Economics

Lecture #8

Regression with a Binary Dependent Variable, Part II

Announcements: **No class Mon 10/8; PS4 due Tue 10/9, 9am**

Outline

1. Probit with multiple regressors
2. Logit
3. Logit and probit example: *HMDA data*
4. Maximum likelihood estimation
5. Ordered probit (ordered categorical data)

Probit with multiple regressors

```
. probit deny p_irat black, r

Iteration 0:   log likelihood = -872.0853
Iteration 1:   log likelihood = -800.88504
Iteration 2:   log likelihood = -797.1478
Iteration 3:   log likelihood = -797.13604

Probit estimates                                  Number of obs   =       2380
                                                  Wald chi2(2)    =     118.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -797.13604                       Pseudo R2       =     0.0859

------------------------------------------------------------------------------
             |               Robust
        deny |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      p_irat |   2.741637   .4441633     6.17   0.000     1.871092    3.612181
       black |   .7081579   .0831877     8.51   0.000      .545113    .8712028
       _cons |  -2.258738   .1588168   -14.22   0.000    -2.570013   -1.947463
------------------------------------------------------------------------------
```

We'll go through the estimation details later…

STATA Example, ctd.: predicted probit probabilities

```
. probit deny p_irat black, r

Probit estimates                                  Number of obs   =       2380
                                                  Wald chi2(2)    =     118.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -797.13604                       Pseudo R2       =     0.0859

------------------------------------------------------------------------------
             |               Robust
        deny |      Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      p_irat |   2.741637   .4441633     6.17   0.000     1.871092    3.612181
       black |   .7081579   .0831877     8.51   0.000      .545113    .8712028
       _cons |  -2.258738   .1588168   -14.22   0.000    -2.570013   -1.947463
------------------------------------------------------------------------------

.  scalar z1 = _b[_cons]+_b[p_irat]*.3+_b[black]*0

.  display "Pred prob, p_irat=.3, white: " normprob(z1)

Pred prob, p_irat=.3, white: .07546603
```

NOTE: _b[_cons] is the estimated intercept (-2.258738)
_b[p_irat] is the coefficient on p_irat (2.741637)
scalar creates a new scalar which is the result of a calculation
display prints the indicated information to the screen
normprob(z1) computes the cumulative normal probability ≤ z1

STATA Example, ctd.

$$\Pr(deny = 1 \mid P/I, black)$$

$$= \Phi(-2.26 + 2.74 \times P/I\ ratio + .71 \times black)$$

$$\quad (.16) \quad (.44) \qquad\qquad (.08)$$

- Is the coefficient on *black* statistically significant?
- Estimated effect of race for *P/I ratio* = .3:

$$\Pr(deny = 1 \mid .3, 1) = \Phi(-2.26 + 2.74 \times .3 + .71 \times 1) = \Phi(-0.73) = .233$$

$$\Pr(deny = 1 \mid .3, 0) = \Phi(-2.26 + 2.74 \times .3 + .71 \times 0) = \Phi(-1.44) = .075$$

- Difference in rejection probabilities = .158 (15.8 percentage points)
- *Still plenty of room still for omitted variable bias!*

STATA Example: HMDA data – Logit regression

```
. logit deny p_irat black, r;

Iteration 0:   log likelihood =  -872.0853      Later…
Iteration 1:   log likelihood =  -806.3571…

Logit estimates                                   Number of obs   =       2380
                                                  Wald chi2(2)    =     117.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -795.69521                       Pseudo R2       =     0.0876

------------------------------------------------------------------------------
             |               Robust
        deny |      Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      p_irat |   5.370362   .9633435     5.57   0.000     3.482244    7.258481
       black |   1.272782   .1460986     8.71   0.000     .9864339     1.55913
       _cons |  -4.125558    .345825   -11.93   0.000    -4.803362   -3.447753
------------------------------------------------------------------------------

.  dis "Pred prob, p_irat=.3, white: "
>     1/(1+exp(-(_b[_cons]+_b[p_irat]*.3+_b[black]*0)));
Pred prob, p_irat=.3, white: .07485143
   NOTE:  the probit predicted probability is .07546603
```

Predicted probabilities from estimated probit and logit models usually are very close.

The loan officer's decision

- Loan officer uses key financial variables:
  - *P/I ratio*
  - housing expense-to-income ratio
  - loan-to-value ratio
  - personal credit history
- The decision rule is nonlinear:
  - loan-to-value ratio > 80%
  - loan-to-value ratio > 95%
  - credit score
- Illegal to use "protected class" information (gender, race…)

TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

Variable	Definition	Sample Average
Financial Variables		
P/I ratio	Ratio of total monthly debt payments to total monthly income	0.331
housing expense-to-income ratio	Ratio of monthly housing expenses to total monthly income	0.255
loan-to-value ratio	Ratio of size of loan to assessed value of property	0.738
consumer credit score	1 if no "slow" payments or delinquencies 2 if one or two slow payments or delinquencies 3 if more than two slow payments 4 if insufficient credit history for determination 5 if delinquent credit history with payments 60 days overdue 6 if delinquent credit history with payments 90 days overdue	2.1
mortgage credit score	1 if no late mortgage payments 2 if no mortgage payment history 3 if one or two late mortgage payments 4 if more than two late mortgage payments	1.7
public bad credit record	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074

Additional Applicant Characteristics		
denied mortgage insurance	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
self-employed	1 if self-employed, 0 otherwise	0.116
single	1 if applicant reported being single, 0 otherwise	0.393
high school diploma	1 if applicant graduated from high school, 0 otherwise	0.984
unemployment rate	1989 Massachusetts unemployment rate in the applicant's industry	3.8
condominium	1 if unit is a condominium, 0 otherwise	0.288
black	1 if applicant is black, 0 if white	0.142
deny	1 if mortgage application denied, 0 otherwise	0.120

TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* **= 1 If mortgage application is denied, = 0 if accepted; 2380 observations.**

Regression Model	*LPM*	*Logit*	*Probit*	*Probit*	*Probit*	*Probit*
Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
black	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
P/I ratio	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
housing expense-to-income ratio	−0.048 (.110)	−0.11 (1.29)	−0.18 (0.68)	−0.30 (0.68)	−0.50 (0.70)	−0.54 (0.74)
medium loan-to-value ratio (0.80 ≤ *loan-value ratio* ≤ 0.95)	0.031* (0.013)	0.46** (0.16)	0.21** (0.08)	0.22** (0.08)	0.22** (0.08)	0.22** (0.08)
high loan-to-value ratio (*loan-value ratio* ≥ 0.95)	0.189** (0.050)	1.49** (0.32)	0.79** (0.18)	0.79** (0.18)	0.84** (0.18)	0.79** (0.18)
consumer credit score	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
mortgage credit score	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
public bad credit record	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
denied mortgage insurance	0.702** (0.045)	4.55** (0.57)	2.56** (0.30)	2.59** (0.29)	2.59** (0.30)	2.59** (0.29)

Table 11.2, ctd.

self-employed	0.060** (0.021)	0.67** (0.21)	0.36** (0.11)	0.35** (0.11)	0.34** (0.11)	0.35** (0.11)
single				0.23** (0.08)	0.23** (0.08)	0.23** (0.08)
high school diploma				−0.61** (0.23)	−0.60* (0.24)	−0.62** (0.23)
unemployment rate				0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
condominium					−0.05 (0.09)	
black × P/I ratio						−0.58 (1.47)
black × housing expense-to-income ratio						1.23 (1.69)
Additional credit rating indicator variables	no	no	no	no	yes	no
constant	−0.183** (0.028)	−5.71** (0.48)	−3.04** (0.23)	−2.57** (0.34)	−2.90** (0.39)	−2.54** (0.35)

(Table 11.2 continued)

Table 11.2, ctd.

(Table 11.2 continued)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
Applicant single; HS diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
Additional credit rating indicator variables					1.22 (0.291)	
Race interactions and black						4.96 (0.002)
Race interactions only						0.27 (0.766)
Difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients and p-values are given in parentheses under the F-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

Ordered Probit: Course Evaluations and Beauty

We have the original continuous *Y* data (course evaluations) so we don't need to use these methods, but to illustrate ordered probit we construct artificially categorized data.

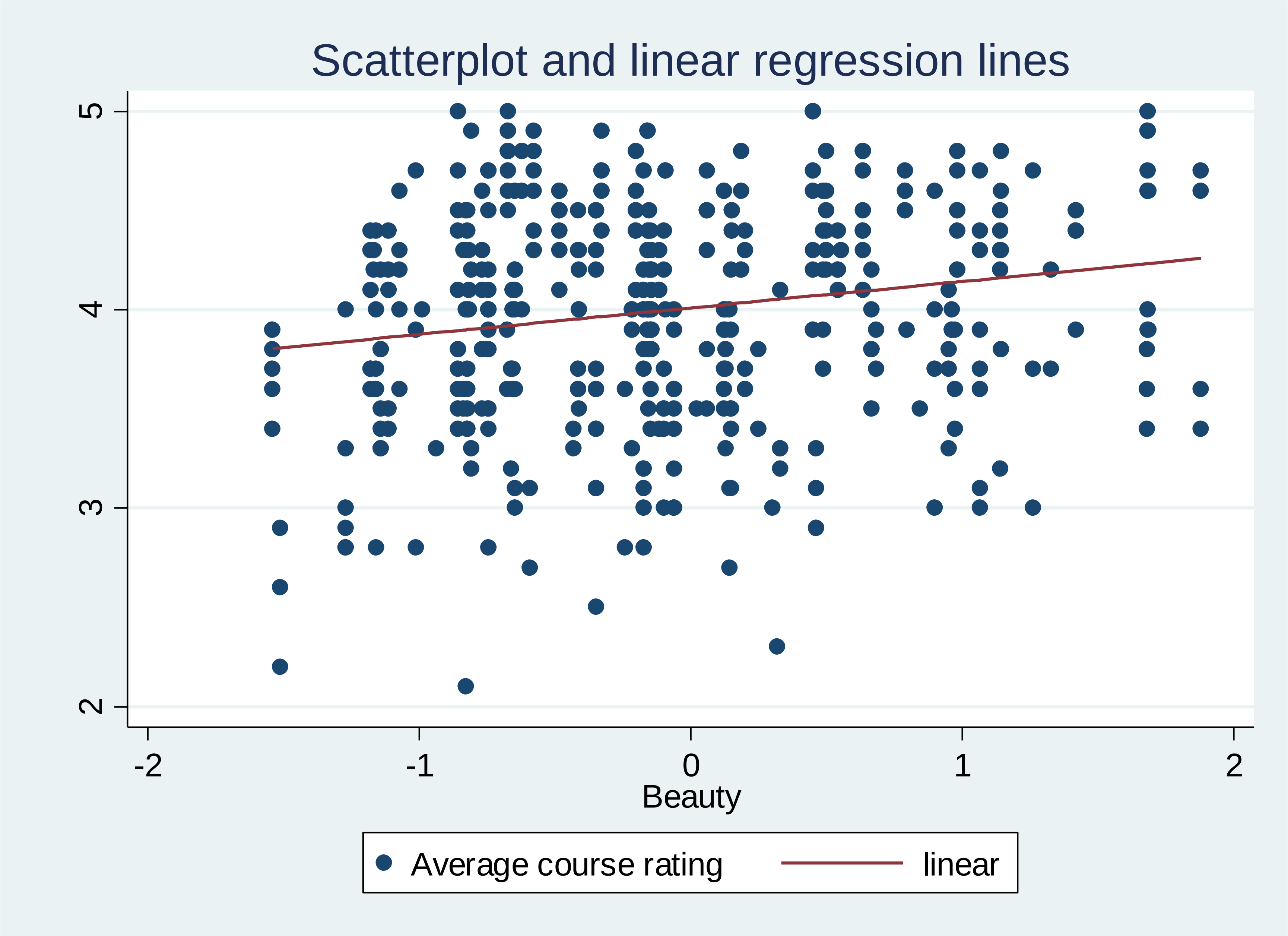
Artificial binary variable

$$eval\_q234 = \begin{cases} 0 \text{ if } courseevaluation \text{ is in first quartile} \\ 1 \text{ if } courseevaluation \text{ is in top three quartiles} \end{cases}$$

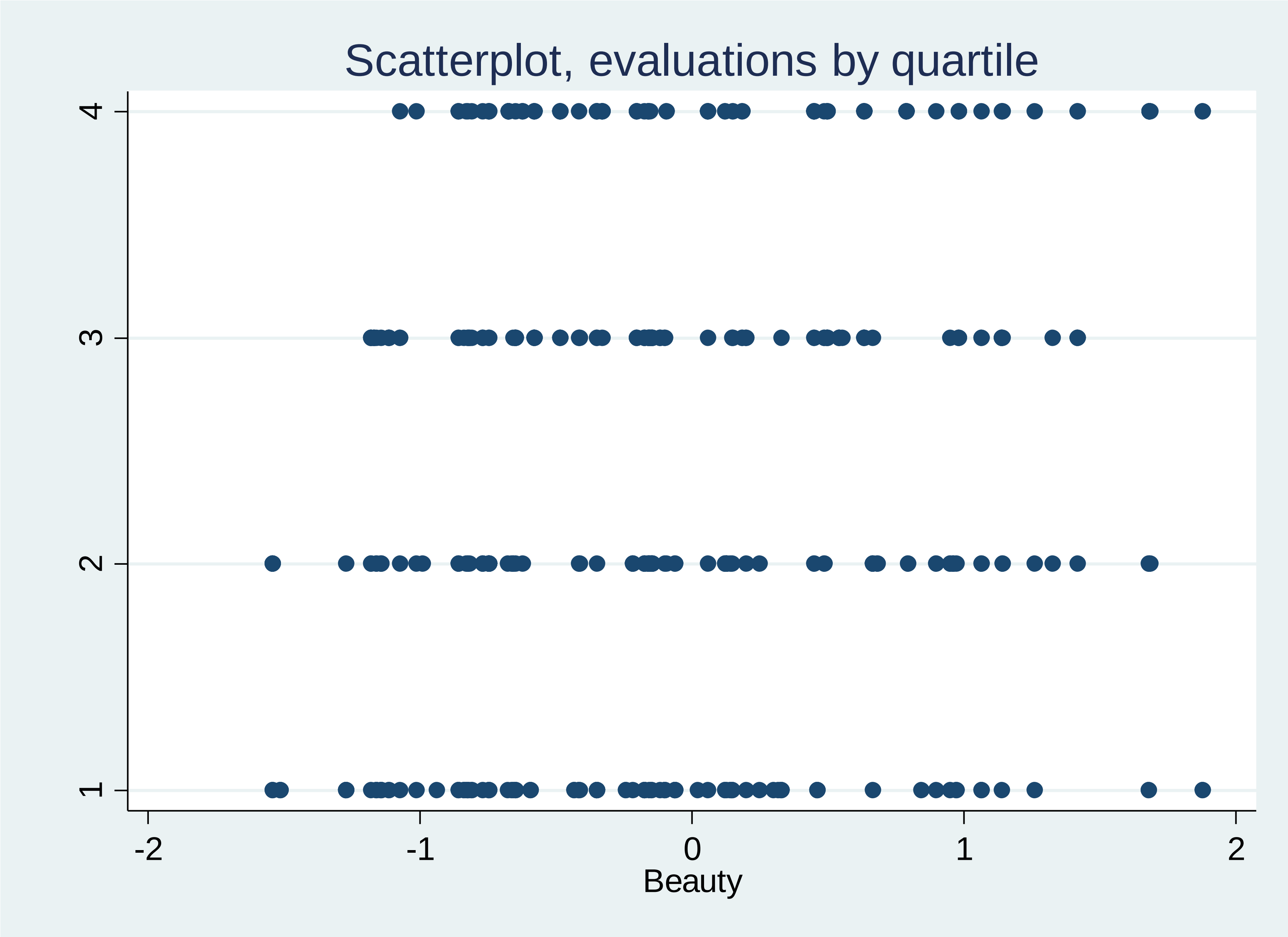
Artificial ordered categorical data

$$eval\_ord = \begin{cases} 1 \text{ if } courseevaluation \text{ is in first quartile} \\ 2 \text{ if } courseevaluation \text{ is in second quartile} \\ 3 \text{ if } courseevaluation \text{ is in third quartile} \\ 4 \text{ if } courseevaluation \text{ is in fourth quartile} \end{cases}$$

Original data with linear regression:



Categorical course evaluation data (categorized by quartile)



STATA implementation – create variables; probit; ordered probit

```
. su courseevaluation, d;

                    Average course rating
-------------------------------------------------------------
      Percentiles      Smallest
 1%          2.6            2.1
 5%            3            2.2
10%          3.3            2.3       Obs                 463
25%          3.6            2.5       Sum of Wgt.         463

50%            4                      Mean           3.998272
                        Largest       Std. Dev.      .5548656
75%          4.4              5
90%          4.7              5       Variance       .3078758
95%          4.8              5       Skewness      -.4658753
99%            5              5       Kurtosis       2.881628

. gen evalq2 = (courseevaluation>r(p25))*(courseevaluation<=r(p50));

. gen evalq3 = (courseevaluation>r(p50))*(courseevaluation<=r(p75));

. gen evalq4 = (courseevaluation>r(p75));

. gen eval_q234 = evalq2 + evalq3 + evalq4;

. gen eval_ord = 1 + evalq2 + 2*evalq3 + 3*evalq4;
```

```
. reg courseevaluation btystdave, r;
Linear regression                                      Number of obs =     463
                                                       F(  1,   461) =   16.94
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0357
                                                       Root MSE      =  .54545

------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1330014   .0323189     4.12   0.000     .0694908    .1965121
       _cons |   4.010023   .0253299   158.31   0.000     3.960246    4.059799
------------------------------------------------------------------------------

. reg eval_q234 btystdave, r;
Linear regression                                      Number of obs =     463
                                                       F(  1,   461) =    9.51
                                                       Prob > F      =  0.0022
                                                       R-squared     =  0.0194
                                                       Root MSE      =  .43833

------------------------------------------------------------------------------
             |               Robust
   eval_q234 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |    .078026   .0253052     3.08   0.002     .0282982    .1277538
       _cons |   .7412348   .0201643    36.76   0.000     .7016095    .7808601
------------------------------------------------------------------------------
```

```
. probit eval_q234 btystdave, r;

Iteration 0:   log pseudolikelihood = -268.02744
Iteration 1:   log pseudolikelihood = -263.43691
Iteration 2:   log pseudolikelihood = -263.42781
Iteration 3:   log pseudolikelihood = -263.42781

Probit regression                                 Number of obs   =        463
                                                  Wald chi2(1)    =       8.52
                                                  Prob > chi2     =     0.0035
Log pseudolikelihood = -263.42781                 Pseudo R2       =     0.0172

------------------------------------------------------------------------------
             |               Robust
   eval_q234 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .2471247   .0846581     2.92   0.004      .081198    .4130515
       _cons |   .6597471   .0647791    10.18   0.000     .5327825    .7867117
------------------------------------------------------------------------------
```

```
. * ordered probit;
. oprobit eval_ord btystdave, r;

Iteration 0:   log pseudolikelihood = -641.41106
Iteration 1:   log pseudolikelihood = -633.59498
Iteration 2:   log pseudolikelihood = -633.59449

Ordered probit regression                         Number of obs   =        463
                                                  Wald chi2(1)    =      15.19
                                                  Prob > chi2     =     0.0001
Log pseudolikelihood = -633.59449                 Pseudo R2       =     0.0122

------------------------------------------------------------------------------
             |               Robust
    eval_ord |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .2549661   .0654143     3.90   0.000     .1267564    .3831759
-------------+----------------------------------------------------------------
       /cut1 |  -.6604092   .0638122                     -.7854789   -.5353394
       /cut2 |   .0227324   .0594761                     -.0938386    .1393034
       /cut3 |   .7111037   .0644798                      .5847256    .8374819
------------------------------------------------------------------------------
```

Calculation of effects – ordered probit

Predicted probabilities for ordered probit (4 categories):

$$\Pr[Y_i = 0|X_i] = \Phi[c_1 - \beta_1 X_i]$$

$$\Pr[Y_i = 1|X_i] = \Phi[c_2 - \beta_1 X_i] - \Phi[c_1 - \beta_1 X_i]$$

$$\Pr[Y_i = 2|X_i] = \Phi[c_3 - \beta_1 X_i] - \Phi[c_2 - \beta_1 X_i]$$

$$\Pr[Y_i = 3|X_i] = 1 - \Phi[c_3 - \beta_1 X_i]$$

What is effect of increasing *btystdave* from -1 to 0 on probability of being in category 3?

x = -1: $\widehat{PR}[Y_i = 2|X_i = -1] = \Phi[\hat{c}_3 - \hat{\beta}_1 \times (-1)] - \Phi[\hat{c}_2 - \hat{\beta}_1 \times (-1)]$

$$= \Phi[.711 - \mathbf{.255} \times (-1)] - \Phi[.023 - \mathbf{.255} \times (-1)]$$

$$= \Phi[.966] - \Phi[.278]$$

$$= .833 - .609 = .224$$

$x = 0$: $PR[Y_i = 2|X_i = 0] = \Phi[\hat{c}_3 - \hat{\beta}_1 \times 0] - \Phi[\hat{c}_2 - \hat{\beta}_1 \times 0]$

$$= \Phi[.711 - \mathbf{.255} \times 0] - \Phi[.023 - \mathbf{.255} \times 0]$$

$$= \Phi[.711] - \Phi[.023]$$

$$= .761 - .509 = .252$$

An increase in *btystdave* from -1.0 to 0 is associated with an increase in the probability of being in the third quartile from .224 to .252, an increase of .028 percentage points

STATA .do file for Beauty example (probit, logit, ordered probit)

```
clear
capture log close
**************************************************************
*  beauty_3_lect9.do
*  Ec1123
*    probit, ordered probit, illustrations
**************************************************************
set more off
log using beauty_3_oprobit_exs.log, replace
**************************************************************
* read in data
use hamermesh_beauty
desc
su
*
gen male = 1-female
gen bty2 = btystdave*btystdave
gen bty3 = btystdave*btystdave*btystdave
gen bty_male = btystdave*male
*
* create data for ordered probit - quartiles
su courseevaluation, d
gen evalq2 = (courseevaluation>r(p25))*(courseevaluation<=r(p50))
gen evalq3 = (courseevaluation>r(p50))*(courseevaluation<=r(p75))
gen evalq4 = (courseevaluation>r(p75))
gen eval_q234 = evalq2 + evalq3 + evalq4
```

```
gen eval_ord = 1 + evalq2 + 2*evalq3 + 3*evalq4
*
list courseevaluation eval_q234 eval_ord
*************************************************************
*    graphs
*************************************************************
reg courseevaluation btystdave, r
 predict peval
 label var peval "linear"
twoway scatter courseevaluation peval btystdave, ///
 ms(O i i i) connect(. l l l) sort(btystdave) ///
 title("Scatterplot and linear regression lines") ///
 xtitle("Beauty") ytitle("Course Overall") yscale(r(2 5))
graph export "beauty_3a.png", replace
*************************************************************
*    probit, logit regressions - one regressor
*************************************************************
reg courseevaluation btystdave, r
* linear probability model
reg eval_q234 btystdave, r
* probit
probit eval_q234 btystdave, r
* logit
logit eval_q234 btystdave, r
*
*************************************************************
*    ordered probit regressions - one regressor
*************************************************************
* ordered probit
oprobit eval_ord btystdave, r
```

```
twoway scatter eval_ord btystdave, ///
 ms(O i i i) connect(. l l l) sort(btystdave) ///
 title("Scatterplot, evaluations by quartile") ///
 xtitle("Beauty") ytitle("Course Overall")
graph export "beauty_3b.png", replace
sca a2 = _b[/cut2] - _b[btystdave]*(-1)
sca a3 = _b[/cut3] - _b[btystdave]*(-1)
dis a2 a3 normprob(a2) normprob(a3) normprob(a3)-normprob(a2)
sca b2 = _b[/cut2] - _b[btystdave]*(0)
sca b3 = _b[/cut3] - _b[btystdave]*(0)
dis b2 b3 normprob(b2) normprob(b3) normprob(b3)-normprob(b2)
************************************************************
log close
clear
exit
```