Nominal logistic regression

Contents

1. Introduction 1

2. Nominal logistic model 4

1. Introduction

In this example we consider data with a mixture of continuous, ordinal and nominal variables. A sample of 200 high school students' socio-economic status, test scores on a writing test and the type of program they were enrolled in is considered. The research question of interest is whether the writing scores and socio-economic status can be used to predict the type of program chosen by a student. Information on the first 15 students as contained in **programchoice.lsf** is shown below.

programchoice.LSF

	ses	program	write
1	1.00	3.00	35.00
2	2.00	1.00	33.00
3	3.00	3.00	39.00
4	1.00	3.00	37.00
5	2.00	3.00	31.00
6	3.00	1.00	36.00
7	2.00	3.00	36.00
8	2.00	3.00	31.00
9	2.00	3.00	41.00
10	2.00	3.00	37.00
11	1.00	3.00	44.00
12	3.00	2.00	33.00
13	1.00	3.00	31.00
14	2.00	3.00	44.00
15	1.00	3.00	35.00

The ordinal variable ses represents a student's socio-economic status, with 1 representing a low SES and 3 a high SES. The nominal variable program indicates the type of program a student is enrolled in. It is coded as follows:

1 = vocational program

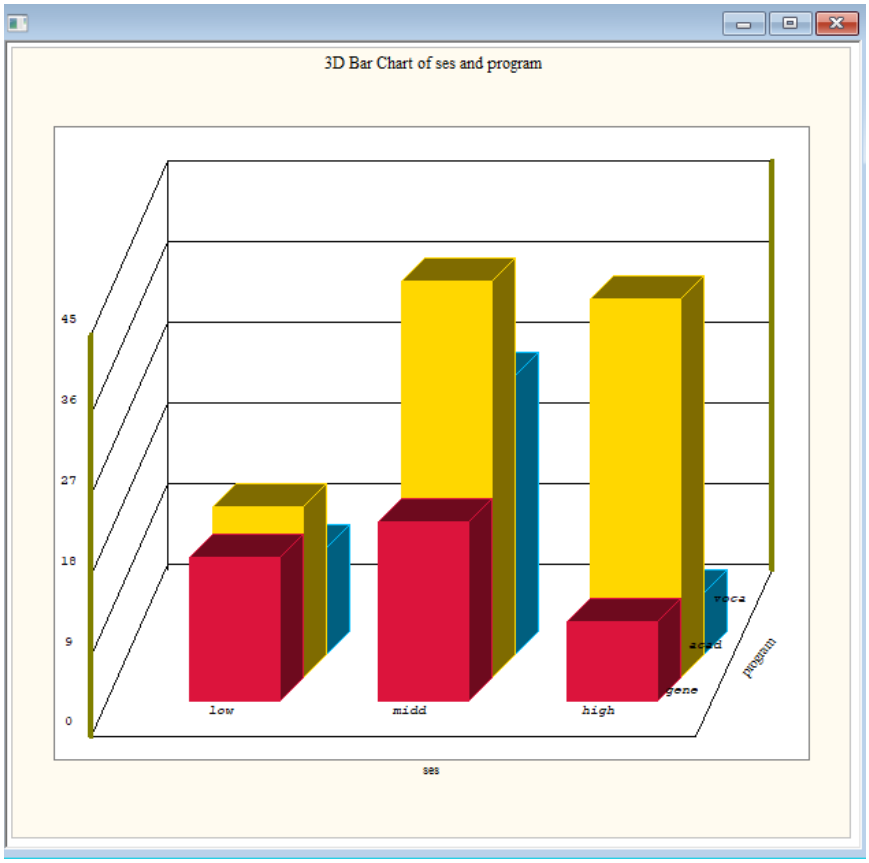
2 = general program

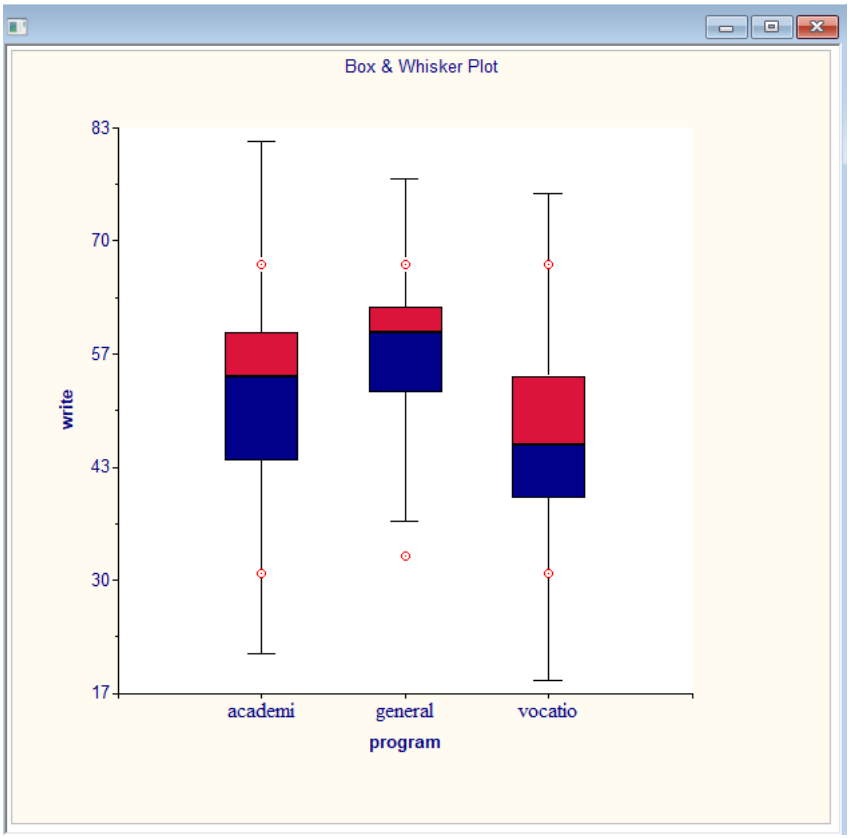
3 = academic program

Exploratory data analysis is performed using the **Data Screening** option from the **Statistics** menu.

```
Univariate Distributions for Ordinal Variables

      ses Frequency Percentage Bar Chart
    low       47        23.5    ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒
   midd       95        47.5    ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒
   high       58        29.0    ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒

  program Frequency Percentage Bar Chart
   gene       45        22.5    ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒
   acad      105        52.5    ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒
   voca       50        25.0    ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒

 Univariate Summary Statistics for Continuous Variables

 Variable     Mean  St. Dev.  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 --------     ----  --------   -------  --------  ------- -----  ------- -----
    write   52.775     9.479    -0.482    -0.750   31.000     4   67.000     7
```

A 3-D bar chart of the variables ses and program is given below. In all three SES groups, enrollment is mostly in the general program. However, in the middle SES group, a large number of students are enrolled in the academic program.





A box-and-whisker plot of writing scores and program choices show that distribution is negatively skewed, particularly for the general and academic group.

2. Nominal logistic model

An appropriate model for a dependent nominal variable such as program choice is nominal logistic regression. Suppose that π_1 represents the probability of an answer in category *j*. Since $\pi_1 + \pi_2 + \pi_3 = 1$ for the three possible program choices in this data, we need to select a reference category. Supposing we use the vocation group (the first category) as reference group. The model for nominal logistic regression can be expressed as

$$\ln(\pi_j / \pi_1) = \alpha_j + \gamma_{j1}x_1 + ... + \gamma_{jq}x_q, \quad j = 2,3$$

in the case of *q* regressors. There is a separate equation for each category of the outcome variable except for the first, i.e. the reference, category.

The syntax file below shows the commands required to set up a nominal logistic regression analysis using program choice as the outcome, and it's second category (general program) as the reference category. The two predictors write and ses are included in the CoVars statement and the ordinal nature of ses is indicated by the additional $ next to the variable name. Data and syntax files can be found in the **MVABOOK\Chapter3** folder.

programchoice1.PRL

```
GlimOptions RefCatCode=0 Output=All;
Title=Students Program Choices;
SY=programchoice.LSF;
Distribution=MUL;
Link=LOGIT;
Intercept=Yes;
DepVar=program Refcat=2;
CoVars=write ses$;
```

Results are given below. Given the second category of program has been selected as the reference category, the model estimated can be expressed as

$$\ln(\pi_1 / \pi_2) = \alpha_1 + \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3$$
$$\ln(\pi_3 / \pi_2) = \alpha_3 + \gamma_{31}x_1 + \gamma_{32}x_2 + \gamma_{33}x_3$$

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept 1	1.6894	1.2269	1.3769	0.1685
intcept 3	4.2355	1.2047	3.5159	0.0004
write 1	-0.0579	0.0214	-2.7056	0.0068
write 3	-0.1136	0.0222	-5.1127	0.0000
ses1 1	1.1628	0.5142	2.2614	0.0237
ses1 3	0.9827	0.5956	1.6500	0.0989
ses2 1	0.6295	0.4650	1.3538	0.1758
ses2 3	1.2741	0.5111	2.4927	0.0127

As both estimated effects for write are negative, the model indicates a decrease in the relative odds of general and vocational programs choice vs the academic program. Should ses increase from 1 to 3 (that is, from low to high) the relative odds of choosing the vocational program vs the academic program increases by 1.1628. Similarly, the relative odds of choosing the

vocational program compared to the academic program increases by 0.9827 for a change in ses from 1 to 3. Should ses increase from 2 to 3 (medium to high SES), the relative odds of choosing the vocation program vs the academic program increases by 0.630.

The goodness-of-fit statistics for this model indicates that we have a highly significant chi-square. To verify that, one could run the intercept-only model only to obtain another chi-square. The syntax file programchoice2.prl may be used for this purpose, though the result would only be a comparison between the current model and the intercept-only model, not a test of whether this model fits the data.

Goodness of Fit Statistics

Statistic	Value	DF	Ratio
Likelihood Ratio Chi-square	359.9524	192	1.8748
Pearson Chi-square	410.2703	192	2.1368
-2 Log Likelihood Function	359.9635		
Akaike Information Criterion	375.9635		
Schwarz Criterion	402.3500		