Logistic regression

Contents

1. Introduction

In a logistic regression the outcome variable is binary in nature. It quite frequently happens that the predictors of interest are categorical in nature. An example may be the response of residents in favor or against the building of a facility in their neighborhood. In such a case, results are frequently tabulated in a format such as the table below, with success defined as being in favor of the development.

Subgroup	In favor of	Against	Total
Males under 40	59 (y_1)	32 ($n_1 - y_1$)	91 (n_1)
Males above 40	44 (y_2)	47 ($n_2 - y_2$)	91(n_2)
Females under 40	66 (y_3)	20 ($n_3 - y_3$)	86 (n_3)
Females above 40	40 (y_4)	53 ($n_4 - y_4$)	93 (n_4)

Assuming that the random variables y_1 to y_4 are independently distributed with the same π, the proportion of successes in the subgroups can be expressed as $p_i = y_i / n_i$ and $E(y_i) = \pi_i$. These data may be viewed as frequencies for *N* binomial distributions (*N* = 4 in this case).

The logistic model for the response probabilities as functions of the predictors can be expressed as

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right]=\alpha+\boldsymbol{\gamma}'\mathbf{x}_i \qquad (1)$$

while the probit model may be written as

$$\Phi^{-1}\left[\pi(\mathbf{x})\right]=\alpha+\boldsymbol{\gamma}'\mathbf{x}_i \ . \qquad (2)$$

When rewritten in terms of π_i, we have

$$\pi_i(\mathbf{x})=\frac{e^{\alpha+\boldsymbol{\gamma}'\mathbf{x}_i}}{1+e^{\alpha+\boldsymbol{\gamma}'\mathbf{x}_i}}$$

for the logistic model and

$$\pi_i(\mathbf{x})=\Phi\left(\alpha+\boldsymbol{\gamma}'\mathbf{x}_i\right).$$

2. Logistic regression

In this example, we use tabulated data from Radelet and Pierce (1991) that report the number of death penalty verdicts for cases involving multiple murders in Florida during the time period 1976 to 1987. The number of death penalties is given by two additional categorical predictors, ethnicity of the defendant and ethnicity of the victim.

These data are given in **Dpv.lsf.** This file can be found in the **MVABOOK Examples\Chapter 2** folder. The first few lines of this file are shown below.

dpv.LSF

	COUNT	DP	VR	DR
1	53.00	1.00	0.00	0.00
2	11.00	1.00	0.00	1.00
3	0.00	1.00	1.00	0.00
4	4.00	1.00	1.00	1.00
5	414.00	0.00	0.00	0.00
6	37.00	0.00	0.00	1.00
7	16.00	0.00	1.00	0.00
8	139.00	0.00	1.00	1.00

The outcome variable of interest is the variable DP. The variable VR indicates the victim ethnicity and DR the defendant ethnicity, both assuming the value 1 if white, 0 otherwise.

We request a logistic regression of these two variables on DP using PRELIS syntax as shown below. The syntax for logistic regression in this analysis is rather different from that in the previous logistic regression analysis we fitted, where predictors were assumed to be continuous.

dpv1.prl

```
!Logistic Regression of Death Penalty Verdicts
sy=dpv.lsf
lr DP on VR DR
ou
```

Univariate summary statistics for the variables are given below. We note that the outcome variable is skewed and that most of the borrowers in the sample did not default on their loans.

```
Univariate Distributions for Ordinal Variables

       DP Frequency Percentage Bar Chart
      0     606        89.9    □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
      1      68        10.1    □□□□□

 Univariate Summary Statistics for Continuous Variables

 Variable     Mean  St. Dev.  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 --------     ----  --------   -------  --------  ------- -----  ------- -----
       VR    0.236     0.425     1.247    -0.447    0.000   515    1.000   159
       DR    0.283     0.451     0.964    -1.075    0.000   483    1.000   191


Test of Univariate Normality for Continuous Variables

             Skewness         Kurtosis      Skewness and Kurtosis

 Variable Z-Score P-Value   Z-Score P-Value   Chi-Square P-Value

       VR  10.619   0.000    -3.012   0.003      121.837   0.000
       DR   8.795   0.000   -14.423   0.000      285.380   0.000
```

The estimate regression equation is

```
Univariate Logit Regression for DP
       DP =  - 2.059  - 2.404*VR + 0.868*DR
 Standerr     (0.146)  (0.601)    (0.367)
 z-values     -14.121  -4.003      2.364
 P-values      0.000    0.000      0.018
```

The *p*-value of the estimated effect of victim ethnicity on whether a death penalty was handed down is highly significant, indicating that the ethnicity of the victim may have an effect on the probability of a death penalty. This does not seem to be the case for the defendant ethnicity.

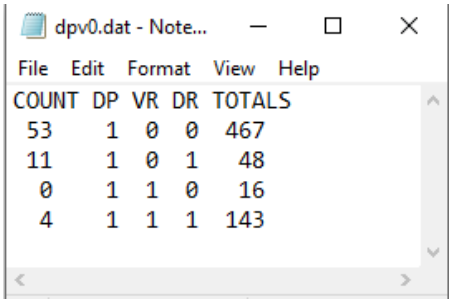
```
 -2lnL for Full Model                                              418.957
 -2lnL for Intercept-Only Model                                    440.843
 Chi-Square for Testing Intercept-Only Model                        21.886
 Degrees of Freedom                                                      2

 Pseudo-R²
 ---------
 McFadden                                                            0.050
 McFadden Adjusted                                                   0.041
 Cox & Snell                                                         0.032
 Nagelkerke                                                          0.067
```

The log-likelihood value reported for the intercept-only model is for the model that assumes that the probability of a cell is the same regardless of ethnicity of either victim of defendant. When compared with the log-likelihood value for the full model, as expressed by the chi-square of 21.886, we conclude that the intercept-only model does not describe the data as well as the full model. It indicates that ethnicity plays a role, which is in agreement with the results of the estimated

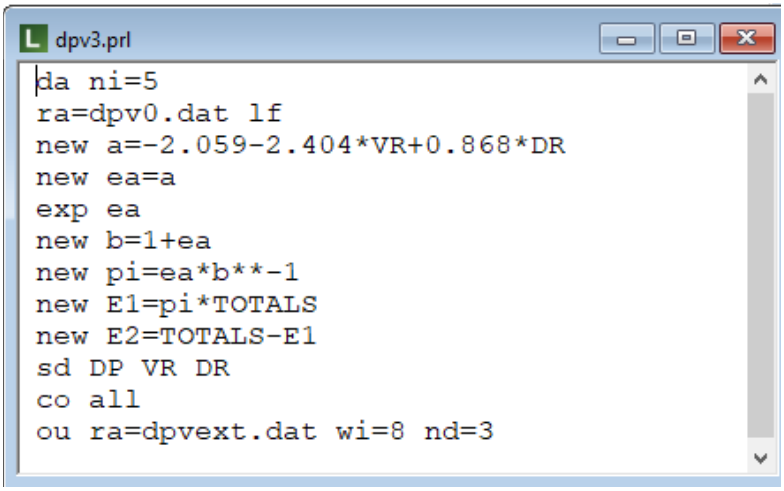
regression equation where we noted a statistically significant victim ethnicity effect. Note that these results do not necessarily indicate that the full model describes the data well, simply that the full model provides a better description than the intercept-only model.

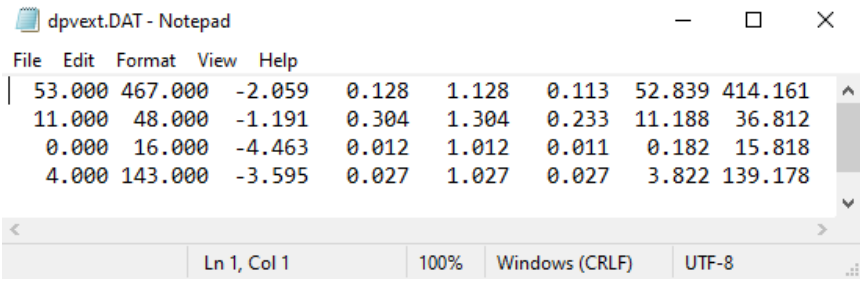
3. Calculating probabilities and expected frequencies

To take a closer look at the fit of the full model considered so far, we use LISREL to calculate the probabilities and expected frequencies for each cell in the data table. To do so, we create a small data matrix (**dpv0.dat**):



We next create a PRELIS syntax file (**dpv3.prl**). The *.**dat** file is used as input and the estimated effects of VR and DR obtained from the logistic regression model is used in calculation new variables E1 and E2 (in steps). Results are written to the file **dpvtext.dat** as requested on the OU line.



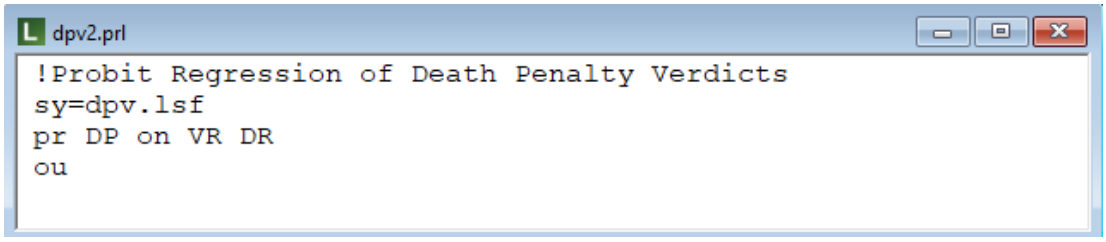


The table below shows the expected frequencies obtained under the logistic regression model. When compared to the observed frequencies in the LSF file, we see that the logistic model describes the data very well.

		Victim's race			
		White		Black	
		White defendant	Black defendant	White defendant	Black defendant
Death penalty	Yes	53	11	0	4
	No	414	37	16	139
Totals		467	48	16	143

4. Probit regression

We may also wish to examine how well a probit regression model fits these data. The syntax file dpv2.prl contains the syntax for fitting this model.



Note that the syntax is exactly the same as for the logit regression, with only one small change: exchanging LR with PR on the third line of the syntax file.

The estimated regression equation obtained is given below. As before, the estimated coefficient associated with the ethnicity of the victim (VR) is statistically significant, but that for the ethnicity of the defendant is not.

```
 Univariate Probit Regression for DP
       DP =  - 1.210  - 1.200*VR + 0.483*DR
 Standerr     (0.0762) (0.284)    (0.209)
 z-values     -15.880  -4.233      2.307
 P-values      0.000    0.000      0.021

-2lnL for Full Model                                    418.844
 -2lnL for Intercept-Only Model                         440.843
 Chi-Square for Testing Intercept-Only Model             21.999
 Degrees of Freedom                                           2

 Pseudo-R²
 ---------
 McFadden                                                 0.050
 McFadden Adjusted                                        0.041
 Cox & Snell                                              0.032
 Nagelkerke                                               0.067
```

When the fit statistics obtained are compared with those obtained for the logistic regression model, we find that they are essentially the same. We conclude that both these models fit the data equally well.