European Study Croup with Industry'113 (ESGI'113)
September 14–18, 2015, Sofia, Bulgaria

Problem 2. Development of mathematical algorithm for direct ascription of missing values in survey research data

GemSeek, www.gemseek.com

Martin Dimov, martin.dimov@gemseek.com

Company's overview. GemSeek is a market intelligence and consulting company. It helps business leaders with decision support analytics that have a direct impact on bottom line and competition. Company's services are organized around Data science and predictive analytics, Market & Industry Intelligence, Customer Insight & Brand Analytics, Advanced Visualization Solutions and Competitive Intelligence.

Definition of the problem. One of Gemseek's core activities is developing and implementing marketing survey research studies among different target groups both on local market and across the world. The results serve as basic foundation for further analysis on customer perceptions, behavior and brand affiliation. Hence, the necessity of complete datasets is a prerequisite for sustainable analyses, robust analytics and unbiased interpretation of results. One of the biggest challenges for company was dealing with "blank spots" in the data i.e. places where respondents refrain from providing correct answering due to various reasons. Some of these include difficulty to find correct answer, too long questionnaires, unwillingness to disclose sensitive personal information (income, age etc.), too many options to choose from etc.

Since most statistical analysis methods assume the absence of missing data and are only able to include observations in which every variable is measured, GemSeek is in need of a robust mathematical approach that could impute incomplete data sets so that analyses which require complete observations can appropriately use all the information present in a dataset without missingness. In this case the level bias and incorrect uncertainty estimates will be avoided.

Task description. In 2014 the company has performed a study among 600 customers of the biggest supermarket chains in Bulgaria. The methodology used

random sampling procedure among population in Bulgaria's top 8 cities. The variables were measured with different type of scales: nominal, ordinal and continues in some of the cases. As a result the final dataset contained a large number of missing cases and "no answers" across variables ranging from 5% to around 50% of all respondents interviewed.

Since all methods for stimulating response rate were exhausted GemSeek is looking for a **computational algorithm** that could use the information from already completed cases and recursively assign values to missing data in every variable controlling for the type of scale and distribution of "real" values. For this exercise we assume that all missing values are of type: Missing At Random (MAR).

Expected results

- Brainstorm on various methods of solving the task;
- Presentation of different algorithms, stating pros and cons for each one;
- Used variables, predictors, distance measures, parameter estimates etc.;
- Suggestions of appropriate software and tools, complete scripts and developer codes for completing the task;
- Discussion of the results with bigger audience.

Assessment criteria. All suggestions for algorithms will be closely reviewed and assessed by Gemseek team. Following criteria will be used when choosing the most effective method:

- Accuracy – measured as percentage of accurately imputed cases vs.real cases;
- ROC curves and Confusion matrices – as a way of graphical visualization of accuracy;
- AIC and BIC – as a method for comparing different methods and their efficiency;
- Statistical significance and hypothesis testing – as non-parametric estimation of results.

Materials provided

- Incomplete data set of survey results in CSV format which could be used for imputation methods and comparing the results;
- Complete questionnaire containing names and labels of all variables in the dataset.

Other information and resources and consultations will be readily available from Gemseek team on request.