**Ordinal logistic regression and its assumptions**

**Introduction**

In this dataset there are 50 different organisations with 50 rows and four columns. The four columns are organisation, ICT, REC and mentorship. Here the organisation column has the overall median ranking of that particular organisation after the tools have been applied form the second phase of the survey. The tools are ICT, REC and mentorship are coded as 0 and 1. Where 1 means it was applied to that particular organisation and 0 means not applied. The dataset has 3 predictor variables and 1 response variable.

**The purpose of this analysis is to discover which tool has most effect on the change in median score of different organisations.** Here, the median score can be from 0 to 5. Where 0 stands for Neutral, 1 stands for Strongly Disagree, 2 stands for Disagree, 3 stands for Neither Agree or Disagree, 4 stands for Agree and 5 stands for Strongly Agree.

Method.

Here the outcome variable is categorised and ranked so we can perform Ordinal logistic regression on the dataset. We have set the alpha level =0.05.

Null Hypothesis: There is no significant effect of a particular tool on organisation’s inclusivity.

Alternate Hypothesis: There is at least one tool affecting the organisation’s inclusivity significantly.

After cleaning the dataset (Glimpse)

Table

Description automatically generated

**Ordinal Logistic Regression**

Here in the dataset, we have DV is Categorical and ordered. The DV is Organisation with 6 ranked levels- Neutral, Strongly Disagree, Disagree, Neither Agree or Disagree, Agree and Strongly Agree. There are three IV’s Mentorship program, DISAB and REC.

**Applying the model.**



Table

Description automatically generated

Getting the P-values and storing it into coefficient tables.

Text

Description automatically generated

Here, we can tell which variables are statistically significant from the coefficient table by simply comparing the absolute value of coefficients. The variable with the largest value would be the most influential factor and in our case, it is REC (2.3585). This variable also has smaller p-value compared to other variables. Since there is at least one variable that is statistically significant, **we reject the H0 and we fail to reject the H1.**

Also, here the coefficients in the table are scaled in terms of logs and it is interpreted as “for one unit increase in REC, the log odds of having higher agreement for the median rank increases by 2.3585”. At times it becomes difficult to interpret, generally it is converted from log of odds to odds ratio.

Text

Description automatically generated

The above result is the conversion of coefficient parameters to proportional odds ratio and their 95% confidence interval. The interpretation of above result is “for one unit increase in REC, the odds of moving from N/A to Strongly Disagree or Disagree or Neither Agree or Disagree or Agree or Strongly Agree are **10.57** times greater, given that the other variables in the model are held constant.

The most statistically significant variable here is REC (Recruitment process) as the odds of moving from one rank to another is the highest with 10.57 times greater by holding other variables constant. In the explanation it means the respondents had the highest change in answering the ranks in the second phase of the survey.

Once we are done assessing whether the assumptions of our model hold, we can obtain predicted probabilities, which are usually easier to understand than either the coefficients or the odds ratios. For example, we can vary **each of these mentor, REC and DISAB** to calculate the probability which tools falls with what outcome. We do this by creating a new dataset of all the values to use for prediction.

Text

Description automatically generated

* So, to find out probabilities in terms of which tool has higher implications on getting A, DA, NA, SA or SDA.

Graphical user interface, text, application

Description automatically generated

**Assumptions of Ordinal Logistic regression**

Here after fitting the model we need to check the assumption to ensure that it is a valid model.

1. The dependent variable is ordered.
2. One or more of the independent variables are either continuous, categorical or ordinal.
3. No multi-collinearity
4. Proportional odds

**Multi-Collinearity**

Our dataset has satisfied first 2 assumptions. We will check for the third assumption now.

Text

Description automatically generated

Chart, bubble chart

Description automatically generated

The threshold for checking the multicollinearity is that the value should be <0.8 to accept that there is no high correlation. From the correlation plot we can see that there is not high correlation between the variables and hence there is no multicollinearity and assumption 3 is met.

**Proportional Odds**

Now we would use the Brant test to test the last assumption about proportional odds. This assumption basically means that the relationship between each pair of outcome group has to be the same.

Table

Description automatically generated

From the above result of the Brant test we conclude that the parallel assumption holds. The output also contains an Omnibus variable, which stands for the whole model, and it is still greater than 0.05. Therefore, the proportional odds assumption is not violated and the model is a valid model for this dataset.

**Conclusion**

The preliminary analysis and Ordinal logistic Regression analysis for WICE survey dataset. Based on the result, we conclude that the Recruitment Process (Tool applied) is the most influential factor that affects the median ranking of the organisation. For one unit increase in REC, the odds of moving from N/A to Strongly Disagree or Disagree or Neither Agree or Disagree or Agree or Strongly Agree are **10.57** times greater, given that the other variables in the model are held constant.