Direct Ascription of Missing Categorical Values in Survey Research Data

Vasil Kolev, Veska Noncheva, Venelin Valkov
Elica Ilieva, Maria Dobreva

Abstract

The complete datasets are a prerequisite for sustainable analyses, robust analytics and unbiased interpretation of results. Missing values in a survey occur when no data value is stored for the variable in an observation. Missing data can have a significant effect on the conclusions that can be drawn from the data. Direct ascription is the process of replacing missing data with predicted values. The aim of this work is to describe an approach to direct ascription of missing categorical values in survey research data based both on the assumption that values in a data set are missing at random and on the implementation of the correspondence analysis.

Key words: correspondence analysis, supplementary points

1. Introduction

One of the biggest challenges in marketing survey research studies is 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, every company developing and implementing marketing survey research studies 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.

Until the 1970s, missing values were handled primarily by editing. Rubin developed a framework of inference from incomplete data that remains in use today [7]. The formulation of the expectation-maximization (EM) algorithm made it feasible to compute maximum likelihood (ML) estimates in many missing-data problems [1]. Rather that deleting or filling in incomplete cases, ML treats

the missing data as random variables to be removed from (i.e., integrated out of) the likelihood function as if they were never sampled. Many examples of EM were described by Little and Rubin [4]. Their book also documented the shortcomings of case deletion and single imputation, arguing for explicit models over informal procedures. About the same time, in [8] Rubin introduced the idea of multiple imputation (MI), in which each missing value is replaced by two or more simulated values prior to analysis [3]. Creation of MIs was facilitated by computer technology and new methods for Bayesian simulation discovered in the late 1980s [9]. ML and MI are now becoming standard because of implementation in free and commercial software [10].

2. Definition of the problem

In 2014 a market intelligence and consulting 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 the company 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 study we assume that all missing values are of type: Missing At Random (MAR).

3. Introduction to correspondence analysis

Correspondence analysis (CA) represents yet one more method for analyzing data in contingency tables and can be regarded as a special kind of canonical correlation analysis [2]. The main purpose of CA is to reveal the structure of complex data matrix by replacing the raw data with a more simple data matrix without losing essential information. CA makes it possible to present the results visually, that is, as points within a space, which facilities interpretation. CA is a method especially for analysis of large contingency tables. The technique is a tool to analyze the association between 2 or more categorical variables by representing the categories of the variables as points in 2D or 3D.

Correspondence analysis was developed in France and is more commonly used in Europe than in North America. Correspondence analysis is a descrip-

tive/exploratory technique designed to analyze two-way and multi-way tables containing measures of correspondence between the row and column variables. The results produced by correspondence analysis provide information which is similar to that produced by principal components or factor analysis. They allow one to explore the structure of the categorical variables included in the table. Correspondence analysis seeks to represent the relationships among the categories of row and column variables with a smaller number of latent dimensions. It produces a graphical representation of the relationships between the row and column categories in the same space.

Correspondence analysis was initially proposed as an inductive method for analyzing linguistic data. From a philosophy standpoint, correspondence analysis simultaneously processes large sets of facts, and contrasts them in order to discover global order; and therefore it has more to do with synthesis (etymologically, to synthesize means to put together) and induction. On the other hand, analysis and deduction (viz., to distinguish the elements of a whole; and to consider the properties of the possible combinations of these elements) have become the watchwords of data interpretation. It has become traditional now to speak of data analysis and correspondence analysis, and not data synthesis or correspondence synthesis.

Correspondence analysis is applied to two-way tables of counts. CA can be seen as a special case of canonical correlation analysis. It seeks scores for the rows and columns which are maximally correlated. As in principal component analysis, the aim of correspondence analysis is to reduce the dimensionality of a data matrix in order to visualize it in a subspace of low dimensionality, commonly two- or three- dimensional ([2], [5], [6]).

To summarize the theory of CA, first divide the $\mathbf{I} \times \mathbf{J}$ data matrix, denoted by $\mathbf{N}$, by its grand total n to obtain the so-called correspondence matrix $\mathbf{P} = \mathbf{N}/n$. Let the row and column marginal totals of $\mathbf{P}$ be the vectors $\boldsymbol{r}$ and $\boldsymbol{c}$ respectively, that is the vectors of row and column masses $\boldsymbol{r} = \mathbf{P}1$, $\boldsymbol{c} = \mathbf{P}^\top 1$, where the notation 1 is used for a vector of ones of length that is appropriate to its use. Let $\mathbf{D}_r = \mathrm{diag}(\boldsymbol{r})$ and $\mathbf{D}_c = \mathrm{diag}(\boldsymbol{c})$ be the diagonal matrices of row and column masses.

The computational algorithm to obtain coordinates of the row and column profiles with respect to principal axes, using the singular-value decomposition (SVD), is as follows:

(1) Calculate the matrix of standardized residuals: $\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \boldsymbol{r}\boldsymbol{c})\mathbf{D}_c^{-\frac{1}{2}}$.

(2) Calculate the SVD: $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\top$, where $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$.

(3) Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\alpha$.

(4) Principal coordinates of columns: $\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\alpha$.

(5) Standard coordinates of rows: $\mathbf{X} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}$.

(6) Standard coordinates of columns: $\mathbf{Y} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$.

The total variance of the data matrix is measured by the inertia which is calculated on relative observed and expected frequencies.

The rows of the coordinate matrices in (3)–(6) above refer to the rows or columns of the original table. The columns of these matrices refer to the principal axes, or dimensions, of the solution. The row and column principal coordinates are scaled in such a way that $\mathbf{F}\mathbf{D}_r F^\top = \mathbf{G}\mathbf{D}_c\mathbf{G}^\top = \mathbf{D}_\alpha^2$. The standard coordinates have weighted sum-of-squares equal to 1: $\mathbf{X}\mathbf{D}_r\mathbf{X}^\top = \mathbf{Y}\mathbf{D}_c\mathbf{Y}^\top = \mathbf{I}$.

Package *ca* in R implements CA. The output of function *ca*() is structured as a list-object. The *ca*() output contains the eigenvalues and percentages of explained inertia for all possible dimensions. Values for the rows and columns (masses, chi-squared distances of points to their average, inertias and standard coordinates) are also given.

Eigenvalues and relative percentages of explained inertia are given for all available dimensions. Additionally, cumulated percentages and a scree plot are shown. The items given in rows and columns of *summary*() include the principal coordinates for the first two dimensions ($k = 1$ and $k = 2$). Squared correlations and contributions for the points are displayed next to the coordinates. Notice that the quantities in these tables are multiplied by 1000 (e.g., the coordinates and masses).

The rows and columns of a data table analyzed by CA are called active points. These are the points that determine the orientation of the principal axes.

It happens that there are additional rows and columns of data that are not the primary data of interest but that are useful in interpreting features discovered in the primary data. Any additional row or column of a data matrix can be positioned on an existing map. These additional rows or columns that are added to the map are called supplementary points.

Supplementary variables have no impact on the computation. They are projected onto the solution space afterwards. Thus, contributions are not applicable for this case. Squared correlations as a measure of how well a point is represented

by the axes are still meaningful for the case of supplementary variables and thus are included in the output.

The results from CA can be visualized in the following way. The graphical solution can be restricted to two dimensions–first principal axis to be displayed horizontally (the x-axis) and the second principal axis to be displayed vertically (the y-axis). Usually the first two dimensions are plotted. However, eigenvalues are known for all possible dimensions. The supplementary variables can be added to the plot with a different symbol.

A three-dimensional display of the CA can also be created. This type of display offers the advantage that one can zoom and navigate using the mouse.

4. Results from CA

CA is performed on the provided survey dataset from GemSeek, Bulgaria. For this aim a data submatrix without missing data is extracted. This matrix contains 264 rows and 4 columns.

The shopping behavior is cross-tabulated according to how often clients shop food and grocery products in supermarket (six levels: daily, OnceTwiceAWeek, OnceEveryFewDays, OnceEvery2-4weeks, OnceEvery1-3months, OnceEvery3-6months) and most important factors for clients when deciding from which hypermarket to shop from (16 levels). The contingency table is reproduced in Table 1.

In Table 2 two supplementary rows are added. First supplementary row is "household's monthly combined income" and it contains five possible answers (Less5K, 5K-10K, 10K-15K, 15K-20K, 25K-30K, MoreThan30K, IDoNotWant-ToDeclare). Second supplementary row is the age with the following levels: 20-24, 25-29, 30-34, 35-39, 40-44, 45-50.

One cannot visualize the profiles exactly, since they are points situated in a four-dimensional space. CA identifies a low-dimensional subspace, which approximately contains the profiles. It reduces the dimensionality of the cloud of points so that we can visualize their relative positions. However, CA gives the coordinates of row and column points for all possible dimensions. This gives us the key to the interpretation of the association between the points.

The algorithm for direct ascription of missing categorical values is based on the association between levels of categorical variables. All associations deduced in this task are presented in Table 3, Table 4, and Table 5.

Table 1. Contingency table with active points

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**17**
daily	15	3	9	9	9	11	4	4	8	8	4	3	2	2	3	2
OnceEvery1-3months	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OnceEvery2-4weeks	6	3	0	3	1	3	5	1	0	0	0	0	1	1	0	0
OnceEvery3-6months	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
OnceEveryFewDays	11	4	4	8	10	13	4	11	6	8	0	3	0	1	0	0
OnceTwiceAWeek	7	7	5	5	6	4	6	3	2	3	2	4	2	1	0	0

Table 2. Contingency table with active and supplement points

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**17**
10K-15K	6	2	3	5	7	4	3	1	2	1	1	2	1	1	0	1
15K-20K	1	1	2	1	0	1	1	0	1	0	1	0	0	0	1	0
25-29	7	2	1	7	2	14	2	1	2	6	1	3	2	2	1	0
25K-30K	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30-34	11	6	4	6	6	4	4	5	4	2	0	1	2	0	1	1
35-39	6	3	5	4	7	8	4	2	4	2	0	1	0	1	1	0
40-44	8	3	6	4	9	1	6	5	3	4	2	3	1	1	1	1
45-50	8	3	2	4	3	5	3	6	3	5	3	2	0	1	0	0
5K-10K	21	5	4	13	15	8	11	11	8	13	3	4	0	3	2	0
daily	15	3	9	9	9	11	4	4	8	8	4	3	2	2	3	2
IDoNotWantToDeclare	4	5	5	1	4	5	2	3	1	3	0	2	1	0	1	1
Less5K	7	4	4	5	1	13	2	4	4	2	2	1	3	1	0	0
MoreThan30K	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OnceEvery1-3months	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OnceEvery2-4weeks	6	3	0	3	1	3	5	1	0	0	0	0	1	1	0	0
OnceEvery3-6months	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
OnceEveryFewDays	11	4	4	8	10	13	4	11	6	8	0	3	0	1	0	0
OnceTwiceAWeek	7	7	5	5	6	4	6	3	2	3	2	4	2	1	0	0

We have detected strong connection between the following levels (see Table 3):

- Level 30-34 of Q2 (How old are you?) and level 4 (Brands available) of Q9.
- Level 25-29 of Q2 (How old are you?) and level 6 (Store spaciousness and organization) of Q9.

It means that if in an observation the level of Q9 is missing (NA) and the level of Q2 is 30–34 then NA values of Q9 can be estimated (imputed) by value 4 (Brands available) and vice versa – the missing values of Q2 can be estimated by 30-34 when the value of Q9 is 4. If the level of Q9 is missing (NA) and the level of Q2 is 25-29, then NA values of Q9 can be estimated by value 6 (Store spaciousness and organization) and vice versa.

Table 4 and Table 5 present the following possible imputations:

- If the level of Q9 is missing (NA) and the level of Q28 is 5 000-10 000, then NA values of Q9 can be estimated (imputed) by value 4 (Brands available) and vice versa.

Table 3. Relationships between some levels of Q2 and Q9

Q2: How old are you?	Q9: Which of these factors is most important to you when deciding from which hypermarket to shop from?
30-34	4 (Brands available)
25-29	6 (Store spaciousness and organization)

Table 4. Relationships between some levels of Q28 and Q9

Q28: What is your households monthly combined income?	Q9: Which of these factors is most important to you when deciding from which hypermarket to shop from?
5 000-10 000 HRK	4 (Brands available)

Table 5. Relationship between some levels of Q4 and Q9

Q4: How often do you shop food and grocery products in supermarket/ hypermarket?	Q9: Which of these factors is most important to you when deciding from which hypermarket to shop from?
daily	9 (Product promotions like buy one get one free)
OnceTwiceAWeek	2 (Diversity of goods sold in the store)
OnceEveryFewDays	8 (Benefits from loyalty program)

- If the level of Q9 is missing (NA) and the level of Q4 is daily, then NA values of Q9 can be estimated (imputed) by value 9 (Product promotions like buy one get one free) and vice versa.

- If the level of Q9 is missing (NA) and the level of Q4 is OnceTwiceAWeek, then NA values of Q9 can be estimated (imputed) by value 2 (Diversity of goods sold in the store) and vice versa.

- If the level of Q9 is missing (NA) and the level of Q4 is OnceEveryFewDays, then NA values of Q9 can be estimated (imputed) by value 8 (Benefits from loyalty program) and vice versa.

New data submartix without missing data can be extracted from the dataset, provided by GemSeek. New couples of associations can be discovered using the same approach.

5. Final words

A typical example in the survey research is the use of ubiquitous chisquare test for association in a cross-tabulation. This test is not a tool detecting which parts of the table are responsible for this association.

Our approach is based on the association between levels of categorical variables.

Our algorithm for direct ascription of missing categorical values is based on the association between row points and column points discovered by correspondence analysis.

An open question is how to extract association between combinations of levels of categorical variables. The following features of the correspondence analysis: the percentages of inertia and squared correlations should also be involved in a machine learning algorithm.

Acknowledgements. The work of Veska Noncheva and Maria Dobreva is partially supported by the Fund NPD, Plovdiv University "Paisii Hilendarski", under Grant SP15-FMIIT-015.

References

[1] Dempster A. P., Laird N. M., Rubin D. B. (1997). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

[2] Greenacre M. (2007). Correspondence Analysis in Practice. Second Edition. London: Chapman & Hall / CRC.

[3] Heitjan D.F., Rubin D.B. (1991). Ignorability and coarse data. *Annals of Statistics*, **19**, 2244–2253.

[4] Little R. J. A., Rubin D. B. (1987). Statistical analysis with missing data. New York: Wiley.

[5] Murtagh F. (2005). Correspondence Analysis and Data Coding with Java and R, Chapman & Hall/CRC.

[6] Nenadic O., Greenacre M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, **20** (3).

[7] Rubin D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

[8] Rubin D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

[9] Schafer J. L. (1997). Analysis of incomplete multivariate data. London: Chapman & Hall.

[10] Schafer J. L., Graham J.W. (2002) Psychological Methods, Vol. 7, No. 2, 147–177