Modeling with Structurally Missing Data by OLS and Shapley Value Regressions

Stan Lipovetsky
GfK Custom Research North America
(stan.lipovetsky@gfk.com)

Ewa Nowakowska
GfK Polonia, Poland
(ewa.nowakowska@gfk.com)

Volume 19, *Number* 3
September 2013, *pp.* 169-178

This paper considers construction and estimation of the predictors' importance in multiple regressions using data with missing observations in any predictor variable. A dataset could have such structurally missing observations due to the nature of the problem. For instance, in a marketing research customer satisfaction study with a bank, questions concerning the quality of the call center, or the convenience of internet operations could be asked only if a respondent has been using these facilities. These are structurally missing values, which in contrast to randomly missing observations cannot be filled by imputed values. Deleting of such multivariate observations can significantly reduce the database, and make modeling simply impossible. It raises the question of constructing special models, which can incorporate all the data, with and without structurally missed observations among the attributes. We consider two approaches – using mixed-effect variables with dummies corresponding to whether or not the observation is available, and using the Gifi system of dichotomized variables, simultaneously with the application of the so-called Shapley Value regression, which produces meaningful individual coefficient estimates in the presence of multicollinear binary predictors. Numerical results are compared for marketing research data with structurally missing observations.

Keywords: **S**tructural Missing Data, Mixed-effect Model, Gifi System, Shapley Regression.

1. Introduction

The problem of missing values or non-available (NA) observations in statistical analysis has a long story and a wide range of remedies to fill blanks in the data. NA occurs in most empirical studies, for instance in marketing research due to various reasons. For instance, an interviewer may loose contact with some respondents during phone questioning, respondents may not know how to answer some questions, and elicited information could be unreliable. Also, variables such as attitudes, beliefs, opinions, income, even demographics can be sensitive and produce a negative reaction resulting in non-response data. Although NA is a natural result of data sampling, application of statistical software usually requires data reduction by deleting records with missing values in any variable, which may lead to biased and inefficient results (Little and Rubin 2002). Some problems can be solved with incomplete data, for instance, a technique developed in (Conklin et al. 2004) for finding key drivers in customer satisfaction and loyalty analysis or in Thurstone

scaling when each respondent makes a ranking or a paired comparison on only some of the numerous attributes, an approach based on order statistics can be applied (Lipovetsky 2007a). But most standard statistical techniques require complete data for estimations. A wide range of the methods has been developed to fill in these gaps in the data using surrogates – these range from simply the mean of the variable's available observations to the maximum likelihood and Bayesian imputations using the whole data set. Reviews and descriptions of various imputation procedures can be found, for instance, in (Draper and Stoneman 1964; Cochran 1977; Rubin 1987; Box 1990; Sarndal et al. 1992; Schafer 1997; Little and Rubin 2002; Lehtonen and Pahkinen 2004; Howell 2007; Fernandez et al. 2010). Sample balance weighting and propensity scoring procedures are often used in adjusting samples for non-response in survey studies (Rosenbaum and Rubin 1983; Rubin 2007; Lipovetsky 2007b). Due to the known classification (Rubin1976), there are so-called missing completely at random, missing at random, and missing not at random variables. Maximum likelihood and multiple imputation techniques for filling NA by many variables (those can be used for the dependent and independent variables in regression modeling) are mostly developed for the first two cases(see, for instance, Package "MICE" 2011, and the references within). The case of non-random NA is more complicated and needs a special model accounting for the missing data (Dunning and Freedman 2007).

An example of non-random but structural NA can be seen in the data where different segments of the respondents could be asked different sections of the questions from the whole questionnaire. For instance, in a study on satisfaction with a hotel system, questions about the breakfast quality can be asked of only those respondents who actually ate breakfast there. In a customer satisfaction study with a bank, some questions about the call center, or internet operations could be asked only if a respondent has been using those facilities. This raises the issue of constructing the regression model by all respondents, with and without observations by some attributes. A common approach consists in using a dummy 0-1 binary variable where the data is not-available or available, respectively, and its product with the original variable/s with the structural NA. Such a model permits us to consider absent values as zeros in the mixed-effect variables with these dummy indicators changing the regression slopes and intercept (Abraham and Ledolter 1983; Weisberg 1985). However, if the main variables are measured in numerical or rating scales (like a Likert scale from 1 to 5, or from 1 to 10, measuring a rated quality from the worst to the best levels), this approach produces a numerical zero value in the mixed-effect variable instead of a categorical zero level. In other words, if we add 0-value to the Likert five-point scale we increase the variability by the mixed-effect variable.

To escape using the artificial zero value of the mixed-effect variables, we suggest another approach. It is based on substituting each variable with NA to the so-called Gifi system of multi-variables (Gifi 1990; Michailidis and de Leeuw 1998; Mair and de Leeuw 2010). Its aim is to split each predictor by its levels into a system of binary variables with missing values corresponding to the reference level. For instance, a variable with a 5-point scale with NA values is represented as six binary variables each of which identifies one of the six codes, and only the first five binary variables defining the numerical codes can be used in the regression model. Sometimes it is possible to consider a fewer number of Gifi binaries combined by their meaning. For example, in key dissatisfier analysis (Conklin et al. 2004), it is sufficient to use only

three binary variables – those of dissatisfaction (lower levels), neutral (middle), and of satisfaction (upper levels).

In both models of mixed-effect dummy variables and the Gifi system of multi-variables,many of the binary variables can be highly correlated. Although the ordinary least squares (OLS) regression is good to fit and predict the data, it is not useful for the analysis of the individual predictor's impact under the conditions of multicollinearity. Multicollinearity is a high level of connection between two or more predictors, which makes estimation of the individual coefficients of the model too prone to small changes in the sample. Because of multicollinearity, the presumably useful attributes can easily receive a negative sign or close to zero value as their coefficients in regression. It could easily lead to the wrong conclusions and errant decisions. To reduce the effects of multicollinearity in regression modeling, various approaches have been developed, for instance, ridge-regression and its modifications (Hoerl and Kennard 1970, 2000; Lipovetsky 2010). Wirth and Hedler (2010) suggest applying mixed-effect variables in the so-called Shapley value regression (SVR) which has been developed earlier on the basis of cooperative game theory and is known for its robust and easily interpreted results (Lipovetsky and Conklin 2001,2005; Nowakowska 2010).In this current paper, we construct and compare OLS and SVR by the binary mixed-effect variables and by Gifi multi-variables, and evaluate the impact of each original variable in the model, using the net effects, or shares in the coefficient of multiple determination.

The paper is organized as follows: Sections II describes regressions with structural missing values via mixed-effect with dummy variables, and by Gifi systems of binary multi variables in OLS and SVR models. Section III presents numerical results for marketing research data with structurally missing observations. Section IV summarizes.

2. Regression Models by Structurally Missed Data

The multiple linear regression model can be written as

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_n x_{in} + e_i , \quad (1)$$

where y_i and x_{ij} are the i-th observations (i=1,...,N) by the dependent variable y and by each j-th independent variable x_j (j=1, 2,...,n), a_j are coefficients of the regression, and e_i are the error terms. If a predictor x_j contains missing observations, we can introduce a dummy variable d_j which has 1 in the positions of the non-missing and 0 in the positions of the missing values. Such dummy dichotomous variables, also called indicator variables, are often used in applied statistical modeling for catching a change in the level or in the slope of the dependent variable due to the change in some predictors (see, for instance, Weisberg 1985, chapter 7). For data with missing values, the model (1) can be extended in a general case by the cross-product of each variable with the corresponding dummy variable and the dummy variables themselves as follows:

$$y_i = a_0 + a_1 (x_{i1} d_{i1}) + ... + a_n (x_{in} d_{in}) + b_1 d_{i1} + ... + b_n d_{in} + e_i . \quad (2)$$

The mixed-effect variables $x_j d_j$ can be considered as the new variables constructed from the original x_j with missing values substituted with zero values. If in any x_j all the values are non-missing then d_j is not used. For a block of the same missing observations by many variables only one dummy variable is used. As discussed in

the introduction, there is a problem of interpretability in using such a "categorical" zero level of the dummy switch variable d_j which produces "numerical" zero values in the mixed-effect x_jd_j variables in the model (2). It can be unnoticeable in some experimental studies with numerical predictors varying in the range covering the zero value. But in the social studies which use variables measured in rating scales (like a Likert scale from 1 to 5, or from 1 to 10, measuring a rated quality from the worst to the best levels) such an approach yields an actual numerical zero value in the mixed-effect variable instead of a categorical zero level. In other words, if we add 0-value to the Likert five-point scale we increase the variability by the mixed-effect variable which can lead to inadequate pair correlations and regression coefficients.

To solve this problem, we suggest splitting each x_j with missing values into a system of binary variables, or Gifi system. If x_j is measured by a several-point Likert scale, each binary variable of a Gifi system identifies each of the levels of the scale by using the value of 1, and all the other levels including NA equal zero. Only the binary variables of the numerical levels are needed, and the missing values serve as a reference. Even if x_j is a continuous numerical variable, it can be quantified into several ordinal levels, and the same approach can be used. For instance, suppose each predictor has missing values, and we split it to three levels of lower, middle and upper values. Then the model (1) can be represented as follows:

$$y_i = a_0 + (a_1^{(1)}v_{i1} + a_1^{(2)}u_{i1} + a_1^{(3)}w_{i1}) + ... + (a_n^{(1)}v_{in} + a_n^{(2)}u_{in} + a_n^{(3)}w_{in}) + e_i \quad (3)$$

where v_j, u_j , and w_j are three binary variables produced by each x_j variable, and all the parameters $a_j^{(k)}$ present coefficients of regression for each *j*-th predictor at its *k*-th levels.

As it is well known in regression analysis, in the presence of multicollinearity among the predictors, important variables can have insignificant coefficients in the model, or the coefficients of the model can receive the opposite sign from their pair relation of the predictors with the dependent variable. Such models can be used for prediction, but they are meaningless in the analysis of the variables' importance. And the mixed-effect cross-products (2), as well as the binary Gifi multi variables (3) are often multicollinear. Indeed, any dummy variable d_j and its mixed-effect with a predictor x_jd_j would show a high correlation in the model (2). And there are more multicollinear variables x_jd_j if the same dummy is used in identification of the block of NA. A similar situation occurs with the Gifi system in model (3): the binary variables of the lower and upper values v_j and w_j are usually strongly negatively correlated, and there are more multicollinear predictors for the missing values in a block of observations. Let us consider how to overcome such difficulties.

From now on, for simplicity of presentation let us use the standardized variables, centered and divided by the standard deviations. For any of the models (1)-(3), the Ordinary Least Square (OLS) regression corresponds to minimizing the objective:

$$S^2 = \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{n} x_{ij}\beta_j\right)^2 = (y - X\beta)'(y - X\beta) = 1 - 2r'\beta + \beta'C\beta\,, \quad (4)$$

where y denotes the vector of standardized observations by the dependent variable, X is the design matrix with n standardized x-variables, β is the vector of standardized beta-coefficients of the OLS multiple regression, prime denotes transposition, and

the variance of the standardized y is $y'y = 1$. With the notations $C = X'X$ and $r = X'y$ for the matrix and vector of correlations among the predictors and with the dependent variable, respectively, minimization of the objective (4) by the vector of coefficients yields the normal system of equations and its solution as follows:

$$C\beta = r\,, \quad \beta = C^{-1}r\,, \tag{5}$$

where C^{-1} is the inverse correlation matrix. A well-known measure of the regression quality is the coefficient of multiple determinations which can be expressed via (4) as:

$$R^2 = 1 - S^2 = \beta'(2r - C\beta) = \sum_{j=1}^{n} \beta_j \left(2r_j - (C\beta)_j\right), \tag{6}$$

where the correlations r_j are the elements of the vector r. Due to the relations (5) the maximum of the coefficient of multiple determination (6) equals:

$$R^2 = \beta' r = \sum_{j=1}^{n} \beta_j r_j = \sum_{j=1}^{n} NEF_j\,. \tag{7}$$

The items of total R^2 define the so-called net effects of each regressor, $NEF_j = \beta_j r_j$, which are used for estimation of individual predictors' importance in the model.

Multicollinearity can change the sign of any coefficient β_j in comparison with the pair correlation r_j, which yields a negative net effect in (7). A better measure of comparative importance of predictors is the utility U_j of each regressor defined as the increment of multiple determinations in the models with and without each particular x_j:

$$U_j = R^2 - R^2_{-j}\,, \tag{8}$$

where R^2_{-j} denotes the multiple determination in the model without x_j. Going in depth with estimated increments by each one, two, three, etc. variables leads to the estimation of each predictor's total share in the regression model. This approach can be formulated via the Shapley Value imputation for the net effects (Lipovetsky and Conklin 2001, 2005).

The Shapley Value (SV) was developed to evaluate the worth of participants in a multi-player cooperative game (Shapley 1953; Roth, 1988) over all possible combinations of the players. The SV is defined as each j-th participant's input to a coalition:

$$SV_j = \sum_{all\ M} \gamma_n(M)\left[\upsilon(M \cup \{j\}) - \upsilon(M)\right], \tag{9}$$

With weights to enter into a coalition M defined as:

$$\gamma_n(M) = m!(n-m-1)!/n!\,, \tag{10}$$

where n is the total number of all participants, m is the number of participants in the M-th coalition, and $\upsilon(.)$ is the characteristic function used for estimation of utility for each coalition. By $M \cup \{j\}$ a set of participants, which includes the j-th

participant is denoted, when M means a coalition without the j-th participant.

In the regression modeling, the participants of the coalition are the predictors in the model, and in the problem of comparative usefulness of predictors, SV assigns a value for each predictor calculated over all possible combinations of predictors in the model, so it includes the competitive influence of any subsets of predictors in the analysis.

The characteristic functions υ and their differences in (9) are defined via R^2 values and their differences in the utilities (8). The total of SV by all variables equals the value of R^2 in the model with all the predictors together:

$$R^2 = \sum_{j=1}^{n} SV_j \,. \tag{11}$$

The relation (11) presents the net effects (7) estimated in the SV approach, when each item is the incremental average across all possible models with different subsets of the predictors. In contrast to (7), the SV net effects in (11) are always positive. Besides estimation of the predictors' importance (11), the SV technique can be used for adjusting the coefficients of the regression model itself (more detail see in Lipovetsky and Conklin 2001, 2010a,b).

3. Numerical Example

For numerical comparison, consider a real marketing research project with the dependent variable of customer satisfaction with a bank. There are 916 observations by overall satisfaction with the bank and 33 independent variables briefly listed in Table 1. All the variables are measured on a Likert scale from 1 to 10 from the worst to the best estimate. Each predictor contains a large portion of missing observations – NA numbers are shown in the first numerical column in Table 1. The next three columns of Table 1 present the OLS net effects (7) calculated by the model (1) with NA substituted by mean values for each variable, model (2) with subtotal net effects for each pairing of a dummy variable and its mixed-effect with a predictor, and model (3) with the Gifi split of each variable and the subtotals of the net effects within each original predictor. We can see that the OLS models produce many negative estimates of the predictors' contribution in the models, so such results are hardly useful for interpretation or for using the predictors to improve the bank relationship with clients.

The last three columns of Table 1 present the Shapley value net effects (11) calculated for the same models (1)-(3). In striking contrast to the OLS results, all the SV net effects are positive and the importance of each variable can be easily estimated. The last two rows in Table 1 show the coefficients of multiple determination and its adjustment by the degrees of freedom. All the models demonstrate similar quality of fit, and the net effects are close across all SVR models. Although the model with NA substituted by means in (1) and the mixed-effects model (2) yield larger values for the maximally important variables, while Gifi estimates are more conservative. The obtained results show that the adequate method of estimation – that is SVR in contrast to OLS – plays the more essential role in evaluation of the predictors' importance than the specifics of data coding – filling missing values with the means of the variables, using the binary mixed-effects, or by applying the Gifi system.

Table 1 *Predictors' Contribution (%) by Several Models*

Predictors	How many NA	OLS Net Effects			SV Net Effects		
		NA as means	Dummy Mixed-Effects	Gifi Binary System	NA as Means	Dummy Mixed-Effects	Gifi Binary System
Competitive Interest Rates	567	-0.14	0.05	1.00	0.06	0.92	0.88
Ability of Automatic Payments	422	-1.29	-0.47	-0.41	0.30	1.18	0.98
Ability to Have Direct Deposit	217	-0.81	0.13	0.27	0.06	1.47	1.05
Receive Canceled Checks	277	2.75	2.53	1.10	1.52	1.75	1.56
Multi-Ways to Access Account	311	-1.46	-1.71	-2.85	0.91	0.44	0.76
Variety of Features	231	4.08	3.89	5.62	3.38	3.67	3.23
Ability to Use Debit Card wherever	308	-2.76	-3.05	-2.79	0.76	0.64	0.83
ATM. Check Card to Pay	329	6.07	6.41	5.49	2.60	2.33	2.17
Overdraft Protection	392	-0.15	0.1	0.90	1.20	0.74	1.14
Online Banking	578	3.59	3.1	2.61	1.65	1.75	1.37
Banking by Telephone	410	0.89	1.13	2.13	0.81	0.91	1.31
ATM Fees	361	-1.10	-0.61	0.86	0.48	1.10	0.98
Fees for Routine Transactions	452	-0.43	-0.24	-0.21	1.58	1.21	1.25
No Fee for Minimum Balance	159	1.47	3.42	1.31	1.32	1.80	2.76
Fees for Overdrafts, Stop Payment	354	7.21	7.91	4.59	4.01	3.91	2.98
Fees for Telephone Banking	542	2.22	0.79	2.08	1.98	1.66	2.25
Fees for New Checks	276	0.11	-0.34	0.49	1.21	1.04	1.55
Communicate on Checking Account	361	0.75	1.19	1.99	2.55	1.85	2.66
Communicate on Changes in Fees	223	4.05	2.68	2.78	4.56	3.53	4.00
Customer Service for Checking	25	22.39	18.41	17.84	14.19	13.89	12.70
Explanations of Features	87	14.63	13.79	9.41	10.27	9.43	7.42
Kept Informed of Changes	70	9.39	7.78	9.40	8.23	7.88	7.79
Convenient Branch Locations	6	-2.66	-2.04	-1.53	1.39	1.80	1.83
Convenient ATM Locations	133	-1.75	-2	-1.02	0.33	0.38	1.02

Error Free Checking	20	4.16	3.99	2.23	4.50	5.17	3.92
Representative Solves Problems	87	10.80	11.13	9.98	9.42	9.50	9.22
Clear Comprehensive Statements	12	-3.56	-3.16	3.08	3.16	4.21	5.00
ATM Availability	326	5.84	6.75	6.51	3.71	3.35	2.82
Teller Window	250	9.42	10.07	10.85	7.41	6.09	6.83
Drive-Through	486	3.56	3.43	2.24	2.40	2.17	3.19
Automated Phone System	522	1.04	1.73	-0.06	1.21	1.07	1.13
Telephone CSR	577	2.35	2.79	3.45	2.62	2.35	2.45
Online Access	684	-0.67	0.39	0.65	0.20	0.81	0.98
R^2		0.39	0.41	0.41	0.39	0.41	0.41
R^2 adjusted		0.36	0.37	0.34	0.36	0.37	0.34

4. Summary

Mixed-effect predictors with dummy binaries and the Gifi system of binary multi-variables are considered for models with structurally missing observations. In contrast to the ordinary least squares models, Shapley regression produces meaningful results for data with many collinear binary predictors, and can be successfully used with many different missing value coding techniques. The suggested approach enriches the theoretical regression modeling and is useful for managers in practical decision making.

Acknowledgment: We thank our colleague Norbert Wirth for attracting our attention to this problem.

5. References

1. Abraham, B, Ledolter, J. (1983). *Statistical Methods for Forecasting*, Wiley, New York.
2. Allison, P.D. (2000). Multiple imputation for missing data, *Sociological Methods & Research*, 28: 301-309.
3. Box, G. E. P. (1990). A simple way to deal with missing observations, Quality Engineering, 3: 249-254.
4. Cochran, W.G. (1977).*Sampling Techniques*, Wiley, New York.
5. Conklin, M., Powaga, K., Lipovetsky, S. (2004). Customer satisfaction analysis: Identification of key drivers, *European Journal of Operational Research*, 154: 819-827.
6. Draper N.R., Stoneman, D.M. (1964). Estimating missing values in unreplicated two-level factorial and fractional factorial designs, *Biometrics*, 20: 443-458.
7. Dunning, T., Freedman, D.A. (2007). Modeling section effects. In: Outhwaite, W., Turner, S. (eds), *Handbook of Social Science Methodology*, Sage, London.
8. Fernandez, A., Nielsen, J.D., Salmeron, A. (2010). Learning Bayesian networks for regression from incomplete databases, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 18: 69-86.

9. Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, Chichester, England.
10. Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12: 55-67.
11. Hoerl, A.E., Kennard, R.W. (2000). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 42: 80-86.
12. Howell, D.C. (2007). The analysis of missing data. In: Outhwaite, W., Turner, S. *Handbook of Social Science Methodology*, Sage, London.
13. R. Lehtonen R., Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, Wiley, New York.
14. Lipovetsky, S. (2007a). Thurstone scaling in order statistics, *Mathematical and Computer Modelling*, 45: 917-926.
15. Lipovetsky, S. (2007b). Ridge regression approach to sample balancing with maximum effective base, *Model Assisted Statistics and Applications*, 2: 17-26.
16. Lipovetsky, S., Enhanced ridge regressions (2010).*Mathematical and Computer Modelling*, 51: 338-348.
17. Lipovetsky, S., Conklin, M. (2001). Analysis of regression in game theory approach, *Applied Stochastic Models in Business and Industry*, 17: 319-330.
18. Lipovetsky, S., Conklin M.(2005). Incremental net effects in multiple regression, *International Journal of Mathematical Education in Science and Technology*, 36: 361-373.
19. Lipovetsky S., Conklin, M. (2010a) Reply to the paper 'Do not adjust coefficients in Shapley value regression' *Applied Stochastic Models in Business and Industry*, 26, 203-204.
20. Lipovetsky S., Conklin M. (2010b). Meaningful regression analysis in adjusted coefficients Shapley value model, *Model Assisted Statistics and Applications*, 5: 251-264.
21. Little, R.J.A., Rubin, D.B. (2002).Statistical Analysis with Missing Data, Wiley, New York.
22. Mair, P, de Leeuw, J. (2010). A general framework for multivariate analysis with optimal scaling: The R package aspect, *Journal of Statistical Software*, 32: 1-23.
23. Michailidis, G., de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis, *Statistical Science*, 13: 307-336.
24. Nowakowska, E. (2010). Modeling in a multicollinear setup: Determinants of SVR advantage, *Model Assisted Statistics and Applications*, 5: 219-233.
25. Package "MICE" (2011).Multivariate Imputation by Chained Equations, TNO Prevention and Health, Department of Statistics, Leiden, http://www.multiple-imputation.com/
26. Rosenbaum, O., and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*; 70: 41-55.
27. Roth, A.E., ed. (1988).*The Shapley Value - Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, Cambridge.
28. Rubin, D.B. (1976). Inference and missing data, *Biometrika*, 63: 581–592.
29. Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
30. Rubin, D.B. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials, *Statistics in Medicine*, 26: 20–36.

31. Sarndal, C.E., Swensson, B.,Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.
32. Schafer, J.L. (1997).*Analysis of Incomplete Multivariate Data,* Chapman & Hall, London.
33. Shapley, L.S. (1953). A Value for n-Person Games. In Kuhn H.W., Tucker A.W. (Eds.) *Contribution to the Theory of Games*, II, Princeton University Press, Princeton, NJ, 307-317.
34. Weisberg, S. (1985). *Applied Linear Regression*, Wiley, New York.
35. Wirth, N., Hedler, F. (2010), A robust switch model for informative structural missing data in customer experience records, *ART Forum, American Marketing Association*, June 6-9, San Francisco, CA.

About Authors

Stan Lipovetsky, Ph.D., Senior Research Director, GfK Research Center for Excellence, Marketing Sciences, GfK Custom Research North America. Numerous publications in multivariate statistics, multiple criteria decision making, econometrics, microeconomics, and marketing research. Member of editorial board of International J. of Operations and Quantitative Management, J. of Electronic Modeling, J. of Model Assisted Statistics and Applications; regular reviewer in dozens professional journals; member of the American Statistical Association (ASA), Mathematical Association of America (AMA), Institute for Operations Research and Management Sciences (INFORMS), International Society on Multiple Criteria Decision Making (MCDM).

Ewa Nowakowska received her M.Sc. degrees in mathematics (2007) and psychology (2006) from the University of Warsaw, she also holds PhD in computer science from Polish Academy of Sciences (2012). Her research interests focus on Bayesian modeling (hierarchical Bayes, discrete choice modeling) as well as classical multivariate data analysis (unsupervised learning/clustering, cluster ability assessment, semi-supervised learning, classification and regression, structural modeling). Since 2005 he is with Marketing Sciences at GfK Polonia.