Lecture 4: Regression specification

- Dummy / indicator variables
- Interactions with indicator variables
- Interactions with "continuous" variables
- Functional form: log-linear regressions
- Applications in Assignment #2
- Chapter 8 in S&W

Regression with indicator variables

- Suppose we are interested in calculating the percent difference in earnings between males and females

- Y_i = ln weekly earnings of person i
- D_{1i} = **1**(if person i is female)

Note: in regressions with only indicator variables, the intercept has a clear interpretation

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

- β_0 = Average log weekly earnings of males

- β_1 = Difference in average log weekly earnings between females and males (≈ percent difference in weekly wages between females and males – see slides at end)

Algebra:

- β_0 = Average log weekly earnings of males
- Since $\beta_0 = E[Y_i|D_{1i}=0]$

- β_1 = Difference in average log weekly earnings between females and males
- Since $E[Y_i|D_{1i}=1] = \beta_0 + \beta_1$

- $\Rightarrow \beta_1 = E[Y_i|D_{1i}=1] - E[Y_i|D_{1i}=0]$

STATA application (CPS data)

```
. regress lwkearn female, robust;

Linear regression                                      Number of obs =    8454
                                                       F(  1,  8452) =  487.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0542
                                                       Root MSE      =  .54535

------------------------------------------------------------------------------
             |               Robust
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2613968    .0118448   -22.07   0.000    -.2846155   -.2381782
       _cons |   6.799713     .008433   806.32   0.000     6.783182    6.816244
------------------------------------------------------------------------------
```

This means that on average, women earn about 26% less than males (not conditional on other attributes, like job type, hours worked)

Current Population Survey (CPS)

- Monthly survey of about 60,000 households
- Administered by U.S. Census Bureau for the Bureau of Labor Statistics (BLS)

- Sample represents the civilian noninstitutional U.S. population

- The survey asks about the employment status of each member of the household 15 years of age or older in the reference week

- ⇒ Used to construct official unemployment rate series

- Also ask about demographics, education, wages and income (some months), industry, etc

Adding more indicator variables to the model

- D_{2i} = **1**(if person i has a college degree)
- $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$
- β_0 = Average log weekly earnings of males without a college degree
- β_1 = Percent difference in average weekly earnings between females and males, for those <u>without</u> a college degree
- β_2 = Percent difference in average weekly earnings between college graduates and non-college graduates, irrespective of gender

- <u>Algebra:</u>
- $\mu_{00} = E[Y|D_1=0, D_2=0] = \beta_0$ [omit i subscript]
- $\mu_{10} = E[Y|D_1=1, D_2=0] = \beta_0 + \beta_1$
- $\mu_{01} = E[Y|D_1=0, D_2=1] = \beta_0 + \beta_2$
- $\mu_{11} = E[Y|D_1=1, D_2=1] = \beta_0 + \beta_1 + \beta_2$
- So:
- $\beta_0 = \mu_{00}$
- $\beta_1 = \mu_{10} - \mu_{00}$
- $\beta_2 = \mu_{01} - \mu_{00} = \mu_{11} - \mu_{10}$

STATA application

```
.  regress lwkearn female college, robust;

Linear regression                                      Number of obs =    8454
                                                       F(  2,  8451) = 1251.87
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2318
                                                       Root MSE      =  .49154

------------------------------------------------------------------------------
             |               Robust
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2794053   .0106736   -26.18   0.000    -.3003283   -.2584824
     college |   .4965453   .0115014    43.17   0.000     .4739998    .5190909
       _cons |   6.635888   .0083724   792.59   0.000     6.619476      6.6523
------------------------------------------------------------------------------
```

This means that on average, non-college graduate women earn about 28% less than non-college graduate males, and the average return to college is about 50%

Adding an interaction to this model

- There is no reason to restrict the "return to college" is the same for males and females

- Add an interaction between D_{1i} and D_{2i} to the model:

- $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} * D_{2i}) + u_i$

- Called "saturated" or "fully-interacted" model (no other functions of D_1 and D_2 can be included in model)
  - Otherwise perfect multicollinearity

- Example of the "difference-in-difference" estimator (to come later

Interpretation

- β_0 = Average log weekly earnings of males without a college degree

- β_1 = Percent difference in average weekly earnings between females and males without a college degree

- β_2 = Percent difference in average weekly earnings of males with and without a college degree (i.e., “male college wage premium”)

- β_3 = Female-male difference in the return to college

- Algebra:
- $\mu_{00} = E[Y|D_1=0, D_2=0] = \beta_0$ [omit i subscript]
- $\mu_{10} = E[Y|D_1=1, D_2=0] = \beta_0 + \beta_1$
- $\mu_{01} = E[Y|D_1=0, D_2=1] = \beta_0 + \beta_2$
- $\mu_{11} = E[Y|D_1=1, D_2=1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$
- So:
- $\beta_0 = \mu_{00}$
- $\beta_1 = \mu_{10} - \mu_{00}$
- $\beta_2 = \mu_{01} - \mu_{00}$
- $\beta_3 = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$

STATA application

```
. regress lwkearn female college fcollege, robust;

Linear regression                                      Number of obs =    8454
                                                       F(  3,  8450) =  838.19
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2319
                                                       Root MSE      =  .49153

------------------------------------------------------------------------------
             |               Robust
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2702025    .012817   -21.08   0.000     -.295327    -.245078
     college |   .5095113   .0165479    30.79   0.000     .4770735    .5419492
    fcollege |  -.0264149   .0229823    -1.15   0.250    -.0714659    .0186361
       _cons |    6.63161   .0091209   727.08   0.000     6.613731    6.649489
------------------------------------------------------------------------------
```

This means that on average, non-college graduate women earn about 27% less than non-college graduate males, that the average return to college for males is about 51%, and that there is no statistically significance F-M difference in the return to college

Interactions between indicator variables and 'continuous' variables

- Consider a slightly different model:

- Y_i = log weekly earnings of person i
- D_{1i} = **1**(if person i is female)
- X_{1i} = years of education of person i

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

- $\beta_0 = E[Y_i | D_{1i}=0, X_{1i}]$

- $\beta_1 = E[Y_i | D_{1i}=1, X_{1i}] - E[Y_i | D_{1i}=0, X_{1i}]$

- β_2 = percent increase in weekly earnings associated with an 1 additional year of education ("return to education"), restricted to be the same for males and females

STATA application

```
.  regress lwkearn female yrseduc, robust;

Linear regression                                      Number of obs =    8454
                                                       F(  2,  8451) = 1460.34
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2731
                                                       Root MSE      =  .47814

------------------------------------------------------------------------------
             |               Robust
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2953517   .0104344   -28.31   0.000    -.3158056   -.2748978
     yrseduc |   .1032773   .0021292    48.51   0.000     .0991036     .107451
       _cons |   5.390107   .0293252   183.80   0.000     5.332623    5.447592
------------------------------------------------------------------------------
```

This means that on average, women earn about 30% less than males, and that the average return to an additional year of education is about 10%

Allowing the return to education to vary by gender

- $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 X_{1i} + \beta_3 (D_{1i} * X_{1i}) + u_i$

- $\beta_0 = E[Y_i | D_{1i} = 0, X_{1i}]$

- $\beta_1 = E[Y | D_{1i} = 1, X_{1i}] - E[Y | D_{1i} = 0, X_{1i}]$

- β_2 = male return to education

- β_3 = female-male difference in the return to education

STATA application

```
. regress lwkearn female yrseduc fyrseduc, robust;

Linear regression                                      Number of obs =    8454
                                                       F(  3,  8450) =  982.28
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2732
                                                       Root MSE      =  .47814

------------------------------------------------------------------------------
             |               Robust
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2415737   .0597732    -4.04   0.000    -.3587438   -.1244036
     yrseduc |   .1050051   .0027211    38.59   0.000      .099671    .1103392
    fyrseduc |  -.0038881   .0043246    -0.90   0.369    -.0123653    .0045891
       _cons |   5.366525   .0369454   145.26   0.000     5.294103    5.438947
------------------------------------------------------------------------------
```

Key Concept 8.4

Interactions Between Binary and Continuous Variables

Through the use of the interaction term $X_i \times D_i$, the population regression line relating Y_i and the continuous variable X_i can have a slope that depends on the binary variable D_i. There are three possibilities:

1. Different intercept, same slope (Figure 8.8a):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i;$$

2. Different intercept and slope (Figure 8.8b):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i;$$

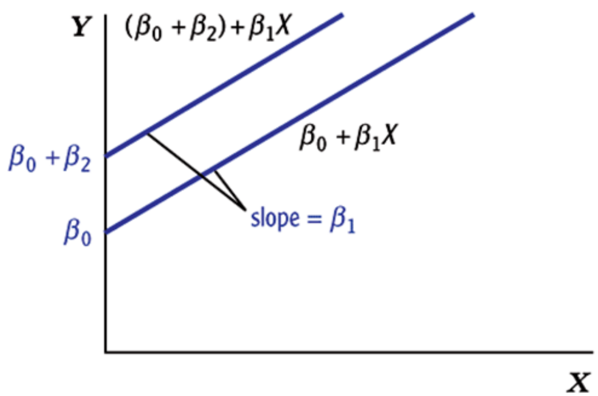
3. Same intercept, different slope (Figure 8.8c):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i.$$

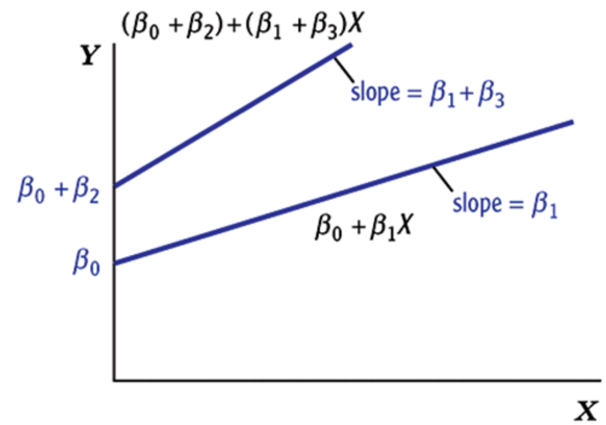
Note: "intercept" in this slide does not refer to the regression intercept (β_0) per se, but to level difference in the population regression function that comes from the group indicator D_i

The intercept for the group where D_i=0 is β_0, while the intercept for the group where D_i=1 is β_0+ β_2

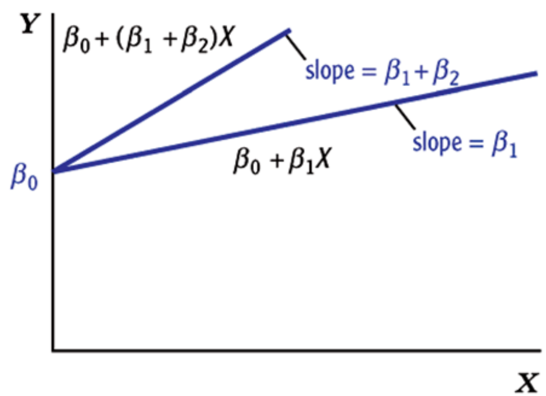
FIGURE 8.8 **Regression Functions Using Binary and Continuous Variables**



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope; (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ allows for different intercepts and different slopes; and (c) $\beta_0 + \beta_1 X + \beta_2 (X \times D)$ has the same intercept but allows for different slopes.

Interactions between two 'continuous' variables

- Y_i = log weekly earnings of person i
- X_{1i} = years of education of person i
- X_{2i} = age of person i

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} * X_{2i}) + u_i$

- The interaction allows the return to education to vary by age:

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 + \beta_3 X_{2i}$$

- This allows marginal effects to vary by value of X_{2i}

STATA application

```
. regress lwkearn age educ age_yrseduc, robust;

Linear regression                                      Number of obs =    8454
                                                       F(  3,  8450) =  709.70
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2068
                                                       Root MSE      =  .49951

------------------------------------------------------------------------------
             |               Robust
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0053146   .0034072     1.56   0.119    -.0013643    .0119935
        educ |   .1058259   .0113045     9.36   0.000     .0836664    .1279854
 age_yrseduc |  -.0001394   .0002455    -0.57   0.570    -.0006207    .0003419
       _cons |   5.061251   .1574626    32.14   0.000     4.752585    5.369916
------------------------------------------------------------------------------
```

For linear marginal effects in STATA use “lincom”. Here 44.92 is the average age in the sample, so this gives the return to education evaluated at the average age

```
. lincom educ + age_yrs*44.92

 ( 1)  educ + 44.92*age_yrseduc = 0

------------------------------------------------------------------------------
     lwkearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0995637   .0021612    46.07   0.000     .0953273    .1038002
------------------------------------------------------------------------------
```

(Natural) Logarithmic regression

- By far, this is the most frequently used "nonlinear" regression model

- Why:
- Logs convert changes into percentage change
- A log-log regression yields estimates of elasticities
- Often fits data better

- Always use the natural log: x=ln(exp(x)), where e = 2.71828

- I will write "log" but mean "ln" (also true with Stata)

Review: Properties of 'ln' function

- (a) log(1/x) = -log(x)
- (b) log(a*x) = a*log(x)
- (c) log(x/a) = log(x) – log(a)
- (d) ∂log(x)/ ∂x = 1/x

- $\Rightarrow$ Natural Log transformation models relations in "percentage" terms, rather than in natural units (linearly)

- *Here's why*: $\ln(x+\Delta x) - \ln(x) = \ln\left(1+\frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$

I. Linear-log regression

- $Y_i = \beta_0 + \beta_1 \log(X_{1i}) + u_i$

- $\beta_1 \equiv \partial Y_i / \partial \log(X_{1i})$

- $\Rightarrow \beta_1$ measures the unit change in Y arising from a proportionate change in X_{1i}

- Why:

$$\frac{\partial Y_i}{\partial \log(X_{1i})} = \beta_1 = \frac{\partial Y_i}{\partial X_{1i}} \frac{\partial X_{1i}}{\partial \log(X_{1i})} = X_{1i} \frac{\partial Y_i}{\partial X_{1i}} \boxed{\approx \frac{\Delta Y_i}{\Delta X_{1i}/X_{1i}}}$$

II. Log-linear regression

- $\log(Y_i) = \beta_0 + \beta_1 X_{1i} + u_i$
- $\beta_1 \equiv \partial \log(Y_i)/\partial X_{1i}$

- Measures the proportionate change in Y arising from a 1-unit change in X_{1i}

- Why:

$$\frac{\partial \log(Y_i)}{\partial X_{1i}} = \beta_1 = \frac{\partial \log(Y_i)}{\partial Y_i}\frac{\partial Y_i}{\partial X_{1i}} = \frac{1}{Y_i}\frac{\partial Y_i}{\partial X_{1i}} \boxed{\approx \frac{\Delta Y_i / Y_i}{\Delta X_{1i}}}$$

- In other words β_1 measures the percent effect on Y of a 1-unit change in X_1

- Classic example: Estimating the return to schooling

III. Log-log regression

- $\log(Y_i) = \beta_0 + \beta_1 \log(X_{1i}) + u_i$
- $\beta_1 \equiv \partial\log(Y_i)/\partial\log(X_{1i})$

- β_1 = proportionate change in Y_i arising from a proportionate change in X_{1i}

- Measures the <u>elasticity</u> of Y_i with respect to X_{1i}

- <u>Why</u>:

$$\frac{\partial\log(Y_i)}{\partial\log(X_{1i})} = \beta_1 = \frac{\partial\log(Y_i)}{\partial Y_i}\partial Y_i \frac{1}{\partial\log(X_{1i})}\frac{\partial X_{1i}}{\partial X_{1i}} = \frac{\partial Y_i}{Y_i}\frac{X_{1i}}{\partial X_{1i}} \boxed{\approx \frac{\Delta Y_i/Y_i}{\Delta X_{1i}/X_{1i}}}$$