SAS Handout: LOGIT and PROBIT MODELS

To illustrate the estimation of logit/probit models in SAS, we can consider the following example:

In 1973 in Troy, Michigan there was a referendum on whether the local tax rate should be increased to provide additional funding for schools. Rubinfeld (1977, "Voting in a local school election: A Micro Analysis", *Review of Economics and Statistics,* pp.30-42) took a survey of 425 voters in the school district and estimated the probability of voting yes to the increase as a function of various individual characteristics. For this handout, we undertake a similar exercise using a sample of 95 people chosen at random from the original 425.

Our analysis is based on the following model:

P(YESVM) = f(PUB12, PUB34, PUB5, PRIV, YEARS, SCHOOL, LOGINC, PTCON) (1)

Where

The data set contains a sample of 95 individuals. The variables are:

1. YESVM is a binary variable equal to 1 if the individual voted yes in the election; 0 if individual voted no.
2. PUB12 = 1 if 1 or 2 children in public school; = 0 otherwise
3. PUB34 = 1 if 3 or 4 children in public school; = 0 otherwise
4. PUB5 = 1 if 5 or more children in public school; = 0 otherwise
5. PRIV = 1 if 1 or more children in private school; = 0 otherwise
6. YEARS = number of years living in the community
7. SCHOOL = 1 if individual is employed as a teacher; = 0 otherwise
8. LOGINC = logarithm of annual household income (in dollars)
9. PTCON = logarithm of property taxes (in dollars) paid per year

In this handout, we will review the results from estimating (1) using the linear probability, logit, and probit models.

1. The Linear Probability Model

Since the main objective of this handout is to discuss estimation of the LOGIT and PROBIT models and to contrast the results, we willnot discuss the LPM estimates here. One concern about the LPM is that the predicted "probabilities" lie outside the [0,1] interval. For this model, there are 6 predicted probabilities outside the admissible range.

```
/*Estimation of the linear probability model*/
proc reg data=f.vote;
model yesvm=PUB12 PUB34 PUB5 PRIV YEARS SCHOOL logINC PTCON/p;
output;
run;
```

2. The Logit and Probit Models

Both these models can be estimated using PROC LOGISTIC in SAS. However, to ensure that SAS gives the output in the form we want it is necessary to sort the data. The command is

```
/* Sort the data*/
proc sort data=f.vote;
by descending yesvm;
run;
```

The above command sorts the data by the variable yesvm. The inclusion of descending guarantees that the observations for which y_t=1 (i.e. the event occurs) are listed first. This is crucial! The only difference in the syntax for these two estimates is that for the PROBIT we must include LINK=NORMIT as an option in the model statement.

These estimates can be used to test which variables are significantly affecting P(voting yes) and to indicate the nature of the relationship (i.e. positive or negative). In many cases we may also be interested in the impact of a change in an explanatory variable.

For a continuous variable this is given by

$$\delta P(y_t=1) / \delta x_{it} = (\exp(x_t'\beta) / [1+ \exp((x_t'\beta)]^2) * \beta_i$$

and for a dummy variable this is given by

$$P(y_t=1|x_{it}=1) - P(y_t=1|x_{it}=0)$$

In general both (2) and (3) depend on x_t. One solution is to evaluate them at the means of the continuous variables and at particular values of the dummy variables. To illustrate this I have calculated:

$\delta P(y_t=1) / \delta$LOGINC evaluated at PUB12=1, PUB34=PUB5=PRIV=0, YEARS=AVERAGE(YEARS), LOGINC=AVERAGE(LOGINC), PTCON=AVERAGE(PTCON).

This represents the marginal change in P(yt=1 resulting from an increase in LOGINC for a family with 1 or 2 childern in public school, no children in private school, and who pay the average amount of taxes and have lived in the community for the average period of time.

$P(y_t=1|$PUB12=1$) - P(y_t=1|$PUB12=0) evaluated at PUB34=PUB5=PRIV=0, YEARS =AVERAGE(YEARS), LOGINC=AVERAGE(LOGINC), PTCON=AVERAGE(PTCON). This figure represents the difference in the probability of voting yes between a family with 1 or 2 children in public school and a family with no children in public school and the characteristics specified above.

Now consider the PROBIT results. Please note that logit and probit estimates are not directly comparable because of the different variances of the distribution. For the probit model, to calculate the marginal effects if the variable is continuous

$$\delta P(y_t=1) / \delta x_{it} = \varphi(x_t'\beta)\, \beta$$

```
/*Estimation of the logit model*/
proc logistic data=f.vote order=data outest=f.logit_param;
model yesvm = PUB12 PUB34 PUB5 PRIV YEARS SCHOOL logINC PTCON/link=logit;
output out=f.logitprobs p=logit_phat;
run;

/*Estimation of the probit*/
proc logistic data=f.vote order=data outtest=f.probit_param;
model yesvm=PUB12 PUB34 PUB5 PRIV YEARS SCHOOL logINC PTCON/link=normit;
output out=f.probitprobs p=probit_phat;
run;

/*Merge the predicted probabilities of both models*/
data f.probs;
merge f.logitprobs f.probitprobs;
keep yesvm logit_phat probit_phat;
run;

proc print data=f.probs;
run;

/*Calculate the derivative of the probability that an individual votes YES evaluated at the mean of
the continuous variables*/
data f.stat1;
set f.stat;
keep my mp ml;
run;

data f.diflogit;
merge f.logit_param f.stat1;
x12 = intercept + PUB12 + my*YEARS + ml*loginc + mp*ptcon;
x0 = intercept + my*YEARS + ml*loginc + mp*ptcon;
difl = loginc*exp(x12)/(1+exp(x12))**2;
dif12= exp(x12)/(1+exp(x12)) - exp(x0)/(1+exp(x0));
run;
proc print data=f.diflogit;
var difl dif12;
run;

/*Same for the probit model*/
data f.difprobit;
merge f.probit_param f.stat1;
run;

data f.difprobit;
merge f.probit_param f.stat1;
x12 = intercept + PUB12 + my*YEARS + ml*loginc + mp*ptcon;
x0 = intercept + my*YEARS + ml*loginc + mp*ptcon;
difl = loginc*sqrt(7/44)* exp(-(x12**2)/2);
dif12= probnorm(x12) - probnorm(x0);
run;
```

Use the data in PNTSPRD.RAW for this exercise.

(i) The variable favwin is a binary variable if the team favored by the Las Vegas point spread wins. A linear probability model to estimate the probability that the favored team wins is

$$P(\text{favwin} = 1 | \text{spread}) = \beta_0 + \beta_1 \text{spread}.$$

Explain why, if the spread incorporates all relevant information, we expect) $\beta_0 = .5$.

(ii) Estimate the model from part (i) by OLS. Test H0:) $\beta_0 = .5$ against a two-sided alternative. Use the usual heteroskedasticity-robust standard errors.

(iii) Is spread statistically significant? What is the estimated probability that the favored team wins when spread = 10?

(iv) Now, estimate a probit model for P(favwin = 1| spread). Interpret and test the null hypothesis that the intercept is zero. [Hint: Remember that $\phi(0) = .5$]

(v) Use the probit model to estimate the probability that the favored team wins when spread = 10. Compare this with the LPM estimate from part (iii).

Mar 29

We want the function $F(z)$ to have the following 3 properties

① $F(z)$ should be a monotone increasing function of z,
If $z_1 < z_2$, then $F(z_1) < F(z_2)$

② $\lim_{z \to \infty} F(z) = 1$

③ $\lim_{z \to -\infty} F(z) = 0$

Probit

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} z^2\right) dz$$

Logit

$$F(z) = \frac{\exp(z)}{1 + \exp(z)}$$