Introduction to Multinomial Logistic Regression

Contents

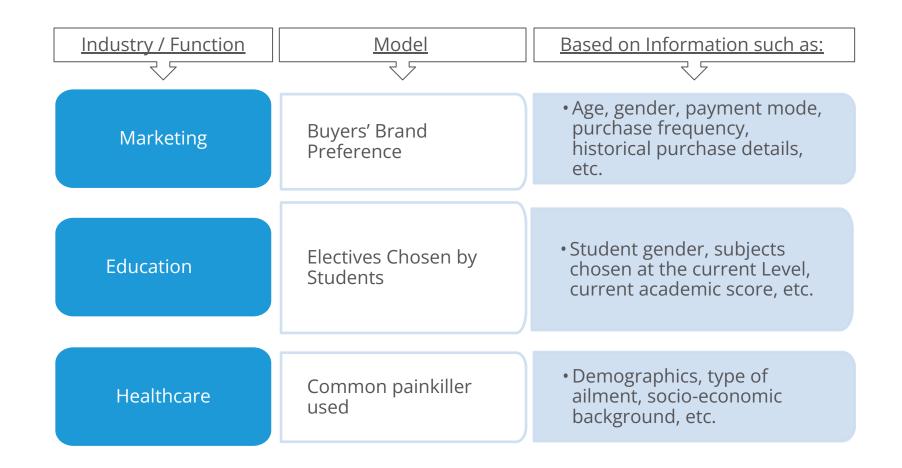
- 1. Basics of Multinomial Logistic Regression
- 2. Application areas
- 3. Statistical Model
- 4. Case Study
- 5. Model fitting in R
- 6. Predicted Probabilities and Classification Table

Multinomial Logistic Regression



• If there are k categories for the dependent variable, then (k-1) logit functions are defined with remaining 1 category as base level.

Application Areas



Statistical Model

- Let Y be the dependent variable with 3 categories as A,B,C and X1,X2,...Xk are k Independent variables.
- There will be 2 logit functions: one for Y=B versus Y=A and other Y=C versus Y=A Assuming A as the base category.

$$\begin{array}{l} g_1(x) = \mbox{ logit function for Y=B versus Y=A} \\ g_1(x) = \mbox{log}\left(\frac{P_B}{P_A}\right) \\ = \mbox{b}_{01} + \mbox{b}_{11}x_1 + \mbox{b}_{21}x_2 + + \mbox{b}_{k1}x_k \end{array} \qquad \begin{array}{l} \mbox{where,} \\ P_B = P \left[\mbox{ Y=B } \mid \times \mbox{ } \right] \\ P_A = P \left[\mbox{ Y=A } \mid \times \mbox{ } \right] \end{array}$$

$$\begin{aligned} g_2(x) &= \text{ logit function for Y=C versus Y=A} \\ g_2(x) &= \log \big(\frac{P_C}{P_A}\big) \\ &= b_{02} + b_{12}x_1 + b_{22}x_2 + + b_{k2}x_k \end{aligned} \qquad \text{where.}$$

Parameters of the model are estimated by the Maximum Likelihood Estimation(MLE)
 Method.

Case Study – High School Program Choice

Background

• At the time of entering high school, students make program choices among general program, vocational program and academic program. Their choice can be modeled using their writing score and their socio-economic status.

Objective

To model student's choice of programs.

Available Information

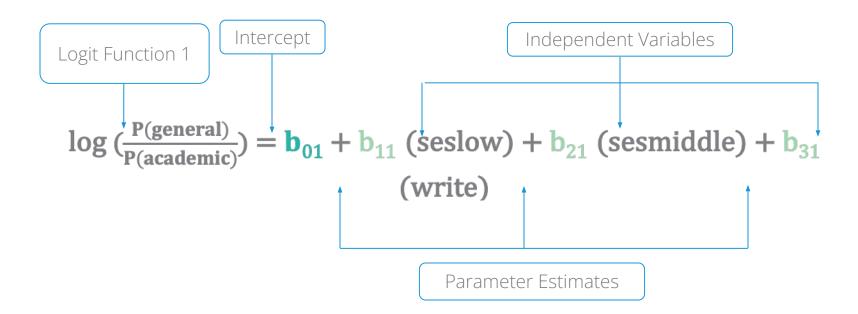
- Data source: https://stats.idre.ucla.edu/
- Sample size is 200
- Independent Variables: Socio-Economic Status (SES) and Writing Score.
- Dependent Variable: Program Chosen (General, Vocational or Academic)

Data Snapshot

High School Data		l	Independent Variables Dependent Variable					
		sn	id	E STATE OF S	write	prog		
		2		middle	50000	vocation general		
	Column	Descript	ion	Туре	Measu	rement	Possible Values	
	sn	serial nun	nber	numeric		_	-	
	id	student	id	numeric		-	-	
	ses	scoio-ecor status	(Categorical	low, mic	ldle, high	3	
	write	writing score studen	(continuous		-	positive value	
	prog	program cho studen	J /	categorical	gen	tional, eral, łemic	3	

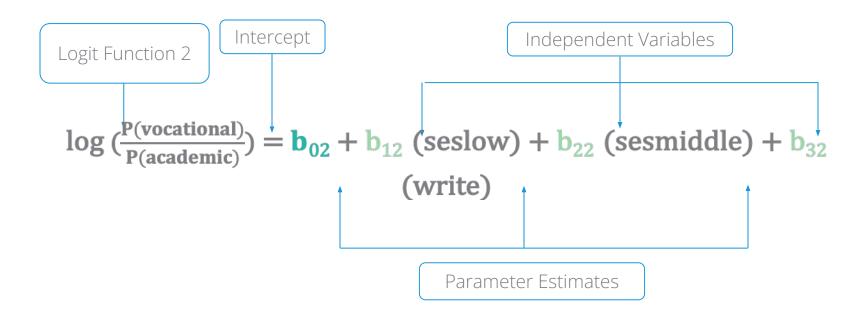
Model for the case study

- There are two categorical variables in the data: 'prog' and 'ses'.
 - For the Dependent variable 'prog', 'academic' is taken as base category.
 - For the Independent variable 'ses', high' is taken as base category.
- Model for the **general vs academic** is given as:



Model for the case study

• Model for the **vocational vs academic** is given as:



Maximum Likelihood Estimates of Parameters

Coefficients					
	Intercept	seslow	sesmiddle	write	
general	1.689478	1.1628411	0.6295638	-0.05793086	
vocation	4.235574	0.9827182	1.2740985	-0.11360389	

Standard Errors						
	Intercept	seslow	sesmiddle	write		
general	1.226939	0.5142211	0.4650289	0.02141101		
vocation	1.204690	0.5955688	0.5111119	0.02222000		

$$log(\frac{P(general)}{P(academic)}) = 1.689478 + 1.1628411(seslow) + 0.629568$$
(sesmiddle) + (-0.05793086) (write)

• Similar to this, there will be another model equation for the category 'vocation' with 'academic' as base category.

Model Fitting in R

#Import the data

```
data<-read.csv("High School Data.csv", header=TRUE)

data$prog<-relevel(data$prog, ref="academic")

# Install and load package 'nnet'.
install.packages("nnet")
library(nnet)</pre>
```

relevel() tells R to re-order levels of a factor so that the level specified by ref is first and the others are moved down. First level is then taken as reference (base) category.

Model Fitting in R

#Run Multinomial Logistic Model

choicemodel<-multinom(prog~ses+write,data=data)
m<-summary(choicemodel)</pre>

- mulinom() fits a Multinomial Logistic Regression.
 Dependent variable is followed by '~' and independent variables are separated by plus signs.
- The output of multinom() function does not contain all the parameters required for further testing.
- In order to be able to extract specific components from the output and perform more actions on them, an object is created from summary().

Model Fitting in R

Output

```
> choicemodel<-multinom(prog~ses+write,data=data)</p>
# weights: 15 (8 variable)
initial value 219.722458
iter 10 value 179.983731
final value 179.981726
converged
> m<-summary(choicemodel)
> m
call:
multinom(formula = prog ~ ses + write, data = data)
Coefficients:
                        seslow sesmiddle
         (Intercept)
                                                write
general
            1.689478 1.1628411 0.6295638 -0.05793086
vocation
            4.235574 0.9827182 1.2740985 -0.11360389
Std. Errors:
                        seslow sesmiddle
         (Intercept)
                                               write
            1.226939 0.5142211 0.4650289 0.02141101
general
            1.204690 0.5955688 0.5111119 0.02222000
vocation
Residual Deviance: 359.9635
AIC: 375.9635
```

 Output gives coefficients and standard errors of variables for 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): $b_{i1} = 0$ (for 1st logit) Alternate Hypothesis (H_1): $b_{i1} \neq 0$ ((for 1st logit) i=1,2...,kNull Hypothesis (H_0): $b_{12} = 0$ (for 2^{nd} logit) Alternate Hypothesis (H_1): $b_{12} \neq 0$ (for 2^{nd} logit) i=1,2...,k $Z^2 = (b_{i1} / Std. Error of b_{i1})^2$ **Test Statistic** Under H0, $Z^2 \sim \chi^2$ Decision Reject the null hypothesis if p-value < 0.05 Criteria

Individual Testing- Case study

Table of p-values							
	Intercept	seslow	sesmiddle	write			
general	0.16851638	0.023736	0.17579	6.816914e-			
	93	73	49	03			
vocational	0.00043826	0.098932	0.01267	3.176088e-			
	01	76	41	07			

• p-value for seslow (general), sesmiddle (vocational) and write (general and vocational) < 0.05

Interpretation of Results

Coefficients						
	Intercept	seslow	sesmiddle	write		
general	1.689478	1.1628411	0.6295638	-0.05793086		
vocational	4.235574	0.9827182	1.2740985	-0.11360389		
P-values						
general	0.1685163893	0.02373673	0.1757949	6.816914e-03		
vocational	0.0004382601	0.09893276	0.0126741	3.176088e-07		

- 'write' is a significant variable. Higher the writing score, less preference to 'general' or 'vocational'(as academic is base category and coefficient sign is negative).
- 'Low' SES category prefer 'general' over 'academic' more than 'high' SES category (as high SES is base category).
- 'middle' SES category prefer 'vocation' over 'academic' more than 'high' SES category.

Individual Testing in R

#Individual Testing

```
z<-m$coefficients/m$standard.errors

pvalue <-1-pchisq(z^2,df=1)

pvalue</pre>
```

- 'z' creates a dataframe of Z values as coefficients divided by standard errors
- pchisq() is used to calculate p-values using square of Z and degrees of freedom as arguments
- pvalue stores table of p-values.

Individual Testing in R

Output:

```
(Intercept) seslow sesmiddle write general 0.1685163893 0.02373673 0.1757949 6.816914e-03 vocation 0.0004382601 0.09893276 0.0126741 3.176088e-07
```

Interpretation:

seslow(general), write(general), sesmiddle (vocation), write(vocation) are significant, as p-value <0.05.

Classification Table

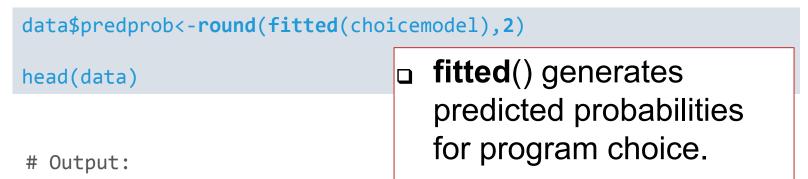
- Cross tabulation of observed values of Y and estimated values of Y is called as Classification Table.
- The predictive success of the logistic regression can be assessed by looking at the classification table

Classification					
	Predicted				
Observed	academic general		vocation	Percent Correct	
academic	92	4	9	87.61%	
general	27	7	11	15.56%	
vocation	23	4	23	46.00%	
Overall Percentage	71.0%	7.5%	21.5%	61.0%	

• Table shows that, model is predicting 61%=(92+7+23)/ 200 correctly.

Predicted Probabilities and Classification Table in R

Predicted Probabilities



	sn	id	ses	write	prog	predprob.academic	predprob.general	predprob.vocation
1	1	45	1ow	35	vocation	0.15	0.34	0.51
2	2	108	middle	33	general	0.12	0.18	0.70
3	3	15	high	39	vocation	0.42	0.24	0.34
4	4	67	low	37	vocation	0.17	0.35	0.48
5	5	153	middle	31	vocation	0.10	0.17	0.73
6	6	51	high	36	general	0.35	0.24	0.41

Predicted category is Vocation since it has highest probability 0.51

Interpretation:

- Predicted probabilities are given for each outcome (academic, general, vocation).
- ☐ Category of the maximum of these probabilities is taken as predicted category of that observation.

Predicted Probabilities and Classification Table in R

Classification Table

```
ctable<-table(data$prog,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 to deciding power of the model.</pre>
```

Output:

	expected		
	academic	general	vocation
academic	92	4	9
general	27	7	11
vocation	23	4	23

Interpretation:

Classification table of predicted and expected counts.

Quick Recap

In this session, we learned about **Multinomial Logistic Regression**:

Multinomial Logistic Regression

Multinomial Logistic regression in R

- Dependent variable is nominal with more than two categories and independent variables are categorical or continuous or mix of both.
- Parameters are estimated using MLE.
- If there are k categories for the dependent variable then (k-1) logit functions are defined with remaining 1 category as base level.
- relevel() used to define base category.
- nnet() library required for multinomial regression
- multinom() performs multinomial logistic regression
- Use summary() function to extract more details from multinom() function.