Ordinal logistic regression

Contents

1. Introduction

In the data file below, we show data from McFadden et. al. (2000) related to the preferences of drivers for various features in cars. Data and syntax files can be found in the **MVABOOK\Chapter3** folder.

carpref.LSF

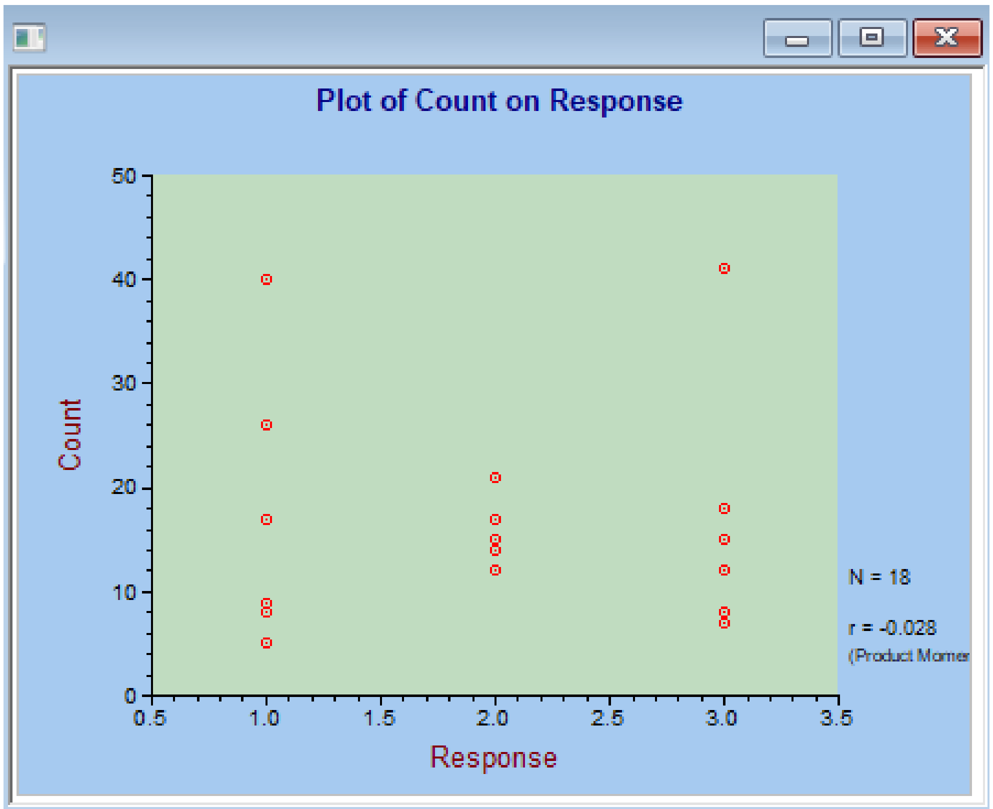
	Men	Agegrp	Response	Count
1	0.00	1.00	1.00	26.00
2	0.00	1.00	2.00	12.00
3	0.00	1.00	3.00	7.00
4	0.00	2.00	1.00	9.00
5	0.00	2.00	2.00	21.00
6	0.00	2.00	3.00	15.00
7	0.00	3.00	1.00	5.00
8	0.00	3.00	2.00	14.00
9	0.00	3.00	3.00	41.00
10	1.00	1.00	1.00	40.00
11	1.00	1.00	2.00	17.00
12	1.00	1.00	3.00	8.00
13	1.00	2.00	1.00	17.00
14	1.00	2.00	2.00	15.00
15	1.00	2.00	3.00	12.00
16	1.00	3.00	1.00	8.00
17	1.00	3.00	2.00	15.00
18	1.00	3.00	3.00	18.00

The first variable in the data set is Men, an indicator for gender coded 1 for males, and 0 for females. The age of a respondent is available as potential covariate Agegrp and is an ordinal variable with 3 categories. Ages 18 to 23 were assigned a value of 1 on this variable, ages 24 to 40 were assigned a value of 2, and a value of 3 was assigned for respondents over the age of 40.

The variable of interest is their response to questions regarding the desirability of features such as power steering or air conditioning in vehicles. Respondents rated features on this on a three-point scale, captured in the Response variable. A value of 1 was assigned if a feature was deemed on little importance, 2 if it was considered important, and a value of 3 was assigned if a respondent rated these features are very important. These data were originally published by Dobson and Barnett (2008, Table 8.1).

The final variable in the data set, Count, is the frequency of response in a cell in a frequency table of gender by response category.

A scatterplot of the frequency of response for each value of response is given below. There is considerable variation in the observed frequencies for all three response options. We will now explore what role gender and age of the respondents play in predicting their most likely response.



2. Ordinal logistic model

We fit an ordinal logistic model to these data and use the last category of the outcome ("Very important") as the reference category. The first (youngest) age group is used as the reference category for the ordinal variable Agegrp. We can express the model with the proportional odds parameterization as follows:

$$\ln\left(\frac{\pi_1}{\pi_2+\pi_3}\right)=\alpha_1+\gamma_1 x_1+\gamma_2 x_2+\gamma_3 x_3,$$

$$\ln\left(\frac{\pi_1+\pi_2}{\pi_3}\right)=\alpha_2+\gamma_1 x_1+\gamma_2 x_2+\gamma_3 x_3,$$

where x_1 represents males, x_2 represents an indicator variable for the age category 24 – 40, and x_3 is an indicator variable that assumes the value 1 for any respondent over the age of 40, and 0 otherwise. To request LISREL to automatically create the latter two indicator variables for the variable Agegrp, a "$" is used in the syntax file on the COVAR statement.

carpref1.prl

```
GlimOptions RefCatCode=0 Output=Residuals;
Title=car preferences;
SY=carpref.lsf;
Distribution=MUL;
Link=OLOGIT;
DepVar=Response Refcat = 3;
CoVars=Men Agegrp$;
Refcats=1;
Freq=Count;
```

Results are given below.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
Alpha1	0.0435	0.2303	0.1890	0.8501
Alpha2	1.6550	0.2536	6.5259	0.0000
Men	0.5762	0.2261	2.5484	0.0108
Agegrp2	-1.1471	0.2773	-4.1371	0.0000
Agegrp3	-2.2325	0.2904	-7.6871	0.0000

Using the model specification and the estimated coefficients

$$\ln\left(\frac{\hat{\pi}_1}{\hat{\pi}_2+\hat{\pi}_3}\right)=\hat{\alpha}_1+\hat{\gamma}_1 x_1+\hat{\gamma}_2 x_2+\hat{\gamma}_3 x_3,$$

$$\ln\left(\frac{\hat{\pi}_1+\hat{\pi}_2}{\hat{\pi}_3}\right)=\hat{\alpha}_2+\hat{\gamma}_1 x_1+\hat{\gamma}_2 x_2+\hat{\gamma}_3 x_3,$$

so that

$$\frac{\hat{\pi}_1}{\hat{\pi}_2+\hat{\pi}_3}=\exp\left[\hat{\alpha}_1+\hat{\gamma}_1 x_1+\hat{\gamma}_2 x_2+\hat{\gamma}_3 x_3\right],$$

$$\frac{\hat{\pi}_1+\hat{\pi}_2}{\hat{\pi}_3}=\exp\left[\hat{\alpha}_2+\hat{\gamma}_1 x_1+\hat{\gamma}_2 x_2+\hat{\gamma}_3 x_3\right],$$

We note that the odds ratio $\pi_1/(\pi_2+\pi_3)$ decrease by exp (1.1471) for respondents in the age group 25 to 40, and the odds ratio $(\pi_1+\pi_2)/\pi_3$ decrease by exp (-2.2325) for the respondents older than 40 relative to the young respondents (18-24).

We conclude that these extra features are more important to the younger drivers than to the older drivers. From the estimate for gender it seems that women are more inclined to rate these additional features as important relative to men.

By using these equations, we can calculate the expected probability of any respondent of certain age and gender picking one of the three outcomes. Or, if we want LISREL to do these rather tedious calculations for us, this can be accomplished by requestion the residual file using the Output = Residuals option on the GlimOptions command.

Contents of the residual file obtained as part of this analysis is shown below:

carpref1_RES.LSF

	)bservec	Prob.	Fitted	Raw
1	26.00	0.51	22.99	3.01
2	12.00	0.33	14.79	-2.79
3	7.00	0.16	7.22	-0.22
4	9.00	0.25	11.21	-2.21
5	21.00	0.38	16.89	4.11
6	15.00	0.38	16.91	-1.91
7	5.00	0.10	6.04	-1.04
8	14.00	0.26	15.53	-1.53
9	41.00	0.64	38.43	2.57
10	40.00	0.65	42.26	-2.26
11	17.00	0.25	16.44	0.56
12	8.00	0.10	6.30	1.70
13	17.00	0.37	16.33	0.67
14	15.00	0.38	16.55	-1.55
15	12.00	0.25	11.12	0.88
16	8.00	0.17	6.81	1.19
17	15.00	0.33	13.67	1.33
18	18.00	0.50	20.51	-2.51

The calculated expected cell probabilities for the subgroups are shown in the table below by gender and age.

Gender	Age	Response		
		Little importance	Important	Very important
Female	18-23	0.51	0.33	0.16
	24-40	0.25	0.38	0.38
	40+	0.10	0.26	0.64
Male	18-23	0.65	0.25	0.10
	24-40	0.37	0.38	0.25
	40+	0.17	0.33	0.50

We see that for women, the older respondents were most likely to rate these features as very important (probability is 0.64). Women under 24 did not think the features important factors in purchasing, with a probability of 0.51. The same trend is observed for males: the older respondents were more likely to rate features as very important, while your males did not (0.50 and 0.65 probabilities respectively). As the age of both genders increase, the probability of a response being “little importance” decreases over the three age categories, while the probability of a response being “very important” increases. While there are differences in the probabilities of responding in the “important” category, this category seems to capture roughly a third of responses regardless of age group.

A scatter plot of the fitted and observed frequencies shows a reasonably straight line, indicating that the current model does a reasonable job describing these data.