Summary of Missing values handling for machine learning portfolios

JFE 2024

Feng Lixuan 20241016

1) What are the research questions ?

- How to effectively handle missing values in machine learning portfolios?

2) Why are the research questions interesting?

- Applies ML to asset pricing often encounters missing values problem.
- Drop or impute missing values are often untenable or dangerous.
- Selecting an appropriate method to handle missing values is crucial.

3) What is the paper's contribution?

- Literature on missing values in cross-sectional predictor data.
  - Prior: avoid modeling the missingness process.
  - This study: discussing the origins of missingness.
- Literature on predictor and stock return datasets.
  - Prior: 80+ predictors from the CZ open-source dataset,...
  - This study: size and breadth of datasets, diversity of statistical methods.
- Literature on using mean imputation in ML.
  - Prior: Combine large sets of predictors.
  - This study: simple imputation algorithm is helpful for transparency.

4) What hypotheses are tested in the paper?

- Hypotheses
  - Simple mean imputation for handling missing values outperforms the Expectation Maximization (EM) algorithm.
- The logic of hypotheses
  - Structure of missingness; low cross-sectional correlations; estimation noise.

5) Sample

- 159 predictors include CRSP, Compustat, and IBES analyst forecast data.

6) Regression/prediction model specification

- Baseline forecasts: principal component regressions (PCR).

7) What difficulties arise in drawing inferences from the empirical work?

- The existing results still struggle to explain the issue of dimensionality.

8) Describe at least one publishable and feasible extension of this research.

- Research on missing value handling methods based on panel data.