CALCULATING MARGINAL PROBABILITIES IN PROC PROBIT

Guy Pascale, Memorial Health Alliance

Introduction

The PROBIT procedure within the SAS system provides a simple method for estimating discrete choice variables (i.e. dichotomous or polychotomous). The difficulty with the procedure is that the parameter estimates are difficult to interpret. One way to ease this interpretation issue is to calculate marginal probabilities for each parameter estimate.

The purpose of this paper is to illustrate the use of marginal probabilities in the context of PROC PROBIT. This is done using the normal, logistic, and gompertz distributions that are available in the PROBIT procedure. The derivation of the marginal probability is illustrated for each distributional assumption. The predicted probabilities from each model are outputted and the marginal probabilities are calculated. The linear probability model (LPM) is also used to provide a baseline for comparisons across the distributions.

Derivation of Marginal Probabilities

In order to get the marginal probabilities, we must take the first derivative for each of the distributional assumptions. For the LPM, this is quite simple:

$$(\partial/\partial X_{ik})\ X_i\beta = \beta_k$$

Thus, the marginal probability in the LPM is the parameter estimate from the regression analysis. No additional manipulation is necessary. The same cannot be said for the normal, logistic, and gompertz distributions.

The first derivative for the normal distribution is as follows:

$$(\partial/\partial X_{ik})\ \Phi(X_i\beta) = \phi(X_i\beta)\beta_k$$

where Φ is the cumulative density function (CDF) for the normal and ϕ is the probability density function (PDF) for the normal. Thus the marginal probability assuming a normal distribution is the parameter estimate from the PROBIT multiplied by a standardization factor.

The first derivative for the logistic distribution is as follows:

$$(\partial/\partial X_{ik})\ L(X_i\beta) =$$

$$\frac{e^{x_i b}}{(1+e^{x_i b})^2} b_k$$

where L is the logistic distribution and equals 1/(1+exp(-x)).

We see that the marginal probability for the logistic distribution is the parameter estimate for the PROBIT multiplied by a standardization factor. This factor is the probability of being a 1 multiplied by the probability of being a 0.

The first derivative of the gompertz distribution is as follows:

$$(\partial/\partial X_{ik})\ G(X_i\beta) =$$

$$e^{x_i b} e^{-e^{x_i b}} b_k$$

where G is the gompertz distribution and equals 1-exp(-exp(x)).

Once again, we see that the marginal probability is equal to the estimated coefficient multiplied by a standardization factor.

Calculating Marginal Probabilities

There are two ways to calculate the marginal probabilities. One, the correction factor could be evaluated at the sample means.

For example, if the mean age were 35.2 years, the average level of educational attainment were 12.6 years, and the mean income level were $33,000, then standardization factor is calculated based on these means. We will not use this tact since no one in the sample will have these "average" characteristics.

For our purposes, we will calculate the value of the first derivative for each observation and then take the average of the standardization factor for the entire sample. This mean score, when multiplied by the parameter estimates from the PROBIT model, will give the marginal probabilities.

For the normal distribution, the standardization factor (in terms of SAS coding) is as follows:

```
Pdfnorm =
exp(.5*xbeta*xbeta)/sqrt(2*3.1459);
```

where xbeta is the predicted probabilities from the PROBIT procedure that specifies the normal distribution. The mean of Pdfnorm is the standardization factor for the normal distribution.

For the logistic distribution, the correction factor (in terms of SAS code) is as follows:

```
Problog1 = 1/(1+exp(-xbeta));
```

```
Problog0 = exp(-xbeta)/(1+exp(-
xbeta));
Prob0X1 = Problog1*Problog0;
```

Where xbeta is the predicted probabilities from the PROBIT procedure that assumes a logistic distribution. Problog1 is the probability of being a 1, Problog0 is the probability of being a 0, and Prob0X1 is the probability of being a 0 multiplied by the probability of being a 1. The mean of Prob0X1 is the standardization factor for the logistic distribution.

For the gompertz distribution, the standardization factor (in terms of SAS code) is as follows:

```
Pdfgomp = exp(xbeta)*exp(-
1*exp(xbeta));
```

Where xbeta, once again, is the predicted probabilities from the PROBIT procedure. The mean of Pdfgomb is the standardization factor for the gompertz distribution.

An Illustrative Example

To illustrate the calculation of the marginal probabilities, we will look at the factors that influence the probability of response to a patient satisfaction survey at Memorial Hospital of Burlington County. The hospital sends the satisfaction surveys to all in-patients excluding those who expired in the hospital, were patients on the Mental Health Unit, or were newborns.

The surveys have the patient's hospital ID number printed on them so that the surveys could be linked to hospital billing and clinical information. Patients were considered to be non-respondents if they failed to respond, scratched out there ID number on the survey, or failed to respond to how likely they were to recommending the hospital friends or family.

21,449 surveys were sent patients discharged between 4/23/96 and 10/23/97 with 4,474 useable surveys returned. The probability of response was modeled using length of stay, gender, the specialty of the service the patient received, type of insurance coverage, and whether the patient gave birth.

The probability of response to the survey is estimated four different ways: via the linear probability model using regression analysis and using the PROBIT procedure and separately specifying the normal, logistic, and gompertz distributions.

The calculation of the marginal probabilities entails to steps. First, the model is estimated with the predicted probabilities outputted to a separate data set. Second, the outputted probabilities are used to calculate the standardization factor for each observation. The mean of the factor is calculated and utilized to standardize the parameter estimate from the PROBIT. The results of the four models is presented in Table 1. The SAS programming is given in the Appendix

Three things stand out when looking at Table 1. First, the same variables are significant regardless of model specification. Second, the sign of the parameter estimates are identical and finally, the magnitude of the calculated marginal probabilities are similar.

For example, patients who have a length of stay of greater than 5 days are 4.6% less likely to respond than patients with shorter length of stay when using the LPM. When utilizing PROC PROBIT and specify the normal, logistic, or gompertz distributions, we note that patients with longer stays in the hospital have a 4.8%, 4.8%, and 4.7% lower probability of response, respectively.

The other significant variables also have similar marginal probabilities. General surgery patients have a 7.1%, 6.5%, 6.3%, and a 6.1% higher probability of response for the LPM, Probit, Logit, and Gompit. The results are similar for the remainder of the variables.

All the models find no relationship between gender, cardiology patients, gastro-enterology patients, neurology patients, having commercial insurance, CHAMPUS, NJ Blue Cross, or PPO and the likelihood of response.

The marginal probabilities do not match exactly for two reasons. One the LPM assumes a linear relationship. Thus, we expect to see differences between the LPM estimates and the maximum likelihood estimates (MLE) which are non-linear. Two, some slight differences are expected for the MLE estimates because the distributions are slightly different. This is especially true for the gompertz which is an extreme value function.

Conclusion

The methodology listed in the paper provides a simple method for easing the interpretation of parameter estimates in PROC PROBIT. By outputting the predicted probabilities from the normal, logistic, or gompertz distributions, a simple standardization factor can convert the parameter estimates to marginal probabilities.

Table 1
Comparison of Marginal Probability Calculations

Variable	LPM	Probit	Logit	Gompit
DAYS > 5	-.04603**	-.04774**	-.04765**	-.04740**
MALE	.00664	.00719	.00673	.00610
CARDIOLOGY	-.00588	-.00532	-.00609	-.00653
GASTRO-ENTEROLOGY	-.01121	-.01005	-.01163	-.01258
GENERAL SURGERY	.07073**	.06505**	.06321**	.06123**
NEUROLOGY	-.00741	-.00682	-.00762	-.00832
OBSTETRICS	-.10311**	-.11891**	-.12399**	-.12793**
ORTHOPEDICS	.02815*	.02563*	.02577*	.02613*
PULMONARY	-.03033**	-.03270**	-.03343**	-.03420**
COMMERCIAL INSURANCE	.02989	.02735	.02717	.02751
CHARITY CARE	-.10031**	-.11445**	-.11733**	-.12047**
CHAMPUS	.02061	.01914	.01957	.01969
HMO OTHER	-.01457	-.01422	-.01392	-.01376
MEDICARE	-.02631*	-.02625*	-.02585*	-.02587*
NJ BLUE CROSS	.02387	.02263	.02229	.02207
PPO	.01529	.01477	.01457	.01451
SELF-PAY	-.17717**	-.27440**	-.30858**	-.32983**
US HEALTHCARE	.06545**	.05813**	.05744**	.05693**
MEDICAID	-.10649**	-.13192**	-.14069**	-.14716**
DRG 371	.11682**	.13149**	.13592**	.13928**
DRG 372	.06104*	.07832**	.08169**	.08365*
DRG 373	.09009**	.10763**	.11195**	.11518**

** Significant at the 99% level
* Significant at the 95% level

Appendix

SAS Coding for Calculating Marginal Probabilities

```
libname guy 'd:\sasdata';

data one;
        set guy.merged6;

data lpm;
        set one;

proc reg s;

title 'OLS Regression (Linear Probability Model)';
title2 'dependent variable (survey) = 1 if person responded to survey';

model survey =  daysgt5 male cardio gastro gensurg neuro obstet ortho pulm fcc
                fcf fcg fch fcm fcn fcp fcs fcu fcx drg371 drg372 drg373;

run;


data probit;
        set one;

if survey eq 1 then nsurvey = 0; else nsurvey = 1;

proc probit;
title 'PROBIT with (nsurvey) = 0 if person responded to survey';

class nsurvey;
        model nsurvey =  daysgt5 male cardio gastro gensurg neuro obstet ortho
                         pulm fcc fcf fcg fch fcm fcn fcp fcs fcu fcx drg371
                         drg372 drg373 / converge = .00001;
        output out = probit2 xbeta = xbpr prob = probpr;


data probit2;
        set probit2;

pdf1norm = exp (-.5*xbpr*xbpr)/sqrt(2*3.1459);
probpr1 = 1-probnorm(-xbpr);
probpr0 = probnorm(-xbpr);
millslo = pdf1norm/(1-probnorm(-xbpr));
millshi = pdf1norm/probnorm(-xbpr);

proc means;
var pdf1norm probpr1 probpr0 millslo millshi;
run;


data logit;
        set one;

if survey eq 1 then nsurvey = 0; else nsurvey = 1;

proc probit;
title 'LOGIT with (nsurvey) = 0 if person responded to survey';

class nsurvey;
        model nsurvey =  daysgt5 male cardio gastro gensurg neuro obstet ortho
                         pulm fcc fcf fcg fch fcm fcn fcp fcs fcu fcx drg371
                         drg372 drg373
                         / d=logistic converge = .00001;
        output out = logit2 xbeta = xblog prob = problog;

data logit2;
        set logit2;

problo1 = 1/(1+exp(-xblog));
problo0 = exp(-xblog)/(1+exp(-xblog));
```

```
prob0X1 = problo1*problo0;
millslo = prob0X1/(1-problo0);
millshi = prob0X1/problo0;

proc means;
var problo1 problo0 prob0X1 millslo millshi;
run;



data gompit;
        set one;

if survey eq 1 then nsurvey = 0; else nsurvey = 1;

proc probit;

title 'gombit with (nsurvey) = 0 if person responded to survey';

class nsurvey;
        model nsurvey =  daysgt5 male cardio gastro gensurg neuro obstet ortho
                         pulm fcc fcf fcg fch fcm fcn fcp fcs fcu fcx drg371
                         drg372 drg373
                         / d=gompertz converge = .00001;
        output out = gompit2 xbeta = xbgomp prob = probgomp;


data gompit2;
        set gompit2;

pdfgomp = exp(xbgomp)*exp(-1*exp(xbgomp));
prgomp1 = 1-exp(-1*exp(xbgomp));
prgomp0 = exp(-1*exp(xbgomp));

proc means;
var pdfgomp prgomp1 prgomp0;
run;
```