

Part 4

Advanced Cross-sectional  
Models

Lecture 9

Binary Choice Models

Read	Wooldridge	ch 15
	Green	ch 21
	Verbeek	ch 7.1

Also, panel choice models

# Discrete Choice Models

①

Dependent variable is a dummy variable.

Eg) Mortgage rate choice  $\begin{cases} 1 = \text{fixed rate} \\ 0 = \text{adj. rate} \end{cases}$

Eg) Political vote  $\begin{cases} 1 = \text{Republican} \\ 0 = \text{Democratic} \end{cases}$

Choice models try to explain determinants of choice.

Eg)  $Y_i = \alpha + \beta_1 \text{Income}_i + \beta_2 \text{Edu}_i + \beta_3 \text{Race}_i + \dots + \epsilon_i$

$\begin{cases} 1 = \text{Rep} \\ 0 = \text{Demo} \end{cases}$

How do we estimate?

1) Can use OLS

i)  $\hat{y}_i = \hat{p}_i = P(y_i = 1)$  "probability"  
then called linear prob. model (LPM).

ii) Marginal effect

$\hat{\beta}_j = \frac{\Delta \hat{p}}{\Delta X_j} = \frac{\Delta \hat{f}}{\Delta X_j}$  change in prob. induced by 1 unit change in  $X_j$ .  
(cov difference)

iii) usual tests (t-test, F-test, LR test) can be used.

Problems

① Predicted prob  $\hat{p}_i$  can be  $> 1$  or  $< 0$ .

② Heteroskedasticity.  $\text{Var}(\epsilon_i) = p_i(1-p_i) \neq \text{constant}$   
can do OLS, though.

Note these problems may not be serious.

(2)

Note Advantages of the LPM

① Good approximation for common values of the covariates

(Problem occurs only at extreme cases;  $\hat{\beta}_i > 1$  or  $< 0$ )

- Good estimates of constant partial effects near the center of  $x$ .

② Can use robust std error or WLS (weighted LS : GLS) to correct for heteroskedasticity

③ More useful if most  $X_i$ 's take on only a few values or saturated.

i No reason to worry about the predicted probability lying outside the bound.

④ Easier to use in the panel data models with the LPM.

eg) Dynamic panel choice model  
(no clear solutions available for the FE or RE logit or probit models)

- Panel choice models often hard to estimate.

(3)

To guarantee  $0 \leq \hat{p}_i \leq 1$ , we consider two different transformations using cdf of

i) std. normal dist  $\rightarrow$  probit model

ii) logistic dist  $\rightarrow$  logit model

note other distributions can be used as well.

## 2) LOGIT Model

$$p_i = G(\beta'x_i) + \epsilon_i$$

where  $\beta'x_i = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$

$G$  is the cdf of the logistic dist.

s.t.

$$G(\beta'x_i) = \frac{1}{1 + e^{-\beta'x_i}} \stackrel{\text{let}}{=} p_i$$

Rewriting gives:

$$\log \frac{p_i}{1-p_i} = \beta'x_i + \epsilon_i$$

log of "odds"

$$\text{or } \text{logit } p_i = \beta'x_i + \epsilon_i$$

"log  $\frac{p_i}{1-p_i}$ "

Estimation : MLE

$$L = \prod_{i=1}^n p_i^{y_i} (1-p_i)^{1-y_i}$$

$$\log L = \sum_{i=1}^n (y_i \log p_i + (1-y_i) \log(1-p_i))$$

where  $p_i = \frac{1}{1 + e^{-\beta'x_i}}$

### Interpretation

- i) sign  $\hat{\beta}_j$ : direction of the effect of  $X_j$  on  $P$
- ii) Marginal effect (note it's not  $\hat{\beta}_j$ )

Report at means of  $X$ 's

$$\frac{d\hat{P}_i}{dx_{ij}} = g(\hat{\beta}'X_i) \cdot \hat{\beta}_j \quad \text{of course not constant.}$$

- ... depends on <sup>1</sup> the coeff. of  $X_j$  (ie  $\hat{\beta}_j$ )
- as well as <sup>2</sup> all other coefficients and
- <sup>3</sup> all (other) indep. var. ( $\hat{\beta}'X_i$ )

where  $g$  is the pdf of  $G$  st.  $g = G'$

$$g(\hat{\beta}'X_i) = \frac{e^{-\hat{\beta}'X_i}}{(1 + e^{-\hat{\beta}'X_i})^2}$$

iii) prediction

$$\hat{P}_i = \frac{1}{1 + e^{-\hat{\beta}'X_i}}$$

iv) testing hypothesis

$R^2$  is hardly defined  $\rightarrow$  thus No F-test  
 $\log L$  is well defined  $\rightarrow$  thus LR test is used.

Note t-test can be used on each coeff. (Fine.)

Note Standard error of Marginal effect can be computed by the delta method.  
 "partial derivatives of probabilities" at means of indep. variables.

### Distinguish!

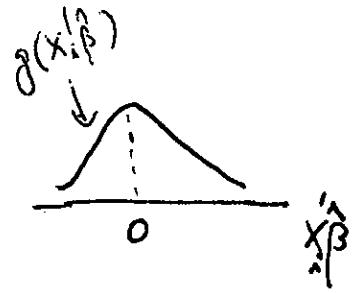
$$\hat{\beta} \rightarrow s(\hat{\beta})$$

$$\frac{d\hat{P}_i}{dx_{ij}} \rightarrow s\left(\frac{d\hat{P}_i}{dx_{ij}}\right)$$

(marginal effect)  $C g(\hat{\beta}'X_i) \hat{\beta}_j$   
 delta method (next page)

(More on Marginal effect)

$$i) \frac{d\hat{p}_i}{dx_{ij}} = g(\hat{\beta}'x_i) \cdot \hat{\beta}_j$$



: Maximized when  $x_i'\hat{\beta} = 0$

and diminishes as  $x_i'\hat{\beta}$  increases (or decreases)

ii) In the logit models, the relative effects do not depend on  $x_s$ .

$$\left(\frac{d\hat{p}}{dx_j}\right) / \left(\frac{d\hat{p}}{dx_k}\right) = \hat{\beta}_j / \hat{\beta}_k$$

iii) If  $x_k$  is binary (dummy or discrete),  $\hat{\beta}_k$  is NEVER a partial effect. Instead,

partial effect is to be evaluated as

$$= G(x_{k-1}\hat{\beta}_{k-1} + \hat{\beta}_k \cdot 1) - G(x_{k-1}\hat{\beta}_{k-1} + \hat{\beta}_k \cdot 0)$$

$(x_k=1)$   $(x_k=0)$

thus it also depends on all other  $\hat{\beta}$ 's

: Actually, this method is general, and can be applied to cases of continuous  $x$ 's

: This is obvious, but it might be one common mistake, when  $x_j$  is discrete.

\* iv) Std. error of  $\frac{d\hat{p}_i}{dx_{ij}}$  can be obtained by

the delta method;  $var\left(\frac{d\hat{p}}{dx}\right) = \frac{dg(\cdot)}{d\hat{\beta}} \cdot var(\hat{\beta}) \cdot \frac{dg(\cdot)}{d\hat{\beta}}$

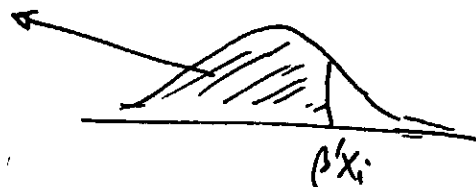
### 3) PROBIT Model

(6)

let

$$p_i = \Phi(\beta'X_i)$$

where  $\Phi(\beta'X_i)$  is the cdf of the std normal dist.



then

$$\Phi^{-1}(p_i) = \beta'X_i \quad \text{i.e.}$$

$$\uparrow = \alpha + \beta'X_i + \varepsilon_i$$

what is this: Inverse function of  $\Phi(\cdot)$   
there is no closed form solution,  
thus hard to express it  
(people use table, instead).

interpretation

i) sign of  $\hat{\beta}_j$

ii) Marginal effect

$$\frac{d\hat{p}_i}{dx_{ij}} = \phi(\hat{\beta}'X_i) \hat{\beta}_j$$

where  $\phi(\hat{\beta}'X_i)$  is the pdf. of the std. normal dist

$$\phi(\hat{\beta}'X_i) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\hat{\beta}'X_i)^2\right]$$

iii) prediction

$\hat{p}_i = \Phi(\hat{\beta}'X_i)$  using the table of the std. normal dist.

i) Testing hypothesis

(5)

No F-test. Use LR test.

t-test can be used for each coefficient.  
(same as logit models)

(Exercise)

Examples labor participation of women.

sel: choice-binary-log output.

use each of probit & logit model

i) Interpret the estimated coefficient.  
(sign)

ii) Marginal effect.

what is the effect of 2 additional years  
of education? (on average)

a) on average

b) of the individual who is a college  
graduate 25 years old female.

iii) Predicted probability

of a woman who is a college graduate  
and 25 years old.

iv) Test on significance of each coeff.

v) Test on the joint significance of  
AGE & GENDER.

(A)

Solution Here, 25 yrs old, 16 yrs of ed.

Gender = 0

First, Logit

$$\Rightarrow \text{pred } \hat{P}_i = \frac{1}{1 + \exp(-\hat{\beta}'X_i)}$$

$$\begin{aligned} \text{where } \hat{\beta}'X_i &= -5.266 - .0098 \text{ Age} \\ &\quad + .674 \text{ Edu} - 2.608 \text{ Gender} \\ &= -5.266 - .0098(25) \\ &\quad + .674(16) - 2.608(0) \\ &= 4.599 \end{aligned}$$

$$\hat{P}_i = \frac{1}{1 + \exp(-4.599)} = .990$$

$\therefore$  Two more yrs of edu.

$$= 2 \cdot \frac{\Delta \hat{P}_i}{\Delta \text{Edu}} = 2 \cdot g(\hat{\beta}'X_i) \cdot \hat{\beta}_j$$

$\downarrow 4.599$   
 $\downarrow .674$

$$\text{where } g(\hat{\beta}'X_i) = \frac{e^{-4.599}}{(1 + e^{-4.599})^2} = .0098$$

$$= 2 \times (.0098) \times .674 = .013$$

$\therefore$  hard to improve prob., as the prob is already close to 1.0

Second, probit

(9)

i) Prediction

$$\begin{aligned}\hat{p}_i &= -3.048 - .00613 \text{ Age} + .3869 \text{ Edu} \\ &\quad - 1.454 \text{ Gender} \\ &= -3.048 - .00613(25) + .3869(16) - 1.454(0) \\ &= 2.603\end{aligned}$$

$$\hat{p}_i = .9953 \text{ using } z\text{-table}$$



ii) Two more yrs of edu

$$= 2 \cdot \frac{\Delta \hat{p}_i}{\Delta \text{Edu}} = 2 \cdot \phi(\hat{p}_i X_i) \cdot \hat{\beta}_j$$

$\downarrow$   
 $2.603$   
 $\downarrow$   
 $\hat{p}_i X_i$   
 $\downarrow$   
 $.3869$   
 $\hat{\beta}_j$

$$\begin{aligned}\text{where } \phi(2.603) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(2.603)^2\right] \\ &= .0054\end{aligned}$$

$$= 2 \times (.0054) \times .3869 = .00415$$

Exercise, Repeat for a 34 yrs old, 14 yrs of educated man.

$$\begin{aligned}\text{Ans) } \hat{p} &= .7733 \text{ or } .7602 \quad (\text{logit w probit}) \\ 2 \times \text{way effect} &= .2363 \text{ or } .24 \quad ( " )\end{aligned}$$

(1)

```

-----
* File: choice_binary.do

log using choice_binary.log, replace
use choice_binary.dta
set more off

list job age school gender

regress job age school gender
predict pred_lpm, xb

logit job age school gender
predict pred_logit, p
mfx compute
*dlogit2 job age school gender

logistic job age school gender
predict pred_logistic, p
mfx compute

probit job age school gender
predict pred_probit, p
mfx compute
*dprobit job age school gender

list pred_lpm pred_probit pred_logit pred_logistic

log close

```

*marginal effect*  
 → (dlogit2 computes marginal effects)

→ "logistic" reports odd ratio of  $\hat{\beta}^*$

$$\hat{\beta}^* = \exp(\hat{\beta})$$

*marginal effects*

→ (dprobit computes marginal effects)

```

-----
log type: text
opened on: 21 Oct 2004, 16:19:18

. use choice_binary.dta

. set more off

. list job age school gender

```

	job	age	school	gender
1.	1	31	16	0
2.	1	34	14	1
3.	1	41	16	1
4.	0	67	9	0
5.	1	25	12	0
6.	0	58	12	1
7.	1	45	14	0
8.	1	55	10	0
9.	0	43	12	0

(ii)

10.	1	55	8	0
11.	1	25	11	0
12.	1	41	14	0
13.	0	62	12	1
14.	1	51	13	1
15.	0	39	9	1
16.	1	35	10	0
17.	1	40	14	1
18.	0	43	10	1
19.	0	37	12	1
20.	1	27	13	0
21.	1	28	14	0
22.	1	48	12	1
23.	0	66	7	1
24.	0	44	11	1
25.	0	21	12	1
26.	1	40	10	1
27.	1	41	15	0
28.	0	23	10	1
29.	0	31	11	1
30.	1	44	12	1

. regress job age school gender

Source	SS	df	MS	Number of obs =	30
Model	2.62055889	3	.87351963	F( 3, 26) =	4.96
Residual	4.57944111	26	.17613235	Prob > F =	0.0075
Total	7.2	29	.248275862	R-squared =	0.3640
				Adj R-squared =	0.2906
				Root MSE =	.41968

job	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0009682	.0066986	-0.14	0.886	-.0147373	.012801
school	.0911182	.0375996	2.42	0.023	.0138312	.1684052
gender	-.3803107	.1562374	-2.43	0.022	-.7014612	-.0591602
_cons	-.2227047	.6153338	-0.36	0.720	-1.487541	1.042132

. predict pred\_lpm, xb

. logit job age school gender

Iteration 0: log likelihood = -20.19035  
Iteration 1: log likelihood = -14.08727  
Iteration 2: log likelihood = -13.333675



Marginal effects after logistic  
 y = Pr(job) (predict)  
 = .69517648

variable	dy/dx	Std. Err.	z	P> z	[ 95% C.I. ]	X
age	-.0020954	.00821	-0.26	0.799	-.018188 .013997	41.3333
school	.1428596	.06298	2.27	0.023	.019429 .26629	11.8333
gender*	-.4849757	1.19975	-0.40	0.686	-2.83643 1.86648	.566667

(\*) dy/dx is for discrete change of dummy variable from 0 to 1

. probit job age school gender

Iteration 0: log likelihood = -20.19035  
 Iteration 1: log likelihood = -13.937494  
 Iteration 2: log likelihood = -13.305501  
 Iteration 3: log likelihood = -13.267639  
 Iteration 4: log likelihood = -13.26742

Probit estimates	Number of obs	=	30
	LR chi2(3)	=	13.85
	Prob > chi2	=	0.0031
Log likelihood = -13.26742	Pseudo R2	=	0.3429

job	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	-.0061331	.0237865	-0.26	0.797	-.0527538 .0404876
school	.3869652	.1759277	2.20	0.028	.0421532 .7317773
gender	-1.45382	.6299747	-2.31	0.021	-2.688547 -.2190919
_cons	-3.047916	2.399211	-1.27	0.204	-7.750285 1.654452

. predict pred\_probit, p

. mfx compute

Marginal effects after probit  
 y = Pr(job) (predict)  
 = .67502799

variable	dy/dx	Std. Err.	z	P> z	[ 95% C.I. ]	X
age	-.0022073	.00854	-0.26	0.796	-.018947 .014532	41.3333
school	.1392695	.06012	2.32	0.021	.021432 .257107	11.8333
gender*	-.4692287	.16282	-2.88	0.004	-.788354 -.150103	.566667

(\*) dy/dx is for discrete change of dummy variable from 0 to 1

. \*dprobit job age school gender

(V)

. list pred\_lpm pred\_probit pred\_logit pred\_logistic

	pred_lpm	pred_p~t	pred_l~t	pred_l~c
1.	1.205173	.9984285	.9945883	.9945883
2.	.6397215	.7602951	.7733448	.7733448
3.	.8151807	.9248186	.9245898	.9245898
4.	.532491	.5095153	.5346048	.5346048
5.	.8465095	.9253966	.9293296	.9293296
6.	.4342487	.4153232	.4113591	.4113591
7.	1.009382	.9818525	.9765006	.9765006
8.	.6352275	.6859546	.7173721	.7173721
9.	.8290821	.9085606	.9167091	.9167091
10.	.452991	.3860938	.3972671	.3972671
11.	.7553912	.8543728	.8701486	.8701486
12.	1.013255	.9829184	.9773914	.9773914
13.	.4303759	.4057835	.4018164	.4018164
14.	.5321442	.5855156	.5950862	.5950862
15.	.1792894	.1041526	.1003838	.1003838
16.	.6545911	.7281004	.7556962	.7556962
17.	.6339124	.7487157	.7627772	.7627772
18.	.266535	.185178	.1738827	.1738827
19.	.4545805	.4660999	.4623967	.4623967
20.	.9356914	.9653944	.961978	.961978
21.	1.025841	.9860208	.9800642	.9800642
22.	.4439305	.4393799	.4354945	.4354945
23.	-.029088	.013983	.0217048	.0217048
24.	.356685	.3032854	.2902687	.2902687
25.	.4700715	.5052068	.5018782	.5018782
26.	.2694395	.1901327	.1781852	.1781852
27.	1.104373	.9938793	.9883499	.9883499
28.	.2858986	.2197184	.204144	.204144
29.	.3692714	.3316926	.3174461	.3174461
30.	.4478032	.4490708	.4452417	.4452417

. log close

log type: text

closed on: 21 Oct 2004, 16:22:10

### Review Questions on Binary Choice Models

1. What are possible limitations of using the OLS estimator when the dependent variable is a dummy variable? Why is it called the LPM?
2. Discuss about how we use each of the LPM, Probit and Logit models, for the following.
  - (i) prediction equation
  - (ii) marginal effect (Is it constant over different individuals? If not, what is the term for this?)
  - (iii) testing hypothesis (Which tests can be used?)
3. Using the estimation result of the example on labor participation, for each of LPM, Probit and logit models,
  - (a) What is the predicted probability of labor participation for a 34 years old man with 14 years of education?
  - (b) If he is not employed, what will be the marginal effect on the increased probability of labor participation when he receives two more years of education?

### Exercise on Choice Models

**Objectives:** Analyze Qualitative Choice Models using the aacsb.xls file (web site).

**Data Background:**

The data file aacsb.txt contains data for 404 graduate business school programs on six variables, the names of which are contained in the first row of the file. The variable **AACSB** is a 1,0 dummy variable that indicates whether or not the American Assembly of Collegiate Schools of Business (AACSB) accredits the business school program. Three other variables indicate the number of students enrolled (**SIZE**), the average **GMAT** score of enrollees, and the proportion of faculty with doctorates (**FACPHD**). Two additional 1,0 dummy variables indicate whether the program granted a doctoral business degree (**PHD**), and whether the program was a public university (**STATE**). The data for this file come from a study of accrediting practices published by Jantzen and Pendleton ("Preferences of the American Assembly of Collegiate Schools of Business," *Journal of Education for Business*, Vol. 70, Sept/Oct 1994, pp. 6-11). About two-fifths of all graduate business programs were accredited at the time the data were collected (1992).

**Assignment:**

- (a) Read the data in. The appropriate STATA command is:

*insheet using aacsb.txt* (saved as a text file at the right folder)

- (b) To examine how the probability of being accredited is likely to be affected by a graduate business program's size, student GMAT scores, faculty doctorates, the presence of a doctoral business program, or being a publicly funded program, run the following ordinary least squares (OLS) regression:

*regress aacsb size gmat facphd phd state*

Interpret the regression coefficients, and conduct appropriate t-tests.

- (c) For how many observations, will the predicted probability lie outside the bound or 1 and 1?

*redict pred\_lpm, xb*

*list pred\_lpm*

- (d) Estimate a logit regression model explaining which schools became accredited by the AACSB and which were not. Then, conduct appropriate t-tests on the logit regression coefficients. Repeat for probit models.

*logit aacsb size gmat facphd phd state*

*probit aacsb size gmat facphd phd state*

- (e) Conduct a likelihood ratio test on whether or not the coefficients on the FACPHD and the PHD variables are both zeros each of the logit and probit models.

*logit aacsb size gmat state*

*probit aacsb size gmat state*

- (f) Using the marginal effects at means, determine which variables significantly affect the probability of being AACSB accredited in each of the LPM, logit and probit models.

*dlogit2 aacsb size gmat facphd phd state*

*dprobit aacsb size gmat facphd phd state*

[Note: dlogit2 needs to be installed. Help-search-all-dlogit2-install]

- (g) Find the partial effect of PHD, at means of regressors, each of the LPM, logit and probit models.

*dlogit2 aacsb size gmat facphd phd state*

*dprobit aacsb size gmat facphd phd state*

Suppose that one has collected the following data for Greene University:

SIZE	GMAT	FACPHD	PHD	STATE
300.	540.	90	0	1

- (h) Using each of the LPM, Logit and Probit models, find the probability that AACSB will accredit the business school program of the Greene University.
- (i) Using each of (i) LPM and (ii) Logit and (iii) Probit models, determine how much the probability will be changed if the Greene University implements a Ph.D. program.

# Choice Model: Motivation

(10)

$$\left\{ \begin{array}{l} y_i^* = \beta'X_i + \varepsilon_i \quad \text{"latent regression"} \\ \text{where } y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \quad \text{"random utility models (see Green p670)} \end{array} \right.$$

Note  $y_i^* = \text{benefit} - \text{cost}$ ; latent variable

- We do not know what are the factors separately for benefit or cost. We do not observe  $y_i^*$ , either.
- We only observe  $y_i$  and  $X_i$ .

eg) I observe  $y_i = 1$  if a student takes this class.  
I do not observe his (her) benefit or cost. Simply I believe that  $y_i^* > 0$ .  
I also observe  $X_i$  (gender, age, GPA, ...) which will determine  $y_i^*$ .

Note Common mistake in explaining the choice model

$$y_i = \beta'X_i + \varepsilon_i, \quad y_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$$

↑  
this should be  $y_i^*$ .

Note  $P(y_i^* > 0) = \Phi(\beta'X_i)$  or  $G(\beta'X_i)$  or  $\alpha + \beta X_i$

↑  
probit                      logit                      LPM

$P(y_i = 1) = P(y_i^* > 0)$

MLE (more)

Let  $G(x_i'\beta)$  be the cdf of the std normal or logistic dist.

$$\mathcal{L} = \prod_{i=1}^n [G(x_i'\beta)]^{y_i} [1 - G(x_i'\beta)]^{1-y_i}$$

$$\log \mathcal{L}_i = y_i \log [G(x_i'\beta)] + (1-y_i) \log [1 - G(x_i'\beta)]$$

$$\log \mathcal{L} = \sum_{i=1}^n \log \mathcal{L}_i$$

- foc (Score vector)

$$S_i(\beta) = \left[ \frac{g(x_i'\beta) \cdot [y_i - G(x_i'\beta)]}{G(x_i'\beta) [1 - G(x_i'\beta)]} \right] \cdot x_i = (\text{generalized residual}) \cdot x_i$$

- soc (Hessian, information matrix)

$$-E[H_i(\beta) | x_i] = \frac{[g(x_i'\beta)]^2 x_i \cdot x_i'}{G(x_i'\beta) [1 - G(x_i'\beta)]} \equiv A(x_i, \beta)$$

$$\text{Var}(\hat{\beta}) = \left[ \sum_{i=1}^n A(x_i, \hat{\beta}) \right]^{-1} \equiv \hat{V}$$

Note Q-MLE and Huber-White sandwich estimator

$$\text{Var}(\hat{\beta}) = \left( \sum_i \hat{A}_i \right)^{-1} \left( \sum_i \hat{S}_i \hat{S}_i' \right) \left( \sum_i \hat{A}_i \right)^{-1}$$

This can be used when the model is suspected to be mis-specified. eg  $\hat{p} = \Phi(x_i'\beta)$ ;  $\hat{p} = G(x_i'\beta)$  logistic

Then, it's QMLE. ( $\log p = x_i'\beta$ )

But, it's not imperative to use the sandwich estimator for  $\text{var}(\hat{\beta})$

STATA: option  $\Rightarrow$  robust

Exercise (a) show that in the logit model, the generalized residual is given as (12)

$$\left[ y_i - \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)} \right]$$

(b) let  $q = 2y - 1$ . show that

$$\ln L = \sum_i \ln G(q_i; X_i; \beta), \quad S_i(\hat{\beta}) = \frac{q_i \cdot g(q_i; X_i; \hat{\beta})}{G(q_i; X_i; \hat{\beta})} \cdot X_i$$

Note we can write

$$\begin{aligned} S(\hat{\beta}) &= \sum_{i=1}^n S_i(\hat{\beta}) = \sum_{i=1}^n \left[ y_i - \frac{\exp(X_i'\hat{\beta})}{1 + \exp(X_i'\hat{\beta})} \right] \cdot X_i \\ &= \sum_{i=1}^n (y_i - \hat{p}_i) X_i = 0 \end{aligned}$$

this implies

$$\sum \hat{p}_i X_i = \sum y_i X_i$$

### Testing hypothesis in binary choice models

$H_0: \gamma = 0$  (subset of parameters) : exclusion restrictions

∴ LR test

$$LR = 2(\log L_u - \log L_R) \sim \chi_m^2$$

$m = \#$  of restrictions

$\log L_R =$  restricted  $\log L$  imposing  $\gamma = 0$

∴ LM test

let  $y_i^* = X_i'\beta + \gamma_i + \varepsilon_i$ , and test  $H_0: \gamma = 0$ .  
(max1)

let  $\hat{\beta}$  be the probit estimator by imposing  $\gamma = 0$

$$y_i^* = X_i'\beta + u_i \Rightarrow \hat{\beta}, \hat{u}_i$$

then, obtain  $\hat{u}_i$  from the restricted model

(13)

$$\hat{u}_i = y_i - G(x_i; \hat{\beta})$$

Regress

$$\frac{\hat{u}_i}{\sqrt{\hat{G}_i(1-\hat{G}_i)}} \quad \text{on} \quad \frac{\hat{y}_i}{\sqrt{\hat{G}_i(1-\hat{G}_i)}} x_i, \quad \frac{\hat{y}_i}{\sqrt{\hat{G}_i(1-\hat{G}_i)}} z_i$$

LM =  $nR^2$  from this regression

$$\sim \chi^2_n$$

$\therefore$  Wald test

$$H_0: R\beta = r \quad \text{or} \quad r = 0$$

$$\text{Wald} = \hat{\beta}' \hat{V}^{-1} \hat{\beta} \quad \text{where } \hat{V} \text{ is given in page 11.}$$

$$\text{or } (R\hat{\beta} - r)' (R\hat{V}R)' (R\hat{\beta} - r)$$

$$\text{or } (R\hat{\beta} - r)' [J(\hat{\beta}) \hat{V} J(\hat{\beta})']^{-1} (R\hat{\beta} - r) \quad \text{if } R\beta = r \text{ is of a non linear form.}$$

$$\text{with } J(\hat{\beta}) = \frac{dR\hat{\beta}}{d\hat{\beta}'}$$

Testing for heteroskedasticity, solution of the issue

point: the probit (logit) estimator becomes inconsistent in the presence of heteroskedasticity.

If  $\text{Var}(e_i) = (\exp(z_i'\beta))^2$ , say, we can modify

$$\text{LNL} = \sum \left[ y_i \ln \left( \frac{x_i'\beta}{\exp(z_i'\beta)} \right) + (1-y_i) \ln \left( 1 - \frac{x_i'\beta}{\exp(z_i'\beta)} \right) \right]$$

... statz "HETPROB" procedure, eg.

Also, one can use LM test for testing heteroskedasticity

$$y_i^* = x_i' \beta + [-(x_i' \hat{\beta}) z_i] \gamma + \epsilon_i$$

$$H_0: \gamma = 0$$

and do the LM test : see Green, p.681.

Note Wooldridge (book, p.479) provides some insights on heteroskedasticity & non-normality.

$$P(y_i=1 | x_i) = G(x_i' \beta)$$

$$y_i^* = x_i' \beta + \epsilon_i, \quad \underbrace{\text{Var}(\epsilon_i) \neq \sigma^2}_{\text{heteroskedasticity}} \quad \text{or} \quad \underbrace{\epsilon_i \text{ is non-normal.}}_{\text{non-normality}}$$

- we're more interested in the 1st equation. If this is invalid, we lose the whole point. But QMLE is useful!
- we're not much interested in  $E(y_i^* | x_i)$ , where the heteroskedasticity issue arises.
- If  $\epsilon_i \sim$  logistic, but probit (std. normal) is used. Even in this case, the marginal effects of the probit model are reasonable. Thus, non-normality may not be a big issue.

Choice-based Sampling : Green, p673

eg) 'loan default' Our sample has more observations with default than the true population.

population :	$w_1$	$w_0$	
sample :	$p_1$	$p_0$	$(p_1 \gg w_1)$
	$(y_i=1)$	$(y_i=0)$	

$$\text{let } w_i = y_i \left( \frac{w_1}{p_1} \right) + (1-y_i) \left( \frac{w_0}{p_0} \right)$$

$$\log L = \sum (w_i) \ln G(q_i' \beta / x_i) \quad : \quad \frac{\text{weighted sampling}}{\text{MLE}}$$

### Goodness of fit

there are pseudo  $R^2$  and McFadden  $R^2$

$$1 - \frac{1}{1 + 2(\log L_1 - \log L_0)/n} \quad \swarrow \quad 1 - \log L_1 / \log L_0$$

where  $\log L_0$  is from the model with a constant only.

### Frequencies of actual & predicted outcomes

		Actual	
		1	0
Predicted	1	14	3
	0	4	9

threshold = 0.5  
7 obs mis-predicted.

stat: "1stat"

Note  $\beta$  and  $\sigma$  are not separately identified since  $y^*$  does not have a well defined unit of measurement. Thus, we obtain  $\hat{\beta}/\hat{\sigma}$  from usual softwares.

Note  $\hat{\beta}$ 's from the logit model & probit models are different. But the marginal effects  $d\hat{p}/dx_j$  appear similar.

### \* Proportions Data

grouped data  $P_i = \# \text{ of } 1\text{'s out of } n_i \text{ observation of group } i$

$$0 \leq P_i \leq 1$$

$$\log L = \sum_{i=1}^n n_i \{ P_i \ln G(x_i/\beta) + (1 - P_i) \ln [1 - G(x_i/\beta)] \}$$

or  $\log L = \sum_{i=1}^n \{ y_i \ln G(x_i/\beta) + (1 - y_i) \ln [1 - G(x_i/\beta)] \}$  "Waldridge fractional logit."  
where  $y_i = \text{proportions (not 1 or 0)}$

## Other estimators

(16)

- Semi-parametric estimators of the slope parameters (non-parametric kernel estimate)

... We do not need the function  $G(\cdot)$ .

and can estimate  $\hat{\beta}$ .

But, we cannot predict probabilities.

: Stoker (1986) & others

- Median estimator; Minimum score estimator by Manski (1975)

$$\text{Min} \sum_{i=1}^n |y_i - \mathbb{I}(x_i \beta > 0)| \quad \text{"absolute value"}$$

## Endogeneity problem of 2SLS

(17)

$$y_1^* = z_1 \delta_1 + \gamma \underbrace{(y_2)} + u_1$$

endogenous (continuous)

We simply ASSUME that a reduced form for  $y_2$  exists.

(if  $y_1$  appears in the  $y_2$  equation, there is no reduced form equation for  $y_2$ ; due to non-linearity)

$$y_2 = z_1 c_1 + z_2 c_2 + v_2 = zc + v_2 \quad z = (z_1, z_2)$$

( $u_1$  &  $v_2$  are correlated)

1) Testing for endogeneity in probit or logit models.

write  $u_1 = \theta_1 v_2 + e_1$ . then  $\theta_1 = \frac{\text{cov}(u_1, v_2)}{\text{var}(v_2)}$

$$y_1^* = z_1 \delta_1 + \gamma y_2 + (\theta_1 v_2 + e_1)$$

where  $\text{var}(e_1) = \text{var}(u_1) - \text{corr}(u_1, v_2)^2$

Thus, we can set up the test

① Run OLS on the reduced form of  $y_2$

$$y_2 = z\hat{c} + \hat{v}_2$$

and obtain the residual,  $\hat{v}_2$ .

② Run Probit  $y_1$  on  $z_1, y_2$  and  $\hat{v}_2$

$$y_1^* = z_1 \delta_1 + \gamma y_2 + \theta_1 \hat{v}_2 + e_1$$

Test  $H_0: \theta_1 = 0$  implying  $\text{corr}(u_1, v_2) = 0$

(t-test)

## 2) Estimation

i) 2-step estimation (the same as the previous testing procedure)

... do not use  $\hat{y}_2$ . Just add  $\hat{v}_2$  in the probit or logit estimation.

If  $\theta_1 = 0$  is rejected (t-test), then reported  $S(\hat{\theta}_1)$  is not valid.

... when  $S(\hat{\theta}_1)$  is obtained, the correlation between  $(u_1, v_2)$  is not accounted for.

... see Rivers & Vuong (1988); textbook p. 474

p. 361  
Correction

ii) full MLE

see textbook, p. 476

where

$$w_i = \frac{z_{1i}\delta_1 + \alpha_2 y_{2i} + (\rho/\sigma_2)(y_{2i} - z_{2i}\delta_2)}{(1-\rho^2)^{\frac{1}{2}}}$$

$$f(y_1, y_2) = \underbrace{f(y_1 | y_2)}_{\Phi(w)} \underbrace{f(y_2)}_{\text{normal}}$$

$$= (\Phi(w))^{y_1} (1 - \Phi(w))^{1-y_1} \cdot f(y_2)$$

$$L = \prod_{i=1}^n f(y_{1i}, y_{2i})$$

the MLE accounts for  $\text{CORR}(u_1, v_2) \equiv \rho$

but this parameter is often troublesome.

( $\rho \rightarrow 1$  or  $-1$ ) not converging, if so.

# Binary Endogeneous Regressor

$$y_1^* = z_1 \beta_1 + \gamma(y_2) + u_1 \quad = y_2 \text{ is binary, too.}$$

$$y_2^* = z_2 \beta_2 + v_2$$

$$z = (z_1, z_2)$$

eg)  $y_1 = \begin{cases} 1 & \text{dividend pay} \\ 0 & \text{o/w} \end{cases} \quad y_2 = \begin{cases} 1 & \text{bank ownership} \\ 0 & \text{o/w} \end{cases}$

eg)  $y_1 = \begin{cases} 1 & \text{Catholic school} \\ 0 & \text{o/w} \end{cases} \quad y_2 = \begin{cases} 1 & \text{Catholic} \\ 0 & \text{o/w} \end{cases}$

One tempting but incorrect approach:

1st Do probit of  $y_2^*$ , and obtain  $\Phi(z\hat{\beta})$

2nd Replace  $y_2$  with  $\Phi(z\hat{\beta})$

... this is an example of "forbidden regression"

(see WI, p. 478 and p 230-235)

"the expectation cannot pass through non-linear indicator functions"

MLE is easier (really?) and more efficient.

Combine 4 possible outcomes of  $(y_1, y_2)$  and

construct log L using  $P(y_1 = i | y_2 = j)$

$i, j = 1, 0$ ; see textbook p 478  
WI

# Panel choice Models

20

1) pooling probit or logit

$$\text{Max}_{\beta} \sum_{i=1}^N \sum_{t=1}^T \left[ y_{it} \log G(X_{it}\beta) + (1-y_{it}) \log \Phi(X_{it}\beta) \right]$$

$$y_{it}^* = X_{it}'\beta + u_{it}$$

Note Testing for "dynamic" models

(possible lagged dep. variables)

① Do pooling probit or logit, and obtain residuals.

$$y_{it} - \Phi(X_{it}'\hat{\beta}) \equiv \hat{u}_{it}$$

② Do pooling probit or logit with  $\hat{u}_{i,t-1}$  added.

$$P(y_{it}=1) = \Phi(X_{it}'\beta + \delta_1 \hat{u}_{i,t-1})$$

Test  $H_0: \delta_1 = 0$  (no dynamics)

If  $H_0$  is rejected, consider a dynamic model; say

$$y_{it}^* = X_{it}'\beta + \alpha y_{i,t-1} + u_{it}$$

∴ still, can do dynamic LPM.

## 2) Panel choice models

$$P(y_{it} = 1) = \Phi(X_{it}'\beta + c_i)$$

$\underbrace{\hspace{10em}}$   
 unobserved heterogeneity.

### ① FE probit (seldom)

$c_i$  is the parameter to estimate.

- 'demeaning' is not possible;  $y_{it}^* = y_{it} - \bar{y}_i$   
 works only for linear models.

- incidental parameter problems  
 (too many parameters to estimate)

⇒ FE probit is "rarely" used.

### ② RE probit (hard) (but commonly used)

$$f(y_1, \dots, y_T) = \int_{-b}^b f(y_1, \dots, y_T, c) dc$$

... like  $f(y) = \int f(x, y) dx$

$$= \int_{-b}^b \left[ \prod_{t=1}^T f(y_t | c) \right] f(c) dc$$

... like  $f(x, y) = f(y | x) f(x)$

this is not trivial; Butler & Moffitt (1982)  
 provides an approximation of the above  
 integration.

Note "Integrating out" method for RE models  
 (Butler & Moffitt's Gaussian-Hermite  
quadrature method) Greene, p 690

(22)

$$\varepsilon_{it} = v_{it} + \alpha_i$$

$$\text{Var}(\varepsilon_{it}) = \sigma_v^2 + \sigma_\alpha^2 \quad (\text{assume } \sigma_v^2 = 1)$$

$$L_i = P(y_{i1}, \dots, y_{iT_i}) = \int_{L_{i1}}^{u_{i1}} \dots \int_{L_{iT_i}}^{u_{iT_i}} f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i}) d\varepsilon_{i1} \dots d\varepsilon_{iT_i}$$

where  $L_i \rightarrow U_i$  implies  $\begin{cases} (-\infty, -x_i/\beta) & \text{if } y_i = 0 \\ (-x_i/\beta, \infty) & \text{if } y_i = 1 \end{cases}$

and

$$f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i}, \alpha_i)$$

$$= f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i} | \alpha_i) f(\alpha_i)$$

$$f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i}) = \int_{-\infty}^{\infty} f(\varepsilon_{i1}, \dots, \varepsilon_{iT_i} | \alpha_i) f(\alpha_i) d\alpha_i$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} | \alpha_i) f(\alpha_i) d\alpha_i$$

thus,

$$L_i = \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \left( \int_{L_{it}}^{U_{it}} f(\varepsilon_{it} | \alpha_i) \right) \right] f(\alpha_i) d\alpha_i \quad (*)$$

$$= \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} | x_{it}(\beta + \alpha_i)) \right] \frac{1}{\sigma_\alpha \sqrt{2\pi}} e^{-\frac{\alpha_i^2}{2\sigma_\alpha^2}} d\alpha_i$$

$$\text{Let } \frac{\alpha_i^2}{2\sigma_\alpha^2} = r_i^2 \Rightarrow \begin{cases} \alpha_i = (\sigma_\alpha \sqrt{2}) r_i \\ d\alpha_i = (\sigma_\alpha \sqrt{2}) dr_i \end{cases}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} \left[ \prod_{t=1}^{T_i} \Phi(y_{it} + (x_{it}(\beta + (\sigma_\alpha \sqrt{2}) r_i)) \right] dr_i$$

$\Rightarrow$  transform the integral problem into  
 a summation problem via Gaussian quadrature.



\* (3) FE logit (Common) see WI, p490-492

Find the joint dist of  $(y_{it} \dots y_{iT})$ , conditional on  $c_i$  and  $\eta_i = \sum_{t=1}^T y_{it}$

$$P(y_{it}=1) = \frac{e^{\alpha_i + \beta'X_{it}}}{1 + e^{\alpha_i + \beta'X_{it}}}$$

then the conditional ML does not depend on  $c_i$ . ( $c_i$  cancelled out.)

⇒ CMLE (Chamberlain's method)

Not possible to sweep out  $\alpha_i$  by taking differences or deviation from group means

- We can obtain log-odd expressions, but cannot obtain partial effects since  $c_i$  is not estimated.

- Requires conditional independence, thus FE logit cannot allow for dynamic models.

(4) RE logit (not popular, but can be done as in RE probit.)

- no feasible simple form, but can be done as in RE probit.

Remark: Conditional likelihood function (FE logit) Chamberlain's method. (25)

A FE logit model is

$$P(y_{it}=1 | x_{it}) = \frac{e^{\alpha_i + x_{it}'\beta}}{1 + e^{\alpha_i + x_{it}'\beta}}$$

unconditional likelihood

$$L = \prod_i \prod_t (F_{it})^{y_{it}} (1 - F_{it})^{1 - y_{it}}$$

conditional likelihood

$$L^c = \prod_{i=1}^n P(y_{i1}=y_{i1}, \dots, y_{iT}=y_{iT} | \sum_{t=1}^T y_{it})$$

(∵ this is free of  $\alpha_i$ )

$$= \prod_{i=1}^n \frac{\exp(\sum_{t=1}^T y_{it} x_{it}'\beta)}{\sum_{\text{all set}} \exp(\sum_{t=1}^T d_{it} x_{it}'\beta)}$$

where the denominator is summed over

the set of all  $T_i C_{Si}$  sequences

with  $S_i = \sum_{t=1}^T y_{it}$ .

ex)  $T_i = 2$

$$\text{i) } y_{i1}=0, y_{i2}=0 \Rightarrow P(0,0 | \text{sum}=0) = 1$$

$$\text{ii) } y_{i1}=1, y_{i2}=1 \Rightarrow P(1,1 | \text{sum}=2) = 1$$

$$\text{iii) } y_{i1}=0, y_{i2}=1 \Rightarrow P(0,1 | \text{sum}=1)$$

$$= \frac{P(0,1)}{P(0,1) + P(1,0)}$$

this is given as

$$\frac{1}{1 + \exp(\alpha_i + x_{i1}'\beta)} \cdot \frac{\exp(\alpha_i + x_{i2}'\beta)}{1 + \exp(\alpha_i + x_{i2}'\beta)} = \frac{e^{x_{i2}'\beta}}{e^{x_{i1}'\beta} + e^{x_{i2}'\beta}}$$

$$\text{(top) } + \frac{\exp(\alpha_i + x_{i1}'\beta)}{1 + \exp(\alpha_i + x_{i1}'\beta)} \cdot \frac{1}{1 + \exp(\alpha_i + x_{i2}'\beta)}$$

... free of  $\alpha_i$

thus, by conditioning on the sum of two observations, we can remove  $\alpha_i$ .

Note i) One can test CML vs ML.

Hausman test

$$(\hat{\beta}_{CML} - \hat{\beta}_{ML})' [Var(CML) - Var(ML)]^{-1} (\hat{\beta}_{CML} - \hat{\beta}_{ML})$$

∴) We cannot estimate the marginal effects on the predicted probabilities unless we can plug  $\alpha_i$  (which was cancelled out!)

### Dynamic choice Models (Panel)

$$P(y_{it} = 1 | y_{i,t-1}, \dots, y_{i0}, x_{it}, \alpha_i) \\ = G(x_{it}'\beta + \rho y_{i,t-1} + \alpha_i)$$

$H_0: \rho = 0$  (state dependence is absent)

- FE cannot be used
- RE version is possible (integrating out  $\alpha_i$ )

$$P(y_{it} = 1 | \dots) = G(\underbrace{\phi + x_{it}'\beta + \rho y_{i,t-1} + \psi_0 y_{i0} + x_{i0}'\gamma}_{\text{added}} + \underbrace{\alpha_i + e_{it}}_{\substack{\downarrow \\ \text{integrate out}}})$$

Exercise Wooldridge 15.19 (p. 515)

(a) ~ (e).

Logit

logit fits maximum-likelihood logit models. depvar==0 indicates a negative outcome; depvar!=0 & depvar!=. (typically depvar==1) indicates a positive outcome.

Also see help logistic; many users prefer the logistic command to logit. Results are the same regardless of which you use, but logistic reports odds ratios rather than coefficients by default and some people simply prefer the name logistic to logit. In Stata, both are the maximum-likelihood estimator. A number of commands are documented under help logistic that may be run after logit or logistic.

If estimating on grouped data, see help glogit.

Logistic

logistic fits maximum-likelihood logistic regression models. depvar==0 indicates a negative outcome; depvar!=0 & depvar!=. (typically depvar==1) indicates a positive outcome.

logistic reports odds ratios. You can type "logit" after logistic estimation to obtain the coefficients. In addition, there are a number of other commands that can be used after logistic to explore the nature of the fit:

- help lfit Perform goodness-of-fit tests
- help lstat Report summary statistics including classification table
- help lroc Graph the ROC curve
- help lsens Graph sensitivity and specificity vs. P cutoff

probit

probit fits maximum-likelihood probit models.

dprobit also fits maximum-likelihood probit models. Rather than reporting coefficients, dprobit reports the change in the probability for an infinitesimal change in each independent, continuous variable and, by default, the discrete change in the probability for dummy variables. probit may be typed without arguments after dprobit estimation to see the model in coefficient form.

depvar==0 indicates a negative outcome; depvar!=0 & depvar!=. (typically depvar==1) indicates a positive outcome.

If you are estimating on grouped data, see help glogit (sic).

**blogit, bprobit, glogit and gprobit**

blogit and bprobit produce maximum-likelihood logit and probit estimates on grouped ("blocked") data; glogit and gprobit produce weighted least-squares estimates. In the syntax diagrams, pos\_var and pop\_var refer to variables containing the total number of positive responses and the total population.

See help logit and help probit for obtaining maximum-likelihood estimates on ungrouped (individual or micro) data.

**xtprobit**

xtprobit fits random-effects (re) and population-averaged (pa) probit models for cross-sectional time-series datasets.

For the random-effects model, the likelihood (for an independent unit  $i$ ) is expressed as an integral which is computed using Gauss-Hermite quadrature. After fitting your final model, you may want to run quadchk to check the numerical soundness of the Gauss-Hermite quadrature approximation; see help quadchk and [XT] quadchk for details.

By default, the population-averaged model is an equal-correlation model; that is, xtprobit, pa assumes corr(exchangeable). See help xtgee for details on how to fit other population-averaged models.

**Xtlogit**

xtlogit fits a fixed-effects (fe), an random-effects (re), or a population-averaged (pa) logit model for cross-sectional time-series datasets. Whenever we refer to a fixed-effects model, we mean the conditional fixed-effects model.

For the random-effects model, the likelihood (for an independent unit  $i$ ) is expressed as an integral which is computed using Gauss-Hermite quadrature. After fitting your final model, you may want to run quadchk to check the numerical soundness of the Gauss-Hermite quadrature approximation; see help quadchk and [XT] quadchk for details.

By default, the population-averaged model is an equal-correlation model; that is, xtlogit, pa assumes corr(exchangeable). See help xtgee for details on how to fit other population-averaged models.

# STATA<sup>®</sup> Statistical Software for Professionals

>> Home >> Resources & Support >> FAQs >> Advantages  
of the robust variance estimator

Products Resources & Support Company Search

## What are the advantages of using the robust variance estimator over the standard maximum-likelihood variance estimator in logistic regression?

Title Advantages of the robust variance estimator  
Author Bill Sribney, StataCorp  
Date January 1998

I once overheard a famous statistician say that the robust variance estimator for (unclustered) logistic regression is stupid. His reason was that if the outcome variable is binary then it's got to be a Bernoulli distribution. It's not like linear regression with data that stands a good chance of being nonnormal and heteroscedastic.

Well, it's not as simple as this; there's a bunch of subtle nuisances here. Let me lay them out here. I'm sure that the famous statistician is aware of them, but they don't necessarily lead to his conclusion.

It's true that it's got to be a Bernoulli distribution. That is, if  $Y_i$  is a random variable for the outcome of the  $i$ -th unit, then

$$P(Y_{i=1}) = p_i$$

or equivalently,

$$E(Y_i) = p_i$$

This *has* to be true. Note how I indexed the RHS by  $i$ . The term  $p_i$  is dependent on  $i$ . It's certainly not true in general that  $P(Y_{i=1}) = p$ , where  $p$  is a constant independent of  $i$ .

In logistic regression, we model  $p_i$  with a likelihood that assumes

$$\text{logit}(p_i) = x_i * b$$

So these are our assumptions:

1. That the **logit** link function is correct.
2. That **logit**( $p_i$ ) =  $x_i * b$ ; i.e., that the relation is linear and all necessary predictors are in the model; i.e., that the model is correctly specified.
3. That the  $i=1, \dots, N$  observations are independent.

FAQs  
What's new  
Statistics  
Data management  
Graphics  
Programming  
Resources  
Internet connectivity  
Stata for Windows  
Stata for Macintosh  
Stata for Linux  
Technical support

Resources  
Support  
FAQs  
Technical support  
NetCourses  
Short courses  
Users groups  
Statalist  
Links  
Software  
Software updates  
Customized  
Manuals  
Supplements  
Stata Journal  
STB  
Stata News  
Plugins  
Stata corporate

Site overview  
Products  
Resources  
Company  
Site index

The robust variance estimator is robust to the assumptions (1) and (2). It does require (3), but you can specify clusters and just assume independence of the clusters if you wish.

The MLE is also quite robust to (1) being wrong. If the link function is really probit and you estimate a logit, everything's almost always fine.

Now if (2) is wrong, strictly speaking, you are in trouble with the interpretation of the point estimates of your model, never mind the variance estimates. Imagine that  $\text{logit}(p_i)$  is truly quadratic in  $x_i$ , but you fit it linear in  $x_i$ . What's the interpretation of the coefficient? It's the best fit of a straight line to something that's not straight! It's got some meaning, but it's somewhat problematic.

Well, the robust variance estimator will do a good job of giving you variance estimates and confidence intervals for this problematic case of a misspecified model. That is, if one imagines resampling the data and each time fitting the same misspecified model, then you get good coverage probabilities w.r.t. the "true" population parameters of the misspecified model, i.e., the besting fitting straight line in the population to something that's not straight.

On the other hand, if you have confidence that your model is not misspecified, then the ML variance estimator is theoretically more efficient.

## Advice

In summary, my personal advice (and I have respect for conflicting opinions) is

- I never worry about whether (1) is true. I assume the **logit** link is OK.
- If I think that the model is reasonably specified, I use the ML variance estimator for logistic regression.
- Only if I have good reason to believe that the model is poorly specified would I use the robust variance estimator. That is, if the model fails goodness-of-fit tests, etc. Sometimes one just has to live with missing predictors and badly fitting models because data was only collected for a few predictors. In this case, I'd use the robust variance estimator.

And, obviously, I'd use the robust variance estimator if I had clustered data.

This is in contrast to the advice I'd give for linear regression when I'd say *always* use the robust variance estimator.

## The robust variance estimator is only approximate for ML models

Note that there are also other theoretical reasons to be keener on the robust variance estimator for linear regression than for general ML models. The robust variance estimator uses a one-term Taylor series approximation. In linear regression, the coefficient estimates  $\mathbf{b}$  are a linear function of  $\mathbf{y}$ ; namely,

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

Thus, the one-term Taylor series is exact and not an approximation.

For logistic regression and other MLEs, the ignored higher-order terms in the Taylor series are nonzero. So it's truly an approximation in these cases.

One can come up with a robust variance estimator that uses a second order correction. It's been worked out for logistic regression, and it will likely be implemented in Stata at some point in the future.

## Follow-up question

What are the "true" population parameters for which the robust variance estimator gives good coverage properties?

## Linear regression model

First let me assume a linear regression model, then later I'll discuss MLEs.

Consider the entire population from which the sample is drawn. Let  $\mathbf{X}$  be the matrix of independent variables and  $\mathbf{Y}$  the vector of dependent variables for the *entire population*. Then consider

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

The parameter  $\mathbf{B}$  is the coefficient vector for the linear model for the entire population.  $\mathbf{Y}$  may be linear in  $\mathbf{X}$  or it may not.  $\mathbf{B}$  is simply the best least-squares coefficients for the entire population.  $\mathbf{B}$  is what I was referring to when I said "the 'true' population parameters" in my above explanation.

The parameter  $\mathbf{B}$  is what the robust variance estimator considers you to be estimating. The sample estimate is

$$\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y}$$

where  $\mathbf{x}$  is the matrix of the sample values of the independent variables and  $\mathbf{y}$  is the vector of sample dependent variables. If there were sampling weights, the above equation would have weights added in the appropriate places.

Anyhow,  $\mathbf{b}$  is an estimate of  $\mathbf{B}$ . The robust variance estimator estimates  $\mathbf{V}(\mathbf{b})$  such that nominal (1-alpha) confidence intervals constructed from it have  $\mathbf{B}$  in the interval about (1-alpha) of the time if one was to repeatedly resample from this population.

If  $\mathbf{Y}$  is not linear in  $\mathbf{X}$  due to incorrect functional form or missing predictors, then the interpretation of  $\mathbf{B}$  is problematic.  $\mathbf{B}$  can be considered to be the best least-squares linear fit for this set of predictors.  $\mathbf{b}$  and  $\mathbf{V}(\mathbf{b})$  are "robust to misspecification" in that  $\mathbf{b}$  estimates  $\mathbf{B}$  and that  $\mathbf{V}(\mathbf{b})$  is a valid estimate of the variance of  $\mathbf{b}$  even though misspecification is present.

Note that this theory requires no assumptions about the distribution of  $\mathbf{Y}$ .

Contrast this to the case of OLS estimates, which do not give valid variance estimates, in general, under misspecification, and which do require distributional assumptions on  $\mathbf{Y}$  — i.e., normality and homoscedasticity.

## ML Models

For ML models, consider  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$ , an arbitrary likelihood function with data  $\mathbf{Y}, \mathbf{X}$  for the entire population. Let  $\mathbf{B}^*$  give the maximum of  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$ . The sample estimate  $\mathbf{b}^*$  is the maximum of  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$ . If there are weights, we add weights to the likelihood function so that  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$  estimates  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$ . Because  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$  estimates  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$ ,  $\mathbf{b}^*$  estimates  $\mathbf{B}^*$ . The robust variance estimator produces correct variance estimates  $\mathbf{V}(\mathbf{b}^*)$  for  $\mathbf{b}^*$  in the same sense discussed above: nominal (1-alpha) confidence intervals constructed from it have  $\mathbf{B}^*$  in the interval about (1-alpha) of the time if one was to repeatedly resample from this population.

Note that  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$  is not necessarily the true likelihood for the population; i.e., it is not necessarily the correct distribution of  $\mathbf{Y}|\mathbf{X}$ . The theory doesn't require it; it can be any function.

Standard MLE theory, on the other hand, requires  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$  to be the true distribution function for the sample.

Note that when there are sampling weights or clustering,  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$  is in no sense a valid likelihood; it's clearly not the distribution of the sample when there are weights or cluster sampling.  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$  merely has to estimate the arbitrary  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$  for our theory to hold. This is why the survey theorists call  $L(\mathbf{b}; \mathbf{y}, \mathbf{x})$  a pseudo-likelihood, and it's also why you can't do standard likelihood ratio tests with it.

However, if  $L(\mathbf{B}; \mathbf{Y}, \mathbf{X})$  is not close to the true distribution, its interpretation is problematic, just as in the case of a misspecified linear regression.

---

© Copyright 2004 StataCorp LP | [Terms of use](#) | [Privacy](#) | [Contact us](#) | [What's new](#) | [Site index](#)