Summary of *Missing values handling for machine learning portfolios*

Andrew Y. Chena, Jack McCoy, JFE, 2024

2024.10.16 喻清言

1. **What are the research questions?**
   - Which method should be applied when handling missing value to machine learning asset pricing?
2. **Why are the research questions interesting?**
   - When learning from large scale of predictors to analyzing asset pricing, the standard practice of dropping stocks with missing value is often untenable.
   - Since ML researchers have no choice but to impute missing value, it is essential to analyze which imputation method should be applied.
3. **What is the paper' s contribution?**
   - the methodology of imputation missing values for machine learning portfolios.
     - Existing practice: applying ad-hoc imputation or data adjustments, with little discussion of their motivation or study of alternatives..
     - Extension：recommending using simple mean imputation for ML studies.
4. **What hypotheses are tested in the paper?**
   - Simple cross-sectional mean imputation method outperforms EM method.

   a) Do these hypotheses follow from and answer the research questions?
   - Yes, it is comparing the two methods of imputing missing values.

   b) Do these hypotheses follow from theory? Explain logic of the hypotheses.
   - Three facts about predictor data reveals that observed predictors provide little information about the missing predictors, thus EM imputation may introduce extra estimation noise than simple mean imputation.
5. **Sample: comment on the appropriateness of the sample selection procedures.**
   - The structure of the sample data is thoroughly analyzed to strengthen the conclusion and provide evidence for explanation.
6. **What difficulties arise in drawing inferences from the empirical work?**
   - The comparison between two imputation methods is multidimensional thus the conclusion is tenable. For robustness, four other imputation methods are added to the comparison.

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

    - The method can be applied to imputing missing values on all large-scale cross-sectional analysis on asset pricing.