

# Intermediate Social Statistics Hilary 2011 Lecture 4: Bi-variate Probit and Ordered Probit

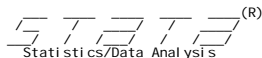
Raymond Duch

Nuffield College

February 8, 2011

# Recall Logit Model from Lecture 3

Monday February 8 14:57:51 2010 Page 1



Logistic regression

Number of obs = 785  
 LR chi2(7) = 128.17  
 Prob > chi2 = 0.0000  
 Pseudo R2 = 0.1262

Log Likelihood = **-443.69615**

incumvote	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
retnat	-.7038763	.1034497	-6.80	0.000	-.906634 -.5011186
class	-.3854525	.0861783	-4.47	0.000	-.5543589 -.216546
union	.5060247	.1915331	2.64	0.008	.1306267 .8814226
southwest	-.933915	.3541623	-2.64	0.008	-1.62806 -.2397697
urban	.3036703	.1060421	2.86	0.004	.0958315 .5115091
lrsel f	-.1339321	.0369293	-3.63	0.000	-.2063122 -.0615519
own	-.5231139	.1897608	-2.76	0.006	-.8950382 -.1511896
_cons	1.97143	.4262088	4.63	0.000	1.136076 2.806784

# Deriving The Likelihood Function for Probit

- We are modeling actually a latent quantity that gives rise to the observed discrete outcomes.
- Think of this underlying latent quantity as a "probability" or a "random utility".
- We model the unobserved net utilities  $y^*$  of the choices  $y$  via the model:

$$y^* = \mathbf{x}_i\beta + \varepsilon_i \quad (1)$$

## Deriving The Likelihood Function

- With either independent and identically distributed (iid)  $N(0,1)$  for probit.
- We don't observe the net utility of the choice , just whether it was made or not.
- So we observe
  - ▶ whether a person voted for Labour or for the Opposition;
  - ▶ whether an individual made a campaign contribution or did not;
  - ▶ whether a country went to war or did not;
  - ▶ whether someone died or did not.
- We posit that

$$\begin{aligned}y_i &= 0 \text{ if } y_i^* \leq 0 \\ &= 1 \text{ if } y_i^* > 0\end{aligned}$$

# The Probit Estimator

Some simple algebra gives us our estimator.

$$\begin{aligned} \text{Prob}(y_i = 1 | \mathbf{x}_i) &= \text{Prob}(y_i^* > 0) \\ &= \text{Prob}(\mathbf{x}_i\beta + \varepsilon_i > 0) \\ &= \text{Prob}(\varepsilon_i > -\mathbf{x}_i\beta) \\ &= \text{Prob}\left(\frac{\varepsilon_i}{\sigma} > \frac{-\mathbf{x}_i\beta}{\sigma}\right) \end{aligned} \tag{2}$$

# The Probit Estimator

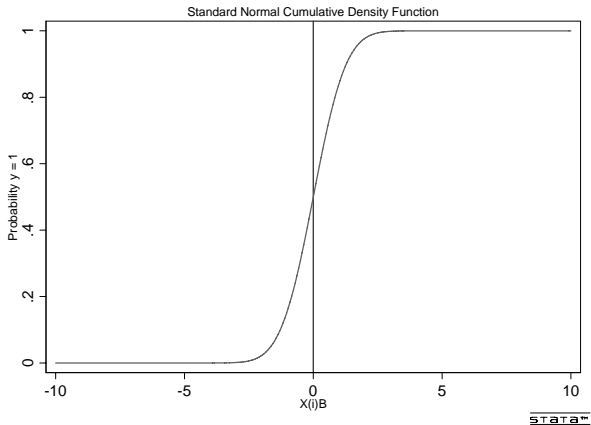
Because the error has a normal distribution this becomes

$$Prob(y_i = 1|\mathbf{x}_i) = 1 - \Phi\left(\frac{-\mathbf{x}_i\beta}{\sigma}\right) = \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right) \quad (3)$$

Similarly,

$$Prob(y_i = 0|\mathbf{x}_i) = 1 - \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right) \quad (4)$$

This is simply the standard normal cumulative density function and as you can see it closely resembles the logistic CDF that we examined last week.



# Predictions in Logit and Probit

```
. probit incumvote retnat class union southwest urban lrself own
```

```
Iteration 0:  log likelihood = -507.77976
Iteration 1:  log likelihood = -446.40068
Iteration 2:  log likelihood = -446.10547
Iteration 3:  log likelihood = -446.10541
Iteration 4:  log likelihood = -446.10541
```

```
Probit regression                Number of obs   =       785
                                LR chi2(7)        =       123.35
                                Prob > chi2         =       0.0000
                                Pseudo R2          =       0.1215

Log likelihood = -446.10541
```

```
-----+-----
incumvote |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
    retnat |   -.4175061   .0608463    -6.86   0.000   -.5367627   -.2982494
     class |   -.2294566   .0509206    -4.51   0.000   -.3292591   -.1296541
     union |    .2947863   .1154144     2.55   0.011    .0685783    .5209943
southwest |   -.526983    .1965508    -2.68   0.007   -.9122154   -.1417506
     urban |    .2153416   .1079612     1.99   0.046    .0037415    .4269417
     lrself |   -.0800539   .0215398    -3.72   0.000   -.122271    -.0378367
         own |   -.3303976   .1131544    -2.92   0.004   -.5521762   -.1086189
         _cons |    1.39502    .2285129     6.10   0.000    .9471429    1.842897
-----+-----
```

# Uncertainty in Probit Predictions

$$\text{Prob}(y_i = 1 | \mathbf{x}_i) = \Phi\left(\frac{\mathbf{x}_i \hat{\beta}}{\sigma}\right) \quad (5)$$

- note that  $\hat{\beta}$  is an estimate
- there is uncertainty associated with this estimate that is captured by its covariance matrix  $\hat{\Sigma}$
- how much confidence can we have in the particular point prediction we have generated?
- simulation methods are typically currently employed to estimate a confidence interval for  $p$

# Uncertainty in Probit Predictions

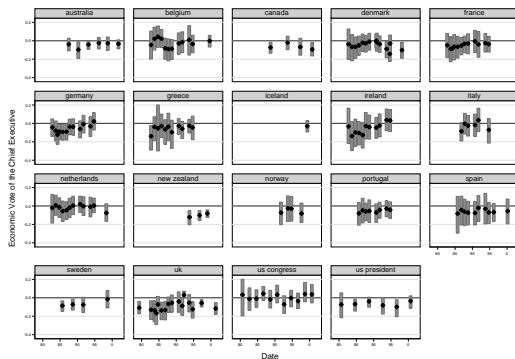
$$\text{Prob}(y_i = 1 | \mathbf{x}_i) = \Phi\left(\frac{\mathbf{x}_i \hat{\beta}}{\sigma}\right) \quad (6)$$

- note that  $\hat{\beta}$  is an estimate
- there is uncertainty associated with this estimate that is captured by its covariance matrix  $\hat{\Sigma}$
- how much confidence can we have in the particular point prediction we have generated?
- simulation methods are typically currently employed to estimate a confidence interval for  $\hat{p}$

## Steps for Estimating a Confidence Interval for $p_i$

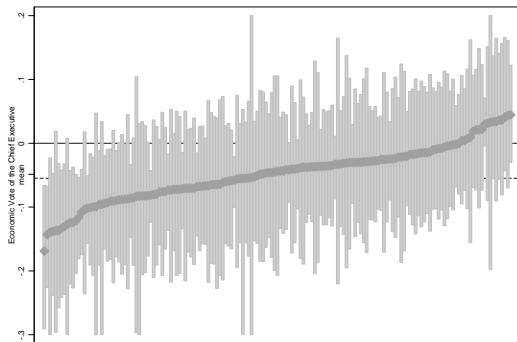
- Estimate  $\hat{\beta}$  and  $\hat{\Sigma}$
- Repeat the following steps  $S$  times (where  $S = 5000$ , for example), indexing each iteration by  $l$ :
  - ▶ Draw a vector  $\hat{\beta}$  from a normal distribution with mean vector  $\tilde{\beta}$  and covariance matrix  $\hat{\Sigma}$
  - ▶ Calculate  $\hat{p}_i^l = \Phi\left(\frac{\mathbf{x}_i \hat{\beta}}{\sigma}\right)$
- Use the percentiles of the histogram of  $\hat{p}_i^l$ ,  $l = 1, \dots, S$ , to form a confidence interval for  $\hat{p}_i$

# Further Illustrations: The Economic Vote



STATA™

## Further Illustrations: The Economic Vote



Confidence bounds greater than .2 and less than -.3 truncated for display

STATA™

# Example: CCAP German Election Study

```
gen CDU_vote=vote_2
recode CDU_vote 2=1 1=0 3/6=0
preserve
keep if savvy_type==1
estsimp probit CDU_vote LR_Self_1 retnat_2 educate female religiosity
setx mean
simqi, fd(prval(1)) changex(retnat_2 3 4)
drop b*

restore

preserve

keep if savvy_type==2
estsimp probit CDU_vote LR_Self_1 retnat_2 educate female religiosity
setx mean
simqi, fd(prval(1)) changex(retnat_2 3 4)
drop b*

restore
```

# Example: CCAP German Election Study

```
. tabulate vote_2
```

vote_2	Freq.	Percent	Cum.
SPD	447	19.81	19.81
CDU/CSU	549	24.32	44.13
FDP	472	20.91	65.04
Green	314	13.91	78.95
Linke	475	21.05	100.00
Total	2,257	100.00	

# Example: Low Coalition Reasoning

```
. keep if savvy_type==1
(3790 observations deleted)
```

```
. estsimp probit CDU_vote LR_Self_1 retnat_2 educate female religiosity
```

```
Iteration 0: log likelihood = -430.51297
Iteration 1: log likelihood = -362.53155
Iteration 2: log likelihood = -360.42229
Iteration 3: log likelihood = -360.41097
Iteration 4: log likelihood = -360.41096
```

Probit regression

```
Number of obs = 771
LR chi2(5) = 140.20
Prob > chi2 = 0.0000
Pseudo R2 = 0.1628
```

Log likelihood = -360.41096

CDU_vote	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
LR_Self_1	.2175642	.0278096	7.82	0.000	.1630584	.27207
retnat_2	-.2285589	.0586011	-3.90	0.000	-.3434149	-.1137028
educate	.0170922	.0385282	0.44	0.657	-.0584217	.0926061
female	.1554171	.1109873	1.40	0.161	-.0621141	.3729483
religiosity	-.2208956	.05774	-3.83	0.000	-.3340638	-.1077274
_cons	-.5897386	.3851742	-1.53	0.126	-1.344666	.1651889

# Example: Moderate Coalition Reasoning

```
. keep if savvy_type==2
(4926 observations deleted)

. estsimp probit CDU_vote LR_Self_1 retnat_2 educate female religiosity
```

```
Iteration 0:  log likelihood = -246.25268
Iteration 1:  log likelihood = -187.97437
Iteration 2:  log likelihood = -183.90439
Iteration 3:  log likelihood = -183.80212
Iteration 4:  log likelihood = -183.80203
```

```
Probit regression                Number of obs   =          477
                                LR chi2(5)         =          124.90
                                Prob > chi2         =          0.0000
Log likelihood = -183.80203      Pseudo R2       =          0.2536
```

```
-----+-----
      CDU_vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      LR_Self_1 |   .3546698   .0399453     8.88   0.000     .2763785   .4329611
      retnat_2 |  -.2832845   .0895562    -3.16   0.002    - .4588114  -.1077577
      educate   |  -.0042755   .0498819    -0.09   0.932    - .1020423  .0934912
      female    |   .2251718   .1517415     1.48   0.138    - .0722361  .5225797
religiosity    |   .0278492   .0862295     0.32   0.747    - .1411576  .196856
      _cons    |  -2.04437    .5363551    -3.81   0.000    -3.095607  -.9931335
-----+-----
```

# Example: High Coalition Reasoning

```
. keep if savvy_type==3
(5878 observations deleted)

. estsimp probit CDU_vote LR_Self_1 retnat_2 educate female religiosity
```

```
Iteration 0: log likelihood = -76.377366
Iteration 1: log likelihood = -53.583866
Iteration 2: log likelihood = -51.794099
Iteration 3: log likelihood = -51.720703
Iteration 4: log likelihood = -51.720515
```

```
Probit regression                               Number of obs =      129
                                                LR chi2(5)       =      49.31
                                                Prob > chi2      =      0.0000
Log likelihood = -51.720515                    Pseudo R2       =      0.3228
```

```
-----+-----
      CDU_vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
  LR_Self_1 |   .4157707   .0827566     5.02   0.000   .2535708   .5779707
  retnat_2 |   .0765187   .1945564     0.39   0.694   -.3048049   .4578423
  educate |   .0543228   .1096873     0.50   0.620   -.1606605   .269306
  female |   .5957627   .3364713     1.77   0.077   -.063709   1.255234
religiosity |  -.3214772   .155906     -2.06   0.039   -.6270474   -.015907
   _cons |  -2.821291   1.209299     -2.33   0.020   -5.191474   -.4511092
-----+-----
```

# Example: CCAP German Election Study

```
keep if savvy_type==1
(3790 observations deleted)
```

```
Simulating main parameters. Please wait....
% of simulations completed: 16% 33% 50% 66% 83% 100%
```

```
Number of simulations : 1000
Names of new variables : b1 b2 b3 b4 b5 b6
```

```
. setx mean
```

```
. simqi, fd(prval(1)) changex(retnat_2 3 4)
```

```
First Difference: retnat_2 3 4
```

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
dPr(CDU_vote = 1)	-.0681765	.0185055	-.1046059	-.0325488

# Example: CCAP German Election Study

```
keep if savvy_type==2
(4926 observations deleted)
```

```
Simulating main parameters. Please wait....
% of simulations completed: 16% 33% 50% 66% 83% 100%
```

```
Number of simulations : 1000
Names of new variables : b1 b2 b3 b4 b5 b6
```

```
. setx mean
```

```
. simqi, fd(prval(1)) changex(retnat_2 3 4)
```

```
First Difference: retnat_2 3 4
```

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
dPr(CDU_vote = 1)	-.0702687	.0237353	-.1194669	-.0270208

# Example: CCAP German Election Study

```
. keep if savvy_type==3
(5878 observations deleted)
```

Simulating main parameters. Please wait....

Note: Clarify is expanding your dataset from 327 observations to 1000 observations in order to accommodate the simulations. This will append missing values to the bottom of your original dataset.

% of simulations completed: 16% 33% 50% 66% 83% 100%

```
Number of simulations : 1000
Names of new variables : b1 b2 b3 b4 b5 b6
```

```
. setx mean
```

```
. simqi, fd(prval(1)) changex(retnat_2 3 4)
```

First Difference: retnat\_2 3 4

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
-----+-----				
dPr(CDU_vote = 1)	.0149392	.049925	-.0916377	.1004407

```
. drop b*
```

# Fitted Probabilities and Percent Correctly Predicted (PCP)

PCP is typically defined by the following procedure:

- Estimate  $\hat{\beta}$  and for each observation  $i$ , calculate  $\hat{p}_i$  using the equation above
- For each observation with  $\hat{p}_i > 0.5$  set  $\hat{y}_i = 1$ , otherwise set  $\hat{p}_i > 0.5$  set  $\hat{y}_i = 0$
- Call each observation with  $\hat{y}_i = y_i$  a correct prediction.
- PCP is defined as the percentage of observations that are correctly predicted.

## Fitted Probabilities and Percent Correctly Predicted (PCP)

```
predict yhat_logit

gen predict_vote=0
replace predict_vote=1 if yhat_logit>.499999

gen correct_0=0
gen correct_1=0
replace correct_0=1 if (predict_vote==0 & incumvote==0)
replace correct_1=1 if (predict_vote==1 & incumvote==1)
summarize correct_1 correct_0
```

# Predictions in Logit and Probit

	yhat_probitt	yhat_logit	incumvote	predict_vote
1.	.3928562	.3869703	1	0
2.	.560674	.5643634	.	1
3.	.3432808	.3406143	1	0
4.	.5578907	.5597182	.	1
5.	.560674	.5643634	.	1
6.	.4942218	.4918689	0	0
7.	.5892738	.5929783	.	1
8.	.7399474	.7458074	0	1
9.	.4524617	.4533052	0	0
10.	.3928562	.3869703	0	0
11.	.3445356	.3425307	1	0
12.	.2229143	.2179674	0	0
13.	.7816748	.7864491	0	1
14.	.4901605	.4919069	.	0
15.	.3007564	.2955182	.	0
16.	.	.	1	1
17.	.2117439	.2085973	.	0
18.	.	.	.	1
19.	.2395462	.2338773	0	0
20.	.4267031	.4243549	.	0

# Predictions in Logit and Probit

```
. summarize correct_1 correct_0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
correct_1	1500	.114	.3179173	0	1
correct_0	1500	.2993333	.4581188	0	1

PCP versus ePCP

$$ePCP = \frac{1}{N} \left( \sum_{y_i=1}^N p_i + \sum_{y_i=0}^N (1 - p_i) \right) \quad (7)$$

# Ordered Probit

We assume that individual opinion is continuous, but unobserved. So just as we did with probit, we treat this as a latent regression model, where:

$$y^* = \mathbf{x}_i\beta + \varepsilon_i \quad (8)$$

# Ordered Probit

- Again the  $y^*$  is unobserved, but we do observe discrete outcomes.
- Here I work out the example of dependent variable with four ordered values
- It has four discrete ordered outcomes: 1, 2, 3 and 4.

## Ordered Probit

$$\begin{aligned} y_i &= 1 \text{ if } y_i^* < \gamma_1 \\ &= 2 \text{ if } \gamma_1 \geq y_i^* < \gamma_2 \\ &= 3 \text{ if } \gamma_2 \geq y_i^* < \gamma_3 \\ &= 4 \text{ if } \gamma_3 \leq y_i^* \end{aligned} \tag{9}$$

# Ordered Probit

So, thinking of this as a probit, we have:

$$\begin{aligned} \text{Prob}(y_i = 1 | \mathbf{x}_i) &= \text{Prob}(y_i^* < \gamma_1) \\ &= \text{Prob}(\mathbf{x}_i\beta + \varepsilon_i < \gamma_1) \\ &= \text{Prob}(\varepsilon_i < \gamma_1 - \mathbf{x}_i\beta) \\ &= \Phi(\gamma_1 - \mathbf{x}_i\beta) \end{aligned} \tag{10}$$

## Ordered Probit

$$\begin{aligned} \text{Prob}(y_i = 2 | \mathbf{x}_i) &= \text{Prob}(\gamma_1 \leq y_i^* < \gamma_2) && (11) \\ &= \text{Prob}(\gamma_1 \leq \mathbf{x}_i \beta + \varepsilon_i < \gamma_2) \\ &= \text{Prob}(\varepsilon_i < \gamma_2 - \mathbf{x}_i \beta) - \text{Prob}(\varepsilon_i < \gamma_1 - \mathbf{x}_i \beta) \\ &= \Phi(\gamma_2 - \mathbf{x}_i \beta) - \Phi(\gamma_1 - \mathbf{x}_i \beta) \end{aligned}$$

## Ordered Probit

$$\begin{aligned} \text{Prob}(y_i = 3 | \mathbf{x}_i) &= \text{Prob}(\gamma_2 \leq y_i^* < \gamma_3) && (12) \\ &= \text{Prob}(\gamma_2 \leq \mathbf{x}_i\beta + \varepsilon_i < \gamma_3) \\ &= \text{Prob}(\varepsilon_i < \gamma_3 - \mathbf{x}_i\beta) - \text{Prob}(\varepsilon_i < \gamma_2 - \mathbf{x}_i\beta) \\ &= \Phi(\gamma_3 - \mathbf{x}_i\beta) - \Phi(\gamma_2 - \mathbf{x}_i\beta) \end{aligned}$$

## Ordered Probit

$$\begin{aligned} \text{Prob}(y_i = 4 | \mathbf{x}_i) &= \text{Prob}(y_i^* \geq \gamma_3) \\ &= \text{Prob}(\mathbf{x}_i \beta + \varepsilon_i \geq \gamma_3) \\ &= \text{Prob}(\varepsilon_i \geq \gamma_3 - \mathbf{x}_i \beta) \\ &= 1 - \Phi(\gamma_3 - \mathbf{x}_i \beta) \end{aligned} \tag{13}$$

# Predictions in Logit and Probit

```
. tabulate savvy_type
```

savvy_type	Freq.	Percent	Cum.
1	2,415	60.06	60.06
2	1,279	31.81	91.87
3	327	8.13	100.00
Total	4,021	100.00	

# Predictions in Logit and Probit

```
ologit savvy_type coalKnow_w2 anticipation_w3 interest_w4 educate income_w1 female v4_6361
```

```
Iteration 0: log likelihood = -487.82214
Iteration 1: log likelihood = -465.5042
Iteration 2: log likelihood = -465.19106
Iteration 3: log likelihood = -465.191
Iteration 4: log likelihood = -465.191
```

Ordered logistic regression

```
Number of obs = 555
LR chi2(7) = 45.26
Prob > chi2 = 0.0000
Pseudo R2 = 0.0464
```

Log likelihood = -465.191

savvy_type	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
coalKnow_w2	.2863765	.1254063	2.28	0.022	.0405848	.5321682
anticipati~3	.0461033	.1849815	0.25	0.803	-.3164537	.4086603
interest_w4	.2914186	.114078	2.55	0.011	.0678299	.5150073
educate	.0536963	.063295	0.85	0.396	-.0703596	.1777522
income_w1	.0041216	.0427418	0.10	0.923	-.0796507	.0878939
female	-.2434122	.193061	-1.26	0.207	-.6218048	.1349805
v4_6361	.1311079	.0532742	2.46	0.014	.0266923	.2355234
/cut1	3.027085	.5627353			1.924144	4.130026
/cut2	4.981767	.5874418			3.830402	6.133132

# Ordered Probit Predictions in Clarify

Simulating main parameters. Please wait...

% of simulations completed: 11% 22% 33% 44% 55% 66% 77% 88% 100%

Number of simulations : 1000

Names of new variables : b1 b2 b3 b4 b5 b6 b7 b8 b9

.  
. setx mean

. simqi, pr

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
Pr(savvy_~e=1)	.6204754	.0214185	.5777605	.6601951
Pr(savvy_~e=2)	.2994281	.0200014	.2580127	.3387058
Pr(savvy_~e=3)	.0800965	.0114185	.0595617	.1042361

# Ordered Probit Predictions in Clarify

coalKnow_w2	Freq.	Percent	Cum.
0	204	5.07	5.07
1	446	11.09	16.17
2	1,132	28.15	44.32
3	2,239	55.68	100.00
Total	4,021	100.00	

```
. simqi, fd(prval(1)) changex(coalKnow_w2 2 3)
```

```
First Difference: coalKnow_w2 2 3
```

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
dPr(savvy_~e = 1)	-.0674182	.0288944	-.1211323	-.0109853

```
. simqi, fd(prval(2)) changex(coalKnow_w2 2 3)
```

```
First Difference: coalKnow_w2 2 3
```

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
dPr(savvy_~e = 2)	.0456463	.0195708	.0075696	.0834127

```
. simqi, fd(prval(3)) changex(coalKnow_w2 2 3)
```

```
First Difference: coalKnow_w2 2 3
```

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
dPr(savvy_~e = 3)	.0217719	.0099529	.0030139	.0418104