AFRE 835: Introductory Econometrics

Chapter 7: Multiple Regression Analysis: Discrete Variables

Spring 2017

Introduction

- For the most part, we have focussed our attention on continuous (or quantitative) dependent and independent variables thus far.
- However, we are often interested in how a dependent variable changes over qualitatively different (or discrete) subgroups; e.g.,
  - How housing prices vary by region of the country;
  - How wages vary by gender or race;
  - How educational outcomes vary by private versus public education systems;
  - How regional economic growth varies by EPA non-attainment designation;
- We are also often interested in how discrete outcomes are impacted by different discrete or continuous independent variables; e.g.,
  - How female participation in the labor force is impacted family size and composition;
  - How promotion varies by gender or race;
  - How arrests are impacted by law enforcement expenditures and sentencing policies;

Outline

Discrete Variables

- A *discrete* variable is one that takes on at most a finite or countably infinite number of values.
- The most common discrete variable is a *binary* (or *dummy*) variable, that takes on only one of two values (0 and 1).
- Examples include
  - *female*=1 for women; =0 for men;
  - *married*=1 for a married individual; =0 for a single individual;
  - *hispanic*=1 for a hispanic individual; =0 for for a non-hispanic individual;
  - *jobtraining*=1 for an individual who participated in a job training program; =0 for those who did not;
  - *constillage*=1 for a farmer employing conservation tillage; =0 for farmers using conventional tillage;

A Single Binary Independent Variable

- The simplest case is when we add a single binary variable to a simple regression model, with

$$y = \beta_0 + \delta_0 D + \beta_1 x + u \tag{1}$$

where D is our binary (or dummy) variable.

- The binary variable represents a *shift* in the intercept for those individuals in the group represented by $D = 1$.
- To see this, note that under the zero conditional mean assumption:

$$E(y|x, D = 0) = beta_0 \qquad + beta_1 x \tag{2}$$

while

$$E(y|x, D = 1) = beta_0 + \delta_0 + beta_1 x \tag{3}$$

- The group represented by $D = 0$ is typically referred to as the *base group*.

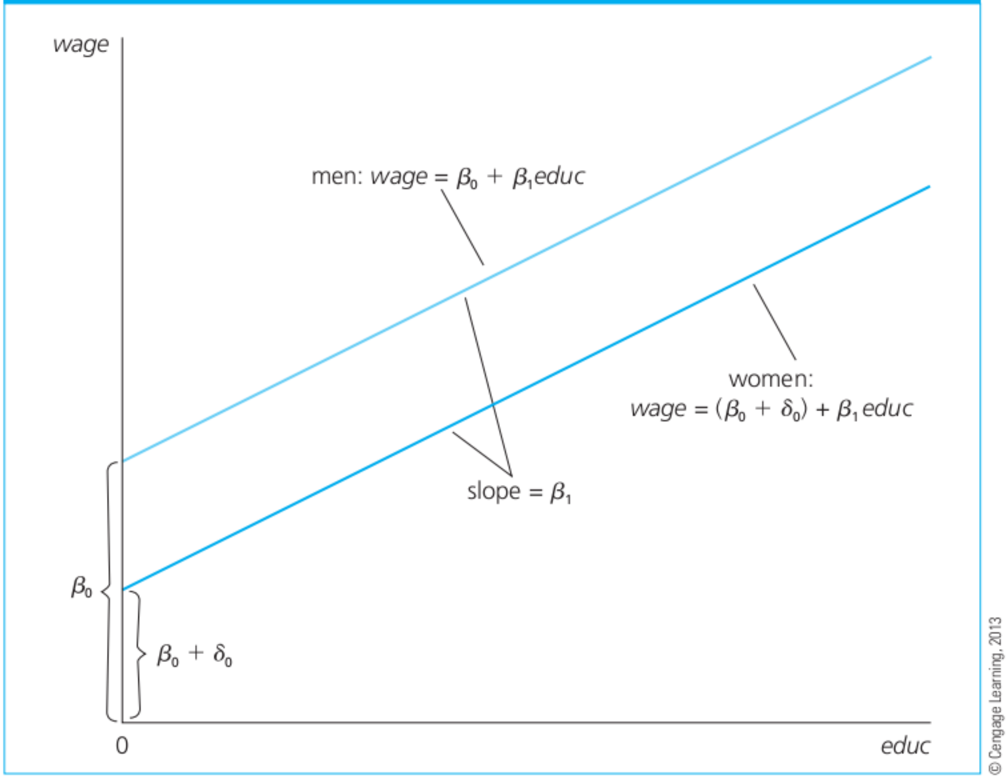
The Wage Model Example

- Wooldridge uses an example modeling wage, including a binary variable for females; i.e.,

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u \tag{4}$$

- Here we are modeling the impact of education on wages, allowing for difference between male and female wages.
- Males are the base group.
- The model allows for an overall gender effect on wages, but assumes the returns to education is the same for both men and women.
- In this case, we would expect that $\beta_1 > 0$ and $\delta_0 < 0$.

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0\, female + \beta_1\, educ$ for $\delta_0 < 0$.



Using the WAGE1 Data Set

```
. reg wage female educ

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(2, 523)       =     91.32
       Model |  1853.25304         2  926.626518   Prob > F        =    0.0000
    Residual |  5307.16125       523  10.1475359   R-squared       =    0.2588
-------------+----------------------------------   Adj R-squared   =    0.2560
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592     .605445
       _cons |   .6228168   .6725334     0.93   0.355     -.698382    1.944016
------------------------------------------------------------------------------
```

Controlling for Additional Variables

Wages

	Female	Add Educ	Add Exper & Tenure
female	-2.5118	-2.2734	-1.8109
	(0.3034)**	(0.2790)**	(0.2648)**
educ		0.5065	0.5715
		(0.0504)**	(0.0493)**
exper			0.0254
			(0.0116)*
tenure			0.1410
			(0.0212)**
_cons	7.0995	0.6228	-1.5679
	(0.2100)**	(0.6725)	(0.7246)*
R^2	0.12	0.26	0.36
N	526	526	526

$p < 0.05$, $p < 0.01$

Interpreting Coefficients on Binary Variables with $log(y)$

- In a model with:

$$log(y) = \beta_0 + \delta_0 D + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{5}$$

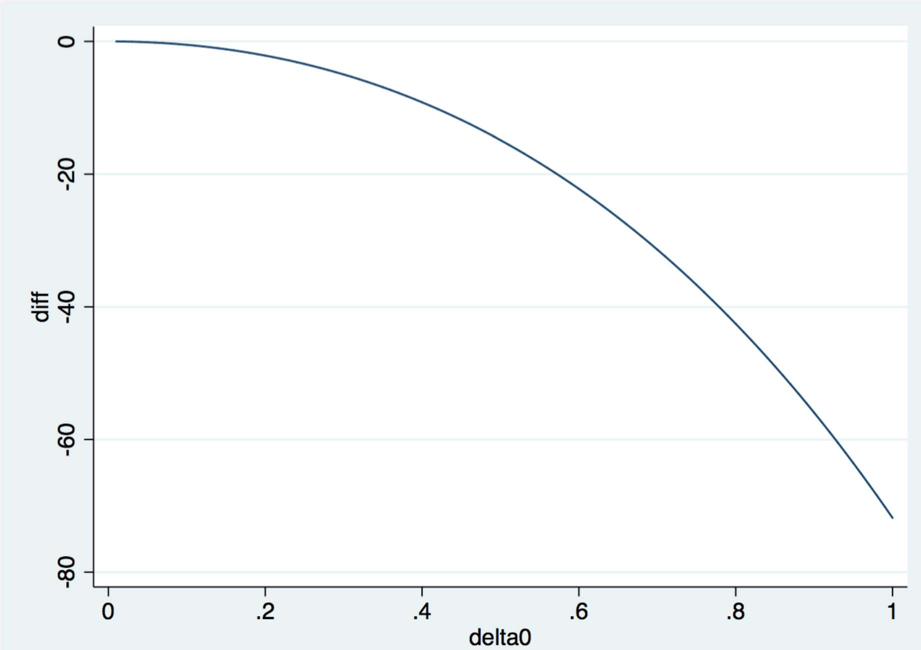
  $100 \cdot \delta_0$ is *roughly* interpreted as the percentage change in y given a change in D from $D = 0$ to $D = 1$, holding everything else fixed.
- The *exact* percentage change would be

$$100 \cdot [exp(\delta) - 1] \tag{6}$$

- The difference between the two will depend on the size of δ_0

Error

The following figure graphs $diff = \{100 \cdot \delta_0\} - \{100 \cdot [exp(\delta) - 1]\}$; indicating the bias in using the simple percentage change interpretation of $100 \cdot \delta_0$.



Dummy Variables for Catagorical/Ordinal Variables

- There are many settings in which explanatory variables are ordinal or categorical in nature;
  - Credit ratings for cities or individuals;
  - Water quality evaluations (e.g., the Water Quality Ladder: *boatable, fishable, swimmable, drinkable*);
  - Schooling (e.g., high school graduate, some college, college graduate, some graduate work, professional degree, etc.)
  - Qualitative survey questions with response categories such as *strongly agree, agree, disagree, strongly dissagree* and the like;
  - Qualitative evaluations (e.g., attractiveness: *homely, quite plain, average, good looking, strikingly beautiful or handsome*.)
- In these settings, it typically makes little sense to incorporate these variables directly in the model.

Wage Example

- Suppose we have the following variable on schooling:

$$schooling = \begin{cases} 1 & \text{grade school or less} \\ 2 & \text{grade school graduate} \\ 3 & \text{some high school} \\ 4 & \text{high school graduate} \\ 5 & \text{some college} \\ 6 & \text{college graduate} \\ 7 & \text{some graduate education} \end{cases} \quad (7)$$

- Consider simply including *schooling* in a simple regression model

$$ln(wages) = \beta_0 + \beta_1 schooling + u \quad (8)$$

- This model would assume that each step in schooling has the same impact of wages.

  ...The marginal effect of graduating grade school is the same as the marginal effect of graduating college


```
. reg lwage schooling

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(1, 524)       =    127.07
       Model |   28.948833         1   28.948833   Prob > F        =    0.0000
    Residual |  119.380929       524    .2278262   R-squared       =    0.1952
-------------+----------------------------------   Adj R-squared   =    0.1936
       Total |  148.329762       525   .28253288   Root MSE        =    .47731

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |   .1758321   .0155986    11.27   0.000     .1451887    .2064755
       _cons |   .8637806   .0705173    12.25   0.000     .7252492    1.002312
------------------------------------------------------------------------------
```

Each category of school increases wages by roughly 18 percent.

Binary Variable Categories

- A better approach would be to create a series of binary variables to represent each category; e.g.,

$$D_j = \begin{cases} 1 & schooling = j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

- Each category would then be included as a distinct binary variable in our regression model;

$$ln(wages) = \beta_0 + \delta_2 D_2 + \cdots + \delta_7 D_7 + u. \tag{10}$$

- β_0 represents the average log wage for those not completing grade school, while δ_j represents the increased wages *relative to those in group 1* from having completed education level j.


```
. reg lwage d2 d3 d4 d5 d6 d7
```

Source	SS	df	MS
Model	34.5640494	6	5.76067491
Residual	113.765712	519	.219201758
Total	148.329762	525	.28253288

Number of obs	= 526
F(6, 519)	= 26.28
Prob > F	= 0.0000
R-squared	= 0.2330
Adj R-squared	= 0.2242
Root MSE	= .46819

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
d2	.1836909	.1488005	1.23	0.218	-.1086344	.4760163
d3	-.0240041	.1227278	-0.20	0.845	-.2651084	.2171002
d4	.2614643	.1152604	2.27	0.024	.03503	.4878985
d5	.3772034	.118818	3.17	0.002	.1437801	.6106268
d6	.6789231	.1241025	5.47	0.000	.4351181	.922728
d7	.9812598	.1387404	7.07	0.000	.708698	1.253822
_cons	1.293997	.1103534	11.73	0.000	1.077203	1.510792

Why don't we include a binary variable for group 1 (i.e., D_1)?

Estimating Incremental Gains

- Having a comparison to group 1 can be useful, but we may instead be interested in the incremental gains from moving up a group.
- In this case we might form the group binary variable as follows:

$$\tilde{D}_j = \begin{cases} 1 & schooling \geq j \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

- Our regression model;

$$ln(wages) = \beta_0 + \delta_2 \tilde{D}_2 + \cdots + \delta_7 \tilde{D}_7 + u. \tag{12}$$

- In this case:

$$E[ln(wage)|D_j = 1] = \beta_0 + \delta_2 + \cdots + \delta_j \tag{13}$$

$$\Rightarrow E[ln(wage)|D_j = 1] - E[ln(wage)|D_{j-1} = 1] = \delta_j \tag{14}$$


```
. reg lwage td2 td3 td4 td5 td6 td7
```

Source	SS	df	MS			
Model	34.5640494	6	5.76067491	Number of obs	=	526
Residual	113.765712	519	.219201758	F(6, 519)	=	26.28
				Prob > F	=	0.0000
Total	148.329762	525	.28253288	R-squared	=	0.2330
				Adj R-squared	=	0.2242
				Root MSE	=	.46819

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
td2	.1836909	.1488005	1.23	0.218	-.1086344	.4760163
td3	-.207695	.1133488	-1.83	0.067	-.4303739	.0149838
td4	.2854684	.0631768	4.52	0.000	.1613546	.4095821
td5	.1157391	.0551989	2.10	0.036	.0072984	.2241799
td6	.3017196	.0718568	4.20	0.000	.1605538	.4428855
td7	.3023368	.1014622	2.98	0.003	.1030097	.5016639
_cons	1.293997	.1103534	11.73	0.000	1.077203	1.510792

Finishing degrees has a bigger impact than the longer incremental years.

Binning

- In the case of continuous variables, we often want to represent the impact of a variable in a flexible fashion.
- Including quadratic or cubic terms may not suffice.
- One solution is to include binary variables representing discrete ranges or "bins" of the variable.
- This is commonly done for weather variables, where we might include, say, average temperature in the form of discrete bins:

$$yields = \beta_0 + \delta_2 D_{T2} + \cdot + \delta_B D_{TB} + \text{other factors} + u \tag{15}$$

where B denotes the number of temperature bins and

$$D_{Tb} = \begin{cases} 1 & AvgTemp \in (T_b, T_{b+1}] \end{cases} \tag{16}$$

with T_b and T_{b+1} denoting the lower and upper bounds, respectively, of temperature bin.

Interacting Two Binary Variables

- We are often interesting in the combined effect of two binary variables,
- For example, we might be interested if there are differences in wage by both gender (*female*) and region (*south*), as well as in combination.

$$ln(wages) = \beta_0 + \delta_1 female + \delta_2 south + \delta_3 female \times south + u \tag{17}$$

- In this setting, and given the zero conditional mean assumption,

$$E[ln(wages)|female = 1] = \beta_0 + \delta_1 + \delta_2 south + \delta_3 south$$
$$E[ln(wages)|female = 0] = \beta_0 \qquad + \delta_2 south$$
$$\Rightarrow E[ln(wages)|female = 1] - E[ln(wages)|female = 0] = \delta_1 + \delta_3 south$$

- This allows us to test whether gender discrimination differs by region.

```
. reg lwage female south fsouth

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     31.67
       Model |  22.840836         3  7.61361201   Prob > F        =    0.0000
    Residual |  125.488926       522  .240400241   R-squared       =    0.1540
-------------+----------------------------------   Adj R-squared   =    0.1491
       Total |  148.329762       525   .28253288   Root MSE        =    .49031

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.4443875   .0532616    -8.34   0.000     -.549021    -.339754
       south |  -.1745113   .0611542    -2.85   0.004    -.2946499   -.0543728
      fsouth |   .1192185   .0896253     1.33   0.184    -.0568521    .2952891
       _cons |   1.879171   .0374947    50.12   0.000     1.805512     1.95283
------------------------------------------------------------------------------
```

In this case, there does not appear to be a regional difference in gender discrimination.

Interacting a Binary and Continuous Variable

- We many also want to interact binary and continuous variables.
- For example if we want to understand the returns to experience and whether these returns differ by gender

$$
\begin{aligned}
ln(wage) &= \beta_0 + \delta_0 female + \beta_1 exper + \delta_1 exper \times female + u \\
&= (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) \times exper + u \qquad (18)
\end{aligned}
$$

- δ_0 capture the shift in the intercept for females, while δ_1 capture the shift in the slope, or marginal impact of *exper* on $ln(wages)$.
- Also

$$
\begin{aligned}
&E[ln(wage)|female = 1] = \beta_0 + \delta_0 + (\beta_1 + \delta_1)exper \\
&E[ln(wage)|female = 0] = \beta_0 \qquad + \beta_1 exper \\
&\Rightarrow E[ln(wage)|female = 1] - E[ln(wage)|female = 0] = \delta_0 + \delta_1 exper
\end{aligned}
$$

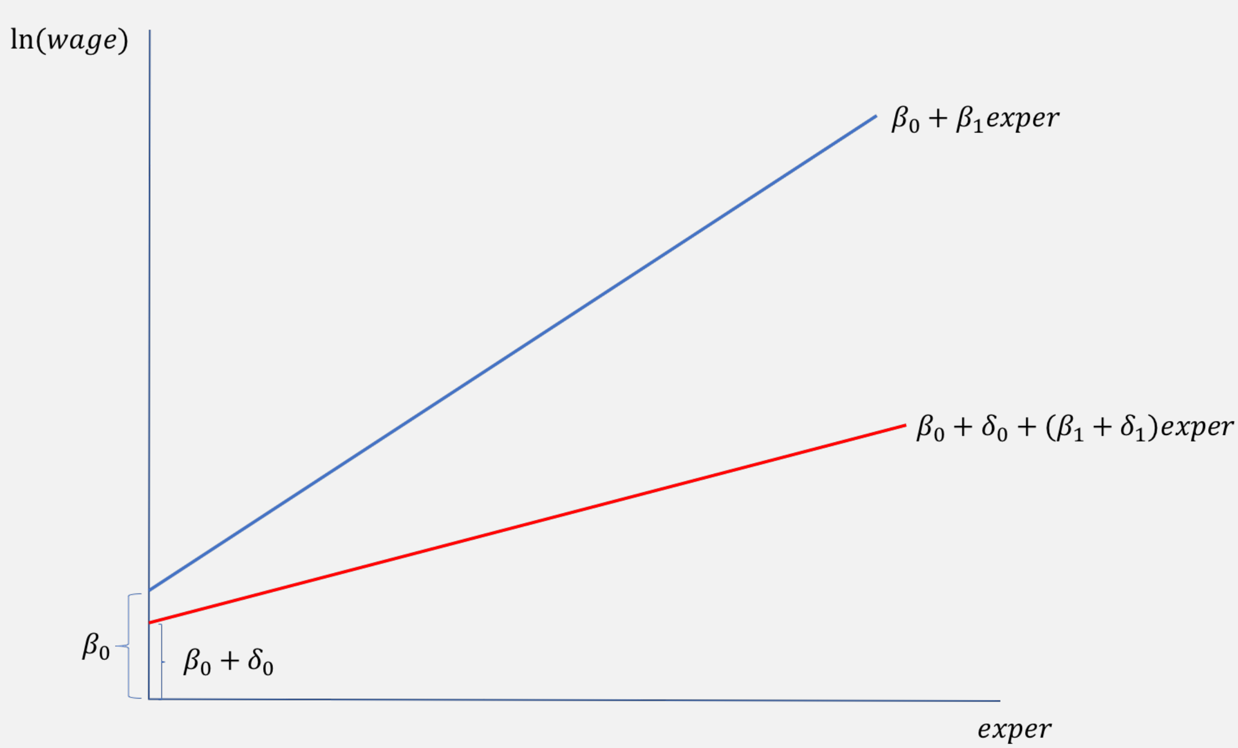
```
. gen femaleexper = female*exper

. reg lwage female exper femaleexper

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     31.78
       Model |  22.9051679         3  7.63505595   Prob > F        =    0.0000
    Residual |  125.424594       522     .240277   R-squared       =    0.1544
-------------+----------------------------------   Adj R-squared   =    0.1496
       Total |  148.329762       525   .28253288   Root MSE        =    .49018

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2934318   .0685958    -4.28   0.000    -.4281896    -.158674
       exper |   .0066007   .0021976     3.00   0.003     .0022835    .0109179
 femaleexper |  -.0058634   .0031567    -1.86   0.064    -.0120649     .000338
       _cons |   1.697672   .0486394    34.90   0.000     1.602119    1.793225
------------------------------------------------------------------------------
```

In this case, the gap between wages for men versus women appears to grow with experience.



Testing for Differences in Regression Functions Across Groups

- In the previous example, we might want to test whether the gender effects are statistically significant; i.e., whether or not the $ln(wage)$ regression equations differ for the two groups.
- One way to do this is to test the hypothesis $H_0 : \delta_0 = 0$ and $\delta_1 = 0$ versus the alternative hypothesis $H_A : H_0$is not true.
- Using our $F - test$ from chapter 4, we have

$$F = \frac{\frac{SSR_c - SSR_{uc}}{q}}{\frac{SSR_{uc}}{n-k-1}} = \frac{\frac{146.5-125.4}{2}}{\frac{125.4}{521}} = \frac{10.55}{0.24} = 44.0 \sim F_{2,521} \tag{19}$$

- The corresponding 1% critical level is 4.61, so we clearly reject this restriction.

The Chow Test

- An alternative way of constructing this $F - statistic$ is to obtain $SSR_{uc} = SSR_f + SSR_m$, where SSR_f denotes the SSR from regressing $ln(wage)$ on $exper$ for the $female = 1$ subpopulation and SSR_m denotes the SSR from regressing $ln(wage)$ on $exper$ for the $female = 0$ subpopulation.
- This will yield the same $F - stat$ and the same result.
- It's just a different way of constructing the components.
- The approach can readily be generalized to more than two groups.

```
. reg lwage exper if female==0

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(1, 272)       =      7.77
       Model |  2.16773686         1  2.16773686   Prob > F        =    0.0057
    Residual |  75.9164134       272  .279104461   R-squared       =    0.0278
-------------+----------------------------------   Adj R-squared   =    0.0242
       Total |  78.0841503       273  .286022529   Root MSE        =     .5283

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0066007   .0023685     2.79   0.006     .0019378    .0112636
       _cons |   1.697672   .0524223    32.38   0.000     1.594467    1.800877
------------------------------------------------------------------------------

. reg lwage exper if female==1

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(1, 250)       =      0.13
       Model |  .025431397         1  .025431397   Prob > F        =    0.7204
    Residual |  49.5081804       250  .198032722   R-squared       =    0.0005
-------------+----------------------------------   Adj R-squared   =   -0.0035
       Total |  49.5336118       251  .197345067   Root MSE        =    .44501

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0007373   .0020574     0.36   0.720    -.0033147    .0047893
       _cons |   1.404241   .0439119    31.98   0.000     1.317756    1.490725
------------------------------------------------------------------------------
```


Program Evaluation

- Researchers are often interested in evaluating the success of a policy, such as the impact of :
  - A voluntary rate program on energy usage;
  - A job training program on wages;
  - Summer school on student achievement;
  - Financial aid packages on students recruitment;
  - Military service on wages;
  - Food aid programs on health outcomes;
- In each of these cases, it is tempting to usage a simple model, with

$$y = \beta_0 + \beta_1 partic + u \tag{20}$$

  where *partic* denotes participation in the program in question.
- The key issue here is *self-selection*.

  . . . Individuals select to participate in the program and that selection process is likely correlated with unobserved factors represented by u, requiring special techniques to avoid biased estimates.

Binary Dependent Variables

- Thus far, we have focused on continuous dependent variables.
- We might be interesting in modeling binary outcomes, such as
  - whether or not an individual works outside the home;
  - whether or not a student graduates;
  - whether or not a home loan is approved;
  - whether or not an individual is arrested;
  - whether or not a farmer participates in a conservation program;
  - whether or not some one smokes.
- However, none of the theoretical results regarding OLS (unbiasedness, consistency, asymptotic normality, etc.) required y to be continuous.
- All the results that we have seen carry forward if y is discrete.
- Interpreting OLS results in these cases, however, requires additional care.

The Linear Probability Model (LPM)

- Consider the standard multiple regression model, but with y representing a binary variable:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{21}$$

- Under the zero conditional mean assumption

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{22}$$

- Interpreting the model in this case is helped by noting that

$$\begin{aligned} E(y|\mathbf{x}) &= 1 \cdot P(y=1|\mathbf{x}) + 0 \cdot P(y=1|\mathbf{x}) \\ &= P(y=1|\mathbf{x}) \end{aligned} \tag{23}$$

- Thus, in the case of a binary dependent variable, we can interpret our population regression model as implying a *Linear Probability Model (LPM)*; i.e.,

$$P(y=1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{24}$$

Interpreting the LPM

- In the LPM, the parameter β_j measures how much the probability of "success" (i.e., $Pr[y = 1|\boldsymbol{x}]$) changes when x_j changes, holding all other factors fixed.

$$\beta_j = \frac{\partial P(y = 1|\boldsymbol{x})}{\partial x_j} \tag{25}$$

- As was the case for a continuous y, we can include nonlinear functions of our regressors in the model, as well as qualitative and discrete independent variables.

Binary Dependent Variables: The Linear Probability Model

Example: Labor Force Participation

- Wooldridge examines labor force participation for women in 1975.
- A simpler version of his model would be:

$$inlf = \beta_0 + \beta_1 kidslt6 + \beta_2 kidsge6 + \beta_3 educ \tag{26}$$

- In this case

$$\beta_3 = \frac{\partial P(inlf = 1|\boldsymbol{x})}{\partial educ} = \frac{\Delta P(y = 1|\boldsymbol{x})}{\Delta educ} \tag{27}$$

denotes the change in the probability of labor force participation for each additional year of education.

- Similarly,

$$\beta_1 = \frac{\Delta P(inlf = 1|\boldsymbol{x})}{\Delta kidslt6} \tag{28}$$

denotes the change in the probability of labor force participation for each additional child less than 6.

```
. reg inlf kidslt6 kidsge6 educ

      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(3, 749)       =     25.14
       Model |  16.8988576         3  5.63295254   Prob > F        =    0.0000
    Residual |  167.828898       749  .224070625   R-squared       =    0.0915
-------------+----------------------------------   Adj R-squared   =    0.0878
       Total |  184.727756       752  .245648611   Root MSE        =    .47336

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kidslt6 |  -.2267395     .03328    -6.81   0.000    -.2920726   -.1614064
     kidsge6 |   .0114245   .0131559     0.87   0.385    -.0144024    .0372514
        educ |   .0467749   .0076333     6.13   0.000     .0317898    .0617601
       _cons |   .0321156   .0975068     0.33   0.742    -.1593037    .2235348
------------------------------------------------------------------------------
```


Results

- The results suggest that women with more education are more likely to enter the labor force, with each additional year increasing the probability by roughly 4.7 percentage points.
- Each additional child reduces the labor force participation probability by 22.7%.
- Older children have neither a statistically significant impact nor a substantive impact.
- The results illustrate one limitation of the LPM.

  . . . it assumes a constant marginal impact of each independent variable, whereas there may be nonlinear results.

  Also, note that:

$$\begin{aligned} &P(inlf = 1|kidslt6 = 3, kidsge6 = 0, educ = 8) \\ =&0.032 + (-0.227)3 + (0.011)0 + (0.032)8 = -0.393 \end{aligned} \tag{29}$$

  The LPM can yield probabilities outside the unit interval.

Binning Education

- The problem of constant marginal effects can be somewhat mitigated by using binning or quadratic variables.
- Consider in the case of education, subdividing education into two year increments/bins with

$$\begin{aligned}
D_{8,9} &= 1 \text{ for } 7 < educ \leq 9; = 0 \text{ otherwise} \\
D_{10,11} &= 1 \text{ for } 9 < educ \leq 11; = 0 \text{ otherwise} \\
D_{12,13} &= 1 \text{ for } 11 < educ \leq 13; = 0 \text{ otherwise} \\
D_{14,15} &= 1 \text{ for } 14 < educ \leq 15; = 0 \text{ otherwise} \\
D_{16+} &= 1 \text{ for } 15 < educ; = 0 \text{ otherwise}
\end{aligned}$$


```
. reg inlf kidslt6 kidsge6 d89 d1011 d1213 d1415 d16p

      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(7, 745)       =     10.46
       Model |  16.5300232         7  2.36143189   Prob > F        =    0.0000
    Residual |  168.197732       745  .225768768   R-squared       =    0.0895
-------------+----------------------------------   Adj R-squared   =    0.0809
       Total |  184.727756       752  .245648611   Root MSE        =    .47515

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kidslt6 |  -.2228721   .0334173    -6.67   0.000    -.2884753   -.1572689
     kidsge6 |   .0120482   .0132595     0.91   0.364    -.0139821    .0380786
         d89 |   .0568961   .1292979     0.44   0.660    -.1969355    .3107277
       d1011 |   .1505106   .1231388     1.22   0.222    -.0912297    .3922509
       d1213 |   .2243499   .1144714     1.96   0.050     -.000375    .4490748
       d1415 |   .2921095   .1266896     2.31   0.021     .0433984    .5408206
        d16p |   .4450079   .1214317     3.66   0.000     .2066188     .683397
       _cons |   .3708122   .1129393     3.28   0.001     .1490952    .5925293
------------------------------------------------------------------------------
```

Heteroskedasticity

- Another limitation of the LPM is that it necessarily violates the homoskedasticity assumption.
- In particular, one can show that

$$Var(y|\boldsymbol{x}) = P(y = 1|\boldsymbol{x})[1 - P(y = 1|\boldsymbol{x}] \tag{30}$$

- This doesn't mean the OLS estimator is suddenly biased or inconsistent, but traditional standard error calculations will be inconsistent and should be used with caution.

Interpreting Discrete Dependent Variables

- We've seen in interpreting a LPM, that we need to be careful in interpreting the implications of the model.
- The same is true when the dependent variable is discrete (such as the number of children in a family or the number of trips taken by a household.
- The marginal effect being measured are the marginal effects of the expected value of the dependent variable.

  . . . Just because the dependent variable is discrete it doesn't mean that these marginal effects have to be discrete.