Missing Data in Asset Pricing Panels*

Joachim Freyberger† Björn Höppner‡

Andreas Neuhierl§ Michael Weber¶

September 2021

Abstract

Missing data for return predictors is a common problem in cross sectional asset pricing studies. Most papers do not explicitly discuss how they treat missing data but conventional treatments focus on complete cases for all predictors or impute the unconditional mean for the missing predictor. Both methods have undesirable properties - they are either inefficient or lead to biased estimators and incorrect inference. We propose a simple and computationally attractive alternative approach using conditional mean imputations and weighted least squares. This method allows us to use all sample points with observed returns, it results in valid inference, and it can be applied in nonlinear and high-dimensional settings. We map our estimator into a GMM framework to study its relative efficiency and find that it performs almost as well as the efficient but computationally costly GMM estimator in many cases. We apply our procedure to a large panel of return predictors and find that it leads to improved out-of-sample predictability.

JEL classification: C14, C58, G12

Keywords: Cross Section of Returns, Missing Data, Expected Returns, Generalized Method of Moments

*We thank Bruce Carlin, Zhuo Chen, Alex Chinco, Kevin Crotty, Wayne Ferson, Todd Gormely, Lena Janys, Andrew Karolyi, Soohun Kim, Hugues Langlois, Yan Liu, Asaf Manela, Markus Pelger, Jan Scherer, Guofu Zhou and seminar participants at Washington University in St. Louis and Rice University for helpful comments and discussions.

†University of Bonn. e-Mail: freyberger@uni-bonn.de

‡University of Bonn. e-Mail: b.hoeppner@uni-bonn.de

§Washington University in St. Louis, Olin School of Business. e-Mail: andreas.neuhierl@wustl.edu

¶Booth School of Business, the University of Chicago and NBER. e-Mail: michael.weber@chicagobooth.edu.

1 Introduction

Missing data is a common problem in cross-sectional asset pricing studies. While the problem of missing return observations has received some attention and is typically treated by utilizing the so-called delisting returns (Shumway (1997), Beaver et al. (2007)), the problem of missing covariates is typically only addressed implicitly. A large and growing number of papers utilize these covariates, such as firm characteristics, to predict future returns cross-sectionally or use such information for building factor portfolios to explain the cross section of returns. In these studies, by far the most common procedure to deal with missing covariates is to exclude an observation altogether if any covariate is missing and conduct the subsequent analysis only on the cases for which no covariates or outcomes are missing (complete cases analysis). This approach typically neglects a substantial subset of the data. For example, in this paper, we use the data set of Chen and Zimmermann (2021) containing more than 3 million return observations and 40 covariates. For around 2 million of these observations between 1 and 5 covariates are missing. These observations would then all be excluded from the analysis, even though they contain useful information. This is in contrast to what Zhang et al. (2005) call "one of statistics' first principles" – "thou shall not throw data away". Moreover, the complete case approach has an additional drawback that may be overlooked at first sight. By conditioning on firms, for which all covariates are available, we might inadvertently ignore an interesting part of the *return distribution* and thus preclude us from forming better portfolios.

To harness the additional power from studying all firms with valid return observations, we propose a simple approach to impute the missing observations in the covariates. At an intuitive level, our approach works by replacing the missing covariates with suitable estimates and accounting for the estimation error in subsequent inference. In addition, we also "down-weigh" the observations for which data was imputed, thereby adjusting for the fact that these are not truly observed data points. In general, the more covariates are imputed, the less weight an observation receives. Our approach therefore allows us to use all firms with valid return observations, while enabling feasible and correct inference. We can obtain

suitable replacements of the missing values from the (observed) cross-section or from the time-series of past observations. The method can be used if the main model of interest is parametric or nonparametric and does not require us to specify the entire distribution of the missing covariates. We show that it can be cast into a generalized method of moments (Hansen (1982)) setting, which allows us to easily study its statistical properties. This enables us to account for the imputation step in conducting inference and also understand the efficiency gains of the proposed approach. Contrary to many Bayesian and likelihood-based approaches that address missing data issues, such as multiple imputation or the EM algorithm, our method is computationally inexpensive and places fewer assumptions on the data generating process. However, we do need to impose certain assumptions on why observations are missing. Specifically, similar to the complete case and many other approaches, we cannot allow the probability that a particular observation is missing to depend on the outcome variable, once we condition on observed covariates. We characterize the conditions under which we obtain consistent estimates and correct inference, and we argue that these conditions are plausible in many empirical asset pricing studies.

In recent years, many asset pricing papers aim to respond to Cochrane (2011)'s multidimensional challenge. In such a setting the number of possible predictors naturally grows and the missing data problem is aggravated. In an attempt not to throw away too many observations, some researchers replace missing values of the covariates with their cross-sectional mean (mean imputation) of that period. We wholeheartedly agree with the aim to use as many return observations as possible. However, we also show that mean imputation is rarely desirable. Mean imputation (typically) leads to inconsistent estimates and incorrect standard errors. Intuitively, mean imputation leads to an underestimation of (co)variance and hence over-rejection of null hypotheses, i.e. in the context of cross-sectional asset pricing, we would find too many successful cross-sectional predictors.

We illustrate the finite sample properties of our approach in an extensive simulation study and find that it performs well in realistic sample sizes. The simulations also help illustrate when the ad-hoc approaches, such as mean imputation and complete case analysis are (and

are not) problematic.

We also apply our method to the CRSP/Compustat sample. We find that it is indeed desirable to use all return observations because conditioning on complete cases ignores an interesting part of the return distribution. Portfolios sorted on the return prediction achieve higher out-of-sample returns when using the full sample, where the missing predictors are imputed using our method. We also document that mean imputation can lead to some incorrect inferences. In addition, we carry out a model selection analysis over the full sample to determine the most important predictors. The complete cases analysis discards many predictors and even well established predictors such as size or value are found to be irrelevant. We illustrate that our simple approach can be widely applied as it allows the use of all valid return observations, while providing correct standard errors for inference at the same time.

1.1 Related Literature

The problem of missing data is ubiquitous in empirical analyses. For example, clinical trials routinely have to confront the problem that some patients do not show up for follow-up examinations. A related problem occurs in surveys, where respondents often leave questions blank, sometimes by accident and at other times because they feel uncomfortable answering. Regardless of the reason, the result is missing data. Either explicitly or implicitly, researchers have to make assumptions about how to proceed with the empirical analysis in such situations. The problem of missing data and issues like the ones listed here have long been recognized in the applied and methodological literature. Consequently, researchers have proposed many different procedures to deal with missing data in a variety of settings.

The general literature on missing data is too vast to be summarized here and we refer to Molenberghs et al. (2015) and Little and Rubin (2020) for textbook introductions to the most common approaches to deal with missing data in different situations. We will therefore only review the most common approaches that are closely related to ours and place special emphasis on the treatment of missing data in asset pricing. In general, there is no single procedure that can be successfully applied to all missing data problems. Dealing

with missing data successfully requires taking a stance on *why* the data is missing – the so called missing mechanism.[1] If the probability that a particular observation is missing depends on the outcome variable (even after conditioning on observables), this is typically labeled as not missing at random. In this case, the missing mechanism has to be modeled explicitly, for example through a selection model, such as the Heckman selection estimator (Heckman (1979)). Since we do not pursue such an approach, we will not elaborate on this literature further.[2]

In situations where the probability of observing a sample point does not depend on the outcome variable itself, but may depend on observed covariates, the literature has proposed several general approaches to deal with missing data. Some of these approaches rely on strong distributional assumptions on unobservables (i.e. likelihood approaches and Bayesian methods) that we do not want to impose. Instead, we use a method based on moment restrictions and imputation, i.e. replacing the missing variables with suitable estimates. Imputation has a long history and is studied, among others, in Yates (1933), Dagenais (1973), Rubin (1978), Nijman and Palm (1988), Little (1992), and Rao and Toutenburg (1999). Just like we do, some of these approaches also down-weigh observations with missing values, but these studies typically only allow for one missing pattern, which means that either all variables are observed or one particular subset of the variables is missing. We extend these ideas (specifically the weighting approach of Dagenais (1973)) and allow for general missing patterns. One challenge that arises with imputation methods is how to account for the "imputation uncertainty" in inference, because the imputations are estimates themselves. This idea goes back to Gourieroux and Monfort (1981) who also just have one missing pattern. One way to approach this issue is to cast the imputation model and main model in a generalized method of moments (GMM) setting (Hansen (1982)) and thereby obtain standard errors that are corrected for the uncertainty from the imputation step. Following this route, Abrevaya and Donald (2017) study the efficient estimator with one missing pattern. One drawback of the

[1] We review the most commonly used missing mechanisms in Section A.1.

[2] In finance, studies of fund performance are examples in which such a situation arises as noted by Brown et al. (1992).

optimal GMM estimator is that it can be computationally very costly as it amounts to solving a nonlinear optimization problem. In our application with general missing patterns and many covariates, the efficient GMM estimator is computationally infeasible. These problems are also well-documented in macro finance applications, e.g. Hansen et al. (1996). We show that our estimator can be interpreted as a GMM estimator with a specific weight matrix.[3] This estimator is available in closed form, computationally much less costly than the efficient estimator, and simulation show that the loss in efficiency is small. Importantly, we can use standard GMM results to compute standard errors.

Another estimation approach that relies on moment restrictions is inverse probability weighting (IPW), i.e. re-weighting the complete case sample such that it more closely mirrors the population (Robins et al. (1994), Wooldridge (2007)), in which case we typically need to model the probability that a particular case is observed. The IPW approach relaxes important assumptions relative to the (unweighted) complete case, but does not use all available data. A considerable generalization is the class of augmented IPW (AIPW) estimators, which uses the whole sample. Under certain assumptions, which differ slightly from our setup, Robins et al. (1994) show that the AIPW estimator is semiparameteric efficient. However, similar to the optimal GMM estimator, the efficient AIPW estimator is generally not available in closed form and computationally prohibitive in our application. For comprehensive results on AIPW estimators see for example Tsiatis and Davidian (2015).

To our knowledge, no paper deals explicitly with the problem of missing predictors in multivariate (cross sectional) asset pricing studies and studies the consequences of different assumptions. Nonetheless, the problem of missing data has been recognized by empirical asset pricing researchers. In early an contribution, Haugen and Baker (1996) worry if a potential bias may arise from using only the fully observed cases. While most papers do not explicitly state this, using only cases for which all covariates (and the outcome) are observed, appears to be the most commonly employed approach to deal with missing data

[3]Zhou (1994) uses an alternative weight matrix to derive analytical GMM tests in the context of linear factor models. More recently, Liao and Liu (2020) also propose a two-step approach to test linear factor models – notably, they obtain optimality results in this case.

in asset pricing studies. Recent examples who use the complete case method are Lewellen (2015), Freyberger et al. (2020), Kelly et al. (2019), Kim et al. (2021). Other papers, follow a special imputation approach and replace the missing covariate values with the cross-sectional mean or median, see e.g. Light et al. (2017), Kozak et al. (2020), Gu et al. (2020).

Connor and Korajczyk (1987), Xiong and Pelger (2019), Kim and Skoulakis (2018) and Blanchet et al. (2021) are concerned with different missing data problems relative to us, but they deserve special mention as part of the few papers in finance that recognize missing data as an issue to be dealt with in empirical studies. Other recent papers that deal with missing data in factor models include Bai and Ng (2021), Cahan et al. (2021), and Jin et al. (2021). Lastly, Harvey et al. (2016) recognize that unreported tests for the significance of a cross-sectional predictors can be interpreted as a missing data problem. They estimate the number of unreported (and thus missing) tests and then suitably adjust their proposed multiple testing thresholds.

2 Model

2.1 Simple example

We start by illustrating the main idea of our approach using a simple example with cross-sectional data. In the next subsection, we introduce the general panel data model. Let Y_i be the return of firm i. Let $X_i \in \mathbb{R}^2$ be a vector of two characteristics. In this example, we use the linear regression model

$$Y_i = \beta_0 + X_{i,1}\beta_1 + X_{i,2}\beta_2 + \varepsilon_i, \qquad E[\varepsilon_i \mid X_i] = 0$$

The parameters of interest are β_0, β_1, and β_2.

Suppose that for a subset of the data $X_{i,2}$ is not observed, but $X_{i,1}$ and Y_i are always observed. Define $D_i = 0$ if observation i is complete and let $D_i = 1$ if $X_{i,2}$ is missing. We allow data to be missing systematically, but we essentially assume that the data is missing

at random once we condition on the observed characteristics. This assumption consists of two parts. First, we assume that

$$E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}, D_i = 0\right] = 0$$

Since we also assume that $E[\varepsilon_i \mid X_i] = 0$, a sufficient condition for this assumption is that ε_i is independent of D_i conditional on X_i. Notice that this assumption is also implicitly imposed when using the complete subset of observations only. Second, we assume that

$$E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right] = E\left[X_{i,2} \mid X_{i,1}, D_i = 1\right].$$

That is, the conditional mean of $X_{i,2} \mid X_{i,1}$ is the same for the complete and the incomplete subset of the observations. Hence, while D_i may depend on $X_{i,1}$, it cannot depend on $X_{i,2}$.

In the full model, we allow D_i to depend on all variables that are always observed. In particular, in our sample we always observe 18 firm characteristics, including size, book-to-market, beta, idiosyncratic risk, and the return of the previous month, and the probability that an observation is incomplete can be a function these characteristics (see Section 4 for a detailed description of the data and a full list of characteristics). For example, smaller firms may be more likely to have missing values. However, conditional on all of these characteristics, we essentially assume that the data is missing at random. While these assumptions are not directly testable, as explained below, we can test the implications of the assumptions that we use to construct our estimator.

The first part of the assumption implies that

$$E\left[Y_i \mid X_{i,1}, X_{i,2}, D_i = 0\right] = \beta_0 + X_{i,1}\beta_1 + X_{i,2}\beta_2$$

which means that we could estimate the parameters using the subset of complete observations only. This approach is inefficient because it neglects a part of the data that contains both

Y_i and $X_{i,1}$. For this part of the sample, the best predictor of Y_i given $X_{i,1}$ is

$$\begin{aligned} E\left[Y_i \mid X_{i,1}, D_i = 1\right] &= \beta_0 + X_{i,1}\beta_1 + E\left[X_{i,2} \mid X_{i,1}, D_i = 1\right]\beta_2 + E\left[\varepsilon_i \mid X_{i,1}, D_i = 1\right] \\ &= \beta_0 + X_{i,1}\beta_1 + E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right]\beta_2 \end{aligned}$$

In the second line, we used that $E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}\right] = E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}, D_i = 0\right] = 0$ implies that $E\left[\varepsilon_i \mid X_{i,1}, D_i = 1\right] = 0$ given that $D_i = 1$ with positive probability. Notice that $E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right]$ can be estimated using the complete subset of the sample. In this example, we assume that

$$E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right] = \gamma_0 + X_{i,1}\gamma_1$$

in which case

$$E\left[Y_i \mid X_{i,1}, D_i = 1\right] = \beta_0 + X_{i,1}\beta_1 + \left(\gamma_0 + X_{i,1}\gamma_1\right)\beta_2$$

To summarize, we now have the three conditional moment restrictions

$$\begin{aligned} E\left[Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,1}\beta_2 \mid X_{i,1}, X_{i,2}, D_i = 0\right] &= 0 \\ E\left[Y_i - \beta_0 - X_{i,1}\beta_1 - \left(\gamma_0 + X_{i,1}\gamma_1\right)\beta_2 \mid X_{i,1}, D_i = 1\right] &= 0 \\ E\left[X_{i,2} - \gamma_0 - X_{i,1}\gamma_1 \mid X_{i,1}, D_i = 0\right] &= 0 \end{aligned}$$

and the corresponding unconditional moments

$$\left.\begin{aligned} &E\left[\mathbf{1}(D_i = 0)\left(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2\right)\right] = 0 \\ &E\left[\mathbf{1}(D_i = 0)\left(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2\right)X_{i,1}\right] = 0 \\ &E\left[\mathbf{1}(D_i = 0)\left(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2\right)X_{i,2}\right] = 0 \end{aligned}\right\} \text{1st set}$$

$$\left.\begin{aligned} &E\left[\mathbf{1}(D_i = 1)\left(Y_i - \beta_0 - X_{i,1}\beta_1 - (\gamma_0 + X_{i,1}\gamma_1)\,\beta_2\right)\right] = 0 \\ &E\left[\mathbf{1}(D_i = 1)\left(Y_i - \beta_0 - X_{i,1}\beta_1 - (\gamma_0 + X_{i,1}\gamma_1)\,\beta_2\right) X_{i,1}\right] = 0 \end{aligned}\right\} \text{ 2nd set}$$

$$\left.\begin{aligned} &E\left[\mathbf{1}(D_i = 0)\left(X_{i,2} - \gamma_0 - X_{i,1}\gamma_1\right)\right] = 0 \\ &E\left[\mathbf{1}(D_i = 0)\left(X_{i,2} - \gamma_0 - X_{i,1}\gamma_1\right) X_{i,1}\right] = 0 \end{aligned}\right\} \text{ 3rd set}$$

The first and third set of moments point identify β and γ, respectively, and they are based on the complete subset of the data only. The second set of moments uses the incomplete part of the data, is derived from our additional assumptions, and leads to overidentifying restrictions. These overidentifying restrictions are testable and will do so using a modified version of the J-test (see Section A.6 for a derivation of the test statistic in the general model and Section 4 for the test results).

It is also important to mention that the assumption that $E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right]$ is a linear function is not required to derive our unconditional moment conditions. To avoid it, we can use an alternative derivation based on projections, which is less intuitive and discussed in Section A.3 in the appendix.

Based on the moments, there are different ways to estimate the parameters $(\beta_0, \beta_1, \beta_2)$:

1. Use the complete subset of the data and thus, the first set of moments only.

2. Use the optimal GMM estimator that pools all moments and estimates the parameters jointly.

3. Use the third set of moments to estimate γ_0 and γ_1. Then, using the estimated values and the first two sets of moments, estimate β_0, β_1, and β_2. The estimator will depend on the GMM weighting matrix in the second step due to the overidentifying restrictions.

Clearly, option 1 does not use all information contained in the data, while the second option yields the most efficient estimator. However, the moments are nonlinear in the parameters and the optimal GMM estimator does not have a closed form solution. It can therefore be computationally very demanding in large samples and with a large number of predictors, especially when the parameters are estimated for many different time periods. We will now

explain that the third option is an appealing alternative, which is easy to implement and has very good finite sample properties in our simulations.

To gain some intuition, first that suppose γ is known. It then turns out that the optimal GMM estimator based on the first two sets of moments minimizes

$$\sum_{i=1}^{n}\left((1-D_i)\frac{(Y_i-\beta_0-X_{i,1}\beta_1-X_{i,1}\beta_2)^2}{var(\varepsilon_i)}+D_i\frac{(Y_i-\beta_0-X_{i,1}\beta_1-(\gamma_0+X_{i,1}\gamma_1))^2}{var(\varepsilon_i)+var(X_{2_i}-\gamma_0-X_{i,1}\gamma_1)\beta_2^2}\right)$$

and the denominators of the two fractions can be replaced with consistent estimators. We prove this equivalence in a more general setting in Appendix A.4. An alternative way to obtain the estimator is therefore to impute missing values of $X_{i,2}$ with the conditional mean $\gamma_0+X_{i,1}\gamma_1$ and then estimate $(\beta_0,\beta_1,\beta_2)$ using the generalized least squares (GLS) estimator. This estimator then places less weight on observations where $X_{i,2}$ has been imputed. To better understand the reason for down-weighting observations with a missing regressor, define $Z_i=X_{2,i}$ if $D_i=0$ and $Z_i=E[X_{2,i}\mid X_{1,i}]$ if $D_i=1$. We can then write our outcome equation as

$$Y_i=\beta_0+X_{i,1}\beta_1+Z_{i,2}\beta_2+u_i$$

where

$$u_i=\begin{cases}\varepsilon_i & \text{if } D_i=0\\ \varepsilon_i+(X_{i,2}-\gamma_0-X_{i,1}\gamma_1)\,\beta_2 & \text{if } D_i=1\end{cases}$$

Hence, observations with a missing regressor have an unobservable with a larger variance due to the imputation error. The GMM estimator with the estimated optimal weighting matrix is simply the feasible GLS estimator.

When γ_0 and γ_1 have to be estimated as well, the GLS estimator with imputed values is no longer equivalent to the optimal GMM estimator, but it is much easier to implement. We study the loss of efficiency in simulations and find that it is generally small.

The usual GLS standard errors for $(\beta_0,\beta_1,\beta_2)$ are not valid with estimated γ_0 and γ_1.

Instead, we can interpret the GLS estimator as a GMM estimator with a specific weighting matrix and derive the corresponding standard errors.

Yet another alternative is to impute the conditional mean and use the OLS instead of the GLS estimator. This estimator simply ignores the additional variance due to imputation and is also a GMM estimator with a specific weighting matrix. Our simulations suggests that this approach may lead worse statistical properties than the complete case estimator, even when a substantial subset of the data contains missing values. These results are in line with Gourieroux and Monfort (1981) and Nijman and Palm (1988) who find that the GLS estimator is more efficient than the OLS estimator in the presence of one missing pattern.

Finally, a popular approach is to impute the unconditional mean instead of the conditional mean and then estimate $(\beta_0, \beta_1, \beta_2)$ by OLS. Such an approach uses invalid moment conditions and yields a biased estimator, even in this simple example. To see why, write

$$\begin{aligned} Y_i &= \beta_0 + X_{i,1}\beta_1 + X_{i,2}\beta_2 + \varepsilon_i \\ &= \beta_0 + X_{i,1}\beta_1 + E[X_{i,2}]\beta_2 + (X_{i,2} - E[X_{i,2}])\beta_2 + \varepsilon_i \end{aligned}$$

When $D_i = 1$ and $E[X_{i,2}]$ is imputed, the unobservable becomes $(X_{i,2} - E[X_{i,2}])\beta_2 + \varepsilon_i$. But

$$E[(X_{i,2} - E[X_{i,2}])\beta_2 + \varepsilon_i \mid X_{i,1}, D_i = 1] = E[(X_{i,2} - E[X_{i,2}]) \mid X_{i,1}]\beta_2$$

which is not 0 unless $\beta_2 = 0$ or $X_{i,2}$ is mean independent of $X_{i,1}$.

2.2 General Model

We now consider the general panel data model. Let Y_{it} be the return of firm i at time t and let $X_{it} \in \mathbb{R}^K$ be a vector of characteristics. We assume that

$$Y_{it} = \sum_{k=1}^{K} X_{it,k}\beta_{t,k} + \varepsilon_{it}, \qquad E[\varepsilon_{it} \mid X_{it}] = 0$$

That is,

$$E[Y_{it} \mid X_{it}] = \sum_{k=1}^{K} X_{it,k}\beta_{t,k}$$

While the conditional mean function is linear in the parameters, the regressors may include nonlinear functions of the characteristics. Also notice that the vector X_{it} contains a constant. In this model all parameters may depend on t and can be estimated period by period. When the parameters are time invariant, an alternative is to pool data from different time periods.

We allow the subset of observed regressors to vary by observation. Specifically, we assume that there are L different missing patterns where for each missing pattern a different subset of regressors is observed. Let $D_{it} = l$ if observation i at time t has missing pattern l. In this case we denote by $X_{it}^{(l)} \subseteq X_{it}$ the subvector of observed characteristics and by $I_t^{(l)} \subseteq \{1, \ldots, K\}$ the corresponding indices. As before, for complete observations we use $D_{it} = 0$, and in this case $X_{it}^{(0)} = X_{it}$.

Similar to the simple example, we can allow data to be missing systematically, but we impose two conditions. First, we assume that

$$E\left[\varepsilon_{it} \mid X_{it}^{(l)}, D_{it} = l\right] = 0$$

for all $l = 0, 1, \ldots, L$, which implies that

$$E\left[Y_{it} \mid X_{it}^{(l)}, D_{it} = l\right] = \sum_{k=1}^{K} E[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l]\beta_{t,k}$$

Second, we assume that

$$E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l\right] = E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = 0\right]$$

for all $l = 0, 1, \ldots, L$. As discussed above, these assumptions allow D_{it} to depend on regressors that are always observed, and since we observe 18 important firm characteristics, these assumptions seems to be reasonable in our empirical application (see Section 4 for further

discussions). We can relax these assumptions by conditioning on additional characteristics that D_{it} may depend on, such as industry dummies (see Sections 2.3.2 for more details).

Now define

$$Z_{it,k}^{(l)} = E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l\right] = E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = 0\right]$$

Notice that if $X_{it,k} \subseteq X_{it}^{(l)}$, then $Z_{it,k}^{(l)} = X_{it,k}$ is observed. If $X_{it,k} \not\subseteq X_{it}^{(l)}$, then $Z_{it,k}^{(l)}$ is not observed and needs to be estimated, which we do using the subset of observation at time t where all characteristics are observed (i.e. the complete cases at time t). Under certain assumptions, we could also use observed covariates from other time periods for imputation, as discussed in Section 2.3.2. Here, we assume that for all $l = 1, \ldots, L$ and $k \notin I_t^{(l)}$

$$E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l\right] = X_{it}^{(l)\prime}\gamma_t^{(l,k)}.$$

Alternatively, we could interpret $X_{it}^{(l)\prime}\gamma_t^{(l,k)}$ as a linear projection in which case we do not require a parametric conditional mean assumption. Using our assumptions, we can write for all $l = 1, \ldots, L$ and $k \notin I_t^{(l)}$

$$E\left[X_{it,k} - X_{it}^{(l)\prime}\gamma_t^{(l,k)} \mid X_{it}^{(l)}, D_{it} = 0\right] = 0.$$

In addition, for $l = 0$ we have

$$E\left[Y_{it} - \sum_{k=1}^{K} \beta_{t,k} X_{it}^{(0)} \mid X_{it}^{(0)}, D_{it} = 0\right] = 0$$

because in this case all characteristics are observed. Finally, for $l = 1, \ldots, L$, we have

$$E\left[Y_{it} - \sum_{k=1}^{K} \beta_{t,k} Z_{it,k}^{(l)} \mid X_{it}^{(l)}, D_{it} = l\right] = 0$$

where

$$Z_{it,k}^{(l)} = E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = 0\right] = \begin{cases} X_{it,k} & \text{if } k \in I_t^{(l)} \\ X_{it}^{(l)\prime}\gamma_t^{(l,k)} & \text{if } k \notin I_t^{(l)} \end{cases}$$

To estimate the parameters, we use the following unconditional moments:

$$E\left[\mathbf{1}(D_{it}=0)\left(Y_{it} - \sum_{k=1}^{K}\beta_{t,k}X_{it,k}^{(0)}\right)X_{it}^{(0)}\right] = 0 \tag{1}$$

$$E\left[\mathbf{1}(D_{it}=l)\left(Y_{it} - \sum_{k=1}^{K}\beta_{t,k}Z_{it,k}^{(l)}\right)X_{it}^{(l)}\right] = 0 \qquad l=1,\ldots,L \tag{2}$$

$$E\left[\mathbf{1}(D_{it}=0)\left(X_{it,k} - X_{it}^{(l)\prime}\gamma_t^{(l,k)}\right)X_{it}^{(l)}\right] = 0 \qquad l=1,\ldots,L \text{ and } k \notin I_t^{(l)} \tag{3}$$

These three sets of moment conditions are analogous to the ones we used in the simple example. The moment conditions in (1) and (3) point identify β_t and $\gamma_t^{(l,k)}$, respectively, and are based on the complete subset of the data only. The moment conditions in (2) are additional restrictions that yield efficiency gains.

Just as in the simple example, there are different ways to estimate the parameters. One option that we pursue in the application is to estimate $\gamma_t^{(l,k)}$ using the third set of moments and then use the first two sets of moments, along with the estimates of $\gamma_t^{(l,k)}$, to estimate β_t. In the second step, we use the weight matrix that is optimal with known $\gamma_t^{(l,k)}$. As before, this estimator is equivalent to the GLS estimator where missing value are replaced with the estimated mean, conditional on the set of observed regressors. The estimator accounts for the additional variance due to imputation. In general, the more regressors are imputed, the less weight is placed on an observation. We derive the large sample distribution of the estimator in the Appendix A.5 and provide plug-in estimators for the standard errors.

A potential alternative is the optimal GMM estimator, which can be hard to compute in practice because the objective function is not quadratic in the parameters. In fact, in our empirical application with large numbers of observations and regressors, this estimator is computationally infeasible.

2.3 Extensions

2.3.1 High-Dimensional and Nonlinear Models

Our two-step estimator can also be applied in high-dimensional and nonlinear models. Recall that we estimate conditional mean functions in the first step. Instead of using a linear regression model, we could also employ machine learning methods, such as a neural networks or random forests. Within the linear framework, but with a number large of regressors, we could also use a penalized estimator such as the LASSO estimator or the Ridge estimator.

Constructing a consistent estimator in the second step is more complicated. To illustrate potential problems, let's return to the simple cross-sectional example, and suppose that

$$Y_i = \beta_0 + X_{i,1}\beta_1 + X_{i,1}^2\beta_2 + X_{i,2}\beta_3 + X_{i,2}^2\beta_4 + \varepsilon_i, \qquad E[\varepsilon_i \mid X_i] = 0$$

As before, $X_{i,1}$ is always observed, but $X_{i,2}$ is not, and $D_i = 1$ denotes the case where $X_{i,2}$ is missing. We then have

$$E[Y_i \mid X_{i,1}, D_i = 1] = \beta_0 + X_{i,1}\beta_1 + X_{i,1}^2\beta_2 + E[X_{i,2} \mid X_{i,1}]\beta_3 + E[X_{i,2}^2 \mid X_{i,1}]\beta_4$$

Hence, we could impute estimates of $E[X_{i,2} \mid X_{i,1}]$ and $E[X_{i,2}^2 \mid X_{i,1}]$ for $X_{i,2}$ and $X_{i,2}^2$, respectively, and estimate the parameters by GLS.

A potential alternative could be to define to $Z_i = X_{2,i}$ if $D_i = 0$ and $Z_i = E[X_{2,i} \mid X_{1,i}]$ if $D_i = 1$ and regress Y_i on $X_{1,i}$, $X_{1,i}^2$, Z_i, and Z_i^2. However, since $Z_i^2 = E[X_{i,2} \mid X_{i,1}]^2 \neq E[X_{i,2}^2 \mid X_{i,1}]$, the resulting estimator is inconsistent. These issues carry over to other nonlinear models.

One possibility to allow for nonlinearities and models selection simultaneously, which we use in our application, is the group LASSO estimator of Freyberger et al. (2020). Similar to the simple example above, in the first step one needs to impute conditional expectations of nonlinear transformations of the regressors (such as polynomials or splines). The second step is then simply the estimator of Freyberger et al. (2020), with the possibility of down-

weighting observations with imputed values. This approach not only allows for nonlinearities but also pre-specified interactions.

2.3.2 Additional covariates

We could use additional covariates to relax our missing at random assumptions or to obtain better imputations. In our application, these variables might include additional firm characteristics or characteristics from other periods. We now briefly describe different approaches using our introductory example and discuss the details in Section A.2 in the appendix.

Consider again the simple model

$$Y_i = \beta_0 + X_{i,1}\beta_1 + X_{i,2}\beta_2 + \varepsilon_i, \qquad E[\varepsilon_i \mid X_i] = 0$$

where $X_{i,1}$ is a always observed, but $X_{i,2}$ might be missing. Let $D_i = 0$ if observation i is complete and let $D_i = 1$ if $X_{i,2}$ is missing. To derive the estimator, our two main assumptions on the missing patterns are:

$$E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}, D_i = 0\right] = 0$$

and

$$E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right] = E\left[X_{i,2} \mid X_{i,1}, D_i = 1\right],$$

and a sufficient condition for these assumptions is

$$D_i \perp\!\!\!\perp Y_i, X_{i,2} \mid X_{i,1}$$

Let V_i be an additional vector of covariates that is always observed, such as industry dummies, which do not have a direct effect on the outcomes. We can then weaken the conditional independence assumption to

$$D_i \perp\!\!\!\perp Y_i, X_{i,2} \mid X_{i,1}, V_i.$$

One can then show that

$$E\left[\frac{(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2)}{P(D_i = 0 \mid X_{i,1}, V_i)} \mid X_{i,1}, X_{i,2}, D_i = 0\right] = 0$$

and

$$E\left[\frac{(Y_i - \beta_0 - X_{i,1}\beta_1 - E[X_{i,2} \mid X_{i,1}, V_i, D_i = 0]\beta_2)}{P(D_i = 1 \mid X_{i,1}, V_i)} \mid X_{i,1}, D_i = 1\right] = 0$$

Hence, we impute $X_{i,2}$ using both $X_{i,1}$ and V_i and then use moments as before, but weighted by the inverse of the conditional probability of D_i (inverse propensity score weighting).

This previous approach does not require an assumption on how V_i relates to ε_i. Now suppose we also assume that $E[\varepsilon_i \mid X_i, V_i] = 0$, which might be reasonable for industry dummies and characteristics from other time periods. It can then be shown that

$$E\left[Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2 \mid X_{i,1}, X_{i,2}, V_i, D_i = 0\right] = 0$$

and

$$E\left[Y_i - \beta_0 - X_{i,1}\beta_1 - E[X_{i,2} \mid X_{i,1}, V_i, D_i = 0]\beta_2 \mid X_{i,1}, V_i, D_i = 1\right] = 0$$

Again, we can impute $X_{i,2}$ using both $X_{i,1}$ and V_i, but with the additional assumption, the moments are more informative (because we also condition on V_i) and we avoid inverse propensity score weighting.

3 Simulations

We now illustrate the statistical properties of our estimator and alternative approaches in various Monte Carlo simulations. We start with a low-dimensional setting and mainly focus on efficiency and inference. We then consider a high-dimensional setting and discuss model selection and out-of-sample predictions.

3.1 Low-dimensional setting

We start with the model

$$Y_i = \sum_{k=1}^{K} X_{i,k}\beta_k + \varepsilon_i, \qquad E[\varepsilon_i \mid X_i] = 0$$

where $K = 5$ and $X_{i,1} = 1$. We let $X_{i,2}, \ldots, X_{i,K}$ be jointly normally distributed with means of 0 and $cov(X_{i,k}, X_{i,j}) = 0.9^{|k-j|}$ and $\varepsilon_i \sim N(0,1)$. The true values of the coefficients are $\beta = (1, 0.5, 1, -1, 3)'$.

We consider three different types of missing patterns given in Figure 1. In the first setup, there is one subset of complete observations ($l = 0$) and one subset where $X_{i,3}$ is missing ($l = 1$). In the second setup, there is one subset of complete observations ($l = 0$) and one

Figure 1: Missing Patterns

This figure shows examples for missing patterns. For all settings $l = 0$ denote the complete case, i.e. the fraction of that data for which all covariates (and the outcome) are observed. In Setup 1, some part of the data are missing the third covariate ($X_{i,3}$). In Setup 2, some part of the data are missing three covariates ($X_{i,2},\ X_{i,4},\ X_{i,5}$). Setup 3 is a general missing pattern, some part of the data are missing the third covariate ($X_{i,3}$), another part are missing the fifth covariate ($X_{i,3}$) and another part of the data are missing the fourth and the fifth covariate ($X_{i,4},\ X_{i,5}$).

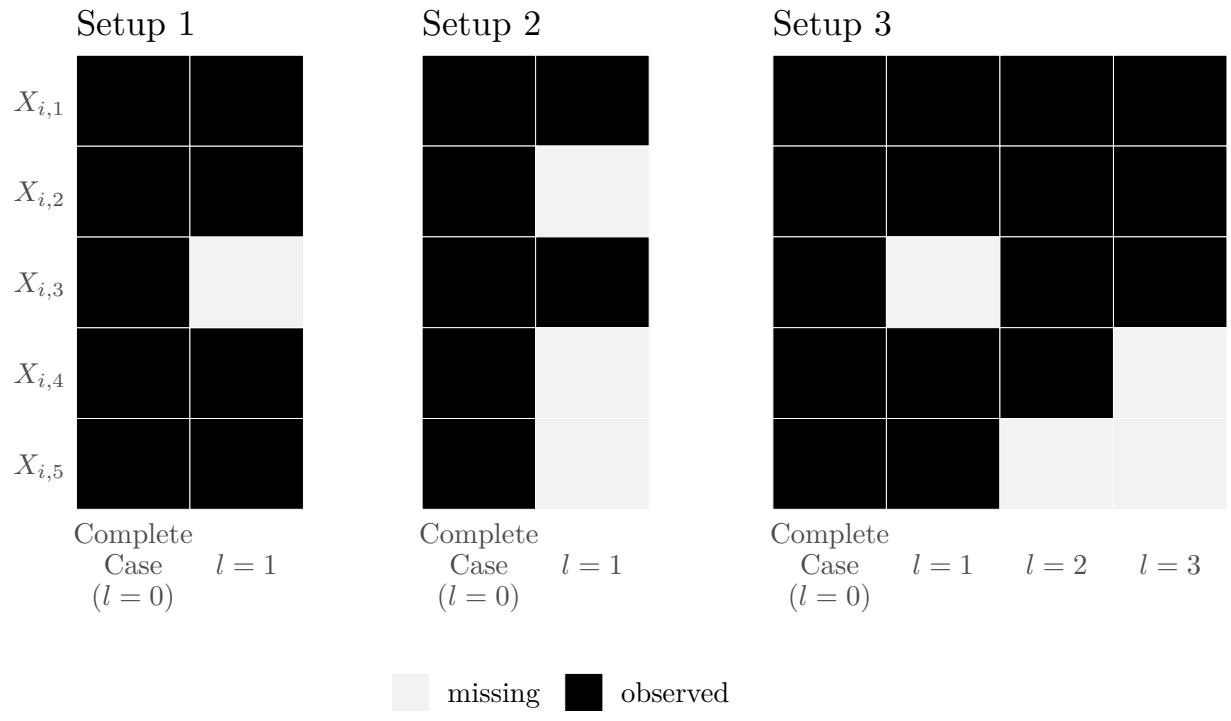


Table 1: Simulation - Coverage and Length of Confidence Intervals

This table shows the coverage probabilities of 90% confidence intervals and the length of the confidence intervals for the three sets of missing patterns described in Figure 1 when 50% of the data are missing at random.

	Complete Case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. Mean	
	Cover	Length	Cover	Length	Cover	Length	Cover	Length	Cover	Length
	Setup 1									
β_1	0.908	0.147	0.888	0.109	0.885	0.109	0.891	0.110	0.890	0.109
β_2	0.893	0.337	0.905	0.291	0.903	0.291	0.902	0.293	0.000	0.193
β_3	0.896	0.453	0.902	0.452	0.902	0.452	0.896	0.454	0.000	0.208
β_4	0.905	0.453	0.901	0.367	0.900	0.368	0.905	0.370	0.000	0.297
β_5	0.907	0.337	0.903	0.249	0.900	0.249	0.905	0.252	0.909	0.251
	Setup 2									
β_1	0.908	0.147	0.897	0.136	0.909	0.139	0.893	0.190	0.824	0.160
β_2	0.893	0.337	0.892	0.337	0.894	0.337	0.893	0.338	0.000	0.399
β_3	0.896	0.453	0.896	0.449	0.895	0.451	0.888	0.470	0.000	0.214
β_4	0.905	0.453	0.897	0.453	0.902	0.453	0.905	0.455	0.000	0.614
β_5	0.907	0.337	0.906	0.337	0.905	0.337	0.908	0.338	0.979	0.518
	Setup 3									
β_1	0.905	0.147	0.907	0.121	0.901	0.122	0.912	0.150	0.862	0.164
β_2	0.895	0.337	0.893	0.293	0.891	0.297	0.903	0.369	0.001	0.311
β_3	0.901	0.453	0.903	0.424	0.903	0.428	0.914	0.516	0.481	0.331
β_4	0.902	0.454	0.896	0.402	0.892	0.404	0.900	0.439	0.000	0.378
β_5	0.905	0.337	0.901	0.294	0.900	0.295	0.906	0.299	0.000	0.341

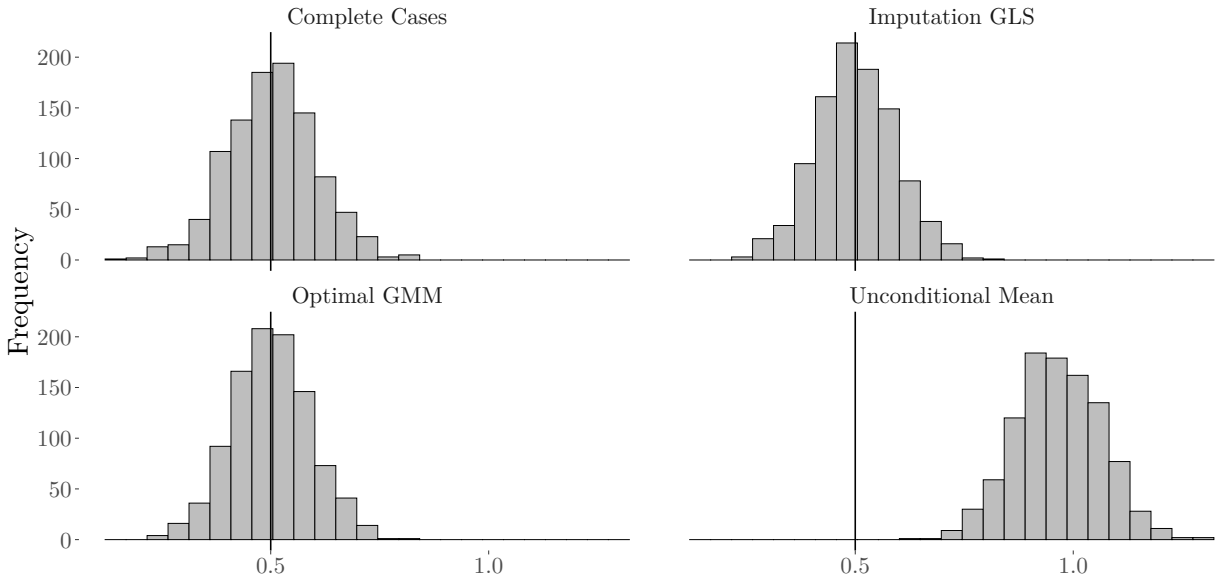
subset where all regressors except for $X_{i,3}$ are missing ($l = 1$). In the last setup, there are four subsets of the data with different missing patterns.

Table 1 shows coverage rates and average lengths of 90% confidence intervals for the different setups when the complete sample contains 50% of the observations. The sample size is $n = 1,000$ and we ran $1,000$ Monte Carlo simulations. We report results for the estimator that only uses the complete subset, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.

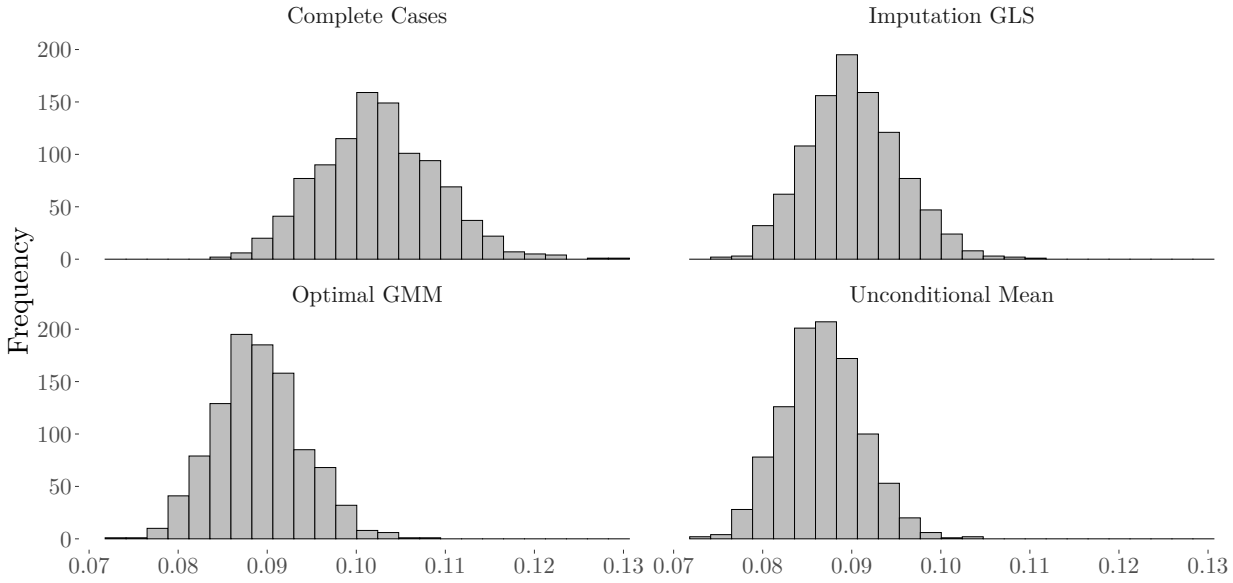
In setup 1 only $X_{i,3}$ may be missing. Comparing the complete case and the optimal GMM estimator, we can see that for all coefficients, except for β_3, the average lengths of the confidence intervals decreases substantially. Both the GLS and the OLS estimator with conditional mean imputation perform almost as well as the optimal GMM estimator. All of these four estimators have coverage probabilities close to 90%. The estimator based

Figure 2: Simulation - Histograms for Setup 3

This figure shows histograms of the repeated sample distribution for estimates of β_2 (panel a) and standard errors of $\hat{\beta}_2$ (panel b) in the general missing pattern (Setup 3 of Figure 1). The vertical bar indicates the correct value for the parameter, $\beta_2 = 0.5$.



(a) Estimates of β_2



(b) Standard Errors of $\hat{\beta}_2$

on unconditional mean imputation has low coverage rates, which is due to the bias of the estimator (to be discussed in more detail below). Interestingly, the confidence intervals can be much narrower than those of the optimal GMM estimator. The reason is that the regressors appear less correlated once the unconditional mean is imputed. In setup 2, more regressors are missing and the gains from imputation are lower. Setup 3 has more complicated missing patterns, but the results are overall similar to those of setup 1. One difference is that the GLS estimator performs much better than the OLS estimator. In fact the average length of the confidence intervals of the OLS estimator can be larger than those of the complete case estimator. While the OLS estimator uses more moment conditions, it combines them in an inefficient way. To illustrate these points further, Figure 2 shows histograms of the estimates of β_2 and the corresponding standard errors for setup 3. We can see that the imputation estimator and the optimal GMM estimator perform very similarly and are both more efficient than the estimator based on the complete case. In addition, mean imputation yields both a bias and artificially small standard errors. Table 2 shows results for setup 3 with a different

Table 2: Simulation - Coverage and Length of Confidence Intervals for Varying Missing Percentage

This table shows the coverage probabilities of 90% confidence intervals and the length of the confidence intervals for the general missing pattern (setup 3) in Figure 1 when 75% and 25% of the data are missing at random.

	Complete Case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. Mean	
	Cover	Length	Cover	Length	Cover	Length	Cover	Length	Cover	Length
	25% complete									
β_1	0.904	0.208	0.881	0.141	0.876	0.149	0.902	0.204	0.758	0.173
β_2	0.892	0.475	0.872	0.366	0.881	0.397	0.894	0.536	0.000	0.316
β_3	0.891	0.642	0.879	0.563	0.884	0.605	0.895	0.792	0.004	0.332
β_4	0.883	0.642	0.893	0.501	0.896	0.519	0.915	0.610	0.000	0.365
β_5	0.906	0.478	0.887	0.348	0.892	0.353	0.907	0.362	0.000	0.373
	75% complete									
β_1	0.889	0.120	0.890	0.110	0.887	0.111	0.888	0.123	0.887	0.145
β_2	0.889	0.275	0.893	0.258	0.890	0.259	0.910	0.290	0.062	0.294
β_3	0.918	0.370	0.905	0.359	0.905	0.360	0.891	0.394	0.792	0.338
β_4	0.898	0.370	0.888	0.352	0.891	0.352	0.890	0.366	0.000	0.375
β_5	0.907	0.275	0.901	0.261	0.899	0.261	0.907	0.263	0.000	0.310

percentage of complete observations. When the fraction of complete observations is low, the relative gains from imputation are larger and the differences between OLS and GLS are much more striking.

The poor coverage probability obtained from imputing unconditional means is due to the large bias of the estimator. We show the biases for setup 3, again with a different percentage of complete observations, in Table 3. Even when 75% of the sample is complete, the bias is substantial. The biases of all other estimators are negligible. In that table, we also report the root mean squared errors (RMSE) of the different estimators. We can see that the optimal GMM estimator can be much more precise than the estimator based on the complete sample. The GLS estimator is almost as precise as the optimal GMM estimator and generally much more precise than the OLS estimator.

Table 4 shows results for setup 3 with independent regresssors when the complete sample

Table 3: Simulation - Bias and Model Fit for a General Missing Pattern

This tables shows the bias in the estimated coefficients and the root mean-squared error for the general missing pattern (Table 1, setup 3) when different percentages of the data are missing.

	Complete Case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. Mean	
	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
	25% complete									
β_1	0.064	0.003	0.045	0.001	0.047	0.000	0.063	0.000	0.072	0.000
β_2	0.146	0.006	0.121	-0.009	0.131	-0.003	0.170	-0.007	0.516	0.507
β_3	0.195	-0.005	0.179	0.018	0.190	0.008	0.240	0.015	0.428	0.417
β_4	0.201	-0.004	0.158	-0.015	0.157	-0.001	0.175	-0.002	1.355	1.350
β_5	0.144	0.000	0.108	0.005	0.107	-0.003	0.106	-0.006	1.357	-1.353
	50% complete									
β_1	0.045	0.001	0.037	0.000	0.037	0.000	0.045	0.000	0.056	0.002
β_2	0.101	0.004	0.090	0.001	0.091	0.001	0.111	-0.001	0.476	0.466
β_3	0.135	-0.005	0.126	-0.003	0.128	-0.003	0.150	-0.002	0.204	0.175
β_4	0.137	-0.001	0.126	0.000	0.126	0.002	0.135	0.003	1.125	1.115
β_5	0.102	0.001	0.090	0.001	0.091	0.000	0.090	-0.001	1.255	-1.250
	75% complete									
β_1	0.037	0.000	0.035	0.000	0.035	0.000	0.038	0.000	0.046	0.002
β_2	0.084	0.000	0.080	0.000	0.080	0.000	0.087	0.001	0.309	0.294
β_3	0.110	-0.001	0.110	0.000	0.110	0.000	0.120	0.000	0.130	0.053
β_4	0.115	0.000	0.111	0.001	0.111	0.001	0.114	0.001	0.717	0.701
β_5	0.083	-0.001	0.081	-0.002	0.081	-0.002	0.080	-0.002	0.815	-0.806

contains 50% of the observations. In this case, the conditional expectations of the regressors are equal to the unconditional ones and thus, imputing unconditional means leads to valid moment conditions. However, the moment conditions are combined in an inefficient way because observations with missing regressors have the same weight as complete observations. Using the GLS estimator or the optimal GMM estimator leads to a much better performance. Moreover, the standard errors with unconditional mean imputation are incorrect because they do not account for the fact that the imputed means are estimated.

One setting where unconditional mean imputation outperforms the other methods is when all regression coefficients in front of regressors that have missing values are equal to 0. In this case, unconditional mean imputation leads to correct moment conditions, as discussed at the end of section 2.1. Moreover, imputing the conditional or the unconditional mean does not increase the variance of the error term and thus, there are no benefits from using GLS. We show simulation results in Table 5 for setup 3 when $\beta = (1, 0.5, 0, 0, 0)'$ and the complete sample contains 50% of the observations. In this case, imputing the unconditional means decreases the correlation between the regressors, which reduces the variance of the estimated coefficients and the length of the confidence intervals. Since the moment conditions are valid, the estimator is also asymptotically unbiased. Consequently, it also has a lower mean squared error compared the estimators that impute conditional means. Clearly, in applications we do not know a priori if coefficients are equal to 0, and we should therefore not rely on the mean

Table 4: Simulation - Coverage and Length of Confidence Intervals with Independent Regresssors

This table shows the coverage probabilities of 90% confidence intervals and the length of the confidence intervals for the general missing pattern (setup 3) in Figure 1 when all regressors are independent.

	Complete Case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. Mean	
	Cover	Length	Cover	Length	Cover	Length	Cover	Length	Cover	Length
β_1	0.905	0.147	0.899	0.135	0.895	0.136	0.877	0.266	0.805	0.217
β_2	0.910	0.147	0.891	0.134	0.894	0.136	0.920	0.265	0.907	0.216
β_3	0.889	0.147	0.882	0.143	0.889	0.145	0.913	0.309	0.906	0.250
β_4	0.897	0.147	0.902	0.135	0.902	0.136	0.905	0.220	0.910	0.197
β_5	0.906	0.147	0.897	0.136	0.895	0.138	0.898	0.149	0.898	0.143

Table 5: Simulation - Coverage and Length of Confidence Intervals when $\beta = (1, 0.5, 0, 0, 0)'$

This table shows the coverage probabilities of 90% confidence intervals and the length of the confidence intervals for the general missing pattern (setup 3) in Figure 1 when all regressors that may be missing do not affect the outcome.

	Complete Case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. Mean	
	Cover	Length	Cover	Length	Cover	Length	Cover	Length	Cover	Length
β_1	0.905	0.147	0.881	0.103	0.880	0.103	0.897	0.104	0.898	0.104
β_2	0.895	0.337	0.889	0.246	0.894	0.246	0.897	0.249	0.883	0.197
β_3	0.901	0.453	0.902	0.366	0.904	0.366	0.914	0.370	0.888	0.210
β_4	0.902	0.454	0.894	0.377	0.897	0.377	0.895	0.381	0.887	0.239
β_5	0.905	0.337	0.892	0.290	0.893	0.290	0.902	0.293	0.900	0.216

imputation to deliver satisfactory results. As discussed below, to determine which regressors are irrelevant, we can carry out model selection to obtain a smaller model.

3.2 High-dimensional setting

We now again simulate data from the linear model

$$Y_i = \sum_{k=1}^{K} X_{i,k}\beta_k + \varepsilon_i, \qquad E[\varepsilon_i \mid X_i] = 0$$

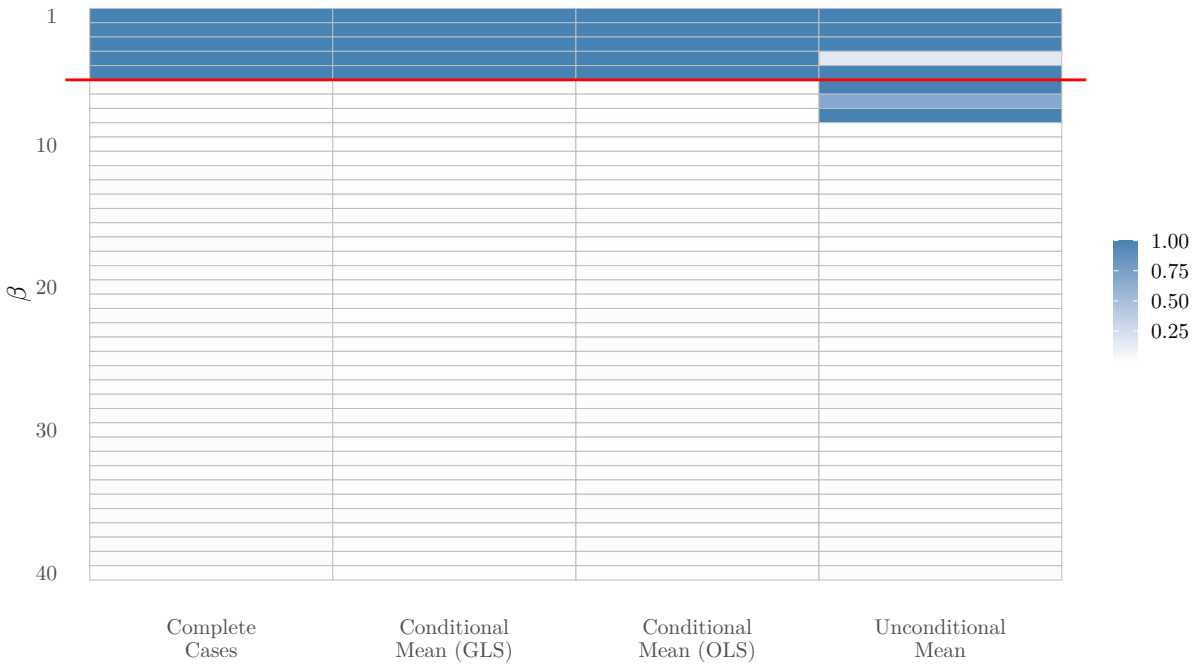
but we use $K = 40$ regressors. As before, $X_{i,1} = 1$, $X_{i,2}, \ldots, X_{i,K}$ are jointly normally distributed with means of 0 and $cov(X_{i,k}, X_{i,j}) = 0.9^{|k-j|}$, and $\varepsilon_i \sim N(0, 1)$.

We also again choose the first five elements of β to be $(1, 0.5, 1, -1, 3)'$ and the remaining 35 elements are all equal to 0. For the first five regressors, we use the same missingness pattern as in Setup 3 above with $L = 3$. When $l = 0$ or $l = 3$, all other regressors are observed as well. When $l = 1$, $X_{i,36}, X_{i,37}, \ldots, X_{i,40}$ are not observed and when $l = 2$, $X_{i,6}$ and $X_{i,7}$ are not observed. The probability that an observation is missing now varies with $X_{i,2}$, which is always observed. In particular, observations with high values of $X_{i,2}$ are more likely to be complete.

We now consider four different estimators, namely the estimator that only uses the complete subset, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean. For all estimators, we estimate the parameters

Figure 3: Model Selection - Sparse Model

This figure shows the model selection results for the sparse example (Section 3). The darker the color, the more frequent a particular model estimates a non-zero β_i. In the true model, the first five betas are non-zeros (above the red line), whereas the rest is equal to zero.

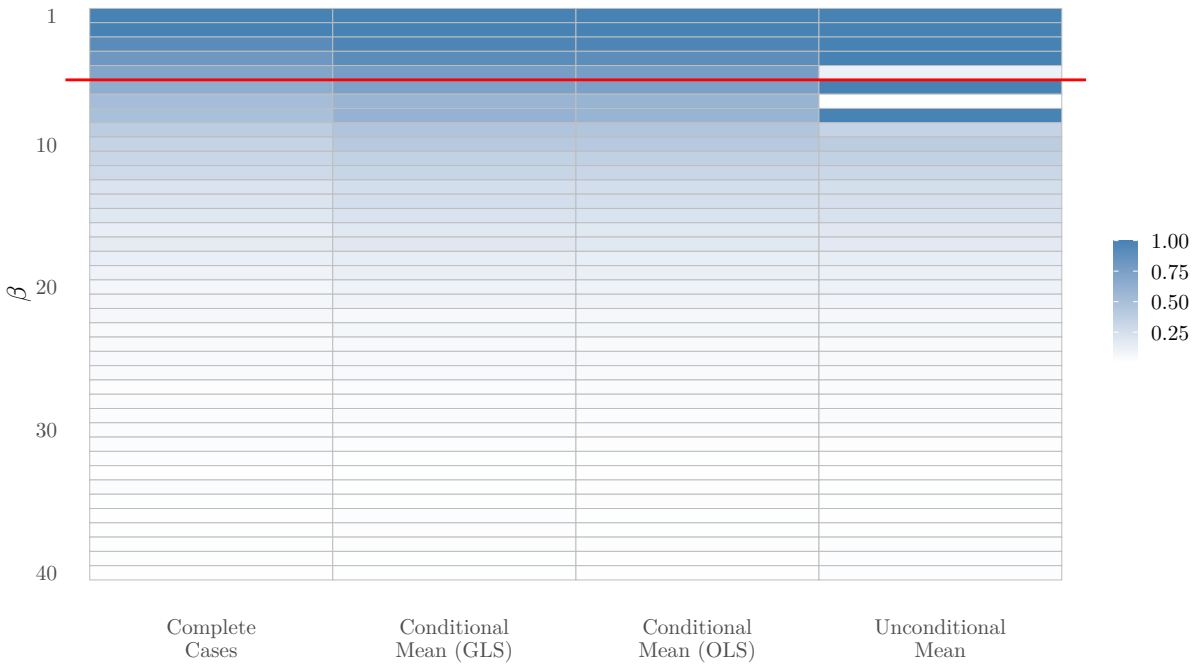


$\beta_1, \beta_2, \ldots, \beta_{40}$ using the adaptive-post-LASSO method and choose the penalty parameter based on the BIC. We use the same LASSO procedure for the imputation step. All estimator are very easy to implement using standard software.

Figure 3 illustrates the frequency with which the different methods select regressors. The complete case estimator and both conditional mean imputation estimators select the variables with nonzero coefficients with very large probability and typically set coefficients of irrelevant variables to 0. Unconditional mean imputation tends to set the estimated value of β_4 to 0 and instead frequently includes three of the irrelevant regressors. The mean squared prediction errors (MSPEs) of the four methods are 1.0187, 1.0105, 1.0141, and 1.7059, respectively, showing that the imputation GLS estimator performs best and unconditional mean imputation performs worst. For the out-of-sample predictions, we generate a new sample of complete observations with a sample size of $5,000$.

Figure 4: Model Selection - Dense Model

This figure shows the model selection results for the non-sparse example (Section 3). The darker the color, the more frequent a particular model estimates a non-zero β_i. In the true model, the first five betas are non-zeros (above the red line), whereas the rest is equal to zero.

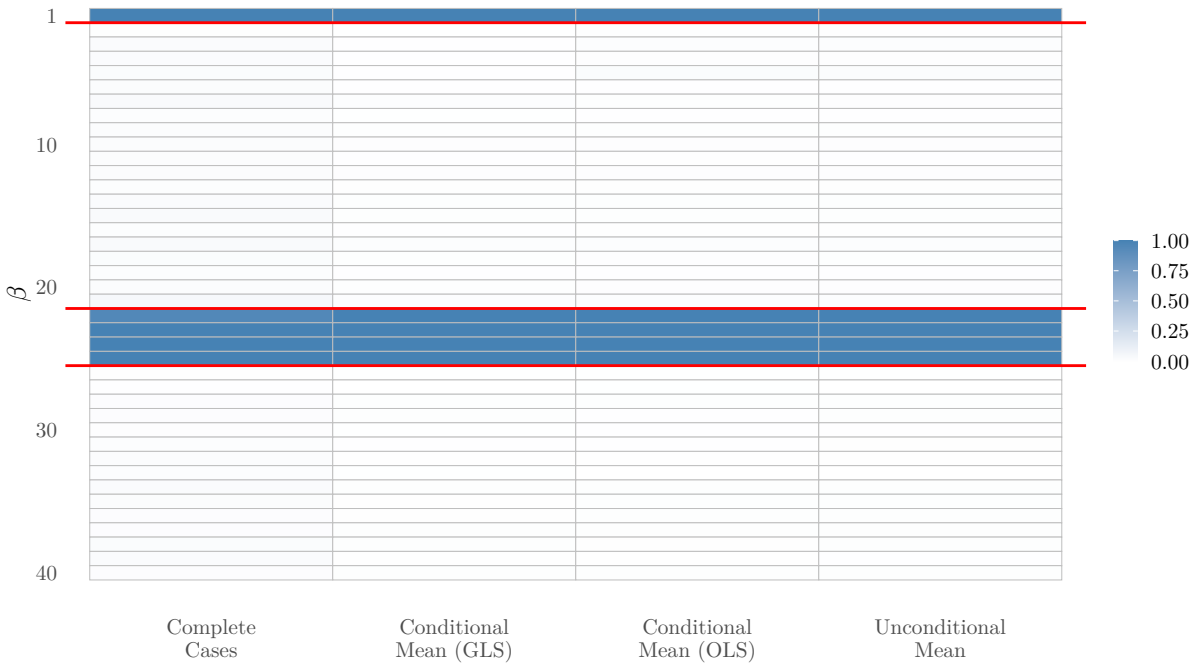


The previous case was sparse, in the sense that only 5 coefficients are not equal to 0. We now assume that $\beta_j = 0.8^j$, but leave all other features of the data generating process unchanged. The selection results are illustrated in Figure 4. We can see that for the complete case estimator and both conditional mean imputation estimators, the larger a coefficient, the more likely it is not set to 0. This monotonicity does not hold for unconditional mean imputation. Here the estimated value of β_5 is often set to 0, but regressors with smaller coefficients are included much more frequently. The MSPEs of the four methods are 1.0552, 1.0345, 1.0343, and 1.0887, respectively.

Another case where imputation works particularly well is when regressors with missing values to do not have an impact on the outcome. To illustrate this situation, again consider the sparse setting, but let $\beta_1 = 1$, $\beta_2 = \beta_3 = \cdots = \beta_{21} = 0$, $(\beta_{22}, \ldots, \beta_{25}) = (0.5, 1, -1, 3)$, and the remaining 15 elements be all equal to 0. The results are reported in Figure 5. In this

Figure 5: Model Selection - Missing Regressor Irrelevant

This figure shows the model selection results for a non-sparse example (Section 3) when none of the potentially missing regressors affect the outcome The darker the color, the more frequent a particular model estimates a non-zero β_i. The non-zero coefficients of the true model are separated with red lines.

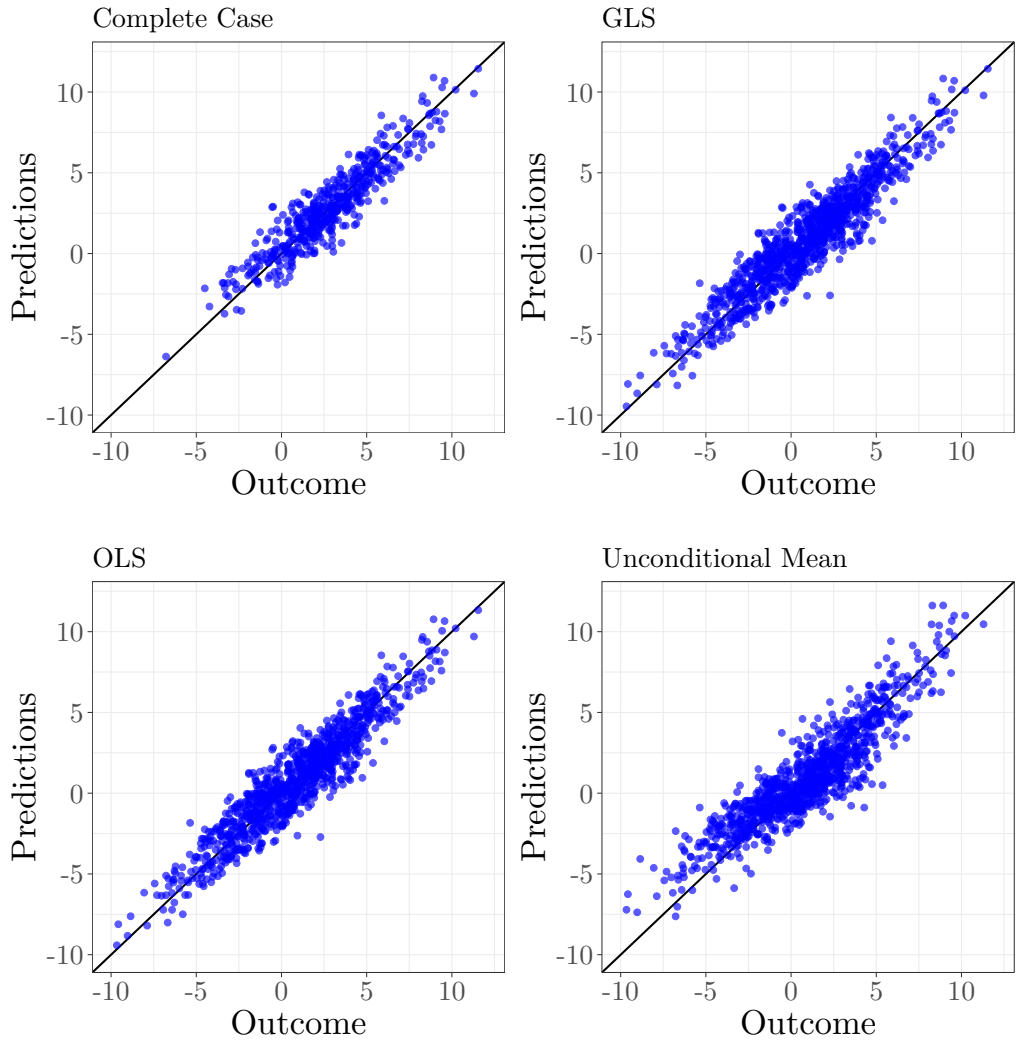


case, the imputation methods mostly ignore regressors with missing values, but can make use of the full data set. The MSPEs of the four methods are 1.0206, 1.0076, 1.0075, and 1.0072 respectively. Therefore, all imputation methods perform similarly well and outperform the complete case.

One advantage of imputations is that the whole sample can be used for predictions. The MSPE for the subset of non-complete observations is typically higher than for the complete observations, but the complete case might miss particularly interesting parts of the conditional distribution of outcomes. To illustrate this feature, Figure 6 plots the out-of-sample outcomes against the predictions obtained with the different methods. Recall that the probability that an observation is completely observed depends on $X_{i,2}$. When using imputations, we make predictions for all outcomes, even when some regressors are missing. Comparing panels (a) and (b) we can see that the observations with missing regressors tend

Figure 6: Outcomes versus predictions

This figure shows out-of-sample outcomes against the predictions when the probability of an observation being complete depends on the regressors.



to have lower outcomes. There are two important implications for out-of-sample portfolio sorts that we will discuss in more detail in our application. First, when using imputations, we have a large number of observations we can form portfolios with. Therefore, the number of observations corresponding to the 10% highest and lowest predictions is much higher when using imputations, and portfolio variances will be lower. When we instead fix the number of observations in each portfolio (instead of the %), we will observe a large difference in portfolio returns. Second, when the probability that an observation is missing depends on the observed covariates, the complete case misses a systematically different part of the

distribution of outcomes, and not just a random sample. In this case, differences in portfolio returns will be even more distinctive. Finally, panel (d) shows that imputing unconditional means yields biased predictions. However, since predictions and outcomes are still positively related, portfolio formed based on these predictions will be very similar to those obtained with conditional mean imputation.

4 Data

We use stock returns, volume and price data from the Center for Research in Security prices (CRSP) monthly stock file. Following standard conventions in the literature, we restrict the analysis to common stocks of firms incorporated in the US trading on NYSE, Nasdaq or Amex. Balance sheet data is obtained from Compustat.

In order to avoid potential forward looking biases, we lag all characteristics that build on Compustat annual by at least six months and all that build on Compustat quarterly by at least four months. Our main dataset is obtained from Chen and Zimmermann (2021) and consists of 40 firm characteristics that are available from 1965 - 2020. The firm characteristics feature a combination of accounting information as well as versions of momentum and functions of trading volume. Table A.1 provides an overview of the characteristics we use in our main empirical analysis. It should be noted that these predictors are not a randomly selected set of features, but have been found to be successful cross-sectional predictors in the literature.

Table A.1 also shows the fraction of missing values per characteristic. Overall, we have a total of 3,051,103 firm month observations. Fama and French (1992) define the benchmark for empirical analysis of the cross-section of expected returns. We follow them and require that a minimum of information is available for each firm. As Fama-French we require the inputs (market beta, size, and book-to-market) of the Fama-French 3 factor model to be available for all firms. When we condition on firms having `beta`, `bm` and `size` available, we have 2,315,566 firm month observations. The complete case consists of only 243,443

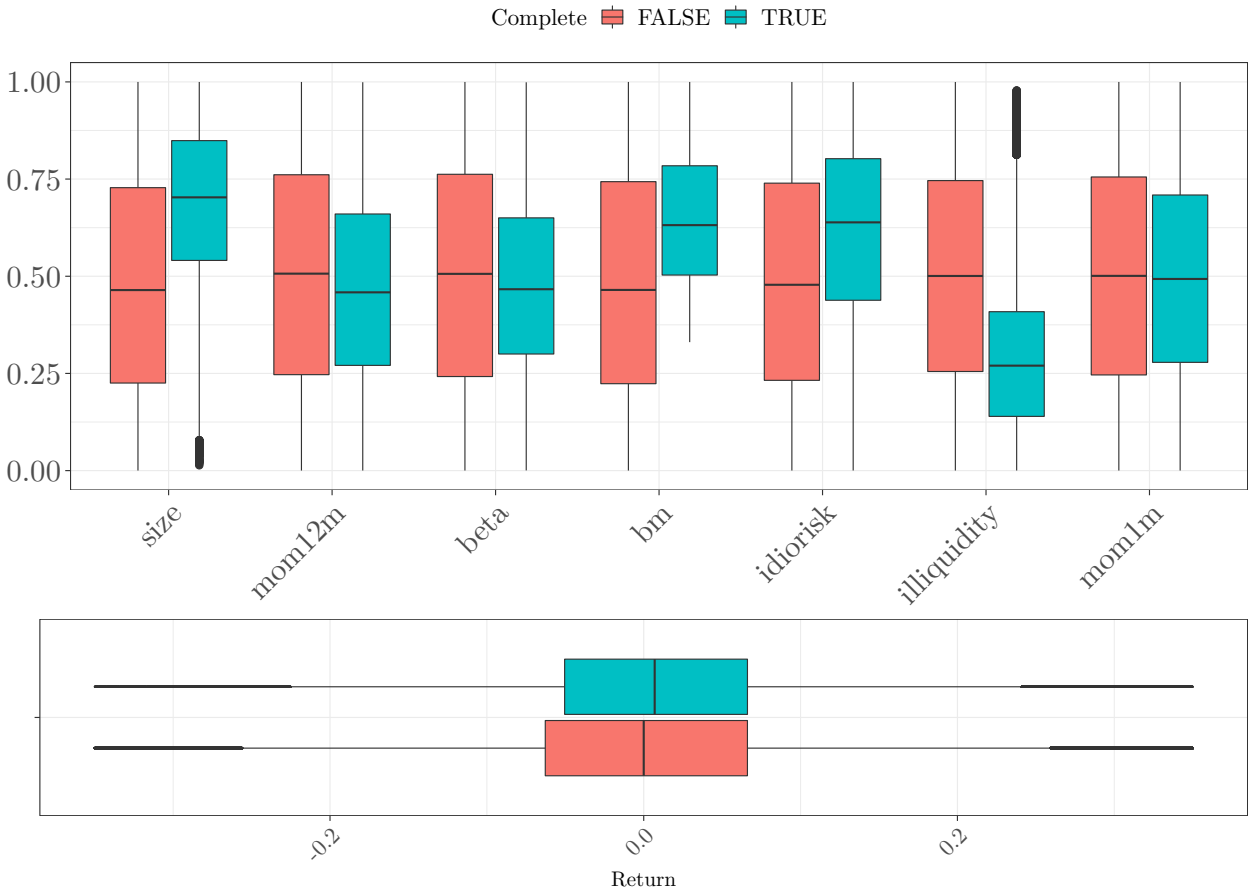
firm month observations, i.e. the complete case would discard around 90% of all available return observations. This makes the complete case analysis rather inefficient. Recall that we assume that the data is missing at random conditional on the characteristics that are always observed. If we drop an additional 17,695 observations, we always observe the following characteristics: `exchswitch`, `divinit`, `divomit`, `opleverage`, `leverage`, `high52`, `indmom`, `mom12m`, `mom6m`, `mom1m`, `intmom`, `beta`, `bm`, `am`, `idiorisk`, `maxret`, `coskewness` and `size`. We then only loose very few additional observations, but our main assumption is much more palatable. Hence, our data set has a total of 2,297,871 firm month observations. In the empirical analysis we always apply the rank-transformation as in Freyberger et al. (2020) such the the continuous characteristics are always uniformly distributed on $[0, 1]$, a standard transformation, which is also applied in Kozak et al. (2020), Gu et al. (2020) and many others.

To gain some more intuition about characteristics of firms for which some characteristic is missing and for which none is missing, we plot the distribution of some characteristics in Figure 7. We can see from the figure that there are systematic differences in the unconditional distribution of the characteristics between the cases for which something is missing and the complete case. In particular, Figure 7 illustrates that firms for which some characteristic is missing tend to be smaller, have higher betas, lower book-to-market ratios. The returns for the incomplete case are more dispersed relative to the complete case. It is important to note that this is not a violation of our missing at random assumption. The assumption allows for unconditional differences in the characteristics but not conditional on the always observed characteristics.

To further investigate the conditional difference we follow an indirect route. As mentioned in Section 2, while our missing at random assumption is not directly testable, we can test the implications of the assumption that we use to construct our estimator. In particular, our imputation based estimator uses additional moment restrictions (relative to the complete case estimator) that are derived from the missing at random assumption. Since these moments

Figure 7: Comparison of Complete and Incomplete Samples

This figure illustrates unconditional difference for some characteristics (upper panel) and return (lower panel) between the stock for which the observations are complete, relative to the stocks which have some items missing.



conditions are over-identifying, we can test the null hypothesis that they hold, i.e.

$$H_0 : E\left[\mathbf{1}(D_{it} = l)\left(Y_{it} - \sum_{k=1}^{K} \beta_{t,k} Z_{it,k}^{(l)}\right) X_{it}^{(l)}\right] = 0 \qquad l = 1, \ldots, L$$

using a J-test. The technical details are explained in Section A.6. We implement this test on a period-by-period basis and find that it rejects the null hypothesis that the over-identifying restrictions hold in only 0.94% of the time periods with a 1% significance level, in 4.09% of the time periods with a 5% significance level and in 8.96% of the time periods with a 10% significance level. These are exactly the rejection probabilities we would expect if the null hypothesis was true in all time periods and if the tests were independent.

5 Empirical Application

In this section we illustrate the empirical effect of different choices for treating missing data in several applications: out of sample return prediction and determining which characteristics provide incremental information.

5.1 Out-of sample predictions

We first illustrate the different ways of treating missing regressors in a classic empirical asset pricing application - cross-sectional out-of-sample return predictions. We report results for three different methods, namely (1) estimate the prediction model only on the completely observed data, (2) estimate the model on the data for which we imputed the conditional mean with the GLS weighting scheme, and (3) estimate the model on the data for which we imputed the unconditional mean with OLS. Afterwards, we also illustrate how to use our approach in a regularized additive model, in which we carry out model selection using the adaptive group LASSO of Huang et al. (2010) and Freyberger et al. (2020).

Throughout, we make rolling out-of-sample predictions for the next month using an estimation window of 120 months. We then sort stocks into portfolios based on the predicted return. We consider two portfolio implementations, where we go long the stocks with highest 10% (50%) predicted returns and go short the stocks with lowest 10% (50%) predicted return. We then record the return for the out-of-sample month, slide the estimation window forward and repeat the portfolio formation exercise throughout the sample. Our out-of-sample period is 1990 through 2018. The results are summarized in Table 6.

Panel A of Table 6 shows the results for the linear model using all characteristics. For the 50-50 portfolio, we see that the complete-case analysis yields the lowest average returns at about the same level of volatility as the conditional mean and unconditional mean imputation. The complete case method simply does not use an interesting subset of securities that could otherwise be invested in. While the conditional mean method delivers slightly better performance on average, the differences between the two appear to be relatively small

Table 6: Performance Statistics For Out-of-Sample Predictions

This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction. We differentiate between the complete case method, conditional mean imputation and GLS weighting and unconditional mean imputation. Long Pf. and Short Pf. denote the annualized average return of the long and short leg respectively. Skewness and kurtosis are the sample statistics of the monthly returns and maximum drawdown is measured from peak to trough. The implementation of the linear and polynomial model is detailed in Section 5.1. The sample period is 1990-2018.

	Mean (%)	Standard Deviation (%)	Sharpe Ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis	Maximum Drawdown
Panel A: Linear Model								
Long (short) 50% highest (lowest) predicted returns								
Complete Case	6.07	6.41	0.95	17.13	11.06	-0.25	2.62	0.17
Cond. Mean (GLS)	13.14	7.20	1.82	20.70	7.56	1.16	8.64	0.19
Uncond. Mean	12.13	7.33	1.65	20.19	8.06	1.11	8.60	0.19
Long (short) 10% highest (lowest) predicted returns								
Complete Case	9.85	14.59	0.68	19.80	9.95	-0.08	3.20	0.53
Cond. Mean (GLS)	35.19	16.99	2.07	33.10	-2.09	1.05	6.78	0.40
Uncond. Mean	33.64	16.99	1.98	31.87	-1.77	1.05	6.67	0.39
Panel B: Nonlinear Model								
Long (short) 50% highest (lowest) predicted returns								
Complete Case	5.22	5.84	0.90	16.71	11.48	-0.06	3.11	0.16
Cond. Mean (GLS)	14.65	6.21	2.36	21.45	6.80	1.56	11.07	0.12
Uncond. Mean	13.83	6.07	2.28	21.04	7.21	1.08	8.35	0.12
Long (short) 10% highest (lowest) predicted returns								
Complete Case	12.95	15.35	0.84	21.30	8.35	-0.41	5.19	0.41
Cond. Mean (GLS)	46.02	17.58	2.62	42.07	-3.94	1.89	9.56	0.21
Uncond. Mean	44.70	17.24	2.59	40.66	-4.03	1.60	7.15	0.19
Panel C: Regularized Nonlinear Model								
Long (short) 50% highest (lowest) predicted returns								
Complete Case (LASSO)	3.37	6.24	0.54	15.78	12.41	-1.35	13.08	0.31
Cond. Mean (GLS / LASSO)	14.11	5.85	2.41	21.18	7.07	1.06	5.91	0.09
Uncond. Mean (LASSO)	12.97	6.06	2.14	20.61	7.64	1.05	8.44	0.11
Long (short) 10% highest (lowest) predicted returns								
Complete Case (LASSO)	9.43	17.38	0.54	20.63	11.20	-1.94	14.43	0.91
Cond. Mean (GLS / LASSO)	45.70	17.13	2.67	41.62	-4.08	1.62	7.24	0.17
Uncond. Mean (LASSO)	43.71	17.20	2.54	39.80	-3.91	1.36	5.21	0.17

in the prediction application. The differences between the complete case method and the imputation approaches are even larger in the "10-1" portfolio.

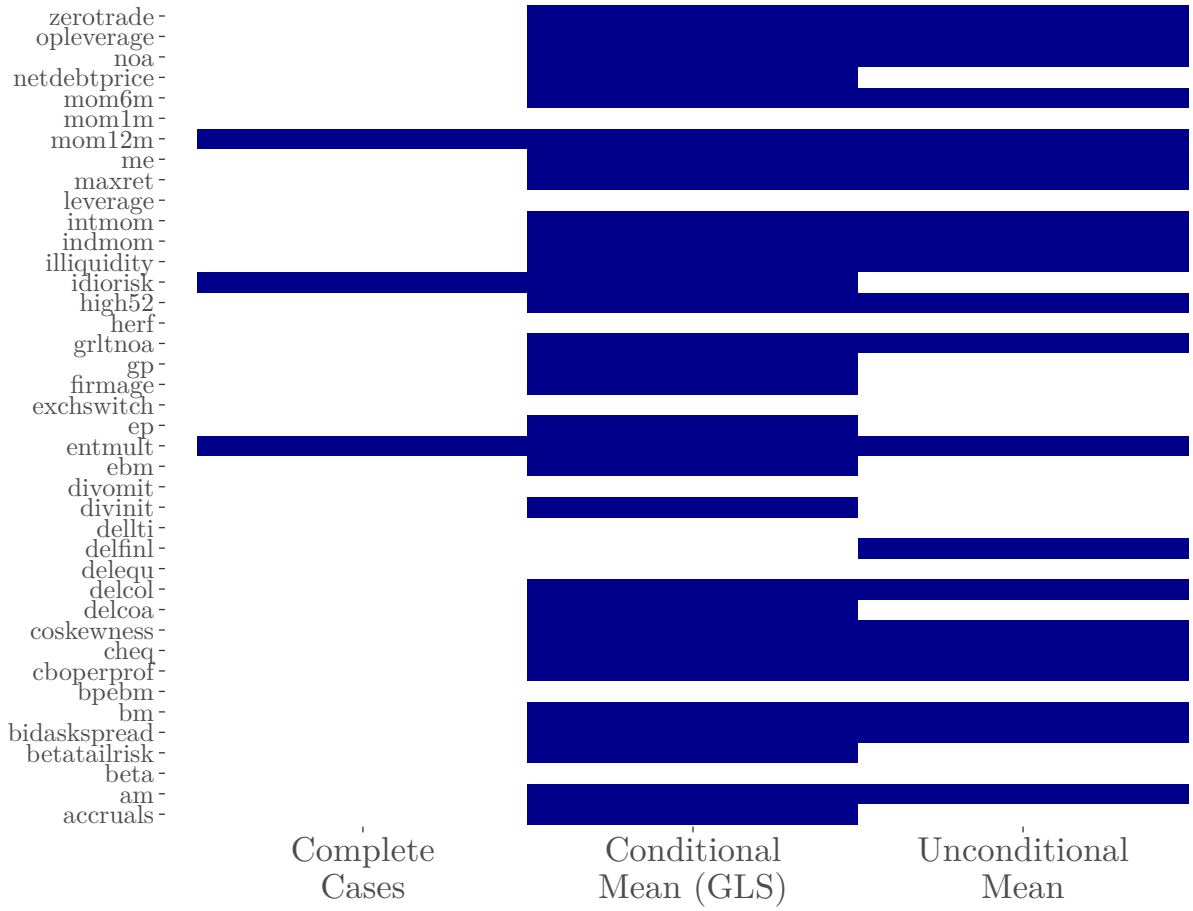
The Panel B and Panel C illustrate the results for a nonlinear model, i.e. an additive model as outline in Section 2.3.1. In Panel B we present the results for the additive model using all 40 characteristics. As in the case of the linear model, using only the complete cases results in very low returns relative to the conditional and unconditional mean imputation. Both Panel B and Panel C show that modeling returns as a nonlinear function of character-

istics yields much better predictions. Notably, the difference between the linear model and nonlinear models is more pronounced for the "10-1" portfolio as most of the nonlinearities in the predictive relationship occurs in the extremes of the characteristics distributions.

For the results in Panel C, we first carry out a model selection step over the period from 1965 through 1989. We apply the adaptive group LASSO as in Freyberger et al. (2020) to select the most important characteristics over the first part of the sample and then, exactly as for the other methods, make rolling one-month predictions using an estimation window of 120 months. Overall, the results are very similar to those in Panel B. Again note that the predictors are known to predict cross-sectional predictors a priori and it is therefore not surprising that including all of them in the model may yield favorable results.

Figure 8: Selected Characteristics with the Group LASSO Procedure.

This figure shows the selected characteristics using the group LASSO procedure for each of the three methods.



Model selection will play a more important role in other data sets with a large number of characteristics, in which some are irrelevant or have only very small influence. While the imputation methods perform similarly with and without regularization, in the complete case, the regularized model leads to considerably worse performance. The reason is that the complete case contains much fewer observations and is consequently less likely to detect significant return predictors. As can see from Figure 8, we only select 4 characteristics using the complete case, and we select considerably more with the other two methods.

5.2 Incremental Information

We now re-visit the the classic question if a characteristic contains incremental information relative to previously discovered characteristic. Cochrane (2011) raises this question in his presidential address. While the previous literature mostly proceeded in a "univariate fashion", i.e. analyzing one characteristic at a time, recent papers e.g. by Green et al. (2017), Freyberger et al. (2020), Kozak et al. (2020) and Gu et al. (2020) make it abundantly clear that we need to consider characteristics jointly and to determine if a characteristics provides incremental information, we need to control for the ones that were previously discovered.

The more characteristics we want to consider within the same model, the more our choices about missing data may affect the results. We illustrate this by studying the characteristics listed in Table A.1 in the appendix. For each characteristic, we consider if it should have been recognized as containing incremental information at the time of discovery (based on the publication dates in Table A.1) when previous characteristics are taken into account. Throughout we compare the three approaches to treating missing data, the complete case approach, the conditional mean imputation with GLS weighting, and the unconditional mean imputation. We then estimate the following linear model

$$Y_{it} = \beta_0 + \underbrace{\beta_1 X_{it,1} + \beta_2 X_{it,2} + \ldots + \beta_{k-1} X_{it,k-1}}_{\text{previously published characteristics}} + \underbrace{\beta_k X_{it,k}}_{\text{new candidate}} + \varepsilon_{it}. \tag{4}$$

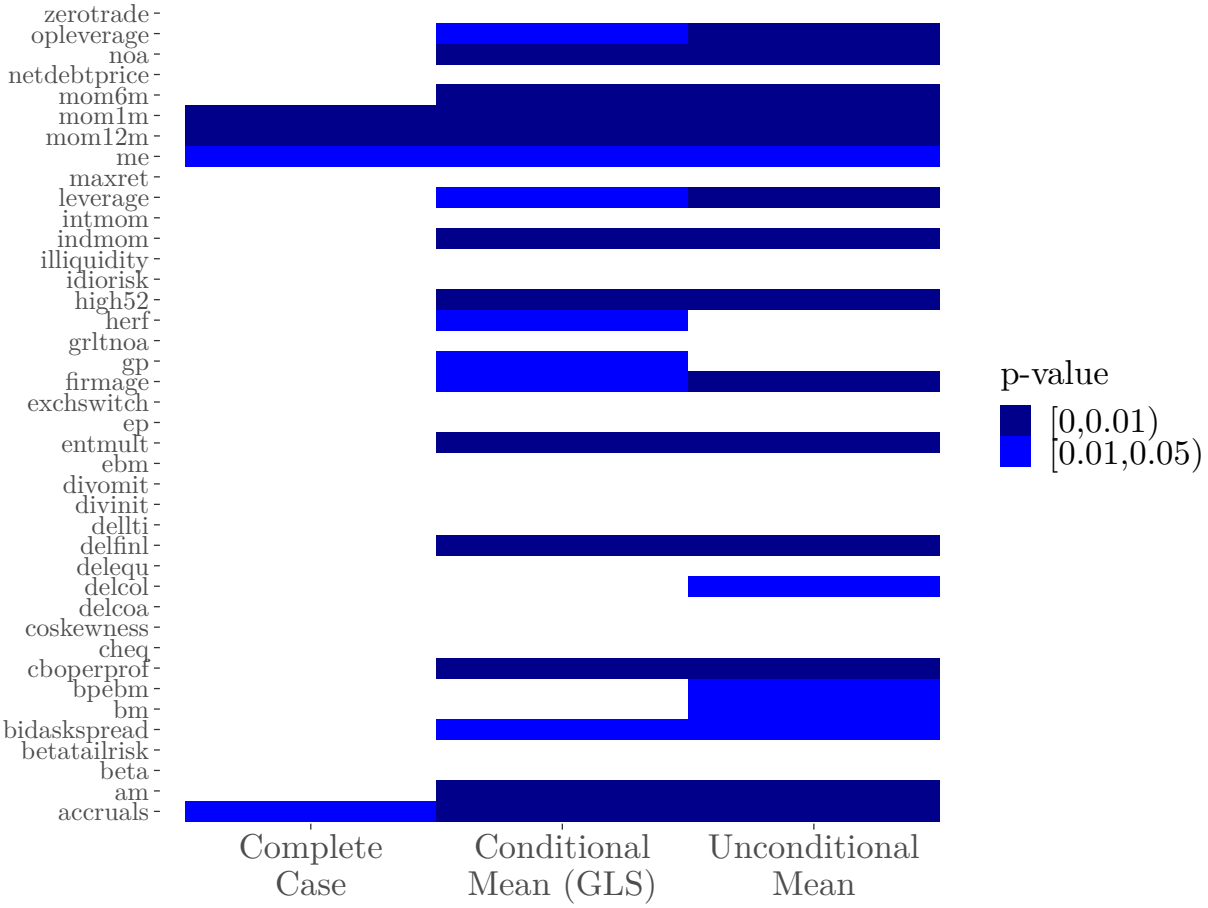
Table 7: Significance test results

This table shows the estimates, standard errors, and adjusted p-value for each new characteristics.

	Complete case			Conditional mean (GLS)			Unconditional mean		
Characteristic	Est.	se	p-value	Est.	se	p-value	Est.	se	p-value
zerotrade	-0.0092	0.0063	1.0000	-0.0034	0.0038	1.0000	0.0030	0.0044	1.0000
opleverage	0.0024	0.0026	1.0000	0.0049	0.0016	0.0186	0.0062	0.0016	0.0014
noa	0.0042	0.0016	0.1945	0.0096	0.0017	0.0000	0.0102	0.0016	0.0000
netdebtprice	-0.0050	0.0183	1.0000	0.0223	0.0124	0.4535	0.0013	0.0023	1.0000
mom6m	-0.0023	0.0029	1.0000	-0.0103	0.0026	0.0014	-0.0091	0.0025	0.0042
mom1m	-0.0150	0.0026	0.0000	-0.0249	0.0024	0.0000	-0.0243	0.0024	0.0000
mom12m	0.0138	0.0030	0.0005	0.0163	0.0026	0.0000	0.0180	0.0025	0.0000
me	-0.0155	0.0045	0.0297	-0.0151	0.0047	0.0155	-0.0152	0.0047	0.0139
maxret	0.0035	0.0038	1.0000	-0.0043	0.0030	0.8715	-0.0046	0.0030	0.6614
leverage	0.0028	0.0028	1.0000	0.0060	0.0019	0.0182	0.0068	0.0020	0.0074
intmom	0.0078	0.0035	0.3896	0.0013	0.0022	1.0000	0.0010	0.0023	1.0000
indmom	0.0009	0.0016	1.0000	0.0097	0.0018	0.0000	0.0096	0.0018	0.0000
illiquidity	-0.0304	0.0141	0.4131	-0.0075	0.0111	1.0000	-0.0161	0.0096	0.5754
idiorisk	-0.0028	0.0044	1.0000	-0.0066	0.0034	0.3621	-0.0055	0.0029	0.3980
high52	0.0025	0.0039	1.0000	0.0191	0.0047	0.0011	0.0202	0.0045	0.0002
herf	-0.0002	0.0020	1.0000	0.0045	0.0016	0.0402	0.0036	0.0013	0.0624
grltnoa	0.0002	0.0012	1.0000	0.0022	0.0011	0.3782	0.0019	0.0011	0.6123
gp	0.0007	0.0039	1.0000	0.0093	0.0029	0.0155	0.0062	0.0024	0.0850
firmage	-0.0264	0.0090	0.0937	-0.0253	0.0089	0.0402	-0.0109	0.0031	0.0061
exchswitch	0.0020	0.0020	1.0000	0.0029	0.0014	0.2880	0.0034	0.0014	0.0907
ep	0.0087	0.0040	0.3896	0.0087	0.0040	0.2228	0.0088	0.0040	0.1953
entmult	0.0064	0.0054	1.0000	0.0096	0.0021	0.0002	0.0088	0.0014	0.0000
ebm	0.0008	0.0015	1.0000	-0.0017	0.0011	0.8715	-0.0010	0.0011	1.0000
divomit	-0.0001	0.0032	1.0000	0.0016	0.0023	1.0000	0.0015	0.0023	1.0000
divinit	0.0010	0.0033	1.0000	0.0054	0.0026	0.2880	0.0055	0.0026	0.2578
dellti	0.0003	0.0009	1.0000	0.0008	0.0008	1.0000	0.0010	0.0008	1.0000
delfinl	0.0026	0.0011	0.3896	0.0047	0.0009	0.0000	0.0049	0.0009	0.0000
delequ	0.0016	0.0035	1.0000	0.0023	0.0021	1.0000	0.0008	0.0020	1.0000
delcol	-0.0030	0.0022	1.0000	-0.0039	0.0014	0.0638	-0.0039	0.0014	0.0471
delcoa	-0.0008	0.0019	1.0000	0.0011	0.0016	1.0000	0.0011	0.0016	1.0000
coskewness	0.0018	0.0025	1.0000	0.0029	0.0016	0.4535	0.0030	0.0016	0.4380
cheq	0.0179	0.0081	0.3896	0.0046	0.0033	0.9121	0.0050	0.0024	0.2407
cboperprof	0.0137	0.0070	0.6002	0.0118	0.0024	0.0000	0.0123	0.0017	0.0000
bpebm	0.0084	0.0030	0.1203	-0.0067	0.0025	0.0654	-0.0071	0.0024	0.0337
bm	0.0059	0.0037	1.0000	0.0034	0.0037	1.0000	0.0097	0.0031	0.0186
bidaskspread	-0.0204	0.0065	0.0628	-0.0319	0.0097	0.0140	-0.0265	0.0084	0.0166
betatailrisk	-0.0070	0.0072	1.0000	0.0052	0.0046	1.0000	0.0042	0.0036	1.0000
beta	-0.0075	0.0084	1.0000	-0.0075	0.0084	1.0000	-0.0075	0.0084	1.0000
am	0.0046	0.0056	1.0000	0.0187	0.0048	0.0015	0.0288	0.0046	0.0000
accruals	0.0045	0.0013	0.0297	0.0040	0.0011	0.0055	0.0039	0.0011	0.0050

Figure 9: p-values of hypothesis test

This figure illustrates which of the tests yield significant results using the p-values from Table 7. We use different colors for different significance levels.

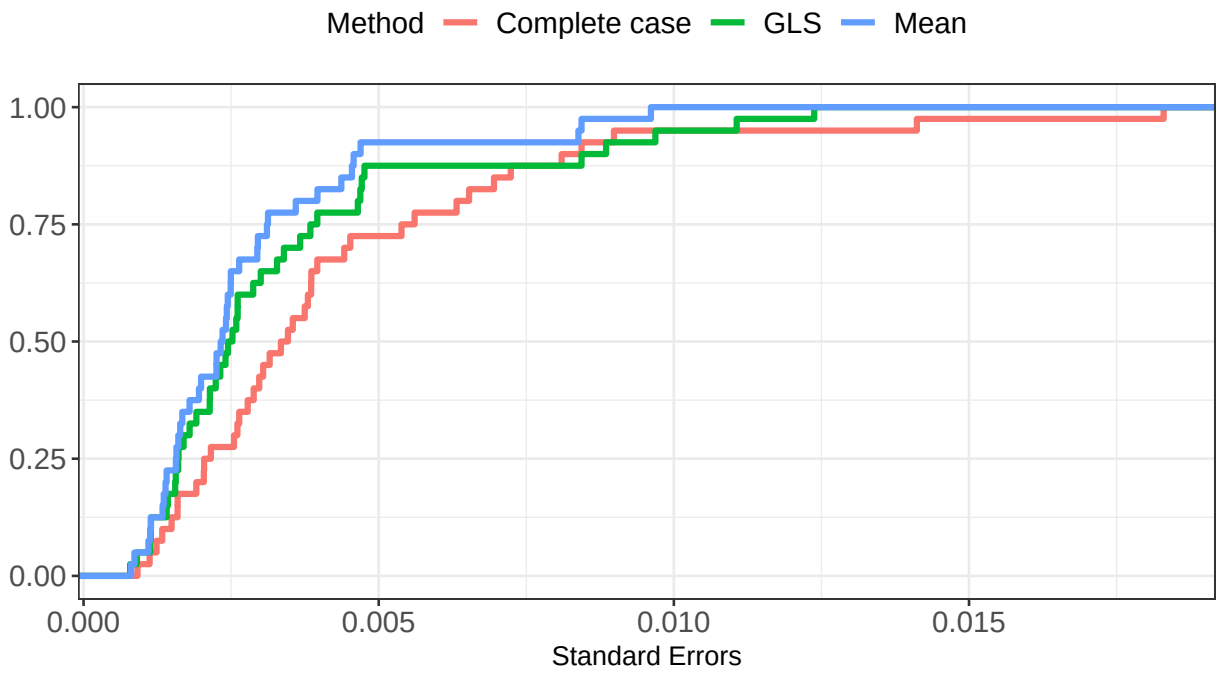


We estimate this model using only the data up until the publication date of the new candidate predictor, not the full sample. To determine if a characteristic is significant, we test $H_0 : \beta_k = 0$ using a two-sided t-test. We allow for cross-sectional dependence of the error terms by using clustered standard errors. Since, we have 40 characteristics in total and thus perform 40 separate t-tests, we use p-values adjusted for the false discovery rate to take the multiple testing problem into account (see Benjamini and Yekutieli (2001) and Green et al. (2017)).[4] These p-values might be larger than 1 in which case we set them to 1

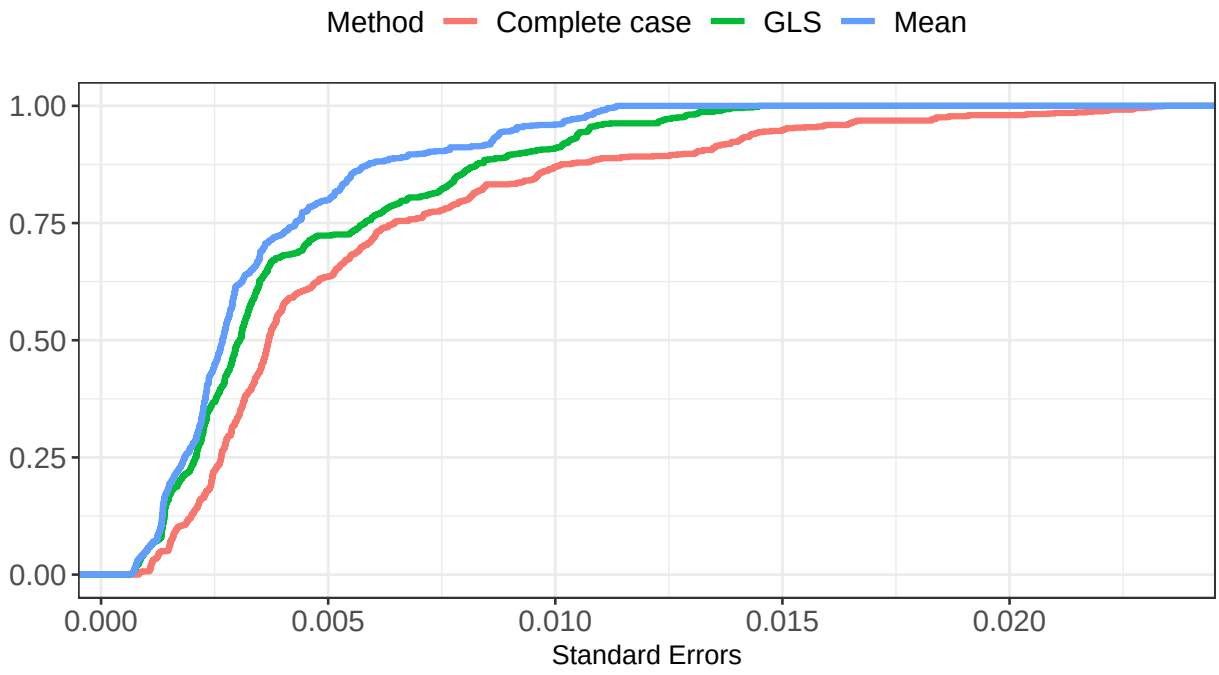
[4]Specifically, let p_i denote the standard p-value of the ith test and assume that the p-values have been ordered, such that $p_1 \leq p_2 \leq \cdots \leq p_{40}$. The adjusted false discovery rate p-values are $\tilde{p}_{40} = \left(\sum_{i=1}^{40}(1/i)\right) p_{40}$ and $\tilde{p}_i = \min\left\{\tilde{p}_{i+1}, \left(\sum_{j=1}^{40}(1/j)\right)(40/i)p_i\right\}$ for all $i < 40$.

Figure 10: Empirical cdfs of standard errors of new characteristics

This figure shows the empirical distribution function of the standard errors of the estimated coefficients for the three different methods. Panel (a) shows the standard errors for all new characteristics (i.e. those in Table 7). Panel (b) shows the standard errors for all estimated coefficients in equation (4) and for each time period.



(a) New characteristics



(b) All characteristics

when presenting our results. Table 7 shows the estimates, standard errors, and the adjusted p-values.

Figure 9 illustrates which characteristics have a significant effect on returns. It shows that we select very few characteristics in the complete case. The efficiency loss is indeed large, because we discard too much data and thus do not make use of the available information. Conditional mean imputation using the GLS adjustment selects more characteristics, but still fewer than the unconditional mean imputation. Most notably the selection of characteristics between the conditional mean and unconditional mean imputation is quite different. This is due to the interaction of two effects highlighted in Section 3. First, mean imputation yields biased estimators and estimated coefficients may be either too larger or too close to 0. As a specific example, consider the book-to-market ratio (bm) in Table 7. The coefficients in the complete case and with conditional mean imputation are quite similar (0.0059 and 0.0034, respectively), while unconditional mean imputation yields a much larger estimated coefficient (0.0097) that is significantly different from 0. Second, with unconditional mean imputation, we underestimate the covariance between the characteristics and therefore get artificially small standard errors. To illustrate this difference, Figure 10 shows empirical cdfs of the standard errors obtained using the different methods. In panel (a), we plot the cdf for all new characteristics (i.e. for the standard errors in Table 7). Panel (b) shows the cdf of the standard errors for all estimated coefficients in equation (4) and for each time period. We can see that the complete case yields the largest standard errors because it only makes use of a subset of the data, and the standard errors with mean imputation tend to be the smallest.

6 Conclusion

Missing data occur in virtually all cross-sectional empirical asset pricing studies. The primary goal of this paper is to provide empirical researchers with an easy approach to address this problem more systematically. Our proposed approach can be implemented with

standard statistical packages and is computationally tractable even in high dimensions and for very large panels.

Our results show that the complete case method, despite its intuitive appeal, neglects and important part of the return distribution. We therefore advocate the use of imputation. Moreover, since unconditional mean imputation leads to bias in the estimation and incorrect inference, we urge researchers not to use it. Instead, researchers should use conditional mean imputation and adjust for the estimation error in subsequent inference.

The two step approach enjoys broad appeal and can be applied in other common areas of research such as estimating the stochastic discount factors, illustrated in A.2, characteristic based factor models, and international studies. These items are left for future research.

References

Abrevaya, J. and S. G. Donald (2017). A gmm approach for dealing with missing data on regressors. *Review of Economics and Statistics 99*(4), 657–662.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets 5*(1), 31–56.

Amihud, Y. and H. Mendelson (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics 17*(2), 223–249.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *The Journal of Finance 61*(1), 259–299.

Bai, J. and S. Ng (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association 0*, 1–50.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics 99*(2), 427 – 446.

Ball, R., J. Gerakos, J. T. Linnainmaa, and V. Nikolaev (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics 121*(1), 28–45.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics 9*(1), 3–18.

Barry, C. B. and S. J. Brown (1984). Differential information and the small firm effect. *Journal of Financial Economics 13*(2), 283–294.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance 32*(3), 663–682.

Beaver, W., M. McNichols, and R. Price (2007). Delisting returns and their effect on accounting-based market anomalies. *Journal of Accounting and Economics 43*(2-3), 341–368.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165 – 1188.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance 43*(2), 507–528.

Blanchet, J., F. Hernandez, V. A. Nguyen, M. Pelger, and X. Zhang (2021). Time-series imputation with wasserstein interpolation for optimal look-ahead-bias and variance tradeoff. *arXiv preprint arXiv:2102.12736*.

Brown, S. J., W. Goetzmann, R. G. Ibbotson, and S. A. Ross (1992). Survivorship bias in performance studies. *The Review of Financial Studies 5*(4), 553–580.

Cahan, E., J. Bai, and S. Ng (2021). Factor-based imputation of missing values and covariances in panel data of large dimensions. *Working Paper*.

Chen, A. Y. and T. Zimmermann (2021). Open source cross sectional asset pricing. *Critical Finance Review, Forthcoming*.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance 66*(4), 1047–1108.

Connor, G. and R. A. Korajczyk (1987). Estimating pervasive economic factors with missing observations. *Available at SSRN 1268954*.

Dagenais, M. G. (1973). The use of incomplete observations in multiple regression analysis: A generalized least squares approach. *Journal of Econometrics 1*(4), 317–328.

Dharan, B. G. and D. L. Ikenberry (1995). The long-run negative drift of post-listing stock returns. *The Journal of Finance 50*(5), 1547–1574.

Fairfield, P. M., J. S. Whisenant, and T. L. Yohn (2003). Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review 78*(1), 353–371.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*(2), 427–465.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Fitzmaurice, G. M., M. G. Kenward, G. Molenberghs, G. Verbeke, and A. A. Tsiatis (2015). Missing data: Introduction and statistical preliminaries. In G. Molenberghs, G. M. Fitzmaurice, M. G. Kenward, A. A. Tsiatis, and G. Verbeke (Eds.), *Handbook of Missing Data Methodology* (1 ed.)., pp. 3–22. Boca Raton: CRC Press, Taylor & Francis Group.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

George, T. J. and C.-Y. Hwang (2004). The 52-week high and momentum investing. *The Journal of Finance 59*(5), 2145–2176.

Gourieroux, C. and A. Monfort (1981). On the problem of missing data in linear models. *The Review of Economic Studies 48*(4), 579–586.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–280.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies 29*(1), 5–68.

Harvey, C. R. and A. Siddique (2000). Conditional skewness in asset pricing tests. *The Journal of Finance 55*, 1263–1295.

Haugen, R. A. and N. L. Baker (1996). Commonality in determinants of expected stock returns. *Journal of Financial Economics 41*(3), 401–439.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Hirshleifer, D., K. Hou, S. H. Teoh, and Y. Zhang (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics 38*, 297–331.

Hou, K. and D. T. Robinson (2006). Industry concentration and average stock returns. *The Journal of Finance 61*(4), 1927–1956.

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of Statistics 38*(4), 2282–2313.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*, 65–91.

Jin, S., K. Miao, and L. Su (2021). On factor models with random missing: Em estimation, inference, and cross validation. *Journal of Econometrics 222*(1), 745–777.

Kelly, B. and H. Jiang (2014). Tail risk and asset prices. *The Review of Financial Studies 27*(10), 2841–2871.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2021). Arbitrage portfolios. *The Review of Financial Studies 34*(6), 2813–2856.

Kim, S. and G. Skoulakis (2018). Ex-post risk premia estimation and asset pricing tests using large cross sections: The regression-calibration approach. *Journal of Econometrics 204*(2), 159–188.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Lewellen, J. (2015). The cross section of expected stock returns. *Critical Finance Review 4*(1), 1–44.

Liao, Z. and Y. Liu (2020). Optimal cross-sectional regression. *Available at SSRN*.

Light, N., D. Maslov, and O. Rytchkov (2017). Aggregation of information about the cross section of stock returns: A latent variable approach. *The Review of Financial Studies 30*(4), 1339–1381.

Little, R. J. A. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association 87*(420), 1227–1237.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika 81*(3), 471–483.

Little, R. J. A. and D. B. Rubin (2020). *Statistical Analysis with Missing Data* (3 ed.). John Wiley & Sons, Inc.

Liu, W. (2006). A liquidity-augmented capital asset pricing model. *Journal of Financial Economics 82*(3), 631–671.

Lockwood, L. and W. Prombutr (2010). Sustainable growth and stock returns. *Journal of Financial Research 33*(4), 519–538.

Loughran, T. and J. W. Wellman (2011). New evidence on the relation between the enterprise multiple and average stock returns. *Journal of Financial and Quantitative Analysis 46*(6), 1629–1650.

Manski, C. F. (2005). Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning 39*(2), 151–165. Imprecise Probabilities and Their Applications.

Michaely, R., R. H. Thaler, and K. L. Womack (1995). Price reactions to dividend initiations and omissions: Overreaction or drift? *The Journal of Finance 50*(2), 573–608.

Molenberghs, G., G. M. Fitzmaurice, M. G. Kenward, A. A. Tsiatis, and G. Verbeke (2015). *Handbook of Missing Data Methodology* (1 ed.). Boca Raton: CRC Press, Taylor & Francis Group.

Moskowitz, T. J. and M. Grinblatt (1999). Do industries explain momentum? *The Journal of Finance 54*(4), 1249–1290.

Nijman, T. and F. Palm (1988). Efficiency gains due to using missing data procedures in regression models. *Statistical Papers 29*(1), 249–256.

Novy-Marx, R. (2011). Operating leverage. *Review of Finance 15*(1), 103–134.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics 103*(3), 429–453.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics 108*(1), 1–28.

Penman, S. H., S. A. Richardson, and I. Tuna (2007). The book-to-price effect in stock returns: Accounting for leverage. *Journal of Accounting Research 45*(2), 427–467.

Rao, C. R. and H. Toutenburg (1999). *Linear Models: Least Squares and Alternatives* (2 ed.). Springer.

Richardson, S. A., R. G. Sloan, M. T. Soliman, and I. Tuna (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics 39*(3), 437–485.

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.

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

Rubin, D. B. (1978). Multiple imputations in sample surveys - a phenomenological bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Volume 1, pp. 20–34. American Statistical Association.

Shumway, T. (1997). The delisting bias in crsp data. *The Journal of Finance 52*(1), 327–340.

Sloan, R. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review 71*(3), 289–315.

Tsiatis, A. A. and M. Davidian (2015). Missing data methods: A semi-parametric perspective. In G. Molenberghs, G. M. Fitzmaurice, M. G. Kenward, A. A. Tsiatis, and G. Verbeke (Eds.), *Handbook of Missing Data Methodology* (1 ed.)., pp. 149–184. Boca Raton: CRC Press, Taylor & Francis Group.

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

Xiong, R. and M. Pelger (2019). Large dimensional latent factor modeling with missing observations and applications to causal inference. arxiv eprint. *arXiv preprint arXiv:1910.08273*.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture 1*(2), 129–142.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association 100*(472), 1394–1411.

Zhou, G. (1994). Analytical gmm tests: Asset pricing with time-varying risk premiums. *The Review of Financial Studies 7*(4), 687–709.

Appendix:
Missing Data in Asset Pricing Panels

Table A.1: Overview of the Characteristics

This Table gives an overview of the characteristic used in the empirical analysis. They are obtained from Chen and Zimmermann (2021). We refer to their paper and the companion website for the precise construction.

Acronym	Description	Publication Year	Reference	% missing
Accruals	Accruals	1996	Sloan (1996)	4.26
AM	Total assets to market	1992	Fama and French (1992)	0.00
Beta	CAPM beta	1973	Fama and MacBeth (1973)	0.00
BetaTailRisk	Tail risk beta	2014	Kelly and Jiang (2014)	28.89
BidAskSpread	Bid-ask spread	1986	Amihud and Mendelson (1986)	10.29
BM	Book to market	1992	Fama and French (1992)	0.00
BMent	Enterprise component of BM	2007	Penman et al. (2007)	0.30
BPEBM	Leverage component of BM	2007	Penman et al. (2007)	0.30
CBOperProf	Cash-based operating profitability	2016	Ball et al. (2016)	15.04
ChBE	Sustainable Growth	2010	Lockwood and Prombutr (2010)	4.93
ChBEtoA	Change in equity to assets	2005	Richardson et al. (2005)	4.24
ChCOA	Change in current operating assets	2005	Richardson et al. (2005)	3.88
ChCol	Change in current operating liabilities	2005	Richardson et al. (2005)	4.26
ChFinLiab	Change in financial liabilities	2005	Richardson et al. (2005)	4.39
ChLTI	Change in long-term investment	2005	Richardson et al. (2005)	3.88
Coskewness	Coskewness	2000	Harvey and Siddique (2000)	0.00
DivInit	Dividend Initiation	1995	Michaely et al. (1995)	0.00
DivOmit	Dividend Omission	1995	Michaely et al. (1995)	0.00
EntMult	Enterprise Multiple	2011	Loughran and Wellman (2011)	15.56
EP	Earnings-to-Price Ratio	1977	Basu (1977)	24.34
ExchSwitch	Exchange Switch	1995	Dharan and Ikenberry (1995)	0.00
FirmAge	Firm Age	1984	Barry and Brown (1984)	65.61
GrLTNOA	Growth in Long term net operating assets	2003	Fairfield et al. (2003)	4.56
GrossProf	gross profits / total assets	2013	Novy-Marx (2013)	15.70
Herf	Industry concentration (Herfindahl)	2006	Hou and Robinson (2006)	4.49
High52	52 week high	2004	George and Hwang (2004)	0.00
IdioRisk	Idiosyncratic risk	2006	Ang et al. (2006)	0.00
Illiquidity	Amihud's illiquidity	2002	Amihud (2002)	7.45
IndMom	Industry Momentum	1999	Moskowitz and Grinblatt (1999)	0.00
InterMom	Intermediate Momentum	2012	Novy-Marx (2012)	0.00
Leverage	Market leverage	1988	Bhandari (1988)	0.00
MaxRet	Maximum return over month	2011	Bali et al. (2011)	0.00
Mom12m	Momentum (12 month)	1993	Jegadeesh and Titman (1993)	0.00
Mom1m	Short term reversal	1990	Jegadeesh and Titman (1993)	0.00
Mom6m	Momentum (6 month)	1993	Jegadeesh and Titman (1993)	0.00
NetDebtPrice	Net debt to price	2007	Penman et al. (2007)	51.79
NOA	Net Operating Assets	2004	Hirshleifer et al. (2004)	3.97
OperLeverage	Operating Leverage	2011	Novy-Marx (2011)	0.00
Size	Market value of equity	1981	Banz (1981)	0.00
ZeroTrade	Days with zero trades	2006	Liu (2006)	6.70

A.1 Additional Definitions

We briefly recall some basic notions relevant to missing data treatment. Introductory treatments can be found for example in Little and Rubin (2020), Fitzmaurice et al. (2015).

A.1.1 Missing patterns

A missing pattern describes which data are missing. Figure 1 shows examples of missing patterns. In our application, we cannot assume that we are confronted with a particular missing pattern, and instead deal with general missing patterns. Our theoretical results require a non-negligible part of the data to be complete. Generalizing these results would require much stronger assumptions and does not occur in our empirical application.

A.1.2 Missing mechanisms

The missing mechanism describes why data are missing, i.e. it describes the relationship between the missingness and the values of the observed (and possibly unobserved) variables. Rubin (1976) introduces three formal definitions for missing mechanisms that have become standard in the literature. He differentiates between missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). We recall these basic definitions, using our notation from Section 2 below.

In Section 2 (and with only cross-sectional data) the missing pattern of observation i is denoted by D_i. The outcome is Y_i and the regressors are X_i. Let $X_i^{(o)}$ be the subset of X_i that is observed under all missing patterns. Let V_i be a vector of observed additional characteristics (as in section 2.3.2). We refer to the analysis based on the cases that are completely observed as the complete case analysis. This is in contrast to the "complete data analysis" which is based on the hypothetically observed data in the absence of any missing data.

The data is **MCAR** if $D_i \perp\!\!\!\perp Y_i, X_i, V_i$, i.e. whether an observation is missing does not depend on the other variables. When the data is MCAR, the complete case analysis yields valid inference, but there is a loss of efficiency relative to the complete data analysis due to the decreased sample size (Fitzmaurice et al. (2015)).

The data is **MAR**[1] if $D_i \perp\!\!\!\perp Y_i, X_i \mid X_i^{(o)}, V_i$. That is, missing is only random once we condition on observed covariates. We rely on this type of assumption (but based solely on conditional moments) in our analysis. When the data is MAR, the complete case analysis generally yields valid inference, but might require an estimator based on inverse propensity weighting (as in section 2.3.2). Again, neglecting a part of the sample results in an inefficient estimator.

Data is **NMAR**, sometimes also referred to as missing not at random, if D_i depends on unobserved regressors or the outcome. In this case, the missing data mechanism cannot be ignored. One approach could then be to model it explicitly as in selection models (Heckman (1979)) or pattern-mixture models (Little (1994)). Alternatively, one could use a partial identification approach (Manski (2005)).

A.2 Extensions

A.2.1 Stochastic Discount Factor Estimation

In this section we briefly explain how our proposed method can be used to estimate the stochastic discount factor when covariates might be missing. We start with the standard moment condition

$$E\left[M_{t+1}R^e_{it+1} \mid X_{it}\right] = 0$$

for all $i = 1, 2, \ldots, n$ and $t = 1, \ldots, T$, where M_{t+1} is the stochastic discount factor, R^e_{it+1} are excess returns, and X_{it} are variables known at time t. The discount factor is a linear combination of the excess returns and we assume that the weights are a parametric function of $X_{it} \in \mathbb{R}^K$. That is

$$M_{t+1} = 1 - \sum_{j=1}^{n} \omega\left(X_{jt}, \beta\right) R^e_{t+1,j}$$

where

$$\omega\left(X_{jt}, \beta\right) = \sum_{k=1}^{K} \beta_k X_{jt,k}.$$

[1] MCAR is a special case of MAR.

Combining the previous three equations we get

$$E\left[\left(1-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k X_{jt,k}\right)R^e_{jt+1}\right)R^e_{it+1}\mid X_{it}\right]=0$$

As before, assume we have L missing patterns. Let $D_{it}=l$ for missing pattern l, and let $X^{(l)}_{it,k}$ be the corresponding subset of observed element of X_{it}. Then, under an analogous MAR assumption as before,

$$\begin{aligned}
0 &= E\left[\left(1-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k X_{jt,k}\right)R^e_{jt+1}\right)R^e_{it+1}\mid X^{(l)}_{it}\right]\\
&= E\left[\left(1-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k X_{jt,k}\right)R^e_{jt+1}\right)R^e_{it+1}\mid X^{(l)}_{it},D_{it}=l\right]\\
&= E\left[R^e_{it+1}-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k X_{jt,k}R^e_{jt+1}R^e_{it+1}\right)\mid X^{(l)}_{it},D_{it}=l\right]
\end{aligned}$$

For k such that $X_{jt,k}\subseteq X^{(l)}_{jt}$, let

$$Z^{(l)}_{it,jk}=\begin{cases}X_{jt,k}R^e_{jt+1}R^e_{it+1} & \text{if } k\in I^{(l)}_t\\ E[X_{jt,k}R^e_{jt+1}R^e_{it+1}\mid X^{(l)}_{it},D_{it}=0] & \text{if } k\notin I^{(l)}_t\end{cases}$$

Assuming that

$$E[X_{jt,k}R^e_{jt+1}R^e_{it+1}\mid X^{(l)}_{it},D_{it}=l]=E[X_{jt,k}R^e_{jt+1}R^e_{it+1}\mid X^{(l)}_{it},D_{it}=0]$$

we obtain the conditional moment restrictions

$$E\left[R^e_{it+1}-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k Z^{(l)}_{jt,t}\right)\mid X^{(l)}_{it},D_{it}=l\right]=0$$

To impute missing values, let

$$E[X_{jt,k}R^e_{jt+1}R^e_{it+1}\mid X^{(l)}_{it},D_{it}=0]=h\left(X^{(l)}_{it},\gamma^{(l,k)}\right)$$

where h is a flexible parametric function of $X_{it}^{(l)}$ with parameter vector $\gamma^{(l,k)}$. Finally, let $g(X_{it}^{(l)})$ be a vector of transformations of $X_{it}^{(l)}$. We can then estimate the parameters based on the following unconditional moments:

$$E\left[\mathbf{1}(D_{it}=0)\left(R_{it+1}^{e}-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k X_{jt,k}R_{jt+1}^{e}R_{it+1}^{e}\right)\right)g(X_{it}^{(l)})\right]=0 \tag{A.1}$$

$$E\left[\mathbf{1}(D_{it}=l)\left(R_{it+1}^{e}-\sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k Z_{it,jk}^{(l)}\right)\right)g(X_{it}^{(l)})\right]=0 \qquad l=1,\ldots,L \tag{A.2}$$

$$E\left[\mathbf{1}(D_{it}=0)\left(X_{jt,k}R_{jt+1}^{e}R_{it+1}^{e}-h\left(X_{it}^{(l)},\gamma^{(l,k)}\right)\right)g(X_{it}^{(l)})\right]=0 \qquad l=1,\ldots,L \quad k\notin I_t^{(l)} \tag{A.3}$$

A.2.2 Derivation with additional covariates

Consider the simple model

$$Y_i=\beta_0+X_{i,1}\beta_1+X_{i,2}\beta_2+\varepsilon_i,$$

where $X_{i,1}$ is a always observed, but $X_{i,2}$ might be missing. Let $D_i=0$ if observation i is complete and let $D_i=1$ if $X_{i,2}$ is missing. We now derive moment conditions under the conditional independence assumption

$$D_i \perp\!\!\!\perp Y_i, X_{i,2} \mid X_{i,1}, V_i$$

where V_i is a observed covariate. In this case, we get

$$\begin{aligned}
0 &= E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}\right] \\
&= E\left[E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}, V_i\right] \mid X_{i,1}, X_{i,2}\right] \\
&= E\left[E\left[\varepsilon_i \mid X_{i,1}, X_{i,2}, V_i, D_i=0\right] \mid X_{i,1}, X_{i,2}\right] \\
&= E\left[E\left[\mathbf{1}(D_i=0)\varepsilon_i \mid X_{i,1}, X_{i,2}, V_i\right]\frac{1}{P(D_i=0 \mid X_{i,1}, X_{i,2}, V_i)} \mid X_{i,1}, X_{i,2}\right] \\
&= E\left[E\left[\mathbf{1}(D_i=0)\varepsilon_i \mid X_{i,1}, X_{i,2}, V_i\right]\frac{1}{P(D_i=0 \mid X_{i,1}, V_i)} \mid X_{i,1}, X_{i,2}\right] \\
&= E\left[\frac{1}{P(D_i=0 \mid X_{i,1}, V_i)}\mathbf{1}(D_i=0)\varepsilon_i \mid X_{i,1}, X_{i,2}\right]
\end{aligned}$$

$$= E\left[\frac{1}{P(D_i = 0 \mid X_{i,1}, V_i)}\mathbf{1}(D_i = 0)(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2) \mid X_{i,1}, X_{i,2}\right]$$

Similarly, it can shown that

$$E\left[\frac{1}{P(D_i = 1 \mid X_{i,1}, V_i)}\mathbf{1}(D_i = 1)\left(Y_i - \beta_0 - X_{i,1}\beta_1 - E[X_{i,2} \mid X_{i,1}, V_i, D_i = 0]\beta_2\right) \mid X_{i,1}\right] = 0$$

We then have a similar structure as before because we can impute $X_{i,2}$ with an estimate of $E[X_{i,2} \mid X_{i,1}, V_i, D_i = 0]$ and use an inverse probability weighted estimator with an estimate of the nuisance functions are $P(D_i = 0 \mid X_{i,1}, V_i)$.

This previous approach does not require an assumption on how V_i relates to ε_i. Now suppose we also assume that

$$E[\varepsilon_i \mid X_i, V_i] = 0$$

Using the previous arguments, it is easy to derive the unconditional moments

$$E\left[(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2) \mid X_{i,1}, X_{i,2}, V_i, D_i = 0\right] = 0$$

and

$$E\left[(Y_i - \beta_0 - X_{i,1}\beta_1 - E[X_{i,2} \mid X_{i,1}, V_i, D_i = 0]\beta_2) \mid X_{i,1}, V_i, D_i = 1\right] = 0$$

A.3 Projection

We now briefly discuss how to allow for $E\left[X_{it,k}|X_{it}^{(l)}, D_{it} = l\right] \neq X_{it}^{(l)\prime}\gamma_t^{(l,k)}$ by using arguments based on projections. In this case $Z_{it,k}^{(l)} = X_{it}^{(l)\prime}\gamma_t^{(l,k)}$ can be interpreted as the linear projection of $X_{it,k}$ onto $X_{it}^{(l)}$ under missing pattern l, based on the complete subset of the data. By definition of a linear projections, it then holds that

$$E\left[\mathbf{1}(D_{it} = 0)u_{it,k}^{(l)}X_{it}^{(l)}\right] = 0$$

for all $l = 0, 1, \ldots, L$ and $k \notin I_t^{(l)}$ and with $u_{it,k}^{(l)} = X_{it,k} - Z_{it,k}^{(l)}$. These are exactly the moment condition in equation (3). The moment conditions in equation (1) hold as long as $E[\varepsilon_{it} \mid X_{it}^{(0)}, D_{it} =$

0] = 0, which follows from our previously imposed MAR assumption. Finally, for the moment conditions in equation (2), we can use our assumption $E[\varepsilon_{it} \mid X_{it}^{(l)}, D_{it} = 0] = 0$ to write

$$\begin{aligned} E\left[\mathbf{1}(D_{it} = l)(Y_{it} - \sum_{k=1}^{K} \beta_{t,k} Z_{it,k}^{(l)}) X_{it}^{(l)}\right] &= E\left[\mathbf{1}(D_{it} = l)\left(\varepsilon_{it} + \sum_{k=1}^{K} \beta_{t,k} u_{it,k}^{(l)}\right) X_{it}^{(l)}\right] \\ &= \sum_{k=1}^{K} \beta_{t,k} E\left[\mathbf{1}(D_{it} = l) u_{it,k}^{(l)} X_{it}^{(l)}\right] \end{aligned}$$

Hence, the moment conditions hold as longs as

$$E\left[\mathbf{1}(D_{it} = l) u_{it,k}^{(l)} X_{it}^{(l)}\right] = 0$$

for all $l = 0, 1, \ldots, L$, which we can also write as

$$E\left[\mathbf{1}(D_{it} = l) u_{it,k}^{(l)} X_{it}^{(l)}\right] = \left[\mathbf{1}(D_{it} = 0) u_{it,k}^{(l)} X_{it}^{(l)}\right]$$

This equation holds as long the linear projection of $X_{it,k}$ on $X_{it}^{(l)}$ does not depend on D_{it}, which is analogous to the second part of the previous MAR assumption, namely

$$E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l\right] = E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = 0\right].$$

A.4 Equivalence GLS and Optimal GMM

Consider the moment conditions

$$E\left[\mathbf{1}(D_{it} = 0)\left(Y_{it} - \sum_{k=1}^{K} \beta_{t,k} X_{it,k}^{(0)}\right) X_{it}^{(0)}\right] = 0 \tag{A.4}$$

$$E\left[\mathbf{1}(D_{it} = l)\left(Y_{it} - \sum_{k=1}^{K} \beta_{t,k} Z_{it,k}^{(l)}\right) X_{it}^{(l)}\right] = 0 \qquad l = 1, \ldots, L \tag{A.5}$$

$$E\left[\mathbf{1}(D_{it} = 0)\left(X_{it,k} - X_{it}^{(l)\prime} \gamma_t^{(l,k)}\right) X_{it}^{(l)}\right] = 0 \qquad l = 1, \ldots, L \text{ and } k \notin I_t^{(l)} \tag{A.6}$$

To show equivalence of the GLS and the optimal GMM estimator, we impose the following additional assumptions:

- γ_t is known.
- $E\left[\varepsilon_{it}^2 \mid X_{it}^{(0)}, D_i = l\right] = \sigma_{\varepsilon,t}^2$
- $E\left[u_{it}^{(l)} u_{it}^{(l)\prime} \mid X_{it}^{(l)}, D_{it} = l\right] = \Sigma_t^{(l)}$ for all $l = 1, \dots L$, where $u_{it,k}^{(l)} = X_{it,k} - X_{it}^{(l)\prime}\gamma_t^{(l,k)}$ for all $k \notin I_t^{(l)}$.

The last two conditions assume that the unobservables are homoskedastic.

We start by analyzing the GMM estimator. Since γ_t is known, we can ignore the moment conditions in (A.6). Now define

$$g_{it}(\beta_t) = \begin{pmatrix} \mathbf{1}(D_{it} = 0)\left(Y_{it} - \sum_{k=1}^K \beta_{t,k} X_{it,k}^{(0)}\right) X_{it}^{(0)} \\ \mathbf{1}(D_{it} = 1)\left(Y_{it} - \sum_{k=1}^K \beta_{t,k} Z_{it,k}^{(1)}\right) X_{it}^{(1)} \\ \vdots \\ \mathbf{1}(D_{it} = L)\left(Y_{it} - \sum_{k=1}^K \beta_{t,k} Z_{it,k}^{(L)}\right) X_{it}^{(L)} \end{pmatrix}$$

The GMM estimator minimizes the sample analog of $E[g_{it}(\beta_t)]' W E[g_{it}(\beta_t)]$. The efficient matrix is the block-diagonal matrix

$$\begin{aligned} W &= E[g_{it}(\beta_t) g_{it}(\beta_t)']^{-1} \\ &= \text{diag}\left(w^{(l)}\right)^{-1} \end{aligned}$$

where $w^{(l)}$ is the $dim(X_{it}^{(l)}) \times dim(X_{it}^{(l)})$ matrix

$$w^{(l)} = E\left[\mathbf{1}(D_{it} = l) X_{it}^{(l)} X_{it}^{(l)\prime} \left(Y_{it} - \sum_{k=1}^K \beta_{t,k} Z_{it,k}^{(l)}\right)^2\right]$$

The remaining element are zero because $\mathbf{1}(D_{it} = k)\mathbf{1}(D_{it} = l) = 0$ for $k \neq l$. The first diagonal block, $w^{(0)}$, can be expressed as

$$w^{(0)} = E\left[\mathbf{1}(D_{it} = 0) X_{it}^{(0)} X_{it}^{(0)\prime} \varepsilon_{it}^2\right]$$

Using $E\left[\varepsilon_{it}^2 \mid X_{it}^{(0)}, D_i = 0\right] = \sigma_{\varepsilon,t}^2$ we can write it as

$$E\left[\mathbf{1}(D_{it}=0)X_{it}^{(0)}X_{it}^{(0)\prime}\varepsilon_{it}^2\right] = \sigma_{\varepsilon,t}^2 E\left[\mathbf{1}(D_{it}=0)X_{it}^{(0)}X_{it}^{(0)\prime}\right]$$

For the other blocks, we can write

$$\begin{aligned} w^{(l)} &= E\left[\mathbf{1}(D_{it}=l)X_{it}^{(l)}X_{it}^{(l)\prime}\left(Y_{it} - \sum_{k=1}^{K}\beta_{t,k}Z_{it,k}^{(l)}\right)^2\right] \\ &= E\left[\mathbf{1}(D_{it}=l)X_{it}^{(l)}X_{it}^{(l)\prime}\left(\varepsilon_{it} + \sum_{k=1}^{K}\beta_{t,k}u_{it,k}^{(l)}\right)^2\right] \end{aligned}$$

Let $\beta_t^{(l)}$ be the subvector of β_t with entries $\beta_{t,k}$ with $k \notin I_t^{(l)}$. Our assumptions above then imply that

$$\begin{aligned} E\left[\left(\varepsilon_{it} + \sum_{k=1}^{K}\beta_{t,k}u_{it,k}^{(l)}\right)^2 \mid X_{it}^{(l)}, D_{it}=l\right] &= E\left[\varepsilon_{it}^2 \mid X_{it}^{(l)}, D_{it}=l\right] \\ &\quad + 2E\left[\varepsilon_{it}\left(\sum_{k=1}^{K}\beta_{t,k}u_{it,k}^{(l)}\right) \mid X_{it}^{(l)}, D_{it}=l\right] \\ &\quad + E\left[\left(\sum_{k=1}^{K}\beta_{t,k}u_{it,k}^{(l)}\right)^2 \mid X_{it}^{(l)}, D_{it}=l\right] \\ &= \sigma_{\varepsilon,t}^2 + \beta_t^{(l)\prime}\Sigma_t^{(l)}\beta_t^{(l)} \end{aligned}$$

The cross terms are 0 because

$$E\left[\varepsilon_{it}\left(\sum_{k=1}^{K}\beta_{t,k}u_{it,k}^{(l)}\right) \mid X_{it}^{(l)}, D_{it}=l\right] = \sum_{k=1}^{K}\beta_{t,k}E\left[u_{it,k}^{(l)}E\left(\varepsilon_{it} \mid X_{it}, D_{it}=l\right) \mid X_{it}, D_{it}=l\right] = 0$$

It then follows that

$$w^{(l)} = \left(\sigma_{\varepsilon,t}^2 + \beta_t^{(l)\prime}\Sigma_t^{(l)}\beta_t^{(l)}\right)E\left[\mathbf{1}(D_{it}=l)X_{it}^{(l)}X_{it}^{(l)\prime}\right]$$

for $l = 1, \ldots, L$.

The feasible optimal GMM estimator minimizes $\bar{g}(\beta_t)'\hat{W}\bar{g}(\beta_t)$ where $\bar{g}(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_{it}(\beta)$ and

$\hat{W} = \text{diag}\left(\hat{w}^{(l)}\right)^{-1}$ with

$$\hat{w}^{(0)} = \hat{\sigma}^2_{\varepsilon,t} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(D_{it} = 0) X_{it}^{(0)} X_{it}^{(0)\prime}$$
$$\hat{w}^{(l)} = \left(\hat{\sigma}^2_{\varepsilon,t} + \left(\hat{\beta}_t^{(l)}\right)' \hat{\Sigma}_t^{(l)} \hat{\beta}_t^{(l)} \right) \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(D_{it} = l) X_{it}^{(l)} X_{it}^{(l)\prime}$$

We require that $\hat{\sigma}^2_{\varepsilon,t} \xrightarrow{p} \sigma^2_{\varepsilon,t}$, $\hat{\beta}_t^{(l)} \xrightarrow{p} \beta_t^{(l)}$ and $\hat{\Sigma}_t^{(l)} \xrightarrow{p} \Sigma_t^{(l)}$, which can be achieved by estimating the parameters using the complete case. We then get $\hat{W} \xrightarrow{p} W$.

The first-order conditions are

$$\frac{\partial}{\partial \beta_t} \bar{g}(\beta_t)' \hat{W} \bar{g}(\beta_t) = 0$$

with

$$\frac{\partial}{\partial \beta_t} \bar{g}(\beta_t) = \begin{pmatrix} -\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(D_{it} = 0) X_{it}^{(0)} X_{it}^{(0)\prime} \\ -\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(D_{it} = 1) X_{it}^{(1)} Z_{it}^{(1)\prime} \\ \vdots \\ -\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(D_{it} = L) X_{it}^{(L)} Z_{it}^{(L)\prime} \end{pmatrix}$$

Solving the first order conditions yields the following closed-form expression for the optimal GMM estimator:

$$\hat{\beta}_{t,GMM} = \left(\frac{\partial}{\partial \beta_t} \bar{g}(V_{it}, \hat{\beta}_t)' \hat{W} \frac{\partial}{\partial \beta_t} \bar{g}(\beta_t) \right)^{-1} \frac{\partial}{\partial \beta_t} \bar{g}(V_{it}, \hat{\beta}_t)' \hat{W} \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{1}(D_{it} = 0) X_{it}^{(0)} Y_{it} \\ \mathbf{1}(D_{it} = 1) X_{it}^{(1)} Y_{it} \\ \vdots \\ \mathbf{1}(D_{it} = L) X_{it}^{(L)} Y_{it} \end{pmatrix}$$

We will now rewrite this estimator to relate it to the GLS estimator. Consider

$$\left(\frac{\partial}{\partial\beta_t}\bar{g}(V_{it},\hat{\beta}_t)'\hat{W}\right)' = \begin{pmatrix} -\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=0)X_{it}^{(0)}X_{it}^{(0)'}\left(\hat{w}^{(0)}\right)^{-1} \\ -\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=1)Z_{it}^{(1)}X_{it}^{(1)'}\left(\hat{w}^{(1)}\right)^{-1} \\ \vdots \\ -\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=L)Z_{it}^{(L)}X_{it}^{(L)'}\left(\hat{w}^{(L)}\right)^{-1} \end{pmatrix}$$

The first element is simply

$$-\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=0)X_{it}^{(0)}X_{it}^{(0)'}\left(\hat{w}^{(0)}\right)^{-1} = -\left(\hat{\sigma}^2_{\varepsilon,t}\right)^{-1} I_{K\times K}$$

Next, we assume without loss of generality that the elements in $Z_{it}^{(l)}$ are ordered such that $Z_{it}^{(l)} = \left(X_{it}^{(l)'}, X_{it}^{(l)'}\gamma_t^{(l)'}\right)'$. Define $J_t^{(l)} = \left|(I_t^{(l)})^c\right|$. Then for the l-th element

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=l)Z_{it}^{(l)}X_{it}^{(l)'}\left(\hat{w}^{(l)}\right)^{-1} &= -\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=l)\begin{pmatrix} X_{it}^{(l)} \\ \gamma_t^{(l)}X_{it}^{(l)} \end{pmatrix} X_{it}^{(l)'}\left(\hat{w}^{(l)}\right)^{-1} \\ &= -\begin{pmatrix} I_{(K-J_t^{(l)})\times(K-J_t^{(l)})} \\ \gamma_t^{(l)} \end{pmatrix}\frac{1}{n}\sum_{i=1}^n \mathbf{1}(D_{it}=l)X_{it}^{(l)}X_{it}^{(l)'}\left(\hat{w}^{(l)}\right)^{-1} \\ &= -\left(\hat{\sigma}^2_{\varepsilon,t} + \left(\hat{\beta}_t^{(l)}\right)'\hat{\Sigma}_t^{(l)}\hat{\beta}_t^{(l)}\right)^{-1}\begin{pmatrix} I_{(K-J_t^{(l)})\times(K-J_t^{(l)})} \\ \gamma_t^{(l)} \end{pmatrix} \end{aligned}$$

It follows that

$$\frac{\partial}{\partial\beta_t}\bar{g}(V_{it},\hat{\beta}_t)'\hat{W}\frac{\partial}{\partial\beta_t}\bar{g}(\beta_t) = -\frac{1}{n}\sum_{i=1}^n\left\{\frac{\mathbf{1}(D_{it}=0)Z_{it}^{(0)}Z_{it}^{(0)'}}{\hat{\sigma}^2_{\varepsilon,t}} + \sum_{l=1}^L \frac{\mathbf{1}(D_{it}=l)Z_{it}^{(l)}Z_{it}^{(l)'}}{\hat{\sigma}^2_{\varepsilon,t} + \left(\hat{\beta}_t^{(l)}\right)'\hat{\Sigma}_t^{(l)}\hat{\beta}_t^{(l)}}\right\}$$

where $X_{it}^{(0)} = Z_{it}^{(0)}$. Define

$$(\hat{\sigma}_t^{(l)})^2 := \begin{cases} \hat{\sigma}^2_{\varepsilon,t} & \text{if } l=0 \\ \hat{\sigma}^2_{\varepsilon,t} + \left(\hat{\beta}_t^{(l)}\right)'\hat{\Sigma}_t^{(l)}\hat{\beta}_t^{(l)} & \text{otherwise} \end{cases}$$

Then

$$\frac{\partial}{\partial \beta_t}\bar{g}(V_{it},\hat{\beta}_t)'\hat{W}\frac{\partial}{\partial \beta_t}\bar{g}(\beta_t) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Z_{it}^{(l)\prime}}{(\hat{\sigma}_t^{(l)})^2}$$

Using the same arguments we can also write

$$\hat{W}\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{1}(D_{it}=0)X_{it}^{(0)}Y_{it}\\ \mathbf{1}(D_{it}=1)X_{it}^{(1)}Y_{it}\\ \vdots \\ \mathbf{1}(D_{it}=L)X_{it}^{(L)}Y_{it}\end{pmatrix} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Y_{it}}{(\hat{\sigma}_t^{(l)})^2}$$

Hence

$$\hat{\beta}_{t,GMM} = \left(\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Z_{it}^{(l)\prime}}{(\hat{\sigma}_t^{(l)})^2}\right)^{-1}\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Y_{it}}{(\hat{\sigma}_t^{(l)})^2}$$

Next, consider the GLS estimator, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{(Y_{it}-Z_{it}^{(l)\prime}\beta_t)^2}{(\hat{\sigma}_t^{(l)})^2}$$

The first-order conditions are

$$\begin{aligned}0 &= \sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Y_{it}-Z_{it}^{(l)}Z_{it}^{(l)\prime}\hat{\beta}_{t,GLS}}{(\hat{\sigma}_t^{(l)})^2}\\ \Leftrightarrow\ \hat{\beta}_{t,GLS} &= \left(\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Z_{it}^{(l)\prime}}{(\hat{\sigma}_t^{(l)})^2}\right)^{-1}\sum_{i=1}^{n}\sum_{l=0}^{L}\mathbf{1}(D_{it}=l)\frac{Z_{it}^{(l)}Y_{it}}{(\hat{\sigma}_t^{(l)})^2}\end{aligned}$$

Therefore

$$\hat{\beta}_{t,GMM} = \hat{\beta}_{t,GLS}.$$

A.5 Large Sample Distribution

Let $\gamma_t = \{\gamma_t^{(l,k)}\}_{l=1,...,L,k\notin I_t^{(l)}}$ and define

$$g_{it,1}(\beta_t,\gamma_t) = \begin{pmatrix} \mathbf{1}(D_{it}=0)\left(Y_{it} - \sum_{k=1}^K \beta_{t,k} X_{it}^{(0)}\right) X_{it}^{(0)} \\ \mathbf{1}(D_{it}=1)\left(Y_{it} - \sum_{k=1}^K \beta_{t,k} Z_{it,k}^{(1)}\right) X_{it}^{(1)} \\ \vdots \\ \mathbf{1}(D_{it}=L)\left(Y_{it} - \sum_{k=1}^K \beta_{t,k} Z_{it,k}^{(L)}\right) X_{it}^{(L)} \end{pmatrix}$$

and

$$g_{it,2}(\gamma_t) = \begin{pmatrix} \left\{\mathbf{1}(D_{it}=0)\left(X_{it,k} - X_{it}^{(1)\prime}\gamma_t^{(1,k)}\right) X_{it}^{(1)}\right\}_{k\notin I_t^{(1)}} \\ \vdots \\ \left\{\mathbf{1}(D_{it}=0)\left(X_{it,k} - X_{it}^{(L)\prime}\gamma_t^{(L,k)}\right) X_{it}^{(L)}\right\}_{k\notin I_t^{(L)}} \end{pmatrix}$$

We will derive the large sample distribution of any GMM estimator which minimizes a sample analog estimator of

$$\begin{pmatrix} E[g_{it,1}(\beta_t,\gamma_t)] & E[g_{it,2}(\gamma_t)] \end{pmatrix} \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix} \begin{pmatrix} E[g_{it,1}(\beta_t,\gamma_t)] \\ E[g_{it,2}(\gamma_t)] \end{pmatrix}$$

We then show that both the two-step OLS and GLS estimators are special cases for particular choices of W_1 and W_2. In particular, we will take $W_2 = \frac{1}{w_2} I_{dim(g_{it,2})\times dim(g_{it,2})}$, $w_2 \to 0$, and $I_{dim(g_{it,2})\times dim(g_{it,2})}$ is an identity matrix. Intuitively, we put infinite weight on the second set of moment conditions, which implies that we solve the sample analog exactly. We show that the limit is well defined and derive an expression for the corresponding standard errors.

Define

$$\bar{g}_1(\beta_t,\gamma_t) = \frac{1}{n}\sum_{i=1}^n g_{it,1}(\beta_t,\gamma_t)$$

and

$$\bar{g}_2(\gamma_t) = \frac{1}{n}\sum_{i=1}^n g_{it,2}(\gamma_t)$$

The objective function is then

$$\bar{g}_1(\beta_t,\gamma_t)' W_1 \bar{g}_1(\beta_t,\gamma_t) + \bar{g}_2(\gamma_t)' W_2 \bar{g}_2(\gamma_t)$$

and the first order conditions are

$$\left(\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)=0$$

and

$$\left(\frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)+\left(\frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\right)' W_2\bar{g}_2\left(\hat{\gamma}_t\right)=0$$

Using $W_2=\frac{1}{w_2}I_{dim(g_{it,2})\times dim(g_{it,2})}$, we can then write the first order condition as

$$\begin{pmatrix}\left(\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & 0\\ w_2\left(\frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & \left(\frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\right)'\end{pmatrix}\begin{pmatrix}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\\ \bar{g}_2\left(\hat{\gamma}_t\right)\end{pmatrix}=0$$

Notice that when $w_2=0$, these are the first order conditions corresponding to the two-step GLS estimator, which we derived in Section A.4 where $W_1=\text{diag}\left(\hat{w}^{(l)}\right)^{-1}$ and expressions for $\hat{w}^{(l)}$ are provided in Section A.4. We obtain the two-step OLS estimator when W_1 is an identity matrix. Using a first-order Taylor expansion, we get

$$\begin{aligned}0=&\begin{pmatrix}\left(\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & 0\\ w_2\left(\frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & \left(\frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\right)'\end{pmatrix}\begin{pmatrix}\bar{g}_1\left(\beta_t,\gamma_t\right)\\ \bar{g}_2\left(\gamma_t\right)\end{pmatrix}\\ &+\begin{pmatrix}\left(\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & 0\\ w_2\left(\frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & \left(\frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\right)'\end{pmatrix}\begin{pmatrix}\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right) & \frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\\ 0 & \frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\end{pmatrix}\begin{pmatrix}\hat{\beta}-\beta\\ \hat{\gamma}-\gamma\end{pmatrix}\\ &+o_p(1/\sqrt{n})\end{aligned}$$

or

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\hat{\beta}_t-\beta_t\\ \hat{\gamma}_t-\gamma_t\end{pmatrix}=&\left(-\begin{pmatrix}\left(\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & 0\\ w_2\left(\frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & \left(\frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\right)'\end{pmatrix}\begin{pmatrix}\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right) & \frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\\ 0 & \frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\end{pmatrix}\right)^{-1}\\ &\times\begin{pmatrix}\left(\frac{\partial}{\partial\beta_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & 0\\ w_2\left(\frac{\partial}{\partial\gamma_t}\bar{g}_1\left(\hat{\beta}_t,\hat{\gamma}_t\right)\right)' W_1 & \left(\frac{\partial}{\partial\gamma_t}\bar{g}_2\left(\hat{\gamma}_t\right)\right)'\end{pmatrix}\sqrt{n}\begin{pmatrix}\bar{g}_1\left(\beta_t,\gamma_t\right)\\ \bar{g}_2\left(\gamma_t\right)+o_p(1)\end{pmatrix}\end{aligned}$$

We know that

$$\sqrt{n}\begin{pmatrix}\bar{g}_1(\beta_t,\gamma_t)\\ \bar{g}_2(\gamma_t)\end{pmatrix}\xrightarrow{d}N(0,\Omega_t)$$

where

$$\Omega_t = E\left[\begin{pmatrix}g_{it,1}(\beta_t,\gamma_t)\\ g_{it,2}(\gamma_t)\end{pmatrix}\begin{pmatrix}g_{it,1}(\beta_t,\gamma_t) & g_{it,2}(\gamma_t)\end{pmatrix}\right]$$

and thus

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_t-\beta_t\\ \hat{\gamma}_t-\gamma_t\end{pmatrix}\xrightarrow{d}N(0,\Sigma_t)$$

where

$$\Sigma_t = \left(D_t'Q_t\right)^{-1}D_t'\Omega_t D_t\left(Q_t'D_t\right)^{-1}$$

where

$$D_t' = \begin{pmatrix}\left(\frac{\partial}{\partial\beta_t}E[g_1(V_{it},\beta_t,\gamma_t)]\right)'W_1 & 0\\ w_2\left(\frac{\partial}{\partial\gamma_t}E[g_1(V_{it},\beta_t,\gamma_t)]\right)'W_1 & \left(\frac{\partial}{\partial\gamma_t}E[g_2(V_{it},\gamma_t)]\right)'\end{pmatrix}$$

and

$$Q_t = \begin{pmatrix}\frac{\partial}{\partial\beta_t}E[g_1(V_{it},\beta_t,\gamma_t)] & \frac{\partial}{\partial\gamma_t}E[g_1(V_{it},\beta_t,\gamma_t)]\\ 0 & \frac{\partial}{\partial\gamma_t}E[g_2(V_{it},\gamma_t)]\end{pmatrix}$$

All these matrix can be estimated using sample analogs. As already mentioned, for the two-step GLS estimator, we simply set $w_2 = 0$ and use W_1 as defined above.

A.6 J-test

Let $\gamma_t = \{\gamma_t^{(l,k)}\}_{l=1,\ldots,L,k\notin I_t^{(l)}}$ and define

$$g_{it,11}(\beta_t) = \left(\mathbf{1}(D_{it}=0)\left(Y_{it}-\sum_{k=1}^K\beta_{t,k}X_{it}^{(0)}\right)X_{it}^{(0)}\right)$$

$$g_{it,12}(\beta_t,\gamma_t) = \begin{pmatrix}\mathbf{1}(D_{it}=1)\left(Y_{it}-\sum_{k=1}^K\beta_{t,k}Z_{it,k}^{(1)}\right)X_{it}^{(1)}\\ \vdots\\ \mathbf{1}(D_{it}=L)\left(Y_{it}-\sum_{k=1}^K\beta_{t,k}Z_{it,k}^{(L)}\right)X_{it}^{(L)}\end{pmatrix}$$

and

$$g_{it,2}\left(\gamma_t\right) = \begin{pmatrix} \left\{\mathbf{1}(D_{it}=0)\left(X_{it,k} - X_{it}^{(1)\prime}\gamma_t^{(1,k)}\right)X_{it}^{(1)}\right\}_{k\notin I_t^{(1)}} \\ \vdots \\ \left\{\mathbf{1}(D_{it}=0)\left(X_{it,k} - X_{it}^{(L)\prime}\gamma_t^{(L,k)}\right)X_{it}^{(L)}\right\}_{k\notin I_t^{(L)}} \end{pmatrix}$$

Let $\hat{\beta}_t$ be the estimator that solves

$$\sum_{i=1}^{n} g_{it,11}\left(\hat{\beta}_t\right) = 0$$

which is our estimator based on the complete case. Let $\hat{\gamma}_t$ be the estimator that solves

$$\sum_{i=1}^{n} g_{it,2}\left(\hat{\gamma}_t\right) = 0$$

which is our standard, period-by-period imputation estimator.

To test our overidentifying restrictions, we test

$$H_0 : E[g_{it,12}\left(\beta_t, \gamma_t\right)] = 0$$

for the values of β_t and γ_t that are identified through the first and third set of moments, respectively. The test statistic will be a quadratic version of the sample analog of these moment conditions.

To derive the test statistic, let $\delta_t = (\beta_t, \gamma_t)$ and write

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} g_{it,12}\left(\hat{\delta}_t\right) &= \frac{1}{n}\sum_{i=1}^{n} g_{it,12}\left(\delta_t\right) + \frac{1}{n}\sum_{i=1}^{n}\left(g_{it,12}\left(\hat{\delta}_t\right) - g_{it,12}\left(\delta_t\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} g_{it,12}\left(\delta_t\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\delta}g_{it,21}\left(\delta_t\right)\right)\left(\hat{\delta}_t - \delta_t\right) + o_p(1/\sqrt{n}). \end{aligned}$$

Hence,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{it,12}\left(\hat{\delta}_t\right) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{it,12}\left(\delta_t\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\delta}g_{it,12}\left(\delta_t\right)\right)\sqrt{n}\left(\hat{\delta}_t - \delta_t\right) + o_p(1).$$

Under the null hypothesis it holds that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{it,12}\left(\delta_t\right) \xrightarrow{d} N(0, E[g_{it,12}\left(\delta_t\right) g_{it,12}\left(\delta_t\right)'])$$

For the second term, it is easy to show that we can write

$$\sqrt{n}\left(\hat{\delta}_t - \delta_t\right) = \begin{pmatrix} \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta} g_{it,11}\left(\beta_t\right)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n g_{it,11}\left(\beta_t\right) \\ \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\gamma} g_{it,2}\left(\gamma_t\right)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n g_{it,2}\left(\gamma_t\right) \end{pmatrix}$$
$$= G_t^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} g_{it,11}\left(\beta_t\right) \\ g_{it,2}\left(\gamma_t\right) \end{pmatrix} + o_p(1)$$

where

$$G_t = \begin{pmatrix} E\left[\frac{\partial}{\partial\beta} g_{it,11}\left(\beta_t\right)\right] & 0 \\ 0 & E\left[\frac{\partial}{\partial\gamma} g_{it,2}\left(\gamma_t\right)\right] \end{pmatrix}$$

Hence

$$\sqrt{n}\left(\hat{\delta}_t - \delta_t\right) \xrightarrow{d} N(0, \Sigma_t)$$

where

$$\Sigma_t = G_t^{-1} E\left[\begin{pmatrix} g_{it,11}\left(\beta_t\right) \\ g_{it,2}\left(\gamma_t\right) \end{pmatrix}\begin{pmatrix} g_{it,11}\left(\beta_t\right) \\ g_{it,2}\left(\gamma_t\right) \end{pmatrix}'\right](G_t')^{-1}$$

It follows that

$$\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)\right)\sqrt{n}\left(\hat{\delta}_t - \delta_t\right) \xrightarrow{d} N\left(0, E\left[\frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)\right]\Sigma_t E\left[\frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)'\right]\right)$$

The two normals are independent because they are based on different subsets of the data. Hence,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n g_{it,12}\left(\hat{\delta}_t\right) \xrightarrow{d} N\left(0, E[g_{it,12}\left(\delta_t\right) g_{it,12}\left(\delta_t\right)'] + E\left[\frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)\right]\Sigma_t E\left[\frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)'\right]\right)$$

Let $\hat{\Omega}_t$ be a consistent estimator of $E[g_{it,12}\left(\delta_t\right) g_{it,12}\left(\delta_t\right)'] + E\left[\frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)\right]\Sigma_t E\left[\frac{\partial}{\partial\delta} g_{it,12}\left(\delta_t\right)'\right]$. Then

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n g_{it,12}\left(\hat{\delta}_t\right)\right)' \hat{\Omega}_t^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n g_{it,12}\left(\hat{\delta}_t\right)\right) \xrightarrow{d} \chi^2_{d_{12}}$$

where d_{12} is the dimension of $g_{it,12}(\delta_t)$. We therefore reject the null hypothesis if

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{it,12}\left(\hat{\delta}_t\right)\right)' \hat{\Omega}_t^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{it,12}\left(\hat{\delta}_t\right)\right)$$

is larger than the $1-\alpha$ quantile of the $\chi^2_{d_{12}}$ distribution.