\beta,\gamma}_{\pi_{J+1}} & \cdots & 0 \\ 0 & \cdots & 0 & 0 & H^{\beta,\gamma}_{\pi_{J+1}} & \cdots & 0 & 0 & H^{\gamma}_{\pi_{J+1}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & H^{\beta}_{\pi_{(J-1)^T}} & 0 & 0 & \cdots & H^{\beta,\gamma}_{\pi_{(J-1)^T}} \\ 0 & \cdots & 0 & 0 & 0 & \cdots & H^{\beta,\gamma}_{\pi_{(J-1)^T}} & 0 & 0 & \cdots & H^{\gamma}_{\pi_{(J-1)^T}} \end{bmatrix}$$

which is different from the Hessian for the GMM estimators in Section 5, introduced as $H$ in section A.7.1. The Hessian is different because derivatives for the regression coefficient are now with respect to $(\beta_\pi,\, \pi)$ instead of with respect to $\beta_0$.

The estimator $\hat{\tilde{\theta}}$ is a method of moments estimator, because the number of parameters is the same as the number of moment conditions. Its asymptotic distribution is

$$\sqrt{n}\left(\hat{\tilde{\theta}} - \tilde{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(H'_{MD}\Sigma^{-1}H_{MD}\right)^{-1}\right),$$

where $\tilde{\theta}_0$ is the vector of true values of the parameters targeted by the stacked CMLEs, so that

$$
\begin{aligned}
\tilde{\theta}_0 &= \left[\begin{array}{ccccc}
I_K & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \cdots & 0 \\
I_K & 0 & 0 & \cdots & 0 \\
I_K & 0 & 0 & \cdots & 0 \\
0 & I_{n_{\pi_J}} & 0 & \cdots & 0 \\
I_K & 0 & 0 & \cdots & 0 \\
0 & 0 & I_{n_\pi} & \cdots & 0 \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
I_K & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & I_{n_{\pi_{(J-1)T}}}
\end{array}\right] \begin{pmatrix} \beta_0 \\ \gamma_{\Delta,0} \end{pmatrix} \\
&\equiv R \begin{pmatrix} \beta_0 \\ \gamma_{\Delta,0} \end{pmatrix}.
\end{aligned}
$$

This suggests that $(\beta_0, \gamma_{\Delta,0})$ can be estimated using classical minimum distance. The optimal minimum distance estimator (see Newey and McFadden, 1994, p. 2164), which I will denote by

$$\hat{\theta}^*_{md} = \left(\hat{\beta}^*, \hat{\gamma}^*_\Delta\right) \tag{61}$$

sets $W_{md} = \left(H'_{MD}\Sigma^{-1}H_{MD}\right)$ in the minimization

$$\hat{\theta}_{W,md} = \text{argmin} \left(\hat{\tilde{\theta}} - R\hat{\theta}_{W,md}\right)' W_{md} \left(\hat{\tilde{\theta}} - R\hat{\theta}_{W,md}\right).$$

The resulting estimator is $\hat{\theta}^*_{md} = \left(R' H'_{MD}\Sigma^{-1}H_{MD}R\right)^{-1} R' H'_{MD}\Sigma^{-1}H_{MD}\hat{\theta}$. Inspecting $R$, $H_{MD}$, and $H$ reveals that

$$H_{MD}R = H. \tag{62}$$

It follows that

$$\hat{\theta}^*_{md} = \left(H'\Sigma^{-1}H\right)^{-1} H'\Sigma^{-1}H_{MD}\hat{\tilde{\theta}}$$

and

$$\sqrt{n}\left(\hat{\theta}^*_{md} - \begin{pmatrix} \beta_0 \\ \gamma_{\Delta,0} \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(0, V\right)$$

using (62) once more. In conclusion, the optimal MD estimator from all CMLE's is asymptotically equivalent to the optimal GMM estimator.

### A.7.3 Implementation

This subsection discusses the implementation of the procedure in A.7.2, assuming available software that returns $\hat{\hat{\theta}}$ and its variance. In Stata, each block in

$$\hat{\hat{\theta}} = \begin{bmatrix} \hat{\beta}_{\pi_1} \\ \vdots \\ \hat{\beta}_{\pi_{J-1}} \\ \hat{\beta}_{\pi_J} \\ \hat{\gamma}_{\pi_J,\Delta} \\ \vdots \\ \hat{\beta}_{\pi_{(J-1)T}} \\ \hat{\gamma}_{\pi_{(J-1)T},\Delta} \end{bmatrix}$$

can be computed using **clogit**, as explained in the main text, in the implementation section. Furthermore, Stata's **suest** can be used to compute a consistent estimate of

$$\left(H_{MD}^{'}\Sigma H_{MD}\right)^{-1},$$

say $\hat{\Omega}^{-1}$. Then, a feasible version of the optimal minimum distance estimator can be computed by evaluating

$$\hat{\theta}^{*}_{md} = \left(R^{'}\hat{\Omega}R\right)^{-1} R^{'}\hat{\Omega}\hat{\hat{\theta}}$$

which uses the known matrix $R$ (see previous subsection), the results from **clogit**, and the result from **suest**.