GLOBAL JOURNAL OF PURE AND APPLIED SCIENCES VOL 15, NO. 3&4, 2009: 449-452

 ISSN 1118-057

PROBIT REGRESSION IN PREDICTION ANALYSIS

M. E. NJA

(Received 12, December 2008; Revision Accepted 30, July 2009)

ABSTRACT

To avoid diagnostic surgery, the probability of nodal involvement of prostate cancer is modeled using the probit link function to determine whether the lymph nodes of a patients are infected. X-ray status and level of acid phosphatase in the blood serum of patients are considered as preoperative explanatory variables with the number of patients having nodal involvement as response variable. Within the framework of the probit regression model, the level of nodal involvement is predicted and the probability of nodal involvement obtained.

KEY WORDS: Probit model, Nodal involvement, Standard cumulative normal distribution, Latent variable, Logistic regression.

1. INTRODUCTION

For some dichotomous variables, the response y is actually a proxy for a variable that is continuous (Newsom, 2005). A regression analysis that predicts the underlying latent variable is called the probit regression. It is characterized by the probit link function defined as the inverse of the standard cumulative normal distribution. The standard cumulative normal distribution is the area to the left of the value Z on a standard normal distribution. This function maps the interval (0,1) to the real line.

Let $p_i = P_r(y_i = 0 \mid Z_i > 0)$ be the probability of non-involvement of a patient's lymph nodes in the ith group.

Mathematically, the probit regression is defined as

$$\Phi^{-1}(p_i) = Z_i = \frac{X_i - \mu}{\sigma} = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

where

$$p_i = \Phi(Z_i) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{\alpha+\beta x} \exp\left(-\frac{1}{2} Z_i^2\right) dz_i$$

Φ (Z), the standard cumulative normal distribution is the area under the curve between -∞ and $\alpha + \beta X$ as the standard normal variate Z is defined as $Z = \frac{X - \mu}{\sigma} = \frac{-\mu}{\sigma} + \frac{1}{\sigma} X = \beta_0 + \beta X$. It can also be regarded as the probability that Z lies between -∞ and $\alpha + \beta X$ where $X = (x_1, x_2)'$ is the vector of explanatory variables defined in section 2 below.

Z is the latent or unobserved continuous variable such as the level of success or failure. In our study, Z is the level of non-involvement of the lymph nodes of cancer patients. The dichotomous variable y_i is a proxy variable.

Byron and Brown (1980), in their study to determine which of five preoperative 'variables are predictive of nodal involvement in cancer of the prostate used logistic regression analysis, contingency table analysis, and other techniques in particular. The aim was to determine whether or not an elevated level of acid phosphatase in the blood serum would be of added value in the prediction of whether or not the lymph nodes were affected, given other four more generally used variables. In their study, acid phosphatase level was found to be directly proportional to nodal involvement.

In this work, probit regression is employed to determine the level of non-involvement of prostate cancer using acid phosphatase in the blood serum and X-ray status as preoperative explanatory variables. The level of non-involvement is the unobserved variable whose values are obtained by the probit model. The logistic model used by Byron and Brown does not have this ability. Also determined is the probability of non-involvement for each of the four categories of patients.

The log-likelihood function for probit is $\ln L = \sum w_i \ln \Phi(x_i \beta) + \sum w_i \ln[1 - \Phi(x_i \beta)]$ where w_i denotes optional weights.

M. E. Nja, Department of Mathematics/Statistics, Cross River University of Technology, Calabar, Nigeria

The probit link function is one of the link functions in generalized linear models.
Others are (Fox, 1997 & McCullagh, 1992):

Log link: $\ln \mu$

Inverse link: $\frac{1}{\mu}$

Square root link: $\sqrt{\mu}$

Logit link: $\ln\left(\frac{\pi}{1-\pi}\right)$

Log-log link: $\ln[-\ln(1-\pi)]$

2. The Probit Model

Probit $(p_i) = \Phi^{-1}(p_i) = Z_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

where

p_i = probability of non-involvement of lymph nodes in the ith group
$\Phi(Z)$ = Standard cumulative normal distribution
Z_i = Standard normal variate for the ith group
x_1 = Acid PH (explanatory variable)
x_2 = X-ray status (explanatory variable)

3. Example

The table below is the data on nodal involvement of cancer patients. It is required to model the probability of non-involvement of lymph nodes of patients using probit analysis. It is also required to model the level of involvement.

Table: Nodal involvement of cancer patients

Group 1	Acid PH x_1	X-ray x_2	No involvement y_i	Involvement	Total m_i
1	< 60	Negative	17	2	19
2	< 60	Positive	2	2	4
3	≥ 60	Negative	12	7	19
4	≥ 60	Positive	2	9	11
Total	**-**	**-**	**33**	**20**	**53**

Source: Byron, W. M., Brown, J. R. (1980)

4. Parameter estimates

From Systems Analysis Software (SAS) probit results, the parameter estimates for the probit model are obtained as follows:

β_0 (intercept) = 2.1176
β_1 (Acid PH) = -1.5695
β_2 (X-ray) = -2.0772

5. Probit Analysis

Group 1: Patients with Acid PH < 60, negative X-ray status
Probit $(p_1) = \Phi^{-1}(p_1) = Z_1 = 2.1176$
p_1 = prob ($y = 0 \mid x_1 < 60$, x_2 negative) = probability that a patient from group 1 is not involved
$p_1 = \Phi(Z_1) = \Phi\ (2.1176) = 0.9830$
Thus probability of involvement for group 1 is equal to 0.017

Group 2: Patients with Acid PH < 60, positive X-ray status
Probit $(p_2) = \Phi^{-1}(p_2) = Z_2 = 2.1176 - 2.0772 = 0.0404$
p_2 = prob ($y = 0 \mid x_1 < 60$, x_2 positive)
$p_2 = \Phi(Z_2) = \Phi\ (0.0404) = 0.5160$
Thus probability of involvement for group 2 is equal to 0.484

Group 3: Patients with Acid PH ≥ 60, negative X-ray status
Probit $(p_3) = \Phi^{-1}(p_3) = Z_3 = 2.1176 - 1.5695 = 0.5481$
p_3 = prob ($y = 0 \mid x_1 \geq 60$, x_2 negative)$p_3 = \Phi(Z_3) = \Phi\ (0.5481) = 0.7088$

Thus probability of involvement for group 3 is equal to 0.2912

Group 4: Patients with Acid PH ≥ 60, positive X-ray status
Probit (p_4) = $\Phi^{-1}(p_4)$ = Z_4 = -1.4596
p_4 = prob (y = 0 | $x_1 \geq 60$, x_2 positive)
$p_4 = \Phi(Z_4) = \Phi$ (-1.4596) = 0.0721
Thus probability of involvement for group 4 is equal to 0.9279

These values of probabilities can be compared with the proportions of non-involvement to check the accuracy of the fitted probability values as follows:

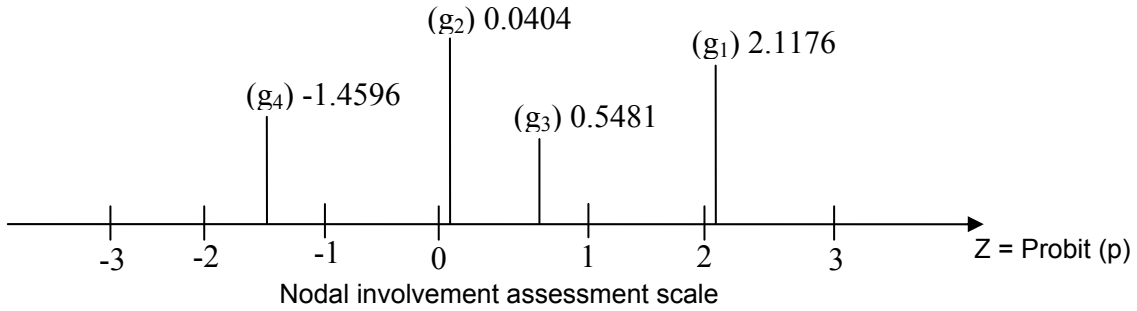
Group 1: Proportion of non-involvement

$$\text{Prop}_1 = \frac{17}{19} = 0.8947 \text{ (cf } P_1 = 0.9830)$$

Group 2: $$\text{Prop}_2 = \frac{2}{4} = 0.5000 \text{ (cf } P_2 = 0.5160)$$

Group 3: $$\text{Prop}_3 = \frac{12}{19} = 0.6316 \text{ (cf } P_3 = 0.7088)$$

Group 4: $$\text{Prop}_4 = \frac{2}{11} = 0.1818 \text{ (cf } P_4 = 0.0721)$$



Nodal involvement assessment scale

6. Discussion

Probit values can be used as a measure of the level of non-involvement of the nodes. Lower values depict higher nodal involvement. The degree of non-involvement is higher in group 1 (probit value 2.1176) followed by group 3 (probit value 0.7088), followed by group 2 (probit value 0.0404) and lastly group 4 (probit value -1.4596).

The probabilities of non-involvement (probabilities of survival) for the four groups comply with the probit assessment. The probability of non-involvement is highest in group 1 (p_1 = 0.9830). This is followed by group 3 (p_3 = 0.7088), followed by group 2 (p_2 = 0.5160) and lastly group 4 (p_4 = 0.0721).

High probit values are associated with groups 1 and 3 which have high levels of non-involvement. Low probit values are associated with groups 2 and 4 which have lower levels of non-involvement.

Going by acid PH classification, the lower level (> 60) is associated with the highest probit level (2.1176) and with a probability of non-involvement of 0.9830. The higher level (≥ 60) is associated with the lowest probit level (-1.4596) and with a probability of non-involvement of 0.0721.

7. Conclusion

From the foregoing discussion, it is concluded that the level of acid PH is directly proportional to the level of involvement (indirectly proportional to the level of non-involvement). The level of non-involvement of lymph nodes, the unobserved variable, has been modeled as a probit, thus making it possible to determine the level of non-involvement for each of the four groups of patients.

References

Byron, W. M., Brown, J. R., 1980. Prediction analysis for binary data. Biostatistics casebook. John Wiley & Sons. New York.

Fox, J., 1997. Applied regression analysis, linear models and related methods. Thousand Oaks: Sage.

Newson, R., 2005. Link Function and Probit Analysis. Data Analysis II. Statalist.

McGullagh, P., Nelder, J. A., 1992. Generalized Linear Models. Champman and Hall Madras.

APPENDIX

The Logistic Procedure

Model Information

Response Variable	non-inv
Number of Response Levels	2
Frequency Variable	Inv
Model	Binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	8
Number of Observations Used	8
Sum of frequencies Read	53
Sum of Frequencies Used	53

Response Profile

Ordered Value	Nodal Inv	Total Frequency
1	0	20
2	1	33

Probability modeled is non-inv = 0

Model Convergence Status
Convergence criterion (GCONV – E-8) satisfied
Deviance and Pearson Goodness-of-fit-Statistics

Criterion	Value	DF	Value/DF	Pr > Chi Sq
Deviance	0.0039	1	0.0039	0.9500
Pearson	0.0039	1	0.0039	0.9500

The Logistic Procedure
Number of unique profiles: 4

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	72.252	59.775
SC	74.222	65.686
-2 log L	70.242	53.775

Testing Global Nuli Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > Chi Sq	Chi-Square
Likelihood Ratio	16.4770	2	0.0003	16.4770
Score	15.3182	2	0.0005	15.3182
Wald	11.5809	2	0.0031	11.5809

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Sq
Intercept	1	2.1176	0.6506	10.5942	0.0011
Acid pH	1	-1.5695	0.7244	4.6945	0.0303
X-ray	1	-2.0772	0.7402	7.8745	0.0050