Missing Data, PLS and Bootstrap:
A Magical Recipe?

Cordeiro, C.; Machás, A. and Neves, M.

The user R Conference, Wien, Austria,
June 15-17, 2006

Research questions...

- Could missing data method change the quality of the results obtained from a Customer Satisfaction market study?
- Could standard or classical imputation methods be applied no matter the rate of non responses?
- Could Bootstrap improve quality of estimates?

Missing Data

- Standard practices to treat non-responses are not statistically justified and could result in biased estimates
- Data imputation methods are used for reconstructing the incomplete data to obtain a complete data set to produce more accurate estimates.
- Most common methods to treat missing data are:
  - Mean imputation
  - Listwise deletion
  - Pairwise deletion
  - Maximum Likelihood

Missing Data Methods

IMPUTATION METHODS	⇨Mean, Modal and Median ⇨Nearest Neighbour (NN)
MODEL BASED METHODS	⇨Multiple Imputation (MI) ⇨Maximum Likelihood (ML) ⇨Expectation Maximization (EM)

Missing Data and Bootstrap

Efron(1994) uses the extensive imputation theory developed by Rubin(1987)

The simplest nonparametric bootstrap approach:

⇨The rows in the original data matrix are resampled with replacement

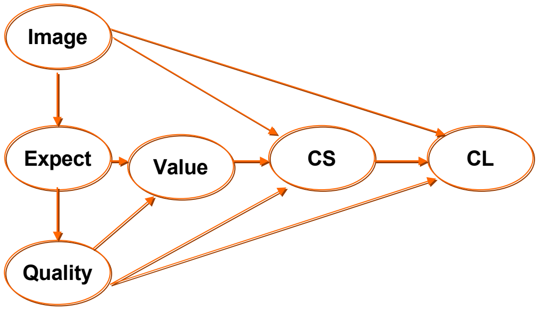
⇨A bootstrap matrix is obtained and a bootstrap estimate is calculated for the parameter in study

So an extensive computer work is performed, repeating the above procedure several times; a large number of estimates are calculated and imputed in the original data.

Case Study: Bootstrap and SEM-PLS on CSM

- ACSI Model for Mobile Telecom (Fornell, C)
- SEM estimated with PLS algorithm (Chin, W)
- Data treatment for missing data: Standard procedure Mean imputation

➢ **STRUCTURAL MODEL**



➢ **MEASUREMENT MODEL** (number of questions)

Image	5
Expectations	3
Quality	8
Value	2
Customer Satisfaction	3
Customer Loyalty	2

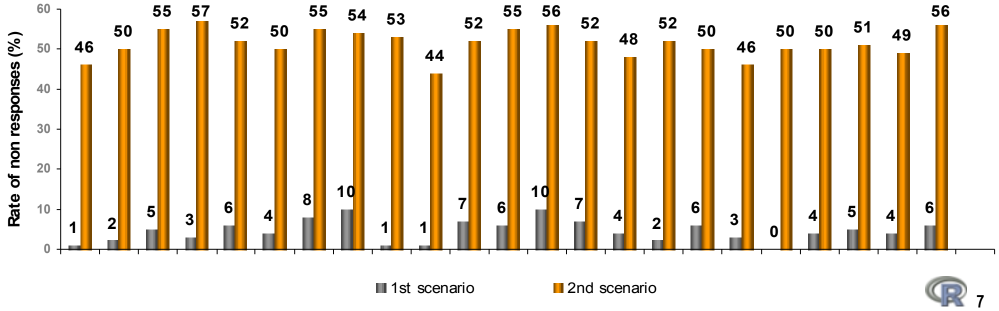
➢ **ACSI Model,** with Image like in **EPSI Model**

Methodological aspects

Compared scenarios:

❖ 10% Rate of non responses from Original Data Matrix X = 1st scenario

❖ 50% Rate of non-responses from Simulated Data Matrix Y= 2nd scenario



Using R

Bootstrap application in R

Step1: matrix rows are resampled with replacement;

Step2: a bootstrap sample is obtained;

Step3: a bootstrap estimate is computed according to the missing data method;

Step4: go to step1.

Using R

⇨This procedure was repeated r=5000 times;

⇨ Missing values in scenarios 1 and 2, are replaced with new estimates generated by 5000 replications;

⇨Then, a new PLS estimation is performed.

Both scenarios, using bootstrap methodology, were compared with the classical situation (CSM estimation based on PLS, where Mean Imputation is the ad hoc procedure adopted for ECSI/EPSI model).

Case Study questions...

? How the classical missing data techniques perform for the two scenarios

? How the Bootstrap perform with the missing data techniques for the two scenarios

? What conclusion based on quality measures of model adjustment like: RSquared, Residual Variance….

Conclusion

⇨ 1st Scenario (10%): Bootstrap methodology doesn't increase the quality of estimates

⇨2nd Scenario (50%): Bootstrap methodology used with Hot Deck Imputation and K Nearest Neighbor achieves good results

Overall, it was seen that for a higher non- response rates, bootstrap is the best method to be adopted in case of missing data completely at random.

The work still goes on...

⇨ Perform an extensive theoretical work

⇨ Improve some performance methods

⇨ Explore other bootstrap approaches to the estimation in the problem of missing

THANK YOU