THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

EFMD EQUIS ACCREDITED

ECON2271

Business Econometrics

Or: Practical Econometrics for Beginners

Week 10: Models for limited and discrete dependent Variables

Topic 4: Limited Discrete Dependent Vars

Agenda and learning outcomes

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Topic 4: Models for limited and discrete dependent variables:

Part (i): Models for binary dependent variables

a) The linear probability model:
b) The logit and probit models:

❖ Text reference: Studenmund Ch. 12 (Gujarati Ch 15 are more detailed)

Part (ii): Other types of non-continuous dependent variables:

c) Models for ordered dependent variables (ordered probit/logit)
   - Understand when and why these models may be preferred to standard linear regression models
d) Models for count data (Poisson models)
   - Understand when and why these models may be preferred to standard linear regression models, and how to interpret model estimates
e) Models for censored or truncated data (Tobit model)
   - Understand when and why these models may be preferred to standard linear regression models.

❖ Text reference: Gujarati Ch 15

3

Topic 4(i): Binary Dependent Variables

Recap

- **3 models for binary dependent variables:**
  1. Linear Probability Model (LPM): $Y_i = b_0 + b_1 X_i + u_i$
     - Assumes a linear relationship between *X* and Pr(*Y* = 1)
     - Treats Y as an unlimited and continuous variable
     - Coefficients can be interpreted as the marginal effect of X on the probability of observing Y=1.
  2. Logit model: $\ln\left(\frac{Y_i}{1-Y_i}\right) = b_0 + b_1 X_i + u_i$
     - Assumes a sigmoid relationship between *X* and Pr(*Y* = 1)
     - Treats Y as a limited variable, bounded at 0 and 1.
     - Coefficients measure the estimated marginal effect of X on the log odds-ratio of observing Y = 1.
  3. Probit model: $Z_i = \Phi^{-1}(P_i) = b_0 + b_1 X_i + u_i$
     - Assumes a sigmoid relationship between *X* and Pr(*Y* = 1)
     - Treats Y as a limited variable, bounded at 0 and 1.
     - Coefficients are hard to interpret directly; you have to calculate Z and refer to the normal distribution table.

4

Topic 4(i): Binary Dependent Variables

Pre-lab exercises

- **Estimating women's probability of living with children (for illustrative purposes)**

```
. reg kids lninc lnwages lnwealth schooling age agesq if female==1

      Source |       SS           df       MS      Number of obs   =     3,770
-------------+----------------------------------   F(6, 3763)      =    194.86
       Model |  211.081664         6  35.1802773   Prob > F        =    0.0000
    Residual |  679.362899     3,763  .180537576   R-squared       =    0.2371
-------------+----------------------------------   Adj R-squared   =    0.2358
       Total |  890.444562     3,769  .236254859   Root MSE        =     .4249

------------------------------------------------------------------------------
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lninc |   .0444408   .0122705     3.62   0.000     .0203833    .0684983
     lnwages |  -.0777184   .0071685   -10.84   0.000    -.0917729   -.0636638
    lnwealth |    .016425   .0060605     2.71   0.007     .0045429    .0283071
   schooling |   .0010717   .0030653     0.35   0.727     -.004938    .0070814
         age |   .0994679   .0056103    17.73   0.000     .0884683    .1104674
       agesq |  -.0013141    .000063   -20.86   0.000    -.0014376   -.0011906
       _cons |  -1.218137   .1680056    -7.25   0.000    -1.547528   -.8887458
------------------------------------------------------------------------------
```

Topic 4(i): Binary Dependent Variables

Pre-lab exercises

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

1. The first model is a standard Linear Probability Model (LPM). Answer these questions:

   a) What estimator is used to generate these results?

   b) What are the potential issues implies with using that estimator?

   c) Notwithstanding these limitations (b), provide a written interpretation for each of the estimated model parameters.

      - lninc: Comparing across females, an increase in the log of income of 1 is associated with a
      - Lnwages: Comparing across females, an increase in the log of wages of 1 is associated with a
      - Lnwealth: (same as lninc, except 1.6 p.p.)
      - Schooling: Comparing across females, an additional year of schooling is associated with a

Topic 4(i): Binary Dependent Variables

Pre-lab exercises

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

1. The first model is a standard Linear Probability Model (LPM). Answer these questions:

   c) Notwithstanding these limitations (b), provide a written interpretation for each of the estimated model parameters. (cont.)

      - Age: comparing across females, the relationship between age and the probability of living with children is nonlinear, with a slope equal about
      - Constant: A female with zero income, wages, wealth and schooling, aged zero, will have a probability of

CRICOS Provider Code: 00126G

7

Topic 4(i): Binary Dependent Variables

Pre-lab exercises

The University of Western Australia
Achieve International Excellence

1. The first model is a standard Linear Probability Model (LPM). Answer these questions:

   d) Calculate the predicted value for *kids* for a female whose annual disposable household income, wages and household wealth are $130,000, $80,000 and $200,000, respectively; who has 16 years of schooling and is aged 40. Interpret this number.

   e) Calculate the predicted value for someone who is aged 65, but otherwise the same. Interpret this number.

8

Topic 4(i): Binary Dependent Variables

Pre-lab exercises

The University of Western Australia
Achieve International Excellence

- **The logit model estimates:**

```
. logit kids lninc lnwages lnwealth schooling age agesq if female==1

Iteration 0:   log likelihood = -2508.0738
Iteration 1:   log likelihood = -1911.4224
Iteration 2:   log likelihood = -1823.9753
Iteration 3:   log likelihood =  -1820.538
Iteration 4:   log likelihood = -1820.5267
Iteration 5:   log likelihood = -1820.5267

Logistic regression                             Number of obs     =      3,770
                                                LR chi2(6)        =    1375.09
                                                Prob > chi2       =     0.0000
Log likelihood = -1820.5267                     Pseudo R2         =     0.2741

------------------------------------------------------------------------------
        kids |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lninc |    .455405   .0958004     4.75   0.000     .2676396    .6431704
     lnwages |  -.5345653   .0506023   -10.56   0.000     -.633744   -.4353865
    lnwealth |   .0663709   .0373191     1.78   0.075    -.0067733     .139515
   schooling |    -.00588   .0183588    -0.32   0.749    -.0418625    .0301025
         age |   1.101115   .0515112    21.38   0.000     1.000155    1.202075
       agesq |  -.0144092   .0006461   -22.30   0.000    -.0156755    -.013143
       _cons |  -20.47227   1.339051   -15.29   0.000    -23.09676   -17.84778
------------------------------------------------------------------------------
```

9

Topic 4(i): Binary Dependent Variables
Pre-lab exercises

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

2. The second model is a logit model. Answer these questions:

a) What estimator is used to generate these results?
b) What is the implied underlying shape of the relationship between the representative explanatory variable and the dependent variable?
c) What do the estimated model parameters represent? (Hint: refer to slide 9)
   - The estimated relationship between the X variables and the log odds ratio of living with children.
   - Lninc: Comparing across females, an increase in the log of income of 1 is associated with an increase in the log odds ratio of living with children of about 0.46 (Note: this is more than 10 times the LPM coefficient! Under ideal conditions, it should be only 4 times… This is a sign that this coefficient is sensitive to the choice of model. Should be explored further.
   - Lnwages: Comparing across females, an increase in the log of wages of 1 is associated with a decrease in the log odds ratio of living with children of about 0.53 (Note: this is about 6 times the LPM coefficient. So a little sensitive.)
   - Lnwealth: Comparing across females, an increase in the log of wealth of 1 is associated with an increase in the log odds ratio of living with children of about 0.066 (Note: this is not far off 4 times the LPM coefficient. Means the coefficient seems robust.)
   - The coefficient for schooling is now negative!! But it is not significant, so doesn't really matter.

10

Topic 4(i): Binary Dependent Variables
Pre-lab exercises

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

d) Calculate the predicted (logit) value for *kids* for a female whose annual disposable household income, wages and household wealth are $130,000, $80,000 and $200,000, respectively; who has 16 years of schooling and is aged 40. What you now have is a logit – a log of the odds of a person living with children. Convert this number into a probability. You can do this by working with the log odds, or you can work from the format: (Here, Z is the logit value you've calculated). $P(Y=1|X)=\frac{1}{1+e^{-Z}}$
   - Z= -20.47 + 0.46*ln(130000) – 0.53*ln(80000) + 0.066*ln(200000) -0.00588*16 + 1.10*40-0.0144*40^2 ≈ 0.53
   - $P(Kids=1|\boldsymbol{X})=\frac{1}{1+e^{-0.53}}$ ≈ 0.63
   -

e) Calculate the predicted (logit) value for someone who is aged 65, but otherwise the same. Interpret this number. Compare with your answer in 1e.
   - Z= -20.47 + 0.46*ln(130000) – 0.53*ln(80000) + 0.066*ln(200000) -0.00588*16 + 1.10*65-0.0144*65^2 ≈ -9.77
   - $P(Kids=1|\boldsymbol{X})=\frac{1}{1+e^{-(-9.77)}}$ ≈ 0.00006.

11

Topic 4(i): Binary Dependent Variables

a) Pre-lab exercises

- We can confirm our estimated probability for (d) using Stata:

```
. margins, at(lninc=11.7753 lnwages=11.2898 lnwealth=12.20607 schooling=16 age=40 agesq=1600)

Adjusted predictions                            Number of obs     =      3,770
Model VCE    : OIM

Expression   : Pr(kids), predict()
at           : lninc           =     11.7753
               lnwages         =     11.2898
               lnwealth        =    12.20607
               schooling       =          16
               age             =          40
               agesq           =        1600

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .6366827   .0181571    35.07   0.000     .6010956    .6722699
------------------------------------------------------------------------------
```

CRICOS Provider Code: 00126G

12

Topic 4(i): Binary Dependent Variables

a) Pre-lab exercises

- We can confirm our estimated probability for (d) using Stata:

```
. margins, at(lninc=11.7753 lnwages=11.2898 lnwealth=12.20607 schooling=16 age=65 agesq=4225)

Adjusted predictions                            Number of obs     =      3,770
Model VCE    : OIM

Expression   : Pr(kids), predict()
at           : lninc           =     11.7753
               lnwages         =     11.2898
               lnwealth        =    12.20607
               schooling       =          16
               age             =          65
               agesq           =        4225

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .0000591   .0000243     2.43   0.015     .0000115    .0001068
------------------------------------------------------------------------------
```

CRICOS Provider Code: 00126G

Topic 4(i): Binary Dependent Variables

Pre-lab exercises

- **Marginal effects:** Calculates dy/dx (i.e. point slopes of the logit function) at mean X values

```
. margins, dydx(*) atmeans post

Conditional marginal effects                    Number of obs     =      3,770
Model VCE    : OIM

Expression   : Pr(kids), predict()
dy/dx w.r.t. : lninc lnwages lnwealth schooling age agesq
at           : lninc           =     11.51531 (mean)
               lnwages         =     10.56226 (mean)
               lnwealth        =     12.86391 (mean)
               schooling       =     13.92679 (mean)
               age             =     43.56764 (mean)
               agesq           =     2016.641 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lninc |   .0815469   .0170246     4.79   0.000     .0481794    .1149144
     lnwages |  -.0957217   .0092877   -10.31   0.000    -.1139253   -.0775181
    lnwealth |   .0118847      .0067     1.77   0.076     -.001247    .0250164
   schooling |  -.0010529   .0032855    -0.32   0.749    -.0074924    .0053866
         age |   .1971707   .0071061    27.75   0.000      .183243    .2110983
       agesq |  -.0025802   .0000853   -30.23   0.000    -.0027475   -.0024129
------------------------------------------------------------------------------
```


Topic 4(i): Binary Dependent Variables

Pre-lab exercises

- **Comparing these with LPM estimates (equivalent in terms of interpretation, at mean X values)**

kids	Coef.	Std. Err.	t	P>|t|
lninc	.0444408	.0122705	3.62	0.000
lnwages	-.0777184	.0071685	-10.84	0.000
lnwealth	.016425	.0060605	2.71	0.007
schooling	.0010717	.0030653	0.35	0.727
age	.0994679	.0056103	17.73	0.000
agesq	-.0013141	.000063	-20.86	0.000
_cons	-1.218137	.1680056	-7.25	0.000

dy/dx	Std. Err.	z	P>|z|
.0815469	.0170246	4.79	0.000
-.0957217	.0092877	-10.31	0.000
.0118847	.0067	1.77	0.076
-.0010529	.0032855	-0.32	0.749
.1971707	.0071061	27.75	0.000
-.0025802	.0000853	-30.23	0.000

- Both models suggest schooling is statistically unrelated to the probability of living with children, for women.
- Other than that, the sign and significance of the estimated coefficients are consistent across the two models.
- However, their magnitudes are not – they vary a fair bit. The effect of income is twice as large in the logit model than in the LPM model. The relationship between age and pr(kids) appears to be quite different too, so this estimate is sensitive to choice of model.
- Note that the logit model estimate only holds for mean values of **X**

15

Topic 4(i): Binary Dependent Variables
Pre-lab exercises

The University of Western Australia
Achieve International Excellence

- **You can also estimate dy/dx at any other values of X:**

```
. margins, dydx(*) at(lninc=11.7753 lnwages=11.2898 lnwealth=12.20607 schooling=16
>  age=40 agesq=1600) post

Conditional marginal effects                    Number of obs     =      3,770
Model VCE    : OIM

Expression   : Pr(kids), predict()
dy/dx w.r.t. : lninc lnwages lnwealth schooling age agesq
at           : lninc           =     11.7753
               lnwages         =     11.2898
               lnwealth        =    12.20607
               schooling       =          16
               age             =          40
               agesq           =        1600

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lninc |   .1053433    .021565     4.88   0.000     .0630767    .1476099
     lnwages |  -.1236544   .0124178    -9.96   0.000    -.1479928   -.0993161
    lnwealth |   .0153528   .0087934     1.75   0.081     -.001882    .0325875
   schooling |  -.0013602   .0042578    -0.32   0.749    -.0097052    .0069849
         age |   .2547075   .0101021    25.21   0.000     .2349078    .2745071
       agesq |  -.0033331   .0001286   -25.93   0.000    -.0035851   -.0030811
------------------------------------------------------------------------------
```

Estimated marginal effects will differ with X values. That's the point with a nonlinear function!

16

Topic 4(ii): Limited and Discrete Y
c) Models for ordered variables

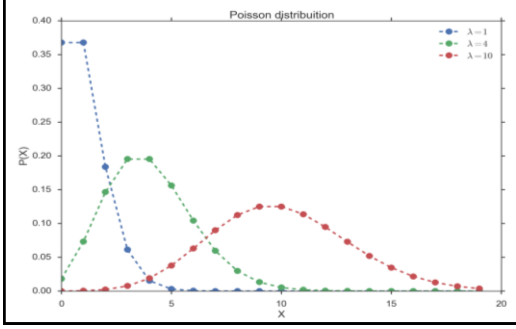
The University of Western Australia
Achieve International Excellence

- **Ordered dependent variables:**
  - These are both discrete and limited.
  - E.g. education levels, Likert scales (e.g. strongly disagree, disagree, neutral, agree, strongly agree)
  - Can be estimated as a standard linear regression model, using OLS, but:
    - This imposes an assumption about linearity between X and Y
    - It implicitly imposes an assumption that people who strongly disagree and disagree are equally different as people who disagree and are neutral, and likewise for all adjacent score-points on the ordered scale. I.e. assumes equidistant score-points.
    - We also have the same issue as for binary dependent variables, in that the model assumes Y is unlimited.
    - If you can argue that the assumptions of linearity and equidistant score-points is reasonable (if not perfect), then OLS will yield reasonable parameters – at least if there is sufficient variation in Y (i.e. if the scale is wide enough)
    - If this seems heroic, you can consider an ordered probability (logit or probit) model instead.
    - Ordered probability variables are not covered in either Gujarati or Studenmund, so we won't cover it in detail here. I just want to you be aware that such models exist as an alternative to OLS, if you should find yourself trying to estimate models with ordered Y.

17

Topic 4(ii): Limited and Discrete Y
d) Models for count data

- **Count data:**
  - These are both discrete and limited, but they are numeric scores, so the underlying measurement scale can usually be treated as ratio quality (fully cardinal).
  - E.g. Number of children, number of marriages/divorces, number of standard drinks per week, etc.
  - Can be estimated as a standard linear regression model, using OLS, but:
    - The main problem is that OLS assumes Y is normally distributed
    - Count data are rarely normally distributed. They tend to follow a Poisson distribution:



- ➢ A discrete probability distribution used to describe the occurrence of unlikely events in large number of independent repeated trials.
- ➢ There is "family" of distributions, defined by the mean (λ or μ), which is bounded at 0. If the mean is quite large (e.g. λ = μ =10), the distribution appears quite 'normal'.
- ➢

18

Topic 4(ii): Limited and Discrete Y
d) Models for count data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Count data:**
  - So, if the dependent variable is a count variable, and especially if it is a "rare event" type variable, where zeros are commonly observed and the mean is low (such as number of children or number of marriages), then OLS will not be optimal.
  - Solution: use a model specifically tailored to this type of underlying distribution – the Poisson Regression Model:

$$Y_i = \mathrm{E}(Y_i) + u_i = \mu_i + u_i \,;\, \mu_i = b_0 + \boldsymbol{b}_{\mathrm{k}} \boldsymbol{X}_{ki}$$

$\Rightarrow$The model to estimate is: $Y_i = \frac{\mu^Y e^{-\mu}}{Y!} + u_i$

$\Rightarrow$Clearly nonlinear! So the estimator is ML

$\Rightarrow$Interpretation of model estimates is now a little different. It can be interpreted as a "semielasticity", such that b1 (the coefficient for X1) reflects the percentage change in Y for every one-unit change in X (See Cameron & Trivedi, 2010, p. 576). However, as for logit (and probit), you can ask Stata to calculate marginal effects at particular values of X (or at means).

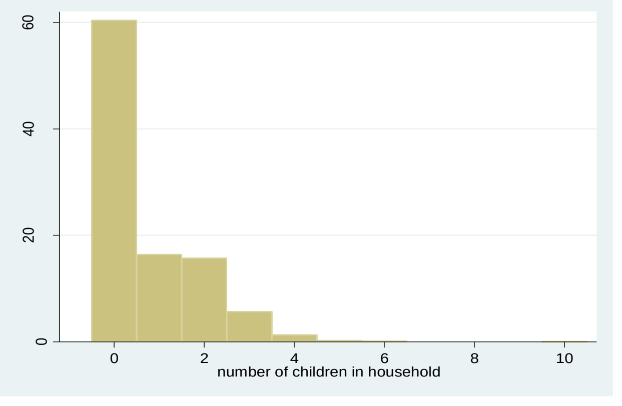
19

Topic 4(ii): Limited and Discrete Y

d) Models for count data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Count data example:** A model for the number of children per household (females only)
  - First, look at the distribution (Stata code: *histogram nkids if female==1, discrete percent*)



20

Topic 4(ii): Limited and Discrete Y

d) Models for count data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Count data example:** A model for the number of children per household (females)
  - ➢ Check the condition that mean = variance

```
. su nkids if female==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       nkids |      6,062    .7291323    1.069074          0         10
```

  - ➢ Hmm… Variance = 1.14
  - ➢ Common problem, but can be fixed
    - ➢ Consequence is that SEs will be estimated as smaller than they really are
    - ➢ The solution is to warn Stata the data is "overdispersed". We'll add that later. First, lets start with the basic model estimation:

– **The Poisson model**: $nkids_i = \mathrm{E}(nkids_i) + u_i = \mu_i + u_i\,;\, \mu_i = b_0 + \boldsymbol{b}_k\boldsymbol{X}_{ki}$

$\boldsymbol{X}$= a vector of explanatory variables including age, marital status (*marrdef* = 1 if married/defacto), schooling, number of siblings and a dummy for first-borns (=*eldest*), and years of schooling.

Topic 4(ii): Limited and Discrete Y

d) Models for count data

The University of Western Australia
Achieve International Excellence

- **Count data example:** A model for the number of children per household (females)

```
. poisson nkids age agesq marrdef schooling nsiblings eldest if female==1

Iteration 0:   log likelihood = -5537.2135
Iteration 1:   log likelihood = -5475.6511
Iteration 2:   log likelihood = -5475.6066
Iteration 3:   log likelihood = -5475.6066

Poisson regression                              Number of obs     =      5,589
                                                LR chi2(6)        =    3012.24
                                                Prob > chi2       =     0.0000
Log likelihood = -5475.6066                     Pseudo R2         =     0.2157

------------------------------------------------------------------------------
       nkids |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5446435   .0198151    27.49   0.000     .5058066    .5834805
       agesq |  -.0074526   .0002505   -29.76   0.000    -.0079435   -.0069617
     marrdef |   .5301556   .0450858    11.76   0.000     .4417891     .618522
   schooling |  -.0431935   .0066957    -6.45   0.000    -.0563168   -.0300702
   nsiblings |    .042854   .0086327     4.96   0.000     .0259342    .0597738
      eldest |   .0951933   .0333036     2.86   0.004     .0299195    .1604671
       _cons |  -9.507813   .3921896   -24.24   0.000    -10.27649   -8.739136
------------------------------------------------------------------------------
```


Topic 4(ii): Limited and Discrete Y

d) Models for count data

The University of Western Australia
Achieve International Excellence

- **Count data example:** Let's now ask for robust SEs

```
. poisson nkids age agesq marrdef schooling nsiblings eldest if female==1, vce(robust)

Iteration 0:   log pseudolikelihood = -5537.2135
Iteration 1:   log pseudolikelihood = -5475.6511
Iteration 2:   log pseudolikelihood = -5475.6066
Iteration 3:   log pseudolikelihood = -5475.6066

Poisson regression                              Number of obs     =      5,589
                                                Wald chi2(6)      =     640.05
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -5475.6066               Pseudo R2         =     0.2157

------------------------------------------------------------------------------
             |               Robust
       nkids |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5446435    .031313    17.39   0.000     .4832711    .6060159
       agesq |  -.0074526   .0004062   -18.35   0.000    -.0082487   -.0066565
     marrdef |   .5301556     .05512     9.62   0.000     .4221224    .6381887
   schooling |  -.0431935   .0071717    -6.02   0.000    -.0572497   -.0291373
   nsiblings |    .042854   .0104992     4.08   0.000      .022276     .063432
      eldest |   .0951933   .0338945     2.81   0.005     .0287614    .1616253
       _cons |  -9.507813   .5940674   -16.00   0.000    -10.67216   -8.343462
------------------------------------------------------------------------------
```

23

Topic 4(ii): Limited and Discrete Y

d) Models for count data

- **Count data example:** Let's now ask for marginal effects:

```
. margins, dydx(*)

Average marginal effects                        Number of obs     =      5,589
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age agesq marrdef schooling nsiblings eldest

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4030499   .0231577    17.40   0.000     .3576616    .4484382
       agesq |  -.0055151   .0003012   -18.31   0.000    -.0061054   -.0049248
     marrdef |   .3923285    .040311     9.73   0.000     .3133203    .4713366
   schooling |  -.0319643   .0053829    -5.94   0.000    -.0425146    -.021414
   nsiblings |    .031713   .0078163     4.06   0.000     .0163934    .0470327
      eldest |   .0704454   .0250748     2.81   0.005     .0212998    .1195911
------------------------------------------------------------------------------
```


24

Topic 4(ii): Limited and Discrete Y

d) Models for count data

- **Count data example:** Compare with standard linear regression estimates:

```
. reg nkids age agesq marrdef schooling nsiblings eldest if female==1

      Source |       SS           df       MS      Number of obs   =     5,589
-------------+----------------------------------   F(6, 5582)      =    261.19
       Model |  1415.92299         6  235.987165   Prob > F        =    0.0000
    Residual |  5043.33341     5,582  .903499356   R-squared       =    0.2192
-------------+----------------------------------   Adj R-squared   =    0.2184
       Total |   6459.2564     5,588   1.1559156   Root MSE        =    .95053

------------------------------------------------------------------------------
       nkids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1107844   .0097177    11.40   0.000      .091734    .1298348
       agesq |   -.001672   .0001078   -15.51   0.000    -.0018833   -.0014606
     marrdef |   .3337081   .0302049    11.05   0.000     .2744948    .3929214
   schooling |  -.0266682   .0053402    -4.99   0.000    -.0371372   -.0161992
   nsiblings |    .018919    .006631     2.85   0.004     .0059196    .0319184
      eldest |   .0801419   .0277644     2.89   0.004     .0257129    .1345709
       _cons |  -.6315384   .2198956    -2.87   0.004    -1.062619   -.2004575
------------------------------------------------------------------------------
```

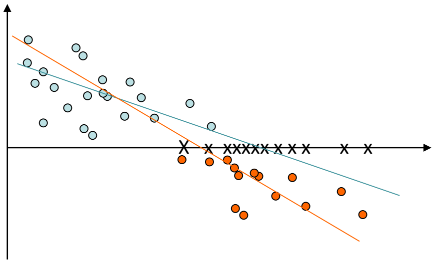
25

Topic 4(ii): Limited and Discrete Y

e) Models for censored data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Censored data:**
  - A censored sample is one where information is only available for some observations.
  - E.g. Estimating the likelihood of having extramarital affairs (Ray Fair's model, see Gujarati section 15.11).
    - ➢ Problem is that you only observe people who admit to having affairs, but you don't observe all the people on the other end of 0: the degree with which people perceive such behaviour as inconceivable.



26

Topic 4(ii): Limited and Discrete Y

e) Models for censored data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Censored data: The Tobit model**
  - Tobit versus OLS, from Gujarati p 619: Z_5 is the variable measuring religiosity

OLS AND TOBIT ESTIMATES OF EXTRAMARITAL AFFAIRS

Explanatory variable	OLS estimate	Tobit estimate
Intercept	5.8720 (5.1622)*	7.6084 (1.9479)†
Z_1	0.0540 (0.1799)	0.9457 (0.8898)
Z_2	−0.0509 (−2.2536)	−0.1926 (−2.3799)
Z_3	0.1694 (4.1109)	0.5331 (3.6368)
Z_4	−0.1426 (−0.4072)	1.0191 (0.7965)
Z_5	−0.4776 (−4.2747)	−1.6990 (−4.1906)
Z_6	−0.0137 (−0.2143)	0.0253 (0.1113)
Z_7	0.1049 (1.1803)	0.2129 (0.6631)
Z_8	−0.7118 (−5.9319)	−2.2732 (−5.4724)
R^2	0.1317	0.1515

*The figures in the parentheses are the *t* values.
†The figures in the parentheses are the *Z* (standard normal) values.
Note: In all there are 601 observations, of which 451 have zero values for the dependent variable (number of extramarital affairs) and 150 have nonzero values.