Journal of the Korean Data & Information Science Society
2016, **27**(1), 217–223

http://dx.doi.org/10.7465/jkdi.2016.27.1.217
한국데이터정보과학회지

Estimation for misclassified data with ultra-high levels

Moonsu Kang[1]

[1]Department of Information Statistics, Gangneung-Wonju National University

Received 4 August 2015, revised 29 October 2015, accepted 15 January 2016

Abstract

Outcome misclassification is widespread in classification problems, but methods to account for it are rarely used. In this paper, the problem of inference with misclassified multinomial logit data with a large number of multinomial parameters is addressed. We have had a significant swell of interest in the development of novel methods to infer misclassified data. One simulation study is shown regarding how seriously misclassification issue occurs if the number of categories increase. Then, using the group lasso regression, we will show how the best model should be fitted for that kind of multinomial regression problems comprehensively.

Keywords: Bayesian, misclassification, multiple imputation.

1. Introduction

In machine learning and other applied statistics, classification is the problem of identifying to which of a set of categories a new observation belongs based on a training set of data containing observations whose category membership is known. It is possible to classify the individuals of a population in multiple ways. Examples include classifying subjects by sex, smoking status, health status (ill/sane), and so on. These examples are based on dichotomous data. Multiple classifications are also possible and more popular, e.g. individuals may be classified in the following groups: (A) single, (B) married, (C) divorced, and (D) widowed. However, when information is collected in the real world, the data do not often reflect the true status of the elements in the sample, that is, the data-generating process is often noisy (misclassification). This fact can happen due to several causes. In consumer surveys, consumers may not remember their previous behaviors accurately, may misunderstand survey questions or may intentionally misreport. In medical diagnosis, test failures and miscoded information are causes of distortion. The underlying statistical problem is known as inference with misclassified data. A widespread hardship in making inference with categorical data comes from misclassification (Chen, 1989). The effects of ignoring misclassification were first noted by Bross (1954), who showed that classical estimators base sampling on a dichotomous process under the assumption of known noise parameters. It was argued within that paper that these parameters are needed to correct the bias resulting from estimation based on the observed proportion. Though methods to adjust for bias in estimates due to a misclassification of (binary) outcome have existed over the past decades, those methods

[1] Assistant professor, Department of Information Statistics, Gangneung-Wonju National University, Gangneung 210-702, Korea. E-mail: mkang@gwnu.ac.kr

are rarely used in practice. Methods to correct for it depend on relationship between the observed outcome and the gold-standard outcome. To assess the extent of the misclassification with a particular method of observation is to compare it to a definitive method which is error free. A portion of the sample is processed by an infallible classifier, and then it is cross-classified by both fallible and infallible classifiers. This fact allows to estimate false positive and false negative rates. This technique is known as double sampling. A maximum likelihood approach to double sampling was presented by Tenenbein (1972) and generalized by Ekholm and Palmgren (1987). However, double sampling is not always the best method because gold standard methods are often not available for many applications, can be prohibitively expensive or can be computationally infeasible. Vianna (1994) and others adopted misclassification matrix for this problem while noise parameters were given in advance or estimated. They, however, did not consider predictor variables but outcome variables only. Polychotomous logistic regression model is fully introduced and developed in the literature (Songyong and Heemo, 2014). Multinomial outcomes with many levels can be challenging to model. Models of response variables with large numbers of outcome categories encounter several difficulties. Foremost is the rate at which the model dimensions expand when adding new covariates. If there are p categories, adding a covariate whose values are specific to the decision-maker adds $p-1$ identifiable regression parameters to the model. No correction for outcome misclassification with a large number of outcomes affects the estimation of regression parameters seriously

2. Outcome model specification

Suppose that we have apparent probabilities $p_1, \ldots, p_m$, have true probabilities $\pi_1, \pi_2, \cdots,$ π_m, where $p_i = P(T_k = i), \pi_i = P(Y_k = i)$ using the random variable X_k having the observed value and the random variable T_k having the true value in the $k-$th sampling unit where $k = 1, 2, \cdots, n$. The data are divided into m disjoint categories denoted by $\theta_1, \theta_2, \ldots, \theta_m$. The main interest is focused on making inferences on the proportions $\pi_1, \pi_2, \cdots, \pi_m$ of individuals belonging to the classes $\theta_1, \theta_2, \cdots, \theta_m$. The noise parameters are characterized by the transition (or misclassification) matrix:

$$\Lambda = \begin{pmatrix} \lambda_{11} & \ldots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{m1} & \ldots & \lambda_{mm} \end{pmatrix}$$

where λ_{ij} denotes the probability that an individual from θ_i is classified in θ_j, $\theta = (\theta_1, \theta_2, \cdots,$ $\theta_m)'$ denotes the vector of true proportions and $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{im})'$ denotes the transition vector for the class θ_i $(i = 1, 2, \cdots, m)$.

Multinomial logistic regression is used when the dependent variable is nominal. Multinomial logit regression is a solution to the classification problem to assume that a linear combination of the observed features and regression parameters can be used to determine the probability of each particular outcome of the dependent variable. The best values of the parameters are in general determined from a training data.

As in other forms of linear regression, multinomial logistic regression uses a linear predictor function $f(j, i), j = 1, 2, \cdots, m, i = 1, 2, \ldots, n$ to predict the probability that observation i has outcome j, of the following form: $f(j, i) = \beta_{0,j} + \beta_{1,j}x_{1,i} + \beta_{2,j}x_{2,i} + \cdots + \beta_{p-1,j}x_{p-1,i}$, where $\beta_{l,j}$ is a regression coefficient associated with the l-th explanatory variable and the j-th outcome. More compactly, $f(j,i) = \boldsymbol{\beta}_j \cdot \mathbf{x}_i$, where $\boldsymbol{\beta}_j$ is the set of regression coefficients associated with outcome j, and $\mathbf{x}_i$ is the set of explanatory variables associated with observation i. Suppose that the number of response levels is m. The corresponding model is given by

$$\begin{aligned}
Pr(Y_i = 1) &= \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{x}_i}}{1 + \sum\limits_{j=1}^{m-1} e^{\boldsymbol{\beta}_j \cdot \mathbf{x}_i}} \\
Pr(Y_i = 2) &= \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{x}_i}}{1 + \sum\limits_{j=1}^{m-1} e^{\boldsymbol{\beta}_j \cdot \mathbf{x}_i}} \\
&\vdots \\
Pr(Y_i = m) &= \frac{1}{1 + \sum\limits_{j=1}^{m-1} e^{\boldsymbol{\beta}_j \cdot \mathbf{x}_i}}.
\end{aligned}$$

Vienna (1994) with many researchers have assumed that misclassification matrix (confusion matrix) should be estimated from a validation study. or be known in advance. Please refer to that paper in more details.

3. Motivating example

We simulate multinomial logistic regression model using the function multinom in R package nnet. That model consists of 7 outcomes (A, B, C, D, E, F, G) with the reference level A and the predictor variable normally distributed random variable $X1$ and the total number of observations is 100. We partition the data into training dataset and test data set which have 80 observations and 20 observations, respectively. Based upon the training dataset, we get a set of 20 predicted probabilities and compare true outcomes in test dataset with the outcomes with the highest predicted probability based upon the fitted model with the training dataset. Now let us describe that fitted model as below.

$$\begin{aligned}
log(B/A) &= 0.7427399 - 0.14767536 * X1 \\
log(C/A) &= 0.6134757 - 0.05831624 * X1 \\
log(D/A) &= 0.4320643 + 0.32611181 * X1 \\
log(E/A) &= 0.4823187 + 0.47068902 * X1 \\
log(F/A) &= 0.2543827 + 0.07211916 * X1 \\
log(G/A) &= 0.6206348 + 0.15067462 * X1
\end{aligned}$$

From the table in Appendix, we see that as the number of outcomes increases, misclassification increases exponentially. In a multinomial logit model, one unit increase in the number of outcomes brings about the increase in the entire number of explanatory variables. As a good illustration, we have 7 outcomes with the 6 correspondent models. What will happen if a lot of outcomes exist? The correspondent model is too complex to be defined. From the simulation study above, we can observe how serious misclassification becomes, that is, 15 out of 20 cases. Thus, the selection of significant outcomes is in priority, more exactly, removing the redundant outcomes with rare chances. And then with that reduced number of outcomes, we utilizes predictor variable selection method, which is much easier way to model selection. Now let us investigate the property of outcomes.

4. Proposed model

As for polychotomous logistic regression composed of ordinal responses, there is the common sense that the model assume the proportionality assumption. The most popular ordinal link function uses every probability in every function by contrasting the lower levels of response variable with the higher levels of response Y. The general model has unequal slopes for the predictors, and we need enough data to estimate a different coefficient for each predictor in each response function. To simplify this model, you can induce an ordering on the linear predictors by using the same slope parameters for each response function and constraining the intercepts to increase or decrease. On the other hand, for nominal response, a logit link function is defined for each probability of $Pr(Y = j)$. A link function should be defined so that each response function contrasts a lower level with the last level: This is the generalized logit link, and it ignores the order of the responses, beyond identifying the last one as the reference response. Thus, the multinomial regression consists of nominal variables without ordering as a definition of a nominal variable. We should deal with each response separately after reduced number of outcomes. Viana (1994) clearly discussed the misclassification matrix for that matter. Please refer to that for more details. With reduced model regarding the number of responses, we fit the data to each model for a response.

Penalized regression is effective specially when the number of regression parameters is huge and those are correlated (SangIn, 2015). Tibshrani (1996) proposed the penalized regression by the lasso penalty. Meier *et al.* (2008) enhanced that regression by the group lasso penalty which is useful for the multinomial regression. In summary, first suppose that we have independent and identically distributed observations $(\mathbf{x}_i, \mathbf{Y}_i), i = 1, \ldots, n$ with a p-dimensional vector $\mathbf{x}_i \in \mathbb{R}^p$ of G predictors and the multinomial response variable vector $1 \times m$ vector $\mathbf{Y}_i^T = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,m})$ for $Y_{i,j} \in \{0, 1\}$. We define df_g as the degree of freedom of the $g-$predictor. $P_{\boldsymbol{\beta}}(\mathbf{x}_i) = P(Y_{i,j} = 1|\mathbf{x}_i)$ should be used as $\log\{\frac{p_{\boldsymbol{\beta}}(\mathbf{x}_i)}{1-p_{\boldsymbol{\beta}}(\mathbf{x}_i)}\} = \nu_{\boldsymbol{\beta}}(\mathbf{x}_i)$. $\boldsymbol{\beta}_g \in \mathbb{R}^{df_g}$ is the parameter vector corresponding to the $g-$th predictor. The group lasso estimator $\hat{\boldsymbol{\beta}}_\lambda$ is minimizer of the convex function

$$S_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} s(df_g)||\boldsymbol{\beta}_g||_2,$$

where $s(df_g)$ of the number $df_g^{1/2}$ to ensure that the penalty term is of the order of the number of the parameters $(df)_g$, $l_{\boldsymbol{\beta}} = \sum_{i=1}^n y_i \nu_{\boldsymbol{\beta}}(\mathbf{x}_i) - \log[1 + \exp(\nu_{\boldsymbol{\beta}}(\mathbf{x}_i)]$ and the whole parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ as $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T)^T$. $\nu_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \sum_{g=1}^G \mathbf{x}_{i,g}^T \boldsymbol{\beta}_g$ and $\mathbf{x}_i = (\mathbf{x}_{i,1}^T, \ldots, \mathbf{x}_{i,G}^T)^T$ with the group of variables $\mathbf{x}_{i,g} \in \mathbb{R}^{(df)_g}, g = 1, \ldots, G$. For easier understanding, we can think of multivariate analysis of variance for a set of dichotomous response variables where $k = \sum_{g=1}^{G} df_g$. *The response matrix* $\mathbf{Y}_{n\times m}$ given by

$$\begin{pmatrix} Y_{11} & \dots & Y_{1m} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \dots & Y_{nm} \end{pmatrix}.$$

The matrix of explanatory variables $\mathbf{X}_i$ with dimension $k \times p$ is $(\mathbf{X}_{i1}^T, \dots, \mathbf{X}_{iG}^T)^T$ as below.

$$\mathbf{X}_i = \begin{pmatrix} X_{i1,1,1} & \dots & X_{i,df_1,p} \\ \vdots & \ddots & \vdots \\ X_{ij,df_j,1} & \dots & X_{i,df_j,p} \\ \vdots & \ddots & \vdots \\ X_{iG,df_G,1} & \dots & X_{i,df_G,p} \end{pmatrix}.$$

The matrix of regression parameters $\boldsymbol{\beta}_{(k+1)\times m}$ given by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{0,1,1} & \dots & \beta_{0,1,m} \\ \beta_{1,1,1} & \dots & \beta_{1,1,m} \\ \vdots & \ddots & \vdots \\ \beta_{1,df_1,1} & \dots & \beta_{1,df_1,m} \\ \vdots & \ddots & \vdots \\ \beta_{G,df_G,1} & \dots & \beta_{G,df_G,m} \end{pmatrix}.$$

Based on the above model, we fit the simulation data in next section.

5. Numerical analysis

We simulate a multinomial logistic regression data based on R grpreg package. There are 100 observations with a set of 5 multinomial response variables. For each response variable, we fit the model with a group of 10 explanatory variables per the same group size 5. The matrix of fitted regression parameters $\hat{\boldsymbol{\beta}}$ are summarized as below.

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0.3294 & 0.3001 & 0.2734 & 0.2491 & 0.2270 & 0.2068 & 0.1885 & 0.1710 & 0.1565 & 0.1426 \\ 0.1299 & 0.1184 & 0.1079 & 0.0983 & 0.0895 & 0.0816 & 0.0743 & 0.0677 & 0.0617 & 0.0562 \\ 0.0512 & 0.0467 & 0.0425 & 0.0388 & 0.0353 & 0.0322 & 0.0293 & 0.0267 & 0.0243 & 0.0222 \\ 0.0202 & 0.0184 & 0.0168 & 0.0153 & 0.0139 & 0.0127 & 0.0116 & 0.0105 & 0.0096 & 0.0087 \\ 0.0080 & 0.0073 & 0.0066 & 0.0060 & 0.0055 & 0.005 & 0.0046 & 0.0042 & 0.0038 & 0.0035 \\ 0.0031 & 0.0029 & 0.0026 & 0.0024 & 0.0022 & 0.0020 & 0.0018 & 0.0016 & 0.0015 & 0.0014 \\ 0.0012 & 0.0011 & 0.0010 & 9e-04 & 9e-04 & 8e-04 & 7e-04 & 6e-04 & 6e-04 & 5e-04 \\ 5e-04 & 4e-04 & 4e-04 & 4e-04 & 3e-04 & 3e-04 & 3e-04 & 3e-04 & 2e-04 & 2e-04 \\ 2e-04 & 2e-04 & 2e-04 & 1e-04 & 1e-04 & 1e-04 & 1e-04 & 1e-04 & 1e-04 & 1e-04 \\ 1e-04 & 1e-04 & 1e-04 & 1e-04 & 1e-04 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}.$$

It should be noted that a group lasso penalty shrinks each parameter close to 0 at once instead of utilizing complex model selection procedures in the literature of a usual linear regression model.

6. Concluding remarks

We have seen how we estimate for misclassified regression data with ultra-high level of response variables. First of all, redundant response variables should be diminished for an easier model. Then, it should be kept in mind that one categorical variable with k levels should be represented as $k-1$ dummy variables. The multinomial outcome with m levels corresponds to a group of m dichotomous variables. Based on those, we can fit the model appropriately. Since we handle too many explanatory variables because of the nature of a multinomial logistic regression, a group lasso penalized regression can be so useful by shrinking each parameter close to 0 automatically.

Appendix

Predicted probabilites

Pr(Y=A)	Pr(Y=B)	Pr(Y=C)	Pr(Y=D)	Pr(Y=E)	Pr(Y=F)	Pr(Y=G)	Predictor	True response	Predicted response
0.18639908	0.1638810	0.1369951	0.14417576	0.11452575	0.1652575	0.08876579	0.005764186	F	A
0.16858519	0.1533320	0.1483103	0.16488754	0.11259153	0.1673830	0.08491050	0.385280401	A	A
0.20465136	0.1739768	0.1258409	0.12542195	0.11575566	0.1621653	0.09218806	-0.370660032	C	A
0.15682700	0.1459786	0.1559858	0.18004046	0.11087645	0.1682226	0.08206913	0.644376549	B	D
0.19730799	0.1700001	0.1302723	0.13268842	0.11534723	0.1635107	0.09087324	-0.220486562	A	A
0.17105584	0.1548371	0.1467174	0.16185988	0.11290612	0.1671467	0.08547690	0.331781964	C	A
0.13720414	0.1329826	0.1690953	0.20836573	0.10714595	0.1684446	0.07676172	1.096839013	A	D
0.16629349	0.1519236	0.1497942	0.16774312	0.11228581	0.1675841	0.08437574	0.435181491	B	D
0.20245622	0.1727999	0.1271574	0.12755632	0.11564539	0.1625811	0.09180362	-0.325931586	A	A
0.13503106	0.1314856	0.1705655	0.21176250	0.10666033	0.1683671	0.07612795	1.148807618	E	D
0.14157630	0.1359590	0.1661469	0.20169668	0.10807747	0.1685354	0.07800828	0.993503856	C	D
0.16114140	0.1487136	0.1531519	0.17433351	0.11154852	0.1679710	0.08314009	0.548396960	A	D
0.17538591	0.1574420	0.1439435	0.15667769	0.11342065	0.1666856	0.08644459	0.548396960	A	A
0.21739476	0.1806103	0.1183381	0.11363811	0.11620401	0.1595367	0.09427808	-0.627906076	G	A
0.12635689	0.1253905	0.1764600	0.22589452	0.10456586	0.1678307	0.07350146	1.360652449	F	D
0.21601620	0.1799089	0.1191380	0.11486449	0.11617091	0.1598380	0.09406349	-0.600259587	F	A
0.09556021	0.1020998	0.1974354	0.28483331	0.09483733	0.1624292	0.06280482	2.187332993	G	D
0.11954105	0.1204637	0.1811120	0.23768600	0.10273595	0.1671362	0.07132513	1.532610626	A	D
0.19804840	0.1704062	0.1298220	0.13193924	0.11539356	0.1633810	0.09100955	-0.235700359	A	A
0.23746460	0.1903823	0.1070224	0.09701788	0.11628681	0.1547299	0.09709614	-1.026420900	F	A

References

Tibshirani, R .(1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B (Methodological)*, **58**, 267-288.

Bross, I. (1954). Misclassification in 2 by 2 tables, *Biometrics*, **10**, 478-486.

Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics*, **10**, 187-202.

Viana, M. A. G. (1994). Bayesian small-sample estimation of misclassified multinomial data. *Biometrics*, **50**, 237-243.

Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, **8**, 1095-1106.

Ekholm, A. and Palmgren, J. (1987). Correction for misclassification using doubly sampled data. *Journal of Official Statistics*, **3**, 419-429.

Meier, L., Geer, S. V. D. and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B (Statistical Methodology)*, **70**, 53-71.

Songyong, S. and Heemo, K. (2014). A polychotomous regression model with tensor product splines and direct sums. *Journal of the Korean Data & Information Science Society*, **25**, 19-26.

SangIn, L. (2015). A note on standardization in penalized regressions=A note on standardization in penalized regressions. *Journal of the Korean Data and Information Science Society*, **26**, 505-516.