Summary of Missing values handling for machine learning portfolios

Andrew Y. Chen , Jack McCoy

Journal of Financial Economics

October 15, 2024 Yanrui Zhou

1. **What are the research questions?**
   - How to handle missing values for portfolios constructed using ML?
2. **Why are the research questions interesting?**
   - Missing value is a serious problem in applying ML methods in asset pricing.
   - Studying how to solve this problem can help researchers study better.
3. **What is the paper's contribution?**
   1. Papers studying missing values in cross-sectional predictor data.(Freyberger et al., 2023)
      - **Past studies:** advocates different imputation algorithms.
      - **Expand:** study the reasons of missing values and whether imputation works well.
   2. Papers using ML and large sets of predictors in finance.(Gu et al., 2020)
      - **Past studies:** imputes predictors with cross-sectional means or medians.
      - **Expand:** provide a rigorous justification for the use of mean imputation.
   3. Debate about the factor structure of the cross-section of returns.(Green et al., 2014)
      - **Past studies:** there are many dimensions of predictability.
      - **Expand:** complement Lopez-Lira and Roussanov (2020).
4. **What hypotheses are tested in the paper?**
   - H1: EM and simple mean work well while imputing missing data.
   - H2: while using other imputation and forecast methods, results hold.

   a) Do these hypotheses follow from and answer the research questions?
      - H1 answered research question well and H2 studies furtherly.

   b) Do these hypotheses follow from theory?Explain logic of the hypotheses.
      - Theory: imputation methods are commonly used to treat data and work well.
      - Logic: H1 tests easy methods and H2 tests complicated methods.
5. **Sample: comment on the appropriateness of the sample selection procedures.**
   - Predictors data begins with the August 2023 release of the CZ dataset.This release contains 212 cross-sectional predictors published in academic journals, it's proper.
6. **Comment on the appropriateness of variable definition and measurement.**
   - The variables are stock return and factors, which are commonly used in papers before.
7. **Comment on the appropriateness of the regress/predict model specification.**
   - The predict model is specified well and could be used for return prediction.
8. **What difficulties arise in drawing inferences from the empirical work?**
   - This paper excludes too many factors which may influence the result of the paper.
9. **Describe at least one publishable and feasible extension of this research.**
   - This article could study whether we can use complicated methods of imputation works while excluding estimation noise?