Summary of *Missing values handling for machine learning portfolios*

Andrew Y. Chen, Jack McCoy (JFE, 2024)

Summarized by Li Ziming

1. **What are the research questions?**
   - Where does missingness come from?
   - How missing values in predictors affect the performance of ML algorithms?
   - Which methods of handling missing values work best for predicting stock returns?

2. **Why are the research questions interesting?**
   - The problem of missing values is serious when applying ML to asset pricing.
   - Standard practice of dropping stocks with missing values is untenable, ML researchers often have no choice but to impute missing values.

3. **What is the paper's contribution?**
   - Contribute to the literature that studies missing values in cross-sectional predictor data.
     - Prior literature: Use moment conditions (Freyberger et al., 2023); latent factor method (Bryzgalova et al., 2023); masked language model (Beckmeyer and Wiedemann, 2023).
     - Extend: Take a more neutral approach that comparing textbook imputation methods with the method used in asset pricing.
   - Contribute to the literature that applies ML to asset pricing.
     - Prior literature: Combine large sets of predictors imputes with means or medians, algorithms are complex (Gu et al., 2020; Freyberger et al., 2020; Kozak et al., 2020).
     - Extend: Provide a rigorous justification for the ubiquitous use of mean imputation.

4. **What hypotheses are tested in the paper?**
   - H1: Cross-sectional mean imputation performs similarly to more sophisticated methods, such as EM, in terms of predicting stock returns.
   - H2: Complex imputation methods introduce estimation noise that can lead to underperformance, particularly in models involving small-cap stocks.

   a) Do these hypotheses follow from and answer the research questions?

   - Yes.

   b) Do these hypotheses follow from theory? Explain logic of the hypotheses.

   - The assumption that mean imputation performs well arises from the observation that predictor correlations are low and missingness occurs in large blocks. Therefore, sophisticated imputations introduce unnecessary noise, which supports the second hypothesis.

5. **Sample: comment on the appropriateness of the sample selection procedures.**

   - The inclusion of 159 predictors based on analyst forecasts and accounting data helps create a realistic scenario for missing data issues, making the findings relevant to a broad range of financial applications.

6. **Comment on the appropriateness of variable definition and measurement.**

   - The paper uses a clear methodology to categorize and impute missing values, ensuring consistency across multiple datasets.

7. **Comment on the appropriateness of the regress/predict model specification.**

   - The use of common models like PCR, neural networks, and gradient boosting trees ensures that the findings are not limited to a specific algorithm.

8. **What difficulties arise in drawing inferences from the empirical work?**

   - The paper assumes that missing data is ignorable and occurs at random. But this assumption may not hold in all contexts, which could affect the validity of some conclusions.

9. **Describe at least one publishable and feasible extension of this research.**

   - How missing data imputation techniques perform in time-series settings, particularly in other asset classes like bonds or commodities. This could extend the understanding of missing value handling beyond stock returns and provide insights into broader financial applications.