Introduction to

Ordinal Logistic Regression

Contents

Ordinal Logistic Regression

DEPENDENT VARIABLE

↓

Ordinal

(With two or more mutually exclusive and exhaustive categories)

INDEPENDENT VARIABLE

↓

Categorical or Continuous

- If there are **k categories** for the dependent variable then **(k-1) logit functions** are defined with remaining 1 category as base level.
- Here **coefficient of the variable is assumed to be same for each logit function** but **intercepts in logit functions differ**.

Ordinal Logistic Regression

Typical Examples of Ordinal and Scaled Variables

Revisiting a hotel	Very Unlikely	Unlikely	Neutral	Likely	Highly Likely
Movie verdict	Superhit	Hit	Average	Below Average	Flop
Quality of life	Very Good	Good	Bad	Very Bad	

Generally ordinal dependent variable is represented numerically, i.e. it is coded.

Eg. Quality of life can be coded as Very Good = 4, Good = 3, Bad = 2 and Very Bad = 1.

* Note that though the difference in Y values, between Very Good & Good and Good & Bad is equal to 1,it does not mean the difference in level of satisfaction is equal.

Application Areas

Industry	Model	Based on Information such as:
Human Resource	Employee Satisfaction	• Gender, Age, Department, Function, Years of Experience etc.
CRM	Customer Satisfaction	• Gender, Age, Buying Frequency, Location etc.
Stock Market	Stock Performance as Good , Average, Poor	• Net sale, Earning per share, Volatility Index etc.

Case Study – Brand Preference

Background

- A study was conducted to understand the customers' preference towards a brand. Data collected was customer demographics and their brand preference on a Likert scale.

Objective

- To study the factors that influence the brand preference.

Available Information

- Sample size is 259
- Independent Variables: Gender, Age, Location
- Dependent Variable: Brand Preference 1- Not Likely, 2-Likely, 3-Most Likely

Data Snapshot

Dependent Variables Independent Variables

Observations

id	Preference	Gender	Location	Age
1	3	MALE	CITY	<=25
2	2	MALE	CITY	25-40
3	1	MALE	CITY	40-55
4	2	FEMALE	CITY	25-40
5	2	FEMALE	CITY	40-55
6	1	FEMALE	CITY	25-40
7	2	FEMALE	CITY	<=25
8	2	FEMALE	SUBURBS	<=25

Column	Description	Type	Measurement	Possible Values
Id	Customer ID	numeric	-	-
Preference	Preference to the Brand	Categorical	1- Not Likely, 2-Likely, 3-Most Likely	3
Gender	Gender	Categorical	Male, Female	2
Location	Location	Categorical	City, Suburbs	2
Age	Age of the Customer	Categorical	<25, 25-40, 40-55	3

Model fitting in R

```
#Import the data
data<-read.csv("Brand Preference Study.csv", header=TRUE)
```

```
data$Preference<-as.ordered(data$Preference)

#Install and load package ‘MASS’
install.packages("MASS")
library(MASS)

prefmodel <- polr(Preference~Gender+Location+Age,data=data,Hess=TRUE)
effect<-summary(prefmodel)
effect
```

- **as.ordered**() tells R to treat variable “Preference” as Ordinal variable.

- **polr**() fits a Proportional Odds Logistic Regression. Dependent variable is followed by a ‘~’ and independent variables are separated by plus signs.
- **Hess=TRUE** ensures that the Hessian (the observed information matrix) is returned.

Model fitting in R

```
# Output:
```

```
Call:
polr(formula = Preference ~ Gender + Location + Age, data = data,
    Hess = TRUE)

Coefficients:
                   Value Std. Error t value
GenderMALE        1.1872     0.3420  3.4710
LocationSUBURBS -2.3863     0.2962 -8.0560
Age25-40        -0.2174     0.3141 -0.6923
Age40-55        -0.7511     0.3531 -2.1268

Intercepts:
    Value   Std. Error t value
1|2 -1.4568  0.3135    -4.6468
2|3  1.1904  0.3063     3.8859

Residual Deviance: 397.5779
AIC: 409.5779
```

- Output gives coefficient, standard error and t value for variables in each logit.

Individual Testing Using Wald's Test

- Individual testing is used for checking significance of each independent variable separately.

Objective	To test the **null hypothesis** that **each variable is insignificant**

Null Hypothesis (H_0): $b_i = 0$

Alternate Hypothesis (H_1): $b_i \neq 0$

$i=1,2\ldots,k$

Test Statistic	$Z^2 = (b_i / \text{Std. Error of } b_i)^2$ Under H0, $Z^2 \sim \chi^2_{(1)}$
Decision Criteria	Reject the null hypothesis **if p-value < 0.05**

Individual Testing in R

```
#Individual Testing
ptable<-data.frame(effect$coefficients)

ptable$pvalue<- 1-pchisq(ptable$t.value^2,df=1)

ptable$pvalue<-round(ptable$pvalue,4)
ptable
```

- ptable stores coefficients along with t values
- **pchisq()** is used to calculate p-values.
- pvalue stores table of p-values.

Individual Testing in R

```
# Output:
```

	Value	Std..Error	t.value	pvalue
GenderMALE	1.1871541	0.3420191	3.4710171	0.0005
LocationSUBURBS	-2.3862520	0.2962095	-8.0559606	0.0000
Age25-40	-0.2174104	0.3140560	-0.6922664	0.4888
Age40-55	-0.7510563	0.3531452	-2.1267637	0.0334
1|2	-1.4567802	0.3135024	-4.6467910	0.0000
2|3	1.1903988	0.3063378	3.8859027	0.0001

Interpretation :

- Gender, Location and age40-55 are significant, as p-value <0.05.

Classification Table

- **Cross tabulation** of observed values of Y and estimated values of Y is called as Classification Table.
- The predictive success of the ordinal logistic regression can be assessed by looking at the classification table

Classification			
Observed	Predicted		
	Not Likely	**Likely**	**Most Likely**
Not Likely	**108**	**24**	**1**
Likely	**34**	**56**	**4**
Most Likely	**2**	**24**	**6**

- Table shows that, model is predicting 66%=(108+56+6)/ 259 correctly.

Predicted Probabilities and Classification Table in R

```
# Predicted Probabilities
```

```r
data$predprob<-round(fitted(prefmodel),2)
head(data)
```

- **fitted**() generates predicted probabilities for brand preference.

```
# Output:
```

	id	Preference	Gender	Location	Age	predprob.1	predprob.2	predprob.3
1	1	3	MALE	CITY	<=25	0.07	0.43	0.50
2	2	2	MALE	CITY	25-40	0.08	0.47	0.45
3	3	1	MALE	CITY	40-55	0.13	0.55	0.32
4	4	2	FEMALE	CITY	25-40	0.22	0.58	0.20
5	5	2	FEMALE	CITY	40-55	0.33	0.54	0.13
6	6	1	FEMALE	CITY	25-40	0.22	0.58	0.20

Predicted category is 3(most likely) since it has highest probability 0.50

Interpretation :

- Predicted probabilities are given for each outcome likely, likely, most likely).
- Category with maximum of these probabilities is taken as predicted category of that observation.

Predicted Probabilities and Classification Table in R

```r
# Classification Table
expected<-predict(prefmodel,data,type="class")

ctable<-table(data$Preference,expected)
ctable
```

- **predict**() returns predicted values.
- **type="class"** returns a factor of classifications based on the responses (frequency). **type="probs"** returns matrix of probabilities.
- **table**() function simply gives the true positive and negative rates of the model (in the form of counts), which are key for deciding power of the model.

```
# Output:
   expected
      1   2   3
  1 108  24   1
  2  34  56   4
  3   2  24   6
```

Interpretation :

- Classification table of predicted and expected shows that, model is predicting 66%=(108+56+6)/259 correctly.

Quick Recap

In this session, we learned about Ordinal Logistic Regression :

Ordinal Logistic Regression	• Generally ordinal dependent variable is represented numerically, i.e. it is coded. • If there are k categories for the dependent variable then (k-1) logit functions are defined with remaining 1 category as base level. • Coefficient of the variable is assumed to be same for each logit function but intercepts in logit function differ.
Ordinal Logistic regression in R	• **`MASS`** library required for ordinal regression • **`polr()`** fits a Proportional Odds Logistic Regression. • **`predict()`** function with **`type=class`** returns predicted category,