Ordinal Logistic Regression

by

Md Asad Uz Jaman

|

**Date of Submission: 13 December, 2018**

# General Purpose and Description

[Ordinal Regression](https://en.wikipedia.org/wiki/Ordinal_regression) Model is a [regression](https://en.wikipedia.org/wiki/Regression_analysis) model for [ordinal](https://en.wikipedia.org/wiki/Levels_of_measurement#Ordinal_type) [dependent variable(s](https://en.wikipedia.org/wiki/Dependent_variable)). This model was first considered by [Peter McCullagh](https://en.wikipedia.org/wiki/Peter_McCullagh). For instance, if one question on a survey is to be answered by a [choice among "poor", "fair", "good", and "excellent"](https://en.wikipedia.org/wiki/Likert_scale), and the purpose of the analysis is to see how well that answer can be predicted by the responses to other questions. It can be thought as an extension of the [logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) model that is used to analyze a dichotomous ('yes' or 'no') outcome.

There are three major uses for Ordinal Regression Analysis: 1) causal analysis, 2) forecasting an effect, and 3) trend forecasting.  Ordinal regression is used to verify the prediction strength of the independent variables on a dependent variable. Firstly, a typical question is, “What is the strength of relationship between dose (low, medium, high) and effect (mild, moderate, severe)?”

Secondly, ordinal regression can forecast the impacts of changes. It shows if one unit changes in independent variable (predictor variable), how much change we can expect on dependent variables. An important question is, “When is the response most likely to jump into the next category?” Finally, the model estimates future values. A typical question is, “If I take medium preparation for exam, what grade (A-F) can I expect?” (statisticssolutions.com, 2018)

# 2. Fundamental Equation

## 2.1. General, Logit and Probit formulas

The logit is the inverse of the standard cumulative logistic distribution function. The normit function, also known as probit, is the inverse of the standard cumulative normal distribution function. (Lemeshow and Hosmer, 2001).

general: g(**χ** **k**) = **θk**+**x**'**β**, **k** = 1, ..., **K**-1

logit: **g**(**χ**) = loge(**χ**/ (1 – **χ**))

probit: **g**(**χ**) = Φ–1(**χ**)

where,

***K*** = number of distinct categories of the response

***χk*** = cumulative probability up to and including category *k*, (*π* 1+ ...+ *πk*)

***g*(*χk*)** = vector of predictor variables

***θk* =** constant associated with the ***k*th**distinct response category

**x** = a vector of predictor variables

**β =** a vector of coefficients associated with the predictors

## 2.2. Formula of Predictive Probabilities

Event probabilities are the π kfor k = 1, 2, ..., K.

Where,

***K*** **=** equals 1, ..., K – 1

**θk** = constant

***β =*** vector of coefficients from the logit equation

## 2.3. Formula of Cumulative Predictive Probabilities

The sum of the probabilities equals 1. The logits of the first K - 1 cumulative probabilities are:

## 2.4. Formula of Coefficient

The coefficient for the predictor indicates that for any fixed k, the estimated change in the logit of the response when predictor is at one level compared to the reference level. The formula of coefficient is same as ‘Predictive Probabilities’. (support.minitab.com, 2018)

## 2.5. Formula of Z Test

Z test determines whether a predictor variable is significantly related to the outcome or dependent variables. The larger values of Z, the more significant relationship is there. (support.minitab.com, 2018)

Z = βi / standard error

The formula for the constant is:

Z = θk / standard error

## 2.6. Formula for Odds Ratio

Proportional odds model is used for ordinal logistic regression. Only one parameter and one odds ratio are calculated for each predictor. The odds ratio utilizes cumulative probabilities and their complements. (support.minitab.com, 2018) For a predictor with 2 levels x 1 and x 2, the cumulative odds ratio is:

## 2.7. Formula of Confidence interval

The large sample confidence interval for **βi** is:

**α** = level of significance

**βi *=*** vector of coefficients from the logit equation

To obtain the confidence interval of the odds ratio, exponentiate the lower and upper limits of the confidence interval. The interval provides the range in which the odds may fall for every unit change in the predictor.

## 2.8. Formula of Log-likelihood Test

For small samples, the likelihood-ratio test may be easy to test the significant relationship of predictor variable with dependent variables. But for large dataset, it is difficult to calculate by hand. For ordinal logistic regression, there are n independent multinomial vectors, each with k categories. These observations are denoted by y 1, ..., y n, where yi = (y i1, ..., yik ) and Σ j yij = mi is fixed for each i. From the ith observation yi , the contribution to the log likelihood is:

**L(πi , yi ) = Σ k yiklog πik**

**πik** = probability of the ith observation for the kth category

## 2.9. Assumption

Under the proportional odds model, the odds ratio assessing the effect of an exposure variable for any of these comparisons will be the same regardless of where the cut-point is made. In other words, the relationship among all groups is indifferent and Indistinguishable. (Kleinbaum and Klein, 2002, p.305)

# 3. A Simple Example

Next, we present an example of the ordinal logistic or proportional odds model using data from the Black/White Cancer Survival Study. Suppose we are interested in assessing the effect of RACE on tumor grade among women with invasive endometrial cancer. RACE, the exposure variable, is coded 0 for white and 1 for black. The disease variable, tumor grade, is coded 0 for well-differentiated tumors, 1 for moderately differentiated tumors, and 2 for poorly differentiated tumors. (Kleinbaum and Klein, 2002, pp. 305-309)

The following table is containing data summary of 288 patients.

|  |  |  |
| --- | --- | --- |
|  | White (E=0) | Black (E=1) |
| Well differentiated (D=0) | 104 | 26 |
| Moderately differentiated (D=1) | 72 | 33 |
| Poorly differentiated (D=2) | 31 | 22 |

## 3.1. Calculation of Odd Ratios

We separate the dataset based on Grade (D) and we will test whether the proportional odds assumption holds or not by calculating odd ratios.

The first table combines the well-differentiated and moderately differentiated levels.

|  |  |  |
| --- | --- | --- |
|  | White (0) | Black (1) |
| Well + Moderately differentiated (D= 0 and 1) | 176 | 59 |
| Poorly differentiated (D = 2) | 31 | 22 |

= = = 2.12

The second table combines the moderately and poorly differentiated levels.

|  |  |  |
| --- | --- | --- |
|  | White | Black |
| Well differentiated | 104 | 26 |
| Poorly differentiated | 103 | 55 |

= = = 2.14

The odds ratios for the first and second table are same. These two ratios represent that the proportional odds assumption is satisfied.

## 3.2. Calculation of Coefficient and Interval

To find coefficient and intercepts, we can use R to summarize the results because the calculation is very complex.

## Summary of Ordinal Logistic Regression

OLR <- polr(grade ~ race, data = data, Hess=TRUE)

summary(OLR)

Output:

Coefficients:

Value Std. Error

race 0.7555 0.2466

Intercepts:

Value Std. Error

Well defined|moderately defined -1.7388 0.1765

moderately defined|poorly defined -0.0089 0.1368

The odd ratio (exponentiated coefficient of race) indicates that the number of black women to have endometrial cancer are 2.13 (twice) times than white woman when data is categorized as poorly differentiated vs. moderately differentiated or well differentiated. To conclude, black women were holding the double possibility of having severe endometrial cancer than white women.

The result yields 2 intercepts, with each intercept corresponding to the log odds of a different inequality (depending on the value of g). Moreover, the log odds of g ≤ D is greater than the log odds of (g + 1) ≤ D (assuming category g is nonzero). This means that α1 > α2 > αG1. (Kleinbaum and Klein, 2002, p.312)

## 3.3. Confidence Interval

95% Confidence Interval for Race

= ]

= (1.31, 3.45)

In confidence intervals contain exponentiated coefficient of race (2.13). We are 95% confident that exponentiated coefficient of race will be between 1.31 and 3.45.

## 3.4. Likelihood Ratio Test (Wald Test)

H0 : [Beta coefficient has no relationship]

HA : [Beta coefficient has relationship]

Z=

*P value* = 0.002

For such *P value,* we reject null hypothesis. We can conclude that Race has significant association with tumor grade at 5% significance level. Wald test of the beta coefficient for Race is 0.002, indicating that Race is significantly associated with tumor grade at the 0.05 level.