Scientific Software International

Logistic and Probit regression

Contents

1. Introduction

One of the most common non-linear regression models is logistic regression. When the outcome variable is binary in nature, the usual assumptions of linear regression are no longer tenable. In a linear regression model, the expected outcome can assume any value. In the case of a binary outcome, the expected outcome is either 0 or 1.

The logistic model can be expressed as

$$E(y \mid \mathbf{x}) = \frac{1}{1+e^{-(\alpha+\boldsymbol{\gamma}'\mathbf{x})}} \qquad (1)$$

Another suitable regression model for a binary variable *y* is the probit regression model. It can be shown mathematically that the results of probit and logistic regression results are approximately related by a scale factor.

2. Logistic regression

To illustrate, we use financial data. The main focus is to determine the credit risk of persons taking loans from a bank. The bank has the following information on this group of borrowers:

- AGE age in years
- EDUC education, measured on a four category scale
- EMPLOY the years of employment at the current job

- ADDRESS the years resident at the current address
- INCOME monthly income from regular position
- ADDINC additional income from other sources
- CREDDPT credit department input
- OTHRDEPT other department input
- DEFAULT indicating default on loan, assumes a value of 1 if the borrower defaulted, 0 otherwise.

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

Credit Risk.lsf

	AGE	EDUC	EMPLOY	ADDRESS	INCOME	ADDINC	CREDDEPT	OTHRDEPT	DEFAULT
1	41.00	3.00	17.00	12.00	176.00	9.30	11.36	5.01	1.00
2	27.00	1.00	10.00	6.00	31.00	17.30	1.36	4.00	0.00
3	40.00	1.00	15.00	14.00	55.00	5.50	0.86	2.17	0.00
4	41.00	1.00	15.00	14.00	120.00	2.90	2.66	0.82	0.00
5	24.00	2.00	2.00	0.00	28.00	17.30	1.79	3.06	1.00
6	41.00	2.00	5.00	5.00	25.00	10.20	0.39	2.16	0.00
7	39.00	1.00	20.00	9.00	67.00	30.60	3.83	16.67	0.00
8	43.00	1.00	12.00	11.00	38.00	3.60	0.13	1.24	0.00
9	24.00	1.00	3.00	4.00	19.00	24.40	1.36	3.28	1.00
10	36.00	1.00	0.00	13.00	25.00	19.70	2.78	2.15	0.00
11	27.00	1.00	0.00	1.00	16.00	1.70	0.18	0.09	0.00
12	25.00	1.00	4.00	0.00	23.00	5.20	0.25	0.94	0.00
13	52.00	1.00	24.00	14.00	64.00	10.00	3.93	2.47	0.00
14	37.00	1.00	6.00	9.00	29.00	16.30	1.72	3.01	0.00
15	48.00	1.00	22.00	15.00	100.00	9.10	3.70	5.40	0.00
16	36.00	2.00	9.00	6.00	49.00	8.60	0.82	3.40	1.00
17	36.00	2.00	13.00	6.00	41.00	16.40	2.92	3.81	1.00
18	43.00	1.00	23.00	19.00	72.00	7.60	1.18	4.29	0.00
19	39.00	1.00	6.00	9.00	61.00	5.70	0.56	2.91	0.00
20	41.00	3.00	0.00	21.00	26.00	1.70	0.10	0.34	0.00

The outcome variable of interest is the variable DEFAULT. We request a logistic regression of all other variables on DEFAULT using PRELIS syntax and the LR command as shown below. The AP option is included on the OU command to request the alternative parameterization for ordinal variables.

Credit Risk1.PRL

```
!Logistic Regression for Credit Risk
System File 'Credit Risk.LSF'
LR DEFAULT ON AGE - OTHRDEPT
OU AP
```

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

  DEFAULT Frequency Percentage Bar Chart
     No      517         73.9   ################################################
    Yes      183         26.1   #################

 Univariate Summary Statistics for Continuous Variables

 Variable      Mean  St. Dev.  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 --------      ----  --------  --------  --------  ------- -----  ------- -----
      AGE    34.860     7.997     0.361    -0.611   20.000     2   56.000     1
     EDUC     1.723     0.928     1.201     0.744    1.000   372    5.000     5
```

EMPLOY	8.389	6.658	0.831	0.233	0.000	62	31.000	3
ADDRESS	8.279	6.825	0.938	0.322	0.000	50	34.000	1
INCOME	45.601	36.814	3.859	26.170	14.000	6	446.000	1
ADDINC	10.261	6.827	1.096	1.219	0.400	1	41.300	1
CREDDEPT	1.554	2.117	3.899	21.980	0.012	1	20.561	1
OTHRDEPT	3.058	3.288	2.728	10.329	0.046	1	27.034	1

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
AGE	3.822	0.000	-4.791	0.000	37.565	0.000
EDUC	10.538	0.000	3.128	0.002	120.830	0.000
EMPLOY	7.978	0.000	1.237	0.216	65.187	0.000
ADDRESS	8.776	0.000	1.620	0.105	79.646	0.000
INCOME	20.277	0.000	14.430	0.000	619.408	0.000
ADDINC	9.865	0.000	4.407	0.000	116.752	0.000
CREDDEPT	20.368	0.000	13.957	0.000	609.665	0.000
OTHRDEPT	17.238	0.000	11.666	0.000	433.251	0.000

The estimate regression equation is

Univariate Logit Regression for DEFAULT

DEFAULT =	- 1.554	+ 0.0344*AGE	+ 0.0906*EDUC	- 0.258*EMPLOY	- 0.105*ADDRESS
Standerr	(0.619)	(0.0174)	(0.123)	(0.0332)	(0.0232)
z-values	-2.509	1.981	0.736	-7.787	-4.521
P-values	0.012	0.048	0.462	0.000	0.000

- 0.00857*INCOME	+ 0.0673*ADDINC	+ 0.626*CREDDEPT	
(0.00796)	(0.0305)	(0.113)	
-1.077	2.205	5.545	
0.282	0.027	0.000	

+ 0.0627*OTHRDEPT
(0.0775)
0.809
0.418

In contrast to the interpretation of the estimated coefficients of the predictors in linear regression, these estimated coefficients represent the logodds of defaulting on a loan associated with each of the predictors. We note that the logodds associated with input from the credit department (CREDDEPT) is considerably larger than for input from another department (OTHRDEPT). The latter is not statistically significant, though the estimated effect for the credit department input is highly significant. Educational level does not seem to have a significant impact (*p*-value is 0.462), though the effects associated with age, years of employment and years resident at the current address are. Surprisingly the impact of additional income on the probability of defaulting is statistically significant although the impact of regular income is not.

3. Probit regression

We now fit a probit regression model to the same data. This can be accomplished by changing the line

```
LR DEFAULT ON AGE – OTHRDEPT
```

in the syntax file to

```
PR DEFAULT ON AGE – OTHRDEPT
```

Results for this analysis are as follows.

```
Univariate Probit Regression for DEFAULT
  DEFAULT =  - 0.954 + 0.0192*AGE + 0.0480*EDUC - 0.142*EMPLOY - 0.0558*ADDRESS
 Standerr     (0.356) (0.0101)     (0.0709)      (0.0178)       (0.0129)
 z-values     -2.682   1.908        0.676        -7.947         -4.326
 P-values      0.007   0.056        0.499         0.000          0.000

             - 0.00395*INCOME + 0.0422*ADDINC + 0.347*CREDDEPT
              (0.00443)        (0.0173)        (0.0619)
              -0.892            2.436           5.602
               0.372            0.015           0.000

             + 0.0223*OTHRDEPT
              (0.0433)
               0.515
               0.606
```

We note that though the results for this analysis are numerically different from those obtained via logistic regression, the same pattern holds for predictors – those deemed to play a statistically significant role in explaining the probability of defaulting previously are the same for this model too.

4. Pseudo R-squares

In addition to the regular output, a logistic regression analysis also provides the following fit statistics.

```
-2lnL for Full Model                                  551.669
-2lnL for Intercept-Only Model                        804.364
Chi-Square for Testing Intercept-Only Model           252.695
Degrees of Freedom                                          8

Pseudo-R²
---------
McFadden                                                0.314
McFadden Adjusted                                       0.294
Cox & Snell                                             0.303
Nagelkerke                                              0.444
```

For the probit regression model, the similar statistics are:

```
-2lnL for Full Model                                  553.591
```

-2lnL for Intercept-Only Model	804.364
Chi-Square for Testing Intercept-Only Model	250.773
Degrees of Freedom	8

Pseudo-R²	
McFadden	0.312
McFadden Adjusted	0.292
Cox & Snell	0.301
Nagelkerke	0.441

The results for the two regressions are very similar. The four statistics under the "pseudo-R^2" heading are called pseudo as the R^2 cannot be calculated in the usual way for these models, as the outcome is binary rather than continuous. All four these pseudo-R^2 s are defined in terms of the estimated likelihoods for the full and intercept-only models, which are denoted by L_1 and L_0 respectively. The definitions of the four pseudo-R^2 values provided by LISREL are given below. In these formulae, *q* indicates the number of *x*-variables and *N* represents the sample size.

McFadden:

$$R^2 = 1 - \frac{\ln L_1}{\ln L_0}$$

McFadden Adjusted:

$$R^2 = 1 - \frac{\ln L_1 - q}{\ln L_0}$$

Cox and Snell:

$$R^2 = 1 - \left(\frac{L_1}{L_0}\right)^{(2/N)}$$

Nagelkerke:

$$R^2 = \frac{1 - \left(\frac{L_1}{L_0}\right)^{(2/N)}}{1 - L_1^{(2/N)}}$$