Lecture 8
Qualitative Information: Dummy Variable

(Chapter 7 in the textbook)

Econ 322 Econometrics

Spring 2025

1. Describing Qualitative Information
2. A Single Dummy Variable
3. Dummy Variables for Multiple Categories
4. Interactions Among Dummy Variables
5. The Linear Probability Model

Table of Contents

1. Describing Qualitative Information
2. A Single Dummy Variable
3. Dummy Variables for Multiple Categories
4. Interactions Among Dummy Variables
5. The Linear Probability Model

Qualitative Information

- How do we describe binary qualitative information? A person is either male or female. A worker belongs to a union or does not. A firm offers a 401(k) pension plan or it does not.
- Can be captured by defining a **binary variable** (or **dummy variable**). Also called a **zero-one** variable to emphasize the two values it takes on.

Qualitative Information

- Must decide which outcome is assigned zero, which is one. Choose the variable name to be descriptive.
- For example, to indicate gender, *female*, which is one if the person is female, zero if the person is male, is a better name than *gender* or *sex* (unclear what *gender* $= 1$ corresponds to).

- From the data set GPA2.DTA

```
. list colgpa sat athlete female in 1/10
```

	colgpa	sat	athlete	female
1.	2.04	920	1	1
2.	4	1170	0	0
3.	1.78	810	1	0
4.	2.42	940	0	0
5.	2.61	1180	0	0
6.	3.03	980	0	1
7.	1.84	880	0	0
8.	3.05	980	0	0
9.	3	1240	0	0
10.	2	1230	0	0

- Of the first 10 people in the data set, two are women and only the first and the third students are athletes.

Qualitative Information

- For distinguishing different types, any two different values would do. But as we will see, zero and one are very convenient for use in regression analysis.
- If we start with two pieces of qualitative information – say, gender and marital status, as captured by *female* and *married* – we can define more than two categories. These would be married male (*marrmale*), married female (*marrfem*), single male (*singmale*), and single female (*singfem*).

Table of Contents

1 Describing Qualitative Information

2 A Single Dummy Variable

3 Dummy Variables for Multiple Categories

4 Interactions Among Dummy Variables

5 The Linear Probability Model

A Single Dummy

- What would it mean to specify a simple regression model where the explanatory variable is binary? Consider

$$wage = \beta_0 + \delta_0 female + u$$

where we assume SLR.4 holds:

$$E(u|female) = 0$$

- Therefore,

$$E(wage|female) = \beta_0 + \delta_0 female$$

A Single Dummy

- There are only two values of *female*, 0 and 1.

$$E(wage|female = 0) = \beta_0 + \delta_0 \cdot 0 = \beta_0$$
$$E(wage|female = 1) = \beta_0 + \delta_0 \cdot 1 = \beta_0 + \delta_0$$

  In other words, the average *wage* for men is β_0 and the average *wage* for women is $\beta_0 + \delta_0$.
- We can write

$$\delta_0 = E(wage|female = 1) - E(wage|female = 0)$$

  as the difference in average *wage* between women and men.
- So δ_0 is not really a slope. It is just a difference in average outcomes between the two groups.

A Single Dummy

- The population relationship is mimicked in the simple regression estimates.

$$\hat{\beta}_0 = \overline{wage}_m$$
$$\hat{\beta}_0 + \hat{\delta}_0 = \overline{wage}_f$$
$$\hat{\delta}_0 = \overline{wage}_f - \overline{wage}_m$$

where $\overline{wage}_m$ is the average wage for men in the sample and $\overline{wage}_f$ is the average wage for women in the sample.

Example: Wage1.dta

```
. sum wage female exper

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        526    5.896103    3.693086        .53      24.98
      female |        526    .4790875     .500038          0          1
       exper |        526    17.01711    13.57216          1         51

. * The mean (average) of a binary variable is the fraction of ones in
. * the sample. So the fraction of females is about .479, or 47.9%
. * Or, we can use the "tabulate" command.

. tab female

      =1 if |
     female |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        274       52.09       52.09
          1 |        252       47.91      100.00
------------+-----------------------------------
      Total |        526      100.00
```

Example: Wage1.dta

```
. reg wage female, r

Linear regression                               Number of obs     =        526
                                                F(1, 524)         =      71.23
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1157
                                                Root MSE          =     3.4763

------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   -2.51183   .2976209    -8.44   0.000    -3.096507   -1.927154
       _cons |   7.099489   .2513859    28.24   0.000     6.605641    7.593337
------------------------------------------------------------------------------
```

- The estimated difference is very large. Women earn about \$2.51 less than men per hour, on average. Of course, there are some women who earn more than some men; this is a difference in averages.

A Single Dummy

- This simple regression allows us to do a simple **comparison of means test**. The null is

$$H_0 : \mu_f = \mu_m$$

where μ_f is the population average *wage* for women and μ_m is the population average *wage* for men.

- The t statistic and confidence interval are directly reported.

$$t_{female} = -8.44$$

which is a very strong rejection of H_0.

A Single Dummy

- The estimate $\hat{\delta}_0 = -2.51$ does not control for factors that should affect wage, such as workforce experience and schooling. If women have, on average, less experience and less schooling, that could explain the difference in average wages.
- If we just control for education, the model written in expected value form is

$$E(wage|female, educ) = \beta_0 + \delta_0 female + \beta_1 educ$$

where now δ_0 measures the gender difference when we hold fixed *educ*.

A Single Dummy

- Another way to write δ_0:

$$\delta_0 = E(wage|female, educ_0) - E(wage|male, educ_0)$$

  where $educ_0$ is any years of schooling that are the same for the woman and man.
- There is still a difference of about \$2.27, which is smaller than when *educ* is not controlled for, but still very large and very statistically significant.

A Single Dummy

```
. reg wage female educ, r

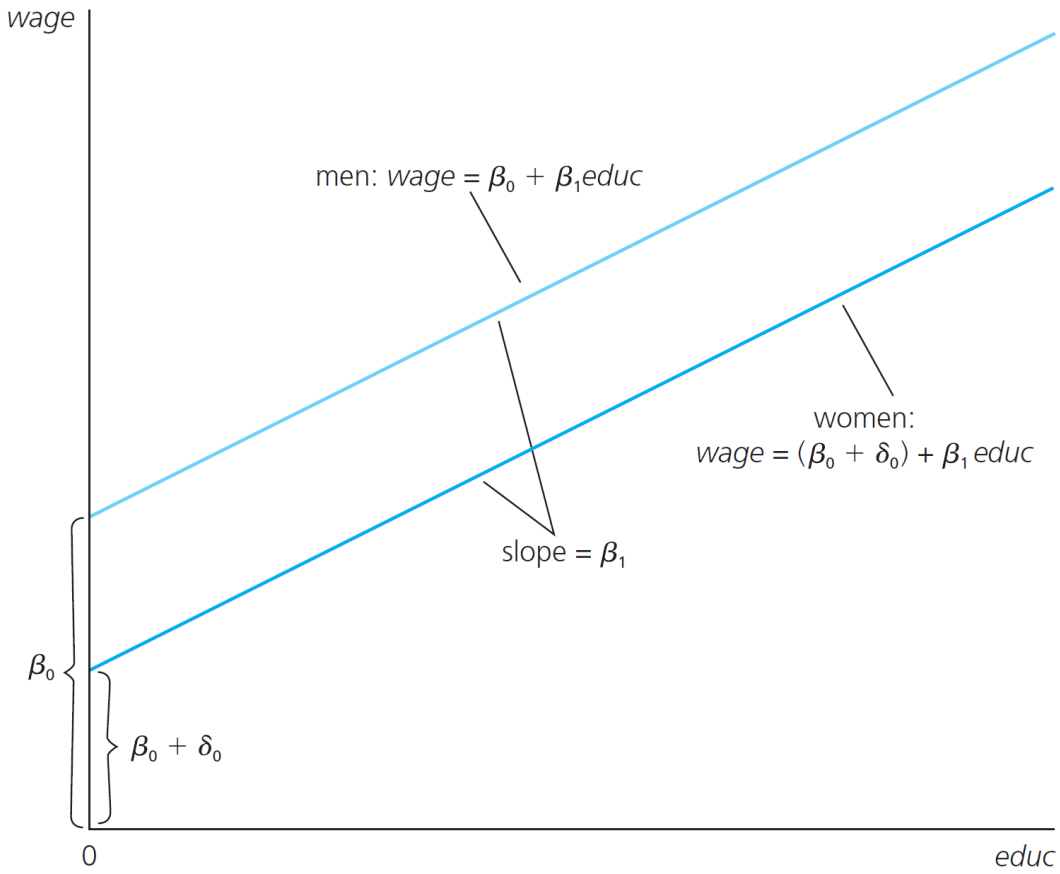
Linear regression                               Number of obs     =        526
                                                F(2, 523)         =      69.10
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2588
                                                Root MSE          =     3.1855

------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2702033    -8.41   0.000    -2.804179   -1.742545
        educ |   .5064521   .0598956     8.46   0.000     .3887867    .6241176
       _cons |   .6228168   .7286843     0.85   0.393    -.8086909    2.054324
------------------------------------------------------------------------------
```

- The model imposes a common slope on *educ* for men and women, β_1, estimated to be 0.506 in this example. It is only the intercepts that are allowed to differ.
- The estimated difference in average wages is the same at all years of schooling: \$2.27.

A Single Dummy

FIGURE 7.1 Graph of *wage* = $\beta_0 + \delta_0$ *female* + β_1 *educ* for $\delta_0 < 0$.



A Single Dummy

- Easy to add other variables (but picture is harder to draw). If we also control for levels of experience, the gap shrinks to about \$2.16, which is still large and very statistically significant.

```
. reg wage female educ exper, r

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      51.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3093
                                                Root MSE          =     3.0781

------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -2.155517   .2591294    -8.32   0.000    -2.664582   -1.646453
        educ |   .6025802   .0641895     9.39   0.000     .4764786    .7286817
       exper |   .0642417    .010044     6.40   0.000     .0445102    .0839733
       _cons |  -1.734481   .8577554    -2.02   0.044    -3.419558   -.0494042
------------------------------------------------------------------------------
```

A Single Dummy

- The previous regressions use males as the **base group** (or **benchmark group** or **reference group**). The coefficient -2.16 on *female* tells us how women do compared with men.
- Of course, we get the same answer if we use women as the base group, which means using a dummy variable for males rather than females.
- Because $male = 1 - female$, the coefficient on the dummy changes sign but must remain the same magnitude.
- The intercept changes because now the base (or reference) group is females.

A Single Dummy

```
. gen male = 1 - female

. reg wage male educ exper, r

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      51.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3093
                                                Root MSE          =     3.0781

------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        male |   2.155517   .2591294     8.32   0.000     1.646453    2.664582
        educ |   .6025802   .0641895     9.39   0.000     .4764786    .7286817
       exper |   .0642417    .010044     6.40   0.000     .0445102    .0839733
       _cons |  -3.889998   .8590559    -4.53   0.000     -5.57763   -2.202367
------------------------------------------------------------------------------
```

- We get just what we had before: the intercept for women is about -3.89 and that for men is $-3.89 + 2.15 \approx -1.74$.

A Single Dummy

- Putting *female* and *male* both in the equation is redundant. We have two groups so need only two intercepts. This will cause the perfect collinearity problem because male + female = 1.
- This is the simplest example of the so-called **dummy variable trap**, which results from putting in too many dummy variables to represent the given number of groups (two in this case). Because an intercept is estimated for the base group, we need only one dummy variable that distinguishes the two groups.

```
. reg wage female male educ exper, r
note: male omitted because of collinearity.

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      51.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3093
                                                Root MSE          =     3.0781

------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -2.155517   .2591294    -8.32   0.000    -2.664582   -1.646453
        male |           0  (omitted)
        educ |   .6025802   .0641895     9.39   0.000     .4764786    .7286817
       exper |   .0642417    .010044     6.40   0.000     .0445102    .0839733
       _cons |  -1.734481   .8577554    -2.02   0.044    -3.419558   -.0494042
------------------------------------------------------------------------------
```

Interpreting Dummy Variable Coefficients with log(y) as the Dependent Variable

- Use the data in "wage1.dta" again but now with $lwage = \log(wage)$ as the dependent variable:

$$\widehat{lwage} = \underset{(.030)}{1.8136} - \underset{(.043)}{.3972}female$$

$$n = 526, R^2 = .14$$

- A rough estimate is that in this sample, the average wage for women is below that of men by 39.7%.

```
. reg lwage female, r

Linear regression                               Number of obs     =        526
                                                F(1, 524)         =      86.36
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1396
                                                Root MSE          =      .4935

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -.3972175   .0427437    -9.29   0.000    -.4811875   -.3132474
       _cons |    1.81357   .0323116    56.13   0.000     1.750094    1.877046
------------------------------------------------------------------------------

. reg lwage female exper educ, r

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      80.28
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3526
                                                Root MSE          =      .42893

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -.3435967   .0374432    -9.18   0.000    -.4171546   -.2700389
       exper |   .0094139   .0014749     6.38   0.000     .0065164    .0123113
        educ |   .0912897     .00809    11.28   0.000     .0753968    .1071827
       _cons |   .4808357   .1174382     4.09   0.000     .2501262    .7115452
------------------------------------------------------------------------------
```

A Single Dummy

- The gap shrinks, but is still substantial, if we control for workforce experience (measured in years) and amount of education (measured in years). With both controls, women on average earns about $exp(-.344) = 70.9\%$ of what men earn. That is, at any given levels of experience and education, a woman is predicted to earn about 29.1% less than a man.

Table of Contents

1 Describing Qualitative Information

2 A Single Dummy Variable

3 Dummy Variables for Multiple Categories

4 Interactions Among Dummy Variables

5 The Linear Probability Model

Multiple Categories

- Suppose in the wage example we have two qualitative variables, gender and maritial status. Call these *female* and *married*.
- We can define four exhaustive and mutually exclusive groups. These are married males (*marrmale*), married females (*marrfem*), single males (*singmale*), and single females (*singfem*).
- Note that we can define each of these dummy variables in terms of *female* and *married*:

$$marrmale = married \cdot (1 - female)$$
$$marrfem = married \cdot female$$
$$singmale = (1 - married) \cdot (1 - female)$$
$$singfem = (1 - married) \cdot female$$

Multiple Categories

- We can allow each of the four groups to have a different intercept by choosing a base group and then including dummies for the other three groups.
- So, if we choose single males as the base group, we include *marrmale*, *marrfem*, and *singfem* in the regression. The coefficients on these variabels are relative to single men.
- With *lwage* as the dependent variable, we can give them a percentage change interpretation.
- Use WAGE1.DTA. Control for education, experience (as a quadratic), and tenure (as a quadratic).

```
. gen marrmale = married*(1 - female)

. gen marrfem = married*female

. gen singmale = (1 - married)*(1 - female)

. gen singfem = (1 - married)*female

. reg lwage marrmale marrfem singfem educ exper expersq tenure tenursq, r

Linear regression                               Number of obs     =        526
                                                F(8, 517)         =      51.70
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4609
                                                Root MSE          =     .39329

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    marrmale |   .2126757   .0571419     3.72   0.000     .1004168    .3249346
     marrfem |  -.1982676     .05877    -3.37   0.001     -.313725   -.0828102
     singfem |  -.1103502   .0571163    -1.93   0.054    -.2225587    .0018583
        educ |   .0789103   .0074147    10.64   0.000     .0643437    .0934769
       exper |   .0268006   .0051391     5.22   0.000     .0167044    .0368967
     expersq |  -.0005352   .0001063    -5.03   0.000    -.0007442   -.0003263
      tenure |   .0290875   .0069409     4.19   0.000     .0154516    .0427234
     tenursq |  -.0005331   .0002437    -2.19   0.029    -.0010119   -.0000544
       _cons |   .3213781    .109469     2.94   0.003     .1063193    .5364368
------------------------------------------------------------------------------
```

Multiple Categories

$$\widehat{lwage} = \underset{(.109)}{.321} + \underset{(.057)}{.213}\,marrmale - \underset{(.059)}{.198}\,marrfem - \underset{(.057)}{.110}\,singfem$$
$$+ \underset{(.007)}{.079}\,educ + \underset{(.005)}{.027}\,exper - \underset{(.00011)}{.00054}\,exper^2$$
$$+ \underset{(.007)}{.029}\,tenure - \underset{(.00024)}{.00053}\,tenure^2$$
$$n = 526, R^2 = .461$$

Using the usual approximation based on differences in logarithms – and holding fixed education, experience, and tenure – a married man is estimated to earn about 21.3% more than a single man. Remember, this compares two men with the same level of schooling, general workforce experience, and tenure with the current employer.

Multiple Categories

- This *marriage premium* for men has long been noted by labor economists. Does marriage make men more productive? Is being married a signal to employers (say, of stability and reliability)? Is there a selection issue in that more productive men are likely to be married, on average? The regression cannot tell us which explanation is correct.
- A married woman, at given levels of the other variables, earns about 19.8% *less* than a single man. A single woman earns about 11.0% less than a comparable single man. (Stata reports $t_{singfem} = -1.93$, so marginally statistically significant at the 5% level with p-value = .054.)

Multiple Categories

- What if we want to compare married women and single women? Just plug in the correct set of zeros and ones.

$$\begin{aligned}\text{intercept for married women} &= .321 - .198\\ \text{intercept for single women} &= .321 - .110\\ \text{difference} &= -.198 - (-.110) = -.088\end{aligned}$$

  so married women earn about 8.8% less than single women (controlling for other factors).
- We cannot tell from the previous output whether this difference is statistically significant.
- Note how the intercept for single men gets differenced away.

- Two approaches: (1) Use the `lincom` or `test` command in Stata. (2) Choose, say, married females as the base group and reestimate the model (including the dummies *marrmale*, *singmale*, and *singfem*.)
- The t statistic for the estimated difference $-.088$ is -1.68, which is significant at the 10% level (but not much lower than that).
- Unlike for men, the marriage "premium" for women is either nonexistent or in fact negative. (8.8% is not a small economic effect. It is a bit more than one year of schooling.)

```
. lincom marrfem - singfem

 ( 1)  marrfem - singfem = 0

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |  -.0879174   .0522744    -1.68   0.093    -.1906138     .014779
------------------------------------------------------------------------------

. test marrfem = singfem

 ( 1)  marrfem - singfem = 0

       F(  1,   517) =    2.83
            Prob > F =    0.0932
```

```
. reg lwage marrmale singmale singfem educ exper expersq tenure tenursq, r

Linear regression                               Number of obs     =        526
                                                F(8, 517)         =      51.70
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4609
                                                Root MSE          =     .39329

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    marrmale |   .4109433   .0454289     9.05   0.000     .3216952    .5001913
    singmale |   .1982676     .05877     3.37   0.001     .0828102     .313725
     singfem |   .0879174   .0522744     1.68   0.093     -.014779    .1906138
        educ |   .0789103   .0074147    10.64   0.000     .0643437    .0934769
       exper |   .0268006   .0051391     5.22   0.000     .0167044    .0368967
     expersq |  -.0005352   .0001063    -5.03   0.000    -.0007442   -.0003263
      tenure |   .0290875   .0069409     4.19   0.000     .0154516    .0427234
     tenursq |  -.0005331   .0002437    -2.19   0.029    -.0010119   -.0000544
       _cons |   .1231105   .1131843     1.09   0.277    -.0992471    .3454681
------------------------------------------------------------------------------
```

- No matter which group we choose as the base group, we include only three of the four dummy variables. If we include all four we fall into the dummy variable trap.

Using Dummy Variables to Incorprate Ordinal Information

- The data set BEAUTY.DTA includes a ranking of physical attractiveness of each man or woman, on a scale of 1 to 5, with 5 being "strikingly beautiful or handsome."
- As we move up the scale from 1 to 5, why should a one-unit increase mean the same amount of "beauty"?
- The "looks" variable is what we call an **ordinal variable**: we know that the order of outcomes conveys information (5 is better than 4, and 2 is better than 1) but we do not know that the difference between 5 and 4 is the same as 2 and 1.

```
. des wage looks belavg abvavg exper expersq female educ

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
wage            float   %9.0g                 hourly wage
looks           byte    %8.0g                 from 1 to 5
belavg          byte    %8.0g                 =1 if looks <= 2
abvavg          byte    %8.0g                 =1 if looks >=4
exper           byte    %8.0g                 years of workforce experience
expersq         int     %8.0g                 exper^2
female          byte    %8.0g                 =1 if female
educ            byte    %8.0g                 years of schooling

. tab looks

from 1 to 5 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         13        1.03        1.03
          2 |        142       11.27       12.30
          3 |        722       57.30       69.60
          4 |        364       28.89       98.49
          5 |         19        1.51      100.00
------------+-----------------------------------
      Total |      1,260      100.00
```

Ordinal Information

- In fact, very few people are at the extreme values 1 and 5 (less than 1% each). It makes sense to combine into three categories: below average (*belavg*), average, and above average (*abvavg*).
- Take average as the base group.
- 12.3% of people are "below average," 30.4% are "above average," and everyone else (57.3%) has $looks = 3$ (labeled "average").

```
. tab belavg

=1 if looks |
       <= 2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,105       87.70       87.70
          1 |        155       12.30      100.00
------------+-----------------------------------
      Total |      1,260      100.00

. tab abvavg

=1 if looks |
        >=4 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        877       69.60       69.60
          1 |        383       30.40      100.00
------------+-----------------------------------
      Total |      1,260      100.00
```

Ordinal Information

- With "average" as the base group, we include *belavg* and *abvavg* in a regression:

```
. reg lwage belavg abvavg, r

Linear regression                               Number of obs     =      1,260
                                                F(2, 1257)        =       8.22
                                                Prob > F          =     0.0003
                                                R-squared         =     0.0125
                                                Root MSE          =     .59124

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      belavg |  -.2087896   .0515497    -4.05   0.000    -.3099225   -.1076566
      abvavg |  -.0454376   .0376333    -1.21   0.228    -.1192687    .0283935
       _cons |   1.698296   .0219465    77.38   0.000      1.65524    1.741351
------------------------------------------------------------------------------
```

Ordinal Information

- Controlling for no other factors, those with below average looks earn about 20.88% less than those with average looks. The t statistic is very significant (p-value is zero to three decimal places).
- Those with above average looks are estimated to earn about 4.54% less than those with average looks, but the p-value is .228. So there is little evidence the effect is different from zero.

Ordinal Information

- Now control for some other factors, including gender and education.
- The effect of having below average looks is now about 15% lower salary (on average). Above average looks is still statistically insignificant and gets smaller in magnitude.
- Good practice to look at all coefficients to see if the signs and magnitudes make sense. They do, although the premium for males is very large.

Ordinal Information

```
. reg lwage belavg abvavg educ exper expersq female, r

Linear regression                               Number of obs     =      1,260
                                                F(6, 1253)        =     120.99
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3598
                                                Root MSE          =     .47683

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      belavg |  -.1542032   .0410068    -3.76   0.000    -.2346528   -.0737536
      abvavg |  -.0066465   .0312705    -0.21   0.832    -.0679948    .0547018
        educ |   .0663221   .0055299    11.99   0.000     .0554732     .077171
       exper |   .0408305   .0042387     9.63   0.000     .0325147    .0491463
     expersq |  -.0006301   .0000946    -6.66   0.000    -.0008156   -.0004445
      female |  -.4532832   .0292106   -15.52   0.000    -.5105904   -.3959761
       _cons |    .558981   .0814393     6.86   0.000     .3992086    .7187534
------------------------------------------------------------------------------
```

Ordinal Information

- Putting in the variable *looks* means that better looks always has to have a positive effect.

```
. reg lwage looks educ exper expersq female, r

Linear regression                               Number of obs     =      1,260
                                                F(5, 1254)        =     141.91
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3568
                                                Root MSE          =     .47774

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       looks |   .0562663   .0202904     2.77   0.006     .0164594    .0960733
        educ |   .0659397   .0055077    11.97   0.000     .0551344     .076745
       exper |    .041089   .0042456     9.68   0.000     .0327598    .0494183
     expersq |  -.0006287   .0000945    -6.65   0.000    -.0008141   -.0004434
      female |  -.4558771   .0291488   -15.64   0.000    -.5130629   -.3986914
       _cons |   .3591104   .1022669     3.51   0.000     .1584773    .5597434
------------------------------------------------------------------------------
```

Ordinal Information

- Equation for women only is suggestive of a positive effect of above average looks, but the estimates are imprecise with only 436 observations. *belavg* is only marginally statistically significant at 5% even though its coefficient is still pretty large, −.1257.

```
. reg lwage belavg abvavg educ exper expersq if female, r

Linear regression                               Number of obs     =        436
                                                F(5, 430)         =      25.85
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2126
                                                Root MSE          =       .468

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      belavg |  -.1257102   .0642751    -1.96   0.051    -.2520427    .0006224
      abvavg |   .0433352   .0517728     0.84   0.403     -.058424    .1450944
        educ |   .0786452   .0085834     9.16   0.000     .0617746    .0955158
       exper |   .0298313   .0072465     4.12   0.000     .0155882    .0440743
     expersq |  -.0004843   .0001952    -2.48   0.014     -.000868   -.0001005
       _cons |   .0407305   .1198284     0.34   0.734    -.1947918    .2762527
------------------------------------------------------------------------------
```

Ordinal Information

- One shortcoming in the previous analysis is that it ignores occupation. Maybe we should allow people to sort into occupation (perhaps partly based on looks) and see if there is a "looks premium" in a given occupation. Biddle and Hamermesh (1998, *Journal of Labor Economics*) study lawyers' looks and earnings and find similar results.
- Variables such as credit ratings, or any variables asked on a scale, are ordered variables. For example, someone may be assigned a credit rating on a scale from 1 to 7, or someone may be asked to rate their "happiness" on a scale of 1 to 5.

Table of Contents

1 Describing Qualitative Information

2 A Single Dummy Variable

3 Dummy Variables for Multiple Categories

4 Interactions Among Dummy Variables

5 The Linear Probability Model

Interactions Between Two Binary Variables

- Consider a regression model that includes two binary variables D_1 and D_2

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

- In this specification
  - β_1 is the effect on Y of changing D_1 from 0 to 1 (keeping D_2 constant)
  - β_2 is the effect on Y of changing D_2 from 0 to 1 (keeping D_1 constant)
- Here the effect of the binary variables is not a function of the value of the other binary variable

Interactions Between Two Binary Variables

- Consider the following regression function with an interaction term included

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{1i} \times D_{2i} + u_i$$

- How do we interpret the coefficients?
- The general rule is that we need to compare all the cases
  1. Expected value of Y when $D_1 = 0$

$$E(Y|D_1 = 0, D_2 = d_2) = \beta_0 + \beta_2 d_2$$

  2. Expected value of Y when $D_1 = 1$

$$E(Y|D_1 = 1, D_2 = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$$

Interactions Between Two Binary Variables

- Thus expected change in Y when D_1 changes from 0 to 1 is

$$\beta_1 + \beta_3 d_2.$$

- Now we see that the effect of D_1 on Y is a function of the value of D_2
- We can therefore interpret β_3 as the increment to the effect of D_1 on Y when $D_2 = 1$

Interactions

- Consider again the *lwage* equation with gender and marital status, allowing for the four different groups. We can achieve the same thing by using an interaction between *female* and *married*.
- If we regress *lwage* on *female*, *married*, *female* · *married* and the education, experience, and tenure variables, we get

$$\widehat{lwage} = \underset{(.109)}{.321} - \underset{(.057)}{.110}\, female + \underset{(.057)}{.213}\, married - \underset{(.072)}{.301}\, female \cdot married + \ldots$$

$$n = 526, R^2 = .461$$

Interactions

- Just plug in the correct combination of zeros and ones to obtain differences among the groups. The intercept corresponds to $female = 0$, $married = 0$, so it is still for single men. For single women, $female = 1$, $married = 0$, and we are left with $.321 - .110 = .211$. We saw earlier that $-.110$ was the difference between single women and single men.
- The intercept for married women is the sum of all coefficients (preserving negative sign):

$$.321 - .110 + .213 - .301 = .123$$

or .088 below that for single women.

Interactions

- One advantage of the equation with the interaction is that it allows us to read off the difference in marriage premium between women and men. For men, the premium is the coefficient on *married*, .213. The difference in the marriage premium is $-.301$.
- Is the difference statistically significant?
- The marriage premium for women is $.213 - .301 = -.088$, as we have seen before.

$$\widehat{lwage} = \underset{(.109)}{.321} - \underset{(.057)}{.110}\, female + \underset{(.057)}{.213}\, married - \underset{(.072)}{.301}\, female \cdot married + \ldots$$

Allowing for Different Slopes

- We can interact dummy variables with quantitative variables to obtain regression models with different slopes, as well as different intercepts.
- Recall that in the model

$$lwage = \beta_0 + \delta_0 female + \beta_1 exper + u,$$

the intercept for men is β_0 and that for women is $\beta_0 + \delta_0$. The slope, β_1, is common across men and women.

- How can we allow the return to experience – that is, the slope in the *lwage* equation – to differ by gender?

- An extended model is

$$lwage = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) \cdot exper + u$$

- For men, $female = 0$. For women, plug in $female = 1$.

	Intercept	Slope
Male	β_0	β_1
Female	$\beta_0 + \delta_0$	$\beta_1 + \delta_1$
Difference ($Female - Male$)	δ_0	δ_1

- For now, we use the Greek letter delta to emphasize that δ_0 and δ_1 are differences.
- How can we estimate the four parameters, $\beta_0, \beta_1, \delta_0$, and δ_1? Write the model as

$$lwage = \beta_0 + \delta_0 female + \beta_1 exper + \delta_1 female \cdot exper + u$$

  so we simply add the interaction term $female \cdot exper$.
- Note that $female \cdot exper$ is zero whenever $female = 0$ (that is, for all men).
- Again use the dataset "wage1"

```
. gen femexper = female*exper

. list female exper femexper in 1/10

     +---------------------------+
     | female   exper   femexper |
     |---------------------------|
  1. |      1       2          2 |
  2. |      1      22         22 |
  3. |      0       2          0 |
  4. |      0      44          0 |
  5. |      0       7          0 |
     |---------------------------|
  6. |      0       9          0 |
  7. |      0      15          0 |
  8. |      1       5          5 |
  9. |      1      26         26 |
 10. |      0      22          0 |
     +---------------------------+
```

```
. reg lwage female exper femexper, r

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      29.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1544
                                                Root MSE          =     .49018

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -.2934319   .0667979    -4.39   0.000    -.4246576   -.1622061
       exper |   .0066007   .0025547     2.58   0.010      .001582    .0116194
    femexper |  -.0058634   .0031461    -1.86   0.063     -.012044    .0003171
       _cons |   1.697672   .0498113    34.08   0.000     1.599817    1.795528
------------------------------------------------------------------------------
```

$$\widehat{lwage} = \underset{(.050)}{1.6977} - \underset{(.067)}{.2934}\,female + \underset{(.0026)}{.0066}\,exper - \underset{(.003)}{.0059}\,female \cdot exper$$

$$n = 526, R^2 = .154$$

- The intercept for men is 1.6977 and the slope is .0066 – about 0.66% for each year of experience.
- The intercept for women is $1.6977 - .2934 = 1.4043$ and the slope is $.0066 - .0059 = .0007$ – about 0.07% for each year of experience.
- The interaction term is marginally statistically significant, with p-value = .063. (So at the 10% level but not the 5%.)

- Must use care to interpret the coefficient on *female* when *female* · *exper* is in the equation. To see why, at any level of experience, the predicted difference in *lwage* between females and males is

$$-0.2934 - .0059\mathit{exper}$$

- Therefore, the coefficient on *female*, $-.2934$, is the predicted difference in *lwage* between a woman and man when $\mathit{exper} = 0$. This is not an especially interesting subset of the population. (No one can have zero years of work experience in this sample.)
- More interesting is the gap at around the mean, say $\mathit{exper} = 17$:

$$-0.2934 - .0059(17) = -.3937$$

or about 39.3% less for women.

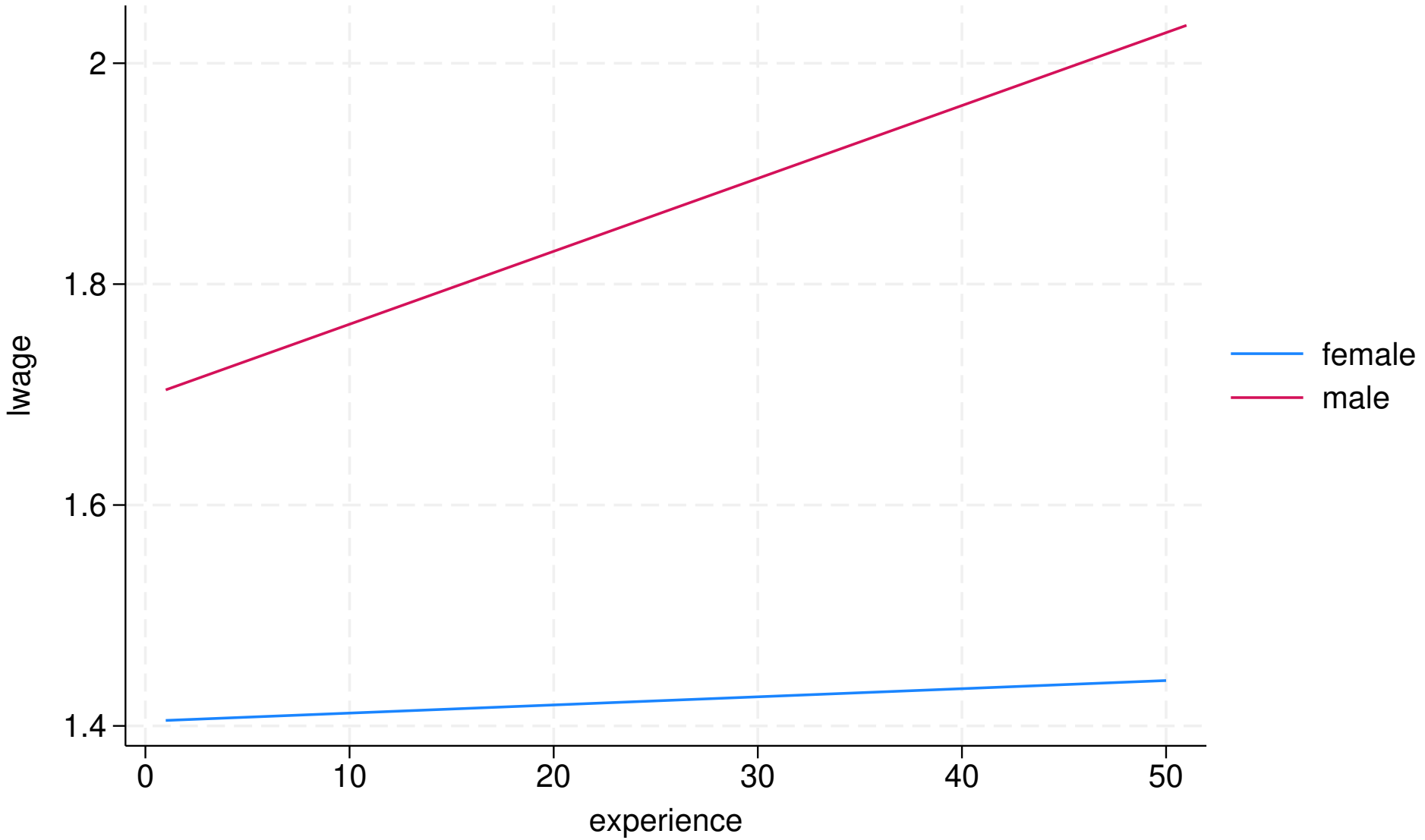
- We can use the same centering scheme as before with interactions. Replace $female \cdot exper$ with $female \cdot (exper - 17)$ so that the coefficient on *female* becomes the difference at 17 years of experience, where 17 is close to the mean value of experience in the sample.

```
. gen femexper_17 = female*(exper - 17)

. reg lwage female exper femexper_17, r

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      29.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1544
                                                Root MSE          =     .49018

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -.3931104   .0421931    -9.32   0.000    -.4759995   -.3102213
       exper |   .0066007   .0025547     2.58   0.010      .001582    .0116194
 femexper_17 |  -.0058634   .0031461    -1.86   0.063     -.012044    .0003171
       _cons |   1.697672   .0498113    34.08   0.000     1.599817    1.795528
------------------------------------------------------------------------------
```


lwage
2
1.8
1.6
1.4
0
10
20
30
40
50
experience
female
male

- Suppose we add the education variable, and also the interaction $female \cdot educ$. Now we see that the coefficient on *female* is much smaller in magnitude and statistically insignificant.
- The coefficient on *female* is now the estimated difference between women and men when $exper = 0$ and $educ = 0$ – not an interesting group, and it is unsurprising we cannot estimate this effect precisely.

```
. gen femeduc = female*educ

. reg lwage female exper femexper educ femeduc, r

Linear regression                               Number of obs     =        526
                                                F(5, 520)         =      50.03
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3698
                                                Root MSE          =     .42398

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   .0850418   .2347587     0.36   0.717    -.3761502    .5462338
       exper |   .0149081    .002355     6.33   0.000     .0102817    .0195345
    femexper |  -.0107923   .0029244    -3.69   0.000    -.0165374   -.0050472
        educ |   .1018128   .0099031    10.28   0.000     .0823577    .1212678
     femeduc |  -.0194987   .0170673    -1.14   0.254    -.0530281    .0140307
       _cons |   .2497944   .1507844     1.66   0.098    -.0464271     .546016
------------------------------------------------------------------------------
```

- We can test

$$H_0\text{: no difference between women and men}$$

by testing the three variables *female*, *femexper*, and *femeduc* jointly. The rejection is very strong despite the two insignificant t statistics!

```
. test female femexper femeduc

 ( 1)  female = 0
 ( 2)  femexper = 0
 ( 3)  femeduc = 0

       F(  3,   520) =   32.06
            Prob > F =    0.0000
```

- A final warning: it is hard to justify omitting the level of a variable but including an interaction that involves that variable. Suppose we drop *educ* but include $female \cdot educ$.
- The model imposes a zero return to education for men – which we cannot justify – and the coefficient on *femeduc* is now a direct estimate of the return to education for women, rather than being the difference in the returns between women and men

Table of Contents

1 Describing Qualitative Information

2 A Single Dummy Variable

3 Dummy Variables for Multiple Categories

4 Interactions Among Dummy Variables

5 The Linear Probability Model

The Linear Probability Model

- We have studied many ways that binary (dummy) explanatory variables can appear in regression analysis. Question: What if the *dependent* variable is a dummy variable?
- That is, we want to explain the outcome of a no/yes or zero/one event.
- For example, suppose we want to study the question of married women's labor force participation. Or, we want to know whether a young man is arrested for a crime during a certain period of time.
- In these cases, the variable y we want to explain is a binary or dummy variable.

The Linear Probability Model

- How do we intepret the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

when y is binary? y can only change from 0 to 1 or 1 to zero. Suppose $\beta_1 = .035$ and $x_1 = educ$. What does it mean for a one year increase in *educ* to increase y by .035?

- Same problem arises for other discrete variables, such as $y = number\ of\ arrests$ or $y = number\ of\ children$. Cannot have a fraction more of a child.

The Linear Probability Model

- Instead, rely on the expected value formulation of linear regression. Under MLR.4,

$$E(y|\mathrm{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

where x is shorthand for $(x_1, x_2, \ldots, x_k)$. Remember, we can interpret β_j as

$$\Delta E(y|\mathrm{x}) = \beta_j \Delta x_j \text{ holding other explanatory variables fixed}$$

- Key relationship when y is binary:

$$E(y|\mathrm{x}) = P(y = 1|\mathrm{x})$$

The Linear Probability Model

- We call $P(y = 1|\mathrm{x})$ the **response probability**. Thus, when we apply the linear model to binary y we are really saying

$$P(y = 1|\mathrm{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

and we call the model the **linear probability model (LPM)**. (Other **binary response models** are covered in Chapter 17 of the textbook.)

The Linear Probability Model

- The important point is that all partial effects are effects on the probability that $y = 1$ (sometimes called a "success"). Since $P(y = 0|\text{x}) = 1 - P(y = 1|\text{x})$, it is the only probability we need.

$$\Delta P(y = 1|\text{x}) = \beta_j \Delta x_j \text{ holding other explanatory variables fixed}$$

The Linear Probability Model

- The sample analog holds as well. When we have the OLS regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k,$$

  $\hat{y}$ is now the predicted probability.
- The intercept is the predicted probability when each x_j is set to zero (which, as with other regression applications, may not make sense).

The Linear Probability Model

- $\hat{\beta}_j$ measures the change in the estimated probability of a "success" when $\Delta x_j = 1$, other factors held fixed.
- Just like in any regression application, we can have explanatory variables in logs, quadratics, and interactions as well as binary regressors.

The Linear Probability Model

EXAMPLE: Married Women's Labor Force Participation (MROZ.DTA)

- The variable *inlf* is one if a woman worked for wage during a certain year, and zero if not. We estimate a linear probability model to see the effects of variables on the probability of being in the labor force.

```
. des inlf nwifeinc educ exper age kidslt6 kidsge6

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
inlf            byte    %9.0g                 =1 if in lab frce, 1975
nwifeinc        float   %9.0g                 (faminc - wage*hours)/1000
educ            byte    %9.0g                 years of schooling
exper           byte    %9.0g                 actual labor mkt exper
age             byte    %9.0g                 woman's age in yrs
kidslt6         byte    %9.0g                 # kids < 6 years
kidsge6         byte    %9.0g                 # kids 6-18
```

```
. sum inlf nwifeinc educ exper age kidslt6 kidsge6

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        inlf |        753    .5683931    .4956295          0          1
    nwifeinc |        753    20.12896     11.6348  -.0290575         96
        educ |        753    12.28685    2.280246          5         17
       exper |        753    10.63081     8.06913          0         45
         age |        753    42.53785    8.072574         30         60
-------------+---------------------------------------------------------
     kidslt6 |        753    .2377158     .523959          0          3
     kidsge6 |        753    1.353254    1.319874          0          8
```

```
. reg inlf nwifeinc educ exper expersq age kidslt6 kidsge6, r

Linear regression                               Number of obs     =        753
                                                F(7, 745)         =      62.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2642
                                                Root MSE          =     .42713

------------------------------------------------------------------------------
             |               Robust
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0034052   .0015249    -2.23   0.026    -.0063988   -.0004115
        educ |   .0379953    .007266     5.23   0.000      .023731    .0522596
       exper |   .0394924     .00581     6.80   0.000     .0280864    .0508983
     expersq |  -.0005963     .00019    -3.14   0.002    -.0009693   -.0002233
         age |  -.0160908    .002399    -6.71   0.000    -.0208004   -.0113812
     kidslt6 |  -.2618105   .0317832    -8.24   0.000    -.3242058   -.1994152
     kidsge6 |   .0130122   .0135329     0.96   0.337     -.013555    .0395795
       _cons |   .5855192   .1522599     3.85   0.000     .2866098    .8844287
------------------------------------------------------------------------------
```

The Linear Probability Model

- The estimated equation is

$$\widehat{inlf} = \underset{(.152)}{.586} - \underset{(.0015)}{.0034}\,nwifeinc + \underset{(.007)}{.038}\,educ + \underset{(.0058)}{.039}\,exper - \underset{(.00019)}{.00060}\,exper^2$$
$$- \underset{(.002)}{.016}\,age - \underset{(.032)}{.262}\,kidslt6 + \underset{(.0135)}{.013}\,kidsge6$$
$$n = 753,\ R^2 = .264$$

The Linear Probability Model

- The intercept is not of interest here. The coefficient on *nwifeinc* (other sources of income) shows a modest effect: if it increases by 20 ($20,000, about two standard deviations), the probability of being in the labor force falls by .068, or 6.8 **percentage points**. The t statistic shows it is statistically significant at the 3% level.
- Each year of education increases the probability by an estimated .038, or 3.8 **percentage points**.

The Linear Probability Model

- Past workforce experience has a positive but diminishing effect. The effect of the first year is about .039, and this diminishes to zero at $exper = .039/(2 \cdot .0006) = 32.5$. (Only 13 women have experience $>$ 32.)
- Having young children has a very large negative effect: being in the labor force falls by .262 for each young child. Almost all of the action is between 0 and 1, with a little from 1 to 2.

The Linear Probability Model

- It is unwise to extrapolate to extreme values when using any linear model, including this one.

```
. tab kidslt6

 # kids < 6 |
      years |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        606       80.48       80.48
          1 |        118       15.67       96.15
          2 |         26        3.45       99.60
          3 |          3        0.40      100.00
------------+-----------------------------------
      Total |        753      100.00
```

The Linear Probability Model: Shortcomings

- Using a linear model for a binary outcome is convenient because estimation is easy and so is interpretation.
- But the LPM does have some shortcomings.

The Linear Probability Model: Shortcomings

1. The fitted values from an OLS regression are never guaranteed to be between zero and one, yet these fitted values are estimated probabilities.

This can be slightly embarrassing but is rarely a big deal. We usually use the LPM to estimate partial effects, not to make predictions.

The Linear Probability Model

```
. qui reg inlf nwifeinc educ exper expersq age kidslt6 kidsge6

. predict inlfh
(option xb assumed; fitted values)

. sum inlfh

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       inlfh |        753    .5683931    .2547633  -.3451103   1.127151

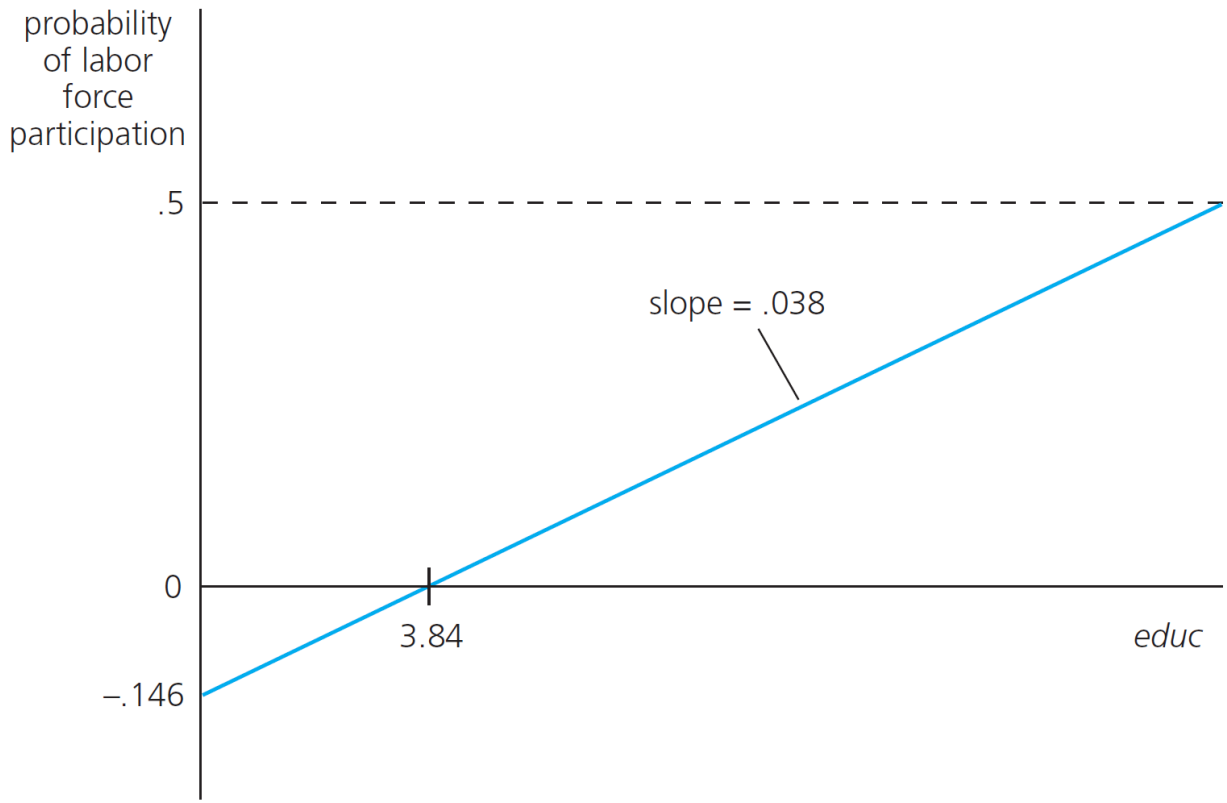
. count if inlfh < 0
  16

. count if inlfh > 1
  17
```

- Only 33 of 753 fitted values are not in [0, 1].

The Linear Probability Model: Shortcomings

FIGURE 7.3 Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



The Linear Probability Model: Shortcomings

2. The estimated partial effects are constant throughout the range of the explanatory variables, possibly leading to silly estimated effects for large changes. (This is related to predicted probabilities possibly being negative or greater than one.)

This is more of a problem because we know that, say, a variable with a positive effect on $P(y = 1|\mathrm{x})$ must eventually have a diminishing effect. But the linear model implies a constant effect (when the variable appears by itself).

The Linear Probability Model: Shortcomings

- For example, take a woman who has no other source of income, 25 years of prior work experience, no children, who is 48 years old. As a function of *educ* the equation looks like

$$\widehat{inlf} = .417 + .038educ$$

At $educ = 12$, the predicted probability is .873, at $educ = 14$ it is .949, and at $educ = 16$, $\widehat{inlf} = 1.025$. For the estimated model to truly represent a probability, the effect of education should be diminishing – that is, the next year of education should increase the probability by less than the previous year so that the estimated probability never goes above one.

The Linear Probability Model: Shortcomings

- Using logarithms does not bound the effect, and using quadratics often does not help with this problem. (In this example, a quadratic in *educ* gives an estimated *increasing* effect, not a diminishing effect.)
- But the LPM does a good job of approximating partial effects if we do not look at extreme values of the explanatory variables.

The Linear Probability Model: Shortcomings

3. Because y is binary – and this really has nothing to do with the LPM per se – the LPM must exhibit heteroskedasticity except in the one case where no x_j affects $P(y = 1|\mathrm{x})$. This follows because for a binary variable,

$$Var(y|\mathrm{x}) = p(\mathrm{x})[1 - p(\mathrm{x})]$$

where $p(\mathrm{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ is the linear response probability.

This is a case where we *know* MLR.5 must fail, and we know how.

Logit Regression

- The problem with the linear probability model is that it models the conditional probability of $Y = 1$ as being linear:

$$P(Y = 1|X) = \beta_0 + \beta_1 X$$

- Instead, we want:
  - $0 \leq P(Y = 1|X) \leq 1$ for all X
  - $P(Y = 1|X)$ to be increasing in X (for $\beta_1 > 0$)
- This requires a nonlinear functional form for the probability.
- How about an "S-curve"

Logit Regression

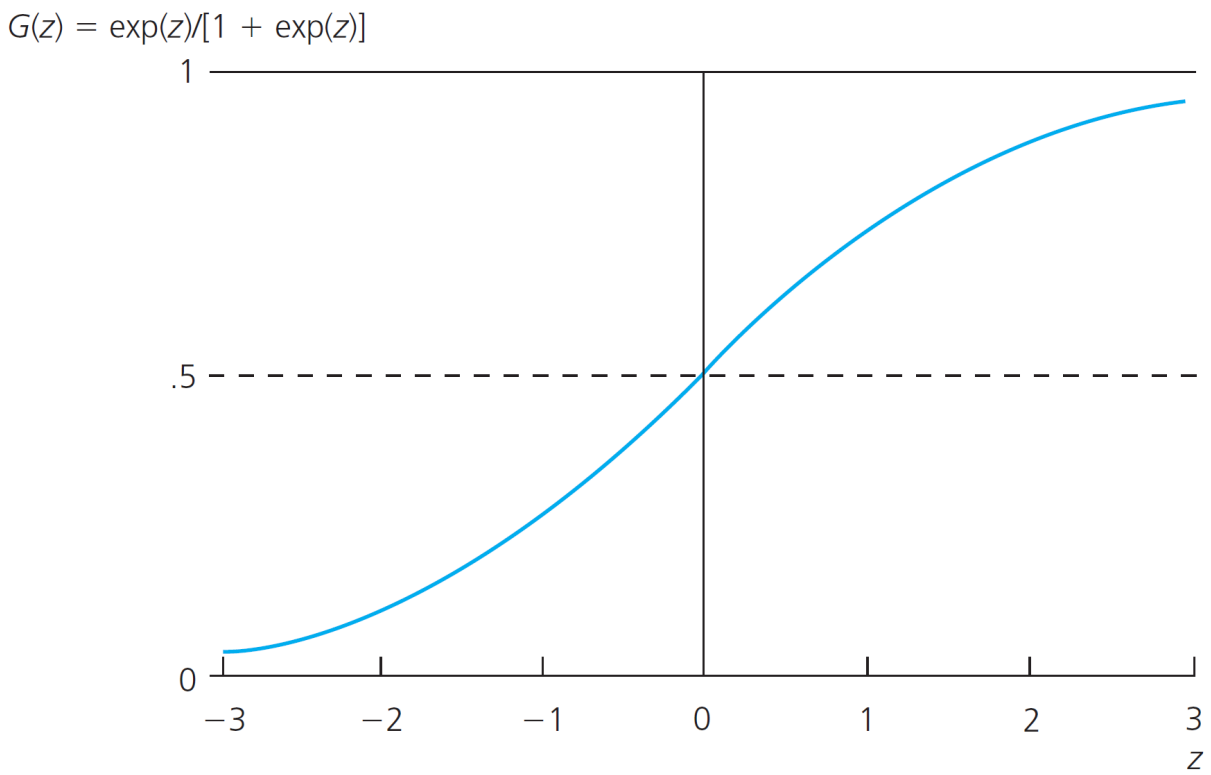
To avoid the LPM limitations, consider a class of binary response models of the form

$$P(Y = 1|X) = G(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

where G is a function taking on values strictly between zero and one: $0 < G(z) < 1$, for all real numbers z. This ensures that the estimated response probabilities are strictly between zero and one.

Logit Regression

FIGURE 17.1 Graph of the logistic function $G(z) = \exp(z)/[1 + \exp(z)]$.



Logit Regression

- The interpretation of the logit regression is more involved with calculus.
- Hence, it is omitted in this lecture note (and the final exam). If you are interested, refer to Chapter 17 in the textbook.