Using Ensamble Learning Techniques for Missing Data Imputation. An Application to Permanent Household Survey, Argentina

Germán Rosati

german.rosati@gmail.com

UNSAM - UNTREF - Digital House

Noviembre de 2017

PyData - San Luis

1. Introduction

- Missing data is a recurrent problem in statistical analisys afecting official statistics (household surveys, census data, etc.), administrative records and any other dataset
- New data sources (from internet traffic, mobile devices applications, web scraping, etc.) take the missing data problema to a new scale, making necessary the generation of tools and methdos which allow to deal efficently with the existance of missing data

2. Objectives

- Present a first imputation model based on Machine Learning Techniques (LASSO regression and Bagging Ensamble)
- Test and show their results and performance in a particular case: imputation of labor income variables in Permanent Household Survey (Argentina, II Quarter 2015) collected by INDEC (National Institute for Statistic and Censuses)
- Outline future lines of research

2. Missing Data Generation Process

- **Missing Completly at Random (MCAR):** the probability for a registry to have a missing data in variable $\boldsymbol{Y}$ is not related neither with the values of $\boldsymbol{Y}$ and the values of the rest of the variables in the matrix $\boldsymbol{X}$. Missing datapoints are a random subsample of the original dataset. This assumption is violated if: a) some group or subgroup have higher probability of presenting missing data; b) some of the values of $\boldsymbol{Y}$ shows higher probability of missing data
- **Missing at Random (MAR):** each registry probability of being missing is independente of the values of $\boldsymbol{Y}$, after conditionning on other variables
- **Non Missing at Random (NMR):** the missing data probability depends both on external $\boldsymbol{X}$ variables and on $\boldsymbol{Y}$ values

3. Usual methods for missing data imputation

- **Unconditioned means:** the mean of valid and complete cases in the variable $\boldsymbol{Y}$ is used as an imputation value. MCAR pattern is assumed. Problems: variance is reduced and hence confidence interval are artificially narrowed
- **Conditioned means:** groups are formed using features correlated with the interest variables. The mean within each group is used as an imputation value. MCAR pattern is assumed. Equivalent to impute values using a linear regression
- **Hot deck:** missing values are replaced in non respondent cases ("receptors") with values observed in valid cases ("donors"). This donors are similar to receptors and can be selected randomly (random hot-deck) or be constrained in someway (deterministic hot-deck)

4. Methods

The proposed method generates several estimations for the missing values and adds them in order to generate the final imputation.

- **LASSO Regresion** It is a linear regression, but instead of perform the minimization of the RSS (Residual Sum of Squares), the aim is to minimize the following function:

$$\sum_{i=1}^{n}(\boldsymbol{y_i} - \beta_0 - \sum_{j=1}^{p}\beta_j \boldsymbol{x_{ij}})^2 + \boldsymbol{\lambda}|\beta_j| = \boldsymbol{RSS} + \boldsymbol{\lambda}|\beta_j| \qquad (1)$$

- **Ensamble Learning: Bagging**
  1. A dataset with complete cases is created *TrS*
  2. Another dataset with missing data is created *TeS*
  3. Number of iterations *rep* is fixed
  4. For r between 1 and *rep*
     1. A bootstrap sample (SRS with replacement and $\boldsymbol{n} = \boldsymbol{n^*}$) is extracted from *TrSet*
     2. In the generated sample a LASSO regression is estimated
     3. With the parameters of the previous step prediction in *TeS* are made
  5. After *rep* iterations *rep* predictions of missing data in *TeS* are generated and aggregated (using median)

5. Tests performed

1. **Test 1:** Results obtained by the Bagging-LASSO method were compared with results in imputation performed by the INDEC (using hot-deck imputator). The parameters used for hot-deck method were unknown
2. **Test 2** Missing data were randomly generated and the performance of both methods were assesed

- A "worst case scenario" approach was used:
  - SRS were applied (no stratifiacation)
  - gross filter of outliers in the first test and no filter in the second test
  - no state-level models were generated
- In both tests the model tested asummed the form

$$\boldsymbol{CF} = \boldsymbol{log_{10}(y_i)} = \beta_0 + \sum_{j=1}^{p}\beta_j \boldsymbol{X_{ij}} + \boldsymbol{e_i} \qquad (2)$$

- $\boldsymbol{y_i}$ labor income of individual and $\boldsymbol{\beta_j}$ regression coefficients
- minimizing $\boldsymbol{RSS} + \boldsymbol{\lambda}|\boldsymbol{\beta_j}|$

6. Main results

- **Test 1**

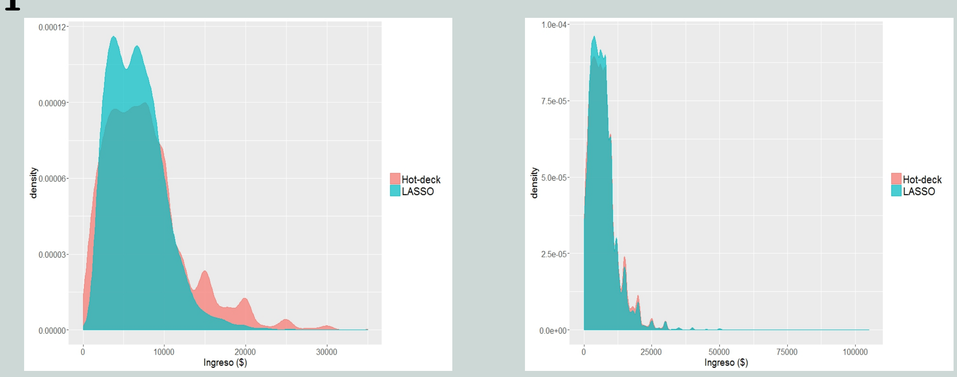
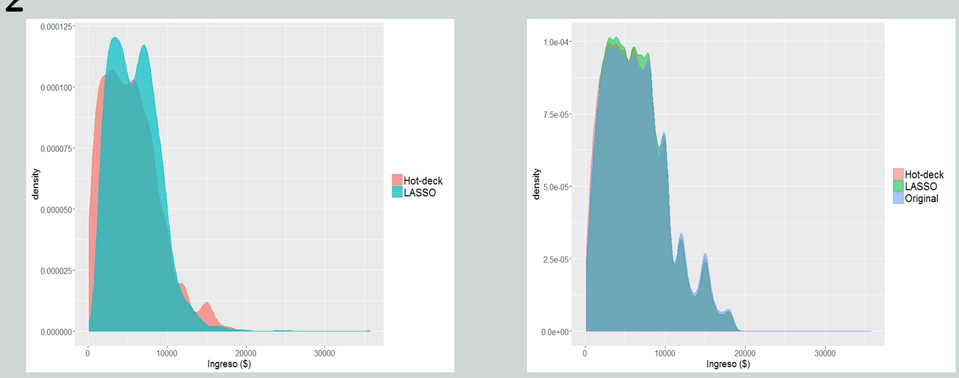


Figure 1. Missing, complete and original data distributions by imputation method

- Missing data distribution: "Thick tails" in hot-deck estimation
- Complete cases distribution: very similar in both methods

Test 2



Test 2. Missing and complete data distributions by imputation method

- k-fold CV ($\boldsymbol{k = 9}$) was performed to evaluate both methods performance (Rooted Mean Squared Error -RMSE-)
  1. RMSE - Bagging-LASSO = \$3.994
  2. RMSE - Hotdeck = \$4.933
- 20% reduction in RMSE. Bagging-LASSO outperforms hot-deck imputation in labor income variable prediction
- Smaller variability in bagging-LASSO imputations
- Smaller differences in central values
- Wider spreads in distribution tails

7. Discusion and future work

- Small difference with imputed case distribution generated by INDEC (with hot-deck method)
- 20% reduction in RMSE using bagging-LASSO method
- Better response of bagging-LASSO in presence of extreme cases
- Automatic feature selection and regularization which allows a more comprehensive theoretical approach
- Bagging-LASSO does not seems so senstive to sampling variability =¿ random selection of features (similar as Random Forest)
- Evaluation of other ensamble algorithms (boosting)
- Development of a library with bagging-LASSO immplementation (R & Scikit-Learn) and improvement in computing performance

Original Paper

- Rosati, G. (2017). Construcción de un modelo de imputación para variables de ingresos con valores perdidos a partir de Ensamble Learning. Aplicación a la Encuesta Permanente de Hogares, *Revista SaberES*, 9 (1), 91-111.