Karki D, Qualitative Response Modeling Using Ordinal Logistic Regression.
JHAS 2010; Vol. 1, No. 1, p. 62-64

Short Communication

Qualitative Response Modeling Using Ordinal Logistic Regression

Dipesh Karki[1*]

[1]Pharmacy Department, Central Institute of Science and Technology College, New Baneshwor, Kathmandu, Nepal.

BACKGROUND

Qualitative Response Analysis is key methodology in carrying out research in field of social sciences, health care and management. Likert type scale that uses 5 point ordinal scaling is extensively used to collect the qualitative response from a sample in population. Once the data is collected then mean of each of the response is then computed. Besides if there is pre-test and post-test survey, t-test as well as ANOVA (Analysis of Variance) can be done to see the change in response. Despite extensive use of the qualitative analysis in this aspect, a very few research has been done to actually devise a model that can predict the qualitative response. For instance: If the question in researchers mind is to evaluate "Will the person contract heart disease?", the answer can be very likely, neutral, unlikely and very unlikely. This response can vary depending on various parameters like age, income, sex etc of the responder. Therefore a qualitative response model has to be devised on the basis of sample data that can truly reflect the population response.

This paper attempts to explore one of such technique known as Ordinal Logistic Regression that can be used for properly modeling the qualitative response. And it further attempts to provide a road map for the future researcher to follow a correct methodology for both predicting and validating the qualitative response generated from population.

Introduction to Logit Model

In statistics a Logistic regression can be employed to model the binary responses where the dependent variable (regressand) can take one of the two values.[1] For instance: The result for query whether person will complete the college level education? Can be of two extreme: i) yes ii) no. And probability of the answer can depend on several factors like – Sex, IQ, and Income.

This can be represented using the logistic regression in following manner:

$$P_i = f(z) = \frac{1}{1+e^{-z}} \quad \text{..... (i)}$$

Where,

$z = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

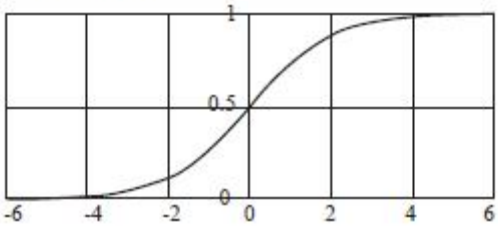
$f(z)$: Logistic function

P_i : Bernoulli probability distribution for whether person will complete the college or not (i.e. 0 or 1)

X_1 : Sex (encoded as female= 0 and male=1)

X_2 : IQ (continuous value)

X_3 : Income (in thousands)



The sigmoid or the logistic function has the ability to converge the domain [-∞, +∞] to [0,1]. As a result Pi will range from 0 to 1. But since the function is non-linear, Ordinary Least Square cannot be used to estimate the parameters (b0, b1, b2, b3).

So the equation (i) has to be modified to make it linear,

$$P_i = f(z) = \frac{e^z}{1+e^z} \quad \text{..... (ii)}$$

$$1-P_i = \frac{1}{1+e^z} \quad \text{..... (iii)}$$

* Correspondence: Dipesh Karki, Pharmacy Department, Central Institute of Science and Technology College, New Baneshwor, Kathmandu, Nepal.

Dividing (ii) by (iii) we obtain the odd ratio in favor of completing the college.

$$\frac{Pi}{1-Pi} = e^{z} \quad \text{..... (iv)}$$

By taking natural log we get,

$$L_i = \text{In}\left(\frac{P_i}{1-P_i}\right) = z = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad \text{.....(v)}$$

The equation v) is known as Logit model or Logistic Regression equation for dichotomous qualitative response. And it has to be noted that since Pi can be either 0 or 1. The above equation will become either In (0/1) or In (1/0) for individual observations. Hence Maximum Likelihood estimation should be used. Meanwhile for grouped data weighted least square method is to be employed for the estimation of parameter.

Extension of Logit Model for Ordinal Responses

Logit model in simplest form can only model dichotomous binary response. But if the response have various categories like the query posed in background section- whether or not person is likely to contract heart disease? And further its response depends on parameters viz- sex, age and income. The response can range Likert type 5 point scale (1- Very Unlikely, 2- Unlikely, 3- Neutral, 4- Likely, 5- Very Likely). These responses further are not nominal but instead are ordinal in nature as there is clear ranking between the categories. Besides, the underlying distribution of variable is not continuous and hence the categories cannot be segregated to particular interval. For example Likely doesn't imply it is twice of Neutral and Very Likely is three time of Neutral. For such scenario the Logit model should be extended and is called Ordinal Logistic Regression.

If the given query is modeled mathematically we get,

$$Z = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad \text{..... (vi)}$$

Where,

x_1 : Sex (encoded as female= 0 and male=1)

x_2 : Age

x_3 : Income (in thousands) and,

For all the possible observation there is latent value Z* such that based on the latent value the response might fall into one of the categories.

Y=1 (Very Unlikely)	if $Z < u_1$
Y=2 (Unlikely)	if $u_1 < Z < u_2$
Y=3 (Neutral)	if $u_2 < Z < u_3$
Y=4 (Likely)	if $u_3 < Z < u_4$
Y=5 (Very Likely)	if $u_4 < Z$

Where u_1, u_2, u_3, u_4 are threshold or cutoff limits for the observation to fall into one distinct category.

Using the previous logistic function, the probability of falling into each category can be computed using the relation,

$$P(Y=1) = \frac{1}{1+e^{-(z-u_2)}}$$

$$P(Y=2) = \frac{1}{1+e^{-(z-u_2)}} - \frac{1}{1+e^{-(z-u_1)}}$$

$$P(Y=3) = \frac{1}{1+e^{-(z-u_3)}} - \frac{1}{1+e^{-(z-u_2)}}$$

$$P(Y=4) = \frac{1}{1+e^{-(z-u_4)}} - \frac{1}{1+e^{-(z-u_3)}}$$

$$P(Y=5) = 1 - \frac{1}{1+e^{-(z-u_4)}}$$

In general if there are 'j' categories and 'k' regressor then

$$P(Y=j) = P(u_j - 1 < z < u_j) = F(u_j - z) - F(u_{j-1} - z) \quad \text{...(vii)}$$

Where,

F : Logistic function

$Z = \Sigma^{k}_{i=1} b_i x_i$

In Ordinal logit form it can be represented as:

$$L_i = \text{In}\left(\frac{P(Y < j / xk}{1 - P(y < j / xk)}\right) = uj - \Sigma^{k}_{i=1} b_i x_i \quad \text{..... (viii)}$$

Methodology for Using Ordinal Logistic Regression

Since the estimation of ordinal logistic regression using the equation viii) is extremely tedious we employ standard statistical package such as SPSS or SAS to evaluate both parameters (b_1b_1, b_2b_2...b_ib_i...b_kb_k) and the threshold points (b_1, b_2...,b_i...b_k...,). In SPSS there is the PLUM (Polytomous Universal Model) routine that easily evaluates the ordinal regression if data is fed properly in .sav format. For instance if hypothetical data in our research is some what as illustrated in table below.

S. No	Condition	Sex	Age	Income
1	1	1	75	100
2	4	0	65	20
3	5	1	14	5

Where, Condition specifies the response in five point scale. The entire data is stored in case.sav file

Then, by issuing following command in SPSS:

GET

FILE = C:\case.sav'.

PLUM

Condition WITH Sex Age Income

/CRITERIA = CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)

PCONVERGE(1.0E-6) SINGULAR(1.0E-8) / LINK = LOGIT

/PRINT = FIT PARAMETER SUMMARY TPARALLEL

/SAVE = ESTPROB (Plum)

This will yield the entire parameter coefficient as well as the threshold value. Besides the positive value for parameter coefficient will indicate strong correlation of the parameter with outcome where as negative value will indicate negative correlation.

After obtaining all the parameters we can forecast the response for out of sample response using the equation viii). Once the forecasted values are obtain then they can be validated with the validation set using any standard hypothesis testing means such as t-test or Coefficient of Determination.

A famous example of use of ordinal logistic regression is by Hamilton[2] where he showed the possibility of Challenger Space Shuttle crashing during launch based on the sample data received from previous flight. Had his model been used prior to the launch in January 28, 1986 the catastrophe could have been avoided.

CONCLUSION

Ordinal logistic regression technique can be used to properly determine and model the ordinal qualitative response that is used extensively in social science research. It can be used to predict future outcomes as well as provide the framework to correlate the independent parameters with the likely response.

REFERENCES

1 Gujarati, N.D. Basic Econometrics. 4th Ed. The McGraw Hill. 2007.

2 Hamilton, Statistics with STATA.