# Linear Estimation of Structural and Causal Effects for Nonseparable Panel Data

Victor Chernozhukov (MIT), Ben Deaner (UCL), Ying Gao (UBC)
Jerry Hausman (MIT and NBER), Whitney K. Newey (MIT and NBER)*

May 15, 2025

**Abstract**

This paper develops linear estimators for structural and causal parameters in nonparametric, nonseparable models using panel data. These models incorporate unobserved, time-varying, individual heterogeneity, which may be correlated with the regressors. Estimation is based on an approximation of the nonseparable model by a linear sieve specification with individual-specific parameters. Effects of interest are estimated by a bias corrected average of individual ridge regressions. We demonstrate how this approach can be applied to estimate causal effects, counterfactual consumer welfare, and averages of individual taxable income elasticities. We show that the proposed estimator has an empirical Bayes interpretation and possesses a number of other useful properties. We formulate Large-$T$ asymptotics that can accomodate discrete regressors and which bypass partial identification in this case. We employ the methods to estimate average equivalent variation and deadweight loss for potential price increases using data on grocery purchases.

## 1 Introduction

Panel data provide a valuable means of identifying and estimating economic effects when there is dependence between variables of interest and unobservable heterogeneity. Specifically, in "fixed effects" type models, time-invariance of heterogeneity like tastes or technology can be combined with time variation in variables of interest to identify and estimate economic effects of those variables. Similarly, in causal models panel data can be used to identify and estimate counterfactual effects of interest when treatment varies over time and unobserved confounders do not. Ideally, these structural or causal parameters are identified in nonparametric outcome models that are not additively separable in observables and unobservables. Such models provide very general specifications for economic and causal models, with heterogeneity representing tastes and/or technology for economic models or counterfactual outcomes in causal models.

This paper develops linear methods of identifying and estimating economic or treatment effects for nonseparable models using panel data. The basic identifying assumption is a "fixed effects" condition, referred to henceforth as time-stationarity, that the conditional distribution of heterogeneity in each time period given observed regressors does not depend on the time period, as in Manski (1987). Time effects are also allowed via regressors whose magnitude varies with time. To use linear methods we approximate a smooth nonseparable model by a linear model with coefficients that depend only on the unobserved heterogeneity, and so inherit the time-stationarity property.

*The present paper formed the basis of the Fisher–Schultz Lecture given by Whitney Newey at the 2023 European Meeting of the Econometric Society in Barcelona. This research was supported by NSF Grants 1757140 and 224247. Helpful comments were provided by I. Fernandez-Val, B. Graham, G. Imbens, R. Matzkin. Any inquiries email wnewey@mit.edu or b.deaner@ucl.ac.uk.

Least squares for each individual can then be used to estimate individual-specific coefficients. We regularize using ridge regression to allow the individual coefficients to be weakly identified and bias correct the average ridge estimator.

The effects of interest we consider are averages of individual-specific effects. Examples of these effects given here are average policy or treatment effects, average equivalent variation and deadweight loss for demand, and average tax effects for nonlinear budget sets. We estimate these as bias corrected averages of individual-specific linear combinations of the individual ridge regression coefficients. The bias correction makes the estimator of average effects be unbiased when true slope coefficients, i.e. coefficients of nonconstant regressors, do not vary over individuals. We give an empirical Bayes interpretation of the bias corrected average of individual ridge regression coefficients. We show that as the ridge parameter goes to infinity the debiased average parameter approaches the fixed effects estimator that imposes that slopes are constant across individuals. We also find that when all individuals have a nonsingular second moment matrix of regressors, i.e., all individual-specific coefficients are identified, the bias of the estimator of average effects goes to zero as the ridge parameter does.

We provide means of quantifying the extent of identification across individuals of effects of interest. We compare the pre-specified, true linear combination for each individual with regularized linear combinations implied by debiased ridge. We measure the extent of identification using the distribution across individuals of the discrepancy between true linear combinations and regularized counterparts.

We apply our methods to estimate bounds on average equivalent variation and deadweight loss for consumer demand. Here, panel data approximately controls for price endogeneity from imperfect competition with many consumers when time varying individual heterogeneity is independent of unobserved supply shocks, as discussed in Section 4. The methods are applied to scanner data with outcome variable specified as expenditure share and regressors as powers of the log of prices and total expenditure. Average equivalent variation and deadweight loss for price increases on soda or milk are estimated.

We develop large sample theory where the number $T$ of time periods grows with the number of individuals. The number of regressors is allowed to grow with $T$ so that the approximation error for a general nonseparable model is small enough for accurate inference. We also allow for nonidentifiability of effects of interest for a fixed number of time periods $T$, but specify that the identified set shrinks at some power of $T$ that is sufficient for asymptotic inference using the regularized ridge estimates. Conditions for such shrinkage rates for the identified set are known from Chernozhukov *et al.* (2013).

Chamberlain (1982), Chamberlain (1992), Pesaran & Smith (1995), Wooldridge (2005), Arellano & Bonhomme (2012), and Graham & Powell (2012) have previously considered estimation of panel models with individual-specific slope coefficients. Manski (1987), Honore (1992), Abrevaya (2000), Chernozhukov *et al.* (2013), Hoderlein & White (2012), Shi *et al.* (2018), and Pakes & Porter (2024) have all considered estimation of nonseparable panel models under time-stationarity. Altonji & Matzkin (2005) considers identification via control functions as does Semenova *et al.* (2023) while also allowing for sparse, additive individual-specific effects. We innovate in approximating a general nonseparable model by a linear one, in the use of bias corrected ridge regularization, in providing methods for evaluating the extent of identification across individuals, and in providing asymptotic theory for growing $T$ under partial identification for fixed $T$.

In Section 2 we describe the model and effects we consider and give the linear approximation to the conditional mean of a nonseparable outcome function. Section 3 gives the bias corrected average ridge estimator of structural and causal effects and describes its properties. Section 4 describes how the general model and methods can be applied to demand analysis. Section 5 gives an application to scanner data. Section 6 gives large sample theory.

## 2 Model and Parameters of Interest

We consider panel data made up of observations across $n$ individuals, indexed by $i$, and $T$ time periods indexed by $t$. The individuals are drawn independently and identically from some population. Each observation consists of a scalar outcome variable $S_{it}$ and a vector of regressors $X_{it}$. Let $X_i = (X_{i1}', ..., X_{iT}')'$ be the full history of regressors. We assume that $S_{it}$ satisfies a nonseparable model:

$$E[S_{it}|X_i, \eta_{it}] = s(X_{it}, \eta_{it}), \ (t = 1, ..., T; \ i = 1, ..., n), \tag{2.1}$$

where $\eta_{it}$ are individual and time specific unobserved variables representing preferences, technology, or heterogeneity in potential outcomes. A leading case is that where $\eta_{it}$ represents all unobserved heterogeneity, so that $S_{it} = s(X_{it}, \eta_{it})$. The additional generality of (2.1) is important with discrete outcomes because it allows for $s(\cdot, \eta_{it})$ to be a smooth function, via integration over unobserved heterogeneity additional to $\eta_{it}$.

We do not restrict the dimension of $\eta_{it}$ and so allow for the possibility that it is infinite-dimensional. The model is nonseparable in the sense that it allows for general interactions between the observed variables $X_{it}$ and the unobserved $\eta_{it}$. We take the model (2.1) to have a structural or causal interpretation where $s(x_{it}, \eta_{it})$ represents the mean potential outcome when $x_{it}$ differs from the observed $X_{it}$.

We use conditional time-stationarity of $\eta_{it}$ to identify and estimate objects of interest.

**Assumption 1 (Time-Stationarity).** The distribution of $\eta_{it}$ conditional on $X_i$ does not depend on $t$.

This condition allows for endogeneity where the conditional distribution of $(\eta_{i1}, ..., \eta_{iT})'$ given $X_i$ may depend on $X_i$. Such endogeneity is present when $X_{it}$ includes choice or equilibrium values that are determined by the preferences or technology represented by $\eta_{it}$. Time-varying preferences and technology are often allowed for in panel data applications and can be empirically important, as we discuss further in the context of our objects of interest. Assumption 1 requires that the time-varying components of $\eta_{it}$ have the same distribution in each period conditional on the history of regressors $X_i$. It is common to decompose $\eta_{it}$ into time constant components $\alpha_i$ and time varying components $v_{it}$. Such an $\alpha_i$ trivially satisfies Assumption 1, so that Assumption 1 only restricts $v_{it}$. Assumption 1 does allow for systematic variation over time in $S_{it}$ via regressors. i.e. elements of $X_{it}$, with magnitudes that vary with time, such as time trends or seasonal indicators.

Our objects of interest are differences in weighted means of counterfactual outcomes. Let $X_{it}^+$ and $X_{it}^-$ be counterfactual regressors which can differ from the factual value of the regressors $X_{it}$. Let $H_{it}^+$ and $H_{it}^-$ be weights specified by the researcher. While we refer to these as 'weights' they may be both positive and negative and they need not sum to unity. We identify and estimate objects of the form

$$\theta_0 = E\big[\frac{1}{T}\sum_{t=1}^{T}\big(H_{it}^+ s(X_{it}^+, \eta_{it}) - H_{it}^- s(X_{it}^-, \eta_{it})\big)\big] \tag{2.2}$$

We assume throughout that $X_{it}^+$, $X_{it}^-$, $H_{it}^+$, and $H_{it}^-$ are functions of $X_i$ and possibly some random noise whose distribution is known to the researcher. More precisely, we require the following condition:

**Assumption 2 (Counterfactuals).** $X_{it}^+$, $X_{it}^-$, $H_{it}^+$, and $H_{it}^-$ are jointly independent of $\eta_{it}$ conditional on $X_i$.

A number of important policy-relevant quantities may be written in the form of $\theta_0$. Three examples we consider are average causal effects of alternative treatment regimes, bounds on average equivalent variation and on deadweight loss in demand analysis, and taxable income effects with nonlinear budget sets.

**Example 1: Average Effects of Alternative Treatment Regimes**

In order to define causal effects within the model we let $s(x_{it}, \eta_{it})$ represent the potential outcome from treatment $x_{it}$, for individual $i$ in period $t$, when $S_{it} = s(X_{it}, \eta_{it})$. The heterogeneity $\eta_{it}$ then captures the variation in potential outcomes, e.g. similarly to Imbens & Newey (2009).

When $S_{it}$ is discrete we let $s(x_{it}, \eta_{it})$ represent the average potential outcome assuming that $\eta_{it}$ captures all confounding, so that $s(x_{it}, \eta_{it})$ is an average over time-stationary unobserved heterogeneity that is independent of $X_i$.

Consider two counterfactual treatment assignment processes. In the first, an individual $i$ in period $t$ receives a random treatment $X_{it}^-$ and in the second they receive treatment $X_{it}^+$. These counterfactual treatments may depend on $X_i$. For example, we may wish to compare mean outcomes under the counterfactual assignments $X_{it}^+$ with the factual treatments, in which case we can set $X_{it}^- = X_{it}$. The difference in expected time-average outcomes between the two regimes is

$$\theta_0 = E[\frac{1}{T}\sum_{t=1}^{T}\{s(X_{it}^+, \eta_{it}) - s(X_{it}^-, \eta_{it})\}].$$

The object above is of the form (2.2) with both the weights $H_{it}^+$ and $H_{it}^-$ set to unity. A major challenge for inference in causal models is the possibility of unmeasured confounding. That is, there may be latent factors that jointly determine the treatment assignment $X_{it}$ and the variation in potential outcomes $\eta_{it}$. Assumption 1 allows for the possibility of unobserved confounding under suitable conditions on the time-dependence structure.

We can motivate Assumption 1 in this context using a nonparametric structural model for the treatment assignments and the heterogeneity in potential outcomes. Let $\alpha_i$ be a vector of time-invariant confounding factors and consider the following model where the time-varying innovations $\{u_{it}\}_{t=1}^T$ and $\{v_{it}\}_{t=1}^T$ are each jointly independent of $\alpha_i$:

$$\eta_{it} = e(\alpha_i, u_{it}),\ X_{it} = x(\alpha_i, v_{it})$$

The first equation decomposes the heterogeneity in potential outcomes into variation between individuals, captured in $\alpha_i$, and variation over time $u_{it}$. Let us suppose that the innovations $\{u_{it}\}_{t=1}^T$ are jointly independent of $\{v_{it}\}_{t=1}^T$. This implies $u_{it}$ is independent of the history of treatments $X_i$. In addition, suppose the marginal distribution of $u_{it}$ (but not necessarily $v_{it}$), is time-invariant. Under these conditions Assumption 1 holds.

In this model the time-invariant factors $\alpha_i$ are akin to fixed-effects. They are individual-specific characteristics that explain the confounding between treatments and outcomes but which do not vary over time. The condition that the temporal variation in potential outcomes $u_{it}$ is independent of the history of treatment assignments is akin to strict exogeneity. Unlike in the classic fixed effects model, $\alpha_i$ may enter non-separably into the possibly non-linear model for the outcome $S_{it}$. Similar panel data treatment effect models were explicitly formulated in Chernozhukov *et al.* (2013) and Torgovitsky (2019).

**Example 2: Average Equivalent Variation and Deadweight Loss Bounds**

The average equivalent variaton and deadweight loss of a price change are important objects of interest in empirical demand analysis. Obtaining bounds on these quantities is a crucial step in assessing the welfare impact of a policy that may alter consumer prices, such as a sales tax. Suppose $S_{it} = s(X_{it}, \eta_{it})$ is the expenditure share of some commodity where $X_{it} = (P_{it}, Z_{it})$, $P_{it}$ is the product price, and $Z_{it}$ is a vector of covariates that includes total expenditure $Y_{it}$ and the prices of other goods.

In order to define the welfare effects of a price change we must choose an initial price for good $i$ at time $t$, which we denote by $P_{it}^-$. This starting price may depend on $X_i$. For example, $P_{it}^-$ could

simply be $P_{it}$, the price paid in period $t$ by individual $i$ for the good. Let $\Delta_{it}$ denote the change in the price of the good for individual $i$ in period $t$ that also may depend on $X_i$. Let $\omega_t(X_i)$ be some weighting that may depend upon $X_i$. Using Hausman & Newey (2016), we obtain bounds on the weighted average equivalent variation from a price change from $P_{it}^-$ to $P_{it}^- + \Delta_{it}$. Let $\pi$ be an upper or lower bound on the income effect for every individual and let $U_i = (U_{i1}, ..., U_{iT})'$ be a vector of $T$ random variables that are uniformly distributed on $(0, 1)$ and independent of the data, i.e. that are simulation draws from the standard uniform distribution. Taking $S_{it} = s(p_{it}, Z_{it}, \eta_{it})$ to be the counterfactual demand at price $p_{it}$, a bound on the weighted average equivalent variation is

$$\theta_{EV} = E[\frac{1}{T}\sum_{t=1}^{T} H_{it}^+ s(P_{it}^- + \Delta_{it}U_{it}, Z_{it}, \eta_{it})] \tag{2.3}$$

$$H_{it}^+ = \omega_t(X_i) \exp\big(-\pi(P_{it}^- + \Delta_{it}U_{it})\big)\Delta_{it}\frac{Y_{it}}{P_{it}^- + \Delta_{it}U_{it}}, \tag{2.4}$$

If $\pi$ is a lower (upper) bound on the income effect for every individual then, by Hausman & Newey (2016), $\theta_{EV}$ is an upper (lower) bound on average over time and individuals of the equivalent variation for a change from $P_{it}^-$ to $P_{it}^- + \Delta_{it}$, weighted by $\omega_t(X_i)$. The weights $\omega_t(X_i)$ allow us to assess the welfare impact on particular sub-populations, such as those in a low income bracket or with a certain family size.

A corresponding deadweight loss bound can be obtained by subtracting the weighted average change in final demand as below.

$$\theta_{DWL} = \theta_{EV} - E\big[\frac{1}{T}\sum_{t=1}^{T} H_{it}^- s(P_{it}^- + \Delta_{it}, Z_{it}, \eta_{it})\big] \tag{2.5}$$

$$H_{it}^- = \omega_t(X_i)\Delta_{it}\frac{Y_{it}}{P_{it}^- + \Delta_{it}} \tag{2.6}$$

If $\pi$ is a lower (upper) bound on the income effect then $\theta_0$ will be an upper (lower) bound for weighted dead weight loss averaged over all time periods and individuals.

Both $\theta_{EV}$ and $\theta_{DWL}$ are of the form in (2.2). In particular, $\theta_{EV}$ corresponds to the case with $H_{it}^+$ defined by (2.4), $H_{it}^- = 0$, and $X_{it}^+ = (P_{it}^- + \Delta_{it}U_{it}, Z_{it}')'$. The deadweight loss shares these choices of $H_{it}^+$ and $X_{it}^+$, but in this case $H_{it}^-$ is set as in (2.6) and $X_{it}^- = (P_{it}^- + \Delta_{it}, Z_{it}')'$. This example is discussed further in Section 5, where an application to scanner data is given.

**Example 3: Average Heterogeneous Taxable Income Elasticities**

In some structural economic models the nonseparable model may be a random coefficients model where the expectation of one or more of the coefficients is of interest. An example is a parametric version of the panel budget set regression of Blomquist *et al.* (2024). An isoelastic utility function with individual specific elasticity and productivity together with scale heterogeneity varying identically over time and independently of the elasticity, growth rate, and budget set gives a budget set regression of the form

$$s(X_{it}, \eta_{it}) = \beta_{1i} + \sum_{j=2}^{4} X_{jit}\beta_{ji}. \tag{2.7}$$

Here $\beta_{2i}$ is the taxable income elasticity for individual $i$, $X_{2it}$ is the log of the slope of the last budget segment for individual $i$ in period $t$, $X_{3it}$ is the difference of logs of the slope of the first and last segment, and $X_{4it} = t$. This is a panel version of the budget set regression of Blomquist *et al.* (2024) that gives taxable income as a function of the budget set and unobserved heterogeneity. In this model the dependence of each $\beta_{ji}$ on $i$ allows for individual heterogeneity of preferences and for heterogeneous productivity growth.

The panel data setting of this model allows for endogeneity of budget sets where the slopes and intercepts of segments may be correlated with preferences and productivity growth. A parameter of interest in this model is

$$\theta_0 = E[\beta_{2i}] = E[\frac{1}{T}\sum_{t=1}^{T}\{s(X_{it}+e_2,\eta_{it}) - s(X_{it},\eta_{it})\}],$$

where $e_2$ is a unit vector with 1 in the second position and zeros elsewhere. This parameter is again of the form in (2.2). In this case $H_{it}^+$ and $H_{it}^-$ are both set to unity, $X_{it}^+ = X_{it} + e_2$, and $X_{it}^- = X_{it}$. There are a wide variety of estimates of a taxable income elasticity that is common across individuals; see Blomquist *et al.* (2024) for references. This model allows for heterogeneous taxable income elasticities where the parameter of interest is the average of those across individuals.

# 3 Linear Approximation and Estimation

## 3.1 Approximation

Our estimation methods are based upon a series approximation for the unknown, non-separable conditional function $s(x_t, \eta_t)$. For each fixed value of $\eta_t$, we consider an approximation that is linear in a $J \times 1$ vector of known basis functions $b(x_t) = (b_1(x_t), ..., b_J(x_t))'$. We assume throughout that $b_1(x_t) = 1$ so that the approximation includes an intercept. The coefficients in the approximation depend on the value of $\eta_t$. We denote them by $\beta(\eta_t) = (\beta_1(\eta_t), ..., \beta_J(\eta_t))'$. The approximation for the function $s(x_t, \eta_t)$ is

$$s(x_t, \eta_t) \approx b(x_t)'\beta(\eta_t). \tag{3.1}$$

In effect, we approximate $s(\cdot, \eta_t)$ separately for each value of $\eta_t$ by $b(\cdot)'\beta(\eta_t)$. To ensure a small approximation error uniformly in $(x_t, \eta_t)$ it suffices that $x_t$ be bounded and that $s(x_t, \eta_t)$ have derivatives with respect to $x_t$ of high enough order that are bounded uniformly $(x_t, \eta_t)$. For example there are Jackson theorems that give such approximations for polynomial $b(x_t)$. Other choices of $b(x_t)$ also give such approximations. We provide a more formal analysis of the approximation error in Section 6. Analogous approximations for demand functions were given in Hausman & Newey (2016) for $\eta_{it}$ independent of prices and total expenditure.

We have specified that $s(x_t, \eta_t)$ is a conditional mean so that this approximation can be valid when $S_{it}$ is discrete. To help explain consider a binary choice model where

$$S_{it} = 1(\Delta U(X_{it}, \alpha_i) + \tilde{v}_{it} > 0),$$

$\Delta U(x_{it}, \alpha_i)$ is a utility difference that is smooth in $x_{it}$, and $-\tilde{v}_{it}$ is independent of $X_i$ and $\alpha_i$ with unknown marginal CDF $G(v)$. Then specifying that $\eta_{it} = \alpha_i$ we have

$$E[S_{it}|X_i, \eta_{it}] = s(X_{it}, \eta_{it}) = G(\Delta U(x_{it}, \alpha_i)),$$

so that Assumption 1 is satisfied. Here $s(x_{it}, \eta_{it}) = G(\Delta U(x_{it}, \alpha_i))$, is a binary choice probability that will be smooth in $x_{it}$ as long as $G(\tilde{v})$ is smooth in $\tilde{v}$ and $\Delta U(x_{it}, \alpha_i)$ is smooth in $x_{it}$. Similarly, for $S_{it}$ that are discrete but not binary, we can formulate models so that equation (2.1) and Assumption 1 are satisfied for a conditional expectation $s(x_{it}, \eta_{it})$ that is smooth in $x_{it}$ by integrating over time-stationary unobservables that are independent of $X_i$ and $\alpha_i$.

Under time-stationarity as in Assumption 1 and the approximation in equation (3.1), for any $x_i$ that is conformable with $X_i$,

$$E[s(x_{it}, \eta_{it})|X_i] \approx E[b(x_{it})'\beta(\eta_{it})|X_i] = b(x_{it})'E[\beta(\eta_{it})|X_i] = b(x_{it})'\beta_i, \tag{3.2}$$

where $\beta_i$ is defined as

$$\beta_i := E[\beta(\eta_{it})|X_i]. \tag{3.3}$$

This approximation implies that the conditional expectation of the outcome $S_{it}$ has approximately a correlated random coefficients form. From (3.2), (3.3), and iterated expectations we get

$$E[S_{it}|X_i] = E[s(X_{it}, \eta_{it})|X_i] \approx b(X_{it})'\beta_i \tag{3.4}$$

The above demonstrates the crucial role that time-stationarity plays in our analysis. The assumption ensures that the coefficients in the approximate random-coefficients model are time-invariant. Thus they can be estimated using variation in the regressors over time for each individual.

The approximate model (3.4) has a structural/causal interpretation, with

$$b(x_{it})'\beta_i = E[b(x_{it})'\beta(\eta_{it})|X_i] \approx E[s(x_{it}, \eta_{it})|X_i] \tag{3.5}$$

being an approximate average potential outcome at $x_{it}$ conditional on $X_i$. Consequently our objects of interest are approximately averages of counterfactual linear combinations of $\beta_i$. For exposition consider Example 1. Suppose Assumption 1 holds and that $X_{it}^+$ and $X_{it}^-$ are functions of $X_i$, then applying the approximation in equation (3.2) we obtain

$$\theta_0 = E[\frac{1}{T}\sum_{t=1}^{T}\{s(X_{it}^+, \eta_{it}) - s(X_{it}^-, \eta_{it})\}] \approx E[a_i'\beta_i],\ a_i = \frac{1}{T}\sum_{t=1}^{T}\{b(X_{it}^+) - b(X_{it}^-)\}$$

Thus in Example 1, the parameter of interest is approximately the expected inner product of the known vector $a_i$ with the random coefficients $\beta_i$. More generally, we can approximate parameters of the form $\theta_0$ given in equation (2.2) using the formula

$$\theta_0 \approx E[a_i'\beta_i],\ \ a_i = \frac{1}{T}\sum_{t=1}^{T}\{H_{it}^+ b(X_{it}^+) - H_{it}^- b(X_{it}^-)\}. \tag{3.6}$$

Therefore, the parameter of interest $\theta_0$ is approximately an expectation of the inner product of a known vector $a_i$ with $\beta_i$.

## 3.2 Estimation

The approximation of $\theta_0$ above motivates an estimator. The expectation could be replaced with a sample average, and the unknown random coefficients replaced with estimates. The correlated random coefficients approximation (3.4) suggests that the unknown coefficients $\beta_i$ could estimated from a linear regression of $S_i = (S_{i1}, ..., S_{iT})'$ on $B_i = \big(b(X_{i1}), ..., b(X_{iT})\big)'$.

In practice, there could be high multicollinearity in this regression, particularly if $T$ is not much larger than the number of basis functions. Here we address this problem using individual specific ridge regression. Let $Q_i = B_i'B_i/T$, $D_i$ be a diagonal matrix with 0 as its upper left entry and all other diagonal entries strictly positive, and $\lambda$ a positive constant. A ridge regression estimator of $\beta_i$ is defined as

$$\hat{\beta}_i = (Q_i + \lambda D_i)^{-1} B_i' S_i/T. \tag{3.7}$$

The zero in the top left entry of $D_i$ ensures that we do not penalize the intercept in the ridge regression. By allowing $D_i$ to be individual-specific we can accommodate individual-level re-scaling of the regressors.

These individual ridge estimators are biased, as usual for ridge regression. In particular, ridge regression tends to shrink estimate towards zero. It is possible to mitigate this ridge bias in the estimation of $\theta_0 = E[a_i'\beta_i]$. Let $A_i$ denote a square $T$-dimensional matrix with $a_i'$ as its first row and its other rows being distinct rows of an identity matrix of dimension $J$ with the missing row not being orthogonal to $a_i$. Also, let

$$W_i = (Q_i + \lambda D_i)^{-1} Q_i \tag{3.8}$$

A debiased average ridge estimator of $\theta_0$ is then as follows.

$$\hat{\theta} = \bar{a}'(\overline{AW})^{-1}\overline{A\beta},\ \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i,\ \overline{AW} = \frac{1}{n}\sum_{i=1}^{n} A_i W_i,\ \overline{A\beta} = \frac{1}{n}\sum_{i=1}^{n} A_i\hat{\beta}_i \tag{3.9}$$

As we discuss in detail in Section 6, the estimator above tends to have smaller bias than the plug-in estimator $\frac{1}{n}\sum_{i=1}^{n} a_i'\hat{\beta}_i$ and is less sensitive to the choice of penalty parameter $\lambda$.

An estimator $\hat{V}$ for the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ can be obtained via the delta method as

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i^2,\ \hat{\psi}_i = (a_i - \bar{a})'(\overline{AW})^{-1}\overline{A\beta} + \bar{a}'(\overline{AW})^{-1}A_i[\hat{\beta}_i - W_i(\overline{AW})^{-1}\overline{A\beta}]$$

Example 3 illustrates that parameters of interest may include elements of the vector $E[\beta_i]$. Each component of this vector has the form $E[a_i'\beta_i]$ where $a_i$ is a unit vector. When $a_i$ is constant, the formula for the debiased estimator simplifies so that $A_i$ cancels out. Thus we obtain an estimator $\hat{\theta}$ of $E[\beta_i]$ and a corresponding estimator $\hat{V}$ of the asymptotic variance, for $\overline{W} := \sum_{i=1}^{n} W_i/n$,

$$\hat{\theta} = \overline{W}^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i,\ \hat{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i',\ \hat{\psi}_i = \overline{W}^{-1}(\hat{\beta}_i - W_i\hat{\theta}) \tag{3.10}$$

This estimator has a straight-forward interpretation. The term $\sum_{i=1}^{n}\hat{\beta}_i/n$ is the sample average of individual ridge estimates that suffers from ridge bias. Multiplying by $\overline{W}^{-1}$ effectively undoes the ridge bias on average.

## 3.3 Summary of Properties

The debiased panel ridge estimator has several interesting characteristics that help explain its form and how it may be used and interpreted. Here we provide a brief summary of these properties, with a more general discussion deferred to Section 6 along with our asymptotic analysis.

Firstly, the average coefficient estimator $\hat{\theta}$ of equation (3.10) can be interpreted as an empirical Bayes estimator. Suppose that $S_{it}$ is Gaussian and that each $\beta_i$ has a Gaussian prior with common nonzero mean. It turns out that $\hat{\theta}$ is the average of Maximum a Posterior estimators of $\beta_i$ when $\hat{\theta}$ is the common prior mean. In this way $\hat{\theta}$ has a self consistency property as a prior mean determined by the data and the average of Bayesian estimators. In Section 6 we derive this interpretation and a corresponding empirical Bayes interpretation of the debiased ridge estimator of $E[a_i'\beta_i]$ in equation (3.9).

The relatively simple example of the average coefficient of a single regressor helps describe other properties of the debiased ridge estimator. Suppose that $X_{it}$ is a scalar, $J = 2$, and that the approximation in equation (3.1) is exact. Let $Z_{it} := b_2(X_{it})$, $\tilde{Z}_{it} := Z_{it} - \sum_{s=1}^{T} Z_{is}/T$, and $\tilde{Q}_i := \sum_{s=1}^{T}\tilde{Z}_{is}^2/T$. Then the debiased average panel ridge estimator of equation (3.10) is

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(\tilde{Q}_i + \lambda)^{-1}\sum_{s=1}^{T}\tilde{Z}_{is}S_{is}/T}{\sum_{i=1}^{n}(\tilde{Q}_i + \lambda)^{-1}\tilde{Q}_i}.$$

Multiplying numerator and denominator by $\lambda$ it follows that as $\lambda$ goes to infinity $\hat{\theta}$ converges to the fixed effects estimator $[\sum_{i=1}^{n}\tilde{Q}_i]^{-1}\sum_{i=1}^{n}\tilde{Z}S_i$. Conversely, as $\lambda$ shrinks to zero $\hat{\theta}$ converges to the average of individual OLS estimators if $\tilde{Q}_i > 0$ for all individuals. Thus the choice of penalty parameter $\lambda$ allows a smooth transition between these two extremes. Also, by taking the conditional expectation we have

$$E[\hat{\theta}|X_1, ..., X_n] = \sum_{i=1}^{n} w_i\beta_{2i}/\sum_{i=1}^{n} w_i, w_i = \frac{\tilde{Q}_i}{\tilde{Q}_i + \lambda}.$$

Here we see that the debiased ridge estimator is unbiased when $\beta_{2i}$ does not vary with $i$. This property holds regardless of the choice of $\lambda$ and applies in cases in which $\tilde{Q}_i = 0$ for some individuals so that $\beta_i$ is not identified.

All of the properties described in the previous paragraph are shared by the general debiased ridge estimator of equation (3.9), with the condition $\tilde{Q}_i > 0$ replaced by $Q_i$ nonsingular and $\tilde{Q}_i = 0$ replaced by $Q_i$ singular, as shown in Section 6. One interesting property that is special to the example of one nonconstant regressor is that $E[\hat{\theta}|X_1, ..., X_n]$ is a weighted average of individual specific $\beta_i$, with larger weight $w_i$ given to observations where $\beta_i$ is more strongly identified, i.e. $\tilde{Q}_i$ is larger. In general the conditional expectation of the debiased ridge estimator is a matrix weighted average with weights discussed more fully in Section 6.

The debiased average ridge estimator belongs to a general class of regularized panel estimators based on a regularized version $Q_i^{--}$ of $Q_i^{-1}$ that is well defined even when $Q_i$ is singular. This general class consists of estimators obtained from equations (3.7)-(3.9) with $Q_i^{--}$ replacing $(Q_i + \lambda D_i)^{-1}$. The average coefficient estimator of Graham & Powell (2012) is a member of this general class where $Q_i^{--} = 1(det(Q_i) > k)Q_i^{-1}$, understood to be zero when $det(Q_i) \leq k$ and $a_i$ is constant. When there is more than one non-constant regressor the debiased average ridge estimator has the advantage that informative observations may be used when they are not by the Graham & Powell (2012) estimator. For example suppose $Q_i$ is singular because a regressor other than the first one is constant over time. The $ith$ observation may still be informative for the coefficients of other regressors and would be used by ridge estimator, but not by the Graham & Powell (2012) estimator.

When $a_i$ varies with $i$ the $\hat{\theta}$ of equation (3.9) is an innovation of our paper. The presence of $\bar{a}$ in our estimator is essential to unbiasedness of $\hat{\theta}$ when the slope coefficients in $\beta_i$ do not vary with $i$.

### 3.4 Evaluating the Extent of Identification

Identification of the average effect $\theta_0 = E[a_i'\beta_i]$ depends on how informative each observation $i$ is for $a_i'\beta_i$. One way to measure this informativeness is to compare $a_i$ with a regularized version $\hat{a}_i$ that determines the conditional mean of the debiased ridge estimator. Taking the conditional expectation we have

$$E[\hat{\theta}|X_1, ..., X_n] = \frac{1}{n}\sum_{i=1}^{n} \hat{a}_i'\beta_i, \hat{a}_i = \bar{a}'(A\bar{W})^{-1}A_iW_i.$$

This $\hat{a}_i$ will tend to differ from $a_i$ depending on whether $Q_i$ is singular. The vector $\bar{a}'(A\bar{W})^{-1}$ is close to the first unit vector, and hence $\hat{a}_i$ is close to $a_i'W_i$, when the great majority of $Q_i$ are nonsingular and $\lambda$ is small. Also, as $\lambda$ shrinks to zero $a_i'W_i$ converges to $a_i$ if $Q_i$ is nonsingular but need not otherwise. Thus we can compare $\hat{a}_i$ and $a_i$ to evaluate identification. We focus on this comparison because $\hat{a}_i$ is closely related the contribution of the ith observation to $\hat{\theta}$.

One way to measure the difference between $a_i$ and $\hat{a}_i$ to use

$$\zeta_i = \frac{\|\hat{a}_i - a_i\|}{\sqrt{2\|\hat{a}_i\|^2 + 2\|a_i\|^2}}.$$

This $\zeta_i$ is constrained to be in the unit interval. We can use a quantile plot of this object to evaluate the extent of identification, with departures from zero indicating a lack of identification.

## 4 Nonparametric, Nonseperable Demand Models for Panel Data

The nonseparable, nonparametric model $E[S_{it}|X_i, \eta_{it}] = s(X_{it}, \eta_{it})$ (of equation (2.1)) provides a very general specification of individual demand. In the application we take the outcome variable $S_{it}$ to be expenditure share of a good, which has long been a useful specification, as in Deaton

& Muellbauer (1980b), Chaudhuri *et al.* (2006), and Hsiao (2021), but other choices of outcome variable will also do. Discrete choice is included as a special case where $S_{it}$ is the number of units of a particular good purchased by an individual in time period $1t$ and the outcome model is specified analogous to that in Section 3.1.

The model allows unobserved heterogeneity $\eta_{it}$ to affect demand in very general ways. The $\eta_{it}$ is allowed to be infinite dimensional corresponding to stochastic revealed preference as in McFadden & Richter (1990), McFadden (2005), and Kitamura & Stoye (2018) with demand restricted to be single valued. Such choice specifications have been considered by Lewbel (2001), Blomquist *et al.* (2014), Blundell *et al.* (2014), Hoderlein & Stoye (2014), Bhattacharya (2015), Dette *et al.* (2016), and Hausman & Newey (2016). In addition $\eta_{it}$ may include product specific unobserved characteristics as in Berry (1994) and Berry *et al.* (1995). The presence of such could create correlation across individuals in $\eta_{it}$. Alternatively, if $\eta_{it}$ is tastes by an individual for unobserved product characteristics and preferences are independent across individuals then correlation across individuals need not be present.

To help this model relate to existing demand models it is helpful to decompose the heterogeneity $\eta_{it}$ into a component $\alpha_i$ that does not vary with $t$ and a time varying component $v_{it}$. This decomposition is common in panel demand models, including discrete choice, as in Chamberlain (1984). Here $\alpha_i$ represents preference features that are stable over time for a given individual while $v_{it}$ allows some time variation in demand. For example, $v_{it}$ could represent a taste for variety that is not observable to the econometrician. Tastes for variety could also be incorporated by including functions of $t$ in $b(X_{it})$. Generally it is quite common to incorporate time varying heterogeneity as represented by $v_{it}$ in nonlinear panel data models.

An important feature of panel demand data is that prices are common across consumers and are determined in market equilibrium. As a result prices will generally be endogenous in being related to individual preferences. Restrictions on $v_{it}$ mitigate potential price endogeneity. If $v_{it}$ is i.i.d. over time and independent of unobserved supply (or markup) shocks and there are many consumers in the market then bias from price endogeneity will be small, as shown by Moon & Newey (2024). Intuitively, price effects will be (nearly) identified from movement of prices over time because supply shocks are independent of time variation in preferences. This independence seems plausible when $v_{it}$ is a stochastic taste for variety of an individual and variation in supply is due to cost shocks. Also, Hausman (1997) found that the use of prices from other markets as instruments did not change demand estimates for scanner data, providing evidence that relying on time variation in prices for identification of price effects is consistent with scanner data.

The interpretation of $v_{it}$ as preference heterogeneity means that preferences are allowed to change over time, even being correlated over time in order to represent a taste for variety. Demand specifications with time varying, unobserved preference effects $v_{it}$ are common in panel data, discrete choice demand being a prime example. The presence of $v_{it}$ helps demand and other models fit the data better. It allows for departures from the weak axiom of revealed preference in the choice of an individual over time, as has been found in empirical work, for example Crawford (2019).

Time varying preferences have little effect on the interpretation of welfare calculations. The average equivalent variation and deadweight loss calculations just average over time. As such they estimate the expected value of welfare integrated over as $v_{it}$, similar to welfare estimates for discrete choice panel data. Such time average welfare meausres are consistent with utility maximization over time if there are no dynamic linkages in goods. Of course such preferences are not consistent with stockpiling models like that of Hendel & Nevo (2006). In the application we take one month as the time unit and focus on goods with little potential for stockpiling to avoid this concern.

Allowing for zero demand is important in demand modeling. For example, consumer data which considers alcohol or tobacco consumption will have many individuals with zero consumption. Including zero demand observations in the data correctly accounts for zero consumption in calculations of average equivalent variation and deadweight loss, as shown in Hausman & Newey (2016), Theorem 3. Intuitively, there is no effect of a price change on the welfare of a consumer who never purchases

a product and the average is also correct when the product in only purchased sometimes. The nonseparable, nonparametric specification gives the demand equations flexibility to allow for zeros while being consistent with utility maximization.

An observation arising from economic theory is that often, but not always, the policy question of interest depends on only one, or a very few, price effects. For example, estimation of individual welfare effects typically depends only on the own price effect when all other prices are held constant, Hausman (1981). Also, small cross-price effects will mitigate market equilibrium effects of changing one price. Price changes for one good will shift demand for other goods by small amounts so that equilibrium welfare effect from changing only one price can be well approximated by the effect of just that price on average demand.

Computational simplicity is an important virtue of demand analysis in panel data based on linear in coefficients approximation to nonseparable, nonparametric demand. Average equivalent variation and deadweight loss are estimated by a debiased average of individual specific linear combinations of ridge regressions. Simulation is used to approximate the integrals in the welfare estimates. Simple inference is based on independence of estimates across individuals. All of these features make this approach to demand estimation simple to implement, even in very large data sets.

# 5 Application to Scanner Data

We apply our methods to estimate price elasticities for groceries and to analyze the impact of counterfactual tax changes on consumer welfare. In this context, the outcome variable $S_{it}$ is share of expenditure on a particular class of goods, and the regressors $X_{it}$ include the natural log of prices and total expenditure. Our specification generalizes the popular AIDs demand system of Deaton & Muellbauer (1980a) to approximate a nonparametric, fully nonseparable demand model as we describe in Section 4.

Given this specification, our debiasing method has important implications for our elasticity estimates and consequently, our estimates of counterfactual welfare. In the absence of debiasing, ridge regression tends to shrink parameters to zero. Therefore, in an AIDs-type specification, ridge would shrink the own price elasticity towards $-1$, the cross-price elasticities toward 0, and the expenditure elasticity towards 1. The debiasing mitigates the effects of shrinkage and results in estimates that are less sensitive to the choice of penalty parameter. However, we note that shrinkage of the cross-price elasticities may be appropriate in consumer demand panel datasets, where small cross-price effects often found in the literature Burda *et al.* (2008) and Burda *et al.* (2012). Indeed, cross-sectional OLS regressions in our empirical setting recover small cross-price elasticities, as we report in Table IV in Appendix A.

Rather than estimate demand for particular products, we instead focus on the demand for classes of goods. In effect, we model demand at an intermediate level of multi-stage budgeting to estimate welfare effects of price changes for good types. Here, the consumer decides how much to spend on a class of goods based on individual- and type-specific second-order flexible price indices, and on the total expenditure on all included classes of goods.

Modeling demand for good types can be justified by certain conditions on the separability of preferences, as in for example, Gorman (1959), Gorman (1981), Deaton & Muellbauer (1980a), and Blundell & Robin (2000). An alternative motivation relies on statistical aggregation for the many prices into a price index which is independent of consumer preferences, as in Hoderlein & Lewbel (2012). For the intermediate level of commodities we consider (e.g., soda) it may be important to allow for more general substitution patterns across the dissimilar kinds of goods. The flexibility in allowing for general cross-price effects provided by AIDs demand system of Deaton & Muellbauer (1980a), may be useful here as it is in Chaudhuri *et al.* (2006) and Hsiao (2021).

We use NielsenIQ retail scanner data to construct price indices, and the NielsenIQ Homescan

Panel to track purchases and household characteristics.[1]

The data include 2585 households with Houston-area ZIP codes in the years 2010-2014. The number of monthly observations for each household ranges from 1 to 60, and we restrict our analysis to the households included for at least 12 months.[2]

We construct the price indices for each consumer from data on the monthly total expenditures per good category, and on the quantity purchased per month. The original data had time-stamps for purchases. The price indices span 15 aggregated groups of goods: soda, milk, soup, water, butter, cookies, eggs, orange juice, ice cream, bread, chips, salad, yogurt, coffee, and cereal. As in Burda *et al.* (2008) and Burda *et al.* (2012) we chose these groups because they made up a relatively large proportion of total grocery expenditure. The data also includes demographics such as race, marital status, household size and composition, and employment status.

The price index for each group of goods is computed as a weighted geometric average of the actual purchase prices (expenditure divided by quantity) over all purchases made by the household in the month, with weights equal to the proportion of expenditure on a specific item associated with a unique item code. The price index $P_{g,it}$ for household $i$ at time $t$, for the $g^{th}$ group of goods is specified by

$$\ln(P_{g,it}) = \sum_{j=1}^{J_g} w_{gj,it} \ln(P_{gj,it}),$$

where $j$ denotes a particular item code, $J_g$ is the number of codes for the $g^{th}$ commodity, $w_{gj,it}$ is the proportion of expenditure on commodity $g$ that is spent on code $j$, and $P_{gj,it}$ is expenditure by household $i$ on code $j$ divided by quantity of code $j$ in month $t$. This is a Törnqvist price index which was shown by Diewert (1976) to be exact for a quadratic utility specification and a second order approximation to the exact price index for any utility. Deaton & Muellbauer (1980b) (pp. 132-133) showed that with weak separability this price index appears in share equations for a Rotterdam demand specification (i.e., log quantity as a linear function of log prices and log expenditure) and suggest that it could lead to a good approximation when prices within a group tend to move together.

The price indices may be endogenous because the amount spent on a particular item in a group of goods is a choice of the consumer. Price endogeneity could be particularly important when a group of goods contains commodities of varying quality, such as organic and non-organic milk, or fresh and frozen orange juice. As we discuss in the previous section, our approach can accommodate such endogeneity, provided that the unobserved heterogeneity satisfies the time-invariance condition formalized in Assumption 1.

As stated above, we construct price indices using prices actually paid by each household. Including zero expenditures makes it necessary to impute price indices for time periods where an individual purchased none of a particular good. If a household had purchased the good before, then price indices are imputed as the most recent price faced by the household in a past purchase. Rarely, a good is never purchased prior to a given month, in which case its imputed price is the average price of the same good within a subset of stores similar to those at which the household shops.[3] The frequency of household-month observations with zero total expenditures varies by good: for some goods, most households record purchases each month, while other goods, such as orange juice and ice cream, are

[1]The empirical work is researchers' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[2]We checked for differences in results between using all households and the 2197 that were present for at least a year and found no statistically significant differences. The insensitivity to panel length suggests that attrition bias does not play a large role in this data.

[3]Specifically, we group retailers in the Houston area into 4 categories, and assign households to their most-visited retailer category each year. Then we construct monthly price indices for each retailer category and each good, which are used to fill in missing prices.

purchased more infrequently. Our analysis focuses on estimating demand for the goods for which we have the most reliable data, namely, soda and milk.

The inclusion of prices for all 15 categories of goods allows estimation of cross-price demand effects. This gives us 16 price and expenditure regressors. This is too large a number of regressors for standard nonparametric estimation, such as kernel regression, where it is thought to be impractical to use more than five or six regressors. For panel estimation, 16 regressors may also be excessively large. The large number of regressors with small coefficients for the many cross-price effects motivates our use of ridge regularization.

In total, our analysis uses 86,122 observations across the households. As a baseline, we consider the log-linear AIDs-type specification below.

$$S_{it} = \alpha_i + \gamma_i \log \text{Exp}_{it} + \sum_g \beta_{g,i} \log P_{g,it} + u_{it} \tag{5.1}$$

where $S_{it}$ is the share of expenditue by household $i$ in month $t$ on a particular class of goods. $Exp_{it}$ is that household's total monthly expenditure over the 15 categories of goods, and $P_{g,it}$ is the household's price index for good $g$ in that month. $\alpha_i$, $\gamma_i$, and $\beta_{g,i}$ are individual-specific coefficients and $u_{it}$ a time-varying residual. We estimate separate models for soda and milk, with no restrictions that the coefficients in each case are the same.

In order to more precisely approximate a possibly non-linear and non-separable underlying demand model, in some of our analyses we enrich the specification (5.1) by including some powers and interactions of log prices and total expenditure.

Table I contains elasticity estimates for both soda and milk. We employ the model (5.1) and compare three methods for estimation. These are cross-sectional OLS, fixed-effects estimates, individual-specific ridge without debiasing, and estimates that employ our debiased individual-specific ridge method.

In order to perform individual-ridge, we must select the matrix $D_i$ in the formula (3.7). We let $D_i$ be the identity matrix with its first diagonal entry set to zero. We carry out ridge using two alternative choices for the penalty parameter $\lambda$. As a robustness check, we carry out the analysis with and without the inclusion of seasonal dummy variables. Seasonal variation in both price and tastes could be problematic for our analysis as it suggests that heterogeneity in preferences is time-varying given prices, which would contradict Assumption 1.

Table 1: Estimates of own-price elasticity, with (top) and without (bottom) season dummies.

| | OLS | FE | Ridge 0.05 | Ridge 0.0005 | DBR 0.05 | DBR 0.0005 |
|---|---|---|---|---|---|---|
| soda | -0.795 | -0.815 | -0.829 | -0.790 | -0.775 | -0.777 |
| | (0.003) | (0.004) | (0.007) | (0.016) | (0.009) | (0.016) |
| milk | -1.206 | -0.607 | -0.843 | -0.480 | -0.445 | -0.349 |
| | (0.012) | (0.016) | (0.011) | (0.038) | (0.046) | (0.037) |
| soda | -0.795 | -0.815 | -0.823 | -0.768 | -0.770 | -0.756 |
| | (0.003) | (0.003) | (0.007) | (0.017) | (0.009) | (0.017) |
| milk | -1.206 | -0.608 | -0.838 | -0.454 | -0.454 | -0.347 |
| | (0.012) | (0.016) | (0.008) | (0.041) | (0.041) | (0.041) |

The columns in Table I respectively contain estimates from cross-sectionl OLS, fixed-effects, individual-ridge without debiasing and penalty parameters 0.05 and 0.0005, and our debiased ridge estimates (abbreviated to 'DBR') with those same penalties. These methods are used to estimate the average of the coefficient on log own-price in specification (5.1). We then obtain elasticities by dividing the estimates by the average (over all individuals and time periods) of the expenditure share of the relevant good and subtracting unity. In order to account for dependence between the coefficient estimates and the mean expenditure share, standard errors are calculated by bootstrap.

In all cases, the estimates are insensitive to the inclusion of seasonal dummies. The ridge estimates without debiasing are sensitive to the choice of penalty parameter, particular in the case of milk. As we discuss above, in our specification, the shrinkage associated with ridge will tend to bias elasticity estimates towards $-1$. Indeed, when we do not debias, the elasticity estimates from ridge are closer to $-1$ when we employ a higher penalty than with a smaller penalty. this is particularly striking for milk. By contrast, when we debias, which mitigates the shrinkage associated with ridge, we obtain elasticity estimates that are much less sensitive to the choice of penalty.

The elasticity estimates from our debiased ridge method roughly align with those found in the previous literature (see for example, the meta-analysis of Andreyeva *et al.* (2010)).

We apply our methods to estimate bounds on the average equivalent variation consumer surplus and deadweight loss from a 10% increase in price for both soda and milk. This increase is relative to the actual price faced by each household in a particular period. The bounds follow the formulas in Hausman & Newey (2016) as detailed in Example 2 in Section 2. The formulas require we impose lower and upper bounds on the income effect. We take our lower bound to be 0 which corresponds to the assumption that soda and milk are normal goods. This lower bound on the income effects corresponds to an upper bound on the welfare loss. In order to obtain a lower bound on the welfare we require an upper bound on the income effect. For our upper bound, we use two times the fixed effects estimates of the derivative of quantity with respect to total expenditure (over the 15 classes of goods) at the mean quantity and total expenditure. Thus we obtain conservative upper bounds for the income effect of $\approx 3.082$ for soda and $\approx 6.241$ for milk.

In order to provide some distributional analysis, we estimate the welfare bounds separately for households in three different income groups. In particular, for those whose household income (averaged over all periods for which there is data on that household) is in the bottom quartile, top quartile, and for all households.

Tables II and III contain our estimation results for the welfare upper bounds. We report lower bounds in the appendix. We applied our analysis for both the log-linear specification (5.1) and a cubic specification which supplements the regressors in the linear model with all powers and interactions of the log own-price and total expenditure up to order three. We provide results for various choices of the penalty parameter $\lambda$. The welfare estimates have been annualized, that is, the numbers represent the welfare change over the course of a year.

Table 2: Soda Welfare Upper Bounds

| | Deadweight Loss (Linear) | | | Deadweight Loss (Cubic) | | |
|---|---|---|---|---|---|---|
| | Income Quartiles | | | Income Quartiles | | |
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 0.367 | 0.407 | 0.399 | 0.365 | 0.408 | 0.398 |
| | (0.028) | (0.033) | (0.013) | (0.030) | (0.035) | (0.014) |
| 0.0005 | 0.359 | 0.407 | 0.394 | 0.404 | 0.409 | 0.400 |
| | (0.029) | (0.041) | (0.015) | (0.042) | (0.041) | (0.020) |
| | Consumer Surplus (Linear) | | | Consumer Surplus (Cubic) | | |
| | Income Quartiles | | | Income Quartiles | | |
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 10.13 | 10.54 | 10.64 | 10.12 | 10.58 | 10.66 |
| | (0.682) | (0.707) | (0.281) | (0.680) | (0.707) | (0.280) |
| 0.0005 | 10.15 | 10.57 | 10.66 | 10.09 | 10.59 | 10.66 |
| | (0.683) | (0.707) | (0.281) | (0.679) | (0.709) | (0.282) |

The estimates of average deadweight loss and consumer surplus for the full set of households are remarkably stable, both between the linear and cubic specifications, and for different values of the penalty parameter. In part, this may reflect the tendency of debiasing to mitigate the shrinkage

Table 3: Milk Welfare Upper Bounds

| | Deadweight Loss (Linear) Income Quartiles | | | Deadweight Loss (Cubic) Income Quartiles | | |
|---|---|---|---|---|---|---|
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 0.178 | 0.148 | 0.158 | 0.142 | 0.140 | 0.121 |
| | (0.017) | (0.014) | (0.009) | (0.026) | (0.023) | (0.017) |
| 0.0005 | 0.120 | 0.108 | 0.120 | 0.190 | 0.172 | 0.136 |
| | (0.024) | (0.021) | (0.013) | (0.043) | (0.046) | (0.031) |
| | Consumer Surplus (Linear) Income Quartiles | | | Consumer Surplus (Cubic) Income Quartiles | | |
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 8.09 | 6.73 | 7.41 | 8.15 | 6.71 | 7.43 |
| | (0.490) | (0.390) | (0.170) | (0.495) | (0.394) | (0.171) |
| 0.0005 | 8.15 | 6.76 | 7.44 | 8.12 | 6.70 | 7.43 |
| | (0.495) | (0.394) | (0.171) | (0.494) | (0.396) | (0.173) |

induced by regularization, and thus to reduce sensitivity to the choice of penalty parameter $\lambda$.

We estimate that the deadweight loss from a price increase for soda is markedly higher than for milk. This is not surprising given that milk, unlike soda, is a staple food and so demand for this product may be relatively inelastic. Indeed, this aligns with our elasticity estimates in Table I.

Harding & Lovenheim (2017) analyze the role of prices in determining food purchases and nutrition and estimate the impact of taxes on nutrition and individual welfare. Allcott *et al.* (2019) and Dubois *et al.* (2020) have also considered the welfare effects of taxing soda. Like Dubois *et al.* (2020) our panel approach estimates individual-specific demands. Our approach is simpler in that it is based on continuous demand modeling and individual ridge regression with total expenditure included in the demand function. Also, our application averages over on-the-go and individual that purchase soda and those that don't. We obtain substantially larger estimates of average equivalent variation than their compensating varation which is to be expected because we model household demand and they model individual demand.

Figure 1 plots the quantiles of the scaled distance between the implied and true $a_i$ for our deadweight loss and consumer surplus upper bound estimates. In particular, we plot quantiles of $\zeta_i$ which has formula given in Section 3.4.

We see from the figures that for a large proportion of individuals the discrepency between the true and implied $a_i$s is relatively small. This is particularly clear in the case of our consumer surplus estimates, for which the quantiles are almost identically zero.

# 6 Theoretical Results

We now turn to a formal analysis of the properties of our estimation procedure. For this purpose, we will be explicit about allowing the number of periods for which we have observations to vary between individuals. In particular, we let $T_i$ denote the number of periods for which we observe data on individual $i$. In addition, we explicitly define the approximation error that results from the use of a linear sieve space. Recall that we employ an approximation $s(x,\eta) \approx b(x)'\beta(\eta)$. To explicitly define $\beta(\eta)$, we suppose that $\frac{1}{T_i}\sum_{t=1}^{T_i} E[b(X_{it})b(X_{it})']$ is non-singular. For each fixed value $\eta$ in the support of $\eta_{it}$, define the function $\beta(\eta)$ as follows.

$$\beta(\eta) = \left(E[\frac{1}{T_i}\sum_{t=1}^{T_i} b(X_{it})b(X_{it})']\right)^{-1} E[\frac{1}{T_i}\sum_{t=1}^{T_i} b(X_{it})s(X_{it},\eta)]$$

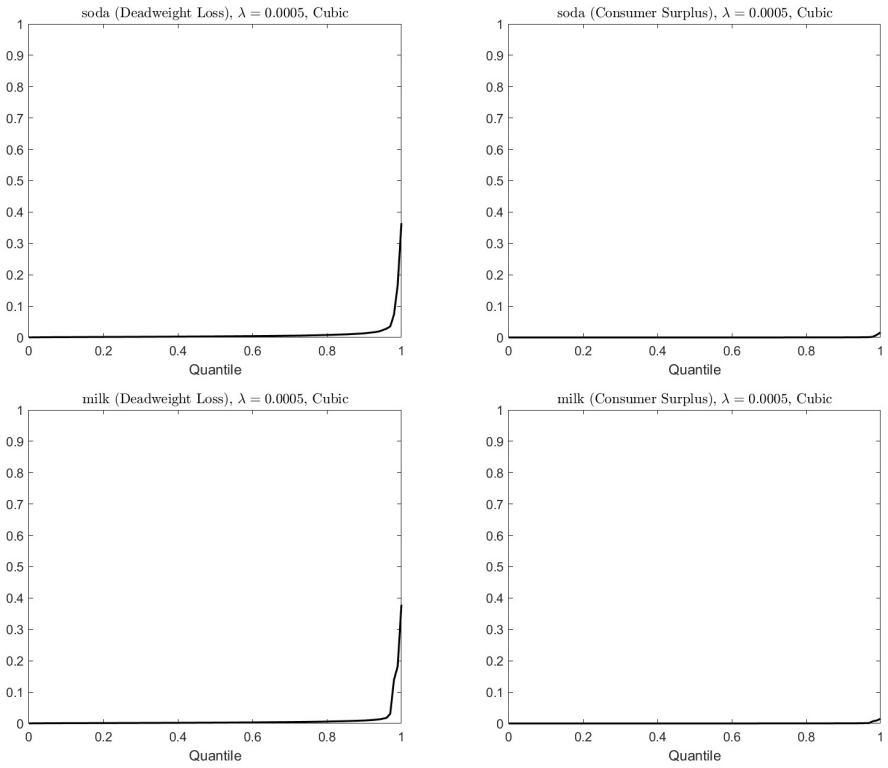


Figure 5.1: Scaled Distance Between Implied and True $a_i$

That is, $\beta(\eta)$ is the vector of coefficients from a best approximation of $s(\cdot, \eta)$ by a linear combination of the basis functions $b(\cdot)$. We define the approximation error $r(x, \eta)$ as below:

$$r(x, \eta) = s(x, \eta) - b(x)'\beta(\eta)$$

We then let $r_{it} = r(X_{it}, \eta_{it})$. Under Assumption 1, $E[\beta(\eta_{it})|X_i]$ does not depend on $t$, so we write $\beta_i = E[\beta(\eta_{it})|X_i]$. We also define a residual $u_{it}$ as follows:

$$u_{it} = S_{it} - E[S_{it}|X_{it}, \eta_{it}] + b(X_{it})'[\beta(\eta_{it}) - \beta_i]$$

Let $r_i$ and $u_i$ be the length-$T_i$ column vectors whose $t$-th entries are $r_{it}$ and $u_{it}$ respectively. Thus we obtain the following model:

$$S_i = B_i\beta_i + u_i + r_i,\ E[u_i|X_i] = 0 \tag{6.1}$$

Thus we obtain an approximate random correlated coefficients model for the outcome $S_{it}$ with the approximation error captured in $r_i$ and the individual-specific parameters closely related to our objects of interest.

It is helpful to introduce some notation. Let $D_{1,i}$ be equal to $D_i$ but with the first row and column removed (recall that by definition, the first row and column of $D_i$ contain only zeros). $J$ is the length of the vector $b(X_{it})$, and let $B_{1,it}$ denote $b(X_{it})$ with its first entry (the constant) removed. Finally, define $\bar{S}_i = \frac{1}{T_i}\sum_{t=1}^{T} S_{it}$, $\bar{B}_i = \frac{1}{T_i}\sum_{t=1}^{T} B_{1,it}$, and $\tilde{Q}_i = \frac{1}{T_i}\sum_{t=1}^{T} B_{1,it}B_{1,it}' - \bar{B}_i\bar{B}_i'$.

Recall that Section 3 defines $D_i$ to be diagonal with first diagonal entry zero and the others non-zero. In this section we allow for more general choices of $D_i$: we retain the condition that the first row and column of $D_i$ contain only zeroes, but unless we state otherwise $D_{1,i}$ can be any strictly positive-definite matrix. We implicitly assume throughout that the estimator (3.9) is well defined, that is, $\overline{AW}$ is non-singular.

## 6.1 Properties of the Estimator

Before we turn to the asymptotic behavior of our estimation procedures, we elaborate on a number of notable properties of the estimator outlined in Section 3. In particular, we consider the sense in which the estimator eliminates regularization bias, its limiting behavior under large and small values of the penalty parameter, and its motivation as an empirical Bayes estimator.

### Property A: No Regularization Bias Under Exogenous Effects

Our estimator is based on individual-level ridge regressions. Ridge estimates typically suffer from 'regularization' bias. The form of our estimator (3.9) is designed to mitigate, and in some cases entirely eliminate, regularization bias. In particular, in the case in which $\beta_i$ is constant, apart from the intercept and there is no approximation error.

For some insight into the bias properties of the estimator, it is helpful to compare our method with a plug-in estimator based on individual-specific OLS. An individual OLS estimate of $\beta_i$ is given below where $Q_i^\dagger$ is the pseudo-inverse of $Q_i$ and is well-defined even if $Q_i$ is singular:

$$\tilde{\beta}_i = Q_i^\dagger B_i' S_i / T_i$$

A plug-in OLS estimator of $\theta_0$ is then $\frac{1}{n}\sum_{i=1}^n a_i'\tilde{\beta}_i$. Suppose Assumptions 1 and 2 hold. If $Q_i$ is non-singular for all $i$, and the mean of $a_i'\tilde{\beta}_i$ is finite, then the plug-in OLS estimator is unbiased (up to approximation error) for $\theta_0$. This is because, in the absence of approximation error, $\tilde{\beta}_i$ is a conditionally (on $X_i$) unbiased estimate of $\beta_i$. Moreover, by Assumption 2, $\tilde{\beta}_i$ is independent of $a_i$ conditional on $X_i$. Thus if $E[|a_i'\tilde{\beta}_i|] < \infty$ then we can apply the law of iterated expectations and we see that

$$E[a_i'\tilde{\beta}_i] = E\big[a_i' E[\tilde{\beta}_i | X_i, a_i]\big] = E\big[a_i' E[\tilde{\beta}_i | X_i]\big] = E[a_i'\beta_i].$$

If individuals are drawn independently and identically from the population, then consistency of the plug-in OLS estimator follows by the law of large numbers. However, if $Q_i$ is singular with positive probability, this argument fails because an $\tilde{\beta}_i$ is (in general) biased when $Q_i$ is singular. Moreover, the assumption that $E[|a_i'\tilde{\beta}_i|]$ is finite is crucial. If this moment is infinite, then one cannot apply the law of iterated expectations, nor the law of large numbers. The first moment may be infinite even if $Q_i$ is non-singular almost surely. Graham & Powell (2012) acknowledge that the finite mean condition may fail, particularly if the number of regressors is close to the number of time periods. In the case of $a_i$ constant, this situation coincides with the case in which the information bound derived in Chamberlain (1992) is infinite, and thus regular estimation is impossible with the number of time periods fixed.

In contrast to OLS, individual-specific ridge estimates have finite expectation under weak conditions. Proposition 1, which provides conditions under which the expectation of $a_i'\hat{\beta}_i$ is finite, applies even if $Q_i$ is singular with positive probability.

**Proposition 1.** *Suppose $\lambda > 0$, $D_{i,1}$ has eigenvalues bounded below by $c > 0$, $\|B_i\|$ and $\|a_i\|$ are uniformly bounded, and $E[|S_{it}|]$ is finite. Then $E[|a_i'\hat{\beta}_i|] < \infty$.*

As we discuss in Section 3, individual ridge estimates are conditionally biased, even in the absence of approximation error. This in turn suggests that the sample average $\frac{1}{n}\sum_{i=1}^n a_i'\hat{\beta}_i$ is biased for $E[a_i'\beta_{it}]$. This motivates our debiasing strategy. Proposition 2 shows if effects are exogenous, then the debiased estimator is exactly unbiased up to approximation error. Note that the theorem holds for any $\lambda > 0$ and also applies when $Q_i$ is singular with positive probability. This contrasts with the average of plug-in OLS, which is in general conditionally biased when $Q_i$ is singular with positive probability. By 'effects are exogenous', we mean that $\beta_2(\eta_{it})$ is mean independent of $X_i$, where $\beta_2(\eta)$ is the subvector of $\beta(\eta)$ formed by removing its first component.

**Proposition 2.** *Suppose Assumptions 1 and 2 hold, $r_{it} = 0$ almost surely, and $\beta_2(\eta_{it})$ is mean independent of $X_i$. In addition, suppose $D_i$ is a function of $X_i$. If $\frac{1}{n}\sum_{i=1}^n A_i W_i$ is non-singular then*

$$E[\hat{\theta}|X_1, X_2, ..., X_n] = \frac{1}{n}\sum_{i=1}^n E[a_i'\beta_i|X_1, X_2, ..., X_n].$$

*If the above holds and $E[|\hat{\theta}|] < \infty$ then $E[\hat{\theta}] = E[a_i'\beta_i]$.*

Proposition 2 requires that $\beta_2(\eta_{it})$ is mean independent of $X_i$. The first component of $\beta(\eta_{it})$ is unrestricted. We do not need to restrict this component because we do not penalize the intercept in our individual-ridge regressions (this is why the first row and column of $D_i$ are composed of zeros).

In fact, if we strengthen the condition that effects are exogenous so that the entire vector $\beta(\eta_{it})$ is mean independent of $X_i$, then Proposition 2 applies for a general class of estimators. Consider that we can rewrite the estimator (3.9) as follows:

$$\hat{\theta} = \bar{a}'(\overline{AW})^{-1}\overline{AW\beta},\ \bar{a} = \frac{1}{n}\sum_{i=1}^n a_i,\ \overline{AW} = \frac{1}{n}\sum_{i=1}^n A_i W_i,\ \overline{AW\beta} = \frac{1}{n}\sum_{i=1}^n A_i W_i \tilde{\beta}_i \tag{6.2}$$

If we replace $W_i := (Q_i + \lambda D_i)^{-1} Q_i$ with some other conformable matrix that depends only on the regressors, then we obtain an alternative estimate of $\theta_0$. Consider the special case in which $a_i$ is constant and let $W_i$ be an indicator that the determinant of $Q_i$ exceeds a cut-off $h$, then the formula yields the estimator of Graham and Powell (absent adjustment for time-effects). If $\beta(\eta_{it})$ is mean independent of $X_i$, Proposition 2 applies for any estimator of the form above so long as $W_i Q_i^{\dagger} Q_i = W_i$, which holds both for our choice of $W_i$ as well as that of Graham and Powell.

**Property B: Convergence to Fixed Effects with Large Penalty**

As the penalty parameter grows to infinity, our estimator converges to a plug-in fixed effects or generalized fixed effects estimator. To state this formally, let us first note that the standard fixed effects estimate $\hat{\beta}_{FE,i}$ may be expressed as follows. The first component of this vector is an individual intercept given by $\bar{S}_i - \bar{B}_i'\hat{\beta}_{FE,1}$, where $\hat{\beta}_{FE,1}$ is a vector of shared slope parameters and constitute the remaining components of $\hat{\beta}_{FE,i}$. The slopes are given by

$$\hat{\beta}_{FE,1} = (\frac{1}{n}\sum_{i=1}^n \tilde{Q}_i)^{-1}\frac{1}{n}\sum_{i=1}^n \frac{1}{T_i}\sum_{t=1}^{T_i}(B_{1,i,t} - \bar{B}_i)S_{i,t}.$$

$\hat{\beta}_{FE,i}$ is a special case of a generalized fixed-effects estimator $\hat{\beta}_{GFE,i}$. Again, the first component of $\hat{\beta}_{GFE,i}$ is an individual intercept, in this case $\bar{S}_i - \bar{B}_i'\hat{\beta}_{GFE,1}$, where $\hat{\beta}_{GFE,1}$ is a vector of shared slopes. Let $G_i$ be a non-singular weighting matrix, then the corresponding vector of slope parameters $\hat{\beta}_{GFE,1}$ is defined as follows

$$\hat{\beta}_{GFE,1} = (\frac{1}{n}\sum_{i=1}^n G_i\tilde{Q}_i)^{-1}\frac{1}{n}\sum_{i=1}^n \frac{1}{T_i}\sum_{t=1}^{T_i} G_i(B_{1,i,t} - \bar{B}_i)S_{i,t}.$$

A plug-in fixed-effects estimate of $\theta_0$ is given by $\frac{1}{n}\sum_{i=1}^n a_i'\hat{\beta}_{FE,i}$ and plug-in generalized fixed-effects estimator by $\frac{1}{n}\sum_{i=1}^n a_i'\hat{\beta}_{GFE,i}$.

**Proposition 3.** $\lim_{\lambda\to\infty}\hat{\theta} = \frac{1}{n}\sum_{i=1}^n a_i'\hat{\beta}_{GFE,i}$ *where $G_i = D_{1,i}$. If $D_i$ does not vary with $i$, then* $\lim_{\lambda\to\infty}\hat{\theta} = \frac{1}{n}\sum_{i=1}^n a_i'\hat{\beta}_{FE,i}$.

The proposition states that as the penalty parameter grows towards infinity, the debiased panel ridge estimator converges to the plug-in generalized fixed-effects estimator with $G_i$ equal to $D_{1,i}$. In the special case in which $D_i$ does not depend on $i$, this is identical to the standard plug-in fixed effects estimator.

**Property C: Convergence Under Small Penalty**

Suppose that for each $i$, the matrix $Q_i$ is non-singular. Then as $\lambda$ goes to zero, the matrix $(Q_i + \lambda D_i)^{-1}$ converges to $Q_i^{-1}$, for every $i$. As such, our estimate converges to the plug-in average of individual OLS estimates $\frac{1}{n}\sum_{i=1}^n a_i'\tilde{\beta}_i$. Note the contrast with Property B. As $\lambda \to \infty$ our estimator converges to a plug-in average of (generalized) fixed effects estimates, with individual-specific intercepts and shared slope parameters. If $Q_i$ is non-singular for all $i$, then as $\lambda \to 0$, the estimator converges instead to the plug-in average of OLS estimates, which have both individual-specific intercepts and slopes. The choice of $\lambda$ thus allows us to smoothly transition between these two estimators.

When $Q_i$ is singular for some individuals, our estimator generally does not converge to plug-in individual OLS, but nonetheless it has an interpretable limit. Proposition 4 considers the special case in which one of the regressors, denoted by $Z_{it}$, is discretely distributed. Because $Z_i$ is discrete, it may be constant over time for some individual $i$, in which case $Q_i$ is singular.

**Proposition 4.** *Suppose $b(X_{it}) = (1, Z_{it}, B_{2,it}')'$, where $Z_{it}$ is a discrete scalar and the vector $B_{2,it}$ is continuously distributed. Let $C_i$ be equal to 1 if $Z_{i1} = Z_{i2} = ... = Z_{iT_i}$ and zero otherwise and let $\hat{p} = \frac{1}{n}\sum_{i=1}^n C_i$.*

*Suppose i. $D_i$ is diagonal, ii. if $C_i \neq 0$ then $Q_i$ is non-singular, and iii. the submatrix of $\tilde{Q}_i$ formed by removing its first row and column is non-singular. Define estimates*

$$\tilde{\beta}_{i,1}^* = \bar{S}_i - \bar{Z}_i\tilde{\beta}_{i,2}^* - \bar{B}_{2,i}'\tilde{\beta}_{i,3}$$
$$\tilde{\beta}_{i,2}^* = (1 - C_i)\tilde{\beta}_{i,2} + C_i\frac{1}{\hat{p}n}\sum_{i=1}^n(1 - C_i)\tilde{\beta}_{i,2},$$

*where $\tilde{\beta}_{i,2}$ and $\tilde{\beta}_{i,3}$ are respectively the individual OLS coefficients on $Z_{it}$ and $B_{2,it}$. Let $\tilde{\beta}_i^* = (\tilde{\beta}_{i,1}^*, \tilde{\beta}_{i,2}^*, \tilde{\beta}_{i,3}')'$. Then $\lim_{\lambda\to 0}\hat{\theta} = \frac{1}{n}\sum_{i=1}^n a_i'\tilde{\beta}_i^*$.*

The limit in Proposition 4 differs from plug-in individual OLS in that the OLS coefficient on $Z_{it}$ is replaced with the alternative estimate $\tilde{\beta}_{i,2}^*$ and the intercept is adjusted accordingly. If $Z_{it}$ varies for individual $i$ , then $\tilde{\beta}_{i,2}^*$ is equal to the individual OLS estimate $\tilde{\beta}_{i,2}$. However, if $Z_{it}$ does not vary, then $\tilde{\beta}_{i,2}^*$ is equal to $\frac{1}{\hat{p}n}\sum_{i=1}^n(1 - C_i)\tilde{\beta}_{i,2}$ which is the average of the OLS coefficients among the individuals for whom $Z_{it}$ does vary. In other words, for individuals without variation in $Z_{it}$ we impute the value of this coefficient as the average among individuals for whom $Z_{it}$ varies.

**Proposition 5.** *Define $\tilde{B}_{it} = D_{1,i}^{-1/2}(B_{1,it} - \bar{B}_i)$ and let $\tilde{B}_i = (\tilde{B}_{i1}, \tilde{B}_{i2}, ..., \tilde{B}_{iT_i})'$. Define the projection matrix $P_i = D_{1,i}^{-1/2}\tilde{B}_i^\dagger\tilde{B}_i D_{1,i}^{1/2}$ and the following vectors of coefficients*

$$\tilde{\beta}_{i,2}^\circ = D_{1,i}^{-1/2}\tilde{B}_i^\dagger S_i/T_i$$
$$\tilde{\beta}_{i,2}^* = \tilde{\beta}_{i,2}^\circ + (I - P_i)(\frac{1}{n}\sum_{i=1}^n P_i)^{-1}\frac{1}{n}\sum_{i=1}^n\tilde{\beta}_{i,2}^\circ.$$

*In addition, let $\tilde{\beta}_{i,1}^* = (\bar{S}_i - \bar{B}_{1,i}'\tilde{\beta}_{i,2}^*)$ and define $\tilde{\beta}_i^* = (\tilde{\beta}_{i,1}^*, \tilde{\beta}_{i,2}^{*\prime})'$, then $\lim_{\lambda\to 0}\hat{\theta} = \frac{1}{n}\sum_{i=1}^n a_i'\tilde{\beta}_i^*$.*

Proposition 5 considers the small $\lambda$ limit of our estimator in the general case. If $Q_i$ has full rank then $\tilde{\beta}_i^* = \tilde{\beta}_i$, the individual OLS estimate. To interpret $\tilde{\beta}_i^*$ when $Q_i$ is singular, first note that $P_i$ is the orthogonal (with respect to the inner-product $\langle a, b\rangle := a'D_i b$) projection onto the range of $\tilde{B}_i'$. Thus $I - P_i$ is the orthogonal projection onto the null space of $\tilde{B}_i$. If $Q_i$ is singular, this null space is non-trivial. This is problematic because, by construction, $(I - P_i)\tilde{\beta}_{i,2}^\circ = 0$. That is, the projection of the coefficients $\tilde{\beta}_{i,2}^\circ$ onto this subspace is zero and so, loosely speaking, this part

of $\tilde{\beta}^{\circ}_{i,2}$ is missing. The second term in the definition of $\tilde{\beta}^{*}_{i,2}$ adjusts for this by, in effect, replacing the missing part of $\tilde{\beta}^{\circ}_{i,2}$ with the projection of $(\frac{1}{n}\sum_{i=1}^{n} P_i)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\beta}^{\circ}_{i,2}$ onto this subspace. In the special case in Proposition 4, when $Q_i$ is singular the null space consists simply the vectors of the form $(0,1,0,0,...,0)'$, which correspond to the coefficients on $Z_{it}$.

**Property D: Empirical Bayes Interpretation**

The estimator $\hat{\theta}$ can be expressed as an empirical Bayes estimator. The empirical Bayes strategy imposes a priori that the individual coefficient vectors $\{\beta_i\}_{i=1}^{n}$ are concentrated around a prior mean which we estimate jointly with the individual-level coefficients.

To be more precise, our debiased ridge estimator is a Bayesian maximum a posteriori (MAP) estimate under a conditional Gaussian model for the outcome $S_{it}$ and a Gaussian prior for the individual regression coefficients $\beta_i$. It is well-known that standard ridge regression estimates can be expressed as MAP estimates in which the prior is Gaussian with mean zero. What distinguishes our approach is the manner in which the prior mean for $\beta_i$ is determined by the data. In particular, the prior mean $\bar{\beta}$, is pinned down by the restriction

$$\frac{1}{n}\sum_{i=1}^{n} A_i\beta_i^{\text{Post}} = \frac{1}{n}\sum_{i=1}^{n} A_i\bar{\beta}, \tag{6.3}$$

where $\{\beta_i^{\text{Post}}\}_{i=1}^{n}$ is the posterior mode for $\{\beta_i\}_{i=1}^{n}$. Thus the prior mean for the individual coefficients is fixed by imposing that the prior mode and posterior modes of $\frac{1}{n}\sum_{i=1}^{n} A_i\beta_i$ are identical. Loosely speaking, it ensures that observing the data does not lead us to update (i.e., improve) upon our prior for $\frac{1}{n}\sum_{i=1}^{n} A_i\beta_i$.

To be yet more precise, consider a Gaussian conditional likelihood for the outcomes $S_{it}|X_i \overset{iid}{\sim} N(b(X_{it})'\beta_i, \sigma)$ and prior for the individual slope parameters $\beta_i \overset{iid}{\sim} N(\bar{\beta}, \Sigma)$, where $\bar{\beta}$ is the prior mean and $\Sigma$ the prior variance-covariance matrix. Given this likelihood and prior, the posterior density $g$ satisfies the expression below:

$$ln\big(g(\beta_1,\beta_2,...,\beta_n)\big) \propto -\sum_{i=1}^{n}[\frac{1}{T_i}\|S_i - B_i\beta_i\|^2 + \sigma(\beta_i-\bar{\beta})'\Sigma^{-1}(\beta_i-\bar{\beta})] \tag{6.4}$$

The parameters $\{\beta_i\}_{i=1}^{n}$ that maximize the above (given a fixed $\bar{\beta}$) is the MAP estimate. In order to obtain an empirical Bayes estimate, we estimate the prior mean $\bar{\beta}$ from the data by imposing equation (6.3).

**Proposition 6.** *The estimator $\hat{\theta}$ in (3.9) can be written as $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} a_i'\beta_i^{Post}$ where $\{\beta_i^{Post}\}_{i=1}^{n}$ and $\bar{\beta}$ jointly solve the equations (6.3) and (6.5) below*

$$\{\beta_i^{Post}\}_{i=1}^{n} = \arg\max_{\{\beta_i\}_{i=1}^{n}} -\sum_{i=1}^{n}[\frac{1}{T_i}\|S_i - B_i\beta_i\|^2 + \lambda(\beta_i-\bar{\beta})'D_i(\beta_i-\bar{\beta})]. \tag{6.5}$$

The objective (6.5) is a monotone transformation of a Bayesian posterior (6.4) when $D_i = \frac{\sigma}{\lambda}\Sigma^{-1}$. Thus $\{\beta_i^{\text{Post}}\}_{i=1}^{n}$ are MAP estimates of the individual slopes. The prior is fixed by the second equation.[4] In the special case of $A_i$ the identity, solving the two equations above is equivalent to maximizing the objective in (6.5) jointly over both $\beta_i$ and $\bar{\beta}$.

[4]Note that we take $D_i$ to have first row and columns composed of zeroes, and so $D_i$ is singular. As such, strictly speaking we use flat 'improper' prior for the individual intercepts.

## 6.2 Consistency and Asymptotic Normality

Let us now consider the statistical properties of our estimator. We impose some additional conditions. In the assumptions below, inequalities involving random variables are understood to hold almost surely. Throughout we define $\delta_n := E[\|(Q_i + \lambda D_i)^{-1}\|]$.

**Assumption 3 (Consistency).** For some scalar $0 < c < \infty$, i. $\|A_i\|, \|a_i\|, \|B_i\| \leq c$, ii. $\|(\frac{1}{n}\sum_{i=1}^n A_i)^{-1}\| \leq c$, iii. $\|\beta_i\| \leq c$, iv. $\|E[u_i u_i'|X_i]\| \leq c$, v. $H_t^-(\cdot)$ and $H_t^+(\cdot)$ are uniformly bounded, vi. $\sup_{x,\eta}|r(x,\eta)| \leq \ell_n$ with $\ell_n \to 0$, vii. $T_i \geq T$, viii. The first row and column of $D_i$ contains only zeros and the eigenvalues of $D_{1,i}$ are bounded above by $c$ and below by $1/c$

**Assumption 4 (Asymptotic Normality).** For some finite constants $c, \xi, v, q > 0$ such that $(v-2)(q-2) > 4$, i. $\frac{1}{T}\sum_{t=1}^T E[u_{it}^v] \leq c$, ii. $1/c \leq Var(a_i'\beta_i)$, and iii. we have

$$n^{\frac{1}{v}+\frac{1}{q}-\frac{1}{2}} E[\|(Q_i + \lambda D_i)^{-1}\|^{q/2}]^{1/q} (\delta_n/T)^{\xi} = o(1).$$

**Assumption 5 (Remainders).** $\lambda\delta_n, \ell_n\sqrt{\delta_n}, \ell_n = o(\sqrt{\frac{1}{n}})$, $\frac{\delta_n}{T} = O(1)$, and $\frac{J\lambda^2\delta_n^3}{T} = o(1)$.

Assumption 3 i. and ii. impose conditions on $a_i$, $A_i$, and $B_i$ which are chosen directly by the researcher. 3.iii imposes that the individual-specific mean parameter $\beta_i$ is bounded in norm. iv. concerns the conditional second moments of $u_i$, $u_{it}$ may be dependent across time, but the dependence is restricted so that the norm of $E[u_i u_i'|X_i]$ does not grow with $T$. 3.v restricts $H_t^-(\cdot)$ and $H_t^+(\cdot)$. 3.vi is a condition on the sieve approximation error. Conditions of this form hold for many choices of sieve space used in practice under smoothness conditions on $s(\cdot,\eta)$, see e.g., DeVore & Lorentz (1993) for examples. 3.vii imposes that the number of time periods $T_i$, which can vary between individuals, is bounded below by some $T$. 3.viii is a weak condition on $D_i$ which is chosen by the researcher.

Assumption 4 ensures a normal limiting distribution via a Lyapunov condition. The assumption restricts the $v$-th moment of a random and the $q$-th moment of another. $q$ and $v$ must satisfy be strictly positive and $(v-2)(q-2) > 4$, which implies that $v, q > 2$. The conditions trade-off in the sense that, if $v$ is large, then $q$ need not be much larger than 2 and vice versa. 4.i bounds the average $v$-th moment of $u_{it}$. 4.ii states that the variance of the individual-specific approximation $a_i'\beta_i$ is bounded below. 4.iii requires that a particular sequence is $o(1)$. Each entry in the sequence is a product of three terms. The first is $n^{\frac{1}{v}+\frac{1}{q}-\frac{1}{2}}$. The conditions on $v$ and $q$ imply that this term goes to zero with $n$. The second term is the $q/2$-th moment of $\|(Q_i + \lambda D_i)^{-1}\|$ raised to the power $1/q$, the third is $\delta_n/T$ raised to the power $\xi$. Note that $\xi$ can be any strictly positive constant. As such, in the case of $\delta_n/T \to \infty$, 4.iii holds for a sufficiently large choice of $\xi$ so long as the $q$-th moment of $\|(Q_i + \lambda D_i)^{-1}\|$ grows polynomially with $T$.

Assumption 5 restricts the rates at which various sequences converge to zero. It ensures some terms in the asymptotic expansion of the estimation error are second order.

Theorem 1 provides general asymptotic theory for our estimator (3.9). It applies for both continuous and discrete regressors. The result follows from the more general result in Lemma 1 in the appendix which applies to any estimator of the form (6.2) such that $W_i Q_i^\dagger Q_i = W_i$.

**Theorem 1 ( Asymptotics).** *Suppose Assumptions 1, 2, and 3 hold and $\lambda\delta_n \to 0$.*
***a. (Consistency)***

$$\hat{\theta} - \theta_0 = O_p\left(\sqrt{\frac{1}{n}} + \sqrt{\frac{\delta_n}{nT}} + \lambda\delta_n + \lambda\delta_n\sqrt{\frac{J\delta_n}{nT}} + \ell_n(1+\sqrt{\delta_n})\right)$$

***b. (Asymptotic Normality)***

*In addition, if Assumptions 4 and 5 hold, then* $\hat{\theta}-\theta_0 = O_p(n^{-1/2})$ *and* $\sqrt{n}\sigma_n^{-1}(\hat{\theta}-\theta_0) \sim^a N(0,1)$*, where* $\sigma_n^2$ *is given by:*

$$\sigma_n^2 = E[|\frac{1}{T_i}a_i' W_i Q_i^\dagger B_i' u_i|^2] + Var(a_i'\beta_i) \tag{6.6}$$

Theorem 1 shows the crucial role of $\delta_n$ in the asymptotic behavior of the estimator. To understand what $\delta_n$ represents, suppose for simplicity that $D_i$ is the identity matrix. Then $\|(Q_i+\lambda D_i)^{-1}\|$ is equal to $(\mu_{\min}(Q_i)+\lambda)^{-1}$, where $\mu_{\min}(Q_i)$ is the smallest eigenvalue of $Q_i$. As such, if $Q_i$ is close to singular, then $\|(Q_i+\lambda D_i)^{-1}\|$ is close to $1/\lambda$. In the extreme case, if $Q_i$ is singular with probability $p$, then $p/\lambda \leq \delta_n$. Thus the condition $\lambda\delta_n \to 0$ is only possible if $p$ shrinks to zero with the sample size, which generally requires that $T$ grows with $n$. On the other hand, if $E[1/\mu_{\min}(Q_i)]$ is bounded by a finite constant, then $\delta_n \leq E[1/\mu_{\min}(Q_i)]$, uniformly over $\lambda$. Thus $\lambda\delta_n = o(n^{-1/2})$ so long as $\lambda$ shrinks sufficiently quickly to zero.

To examine this in more detail, we consider two extreme cases below. In the first case, captured in Corollary 1, we suppose that $E[\mu_{\min}(\tilde{Q}_i)^{-1}]$ is bounded above, where $\mu_{\min}(\tilde{Q}_i)$ is the smallest eigenvalue of $\tilde{Q}_i$. This is only possible if all regressors are continuously distributed and $T > J$. In the absence of approximation error (so that $\ell_n = 0$ with $J < T$ fixed), root-$n$ consistency and asymptotic normality do not require that $T$ grows with the sample size. Indeed, if $E[\mu_{\min}(\tilde{Q}_i)^{-1}]$ is finite then the efficiency bound in Chamberlain (1992) is finite and root-$n$ regular estimation is possible.

**Corollary 1 (Continuous Case).** *Let* $\|Q_i\| \leq c$*. Suppose Assumptions 1, 2, and 3 hold. If* $E[\mu_{\min}(\tilde{Q}_i)^{-1}]$ *is bounded above and* $\lambda \to 0$*, then:*

$$\hat{\theta}-\theta_0 = O_p\Big(\sqrt{\frac{1}{n}} + \lambda + \lambda\sqrt{\frac{J}{nT}} + \ell_n\Big)$$

*If in addition Assumption 4 holds,* $\ell_n, \lambda = o(\sqrt{\frac{1}{n}})$ *and* $\frac{J\lambda^2}{T} = o(1)$*, then* $\hat{\theta}-\theta_0 = O_p(n^{-1/2})$ *and* $\sqrt{n}\sigma_n^{-1}(\hat{\theta}-\theta_0) \sim^a N(0,1)$*, where* $\sigma_n^2$ *is given by (6.6).*

At the other extreme, Corollary 2 applies for the case of a single binary regressor. We assume that for a given individual $i$, the probability $X_{it} = 1$ is given by $\pi_i \in (0,1)$ and that, conditional on $\pi_i$, the regressor is independent over time. In this case $\tilde{Q}_i$ must be singular with positive probability because for an individual $i$, the regressors is constant with positive probability. However, as $T_i$ grows, this probability shrinks to zero at a rate that depends on the distribution of $\pi_i$.

**Corollary 2 (Binary Case).** *Suppose Assumptions 1, 2, and 3 hold, and* $b(X_{it}) = (1, X_{it})'$ *where* $X_{it}$ *is binary. Suppose* $P(X_{it} = 1|\pi_i) = \pi_i$ *and the entries of the sequence* $\{X_{it}\}_{t=1}^T$ *are jointly independent conditional on* $\pi_i$*. Let* $\pi_i$ *admit a probability density* $f_\pi$ *so that* $f_\pi(\pi) \leq C(1-\pi)^\omega \pi^\omega$ *where* $\omega > 0$*. Then if* $\lambda \to 0$ *and* $T \to \infty$ *we have:*

$$\hat{\theta}-\theta_0 = O_p\Big(\lambda + T^{-(1+\omega)} + \sqrt{\frac{1}{n}} + \sqrt{\frac{1}{nT}} + \sqrt{\frac{T^{-(2+\omega)}}{\lambda n}}\Big)$$

*In addition, if* $\frac{T^{-(1+\omega)}}{\lambda} = O(1)$*,* $T^{-(1+\omega)}, \lambda = o(\sqrt{1/n})$*, and Assumptions 4.i and 4.ii hold with* $q/2 < \omega$*, then* $\hat{\theta}-\theta_0 = O_p(n^{-1/2})$ *and* $\sqrt{n}\sigma_n^{-1}(\hat{\theta}-\theta_0) \sim^a N(0,1)$*, where* $\sigma_n^2$ *is given by (6.6).*

Corollary 2 establishes root-$n$ consistency only under the condition that $T^{-(1+\omega)} = o(\sqrt{1/n})$. Thus the rate at which $T$ must grow with $n$ depends on the rate at which $f_\pi(\pi)$ goes to zero as $\pi$ goes to zero or one. If $f_\pi(\pi)$ goes quickly to zero, then the probability $X_{it}$ is constant over time goes to zero quickly as $T$ grows, and so $T$ need not increase rapidly with $n$. This phenomenon is

considered in Chernozhukov *et al.* (2013) and tied to the rate at which the identified set shrinks with $T$.

Results for other cases, for example with both discrete and continuous regressors, may also be obtained from Theorem 1. As in the proofs of Corollaries 1 and 2, it would suffice to derive a convergence rate for $\delta_n$ under suitable assumptions.

# References


Abrevaya, Jason. 2000. Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics*, **95**, 1–23.

Allcott, Hunt, Lockwood, Benjamin B., & Taubinsky, Dmitry. 2019. Should We Tax Sugar-Sweetened Beverages? An Overview of Theory and Evidence. *Journal of Economic Perspectives*, **33**, 202–227.

Altonji, Joseph G., & Matzkin, Rosa L. 2005. Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors. *Econometrica*, **73**, 1053–1102.

Andreyeva, Tatiana, Long, Michael W., & Brownell, Kelly D. 2010. The Impact of Food Prices on Consumption: A Systematic Review of Research on the Price Elasticity of Demand for Food. *American Journal of Public Health*, **100**, 216–222.

Arellano, Manuel, & Bonhomme, Stéphane. 2012. Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies*, **79**, 987–1020.

Berry, Steven, Levinsohn, James, & Pakes, Ariel. 1995. Automobile Prices in Market Equilibrium. *Econometrica*, **63**, 841.

Berry, Steven T. 1994. Estimating Discrete-Choice Models of Product Differentiation. *RAND Journal of Economics*, **25**, 242.

Bhattacharya, Debopam. 2015. Nonparametric Welfare Analysis for Discrete Choice. *Econometrica*, **83**, 617–649.

Blomquist, Soren, Kumar, Anil, Liang, Che-Yuan, & Newey, Whitney K. 2014. Individual Heterogeneity, Nonlinear Budget Sets, and Taxable Income. *CEMMAP working paper 21/14*.

Blomquist, Soren, Kumar, Anil, Liang, Che-Yuan, & Newey, Whitney K. 2024. Nonlinear budget set regressions in random utility models: Theory and application to taxable income. *Journal of Econometrics*, 105859.

Blundell, Richard, & Robin, Jean-Marc. 2000. Latent Separability: Grouping Goods without Weak Separability. *Econometrica*, **68**, 53–84.

Blundell, Richard, Kristensen, Dennis, & Matzkin, Rosa. 2014. Bounding quantile demand functions using revealed preference inequalities. *Journal of Econometrics*, **179**, 112–127.

Burda, Martin, Harding, Matthew, & Hausman, Jerry. 2008. A Bayesian mixed logit–probit model for multinomial choice. *Journal of Econometrics*, **147**, 232–246.

Burda, Martin, Harding, Matthew, & Hausman, Jerry A. 2012. A poisson mixture model of discrete choice. *Journal of Econometrics*, **166**, 184–203.

Chamberlain, Gary. 1982. Multivariate regression models for panel data. *Journal of Econometrics*, **18**, 5–46.

Chamberlain, Gary. 1984. *Handbook of Econometrics, Volume 2*. North-Holland. Chap. Chapter 22 Panel data, pages 1247–1318.

Chamberlain, Gary. 1992. Efficiency Bounds for Semiparametric Regression. *Econometrica*, **60**, 567–596.

Chaudhuri, Shubham, Goldberg, Pinelopi K., & Jia, Panle. 2006. Estimating the Effects of Global Patent Protection in Pharmaceuticals: A Case Study of Quinolones in India. *American Economic Review*, **96**, 1477–1514.

Chernozhukov, Victor, Fernández-Val, Iván, Hahn, Jinyong, & Newey, Whitney. 2013. Average and quantile effects in nonseparable panel models. *Econometrica*, **81**(2), 535–580.

Crawford, Ian. 2019. *Nonparametric Analysis of Labour Supply Using Random Fields*.

Deaton, Angus, & Muellbauer, John. 1980a. An almost ideal demand system. *American Economic Review*, **3**, 312–326.

Deaton, Angus, & Muellbauer, John. 1980b. *Economics and Consumer Behavior*. Cambridge:Cambridge University Press.

Dette, Holger, Hoderlein, Stefan, & Neumeyer, Natalie. 2016. Testing multivariate economic restrictions using quantiles: The example of Slutsky negative semidefiniteness. *Journal of Econometrics*, **191**, 129–144.

DeVore, Ronald A., & Lorentz, George G. 1993. *Constructive Approximation*. Springer Berlin, Heidelberg.

Diewert, W. E. 1976. Exact and superlative index numbers. *Journal of Econometrics*, **4**, 115–145.

Dubois, Pierre, Griffith, Rachel, & O'Connell, Martin. 2020. How Well Targeted Are Soda Taxes? *American Economic Review*, **110**, 3661–3704.

Gorman, W. M. 1959. Separable Utility and Aggregation. *Econometrica*, **27**, 469.

Gorman, W. M. 1981. *Essays in the Theory and Measurementof Consumer Behaviour in Honor of Sir Richard Stone*. Cambridge:Cambridge University Press. Chap. Some Engel curves, pages 7–30.

Graham, Bryan S., & Powell, James L. 2012. Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica*, **80**(5), 2105–2152.

Harding, Matthew, & Lovenheim, Michael. 2017. The effect of prices on nutrition: Comparing the impact of product- and nutrient-specific taxes. *Journal of Health Economics*, **53**, 53–71.

Hausman, Jerry. 1997. *The Economics of New Goods*. University of Chicago Press. Chap. Valuation of New Goods under Perfect and Imperfect Competition, pages 209–237.

Hausman, Jerry A. 1981. Exact consumer's surplus and deadweight loss. *American Economic Review*, **4**, 662–676.

Hausman, Jerry A., & Newey, Whitney K. 2016. Individual Heterogeneity and Average Welfare. *Econometrica*, **84**, 1225–1248.

Hendel, Igal, & Nevo, Aviv. 2006. Measuring the Implications of Sales and Consumer Inventory Behavior. *Econometrica*, **74**, 1637–1673.

Hoderlein, Stefan, & Lewbel, Arthur. 2012. REGRESSOR DIMENSION REDUCTION WITH ECONOMIC CONSTRAINTS: THE EXAMPLE OF DEMAND SYSTEMS WITH MANY GOODS. *Econometric Theory*, **28**, 1087–1120.

Hoderlein, Stefan, & Stoye, Jörg. 2014. Revealed Preferences in a Heterogeneous Population. *Review of Economics and Statistics*, **96**, 197–213.

Hoderlein, Stefan, & White, Halbert. 2012. Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, **168**, 300–314.

Honore, Bo E. 1992. Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects. *Econometrica*, **60**, 533–565.

Hsiao, Allan. 2021. *Coordination and Commitment in International Climate Action: Evidence from Palm Oil*.

Imbens, Guido, & Newey, Whitney. 2009. Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, **77**, 1481–1512.

Kitamura, Yuichi, & Stoye, Jörg. 2018. Nonparametric Analysis of Random Utility Models. *Econometrica*, **86**, 1883–1909.

Lewbel, Arthur. 2001. Demand Systems With and Without Errors. *American Economic Review*, **91**, 611–618.

Manski, Charles F. 1987. Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data. *Econometrica*, **55**, 357–362.

McFadden, Daniel, & Richter, Marcel K. 1990. *Preferences, Uncertainty, and Optimality, Essays in Honor of Leo Hurwicz*. Westview Press, Boulder, CO. Chap. Stochastic rationality and revealed stochastic preference, pages 161–186.

McFadden, Daniel L. 2005. Revealed stochastic preference: a synthesis. *Economic Theory*, **26**, 245–264.

Moon, Sarah, & Newey, Whitney. 2024. *Demand in Markets with Many Consumers*.

Pakes, Ariel, & Porter, Jack. 2024. Moment inequalities for multinomial choice with fixed effects. *Quantitative Economics*, **15**, 1–25.

Pesaran, M. Hashem, & Smith, Ron. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, **68**(1), 79–113.

Semenova, Vira, Goldman, Matt, Chernozhukov, Victor, & Taddy, Matt. 2023. Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, **14**, 471–510.

Shi, Xiaoxia, Shum, Matthew, & Song, Wei. 2018. Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity. *Econometrica*, **86**, 737–761.

Torgovitsky, Alexander. 2019. Nonparametric Inference on State Dependence in Unemployment. *Econometrica*, **87**, 1475–1505.

Wooldridge, Jeffrey M. 2005. Instrumental variables estimation with panel data. *Econometric Theory*, **21**(4), 865–869.

# A Additional Empirical Results

Table IV contains cross-sectional OLS expenditure, own-price, and cross-price elasticity estimates for both soda and milk. The figures are estimated using the baseline model in Section 5 without seasonal dummies. Standard errors are to the right of the coefficient and elasticity estimates. For soda the cross-price elasticity with greatest magnitude is for butter, at $-0.2380$ and most of the estimates have magnitude below 0.1. For milk, all cross-price elasticities have magnitude less than 0.2 and many have magnitude below 0.01. Tables V and VI contain lower bounds on counterfactual

Table 4: OLS cross-price elasticities for soda and milk, 2010-2014

| | Soda | | | | Milk | | | |
|---|---|---|---|---|---|---|---|---|
| | Coeff | s.e | Elast | s.e. | Coeff | s.e. | Elast | s.e. |
| exp | 0.0208 | 0.0010 | 1.1452 | 0.0067 | -0.0044 | 0.0008 | 0.9608 | 0.0073 |
| soda | 0.0295 | 0.0005 | -0.7945 | 0.0033 | -0.0024 | 0.0004 | -0.0216 | 0.0034 |
| soup | -0.0030 | 0.0012 | -0.0212 | 0.0084 | -0.0008 | 0.0009 | -0.0071 | 0.0078 |
| water | 0.0007 | 0.0005 | 0.0051 | 0.0034 | 0.0002 | 0.0003 | 0.0022 | 0.0030 |
| butter | -0.0340 | 0.0011 | -0.2370 | 0.0079 | 0.0003 | 0.0008 | 0.0024 | 0.0069 |
| cookies | -0.0031 | 0.0007 | -0.0215 | 0.0051 | -0.0042 | 0.0005 | -0.0373 | 0.0046 |
| eggs | -0.0123 | 0.0019 | -0.0858 | 0.0130 | 0.0003 | 0.0013 | 0.0030 | 0.0120 |
| oj | 0.0113 | 0.0012 | 0.0789 | 0.0083 | -0.0006 | 0.0008 | -0.0054 | 0.0071 |
| ice cream | 0.0113 | 0.0015 | 0.0788 | 0.0104 | -0.0184 | 0.0011 | -0.1630 | 0.0093 |
| bread | -0.0274 | 0.0013 | -0.1911 | 0.0093 | -0.0104 | 0.0009 | -0.0927 | 0.0081 |
| chips | 0.0050 | 0.0014 | 0.0350 | 0.0100 | -0.0047 | 0.0011 | -0.0417 | 0.0095 |
| milk | 0.0136 | 0.0017 | 0.0947 | 0.0119 | -0.0232 | 0.0013 | -1.2058 | 0.0117 |
| salad | -0.0151 | 0.0010 | -0.1049 | 0.0069 | 0.0062 | 0.0007 | 0.0553 | 0.0063 |
| yogurt | -0.0051 | 0.0007 | -0.0354 | 0.0046 | 0.0018 | 0.0005 | 0.0160 | 0.0043 |
| coffee | -0.0004 | 0.0009 | -0.0031 | 0.0060 | 0.0013 | 0.0006 | 0.0112 | 0.0053 |
| cereal | 0.0067 | 0.0015 | 0.0466 | 0.0108 | -0.0107 | 0.0011 | -0.0953 | 0.0097 |

welfare for both soda and milk. These lower are calculated in the same fashion as the upper bounds in Tables II and III albeit using a conservative upper bound on the income effect.

Table 5: Soda Welfare Lower Bounds

| | Deadweight Loss (Linear) Income Quartiles | | | Deadweight Loss (Cubic) Income Quartiles | | |
|---|---|---|---|---|---|---|
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | -0.021 | 0.0945 | 0.0347 | -0.0245 | 0.0939 | 0.0317 |
| | (0.0187) | (0.0221) | (0.0102) | (0.0222) | (0.024) | (0.0121) |
| 0.0005 | -0.0320 | 0.0917 | 0.0281 | 0.0164 | 0.0948 | 0.0330 |
| | (0.0236) | (0.0321) | (0.0126) | (0.0357) | (0.0343) | (0.0192) |
| | Consumer Surplus (Linear) Income Quartiles | | | Consumer Surplus (Cubic) Income Quartiles | | |
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 9.746 | 10.23 | 10.27 | 9.729 | 10.26 | 10.29 |
| | (0.655) | (0.686) | (0.271) | (0.653) | (0.686) | (0.27) |
| 0.0005 | 9.755 | 10.26 | 10.29 | 9.698 | 10.27 | 10.29 |
| | (0.656) | (0.686) | (0.271) | (0.652) | (0.688) | (0.272) |

Table 6: Milk Welfare Lower Bounds

| | Deadweight Loss (Linear) Income Quartiles | | | Deadweight Loss (Cubic) Income Quartiles | | |
|---|---|---|---|---|---|---|
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 0.0918 | 0.0864 | 0.0852 | 0.0559 | 0.0781 | 0.0479 |
| | (0.0127) | (0.0117) | (0.0080) | (0.0243) | (0.0229) | (0.0171) |
| 0.0005 | 0.0335 | 0.0457 | 0.0465 | 0.1046 | 0.1107 | 0.0626 |
| | (0.0228) | (0.0205) | (0.0133) | (0.0427) | (0.0457) | (0.0313) |
| | Consumer Surplus (Linear) Income Quartiles | | | Consumer Surplus (Cubic) Income Quartiles | | |
| $\lambda$ | Upper | Lower | All | Upper | Lower | All |
| 0.05 | 8.00 | 6.67 | 7.33 | 8.07 | 6.65 | 7.36 |
| | (0.485) | (0.387) | (0.169) | (0.489) | (0.390) | (0.170) |
| 0.0005 | 8.07 | 6.70 | 7.37 | 8.03 | 6.63 | 7.35 |
| | (0.490) | (0.391) | (0.170) | (0.489) | (0.393) | (0.171) |

# B Proofs and Supporting Lemmas

*Proof of Proposition 1.* Recall the formula for $\hat{\beta}_i$ is $\hat{\beta}_i = (Q_i + \lambda D_i)^{-1} B_i' S_i / T_i$. By the properties of the matrix norm and Holder's inquality:

$$|a_i' \hat{\beta}_i| \leq \|a_i\| \|(Q_i + \lambda D_i)^{-1}\| \|B_i\| \|S_i\| / T_i$$

We can decompose $(Q_i + \lambda D_i)^{-1}$ into the product of block matrices as follows:

$$(Q_i + \lambda D_i)^{-1} = \begin{pmatrix} 1 & -\bar{B}_i' \\ 0 & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\bar{B}_i & I \end{pmatrix}$$

By the properties of the Euclidean matrix norm we then have:

$$\begin{aligned} \|(Q_i + \lambda D_i)^{-1}\| &\leq (1 + \|\bar{B}_i\|)^2 (1 + \|(\tilde{Q}_i + \lambda D_{1,i})^{-1}\|) \\ &\leq (1 + \|B_i\|)^2 (1 + \|(\tilde{Q}_i + \lambda D_{1,i})^{-1}\|) \end{aligned}$$

. Now, note that by the properties of the matrix norm

$$\begin{aligned} \|(\tilde{Q}_i + \lambda D_{1,i})^{-1}\| &= \|D_{1,i}^{-1/2} (D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} + \lambda I)^{-1} D_{1,i}^{-1/2}\| \\ &\leq \|D_{1,i}^{-1/2}\|^2 \|(D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} + \lambda I)^{-1}\| \\ &\leq \frac{1}{\lambda} \|D_{1,i}^{-1/2}\|^2. \end{aligned}$$

Combining everything so far, we get:

$$|a_i' \hat{\beta}_i| \leq \|a_i\| (1 + \|B_i\|)^3 (1 + \frac{1}{\lambda} \|D_{1,i}^{-1/2}\|^2) \|S_i\| / T_i,$$

and so, since $\|a_i\|$, $\|B_i\|$, and $\|D_{1,i}^{-1/2}\|$ are bounded almost surely and $\lambda > 0$, there is a constant $C$ so that $|a_i' \hat{\beta}_i| \leq C \|S_i\| / T_i \leq C \frac{1}{T_i} \sum_{t=1}^{T_i} |S_{it}|$. It follows that $E[|a_i' \hat{\beta}_i|] \leq C \frac{1}{T_i} \sum_{t=1}^{T_i} E[|S_{it}|]$, which is finite because $E[|S_{it}|]$ is finite by supposition. □

*Proof of Proposition 2.* Recall that under Assumption 1, defining $\beta_i := E[\beta(\eta_{it}) | X_i]$ we have

$$S_i = B_i \beta_i + u_i + r_i,\ E[u_i | X_i] = 0. \tag{B.1}$$

Write $\beta_i = (\beta_{1,i}, \beta_{2,i}')'$, where $\beta_{1,i}$ is the scalar first component of the vector $\beta_i$. If $\beta_2(\eta_{it})$ is mean independent of $X_i$ then $\beta_{2,i}$ does not vary with $i$, and so we can write $\beta_{2,i} = \beta_2$.

Now, using the general formula for $\hat{\theta}$, and (B.1) with $r_i = 0$, we have:

$$\begin{aligned}
E[\hat{\theta}|X_1, X_2, ..., X_n] &= E\Big[\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger B_i' S_i/T_i \Big| X_1, ..., X_n\Big] \\
&= E\Big[\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i Q_i^\dagger \beta_i \Big| X_1, ..., X_n\Big] \\
&+ E\Big[\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i B_i' u_i \Big| X_1, ..., X_n\Big]
\end{aligned}$$

For the first term on the RHS of the final equality, note that $W_i = W_i Q_i Q_i^\dagger$ and so this term simplifies:

$$\begin{aligned}
&E\Big[\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i Q_i^\dagger \beta_i \Big| X_1, ..., X_n\Big] \\
=&E\Big[\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i \beta_i \Big| X_1, ..., X_n\Big] \qquad \text{(B.2)}
\end{aligned}$$

Now, we will show that for our particular choice of $W_i$, if $\beta_{2,i}$ is constant, then

$$\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i \beta_i = \frac{1}{n}\sum_{i=1}^n a_i'\beta_i.$$

For more general choices of $W_i$ the above clearly continues to hold if the entire $\beta_i$ is constant.

To see this, let $a_{0,i}$ denote the first component of $a_i$ and $a_{1,i}$ the vector that contains the remaining components. Note that $W_i = \begin{pmatrix} 1 & \bar{B}_i'(I - (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i) \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i \end{pmatrix}$ and so writing things in terms of block matices and multiplying out the product we see that

$$\begin{aligned}
A_i W_i \beta_i &= \begin{pmatrix} a_{0,i} & a_{1,i}' \\ 0 & I \end{pmatrix}\begin{pmatrix} 1 & \bar{B}_i'(I - (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i) \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i \end{pmatrix}\begin{pmatrix} \beta_{i,1} \\ \beta_{2,i} \end{pmatrix} \\
&= \begin{pmatrix} a_{0,i}\beta_{i,1} + \big(a_{0,i}\bar{B}_i' + (a_{1,i}' - a_{0,i}\bar{B}_i')(\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i\big)\beta_{2,i} \\ (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i\beta_{2,i} \end{pmatrix}.
\end{aligned}$$

Define $\beta_1^* := \frac{\frac{1}{n}\sum_{i=1}^n a_{0,i}\beta_{i,1}}{\frac{1}{n}\sum_{i=1}^n a_{0,i}}$. From the above we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n A_i W_i \beta_i \\
=&\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n a_{0,i}\beta_1^* + \frac{1}{n}\sum_{i=1}^n \big(a_{0,i}\bar{B}_i' + (a_{1,i}' - a_{0,i}\bar{B}_i')(\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i\big)\beta_{2,i} \\ \frac{1}{n}\sum_{i=1}^n (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i\beta_{2,i} \end{pmatrix} \\
=&\frac{1}{n}\sum_{i=1}^n A_i W_i \begin{pmatrix} \beta_1^* \\ \beta_{2,i} \end{pmatrix}.
\end{aligned}$$

Using $\beta_{2,i} = \beta_2$ we then have

$$\frac{1}{n}\sum_{i=1}^n a_i'(\frac{1}{n}\sum_{i=1}^n A_i W_i)^{-1}\frac{1}{n}\sum_{i=1}^n A_i W_i \beta_i = \frac{1}{n}\sum_{i=1}^n a_i'\begin{pmatrix} \beta_1^* \\ \beta_2 \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n a_i'\beta_i.$$

Now, substituting into (B.2), we see that

$$E\Big[\frac{1}{n}\sum_{i=1}^{n} a_i'(\frac{1}{n}\sum_{i=1}^{n} A_iW_i)^{-1}\frac{1}{n}\sum_{i=1}^{n} A_iW_iQ_iQ_i^{\dagger}\beta_i\Big|X_1,...,X_n\Big] = E[\frac{1}{n}\sum_{i=1}^{n} a_i'\beta_i|X_1,...,X_n]$$

For the second term, note that by Assumption 2 and independence of the observations, $u_i$ is jointly independent of $a_i$ (and thus $A_i$) conditional on $X_1, X_2, ..., X_n$, and so:

$$\begin{aligned}
&E\Big[\frac{1}{n}\sum_{i=1}^{n} a_i'(\frac{1}{n}\sum_{i=1}^{n} A_iW_i)^{-1}\frac{1}{n}\sum_{i=1}^{n} A_iW_iQ_i^{\dagger}B_i'u_i\Big|X_1,...,X_n\Big]\\
=&\frac{1}{n}\sum_{i=1}^{n} E\Big[(\frac{1}{n}\sum_{i=1}^{n} a_i)'(\frac{1}{n}\sum_{i=1}^{n} A_iW_i)^{-1}A_iW_iQ_i^{\dagger}B_i'\Big|X_1,...,X_n\Big]E[u_i|X_1,...,X_n]\\
=&\frac{1}{n}\sum_{i=1}^{n} E\Big[(\frac{1}{n}\sum_{i=1}^{n} a_i)'(\frac{1}{n}\sum_{i=1}^{n} A_iW_i)^{-1}A_iW_iQ_i^{\dagger}B_i'\Big|X_1,...,X_n\Big]E[u_i|X_i]\\
=&0
\end{aligned}$$

where we have also used that $W_i$ is a function of $X_i$. So in all

$$E[\hat{\theta}|X_1,...,X_n] = \frac{1}{n}\sum_{i=1}^{n} E[a_i'\beta_i|X_1,...,X_n].$$

If $E[|\hat{\theta}|] < \infty$ then the law of iterated expectations holds $E[\hat{\theta}] = E\big[E[\hat{\theta}|X_1,...,X_n]\big]$. Taking expectations of both sides of the above we get $E[\hat{\theta}] = E[a_i'\beta_i]$. □

*Proof of Proposition 3.* Let $a_{0,i}$ denote the first component of $a_i$ and $a_{1,i}$ the vector that contains the remaining components.

Define the block matrices $\tilde{W}_i := \begin{pmatrix} 1 & \bar{B}_i'(\tilde{Q}_i/\lambda + D_{1,i})^{-1}D_{1,i} \\ 0 & (\tilde{Q}_i/\lambda + D_{1,i})^{-1}\tilde{Q}_i \end{pmatrix}$, $\tilde{A}_i := \begin{pmatrix} a_{0,i} & a_{1,i}'/\lambda \\ 0 & I \end{pmatrix}$, and $\tilde{L}_i = \begin{pmatrix} 1 & -\bar{B}_i'(\tilde{Q}_i/\lambda + D_{1,i})^{-1}/\lambda \\ 0 & (\tilde{Q}_i/\lambda + D_{1,i})^{-1} \end{pmatrix}$. Then with some work, one can show that:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} a_i'(\frac{1}{n}\sum_{i=1}^{n} \tilde{A}_i\tilde{W}_i)^{-1}\frac{1}{n}\sum_{i=1}^{n} \tilde{A}_i\tilde{L}_i\begin{pmatrix} 1 & 0 \\ -\bar{B}_i & I \end{pmatrix}\frac{1}{T_i}B_i'S_i \tag{B.3}$$

Now, using that $D_{1,i}$ is non-singular, it is easy to see that $\lim_{\lambda\to\infty}\tilde{W}_i = \begin{pmatrix} 1 & \bar{B}_i' \\ 0 & D_{1,i}^{-1}\tilde{Q}_i \end{pmatrix}$, $\lim_{\lambda\to\infty}\tilde{A}_i = \begin{pmatrix} a_{0,i} & 0 \\ 0 & I \end{pmatrix}$, and $\lim_{\lambda\to\infty}\tilde{L}_i = \begin{pmatrix} 1 & 0 \\ 0 & D_{1,i}^{-1} \end{pmatrix}$. Passing the limits we get that $\lim_{\lambda\to\infty}\hat{\theta}$ is equal to the expression below.

$$\begin{aligned}
&\lim_{\lambda\to\infty}\frac{1}{n}\sum_{i=1}^{n} a_i'(\frac{1}{n}\sum_{i=1}^{n} \tilde{A}_i\tilde{W}_i)^{-1}\frac{1}{n}\sum_{i=1}^{n} \tilde{A}_i\tilde{L}_i\begin{pmatrix} 1 & 0 \\ -\bar{B}_i & I \end{pmatrix}\frac{1}{T_i}B_i'S_i\\
=&\frac{1}{n}\sum_{i=1}^{n} a_i'\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} a_{0,i} & \frac{1}{n}\sum_{i=1}^{n} a_{0,i}\bar{B}_i' \\ 0 & \frac{1}{n}\sum_{i=1}^{n} D_{1,i}^{-1}\tilde{Q}_i \end{pmatrix}^{-1}\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} a_{0,i} & 0 \\ 0 & D_{1,i}^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\bar{B}_i & I \end{pmatrix}\frac{1}{T_i}B_i'S_i
\end{aligned} \tag{B.4}$$

Multiplying out, we have:

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} a_{0,i} & 0 \\ 0 & D_{1,i}^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\bar{B}_i & I \end{pmatrix}\frac{1}{T_i}B_i'S_i = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i} a_{0,i}S_{i,t} \\ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i} D_{1,i}^{-1}(B_{1,i,t} - \bar{B}_i)S_{i,t} \end{pmatrix}.$$

Applying the formula for the inverse of a block matrix and simplifying:

$$\frac{1}{n}\sum_{i=1}^{n} a_i' \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} a_{0,i} & \frac{1}{n}\sum_{i=1}^{n} a_{0,i}\bar{B}_i' \\ 0 & \frac{1}{n}\sum_{i=1}^{n} D_{1,i}^{-1}\tilde{Q}_i \end{pmatrix}^{-1}$$
$$= \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n} a_{1,i}' \end{pmatrix} \begin{pmatrix} 1 & -(\frac{1}{n}\sum_{i=1}^{n} a_{0,i}\bar{B}_i')(\frac{1}{n}\sum_{i=1}^{n} D_{1,i}^{-1}\tilde{Q}_i)^{-1} \\ 0 & (\frac{1}{n}\sum_{i=1}^{n} D_{1,i}^{-1}\tilde{Q}_i)^{-1} \end{pmatrix}$$

Substituting into (B.4) and multiplying, we get:

$$\lim_{\lambda\to\infty} \hat{\theta} = \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n} a_{1,i}' \end{pmatrix} \begin{pmatrix} 1 & -(\frac{1}{n}\sum_{i=1}^{n} a_{0,i}\bar{B}_i')(\frac{1}{n}\sum_{i=1}^{n} D_{1,i}^{-1}\tilde{Q}_i)^{-1} \\ 0 & (\frac{1}{n}\sum_{i=1}^{n} D_{1,i}^{-1}\tilde{Q}_i)^{-1} \end{pmatrix}$$
$$\times \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i} a_{0,i}S_{i,t} \\ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i} D_{1,i}^{-1}(B_{1,i,t}-\bar{B}_i)S_{i,t} \end{pmatrix}$$
$$= \frac{1}{n}\sum_{i=1}^{n} a_i' \begin{pmatrix} \bar{S}_i - \bar{B}_i'\hat{\beta}_{GFE,1} \\ \hat{\beta}_{GFE,1} \end{pmatrix}.$$

In the special case of $D_{1,i} = D_{1,1}$, for all $i$ we have

$$\hat{\beta}_{GFE,1} = (\frac{1}{n}\sum_{i=1}^{n} D_{1,1}^{-1}\tilde{Q}_i)^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i} D_{1,1}^{-1}(B_{1,i,t}-\bar{B}_i)S_{i,t}$$
$$= (\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i)^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{t=1}^{T_i}(B_{1,i,t}-\bar{B}_i)S_{i,t} = \hat{\beta}_{FE,1}.$$

□

*Proof of Proposition 4.* If $C_i = 1$ then $Z_i$ does not vary, and so $\tilde{Q}_i = \begin{pmatrix} 0 & 0 \\ 0 & \hat{Q}_{1,i} \end{pmatrix}$. By supposition $\hat{Q}_{1,i}$ is non-singular and $D_i$ is diagonal. Let $d_{1,i}$ be the second element of the leading diagonal of $D_i$. From the above we see that in this case

$$(\tilde{Q}_i + \lambda D_{1,i})^{-1} = \begin{pmatrix} \frac{1}{\lambda d_{1,i}} & 0 \\ 0 & (\hat{Q}_{1,i} + \lambda\tilde{D}_{1,i})^{-1}. \end{pmatrix}$$

We can write $W_i$ as $W_i = \begin{pmatrix} 1 & \lambda\bar{B}_i'(\tilde{Q}_i + \lambda D_{1,i})^{-1}D_{1,i} \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i \end{pmatrix}$. Substituting, we get

$$W_i = \begin{pmatrix} 1 & \bar{Z}_i & \lambda\bar{B}_{2,i}'(\hat{Q}_{1,i} + \lambda\tilde{D}_{1,i})^{-1}\tilde{D}_{1,i} \\ 0 & 0 & 0 \\ 0 & 0 & (\hat{Q}_{1,i} + \lambda\tilde{D}_{1,i})^{-1}\hat{Q}_{1,i}, \end{pmatrix}$$

where $\bar{Z}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} Z_i$ and $\bar{B}_{2,i} = \frac{1}{T_i}\sum_{t=1}^{T_i} B_{2,it}$. And so, taking the limit, we obtain $\lim_{\lambda\to 0} W_i = \begin{pmatrix} 1 & \bar{Z}_i & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{pmatrix}$. On the other hand, if $C_i = 0$ and so $Q_i$ is non-singular, then it is easy to see that $\lim_{\lambda\to 0} W_i$ is the identity.

Now, let us consider the limit of the individual ridge estimates. Note we can write these estimates as $\hat{\beta}_i = \begin{pmatrix} \bar{S}_i - (\bar{Z}_i, \bar{B}_{2,i}')\hat{\beta}_{1,i} \\ \hat{\beta}_{1,i}, \end{pmatrix}$, where $\hat{\beta}_{1,i}$ are slope parameters given by

$$\hat{\beta}_{i,1} = (\tilde{Q}_i + \lambda D_{1,i})^{-1}\frac{1}{T_i}\sum_{t=1}^{T_i}(B_{1,it} - \bar{B}_i)S_i.$$

If $Z_{it}$ is constant, then this becomes

$$\begin{aligned}\hat{\beta}_{i,1} &= \begin{pmatrix} \frac{1}{\lambda d_{1,i}} & 0 \\ 0 & (\hat{Q}_{1,i} + \lambda \tilde{D}_{1,i})^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ \frac{1}{T_i} \sum_{t=1}^{T_i} (B_{2,it} - \bar{B}_{2,i}) S_i \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ (\hat{Q}_{1,i} + \lambda \tilde{D}_{1,i})^{-1} \frac{1}{T_i} \sum_{t=1}^{T_i} (B_{2,it} - \bar{B}_{2,i}) S_i \end{pmatrix}.\end{aligned}$$

Because $\hat{Q}_{1,i}$ is non-singular we have

$$\lim_{\lambda \to 0} (\hat{Q}_{1,i} + \lambda \tilde{D}_{1,i})^{-1} \frac{1}{T_i} \sum_{t=1}^{T_i} (B_{2,it} - \bar{B}_{2,i}) S_i = \hat{Q}_{1,i}^{-1} \frac{1}{T_i} \sum_{t=1}^{T_i} (B_{2,it} - \bar{B}_{2,i}) S_i = \tilde{\beta}_{i,3},$$

and so, we see that $\lim_{\lambda \to 0} \hat{\beta}_i = (\bar{S}_i - \bar{B}_{2,i}' \tilde{\beta}_{i,3}, 0, \tilde{\beta}_{i,3}')'$. However, if $i \notin \mathcal{A}$ then, letting $\tilde{\beta}_{i,2}$ and $\tilde{\beta}_{i,3}$ be the coefficients from individual OLS regression, $\lim_{\lambda \to 0} \hat{\beta}_i$ is equal to $(\bar{S}_i - \bar{Z}_i \tilde{\beta}_{i,2} - \bar{B}_{2,i}' \tilde{\beta}_{i,3}, \tilde{\beta}_{i,2}, \tilde{\beta}_{i,3}')'$.

Let $a_{0,i}$ and $a_{1,i}$ be the first and second components of $a_i$, and $a_{2,i}$ the vector of remaining components. Substituting everything, we see

$$\begin{aligned}\lim_{\lambda \to 0} \hat{\theta} =& \frac{1}{n} \sum_{i=1}^{n} a_i' \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} a_{0,i} & \frac{1}{n} \sum_{i \in \mathcal{A}} a_{0,i} \bar{Z}_i + \frac{1}{n} \sum_{i \notin \mathcal{A}} a_{1,i} & \frac{1}{n} \sum_{i=1}^{n} a_{2,i}' \\ 0 & \hat{p} & 0 \\ 0 & 0 & I \end{pmatrix}^{-1} \\ & \times \begin{pmatrix} x \\ \frac{1}{n} \sum_{i=1}^{n} (1 - C_i) \tilde{\beta}_{i,2} \\ \frac{1}{n} \sum_{i=1}^{n} \tilde{\beta}_{i,3} \end{pmatrix},\end{aligned}$$

where $x$ is given below:

$$\begin{aligned}x =& \frac{1}{n} \sum_{i=1}^{n} C_i \Big( a_{0,i} (\bar{S}_i - \bar{B}_{2,i}' \tilde{\beta}_{i,3}) + a_{2,i}' \tilde{\beta}_{i,3} \Big) \\ &+ \frac{1}{n} \sum_{i=1}^{n} (1 - C_i) \Big( a_{0,i} (\bar{S}_i - \bar{Z}_i \tilde{\beta}_{i,2} - \bar{B}_{2,i}' \tilde{\beta}_{i,3}) + a_{1,i} \tilde{\beta}_{i,2} + a_{2,i}' \tilde{\beta}_{i,3} \Big).\end{aligned}$$

Using the formula for the inverse of a block matrix and multiplying out, the above becomes

$$\lim_{\lambda \to 0} \hat{\theta} = x - \frac{1}{\hat{p} n} \sum_{i=1}^{n} C_i (a_{0,i} \bar{Z}_i + a_{1,i}) \frac{1}{n} \sum_{i=1}^{n} (1 - C_i) \tilde{\beta}_{i,2}.$$

Substituting for $x$ and simplifying:

$$\begin{aligned}\lim_{\lambda \to 0} \hat{\theta} =& \frac{1}{n} \sum_{i=1}^{n} C_i a_{0,i} (\bar{S}_i - \bar{B}_{2,i}' \tilde{\beta}_{i,3}) - \frac{1}{n} \sum_{i=1}^{n} C_i (a_{0,i} \bar{Z}_i + a_{1,2,i}) \frac{1}{n \hat{p}} \sum_{i=1}^{n} (1 - C_i) \tilde{\beta}_{i,2} \\ &+ \frac{1}{n} \sum_{i=1}^{n} (1 - C_i) a_{0,i} (\bar{S}_i - \bar{Z}_i \tilde{\beta}_{i,2} - \bar{B}_{2,i}' \tilde{\beta}_{i,3}) + \frac{1}{n} \sum_{i=1}^{n} (1 - C_i) a_{1,i} \tilde{\beta}_{i,2} \\ &+ \frac{1}{n} \sum_{i=1}^{n} (1 - C_i) a_{2,i}' \tilde{\beta}_{i,3} + \frac{1}{n} \sum_{i=1}^{n} C_i a_{2,i}' \tilde{\beta}_{i,3}.\end{aligned}$$

Using the variables defined in the theorem, we can rewrite the limit of $\hat{\theta}$ more succinctly as $\lim_{\lambda \to 0} \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} a_i' \tilde{\beta}_i^*$. □

*Proof of Proposition 5.* First note that

$$(\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i = D_{1,i}^{-1/2}(\tilde{B}_i'\tilde{B}_i + \lambda I)^{-1}\tilde{B}_i'\tilde{B}_i D_{1,i}^{1/2}.$$

By the properties of the Moore-Penrose pseudo-inverse $\lim_{\lambda\to 0}(\tilde{B}_i'\tilde{B}_i + \lambda I)^{-1}\tilde{B}_i' = \tilde{B}_i^{\dagger}$. And so $\lim_{\lambda\to 0}(\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i = P_i$. By definition of $W_i$ we then get

$$\lim_{\lambda\to 0} W_i = \lim_{\lambda\to 0}\begin{pmatrix} 1 & \bar{B}_i'\big(I - (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i\big) \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1}\tilde{Q}_i \end{pmatrix} = \begin{pmatrix} 1 & \bar{B}_i'(I - P_i) \\ 0 & P_i \end{pmatrix}.$$

In addition, using the definition of $\hat{\beta}_{i,1}$, we have

$$\lim_{\lambda\to 0}\hat{\beta}_{i,1} = \lim_{\lambda\to 0} D_{1,i}^{-1/2}\big(\tilde{B}_i'\tilde{B}_i + \lambda I\big)^{-1}\tilde{B}_i' S_i/T_i = D_{1,i}^{-1/2}\tilde{B}_i^{\dagger}S_i/T_i = \tilde{\beta}_{i,1}^{\circ}.$$

And so, since $\hat{\beta}_i = (\bar{S}_i - \bar{B}_{1,i}'\hat{\beta}_{i,1}, \hat{\beta}_{i,1}')'$, we see that $\lim_{\lambda\to 0}\hat{\beta}_i = (\bar{S}_i - \bar{B}_{1,i}'\tilde{\beta}_{i,1}^{\circ}, \tilde{\beta}_{i,1}^{\circ\prime})'$.

Combining and using the definition of $A_i$, after some work simplifying we obtain

$$\begin{aligned}\lim_{\lambda\to 0}\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} a_{it}'\Big(\frac{1}{n}\sum_{i=1}^{n} A_i\begin{pmatrix} 1 & \bar{B}_i'(I - P_i) \\ 0 & P_i \end{pmatrix}\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n} A_i\begin{pmatrix} \bar{S}_i - \bar{B}_{1,i}'\tilde{\beta}_{i,1}^{\circ} \\ \tilde{\beta}_{i,1}^{\circ}\end{pmatrix} \\ &= \frac{1}{n}\sum_{i=1}^{n} a_i'\tilde{\beta}_i^{*}.\end{aligned}$$

□

*Proof of Proposition 6.* From the FOCs for the optimization problem in the proposition we get

$$\beta_i^{\text{Post}} = (Q_i + \lambda D_i)^{-1}\frac{1}{T_i}B_i'S_i + (Q_i + \lambda D_i)^{-1}\lambda D_i\bar{\beta}.$$

Substituting for $\beta_i^{\text{Post}}$ into $\frac{1}{n}\sum_{i=1}^{n} A_i\beta_i^{\text{Post}} = \frac{1}{n}\sum_{i=1}^{n} A_i\bar{\beta}$, we see

$$\frac{1}{n}\sum_{i=1}^{n} A_i(Q_i + \lambda D_i)^{-1}\frac{1}{T_i}B_i'S_i + \frac{1}{n}\sum_{i=1}^{n} A_i(Q_i + \lambda D_i)^{-1}\lambda D_i\bar{\beta} = \frac{1}{n}\sum_{i=1}^{n} A_i\bar{\beta}.$$

Solving for $\bar{\beta}$ and simplifying yields

$$\big(\frac{1}{n}\sum_{i=1}^{n} A_i(Q_i + \lambda D_i)^{-1}Q_i\big)^{-1}\frac{1}{n}\sum_{i=1}^{n} A_i(Q_i + \lambda D_i)^{-1}\frac{1}{T_i}B_i'S_i = \bar{\beta}.$$

Substituting back into $\frac{1}{n}\sum_{i=1}^{n} A_i\beta_i^{\text{Post}} = \frac{1}{n}\sum_{i=1}^{n} A_i\bar{\beta}$ we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} A_i\beta_i^{\text{Post}} &= \big(\frac{1}{n}\sum_{i=1}^{n} A_i\big)\big(\frac{1}{n}\sum_{i=1}^{n} A_i(Q_i + \lambda D_i)^{-1}Q_i\big)^{-1}\frac{1}{n}\sum_{i=1}^{n} A_i(Q_i + \lambda D_i)^{-1}\frac{1}{T_i}B_i'S_i \\ &= \frac{1}{n}\sum_{i=1}^{n} A_i(\overline{AW})^{-1}\overline{A\beta}.\end{aligned}$$

Multiplying both sides by a row vector whose first entry is one and with remaining entries zero gives the result. □

**Lemma 1 (General Asymptotics).** *Suppose that Assumptions 1, 2, and 3.i-vii hold and* $W_i Q_i^\dagger Q_i = W_i$. *Define* $\kappa_n = E[\|W_i - I\|]$ *and* $\gamma_n = E[\|W_i (Q_i^\dagger)^{1/2}\|^2]$ *and suppose* $\kappa_n = o(1)$, *then*

$$\hat{\theta} - \theta_0 = O_p\Big(n^{-1/2} + \kappa_n + \sqrt{\frac{\gamma_n}{nT}} + \sqrt{\frac{\gamma_n J \kappa_n^2}{nT}} + (1+\sqrt{\gamma_n})\ell_n\Big).$$

*If in addition* $\kappa_n, \ell_n, \ell_n\sqrt{\gamma_n} = o(\sqrt{\frac{1}{n}})$, $\frac{\gamma_n J \kappa_n^2}{T} = o(1)$, *and Assumptions 4.i and 4.ii hold for* $\delta, v, q > 0$ *with* $(v-2)(q-2) > 4$ *and:*

$$n^{(\frac{1}{v}+\frac{1}{q}-\frac{1}{2})} E[\|W_i (Q_i^\dagger)^{1/2}\|^q]^{1/q} (\gamma_n/T)^\delta \to 0,$$

*then* $\hat{\theta} - \theta_0 = O_p(n^{-1/2})$ *and we have* $\sqrt{n}\sigma_n^{-1}(\hat{\theta} - \theta_0) \sim^a N(0,1)$. *Where the variance* $\sigma_n^2$ *is equal to* $E[|\frac{1}{T_i} a_i' W_i Q_i^\dagger B_i' u_i|^2] + Var(a_i'\beta_i)$.

*Proof.* Note that under the time-stationary condition in Assumption 1, we have that for all $t = 1, ..., T$ we can define $\beta_i = E[\beta(\eta_{it})|X_i]$, and by Assumption 2 $a_i$ and $u_i$ are independent conditional on $X_i$, So expanding and using the definitions of $r(\cdot,\cdot)$ and $a_i$, it follows that

$$\begin{aligned} &E\big[\frac{1}{T}\sum_{t=1}^T \big(H_{it}^+ s(X_{it}^+, \eta_{it}) - H_{it}^- s(X_{it}^-, \eta_{it})\big)\big|X_i\big] \\ =&E[a_i|X_i]'\beta_i + E[\frac{1}{T}\sum_{t=1}^T H_{it}^+ r(X_{it}^+, \eta_{it})|X_i] - E[\frac{1}{T}\sum_{t=1}^T H_{it}^- r(X_{it}^-, \eta_{it})|X_i]. \end{aligned}$$

By Assumption 3, the terms in the RHS are bounded above in magnitude, and so by iterated expectations we have:

$$\theta_0 = E[a_i'\beta_i] + \frac{1}{T}\sum_{t=1}^T E[H_t^+(X_i) r(X_{it}^+, \eta_{it})] - \frac{1}{T}\sum_{t=1}^T E[H_t^-(X_i) r(X_{it}^-, \eta_{it})]$$

It will be convenient to define the mean-zero random variable $\epsilon_i$ as follows

$$\epsilon_i = a_i'\beta_i - E[a_i'\beta_i] + \frac{1}{n}\sum_{i=1}^n a_i' W_i Q_i^\dagger \frac{1}{T_i} B_i' u_i.$$

Using the expression for $\theta_0$ above, we decompose the estimation error $\hat{\theta} - \theta_0$ into a zero-mean part $\frac{1}{n}\sum_{i=1}^n \epsilon_i$, and a number of remainder terms which are generally not mean-zero:

$$\begin{aligned} \hat{\theta} - \theta_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i =& \frac{1}{n}\sum_{i=1}^n a_i'(I - W_i)\big(\frac{1}{n}\sum_{i=1}^n A_i W_i\big)^{-1} \frac{1}{n}\sum_{i=1}^n A_i W_i \beta_i + \frac{1}{n}\sum_{i=1}^n a_i'(W_i - I)\beta_i \\ &+ \frac{1}{n}\sum_{i=1}^n a_i'(I - W_i)\big(\frac{1}{n}\sum_{i=1}^n A_i W_i\big)^{-1} \frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger \frac{1}{T_i} B_i' u_i \\ &+ \frac{1}{n}\sum_{i=1}^n a_i'\big(\frac{1}{n}\sum_{i=1}^n A_i W_i\big)^{-1} \frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger \frac{1}{T_i} B_i' r_i \\ &- \frac{1}{T}\sum_{t=1}^T E\big[H_t^+(X_i) r(X_{it}^+, \eta_{it})\big] + \frac{1}{T}\sum_{t=1}^T E\big[H_t^-(X_i) r(X_{it}^-, \eta_{it})\big], \end{aligned}$$

where we have used that $a_i = A_i'v$ for some vector $v$ and that $W_iQ_i^\dagger Q_i = W_i$. By the triangle inequality and properties of the matrix norm and $\|a_i\| \leq c$ by Assumption 3.i. we get:

$$\begin{aligned}\|\hat{\theta} - \theta_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i\| \leq & \|\frac{1}{n}\sum_{i=1}^n a_i'(I - W_i)\|\|\big(\frac{1}{n}\sum_{i=1}^n A_iW_i\big)^{-1}\|\|\frac{1}{n}\sum_{i=1}^n A_iW_i\beta_i\| \\ & + \|\frac{1}{n}\sum_{i=1}^n a_i'(W_i - I)\beta_i\| \\ & + \|\frac{1}{n}\sum_{i=1}^n a_i'(I - W_i)\|\|\big(\frac{1}{n}\sum_{i=1}^n A_iW_i\big)^{-1}\|\|\frac{1}{n}\sum_{i=1}^n A_iW_iQ_i^\dagger\frac{1}{T_i}B_i'u_i\| \\ & + c\|\big(\frac{1}{n}\sum_{i=1}^n A_iW_i\big)^{-1}\|\|\frac{1}{n}\sum_{i=1}^n A_iW_iQ_i^\dagger\frac{1}{T_i}B_i'r_i\| \\ & + \|\frac{1}{T}\sum_{t=1}^T E\big[H_t^+(X_i)r(X_{it}^+, \eta_{it})\big]\| + \|\frac{1}{T}\sum_{t=1}^T E\big[H_t^-(X_i)r(X_{it}^-, \eta_{it})\big]\|\end{aligned}$$

The RHS above contains a number of objects which we derive rates for below.

## Step 1: Derive Rates for the Remainder

**1.** $\frac{1}{n}\sum_{i=1}^n a_i'(W_i - I), \frac{1}{n}\sum_{i=1}^n A_i(W_i - I) = O_p(\kappa_n)$

First we establish that $\frac{1}{n}\sum_{i=1}^n a_i'(W_i - I)$ and $\frac{1}{n}\sum_{i=1}^n A_i(W_i - I)$ are both $O_p(\kappa_n)$. Under Assumption 3.i $\|a_i\|, \|A_i\| \leq c$ and so by the triangle inequality and definition of the matrix norm:

$$\begin{aligned}\|\frac{1}{n}\sum_{i=1}^n a_i'(W_i - I)\| &\leq \frac{1}{n}\sum_{i=1}^n \|a_i'(W_i - I)\| \leq c\frac{1}{n}\sum_{i=1}^n \|W_i - I\| = O_p(\kappa_n) \\ \|\frac{1}{n}\sum_{i=1}^n A(W_i - I)\| &\leq \frac{1}{n}\sum_{i=1}^n \|A_i(W_i - I)\| \leq c\frac{1}{n}\sum_{i=1}^n \|W_i - I\| = O_p(\kappa_n)\end{aligned}$$

Where the final equalities both follow by Markov's inequality. By supposition $E[\|W_i - I\|] = o(1)$ and so we see that the terms on the LHSs above are also $o_p(1)$.

**2.** $\big(\frac{1}{n}\sum_{i=1}^n A_iW_i\big)^{-1} = O_p(1)$

By Assumption 3.ii, $\|(\frac{1}{n}\sum_{i=1}^n A_i)^{-1}\| = O_p(1)$ and we have already shown that $\|\frac{1}{n}\sum_{i=1}^n A(W_i - I)\| = O_p(\kappa_n) = o_p(1)$, and so we have:

$$\|\big(\frac{1}{n}\sum_{i=1}^n A_iW_i\big)^{-1} - (\frac{1}{n}\sum_{i=1}^n A_i)^{-1}\| = O_p\big(E[\|W_i - I\|]\big) = o_p(1)$$

From the above and the fact that $\|(\frac{1}{n}\sum_{i=1}^n A_i)^{-1}\| = O_p(1)$, we get from the triangle inequality that $\|\big(\frac{1}{n}\sum_{i=1}^n A_iW_i\big)^{-1}\| = O_p(1)$.

**3.** $\|\frac{1}{n}\sum_{i=1}^n A_iW_i\beta_i\| = O_p(1)$

By the triangle inequality and properties of the matrix norm

$$\|\frac{1}{n}\sum_{i=1}^n A_iW_i\beta_i\| \leq \frac{1}{n}\sum_{i=1}^n \|A_i\|\|W_i\|\|\beta_i\| \leq c^2\frac{1}{n}\sum_{i=1}^n \|W_i\|.$$

The second inequality follows from $\|A_i\|, \|\beta_i\| \leq c$ by Assumption 3.i and 3.iii. Recall that $E[\|W_i - I\|] = o(1)$ and so by the triangle inequality $E[\|W_i\|] = O(1)$. Thus we have by Markov's inequality $\|\frac{1}{n}\sum_{i=1}^n A_i W_i \beta_i\| = O_p(1)$.

**4.** $|\frac{1}{n}\sum_{i=1}^n a_i'(W_i - I)\beta_i| = O_p(\kappa_n)$

Note that by the triangle inequality and properties of the matrix norm

$$|\frac{1}{n}\sum_{i=1}^n a_i'(W_i - I)\beta_i| \leq c^2 \frac{1}{n}\sum_{i=1}^n \|W_i - I\| = O_p(E[\|W_i - I\|]),$$

where the inequality follows from Assumptions 3.i and 3.iii, and the final line by Markov's inequality.

**5.** $\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger \frac{1}{T_i} B_i' u_i = O_p\Big(\sqrt{\frac{J\gamma_n}{Tn}}\Big)$

Note that $E[u_{it}|X_i] = 0$ and thus $A_i W_i Q_i^\dagger B_i' u_i$ is mean zero. Moreover, using standard trace inequalities, and properties of the psuedo-inverse we get:

$$\begin{aligned}
E\big[\|\frac{1}{T_i} A_i W_i Q_i^\dagger B_i' u_i\|^2\big] =& E\big[\frac{1}{T_i^2} tr(A_i W_i Q_i^\dagger B_i' E[u_i u_i'|X_i] B_i Q_i^\dagger W_i' A_i')\big] \\
\leq& cE[\frac{1}{T_i} tr(A_i W_i Q_i^\dagger W_i' A_i')] \\
\leq& JcE[\|A_i W_i Q_i^\dagger W_i' A_i'\|/T_i] \\
\leq& Jc^3 E[\|W_i (Q_i^\dagger)^{1/2}\|^2/T_i]
\end{aligned}$$

Where the first equality uses Assumption 2 and the third line uses Assumption 3.iv which states that $\|E[u_i u_i'|X_i]\| \leq c$ almost surely. And so, by the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger \frac{1}{T_i} B_i' u_i = O_p\Big(\sqrt{\frac{JE[\|W_i (Q_i^\dagger)^{1/2}\|^2/T_i]}{n}}\Big) = O_p\Big(\sqrt{\frac{J\gamma_n}{Tn}}\Big),$$

where the final line follows by Assumption 3.vii.

**6.** $\|\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger \frac{1}{T_i} B_i' r_i\| = O_p(\ell_n \sqrt{\gamma_n})$

Again, using properties of the matrix norm, the triangle inequality, and $\|A_i\| \leq c$,

$$\|\frac{1}{n}\sum_{i=1}^n A_i W_i Q_i^\dagger \frac{1}{T_i} B_i' r_i\| \leq c\frac{1}{n}\sum_{i=1}^n \|W_i Q_i^\dagger \frac{1}{T_i} B_i' r_i\|.$$

Using the properties of the matrix norm and pseudo-inverse, we get

$$\begin{aligned}
\|W_i Q_i^\dagger \frac{1}{T_i} B_i' r_i\|^2 = \|W_i Q_i^\dagger Q_i Q_i^\dagger \frac{1}{T_i} B_i' r_i\|^2 &\leq \|W_i Q_i^\dagger Q_i^{1/2}\|^2 \|Q_i^{1/2} Q_i^\dagger \frac{1}{T_i} B_i' r_i\|^2 \\
&= \|W_i (Q_i^\dagger)^{1/2}\|^2 \frac{1}{T_i}\sum_{t=1}^{T_i} |b(X_{it})' Q_i^\dagger \frac{1}{T_i} B_i' r_i|^2.
\end{aligned}$$

By the properties of least squares projections we have

$$\frac{1}{T_i}\sum_{t=1}^{T_i} |b(X_{it})' Q_i^\dagger \frac{1}{T_i} B_i' r_i|^2 \leq \frac{1}{T_i}\sum_{t=1}^{T_i} r_{it}^2 \leq \ell_n^2,$$

and so:

$$\frac{1}{n}\sum_{i=1}^{n}\|W_iQ_i^{\dagger}\frac{1}{T_i}\sum_{t=1}^{T_i}\frac{1}{T_i}B_i'r_i\| \leq \ell_n\frac{1}{n}\sum_{i=1}^{n}\|W_i(Q_i^{\dagger})^{1/2}\|.$$

By Markov's inequality $\frac{1}{n}\sum_{i=1}^{n}\|W_i(Q_i^{\dagger})^{1/2}\| = O(E[\|W_i(Q_i^{\dagger})^{1/2}\|])$ and by Jensen's ineqality $E[\|W_i(Q_i^{\dagger})^{1/2}\|] \leq \sqrt{\gamma_n}$. And so, using $\|\left(\frac{1}{n}\sum_{i=1}^{n}A_iW_i\right)^{-1}\| = O_p(1)$ established earlier, we see that $\|\frac{1}{n}\sum_{i=1}^{n}A_iW_iQ_i^{\dagger}\frac{1}{T_i}B_i'r_i\| = O_p(\ell_n\sqrt{\gamma_n})$.

**7.** $\|\frac{1}{T}\sum_{t=1}^{T}E\big[H_t^{+}(X_i)r(X_{it}^{+},\eta_{it})\big]\| = O(\ell_n)$, $\|\frac{1}{T}\sum_{t=1}^{T}E\big[H_t^{-}(X_i)r(X_{it}^{-},\eta_{it})\big]\| = O(\ell_n)$

Note that $H_t(X_i^{+})$ is uniformly bounded by Assumption 3.v and $|r(X_{it}^{+},\eta_{it})| \leq \ell_n$ by Assumption 3.vi. Thus $\|\frac{1}{T}\sum_{t=1}^{T}E\big[H_t^{+}(X_i)r(X_{it}^{+},\eta_{it})\big]\| = O(\ell_n)$. and similarly for $\frac{1}{T}\sum_{t=1}^{T}E\big[H_t^{-}(X_i)r(X_{it}^{-},\eta_{it})\big]$.

**Combining:**

Using all of the rates derived above, we get

$$\hat{\theta} - \theta_0 - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i = O_p\Big(\kappa_n + \kappa_n\sqrt{\frac{J\gamma_n}{nT}} + \ell_n + \ell_n\sqrt{\gamma_n}\Big).$$

## Step 2: Derive a Rate for $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i$

Recall $E[\epsilon_i] = 0$. We will derive a rate for the variance of $\epsilon_i$. First note that $u_{it}$ is a function of $X_i$ and $\eta_{it}$, the latter of which is independent of $a_i$ given $X_i$ by Assumption 2, and $E[u_i|X_i] = 0$, so we have:

$$E[\epsilon_i^2] = E[(\frac{1}{T_i}a_i'W_iQ_i^{\dagger}B_i'u_i)^2] + E[(a_i'\beta_i - E[a_i'\beta_i])^2]$$

Using properties of the matrix norm and that $\|a_i\| \leq c$ and $\|E[u_iu_i'|X_i]\| \leq c$ by Assumptions 3.i and 3.iv we have:

$$\begin{aligned}
E[(\frac{1}{T_i}a_i'W_iQ_i^{\dagger}B_i'u_i)^2] =& E[\|E[u_iu_i'|X_i]^{1/2}\frac{1}{T_i}B_iQ_i^{\dagger}W_i'a_i\|^2] \\
\leq& cE[\|\frac{1}{T_i}B_iQ_i^{\dagger}W_i'a_i\|^2] \\
=& cE[\frac{1}{T_i}a_i'W_iQ_i^{\dagger}W_i'a_i] \\
\leq& c^3E[\|W_i(Q_i^{\dagger})^{1/2}\|^2/T_i], \qquad \text{(B.5)}
\end{aligned}$$

where the second line uses the law of iterated expectations. In addition, note that by Assumptions 3.i and iii. $Var(a_i'\beta_i) \leq 2c^2$. So in all: $E[\epsilon_i^2] \leq c^3E[\|W_i(Q_i^{\dagger})^{1/2}\|^2/T_i] + 2c^2$.

So by the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i = O_p\Big(\sqrt{\frac{E[\|W_i(Q_i^{\dagger})^{1/2}\|^2/T_i]}{n}} + n^{-1/2}\Big) = O_p\Big(\sqrt{\frac{\gamma_n}{nT}} + n^{-1/2}\Big)$$

Where the final line follows by Assumption 3.vii.

### Convergence Rate

Putting everything together we see

$$\hat{\theta} - \theta_0 = O_p\Big(n^{-1/2} + \kappa_n + (1+\sqrt{J}\kappa_n)\sqrt{\frac{\gamma_n}{nT}} + (1+\sqrt{\gamma_n})\ell_n\Big).$$

So for root-$n$ convergence it suffices that $\kappa_n, \ell_n, \ell_n\sqrt{\gamma_n} = o(\sqrt{\frac{1}{n}})$, $\frac{\gamma_n J\kappa_n^2}{T}, \sqrt{\frac{\gamma_n}{T}} = o(1)$.

## Applying the Central Limit Theorem

Finally, in order to obtain the last result in the theorem we apply the central limit theorem to $\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i$ using a Lyupunov condition. Given individuals are drawn iid from the population, the Lyupunov condition requires that for some $\delta > 0$, $n^{-\delta/2}E\big[\|\sigma_n^{-1}\epsilon_i\|^{2+\delta}\big]$ goes to zero with $n$, where $\sigma_n$ is the square root of the variance of $\epsilon_i$ and is given by:

$$\sigma_n^2 = E[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^2] + Var(a_i'\beta_i)$$

From Assumption 4.ii we then see $\sigma_n^{-1} = O(1)$. By Jensen's inequality:

$$\begin{aligned} n^{-\delta/2}E\big[\|\sigma_n^{-1}\epsilon_i\|^{2+\delta}\big] \leq\ & 2^{1+\delta}n^{-\delta/2}E\big[|\sigma_n^{-1}(a_i'\beta_i - E[a_i'\beta_i])|^{2+\delta}\big] \\ & + 2^{1+\delta}n^{-\delta/2}E\big[|\sigma_n^{-1}\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\delta}\big] \end{aligned}$$

And so it suffices to show that the following two conditions hold:

$$n^{-\delta/2}E\big[|a_i'\beta_i - E[a_i'\beta_i]|^{2+\delta}\big] \to 0 \tag{B.6}$$

$$n^{-\delta/2}E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\delta}\big] \to 0 \tag{B.7}$$

The first condition follows trivially from Assumptions 3.i and 3.iii. The second condition requires more work. By Holder's inequality we have that for any $0 < \delta \leq \alpha$:

$$\begin{aligned} E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\delta}\big] \leq & E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\alpha}\big]^{\delta/\alpha}E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^2\big]^{1-\delta/\alpha} \\ \leq & E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\alpha}\big]^{\delta/\alpha}c^{2(1-\delta/\alpha)}E[\|W_i(Q_i^\dagger)^{1/2}\|^2/T_i]^{1-\delta/\alpha} \end{aligned}$$

Where we have used our earlier result that $E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^2\big]$ is bounded above by $c^2E[\|W_i(Q_i^\dagger)^{1/2}\|^2/T_i]$, which is $O(E[\|W_i(Q_i^\dagger)^{1/2}\|^2]/T)$ by Assumption 3.vii. Thus for (B.7), it is sufficient that:

$$n^{-1/2}E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\alpha}\big]^{1/\alpha}E[\|W_i(Q_i^\dagger)^{1/2}\|^2/T]^{\frac{1}{\delta}-\frac{1}{\alpha}} \to 0$$

Note that $0 < \delta \leq \alpha \iff \frac{1}{\delta} - \frac{1}{\alpha} \geq 0$, and so the above holds for some $0 < \delta \leq \alpha$ if for some $\delta, \alpha > 0$:

$$n^{-1/2}E\big[|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^{2+\alpha}\big]^{1/\alpha}E[\|W_i(Q_i^\dagger)^{1/2}\|^2/T]^{\delta} \to 0$$

Applying Cauchy-Schwartz we get:

$$|\frac{1}{T_i}a_i'W_iQ_i^\dagger B_i'u_i|^2 \leq \|a_i'W_i(Q_i^\dagger)^{1/2}\|^2\frac{1}{T_i}\|u_i\|^2 \leq c^2\|W_i(Q_i^\dagger)^{1/2}\|^2\frac{1}{T_i}\|u_i\|^2,$$

where the final inequality uses Assumption 3.i. And so, again applying Holder's inequality, for any $p > 1$:

$$E\big[|\frac{1}{T_i}a_i' W_i Q_i^\dagger B_i' u_i|^{2+\alpha}\big] \leq c^{2+\alpha} E\Big[\|W_i(Q_i^\dagger)^{1/2}\|^{2+\alpha}(\frac{1}{T_i}\|u_i\|^2)^{\frac{2+\alpha}{2}}\Big]$$
$$\leq c^{2+\alpha} E\Big[\|W_i(Q_i^\dagger)^{1/2}\|^{(2+\alpha)p}\Big]^{\frac{1}{p}} E\Big[(\frac{1}{T_i}\|u_i\|^2)^{(\frac{2+\alpha}{2})\frac{p}{p-1}}\Big]^{\frac{p-1}{p}}$$

Reparameterizing by $q = (2+\alpha)p$ and $v = (2+\alpha)\frac{p}{p-1}$, in which case $\alpha = \frac{vq-2q-2v}{q+v}$, we get:

$$E\big[|\frac{1}{T_i}a_i' W_i Q_i^\dagger B_i' u_i|^{2+\alpha}\big]^{\frac{1}{\alpha}} \leq \Big(cE\Big[\|W_i(Q_i^\dagger)^{1/2}\|^q\Big]^{\frac{1}{q}} E\Big[(\frac{1}{T_i}\|u_i\|^2)^{\frac{v}{2}}\Big]^{\frac{1}{v}}\Big)^{\frac{vq}{vq-2q-2v}}$$

By Jensen's inequality and using $r > 2$ we have:

$$E[(\frac{1}{T_i}\|u_i\|^2)^{\frac{r}{2}}] = E[(\frac{1}{T_i}\sum_{t=1}^{T_i} u_{it}^2)^{\frac{r}{2}}] \leq E[\frac{1}{T_i}\sum_{t=1}^{T_i} u_{it}^r]$$

By Assumption 4.i the quantity on the RHS is bounded above by $c$, and so:

$$n^{-1/2}E\big[|\frac{1}{T_i}a_i' W_i Q_i^\dagger B_i' u_i|^{2+\alpha}\big]^{\frac{1}{\alpha}} \leq n^{-1/2}\Big[cE[\|W_i(Q_i^\dagger)^{1/2}\|^q]^{1/q}\Big]^{\frac{vq}{vq-2q-2v}}$$

So for (B.7) it suffices that:

$$n^{(\frac{1}{v}+\frac{1}{q}-\frac{1}{2})}E[\|W_i(Q_i^\dagger)^{1/2}\|^q]^{1/q}E[\|W_i(Q_i^\dagger)^{1/2}\|^2/T]^{\delta(1-\frac{2}{v}-\frac{2}{q})} \to 0$$

Now, recall that $q = (2+\alpha)p$ and $v = (2+\alpha)\frac{p}{p-1}$ where $\alpha > 0$ and $p > 1$. We will show that for a given $q$ and $v$, such an $\alpha$ and $p$ exist if $q, v > 0$ and $(q-2)(v-2) > 4$. Fix $q, v > 0$ and consider some $\alpha > 0$. From the expression for $q$ we have $p = q/(2+\alpha)$. Substituting out $p$ from the expression for $v$ and solving for $\alpha$, we get $\alpha = \frac{qv-2q-2v}{q+v}$, and so given $q, v > 0$ , $\alpha > 0$ if and only if $qv - 2q - 2v > 0$ or equivalently, $(q-2)(v-2) > 4$. Moreover, plugging our expression for $\alpha$ back into the expression for $p$, we get $p = (q+v)/v$, which is strictly greater than 1 because $q$ and $v$ are both strictly positive.

Note also that because $qv - 2q - 2v > 0$ and $v, q > 0$, we have that $1 - \frac{2}{v} - \frac{2}{q} > 0$. And since the convergence to zero needs only hold for some fixed $\delta > 0$, we can reparameterize again and we see that it suffices that for some $\delta$:

$$n^{(\frac{1}{v}+\frac{1}{q}-\frac{1}{2})}E[\|W_i(Q_i^\dagger)^{1/2}\|^q]^{1/q}E[\|W_i(Q_i^\dagger)^{1/2}\|^2/T]^{\delta} \to 0,$$

which holds by supposition.

Having confirmed the Lyupunov condition we can apply the central limit theorem to $\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i$. We just need that the squares of the remainder terms go to zero strictly faster than the variance of $\frac{1}{n}\sum_{i=1}^n \epsilon_i$, i.e., each remainder term must be $o(\sqrt{\frac{1}{n}})$. Recall the remainders have the rate below $\kappa_n + \sqrt{\frac{J\gamma_n\kappa_n^2}{nT}} + \ell_n + \ell_n\sqrt{\gamma_n}$. Hence it suffices that $\kappa_n, \ell_n, \ell_n\sqrt{\gamma_n} = o(\sqrt{\frac{1}{n}})$, $\frac{\gamma_n J\kappa_n^2}{T} = o(1)$. □

*Proof of Theorem 1.* For our choice of $W_i$ we have:

$$\|W_i - I\| = \lambda\|(Q_i + \lambda D_i)^{-1}D_i\| \leq \lambda\|(Q_i + \lambda D_i)^{-1}\|\|D_i\| \leq c\lambda\|(Q_i + \lambda D_i)^{-1}\|,$$

and so $E[\|W_i - I\|] = O(\lambda\delta_n)$. Moreover, we have:

$$\|W_i(Q_i^\dagger)^{1/2}\|^2 = \|(Q_i + \lambda D_i)^{-1}Q_i(Q_i + \lambda D_i)^{-1}\| \leq \|(Q_i + \lambda D_i)^{-1}Q_i\|\|(Q_i + \lambda D_i)^{-1}\|$$

Now, with some work one can show that:

$$(Q_i + \lambda D_i)^{-1} Q_i = \begin{pmatrix} 1 & -\bar{B}_i' \\ 0 & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1} \tilde{Q}_i \end{pmatrix} \begin{pmatrix} 1 & \bar{B}_i' \\ 0 & 1 \end{pmatrix},$$

and so by properties of the matrix norm:

$$\|(Q_i + \lambda D_i)^{-1} Q_i\| \leq (1 + \|\bar{B}_i\|)^2 \big(1 + \|(\tilde{Q}_i + \lambda D_{1,i})^{-1} \tilde{Q}_i\|\big)$$

Now, using that $D_{1,i}$ is symmetric and strictly positive definite by Assumption 3.viii,

$$\begin{aligned} \|(\tilde{Q}_i + \lambda D_{1,i})^{-1} \tilde{Q}_i\| &= \|D_{1,i}^{-1/2} \big(D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} + \lambda I\big)^{-1} D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} D_{1,i}^{1/2}\| \\ &\leq \|D_{1,i}^{-1/2}\| \|\big(D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} + \lambda I\big)^{-1} D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2}\| \|D_{1,i}^{1/2}\| \\ &\leq \|D_{1,i}^{-1/2}\| \|D_{1,i}^{1/2}\|. \end{aligned}$$

Where the final line uses that $\|(A + \lambda I)^{-1} A\| \leq 1$ for any positive definite matrix $A$ and $\lambda > 0$. By Assumptions 3.i and 3.viii, $\|\bar{B}_i\|$ $\|D_{1,i}^{-1/2}\|$, $\|D_{1,i}^{1/2}\|$ are all uniformly bounded, and so for some constant $C$, $\|W_i (Q_i^\dagger)^{1/2}\|^2 \leq C \|(Q_i + \lambda D_i)^{-1}\|$. Thus $E[\|W_i - I\|] = O(\lambda \delta_n)$, $E[\|W_i (Q_i^\dagger)^{1/2}\|^2] = O(\delta_n)$, and $E[\|W_i (Q_i^\dagger)^{1/2}\|^q]^{1/q} = O(E[\|(Q_i + \lambda D_i)^{-1}\|^{q/2}]^{1/q})$. Substituting into Lemma 1 then gives the result. □

**Lemma 2.** *Suppose $\lambda > 0$ and with probability 1, the eigenvalues of $D_{1,i}$ are all bounded above by $c$ and below by $1/c$, and $\|\bar{B}_i\| \leq c$. Then for any $\alpha > 0$:*

$$E[\|(Q_i + \lambda D_i)^{-1}\|^\alpha]^{1/\alpha} = O(1 + E[(\mu_{\min}(\tilde{Q}_i) + \lambda)^{-\alpha}]^{1/\alpha})$$

*Thus if $E[\mu_{\min}(\tilde{Q}_i)^{-\alpha}]^{1/\alpha} < c$ then $E[\|(Q_i + \lambda D_i)^{-1}\|^\alpha]^{1/\alpha} = O(1)$.*

*Proof.* We can decompose $(Q_i + \lambda D_i)^{-1}$ into the product of block matrices as follows:

$$(Q_i + \lambda D_i)^{-1} = \begin{pmatrix} 1 & -\bar{B}_i' \\ 0 & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (\tilde{Q}_i + \lambda D_{1,i})^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\bar{B}_i & I \end{pmatrix}$$

By the properties of the Euclidean matrix norm we then have:

$$\|(Q_i + \lambda D_i)^{-1}\| \leq (1 + \|\bar{B}_i\|)^2 (1 + \|(\tilde{Q}_i + \lambda D_{1,i})^{-1}\|)$$

Note that:

$$\begin{aligned} \|(\tilde{Q}_i + \lambda D_{1,i})^{-1}\| &= \|D_{1,i}^{-1/2} \big(D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} + \lambda I\big)^{-1} D_{1,i}^{-1/2}\| \\ &\leq \|\big(D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2} + \lambda I\big)^{-1}\| \|D_{1,i}^{-1/2}\|^2 \\ &\leq \frac{1}{\mu_{\min}\big(D_{1,i}^{-1/2} \tilde{Q}_i D_{1,i}^{-1/2}\big) + \lambda} \times \frac{1}{\mu_{\min}(D_{1,i}^{1/2})^2} \\ &\leq \frac{1}{\mu_{\min}(D_{1,i}^{-1/2})^2 \mu_{\min}(\tilde{Q}_i) + \lambda} \times \frac{1}{\mu_{\min}(D_{1,i}^{1/2})^2} \\ &= \frac{\mu_{\max}(D_{1,i}) / \mu_{\min}(D_{1,i})}{\mu_{\min}(\tilde{Q}_i) + \lambda \mu_{\max}(D_{1,i})} \end{aligned}$$

Where we have used that $D_{1,i}$ is symmetric and so $\mu_{\min}(D_{1,i}^{-1/2})^2 = 1/\mu_{\max}(D_{1,i})$ and $\mu_{\min}(D_{1,i}^{1/2})^2 = \mu_{\min}(D_{1,i})$. Combining, we get:

$$\begin{aligned}\|(Q_i + \lambda D_i)^{-1}\| &\leq (1 + \|\bar{B}_i\|)^2\Big(1 + \frac{\mu_{\max}(D_{1,i})/\mu_{\min}(D_{1,i})}{\mu_{\min}(\tilde{Q}_i) + \lambda\mu_{\max}(D_{1,i})}\Big) \\ &\leq (1+c)^2\Big(1 + \frac{c^2}{\mu_{\min}(\tilde{Q}_i) + \lambda/c}\Big) \\ &= (1+c)^2\Big(1 + \frac{c^3}{\mu_{\min}(\tilde{Q}_i) + \lambda}\Big)\end{aligned}$$

Where the final line uses $\|\bar{B}_i\| \leq c$, $\mu_{\max}(D_{1,i}) \leq c$ and $\mu_{\min}(D_{1,i}) \geq 1/c$ by supposition and we can take $c \geq 1$ without loss of generality. So for any $1 \leq \alpha$:

$$E[\|(Q_i + \lambda D_i)^{-1}\|^\alpha]^{1/\alpha} \leq (1+c)^2 + (1+c)^2 E\Big[\Big(\frac{c^3}{\mu_{\min}(\tilde{Q}_i) + \lambda}\Big)^\alpha\Big]^{1/\alpha},$$

and hence:

$$E[\|(Q_i + \lambda D_i)^{-1}\|^\alpha]^{1/\alpha} = O(1 + E[(\mu_{\min}(\tilde{Q}_i) + \lambda)^{-\alpha}]^{1/\alpha})$$

For the final step simply note that $(\mu_{\min}(\tilde{Q}_i) + \lambda)^{-\alpha} \leq \mu_{\min}(\tilde{Q}_i)^{-\alpha}$. □

*Proof of Corollary 2.* By Lemma 2, $E[\|(Q_i+\lambda D_i)^{-1}\|] = O(1)$. Applying Theorem 1 with $\delta_n = O(1)$ then gives the result. □

**Lemma 3.** *Suppose $\lambda > 0$ and with probability 1, the eigenvalues of $D_{1,i}$ are all bounded above by $c$ and below by $1/c$. Suppose for some $1 \leq \alpha$, $\|\bar{B}_i\| \leq c$ and $E[\mu_i^{-\alpha}]^{1/\alpha} = O(1)$ for some individual-specific random scalar $\mu_i$, then for any fixed $\varepsilon > 0$:*

$$E[\|(Q_i + \lambda D_i)^{-1}\|^\alpha]^{1/\alpha} \leq O\Big(1 + \lambda^{-1} P[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\mu_i]^{1/\alpha}\Big)$$

*Proof.* From Lemma 2:

$$E[\|(Q_i + \lambda D_i)^{-1}\|^\alpha]^{1/\alpha} \leq O(1 + E[(\mu_{\min}(\tilde{Q}_i) + \lambda)^{-\alpha}]^{1/\alpha})$$

Let $\varepsilon \in [0,1]$ and define the binary random variable $\epsilon_i$ by $\epsilon_i = 1\{\mu_{min}(\tilde{Q}_i) \geq (1-\varepsilon)\mu_i\}$. Using this random variable we get:

$$\frac{1}{\mu_{\min}(\tilde{Q}_i) + \lambda} \leq \frac{1}{(1-\varepsilon)\mu_i} + (1-\epsilon_i)\frac{1}{\lambda}$$

Using the triangle inequality and the fact that $\epsilon_i$ is binary we have:

$$E[(\mu_{\min}(\tilde{Q}_i) + \lambda)^{-\alpha}]^{1/\alpha} \leq (1-\varepsilon)^{-1} E[\mu_i^{-\alpha}]^{1/\alpha} + \frac{E[1-\epsilon_i]^{1/\alpha}}{\lambda}$$

By definition we have $E[1-\epsilon_i] = P[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\mu_i]$. Substituting gives the result. □

**Lemma 4.** *Suppose $b(X_{it}) = (1, X_{it})'$ where $X_{it}$ is binary. Suppose $P(X_{it} = 1|\pi_i) = \pi_i$ and the entries of the sequence $\{X_{it}\}_{t=1}^T$ are jointly independent conditional on $\pi_i$. Let $\pi_i$ admit a probability density $f_\pi$ so that $f_\pi(x) \leq C(1-x)^\omega x^\omega$ for some $C$. If $r > 0$ then $E[x_i^{-1}(1-x_i)^{-1}] < \infty$ and:*

$$E[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\pi_i(1-\pi_i)] = O(T^{-(1+\omega)})$$

*If in addition, $\omega > q/2 - 1$ for some $q > 2$, then $E[\pi_i^{-q/2}(1-\pi_i)^{-q/2}] < \infty$.*

*Proof.* Let $\bar{X}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} X_{it}$. In this case $\mu_{min}(\tilde{Q}_i) = \bar{X}_i(1-\bar{X}_i)$ so we have:

$$\begin{aligned}
&P[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\pi_i(1-\pi_i)|\pi_i] \\
=&P[\bar{X}_i(1-\bar{X}_i) \leq (1-\varepsilon)\pi_i(1-\pi_i)|\pi_i] \\
\leq&P[\bar{X}_i \leq \sqrt{1-\varepsilon}\pi_i|\pi_i] + P[(1-\bar{X}_i) \leq \sqrt{1-\varepsilon}(1-\pi_i)|\pi_i]
\end{aligned}$$

Let $\tilde{\varepsilon} = 1 - \sqrt{1-\varepsilon}$. By the multiplicative Chernoff bound, for $0 \leq \tilde{\varepsilon} \leq 1$:

$$P[\bar{X}_i \leq (1-\tilde{\varepsilon})\pi_i|\pi_i] \leq exp(-\tilde{\varepsilon}^2\pi_i T_i/2)$$

By supposition, the pdf of $\pi_i$ is bounded above by $C(1-\pi)^\omega\pi^\omega$ with $\omega > 0$. It is easy to see that $E[\pi_i^{-1}(1-\pi_i)^{-1}] \leq \infty$ and if $\omega > q/2 - 1$ then $E[\pi_i^{-q/2}(1-\pi_i)^{-q/2}] \leq \infty$. Moreover, from the above we get:

$$\begin{aligned}
E\big[P[\bar{X}_i \leq \sqrt{1-\varepsilon}\pi_i|\pi_i]\big] &\leq C\int_0^1 y^\omega(1-y)^\omega exp(-\tilde{\varepsilon}^2 yT/2)dy \\
&\leq C\int_0^1 y^\omega exp(-\tilde{\varepsilon}^2 yT/2)dy \\
&= CT^{-(1+\omega)}(\tilde{\varepsilon}^2/2)^{-(1+\omega)}\int_0^{\tilde{\varepsilon}^2 T/2} u^\omega exp(-u)du \\
&\leq CT^{-(1+\omega)}(\tilde{\varepsilon}^2/2)^{-(1+\omega)}\int_0^{\infty} u^\omega exp(-u)du
\end{aligned}$$

The integral $\int_0^\infty u^\omega exp(-u)du$ is finite for $\omega > -1$ (it is the gamma function evaluated at $\omega+1$), and so we see $E\big[P[\bar{X}_i \leq \sqrt{1-\varepsilon}\pi_i|\pi_i]\big] = O(T^{-(1+\omega)})$.

We can apply the same reasoning for $E\big[P[(1-\bar{X}_i) \leq \sqrt{1-\varepsilon}(1-\pi_i)|\pi_i]\big]$. Combining and using iterated expectations, we get that:

$$E[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\mu_i] = E\big[P[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\mu_i|\pi_i]\big] = O(T^{-(1+\omega)})$$

□

*Proof of Corollary 2 (Binary Regressor).* From Lemma 4 we see $E[\pi_i^{-1}(1-\pi_i)^{-1}] \leq C$. Applying Lemma 3 with $\alpha = 1$, we then get:

$$E[\|(Q_i + \lambda D_i)^{-1}\|] \leq O\Big(1 + \lambda^{-1}P[\mu_{min}(\tilde{Q}_i) \leq (1-\varepsilon)\pi_i(1-\pi_i)]\Big)$$

Using Lemma 4 we then have $E[\|(Q_i + \lambda D_i)^{-1}\|] = O\big(1 + \lambda^{-1}T^{-(1+\omega)}\big)$. Thus we get that if $\lambda = o(1)$ and $T \to \infty$ then $\lambda\delta_n = o(1)$. $J$ is fixed and clearly $r_{it} = 0$ almost surely because the model is exhaustive (so $\ell_n = 0$). Thus combining with Theorem 1 gives the first result (where we have simplified the rate using the fact that $\lambda\kappa_n\sqrt{\frac{\kappa_n}{nT}}$ is dominated). By Lemma 4, if $\omega > q/2$ we have $E[\pi_i^{-q/2}(1-\pi_i)^{-q/2}] \leq C$. Then applying Lemma 3 with $\alpha = q/2$, we get:

$$E[\|(Q_i + \lambda D_i)^{-1}\|^{q/2}]^{\frac{1}{q}} \leq O\Big(1 + \big(\lambda^{-1}T^{-(1+\omega)}\big)^{q/2}\Big)$$

The above is $O(1)$ by supposition and thus the condition in Assumption 4.iii becomes $n^{(\frac{1}{v}+\frac{1}{q}-\frac{1}{2})}(1/T)^\delta = o(1)$. But this holds trivially. In addition, we assume $\frac{T^{-(1+v)}}{\lambda} = O(1)$ and thus $\delta_n = O(1)$ and that $T^{-(1+v)}, \lambda = o(\sqrt{1/n})$, so applying Theorem 1 we are done.

□