Project 3

Bhattacharya Ankita, Bertucci Adrien,
Aline Sherman , Kang Shin Heuk

November 29, 2019

(From left) Ankita, Adrien, Sherman, Shin

Overview

Preliminary work

Purpose of the project

In the case of 40% missing data with MCAR (missing completely at random) mechanism, we want to compare three following imputations: 1) Single regression imputation, 2) Bootstrap multiple imputation, 3) Iterative Principal Component Analysis (PCA) imputation.

Base work before the imputation

- Step 0: Log transform wage (Henceforth, *waget*).
- Step 1: Generate 40% of waget missing data with MCAR missing mechanism $M = 20$ times.
- Step 2: For the comparison, compute the estimated bias and the estimated variance for each method.

$$waget_{missing} = \hat{B}_0 + \hat{B}_1 educ + \hat{B}_2 exper$$

Single regression imputation: Methodology

[IDEA] Use predicted values from the log-linear regression in order to impute the missing values.

Suppose that we predict the missing values of log(wage) - *waget* by linear regression.[1]

- Step 1: We build a model from the observed data.
- Step 2: Predictions for the incomplete cases are then calculated under the fitted model and are imputed in place of the missing data.
  This preserves the relation between log(wage), educ and exper : an advantage over mean imputation. Btw the formula for estd. variance does not look correct to me - maybe I am wrong

$$waget_{missing} = \hat{B}_0 + \hat{B}_1 educ + \hat{B}_2 exper \tag{1}$$

[Imputation performance]

Estimated bias
$\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^m_{educ} - \hat{\beta}^C_{educ}) \approx 0.1340144$

Estimated Variance
$\frac{1}{M}\sum_{m=1}^{M}\widehat{var}(\beta^m_{educ}) \approx 0.07118289$

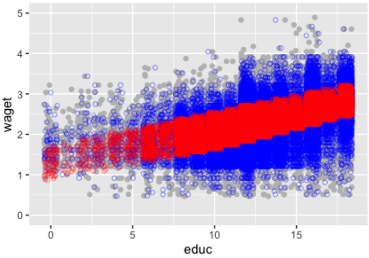
Both bias and variance are relatively large compared to our later tests.

[1]We used *regressionImp* (VIM package) which directly imputes missing values in *waget* by the predicted values. tilde not showing below.
$RegImp\_list < -regressionImp(waget \sim educ + exper, data = amputed\_list)$

Single regression imputation: Drawbacks [Q1]

- In using fitted values, single regression imputation disregards the error term around the coefficients of educ and exper, which leads to an overestimation of the correlation between the explained and explanatory variables. We are thus likely to have biased parameters of regression (Tsikritsis, 2005).
- Naturally, the coefficients from the regression imputation data would have a lower estimated variance as a result.
- There are a few advantages of single regression imputation. For example, it can be used when the data contains highly correlated variables. See Lodder (2013) for details regarding advantages of single regression imputation.

Regression imputation underestimates variance and overestimates correlation. This is visible in our plot, as the imputed values are highly correlated with educ, and their spread is not as large as teh original (blue) data.
We can also observe that regression imputation fails to replicate any heteroskedasticity in the data.



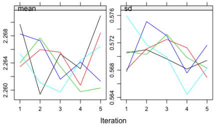
red: imputed data, blue: amputed data, gray: original dataset

Bootstrap Multiple Imputation: Methodology [Q2]

[IDEA] Impute multiple times by chained equations.[2]

MICE, the package we use, uses iteration for each imputation.

- Step 1: Impute using bootstrap regression. A new dataset is created using nonparametric bootstrap.
- Step 2: Run another bootstrap regression on the imputed data, and regression imputation is done for the missing values.
- Step 3: Repeat Steps 1 & 2 until convergence.
  There is no clear method of knowing if MICE algorithm has converged (Buuren), so we run 5 iterations.

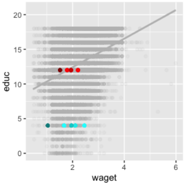


mean & var for each iteration

Multiple Imputation[Q2, Q3]

Multiple imputation creates several datasets of values to replace missing data.

- Benefit: Multiple imputation allows a level of uncertainty for each missing value (Graham, 2009). Variance is measured within a dataset (points with same shade) as well as variance between datapoints (same color).



two sets of imputations

[2]We use "*mice*" package and the command *mice(dataset, m=B, method="norm.boot")*

Bootstrap cont. & Results [Q2]

Nonparametric Bootstrap Regression

- Bootstrap samples: iid samples $X = (X_1, ..., X_n)$of size n are generated by drawing independent observations with replacement from the dataset.
- From each sample, a linear regression is ran, and imputed values are generated.
- Benefit: Nonparametric bootstrap decreases bias induced by patterns in missingness. This is because sampling with replacement creates missingness that is independent.

[Imputation performance]

$\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}_i^m - \hat{\beta}_i^C) \approx -0.081863704$

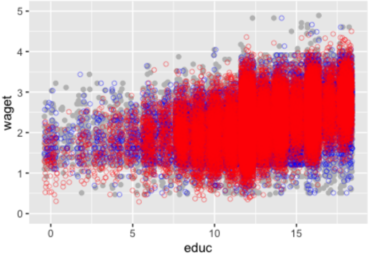
$\frac{1}{M}\sum_{m=1}^{M}\widehat{var}(\beta_i^m) \approx 0.0001871368$

The heteroskedastic variance fits the amputed datapoints better than regression imputation.

Calculation of Variance in Multiple Imputation:

$\beta_i^m = \frac{1}{B}\sum_{b=1}^{B}\hat{Var}(\hat{\beta}_b) + (1+\frac{1}{B})\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_b - \hat{\beta}\right)$

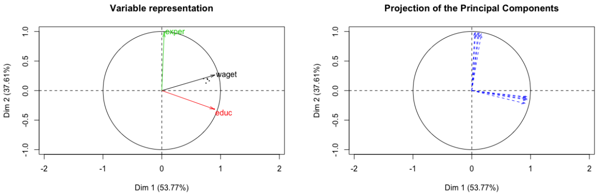
this formula accounts for variance between imputations and variance between observations



Principal Component Analysis [Q3]

[IDEA]

- PCA aims to create a set of components (principal components) such that the variance between components is as large as possible while the distance between components and original data is minimized.
- These components correspond to the imputed values we will generate.



- Observe: if we took the mean of the projected principal components, they would be equal to the waget vector in the variable representation. This is the minimization of distance between the data and imputations.
- Observe: The principal components are as far away from each other as possible while still being correlated with the other variables.
- For high-dimensional data, PCA will use dimensionality lower than that of the data if several variables are highly correlated.

Methodology: Relationship Between PCA and SVD[3] [Q3]

1) PCA: Principal Component Analysis

Principle Component Analysis uses SVD to calculate a set of eigenvectors for covariance, referred to as principle components.

- organize data into an n x m matrix, n=observations, m = variables. (for us, n=74661 m=3)
- subtract mean for each variable, in order to normalize for SVD
- Calculate the SVD

PCA Assumptions

• Linearity
• Low variance is noise, high variance is structural (strong assumption, often incorrect)
• Principle components are orthogonal: allows linear algebra techniques like SVD to be used

2) SVD: Singular Value Decomposition

SVD is a method of computing an orthonormal basis V for the data we have. This orthonormal basis is used to generate values for the missing values.

- Data: X, an $n \times m$ matrix where $X^T X$ has rank r (the number of principal component).
- Find $\hat{v}_i$, a set of orthonormal eigenvectors s.t. $(X^T X)\hat{v}_i = \lambda_i \hat{v}_i$.
- Then $\sigma_i = \sqrt{\lambda_i}$ is the *singular values*.
- Let $\hat{u}_i = \frac{1}{\sigma_i} X \hat{v}_i$, and denote $V = [\hat{v}_1 \quad \hat{v}_2 \dots \hat{v}_r]$ and $U = [\hat{u}_1 \quad \hat{u}_2 \dots \hat{u}_r]$
- Decompose $X = U \Sigma V^T$ where Σ is a diagonal matrix with descending diagonal order $\Sigma_{11} \geq \Sigma_{22} \geq \cdots \geq \Sigma_{rr}$

[3] Theoretically, single value decomposition (SVD) is an ideal method to do PCA on the condition that we only care about the numerical accuracy (Shlens, 2014).

Iterative and Regularized PCA

[IDEA] Iterative PCA

Iterative PCA uses Bayesian probability while iteratively re-generating principal components, allowing it to find components which maximize variance while remaining orthogonal.[4] Process:

- 1) mean imputation is done to generate an initial set of imputated values.
- 2) PCA-imputation is performed on the original data, using the previous imputed values as a posterior distribution.
- 3) Step 2 is repeated until the values converge.

[IDEA] Regularized PCA

Regularized PCA shrinks imputation steps by multiplying them by the percent difference between the singular values and the estimated variance.

PCA: $\hat{\mu}_{ij}^{PCA} = \sum_1^s \lambda_s u_{is} v_{js}$ Regularized: $\hat{\mu}_{ij}^{rPCA} = \sum_1^s \left(\frac{\lambda_s - \hat{\sigma}^2}{\lambda_s}\right) \lambda_s u_{is} v_{js}$

[4] Information contained in the help file for MIPCA function of MissMDA package

Methodology: Dimension Estimation

Algorithm

- In order to apply PCA, we must specify the number of components for the space. In high dimensional datasets this is an important step, as reducing dimensionality can improve performance.
- To estimate component number, we consider 1) The cumulative percentage of variance must be greater than 70%, 2) The eigenvalue of the new component must be greater than 1.

	Eigenvalue	**Cumulative % of variance**
1 component	1.4613	48.71139
2 component	1.1624	87.45710
3 component	0.3763	100.00

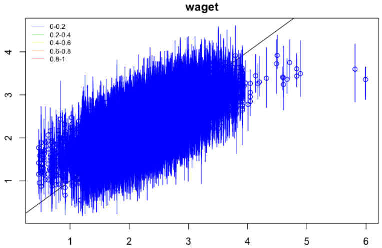
- With determined 2 component ($ncp = 2$), compute eigenvectors from the covariance matrix.

[Imputation performance]

$\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}_i^m - \hat{\beta}_i^C) \approx -0.08186697$ $\frac{1}{M}\sum_{m=1}^{M}\widehat{var}(\beta_i^m) \approx 0.0001671353$

PCA Imputation: Results[5]

X axis: original value Y axis: imputed value and confidence region. Blue means our confidence interval is for 80% confidence interval of or higher



[Advantages]

- PCA can be helpful in identifying patterns for large datasets, thanks to dimensionality reduction.
- Best representation of variance, as it accurately models the variance between individual datapoints.
- PCA is nonparametric, so it has flexibility in use cases.
- Regularized PCA can help prevent overfitting

[Disadvantages]

- Iterative PCA can have overfitting issues when there are too many parameters, or the level of missingness or noise is too high.
- Strict assumptions which may not be true, in which case the results are not valid.

[5] We refer mostly from Shlens (2014).

Conclusion

Performance comparison

We compare the estimated bias and variance of *Educ* coefficient.

Imputation methodology	**Estimated bias**	**Estimated variance**
Regression (single)	0.1340	0.07118
Bootstrap (multiple)	-0.0819	0.00019
Iterative PCA (multiple)	-0.0819	0.00017

Performance (the best to the worst): $PCA \approx Bootstrap \succ Regression$

- Both PCA and bootstrap performs significantly better imputation than single regression imputation.[6]. We conclude that with MCAR mechanism with 40% missing data, using PCA and Bootstrap is preferred to single regression imputation.

Additional remark

- It may be a better idea to not use these imputation especially when a strong prior is known (i.e. Death from horse kick is likely to have Poisson distribution as a strong prior.).
- When there is no time constraint to complete PCA, always choose SVD as it produces more accurate principal components

[6] $100\frac{(0.1334-0.08)}{0.1334} \approx$ 38.6% smaller bias, and $100\frac{(0.07063-0.0001)}{0.07063} \approx$ 99.9% smaller variance.

References

- Graham, J. W. (2009), "Missing data analysis: Making it work in the real world," *Annual review of psychology*, 60, 549–576.
- Jonathon Shlens (2014), "A Tutorial on Principal Component Analysis," *Google Research.*
- Nikos Tsikriktsis (2005), "A Review of Techniques for Treating Missing Data in OM Survey Research", *Journal of Operations Management*, 24, 53-62.
- Stef van Buuren (2012), *"Flexible Imputation of Missing Data"*, Chapman & Hall Interdisciplinary Statistics Series.