Estimation of fixed effects models with missing covariate data, with an application to valuing local water quality[*]

(Job Market Paper Draft)

Most Recent Version [†]

Jessie Coe [‡]

Novvember 2, 2018

Abstract

This paper considers estimation of a linear fixed effects model in which covariate values may be missing. Two inverse probability weighted (IPW) estimators are proposed. The main assumption is a missing at random assumption (MAR) which allows missingness (observation) to be related to the outcome and its shocks, but requires that the probability of observation is not related to the missing values. The inverse of the estimated probability of observation is used to re-weight the estimating equations, which are then estimated in a second stage by either computationally simple pooled OLS, or more asymptotically efficient GMM. Both of the proposed estimators are consistent and $\sqrt{n}$-asymptotically normal, and the asymptotic variance is derived. The main results are developed for the classical linear fixed effects model under strict exogeneity, and the approach generalizes to many panel models, including dynamic linear unobserved effects models.

As an application, the proposed estimator is applied to a hedonic housing price model in which the willingness to pay for local water quality is reflected

[*] I would like to thank my committee members; Jason Abrevaya, Brendan Kline, and Stephen Donald for their invaluable guidance and support; Kuyama, Olmstead, and Zheng for their comments and generously allowing me to use their data. I am also grateful to Sandy Black, Steve Trejo, Richard Coe, members of the UT econometrics writing seminar, and audience members from MEG, TEC, and CESG, for their helpful feedback. All errors are my own.

[†] Also available on my website: https://www.jessiecoe.com

[‡] University of Texas at Austin, email: jessie.coe@utexas.edu

in house prices. Water quality, the main regressor of interest, is missing for many houses in many time periods. Empirical evidence suggests that, in line with the MAR assumption, observation is related to house prices but not water quality itself. The results suggest that accounting for the missing mechanism is important, as the estimated willingness to pay differs in magnitude and statistical significance between the proposed two-step IPW estimator and the more commonly used estimator which drops incomplete observations.

1 Introduction

This paper considers estimation of panel data models with missing covariate values. The main model considered is the classical linear fixed effects model, and the proposed estimation technique extends to many panel data models commonly used in applied work, including the linear fixed effects model under sequential exogeneity and dynamic linear models. The key assumption is the missing at random (MAR) assumption, which allows observation to depend on the outcome variable and its shocks, but assumes that the probability of observation is not directly related to the value of the missing variables. This paper provides the first available estimation approach for dealing with missing at random covariate data in fixed effects models.

The probabilities of observation are used to re-weight the observed data to "recover the balance of a random sample." The resulting inverse probability weighted (IPW) estimator of the slope parameters is consistent and $\sqrt{n}$-asymptotically normal. The proposed estimator is computationally straightforward, flexibly allows for arbitrary patterns of missing values in the covariates over time[1], leaves the linear fixed effects fully flexible, and does not require specification of the distribution of the outcome variable nor any additional structure on the covariates. The proposed estimator is applied to a hedonic house price model that captures homeowners' willingness to pay for water quality in local water amenities (Kuwayama et al. (2018).)

The main question is how to estimate a linear fixed effects model with missing covariate values in the data. By far the most common practice in the face of missing values is dropping observations with missing values and using the resulting complete cases for analysis as if they were a random sample. Implicit in this approach is the assumption that the missingness is *ignorable* (or exogenous[2]), in which case

[1] Referred to as non-monotonic patterns of missing values in the literature, this allows for the case where, for example, a covariate is observed in time period 1, missing in time period 2, and observed in time period 3.

[2] Statistically, exogenous selection assumes that the fully observed sub-sample satisfies the same

estimation using the fully observed sub-sample does not affect consistency[3]. While the inclusion of fixed effects can make the ignorable missingness assumption more palatable[4], the assumption will be violated if missingness is conditionally related to the outcome or shocks to the outcome. This paper considers the MAR assumption (Rubin (1976)), which allows missingness (selection) of the covariates to be related to the outcome, error disturbance, and other fully observed covariates, but not directly related to the missing covariates themselves. Previous approaches to missing data in a panel model have either assumed an outcome model with no time invariant heterogeneity; i.e. fixed effects (Robins et al. (1995), Chen et al. (2010), Moffit et al. (1999).) In the presence of fixed effects, previous literature assumed ignorable missingness (Wooldridge (2010a), Abrevaya (2018)) or missing outcomes (Nijman and Verbeek (1992), Wooldridge (2010b).)

The key to the proposed approach is to apply IPW to the first-differenced estimating equations. While it is well-known that IPW estimators can yield consistent estimators in cross-sectional settings (Rosenbaum and Rubin (1983)), the IPW approach does not directly apply to estimation of fixed effects models. The fixed effects are differenced out, resulting in estimating equations that each involve two time periods. The probability of observation is the joint probability of observing the regressors in both time periods. While estimating equations with multiple time periods are common for panel data models, it is a deviation from the missing data literature, which has focused on a single binary selection variable (Robins et al. (1994), Wooldridge (2007), Chen et al. (2008), Graham et al. (2012)[5].) The model considered here will generally be over-identified with more than two time periods, as there will be more first-differenced moments than slope coefficients. Achieving the main goal of consistent estimation of the slope parameters in the fixed effects model therefore entails two technical contributions to the missing data literature. First, I extend selection to a multivariate binary response model, Second, I consider a population model that is generally over-identified, thereby contributing to the sparse literature on inverse

exogeneity assumptions as the population

[3] There is some difference in the use of the term ignorable missingness in the literature. The statistics literature sometimes uses ignorable missingness to describe any probability of selection that is a function of observed varaibles, such as the missing at random assumption. I maintain the economics usage of ignorable missingness, which refers to any selection process where the unweighted complete case estimator is consistent (Wooldridge (2002)).

[4] For example, if a zero conditional mean assumption is adopted, such as $E[u_{it}|X_{it}, c_i] = 0$, then exogenous selection assumes that the zero condition mean holds when selection is included in the conditioning. This allows selection to be arbitrarily related to the covariates X_{it}, which may not be fully observed, and the unobserved fixed effects c_i.

[5] Chen, Yi, and Cook (Chen et al. (2010)) are a notable exception who consider a panel model without fixed effects, but with missing values in both the outcome and a covariate thus resulting in bivariate selection for each time period. They conclude that accounting for the multivariate structure of observation is important.

probability weighting for missing data in over-identified models[6].

A two-step estimator is proposed, and its asymptotic distribution is derived, where the asymptotics are for fixed T and large N. The first step is estimation of the probability of observation. From the panel setting and first-differencing, selection is a bivariate binary response model. The binary response is observation or not in a given period, and the bivariate structure is from the presence of two time periods in the estimating equations (as opposed to univariate binary selection in a cross-sectional setting.) As is standard in the literature, it is assumed that there are sufficient fully observed variables to consistently model the probability of observing the missing variables. The estimated probabilities, obtained from maximum likelihood estimation of a parametric bivariate binary selection model, are used to correct for non-ignorable selection by re-weighting the observed sample moments. The second step uses the re-weighted moments to estimate the slope parameters. Two versions of the estimator are proposed: (i) a computationally simple pooled OLS estimator that sums the moment functions over time, and (ii) an optimally-weighted GMM estimator that stacks the re-weighted moment functions, thus making use of the over-identification when $T > 2$. The GMM estimator, while computationally more demanding, enjoys better asymptotic efficiency, especially in the case of heteroskedasticity and serial correlation, though the practical gains may be small, a point explored in the simulations.

In cross-sectional models, Wooldridge (Wooldridge (2002), Wooldridge (2007)) formalizes the surprising result that using the estimated selection probabilities from a first-stage conditional MLE of a binary selection model can be more efficient than using the true selection probabilities. In the setting of this paper, when T=2, and thus when the model is just-identified, that result carries over to the IPW estimators proposed here with first-stage conditional MLE from a bivariate binary selection model. Interestingly, the result does not carry over to the panel setting with more than two time periods. It is shown that using the estimated probabilities may result in either lower or higher asymptotic variance, and the direction is not known a priori.

The empirical application applies the proposed IPW-POLS estimator to measure the willingness to pay, as reflected in house prices, for water quality in local water

[6] Chen, Hong, and Tarrozi (Chen et al. (2008)) are one of the few other papers to consider over identification and missing values. They consider a cross-sectional outcome model, but allow for over identification and propose general method of moments estimation, including an IPW-GMM estimator. Their "verify-in-sample" model is similar to the set-up here, except that selection here is a multivariate binary outcome model, whereas they consider selection as a single binary outcome, and the moment functions here may involve different subsets of the data because of the time differencing.

amenities, such as ponds, lakes, and rivers (e.g. Kuwayama et al. (2018), Walsh et al. (2017).) The valuation of a public good or natural resource is important for policy, yet difficult to measure as there often is not a market for such goods. Housing markets offer one way to measure the value of local amenities, in so far as that value is reflected in house prices. Unobserved (to the econometrician) house attributes affect the house price and may be correlated with house or property characteristics, thus the preferred model includes property fixed effects. The identification is off of repeated sales of the same house. Water quality, the main covariate of interest, as measured by the level of dissolved oxygen recorded at nearby monitoring sites, is missing for many properties in many time periods.

The question is why is the water quality measure missing. If the frequency of monitoring is related to the wealth of the area, as captured by the house prices, then ignorable missingness fails, but the missing at random assumption can hold. If instead, water quality is listed as missing when the recording is below some threshold for the measurement device, then missingness is a function of the value of the missing covariate and the missing at random assumption fails, but ignorable missingness may hold. The availability of panel data allows for some suggestive tests of the two assumptions. The empirical evidence suggests the former case, with ignorable missingness failing and MAR holding.

The data are for single family homes in the three counties of the Tampa Bay area (Kuwayama et al. (2018), Zheng (2017).) The data span 17 years, but there are on average only 8 observations of water quality per house. Given the empirical evidence against the ignorable missingness assumption, the unweighted estimates from the literature are suspected to be inconsistent. With suggestive evidence that the missing at random assumption holds, the proposed IPW-POLS is utilized. The results are substantially different from the unweighted first-difference estimator. When the two-step IPW estimator is applied, the estimate of willingness to pay falls and loses significance in Hillsborough county, and increases more than three-fold and becomes highly significant in Pinellas county. The dollar estimates of the value of a 10% increase in dissolved oxygen to the average house differ across estimators by an order of magnitude. As many decisions are made at the county level, accounting for selection appears to have significant implications for local water quality policy.

While the main focus of the paper is the classical linear fixed effects model, the proposed estimation technique applies, under suitable assumptions, to numerous panel models popular in the applied literature, including the linear fixed effects model under sequential exogeneity, and linear dynamic models. Given the popularity of panel models in empirical work, and the prevalence of missing values in data sets, the proposed estimation technique should be of practical importance.

The rest of the paper is organized as follows: the next section presents the classical linear fixed effects model and develops the estimators, the following section formalizes the assumptions and develops the asymptotic theory for the classical linear fixed effects model under strict exogeneity. Section (4) presents some finite sample performance. The empirical application is presented in section (5). Section (6) shows how the estimators may be generalized to handle many panel data models of interest, and the final section briefly concludes.

2 Model and Estimation

This section considers the classical linear fixed effects model under strict exogeneity, in which some of the covariates are not always observed:

$$\begin{aligned} & y_{it} = X_{it}\beta_{x0} + W_{it}\beta_{w0} + c_i + u_{it} \qquad t = 1, ..., T \\ & E[u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i] = 0 \end{aligned} \tag{1}$$

where bold letters denote time histories, for example $\boldsymbol{W}_i = \{W_{i1}, W_{i2}, ..., W_{iT}\}$. Model (1) is assumed to hold for the population and thus holds for a random sample of individuals. Assume $(y_{it}, X_{it}, W_{it})_{i=1,...,N}$ are i.i.d. for each $t = 1, ..., T$, with X_{it} a k_1 vector, W_{it} a k_2 vector where k_2 may equal zero[7], and let $k = k_1 + k_2$. The fixed effects c_i and (X_{it}, W_{it}) may be arbitrarily correlated. In the data, X_{it} is not always observed, but y_{it}, W_{it} are fully observed. If the data are from a survey, the setting is akin to item non-response, as opposed to attrition. Let d_{it} be an indicator that takes the value 1 when X_{it} is fully observed for person i in time period t. Let V_{it} be a vector of not necessarily time-varying variables which are distinct from $(\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{y}_i)$ and are fully observed; for example, V_{it} may include time-invariant variables absorbed in the c_i in model (1). The potentially observed variables are then $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{V}_i)$. This section motivates and defines the two proposed IPW estimators.

The question is how to consistently estimate the slope parameters in (1) when X_{it} is missing at random, as discussed in the introduction, and defined as follows:

Assumption 1. *(MAR)* $\boldsymbol{d}_i|(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{V}_i) \sim \boldsymbol{d}_i|(\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i)$

[7] Though W_{it} is likely non-empty as it will include any time effects.

Assumption (1) states that distribution of observation does not depend directly on the covariates which are not always observed. Observation may depend on the outcome, and thus can indirectly depend on the missing covariate values. The key component of the missing at random assumption is that the probability of observation is not conditionally related to the value of the missing covariates, but is instead a function only of fully observed variables, which can include the y_{it}s. While an exclusion restriction or auxiliary data are not required, additional variables V_{it} influencing selection are allowed; in particular, time-invariant variables that are absorbed in c_i may be good predictors of selection and can be explicitly included in the conditional joint probability.

First-differences are used to eliminate the fixed effects from (1) and obtain estimating equations which are a function of potentially observed variables. First-differencing model (1) yields:

$$\begin{aligned} y_{it} - y_{it-1} &= (X_{it} - X_{it-1})\beta_{x0} + (W_{it} - W_{it-1})\beta_{w0} + u_{it} - u_{it-1} \qquad t = 2, ..., T \\ \Delta y_{it} &= \Delta X_{it}\beta_{x0} + \Delta W_{it}\beta_{w0} + \Delta u_{it} \end{aligned} \tag{2}$$

where one time period is lost to the differencing, and the second equality is standard notation using $\Delta y_{it} = y_{it} - y_{it-1}$..

The first-differenced population moments from the orthogonality conditions implied by the strict exogeneity assumption are[8]:

$$\begin{aligned} E[\Delta X_{it}' \Delta u_{it}] &= 0 \\ E[\Delta W_{it}' \Delta u_{it}] &= 0 \qquad t = 2, ..., T \end{aligned} \tag{3}$$

These are $K(T-1)$ moments for the K slope parameters β_0.

The fully observed data in each time period are $(y_{it}, d_{it}X_{it}, W_{it}, d_{it}, V_{it})_{i=1,...,N}$. The data in the first-differenced moments are fully observed when $d_{it} = 1$ and $d_{it-1} = 1$, the so-called *complete cases*:

$$d_{it}d_{it-1}(\Delta y_{it}, \Delta X_{it}, \Delta W_{it}) \tag{4}$$

The complete cases may not be representative thus the first-differenced moments for the complete cases are not necessarily valid (see remark at the end of this section.)

[8] There are of course infinitely many possible valid moments from the strict exogeneity condition. I focus on the moments standard in the literature.

Consider iterating the expectation on all the potentially observed variables:

$$\begin{aligned} E[d_{it}d_{it-1}\Delta X_{it}'\Delta u_{it}] &= E[E[d_{it}d_{it-1}\Delta X_{it}'\Delta u_{it}|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{V}_i]] \\ &= E[E[d_{it}d_{it-1}|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{W}_i]\Delta X_{it}'\Delta u_{it}] \\ &= E[P[d_{it}=1, d_{it-1}=1|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{V}_i]\Delta X_{it}'\Delta u_{it}] \qquad (5) \end{aligned}$$

where recall that the V_{it} are not necessarily time varying variables that influence observation.

The proposed estimator is an inverse probability weighted estimator. Only the fully observed estimating equations are used, and each estimating equation is re-weighted by the inverse of the probability that equation is fully observed. Let $p_{it}(y, X, W, V) = P[d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{V}_i]$, then, provided the expectation exists, following the same iterated expectations as above yields:

$$E[\frac{d_{it}d_{it-1}}{p_{it}(y, X, W, V)}\begin{pmatrix}\Delta X_{it}' \\ \Delta W_{it}'\end{pmatrix}\Delta u_{it}] = 0 \qquad t = 2, ..., T \qquad (6)$$

These weighted population moment functions are valid for the observed sample. This is the basis for the inverse probability weighted estimators. The completely observed first-differenced estimating equations are re-weighted to make up for the fact that the complete cases themselves may not be a representative panel data set.

For estimation, re-write (6) using (2):

$$E[\frac{d_{it}d_{it-1}}{p_{it}(y, X, W, V)}\begin{pmatrix}\Delta X_{it}' \\ \Delta W_{it}'\end{pmatrix}(\Delta y_{it} - \Delta X_{it}\beta_{x0} - \Delta W_{it}\beta_{w0})] = 0 \qquad t = 2, ...T \qquad (7)$$

The $K(T-1)$ moments (7) will form the basis of estimation.

The population moments (7) require $P(d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{V}_i)$ for $t = 2, ..., T$, which is problematic as this joint probability is generally unknown and is a function of the not fully observed covariates. This is the strength of the MAR assumption as (1) implies:

$$P(d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{V}_i) = P(d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i) \qquad (8)$$

for each $t = 2, ..., T$. The MAR assumption, along with a parametric specification for the joint probability of observation, are the main identifying assumptions used in this paper, and are collected below in (2).

Assumption 2. *(i) For each $t = 2, ..., T$, the joint probability $P(d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{V}_i) = P(d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i)$,*
(ii) For each $t = 2, ..., T$, $P(d_{it} = 1, d_{it-1} = 1|\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i) = G_{t,t-1}(\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i; \delta_{t0})$ for some known function $G_{t,t-1}(\cdot)$ and unknown parameter δ_{t0},
(iii) For each $t = 2, ..., T$ and all $\delta_t \in \mathcal{C}$, there exists a $\kappa \in \mathcal{R}^+$ such that $0 < \kappa \leq G_{t,t-1}(\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i; \delta_t) \leq 1$.

Part (i) is the missing at random (MAR) assumption, which specifies the joint probability as a function of only fully observed variables. Part(i) non-parametrically identifies the selection probabilities. Non-parametric estimation based on part (i) is seldom done in practice, even in the cross-sectional literature, and is less attractive in the panel setting as the curse of dimensionality is even more severe, with T times as many possible right hand side variables as possible in a cross-section[9] Part (ii) is a parametric specification for the joint probability of observation. Part (iii) is the strong overlap assumption which bounds the joint probability away from 0, as the weights used in (7) are the inverse of the probability. This condition guarantees that the moment function in (6) is well-defined.

Assumption (2) contains all the needed structure for the selection process. The formal assumptions in the next section fill in the necessary regularity conditions. The estimators and asymptotic theory are in terms of the parametric function $G_{t,t-1}(\cdot)$. The missing at random assumption allows selection to depend on time-invariant heterogeneity, though not in an arbitrary way. Selection may contain a correlated random effects structure (Mundlak (1978)) as in the simulations. In the simulations, a latent threshold crossing model for selection is used, as is common in the literature, to highlight the time dependence of the bivariate selection mechanism.

I propose two estimators based on the moments (7) and the selection mechanism assumption (2). The first is a pooled estimator which sums the moments over time, resulting in a just-identified system. The second is a GMM estimator which stacks the moments over time, thus taking advantage of over-identifying conditions (when $T > 2$). The optimally-weighted GMM estimator is more asymptotically efficient than the pooled estimator, though may perform worse in finite samples, a point explored in the simulations.

Define the following pooled IPW-POLS estimator:

[9] A notable exception is Chen, Hong, Tarozzi (Chen et al. (2008)), who, among other things, consider first-stage non-parametric estimation of the propensity score. Further extension of the proposed technique to non-parametric estimation of the joint probability of selection is left to future work.

$$\hat{\beta}_{POLS} = (\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_{t,t-1}(Z_{it,it-1};\hat{\delta}_t)} \begin{pmatrix} \Delta X'_{it} \\ \Delta W'_{it} \end{pmatrix} (\Delta X_{it}\, \Delta W_{it}))^{-1} \sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_{t,t-1}(Z_{it,it-1};\hat{\delta}_t)} \begin{pmatrix} \Delta X'_{it} \\ \Delta W'_{it} \end{pmatrix} \Delta y_{it} \tag{9}$$

where, for each $t = 2, ..., T$, $\hat{\delta}_t$ is a conditional MLE estimator of a bivariate binary selection model.

An alternative to the pooled OLS estimator defined above is to stack the moments from (7) and define a GMM estimator. The first step is still the estimation of the $\hat{\delta}_t$s from conditional MLE of a bivariate binary selection model for each $t = 2, ..., T$. Define the K vector $m_{it} = \dfrac{d_{it}d_{it-1}}{G_{t,t-1}(Z_{it,it-1};\hat{\delta}_t)} \begin{pmatrix} \Delta X'_{it} \\ \Delta W'_{it} \end{pmatrix} (\Delta y_{it} - \Delta X_{it}\beta_x - \Delta W_{it}\beta_w)$. Then the $K(T-1)$ vector of sample moments based on the population moments (7) can be written as

$$M_i = (m'_{i2}\, m'_{i3}...m'_{iT})' \tag{10}$$

$$\bar{M} = \frac{1}{N}\sum_i (m'_{i2}\, m'_{i3}...m'_{iT})' \tag{11}$$

The IPW-GMM estimator is defined as:

$$\hat{\beta}_{GMM} = \arg \min_{(\beta_x,\beta_w)} \bar{M}' A_N \bar{M} \tag{12}$$

where A_N is a positive definite matrix and $A_N \xrightarrow{p} A$, for A a positive definite weighting matrix.

2.1 Remark on other appraoches and transformations

The most common approach to missing values is use the complete cases in an unweighted estimator; that is, to use the complete cases as a valid sample. A natural question is, are the moments valid for the complete cases? Without ignorable missingness, the answer is generally no:

$$\begin{aligned} E[d_{it}d_{it-1}\Delta X'_{it}\Delta u_{it}] &= E[E[d_{it}d_{it-1}\Delta X'_{it}\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i, \boldsymbol{d}_i]] \\ &= E[d_{it}d_{it-1}\Delta x'_{it}E[\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i, \boldsymbol{d}_i]] \\ &\neq E[d_{it}d_{it-1}\Delta X'_{it}E[\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i]] = 0 \end{aligned} \tag{13}$$

Ignorable, or exogenous missingness is exactly the assumption that strict exogeneity holds conditional on the observation vector as well: $E[\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i, \boldsymbol{d}_i] = E[\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i] = 0$. In particular, when observation is directly related to the shocks u_{it} or the outcome y_{it}, ignorable missingness will fail, and the difference between $E[\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i, \boldsymbol{d}_i]$ and $E[\Delta u_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i, c_i](= 0)$ is the source of the inconsistency of the unweighted complete case estimator when ignorable missingness does not hold.

Other common approaches to estimation of a linear fixed effects model include the within transformation, and a Chamberlain-Mundlak correlated random effects structure for the c_i. Both the Mundlak-Chamberlain device and the within transformation introduce the entire time history of the covariates. The estimating equations will thus be fully observed only when the covariates are observed for all time periods. The effective sample size is $N_{obs} = \sum_i \prod_t d_{it}$, and any agent i with $d_{it} = 0$ for some t will be dropped from analysis. Using moments which involve a strict subset of the total time periods available allows for use of more of the data, even within complete case analysis, as the complete cases for each moment function depend on the time periods included in that moment and may include observations with missing values in other time periods.

The strong overlap condition in assumption (2) is another opportunity to see the benefits of moments which are a function of as few time periods as possible since $P(d_{i1} = 1, ..., d_{iT} = 1) < P(d_{it} = 1, d_{it-1} = 1)$ and inverse probability weighting is known to suffer as κ decreases. Furthermore, first-differencing can be used, with modification, for sequential exogeneity and dynamic models, a point revisited later.

3 Asymptotic Theory

This section presents the asymptotic theory for the IPW-POLS and the IPW-GMM estimators developed in the previous section. The arguments of the previous section are formalized in the following assumptions and theorems. The limiting behavior is for $N \to \infty$ and finite T.

The general properties of two-step estimators for large N and fixed T are known (see Newey and McFadden (1994)). The cross-sectional two-step IPW estimator where the first step is conditional maximum likelihood estimation from a binary selection model has been extensively studied (Wooldridge (2002), Wooldridge (2007), Graham et al. (2012)). Extending the IPW machinery to a panel model, and thus consistently estimating slope parameters in the presence of missing at random covariate data, is the main contribution of this paper. The presence of multiple time periods has some notable consequences for the IPW estimator. One notable consequence (see corollary (1)) is that it is no longer necessarily true that using the estimated probabilities yields a more efficient estimator than using the true probabilities, as is true in the cross-section (Wooldridge (2002), Wooldridge (2007)). One notable distinction in the mechanics of the estimator is that selection is no longer a single binary response, but rather a bivariate binary response due to the time differencing used to eliminate the fixed effects. Extending selection to a bivariate binary response model for the first stage estimation is a technical contribution to the IPW literature.

For each t, let $L_t = (X_t, W_t)$, where the i subscript is omitted. Recall that bold letters denote time histories; e.g. $\boldsymbol{L} = \{L_{i1}, L_{i2}, ..., L_{iT}\}$. The following assumptions will be adopted throughout:

Assumption 1. *(i) For each time period t=1,...,T, the random variables $y_t \in \mathbb{R}$, and $L_t \in \mathbb{R}^K$ (vector) have finite first and second moments (finite means, variances, and covariances.) Second moment for the vector L_t is $E[L_t' L_t]$ (as opposed to component-wise.) For each i, c_i is a random draw from an unknown distribution $\mathcal{F}_c$ with finite first and second moments. (ii) For each time period t, the random variables y_t, L_t, and c have finite third and fourth moments.*

Assumption (1) is a standard assumption, which gives primitive conditions on the random variables sufficient for consistency ((1)(i)) and asymptotic normality ((1)(ii)) of the estimators [10].

Assumption 2. *$\beta_0 \in \mathcal{B}$, a compact subset of $\mathbb{R}^k$.*

Compactness is not needed for the linear model, but is innocuous in applications and simplifies the analysis.

Assumption 3. *(Strict exogeneity)* $E[u_t|X_1, ..., X_T, W_1, ..., W_T, c] := E[u_t|L_1, ..., L_T, c] = 0$ *for all t.*

[10] Finite variance and higher moments give tractable primitive conditions, but are stronger conditions than the weakest possible conditions necessary for consistency and asymptotic normality, respectively (see Newey and McFadden (1994).)

The main model of interest is the classic linear fixed effects model (2), under strict exogeneity, assumption(3). The next section shows how the estimation strategy can be readily adapted to handle sequential exogeneity, as well as other models of interest.

Let Z_t be a vector-valued function of the fully observed random variables, and z_t a realization. For example, Z_t may equal $(y_t, y_{t-1}, \bar{y}, W_t, W_{t-1}, W_t^2, V_t)$. As discussed in the previous section, V_t may include time invariant random variables, or time-varying variables that affect selection but are conditionally unrelated to the outcome variable y_t, or V_t may be empty. Z_t are the predictors of the joint probability of observation, as formalized below.

Definition 1. *Let d_t be an indicator which takes value 1 when X_t is fully observed. For each $t = 2, ..., T$, let $p_{t,t-1} = P(d_t = 1, d_{t-1} = 1|\boldsymbol{X}, \boldsymbol{Z})$, the joint (conditional) probability of observing X in periods t and $t-1$.*

The next three assumptions repeat the selection mechanism assumptions gathered in the previous section, and are repeated here for completeness.

Assumption 4. *(MAR) $P(d_t = 1, d_{t-1} = 1|\boldsymbol{X}, \boldsymbol{Z}) = P(d_t = 1, d_{t-1} = 1|\boldsymbol{Z})$*

Assumption 5. *(Parametric specification) For each $t, t-1$, there exists a vector $\delta_{t0} \in \mathbb{R}^{|\delta_{t0}|}$ such that $p_{t,t-1} = G_{t,t-1}(z_t, \delta_{t0})$ for a known function G and unknown parameter δ_{t0}.*

Assumption 6. *(Strong overlap) For all t, and all z_t, δ_t in their supports, $p_{t,t-1}(z_t, \delta_t) \geq \kappa > 0$ for some $\kappa \in \mathbb{R}$.*

Assumption (5) allows for a different parameter δ_{t0} in each $(t, t-1)$ for $t = 2, ..., T$. While that is the preferred specification in the interest of accurately predicting $p_{t,t-1}$, with many time periods, this may result in too severe a loss of degrees of freedom. In which case, it may be more tractable to adopt a pooled specification of the form $G_{t,t-1}(z_t, \delta_0)$. The analysis carries through with a slight change to the form of the asymptotic variance, as will be noted.

The previous three assumptions are the main identifying assumptions of the paper. Assumption (4) is the missing at random assumption, which states that the joint probability of observing X is unrelated to the values of X, conditional on the fully observed variables Z. Assumption (6) uniformly bounds the probability of selection away from zero so that the inverse probability is well-defined and well-behaved in the limit. Assumption (5) assumes a correctly specified parametric form for the joint probability of observation, thus ensuring that the joint probability can be consistently estimated.

The remaining assumptions fill out the technical details needed for identification of β_0, and the limiting behavior of the estimators. The assumptions are standard rank ((7), (8)) and smoothness ((9), (10)) assumptions. A brief discussion of how the technical assumptions are used follows each assumption, while the full proofs are in the appendix. The uninterested reader may skip to the next subsection.

Assumption 7. *$E[\Delta L_t' \Delta L_t]$ is non-singular for $t = 2, ..., T$.*

Assumption 8. *$E[\sum_t \frac{d_t d_{t-1}}{p_{t,t-1}} \Delta L_t \Delta L_{t-1}']$ is non singular.*

Assumption (7) is the standard rank condition for the first-differenced estimator. Assumption (8) says that there is enough variation in the fully observed population so that the rank condition still holds.

An entire joint distribution for the random variables (d_t, d_{t-1}) conditional on z_t is specified, as detailed in assumption (9) and (10).

Assumption 9. *(Parametric specification) For each $t, t-1$, $P(d_t = d, d_{t-1} = d'|z_t) = G_{t,t-1}(d, d'|z_t; \delta_{t0})$ (where $d, d' \in \{0, 1\}$) such that:*
(i) $\delta_{t0} \in \mathcal{D}$, a compact subset of $\mathbb{R}^{|\delta_{t0}|}$,
(ii) $\exists z_t \in support(Z_t)$ with $P(d_t = d, d_{t-1} = d'|z_t) > 0$ such that $P(d_t = d, d_{t-1} = d'|z_t) \neq G_{t,t-1}(d, d'|z_t; \delta_t)$ for $\delta_t \neq \delta_{t0}$,
(iii) $G_{t,t-1}$ is continuous in δ for each d, d', z_t,
(iv) $E[sup_\delta |ln(G_{t,t-1}(d, d', z_t; \delta_t))|] < \infty$.

Assumption (9) guarantees that the conditional maximum likelihood estimator (CLME) is consistent for δ_{t0} for each t (see theorem 13.1 in Wooldridge Wooldridge (2010b), or theorem 2.5 in Newey, McFadden Newey and McFadden (1994)).

The δ_{t0} are estimated via conditional maximum likelihood; i.e. $\hat{\delta}_t$ is the solution to:

$$\sum_i s_{t,t-1}(z_{it};\delta_t) = 0 \tag{14}$$

$$s_{t,t-1} = \nabla_{\delta_t}\mathcal{L}(\delta_t) \tag{15}$$

for each $t = 2, ..., T$, where $s_{t,t-1}$ is the score of the log likelihood of the bivariate binary selection model. The log likelihood is:

$$\begin{aligned}\mathcal{L}(\delta_t) = \sum_i d_{it}d_{it-1}ln(G_{t,t-1}(1,1|z_{it};\delta_t))\\ + (1-d_{it})d_{it-1}ln(G_{t,t-1}(0,1|z_{it};\delta_t))\\ + d_{it}(1-d_{it-1})ln(G_{t,t-1}(1,0|z_{it};\delta_t))\\ + (1-d_{it})(1-d_{it-1})ln(G_{t,t-1}(0,0|z_{it};\delta_t))\end{aligned}$$

The addition of the next assumption fills in the technical details that yield asymptotic normality and efficiency of the CMLE estimator for each t (see theorem 13.2 in Wooldgridge Wooldridge (2010b), or theorem 3.4 in Newey, McFadden Newey and McFadden (1994)).

Assumption 10. *(Smoothness) For each* $t, t-1$, $P(d_t = d, d_{t-1} = d'|z_t) = G_{t,t-1}(d, d'|z_t;\delta_{t0})$ *such that:*
(i) $\delta_{t0} \in int(\mathcal{D})$,
(ii) In a neighborhood of δ_{t0}, $G_{t,t-1}(d, d', z_t;\delta_t)$ *is twice continuously differentiable in* δ *and* $G_{t,t-1}(d, d', z_t;\delta_t) > 0$,
(iii) $\int sup_\delta||\nabla_\delta G_{t,t-1}(d, d'|z_t;\delta_t)||dz < \infty$ *and* $\int sup_\delta||\nabla_{\delta\delta} G_{t,t-1}(d, d'|z_t;\delta_t)||dz < \infty$,
(iv) $E[\nabla_\delta G_{t,t-1}(d, d'|z_t;\delta_{t0})\{\nabla_\delta G_{t,t-1}(d, d'|z_t;\delta_{t0})\}']$ *exists and is non-singular.*

Conditions (i) and (ii) guarantee that the score (the partial of the log-likelihood $lnG_{t,t-1}(d1, d2|z_t;\delta_t)$ w.r.t δ) admits a mean-value expansion around the true δ_{t0}. Conditions (iii) and (iv) guarantee the exchange of integration (summation w.r.t. distribution of (d_t, d_{t-1})) and differentiation w.r.t δ are allowed. This yields the results that the score has $E[s_t|z_t] = 0$, and the conditional information matrix equality (CIME) can be established for each t. Condition (iv) guarantees that the sample scores evaluated at the true δ_{t0} are root-n asymptotically normal (as the asymptotic variance is given by the inverse of (iv)) and condition (iii) plus assumption (1) gives uniform convergence of the sample Hessian.

3.1 IPW-POLS estimator

This section establishes identification of β_0, as well as consistency and the asymptotic distribution of the IPW-POLS estimator.

From assumptions (2) and (3), $E[\Delta L_t' \Delta y_t] = E[\Delta L_t' \Delta L_t]\beta$. From assumption (7), β_0 is then uniquely identified in the population as $\beta_0 = E[\Delta L_t' \Delta L_t]^{-1} E[\Delta L_t' \Delta y_t]$ or $\beta_0 = E[\sum_t \Delta L_t' \Delta L_t]^{-1} E[\sum_t \Delta L_t' \Delta y_t]$. Unfortunately, the random sample from the population is subject to missing values. Define the *observable population* as the population defined by the random variables $d_t(y_t, L_t)$. It remains to show that β_0 is identified in the observable population.

Theorem 1. *(Identification) Under assumptions* (1), (2), (3), (4), (5), (6), *and* (8)*(a),* β_0 *is uniquely identified in the observable population:*

$$\beta_0 = E[\sum_t \frac{d_t d_{t-1}}{p_{t,t-1}(z_t, \delta_{t0})} \Delta L_t' \Delta L_t]^{-1} E[\sum_t \frac{d_t d_{t-1}}{p_{t,t-1}(z_t, \delta_{t0})} \Delta L_t' \Delta y_t] \tag{16}$$

Theorem 2. *(Consistency of IPW-POLS) Under assumptions* (1), (2), (3), (4), (5), (6), (8), *and* (9),

$$\hat{\beta}_{IPW-POLS} \xrightarrow{p} \beta_0 \tag{17}$$

The detailed proof is presented in the appendix, and a sketch of the main points is given below.

Proof sketch.

$$\hat{\beta}_{POLS} - \beta_0 = (\frac{1}{N} \sum_i \sum_{t=2} \frac{d_{it} d_{it-1}}{G_{t,t-1}(Z_{it}; \hat{\delta}_t)} \left(\Delta L_{it}'\right) (\Delta L_{it}))^{-1} \frac{1}{N} \sum_i \sum_{t=2} \frac{d_{it} d_{it-1}}{G_{t,t-1}(Z_{it}; \hat{\delta}_t)} \left(\Delta L_{it}'\right) \Delta u_{it}$$

Assumption (9) gives $\hat{\delta}_t \xrightarrow{p} \delta_{t0}$ for each t, from known MLE results. The regularity conditions asume enough smoothness to apply the WLLN and continuous mapping theorem. $\hat{\beta}_{IPW-POLS} - \beta_0 \rightarrow E(\sum_{t=2} \frac{d_{it} d_{it-1}}{G_{t,t-1}} \Delta L_{it}' (\Delta L_{it})^{-1} E[\sum_{t=2} \frac{d_{it} d_{it-1}}{G_{t,t-1}} \Delta L_{it}' \Delta u_{it}]$. The result then follows from lemma (1). □

Theorem 3. *Under assumptions (1), (2), (3), (4), (5), (6), (8), (9), and (10),*

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma) \tag{18}$$

where $\Sigma = A_0^{-1} B_0 A_0^{-1}$ *with*

$$A_0 = E[\sum_t \Delta L_t' \Delta L_t] \tag{19}$$

$$B_0 = E[(\sum_t r_{it})(\sum_t r_{it}')] \tag{20}$$

$$r_{it} = m_{it} + C_{t0} M_t^{-1} s_{it} \tag{21}$$

$$m_{it} = \frac{d_{it} d_{it-1}}{G_{t,t-1}} \Delta L_{it}' \Delta u_{it} \tag{22}$$

$$C_{t0} = E[\nabla_\delta m_t(\delta_{t0}, \beta_0)] \tag{23}$$

$$M_t = E[s_t s_t'] \tag{24}$$

and s_{it} *is the* $|\delta_t|$ *by* 1 *score vector of the bivariate binary selection log-likelihood (see* (15)*,) evaluated at the true* δ_{t0}*.*

The subscript t is used instead of $t, t-1$ for clarity of notation, but note that all of the objects in (3) are functions of (at least) two time periods.

In the case of a pooled specification, the above expressions are valid with C_{t0} replaced by $C_0 = \sum_t E[\nabla_\delta m_t(\delta_0, \beta_0)]$, and M_t replaced by $M = \sum_t E[s_t s_t']$.

A notable consequence of the form of the asymptotic variance in Theorem (3), is that the well known result from cross-sectional IPW analysis that using the estimated probabilities of observation is asymptotically more efficient than if the true probabilities were used (Robins et al. (1994), Wooldridge (2002), Wooldridge (2007)) does not, in general, carry over to the panel setting:

Corollary 1. *Consider*

$$\tilde{\beta} = (\sum_i \sum_{t=2} \frac{d_{it} d_{it-1}}{G_{t,t-1}} \left(\Delta L_{it}'\right) (\Delta L_{it}))^{-1} \sum_i \sum_{t=2} \frac{d_{it} d_{it-1}}{G_{t,t-1}} \left(\Delta L_{it}'\right) \Delta y_{it} \tag{25}$$

which uses the true joint probabilities of observation $G_{t,t-1}$*. For* $T = 2$*,* $avar(\tilde{\beta}) \geq avar(\hat{\beta}_{IPW-POLS})$*.*
For $T > 2$*,* $avar(\hat{\beta}_{IPW-POLS})$ *could be larger or smaller (in a positive definite sense) than* $avar(\tilde{\beta})$*.*

where $avar(\cdot)$ is asymptotic variance and a positive definite matrix A is said to be smaller than a positive definite matrix B if the matrix B-A is positive definite.

Proof sketch.

$$\sqrt{N}(\tilde{\beta}-\beta_0) \xrightarrow{p} \mathcal{N}(0, A_0^{-1}\tilde{B}_0 A_0^{-1})$$

where $\tilde{B}_0 = E[(\sum_t m_t)(\sum_t m_t')]$ where A_0 and m_t are as in theorem (3).

$$\begin{aligned}
E[(\sum_t r_{it})(\sum_t r_{it}')] &= E[(\sum_t m_{it})(\sum_t m_{it'})] + E[(\sum_t m_{it})(\sum_t s_{it}'M_t^{-1}C_{t0}')] \\
&\quad + E[(\sum_t C_{t0}M_t^{-1}s_{it})(\sum_t m_{it}')] + E[(\sum_t C_{t0}M_t^{-1}s_{it})(\sum_t s_{it}'M_t^{-1}C_{t0}')] \\
&= E[(\sum_t m_{it})(\sum_t m_{it'})] + \sum_t E[m_t s_t']M_t^{-1}C_{t0}' + \sum_t C_{t0}M_t^{-1}E[s_t m_t] + \sum_t C_{t0}M_t^{-1}E[s_t s_t'] \\
&\quad + \sum_t \sum_{\tau\neq t} E[m_t s_\tau]M_\tau^{-1}C_{\tau 0}' + \sum_t \sum_{\tau\neq t} C_{t0}M_t^{-1}E[s_t m_\tau'] + \sum_t \sum_{\tau\neq t} C_{t0}M_t^{-1}E[s_t s_\tau']M_\tau \qquad (26) \\
&= E[(\sum_t m_{it})(\sum_t m_{it'})] - \sum_t C_{t0}M_t^{-1}C_{t0} \qquad (27) \\
&\quad + \{t, \tau\, terms\}
\end{aligned}$$

Where (27) follows from::

$$\begin{aligned}
C_{t0} = E[\nabla_\delta m_{it}] &= -E[\frac{d_t d_{t-1}}{G_{t,t-1}^2}\Delta L_t' \Delta u_t \nabla_\delta G_{t,t-1}] \\
E[m_{it}s_{it}] &= E[\frac{d_t d_{t-1}}{G_{t,t-1}}\Delta L_t' \Delta u_t \frac{d_t d_{t-1}}{G_{t,t-1}}\nabla_\delta G_{t,t-1}] \\
&= -C_{t0} \qquad (28)
\end{aligned}$$

which says that, for a given t, when evaluated at the truth, the partial of the FOC for β w.r.t the nuisance parameter δ is the negative of the correlation between the FOC for β and the FOC for δ_t. This has a similar flavor as the conditional information matrix equality (CIME) for maximum likelihood estimation. Thus, for a given t,

$$E[r_{it}r_{it'}] = E[m_{it}m_{it'}] - C_{t0}M_t^{-1}C_{t0}$$

the second term is positive definite and thus $E[m_{it}m_{it'}] - E[r_{it}r_{it'}] = C_{t0}M_t^{-1}C_{t0} \geq 0$. This is where the result comes from in the cross-sectional case (and when T=2). In the panel setting with $T > 2$, while we can conclude that $\sum_t C_{t0}M_t^{-1}C_{t0} \geq 0$, no such conclusion can be made about the t, τ cross terms (26). □

The cross-sectional result has been shown for univariate selection, and carries over to the case of bivariate selection when the first-differenced model is just-identified. The second part of corollary (1) marks a stark departure from the previous results for two-step IPW estimators with first stage CMLE. The intuition for this is similar to the intuition for why partial maximum likelihood estimation is not efficient (see 13.8 in Wooldridge (2010b).) While a correct specification of the distribution of (d_t, d_{t-1}) was assumed for each $t, t-1$, the correct distribution for $\boldsymbol{d} = (d_1, ..., d_T)$ is not specified, and thus, while the CIME holds in a given $(t, t-1)$, it does not hold generally. Even if we adopt the strong assumption that the scores are serially uncorrelated (and thus recover the conditional information matrix equality for maximum likelihood, so the third term in (26) is zero), we are still left with the failure of an equality like (28) for the t, τ cross terms.

3.2 IPW-GMM estimator

Now consider the IPW-GMM estimator defined by (12). Consistency and the asymptotic distribution are combined in the following theorem:

Theorem 4. *(1), (2), (3), (4), (5), (6), (8), (9), and (10), for the IPW-GMM estimator defined by (12) with Σ_1 a positive definite weight matrix,*

$$\sqrt{N}(\hat{\beta}_{GMM} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_2) \tag{29}$$

where $\Sigma_2 = (A_{00}'\Sigma_1 A_{00})^{-1} A_{00}'\Sigma_1 \Omega \Sigma_1 A_{00}'(A_{00}' W A_{00})^{-1})$ *with*

$$A_{00} = (A_{20}' A_{30}' ... A_{T0}')' \tag{30}$$

$$A_{t0} = E\left(\frac{d_{it} d_{it-1}}{G_{t,t-1}(z_t; \delta_{t0})} \Delta L_t' \Delta L_t\right) \tag{31}$$

$$\Omega = E[M_i M_i'] - C_0 E[s_i s_i']^{-1} C_0' \tag{32}$$

$$C_0 = (C_{20}, C_{30}, ..., C_{T0}) \tag{33}$$

$$M_i = (m_{i2}' \, m_{i3}' ... m_{iT}')' \tag{34}$$

$$s_i = (s_{i1,i2}', s_{i2,i3}', ..., s_{iT-1,iT}')' \tag{35}$$

As with the IPW-POLS estimator, the accounting for the first step estimation in the asymptotic variance is in the Ω as neither the A_{00} nor the weight matrix Σ_1 involve the first step estimation. Notice that A_0 from the IPW-POLS estimator's asymptotic variance is $A_0 = \sum_t A_{t0}$.

The optimally-weighted GMM estimator uses a consistent estimate of Ω^{-1} for the weighting matrix. With $\Sigma_1 = \Omega^{-1}$, the asymptotic variance is given in the next corollary:

Corollary 2. *For* $\Sigma_1 = \Omega^{-1}$,

$$\sqrt{N}(\hat{\beta}_{GMMop} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{op}) \tag{36}$$

where $\Sigma_{op} = (A_0'\Omega^{-1}A_0)^{-1}$.

The IPW-GMM estimator defined by (12) with the optimal weight matrix will have smaller asymptotic variance than the IPW-POLS estimator. That the optimally weighted IPW-GMM estimator has lower asymptotic variance than the IPW-POLS estimator can be seen from observing that the IPW-POLS estimator is an IPW-GMM estimator with weight matrix $\Sigma_1 = \iota'\iota$ where ι is a $T-1$ vector of 1s. The result then follows by standard GMM theory and is summarized in the following corollary:

Corollary 3.

$$avar(\hat{\beta}_{GMMop}) \leq avar(\hat{\beta}_{POLS}) \tag{37}$$

The following section gives some finite sample results where the IPW estimators are compared to the unweighted estimator. The trade-offs between the IPW-POLS and IPW-GMM estimators also are explored.

4 Simulations

This section presents some finite sample performance of the IPW-POLS and IPW-GMM estimators developed in the previous sections. The outcome model of interest is:

$$y_{it} = X_{it}\beta_x + W_{it}\beta_w + c_i + u_{it} \tag{38}$$

which is specified as:

$$y_{it} = x_{it} + \alpha_t + w_{it} + c_i + u_{it} \tag{39}$$

for all the simulations. The missing values are in x_{it} and $\beta_{x0} = \beta_{w0} = 1$. The time intercepts α_t are set to $\alpha_t = t$, though the simulations are not sensitive to these values, and the estimation of $\alpha_t - \alpha_{t-1}$ is often not of paramount interest. The unobserved time-invariant heterogeneity c_i is specified as $c_i = 0.3\bar{x}_i + 0.2\bar{w}_i + 0.5b_i$ where $b_i \sim Bernoulli(0.6)$, though the results are not sensitive to other specifications of the c_i. Other results, as well as reproducible code, are available from the author upon request. All simulations are done in R.

4.1 Example of MAR assumption

Assumption (2) includes the missing at random assumption, as well as a parametric assumption on the form of the joint probability of observation. As an example of a data generating process which satisfies the assumptions, a latent variable threshold crossing mechanism for selection in each time period is adopted. A common specification is:

$$\begin{aligned} d^*_{it} &= \pi_t + y_{it}\pi_{yt} + w_{it}\pi_{wt} + V_{it}\pi_{vt} + \xi_i - \nu_{it} \qquad (40)\\ \nu_{it} &\perp (\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{V}_i)\\ d_{it} &= \mathbf{1}[d^*_{it} > 0] \qquad (41) \end{aligned}$$

The time invariant ξ_i allows for unobserved individual heterogeneity in the propensity of observation, such as worker productivity, a respondent's preference for privacy, or unobserved components of the house for sale. Since estimation of the fitted probabilities is required, the unobserved ξ_i cannot be left fully flexible. An additional assumption, which may be application specific, is needed. A common technique, and the one adopted here, is a Chamberlain-Mundlak-Wooldridge correlated random effects structure (Mundlak (1978), Chamberlain (1984), Wooldridge (2010b)):

$$\begin{aligned} \xi_i &= \gamma + \bar{y}_i\gamma_y + \bar{w}_i\gamma_w + \bar{V}_i\gamma_v - a_i \qquad (42)\\ a_i &\perp (\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i)\\ a_i &\not\perp \nu_{it} \end{aligned}$$

Models (40) and (42) can be subsumed as a specific example of the general latent variable model:

$$\begin{aligned} d_{it}^* &= Z_{it}\Pi_t - \eta_{it} \\ \eta_{it} &\perp (\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{V}_i) \\ d_{it} &= \mathbf{1}[d_{it}^* > 0] \end{aligned} \tag{43}$$

where Z_{it} is a vector of functions of the $\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i$s that may vary with t. The slope parameters Π_t are allowed to be time period specific.

The missing at random assumption (2)(i) is imposed through the exclusion of $\boldsymbol{X}_i$ in Z_{it}, and the independence of the error terms η_{it} and $\boldsymbol{X}_i$. The failure of ignorable missingness is in the inclusion of functions of $\boldsymbol{y}_i$ in the Z_{it}. Additionally, though unrelated to the selection assumption (2), serial correlation is allowed in the η_{it}.

Given the latent variable model (43), the conditional joint probability of observation of d_{it} and d_{it-1} is:

$$\begin{aligned} P(d_{it} = 1, d_{it-1} = 1 | \boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i) &= P(d_{it}^* > 0, d_{it-1}^* > 0 | \boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i) \\ &= P(\eta_{it} < Z_{it}\Pi_t, \eta_{it-1} < Z_{it-1}\Pi_{t-1} | \boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i) \\ &= P(\eta_{it} < Z_{it}\Pi_t, \eta_{it-1} < Z_{it-1}\Pi_{t-1}) \\ &= F_{\eta_{it},\eta_{it-1}}(Z_{it}\Pi_t, Z_{it-1}\Pi_{t-1}) \end{aligned} \tag{44}$$

The parametric assumption (2)(ii) is then

$$F_{\eta_{it},\eta_{it-1}}(Z_{it}\Pi_t, Z_{it-1}\Pi_{t-1}) = G_{t,t-1}(Z_{it}, Z_{it-1}; \delta_{t0}) \tag{45}$$

where $\delta_{t0} = (\Pi_t' \quad \Pi_{t-1}' \quad \rho_t)$, with $\rho_t = corr(\eta_{it}, \eta_{it-1})$. The dependence of the slope parameters Π_t on t is extremely flexible, but requires not only finite T but relatively small T. With a very small T, a more flexible Chamberlain assumption, $Z_{it} = (\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{V}_i)$ with time varying coefficients, may be tractable. With a large T, it may be more tenable to assume time invariant coefficients in a pooled binary choice model, $d_{it}^* = Z_{it}\Pi - \eta_{it}$.

For the simulations, $Z_{it} = (1, y_{it}, w_{it}, \bar{y}_i, \bar{w}_i, v_i)$. The structure of the error term varies by simulation as discussed below.

4.2 Finite sample performance

In each simulation, the probability of observation for a single time period is approximately 0.72, the joint probability for two time periods is 0.55, meaning the proposed

IPW estimators are using more than half of the data [11]. The estimators considered are all first difference estimators; the oracle full data estimators, the first difference complete cases, and the proposed IPW estimators using both the true probabilities, and the estimated probabilities where the estimation is from a correctly specified bivariate probit model.

The simulation set-ups detailed above are summarized below.

Simulations 1 and 2: Outcome model

$$
\begin{aligned}
y_{it} &= t + x_{it} + w_{it} + c_i + u_{it} \qquad t = 1, 2, 3 \\
&(u_{it}, x_{it}, w_{it}) \sim \mathcal{N}(0, \sigma^2) \\
&corr(x_{it}, x_{it-1}) = 0.5,\ corr(w_{it}, w_{it-1}) = 0.7 \\
&corr(x_{it}, w_{it}) = -0.35 \\
c_i &= 0.3 * \bar{w}_i + 0.2 * \bar{x}_i + 0.5 v_i \\
v &\sim Bernoulli(0.6)
\end{aligned}
$$

Simulation 3: Outcome model

$$
\begin{aligned}
&y_{it} = \alpha_t + x_{it}\beta_x + w_{it}\beta_w + c_i + u_{it} \qquad t = 1, 2, 3 \\
&u_{it} \sim \mathcal{N}(0, \Sigma) \\
&\quad var(u_{it}) = t \\
&\quad corr(u_{i1}, u_{12}) = 0.3\ corr(u_{i1}, u_{i3}) = 0.15\ corr(u_{i2}, u_{i3}) = 0.25 \\
&(x_{it}, w_{it}) \sim \mathcal{N}(0, \sigma^2) \\
&\quad corr(x_{it}, x_{it-1}) = 0.5,\ corr(w_{it}, w_{it-1}) = 0.7 \\
&\quad corr(x_{it}, w_{it}) = -0.35 \\
&\ c_i = 0.3 * \bar{w}_i + 0.2 * \bar{x}_i + 0.5 v_i \\
&\ v \sim Bernoulli(0.6)
\end{aligned}
$$

[11] The joint probability for all three periods is roughly 0.4, so a full complete case estimator would use 40% of the data.

Simulation 1 and 3: Selection mechanism

$$
\begin{aligned}
d_{it}^{*} &= Z_{it}\pi_t - \eta_{it} \\
Z_{it} &= (1, y_{it}, w_{it}, \bar{y}_i, \bar{w}_i, v_i) \\
d_{it} &= \mathbf{1}[d_{it}^{*} > 0] \\
\eta_{it} &\sim \mathcal{N}(0,1) \\
corr(\eta_{it}, \eta_{it-1}) &= 0.5
\end{aligned}
$$

Simulation 2: Selection mechanism (Serial correlation and correlation with outcome error)

$$
\begin{aligned}
d_{it}^{*} &= Z_{it}\pi_t - \eta_{it} \\
Z_{it} &= (1, y_{it}, w_{it}, \bar{y}_i, \bar{w}_i, v_i) \\
d_{it} &= \mathbf{1}[d_{it}^{*} > 0] \\
\eta_{it} &= 0.2u_{it} + e_{it} \\
e_{it} &\sim \mathcal{N}(0,1) \\
corr(e_{it}, e_{it-1}) &= 0.5 \quad corr(v_{it}, v_{it-2}) = 0.3
\end{aligned}
$$

The main take-away from the simulations is that with missing at random covariate values where the missingness depends on the outcome, the unweighted estimator (Complete) preforms poorly, with a bias around 25% and rmse over 50% larger than that of the IPW estimators.

The simulations allow some comparison of the two proposed estimators. The GMM estimator in simulations 1 and 2 is a one-step estimator with the identity weight matrix. Simulation 3 uses that one-step GMM, optimally weighted GMM using the one-step estimator to calculate the weighting matrix, and the POLS estimator. The optimal weighting matrix may perform quite poorly in finite samples. In simulations 1 and 2, the error u_{it} was homoskedastic (and independent of the regressors.) In this case, we expect POLS to preform fine. In simulation 3, I introduce time heteroskedasticity and serial correlation. Results look like with total independence of errors, POLS has lower variance, with just time heterogeneity (results not shown, 100 replications run) still looks like POLS might perform better than both GMM estimators (in terms of variance), though the optimal GMM estimator has lower variance than the one-step estimator. With time heterogeneity and serial correlation, as in simulation 3, there's some variance improvement for optimal GMM over POLS, though it seems to come at the cost of higher bias.

Table 1: Simulation 1: N=1,000, $N_{obs} \sim 545$, R=500, $\beta_x = \beta_w = 1$

Estimation	$E[\hat{\beta_x}]$	rmse	$var(\hat{\beta_x})$	$E[\hat{\beta_w}]$	rmse	$var(\hat{\beta_w})$
Full GMM	0.998	0.054	0.003	0.996	0.072	0.006
Full POLS	0.999	0.056	0.003	0.997	0.073	0.005
Complete GMM	0.738	0.228	0.008	0.814	0.186	0.012
Complete POLS	0.750	0.209	0.007	0.829	0.163	0.011
IPW-GMM -p_o	1.036	0.123	0.016	1.022	0.152	0.023
IPW-POLS -p_o	1.032	0.116	0.013	1.018	0.139	0.019
IPW-GMM - $\hat{p}$	0.937	0.125	0.012	0.958	0.143	0.019
IPW-POLS - $\hat{p}$	0.938	0.115	0.010	0.960	0.132	0.016

Table 2: Simulation 2: N=1,000, $N_{obs} \sim 545$, R=500, $\beta_x = \beta_w = 1$

Estimation	$E[\hat{\beta_x}]$	rmse	$var(\hat{\beta_x})$	$E[\hat{\beta_w}]$	rmse	$var(\hat{\beta_w})$
Full GMM	1.0001	0.058	0.003	0.9999	0.074	0.005
Full POLS	1.0009	0.056	0.003	0.9995	0.073	0.005
Complete GMM	0.768	0.244	0.008	0.839	0.187	0.012
Complete POLS	0.782	0.229	0.008	0.858	0.167	0.011
IPW-GMM -p_o	0.986	0.110	0.012	1.002	0.134	0.018
IPW-POLS -p_o	0.981	0.101	0.010	0.996	0.122	0.015
IPW-GMM - $\hat{p}$	0.964	0.112	0.012	1.004	0.129	0.017
IPW-POLS - $\hat{p}$	0.961	0.105	0.010	1.001	0.119	0.015

Table 3: Simulation 3: N=1,000, $N_{obs} \sim 545$, R=500, $\beta_x = \beta_w = 1$

Estimation	$E[\hat{\beta_x}]$	rmse	$var(\hat{\beta_x})$	$E[\hat{\beta_w}]$	rmse	$var(\hat{\beta_w})$
Full GMM	1.001	0.065	0.004	0.999	0.082	0.0067
Full opt GMM	1.0015	0.0619	.00382	1.000	0.0786	0.0062
Full POLS	1.0015	0.0623	0.00388	0.9999	0.0797	0.0064
Complete GMM	0.717	0.294	0.0102	0.799	0.225	0.0137
Complete opt GMM	0.729	0.282	0.0104	0.820	0.203	0.0137
Complete POLS	0.729	0.282	0.0096	0.814	0.209	0.0130
IPW-GMM -p_o	0.930	0.146	0.017	0.956	0.161	0.024
IPW-opt GMM -p_o	0.913	0.143	0.013	0.942	0.150	0.020
IPW-POLS -p_o	0.924	0.139	0.014	0.949	0.151	0.021
IPW-GMM - $\hat{p}$	0.910	0.160	0.018	0.954	0.164	0.026
IPW-opt GMM - $\hat{p}$	0.900	0.150	0.0136	0.946	0.147	0.020
IPW-POLS - $\hat{p}$	0.907	0.150	0.0145	0.951	0.148	0.021

5 Empirical application: hedonic housing price model

5.1 Introduction

The empirical application revisits the setting of Kuwayama, Olmstead, and Zheng (Kuwayama et al. (2018)) and Zheng ((Zheng, 2017)) which considers the willingness to pay for water quality in local water amenities[12]. This is one of many settings in which the proposed estimation technique can be used; namely, the specification is a linear fixed effects model, there are missing covariate values, and, as will be shown, empirical evidence suggests that the missing values are not ignorable but can be missing at random.

Quantifying the value of a public good or natural resource is important for policy but difficult in practice as there is often not a market for such goods. The value of a public good or natural resource can be multi-faceted, with use value coming from recreational opportunities for example, but also passive value like existence or non-use value (Krutilla (1967), McConnell and Walls (2005).) Quantifying the value of water quality by looking at housing markets has a long history in the literature (Poor et al. (2007), Walsh et al. (2017), Keiser and Shapiro (2017),) and is relevant for policy as legislation like the Clean Water Act is costly. Willingness to pay is captured by a hedonic house price model, which specifies the price of a house as a function of the house attributes (Rosen (1974).) The attribute of interest is the water quality in nearby water amenities, which include lakes, rivers, ponds, and canals (Kuwayama et al. (2018).) Water quality is measured by the level of dissolved oxygen in mg/L, which is one of the common measures of ambient water quality (Kuwayama et al. (2018), Keiser and Shapiro (2017)) as aquatic species need dissolved oxygen to live, and the level of dissolved oxygen is correlated with other common measures of water quality, such as clarity (Kuwayama et al. (2018).)

A concern is that the wealth of an area may be related to the frequency of monitoring, regardless of the level of the water quality or the characteristics of the house. This would result in non-ignorable yet missing at random missingness[13] as defined in this paper.

[12] Kuwayama et al also consider the recreational benefit from improvements to a large, regional water body. While the recreational benefit is significant, the estimate of the willingness to pay for local, ambient water quality is largely unchanged by the inclusion or exclusion of the recreational measure in the specification (Kuwayama et al. (2018).) A specification with only the local water amenities is adopted here, though the results including a measure of the large, regional water body are similar and available upon request.

[13] This is different than the situation where the water quality is measured at a plant by electronic equipment that only records the level between some threshold (Muehlenbachs et al. (2015).) In that case, observation is not at random as it is directly related to the level of the missing variable.

There are various reasons to suspect that monitoring frequency is related to wealth of an area, and thus house prices. If areas with higher house prices have more political leverage, those areas may receive more frequent monitoring regardless of water quality. Conversely, if lower income areas are targeted as at risk, they may be monitored more closely. If monitoring, and thus observation, is conditionally related to house price then dropping observations with missing values could lead to inconsistent parameter estimates. Without exact information on the motivations for monitoring, the question is empirical.

There are many attributes of a property that affect its sale price, some of which are unobserved to the econometrician. Kuwayama, Olmstead, and Zheng (Kuwayama et al. (2018), hereafter KOZ) show that many observed attributes, like age, number of bedrooms, and distance to the water, are correlated with the dissolved oxygen level in local water amenities, and therefore it is suspected that unobservable attributes that affect price may also be correlated with dissolved oxygen level. That concern leads to the inclusion of time-invariant property fixed effects which affect sale price and may be correlated with the time-varying regressors; namely, dissolved oxygen and age. The resulting linear fixed effects specification for house price is in line with the theoretical derivations in the previous sections. Identification is off of repeated sales of the same house ((Livy et al., 2013).)

Access to panel data allows from some tests of the differing implications of the ignorable missingness assumption and the missing at random assumption. Evidence is presented that suggests that indeed observation is conditionally related to house price, contrary to the ignorable missingness assumption. Additionally, there is some empirical evidence to support the missing at random assumption; namely, that observation is not conditionally related to water quality.

5.2 Data

The data consist of single-family house sales in the Tampa Bay area, which consists of three counties; Hillsborough, Pinellas, and Manatee. Dissolved oxygen data is merged with sale price data to create an unbalanced panel of house sales over time. "Local" is defined as properties within 1 km of a monitoring station[14].

The data span from 1998 to 2014, during which time there were numerous policy

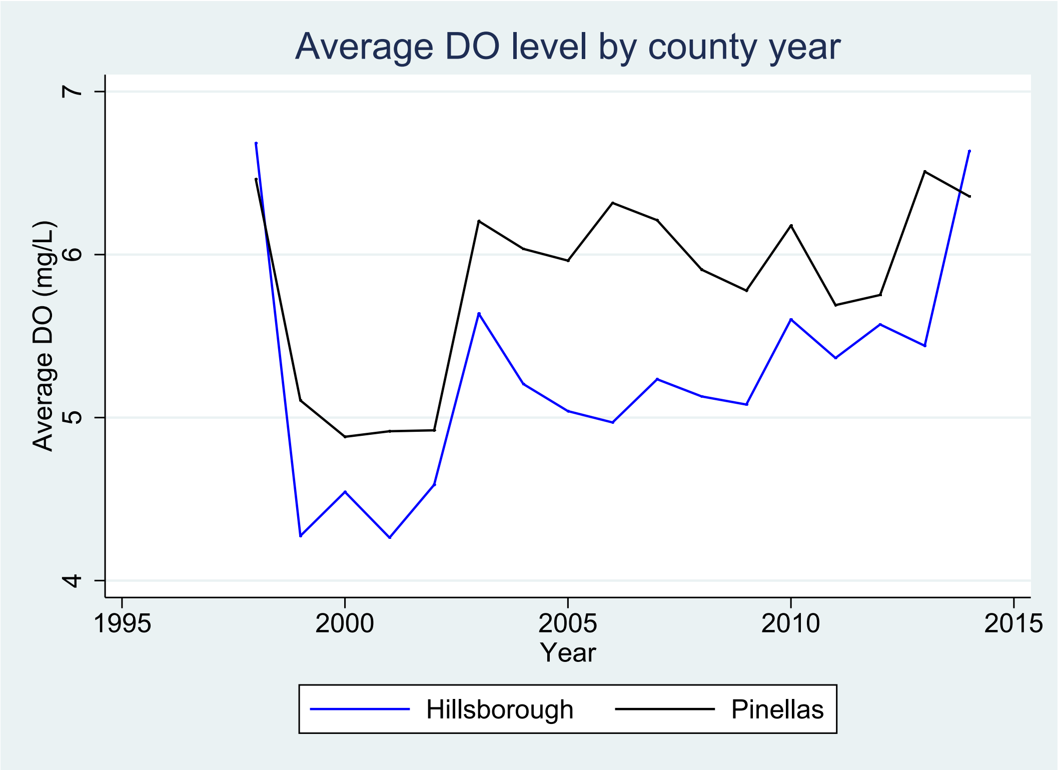
[14] The main specification in Kuwayama et al. (2018) uses a 3 km radius, though they include results for 500m and 1km. The 1 km radius used here is to focus on the immediate proximity, and follows the findings of Walsh et al. (2011) that both the magnitude and precision of the point estimate on willingness to pay (in their case for the water quality of lakes in the Orlando, FL area) decline beyond a radius of 1000m.

changes affecting water quality standards. While there are 17 years of data, each monitoring site has an average of 8 years of recorded measures of dissolved oxygen, resulting in an observation rate of roughly 45%. As many policy decisions are made at the county level, aggregating the results from the three counties may not be the most policy relevant approach. Much monitoring is conducted at the county level; for example, the Environmental Protection Commission of Hillsborough county oversees over 250 monitoring stations located within Hillsborough county (of Hillsborough County (2018).) Aggregating counties also may obfuscate important differences in the value of water quality across the counties. Walsh et. al. (Walsh et al. (2017)) consider 14 counties in the Chesapeake Bay area and find, among other things, heterogeneous returns to water quality across county. Hillsborough and Pinellas county, which are the principle counties in the Tampa-St. Pete MSA, have enough repeated observations to be considered separately. Manatee county, which is part of the North Port-Sarasota-Bradenton MSA and for which sales data is not available until 2005, does not have enough observations to analyze on its own. Results are presented separately for Hillsborough county, Pinellas county, and, as in KOZ, the three counties combined.

Figure (1) plots the average recorded dissolved oxygen level over time for Hillsborough and Pinellas counties. Table (4) presents summary statistics for Hillsborough and Pinellas counties. The three samples considered are the full sample of houses, the sub sample with at least one recording of dissolved oxygen, and the subsample with dissolved oxygen values across consecutive sales, which is the subsample used in first-difference estimation. While there are over 50,000 observations from over 20,000 properties in the full sample for each county, there are fewer than 16,000 observations and fewer than 7,000 properties in the sample with repeated observations of dissolved oxygen. On average, properties in the estimation sample in Hillsborough county appear to be cheaper, older, farther from a water body though closer to Tampa Bay, and closer to a boat ramp. In Pinellas county, properties in the estimation sample appear to be similar on average, though closer to a water body and a boat ramp.

Figure (2) presents the distribution of house prices (in natural log as will be used in the specification) for the full sample compared to the estimation sample, which has repeated observations of dissolved oxygen. In Hillsborough county (2a) the distribution of prices when dissolved oxygen is observed is shifted left with a lower mean (as seen in the summary statistics) and more weight in the lower part of the price distribution. Conversely, the distribution of prices in Pinellas county when dissolved oxygen is observed (2b), while having a similar mean, has more bulk in the distribution above the mean. This suggests that the biases in the two counties are

Figure 1: Average dissolved oxygen level (DO) over time



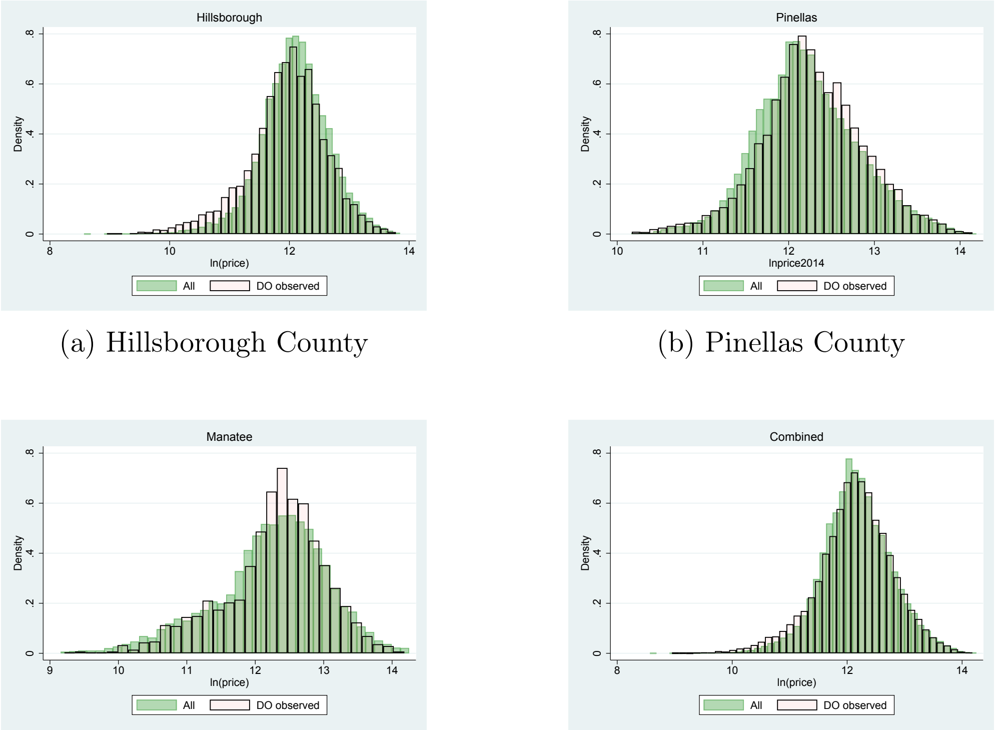
in opposite directions, as the estimation sample in Hillsborough county over samples from lower priced houses, and the estimation sample in Pinellas county over samples from higher priced houses. The seemingly representative sample from figure (2d) may result from the biases at the county level canceling out in the aggregate. The histograms for Hillsborough and Pinellas county are not in line with an ignorable missingness assumption as they suggest that observation of water quality may be related to house price. This is only suggestive as these are summary statistics, aggregated over time, and whichout any control variables.

5.3 Model and Estimation

Let $price_{it}$ be the price of house i in year t. The price is a function of household attributes X_{it}, water quality, as measured by dissolved oxygen level DO_{it}, and unobserved house components that do not vary over time c_i. The specification is[15]:

$$ln(price_{it}) = \beta_0 + \beta_1 ln(DO_{it}) + \beta_2 Age_{it} + \alpha_t + c_i + u_{it} \tag{46}$$

[15] I have modified the specification by omitting the census block by year affects to conserve degrees of freedom, omitting the recreational demand index as discussed in a previous footnote, and using a 1 km radius as discussed in a previous footnote. For more details, see Kuwayama et al. (2018)



(a) Hillsborough County

(b) Pinellas County

(c) Manatee County

(d) Combined Area

Figure 2: Distribution of house prices

where the α_t are year effects. The data span 1998 to 2014, which I break into three periods with respect to the great recession; *pre*: 1998-2006, *during*: 2007-2010, and *post*: 2011-2014. Each era is allowed a different level and linear trend, resulting in 6 time parameters. The specification is then:

$$\begin{aligned} ln(price_{it}) = \beta_1 ln(DO_{it}) + \beta_2 Age_{it} + \alpha_1 \mathbf{1}_{pre} + \alpha_2 \mathbf{1}_{during} + \alpha_3 \mathbf{1}_{post} \\ + \alpha_4 \mathbf{1}_{pre} * t + \alpha_5 \mathbf{1}_{during} * (t - 2006) + \alpha_6 \mathbf{1}_{post} * (t - 2010) + c_i + u_{it} \end{aligned} \tag{47}$$

where the overall intercept is dropped due to collinearity.

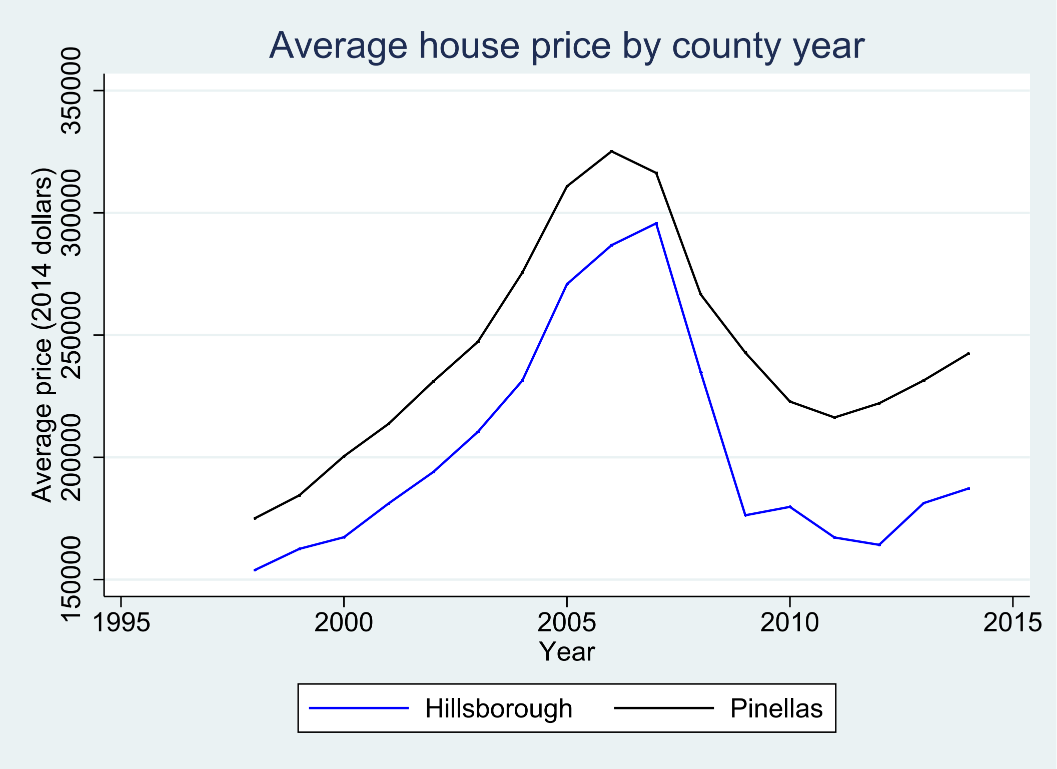
Figure (3) shows the time trend of prices in Hillsborough and Pinellas counties, and gives credence to the three-era piecewise linear specification adopted.

The following strict exogeneity assumption is maintained:

$$E[u_{it}|\boldsymbol{DO}_i, \boldsymbol{Age}_i, \alpha_t, c_i] = 0 \tag{48}$$

The property fixed effects c_i absorb the time invariant house characteristics (number of bedrooms, number of bathrooms, lot size,...). The inclusion of house specific effects in (47) focuses the identifying variation to within house, across time. The parameters are identified by repeat sales of the same house in different years. The

Figure 3: Average house prices aver time



total number of sales ranges from 2 to 7.

It is illustrative to re-write (47) taking into account the unbalanced nature stemming from considering consecutive house sales, as opposed to just consecutive years. Let t_s denote the year t of sale number s. For example, if house i was sold in 1999, 2005, and 2008, then $t_1 = 1999$, $t_2 = 2005$, and $t_3 = 2008$. Model (47) is re-written as:

$$\begin{aligned} ln(price_{it_s}) = \beta_1 ln(DO_{it_s}) + \beta_2 Age_{it_s} + \alpha_1 pre_{it_s} + \alpha_2 during_{it_s} + \alpha_3 post_{it_s} \\ + \alpha_4 pre_{it_s} * t + \alpha_5 during_{it_s} * (t - 2006) + \alpha_6 post_{it_s} * (t - 2010) + +c_i + u_{its} \end{aligned} \tag{49}$$

First-differencing (49) over sales requires a bit of care with respect to the time variables. The resulting first-differenced model is:

$$\begin{aligned} \Delta ln(price_{it}) = \beta_1 \Delta ln(DO_{it}) + \delta_1 pretime_{it} + \delta_2 duringtime_{it} + \delta_3 posttime_{it} \\ + \delta_4 \mathbf{1}_{pre\text{-}during} + \delta_5 \mathbf{1}_{pre\text{-}post} + \delta_6 \mathbf{1}_{during\text{-}post} + \Delta u_{it} \end{aligned} \tag{50}$$

where the differencing is across consecutive sales, s. Model (50) specifies the difference in sale price (measured in 2014 dollars) of house i as a function of the difference in dissolved oxygen level between the two times of sale, as well as three time exposure effects. It is assumed the sale number does not itself affect the outcome or the model parameters; i.e. a house sold for the first time in 2000 and the second time in 2005 is considered to follow the same model as a house sold the second time in 2000 and the third time in 2005. The time exposure variables - *pretime, duringtime, posttime* - are defined as the amount of time house i has spent in each era since the previous sale. The basic idea is that the change in age of the house, as measured by the number of years that has passed between sales, is broken up over the three time periods. A house sold in 2000 and resold in 2003 spent 3 years in the *pre* period, and 0 years in both the *during* and *post* periods. A house sold in 2000 and resold in 2010 spent 6 years in the *pre* period, 4 years in the *during* period, and 0 years in the *post* period. Straightforward calculations show that $\delta_1 = \beta_2 + \alpha_4$, $\delta_2 = \beta_2 + \alpha_5$, and $\delta_3 = \beta_2 + \alpha_6$. From this, notice that the slope coefficients on the exposure variables - *pretime, duringtime, posttime* - are picking up both the age effect and the era-specific linear time trend effect, as anticipated. The age effect β_2 is expected to be negative, the pre- and post- trends, α_4 and α_6 respectively, are expected to be positive and thus the signs of δ_1 and δ_3 are a priori ambiguous depending on the strength of the two effects. The during trend α_5 is expected to be negative thus δ_2 is unambiguously a priori negative. The level shifters capture the level shift if time switched eras between the two sales; i.e. $\nVdash_{pre-post} = 1$ is the previous sale was in the *pre* period and the current sale is in the *post* period. These indicators are convoluted and offer little insight with respect to the model (49), except perhaps to note that the intercept drops out if a house is sold and resold in the same era.

5.3.1 Selection Model

The model for joint observation of dissolved oxygen is specified as a bivariate probit model. For a given sale s of house i in year t_s, consider a latent variable threshold-crossing model for observation:

$$d^*_{it_s} = \delta_1 ln(price_{it_s}) + \delta_2 ln(price_{it_{s-1}}) + \delta_3 ave\_ln(price_i) + X_{it_s}\beta + v_{it_s} \tag{51}$$

$$d_{it_s} = \mathbb{1}[d^*_{it_s} > 0] \tag{52}$$

For a flexible yet tractable model, observation is allowed to depend on current sale price, sale price in the previous sale (if feasible), and average sale price across all sales of the property. Additional covariates X_{it_s} include property age in year t_s, the time-invariant property characteristics available by county (as in the summary statistics table (4)) which include the distance variables (distance to water, boat ramp, and Tampa Bay), a full set of year indicators, and a full set of zip code indicators. The goal of the selection model is to yield a *good* prediction of the probability of observation. While a direct effect of number of bedrooms on observation may not make economic sense, the variable is included to further absorb any indirect effects of dissolved oxygen level (which is correlated with number of bedrooms (Kuwayama et al. (2018))) on observation. The distance variables, particularly distance to a boat ramp, capture the ease of either direct monitoring by boat, or access to a monitoring station for repairs. The joint probability of observation of dissolved oxygen is estimated by the predicted values from the county specific bivariate probit models. The estimation results, and more details for the bivariate probit model by county are included in the appendix.

5.4 Results

Table (11) gives the results of estimation of the first-differenced reduced form (50). Table (5) reproduces the first row of table (11) to focus on the estimates of willingness to pay. The unweighted estimate for the combined sample is similar to that of KOZ for the 1 km radius (Kuwayama et al. (2018), Table 9.) The differences between the point estimates from the two estimators is stark, both within county and across county.

Comparing the point estimates from the two estimators within each county, notice that the IPW-POLS estimate for Hillsborough county decreased ten fold compared to the unweighted estimator, and is consequently no longer statistically different from zero. The IPW-POLS estimate for Pinellas county is more than four times larger than the corresponding unweighted point estimate, and, despite the larger standard error, the IPW-POLS estimate is significant at the 1% level.

The estimates from the two estimators have different implications for the relative willingness to pay for water quality across the two counties, though both estimates indicate heterogeneity in willingness to pay across county. The unweighted estimator yields a point estimate for willingness to pay for local water quality in Hillsborough county that is statistically significant and more than twice the statistically insignificant point estimate for Pinellas county, though the difference between the estimates is statistically insignificant (p-value $>$ 0.15). The IPW-POLS estimator, by contrast, yields a statistically significant point estimate in Pinellas county, which is more than ten times as large as the statistically insignificant point estimate in Hillsborough county, and the difference between the two estimates is significant at the 10% level (p-value =.007).

Note that for both estimators, and in each county, the era-exposure variables (*pre period, during, post,*) whose slope coefficients are the combination of the age and economic effect are significant. The coefficient on the *during-exposure* variable is negative as expected. The positive coefficients on the *pre-exposure* and *post-exposure* variables indicate that the economic trends were stronger predictors of price in those periods, which is not surprising given the booming nature of the housing market during the pre period.

The two estimators yield substantially different predictions for the welfare gains from improved water quality in the two counties. The average sale price in Hillsborough county over this period was roughly $207,000 (in 2014 dollars.) The IPW-POLS estimator predicts that a 10% increase in dissolved oxygen level corresponds to roughly a $72.45 increase in average sale price in Hillsborough county, whereas the unweighted estimator predicts that a 10% increase in dissolved oxygen level corresponds to roughly a $662 increase in average sale price in Hillsborough county. In Pinellas county the average sale price was approximately $243,000. For a 10% increase in dissolved oxygen level, the unweighted estimator predicts an increase of approximately $363 in average sale price, whereas the IPW-POLS estimator predicts an increase of over $2,000. Poor et. al. ((Poor et al., 2007)) predict a two and a half standard deviation change in dissolved inorganic nitrogen is worth $17,642. The IPW-POLS estimator predicts that a two and a half standard deviation change in dissolved oxygen is worth almost $500 in Hillsborough county, over $12,500 in Pinellas county, and just under $3,000 in the combined sample.

5.5 Selection

Given the different predictions of the two estimators, the question is then which estimate to accept. The reliability of one estimator or the other hinges on the mechanism governing the missing values. The unweighted estimator requires the missing values to be ignorable, and will be inconsistent if observation of dissolved oxygen is related to house price, conditional on dissolved oxygen level, age, time, and the fixed effects. The IPW-POLS estimator proposed here requires that the missing values be missing at random. The key component of the missing at random assumption is that the missing values do not depend on the level of the variable which is missing, once all other relevant observed variables have been included as controls. This allows the probability of observation to depend on house price (and thus on dissolved oxygen indirectly), but is violated if observation of dissolved oxygen is related to the level of dissolved oxygen, conditional on house price, age, time, and other observable attributes. Without more information about the implementation of monitoring policy and the practice of water monitoring, the question is empirical.

A straightforward test of ignorable missingness is to include the indicator for ob-

Table 4: Summary statistics by county and observation of dissolved oxygen

		Full sample		DO_{it} observed		$DO_{i,t,t-1}$ observed	
	Units	Mean	Std. Dev	Mean	Std. Dev.	Mean	Std. Dev.
Hillsborough County							
Observations							
N	property	21,919	-	12,754	-	6,033	-
NT	property-year	51,189	-	20,573	-	13,755	-
Property Characteristics							
Sale price	2014 dollars	206,739	121,349	207,117	125,786	196,489	125,467
Dissolved oxygen (DO)	mg/L	-	-	5.2	1.8	5.0	1.6
Age	years	29.24	19.90	35.46	22.55	39.11	24.08
Number of bedrooms	-	3.2	0.80	3.1	0.83	3.0	0.84
Number of bathrooms	-	2.1	0.66	2.0	0.69	1.9	0.70
Number of stories	-	1.2	0.38	1.2	0.38	1.2	0.39
Lot size	acres	0.29	0.41	0.27	0.38	0.25	0.33
Spatial Characteristics							
Distance to nearest water body	meters	677	665	788	699	883	717
Distance to Tampa Bay	meters	26,056	18,939	22,092	16,869	19,521	15,259
Distance to boat ramp	meters	13,711	7,678			10,521	7,210
Pinellas County							
Observations							
N	property	26,734	-	15,439	-	6,969	-
NT	property-year	62,050	-	24,058	-	15,382	-
Property Characteristics							
Sale price	2014 dollars	242,873	166,414	244,093	163,595	244,185	163,564
Dissolved oxygen (DO)	mg/L	-	-	5.7	1.8	5.8	1.7
Age	years	36.94	18.27	37.27	18.83	36.89	19.19
Living area	square feet	2,261	951	2,257	987	2,267	1,003
Spatial Characteristics							
Distance to nearest water body	meters	1,091	1,137	991	920	961	848
Distance to Tampa Bay	meters	6,799	5,494	6,574	5,398	6,562	5,365
Distance to boat ramp	meters	8,700	5,037	8,316	4,953	8,267	4,885

Table 5: First-difference estimation results: Unweighted and IPW-POLS

Dependent variable: ln(price)

	Hillsborough	Pinellas	Combined
Unweighted estimator			
ln(DO)	0.032**	0.015	0.017*
	(0.011)	(0.010)	(0.008)
IPW-POLS			
ln(DO)	0.0035	0.083**	0.027†
	(0.017)	(0.024)	(0.014)
N	7,721	8,411	17,905

Standard errors in parentheses

Significance levels : $\dagger$: 10% $*$: 5% $**$: 1%

servation in the outcome equation. In a cross-section, this is not helpful as $d_i = 1$ for the sub sample used in estimation, so the observation indicator is exactly one for every agent. With access to panel data, however, the indicator for observation in a time period may be included in the outcome equation for other time periods (Wooldridge (2010a)). If the missing values are ignorable, so observation is exogenous, then the slope coefficient on any observation indicator would be zero. Table (6) presents the results from including the indicator for the next sale d_{s+1}, and, in a separate regression, including the indicator for the sale from two periods ago d_{s-2} in the reduced form equation (50). The significance of d_{s+1} in Hillsborough and both indicators (included in separate regressions) in Pinellas is strong evidence against the ignorable missingness assumption as it indicates that observation is related to sale price, even after controlling for the other right hand side variables.

While there is no straightforward test of the missing at random assumption with missing covariates, access to panel data allows some informal exploration of the assumption. First, consider treating the data as a cross-section by adding the observation indicators for each property over time; that is, define $n_i = \sum_t d_{it}$, which is a count variable taking values between zero and seven (the max number of sales of a property.) Table (7) gives the results of estimating a count model for d_i as a function of average sale price of the property and average dissolved oxygen level for the property[16]. Control variables (not shown) in Hillsborough county are the

[16] Estimation is done as a Poisson model with exposure equal to the total number of sales of the property as properties with more sales would be expected, ceteris paribus, to have more

distance variables (distance to nearest water body, Tampa Bay, and a boat ramp,) number of bedrooms, number of bathrooms, number of stories, and lot size. Control variables in Pinellas county are the distance variables, number of stories, and living area[17].) The results in table (7) show more evidence against the ignorable missingness assumption as average price is a significant predictor, and also evidence counter to the missing at random assumption, as average water quality is also a significant predictor. Tables (8) and (9) further explore this approach by breaking the 17 years into four time periods; *post, during, pre-first half, pre-second half*, and running the same count model separately for each period. The results suggest that, in each county, there is one period in which average dissolved oxygen level significantly affects number of observations of dissolved oxygen in that period. In Hillsborough county, the the number of observations of dissolved oxygen by property is more sensitive to the average sale price by property than the average level of dissolved oxygen recorded for that property. In Pinellas county, average sale price by property is a statistically significant predictor of the number of observations of dissolved oxygen by property in two time periods, whereas average level of dissolved oxygen is statistically significant in one time period.

Summing the observation indicator within property, over years reduces the 17 years of data into one cross-section. As an alternate exploration, the data can be considered as a time series by summing the observations within year, over property. Define $n_t = \sum_i d_{it}$, which is the total number of recordings of dissolved oxygen in year t. This count variable n_t is run as a function of the average sale price of all houses sold in that year, the average dissolved oxygen level in that year, and the CPI in that year is used as a control for economic circumstances. Table (10) gives the estimation results, which again indicate evidence against ignorable missingness as the coefficient on average price is significant, but show no evidence of violation of the missing at random assumption as the coefficient on average dissolved oxygen level is insignificant.

Taken all together the evidence against ignorable missingness is strong and thus the unweighted estimates are suspect. Additionally, there is some evidence that supports the missing at random assumption. The IPW-POLS estimates are accepted as the more reliable estimates of the willingness to pay for local amenity water quality in the Tampa Bay area counties. Given the numerous differences between the results of the two estimators, this has implications for assessing the benefits to homeowners of increased dissolved oxygen levels in local water amenities.

observations.

[17] The control variables vary by county because of the available data (see Table (4)

Table 6: Testing ignorable missingness

Variable	Coefficient	(Std. Err.)
Hillsborough		
d_{its-2}	0.013	(0.011)
d_{its+1}	0.078^{**}	(0.012)
Pinellas		
d_{its-2}	0.027^{**}	(0.009)
d_{its+1}	0.034^{**}	(0.010)

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

Table 7: Count model for cross-section d_i

	Variable	Coefficient	(Std. Err.)
Hillsborough	ave_$ln(price_i)$	-0.154^{**}	(0.020)
	ave_$ln(DO_i)$	-0.117^{**}	(0.019)
Pinellas	ave_$ln(price_i)$	0.052^{**}	(0.012)
	ave_$ln(DO_i)$	0.085^{**}	(0.009)

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

Table 8: Count model for cross-section d_i by time period: Hillsborough

	Variable	Coefficient	(Std. Err.)
1998-2002	ave_$ln(price_t)$	0.016	(0.011)
	ave_$ln(DO_t)$	-0.007	(0.007)
2003-2006	ave_$ln(price_t)$	-0.058^{**}	(0.010)
	ave_$ln(DO_t)$	$0.014^{\dagger}$	(0.008)
2007-2010	ave_$ln(price_t)$	-0.013	(0.015)
	ave_$ln(DO_t)$	0.006	(0.011)
2011-2014	ave_$ln(price_t)$	0.015	(0.028)
	ave_$ln(DO_t)$	-0.053	(0.040)

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

Table 9: Count model for cross-section d_i by time periods: Pinellas

	Variable	**Coefficient**	**(Std. Err.)**
1998-2002	ave_$ln(price_t)$	0.020**	(0.007)
	ave_$ln(DO_t)$	0.004	(0.009)
2003-2006	ave_$ln(price_t)$	-0.040**	(0.009)
	ave_$ln(DO_t)$	0.053**	(0.012)
2007-2010	ave_$ln(price_t)$	0.016	(0.019)
	ave_$ln(DO_t)$	0.001	(0.017)
2011-2014	ave_$ln(price_t)$	-0.016	(0.029)
	ave_$ln(DO_t)$	0.015	(0.035)

Significance levels : † : 10% $*$: 5% $**$: 1%

Table 10: Count model for time series d_t

	Variable	**Coefficient**	**(Std. Err.)**
Hillsborough	ave_$ln(price_t)$	1.184*	(0.516)
	ave_$ln(DO_t)$	-0.995	(0.708)
	CPI	0.009*	(0.004)
Pinellas	ave_$ln(price_t)$	1.977**	(0.706)
	ave_$ln(DO_t)$	-2.063	(1.758)
	CPI	-0.011	(0.007)

Significance levels : † : 10% $*$: 5% $**$: 1%

Table 11: Estimation results

Dependent variable: ln(price)

	Hillsborough County		Pinellas County		Combined Area	
	Unweighted	IPW-POLS	Unweighted	IPW-POLS	Unweighted	IPW-POLS
ln(DO)	0.032**	0.0035	0.015	0.083**	0.017*	0.027†
	(0.011)	(0.017)	(0.010)	(0.024)	(0.008)	(0.014)
Exposure						
pre period	0.127**	0.094**	0.109**	0.096**	0.119**	0.096**
	(0.002)	(0.004)	(0.001)	(0.003)	(0.001)	(0.002)
during	-0.424**	-0.356**	-0.233**	-0.221**	-0.330**	-0.293**
	(0.011)	(0.018)	(0.008)	(0.011)	(0.007)	(0.009)
post	0.043**	0.029**	0.030**	0.036**	0.037**	0.030**
	(0.004)	(0.006)	(0.006)	(0.010)	(0.003)	(0.006)
N	7,721	7,721	8,411	8,411	17,905	17,905

Standard errors in parentheses

Significance levels : † : 10% * : 5% ** : 1%

6 Generalizations

The intuition from the first differencing of the classical linear fixed effects model extends naturally to other linear unobserved effects models, and generally to slope parameter estimation of panel models. The covariates not subject to missing values W_{it} are suppressed for ease of notation. This section gives a sketch for extending the proposed method to more general panel models. The full details are the subject of ongoing work.

6.1 Linear fixed effects model under sequential exogeneity

Consider the model as in (1) but under sequential exogeneity $E[u_{it}|X_{i1}, ..., X_{it}, c_i] = 0$. Let superscripts denote time histories us to time t, for example, $X_i^{(}t) = \{X_{i1}, ..., X_{it}\}$ and $X_i^T = \boldsymbol{X}_i$. The lagged first differences are commonly used as instruments since the sequential exogeneity assumption implies the moments $E[(X_{it-1} - X_{it-2})(u_{it} - u_{it-1})] = 0$ are valid for a random sample. A minimum of three time periods is needed for identification. If the observation of the X_{it} conditionally depends on the outcomes $\boldsymbol{y}_i$, and perhaps some V_{it} as in the previous sections, but not on the value of the X_{it} themselves; i.e. $E[d_{it}|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{V}_i] = E[d_{it}|\boldsymbol{y}_i, \boldsymbol{V}_i]$ then the missing at random assumption holds. The methods of this paper can then be applied to recover consistent slope parameter estimates. The necessary selection assumption, akin to assumptions (1) and (2), is that $E[d_{it}d_{it-1}d_{it-2}|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{V}_i] = G_{t,t-1,t-2}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1,t-2,0})$ for a known functional form $G_{t,t-1,t-2}$. The selection mechanism is now a trivariate binary outcome model. The re-weighted, fully observed valid moments are:

$$E[\frac{d_{it}d_{it-1}d_{it-2}}{G_{t,t-1,t-2}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1,t-2,0})}\Delta X_{it-1}\Delta u_{it}] = 0 \qquad t = 3, ..., T \tag{53}$$

These moments can be used for estimation as in the IPW-POLS or IPW-GMM techniques described for the linear fixed effects model under strict exogeneity.

6.2 Dynamic panel data model

Consider a dynamic linear unobserved effects model. For example, consider a model with a single lag of the dependent variable:

$$y_{it} = \gamma y_{it-1} + X_{it}\beta + c_i + u_{it} \tag{54}$$
$$E[u_{it}|y_i^{(t-1)}, X_i^{(t)}] = 0 \tag{55}$$

where $y_i^s = (y_{i1}, y_{i2}, ..., y_{is})$.

First-differencing again removes the individual effect c_i, but $\Delta y_{it} = y_{it} - y_{it-1}$ is not orthogonal to $\Delta u_{it} = u_{it} - u_{it-1}$. Anderson-Hsaio suggest using past levels of y as instruments (Anderson and Hsiao (1982)). Using lagged differences of y has also been suggested as lagged level are potentially poor instruments for current differences (Arellano and Bover (1995)). Arellano-Bond suggest using a system of instruments that grows with t as more lags are available (Arellano and Bond (1991)).

First consider the case of missing values in the X_{it}. The desired population moments can be written as:

$$E[f(y_i^{(t-2)})\Delta u_{it}] = 0 \tag{56}$$
$$E[\Delta X_{it-1}\Delta u_{it}] = 0 \qquad t = 3, ..., T \tag{57}$$

where $f(\cdot)$ is a possibly-vector valued function of the time history of the y_is up to period $t-2$; for example, levels of y_{is} for $s \leq t-2$, or first-differences $y_{is} - y_{is-1}$ for $s \leq t-2$. These moments can be shown to hold be iterating the expectation conditional on $(X_i^{(t-1)}, y_i^{(t-2)})$.

If the observation of X_{it} depends on the y_{it}, then selection is non-ignorable and must be addressed to recover consistent estimates of the slope parameters. If the observation of the X_{it} depends on the y_{it}, and potentially some additional fully observed variables V_i^t, but does not conditionally depend on the level of the X_{it}, then the values are missing at random in sense of this paper $\boldsymbol{d}_i|\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{V}_i \sim \boldsymbol{d}_i|\boldsymbol{y}_i, \boldsymbol{V}_i$. The necessary parametric assumption is that the joint probabilities $E[d_{it}d_{it-1}|\boldsymbol{y}_i, \boldsymbol{V}_i] = G_{t,t-1}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1,0})$ and $E[d_{it}d_{it-1}d_{it-2}|\boldsymbol{y}_i, \boldsymbol{V}_i] = G_{t,t-1,t-2}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1,t-2,0})$ for known functional forms $G_{t,t-1}$ and $G_{t,t-1,t-2}$, and unknown parameters δ_0.

The valid re-weighted fully observed moments can be written as:

$$E[\frac{d_{it}d_{it-1}}{G_{t,t-1}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1})} f(y_i^{(t-2)})\Delta u_{it}] = 0 \tag{58}$$
$$E[\frac{d_{it}d_{it-1}d_{t-2}}{G_{t,t-1,t-2}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1,t-2})} \Delta X_{it-1}\Delta u_{it}] = 0 \qquad t = 3, ..., T \tag{59}$$

The case of missing values in the y_{it} can also be handled with the IPW estimators of this paper. The necessary assumption on selection is that observation of the y_{it} depends on the X_{it}, and perhaps some V_{it}, but not conditionally on the y_{it} themselves so $\boldsymbol{d}_i|\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{V}_i \sim \boldsymbol{d}_i|\boldsymbol{X}_i, \boldsymbol{V}_i$. The extent of time periods for which it is necessary to model the joint probability of observation depends on the desired instruments. Using as few time periods as possible, while leaving some efficiency on the table (Holtz-Eakin et al. (1988),Arellano and Bond (1991)), has advantages in terms of data usage when only the complete cases are to be used, as previously discussed for the classical linear fixed effects model. Consider first-differenced instruments that use as few time periods as possible:

$$E[\Delta y_{it-2}\Delta u_{it}] = 0 \qquad t = 4, ..., T \tag{60}$$

$$E[\Delta X_{it-1}\Delta u_{it}] = 0 \qquad t = 3, ..., T \tag{61}$$

where it is assumed $T \geq 4$. The fully observed moments are:

$$E[d_{it}d_{it-1}d_{it-2}d_{it-3}\Delta y_{it-2}\Delta u_{it}] \qquad t = 4, ..., T \tag{62}$$

$$E[d_{it}d_{it-1}\Delta X_{it-1}\Delta u_{it}] \qquad t = 2, ...T \tag{63}$$

With the parametric assumption $E[d_{it}d_{it-1}d_{it-2}d_{it-3}|\boldsymbol{X}_i, \boldsymbol{V}_i] = G_{t,t-1,t-2,t-3}(\boldsymbol{X}_i, \boldsymbol{V}_i; \delta_{t,t-1,t-2,t-3,0})$, $E[d_{it}d_{it-1}|\boldsymbol{X}_i, \boldsymbol{V}_i] = G_{t,t-1}(\boldsymbol{X}_i, \boldsymbol{V}_i; \delta_{t,t-1,0})$ for some known forms $G_{t,t-1,t-2,t-3}$ and $G_{t,t-1}$, then the re-weighted population moments:

$$E[\frac{d_{it}d_{it-1}d_{it-2}d_{it-3}}{G_{t,t-1,t-2,t-3}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1,t-2,t-3})}\Delta y_{it-2}\Delta u_{it}] = 0 \qquad t = 4, ..., T \tag{64}$$

$$E[\frac{d_{it}d_{it-1}}{G_{t,t-1}(\boldsymbol{y}_i, \boldsymbol{V}_i; \delta_{t,t-1})}\Delta X_{it-1}\Delta u_{it}] = 0 \qquad t = 2, ..., T \tag{65}$$

are valid.

Estimation similar in style to the typical GMM estimation of these models (Holtz-Eakin et al. (1988), Arellano and Bond (1991)) then follows along the same steps as in the linear fixed effects IPW-GMM case.

6.3 General form

While the motivation for the estimators in this paper is linear fixed effects models, neither linearity nor the presence of unobserved effects is necessary for the general approach. The general approach can be specified in terms of moment conditions which are functions of variables subject to missing values X_{it}, fully observed variables Y_{it}, and unknown parameters of interest θ_0 in multiple time periods. Let d_{it} be an indicator taking value 1 when X_{it} is fully observed. This encompasses the linear unobserved effects models presented above, as well as non-linear models without fixed effects, and the few cases of non-linear fixed effects models which can be written in terms of moments which are functions of the observable data and the unknown parameters, such as the conditional logit estimator (Chamberlain (1979).)

The object of interest is a finite-dimensional parameter θ_0, which is the unique solution to a vector of L moments conditions. The moment conditions can be divided in sub-vectors based on the time periods of the X_{it} involved. Each sub-vector, m, of size $|m|$, involves a subset S_m of the time periods $\{1, ..., T\}$ of the X_{it}, for a total of M sub-vectors. For example, the linear fixed effect model under strict exogeneity has $\theta_0 = \beta_0$, a K vector, $L = K(T-1)$, $M = (T-1)$, $|m| = K$, and $S_m = \{t, t-1\}$. That is, the $K(T-1)$ moment conditions can be split into $T-1$ groups. Each group is a K vector that involves two time periods, $\{t, t-1\}$.

Consider the dynamic linear unobserved effects model above, with missing covariate values X_{it}. There are moment conditions, where the τ will depend on how many time periods from y_i^{t-2} are used. For example, if $f(y_i^{(t-2)}) = y_{it-2} - y_{it-3}$, then $|f(\cdot)| = 1$ and $\tau = T-4$. The $K(T-2) + |f(\cdot)|\tau$ can be separated into $(T-2)+\tau$ sub-vectors based on the number of time periods of X_{it} involved. Each sub-vector from the $(T-2)$ group is a K vector involving three time periods, $S_m = \{t, t-1, t-2\}$. Each sub-vector from the $(T-1)$ group is a scalar involving two time periods, $S_m = \{t, t-1\}$ of the X_{it}.

For most non-linear models, such as a pooled probit model or Chamberlain's conditional logit estimator (Chamberlain (1979)), the moment conditions involve the covariates from all time periods. In that case, $m = 1$, and $S_m = \{1, ..., T\}$.

In general, the moments are written as:

$$E[\Psi_m(X_{iS_m}; Y_{iS_m}; \theta_0)] = 0 \qquad m = 1, ..., M \tag{66}$$

where Ψ is a $|m| \times 1$ vector-valued function, and the S_m subscript denotes the collection of t in S_m; e.g. $X_{iS_m} = \{X_{it}\}_{t \in S_m}$. For example, in the linear fixed effects model, $\Psi_m = (X_{it} - X_{it-1})(y_{it} - y_{it-1} - (X_{it} - X_{it-1})\beta_0)$ for each m.

It is suspected, by the researcher, that the moments for the observable population:

$$E[D_i^{S_m} \Psi_m(X_{iS_m}; Y_{iS_m}; \theta)] \qquad m = 1, ..., M \tag{67}$$

may not be valid at the true θ_0, where $D_i^{S_m} = \prod_{t \in S_m} d_{it}$.

If the missing at random assumption, $\boldsymbol{d}_i | \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{V}_i = \boldsymbol{d}_i | \boldsymbol{Y}_i, \boldsymbol{V}_i$ holds, and the parametric assumption $E[d_{iS_m} | \boldsymbol{Y}_i, \boldsymbol{V}_i] = G_{S_m}(\boldsymbol{Y}_i, \boldsymbol{V}_i; \delta_{S_m 0})$ is satisfied for each m, where the V_{it} are as discussed for the classical linear fixed effects model, then the reweighted moments for the observable population:

$$E[\frac{D_i^{S_m}}{G_{S_m}(\boldsymbol{Y}_i, \boldsymbol{V}_i; \delta_{S_m 0})} \Psi_m(X_{iS_m}; Y_{i\mathfrak{S}}; \theta_0)] = 0 \qquad m = 1, ..., M \tag{68}$$

are valid. Two-step estimation of the parameter θ_0 can proceed as described in this paper.

7 Conclusion

This paper analyzes linear unobserved effects models subject to missing values in the covariates. A missing at random assumption, in which the probability of observation does not directly depend on the variables subject to missing values, is adopted. The combination of missing at random covariates in a panel model with fixed effects is new in the literature. Given the ubiquity of missing data and the popularity of linear fixed effects models in applied research, the results in this paper are practically important.

Two inverse probability weighted estimators are proposed, their asymptotic theory is derived, and some finite sample performance is presented. The first estimator is a pooled estimator which is computationally simple and familiar from the textbook treatment of fixed effects models. The second estimator is an over-identified GMM estimator which enjoys better asymptotic efficiency. The simulations confirm that the commonly used unweighted estimator performs poorly when selection depends on the outcome variable. The simulations suggest that the efficiency gains from the optimally-weighted GMM estimator may be small.

An environmental application is explored, in which the proposed pooled IPW estimator is applied to a hedonic house price model to measure the willingness to pay for water quality in local water features. Empirical evidence suggests that the ignorable missingness assumption is violated, and the missing at random assumption may hold. Estimation based on the proposed pooled IPW estimator gives significantly different results from the unweighted estimator used in the literature, thus accounting for the selection may have important policy implications.

References

Abrevaya, J. (2018). Missing dependent variables in fixed-effects models. *forthcoming.*

Anderson, T. W. and C. Hsiao (1982). Formulation and estimation of dynamic models using panel data. *Journal of econometrics 18*(1), 47–82.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies 58*(2), 277–297.

Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of econometrics 68*(1), 29–51.

Chamberlain, G. (1979). Analysis of covariance with qualitative data.

Chamberlain, G. (1984). Panel data. *Handbook of econometrics 2*, 1247–1318.

Chen, B., G. Y. Yi, and R. J. Cook (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association 105*(489), 336–353.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics 36*(2), 808–843.

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies 79*(3), 1053–1079.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica: Journal of the Econometric Society*, 1371–1395.

Keiser, D. A. and J. S. Shapiro (2017). Consequences of the clean water act and the demand for water quality. Technical report, National Bureau of Economic Research.

Krutilla, J. V. (1967). Conservation reconsidered. *The American Economic Review 57*(4), 777–786.

Kuwayama, Y., S. Olmstead, and J. Zheng (2018). The value of water quality: Separating amenity and recreational benefits.

Livy, M., A. Klaiber, et al. (2013). Maintaining public goods: Household valuation of new and renovated local parks. In *Association of Environmental and Resource Economists conference, Banff, Alberta, June*, Volume 11.

McConnell, V. and M. A. Walls (2005). *The value of open space: Evidence from studies of nonmarket benefits*. Resources for the Future Washington, DC.

Moffit, R., J. Fitzgerald, and P. Gottschalk (1999). Sample attrition in panel data: The role of selection on observables. *Annales d'Economie et de Statistique*, 129–152.

Muehlenbachs, L., E. Spiller, and C. Timmins (2015). The housing market impacts of shale gas development. *American Economic Review 105*(12), 3633–59.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Nijman, T. and M. Verbeek (1992). Nonresponse in panel data: The impact on estimates of a life cycle consumption function. *Journal of Applied Econometrics 7*(3), 243–257.

of Hillsborough County, E. P. C. (2018). Water monitoring and maps. http://www.epchc.org/divisions/water-management/water-monitoring-maps-and-data.

Poor, P. J., K. L. Pessagno, and R. W. Paul (2007). Exploring the hedonic value of ambient water quality: A local watershed-based study. *Ecological Economics 60*(4), 797–806.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association 90*(429), 106–121.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy 82*(1), 34–55.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Walsh, P., C. Griffiths, D. Guignet, and H. Klemick (2017). Modeling the property price impact of water quality in 14 chesapeake bay counties. *Ecological Economics 135*, 103–113.

Walsh, P. J., J. W. Milon, and D. O. Scrogin (2011). The spatial extent of water quality benefits in urban housing markets. *Land Economics 87*(4), 628–644.

Wooldridge, J. M. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal 1*(2), 117–139.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*(2), 1281–1301.

Wooldridge, J. M. (2010a). Correlated random effects models with unbalanced panels. *Manuscript (version May 2010), Michigan State University*.

Wooldridge, J. M. (2010b). *Econometric analysis of cross section and panel data.* MIT press.

Zheng, J. (2017). The value of local and regional water quality to homeowners in tampa, florida. *working paper*.

8 Appendix

8.1 Empirical Application Appendix

8.1.1 Selection Model

The results from estimation of the bivariate probit model based on (51) are presented below. A full set of year indicators and zip code indicators are included in the model, and omitted in the results. When previous price is not available (for example in the first sale), it is omitted from the specification, as in table (13). For example, for the joint observation of dissolved oxygen in sale 1 and sale 2, observation for sale 2 is allowed to depend on the price at sale 2 and the price at sale 1 (as well as the average price if there are more than 2 sales), and observation for sale 1 depends on the price at sale 1 and the average price.

While the slope coefficients from a bivariate probit model are not particularly illustrative, there are a few things of interest in the full estimation results. Note that, for each county, the correlation of observation across sales is statistically different from zero (note that "athrho," which is a transformation of the correlation, is significant at the one percent level), and (not shown) the p-value for a Wald test of the null hypothesis of no correlation across time, is zero to four decimal places. Accounting for the joint nature of observation to predict the joint probability is important, as using the product of the probabilities in each time period, for example, would only be valid if the correlation were zero.

Table 12: Bivariate probit estimation: Hillsborough County

Variable	**Coefficient**	**(Std. Err.)**
Equation 1 : d_{it}		
$ln(price_{it})$	0.136	(0.083)
$ln(price_{it-1})$	-0.005	(0.075)
property_age	0.006^{**}	(0.001)
ave_$ln(price_i)$	0.125	(0.147)
dist_boatramp	0.000^{**}	(0.000)
dist_water	0.000^{*}	(0.000)
dist_TB	0.000	(0.000)
beds	-0.018	(0.028)
baths	-0.024	(0.041)
stories	-0.134^{**}	(0.051)
units	0.208	(0.169)
Equation 2 : d_{t-1}		
$ln(price_{it-1})$	0.102	(0.096)
$ln(price_{it-2})$	-0.176^{*}	(0.083)
(property_age)$_{t-1}$	0.005^{**}	(0.002)
ave_$ln(price_i)$	0.276^{*}	(0.133)
dist_boatramp	0.000^{**}	(0.000)
dist_water	0.000^{*}	(0.000)
dist_TB	0.000^{*}	(0.000)
beds	-0.066^{*}	(0.029)
baths	0.000	(0.042)
stories	-0.065	(0.052)
units	$0.350^{\dagger}$	(0.191)
Equation 3 : /		
athrho	0.760^{**}	(0.026)
N	7938	
Log-likelihood	-7479.713	
$\chi^2_{(138)}$	3534.431	

Significance levels : $\dagger$: 10% $*$: 5% $**$: 1%

Table 13: Bivariate probit estimation: Hillsborough County, first two sales

Variable	**Coefficient**	**(Std. Err.)**
Equation 1 : d_{it}		
$ln(price_{it})$	0.169*	(0.070)
$ln(price_{it-1})$	-0.111	(0.070)
property_age_t	0.005**	(0.001)
ave_$ln(price_i)$	0.155	(0.125)
dist_boatramp	0.000**	(0.000)
dist_water	0.000**	(0.000)
dist_TB	0.000**	(0.000)
beds	-0.077**	(0.017)
baths	-0.011	(0.024)
stories	0.010	(0.030)
units	0.226*	(0.099)
Equation 2 : d_{it-1}		
$ln(price_{it-1})$	0.060	(0.060)
property_age_{t-1}	0.000	(0.001)
ave_$ln(price_i)$	-0.065	(0.058)
dist_boatramp	0.000**	(0.000)
dist_water	0.000**	(0.000)
dist_TB	0.000**	(0.000)
beds	-0.051**	(0.018)
baths	0.034	(0.026)
stories	-0.036	(0.033)
units	0.089	(0.097)
Equation 3 : /		
athrho	0.616**	(0.016)

N	21332
Log-likelihood	-19578.245
$\chi^2_{(143)}$	9354.155

Significance levels : † : 10% $*$: 5% $**$: 1%

Table 14: Bivariate probit estimation: Pinellas County

Variable	**Coefficient**	**(Std. Err.)**
Equation 1 : d_{it}		
$ln(price_{it})$	-0.240**	(0.084)
$ln(price_{it-1})$	-0.041	(0.084)
property_age	0.007**	(0.001)
ave_$ln(price_i)$	0.217	(0.152)
dist_boatramp	0.000	(0.000)
dist_water	0.000	(0.000)
dist_TB	0.000**	(0.000)
stories	-0.008	(0.049)
living_units	-0.311	(0.752)
house_size	0.000	(0.000)
Equation 2 : d_{it-1}		
$ln(price_{it-1})$	-0.135	(0.091)
$ln(price_{it-2})$	-0.033	(0.071)
property_age_{t-1}	0.001	(0.001)
ave_$ln(price_i)$	-0.042	(0.122)
dist_boatramp	0.000**	(0.000)
dist_water	0.000	(0.000)
dist_TB	0.000**	(0.000)
stories	0.037	(0.047)
living_units	-0.472	(0.769)
house_size	0.000	(0.000)
Equation 3 : /		
athrho	0.590**	(0.022)
N	8971	
Log-likelihood	-9607.584	
$\chi^2_{(138)}$	3044.555	

Significance levels : $\dagger$: 10% $*$: 5% $**$: 1%

Table 15: Bivariate probit estimation: Manatee County

Variable	**Coefficient**	**(Std. Err.)**
Equation 1 : do_obs		
lnprice2014	0.151**	(0.048)
avelnprice	0.261**	(0.057)
property_age	0.007**	(0.001)
dist_boatramp	0.000**	(0.000)
dist_wb	0.000**	(0.000)
dist_tb	0.000**	(0.000)
acreage	-0.117**	(0.034)
square_footage_gross	0.000†	(0.000)
Equation 2 : do_obs_prev		
lnprice_prev	0.025	(0.047)
property_age_prev	0.007**	(0.001)
avelnprice	0.394**	(0.060)
dist_boatramp	0.000**	(0.000)
dist_wb	0.000**	(0.000)
dist_tb	0.000**	(0.000)
acreage	-0.123**	(0.033)
square_footage_gross	0.000	(0.000)
Equation 3 : /		
athrho	1.222**	(0.033)
N	5351	
Log-likelihood	-5755.448	
$\chi^2_{(32)}$	739.948	

Significance levels : $\dagger$: 10% $*$: 5% $**$: 1%

8.2 Technical Proofs

Lemma 1. *(Identification) Under assumptions* (1), (3), (4), (5), *and* (6),

$$E[\frac{d_t d_{t-1}}{p_{t,t-1}(z_t,\delta_{t0})}\Delta L_t' \Delta u_t] = 0 \tag{69}$$

Proof. Proof of Theorem 1.

Let $L_{it} = (\Delta X_{it}\ \Delta W_{it})$

From the derivation of the OLS estimator and (1) we have:

$$(\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)} L_{it}' L_{it})(\hat{\beta}_{POLS} - \beta_0) = (\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)} L_{it}' \Delta u_{it}) \tag{70}$$

$\hat{\delta}_t$ is an MLE estimator so, by assumpiton (2), $\hat{\delta}_t \to \delta_{t0}$ in probability. By the previous lemma, $\hat{\beta} \to \beta_0$ in probability, thus $(\hat{\beta}_{POLS} - \beta_0) = o_p(1)$, and

$$(\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)} L_{it}' L_{it}) = (\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})} L_{it}' L_{it}) + o_p(1) \tag{71}$$

Then:

$$(\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)} L_{it}' L_{it})(\hat{\beta}_{POLS} - \beta_0) = (\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})} L_{it}' L_{it})(\hat{\beta}_{POLS} - \beta_0) + o_p(1 \tag{72}$$

since $o_p(1)o_p(1) = o_p(1)$.

Thus:

$$\hat{\beta}_{POLS} - \beta_0 = (\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})} L_{it}' L_{it})^{-1} (\sum_i \sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)} L_{it}' \Delta u_{it}) + o_p(1) \tag{73}$$

Let

$$A_0 = E(\sum_{t=2} \frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})} L_{it}' L_{it}) = E[\sum_{t=2} L_{it}' L_{it}] \tag{74}$$

where the second equality follows from assumption (2). By the uniform weak law of large numbers,

$$(\frac{1}{N}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})}L'_{it}L_{it}) = A_0 + o_p(1) \tag{75}$$

Thus,

$$\hat{\beta}_{POLS} - \beta_0 = A_0^{-1}(\frac{1}{N}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)}L'_{it}\Delta u_{it}) + o_p(1) \tag{76}$$

since the term multiplying A_0^{-1} is $o_p(1)$, and we have:

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta_0) = A_0^{-1}(\frac{1}{\sqrt{N}}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)}L'_{it}\Delta u_{it}) + o_p(1) \tag{77}$$

A mean-value expansion around δ_{t0} gives:

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)}L'_{it}\Delta u_{it} =& \frac{1}{\sqrt{N}}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})}L'_{it}\Delta u_{it} && (78)\\ &- E[\sum_{t=2}d_{it}d_{it-1}\frac{\nabla_\delta G_t(Z_{it},Z_{it-1};\delta_{t0})}{G_t(Z_{it},Z_{it-1};\delta_{t0})^2}L'_{it}\Delta u_{it}]\sqrt{N}(\hat{\delta}_t - \delta_{t0}) + && (79)\end{aligned}$$

Let the K by P matrix,

$$C_0 = E[\sum_{t=2}d_{it}d_{it-1}\frac{\nabla_\delta G_t(Z_{it},Z_{it-1};\delta_{t0})}{G_t(Z_{it},Z_{it-1};\delta_{t0})^2}L'_{it}\Delta u_{it}] \tag{80}$$

To this point the derivation mirrors the standard derivation for a two-step IPW estimator. The complication arises in the function $G_t(\cdot)$ and its derivatives. Let $s_{it}(\delta_0)$ be the P by 1 score vector of the bivariate binary response log-likelihood. From standard MLE theory,

$$\sqrt{N}(\hat{\delta}_t - \delta_{t0}) = (E[s_{it}s'_{it}])^{-1}[\frac{1}{\sqrt{N}}\sum_i s_{it}(\delta_{t0})] + o_p(1) \tag{81}$$

where the P by P matrix $E[s_{it}s'_{it}]$ is the negative of the Hessian by the information matrix equality. Substituting back in,

$$\frac{1}{\sqrt{N}}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\hat{\delta}_t)}L'_{it}\Delta u_{it} \tag{82}$$

$$=\frac{1}{\sqrt{N}}\sum_i\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})}L'_{it}\Delta u_{it}-C_0[E[s_{it}s'_{it}]]^{-1}[\frac{1}{\sqrt{N}}\sum_i s_{it}(\delta_{t0})]+o_p(1) \tag{83}$$

$$=\frac{1}{\sqrt{N}}\sum_i[\sum_{t=2}\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})}L'_{it}\Delta u_{it}-C_0[E[s_{it}s'_{it}]]^{-1}s_{it}(\delta_{t0})]+o_p(1) \tag{84}$$

$$:=\frac{1}{\sqrt{N}}\sum_i[\sum_{t=2}r_{it}]+o_p(1) \tag{85}$$

where the K by P vector $r_{it}=\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})}\Delta L'_{it}\Delta u_{it}-C_0[E[s_{it}s'_{it}]]^{-1}s_{it}(\delta_{t0})$. Finally,

$$\sqrt{N}(\hat{\beta}_{POLS}-\beta_0)=A_0^{-1}\frac{1}{\sqrt{N}}\sum_i[\sum_{t=2}r_{it}]+o_p(1) \tag{86}$$

and

$$avar(\sqrt{N}(\hat{\beta}_{POLS}-\beta_0))=A_0^{-1}B_0A_0^{-1} \tag{87}$$

where $B_0=E[(\sum_{t=2}r_{it})(\sum_{t=2}r'_{it})]$.

Recall $r_{it}=\frac{d_{it}d_{it-1}}{G_t(Z_{it},Z_{it-1};\delta_{t0})}\Delta L'_{it}\Delta u_{it}-C_{t0}[E[s_{it}s'_{it}]]^{-1}s_{it}(\delta_{t0})$.

□

Proof of Theorem 4:

Proof. Linearity in β of the moment conditions means that the solution to (12) has a closed form, and,

$$\hat{\beta}_{GMM}=\beta_0+(\hat{A}_{00}\hat{\Sigma}_1\hat{A}_{00})^{-1}\hat{A}_{00}\hat{\Sigma}_1\bar{M} \tag{88}$$

Consistency then follows since $\hat{\delta}_t\to p$ and the regularity conditions ensure $\bar{M}\to E[M]=0$. The asymptotic distribution follows similarly from the previous proof by expanding $\bar{M}$ around $\delta_0=(\delta'_{20},...,\delta'_{T0})$. □