SSC Annual Meeting, June 2015
Proceedings of the Survey Methods Section

ANALYSIS OF ORDINAL SURVEY RESPONSES WITH "DON'T KNOW"

Xichen She and Changbao Wu[1]

ABSTRACT

Ordinal responses are frequently involved in social and health survey researches to evaluate performance, attitude, severity of diseases, etc.. It is also a common practice to list "Don't Know" as an option in the responses, especially for questions with sensitive nature. In this talk, we first briefly introduce approaches dealing with regular ordinal data, then explore methods for analyzing ordinal responses with "Don't Know" as part of the response. Consistency and efficiency are compared among alternative estimators and results from a limited simulation study will be discussed.

KEY WORDS: Missing at random, Category probabilities, Regression Analysis, Multiple Imputation, Fractional Imputation, Propensity score adjustment

1. INTRODUCTION

Ordinal responses are one of the widely collected and analyzed types of data in many scientific fields, such as social and behavioural sciences, public health and medical studies. Examples of ordinal responses include variables measuring performance (poor, average, excellent), attitude (disagree, neutral, agree), etc.. A vast literature is devoted to the analysis of ordinal responses, see for example, McCullagh (1980), Peterson and Harrell Jr (1990) and Ananth and Kleinbaum (1997). Agresti (2010) provides a comprehensive review.

It is a common practice to list "Don't Know" as an option of the ordinal response, especially for questions with sensitive nature. These responses are usually treated as missing values in many cases. There exists extensive discussion on this topic of handling data with missing values. Rosenbaum and Rubin (1983) discussed the *propensity score adjusting* method. The multiple imputation method was first proposed by Rubin (1978) and was further discussed in Rubin (1987). The idea of fractional imputation has also received increasing attention since the paper of Kim and Fuller (2004).

In this paper, we consider independently-observed J-level ordinal responses $\{y_{Ri}, \delta_i; i = 1, ..., n\}$, where δ_i is the response indicator for the ith individual. When $\delta_i = 1$, $y_{Ri} = y_i$ is observed; otherwise $y_{Ri} =$ DK (Don't Know) and is treated as missing. A covariate vector $\boldsymbol{x}_i$ is fully observed and consists of two components $(\boldsymbol{w}_i', \boldsymbol{s}_i')'$, where $\boldsymbol{w}$ are covariates of research interest in regression analysis and $\boldsymbol{s}$ are other available covariates or extraneous surrogates (Robins et al. 1994) whose relationship with the response variable might not be part of the inference goal. We further assume that the responses

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1
Corresponding author: Xichen She; Email: xshe@uwaterloo.ca

are missing-at-random (MAR), termed by Little and Rubin (2002), in the sense that $Pr(\delta = 1|y_R, \boldsymbol{x}) = Pr(\delta = 1|\boldsymbol{x}) := \pi(\boldsymbol{x})$.

In section 2, we consider the estimation of mean responses, which are for the ordinal data just category probabilities. In section 3, we focus on regression analysis and discuss in detail two different scenarios depending on the inferential objectives. In each scenario, we consider different estimators with validity carefully examined and efficiency compared. Results from simulation studies are presented in Section 4.

2. ESTIMATION OF MEAN RESPONSE

Define the cumulative indicator vector $\boldsymbol{z}_i = (z_{i1}, ..., z_{iJ-1})^\top$ of y_i, where $z_{ij} = \boldsymbol{I}(y_i \leq j)$. Let $\boldsymbol{p} = (p_1, ..., p_{J-1})$ denote the cumulative probabilities, where $p_j = Pr(\boldsymbol{Y} \leq j)$. It is easy to see that $E(\boldsymbol{z}_i) = \boldsymbol{p}$. When the interest lies in estimating the mean responses, we can simply focus on estimating $\boldsymbol{p}$, because any category probability of y is a linear contrast of $\boldsymbol{p}$.

We now present four alternative estimators of $\boldsymbol{p}$. The naive complete case (CC) estimator $\hat{\boldsymbol{p}}^{cc}$ is defined by

$$\hat{\boldsymbol{p}}^{cc} = \frac{\sum_{i=1}^n \delta_i \boldsymbol{z}_i}{\sum_{i=1}^n \delta_i}. \tag{2.1}$$

For the propensity score adjusting (PSA) method, we propose a parametric model $\pi(\boldsymbol{x}; \boldsymbol{\phi})$ for the propensity score. Then the PSA estimator $\hat{\boldsymbol{p}}^{psa}$ is given by

$$\hat{\boldsymbol{p}}^{psa} = \sum_{i=1}^n \frac{\delta_i \boldsymbol{z}_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})} \Big/ \sum_{i=1}^n \frac{\delta_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})}, \tag{2.2}$$

where $\hat{\boldsymbol{\phi}}$ is the ML estimator of $\boldsymbol{\phi}$.

For imputation-based estimators, a model for the conditional distribution of $y \mid \boldsymbol{x}$ involving all the covariates in the MAR assumption is required. We adopt the *Cumulative Link Model* from McCullagh (1980) and assume $Pr(y_i \leq j \mid \boldsymbol{x}) = \boldsymbol{\gamma}(\boldsymbol{x}; \boldsymbol{\psi})$. A consistent estimator $\hat{\boldsymbol{\psi}}$ can be obtained based on the observed data only (She, 2016), and we then draw independent samples $\{\tilde{\boldsymbol{z}}_{il},\ l = 1, ..., k\}$ from the discrete mass function $f(\boldsymbol{z}|\boldsymbol{x}_i; \hat{\boldsymbol{\psi}})$ for all missing units. The multiple imputation (MI) estimator $\hat{\boldsymbol{p}}^{mi,k}$ under the frequentist's paradigm is defined by

$$\hat{\boldsymbol{p}}^{mi,k} = \frac{1}{k} \sum_{l=1}^k \hat{\boldsymbol{p}}^l, \tag{2.3}$$

where

$$\hat{\boldsymbol{p}}^l = n^{-1} \sum_{i=1}^n \left[\delta_i \boldsymbol{z}_i + (1 - \delta_i) \tilde{\boldsymbol{z}}_{il}\right].$$

Noting that $f(\boldsymbol{z}|\boldsymbol{x}_i; \hat{\boldsymbol{\psi}})$ is a simple discrete distribution with finite categories, it is more appealing to cover all possible choices rather than to take samples. We propose the deterministic fractional imputation (FI) estimator $\hat{\boldsymbol{p}}^{fi}$ given by

$$\hat{\boldsymbol{p}}^{fi} = n^{-1} \sum_{i=1}^n \Big[\delta_i \boldsymbol{z}_i + (1 - \delta_i) \sum_{j=1}^J w_{ij} \boldsymbol{c}_j\Big], \tag{2.4}$$

where $\boldsymbol{c}_j$ is the cumulative indicator vector for $y = j$ and $w_{ij} = Pr(y_i = j \mid \boldsymbol{x}; \hat{\boldsymbol{\psi}})$ is the fractional weight.

We summarize the asymptotic properties of these estimators in the following theorem. Detailed proofs are available in She (2016).

Theorem 1. For estimating category probabilities, under some regularity conditions,

(1) The CC estimator is not valid, unless the responses are missing completely at random (MCAR); whereas the PSA estimator, the MI estimator and the FI estimator are all consistent.

(2) Under MCAR, the CC estimator and the PSA estimator are equivalent.

(3) The FI estimator is equivalent to the MI estimator with $k = +\infty$ and hence is more efficient than the MI estimator for a finite k.

3. REGRESSION ANALYSIS

Suppose the conditional distribution of $y \mid \boldsymbol{w}$ is of primary interest and we impose a parametric model $\gamma(\boldsymbol{w}; \boldsymbol{\theta})$ on the cumulative probabilities $Pr(y \leq j \mid \boldsymbol{w})$. The *Cumulative Link Model*, for example, is a popular choice. With complete data, the maximum likelihood estimator of $\boldsymbol{\theta}$ can be obtained by solving

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{B}_i [\boldsymbol{z}_i - \boldsymbol{\gamma}(\boldsymbol{w}_i; \boldsymbol{\theta})] = n^{-1} \sum_{i=1}^{n} \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{w}_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{3.1}$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\gamma}_i / \partial \boldsymbol{\theta}$ and $\boldsymbol{B}_i = [Var(\boldsymbol{z}_i \mid \boldsymbol{x}_i)]^{-1}$ which has an explicit form (She, 2016).

3.1 The First Scenario

We consider the first scenario when $\boldsymbol{w} = \boldsymbol{x}$, that is, all the covariates enter the analysis model and there are no surrogates.

The CC estimator $\hat{\boldsymbol{\theta}}_{cc}$ is defined by solving

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \delta_i \boldsymbol{S}(\boldsymbol{z}_i, \delta_i, \boldsymbol{w}_i; \boldsymbol{\theta}). \tag{3.2}$$

The PSA estimator $\hat{\boldsymbol{\theta}}_{psa}$ solves

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\boldsymbol{w}_i; \hat{\boldsymbol{\phi}})} \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{w}_i; \boldsymbol{\theta}), \tag{3.3}$$

where $\hat{\boldsymbol{\phi}}$ is the MLE of the parameter in the model $\pi(\boldsymbol{w}; \boldsymbol{\phi})$ on the propensity score.

Lastly, we construct two imputation estimators based on a model for the conditional distribution of the response given all observed covariates. In this case, $\boldsymbol{w} = \boldsymbol{x}$, so the imputation model required is the same as the analysis model, but for explicit distinction in notation we use $\boldsymbol{\psi}$ to denote the parameter in the imputation model. For each missing y_i, the imputed values $\{\tilde{\boldsymbol{z}}_{il},\ l = 1, ..., k\}$ are independently drawn from $f(\boldsymbol{z} | \boldsymbol{w}_i; \hat{\boldsymbol{\psi}})$ with

parameter $\boldsymbol{\psi}$ estimated using the observed data only. The multiple imputation method makes inference on the regression coefficients $\boldsymbol{\theta}$ based on

$$\hat{\boldsymbol{\theta}}_{mi,k} = \sum_{l=1}^{k} \hat{\boldsymbol{\theta}}_l / k, \tag{3.4}$$

where $\hat{\boldsymbol{\theta}}_l$ is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} [\delta_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{w}_i; \boldsymbol{\theta}) + (1 - \delta_i) \boldsymbol{S}(\tilde{\boldsymbol{z}}_{il}, \boldsymbol{w}_i; \boldsymbol{\theta})].$$

We then propose the fractional imputation estimator $\hat{\boldsymbol{\theta}}_{fi}$ which solves

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \left[\delta_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{w}_i; \boldsymbol{\theta}) + (1 - \delta_i) \sum_{j=1}^{J} w_{ij}(\boldsymbol{w}_i) \boldsymbol{S}(\boldsymbol{c}_j, \boldsymbol{w}_i; \boldsymbol{\theta}) \right], \tag{3.5}$$

where $w_{ij}(\boldsymbol{w}_i) = Pr(y_i = j \mid \boldsymbol{w}_i; \hat{\boldsymbol{\psi}})$.

We summarize the asymptotic properties of these estimators in the following theorem; see She (2016) for detailed proofs.

Theorem 2. For regression analysis where the analysis model involves all observed covariates, under some mild regularity conditions,

(1) All four estimators $\hat{\boldsymbol{\theta}}_{cc}$, $\hat{\boldsymbol{\theta}}_{psa}$, $\hat{\boldsymbol{\theta}}_{mi,k}$ and $\hat{\boldsymbol{\theta}}_{fi}$ are consistent.

(2) The CC estimator and the FI estimator are equally efficient, and are both more efficient than the PSA estimator and the MI estimator with a finite k.

(3) Under MCAR, the PSA estimator and CC estimator are equivalent; when $k \to \infty$, the MI estimator and the CC estimator are equivalent.

3.2 The Second Scenario

Another scenario when the analysis model only involves a subset $\boldsymbol{w}$ of covariates and other observed variables are treated as surrogates is also of practical importance. The choice of $\boldsymbol{w}$ is often dictated by the scientific interest of the researcher. In this case, $\boldsymbol{w} \subset \boldsymbol{x} = (\boldsymbol{w}', \boldsymbol{s}')'$ and a model $E(\boldsymbol{z} \mid \boldsymbol{w}) = \boldsymbol{\gamma}(\boldsymbol{w}; \boldsymbol{\theta})$ is posited. We are interested in making inferences on the regression coefficients $\boldsymbol{\theta}$.

The CC estimator is still defined by (3.2), but is no longer valid in this scenario. The propensity score model $\pi(\boldsymbol{x}; \boldsymbol{\phi})$, unlike that in the first scenario, depends on $\boldsymbol{x}$. The PSA estimator $\hat{\boldsymbol{\theta}}_{psa}$ is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})} \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{w}_i; \boldsymbol{\theta}), \tag{3.6}$$

where $\hat{\boldsymbol{\phi}}$ is still the MLE.

The multiple and fractional imputation estimator are still defined by (3.4) and (3.5), but the imputed values are generated from a different conditional distribution. We require

the imputation model to involve the same set of covariates as in the MAR assumption, i.e. $\boldsymbol{x}$, and thereby a new model $E(\boldsymbol{z} \mid \boldsymbol{x}) = \bar{\boldsymbol{\gamma}}(\boldsymbol{x}; \boldsymbol{\psi})$ is postulated. It is shown that $\boldsymbol{\psi}$ can be estimated with observed data only and the imputed values are generated from the conditional distribution of y given $\boldsymbol{x}$ with parameter $\hat{\boldsymbol{\psi}}$.

The validity and efficiency of estimators under this scenario are summarized below; see She (2016) for detailed proofs.

Theorem 3. For regression analysis where the analysis model involves a subset of all observed covariates, under some mild regularity conditions,

(1) The CC estimator $\hat{\boldsymbol{\theta}}_{cc}$ is usually not consistent.

(2) The other three estimator $\hat{\boldsymbol{\theta}}_{psa}$, $\hat{\boldsymbol{\theta}}_{mi,k}$ and $\hat{\boldsymbol{\theta}}_{fi}$ are consistent.

(3) The FI estimator $\hat{\boldsymbol{\theta}}_{fi}$ is equivalent to the MI estimator $\hat{\boldsymbol{\theta}}_{mi,k}$ with $k = +\infty$ and hence is more efficient than the MI estimator with a finite k.

4. SIMULATION STUDIES

Simulation studies are conducted to assess the finite sample performance of these estimators. We consider an ordinal response variable with three categories; for the mean response estimation and the first scenario of regression analysis, only two covariates, a continuous x_1 and a discrete x_2, are observed, while for the second scenario an additional surrogate s is also available. The ordinal responses are simulated following the *Cumulative Link Model* and the propensity score is assumed to follow a logistic model. We also cover three levels of response rates by carefully choosing parameters (only the results for 85% and 50% response rates are reported). Three sample sizes are chosen for comparison, i.e., $n = 200$, 500 and 1000, each replicated 2000 times.

Tables 1, 2 and 3 list the absolute relative bias (ARB, in %) and root mean squared error (RMSE) for estimators presented in Section 2, Sections 3.1 and 3.2, respectively. We can clearly observe the validity and relative efficiency of these estimators under different cases, which confirm our theoretical results.

5. CONCLUSIONS

In practice, while the models dealing with missing observations should always involve all the available covariates to better explain the missing mechanism, the analysis models adopted by end users can be chosen according to different research interests and involve only a subset of the covariates. The validity and efficiency of estimators actually depend on the choice of analysis objectives. The fractional imputation is an ideal tool for handling ordinal data. In all the three cases with different inference objectives we discussed, the fractional imputation estimator is always consistent and is more efficient than the multiple imputation estimator with a fixed k.

ACKNOWLEDGMENTS

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to C. Wu and a research assistantship from the Canadian Statistical Sciences Institute (CANSSI) to X. She.

Table 1: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) for p_1

RR	SS		Comp.	$\hat{p_1}^{cc}$	$\hat{p_1}^{psa}$	$\hat{p_1}^{mi,1}$	$\hat{p_1}^{mi,5}$	$\hat{p_1}^{mi,10}$	$\hat{p_1}^{fi}$
85%	1000	ARB	0.2	11.9	0.3	0.1	0.2	0.2	0.2
		RMSE	(1.4)	(3.4)	(1.6)	(1.6)	(1.6)	(1.6)	(1.5)
	500	ARB	0.1	12.7	0.01	0.1	0.1	0.1	0.1
		RMSE	(1.9)	(3.7)	(2.2)	(2.3)	(2.2)	(2.2)	(2.2)
	200	ARB	0.3	12.4	0.2	0.5	0.5	0.4	0.4
		RMSE	(3.0)	(4.5)	(3.3)	(3.5)	(3.4)	(3.3)	(3.3)
50%	1000	ARB	—	18.1	0.4	0.1	0.2	0.2	0.2
		RMSE	—	(5.1)	(2.0)	(1.8)	(1.7)	(1.7)	(1.7)
	500	ARB	—	17.8	0.4	0.3	0.1	0.1	0.2
		RMSE	—	(5.4)	(2.7)	(2.5)	(2.4)	(2.4)	(2.4)
	200	ARB	—	17.5	0.7	0.4	0.3	0.3	0.2
		RMSE	—	(6.4)	(4.1)	(4.0)	(3.7)	(3.7)	(3.6)

Table 2: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) for β_1.

RR	SS		Comp.	$\hat{\beta}_{cc}$	$\hat{\beta}_{psa}$	$\hat{\beta}_{mi,1}$	$\hat{\beta}_{mi,5}$	$\hat{\beta}_{mi,10}$	$\hat{\beta}_{fi}$
85%	1000	ARB	0.4	0.4	0.4	0.4	0.4	0.4	0.4
		RMSE	(9.6)	(10.4)	(10.5)	(11.2)	(10.6)	(10.5)	(10.4)
	500	ARB	1.2	1.4	1.4	1.6	1.4	1.4	1.4
		RMSE	(14.8)	(16.2)	(16.3)	(17.2)	(16.4)	(16.3)	(16.2)
	200	ARB	1.9	2.3	2.3	2.6	2.3	2.3	2.3
		RMSE	(22.8)	(25.5)	(25.7)	(27.2)	(25.9)	(25.7)	(25.5)
50%	1000	ARB	—	0.6	1.0	1.0	0.6	0.6	0.6
		RMSE	—	(14.4)	(15.9)	(16.1)	(14.7)	(14.5)	(14.4)
	500	ARB	—	1.8	2.5	2.3	1.9	1.8	1.8
		RMSE	—	(22.0)	(23.9)	(24.4)	(22.5)	(22.2)	(22.0)
	200	ARB	—	3.7	5.4	5.0	3.9	3.9	3.7
		RMSE	—	(36.1)	(38.9)	(40.4)	(37.2)	(36.8)	(36.1)

Table 3: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) for β_1.

RR	SS		Comp.	$\hat{\beta}_{cc}$	$\hat{\beta}_{psa}$	$\hat{\beta}_{mi,1}$	$\hat{\beta}_{mi,5}$	$\hat{\beta}_{mi,10}$	$\hat{\beta}_{fi}$
85%	1000	ARB	0.6	4.5	1.0	0.8	0.7	0.7	0.7
		RMSE	(10.1)	(14.7)	(14.5)	(12.1)	(11.4)	(11.4)	(11.2)
	500	ARB	0.9	4.8	1.9	1.1	1.0	1.0	1.0
		RMSE	(14.6)	(19.3)	(19.7)	(17.2)	(16.3)	(16.2)	(16.2)
	200	ARB	2.2	4.0	2.3	2.8	2.5	2.4	2.4
		RMSE	(23.5)	(31.1)	(25.7)	(28.1)	(27.1)	(26.9)	(26.6)
50%	1000	ARB	—	22.9	9.8	1.2	1.0	1.0	1.0
		RMSE	—	(49.7)	(41.5)	(18.1)	(17.2)	(17.1)	(16.9)
	500	ARB	—	23.5	13.1	2.2	1.8	1.8	1.7
		RMSE	—	(54.3)	(51.2)	(26.6)	(25.0)	(24.8)	(24.6)
	200	ARB	—	27.0	21.4	4.8	3.8	3.7	3.7
		RMSE	—	(70.1)	(74.3)	(44.5)	(41.7)	(41.1)	(40.7)

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, second edition, New York: John Wiley & Sons.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression Models for Ordinal Responses: A Review of Methods and Applications. *International Journal of Epidemiology*, **26(6)**, 1323–1333.

Kim, J. K. and Fuller, W. A. (2004). Fractional Hot Deck Imputation. *Biometrika*, **91**, 559–578.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Wiley.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, Series B*, **42**, 109–142.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*, Chapman and Hall.

Peterson, B. and Harrell Jr, F. E. (1990). Partial Proportional Odds Models for Ordinal Response Variables. *Applied Statistics*, **39(2)**, 205–217.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, **89(427)**, 846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70(1)**, 41–45.

Rubin, D. B. (1978). Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley.

She, X. (2016). *Analysis of Ordinal Responses with Missing Observations*, PhD dissertation, Department of Statistics and Actuarial Science, University of Waterloo.