Statistical Analysis of Questionnaire Data via Cumulative Logistic Regression Model

Xiao-na SHENG¹, Long LIU² and Yu-qiu MA¹,*

¹School of Information Engineering, Harbin University, Harbin 150086, China
²School of Mathematical Sciences, Heilongjiang University, Harbin, 150080, China

*Corresponding author

Keywords: Cumulative logistic model, Rank variable, SAS.

Abstract. Based on a real questionnaire data, we build a cumulative logistic model to deal with the mutual relationship of the variables in the data. Aiming at the problem of too much designed variables, we apply statistical method to select those important ones and delete the useless ones without losing too much information. We also provide the analysis results for the data by the software SAS, and give some suggestions for the design of questionnaire. Our analysis also shows that the model built in paper can fit the questionnaire data very well, and it has higher preciseness rate of forecasting.

Introduction

With the incessant development of statistics, the method of logistic regression analysis attracts more and more attention [1,2]. However, as an important component part of logistic regression, the cumulative logistic regression is fairly less concerned in factual researches [3,4]. In 2013, to achieve the optimization configuration of human resource, a questionnaire about the service conduct situation of higher leaders in some direct-units of Heilongjiang province was carried out. Because too much problems and trivial content are designed in the questionnaire, it will bring bigger error if the usual method is used to analyze the data [5]. In this paper, the cumulative logistic regression is employed to integrate and analyze the data, which reduces some unnecessary variables. Then we use the model we built to perform some statistical forecasting, and the results are satisfactory.

There are 7 direct-units concerned in the survey, and 20675 questionnaires were sent out and collected. Two parts of content are designed: a. personal information of attendee; b. detail items on the service conduct situation, where 28 problems are included. Attendees can answer each problems based the factual situations of testing object (Rank: A, B, C, D). For the convenience of studying, we use variables \(x_i\) \((i=1,\ldots,28)\) to denote the scores to the 28 questions, with values 1, 2, 3 and 4, where variables \(x_i\) \((i=1,\ldots,27)\) are the specific evaluations on the service conduct to a testing object, and variable \(x_{28}\) is an integrative evaluation to the testing object.

Method

Cumulative Logistic Regression Model

Let the response variable \(y\) has \(J\) classes, i.e., \(y\) is an ordinal variable, and \(x=\{x_1,\ldots, x_m\}\) are independent variables.

Model

\[
\ln(\frac{p_j}{1-p_j}) = \alpha_j - \sum_{k=1}^{m} \beta_k x_k, \ j = 1,2,\ldots, J-1,
\]

is called cumulative logistic regression model [3], where

\[
p_j = p(y \leq j|x) = \sum_{i=1}^{j} p(y = i|x), \ j = 1,2,\ldots, J-1
\]
are cumulative logistic functions, \( \alpha_j \) \((j = 1, 2, \ldots, J - 1)\) are the latent thresholds of each class of \( y \), \( \beta_k \) \((k = 1, 2, \ldots, m)\) are coefficients of model.

**Pretreatment for the Data**

Search for the invalid questionnaires including blank ones and those whose answers are all 1, all 2, all 3, or all 4, then delete them [5,6]. The attendees who handed in the invalid questionnaires are irresponsible, and those corresponding data is useless to our analysis.

**Model Building**

Because the testing objects come from 7 different units, to diminish its impact on modeling, we combine the sample data from different units and model based on the integrated data. Moreover, all variables concerned in the questionnaire are rank variables (4 ranks), and the variable about integrative evaluation also has 4 ranks, therefore, we consider building a cumulative logistic regression model [7–9]. We consider \( x_{28} \) to be dependent variable, and \( x = \{ x_1, \ldots, x_{27} \} \) to be independent variables. The corresponding cumulative logistic regression model is as follows:

\[
\ln \left[ \frac{p(x_{28} \leq j|x)}{1 - p(x_{28} \leq j|x)} \right] = \alpha_j - \sum_{i=1}^{27} \beta_i x_i, \quad j = 1, 2, 3.
\]

The cumulative probability

\[
p_j = p(x_{28} \leq j|x) = \frac{e^{\alpha_j - \sum_{i=1}^{27} \beta_i x_i}}{1 + e^{\alpha_j - \sum_{i=1}^{27} \beta_i x_i}}, \quad j = 1, 2, 3, 4;
\]

\( p(x_{28} = j|x), (j = 1, 2, 3, 4) \) denotes the probability that the integrative evaluation is A, B, C and D, respectively; \( \alpha_j \) is the integration of \( \mu_j \) and \( \alpha \); \( \mu_j \) \((j = 1, 2, \ldots, J - 1)\) denotes the \( j \)th threshold, where \( \mu_1 < \mu_2 < \cdots < \mu_{J-1} \); \( \beta_k \) are the model coefficients.

There are totally 27 independent variables, however, we should introduce those with larger effects on the dependent variable, which will simplify the analysis of data. In the paper, based on the statistical analysis software SAS [10,11], we use the method of forward selection to select the independent variables at a certain level, and 13 variables are finally retained in the model (see Table 1), i.e., variables \( x_2, x_4, x_5, x_7, x_9, x_{11}, x_{14}, x_{15}, x_{20}, x_{22}, x_{25}, x_{26}, x_{27} \).

| Step | Effect Entered | DF | Number In | Score ChiSq | Pr>ChiSq |
|------|----------------|----|-----------|-------------|----------|
| 1    | \( x_{28} \)   | 1  | 1         | 569.605     | <.0001   |
| 2    | \( x_7 \)      | 1  | 2         | 190.313     | <.0001   |
| 3    | \( x_{27} \)   | 1  | 3         | 128.863     | <.0001   |
| 4    | \( x_{15} \)   | 1  | 4         | 68.0323     | <.0001   |
| 5    | \( x_{22} \)   | 1  | 5         | 36.4079     | <.0001   |
| 6    | \( x_2 \)      | 1  | 6         | 21.8327     | <.0001   |
| 7    | \( x_{14} \)   | 1  | 7         | 14.4900     | 0.0001   |
| 8    | \( x_{25} \)   | 1  | 8         | 12.4476     | 0.0004   |
| 9    | \( x_{20} \)   | 1  | 9         | 7.0513      | 0.0079   |
| 10   | \( x_{9} \)    | 1  | 10        | 2.8678      | 0.0904   |
| 11   | \( x_{11} \)   | 1  | 11        | 2.5856      | 0.1078   |
| 12   | \( x_5 \)      | 1  | 12        | 1.9622      | 0.1613   |
| 13   | \( x_4 \)      | 1  | 13        | 2.1189      | 0.1455   |
Model Fit

We first consider whether the proportion hypothesis holds. If the hypothesis is declined, i.e., the regression lines of different cumulative log-odds are not parallel, the current cumulative logistic model will be not applicable. Otherwise, it will be applicable. The output results about the testing of the hypothesis by the SAS is presented in Table 2.

Table 2. Testing results of proportion hypothesis.

| Chi-Square | DF | Pr>ChiSq |
|------------|----|----------|
| 35.9836    | 12 | 0.0920   |

It can be seen from Table 2 that the value of statistic is 35.9836 (DF=12) with a p-value 0.092, so the test is not significant, i.e., the proportion hypothesis holds and the model we built is applicable. Besides, we also check the goodness of fit from aspects of deviance, Pearson $\chi^2$ and R-Square, and the corresponding results are listed in Table 3.

Table 3. Testing results of goodness of fit.

| Criterion  | Value    | DF | Value/DF | Pr>ChiSq |
|------------|----------|----|----------|----------|
| Deviance   | 649.2196 | 171.8 | 0.3779 | 1.0000  |
| Pearson    | 170285.856 | 171.8 | 99.1187 | <.0001  |
| R-Square   | 0.7678   | Max-rescaled R-Square | 0.8618 |

We find in Table 3 that the deviance test is not significant ($p=1$), the Pearson test is significant ($p<0.0001$), so the fit of model is good. At the same time, R-Square =0.7678, which means that the integrative evaluation variable has strong persuasion in the model. To give reasonable explains of the considered cumulative logistic model, we also provide the corresponding results of likelihood ratio (LR) test and Wald test (see Table 4).

Table 4. $\chi^2$ and Wald test results for the data.

| Test | Chi-Square | DF | Pr>ChiSq |
|------|------------|----|----------|
| LR   | 1410.6027  | 13 | <.0001   |
| Wald | 411.5647   | 13 | <.0001   |

Table 4 shows that the value of LR statistic is 1410.6027 with a small p-value, so the selected independent variables have significant explainable ability. Moreover, the value of Wald statistic is 411.5647 also with a small p-value, which means that the regression coefficients are significant. All the results from Tables 2 to 4 show that the model we built has good fit to the questionnaire data. The final results of regression analysis are listed in Table 5.

Table 5. Logistic regression analysis for integrative evaluation to independent variables.

| Parameter | DF | Estimate | Standard Error | Wald Chi-Sq | Pr>ChiSq |
|-----------|----|----------|----------------|-------------|----------|
| Intercept | 1  | -23.4498 | 1.1991 | 382.4610 | <.0001  |
| Intercept | 1  | -16.9522 | 0.9204 | 339.1963 | <.0001  |
| Intercept | 2  | -11.4942 | 0.7280 | 249.2984 | <.0001  |
| $x_2$     | 1  | 0.6559   | 0.1909 | 11.8111 | 0.0006  |
| $x_3$     | 1  | 0.2968   | 0.2009 | 2.1840  | 0.0395  |
| $x_4$     | 1  | -0.3749  | 0.2124 | 3.1158  | 0.0775  |
| $x_5$     | 1  | 0.8067   | 0.1937 | 17.3422 | <.0001  |
| $x_6$     | 1  | 0.5086   | 0.2111 | 5.8040  | 0.0160  |
| $x_7$     | 1  | -0.3750  | 0.2134 | 3.0875  | 0.0789  |
| $x_{14}$  | 1  | 0.6562   | 0.1981 | 10.9713 | 0.0009  |
| $x_{15}$  | 1  | 0.8663   | 0.1967 | 19.4009 | <.0001  |
| $x_{20}$  | 1  | 0.4989   | 0.1876 | 7.0721  | 0.0078  |
From Table 5 we can see that all the variables $x_2$, $x_4$, $x_5$, $x_7$, $x_9$, $x_{11}$, $x_{14}$, $x_{15}$, $x_{20}$, $x_{22}$, $x_{25}$, $x_{26}$ and $x_{27}$ have significant impact to the integrative evaluation. By the practical study on the problem background, we also find that the selection of these variables is exactly reasonable. So we suggest that only retaining the 13 variables in later questionnaire is enough. Combined with the above analysis indexes, the model we built based on the questionnaire data is as follows:

\[
\ln \left( \frac{p_i}{p_i + p_j + p_k} \right) = -23.449 + 0.656x_i + 0.297x_2 + 0.375x_3 + 0.807x_4 + 0.509x_5 - 0.375x_6 + 0.656x_7 + 0.866x_8 + 0.499x_9 + 0.737x_{10} + 0.652x_{11} + 0.363x_{12} + 6 + 1.511x_{13} \quad (1)
\]

\[
\ln \left( \frac{p_i + p_j}{p_i + p_j + p_k} \right) = -16.952 + 0.656x_1 + 0.297x_2 - 0.375x_3 + 0.807x_4 + 0.509x_5 - 0.375x_6 + 0.656x_7 + 0.866x_8 + 0.499x_9 + 0.737x_{10} + 0.652x_{11} + 0.363x_{12} + 6 + 1.511x_{13} \quad (2)
\]

\[
\ln \left( \frac{p_i + p_j + p_k}{p_i + p_j + p_k} \right) = -11.494 + 0.656x_1 + 0.297x_2 - 0.375x_3 + 0.807x_4 + 0.509x_5 - 0.375x_6 + 0.656x_7 + 0.866x_8 + 0.499x_9 + 0.737x_{10} + 0.652x_{11} + 0.363x_{12} + 6 + 1.511x_{13} \quad (3)
\]

After further checking and validation, the preciseness rate of forecasting by the model we built can attain 86.75%, which is very ideal in practice.

**Factor Analysis for Affecting Integrative Evaluation**

To further make clear the mutual relationships between the dependent variable and each independent variable, we also consider factor analysis, from which we can directly see the contribution of each independent variable to the dependent variable. The analysis results of consistency for the variables are presented in Table 6.

| InV—DV | D—D | C—C | B—B | A—A |
|--------|-----|-----|-----|-----|
| $x_2 - x_{26}$ | n   | 35  | 65  | 233 | 312 | 69.96 |
|          | %   | 76.09 | 63.11 | 62.80 | 326 | 73.09 |
| $x_4 - x_{26}$ | n   | 33  | 61  | 233 | 351 | 78.7 |
|          | %   | 71.74 | 59.22 | 62.80 | 351 | 78.7 |
| $x_5 - x_{26}$ | n   | 33  | 49  | 220 | 351 | 78.7 |
|          | %   | 71.74 | 47.57 | 59.30 | 351 | 78.7 |
| $x_7 - x_{26}$ | n   | 36  | 58  | 248 | 355 | 79.69 |
|          | %   | 78.26 | 58.31 | 66.85 | 355 | 79.69 |
| $x_9 - x_{26}$ | n   | 28  | 54  | 251 | 365 | 81.84 |
|          | %   | 60.86 | 52.43 | 67.65 | 365 | 81.84 |
| $x_{11} - x_{26}$ | n   | 29  | 61  | 259 | 353 | 79.15 |
|          | %   | 63.04 | 59.22 | 69.81 | 353 | 79.15 |
| $x_{14} - x_{26}$ | n   | 33  | 59  | 259 | 373 | 83.63 |
|          | %   | 71.74 | 57.28 | 69.81 | 373 | 83.63 |
| $x_{15} - x_{26}$ | n   | 34  | 62  | 252 | 377 | 84.53 |
|          | %   | 73.91 | 60.19 | 67.92 | 377 | 84.53 |
| $x_{20} - x_{26}$ | n   | 35  | 63  | 239 | 389 | 87.22 |
|          | %   | 76.09 | 61.17 | 64.42 | 389 | 87.22 |
| $x_{22} - x_{26}$ | n   | 38  | 66  | 257 | 365 | 81.84 |
|          | %   | 82.61 | 64.08 | 69.27 | 365 | 81.84 |
| $x_{24} - x_{26}$ | n   | 34  | 63  | 260 | 313 | 70.08 |
|          | %   | 73.91 | 61.17 | 70.08 | 313 | 70.08 |
| $x_{25} - x_{26}$ | n   | 34  | 73  | 277 | 361 | 80.94 |
|          | %   | 73.91 | 70.87 | 74.66 | 361 | 80.94 |
| $x_{27} - x_{26}$ | n   | 37  | 73  | 284 | 368 | 82.51 |
|          | %   | 80.43 | 70.87 | 76.55 | 368 | 82.51 |

Note: * InV—DV denotes independent variable v.s. dependent variable
Table 6 shows that the rank selection of $x_{28}$ has higher consistency with that of each independent variable. When the rank selections between them are not consistent, the evaluation of independent variable usually has higher rank than that of dependent variable, which shows that the attitude of attendee is careful. In the consistency results, we can see that variables $x_2$, $x_4$, $x_5$ and $x_{35}$ each has a consistency rate about 67% with $x_{28}$; and the consistency rate between variables $x_7$, $x_9$, $x_{11}$, and $x_{28}$ is about 78%, which shows that protecting week-population is taken as a fundamental rule when the attendees fill in the questionnaire.

Conclusion

In this paper, we analyze a questionnaire data by building a cumulative logistic model, which deals with the problem of the too much designed problems, and therefore simplifies the content of questionnaire. Of course, the completeness of data is retained to a large degree[3], and the preciseness of analysis can also be guaranteed.

From the evaluation results to the testing objects, we find that people more care those aspects associated with their work and life, and the execution status to the necessary rules is a focus of people's attention. All in all, the first task of service conduct situation is to begin from the people’s livelihood, and only those rules which can be executed authentically in practice are good ones. In fact, there are still problems that are worthy studying further in the questionnaire data. For example, how to deal with the absence problem of some variables, how to compare the scores of each testing object in the same unit [12], and so on. We will go on carrying out our study in future work.

Acknowledgement

This paper is supported by the 3rd National Agricultural Census Leading Group Project of Heilongjiang Provincial Government (no.028).

References

[1] M. Juan, W. Philippe, G. Nermin, et al., An Original Stepwise Multilevel Logistic Regression Analysis of Discriminatory Accuracy: The Case of Neighbourhoods and Health. Plos One, 11(2016): e0153778.
[2] M. C. Simmonds, J. P.Higgins, A general framework for the use of logistic regression models in meta-analysis. Statistical Methods in Medical Research, 25(2016): 2858.
[3] M. H. Joseph. Logistic Regression Models, Chapman & Hall/CRC Press, 2009.
[4] L. F. Huntsinger, N. M. Rousphail, P. Bloomfield, Trip Generation Models Using Cumulative Logistic Regression. Journal of Urban Planning & Development, 139(2013): 176-184.
[5] J. Rattray, Essential elements of questionnaire design and development. Journal of Clinical Nursing, 16(2007): 234-243.
[6] Y. Song, Y. Son, D. Oh, Methodological Issues in Questionnaire Design. Journal of Korean Academy of Nursing, 45(2015): 323-328.
[7] A. Agresti, Analysis of Ordinal Categorical Data. New York: Wiley, 1984.
[8] S. L. Newman, R. Tumin, R. Andridge, et al. Family Meal Frequency and Association with Household Food Availability in United States Multi-Person Households: National Health and Nutrition Examination Survey 2007-2010. Plos One, 10(2015):e0144330.
[9] M. T. Chan, D. Yu & K. K. W. Yau, Multilevel cumulative logistic regression model with random effects: application to british social attitudes panel survey data. Computational Statistics & Data Analysis, 88(2015): 173-186.
[10] M. E. Stokes, C. S. Davis, G. G. Koch, Categorical Data Analysis Using the SAS System (2nd Edition). SAS Institute and Weily, 2003.

[11] Mervyn G. Marasinghe, William J. Kennedy. SAS for Data Analysis, New York: Springer–Verlag, 2008.

[12] T. F., MRCPsych, J. Hutchings, T. Bywater, Strengths and Difficulties Questionnaire Added Value Scores: evaluating effectiveness in child mental health interventions. The British Journal of Psychiatry, 194(2009): 552-558.