WORKING PAPER · NO. 2022-168

Missing Data in Asset Pricing Panels

Joachim Freyberger, Björn Höppner, Andreas Neuhierl, and Michael Weber

DECEMBER 2022

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Missing Data in Asset Pricing Panels*

Joachim Freyberger† Björn Höppner‡

Andreas Neuhierl§ Michael Weber¶

First draft: September 2021
This draft: November 2022

Abstract

Missing data for return predictors is a common problem in cross sectional asset pricing. Most papers do not explicitly discuss how they deal with missing data but conventional treatments focus on the subset of firms with no missing data for any predictor or impute the unconditional mean. Both methods have undesirable properties - they are either inefficient or lead to biased estimators and incorrect inference. We propose a simple and computationally attractive alternative 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 non-linear and high-dimensional settings. In Monte Carlo simulations, we find that it performs almost as well as the efficient but computationally costly GMM estimator in many cases. We apply our procedure to a large panel of return predictors and find that it leads to improved out-of-sample predictability.

JEL classification: C14, C58, G12

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

*We thank 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, Markus Pelger, Christoph Rothe, Oleg Rytchkov, Jan Scherer, Gustavo Schwenkler, Takuya Ura, Guofu Zhou and conference and seminar participants at the University of Bonn, University of Mannheim, 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, Temple University, TU Muenchen, Queen Mary University, Yale University, 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. Weber also gratefully acknowledges financial support from the University of Chicago Booth School of Business, the Fama Research Fund at the University of Chicago Booth School of Business, and the Fama-Miller Center.

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

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

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

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

1 Introduction

Missing data is a common problem in cross-sectional asset pricing studies. While the problem of missing return observations has received some attention and is typically handled by the use of so-called delisting returns (Shumway (1997), Beaver et al. (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 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 works by replacing the missing covariates with suitable estimates and accounting for the estimation error (from generating these estimates) in the subsequent analysis. In addition, we also "down-weigh" the observations for which we imputed data, thereby adjusting 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 of interest is parametric or nonparametric and does not require us to specify the entire distribution of the missing covariates. We show that our proposed method can be cast 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 Cochrane (2011)'s 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 et al., 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 (2021) with 82 covariates, which 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 et al. (2005) call "one of statistics' first principles" – "thou shall not throw data away". Moreover, the complete case approach has an additional drawback that may be overlooked at first sight. By conditioning on firms for which all covariates are available, we might inadvertently ignore an interesting part of the *return distribution*, which might preclude us from forming portfolios with high out of sample Sharpe ratios.

In an attempt not to delete too many observations, some researchers replace missing values of the covariates with their cross-sectional mean (unconditional mean imputation) of that period. We wholeheartedly agree with the aim of using as many return observations as possible. However, we also show that 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 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. 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, once we condition on observed 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 sample of realistic size. The simulations also help illustrate when the ad-hoc approaches, such as unconditional mean imputation and complete case analysis are (and are not) problematic.

Finally, we apply our method to the CRSP/Compustat sample. We document that 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 due to the generically biased estimators and artificially small standard errors.

1.1 Related Literature

The problem of missing data is ubiquitous in empirical analyses. For example, clinical trials routinely have to confront the problem that some patients do not show up for follow-up examinations. A related problem occurs in surveys, where respondents often leave questions blank, sometimes by accident and at other times because they feel uncomfortable answering 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 is missing – the so called missing mechanism.[1] If the probability that a particular observation is missing depends on the outcome variable (even after conditioning on observables), 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 does not depend on the outcome variable itself, but may depend on observed covariates, the literature has proposed several general approaches to deal with missing data. Some of these approaches rely on strong distributional assumptions on unobservables (for example, likelihood-based approaches and Bayesian methods) that we do not want to impose to computational reasons. 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). Just like we do, some of these approaches also down-weigh observations with missing values, but these studies typically only allow for one missing pattern, which means that either all variables are observed or one particular subset of the variables is missing. We extend these ideas (specifically the weighting approach of

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

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

Dagenais (1973)) and allow for general missing patterns.

One challenge that arises with imputation methods is how to account for the "imputation uncertainty" in inference, because the imputations are estimates themselves. 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 model 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 et al. (2008) present a semi-parametrically 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 et al. (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, re-weighting the complete case sample such that it more closely mirrors the population (Robins et al. (1994), Wooldridge (2007)), in which case we typically need to model the probability that a particular case is observed. The IPW approach relaxes important assumptions relative to the (unweighted) complete case, but does not use all available data. A considerable generalization is the class of augmented IPW (AIPW) estimators, which use the whole sample. Under certain assumptions, which differ slightly from our setup, Robins et al. (1994) show that the AIPW estimator is semiparameteric efficient. However, similar to the optimal GMM estimator, the efficient AIPW estimator is

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

generally not available in closed form and computationally prohibitive in our application. For comprehensive results on AIPW estimators see for example Tsiatis and Davidian (2015).

While most papers do not explicitly state how they treat missing data, using only the complete case appears to be the most common approach in asset pricing studies. Recent examples include Lewellen (2015), Freyberger et al. (2020), Kelly et al. (2019), and Kim et al. (2021). Other papers, follow a special imputation approach and replace the missing covariate values with the cross-sectional mean or median, see e.g. Light et al. (2017), Kozak et al. (2020), Gu et al. (2020).

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 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 the imputation step and consider estimation with the imputed sample in a separate step. Bryzgalova et al. (2022) assume a latent factor model for the 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. Their method mainly focuses on imputation and does not discuss potential adjustments for subsequent estimation and inference. Chen and McCoy (2022) 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. In Section A.4 we briefly describe an alternative EM-algorithm, which would allow for valid inference after imputation, but still relies on assuming normality. Beckmeyer and Wiedmann (2022) use a machine learning algorithm borrowed from natural language processing to impute values, but it is unclear what exactly the required assumptions and theoretical properties are. 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 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 et al. (2021), Jin et al. (2021), and Xiong and Pelger (2022). Lastly, Harvey et al. (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.

2 Data

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

In order to avoid potential lock-ahead biases, we lag all characteristics that build on Compustat annual by at least six months and all that build on Compustat quarterly by at least four months. Our main dataset is obtained from Chen and Zimmermann (2021) and consists of 82 firm characteristics that are available from 1978 - 2021. The firm characteristics feature a combination of accounting information as well as versions of momentum and functions of 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 dataset. 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 characteristics that are standard 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 dataset 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.

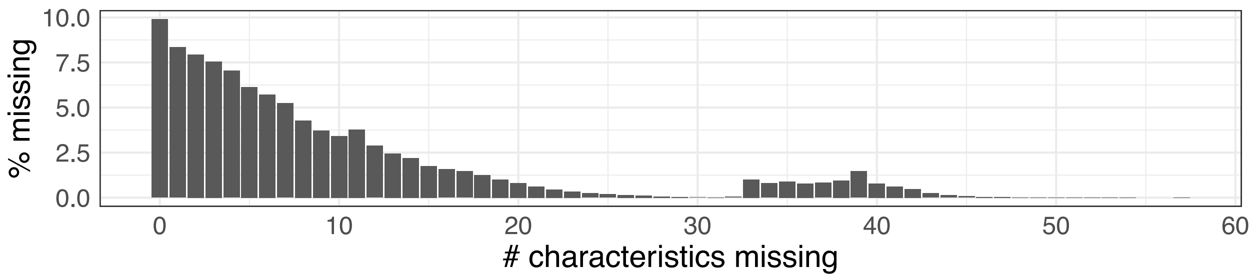
Appendix Table A.1 also shows the fraction of missing values per characteristic. Overall, we have a total of 3,644,484 firm-month observations. Fama and French (1992) define the benchmark for empirical analyses of the cross-section of expected returns. We follow them and require that a minimum of information is available for each firm. As Fama and French, we require the inputs (market beta, size, and book-to-market) of the Fama-French 3 factor model to be available for all firms. When we condition on firms having `Beta`, `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 3, in our proposed method, we essentially assume that the data is 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 assumptions are 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 et al. (2020) such that the continuous characteristics are uniformly distributed on $[0, 1]$, a standard transformation, which is also applied in Kozak et al. (2020), Gu et al. (2020) and many other papers.

2.1 Missing data in CRSP/Compustat

In this section, we provide descriptive statistics to document the prevalence of the missing data problem in a standard dataset for empirical asset pricing but similar conclusions also hold for many studies in empirical corporate finance or international finannce. 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 all observations. Moreover, while many observations are only missing few predictors, a non-trivial 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 one half to two-thirds 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, e.g. about 10% of the observations have no missing characteristic and about 7.5% 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 dataset is obtained from Chen and Zimmermann (2021) and consists of 82 firm characteristics that are available from 1978 - 2021.



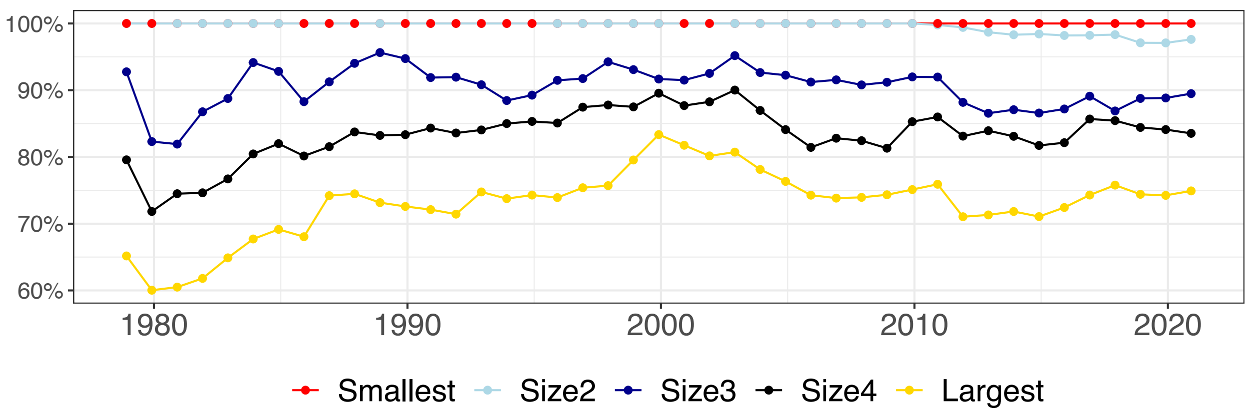
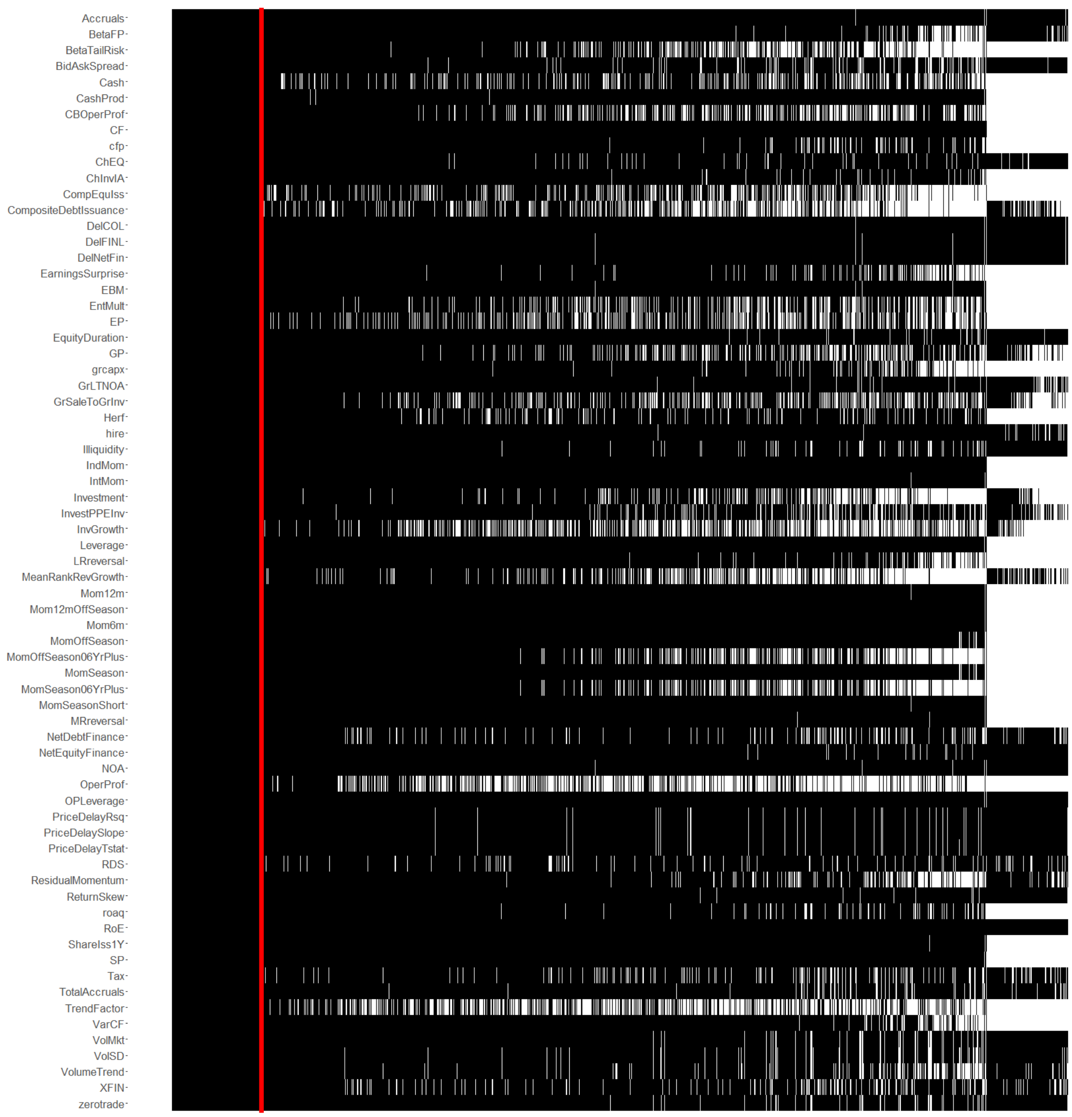


Figure 2 shows the missing (and non-missing) observations for a random sample of approximately 5% of all observations with black indicating an observed observation and white indicating missing data. The roughly 10% of black observations to the left of the red vertical line represent the complete case. 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 dataset is obtained from Chen and Zimmermann (2021) and consists of 82 firm characteristics that are available from 1978 - 2021.

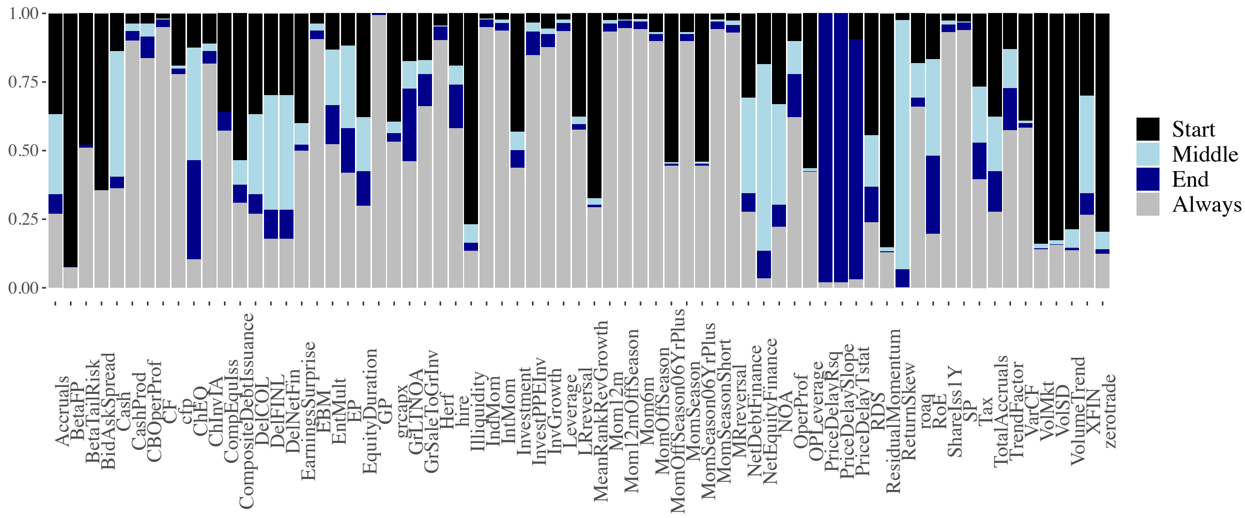


firms have been listed for a while to solve the missing data problem. However, Figure 3 illustrates that this is not the case. In particular, for almost all characteristics we see that they may be missing at the start, that is, when a firm first enters the sample, in the middle, that is, the characteristics was observed when the firm entered the sample and then was no longer observed, but it was observed again later.

It is also possible that a firm characteristic has good availability for most of the sample, but is missing towards 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 due to data requirements, e.g. 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, this might be due to 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

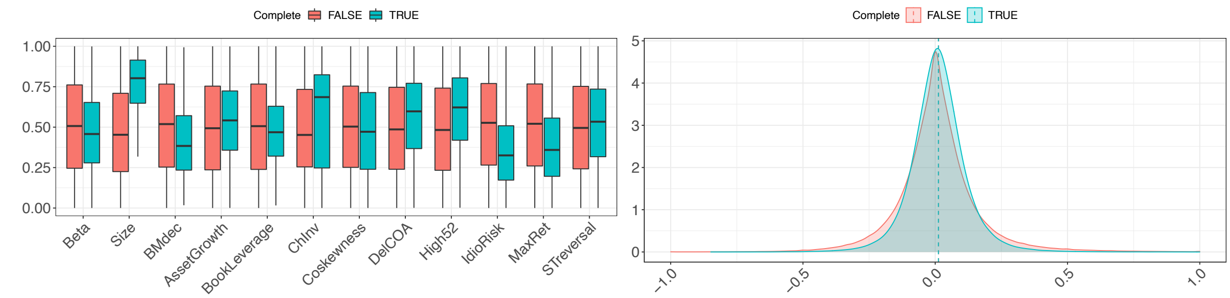
This figure depicts the prevalence of different missingness types. Given a time series of a single firm, a characteristic may be missing at the beginning of the time series but is observed after some point (*Start*), a characteristic may be observed at the beginning of the time series, then it is missing 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.



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. We illustrate this point in Figure 4. 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, e.g. `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 show below.

Figure 4: Complete vs. Incomplete Observations

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



3 Model

In summary, many characteristics are missing in the standard CRSP/Compustat panel such that simply ignoring the problem is possibly 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.

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

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

Since we also assume that $E[\varepsilon_i \mid X_i] = 0$, a sufficient condition for this assumption is that ε_i is independent of D_i conditional on X_i. 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 $X_{i,2}$.

In the full model, we allow D_i to depend on all variables that are always observed. In particular, in our sample we always observe 13 firm characteristics, including size, book-to-market, beta, idiosyncratic risk, and the return of the previous month, and the probability that an observation is incomplete can be a function these characteristics (see Section 2 for a detailed description of the data and a full list of characteristics). For example, smaller firms may be more likely to have missing values. However, conditional on all of these characteristics, we essentially assume that the data is missing at random. While these assumptions are not directly testable, as explained below, we can test the implications of the assumptions that we use to construct our estimator.

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

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

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

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

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 in the Online Appendix for a derivation of the test statistic in the general model and Section 3.2 for the test results).

We want to stress that the assumption that $E\left[X_{i,2} \mid X_{i,1}, D_i = 0\right]$ is a linear function is not required to derive our unconditional moment conditions. To avoid it, we can use an alternative derivation based on projections, which is less intuitive and discussed in Section A.5 in the Online 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 will now explain that the third option is an appealing alternative, which is easy to implement and has very good finite sample properties in our simulations.

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

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

and the denominators of the two fractions can be replaced with consistent estimators. We prove this equivalence in a more general setting in Online Appendix A.6. 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 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 worse statistical properties than the complete case estimator, even when a substantial subset of the data contains missing values. These results are in line with Gourieroux and Monfort (1981) and Nijman and Palm (1988) who find that the GLS estimator is more efficient than the OLS estimator in the presence of one missing pattern.

Finally, a popular approach is to impute the unconditional mean instead of the conditional mean and then estimate $(\beta_0, \beta_1, \beta_2)$ by OLS. Such an approach 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 0 unless in the special cases when $\beta_2 = 0$ or when $X_{i,2}$ is mean independent of $X_{i,1}$. But, even when one of these conditions is close to being satisfied, unconditional mean imputation is still unreliable for inference even if it may yield good out-of-sample predictions.

3.2 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 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, but similar to the simple example, 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

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

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 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 $l = 1, \dots, L$ in which case

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

As discussed above, these assumptions allow D_{it} to depend on regressors that are always observed, and since we always observe 13 important firm characteristics, these assumptions seems to be reasonable in our empirical application (see Section 2 for further discussions).[4]

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

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

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 3.3.2.

We can now collect our conditional moments and transform them to unconditional mo-

[4] We can relax these assumptions by conditioning on additional characteristics that D_{it} may depend on, such as industry dummies (see Sections 3.3.2 for more details).

ments to obtain:

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

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

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

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 Online Appendix A.8. 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]

[5]Incorporating lagged characteristics in the model to estimate $\gamma_t^{(l,k)}$ is straightforward and we will discuss this point in Section 3.3.2. We omit it here to keep notation simple.

2. Estimate $Var(Y_{it} - \sum_{k \in I_t^{(l)}} X_{it,k}\beta_{t,k} - \sum_{k \notin I_t^{(l)}}^{K} X_{it}^{(l)\prime}\gamma_t^{(l,k)}\beta_{t,k} \mid D_{it} = l)$ 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.

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

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

We derive the large sample distribution of the estimator in Online Appendix A.7 and provide plug-in estimators for the standard errors. A potential alternative is the optimal GMM estimator, which can be hard to compute in practice because the objective function is not quadratic in the parameters. In fact, in our empirical application with large numbers of observations and regressors, this estimator is computationally infeasible.

3.3 Extensions

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

3.3.1 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 number large of regressors, we could also use a penalized estimator such as the LASSO estimator or the Ridge estimator.

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

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

As before, $X_{i,1}$ is always observed, but $X_{i,2}$ is not, and $D_i = 1$ denotes the case 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 resulting estimator is inconsistent. These issues carry over to other nonlinear models, which implies that imputations should depend on the model being estimated, which is what we propose here compared to contemporaneous papers that largely focus on imputing characteristics without conditioning on the ultimate model being estimated.

One possibility to allow for nonlinearities and models selection simultaneously, which we use in our application, is the group LASSO estimator of Freyberger et al. (2020). Similar to the simple example above, in the first step 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 et al. (2020), with the possibility of down-weighting observations with imputed values. This approach not only allows for nonlinearities but also pre-specified interactions.

3.3.2 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 Online Appendix A.3.

Consider again the simple model

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

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

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

and

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

and a sufficient condition for these assumptions is

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

Let V_i be an additional vector of covariates that is always observed, such as industry dummies, which do not have a direct effect on the outcomes, 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 3.2 (that is, we do not need inverse propensity score weighting).

4 Simulations

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

4.1 Low-dimensional setting

We start with the model

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

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

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

Table 1 shows coverage rates 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 ran $1,000$ Monte Carlo simulations. 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, for all coefficients the average length of the confidence intervals decreases substantially for the optimal GMM estimator. 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. the unweighted estimator). In fact, the average length of the confidence intervals of the OLS estimator can be larger than those of the complete case estimator. While the OLS estimator uses more moment conditions, it combines them in an inefficient way. All of these four estimators are valid and therefore have coverage probabilities close to 90%. When the fraction of complete observations is low,

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 third covariate ($X_{i,3}$), another part of the data for the fifth covariate ($X_{i,3}$) and another part of the data for the fourth and the fifth covariate ($X_{i,4}$, $X_{i,5}$).

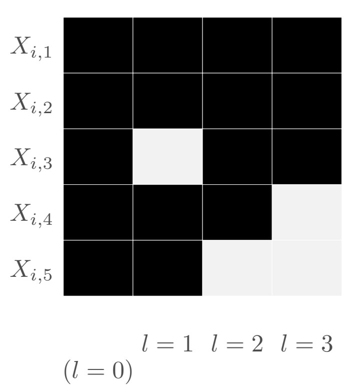


Table 1: Simulation - Coverage and Length of Confidence Intervals for Varying Missing Percentage

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.

	Complete Case		Optimal GMM		Imputation GLS		Imputation OLS		Uncond. Mean	
	Cover	Length	Cover	Length	Cover	Length	Cover	Length	Cover	Length
	25% complete									
β_1	0.904	0.208	0.881	0.141	0.876	0.149	0.902	0.204	0.758	0.173
β_2	0.892	0.475	0.872	0.366	0.881	0.397	0.894	0.536	0.000	0.316
β_3	0.891	0.642	0.879	0.563	0.884	0.605	0.895	0.792	0.004	0.332
β_4	0.883	0.642	0.893	0.501	0.896	0.519	0.915	0.610	0.000	0.365
β_5	0.906	0.478	0.887	0.348	0.892	0.353	0.907	0.362	0.000	0.373
	50% complete									
β_1	0.905	0.147	0.907	0.121	0.901	0.122	0.912	0.150	0.862	0.164
β_2	0.895	0.337	0.893	0.293	0.891	0.297	0.903	0.369	0.001	0.311
β_3	0.901	0.453	0.903	0.424	0.903	0.428	0.914	0.516	0.481	0.331
β_4	0.902	0.454	0.896	0.402	0.892	0.404	0.900	0.439	0.000	0.378
β_5	0.905	0.337	0.901	0.294	0.900	0.295	0.906	0.299	0.000	0.341
	75% complete									
β_1	0.889	0.120	0.890	0.110	0.887	0.111	0.888	0.123	0.887	0.145
β_2	0.889	0.275	0.893	0.258	0.890	0.259	0.910	0.290	0.062	0.294
β_3	0.918	0.370	0.905	0.359	0.905	0.360	0.891	0.394	0.792	0.338
β_4	0.898	0.370	0.888	0.352	0.891	0.352	0.890	0.366	0.000	0.375
β_5	0.907	0.275	0.901	0.261	0.899	0.261	0.907	0.263	0.000	0.310

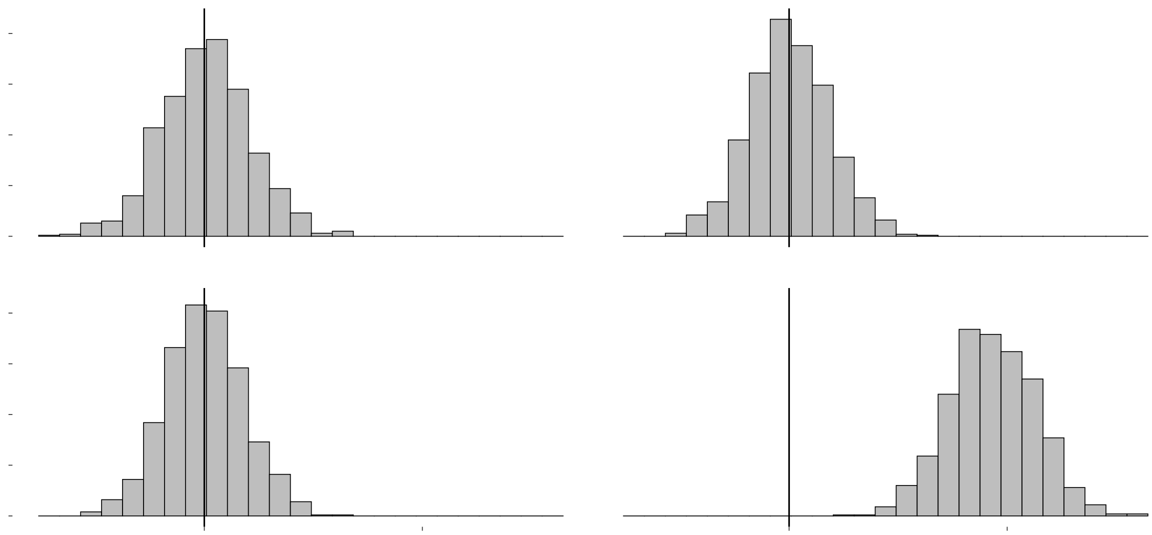
the relative gains from imputation are generally larger and the differences between OLS and GLS are much more striking. The estimator based on unconditional mean imputation has low coverage rates, which is due to the bias of the estimator (more below). Interestingly, the confidence intervals can be much narrower than those of the optimal GMM estimator, see for example those for β_3. The reason is that the regressors appear less correlated once the unconditional mean is imputed.[6]

To illustrate these points further, Figure 6 shows histograms of the estimates of β_2 (Panel (a)) and the corresponding standard errors (Panel (b)) when 50% of the sample is complete.

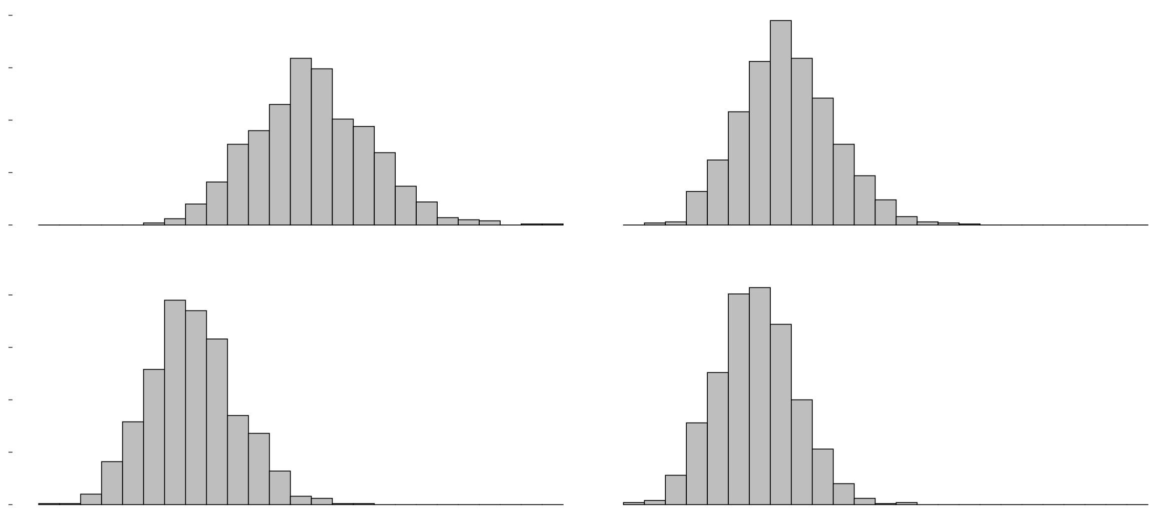
[6]This result can be seen from the following elementary consideration for estimating the covariance between X and Y when the unconditional mean is imputed whenever y_i is missing. $\hat{\sigma}_{X,Y} = \frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\left[\sum_{i:D_i=0}(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i:D_i=1}(x_i - \bar{x})(\bar{y} - \bar{y})\right] < \frac{1}{N_{\text{obs}}}\sum_{i:obs}(x_i - \bar{x})(y_i - \bar{y})$.

Figure 6: Simulation - Histograms for Setup 3

This figure shows histograms of the repeated sample distribution for estimates of β_2 (panel a) and standard errors of $\hat{\beta}_2$ (panel b) when 50% of the observations are complete. The vertical bar indicates the correct value for the parameter, $\beta_2 = 0.5$.



(a) Estimates of β_2



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

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

The poor coverage probability obtained from imputing unconditional means is due to the large bias of the estimator. We show the biases, again with a different percentage of complete observations, in Table 2. Even when 75% of the sample is complete, the bias is substantial for the unconditional mean imputation. The biases of all other estimators are negligible. In the table, we also report the root mean squared 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 almost as precise as the optimal GMM estimator and generally much more precise than the imputation OLS estimator.

Table 3 shows results with independent regressors when the complete sample contains

Table 2: Simulation - Bias and Model Fit for a General Missing Pattern

This tables shows the bias in the estimated coefficients and the root mean-squared error when different percentages of the data are missing.

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

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 0. In this case, unconditional mean imputation leads to correct moment conditions, as discussed in Section 3.1. Moreover, imputing the conditional or the unconditional mean does not increase the variance of the error term and hence, no benefits from using GLS exist. We show simulation results in Table 4 when $\beta = (1, 0.5, 0, 0, 0)'$ and the complete sample contains 50% of the observations. In this case, imputing the unconditional means decreases the correlation between the regressors, which reduces the variance of the estimated coefficients and the length of the confidence intervals. Since the moment conditions are valid, the estimator is also asymptotically unbiased. Consequently, it also has a lower mean squared error compared the estimators that impute conditional means. Clearly, in applications we do not know a priori if coefficients are equal to 0, that is, whether a firm characteristic is a true return predictor and we should therefore not rely on the unconditional mean imputation to deliver satisfactory results. As we discuss below, to determine which regressors are irrelevant,

Table 3: 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 all regressors are independent.

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

Table 4: 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 all regressors that may be missing do not affect the outcome.

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

we can carry out model selection to obtain a smaller model.

4.2 High-dimensional setting

We now again simulate data from the linear model

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

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

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

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

Figure 7: Model Selection - Sparse Model

This figure shows the model selection results for the sparse example (Section 4). The darker the color, the more frequent a particular model estimates a non-zero β_k. In the true model, the first five betas are non-zeros (above the red line), whereas the rest is equal to zero.

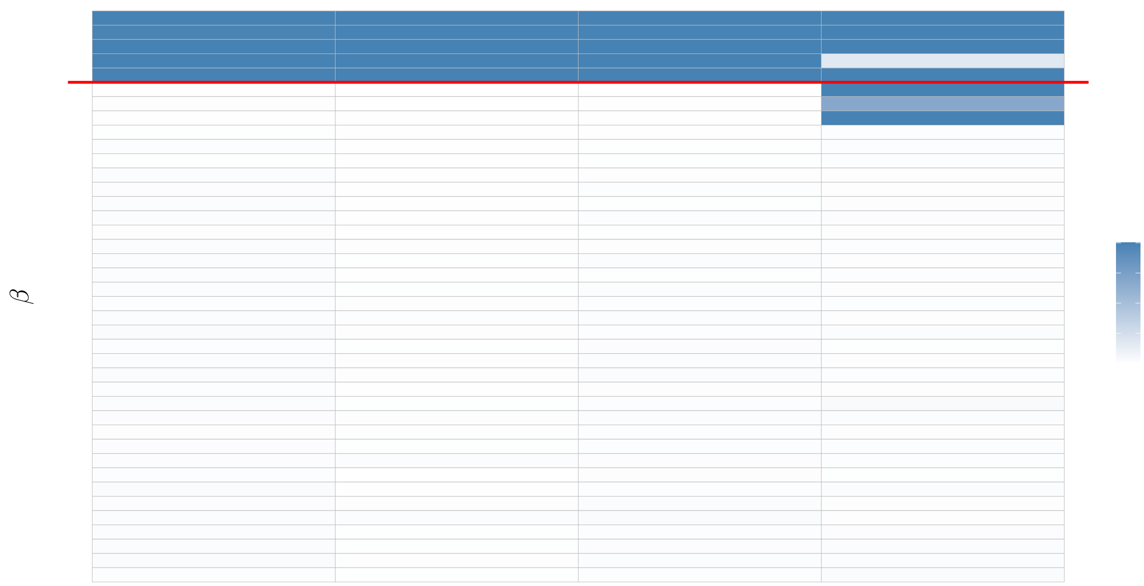
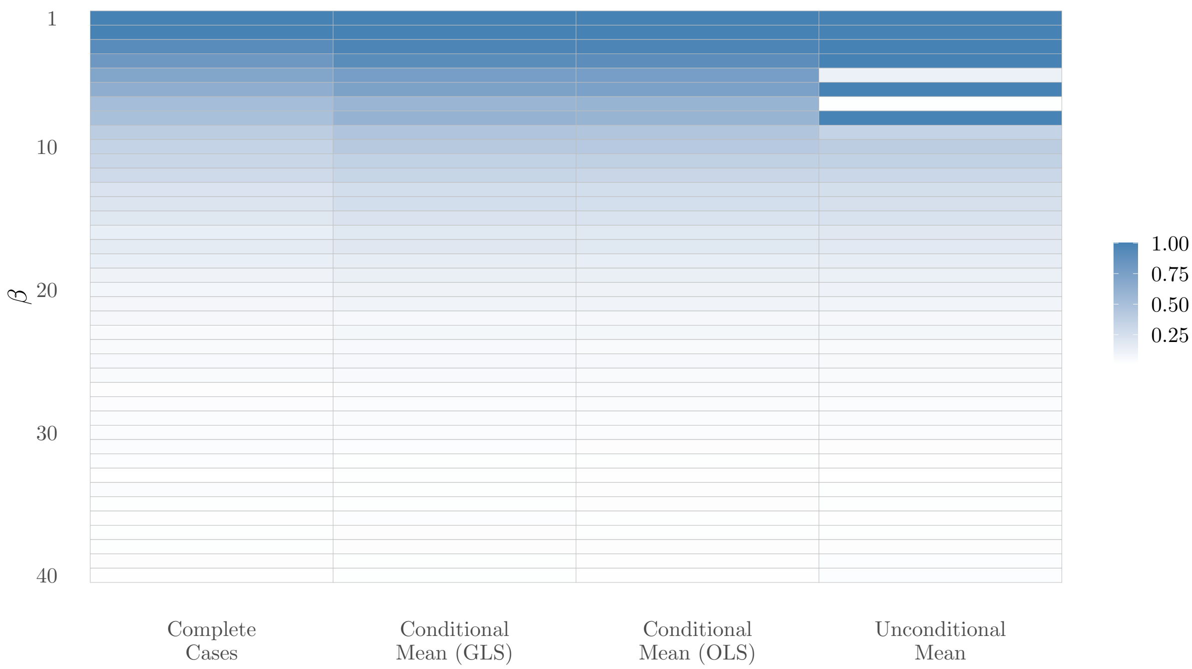


Figure 7 illustrates the frequency with which the different methods select regressors. The darker the color, the more frequent a particular model estimates a non-zero β_k. In the true model, the first five betas are non-zeros (above the red line), whereas the remaining ones are equal to zero. The complete 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 0. Unconditional mean imputation tends to set the estimated value of β_4 to 0 and instead frequently includes three of the irrelevant regressors. The mean squared prediction errors (MSPEs) of the four methods are 1.0187, 1.0105, 1.0141, and 1.7059, respectively, showing the imputation GLS estimator performs best and unconditional mean imputation performs worst.

The previous case was sparse, in the sense that only 5 coefficients were not equal to 0. 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 8. For the complete case estimator

Figure 8: Model Selection - Dense Model

This figure shows the model selection results for the non-sparse example (Section 4). 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 $\beta_k = 0.8^k$.

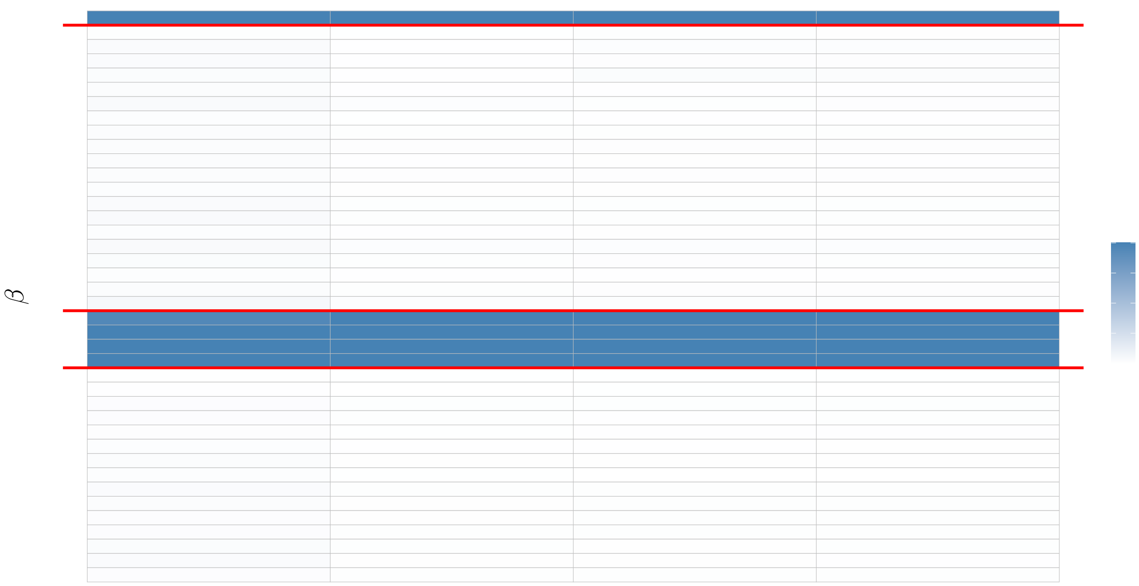


and both conditional mean imputation estimators, the larger a coefficient, the more likely it is not set to 0. This monotonicity does not hold for unconditional mean imputation. Here, the estimated value of β_5 is often set to 0, but regressors with smaller coefficients are included much more frequently. The MSPEs of the four methods are 1.0552, 1.0345, 1.0343, and 1.0887, respectively.

Another case in which imputation works particularly well is when regressors with missing values to do not have an impact on 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 0. The results are reported in Figure 9. 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 1.0206, 1.0076, 1.0075, and 1.0072 respectively. Therefore, all imputation methods perform similarly well and outperform the complete case.

Figure 9: Model Selection - Missing Regressor Irrelevant

This figure shows the model selection results for a sparse example (Section 4) 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.



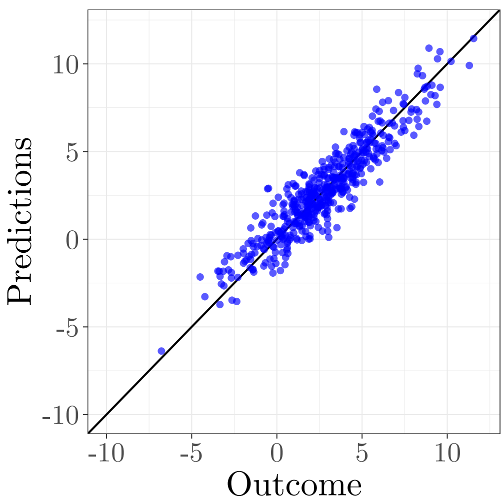
One advantage of imputations is that the whole sample can be used for predictions. The MSPE for the subset of non-complete observations is typically higher than for the complete observations, but the complete case might miss particularly interesting parts of the conditional distribution of outcomes, in our case returns. To illustrate this point, Figure 10 plots the out-of-sample realized returns against the predicted returns obtained with the different methods.[7] Recall the probability that an observation is completely observed depends on $X_{i,2}$. When using imputations, we make predictions for all outcomes, even when some regressors are missing. Comparing panels (a) and (b) we can see the observations with missing regressors tend to have lower returns.

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

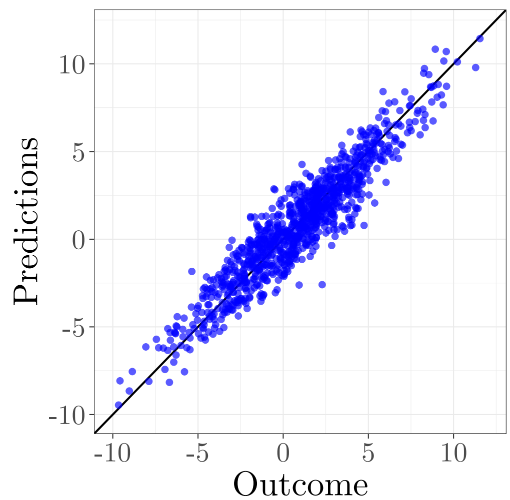
[7]For the out-of-sample predictions, we generate a new sample of complete observations with a sample size of $5,000$.

Figure 10: Outcomes versus predictions

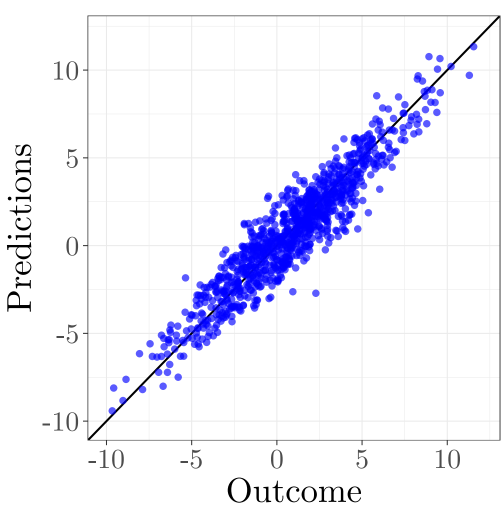
This figure shows out-of-sample outcomes against the predictions when the probability of an observation being complete depends on the regressors.



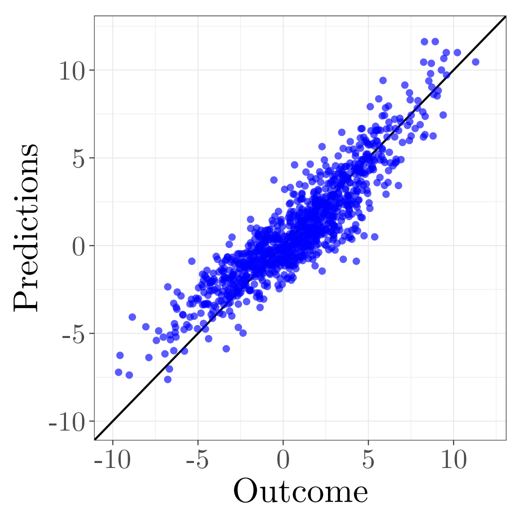
(a) Complete case



(b) GLS



(c) OLS



(d) Unconditional mean

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 large difference in portfolio returns. Second, when the probability that an observation is missing depends on the observed covariates, the complete case misses a systematically different part of the distribution of returns and not just a random sample. In this case, differences in portfolio returns will be even more pronounced. Finally, panel (d) shows that imputing unconditional means yields biased predictions. However, since predictions and outcomes are still positively related, portfolio formed based on these predictions will be very similar to those obtained with conditional mean imputation.

5 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.

5.1 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 not, 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 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 in the data matrix of these complete case observations. 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, both using 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 5 presents the RMSPE for all characteristics with missing values 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.[8] 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.

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 focuses primarily 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 imputation model to a linear conditional mean imputation model with lags reduces the RMSPE from 0.2213 to 0.11908, similar improvements are achieved for `VolumeTrend` and `CompEquIss`. For other characteristics, e.g. `OperProf` or `TotalAccruals` little improvement occurs from using the time series information relative to the purely cross-sectional model.

[8]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 5: Masking the complete case: Out-of-sample prediction error (RMSPE)

This table shows the 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 the 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	0.16130	0.15335	0.15447	0.14596
BetaFP	0.29233	0.19930	0.12656	0.19302	0.12660
BetaTailRisk	0.28832	0.22725	0.12012	0.22133	0.11998
BidAskSpread	0.29054	0.22923	0.21562	0.22436	0.21391
Cash	0.29499	0.23189	0.17492	0.20917	0.17200
CashProd	0.29004	0.16806	0.13283	0.15945	0.13197
CBOperProf	0.28711	0.16921	0.15899	0.16458	0.15455
CF	0.28776	0.14105	0.12489	0.13806	0.12323
cfp	0.29212	0.18823	0.13639	0.18338	0.13573
ChEQ	0.28591	0.17598	0.16921	0.16770	0.16151
ChInvIA	0.28433	0.22813	0.15243	0.22461	0.15153
CompEquIss	0.28513	0.21346	0.14777	0.20599	0.14720
CompositeDebtIssuance	0.28400	0.23291	0.21544	0.22640	0.21185
DelCOL	0.28301	0.18202	0.17691	0.17306	0.16913
DelFINL	0.29226	0.15227	0.15085	0.15034	0.14888
DelNetFin	0.29453	0.16472	0.15960	0.15899	0.15416
EarningsSurprise	0.28514	0.24758	0.20321	0.24218	0.20231
EBM	0.29058	0.20264	0.15735	0.20295	0.15737
EntMult	0.28848	0.15405	0.13559	0.14813	0.13410
EP	0.28528	0.16161	0.13388	0.15560	0.13193
EquityDuration	0.28425	0.16650	0.16549	0.16282	0.16230
GP	0.28540	0.20899	0.14791	0.20031	0.14444
grcapx	0.28822	0.17389	0.17065	0.17265	0.16978
GrLTNOA	0.29293	0.20990	0.20675	0.19784	0.19466
GrSaleToGrInv	0.28237	0.20898	0.20462	0.20304	0.19873
Herf	0.29012	0.27299	0.11431	0.26945	0.11521
hire	0.29036	0.22333	0.22089	0.22220	0.21975
Illiquidity	0.28407	0.11318	0.11068	0.10962	0.11046
IndMom	0.28709	0.27511	0.21835	0.27509	0.21940
IntMom	0.28780	0.21050	0.18632	0.20704	0.18434
Investment	0.29057	0.18117	0.14507	0.17624	0.14352
InvestPPEInv	0.29180	0.18157	0.17369	0.17706	0.16901
InvGrowth	0.28632	0.15011	0.12675	0.14664	0.12467
Leverage	0.29126	0.12722	0.11577	0.12499	0.11483
LRreversal	0.29098	0.20900	0.15263	0.20845	0.15382
MeanRankRevGrowth	0.28559	0.22130	0.11908	0.21632	0.11953
Mom12m	0.29444	0.14598	0.13707	0.14549	0.13616
Mom12mOffSeason	0.28892	0.16075	0.15537	0.16023	0.15551
Mom6m	0.28711	0.18147	0.17921	0.17715	0.17502
MomOffSeason	0.28607	0.18924	0.14549	0.18375	0.14531
MomOffSeason06YrPlus	0.28912	0.25723	0.16048	0.25334	0.15938
MomSeason	0.28727	0.26858	0.26705	0.26482	0.26280
MomSeason06YrPlus	0.28898	0.28893	0.28866	0.29002	0.29024
MomSeasonShort	0.28705	0.23879	0.23982	0.23941	0.24059
MRreversal	0.28912	0.26095	0.21018	0.26019	0.21123
NetDebtFinance	0.28281	0.18372	0.18295	0.17782	0.17784

Table 5: Masking the complete case: Out-of-sample prediction error (RMSPE) *(continued)*

NetEquityFinance	0.28759	0.19888	0.19186	0.19685	0.18893
NOA	0.28768	0.20516	0.18016	0.19255	0.17255
OperProf	0.28955	0.14945	0.14168	0.14149	0.13558
OPLeverage	0.29031	0.15045	0.13106	0.13776	0.12532
PriceDelayRsq	0.29042	0.24420	0.24107	0.21820	0.21549
PriceDelaySlope	0.29261	0.27055	0.27074	0.25555	0.25670
PriceDelayTstat	0.29069	0.28035	0.27744	0.23355	0.23227
RDS	0.29069	0.25025	0.24998	0.24651	0.24749
ResidualMomentum	0.28768	0.19404	0.15714	0.19353	0.15681
ReturnSkew	0.28496	0.21956	0.21812	0.21425	0.21352
roaq	0.28705	0.19720	0.19202	0.19354	0.18946
RoE	0.28562	0.15294	0.14837	0.14815	0.14146
ShareIss1Y	0.28509	0.21810	0.15263	0.21701	0.15338
SP	0.28578	0.12624	0.10651	0.12300	0.10568
Tax	0.28728	0.26373	0.25074	0.25757	0.24573
TotalAccruals	0.29088	0.18600	0.18123	0.18040	0.17394
TrendFactor	0.28447	0.27591	0.26268	0.27831	0.26545
VarCF	0.28942	0.19981	0.12161	0.19454	0.12135
VolMkt	0.28908	0.13890	0.11819	0.13536	0.11719
VolSD	0.28925	0.16072	0.10884	0.15482	0.10859
VolumeTrend	0.28863	0.24126	0.12852	0.23579	0.12837
XFIN	0.29061	0.16954	0.16631	0.16314	0.15916
zerotrade	0.28564	0.14008	0.11625	0.13777	0.11593
Average RMSPE	**0.28828**	**0.20415**	**0.17708**	**0.19838**	**0.17315**

5.2 Out-of sample predictions

We first illustrate the different ways of treating missing regressors in a classic empirical asset pricing application - cross-sectional out-of-sample return predictions. We report results for 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 3.3.1, 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 et al. (2010) and Freyberger et al. (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 vs. 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 6. In the linear setups, both the imputation model and the main model are linear, while in the nonlinear setup both models are additive nonlinear. In Appendix A.2.1, we include a third portfolio implementation going long the stocks with highest 50% predicted returns and shorting the stocks with lowest 50% predicted return.

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

This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction. We differentiate between the complete case method, 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 leg respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and polynomial model is detailed in Section 5.2. The sample period is 1990-2021.

	Mean (%)	Standard Deviation (%)	Sharpe Ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
Panel 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
Panel 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)	54.30	28.33	1.92	41.18	-13.12	-0.73	6.47
Cond. Mean (GLS / LASSO / w.lags)	55.62	28.15	1.98	41.67	-13.95	-0.44	5.00
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)	32.20	20.25	1.59	30.61	-1.59	-0.61	6.27
Cond. Mean (GLS / LASSO / w.lags)	32.85	20.05	1.64	30.78	-2.07	-0.47	5.68
Panel 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
Panel 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)	84.48	36.55	2.31	63.82	-20.66	-0.38	3.38
Cond. Mean (GLS / LASSO / w.lags)	84.28	38.15	2.21	64.92	-19.36	0.09	5.85
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)	41.34	21.35	1.94	36.98	-4.36	-0.64	4.70
Cond. Mean (GLS / LASSO / w.lags)	41.42	21.83	1.90	37.74	-3.69	-0.41	5.44

Panel A of Table 6 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 portfolio 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 in Section 2.1. 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 estimates, we might still get rather similar portfolios as long as the two methods produce a similar ranking of their predicted values.

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 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 $\hat{\sigma}^2_{it}$ presented in Section 3.2 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 11 shows the selected characteristics for each case.

Similar to the standard linear model in Panel A, we find in Panel B of Table 6 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 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.

The Panel C and Panel D illustrate the results for a nonlinear model, that is, an additive model as outlined in Section 3.3.1. 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 Panel C and Panel 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 vs. 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 et al. (2020) to select the most important characteristics over the first part of the sample and then, exactly as for the other methods, make rolling one-month predictions using an estimation window of 60 months. Overall, the results are very similar to those in Panel C. Again note that the characteristics we use are known return predictors and it is therefore not surprising that including all of them in the model 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. While the imputation methods perform similarly with and without regularization,

Figure 11: 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 unconditional 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 3.2. In a next step, we use an adaptive LASSO procedure where the model is selected using the BIC. The initial estimator for β_t we use 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.

in the complete case, the regularized model leads to considerably worse performance. The reason is that the complete case contains much fewer observations and hence, is less likely to detect significant return predictors.

5.3 Incremental Information

We now re-visit the 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, recent papers such as Green et al. (2017), Freyberger et al. (2020), Kozak et al. (2020) and Gu et al. (2020) make it clear that we need to consider characteristics jointly and to determine if a characteristics 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 Online 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 previous 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

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

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 CAPM or possibly later the Fama and French (1993) three-factor model.

Table 7: Incremental Information in Newly Discovered Characteristics

This table shows the estimates, p-values and adjusted 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 non-zero effect on returns given all previously discovered characteristics. Estimates are reported in ×100 %. p-values smaller than 5% are printed in bold. The final two rows of the table display the number of characteristics that are significant at a 5%, 1% significance level 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	0.5006	0.7509	-0.5409	0.5006	0.7509	-0.5409	0.5006	0.7509
ep	0.8178	**0.0089**	**0.0492**	0.8178	**0.0089**	**0.0492**	0.8166	**0.0091**	0.0503
size	-1.4669	**0.0009**	**0.0036**	-1.4008	**0.0024**	**0.0098**	-1.4057	**0.0023**	**0.0098**
earningssurprise	1.5031	**0.0000**	**0.0000**	1.3100	**0.0000**	**0.0000**	1.3344	**0.0000**	**0.0000**
lrreversal	-0.1417	0.6547	1	0.1706	0.5484	1	0.1197	0.6709	1
mrreversal	-0.2512	0.2785	1	-0.3190	0.2157	0.6664	-0.3343	0.2018	0.7325
bidaskspread	-0.0024	0.9929	1	-0.3698	0.2209	0.6862	-0.5272	**0.0377**	0.1639
leverage	0.3287	0.0749	0.6353	0.4485	**0.0162**	0.0823	0.5716	**0.0017**	**0.0087**
streversal	-2.2822	**0.0000**	**0.0000**	-2.4956	**0.0000**	**0.0000**	-2.4484	**0.0000**	**0.0000**
bmdec	1.0577	**0.0000**	**0.0002**	0.7079	**0.0034**	**0.0223**	0.8630	**0.0003**	**0.0022**
bookleverage	0.0055	0.9863	1	-1.0086	**0.0004**	**0.0032**	-1.0810	**0.0000**	**0.0000**
mom12m	1.7553	**0.0000**	**0.0000**	1.7208	**0.0000**	**0.0000**	1.7452	**0.0000**	**0.0000**
mom6m	-0.6392	**0.0077**	0.0504	-1.1884	**0.0000**	**0.0000**	-1.0632	**0.0000**	**0.0001**
cf	1.2370	0.1102	0.6855	0.5691	0.0658	0.3274	0.8669	**0.0001**	**0.0010**
meanrankrevgrowth	0.1385	0.3351	1	0.1855	0.1679	0.8255	-0.0303	0.8062	1
accruals	-0.5112	**0.0077**	0.1126	-0.4800	**0.0000**	**0.0000**	-0.3949	**0.0000**	**0.0000**
roe	0.2555	0.5615	1	0.4645	0.1262	0.7219	0.6325	**0.0000**	**0.0001**
sp	0.4635	**0.0395**	0.3804	0.6160	**0.0035**	**0.0234**	0.6183	**0.0033**	**0.0186**
varcf	-0.0656	0.858	1	0.0550	0.7961	1	0.0228	0.9047	1
volmkt	-0.1371	0.498	1	-0.4880	**0.0315**	0.2009	-0.3346	0.1256	0.6411
volumetrend	-0.3953	**0.0114**	0.1847	-0.3133	**0.0268**	0.1813	-0.2793	**0.0354**	0.2053
chinvia	-0.3913	**0.0001**	**0.0014**	-0.4132	**0.0000**	**0.0000**	-0.4115	**0.0000**	**0.0000**
grsaletogrinv	0.4093	**0.0000**	**0.0001**	0.5180	**0.0000**	**0.0000**	0.5116	**0.0000**	**0.0000**
indmom	0.2513	0.1738	1	0.9742	**0.0000**	**0.0000**	0.9582	**0.0000**	**0.0000**
coskewness	-0.3456	0.1322	1	-0.2424	0.1401	1	-0.2403	0.142	0.8372
volsd	-0.7776	0.1371	1	-0.6515	0.2345	1	0.1971	0.4525	1
chinv	-0.0323	0.8515	1	-0.1918	0.0988	0.7239	-0.3279	**0.0019**	**0.0186**
illiquidity	3.3185	**0.0412**	0.4735	2.4264	**0.0265**	0.2346	-0.0736	0.8915	1
grltnoa	0.0256	0.8475	1	0.0450	0.6164	1	0.0220	0.8004	1
equityduration	-0.7767	0.1492	1	-0.3981	0.0976	0.8121	-0.2999	**0.0291**	0.1915
high52	1.1194	**0.0008**	**0.0170**	2.0976	**0.0000**	**0.0009**	2.0124	**0.0000**	**0.0003**
investment	-0.0117	0.9433	1	0.0821	0.5727	1	-0.0038	0.9706	1
noa	-0.3652	**0.0223**	0.2952	-0.9717	**0.0000**	**0.0000**	-1.0206	**0.0000**	**0.0000**
tax	0.2128	0.0784	0.7583	0.0622	0.5662	1	0.0959	0.3705	1
cfp	0.1788	0.5833	1	0.2275	0.2851	1	0.2631	0.2012	1
delcoa	-0.0287	0.8923	1	-0.0327	0.823	1	-0.0336	0.8246	1
delcol	0.1078	0.4731	1	0.0829	0.5294	1	0.0768	0.4956	1
delfinl	-0.2278	**0.0342**	0.4363	-0.3815	**0.0000**	**0.0003**	-0.3830	**0.0000**	**0.0000**
dellti	-0.1617	0.0772	0.8803	-0.1338	**0.0300**	0.321	-0.1451	**0.0207**	0.1688
delnetfin	-0.2082	0.2387	1	-0.2773	0.0714	0.6912	-0.2764	0.0605	0.5085
pricedelayrsq	-0.2775	0.1388	1	-0.1770	0.3257	1	0.1337	0.3705	1
pricedelayslope	-0.0668	0.5609	1	-0.0398	0.6125	1	-0.0954	0.2263	1
pricedelaytstat	-0.0398	0.6163	1	-0.0087	0.9239	1	-0.0611	0.462	1
totalaccruals	-0.2365	0.2183	1	-0.0192	0.8425	1	-0.0590	0.5201	1
compequiss	-0.3944	0.0676	0.9811	-0.8547	**0.0000**	**0.0004**	-0.7833	**0.0000**	**0.0007**
grcapx	-0.3199	0.1783	1	-0.2343	0.1823	1	-0.0483	0.687	1
herf	-0.1525	0.2572	1	-0.5115	**0.0000**	**0.0004**	-0.5103	**0.0000**	**0.0003**
idiorisk	-0.1309	0.5963	1	-0.1260	0.5727	1	0.1845	0.359	1
netdebtfinance	-0.0753	0.5007	1	-0.2036	**0.0076**	0.1071	-0.1758	**0.0163**	0.167
operprof	0.4086	0.3288	1	0.3386	0.2162	1	1.0198	**0.0000**	**0.0000**

Table 7: Incremental Information in Newly Discovered Characteristics *(continued)*

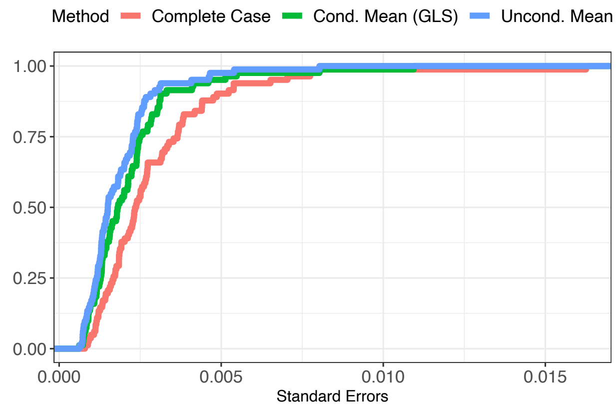
netequityfinance	-0.1300	0.3034	1	-0.1192	0.368	1	-0.1033	0.4239	1
xfin	0.0078	0.9719	1	-0.0609	0.6449	1	-0.0581	0.6241	1
zerotrade	-0.5976	0.12	1	0.2139	0.489	1	0.5222	0.0765	0.8222
ebm	0.1187	0.4793	1	0.1720	0.1345	1	0.2323	**0.0326**	0.3746
assetgrowth	-0.1697	0.7005	1	-0.6540	**0.0002**	**0.0064**	-0.6237	**0.0002**	**0.0042**
compositedebtissuance	0.1441	0.2167	1	-0.1625	0.0716	0.8214	-0.0389	0.6016	1
investppeinv	-0.0924	0.6894	1	0.0868	0.5774	1	-0.0961	0.4596	1
mom12moffseason	-0.0938	0.8435	1	-0.9911	**0.0166**	0.2172	-0.9027	**0.0268**	0.3491
momoffseason	-0.6142	0.2068	1	-1.0601	**0.0000**	**0.0001**	-0.8506	**0.0000**	**0.0001**
momoffseason06yrplus	-0.5856	**0.0099**	0.6449	-0.8756	**0.0000**	**0.0000**	-0.7982	**0.0000**	**0.0000**
momseason	0.6720	**0.0123**	0.7205	0.8469	**0.0000**	**0.0000**	0.9538	**0.0000**	**0.0000**
momseason06yrplus	0.8888	**0.0000**	**0.0002**	0.6639	**0.0000**	**0.0000**	0.6662	**0.0000**	**0.0000**
momseasonshort	0.0311	0.8942	1	0.6371	**0.0015**	**0.0298**	0.6154	**0.0022**	**0.0345**
shareiss1y	0.1736	0.2221	1	0.0220	0.8407	1	-0.0777	0.4831	1
cashprod	-0.1062	0.6973	1	0.5978	**0.0002**	**0.0057**	0.5671	**0.0004**	**0.0076**
cheq	-0.4673	0.0847	1	0.0389	0.7818	1	-0.0717	0.5468	1
maxret	0.0507	0.8945	1	0.4454	0.1541	1	0.4258	0.1682	1
opleverage	-0.2063	0.7517	1	-0.4462	0.1145	1	-0.2699	0.2352	1
roaq	1.1175	**0.0000**	**0.0065**	1.6705	**0.0000**	**0.0000**	1.3882	**0.0000**	**0.0000**
entmult	-0.2862	0.5951	1	-0.5653	**0.0176**	0.2169	-0.5635	**0.0000**	**0.0006**
rds	-0.2272	**0.0433**	1	0.0582	0.4903	1	0.0882	0.3542	1
residualmomentum	0.1587	0.6766	1	0.2274	0.4927	1	-0.0297	0.9116	1
cash	0.8235	**0.0000**	**0.0015**	1.2623	**0.0000**	**0.0000**	1.1832	**0.0000**	**0.0000**
invgrowth	1.0660	0.1303	1	0.4342	0.2895	1	0.2339	0.1189	1
intmom	0.3672	0.3145	1	-0.1285	0.5913	1	-0.0946	0.6968	1
gp	0.3529	0.3405	1	0.4540	**0.0037**	0.0615	0.3423	**0.0086**	0.1311
betafp	-0.3429	0.3316	1	0.0892	0.7209	1	0.3175	0.1561	1
betatailrisk	0.3229	0.3404	1	0.3692	0.1228	1	0.3505	0.1225	1
hire	0.0629	0.7172	1	0.0642	0.4316	1	0.0652	0.3689	1
cboperprof	0.2999	0.4162	1	0.8783	**0.0000**	**0.0004**	0.4308	**0.0027**	**0.0452**
returnskew	0.0628	0.8133	1	0.1748	0.1835	1	0.1350	0.3053	1
trendfactor	1.3712	**0.0000**	**0.0000**	1.4088	**0.0000**	**0.0000**	1.1465	**0.0000**	**0.0003**
# sign. at 5%		**23**	**13**		**38**	**29**		**43**	**34**
# sign. at 1%		**16**	**11**		**31**	**25**		**36**	**30**

We estimate this model using only the data until the publication date of the new candidate predictor, not the full sample. To determine if a characteristic is significant, we test $H_0 : \beta_k = 0$ using a two-sided t-test. We allow for cross-sectional dependence of the error terms by using clustered standard errors. 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) and Green et al. (2017)).[9] These p-values might be larger than 1 in which case we set them to 1 when

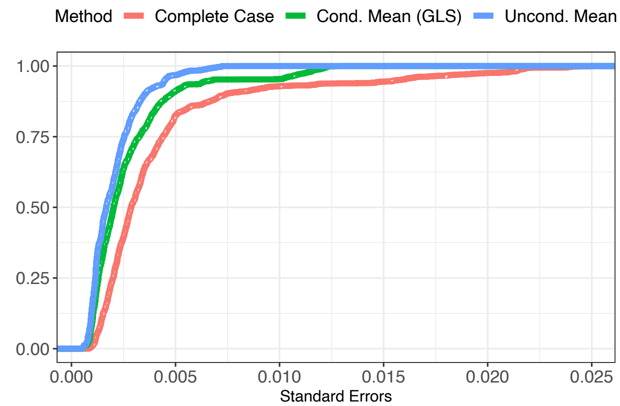
[9]Specifically, let p_i denote the standard p-value of the ith test and assume that the p-values have been ordered, such that $p_1 \leq p_2 \leq \cdots \leq p_K$ where K is the number covariates in the current model. The adjusted

Figure 12: Empirical cdfs of standard errors of new characteristics

This figure shows the empirical distribution function of the standard errors of the estimated coefficients for the three different methods. Panel (a) shows the standard errors for all new characteristics. Panel (b) shows the standard errors for all estimated coefficients in equation (4) across all 82 regression models we estimate.



(a) New characteristics



(b) All characteristics

presenting our results. Table 7 shows the estimates and p-values. In order to interpret the results in Table 7 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.1 in the Online Appendix. In a univariate model, we declare the vast majority of the predictors statistically significant.[10] This result resonates with the findings in Jensen et al. (2021), who document a high degree of replicability in empirical asset pricing studies.

Now, in a multivariate setup, 11 to 23 characteristics are significant in the complete case. This relatively small number is due to a lack of statistical power in the complete case. Even very strong predictors, such as six months momentum or book leverage (`mom6m` and `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 differ-

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 $i < K$.

[10]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. Moreover, the weighted estimator yields almost identical results.

ent. This difference is due to the interaction of two effects highlighted in Section 4. First, unconditional mean imputation yields biased estimators and estimated coefficients may be either too large or too close to 0. As a specific example, consider operating profitability (`operprof`) in Table 7. The coefficients in the complete case and with conditional mean imputation are quite similar (0.4086% and 0.3386%, respectively), while 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 get artificially small standard errors.

To illustrate this difference, Figure 12 shows empirical cumulative distribution functions (cdf) of the standard errors obtained using the different methods.[11] In Panel (a), we plot the cdf for all new characteristics (i.e. for the standard errors in Table 7). Panel (b) shows the cdf of the standard errors for all estimated coefficients in equation (4) across all 82 models we estimate. The complete case yields the largest standard errors because it only makes use of a subset of the data and the standard errors with unconditional mean imputation tend to be the smallest.

6 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

[11] We do not compare the p-values because due to the bias of unconditional mean imputation it is not clear how the p-values of unconditional mean imputation should compare to the p-values of the other methods.

mean imputation and adjust for the estimation error in subsequent inference.

Our proposed two-step approach can be applied in other common areas of research such as estimating the stochastic discount factors, illustrated in A.3, characteristic based factor models, and international studies. These items are left for future research.

References

Abarbanell, J. S. and B. J. Bushee (1998). Abnormal returns to a fundamental analysis strategy. *Accounting Review*, 19–45.

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

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

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

Anderson, C. W. and L. Garcia-Feijoo (2006). Empirical evidence on capital investment, growth options, and security returns. *Journal of Finance 61*(1), 171–194.

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

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

Balakrishnan, K., E. Bartov, and L. Faurel (2010). Post loss/profit announcement drift. *Journal of Accounting and Economics 50*(1), 20–41.

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

Bali, T. G., R. F. Engle, and S. Murray (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

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

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

Barbee Jr, W. C., S. Mukherji, and G. A. Raines (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal 52*(2), 56–60.

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

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

Beckmeyer, H. and T. Wiedmann (2022). Recovering missing firm characteristics with attention-based machine learning. *Available at SSRN 4003455*.

Belo, F. and X. Lin (2012). The inventory growth spread. *Review of Financial Studies 25*(1), 278–313.

Belo, F., X. Lin, and S. Bazdresch (2014). Labor hiring, investment, and stock return predictability in the cross section. *Journal of Political Economy 122*(1), 129–177.

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

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

Blitz, D., J. Huij, and M. Martens (2011). Residual momentum. *Journal of Empirical Finance 18*(3), 506–521.

Bradshaw, M. T., S. A. Richardson, and R. G. Sloan (2006). The relation between corporate financing activities, analysts' forecasts and stock returns. *Journal of Accounting and Economics 42*(1-2), 53–85.

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

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2022). Missing financial data. *Available at SSRN 4106794*.

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

Carhart, M. M., J. N. Carpenter, A. W. Lynch, and D. K. Musto (2002). Mutual fund survivorship. *Review of Financial Studies 15*(5), 1439–1463.

Chandrashekar, S. and R. K. Rao (2009). The productivity of corporate cash holdings and the cross-section of expected stock returns. *Unpublished Manuscript, UT Austin*.

Chen, A. Y. and J. McCoy (2022). Missing values and the dimensionality of expected returns. *arXiv preprint arXiv:2207.13071*.

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

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

Chordia, T., A. Subrahmanyam, and V. R. Anshuman (2001). Trading activity and expected stock returns. *Journal of Financial Economics 59*(1), 3–32.

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

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

Cooper, M. J., H. Gulen, and M. J. Schill (2008). Asset growth and the cross-section of stock returns. *Journal of Finance 63*(4), 1609–1651.

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

Daniel, K. and S. Titman (2006). Market reactions to tangible and intangible information. *Journal of Finance 61*(4), 1605–1643.

De Bondt, W. F. and R. Thaler (1985). Does the stock market overreact? *Journal of Finance 40*(3), 793–805.

Dechow, P. M., R. G. Sloan, and M. T. Soliman (2004). Implied equity duration: A new measure of equity risk. *Review of Accounting Studies 9*(2), 197–228.

Desai, H., S. Rajgopal, and M. Venkatachalam (2004). Value-glamour and accruals mispricing: One anomaly or two? *Accounting Review 79*(2), 355–385.

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

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

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2006). Profitability, investment and average returns. *Journal of Financial Economics 82*(3), 491–518.

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

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

Foster, G., C. Olsen, and T. Shevlin (1984). Earnings releases, anomalies, and the behavior of security returns. *Accounting Review*, 574–603.

Frazzini, A. and L. H. Pedersen (2014). Betting against beta. *Journal of Financial Economics 111*(1), 1–25.

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

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

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

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

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

Han, Y., G. Zhou, and Y. Zhu (2016). A trend factor: Any economic gains from using information over investment horizons? *Journal of Financial Economics 122*(2), 352–375.

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

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

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

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

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

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

Heston, S. L. and R. Sadka (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics 87*(2), 418–445.

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

Hou, K. and T. J. Moskowitz (2005). Market frictions, price delay, and the cross-section of expected returns. *The Review of Financial Studies 18*(3), 981–1020.

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

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

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *Journal of Finance 45*(3), 881–898.

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

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2021). Is there a replication crisis in finance? Technical report, 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*(1), 745–777.

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

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

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

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

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

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation, and risk. *Journal of Finance 49*(5), 1541–1578.

Landsman, W. R., B. L. Miller, K. Peasnell, and S. Yeh (2011). Do investors understand really dirty surplus? *Accounting Review 86*(1), 237–258.

Lev, B. and D. Nissim (2004). Taxable income, future earnings, and equity values. *Accounting Review 79*(4), 1039–1074.

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

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

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

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

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

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

Liu, H., X. Tang, and G. Zhou (2022). Recovering the fomc risk premium. *Journal of Financial Economics 145*(1), 45–68.

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

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

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

Lyandres, E., L. Sun, and L. Zhang (2008). The new issues puzzle: Testing the investment-based explanation. *Review of Financial Studies 21*(6), 2825–2855.

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*(1), 277–307.

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

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

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

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

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

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

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

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics 104*(1), 162–185.

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

Pontiff, J. and A. Woodgate (2008). Share issuance and cross-sectional returns. *Journal of Finance 63*(2), 921–945.

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

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

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

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

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

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

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

Thomas, J. K. and H. Zhang (2002). Inventory changes and future returns. *Review of Accounting Studies 7*(2), 163–187.

Titman, S., K. J. Wei, and F. Xie (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis 39*(4), 677–700.

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

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

Xiong, R. and M. Pelger (2022). Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*.

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

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

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

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 (2021). We refer to their paper and the companion website for the precise construction.

Table A.1: Overview of the Characteristics

Acronym	Description	Publication Year	Reference	% missing
Accruals	Accruals	1996	Sloan (1996)	0.50
AssetGrowth	Asset growth	2008	Cooper et al. (2008)	0.00
Beta	CAPM beta	1973	Fama and MacBeth (1973)	0.00
BetaFP	Frazzini-Pedersen Beta	2014	Frazzini and Pedersen (2014)	6.82
BetaTailRisk	Tail risk beta	2014	Kelly and Jiang (2014)	34.65
BidAskSpread	Bid-ask spread	1986	Amihud and Mendelson (1986)	7.33
BMdec	Book to market using December ME	1992	Fama and French (1992)	0.00
BookLeverage	Book leverage (annual)	1992	Fama and French (1992)	0.00
Cash	Cash to assets	2012	Palazzo (2012)	28.91
CashProd	Cash Productivity	2009	Chandrashekar and Rao (2009)	9.73
CBOperProf	Cash-based operating profitability	2016	Ball et al. (2016)	29.60
CF	Cash flow to market	1994	Lakonishok et al. (1994)	9.10
cfp	Operating Cash flows to price	2004	Desai et al. (2004)	14.34
ChEQ	Growth in book equity	2010	Lockwood and Prombutr (2010)	3.64
ChInv	Inventory Growth	2002	Thomas and Zhang (2002)	0.00
ChInvIA	Change in capital inv (ind adj)	1998	Abarbanell and Bushee (1998)	11.45
CompEquIss	Composite equity issuance	2006	Daniel and Titman (2006)	37.48
CompositeDebtIssuance	Composite debt issuance	2008	Lyandres et al. (2008)	40.26
Coskewness	Coskewness	2000	Harvey and Siddique (2000)	0.00
DelCOA	Change in current operating assets	2005	Richardson et al. (2005)	0.00
DelCOL	Change in current operating liabilities	2005	Richardson et al. (2005)	0.50
DelFINL	Change in financial liabilities	2005	Richardson et al. (2005)	0.79
DelLTI	Change in long-term investment	2005	Richardson et al. (2005)	0.00
DelNetFin	Change in net financial assets	2005	Richardson et al. (2005)	0.79
EarningsSurprise	Earnings Surprise	1984	Foster et al. (1984)	19.17
EBM	Enterprise component of BM	2007	Penman et al. (2007)	9.67
EntMult	Enterprise Multiple	2011	Loughran and Wellman (2011)	27.24
EP	Earnings-to-Price Ratio	1977	Basu (1977)	35.05
EquityDuration	Equity Duration	2004	Dechow et al. (2004)	1.83
GP	gross profits / total assets	2013	Novy-Marx (2013)	19.42
grcapx	Change in capex (two years)	2006	Anderson and Garcia-Feijoo (2006)	18.78
GrLTNOA	Growth in long term operating assets	2003	Fairfield et al. (2003)	2.64
GrSaleToGrInv	Sales growth over inventory growth	1998	Abarbanell and Bushee (1998)	23.15
Herf	Industry concentration (sales)	2006	Hou and Robinson (2006)	16.75
High52	52 week high	2004	George and Hwang (2004)	0.00
hire	Employment growth	2014	Belo et al. (2014)	1.53
IdioRisk	Idiosyncratic risk	2006	Ang et al. (2006)	0.00
Illiquidity	Amihud's illiquidity	2002	Amihud (2002)	4.72
IndMom	Industry Momentum	1999	Moskowitz and Grinblatt (1999)	9.10
IntMom	Intermediate Momentum	2012	Novy-Marx (2012)	9.25
Investment	Investment to revenue	2004	Titman et al. (2004)	25.73
InvestPPEInv	change in ppe and inv/assets	2008	Lyandres et al. (2008)	12.02
InvGrowth	Inventory Growth	2012	Belo and Lin (2012)	40.76
Leverage	Market leverage	1988	Bhandari (1988)	9.33
LRreversal	Long-run reversal	1985	De Bondt and Thaler (1985)	16.00
MaxRet	Maximum return over month	2010	Bali et al. (2011)	0.00
MeanRankRevGrowth	Revenue Growth Rank	1994	Lakonishok et al. (1994)	35.49
Mom12m	Momentum (12 month)	1993	Jegadeesh and Titman (1993)	9.28
Mom12mOffSeason	Momentum without the seasonal part	2008	Heston and Sadka (2008)	9.15
Mom6m	Momentum (6 month)	1993	Jegadeesh and Titman (1993)	9.18
MomOffSeason	Off season long-term reversal	2008	Heston and Sadka (2008)	9.69
MomOffSeason06YrPlus	Off season reversal years 6 to 10	2008	Heston and Sadka (2008)	31.66
MomSeason	Return seasonality years 2 to 5	2008	Heston and Sadka (2008)	9.68
MomSeason06YrPlus	Return seasonality years 6 to 10	2008	Heston and Sadka (2008)	31.53
MomSeasonShort	Return seasonality last year	2008	Heston and Sadka (2008)	9.18
MRreversal	Medium-run reversal	1985	De Bondt and Thaler (1985)	9.32
NetDebtFinance	Net debt financing	2006	Bradshaw et al. (2006)	10.75
NetEquityFinance	Net equity financing	2006	Bradshaw et al. (2006)	0.92
NOA	Net Operating Assets	2004	Hirshleifer et al. (2004)	0.42
OperProf	operating profits / book equity	2006	Fama and French (2006)	56.49

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

OPLeverage	Operating leverage	2010	Novy-Marx (2011)	0.20
PriceDelayRsq	Price delay r square	2005	Hou and Moskowitz (2005)	2.43
PriceDelaySlope	Price delay coeff	2005	Hou and Moskowitz (2005)	2.43
PriceDelayTstat	Price delay SE adjusted	2005	Hou and Moskowitz (2005)	2.71
RDS	Real dirty surplus	2011	Landsman et al. (2011)	6.63
ResidualMomentum	Momentum based on FF3 residuals	2011	Blitz et al. (2011)	13.33
ReturnSkew	Return skewness	2016	Bali et al. (2016)	0.59
roaq	Return on assets (qtrly)	2010	Balakrishnan et al. (2010)	13.83
RoE	net income / book equity	1996	Haugen and Baker (1996)	0.01
ShareIss1Y	Share issuance (1 year)	2008	Pontiff and Woodgate (2008)	9.31
Size	Size	1981	Banz (1981)	0.00
SP	Sales-to-price	1996	Barbee Jr et al. (1996)	9.30
STreversal	Short term reversal	1989	Jegadeesh (1990)	0.00
Tax	Taxable income to income	2004	Lev and Nissim (2004)	11.58
TotalAccruals	Total accruals	2005	Richardson et al. (2005)	4.97
TrendFactor	Trend Factor	2016	Han et al. (2016)	52.96
VarCF	Cash-flow to price variance	1996	Haugen and Baker (1996)	15.51
VolMkt	Volume to market equity	1996	Haugen and Baker (1996)	4.07
VolSD	Volume Variance	2001	Chordia et al. (2001)	5.93
VolumeTrend	Volume Trend	1996	Haugen and Baker (1996)	10.48
XFIN	Net external financing	2006	Bradshaw et al. (2006)	11.43
zerotrade	Days with zero trades	2006	Liu (2006)	4.00

A.1 Additional Definitions

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

A.1.1 Missing patterns

A missing pattern describes which data are missing. Figure 5 shows examples of missing patterns. In our application, we cannot assume that we are confronted with a particular missing pattern, and instead deal with general missing patterns. Our theoretical results require a non-negligible part of the data to be complete. Generalizing these results would require much stronger assumptions and does not occur in our empirical application.

A.1.2 Missing mechanisms

The missing mechanism describes why data are missing, i.e. it describes the relationship between the missingness and the values of the observed (and possibly unobserved) variables. Rubin (1976) introduces three formal definitions for missing mechanisms that have become standard in the literature. He differentiates between missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). We recall these basic definitions, using our notation from Section 3 below.

In Section 3 (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 3.3.2). We refer to the analysis based on the cases that are completely observed as the complete case analysis. This is in contrast to the "complete data analysis" which is based on the hypothetically observed data in the absence of any missing data.

The data is **MCAR** if $D_i \perp\!\!\!\perp Y_i, X_i, V_i$, i.e. whether an observation is missing does not depend on the other variables. When the data is MCAR, the complete case analysis yields valid inference, but there is a loss of efficiency relative to the complete data analysis due to the decreased sample size (Fitzmaurice et al. (2015)). The data is **MAR**[1] if $D_i \perp\!\!\!\perp X_i \mid Y_i, X_i^{(o)}, V_i$. That is, missing

[1]MCAR is a special case of MAR.

is only random once we condition on observed covariates. We rely on this type of assumption (but based solely on conditional moments) in our analysis. When the data is MAR, the complete case analysis generally yields valid inference, but might require an estimator based on inverse propensity weighting (as in section 3.3.2). Again, neglecting a part of the sample results in an inefficient estimator.

Data is **NMAR**, sometimes also referred to as missing not at random, if D_i depends on unobserved regressors. 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 Further empirical results

A.2.1 Prediction exercise

Table A.2: Performance Statistics For Out-of-Sample Predictions

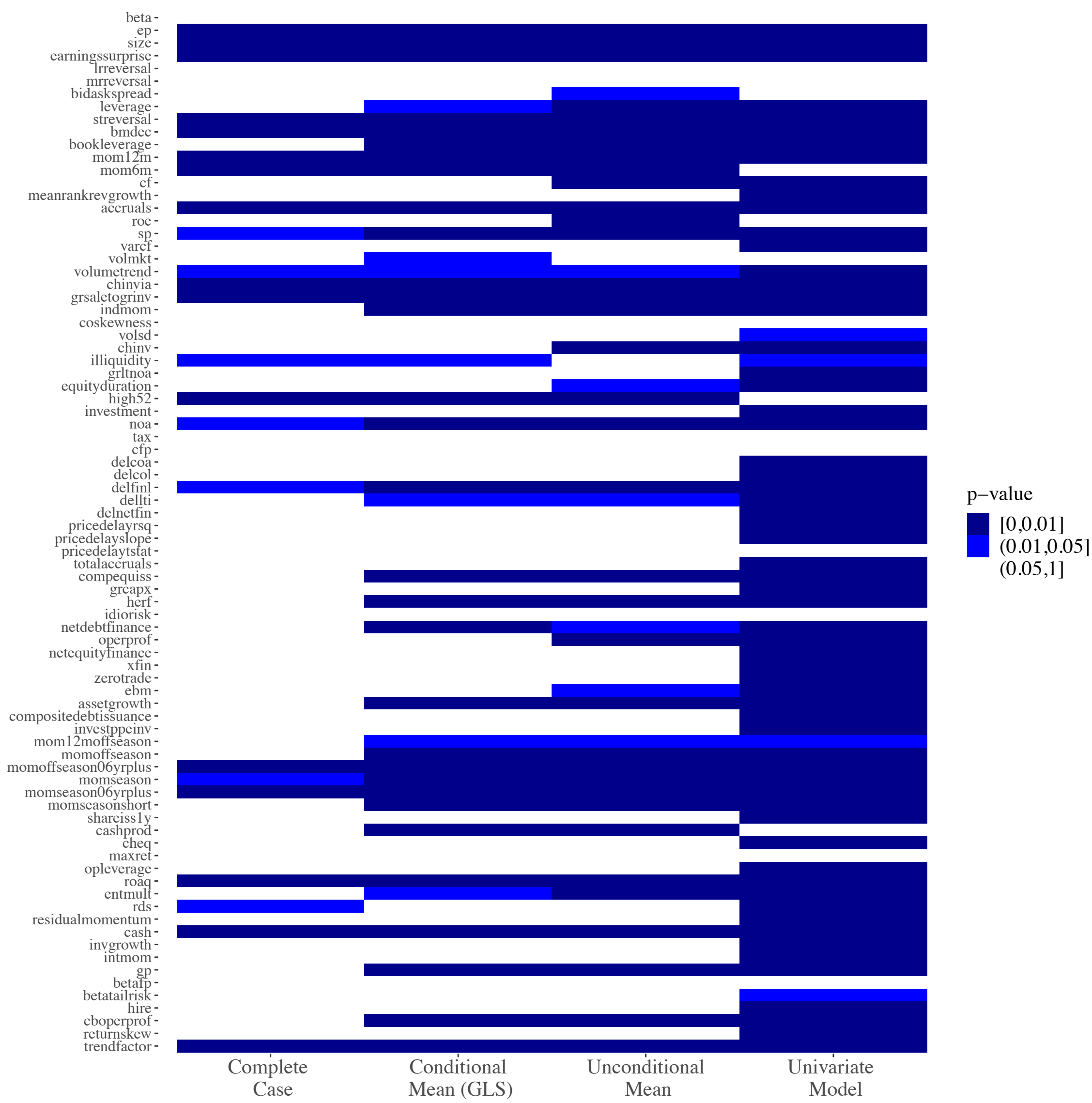
This table shows annualized average returns, standard deviations, Sharpe ratios for portfolios sorted on the out-of-sample return prediction. We differentiate between the complete case method, unconditional mean imputation, and conditional mean imputation with GLS weighting without and with lags. Long Pf. and Short Pf. denote the annualized average return of the long and short leg respectively. Skewness and kurtosis are the sample statistics of the monthly returns. The implementation of the linear and nonlinear model is detailed in Section 5.2. The sample period is 1990-2021.

	Mean (%)	Standard Deviation (%)	Sharpe Ratio	Long Pf. (%)	Short Pf. (%)	Skewness	Kurtosis
Panel 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
Panel 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)	12.50	9.32	1.34	21.05	8.54	-0.64	6.01
Cond. Mean (GLS / LASSO / w.lags)	12.81	9.26	1.38	21.20	8.39	-0.60	6.01
Panel 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
Panel 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)	13.63	8.43	1.62	21.61	7.98	-0.49	4.56
Cond. Mean (GLS / LASSO / w.lags)	14.00	8.65	1.62	21.80	7.79	-0.20	6.12

A.2.2 Incremental information – Univariate model

Figure A.1: 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 characteristics 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, and the weighted estimator yields almost identical results. The characteristics are ordered according to their year of discovery.



A.3 Extensions

A.3.1 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 linear combination of the excess returns and we assume that the weights are a parametric function of $X_{it} \in \mathbb{R}^K$. That is

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

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 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,

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

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

Let

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

Assuming that

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

we obtain the conditional moment restrictions

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

To impute missing values, let

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

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

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

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

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

A.3.2 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.4 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:

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

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 $X_i^{(obs)}$, i.e. $X_i^{(obs)} = (Y_i, X_{i,1}, X_{i,2})$ 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

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

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

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

i.e. 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 must 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 is 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 (2022) 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) allows obtaining the statistical properties and valid standard errors of the parameters of interest.

A.5 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 as long as $E[\varepsilon_{it} \mid X_{it}^{(0)}, D_{it} = 0] = 0$, which follows from our previously imposed MAR assumption. Finally, for the moment conditions in equation (2), we can use our assumption $E[\varepsilon_{it} \mid X_{it}^{(l)}, D_{it} = 0] = 0$ to write

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

Hence, the moment conditions hold as longs as

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

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

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

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 the second part of the previous MAR assumption, namely

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

A.6 Equivalence GLS and Optimal GMM

Consider the moment conditions

$$
\begin{aligned}
E\left[\mathbf{1}(D_{it} = 0)\left(Y_{it} - \sum_{k=1}^{K} \beta_{t,k} X_{it,k}^{(0)}\right) X_{it}^{(0)}\right] &= 0 && && \text{(A.4)} \\
E\left[\mathbf{1}(D_{it} = l)\left(Y_{it} - \sum_{k=1}^{K} \beta_{t,k} Z_{it,k}^{(l)}\right) X_{it}^{(l)}\right] &= 0 && l = 1, \ldots, L && \text{(A.5)} \\
E\left[\mathbf{1}(D_{it} = 0)\left(X_{it,k} - X_{it}^{(l)\prime} \gamma_t^{(l,k)}\right) X_{it}^{(l)}\right] &= 0 && l = 1, \ldots, L \text{ and } k \notin I_t^{(l)} && \text{(A.6)}
\end{aligned}
$$

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

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

The last two conditions assume that the unobservables are homoskedastic.

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

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

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

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

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 0 because

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

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

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

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

The first-order conditions are

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

with

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

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

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

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

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

The first element is simply

$$-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(D_{it}=0)X_{it}^{(0)}X_{it}^{(0)\prime}\left(\hat{w}^{(0)}\right)^{-1} = -\left(\hat{\sigma}_{\varepsilon,t}^{2}\right)^{-1}I_{K\times K}$$

Next, we assume without loss of generality that the elements in $Z_{it}^{(l)}$ are ordered such that $Z_{it}^{(l)} = \left(X_{it}^{(l)\prime}, X_{it}^{(l)\prime}\gamma_t^{(l)\prime}\right)'$. 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

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

It follows that

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

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

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

Using the same arguments we can also write

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

Hence

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

Next, consider the GLS estimator, which minimizes

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

The first-order conditions are

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

Therefore

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

A.7 Large Sample Distribution

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

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

and

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

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

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

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

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

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.6 where $W_1 = \text{diag}\left(\hat{w}^{(l)}\right)^{-1}$ and expressions for $\hat{w}^{(l)}$ are provided in Section A.6. We obtain the two-step OLS estimator when W_1 is an identity matrix. Using a first-order Taylor expansion, we get

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

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

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

and thus

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

where

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

where

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

and

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

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 J-test

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

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

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

and

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

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

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

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

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

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

To test our overidentifying restrictions, we test

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

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

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

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

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

Hence,

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

Under the null hypothesis it holds that

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

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

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

$$= G_t^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} g_{it,11}(\beta_t) \\ g_{it,2}(\gamma_t)\end{pmatrix} + o_p(1)$$

where

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

Hence

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

where

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

It follows that

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

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

Hence,

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

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

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

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

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

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