Categorical Regression Models*

Literature Review

Leon Yuan

12/13/22

This is a brief summary and synthesis on Categorical Regression Models from some books and paper.

Table of contents

*Dr. McGee

Introduction

Continuous response variables have been fully and thoroughly investigated on the regression model, simple linear regression [Bangdiwala, 2018], and multiple linear regression [Kaya Uyanık and Güler, 2013]. However, in the real world, there are many data and fields, such as medicine, healthcare, social science, and economics, coming with categorical response variables regardless of unordered or ordered responses. In this literature review, I will summarize current Categorical Regression Models.

When we consider categorical response variables, the easiest one is called the logistic regression model [Dreiseitl and Ohno-Machado, 2002] and [Cramer, 2003]. However, the traditional logistic regression is for a binary response. For example, a response variable is smoke or not. The logistic regression model gives regression values, and the probability of this response variable is equal to smoke. A threshold is given and used to cut off the categories. As usual, if 0.5 is selected, the estimated regression probability is greater than 0.5, and the predictive response is smoke and vice versa.

The logit model is straightforward to understand. However, the problem arises when the categorical response has more than two values. For example, the response variable is a type of infection. This can include the type of infection 1,2,3, and no infection. In this case, it is hard to set up multiple thresholds to cut off different categories. Because this is very random and subjective way to cut off. But the multinomial categorical response is naturally a generalization of the binary classification.

First of all, a quantitative form of a categorical response is needed in modeling so that the categorical information can be encoded into digit formats and can be used in a mathematical model. Following the way presented in [Tutz, 2011, pp.209-210] and [Agresti, 2012, pp.293-294], one considers a general case; there are k categorical values of response variable Y. The reason why $k-1+1$ is chosen is that one reference category is needed when building modeling and coding the dummy response variable. The way of coding this response is $y = (y_1, \dots, y_{k-1})'$ containing $k-1$ dummy variables.

$$y_r = \begin{bmatrix} 1, & Y = r \\ 0, & \text{otherwise} \end{bmatrix}$$

After building this dummy variable, we can have a vector for each response variable like the following from [Fahrmeir and Tutz, 2001, pp.70-71]: If $Y = j$, then its vector can be $y = (0, \dots, 1_{j-th}, \dots, 0)', \quad j = 1, \dots, k-1$. If this response variable has categorical j, then j_{th} component of this vector is 1, and all other components are 0. In terms of probabilities, we have the following relations:

$$\pi_j = \mathbf{P}(Y = j) = \mathbf{P}\left(y_j = 1\right), \quad j = 1, \dots, k-1$$

Lastly. If Y is the reference category, we have following: $Y = k$,

$$y = (0, ..., 0, ..., 0)', \quad \mathbf{P}(Y = k) = 1 - \pi_1 - ... - \pi_{k-1}$$

because of $\sum_{j=1}^{j=k} \pi_j(x) = 1$.

Unordered Categorical Regression

After these pre-processings paved the way for the formal model, the data structure and model specification can be written out as follows. The Multinomial unordered categorical response regression model is the direct extension and generalization of the binary logistic model. In the binary logit model, the model is trying to regress the probability of $Y = 1$ like the following from [Agresti, 2012, pp.296-299, section.8.1.3]:

$$\mathbf{P}\left(Y_i = 1\right) = \pi_i = \frac{\mathbf{exp}\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}\right)}{1 + \mathbf{exp}\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}\right)} = \frac{\mathbf{exp}\left(x_i'\beta\right)}{1 + \mathbf{exp}\left(x_i'\beta\right)}$$

For Multinomial logit model, the generalized logit model is as following presented from [Tutz, 2011, pp.210-214] because the logit function is the most widely used and effective transformation in such cases as detailed in [Train, 2009, pp.38-44]:

$$\mathbf{P}\left(Y_i = r\right) = \pi_{ir} = \frac{\mathbf{exp}\left(x_i'\beta_r\right)}{1 + \sum_{s=1}^{c} \mathbf{exp}\left(x_i'\beta_s\right)} \quad r = 1, ..., c$$

[McFadden, 1984, p.1403-1411] also presented this multinomial logit model in a very similar manner but also provide a way to illustrate this can derived from a latent variable model. For binary case, the reference probability of

$$P(Y_i = 0) = 1 - P(Y_i = 1)$$

Similarly, in the multinomial case, the reference probability of

$$\pi_{i,c+1} = 1 - \pi_{i1} - ... - \pi_{ic} = \frac{1}{1 + \sum_{s=1}^{c} \mathbf{exp}\left(x_i'\beta_s\right)}$$

The log-odd representation in the binary case is

$$log\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \alpha + x_i'\beta$$

In the multinomial case, the extension of log-odd is

$$\log \frac{\pi_{ir}}{\pi_{i,c+1}} = \alpha_r + x_i'\beta_r \quad \text{or} \quad \frac{\pi_{ir}}{\pi_{i,c+1}} = \exp\left(\alpha_r + x_i'\beta_r\right), \quad r = 1, \dots, c$$

, where $\beta_r = (\beta_{r1}, \dots, \beta_{rk})'$ and α_r are different for each category $r = 1, \dots, c$. Once having log-odd with the reference category, the log-odds with other categories are given by the following:

$$\log \frac{\pi_{i,a}(x_i)}{\pi_{i,b}(x_i)} = \log \frac{\pi_{i,a}(x_i)}{\pi_{i,c+1}(x_i)} - \log \frac{\pi_{i,b}(x_i)}{\pi_{i,c+1}(x_i)}$$

In addition to the above structure, [Tutz, 2011, pp.215] also mentioned that the multinomial model could be considered as a kind of random utility model as the following formula:

$$U_r = fix_r + noise_r$$

where the fix_r is the fixed utility value associated with $r-th$ category and $noise_r$ is the random variable with some distributions.

Ordered Categorical Regression

After the nominal response variable, as we know, in the real world, there are many categorical responses that are ordered shown in [Miot, 2020]. For example, the degree of general things, low, median, high, functional class, educational level, and satisfaction level. In this section, one typical ordered categorical regression is summarized.

The Cumulative Model

The straightforward idea of modeling the ordered categorical response is similar to the cut-off at 0.5 of the binary logit model. When the continuous regression values are greater than 0.5, the categorical response is $Y_i = 1$; when they are lower than 0.5, the response is $Y_i = 0$. However, in the ordered and multiple category case, there are many cut-offs through the continuous axis of the "some latent" variable, u_i presented and illustrated in [McFadden, 1984, p.1403-1411]. The latent variable is initially proposed in the context of continuous variables; however, because of its very effectiveness and usefulness, it is widely adopted and used in the categorical variable analysis shown in [Bishop, 1998]. The idea of an ordered case is that a latent variable is used to determine the categories of the response Y_i. The regression happens on this latent variable u_i. The features of the observation/unit are used to regress on this latent u_i as follows:

$$u_i = -x_i'\beta + \varepsilon_i$$

where β is the effect parameter vector which is invariant to the choice of the category of Y_i. This is an assumption made by people. However, β_j can vary for each category in a complex case. ε_i can have different assumption distributions. As an illustration of this latent continuous variable here, suppose that

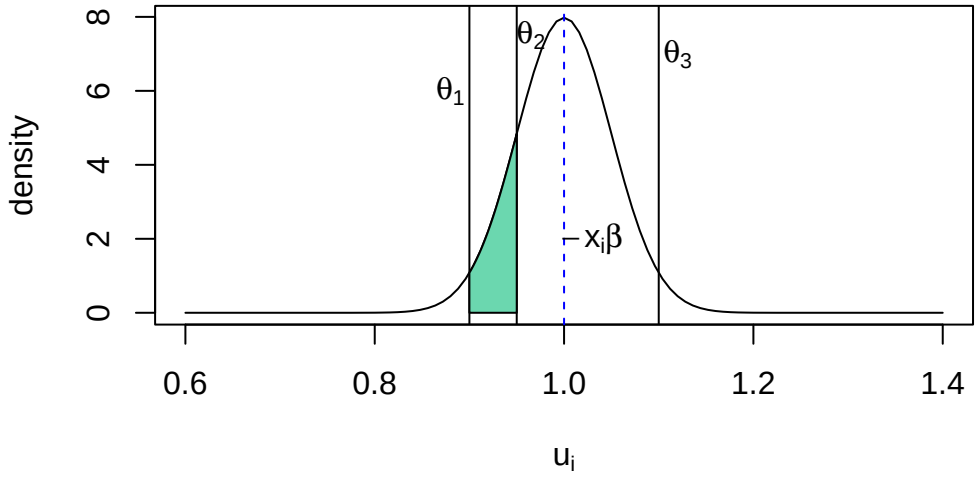
$$\varepsilon \sim N(0, \sigma^2)$$

So due to the linear combination of u_i with ε_i, have the following:

$$u_i \sim N(-x_i'\beta, \sigma^2)$$

For the following example, the response variable Y_i has four categorical values with three finite thresholds on the x-axis. The graph has $-x_i'\beta = 1, \theta_1 = 0.9, \theta_2 = 0.95, \theta_3 = 1.1$. The green filled/shaded area is the probability of $Y_i = 2$. The following figure 1 shows the three cut-off thresholds and the density curve of the observation x_i.

Figure1: density of a latent variable and multiple threshol

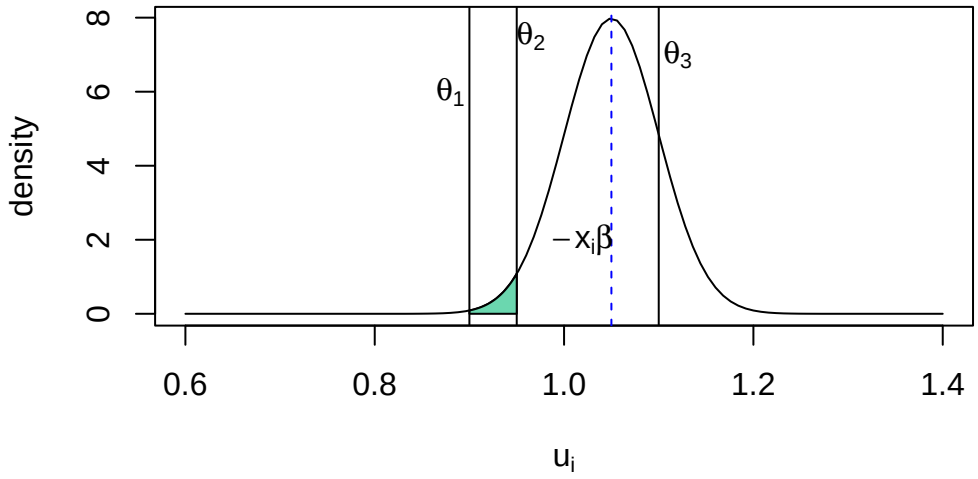


When the $-x_i'\beta = 1.05$ shifted toward right a little bit with $\sigma^2, \theta_1, \theta_2, \theta_3$ constant/fixed. It is obvious that the green shaded area is much less than the above case. The following figure 2 shows the shift, and the green shaded area becomes less. This means when the $-x_i'\beta$ shifts toward the right, the $P(Y_i = 2)$ becomes less. This effect is what we expect. These two graphs briefly explain the logic and reasons behind this cumulative ordered method.

After understanding the logic and under-the-hood reasons, the formal model can be listed out

$$Y_i = r \quad \Longleftrightarrow \quad \theta_{r-1} < u_i \leq \theta_r, \quad r = 1, \ldots, c+1$$

Figure2: density of a latent variable and multiple threshol



where all thresholds having $-\infty = \theta_0 < \theta_1 < \ldots < \theta_{c+1} = \infty$, then the cumulative distribution of the categorical response variable has the following relations:

$$\begin{aligned}
\mathbf{P}\left(Y_i \leq r \middle| x_i\right) &= \mathbf{P}\left(u_i \leq \theta_r | x_i\right) \\
&= \mathbf{P}\left(-x_i'\beta + \varepsilon_i \leq \theta_r | x_i\right) \\
&= \mathbf{P}\left(\varepsilon_i \leq \theta_r + x_i'\beta | x_i\right) \\
&= F\left(\theta_r + x_i'\beta\right), \quad r = 1, \ldots, c+1
\end{aligned}$$

Then for the discrete PMF of each category has the following:

$$\mathbf{P}\left(Y_i = r | x_i\right) = F\left(\theta_r + x_i'\beta\right) - F\left(\theta_{r-1} + x_i'\beta\right), \quad r = 1, \ldots, c+1$$

However, in the above case, the assumption that ε is independent of x is made. This assumption fails when the probability mass function is higher dense in one group than that in other groups. To deal with issue, a dependent τ_x of x is proposed to divide the regression term for scaling it like the following proposed by [Tutz, 2011, pp.257]:

$$P(Y_i \leq r \mid x_i) = F\left(\frac{\theta_r + x_i^T\beta}{\tau_x}\right)$$

By far, the most widely used discrete probability distribution and underlying function are logit shown in [Train, 2009, pp.38-44]. In this ordered category model, the most used function is logit as well. Here, when using logit, it is more straightforward to understand that cumulative

concepts. The following general formula for cumulative models is presented in [Agresti, 2012, pp.301-303]:

$$P(Y_i \leq j|x_i) = \pi_1(x_i) + \pi_2(x_i) + \cdots + \pi_j(x_i), \quad j = 1, \ldots, c+1$$

$$P(Y_i \leq j-1|x_i) = \pi_1(x_i) + \pi_2(x_i) + \cdots + \pi_{j-1}(x_i), \quad j = 1, \ldots, c+1$$

Thus, the cumulative logits are like the following:

$$logit[P(Y \leq j|x)] = \log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} = \log \frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1} + \cdots + \pi_{c+1}(x)}, \quad j = 1, \ldots, c$$

In addition to cumulative ordinal models, there is another type of ordinal model, sequential models. Sequential models are used when the ordinal response variable has to reach each level one by one. An excellent example would be the duration of unemployment. Long-term unemployment is a later outcome than short-term unemployment. A more detailed model and comparison between cumulative and sequential models are presented in [Tutz, 2011, pp.252-256].

Conclusion

Overall, the nominal and ordinal response variable modeling can be considered as a natural generalization of binary classification shown through [Agresti, 2012]. First of all, they are encoded in a format of a vector. Then, a logit function or latent variable model is applied to their covariates and encoded response variables. Cumulative and sequential ordinal models have their own advantages and disadvantages, as suggested in [Tutz, 2011, pp.256-257]. Although there are some categorical regression models, these models are less effective than the continuous regression model because continuous response provides more information and less vague predictions.

Appendix

When modeling the unordered multinomial regression model, it is very useful and imperative to consider multinomial distribution. In this section, briefly introduce it and its maximum likelihood estimation mentioned by most books and paper here [Agresti, 2012, pp.6-18].

Trial	c_1	c_2	c_3	(y_{i1}, y_{i2}, y_{i3})
1	0	0	1	(0,0,1)
2	1	0	0	(1,0,0)
3	0	1	0	(0,1,0)
4	0	0	1	(0,0,1)
5	1	0	0	(1,0,0)
count	$n_1 = 2$	$n_2 = 1$	$n_3 = 2$	$(2, 1, 2)$

Table 1: Multinomial Independent Trial

The counts $(n_1, n_2, \dots, n_c)$ **follows multinomial distribution and** $\pi_j = P(Y_{ij} = 1)$ **is the probability that** Y_i **chooses the category** j**. Then the Probability Mass Function of multinomial distribution is**

$$P(n_1, n_2, \dots, n_c) = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_n^{n_c}$$

where $\sum_{i=1}^{c} n_i = n$**. The**

$$E(n_i) = n \times \pi_i, \quad var(n_i) = n\pi_i(1 - \pi_i)$$

When considering ML to estimate the parameters $(\pi_1, \pi_2, \dots, \pi_c)$ **of this distribution, the counts** $(n_1, n_2, \dots, n_c)$**. Thus, by using the kernel density of the above PMF, the multinomial log-likelihood function is**

$$L(\pi_1, \pi_2, \dots, \pi_c) = \sum_{j=1}^{c} n_j \log \pi_j \quad \text{where all } \pi_j \geq 0 \text{ and } \sum_j \pi_j = 1$$

Differentiate $L(\pi)$ **with respect to** π_j**:**

$$\frac{\partial L(\pi)}{\partial \pi_j} = \frac{n_j}{\hat{\pi}_j} - \frac{n_c}{\pi_c} = 0 \quad \text{with} \quad \sum_j \hat{\pi}_j = 1$$

Lastly, the ML estimation of parameters are

$$\hat{\pi}_j = \frac{n_j}{n} \quad j = 1, 2, \dots, c$$

References

A. Agresti. Categorical Data Analysis. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9780470463635. URL https://books.google.com/books?id=UOrr47-2oisC.

Shrikant I. Bangdiwala. Regression: simple linear. International Journal of Injury Control and Safety Promotion, 25(1):113–115, 2018. doi: 10.1080/17457300.2018.1426702. URL https://doi.org/10.1080/17457300.2018.1426702. PMID: 29400125.

Christopher M. Bishop. Latent Variable Models, pages 371–403. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_13. URL https://doi.org/10.1007/978-94-011-5014-9_13.

J.S. Cramer. The origins of logistic regression, Jan 2003. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=360300.

Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5):352–359, 2002. ISSN 1532-0464. doi: https://doi.org/10.1016/S1532-0464(03)00034-0. URL https://www.sciencedirect.com/science/article/pii/S1532046403000340.

Ludwig Fahrmeir and Gerhard Tutz. Models for Multicategorical Responses: Multivariate Extensions of Generalized Linear Models, pages 69–137. Springer New York, New York, NY, 2001. ISBN 978-1-4757-3454-6. doi: 10.1007/978-1-4757-3454-6_3. URL https://doi.org/10.1007/978-1-4757-3454-6_3.

Gülden Kaya Uyanık and Ne e Güler. A study on multiple linear regression analysis. Procedia - Social and Behavioral Sciences, 106:234–240, 12 2013. doi: 10.1016/j.sbspro.2013.12.027.

Daniel L. McFadden. Chapter 24 econometric analysis of qualitative response models. volume 2 of Handbook of Econometrics, pages 1395–1457. Elsevier, 1984. doi: https://doi.org/10.1016/S1573-4412(84)02016-X. URL https://www.sciencedirect.com/science/article/pii/S157344128402016X.

Hélio Amante Miot. Análise de dados ordinais em estudos clínicos e experimentais. J. Vasc. Bras., 19:e20200185, November 2020.

Kenneth E. Train. Discrete Choice Methods with Simulation. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511805271.

G. Tutz. Regression for Categorical Data. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2011. ISBN 9781139499583. URL https://books.google.com/books?id=hvxuqoxD00kC.