Missing Data in Asset Pricing Panels*

Joachim Freyberger† Bjoern Hoeppner‡

Andreas Neuhierl§ Michael Weber¶

First draft: September 2021
This draft: January 2024

Abstract

We propose a simple and computationally attractive method to deal with missing data in cross-sectional asset pricing using conditional mean imputations and weighted least squares, cast in a generalized method of moments (GMM) framework. This method allows us to use all observations with observed returns; it results in valid inference; and it can be applied in nonlinear and high-dimensional settings. In simulations, we find it performs almost as well as the efficient but computationally costly GMM estimator. We apply our procedure to a large panel of return predictors and find that it leads to improved out-of-sample predictability. (*JEL* C14, C58, G12)

*We thank Gurdip Bakshi, Bruce Carlin, Andrew Chen, Zhuo Chen, Xiaohong Chen, Alex Chinco, Kevin Crotty, Wayne Ferson, Todd Gormely, Lena Janys, Andrew Karolyi, Soohun Kim, Hugues Langlois, Yan Liu, Asaf Manela, Joon Park, Markus Pelger, Christoph Rothe, Oleg Rytchkov, Jan Scherer, Gustavo Schwenkler, Takuya Ura, and Guofu Zhou and conference and seminar participants at the KAIST, University of Bonn, University of Bath, UC Berkeley, Georgia State University, Lund University, University of Gotheburg, University of Oklahoma, University of Mannheim, University of Maryland, University of Oklahoma, SFI Lugano, University of Virginia, University of Washington, Warwick Business School, Washington University in St. Louis, The Ohio State University, Rice University, Stockholm School of Economics, Temple University, Tulane University, TU Muenchen, UCLA, Queen Mary University, Yale University, the 2022 AFA meetings, the NBER Big Data meetings, the World Symposium for Investment Research, HEC-McGill Winter Conference, the European Finance Association Annual Meeting, the VfS Annual Conference 2022, and EcoSta 2022 for helpful comments and discussions. Jakob Juergens provided valuable research assistance. Send correspondence to Michael Weber, michael.weber@chicagobooth.edu.

†University of Bonn, Germany

‡University of Bonn, Germany

§Olin School of Business, Washington University in St. Louis, USA

¶Booth School of Business, the University of Chicago, USA, CEPR, and NBER

Missing data are a common problem in cross-sectional asset pricing studies. Whereas the problem of missing return observations has received some attention and is typically handled by the use of so-called delisting returns (Shumway, 1997; Beaver, McNichols, and Price, 2007), the problem of missing covariates, such as firm characteristics, is typically only addressed implicitly. A large and growing literature uses these covariates to predict future returns cross-sectionally or to build factor portfolios. Most studies in this literature do not explicitly discuss how they handle the case of missing data. For the ones that do, by far the most common procedure to deal with missing covariates is to exclude an observation altogether if any covariate is missing and perform the subsequent analysis only on observations for which no covariates or returns are missing (complete cases analysis). Alternatively, researchers impute the unconditional mean for a missing characteristic from the firms with no missing data (unconditional mean imputation). As we will argue below, both procedures have undesirable properties.

To harness the additional power from studying all firms with valid return observations, we propose a simple approach to impute the missing covariate observations. At an intuitive level, our approach replaces the missing covariates with suitable estimates and accounts for the estimation error (from generating these estimates) in the subsequent analysis. In addition, we also "down-weigh" imputed observations, accounting for the fact that these data points are not truly observed and thus contain less information. In general, the more covariates are imputed, the larger the additional error terms due to imputations, and 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 and/or from the time series of past observations. The method can be used if the main model for return prediction is parametric or nonparametric and does not require us to specify the entire distribution of the missing covariates. Finally, we can cast our method into a generalized method of moments (Hansen (1982)) setting, which allows us to study its statistical properties.

In recent years, many asset pricing papers aim to respond to the Cochrane (2011) multidimensional challenge, that is, identifying which characteristics and factors help predict returns conditional on other predictors. The large number of possible predictors aggravates

the missing data problem (Harvey, Liu, and Zhu, 2016). The complete case analysis typically neglects a substantial subset of the data. For example, in our paper, we use the data set of Chen and Zimmermann (forthcoming) with 82 covariates that contains around 2.4 million observations between 1978 and 2021. Whereas the complete case only consists of around 10% of the overall sample, for almost half of the observations, at most 5 of the 82 covariates are missing. These observations with few missing covariates would then be excluded from the analysis, even though they contain useful information. This exclusion is in contrast to what Zhang, Mykland, and Aït-Sahalia (2005) call "one of statistics' first principles," namely, "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, which might preclude us from forming portfolios with high out-of-sample Sharpe ratios.

To not lose too many observations, some researchers replace missing values of the covariates with their cross-sectional mean (unconditional mean imputation) of that period. We fully agree with the aim of using as many return observations as possible. However, we also show unconditional mean imputation is rarely desirable. First, unconditional mean imputation leads to inconsistent estimators, except in the special cases when the covariates are independent or when covariates with imputed characteristics are no true return predictors. Second, even in these special cases, unconditional mean imputation typically produces incorrect standard errors. Intuitively, unconditional mean imputation leads to an underestimation of (co)variances and therefore standard errors that are often too small.

The mapping of our proposed estimator into a GMM framework allows us to account for the imputation step in conducting inference and also to understand the efficiency gains of the proposed approach. Contrary to many Bayesian and likelihood-based approaches that address the issue of missing data, such as multiple imputation or the EM algorithm, our method is computationally inexpensive and places fewer assumptions on the data generating process (DGP). 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 missing

characteristics, but it can arbitrarily depend on the always observed characteristics. For example, our approach allows for small firms having a higher likelihood of missing characteristics. We characterize the conditions under which we obtain consistent estimators and correct inference and we argue that these conditions are plausible in many empirical asset pricing studies.

We then illustrate the finite sample properties of our approach in an extensive simulation study and find that it performs well in samples of realistic size. The simulations also help understand when the *ad hoc* approaches, such as unconditional mean imputation and complete case analysis, are (and are not) problematic. We also compare our method to other contemporaneous imputation methods, such as the factor model, for characteristics of Bryzgalova et al. (2023) and the EM algorithm of Chen and McCoy (forthcoming) to illustrate the relative advantages and disadvantages.

Finally, we apply our method to the CRSP/Compustat sample. It is desirable to use all firms with valid returns, because conditioning on the complete cases ignores an interesting part of the return distribution. Portfolios going long stocks with high predicted returns and shorting stocks with low predicted returns achieve much higher out-of-sample returns and Sharpe ratios when using the full sample and imputing missing predictors using our method. In addition, we illustrate how our approach can be used for inference by carrying out a model selection analysis over the full sample to determine the most important predictors. Contrary to our method, the inefficient complete case analysis discards many, even well-established predictors, such as size or value, because of a lack of statistical power. We also document that unconditional mean imputation can lead to incorrect inference because of the generically biased estimators and artificially small standard errors.

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 them. 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 related issues 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 summarize 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 methods that are closely related to our proposed method and place special emphasis on the treatment of missing data in asset pricing. In general, no single procedure can be successfully applied to all missing data problems. Dealing with missing data successfully requires taking a stance on *why* the data are missing, that is, the so-called "missing mechanism."[1] If the probability that a particular observation is missing depends on missing variables (even after conditioning on observables), we call the mechanism 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 in which the probability of observing an observation 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 (e.g., likelihood-based approaches and Bayesian methods) that we do not want to impose. Instead, we use a method based on moment restrictions and imputation, that is, 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). As with our approach, 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

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

[2] Brown et al. (1992) and Carhart et al. (2002) are examples of studies in fund performance in which such a situation arises.

uncertainty" in inference, because the imputations are estimates themselves. The idea we follow goes back to Gourieroux and Monfort (1981) who also allow for only a single missing pattern. One way to approach this issue is to cast the imputation and main model in a 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. For a similar setup, Chen, Hong, and Tarozzi (2008) present a semiparametrically efficient estimator that is based on moment restrictions in the presence of missing data. One drawback of the optimal GMM estimator is that it can be computationally very costly as it amounts to solving a nonlinear optimization problem. These problems are also well-documented in macro-finance applications (e.g., Hansen, Heaton, and Yaron (1996)). In our application with general missing patterns and many return predictors, the efficient GMM estimator is computationally infeasible and it does not have the intuitive interpretation of an imputation estimator. 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 simulations 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), that is, reweighting the complete case sample such that it more closely mirrors the population (Robins, Rotnitzky, and Zhao, 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 that uses the whole sample. Under certain assumptions, which differ slightly from our setup, Robins, Rotnitzky, and Zhao (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, e.g., Tsiatis and

[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 (2021) also propose a two-step approach to test linear factor models; notably, they obtain optimality results in this case.

Davidian, 2015).

While most papers do not explicitly state how they treat missing data, using only the complete case seems the most common approach in asset pricing studies. Recent examples include Lewellen (2015), Freyberger, Neuhierl, and Weber (2020), Kelly, Pruitt, and Su (2019), and Kim, Korajczyk, and Neuhierl (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, Maslov, and Rytchkov, 2017; Kozak, Nagel, and Santosh, 2020; Gu, Kelly, and Xiu, 2020).

More recently, some contemporaneous papers also rigorously deal with the problem of missing predictors in multivariate (cross-sectional) asset pricing studies and propose alternative imputation methods. Compared to those paper, an important conceptual difference of our paper is that we consider imputation and estimation of parameters of asset pricing models as a joint problem. As a consequence, how we impute missing characteristics depends on the model being estimated. In a nonlinear model, we directly impute nonlinear functions instead of using nonlinear functions of imputations. Moreover, how the model is estimated directly depends on the quality of the imputations because we down-weigh imputed observations. This joint treatment then allows us to obtain the statistical properties and valid standard errors of the parameters of interest.

Other recent papers mainly focus on imputation and consider estimation with the imputed sample in a separate step. Bryzgalova et al. (2023) assume a latent factor model for firm characteristics to impute missing values. Similar to our setup, their approach allows the imputation models to be flexibly estimated using information from the cross-section and/or time series. Such an approach yields consistent imputed values when the number of characteristics approaches infinity. The paper focuses on imputation and does not discuss potential adjustments for subsequent estimation and inference. Chen and McCoy (forthcoming) use the EM-algorithm for imputation, which requires that the characteristics are jointly normally distributed, and compare out-of-sample predictions to those with unconditional mean imputation. We discuss those two methods in greater detail in our simulation study. Beckmeyer and Wiedmann (2023) use a machine learning algorithm borrowed from natural language processing to impute missing values, but do not spell out the required assumptions

and theoretical properties. In an earlier contribution, Haugen and Baker (1996) worry if a potential bias may arise from using only the fully observed cases.

Connor and Korajczyk (1987), Lynch and Wachter (2013), Kim and Skoulakis (2018), and Liu, Tang, and Zhou (2022), Koh et al. (2022) 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 the general issue of missing data in empirical studies. Similar to our approach, Lynch and Wachter (2013) cast the problem of missing data in an unbalanced panel in a GMM framework but do not follow an imputation-based approach. Other recent papers that deal with missing data in factor models include Bai and Ng (2021), Cahan, Bai, and Ng (2023), Jin, Miao, and Su (2021), and Xiong and Pelger (2023). Lastly, Harvey, Liu, and Zhu (2016) recognize that unreported tests for the significance of 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.

1 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 United States and trading on NYSE, Nasdaq, or Amex. Balance sheet data come from Compustat.

To avoid potential lock-ahead biases, we lag all characteristics that build on Compustat annual by at least 6 months and all that build on Compustat quarterly by at least 4 months. Our main data set comes from Chen and Zimmermann (forthcoming) and consists of 82 firm characteristics that are available from 1978 to 2021. The firm characteristics feature a combination of accounting information as well as functions of past returns and trading volume. Appendix Table A.1 provides an overview of the characteristics we use in our main empirical analysis. The 82 characteristics are a subset of the characteristics in the original data set. If we were to use all available characteristics, no complete case would exist. We select the subset of the available characteristics based on three rationales. First, we keep all standard characteristics in the empirical asset pricing literature, for instance beta, book-to-

market, and size. Second, some characteristics exist multiple times with only minor variation, in which case we only include one of them (e.g., idiosyncratic volatility can be estimated against various factor models). Finally, we exclude characteristics that are rarely observed. For example, long-run seasonality requires 20 years of past observations. Our data set then includes all common characteristics while having a complete case of at least 285 observations in every period. We use data from 1978 because several accounting variables have only been recorded starting in the 1970s, such as net debt financing.

Appendix Table A.1 also shows the fraction of missing values per characteristic. Overall, we have a total of 3,644,481 firm-month observations. Fama and French (1992) define the benchmark for empirical analyses of the cross-section of expected returns. We follow them and require that a minimum of information is available for each firm. As Fama and French (1992), we require the inputs (market beta, size, and book-to-market) of the Fama-French three-factor model to be available for all firms. When we condition on firms having `Beta`, `BMdec` and `Size` available, we have a total 2,408,182 firm-month observations. The complete case sample instead consists of only 238,198 firm-month observations, that is, the complete case would discard around 90% of all available return observations, making the complete case analysis rather inefficient. As we will detail in Section 2, in our proposed method, we essentially assume the data are missing at random conditional on the observed characteristics. If we drop an additional 2,140 observations from the 2,408,182 firm-month observations, we always observe the following characteristics: `AssetGrowth`, `Beta`, `BMdec`, `BookLeverage`, `ChInv`, `Coskewness`, `DelCOA`, `DelLTI`, `High52`, `IdioRisk`, `MaxRet`, `Size` and `STreversal`. Hence, by discarding very few additional observations, our assumption that missing is at random *conditional on the observed characteristics* is more palatable. Our final data set then has a total of 2,406,042 firm-month observations. In the empirical analysis we always apply the rank-transformation as in Freyberger, Neuhierl, and Weber (2020) such that the characteristics are uniformly distributed on $[0, 1]$, a standard transformation, which is also applied in Kozak, Nagel, and Santosh (2020), Gu, Kelly, and Xiu (2020), and many other papers.

1.1 Missing data in CRSP/Compustat

In this section, we provide descriptive statistics to document the prevalence of the missing data problem in a standard data set for empirical asset pricing but similar conclusions also hold for many studies in empirical corporate finance or international finance. Appendix Table A.1 provides a first overview. From this table, we can see that some characteristics are missing more frequently than others. To understand the broader effects and convey more intuition, Figure 1 gives a first graphical overview. Panel A shows we observe all predictors for only 10% of observations. Moreover, whereas many observations are only missing few predictors, a nontrivial fraction of observations are missing approximately 35 to 45 predictors. Hence, even discarding a handful of characteristics with particularly many missing values will not solve the missing data problem. Panel B shows the fraction of incomplete observations over time and separately by size quintiles. On average, larger firms have fewer incomplete observations, however about 60% to 80% of the firms in the largest size quintile are also incomplete. Panel B also illustrates that the problem of missing data does not vanish over time. It is present and severe even for the most recent years.

Figure 1: Overview of missing characteristics

Panel A shows which fraction of the data are missing for a given number of characteristics. For example, about 10% of the observations have no missing characteristic, and about 8% of observations have exactly one missing characteristic. Panel B shows the fraction of incomplete observations separately for each size quintile and over time. Our main data set comes from Chen and Zimmermann (forthcoming) and consists of 82 firm characteristics that are available from 1978 to 2021.

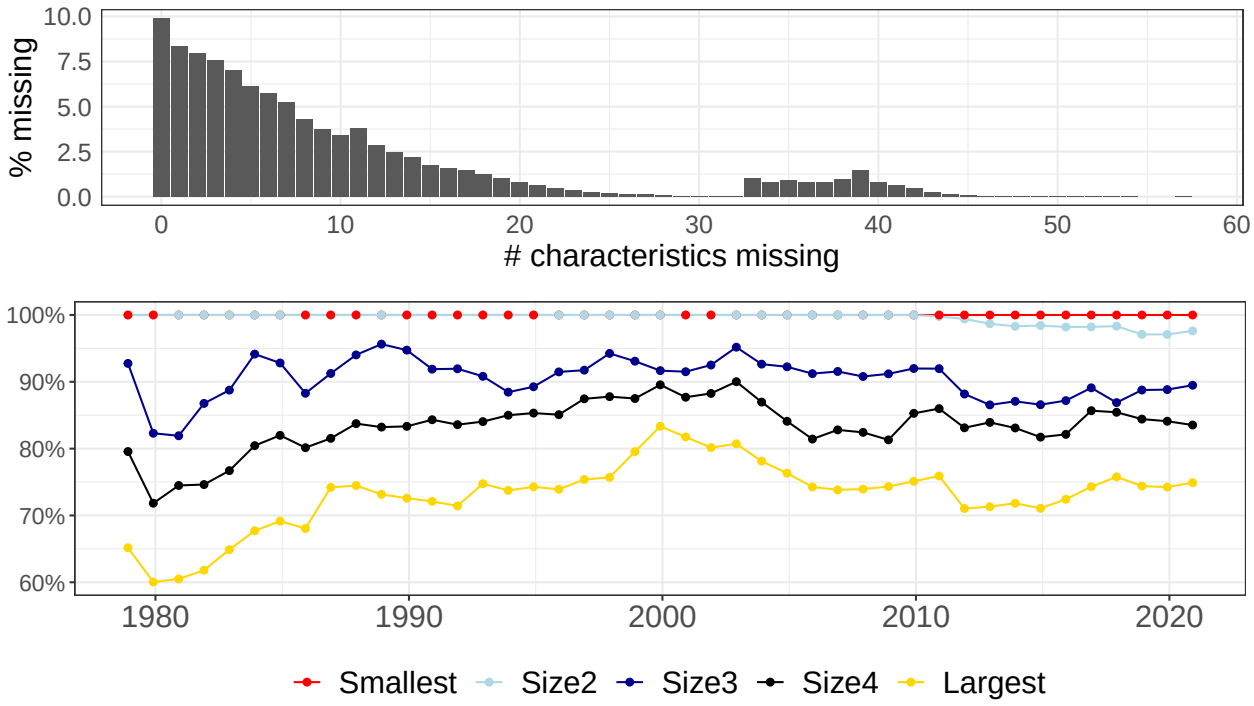
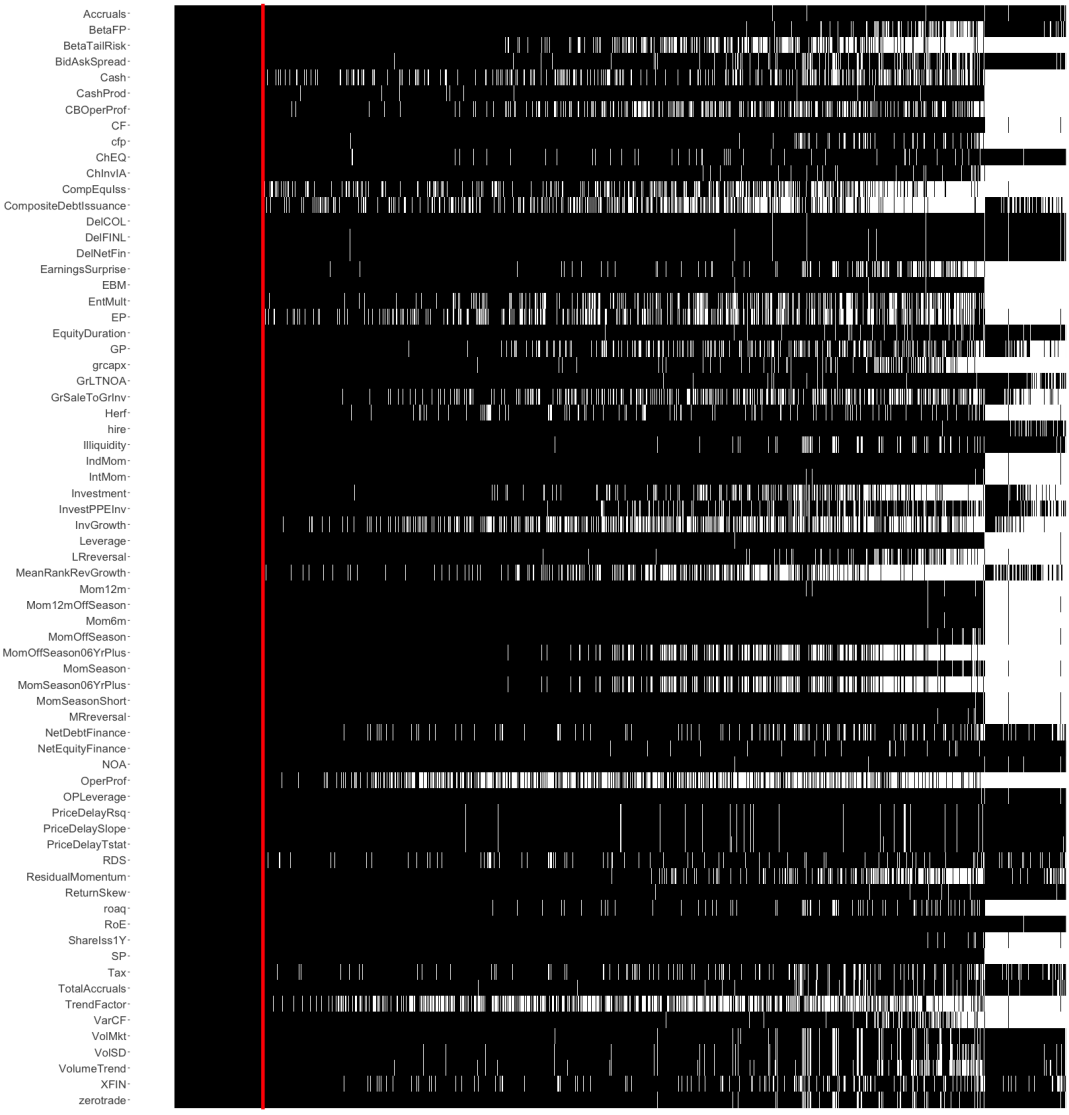


Figure 2 shows the missing (and nonmissing) observations for a random sample of approximately 5% of all observations with black indicating an observed observation and white indicating missing data. The solid black shading to the left of the red vertical line represents the complete case (at roughly 10%). The figure shows no simple pattern for missing characteristics exists such as "white columns," which would indicate that we could simply drop an individual firm-month pair to deal with the missing data problem. Likewise, no "white rows" exist, which would suggest that simply dropping an individual characteristic provides an easy solution of the missing data problem. Instead, missing data are widespread across characteristics, firms, and over time.

Figure 3 additionally illustrates the missing patterns for each of the characteristics over time. In particular, it is perceivable that observations (for each firm) are missing particularly often when the firm enters the sample. In this case it might suffice to simply require that

Figure 2: Complete and incomplete observations for a random subset of firm-month pairs

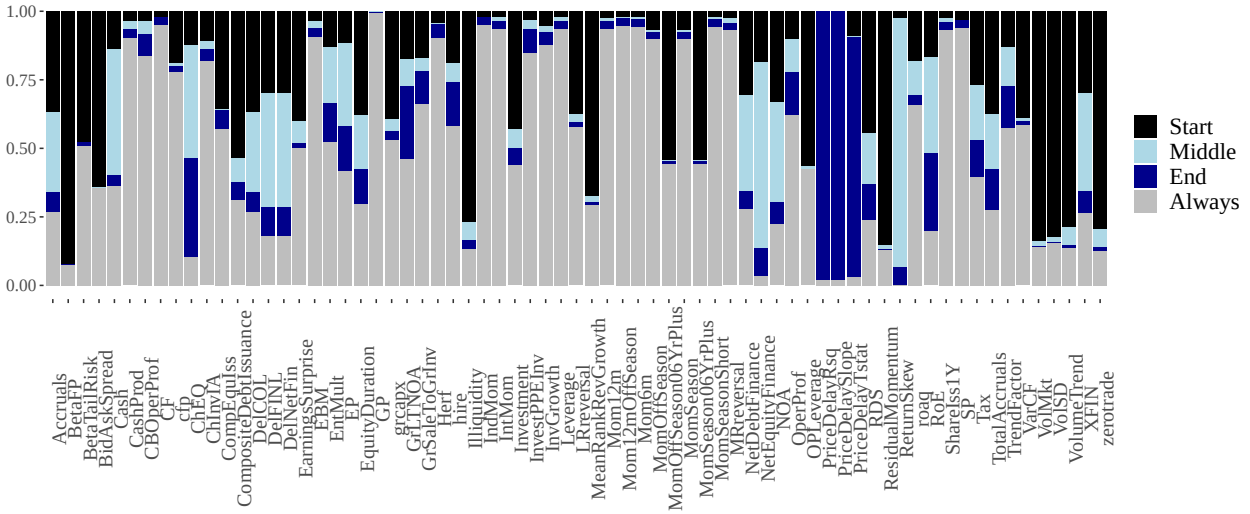
This figure shows a random sample of approximately 5% of the observations. If an item is observed, it is filled in black, whereas it is white if it is not observed. The observations left to the red line depict the complete case, that is, the observation for which no single firm characteristic is missing. We do not include the characteristics which we require to be always observed. Our main data set comes from Chen and Zimmermann (forthcoming) and consists of 82 firm characteristics that are available from 1978 to 2021.



firms have been listed for a while to solve the missing data problem. However, Figure 3 illustrates that this is not the case. In particular, for almost all characteristics we see that they may be missing at the start, that is, when a firm first enters the sample, in the middle, that is, the characteristics was observed when the firm entered the sample and then was no longer observed, but it was observed again later. A firm characteristic could have good availability for most of the sample but be missing toward the end of the sample. Finally, a characteristic might be always be missing for some firms. Figure 3 shows that most characteristics tend to be missing more frequently at the beginning, likely because of data requirements, for example, a certain number of previous observations is required to calculate the characteristics, such as past returns for momentum variables. However, we also see that all patterns are present for all characteristics to some degree. More generally, one could ask why the data are missing in the first place. Typically, this occurs if some item needed for its computation is missing. For accounting items, missingness might be because of a choice on behalf of the firm to not report a certain item, or simply because Compustat did not record it.

Figure 3: Structure of missing characteristics

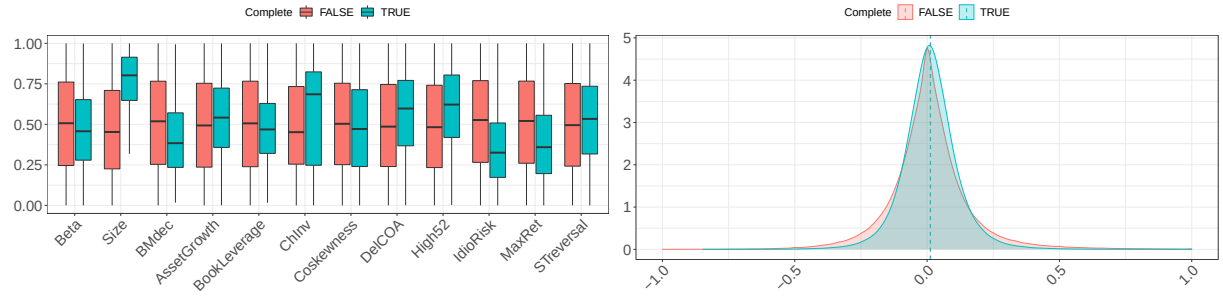
This figure depicts the prevalence of different missingness types. Given a time series of a single firm, a characteristic may be missing at the beginning of the time series but is observed after some point (*Start*), a characteristic may be observed at the beginning of the time series, then it is missing for a few time periods but is observed again (*Middle*), a characteristic may be missing at the end of the time series (*End*), or a characteristic may never be observed (*Always*). For a given characteristic, this figure shows which fraction of the missing observations for this characteristic can be assigned to these missingness types. The statistics are pooled across firms and over time. Our main data set comes from Chen and Zimmermann (forthcoming) and consists of 82 firm characteristics that are available from 1978 to 2021.



In the context of asset pricing, an important additional reason exists why it is undesirable to use the complete case. Firms with missing characteristics may have different properties than firms with no missing characteristics. Figure 4 illustrates this point. The left part of the figure shows a boxplot of important and always observed characteristics contrasting firms with no missing characteristics (green) and those for which at least one other firm characteristics is missing (red). While the distribution of characteristics appears similar for some characteristics, such as `STreversal`, it can be quite different for others, for example, `Size` or `IdioRisk`. The right panel of Figure 4 shows density plots of the returns for firms with no and with missing characteristics. While the mean does not appear drastically different, the incomplete firms have more dispersed return realizations. If a researcher were to focus only on the complete observations, she would ignore an important part of the return distribution. Using these observations may allow to form portfolios with better risk-reward properties, as we will show below.

Figure 4: Complete versus incomplete observations

The left panel shows a boxplot for a subset of characteristics, which we require to be always observed, for the complete (green) versus incomplete observations (red). The right panel shows a density of the returns for the complete (light green) versus incomplete (light red) observations. Our main data set comes from Chen and Zimmermann (forthcoming) and consists of 82 firm characteristics that are available from 1978 to 2021.



2 Model

As many characteristics are missing in the standard CRSP/Compustat panel simply ignoring the problem and using the complete case is inefficient. In the following, we outline our proposed procedure to deal with missing values. The method is flexible enough to use information from both the cross-sectional correlation between characteristics and the temporal relation of a characteristics within a firm to obtain suitable imputations for the missing values.

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, but this simple example contains almost all of the intuition with much simpler notation. 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 are 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. This assumption is also implicitly imposed when using the complete subset of observations only and we can write it as

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

which implies that we could estimate the parameters using the complete observations only.

This approach is inefficient because it neglects a part of the data that contains both Y_i and $X_{i,1}$. To use that part of the sample, we use the second part of the assumption, namely,

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

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 $X_{i,2}$ conditional on $X_{i,1}$.

In the full model, we allow D_i to depend on all variables that we always observe. In particular, in our sample we always observe 13 firm characteristics, including size, book-to-market, beta, idiosyncratic risk, and the return of the previous month. The probability that an observation is incomplete can be a function these characteristics (see Section 1 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 are missing at random. While these assumptions are not directly testable, as we will explain below, we can test the implications of the assumptions

that we use to construct our estimator.

For the incomplete 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 the fact 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 $E\left[\varepsilon_i \mid X_{i,1}, D_i = 1\right] = 0$. Notice that we can estimate $E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right]$ 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

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

and the corresponding unconditional moments

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

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

$$\left.\begin{array}{l} E\left[\mathbf{1}(D_i = 0)\left(X_{i,2} - \gamma_0 - X_{i,1}\gamma_1\right)\right] = 0 \\ E\left[\mathbf{1}(D_i = 0)\left(X_{i,2} - \gamma_0 - X_{i,1}\gamma_1\right)X_{i,1}\right] = 0 \end{array}\right\} \begin{array}{l} \text{3rd set} \\ \gamma \textit{ for imputation model} \end{array} .$$

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 we do so using a modified version of the J-test (see Section A.8.3 in the appendix for a derivation of the test statistic in the general model and Section 2.2 for the test results).

Note we do not require the assumption that $E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right]$ is a linear function to derive our unconditional moment conditions. To avoid it, we can use an alternative derivation based on projections, which is less intuitive and which we discuss in Section A.6 in the appendix.

Based on the moments, different ways exist 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, whereas 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 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 suppose γ is known. The optimal GMM estimator, based

on the first two sets of moments, then turns out to minimize

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

and the denominators of the two fractions can be replaced with consistent estimators. We prove this equivalence and detail the assumption in a more general setting in Appendix A.7. Hence, an alternative way to obtain the estimator is 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 for which $X_{i,2}$ has been imputed. To better understand the reason for down-weighting observations with a missing regressor, define $Z_i = X_{i,2}$ if $D_i = 0$ and $Z_i = E[X_{i,2} \mid X_{i,1}]$ if $D_i = 1$. We can then write our outcome equation as

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

where

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

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 in 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 ordinary least squares (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 suggest that this approach may lead to worse statistical properties than the complete case estimator, even when a substantial subset of the data contains missing values.

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 generally 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 zero unless in the special cases when $\beta_2 = 0$ or when $X_{i,2}$ is mean independent of $X_{i,1}$. Even when one of these conditions is close to being satisfied, unconditional mean imputation is still unreliable for inference, because it ignores the imputation uncertainty.

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, which only contains information known at time $t-1$. 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 note 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 $L+1$ different missing patterns exist, where for each missing pattern we observe a different subset of regressors. 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 $X_{it}^{(0)} = X_{it}$.

As in the simple example, we can allow data to be missing systematically and we impose two conditions. First, we assume

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

for all $l = 0, 1, \ldots, L$, which implies the complete case moment conditions

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

While these moment condition could be used to estimate β_t, we also want to use the incomplete part of the sample. Therefore, for all $l = 1, \ldots, L$, write

$$\begin{aligned} E\left[Y_{it} \mid X_{it}^{(l)}, D_{it} = l\right] &= \sum_{k=1}^{K} E[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l]\beta_{t,k}, \\ &= \sum_{k \in I_t^{(l)}} X_{it,k}\beta_{t,k} + \sum_{k \notin I_t^{(l)}} E[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l]\beta_{t,k}. \end{aligned}$$

Recall, when $D_{it} = l$, $X_{it,k}$ is observed for all $k \in I_t^{(l)}$. Again, we want to replace $E[X_{it,k} \mid X_{it}^{(l)}, D_{it} = l]$ for $k \notin I_t^{(l)}$ by its complete case counterpart that can be estimated. To do so, we impose the second part of our assumption, namely,

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

for all $l = 1, \ldots, L$ in which case

$$E\left[Y_{it} \mid X_{it}^{(l)}, D_{it} = l\right] = \sum_{k \in I_t^{(l)}} X_{it,k}\beta_{t,k} + \sum_{k \notin I_t^{(l)}} E\left[X_{it,k} \mid X_{it}^{(l)}, D_{it} = 0\right] \beta_{t,k}.$$

As discussed above, these assumptions allow D_{it} to depend on regressors that are always observed.[4]

To estimate $E[X_{it,k} \mid X_{it}^{(l)}, D_{it} = 0]$, we again use a linear model. That is, for all $l = 1, \ldots, L$ and $k \notin I_t^{(l)}$, let

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

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. Similar to the simple example, we can then write

$$E\left[Y_{it} \mid X_{it}^{(l)}, D_{it} = l\right] = \sum_{k \in I_t^{(l)}} X_{it,k}\beta_{t,k} + \sum_{k \notin I_t^{(l)}} X_{it}^{(l)\prime}\gamma_t^{(l,k)}\beta_{t,k}$$

and we can interpret $X_{it}^{(l)\prime}\gamma_t^{(l,k)}$ as a replacement for the unobserved covariate $X_{it,k}$, which is based on the observed characteristics and needs to be estimated using the complete cases at time t. Under certain assumptions, we can also use observed covariates from other time periods for imputation, as we discuss in Section 2.3.2.

We can now collect our conditional moments and transform them to unconditional moments to obtain:

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

$$E\left[\mathbf{1}(D_{it} = l)\left(Y_{it} - \sum_{k \in I_t^{(l)}} X_{it,k}\beta_{t,k} - \sum_{k \notin I_t^{(l)}} X_{it}^{(l)\prime}\gamma_t^{(l,k)}\beta_{t,k}\right) X_{it}^{(l)}\right] = 0 \qquad l \geq 1 \tag{2}$$

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

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

These three sets of moment conditions are analogous to the ones 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. These moment conditions are testable using the test statistic described in Appendix A.8.3. In particular, we implement this test pooling data for each quarter and discard missing patterns for which $E[X_{it}X_{it}']$ does not have full rank. These are mostly missing patterns for which we observe less observations than always observed characteristics. We find the test rejects the null hypothesis that the over-identifying restrictions hold in around 1.7% of the time periods with a 5% significance level, providing evidence in favor of the null hypothesis that the moment conditions hold.

Just as in the simple example, different ways to estimate the parameters exists. 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 in which missing values are replaced with the estimated means, 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. Specifically, we implement our estimator using the following steps:

1. Use moment conditions (1) and (3) to estimate β_t and $\gamma_t^{(l,k)}$, respectively, using linear regressions based on the complete case.[5]

2. Estimate $Var(Y_{it} - \sum_{k \in I_t^{(l)}} X_{it,k}\beta_{t,k} - \sum_{k \notin I_t^{(l)}} X_{it}^{(l)\prime}\gamma_t^{(l,k)}\beta_{t,k} \mid D_{it} = l)$ for $l = 1, \ldots, L$ and $Var(Y_{it} - \sum_{k=1}^{K} X_{it,k}\beta_{t,k} \mid D_{it} = 0)$ using the estimated parameters from step 1. Depending on the value of D_{it}, let $\hat{\sigma}_{it}^2$ be the estimated variances for observation i.

[5]Incorporating lagged characteristics in the model to estimate $\gamma_t^{(l,k)}$ is straightforward, and we will discuss this point in Section 2.3.2. We omit it here to keep notation simple.

3. For all i with $D_{it} = l$ and all $l = 0, \ldots, L$ define

$$\hat{Z}_{it,k} = \begin{cases} X_{it,k} & \text{if } k \in I_t^{(l)} \\ X_{it}^{(l)\prime} \hat{\gamma}_t^{(l,k)} & \text{if } k \notin I_t^{(l)}, \end{cases}$$

where $\hat{\gamma}_t^{(l,k)}$ is the estimator from part 1. Now define

$$\hat{\beta}_t = \underset{b \in \mathbb{R}^K}{\arg\min} \sum_{i=1}^{n} \frac{(Y_{it} - \sum_{k=1}^{K} \hat{Z}_{it,k} b_{t,k})^2}{\hat{\sigma}_{it}^2}.$$

We derive the large sample distribution of the estimator in Appendix A.8 and provide plug-in estimators for the standard errors. A potential alternative is the optimal GMM estimator, which can be difficult 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

We now discuss a set of extensions of our baseline models.

2.3.1 High-dimensional and nonlinear models.

We can apply our two-step estimator also in high-dimensional and nonlinear models. Recall 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 large number of regressors, we could also use a penalized estimator, such as the LASSO estimator or the Ridge estimator.

To explain how to construct a consistent estimator in the second step, 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 in which $X_{i,2}$

is missing. We then have

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

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

A potential alternative could be to replace missing values of $X_{it,2}$ and $X_{it,2}^2$ by estimates of $E[X_{i,2} \mid X_{i,1}, D_i = 1]$ and $E[X_{i,2} \mid X_{i,1}, D_i = 1]^2$, respectively. However, since $E[X_{i,2} \mid X_{i,1}, D_i = 1]^2 \neq E[X_{i,2}^2 \mid X_{i,1}, D_i = 1]$, the resultant estimator is inconsistent. These issues carry over to other nonlinear models, which implies that imputations should depend on the model being estimated.

One possibility to allow for nonlinearities and model selection simultaneously, which we use in our application, is the group LASSO estimator of Freyberger, Neuhierl, and Weber (2020). Similar to the simple example above, in the first step we need 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, Neuhierl, and Weber (2020), with the possibility of down-weighting observations with imputed values. This approach not only allows for nonlinearities but also for prespecified 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 lagged values of missing characteristics. We now briefly describe different approaches using our simple example and discuss the details in Appendix A.4.

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 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, that is, returns. We can then change 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 is reasonable for industry dummies and lagged characteristics. 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

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

We can then simply impute $X_{i,2}$ using both $X_{i,1}$ and V_i. All other steps of the estimation procedure are identical to those in Section 2.2 (i.e., we do not need 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 as well as comparisons to the two commonly used methods in practice, namely, mean imputation and using the complete case. We then consider a high-dimensional setting and discuss model selection and out-of-sample predictions. Finally, we compare our method to the factor model of Bryzgalova et al. (2023) and the EM algorithm of Chen and McCoy (forthcoming). We use different DGPs, including a model in which the regressors have a factor structure.

3.1 Low-dimensional setting

We start with the model

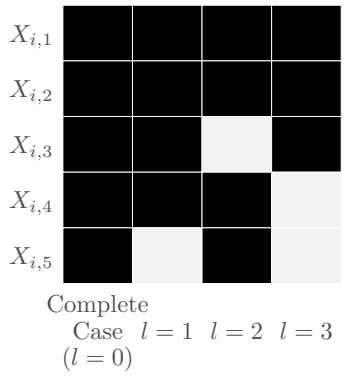
$$Y_i = \sum_{k=1}^{K} X_{i,k}\beta_k + \varepsilon_i, \qquad E[\varepsilon_i \mid X_i] = 0,$$

where $K = 5$ and $X_{i,1} = 1$. We let $X_{i,2}, \ldots, X_{i,K}$ be jointly normally distributed with means of zero 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 $\beta = (1, 0.5, 1, -1, 3)'$.

As a first low-dimensional example, we consider the missing patterns shown in Figure 5. Next to the subset of complete observations ($l = 0$), a subset of the data have missing values for $X_{i,5}$ ($l = 1$), another subset for $X_{i,3}$ ($l = 2$), and another subset for both $X_{i,4}$ and $X_{i,5}$

Figure 5: Missing pattern

This figure shows the missing patterns. $l = 0$ denotes the complete case, that is, the fraction of that data for which all covariates (and the outcome) are observed. In addition, some parts of the data have missing values for the fifth covariate ($X_{i,5}$), another part of the data for the third covariate ($X_{i,3}$), and another part of the data for the fourth and the fifth covariates ($X_{i,4},\ X_{i,5}$).



($l = 3$).

Table 1 shows coverage rates, that is, the fraction of simulations the confidence intervals covers the true value, and average lengths of 90% confidence intervals for different percentages of complete observations. All other missing patters are equally likely. The sample size is $n = 1,000$ and we run $1,000$ Monte Carlo simulations for each design. We report results for the estimator that only uses the complete subset of the data, the optimal GMM estimator, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean. Comparing the complete case and the optimal GMM estimator, we can see that the optimal GMM estimator has substantially smaller confidence intervals across all panels with different degrees of missing. The GLS estimator with conditional mean imputation performs almost as well as the optimal GMM estimator and oftentimes considerably better than the imputation estimator based on OLS (i.e., without downweighting imputed observations). In fact, the average length of the confidence intervals of the OLS estimator can be larger than those of the complete case estimator. Whereas the OLS estimator uses more moment conditions, it combines them in an inefficient way. When the fraction of complete observations is low (top panel), the relative gains from imputation are generally larger and the differences between OLS and GLS are larger.

All of these four estimators are valid and therefore have coverage probabilities close to

Table 1: 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 when 25%, 50%, and 75% of the data are missing at random. The table reports results for the estimator that only uses the complete subset of the data, the optimal GMM estimator, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.

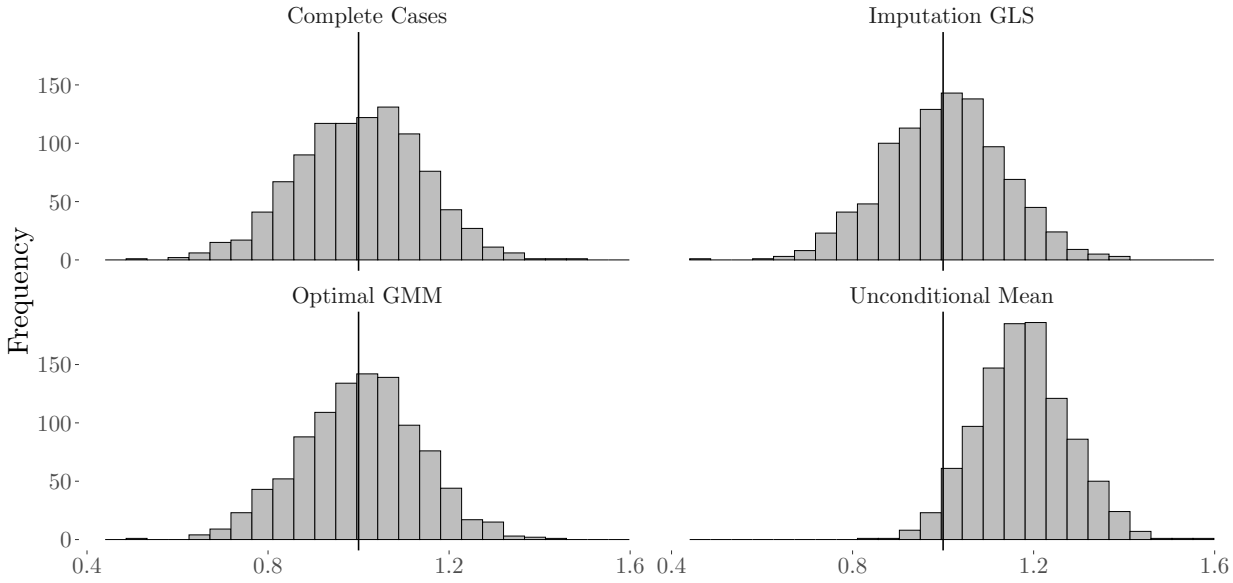
	Complete case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. mean	
	Cover	Length	Cover	Length	Cover	Length	Cover	Length	Cover	Length
25% complete										
β_1	0.912	0.208	0.885	0.141	0.892	0.150	0.893	0.204	0.799	0.172
β_2	0.892	0.477	0.887	0.367	0.893	0.400	0.894	0.539	0.001	0.305
β_3	0.886	0.640	0.873	0.561	0.896	0.607	0.886	0.795	0.003	0.343
β_4	0.888	0.640	0.887	0.498	0.905	0.521	0.896	0.609	0.000	0.415
β_5	0.892	0.476	0.888	0.346	0.899	0.356	0.896	0.361	0.000	0.385
50% complete										
β_1	0.905	0.147	0.892	0.121	0.892	0.123	0.902	0.151	0.864	0.164
β_2	0.909	0.338	0.896	0.294	0.901	0.299	0.909	0.371	0.000	0.309
β_3	0.907	0.454	0.894	0.424	0.902	0.430	0.894	0.518	0.511	0.357
β_4	0.897	0.453	0.891	0.401	0.895	0.406	0.904	0.440	0.000	0.485
β_5	0.903	0.337	0.900	0.294	0.906	0.297	0.905	0.298	0.000	0.399
75% complete										
β_1	0.901	0.120	0.891	0.111	0.895	0.111	0.894	0.124	0.894	0.144
β_2	0.905	0.276	0.904	0.259	0.903	0.261	0.895	0.291	0.065	0.306
β_3	0.914	0.370	0.906	0.359	0.904	0.361	0.898	0.394	0.891	0.404
β_4	0.905	0.370	0.885	0.351	0.901	0.354	0.904	0.365	0.002	0.552
β_5	0.902	0.275	0.904	0.260	0.912	0.262	0.918	0.263	0.000	0.439

90%. To obtain correct coverage probabilities, it is crucial to take the additional variance due to imputations into account. The estimator based on unconditional mean imputation has low coverage rates, which is due to the inherent bias of the estimator (more below). Interestingly, the confidence intervals can be much narrower than those of the optimal GMM estimator (see, e.g., those for β_3 in the top and middle panels). The reason is that the regressors appear less correlated once the unconditional mean is imputed.

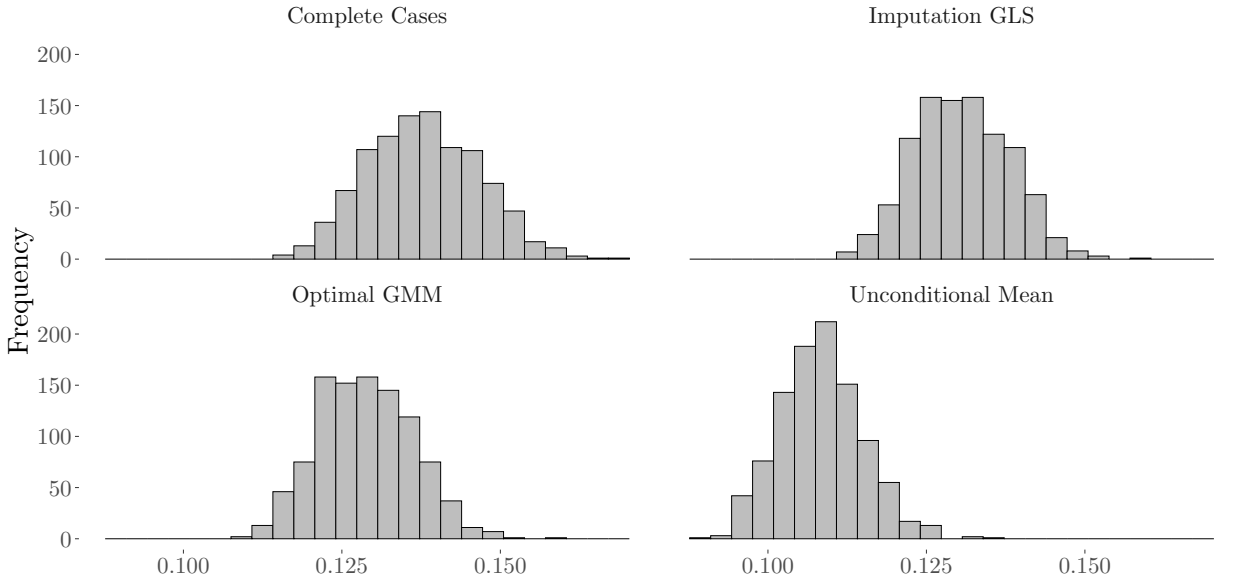
To illustrate these points further, Figure 6 shows histograms of the estimates of β_3 (panel A) and the corresponding standard errors (panel B) when 50% of the sample is complete. The imputation GLS estimator and the optimal GMM estimator perform very similarly and are both more efficient than the estimator based on the complete case. In addition, unconditional

Figure 6: Simulation — histograms

This figure shows histograms of the repeated sample distribution for estimates of β_3 (panel a) and standard errors of $\hat{\beta}_3$ (panel b) when 50% of the observations are complete. The vertical bar represents the correct value for the parameter, $\beta_3 = 1$. The figure reports histograms for the estimator that only uses the complete subset of the data, the optimal GMM estimator, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.



(a) Estimates of β_3



(b) Standard errors of $\hat{\beta}_3$

mean imputation results in both a biased estimator and artificially small standard errors.

Table 2 shows the biases of the estimators when 50% of the sample is complete. With unconditional mean imputation, the biases are substantial, whereas the biases of all other estimators are negligible. We also report the root mean square errors (RMSE) of the different estimators. The optimal GMM estimator can be much more precise than the estimator based on the complete sample. The imputation GLS estimator is as precise as the optimal GMM estimator and generally much more precise than the imputation OLS estimator.

As we discussed in Section 2.1, unconditional mean imputation is based on valid moment conditions in two special cases. First, when all regressors are independent, then the conditional means are equal to the unconditional means. Second, if all regression coefficients in front of regressors that have missing values are equal to zero, imputing any value leads to valid moment conditions. In the latter case, unconditional mean imputation even outperforms the GMM estimator. We illustrate these results in Appendix A.2.1.

3.2 High-dimensional setting

We again simulate data from the linear model

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

Table 2: Simulation of bias and model fit for a general missing pattern

This tables shows the bias in the estimated coefficients and the root mean square error when 50% of the data are complete. The table reports results for the estimator that only uses the complete subset of the data, the optimal GMM estimator, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.

	Complete case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. mean	
	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
β_1	0.044	-0.003	0.038	-0.003	0.038	-0.002	0.046	-0.001	0.055	-0.002
β_2	0.101	0.008	0.091	0.008	0.091	0.008	0.112	0.008	0.473	-0.463
β_3	0.138	-0.002	0.131	-0.004	0.131	-0.004	0.158	-0.006	0.204	-0.177
β_4	0.141	0.003	0.126	0.004	0.125	0.002	0.134	0.004	1.123	-1.114
β_5	0.104	-0.007	0.091	-0.006	0.090	-0.005	0.090	-0.005	1.248	1.243

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 zero 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 zero. For the first five regressors, we use the same missing patterns as above with $L = 3$. When $l = 0$ or $l = 3$, all other regressors are observed as well. When $l = 1$, $X_{i,6}$ and $X_{i,7}$ are not observed and when $l = 2$, $X_{i,36}, X_{i,37}, \ldots, X_{i,40}$ are not observed. The probability that an observation is missing now varies with $X_{i,2}$ and $X_{i,35}$, which are always observed. In particular, observations with high values of $X_{i,2} + X_{i,35}$ are more likely to be complete. Again, 50% of the observations are complete.

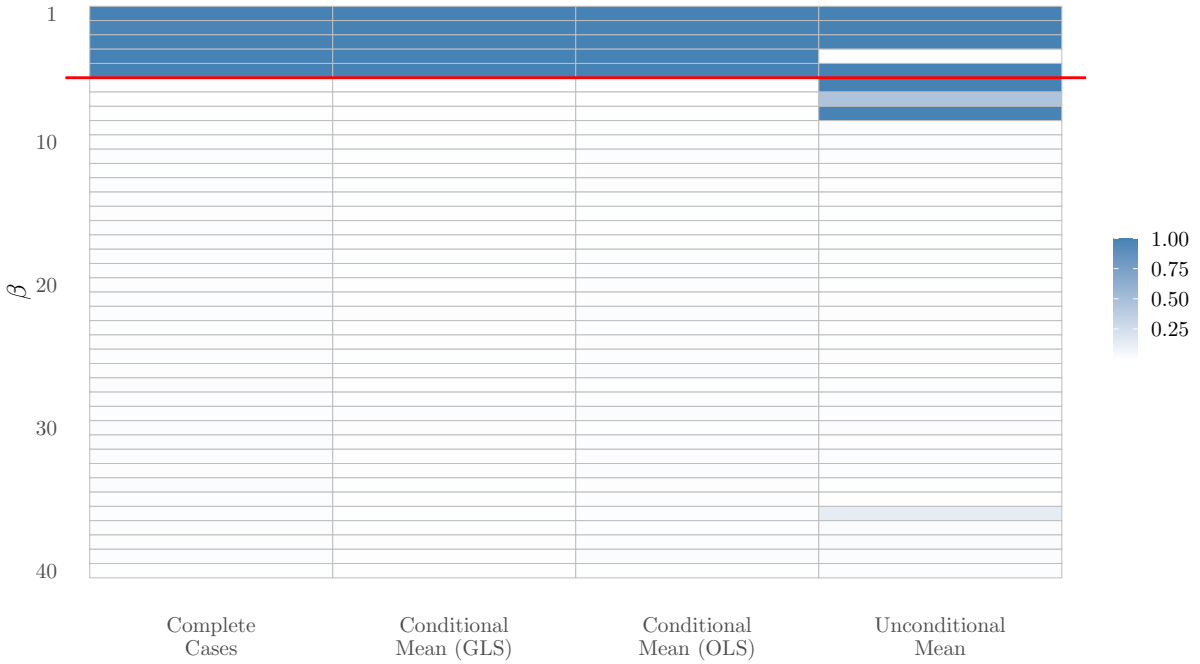
We now consider four different estimators, namely, the estimator that only uses the complete subset of the observations, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean. For all estimators, we estimate the parameters $\beta_1, \beta_2, \ldots, \beta_{40}$ using the post adaptive LASSO method and choose the penalty parameter based on the BIC following Freyberger, Neuhierl, and Weber (2020). We use the same LASSO procedure for the imputation step. All estimators are easy to implement using standard software.

Figure 7 illustrates the frequency with which the different methods select regressors. The darker the color, the more frequent a particular model estimates a nonzero β_k. In the true model, the first five betas are nonzeros (above the red line), whereas the remaining ones are equal to zero. The complete case estimator and both conditional mean imputation estimators select the variables with nonzero coefficients with high probability and typically set coefficients of irrelevant predictors to zero. Unconditional mean imputation tends to set the estimated value of β_4 to zero and instead frequently includes three of the irrelevant regressors. The mean squared prediction errors (MSPEs) of the four methods are 1.0158, 1.009, 1.0121, and 1.6512, respectively, showing the imputation GLS estimator performs best and unconditional mean imputation performs worst.

In Appendix Section A.2.2, we also consider a nonsparse setting. In this setting, our proposed estimator outperforms both the estimator based on unconditional mean imputation and the complete case estimator. In addition, we again show that in the special case when all regressors with missing values do not affect the outcome, that is, are not relevant for return

Figure 7: Model selection — Sparse model

This figure shows the frequency with which the different methods select regressors for the simulation setup with a sparse model. The darker the color, the more frequent a particular model estimates a nonzero β_k. In the true model, the first five betas are nonzeros (above the red line), whereas the rest is equal to zero. The figure shows results for the estimator that only uses the complete subset of the observations, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.



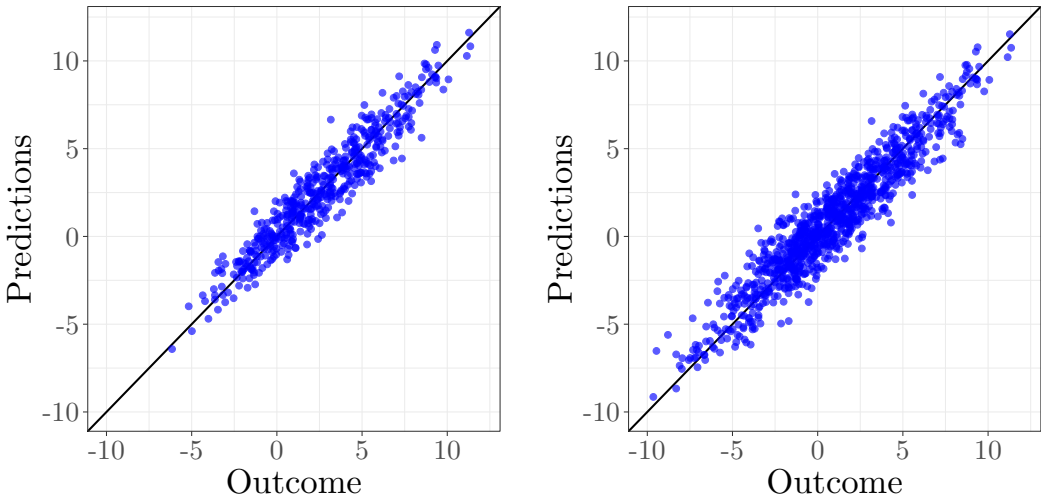
prediction, unconditional mean imputation also consistently selects the relevant regressors.

One advantage of imputations is that we can use the whole sample for predictions. The MSPE for the subset of noncomplete observations is typically higher than for the complete observations, but the complete case might miss particularly interesting parts of the conditional distribution of outcomes, in our case returns. To illustrate this point, Figure 8 plots the out-of-sample realized returns against the predicted returns obtained with the different methods.[6] Recall the probability that an observation is completely observed depends on $X_{i,2} + X_{i,35}$. When using imputations, we make predictions for all outcomes, even when some regressors are missing. Comparing panels A and B, we can see the observations with missing regressors tend to have lower predicted and realized returns.

[6]For the out-of-sample predictions, we generate a new sample of complete observations with a sample size of $1,000$.

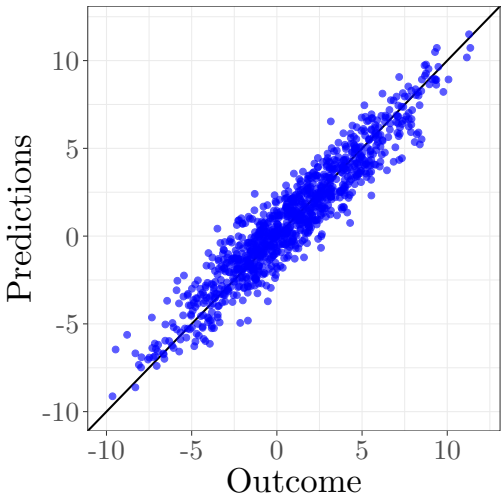
Figure 8: Outcomes versus predictions

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

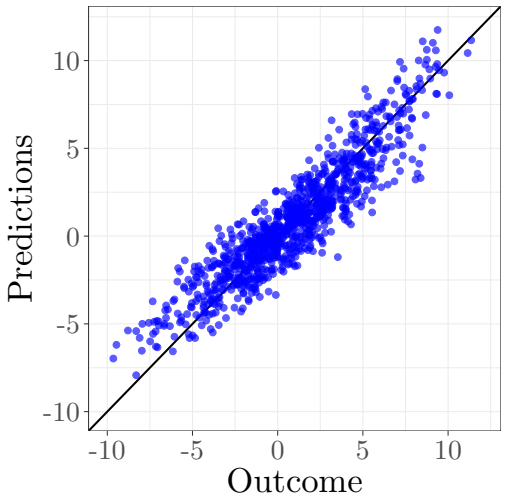


(a) Complete case

(b) GLS



(c) OLS



(d) Unconditional mean

Two important implications for out-of-sample portfolio sorts arise that we will discuss in more detail in our application. First, when using imputations, we have a larger number of observations to form portfolios. 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 larger difference in portfolio returns between the complete case and our method. Second, when the probability that an observation is missing depends on the observed covariates, the complete case misses a systematic part of the distribution of returns and not just a random sample. In this case, differences in portfolio returns will be even more pronounced. Finally, panel D shows imputing unconditional means results in a mild U-shaped pattern between predicted and realized returns, that is, it yields biased predictions. However, since predictions and outcomes are still strongly positively related, portfolios formed based on these predictions will be very similar to those obtained with conditional mean imputation, because the cross-sectional ranking of stocks remains largely intact.

3.3 Comparison to EM algorithm and factor models

The EM algorithm of Chen and McCoy (forthcoming) assumes the regressors are jointly normally distributed and the algorithm essentially imputes missing values based on projections.[7] Just like our approach, these projections require a covariance matrix of the regressors, which we calculate using the complete case. Their EM algorithm on the other hand estimates each entry with all available data (i.e., using all observations for which a given pair of regressors, rather than all regressors, is observed). If the data are missing completely at random, this approach yields efficiency gains, but the estimators may be biased if missingness depends on an always observed regressor.[8]

The factor model of Bryzgalova et al. (2023) instead requires the regressors to have a factor structure. In addition, for the estimator to be consistent, the number of characteristics

[7] We build on their code when implementing the EM algorithm.

[8] We refer to example 8.3 together with example 7.1 in Little and Rubin (2020) for an illustrative argument for why the bias arises.

has to diverge with the sample size. For both of these methods, once missing values are imputed, we estimate the regression coefficients by OLS. It is important to note that the usual OLS standard errors are not valid due to the additional uncertainty that the imputation introduces and it is not clear how to perform inference using these methods. Our comparison therefore focuses on mean squared errors.

We consider different DGPs. We start with the low-dimensional setup of Section 3.1 in which all regressors are normally distributed. We then adapt the setup in two ways by changing the marginal distribution of the regressors. First, we transform them to uniformly distributed random variables by using $\Phi(X_{i,k})$ instead of $X_{i,k}$ as regressors, where Φ denotes the standard normal distribution function. In this case, we take $Var(\varepsilon_i) = 0.25$. Second, we generate the regressors using a factor model, which we describe in detail in Appendix A.2.3. For the estimator of Bryzgalova et al. (2023), we use the true number of factors, whenever the regressors are simulated from a factor model and estimate the number of factors using the cross validation method of Jin, Miao, and Su (2021) if they are not. In addition, we also simulate data using the high-dimensional setup of Section 3.2. Again, as alternatives, we use uniformly distributed regressors and regressors based on a factor model.

Table 3 shows the mean squared errors for the three different estimators and six different DGPs. For each setup and estimator, the table reports the average mean squared error of the estimated coefficients, the average mean squared imputation error of the regressors, and the average mean squared out-of-sample errors for returns. In the low-dimensional case, our GLS estimator and the EM estimator yield very similar imputation errors. However, because of our weighting, the GLS estimator yields more precise estimators of the coefficients. In these low-dimensional cases, the factor model based estimator performs significantly worse, even when the regressors have a factor structure (and a known number of factors).

In the high-dimensional setup, the estimator of Bryzgalova et al. (2023) has the lowest mean squared imputation error when the regressors have a factor structure. However, the lower imputation error does not translate into a lower MSE for the estimated coefficients because imputed and nonimputed observations receive the same weight. When the regressors do not have factor structure, our estimator outperforms their estimator using any of the three criteria. Similarly, the estimator of Chen and McCoy (forthcoming) may result in

Table 3: Simulation of average MSEs of different methods with M(C)AR

This table shows mean squared errors of our GLS estimator, the factor model of Bryzgalova et al. (2023), and the EM algorithm of Chen and McCoy (forthcoming) with different data generating processes. The low- and high-dimensional setups have 5 and 40 regressors, respectively. The marginal distribution of the regressors is normal, uniform, or obtained from a factor model. For each setup and estimator, the table reports the average mean squared error of the estimated coefficients, the average mean squared imputation error of the regressors, and the average mean squared out-of-sample errors for returns.

	Normal DGP			Uniform DGP			Factor DGP		
	GLS	EM	Factor	GLS	EM	Factor	GLS	EM	Factor
Low-dimensional and MCAR									
Coefficients	0.0101	0.0127	0.1217	0.0258	0.0278	0.1378	0.0029	0.0057	0.0441
Regressors	0.0400	0.0400	0.0683	0.0003	0.0003	0.0006	0.3629	0.3619	0.6316
Outcomes	1.6904	1.6930	1.8303	0.3127	0.3128	0.3256	2.7845	2.7917	3.2554
High-dimensional and MAR									
Coefficients	0.0139	0.0190	0.0444	0.0332	0.0369	0.0974	0.0049	0.0136	0.0066
Regressors	0.1473	0.1438	0.3524	0.0012	0.0012	0.0060	0.1566	0.1689	0.1371
Outcomes	1.5109	1.5470	1.6900	0.3035	0.3058	0.3340	2.4076	2.6303	2.3465

lower mean squared imputation errors than our estimator, but our estimator yields more accurate estimated coefficients and predictions of the outcome in all settings.

In Appendix A.2.3, we also discuss results for setups in which regressors are not missing at random and the assumptions are not satisfied to illustrate the limits of the different methods.

4 Empirical Application

In this section we illustrate the empirical relevance of different choices for treating missing data in several applications: out-of-sample return prediction and determining which characteristics provide incremental information. We begin by comparing the different imputation methods with respect to their imputation accuracy in a "masking" exercise.

4.1 Masking the complete case

In general, we can never know how accurate any imputation method is because it requires knowledge of the missing data. We can however aim to get an estimate of the accuracy by assuming that some characteristics that actually were observed are missing, that is, we mask them randomly and then compare the imputed value to the actually observed ones. This exercise allows us to compare the quality of the imputations generated by different imputation methods. Note that high imputation accuracy will not necessarily lead to better predictions as the imputations could be particularly accurate for irrelevant characteristics.

To be more precise, we focus on the 238,198 complete firm-month observations in the data set and randomly delete 1% of the entries of the characteristic data matrix of these complete case observations. With 82 characteristics, the probability of an observation to be complete is then $0.99^{82} \approx 44\%$. Having introduced the missing values, we delete all firm-month observations for which we do not observe the characteristics that we also require to be observed in the entire data set. These characteristics are `AssetGrowth`, `Beta`, `BMdec`, `BookLeverage`, `ChInv`, `Coskewness`, `DelCOA`, `DelLTI`, `High52`, `IdioRisk`, `MaxRet`, `Size`, and `STreversal`. We arrive at a data set of 209,006 observations of which 104,475 observations are complete. In a next step, we impute the now missing characteristic values in this data set using unconditional mean imputation and conditional mean imputation with and without lags, using both a linear model and an additive nonlinear model. To estimate the nonlinear model, we use orthonormal Legendre polynomials up to degree 3. To prevent the number of missing patterns from exploding, the imputation model with lags only includes the lagged value of the missing characteristic we aim to impute if this lagged value is observed. In each period, we estimate the conditional mean imputation models on the complete case of the current and the 59 previous periods. The first 60 periods are jointly imputed. In a final step, we compare the imputed values to the originally observed values that we deleted in the first step. For the nonlinear model, we only consider the RMSPE of the first-degree polynomial values.

Table 4 presents the RMSPE for all characteristics with missing values in the masked data set and the average RMSPE across all characteristics. First, independent from the

conditional mean imputation scheme, the RMSPE error is smaller for conditional mean imputation compared to unconditional mean imputation for almost all characteristics.[9] Hence, if unconditional mean imputation yields consistent estimators in a setup that uses this data set, it is most likely because the coefficients of the missing characteristics are zero.

Table 4: Masking the complete case: Out-of-sample prediction error (RMSPE)

This table shows the root mean square prediction errors (RMSPE) for different imputation methods in the masking exercise across all characteristics. We do not include the characteristics that we require to be always observed. The final row contains the root of the weighted average MSPE across all characteristics, where the weight for a characteristic equals the number of missing values for this characteristic divided by the number of all missing characteristic values. For simplicity, we denote the row with "Average RMSPE." We first randomly delete 1% of entries in the data set of the complete case. Then, we impute the missing characteristic values using unconditional mean imputation and conditional mean imputation with and without lags. In the conditional mean imputation setup we consider a linear and an additive nonlinear model, where we use orthonormal Legendre polynomials of up to degree 3 when estimating the model. In a final step, we calculate the RMSPE by comparing the imputed characteristic values with the initially deleted values.

	Uncond. mean	Cond. mean	Cond. mean (w. lags)	Cond. mean (nonlinear)	Cond. mean (nonlinear / w. lags)
Accruals	0.28803	0.12279	0.12227	0.11286	0.11249
BetaFP	0.29233	0.16759	0.06509	0.15991	0.06470
BetaTailRisk	0.28832	0.20749	0.06241	0.20078	0.06118
BidAskSpread	0.29054	0.21604	0.20041	0.21023	0.19803
Cash	0.29499	0.21910	0.14812	0.18882	0.14374
CashProd	0.29004	0.13658	0.07142	0.12316	0.06934
CBOperProf	0.28711	0.14017	0.13532	0.13267	0.12825
CF	0.28776	0.08818	0.06398	0.08302	0.06071
cfp	0.29212	0.15978	0.09129	0.15219	0.08945
ChEQ	0.28591	0.14558	0.14517	0.13578	0.13520
ChInvIA	0.28433	0.21524	0.12467	0.20986	0.12258
CompEquIss	0.28513	0.19197	0.10561	0.18063	0.10459
CompositeDebtIssuance	0.28400	0.22466	0.19860	0.21643	0.19422
DelCOL	0.28301	0.15160	0.15080	0.13927	0.13856
DelFINL	0.29226	0.10647	0.10644	0.10020	0.09955
DelNetFin	0.29453	0.12020	0.12022	0.10943	0.10954
EarningsSurprise	0.28514	0.24015	0.18235	0.23313	0.18119
EBM	0.29058	0.18650	0.12231	0.18309	0.12253
EntMult	0.28848	0.10978	0.06946	0.09546	0.06460
EP	0.28528	0.12949	0.09007	0.11946	0.08585
EquityDuration	0.28425	0.13542	0.13370	0.13044	0.12880
GP	0.28540	0.18731	0.11139	0.17638	0.10491
grcapx	0.28822	0.14168	0.14147	0.13994	0.13989
GrLTNOA	0.29293	0.18609	0.18606	0.16969	0.16961
GrSaleToGrInv	0.28237	0.19295	0.18952	0.18517	0.18164
Herf	0.29012	0.27071	0.05290	0.26605	0.05435
hire	0.29036	0.21030	0.21014	0.20814	0.20841
Illiquidity	0.28407	0.05307	0.02407	0.04394	0.02235
IndMom	0.28709	0.27236	0.20148	0.27218	0.20232
IntMom	0.28780	0.18864	0.15804	0.18346	0.15508
Investment	0.29057	0.14728	0.09644	0.14182	0.09498
InvestPPEInv	0.29180	0.14191	0.13997	0.13580	0.13380
InvGrowth	0.28632	0.10540	0.08419	0.09931	0.08053

[9]The RMSPE for unconditional mean imputation is around $1/\sqrt{12}$ for each characteristic, which is expected since we rank transform the characteristics to be uniform on $[0, 1]$. The standard deviation of a random variable $X \sim \text{Unif}[0, 1]$ is $1/\sqrt{12}$.

Table 4: Masking the complete case: Out-of-sample prediction error (RMSPE) *(continued)*

Leverage	0.29126	0.06440	0.03644	0.05742	0.03406
LRreversal	0.29098	0.19070	0.11559	0.18997	0.11707
MeanRankRevGrowth	0.28559	0.20807	0.07500	0.20080	0.07573
Mom12m	0.29444	0.08989	0.08492	0.08844	0.08333
Mom12mOffSeason	0.28892	0.11125	0.10850	0.11053	0.10813
Mom6m	0.28711	0.15750	0.15392	0.15125	0.14847
MomOffSeason	0.28607	0.16483	0.10452	0.15682	0.10271
MomOffSeason06YrPlus	0.28912	0.24708	0.13038	0.24182	0.12920
MomSeason	0.28727	0.25979	0.25867	0.25475	0.25389
MomSeason06YrPlus	0.28898	0.28522	0.28529	0.28571	0.28593
MomSeasonShort	0.28705	0.23101	0.23088	0.23088	0.23067
MRreversal	0.28912	0.25159	0.18962	0.25002	0.19006
NetDebtFinance	0.28281	0.15732	0.15707	0.14887	0.14878
NetEquityFinance	0.28759	0.17551	0.17253	0.16733	0.16546
NOA	0.28768	0.18305	0.15326	0.16213	0.13961
OperProf	0.28955	0.11438	0.10833	0.10347	0.09886
OPLeverage	0.29031	0.11719	0.08404	0.10005	0.07456
PriceDelayRsq	0.29042	0.23187	0.23107	0.20244	0.20221
PriceDelaySlope	0.29261	0.26410	0.26416	0.24631	0.24627
PriceDelayTstat	0.29069	0.27714	0.27249	0.22132	0.21701
RDS	0.29069	0.24331	0.24392	0.23840	0.23998
ResidualMomentum	0.28768	0.17208	0.12598	0.17092	0.12537
ReturnSkew	0.28496	0.20472	0.20449	0.19902	0.19902
roaq	0.28705	0.17799	0.17423	0.17307	0.17131
RoE	0.28562	0.10476	0.10421	0.09534	0.09449
ShareIss1Y	0.28509	0.20170	0.11152	0.19831	0.11219
SP	0.28578	0.07113	0.03519	0.06351	0.03296
Tax	0.28728	0.25166	0.23986	0.24321	0.23346
TotalAccruals	0.29088	0.15869	0.15882	0.14800	0.14822
TrendFactor	0.28447	0.26480	0.24964	0.26619	0.25168
VarCF	0.28942	0.17548	0.03975	0.16160	0.03923
VolMkt	0.28908	0.08613	0.04768	0.07810	0.04573
VolSD	0.28925	0.11409	0.03384	0.10491	0.03236
VolumeTrend	0.28863	0.23015	0.07032	0.22318	0.07015
XFIN	0.29061	0.12185	0.12126	0.10939	0.10908
zerotrade	0.28564	0.09636	0.06304	0.09208	0.06254
Average RMSPE	**0.28828**	**0.18286**	**0.15006**	**0.17471**	**0.14425**

Second, including lagged values in the imputation scheme improves the quality of the imputations more than using a nonlinear imputation model, which is why our discussion below primarily focuses on the linear model with and without lags. Especially for characteristics that are correlated over time using lagged values can improve imputations drastically. For instance, for `MeanRankRevGrowth`, going from a linear conditional mean imputation model to a linear conditional mean imputation model with lags reduces the RMSPE from 0.20807 to 0.075, with similar improvements for `VolumeTrend` and `CompEquIss`. For other characteristics, for example, `OperProf` or `TotalAccruals`, little or no improvement occurs from using the time series information relative to the purely cross-sectional model. In Table A.5

in Appendix A.3.2, we present the results of the masking exercise when we include industry dummies as covariates in the linear imputation models. We outline the exact estimation procedure in Appendix A.3.4. With industry dummies, we find a slight improvement of the quality of the imputations in terms of RMSPE.

4.2 Out-of sample predictions

We first illustrate the different ways of treating missing regressors in a classic empirical asset pricing application, namely cross-sectional out-of-sample return predictions. We report results for four different methods, namely, (1) estimate the prediction model only on the completely observed data, (2) estimate the model with OLS on the data for which we imputed the unconditional mean, (3) estimate the model with the GLS weighting scheme on the data for which we imputed the conditional mean using cross-sectional characteristics, and (4) estimate the model with the GLS weighting scheme on the data for which we imputed the conditional mean using cross-sectional and lagged characteristics, where the imputation model only includes the lagged value of the missing characteristic we aim to impute if this lagged value is observed. We consider both linear models and nonlinear models, as presented in Section 2.3.1, using orthonormal Legendre polynomials up to degree 3. In addition, we estimate regularized models based on the adaptive LASSO for the linear models and the adaptive group LASSO, similar to Huang, Horowitz, and Wei (2010) and Freyberger, Neuhierl, and Weber (2020), for the nonlinear models. In each period, the conditional mean imputation models are estimated on the complete case of the current and the 59 previous periods. The first 60 periods are jointly imputed.

Throughout, we make rolling out-of-sample predictions for the next month using an estimation window of 60 months. We then sort stocks into portfolios based on the predicted return. We consider two portfolio implementations. Our first approach follows the standard "10-1" portfolio, in which we go long the stocks with highest 10% predicted returns and short the stocks with the 10% lowest predicted returns. While this portfolio construction is standard in the literature, we need to be careful in our context as the total number of stocks differs in the complete case and the cases in which we impute data. To address this concern, we also form a long-short portfolio with a fixed number of stocks, that is, we buy (sell)

the 100 stocks with highest (lowest) predicted returns. We record the return for the out-of-sample month, move forward the estimation window and repeat the portfolio formation exercise until the end of the sample period. Our out-of-sample period is 1990 through 2021.

We summarize the results in Table 5. In the linear setups, both the imputation model and the main model are linear, whereas in the nonlinear setup, both models are additive nonlinear. Panel A shows the results for the linear model using all characteristics. Each month, we estimate a linear model over 60 months and then make one-month-ahead predictions and sort portfolios on the predicted returns. For the portfolio with 100 stocks on the long and short side, the complete case results in much lower average returns than either of the imputation methods. However, the complete case portfolio also has much lower standard deviation (9.57% annualized vs. approximately 29% for the other methods). Both findings are consistent with the intuition that we presented in Figure 4. However, when we look at the Sharpe ratios, we can see that the additional risk is more than compensated by the higher average returns. All the imputation portfolios have higher Sharpe ratios than the complete case portfolio. When we compare the risk-return properties among the portfolios with imputations, the conditional mean method leads to slightly higher returns at about the same level of risk relative to the unconditional mean imputation. Note also that at least for the purpose of making out-of-sample predictions, adding lags in the imputation step does not seem to make a material difference.

In the case of the classic "10-1" portfolio, we also find that the complete case method produces portfolios with lower average returns and lower standard deviations, but again the Sharpe ratios for the portfolios with imputation are higher. The returns are highest for the conditional mean GLS methods, but the difference relative to the unconditional mean is relatively low. Note that for the portfolio formation, the ranking of the predicted returns is all that matters, not the actually predicted value. Therefore, even if the unconditional mean might lead to biased estimators, we might still get rather similar portfolios as long as the two methods produce a similar ranking of their predicted values, which we confirmed in our simulation exercise.

In panel B, we conduct the analogous exercise, but instead of using all characteristics, we use a regularized linear model, that is, we first carry out a model selection step over the

Table 5: 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. Portfolios are equally weighted. We differentiate between the complete case method, unconditional mean imputation, and conditional mean imputation with GLS weighting without and with lags. To prevent the number of missing patterns to explode, the imputation model with lags only includes the lagged value of the missing characteristic we aim to impute if the lagged value is observed. Long Pf. and Short Pf. denote the annualized average return of the long and short legs, respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and polynomial model is outlined in Section 4.2. The sample period is 1990–2021.

	Mean (%)	Standard deviation (%)	Sharpe ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
A. Linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete case	11.37	9.57	1.19	19.12	7.75	0.22	3.39
Uncond. mean	48.91	29.46	1.66	39.87	-9.05	-0.12	10.38
Cond. mean (GLS)	52.07	29.15	1.79	41.14	-10.93	-0.41	5.04
Cond. mean (GLS / w. lags)	51.89	28.92	1.79	40.67	-11.22	-0.46	5.36
Long (short) 10% highest (lowest) predicted returns							
Complete case	15.74	13.12	1.20	20.70	4.96	0.35	6.07
Uncond. mean	32.11	19.44	1.65	30.94	-1.17	0.54	11.63
Cond. mean (GLS)	33.13	19.77	1.68	31.20	-1.93	-0.13	6.04
Cond. mean (GLS / w. lags)	33.55	19.71	1.70	31.31	-2.24	-0.19	6.16
B. Regularized linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete case (LASSO)	12.47	10.63	1.17	18.96	6.49	-0.45	3.96
Uncond. mean (LASSO)	47.49	28.22	1.68	39.29	-8.20	-0.01	7.04
Cond. mean (GLS / LASSO)	50.37	28.07	1.79	39.48	-10.89	-0.37	3.70
Cond. mean (GLS / LASSO / w. lags)	50.84	27.60	1.84	39.43	-11.42	-0.32	3.33
Long (short) 10% highest (lowest) predicted returns							
Complete case (LASSO)	17.21	14.74	1.17	21.39	4.18	0.32	4.99
Uncond. mean (LASSO)	31.12	20.25	1.54	30.47	-0.66	0.24	11.33
Cond. mean (GLS / LASSO)	31.24	20.31	1.54	30.37	-0.87	-0.45	5.51
Cond. mean (GLS / LASSO / w. lags)	31.49	20.19	1.56	30.44	-1.05	-0.43	5.27
C. Nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete case	11.06	8.57	1.29	18.45	7.39	0.10	1.71
Uncond. mean	85.53	35.05	2.44	65.46	-20.07	1.14	9.07
Cond. mean (GLS)	92.35	32.71	2.82	67.30	-25.05	0.23	1.91
Cond. mean (GLS / w. lags)	92.20	32.33	2.85	66.90	-25.31	0.22	1.68
Long (short) 10% highest (lowest) predicted returns							
Complete case	17.47	13.15	1.33	22.41	4.94	0.94	6.19
Uncond. mean	42.44	18.73	2.27	37.67	-4.77	0.60	7.86
Cond. mean (GLS)	46.17	19.94	2.32	39.84	-6.34	-0.01	5.02
Cond. mean (GLS / w. lags)	46.22	19.93	2.32	39.72	-6.50	0.20	5.48
D. Regularized nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete case (LASSO)	9.31	10.86	0.86	17.50	8.19	0.19	3.51
Uncond. mean (LASSO)	74.29	34.27	2.17	61.97	-12.32	0.99	7.43
Cond. mean (GLS / LASSO)	84.83	35.62	2.38	63.66	-21.17	-0.40	3.17
Cond. mean (GLS / LASSO / w. lags)	85.78	35.42	2.42	65.57	-20.21	-0.48	3.85
Long (short) 10% highest (lowest) predicted returns							
Complete case (LASSO)	14.16	16.14	0.88	19.33	5.18	0.25	4.63
Uncond. mean (LASSO)	38.98	19.29	2.02	36.12	-2.86	0.46	6.65
Cond. mean (GLS / LASSO)	42.50	21.90	1.94	37.88	-4.62	-0.54	5.65
Cond. mean (GLS / LASSO / w. lags)	43.25	21.32	2.03	38.66	-4.59	-0.39	6.15

period from 1978 through 1989. We apply a (weighted version of) the adaptive LASSO to select the most important characteristics over the first part of the sample. Specifically, for conditional mean imputation with and without lags, we weight each observation with the square root of the estimated error variance $\hat{\sigma}_{it}^2$ presented in Section 2.2 to then perform an adaptive LASSO procedure. For the other methods, we directly use the standard adaptive LASSO procedure. In all setups, the initial estimator for β_t is the complete case estimator and the model is selected using the BIC. After model selection, we proceed exactly as in the linear model presented in panel A and make rolling one-month predictions using an estimation window of 60 months. Figure 9 shows the selected characteristics for each case.

Similar to the standard linear model in panel A, we find in panel B of Table 5 the complete case method leads to the lowest average returns compared to all imputation methods. The conditional mean imputation again leads to slightly better predictions than the unconditional mean, with a larger difference for the "100 long / 100 short" portfolio. The average returns of the regularized linear model are relatively similar to the results in panel A. This finding is not too surprising, because all the predictors we consider were successful return predictors during at least parts of our sample period.

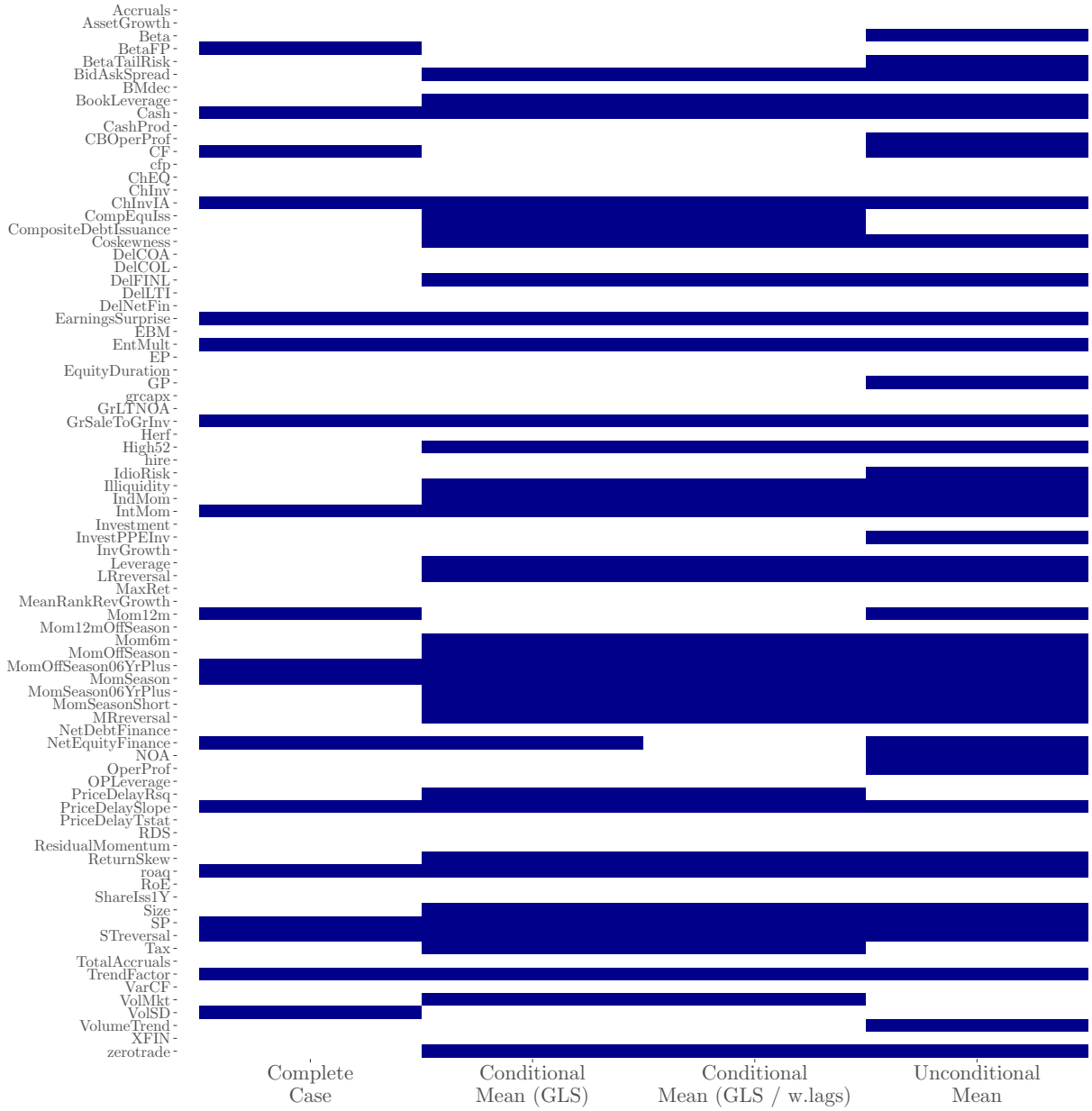
Panels C and D illustrate the results for a nonlinear model, that is, an additive model as outlined in Section 2.3.1. Specifically, we use orthonormal Legendre polynomials up to degree 3. In panel C, we present results for the additive model using all 82 characteristics. As in the case of the linear model, using only the complete cases results in very low returns relative to the conditional and unconditional mean imputation. Both panels C and D show that modeling returns as a nonlinear function of characteristics yields much higher out-of-sample returns compared to the linear models. Notably, the difference between the linear model and nonlinear models is more pronounced for the portfolio that is long (short) in 100 stocks as most of the nonlinearities in the predictive relationship occurs in the extremes of the characteristic distributions. Interestingly, the difference in average returns between the portfolios formed using the conditional mean as compared to the unconditional mean widens in the case of the nonlinear model.

For the results in panel D, we first carry out a model selection step over the period from 1978 through 1989. We apply a weighted version of the adaptive group LASSO discussed in

Freyberger, Neuhierl, and Weber (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

Figure 9: Selected characteristics with the adaptive LASSO procedure (linear model)

This figure shows the selected characteristics using an adaptive LASSO procedure. We perform model selection on the first part of the sample from 1978 to 1989. For conditional mean imputation with and without lags we first weight the observations using the estimated standard deviation of the error terms $\hat{\sigma}_{it}$ as presented in Section 2.2. In a next step, we use an adaptive LASSO procedure where the model is selected using the BIC. The initial estimator for β_t is the complete case estimator. For the complete case analysis and unconditional mean imputation, we skip the weighting step and directly perform model selection via the same adaptive LASSO procedure.



predictions using an estimation window of 60 months. Overall, the results are very similar to those in panel C. Again, note the characteristics we use are known return predictors, and it is thus not surprising that including all of them may yield favorable results. 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 predictive power for returns.

In Appendix A.3.3, we include a third portfolio implementation going long the stocks with highest 50% predicted returns and shorting the stocks with lowest 50% predicted return. In Appendix A.3.4, we present the results of the out-of sample prediction exercise when we additionally include industry dummy variables at the level of the Fama and French 10 industry definition as covariates in the imputation model, but not in the main model. This approach allows missing to differ by industry as we explain in Appendix 2.3.2. Overall, the results with and without industry dummies are very similar.

In the previous analyses, we used all available securities in portfolio formation. This may raise a concern about microcaps dominating the portfolios. To address this concern, we repeat the prediction and portfolio formation step, but we eliminate all firms below and equal to the 20% size percentile at the time of portfolio formation. We report the results in Appendix Table A.7. The level of returns of the long-short portfolio decreases by a nontrivial amount. However, the qualitative conclusions remain the same, that is, imputation models help improve over the complete case in out-of-sample return prediction and we find a modest improvement of using GLS/GMM over mean imputation in most cases.

4.3 Incremental information

We now revisit the classic question if a characteristic contains incremental information relative to previously discovered characteristics. Cochrane (2011) raises this question in his presidential address. The prior literature mostly proceeded in a "univariate fashion," that is, by analyzing one characteristic at a time. However, Green, Hand, and Zhang (2017), Freyberger, Neuhierl, and Weber (2020), Kozak, Nagel, and Santosh (2020), and Gu, Kelly, and Xiu (2020), among other recent papers, show the need to consider characteristics jointly. Hence, to determine whether a characteristic provides incremental information, we need to condition on previously discovered characteristics.

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 issue 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 we take previously published characteristics into account. Throughout, we compare the following three approaches of 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

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

The regression in (4) is not how the literature has progressed, which instead imposed a lower bar by either using no controls or just computing alphas relative to the capital asset pricing model (CAPM) or possibly later the Fama and French (1993) three-factor model.

We estimate this model using only the data until the publication date of the new candidate predictor, not the full sample. To determine whether 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. We report two sets of p-values. The first set is not adjusted for multiple testing, but in the case of the conditional mean imputation does take the additional error from the estimation step into account. The second set of p-values is adjusted for multiple testing. In particular, for each of the 82 models we estimate, we use p-values adjusted for the false discovery rate (see Benjamini and Yekutieli (2001); Green, Hand, and Zhang (2017)).[10] These p-values might be larger than 1 in which case we set them to 1 when presenting our results. Table 6 shows the estimates and p-values. To interpret the results in Table 6, it is easiest to recall how many predictors would be found in a univariate model. We show the results for the univariate model in Figure A.4 in the appendix. In a univariate model, we declare the vast majority of the predictors statistically significant.[11]

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

[11] In the univariate model, we only consider the complete case because the estimators for the parameter

This result resonates with the findings in Jensen, Kelly, and Pedersen (2021), who document a high degree of replicability in empirical asset pricing studies.

Table 6: Incremental information in newly discovered characteristics

This table shows the estimates, p-values, and multiple testing adjusted (false discovery rate) p-values for each new characteristic. The estimation model is the regression model in Equation (4). We test whether a newly discovered characteristic has a significant nonzero effect on returns given all previously discovered characteristics. Estimates are reported in $\times 100$ %. p-values smaller than 5% are in boldface. The final two rows of the table display the number of characteristics that are significant at the 5% and 1% significance levels, respectively. The characteristics are ordered according to their year of discovery.

	Complete case			Cond. mean (GLS)			Uncond. mean		
Characteristic	Est.	p-val	Adj. p-val	Est.	p-val	Adj. p-val	Est.	p-val	Adj. p-val
beta	-0.5409	.5006	.7509	-0.5409	.5006	.7509	-0.5409	.5006	.7509
ep	0.8178	**.0089**	**.0492**	0.8178	**.0089**	**.0492**	0.8166	**.0091**	.0503
size	-1.4669	**.0009**	**.0036**	-1.4008	**.0024**	**.0098**	-1.4057	**.0023**	**.0098**
earningssurprise	1.5031	**.0000**	**.0000**	1.3100	**.0000**	**.0000**	1.3344	**.0000**	**.0000**
lrreversal	-0.1417	.6547	1	0.1706	.5484	1	0.1197	.6709	1
mrreversal	-0.2512	.2785	1	-0.3190	0.2157	0.6664	-0.3343	0.2018	0.7325
bidaskspread	-0.0024	.9929	1	-0.3698	.2209	.6862	-0.5272	**.0377**	.1639
leverage	0.3287	.0749	.6353	0.4485	**.0162**	.0823	0.5716	**.0017**	**.0087**
streversal	-2.2822	**.0000**	**.0000**	-2.4956	**.0000**	**.0000**	-2.4484	**.0000**	**.0000**
bmdec	1.0577	**.0000**	**.0002**	0.7079	**.0034**	**.0223**	0.8630	**.0003**	**.0022**
bookleverage	0.0055	.9863	1	-1.0086	**.0004**	**.0032**	-1.0810	**.0000**	**.0000**
mom12m	1.7553	**.0000**	**.0000**	1.7208	**.0000**	**.0000**	1.7452	**.0000**	**.0000**
mom6m	-0.6392	**.0077**	.0504	-1.1884	**.0000**	**.0000**	-1.0632	**.0000**	**.0001**
cf	1.2370	.1102	.6855	0.5691	.0658	.3274	0.8669	**.0001**	**.0010**
meanrankrevgrowth	0.1385	.3351	1	0.1855	.1679	.8255	-0.0303	.8062	1
accruals	-0.5112	**.0077**	.1126	-0.4800	**.0000**	**.0000**	-0.3949	**.0000**	**.0000**
roe	0.2555	.5615	1	0.4645	.1262	.7219	0.6325	**.0000**	**.0001**
sp	0.4635	**.0395**	03804	0.6160	**.0035**	**.0234**	0.6183	**.0033**	**.0186**
varcf	-0.0656	.858	1	0.0550	.7961	1	0.0228	.9047	1
volmkt	-0.1371	.498	1	-0.4880	**.0315**	.2009	-0.3346	.1256	.6411
volumetrend	-0.3953	**.0114**	.1847	-0.3133	**.0268**	.1813	-0.2793	**.0354**	.2053
chinvia	-0.3913	**.0001**	**.0014**	-0.4132	**.0000**	**0.0000**	-0.4115	**.0000**	**.0000**
grsaletogrinv	0.4093	**.0000**	**.0001**	0.5180	**.0000**	**.0000**	0.5116	**.0000**	**.0000**
indmom	0.2513	.1738	1	.9742	**0.0000**	**.0000**	0.9582	**.0000**	**.0000**
coskewness	-0.3456	.1322	1	-0.2424	0.1401	1	-0.2403	.142	.8372
volsd	-0.7776	.1371	1	-0.6515	0.2345	1	0.1971	.4525	1
chinv	-0.0323	.8515	1	-0.1918	0.0988	0.7239	-0.3279	**.0019**	**.0186**
illiquidity	3.3185	**.0412**	.4735	2.4264	**0.0265**	0.2346	-0.0736	.8915	1
grltnoa	0.0256	.8475	1	.0450	0.6164	1	0.0220	.8004	1
equityduration	-0.7767	.1492	1	-0.3981	0.0976	0.8121	-0.2999	**.0291**	.1915
high52	1.1194	**0008**	**0.0170**	2.0976	**0.0000**	**0.0009**	2.0124	**.0000**	**.0003**
investment	-0.0117	.9433	1	0.0821	0.5727	1	-0.0038	.9706	1
noa	-0.3652	**.0223**	0.2952	-0.9717	**0.0000**	**0.0000**	-1.0206	**.0000**	**.0000**
tax	0.2128	.0784	0.7583	0.0622	0.5662	1	0.0959	.3705	1
cfp	0.1788	.5833	1	0.2275	0.2851	1	0.2631	.2012	1
delcoa	-0.0287	.8923	1	-0.0327	0.823	1	-0.0336	.8246	1
delcol	0.1078	.4731	1	0.0829	0.5294	1	0.0768	.4956	1
delfinl	-0.2278	**.0342**	0.4363	-0.3815	**0.0000**	**0.0003**	-0.3830	**.0000**	**.0000**
dellti	-0.1617	.0772	0.8803	-0.1338	**0.0300**	0.321	-0.1451	**.0207**	.1688
delnetfin	-0.2082	.2387	1	-0.2773	0.0714	0.6912	-0.2764	.0605	.5085

of interest β_1, the slope coefficient, are numerically equivalent for the complete case estimator and the imputation based estimators that do not use weights. Moreover, the weighted estimator yields almost identical results.

Table 6: Incremental information in newly discovered characteristics *(continued)*

pricedelayrsq	-0.2775	.1388	1	-0.1770	0.3257	1	0.1337	.3705	1
pricedelayslope	-0.0668	.5609	1	-0.0398	0.6125	1	-0.0954	.2263	1
pricedelaytstat	-0.0398	.6163	1	-0.0087	0.9239	1	-0.0611	.462	1
totalaccruals	-0.2365	.2183	1	-0.0192	.8425	1	-0.0590	.5201	1
compequiss	-0.3944	.0676	0.9811	-0.8547	**0.0000**	**.0004**	-0.7833	**.0000**	**.0007**
grcapx	-0.3199	.1783	1	-0.2343	.1823	1	-0.0483	.687	1
herf	-0.1525	.2572	1	-0.5115	**.0000**	**.0004**	-0.5103	**.0000**	**.0003**
idiorisk	-0.1309	.5963	1	-0.1260	.5727	1	0.1845	.359	1
netdebtfinance	-0.0753	.5007	1	-0.2036	**.0076**	.1071	-0.1758	**.0163**	.167
operprof	0.4086	.3288	1	0.3386	.2162	1	1.0198	**.0000**	**.0000**
netequityfinance	-0.1300	.3034	1	-0.1192	.368	1	-0.1033	.4239	1
xfin	0.0078	.9719	1	-0.0609	.6449	1	-0.0581	.6241	1
zerotrade	-0.5976	.12	1	0.2139	.489	1	0.5222	.0765	.8222
ebm	0.1187	.4793	1	0.1720	.1345	1	0.2323	**.0326**	.3746
assetgrowth	-0.1697	.7005	1	-0.6540	**.0002**	**.0064**	-0.6237	**.0002**	**.0042**
compositedebtissuance	0.1441	.2167	1	-0.1625	.0716	.8214	-0.0389	.6016	1
investppeinv	-0.0924	.6894	1	.0868	.5774	1	-0.0961	.4596	1
mom12moffseason	-0.0938	.8435	1	-0.9911	**.0166**	.2172	-0.9027	**.0268**	.3491
momoffseason	-0.6142	.2068	1	-1.0601	**.0000**	**.0001**	-0.8506	**.0000**	**.0001**
momoffseason06yrplus	-0.5856	**.0099**	.6449	-0.8756	**.0000**	**.0000**	-0.7982	**.0000**	**.0000**
momseason	0.6720	**.0123**	.7205	0.8469	**.0000**	**.0000**	0.9538	**.0000**	**.0000**
momseason06yrplus	0.8888	**.0000**	**0.0002**	0.6639	**.0000**	**.0000**	0.6662	**.0000**	**.0000**
momseasonshort	0.0311	.8942	1	0.6371	**0.0015**	**.0298**	.6154	**.0022**	**.0345**
shareiss1y	0.1736	.2221	1	0.0220	.8407	1	-0.0777	.4831	1
cashprod	-0.1062	.6973	1	0.5978	**.0002**	**.0057**	0.5671	**.0004**	**.0076**
cheq	-0.4673	.0847	1	0.0389	.7818	1	-0.0717	.5468	1
maxret	0.0507	.8945	1	0.4454	.1541	1	0.4258	.1682	1
opleverage	-0.2063	.7517	1	-0.4462	.1145	1	-0.2699	.2352	1
roaq	1.1175	**.0000**	**0.0065**	1.6705	**.0000**	**.0000**	1.3882	**.0000**	**.0000**
entmult	-0.2862	.5951	1	-0.5653	**.0176**	.2169	-0.5635	**.0000**	**.0006**
rds	-0.2272	**.0433**	1	0.0582	.4903	1	0.0882	.3542	1
residualmomentum	0.1587	.6766	1	0.2274	.4927	1	-0.0297	.9116	1
cash	0.8235	**.0000**	**0.0015**	1.2623	**.0000**	**.0000**	1.1832	**.0000**	**.0000**
invgrowth	1.0660	.1303	1	0.4342	.2895	1	0.2339	.1189	1
intmom	0.3672	.3145	1	-0.1285	.5913	1	-0.0946	.6968	1
gp	0.3529	.3405	1	0.4540	**.0037**	.0615	0.3423	**.0086**	.1311
betafp	-0.3429	.3316	1	0.0892	.7209	1	0.3175	.1561	1
betatailrisk	0.3229	.3404	1	0.3692	.1228	1	0.3505	.1225	1
hire	0.0629	.7172	1	0.0642	.4316	1	0.0652	.3689	1
cboperprof	0.2999	.4162	1	0.8783	**.0000**	**.0004**	0.4308	**.0027**	**.0452**
returnskew	0.0628	.8133	1	0.1748	.1835	1	0.1350	.3053	1
trendfactor	1.3712	**.0000**	**.0000**	1.4088	**.0000**	**.0000**	1.1465	**.0000**	**.0003**
# sign. at 5%		**23**	**13**		**38**	**29**		**43**	**34**
# sign. at 1%		**16**	**11**		**31**	**25**		**36**	**30**

Now, in a multivariate setup, 11 to 23 characteristics are significant in the complete case. A lack of statistical power in the complete case can explain this relatively small number. Even very strong predictors, such as 6-month momentum (`mom6m`) or book leverage (`bookleverage`), would not be significant in the complete case after adjusting for multiple testing. Conditional mean imputation using the GLS adjustment selects more characteristics, but still fewer than unconditional mean imputation. Most notably the selection of

characteristics between the conditional mean and unconditional mean imputation is different. This difference is due to the interaction of two effects highlighted in Section 3. First, unconditional mean imputation yields biased estimators and estimated coefficients may be either too large or too close to zero. As a specific example, consider operating profitability (`operprof`) in Table 6. The coefficients in the complete case and with conditional mean imputation are quite similar (0.4086% and 0.3386%, respectively), whereas unconditional mean imputation yields a much larger estimated coefficient (1.0198%) that is significantly different from 0. Second, with unconditional mean imputation, we underestimate the covariance between the characteristics and therefore tend to obtain artificially small standard errors.

5 Conclusion

Missing data occur in virtually all cross-sectional empirical asset pricing studies, but is also a prevalent problem in empirical corporate finance research, innovation research, international finance, and many other fields of economics. The primary goal of this paper is to provide empirical researchers with an easy approach to address this problem 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 the complete case method, despite its intuitive appeal, neglects an important part of the return distribution. We therefore advocate the use of imputation. Moreover, since unconditional mean imputation leads to biases in the estimation and incorrect inference, we cannot advocate using it. Instead, researchers should use conditional mean imputation and adjust for the estimation error in subsequent inference.

Our proposed two-step approach is applicable in other common areas of research, such as estimating the stochastic discount factors (see Appendix Section A.4), characteristic-based factor models, and international studies. These items are left for future research.

The replication code and data are available in the Harvard Dataverse at `https://doi.org/10.7910/DVN/QR6PHI`.

References

Abrevaya, J., and S. G. Donald. 2017. A gmm approach for dealing with missing data on regressors. *Review of Economics and Statistics* 99:657–62.

Bai, J., and S. Ng. 2021. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association* 116:1746–63.

Beaver, W., M. McNichols, and R. Price. 2007. Delisting returns and their effect on accounting-based market anomalies. *Journal of Accounting and Economics* 43:341–68.

Beckmeyer, H., and T. Wiedmann. 2023. Recovering missing firm characteristics with attention-based machine learning. *Working Paper, University of Muenster* .

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

Brown, S. J., W. Goetzmann, R. G. Ibbotson, and S. A. Ross. 1992. Survivorship bias in performance studies. *Review of Financial Studies* 5:553–80.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger. 2023. Missing financial data. *Working Paper, London Business School* .

Cahan, E., J. Bai, and S. Ng. 2023. Factor-based imputation of missing values and covariances in panel data of large dimensions. *Journal of Econometrics* 233:113–31.

Carhart, M. M., J. N. Carpenter, A. W. Lynch, and D. K. Musto. 2002. Mutual fund survivorship. *Review of Financial Studies* 15:1439–63.

Chen, A. Y., and J. McCoy. forthcoming. Missing values and the dimensionality of expected returns. *Journal of Financial Economics* .

Chen, A. Y., and T. Zimmermann. forthcoming. Open source cross sectional asset pricing. *Critical Finance Review* .

Chen, X., H. Hong, and A. Tarozzi. 2008. Semiparametric efficiency in gmm models with auxiliary data. *Annals of Statistics* 36:808–43.

Cochrane, J. H. 2011. Presidential address: Discount rates. *Journal of Finance* 66:1047–108.

Connor, G., and R. A. Korajczyk. 1987. Estimating pervasive economic factors with missing observations. *Working Paper, London School of Economics & Political Science* .

Dagenais, M. G. 1973. The use of incomplete observations in multiple regression analysis: A generalized least squares approach. *Journal of Econometrics* 1:317–28.

Fama, E. F., and K. R. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47:427–65.

———. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56.

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., 3–22.

Freyberger, J., A. Neuhierl, and M. Weber. 2020. Dissecting characteristics nonparametrically. *Review of Financial Studies* 33:2326–77.

Gourieroux, C., and A. Monfort. 1981. On the problem of missing data in linear models. *Review of Economic Studies* 48:579–86.

Green, J., J. R. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average us monthly stock returns. *Review of Financial Studies* 30:4389–436.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33:2223–73.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54.

Hansen, L. P., J. Heaton, and A. Yaron. 1996. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics* 14:262–80.

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

Haugen, R. A., and N. L. Baker. 1996. Commonality in determinants of expected stock returns. *Journal of Financial Economics* 41:401–39.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61.

Huang, J., J. L. Horowitz, and F. Wei. 2010. Variable selection in nonparametric additive models. *Annals of Statistics* 38:2282–313.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen. 2021. Is there a replication crisis in finance? Working Paper, National Bureau of Economic Research.

Jin, S., K. Miao, and L. Su. 2021. On factor models with random missing: Em estimation, inference, and cross validation. *Journal of Econometrics* 222:745–77.

Kelly, B. T., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics* 134:501–24.

Kim, S., R. A. Korajczyk, and A. Neuhierl. 2021. Arbitrage portfolios. *Review of Financial Studies* 34:2813–56.

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* 204:159–88.

Koh, P.-S., D. M. Reeb, E. Sojli, W. W. Tham, and W. Wang. 2022. Deleting unreported innovation. *Journal of Financial and Quantitative Analysis* 57:2324–54.

Kozak, S., S. Nagel, and S. Santosh. 2020. Shrinking the cross-section. *Journal of Financial Economics* 135:271–92.

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

Liao, Z., and Y. Liu. 2021. Optimal cross-sectional regression. *Working Paper, University of California at Los Angeles* .

Light, N., D. Maslov, and O. Rytchkov. 2017. Aggregation of information about the cross section of stock returns: A latent variable approach. *Review of Financial Studies* 30:1339–81.

Little, R. J. A. 1992. Regression with missing x's: a review. *Journal of the American Statistical Association* 87:1227–37.

———. 1994. A class of pattern-mixture models for normal incomplete data. *Biometrika* 81:471–83.

Little, R. J. A., and D. B. Rubin. 2020. *Statistical analysis with missing data*. 3 ed. Hoboken, John Wiley & Sons, Inc.

Liu, H., X. Tang, and G. Zhou. 2022. Recovering the fomc risk premium. *Journal of Financial Economics* 145:45–68.

Lynch, A. W., and J. A. Wachter. 2013. Using samples of unequal length in generalized method of moments estimation. *Journal of Financial and Quantitative Analysis* 48:277–307.

Manski, C. F. 2005. Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning* 39:151–65.

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.

Nijman, T., and F. Palm. 1988. Efficiency gains due to using missing data procedures in regression models. *Statistical Papers* 29:249–56.

Rao, C. R., and H. Toutenburg. 1999. *Linear models: Least squares and alternatives*. 2 ed. Springer.

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* 89:846–66.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63:581–92.

———. 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*, vol. 1, 20–34. American Statistical Association.

Shumway, T. 1997. The delisting bias in crsp data. *Journal of Finance* 52:327–40.

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., 149–84. Boca Raton: CRC Press, Taylor & Francis Group.

Wooldridge, J. M. 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141:1281–301.

Xiong, R., and M. Pelger. 2023. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics* 233:271–301.

Yates, F. 1933. The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture* 1:129–42.

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* 100:1394–411.

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

Internet Appendix: Missing Data in Asset Pricing Panels

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

Table A.1: Overview of the Characteristics

Acronym	Description	Publication Year	% missing
Accruals	Accruals	1996	0.50
AssetGrowth	Asset growth	2008	0.00
Beta	CAPM beta	1973	0.00
BetaFP	Frazzini-Pedersen Beta	2014	6.82
BetaTailRisk	Tail risk beta	2014	34.65
BidAskSpread	Bid-ask spread	1986	7.33
BMdec	Book to market using December ME	1992	0.00
BookLeverage	Book leverage (annual)	1992	0.00
Cash	Cash to assets	2012	28.91
CashProd	Cash Productivity	2009	9.73
CBOperProf	Cash-based operating profitability	2016	29.60
CF	Cash flow to market	1994	9.10
cfp	Operating Cash flows to price	2004	14.34
ChEQ	Growth in book equity	2010	3.64
ChInv	Inventory Growth	2002	0.00
ChInvIA	Change in capital inv (ind adj)	1998	11.45
CompEquIss	Composite equity issuance	2006	37.48
CompositeDebtIssuance	Composite debt issuance	2008	40.26
Coskewness	Coskewness	2000	0.00
DelCOA	Change in current operating assets	2005	0.00
DelCOL	Change in current operating liabilities	2005	0.50
DelFINL	Change in financial liabilities	2005	0.79
DelLTI	Change in long-term investment	2005	0.00
DelNetFin	Change in net financial assets	2005	0.79
EarningsSurprise	Earnings Surprise	1984	19.17
EBM	Enterprise component of BM	2007	9.67
EntMult	Enterprise Multiple	2011	27.24
EP	Earnings-to-Price Ratio	1977	35.05
EquityDuration	Equity Duration	2004	1.83
GP	gross profits / total assets	2013	19.42
grcapx	Change in capex (2 years)	2006	18.78
GrLTNOA	Growth in long-term operating assets	2003	2.64
GrSaleToGrInv	Sales growth over inventory growth	1998	23.15
Herf	Industry concentration (sales)	2006	16.75
High52	52 week high	2004	0.00
hire	Employment growth	2014	1.53
IdioRisk	Idiosyncratic risk	2006	0.00
Illiquidity	Amihud's illiquidity	2002	4.72
IndMom	Industry Momentum	1999	9.10
IntMom	Intermediate Momentum	2012	9.25
Investment	Investment to revenue	2004	25.73
InvestPPEInv	change in ppe and inv/assets	2008	12.02
InvGrowth	Inventory Growth	2012	40.76
Leverage	Market leverage	1988	9.33
LRreversal	Long-run reversal	1985	16.00
MaxRet	Maximum return over month	2010	0.00
MeanRankRevGrowth	Revenue Growth Rank	1994	35.49
Mom12m	Momentum (12 month)	1993	9.28
Mom12mOffSeason	Momentum without the seasonal part	2008	9.15
Mom6m	Momentum (6 month)	1993	9.18
MomOffSeason	Off season long-term reversal	2008	9.69
MomOffSeason06YrPlus	Off season reversal years 6 to 10	2008	31.66
MomSeason	Return seasonality years 2 to 5	2008	9.68
MomSeason06YrPlus	Return seasonality years 6 to 10	2008	31.53
MomSeasonShort	Return seasonality last year	2008	9.18
MRreversal	Medium-run reversal	1985	9.32
NetDebtFinance	Net debt financing	2006	10.75
NetEquityFinance	Net equity financing	2006	0.92
NOA	Net Operating Assets	2004	0.42
OperProf	operating profits / book equity	2006	56.49
OPLeverage	Operating leverage	2010	0.20
PriceDelayRsq	Price delay r square	2005	2.43

Table A.1: Overview of the Characteristics *(continued)*

PriceDelaySlope	Price delay coeff	2005	2.43
PriceDelayTstat	Price delay SE adjusted	2005	2.71
RDS	Real dirty surplus	2011	6.63
ResidualMomentum	Momentum based on FF3 residuals	2011	13.33
ReturnSkew	Return skewness	2016	0.59
roaq	Return on assets (qtrly)	2010	13.83
RoE	net income / book equity	1996	0.01
ShareIss1Y	Share issuance (1 year)	2008	9.31
Size	Size	1981	0.00
SP	Sales-to-price	1996	9.30
STreversal	Short-term reversal	1989	0.00
Tax	Taxable income to income	2004	11.58
TotalAccruals	Total accruals	2005	4.97
TrendFactor	Trend Factor	2016	52.96
VarCF	Cash-flow to price variance	1996	15.51
VolMkt	Volume to market equity	1996	4.07
VolSD	Volume Variance	2001	5.93
VolumeTrend	Volume Trend	1996	10.48
XFIN	Net external financing	2006	11.43
zerotrade	Days with zero trades	2006	4.00

A.1 Additional Definitions

We briefly lay out some basic definitions relevant to the treatment of missing data. Introductory treatments can be found, for example, in Fitzmaurice et al. (2015) or Little and Rubin (2020).

A.1.1 Missing patterns

A missing pattern describes which data are missing. Figure 5 shows an example of a missing pattern. 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 nonnegligible 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; that is, 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 recast these basic classical definitions, using our notation from Section 2, and present how our MAR missing assumption relates to these classical missing mechanisms.

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 are **MCAR** in the classical sense if $D_i \perp\!\!\!\perp Y_i, X_i, V_i$, that is, whether an observation is missing does not depend on the other variables. When the data are MCAR, the complete case analysis yields valid inference, but there is a loss of efficiency relative to the complete data analysis because of the decreased sample size (Fitzmaurice et al. (2015)). Since we are primarily interested

in regression of Y_i on X_i and Y_i is always observed, we refer to a missing mechanism as MCAR if $D_i \perp\!\!\!\perp X_i, V_i$. The data are **MAR** in the classical sense if $D_i \perp\!\!\!\perp X_i \mid Y_i, X_i^{(o)}, V_i$. That is, missingness is only random once we condition on observed covariates and Y_i. We rely on this type of assumption (but based solely on conditional moments) in our analysis. Specifically, a sufficient condition for the imputations in our estimation procedure to be valid is the MAR-like assumption $D_i \perp\!\!\!\perp X_i \mid X_i^{(o)}, V_i$. Note that we can also allow missingness to depend on Y_i or use Y_i in the imputations, but would then have to weight the moment conditions with an inverse propensity weight, $P(D_i = l \mid Y_i, X_i^{(o)}, V_i)$ (see Section 2.3.2 for inverse propensity weighting).

Data are not missing at random **NMAR** in the classical sense, if D_i depends on unobserved regressors. 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 Simulation Appendix

In this section we present additional simulation results.

A.2.1 Additional results for the low-dimensional setting

Table A.2 shows results for the setup of Section 3.1, but with the regressors being jointly independent. The complete sample 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 imputation 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 in which unconditional mean imputation outperforms the other methods is when all regression coefficients in front of regressors that have missing values are equal to zero. In this case, unconditional mean imputation leads to correct moment conditions, as discussed in Section 2.1. Moreover, imputing the conditional or the unconditional mean does not increase the variance

Table A.2: 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 when 50% of the data are missing at random and all regressors are independent. The table reports results for the estimator that only uses the complete subset of the data, the optimal GMM estimator, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.

	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	0.896	0.135	0.897	0.137	0.894	0.266	0.825	0.217
β_2	0.905	0.147	0.895	0.134	0.902	0.136	0.893	0.265	0.903	0.216
β_3	0.909	0.147	0.893	0.144	0.898	0.146	0.896	0.309	0.901	0.250
β_4	0.900	0.147	0.902	0.135	0.906	0.137	0.902	0.220	0.888	0.198
β_5	0.903	0.147	0.891	0.136	0.898	0.138	0.896	0.149	0.902	0.144

of the error term and hence, no benefits from using GLS exist. We show simulation results in Table A.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 zero, that is, whether a firm characteristic is a true return predictor and we should therefore not rely on the unconditional

Table A.3: 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 when 50% of the data are missing at random and all regressors that may be missing do not affect the outcome. The table reports results for the estimator that only uses the complete subset of the data, the optimal GMM estimator, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.

	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	0.896	0.103	0.897	0.104	0.898	0.104	0.895	0.104
β_2	0.909	0.338	0.905	0.246	0.907	0.250	0.906	0.250	0.898	0.197
β_3	0.907	0.454	0.897	0.366	0.898	0.371	0.898	0.371	0.900	0.209
β_4	0.897	0.453	0.888	0.377	0.891	0.381	0.891	0.381	0.902	0.239
β_5	0.903	0.337	0.897	0.289	0.901	0.292	0.901	0.292	0.914	0.215

mean imputation to deliver satisfactory results. As we discuss in the main text and further below, to determine which regressors are irrelevant, we can carry out model selection to obtain a smaller model.

A.2.2 Additional results for the high-dimensional setting

In Section 3.2, we only considered sparse setups, in the sense that only 5 coefficients were not equal to zero. We now assume that $\beta_k = 0.8^k$, but leave all other features of the data generating process unchanged. The selection results are illustrated in Figure A.1. For the complete case estimator and both conditional mean imputation estimators, the larger a coefficient, the more likely it is not set to zero. This monotonicity does not hold for unconditional mean imputation. Here, the estimated values of β_5 and β_7 are often set to zero, but regressors with smaller coefficients are included much more frequently. The MSPEs of the four methods are 1.0536, 1.0339, 1.0335, and 1.0818, respectively.

As in the low-dimensional case, unconditional mean imputation works particularly well when regressors with missing values to do not affect the outcome. To illustrate this case, 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 set the remaining 15 elements all to zero. The results are reported in Figure A.2. In this 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 1.0187, 1.0069, 1.0069, and 1.0069 respectively. Therefore, all imputation methods perform similarly well and outperform the complete case.

A.2.3 Additional results for alternative models

We first describe the factor model that we use to generate the data for the results in Section 3.3 before illustrating the performance of the estimator when the data are not missing at random.

As before, we set $Var(\varepsilon_i) = 1$. We then generate the regressors as

$$X_{i,k} = F_i' \Lambda_k + e_{i,k}$$

where F_i, Λ_k, e_i and all its elements are mutually independent, $e_{i,k} \sim N(0, 1/3)$ for all $k = 2, \ldots, 40$,

$F_i \sim N(0, (2/3)I_{3\times3})$, and $\Lambda_{k,l} \sim U[0,1]$ for all $l = 1,2,3$ and $k = 2,\ldots,40$. One can show that the DGP implies $Var(X_{i,k}) = 1$ for all k and $cov(X_{i,k}, X_{i,l}) = 0.5$ for all $l \neq k$.

We now present results for three additional simulation setups, which are constructed to show the limits of the different methods. In all setups, we simulate data as in Section 3.2 except that the regressors have a factor structure and the data are not missing at random. We show that, depending on the exact missing mechanism, either the estimator based on the factor model or our approach might perform better.

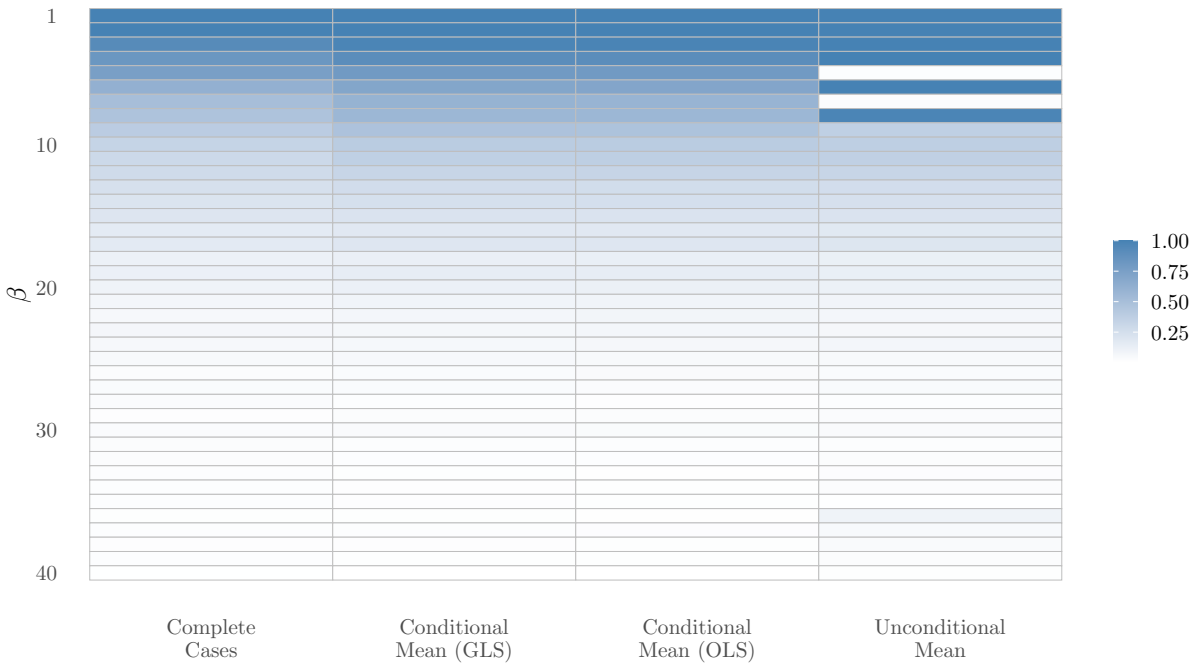
To describe the first of the three setups, let $G_i \sim \text{Bern}(1/2)$ independent across all i, and let

$$F_i \sim \mathbf{1}(G_i = 0)Z_{i1} + \mathbf{1}(G_i = 1)Z_{i2}$$

where $Z_{i1} \sim N(0, (1/3)I_{3\times3})$ and $Z_{i2} \sim N(0, I_{3\times3})$ and Z_{i1} and Z_{i2} are independent. It can be shown that with this DGP $Var(F_i)$ is still equal to $(2/3)I_{3\times3}$. Also as before, we let $\Lambda_{k,l} \sim U[0,1]$,

Figure A.1: Model Selection - Dense Model

This figure shows the frequency with which the different methods select regressors for the simulation setup with a dense model. The darker the color, the more frequent a particular model estimates a non-zero β_k. The true model is dense with β_k decreasing in the index k, specifically $\beta_k = 0.8^k$. The figure shows results for the estimator that only uses the complete subset of the observations, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.



$e_{i,k} \sim N(0, 1/3)$, and

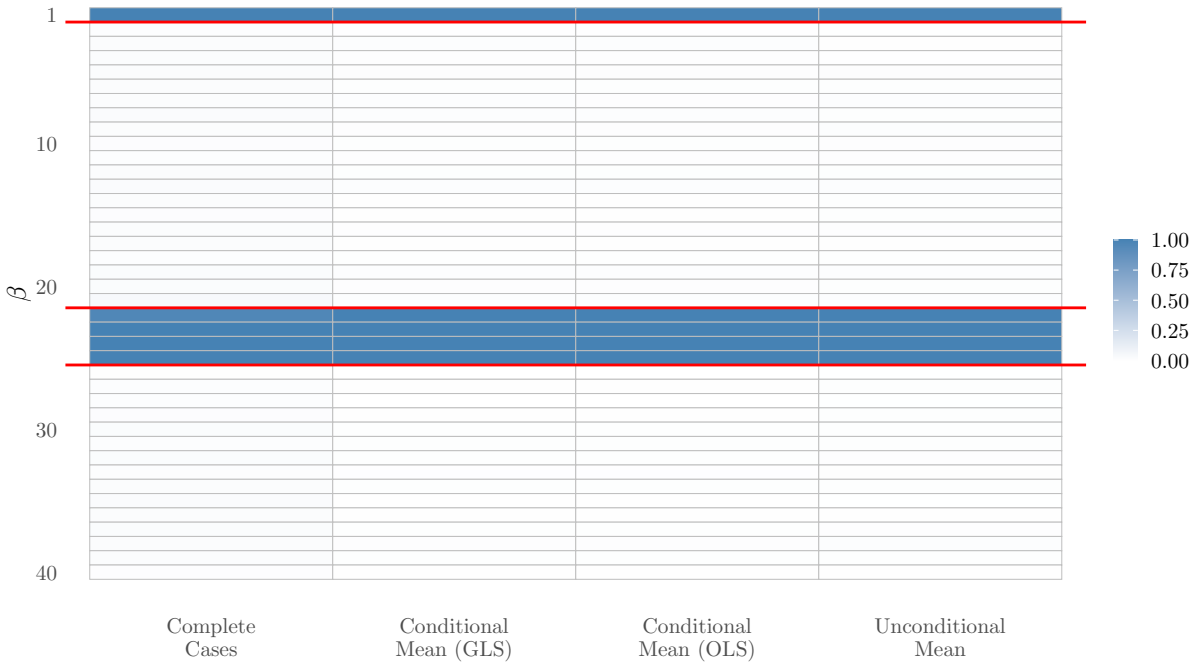
$$X_{i,k} = F_i' \Lambda_k + e_{i,k}$$

The random variables/vectors F_i, Λ_k, G_i, and e_i and all its elements are mutually independent. That is, we have a factor structure and the factors have a mixture distribution.

We now first assign 50% of the observations with $G_i = 0$ and 50% of the observations with $G_i = 1$ randomly to the complete case. Next, $\frac{1}{6} \times 100\%$ of the observations with $G_i = 0$ and $\frac{1}{6} \times 100\%$ of the observations with $G_i = 1$ are assigned to missing pattern $l = 1$, $\frac{31}{96} \times 100\%$ of the observations with $G_i = 0$ and $\frac{1}{96} \times 100\%$ of the observations with $G_i = 1$ are assigned to missing pattern $l = 2$, and $\frac{1}{96} \times 100\%$ of the observations with $G_i = 0$ and $\frac{31}{96} \times 100\%$ of the observations with $G_i = 1$ are assigned to missing pattern $l = 3$. These assignments are constructed in such a way

Figure A.2: Model Selection - Missing Regressor Irrelevant

This figure shows the frequency with which the different methods select regressors for the simulation setup with a sparse model 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_k. The non-zero coefficients of the true model are separated with red lines. The figure shows results for the estimator that only uses the complete subset of the observations, the imputation GLS estimator, the imputation OLS estimator, and the estimator that imputes the unconditional mean.



that we have the same number of observations in each missingness pattern with missing variables. However, importantly, the missingness patterns depend on all regressors through G_i.

In the second of the three setups, the missingness patterns depend on the loadings. In particular, let G_k be a multinomial random variable such that

$$G_k = \begin{cases} 1 & \text{with probability } \frac{1}{3} \\ 2 & \text{with probability } \frac{1}{3} \\ 3 & \text{with probability } \frac{1}{3} \end{cases}$$

Next, let $U_{k,1} \sim U[0, 1/3]$, $U_{k,2} \sim U[1/3, 2/3]$, and $U_{k,3} \sim U[2/3, 1]$, independent of G_k and define

$$\Lambda_{k,l} \sim \mathbf{1}(G_k = 1)U_{k,1} + \mathbf{1}(G_k = 2)U_{k,2} + \mathbf{1}(G_k = 3)U_{k,3},$$

for all k and l. It is easy to show that $\Lambda_{k,l} \sim U[0, 1]$. Moreover, let $e_{i,k} \sim N(0, 1/3)$, $F_i \sim N(0, (2/3)I_{3\times 3})$, and

$$X_{i,k} = F_i'\Lambda_k + e_{i,k}$$

We construct four missingness patterns based on G_k. First, we have the complete case ($l = 0$) to which we randomly assign 50% of the observations. For the remaining 50% of the observations, we randomly select one of the remaining three missing patterns with equal probability. In these three missing patterns, we always observe characteristics $k = 31, \ldots, 40$ and the intercept. Of the remaining characteristics, we do not observe characteristics with $G_k = l$ for missing patterns $l \in \{1, 2, 3\}$.

In the last of the three setups, we use the baseline factor model described at the very beginning of this section, but with missingness being a function of missing covariates. In particular, 20% of the observations are randomly assigned to the complete case, while for the remaining part of the sample, observations with high values of $X_{i,4} + X_{i,5}$ are more likely to be complete. Again, 50% of the observations are complete.

The simulation results for three setups are shown in Table A.4. In terms of the imputation error and the prediction error, the factor model based estimator works best when missingness depends

on the factors, but the GLS estimator yields the lowest MSE of the estimated coefficients. More generally, the factor model estimator is robust to missingness that depends on F_i. However, recall that the estimator crucially relies on the regressors actually having a factor structure. Moreover, as can be seen from Table A.4, the factor model estimator performs very poorly in the setup in which missingness depends on the loadings, and it is then outperformed by our GLS estimator. When selection is based on the regressors, the EM algorithm performs the worst in terms of the imputation error, while the factor model estimator has the lowest imputation error. However, because of the GLS weighting scheme, our estimator has the lowest MSE of the estimated coefficients, and also performs better than the EM estimator in terms of out-of-sample predictions.

Table A.4: Simulation - Average MSEs of different methods with NMAR

This table shows mean squared errors of our GLS estimator, the factor model of Bryzgalova et al. (2023), and the EM algrithm of Chen and McCoy (forthcoming) with different data generating processes. The data are not missing at random and selection is based on the factors, the loadings, or the regressors. The table shows mean squared errors of the three estimators and different DGPs. For each setup and estimator, the table reports the average mean squared error of the estimated coefficients, the average mean squared imputation error of the regressors and the average mean squared out-of-sample errors for the outcomes.

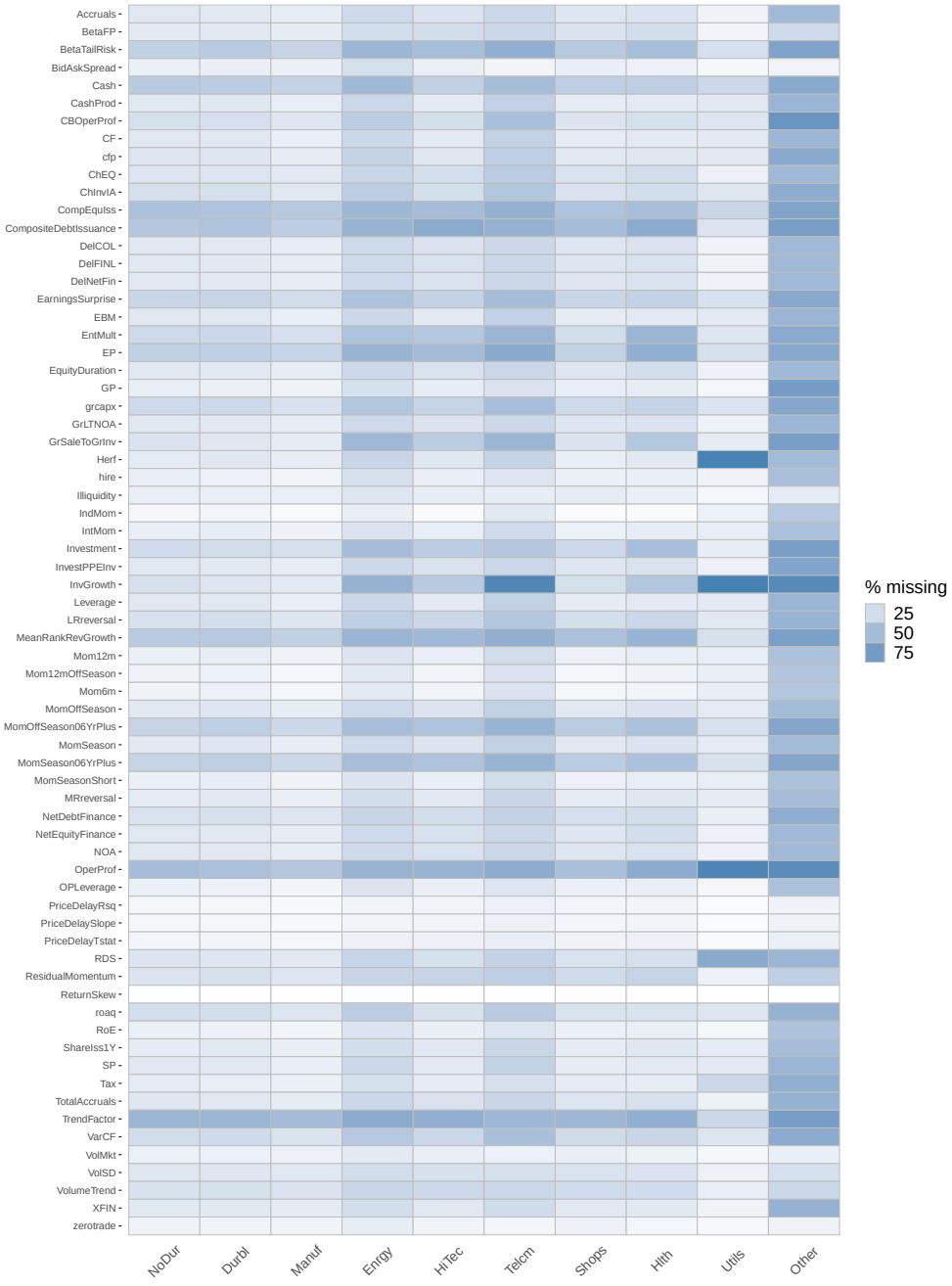
	Selection on factors			Selection on loadings			Selection on regressors		
	GLS	EM	Factor	GLS	EM	Factor	GLS	EM	Factor
Coefficients	0.0048	0.0087	0.0066	0.0049	0.0072	0.0116	0.0050	0.0135	0.0069
Regressors	0.1527	0.1452	0.1364	0.1458	0.1462	0.4898	0.1776	0.2063	0.1424
Outcomes	2.4029	2.4692	2.3623	1.8186	1.8717	2.5759	2.4324	2.6170	2.3060

A.3 Further empirical results

A.3.1 Missingness by FFI10 industry dummies

Figure A.3: Missingness of characteristics by FFI10 industry dummies

The figure summarizes the missing percentage of the different variables by the FFI10 industry dummies. We dropped the always observed characteristics from the figure.



A.3.2 Masking exercise

Table A.5: Out-of-sample prediction error (RMSPE): Masking with FFI10 dummies

This table shows the RMSPE for conditional mean imputation with industry dummies as covariates in the masking exercise across all characteristics. We do not include the characteristics that we require to be always observed. The final row contains the root of the weighted average MSPE across all characteristics, where the weight for a characteristic equals the number of missing values for this characteristic divided by the number of all missing characteristic values. For simplicity we denote the row with "Average RMSPE." We first randomly delete 1% of the entries in the data matrix of the complete case. Then, we impute the missing characteristic values using conditional mean imputation with and without lags, where we include industry dummies in the imputation model. In a final step, we calculate the RMSPE by comparing the imputed characteristic values with the initially deleted values.

	Cond. Mean (GLS)	Cond. Mean (GLS / w. lags)
Accruals	0.12122	0.12071
BetaFP	0.16485	0.06468
BetaTailRisk	0.20491	0.06245
BidAskSpread	0.21533	0.19983
Cash	0.21803	0.14822
CashProd	0.13613	0.07129
CBOperProf	0.13496	0.13168
CF	0.08774	0.06385
cfp	0.15936	0.09104
ChEQ	0.14600	0.14535
ChInvIA	0.21228	0.12415
CompEquIss	0.19133	0.10559
CompositeDebtIssuance	0.22495	0.19889
DelCOL	0.15161	0.15079
DelFINL	0.10655	0.10655
DelNetFin	0.11999	0.12001
EarningsSurprise	0.24055	0.18265
EBM	0.18691	0.12238
EntMult	0.10932	0.06940
EP	0.12925	0.09007
EquityDuration	0.13540	0.13373
GP	0.17364	0.10726
grcapx	0.14168	0.14151
GrLTNOA	0.18619	0.18624
GrSaleToGrInv	0.19265	0.18933
Herf	0.25599	0.04871
hire	0.20955	0.20937
Illiquidity	0.05323	0.02414
IndMom	0.27130	0.20152
IntMom	0.18836	0.15796
Investment	0.14689	0.09644
InvestPPEInv	0.14072	0.13924
InvGrowth	0.10462	0.08375
Leverage	0.06433	0.03651
LRreversal	0.19071	0.11578
MeanRankRevGrowth	0.20657	0.07460
Mom12m	0.08989	0.08493
Mom12mOffSeason	0.11117	0.10846
Mom6m	0.15758	0.15409
MomOffSeason	0.16455	0.10434
MomOffSeason06YrPlus	0.24664	0.13105
MomSeason	0.25957	0.25836
MomSeason06YrPlus	0.28488	0.28500
MomSeasonShort	0.23116	0.23098
MRreversal	0.25149	0.18963

Table A.5: Out-of-sample prediction error (RMSPE) *(continued)*

NetDebtFinance	0.15729	0.15702
NetEquityFinance	0.17500	0.17236
NOA	0.18329	0.15346
OperProf	0.11373	0.10822
OPLeverage	0.11438	0.08283
PriceDelayRsq	0.23137	0.23079
PriceDelaySlope	0.26373	0.26391
PriceDelayTstat	0.27761	0.27279
RDS	0.24280	0.24341
ResidualMomentum	0.17192	0.12596
ReturnSkew	0.20477	0.20464
roaq	0.17749	0.17393
RoE	0.10454	0.10404
ShareIss1Y	0.20179	0.11167
SP	0.07023	0.03507
Tax	0.24954	0.23840
TotalAccruals	0.15899	0.15901
TrendFactor	0.26446	0.24937
VarCF	0.17395	0.03973
VolMkt	0.08600	0.04769
VolSD	0.11228	0.03406
VolumeTrend	0.22816	0.07013
XFIN	0.12176	0.12112
zerotrade	0.09608	0.06299
Average RMSPE	**0.18180**	**0.14982**

A.3.3 Additional out-of-sample predictions

This section collects additional results for the out-of-sample predictions discussed in Section 4.2.

Table A.6: Performance statistics for out-of-sample predictions (all firms)

This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction. The portfolios are equal weighted. We differentiate between the complete case method, unconditional mean imputation, and conditional mean imputation with GLS weighting without and with lags. Long Pf. and Short Pf. denote the annualized average return of the long and short legs, respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and nonlinear model is outlined in Section 4.2. The sample period is 1990–2021.

	Mean (%)	Standard deviation (%)	Sharpe ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
A. Linear model							
Long (short) 50% highest (lowest) predicted returns							
Complete case	6.49	5.84	1.11	16.38	9.89	0.27	3.15
Uncond. mean	12.56	8.52	1.47	21.08	8.51	0.62	13.22
Cond. mean (GLS)	13.41	8.61	1.56	21.50	8.09	-0.13	5.79
Cond. mean (GLS / w. lags)	13.40	8.63	1.55	21.50	8.09	-0.14	5.86
B. Regularized linear model							
Long (short) 50% highest (lowest) predicted returns							
Complete case (LASSO)	6.64	6.23	1.06	16.46	9.82	-0.29	4.55
Uncond. mean (LASSO)	12.36	8.95	1.38	20.97	8.62	0.15	11.81
Cond. mean (GLS / LASSO)	12.44	9.37	1.33	21.02	8.58	-0.50	5.87
Cond. mean (GLS / LASSO / w. lags)	12.55	9.37	1.34	21.07	8.52	-0.50	6.20
C. Nonlinear model							
Long (short) 50% highest (lowest) predicted returns							
Complete case	6.71	5.10	1.32	16.49	9.78	0.32	3.96
Uncond. mean	14.96	7.59	1.97	22.27	7.32	0.85	12.54
Cond. mean (GLS)	16.07	8.12	1.98	22.83	6.76	0.32	6.82
Cond. mean (GLS / w. lags)	16.16	8.01	2.02	22.88	6.71	0.33	6.55
D. Regularized nonlinear model							
Long (short) 50% highest (lowest) predicted returns							
Complete case (LASSO)	5.31	6.55	0.81	15.79	10.48	0.52	4.82
Uncond. mean (LASSO)	13.48	7.90	1.71	21.53	8.06	0.19	9.35
Cond. mean (GLS / LASSO)	14.63	8.57	1.71	22.11	7.48	-0.03	6.45
Cond. mean (GLS / LASSO / w. lags)	15.08	8.20	1.84	22.33	7.26	0.00	5.49

Table A.7: Performance Statistics For Out-of-Sample Predictions (large firms)

This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction for large firms. Specifically, we only include firms with a market capitalization greater than the 20th size percentile at the time of portfolio formation. Portfolios are equally weighted. We differentiate between the complete case method, unconditional mean imputation, and conditional mean imputation with GLS weighting without and with lags. Long Pf. and Short Pf. denote the annualized average return of the long and short legs, respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and nonlinear model is outlined in Section 4.2. The sample period is 1990–2021.

	Mean (%)	Standard Deviation (%)	Sharpe Ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
A. Linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case	11.37	9.57	1.19	19.12	7.75	0.22	3.39
Uncond. Mean	31.35	25.07	1.25	25.93	-5.42	0.01	15.64
Cond. Mean (GLS)	36.83	25.59	1.44	28.39	-8.44	-0.12	7.01
Cond. Mean (GLS / w. lags)	36.40	25.80	1.41	28.27	-8.13	-0.12	7.83
Long (short) 10% highest (lowest) predicted returns							
Complete Case	15.74	13.12	1.20	20.70	4.96	0.35	6.07
Uncond. Mean	22.92	18.13	1.26	23.37	0.45	0.58	14.27
Cond. Mean (GLS)	25.54	18.63	1.37	24.47	-1.07	0.42	8.05
Cond. Mean (GLS / w. lags)	25.56	18.52	1.38	24.42	-1.13	0.45	8.22
B. Regularized linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case (LASSO)	12.47	10.63	1.17	18.96	6.49	-0.45	3.96
Uncond. Mean (LASSO)	31.63	24.79	1.28	27.98	-3.65	-0.30	9.79
Cond. Mean (GLS / LASSO)	33.17	24.74	1.34	26.08	-7.09	-0.38	3.99
Cond. Mean (GLS / LASSO / w. lags)	34.01	24.65	1.38	26.54	-7.47	-0.32	4.65
Long (short) 10% highest (lowest) predicted returns							
Complete Case (LASSO)	17.21	14.74	1.17	21.39	4.18	0.32	4.99
Uncond. Mean (LASSO)	21.92	18.72	1.17	23.00	1.08	0.12	12.79
Cond. Mean (GLS / LASSO)	24.12	18.98	1.27	24.03	-0.10	0.02	6.39
Cond. Mean (GLS / LASSO / w. lags)	24.18	18.95	1.28	23.99	-0.19	-0.04	6.12
C. Nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case	11.06	8.57	1.29	18.45	7.39	0.10	1.71
Uncond. Mean	42.41	25.14	1.69	29.35	-13.06	-0.28	7.09
Cond. Mean (GLS)	45.80	25.56	1.79	30.83	-14.97	-0.82	6.42
Cond. Mean (GLS / w. lags)	45.14	26.25	1.72	30.62	-14.52	-1.01	7.27
Long (short) 10% highest (lowest) predicted returns							
Complete Case	17.47	13.15	1.33	22.41	4.94	0.94	6.19
Uncond. Mean	26.17	17.22	1.52	24.06	-2.11	0.06	12.63
Cond. Mean (GLS)	28.30	17.59	1.61	24.80	-3.50	-0.25	8.32
Cond. Mean (GLS / w. lags)	28.35	17.37	1.63	24.89	-3.46	-0.24	7.93
D. Regularized nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case (LASSO)	9.31	10.86	0.86	17.50	8.19	0.19	3.51
Uncond. Mean (LASSO)	35.69	25.58	1.40	28.05	-7.64	-0.51	6.40
Cond. Mean (GLS / LASSO)	40.94	27.80	1.47	28.33	-12.61	-1.09	6.23
Cond. Mean (GLS / LASSO / w. lags)	41.10	27.19	1.51	28.32	-12.78	-0.94	5.46
Long (short) 10% highest (lowest) predicted returns							
Complete Case (LASSO)	14.16	16.14	0.88	19.33	5.18	0.25	4.63
Uncond. Mean (LASSO)	23.88	17.34	1.38	23.06	-0.83	-0.36	8.91
Cond. Mean (GLS / LASSO)	25.85	19.14	1.35	23.57	-2.28	-0.68	7.30
Cond. Mean (GLS / LASSO / w. lags)	26.03	18.62	1.40	23.74	-2.29	-0.54	6.62

A.3.4 Out-of-sample predictions with industry dummies

In this section, we discuss results when we use the Fama-French 10 industries in the imputation model. In particular, the imputation model for a missing covariate $X_{it,k}$ is then

$$X_{it,k} = X_{it}^{(obs)\prime}\gamma_t + X_{it-1,k} + \sum_{q=2}^{10} \mathbf{1}_{it,q}\delta_q + u_{it,k}$$

where $X_{it}^{(obs)}$ are the observed covariates for observation i at time t and $\{\mathbf{1}_{it,q}\}_{q=2}^{10}$ are the industry dummies with $\mathbf{1}_{it,q} = 1$ if firm i belongs to industry q at time t and zero otherwise.

The out-of-sample prediction results are presented in Tables A.8 and A.9. Overall they are very similar to those without dummies in Section 4.2.

Table A.8: Performance Statistics For Out-of-Sample Predictions (all firms & industry specific imputation)

This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction. We use industry dummies in the imputation step. Portfolios are equally weighted. We differentiate between the complete case method, unconditional mean imputation, and conditional mean imputation with GLS weighting without and with lags. Long Pf. and Short Pf. denote the annualized average return of the long and short legs, respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and nonlinear model is outlined in Sections 4.2 and A.3.4. The sample period is 1990–2021.

	Mean (%)	Standard Deviation (%)	Sharpe Ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
A. Linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case	11.37	9.57	1.19	19.12	7.75	0.22	3.39
Uncond. Mean	48.91	29.46	1.66	39.87	-9.05	-0.12	10.38
Cond. Mean (GLS)	51.93	29.27	1.77	41.01	-10.93	-0.54	4.86
Cond. Mean (GLS / w. lags)	52.04	28.93	1.80	40.65	-11.39	-0.58	5.45
Long (short) 10% highest (lowest) predicted returns							
Complete Case	15.74	13.12	1.20	20.70	4.96	0.35	6.07
Uncond. Mean	32.11	19.44	1.65	30.94	-1.17	0.54	11.63
Cond. Mean (GLS)	32.93	19.77	1.67	30.89	-2.04	-0.21	5.97
Cond. Mean (GLS / w. lags)	33.56	19.72	1.70	31.18	-2.38	-0.30	6.07
B. Regularized linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case (LASSO)	12.47	10.63	1.17	18.96	6.49	-0.45	3.96
Uncond. Mean (LASSO)	47.49	28.22	1.68	39.29	-8.20	-0.01	7.04
Cond. Mean (GLS / LASSO)	50.74	28.77	1.76	40.83	-9.91	-0.45	3.68
Cond. Mean (GLS / LASSO / w. lags)	52.18	28.57	1.83	41.38	-10.80	-0.41	3.63
Long (short) 10% highest (lowest) predicted returns							
Complete Case (LASSO)	17.21	14.74	1.17	21.39	4.18	0.32	4.99
Uncond. Mean (LASSO)	31.12	20.25	1.54	30.47	-0.66	0.24	11.33
Cond. Mean (GLS / LASSO)	31.32	20.27	1.55	30.47	-0.85	-0.49	5.86
Cond. Mean (GLS / LASSO / w. lags)	31.93	19.91	1.60	30.74	-1.19	-0.30	5.20
C. Nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case	11.06	8.57	1.29	18.45	7.39	0.10	1.71
Uncond. Mean	85.53	35.05	2.44	65.46	-20.07	1.14	9.07
Cond. Mean (GLS)	92.56	32.97	2.81	67.64	-24.93	0.25	1.73
Cond. Mean (GLS / w. lags)	92.35	32.77	2.82	66.94	-25.41	0.25	1.86
Long (short) 10% highest (lowest) predicted returns							
Complete Case	17.47	13.15	1.33	22.41	4.94	0.94	6.19
Uncond. Mean	42.44	18.73	2.27	37.67	-4.77	0.60	7.86
Cond. Mean (GLS)	45.42	19.94	2.28	39.12	-6.30	-0.09	4.99
Cond. Mean (GLS / w. lags)	45.54	20.03	2.27	39.21	-6.33	0.10	5.28
D. Regularized nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case (LASSO)	9.31	10.86	0.86	17.50	8.19	0.19	3.51
Uncond. Mean (LASSO)	74.29	34.27	2.17	61.97	-12.32	0.99	7.43
Cond. Mean (GLS / LASSO)	84.00	37.07	2.27	64.59	-19.41	-0.39	3.76
Cond. Mean (GLS / LASSO / w. lags)	84.02	36.33	2.31	64.22	-19.79	-0.42	3.55
Long (short) 10% highest (lowest) predicted returns							
Complete Case (LASSO)	14.16	16.14	0.88	19.33	5.18	0.25	4.63
Uncond. Mean (LASSO)	38.98	19.29	2.02	36.12	-2.86	0.46	6.65
Cond. Mean (GLS / LASSO)	42.71	21.44	1.99	38.45	-4.26	-0.41	6.05
Cond. Mean (GLS / LASSO / w. lags)	43.13	21.66	1.99	38.57	-4.56	-0.46	6.08

Table A.9: Performance Statistics For Out-of-Sample Predictions (large firms & industry specific imputation)

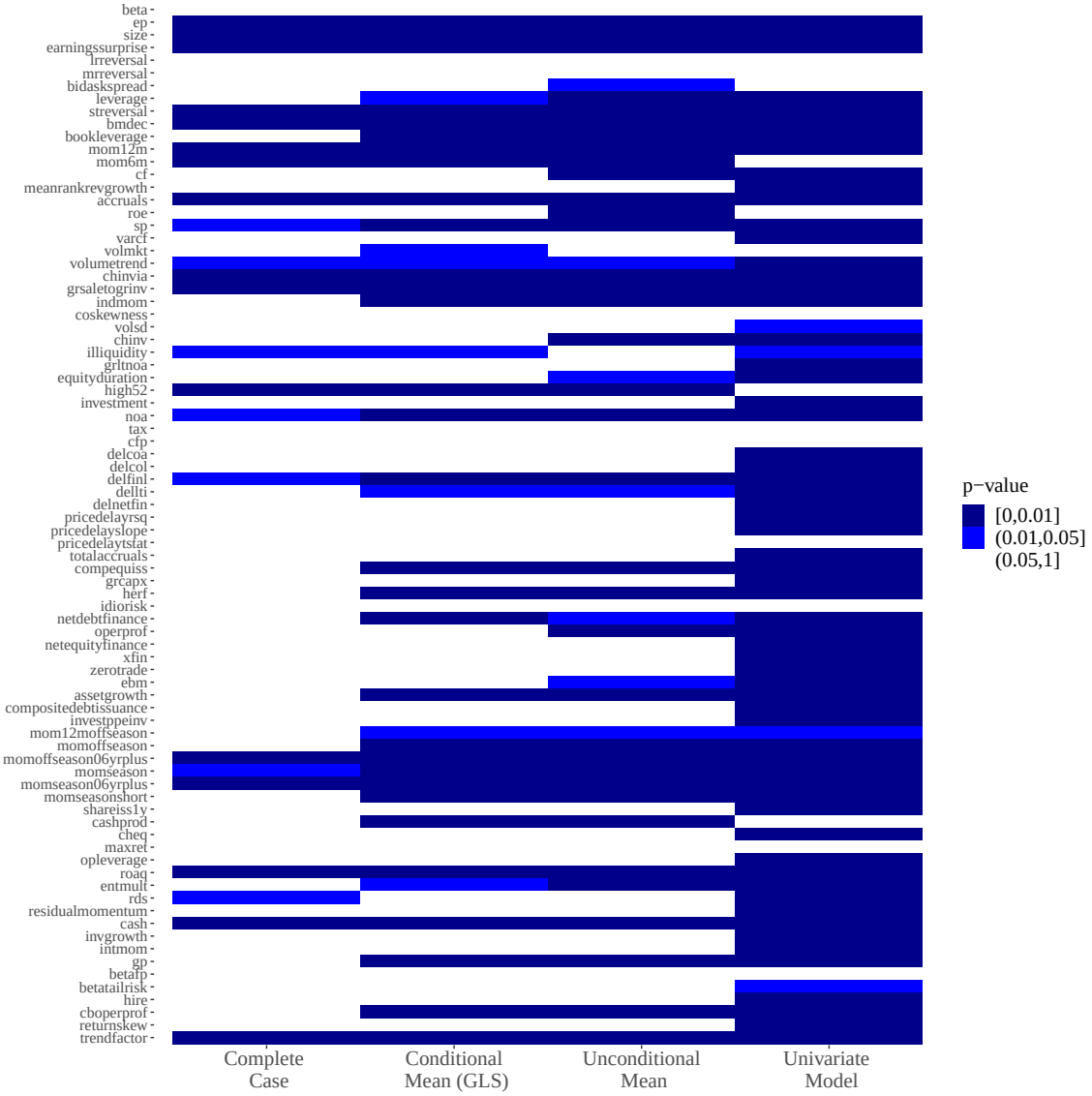
This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction for large firms. Specifically, we only include firms with a market capitalization greater than the 20th size percentile at the time of portfolio formation. We use industry dummies in the imputation step. Portfolios are equally weighted. We differentiate between the complete case method, unconditional mean imputation, and conditional mean imputation with GLS weighting without and with lags. Long Pf. and Short Pf. denote the annualized average return of the long and short legs, respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and nonlinear model is outlined in Sections 4.2 and A.3.4. The sample period is 1990–2021.

	Mean (%)	Standard Deviation (%)	Sharpe Ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
A. Linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case	11.37	9.57	1.19	19.12	7.75	0.22	3.39
Uncond. Mean	31.35	25.07	1.25	25.93	-5.42	0.01	15.64
Cond. Mean (GLS)	35.89	26.06	1.38	27.77	-8.12	-0.06	7.04
Cond. Mean (GLS / w. lags)	35.63	26.34	1.35	27.95	-7.68	-0.10	7.47
Long (short) 10% highest (lowest) predicted returns							
Complete Case	15.74	13.12	1.20	20.70	4.96	0.35	6.07
Uncond. Mean	22.92	18.13	1.26	23.37	0.45	0.58	14.27
Cond. Mean (GLS)	25.47	18.66	1.36	24.51	-0.96	0.27	7.43
Cond. Mean (GLS / w. lags)	25.56	18.66	1.37	24.50	-1.07	0.28	7.47
B. Regularized linear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case (LASSO)	12.47	10.63	1.17	18.96	6.49	-0.45	3.96
Uncond. Mean (LASSO)	31.63	24.79	1.28	27.98	-3.65	-0.30	9.79
Cond. Mean (GLS / LASSO)	33.87	25.35	1.34	26.81	-7.06	-0.55	5.18
Cond. Mean (GLS / LASSO / w. lags)	33.98	25.03	1.36	26.74	-7.24	-0.43	4.78
Long (short) 10% highest (lowest) predicted returns							
Complete Case (LASSO)	17.21	14.74	1.17	21.39	4.18	0.32	4.99
Uncond. Mean (LASSO)	21.92	18.72	1.17	23.00	1.08	0.12	12.79
Cond. Mean (GLS / LASSO)	23.67	18.97	1.25	23.80	0.13	0.03	6.09
Cond. Mean (GLS / LASSO / w. lags)	23.92	18.93	1.26	23.85	-0.07	0.07	6.06
C. Nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case	11.06	8.57	1.29	18.45	7.39	0.10	1.71
Uncond. Mean	42.41	25.14	1.69	29.35	-13.06	-0.28	7.09
Cond. Mean (GLS)	44.56	25.91	1.72	29.91	-14.65	-0.92	6.86
Cond. Mean (GLS / w. lags)	44.99	26.18	1.72	30.25	-14.73	-0.95	6.86
Long (short) 10% highest (lowest) predicted returns							
Complete Case	17.47	13.15	1.33	22.41	4.94	0.94	6.19
Uncond. Mean	26.17	17.22	1.52	24.06	-2.11	0.06	12.63
Cond. Mean (GLS)	27.92	17.70	1.58	24.72	-3.20	-0.25	8.56
Cond. Mean (GLS / w. lags)	28.21	17.66	1.60	24.70	-3.51	-0.26	8.53
D. Regularized nonlinear model							
Long (short) 100 highest (lowest) predicted returns							
Complete Case (LASSO)	9.31	10.86	0.86	17.50	8.19	0.19	3.51
Uncond. Mean (LASSO)	35.69	25.58	1.40	28.05	-7.64	-0.51	6.40
Cond. Mean (GLS / LASSO)	38.51	29.24	1.32	27.63	-10.89	-1.07	7.25
Cond. Mean (GLS / LASSO / w. lags)	42.22	27.80	1.52	27.95	-14.27	-0.93	4.93
Long (short) 10% highest (lowest) predicted returns							
Complete Case (LASSO)	14.16	16.14	0.88	19.33	5.18	0.25	4.63
Uncond. Mean (LASSO)	23.88	17.34	1.38	23.06	-0.83	-0.36	8.91
Cond. Mean (GLS / LASSO)	25.56	18.73	1.36	23.70	-1.86	-0.56	7.24
Cond. Mean (GLS / LASSO / w. lags)	25.98	18.98	1.37	23.53	-2.45	-0.67	7.60

A.3.5 Incremental information: Univariate model

Figure A.4: Incremental Information - Comparison to univariate model

This figure illustrates which characteristics are significant in the growing model described in equation (4) and in a univariate model. The p-values are not adjusted for the false discovery rate because we assume that a researcher is only interested in testing the effect of the newly introduced characteristics at the point in time the new characteristic was discovered so that a multiple testing problem does not arise. As in the main text, the growing model is estimated using the complete case, unconditional mean imputation and conditional mean imputation with weights. The univariate model is estimated on the complete case. In the univariate model, we only consider the complete case because the estimators for the parameter of interest β_1, the slope coefficient, are numerically equivalent for the complete case estimator and the imputation based estimators that do not use weights (the weighted estimator yields almost identical results). The characteristics are ordered according to their year of discovery.



A.4 Extensions

A.4.1 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

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

for all $t = 1, \ldots, T$, where M_{t+1} is the stochastic discount factor, R^e_{t+1} is a vector of n excess returns, and $X_t = (X'_{1t}, \ldots, X'_{nt})'$ where X_{it} are variables known at time t for asset i with $i = 1, 2, \ldots, n$. The discount factor is a portfolio, where the weights are the weights of a mean-variance efficient portfolio (other than the risk-free mimicking portfolio).[1] In addition, similar to Kozak, Nagel, and Santosh (2020), we assume that the weights are a parametric function of $X_{it} \in \mathbb{R}^K$. That is

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

where

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

Combining the previous three equations we get

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

As before, assume we have $L+1$ missing patterns. Let $D_t = l$ for missing pattern l, and let $X_t^{(l)}$ be the corresponding subset of observed elements of X_t. Then, under an analogous MAR assumption as before,

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

[1] If the SDF is not in the span of excess returns, we can work with the projection of the SDF onto the span of excess returns instead.

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

Let

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

Assuming that

$$E[X_{jt,k}R^e_{jt+1}R^e_{t+1} \mid X^{(l)}_t, D_t = l] = E[X_{jt,k}R^e_{jt+1}R^e_{t+1} \mid X^{(l)}_t, D_t = 0]$$

we obtain the conditional moment restrictions

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

To impute missing values, let

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

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

$$E\left[\mathbf{1}(D_t = 0)\left(R^e_{t+1} - \sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k X_{jt,k}R^e_{jt+1}R^e_{t+1}\right)\right)g(X^{(0)}_t)\right] = 0 \tag{A.1}$$

$$E\left[\mathbf{1}(D_t = l)\left(R^e_{t+1} - \sum_{j=1}^{n}\left(\sum_{k=1}^{K}\beta_k Z^{(l)}_{t,jk}\right)\right)g(X^{(l)}_t)\right] = 0 \qquad l = 1, \ldots, L \tag{A.2}$$

$$E\left[\mathbf{1}(D_t = 0)\left(X_{jt,k}R^e_{jt+1}R^e_{t+1} - h\left(X^{(l)}_t, \gamma^{(l,k)}\right)\right)g(X^{(l)}_t)\right] = 0 \qquad l = 1, \ldots, L \tag{A.3}$$
$$k \notin I^{(l)}_t$$

A.4.2 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 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 an observed covariate. In this case, we get

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

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.5 Comparison to the EM-algorithm

We briefly compare the proposed method to the Expectation-Maximization (EM) algorithm. Recall the setup of our simple example:

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

where $X_{i,2}$ is not observed for some subset of the data, while $X_{i,1}$ and Y_i are always observed. For notational convenience, we assume that $X_{i,2}$ is observed for $i = 1, \ldots, r$ and missing for $i = r+1, \ldots, n$ with $r < n$. Define $D_i = 0$ for $i = 1, \ldots, r$ and $D_i = 1$ for $i = r+1, \ldots, n$. To employ the EM algorithm, we have to make some distributional assumptions:

$$\begin{aligned}
\epsilon_i | X_i &\sim N(0, \sigma^2) \\
X_i &\sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right) = N(\mu, \Sigma).
\end{aligned}$$

By joint normality we know that

$$X_{i,2} | X_{i,1} \sim N(\gamma_0 + X_{i,1}\gamma_1, \sigma_X^2),$$

where $\gamma_0 = \mu_2 - \frac{\sigma_{12}}{\sigma_1^2}\mu_1$ and $\gamma_1 = \frac{\sigma_{12}}{\sigma_1^2}$, γ is the least squares estimator, and $\sigma_X^2 = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}$.

We will now discuss the EM algorithm. It maximizes the expectation of the complete data likelihood ($l_n(\beta; Y, X)$), the likelihood we would observe if $X_{i,2}$ was fully observed, conditional on the observed variables and some estimate of parameters of interest θ. θ contains β and other parameters that are required during the estimation process. We denote the observed variables for observation i with $X_i^{(obs)}$, that is, $X_i^{(obs)} = (Y_i, X_{i,1}, X_{i,2})$ for $i \leq r$ and $X_i^{(obs)} = (Y_i, X_{i,1})$ for $i > r$. Starting with some $\theta^{(0)}$, the EM-algorithm iterates through the following procedure updating θ in each iteration. In the k-th iteration we derive (*expectation step*)

$$E\left[l_n(\beta; Y, X)|X^{(obs)}, \theta = \theta^{(k)}\right]$$

to then maximize it with respect to θ (*maximization step*). By the distributional assumptions we know that above is maximized if we maximize

$$\begin{aligned}
& -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} E\left[\frac{1}{2\sigma^2}\left(Y_i - \beta_0 - X_{i,1}\beta_1 - X_{i,2}\beta_2\right)^2 \Big| X_i^{(obs)}, \theta = \theta^{(k)}\right] \\
= & -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} E\left[\frac{1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2 \Big| X_i^{(obs)}, \theta = \theta^{(k)}\right] \\
= & -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{r}\frac{1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2 - \sum_{i=r+1}^{n} E\left[\frac{1}{2\sigma^2}\left(Y_i - X_i'\beta\right)^2 \Big| X_{i,1}, Y_i, \theta = \theta^{(k)}\right]
\end{aligned}$$

Under standard regularity conditions, the interchangeability of integral and differentiation operator, and following standard arguments, we get

$$\begin{aligned}
\hat{\beta}^{(k+1)} = & \left(\sum_{i=1}^{r} X_i X_i' + \sum_{i=r+1}^{n} E\left[X_i X_i' \Big| X_{i,1}, Y_i, \theta = \theta^{(k)}\right]\right)^{-1} \\
& \times \left(\sum_{i=1}^{r} X_i Y_i + \sum_{i=r+1}^{n} E\left[X_i Y_i \Big| X_{i,1}, Y_i, \theta = \theta^{(k)}\right]\right)
\end{aligned}$$

that is, in each iteration the EM algorithm imputes $X_{i,2}$ with $E[X_{i,2}|X_{i,1}, Y_i, \theta = \theta^{(k)}]$ and $X_{i,2}^2$ with $E[X_{i,2}^2|X_{i,1}, Y_i, \theta = \theta^{(k)}]$. It is now also clear that apart from β θ contains the parameters that characterize these conditional means.

For the EM algorithm to lead to valid inference, we make the MAR assumption $D_i \perp\!\!\!\perp X_{i,2}|X_{i,1}, Y_i$.

Then

$$E[X_{i,2}|X_{i,1}, Y_i, D_i, \theta = \theta^{(k)}] = E[X_{i,2}|X_{i,1}, Y_i, \theta = \theta^{(k)}]$$

$$E[X_{i,2}^2|X_{i,1}, Y_i, D_i, \theta = \theta^{(k)}] = E[X_{i,2}^2|X_{i,1}, Y_i, \theta = \theta^{(k)}]$$

If we do not make this assumption, the missing mechanism needs to be modelled explicitly to incorporate that $X_{i,2}$ is missing not at random.

Our estimator is similar to the EM algorithm in that we impute $X_{i,2}$ with a conditional mean, but we only condition on observed covariates and not on the outcome. Moreover, we do not explicitly model the conditional mean of $X_{i,2}^2$. Both are reflected in the asymptotic variance of our estimator, which is larger than the estimator estimated using the EM algorithm. However, for an arbitrarily chosen distribution doing asymptotics with the EM algorithm is not straightforward, whilst the asymptotic distribution of our estimator is readily available and does not require any distributional assumptions.

Chen and McCoy (forthcoming) use an EM-algorithm that only assumes that the covariates are jointly normally distributed. They then impute missing values of covariates with the estimated conditional mean, which only conditions on observed covariates for that observation, and estimate the regression parameters by OLS. An advantage of a joint treatment of the outcome variable Y_i and the covariates (as in the EM algorithm above or as in our estimation method) is that it allows obtaining the statistical properties and valid standard errors of the parameters of interest.

A.6 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 projection, 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 if $E[\varepsilon_{it} \mid X_{it}^{(0)}, D_{it} = 0] = 0$. For the moment conditions in equation (2), we 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 if

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

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

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

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 our 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.7 Equivalence GLS and Optimal GMM

Consider the moment conditions

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

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

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

where

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

To show the equivalence of the GLS and the optimal GMM estimator, we impose A.8.3 stated in Section A.8.1.

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 $E[g_{it}(\beta_t)]'WE[g_{it}(\beta_t)]$.

The efficient weighting 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 elements 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 zero because

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

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

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

We require that $\hat{\sigma}^2_{\varepsilon,t} \xrightarrow{p} \sigma^2_{\varepsilon,t}$, $\hat{\beta}^{(l)}_t \xrightarrow{p} \beta^{(l)}_t$ and $\hat{\Sigma}^{(l)}_t \xrightarrow{p} \Sigma^{(l)}_t$, 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

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

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

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

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

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

The first element is simply

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

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 $J_t^{(l)} = \left|(I_t^{(l)})^c\right|$ and

$$\gamma_t^{(l)} = \left(\gamma_t^{(l,1)}, \ldots, \gamma_t^{(l,J_t^{(l)})}\right)'$$

Then for the l-th element

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

It follows that

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

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

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

Then

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

Using the same arguments we can also write

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

Hence

$$\hat{\beta}_{t,GMM} = \left(\sum_{i=1}^{n} \sum_{l=0}^{L} \mathbf{1}(D_{it} = l) \frac{Z_{it}^{(l)} Z_{it}^{(l)'}}{(\hat{\sigma}_t^{(l)})^2} \right)^{-1} \sum_{i=1}^{n} \sum_{l=0}^{L} \mathbf{1}(D_{it} = l) \frac{Z_{it}^{(l)} Y_{it}}{(\hat{\sigma}_t^{(l)})^2}$$

Next, consider the GLS estimator, which minimizes

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

The first-order conditions are

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \sum_{l=0}^{L} \mathbf{1}(D_{it} = l) \frac{Z_{it}^{(l)} Y_{it} - Z_{it}^{(l)} Z_{it}^{(l)'} \hat{\beta}_{t,GLS}}{(\hat{\sigma}_t^{(l)})^2} \\ \Leftrightarrow \quad \hat{\beta}_{t,GLS} &= \left(\sum_{i=1}^{n} \sum_{l=0}^{L} \mathbf{1}(D_{it} = l) \frac{Z_{it}^{(l)} Z_{it}^{(l)'}}{(\hat{\sigma}_t^{(l)})^2} \right)^{-1} \sum_{i=1}^{n} \sum_{l=0}^{L} \mathbf{1}(D_{it} = l) \frac{Z_{it}^{(l)} Y_{it}}{(\hat{\sigma}_t^{(l)})^2} \end{aligned}$$

Therefore

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

A.8 Regularity Conditions and Asymptotic Results

We collect all regularity conditions here. Some of them are already stated explicitly in the main text or the relevant sections of the appendix. After stating the assumptions, we derive the large sample distribution.

A.8.1 Assumptions

The first two assumptions are also stated in the main text.

Assumption A.8.1 (Mean independence)**.** *For all* $l = 0, \ldots, L$

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

Assumption A.8.2 (Missing at random)**.** *For all* $l = 1, \ldots, L$

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

The following assumption is used to show equivalence between the GLS and the optimal GMM estimator.

Assumption A.8.3 (GLS equivalence)**.** *Let* $I_t^{(l)}$ *be the set of indices of covariates that are observed in missing pattern* l *at time* t*.*

1. $\gamma_t = \left\{ \{\gamma_t^{(l,k)}\}_{k \notin I_t^{(l)}} \right\}_{l=1,\ldots,L}$ *is known*
2. $E[\epsilon_{it} \mid X_{it}, D_{it} = l] = 0$ *for all* $l = 1, \ldots, L$
3. $E[\epsilon_{it}^2 \mid X_{it}, D_{it} = l] = \sigma_{\epsilon,t}^2$ *for all* $l = 0, \ldots, L$
4. $E[u_{it}^{(l)} u_{it}^{(l)\prime} \mid X_{it}^{(l)}, D_{it} = l] = \Sigma_t^{(l)}$ *for all* $l = 1, \ldots, L$ *where* $u_{it,k}^{(l)} = X_{it,k} - X_{it}^{(l)\prime} \gamma_t^{(l,k)}$ *for all* $k \notin I_t^{(l)}$

where all expectations are assumed to exist.

The next assumption (along with the linear conditional mean model and MAR) implies that $\hat{\gamma}_t^{(l,k)} \xrightarrow{p} \gamma_t^{(l,k)}$ for all l and $k \notin I_t^{(l)}$.

Assumption A.8.4 (Consistency of $\hat{\gamma}_t$)**.** *For each $l = 1, \ldots, L$ and $k \notin I_t^{(l)}$*

1. $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(D_{it} = 0) X_{it}^{(l)} X_{it}^{(l)\prime} \xrightarrow{p} E[\mathbf{1}(D_{it} = 0) X_{it}^{(l)} X_{it}^{(l)\prime}]$
2. $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(D_{it} = 0) X_{it}^{(l)} u_{it}^{(l)} \xrightarrow{p} E[\mathbf{1}(D_{it} = 0) X_{it}^{(l)} u_{it}^{(l)}]$

where all expectations are assumed to exist and $u_{it}^{(l)} = \{X_{it,k} - X_{it}^{(l)\prime}\gamma_t^{(l,k)}\}_{k \notin I_t^{(l)}}$.

The next assumption implies that $\hat{\beta}_t \xrightarrow{p} \beta_t$. Recall that γ_t, $\bar{g}_1(\beta_t, \gamma_t)$ and $g_{it,1}(\beta_t, \gamma_t)$ are defined at the beginning of Section A.8.2.

Assumption A.8.5 (Consistency of $\hat{\beta}_t$)**.**

1. $\frac{\partial}{\partial \beta_t}\bar{g}_1(\beta_t, \hat{\gamma}_t) \xrightarrow{p} E\left[\frac{\partial}{\partial \beta_t} g_{it,1}(\beta_t, \gamma_t)\right]$
2. $\bar{g}_1(\beta_t, \hat{\gamma}_t) \xrightarrow{p} E[g_{it,1}(\beta_t, \gamma_t)]$
3. $\hat{W} \xrightarrow{p} W$ *where W ($\hat{W}$) is the (sample) GLS weight matrix as described in section A.6.*
4. $E\left[\frac{\partial}{\partial \beta_t} g_{it,1}(\beta_t, \gamma_t)\right]' W E\left[\frac{\partial}{\partial \beta_t} g_{it,1}(\beta_t, \gamma_t)\right]$ *is invertible*

where all expectations are assumed to exist.

The final assumption is used to derive the large sample distribution of the GLS estimator. Notice that we define γ_t, $\bar{g}_1(\beta_t, \gamma_t)$, $g_{it,1}(\beta_t, \gamma_t)$, $\bar{g}_2(\gamma_t)$, $g_{it,2}(\gamma_t)$, D_t and Q_t in Section A.8.2.

Assumption A.8.6 (Large sample distribution and overidentification test)**.**

1. $\frac{\partial}{\partial \beta_t}\bar{g}_1(\beta_t, \hat{\gamma}_t) \xrightarrow{p} E\left[\frac{\partial}{\partial \beta_t} g_{it,1}(\beta_t, \gamma_t)\right]$
2. $\frac{\partial}{\partial \gamma_t}\bar{g}_1(\hat{\beta}_t, \gamma_t) \xrightarrow{p} E\left[\frac{\partial}{\partial \gamma_t} g_{it,1}(\beta_t, \gamma_t)\right]$
3. $\frac{\partial}{\partial \gamma_t}\bar{g}_2(\gamma_t) \xrightarrow{p} E\left[\frac{\partial}{\partial \gamma_t} g_{it,2}(\gamma_t)\right]$
4. $$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \begin{pmatrix} g_{it,1}(\beta_t, \gamma_t) \\ g_{it,2}(\gamma_t) \end{pmatrix} \xrightarrow{d} N(0, \Omega_t)$$

where

$$\Omega_t = \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{i=1}^{n} \begin{pmatrix} g_{it,1}(\beta_t, \gamma_t) \\ g_{it,2}(\gamma_t) \end{pmatrix} \sum_{i=1}^{n} \begin{pmatrix} g_{it,1}(\beta_t, \gamma_t) \\ g_{it,2}(\gamma_t) \end{pmatrix}'\right]$$

is finite and positive definite. A sample analog estimator $\hat{\Omega}_t$ *is consistent.*

5. $D_t' Q_t$ *is invertible*

where all expectations are assumed to exists.

A.8.2 Large Sample Distribution

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

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

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

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

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.7, where $W_1 = \text{diag}\left(\hat{w}^{(l)}\right)^{-1}$ and expressions for $\hat{w}^{(l)}$ are provided in Section A.7. We obtain the two-step OLS estimator when W_1 is an identity matrix. Using a first-order Taylor expansion, we get

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

$$+ o_p(1/\sqrt{n})$$

or

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

We know that

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

where Ω_t is defined in Assumption A.8.6 in Section A.8.1 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

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

with

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

and

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

where $V_{it} = (Y_{it}, X_{it}')$. 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.8.3 J-test

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

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

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

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

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

To test our overidentifying restrictions, we test

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

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

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

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

Hence,

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

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

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

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 $\hat{\Xi}_t$ 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

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

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

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

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