# Announcements

- Problem Set 5 is posted, it will be graded and is due Friday, March 11 by 5pm
- Problem Set 5 is the last graded problem set, there will be one more ungraded problem set posted next week
- Midterm grades should be posted this evening (solutions are already posted)
- When they are posted, I will send out an email with a rough guide to letter grades

# Outline for the Remaining Lectures

- The last chapters we will cover are Chapters 9, 10 and 11
- The final exam will be cumulative, about 50% of the exam will be on Chapters 9-11, the other 50% will be on the earlier material
- Today we will finish Chapter 9 and start Chapter 10
- On Thursday we will finish Chapter 10 and start Chapter 11
- Next week we will focus on problems that you run into with data analysis and some solutions (Chapter 11 plus additional material not in the textbook)

# Testing Significance of a Subset of Regressors

Testing significance of a subset of regressors:

- Unrestricted model:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + \varepsilon$$

- Restricted model:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_g x_g + \varepsilon$$

# Testing Significance of a Subset of Regressors

$$H_o: \ \beta_{g+1} = 0, \beta_{g+2} = 0, ..., \beta_k = 0$$

$$H_a: \ \text{at least one of } \beta_{g+1}, ..., \beta_k \neq 0$$

$$F^* = \frac{R_u^2 - R_r^2}{1 - R_u^2} \frac{n-k}{k-g}$$

$$p = Pr(F_{k-g,n-k} > F^*) = FDIST(F^*, k-g, n-k)$$

$$c = F_{\alpha,k-g,n-k} = FINV(\alpha, k-g, n-k)$$

- Reject null hypothesis if $p < \alpha$ or $F^* > c$

# Testing Significance of a Subset of Regressors

- If our regression gives us p-values for each individual coefficient, why not just look at those to assess the null hypothesis that $\beta_{g+1}$ through $\beta_k$ are zero?
- It is possible for a subset of regressors to be jointly significant even though none of the regressors are individually significant
- One way this can happen is if two regressors are very highly correlated

# Testing Significance of a Subset of Regressors

- For an example, let's think about what determines a person's income beyond the standard stuff (years of schooling, work experience, etc.)
- We want to test whether including information on parents' education helps predict differences in income
- The problem is, father's and mother's education will be highly correlated (college grads tend to marry college grads)
- How will this affect our regression results? To Excel (joint-significance-simulation.xlsx ...)

# Testing Significance of a Subset of Regressors

- From our regression results, the p-values for the father's education and mother's education coefficients were 0.87 and 0.26, respectively
- So they are nowhere close to being statistically significant
- However, this doesn't mean that father's and mother's education doesn't matter
- The p-value for the joint signficance F test was $2.6 \cdot 10^{-84}$, so father's and mother's education are clearly jointly significant
- We can't get precise estimates of their individual coefficients, but we can say that taken together they matter quite a bit

# Multivariate Data Transformation

- Just as with bivariate data, sometimes we will need to use data transformations with multivariate data
- We can use all of the transformations we have already talked about:
  - Taking the natural log of the dependent variable
  - Taking the natural log of some or all of the regressors
  - Using polynomials for particular regressors
- We also have a couple of new possibilities
  - Multiple dummy variables
  - Interaction terms

# Logs and Multivariate Data

- We use logs with multivariate data for the same reasons as with bivariate data
- Changes in logs can be interpreted as percent changes (eg. elasticities)
- Logs help us deal with a variable for which different observations are on very different scales (eg. population, income)
- Logs can capture exponential growth (with log-linear models)
- It may make sense to take logs of just some variables or to take logs of all variables

# A Classic Example of a Multivariate Log-log Model

- Consider the widely used Cobb-Douglas production function:

$$y = AK^{\alpha}L^{\beta}$$

- Suppose we want to get estimates of $A$, $\alpha$ and $\beta$ using ordinary least squares
- We need to transform this into a linear model:

$$\ln y = \ln(AK^{\alpha}L^{\beta})$$

$$\ln y = \ln A + \ln K^{\alpha} + \ln L^{\beta}$$

$$\ln y = \ln A + \alpha \ln K + \beta \ln L$$

- So if we regress $\ln y$ on $\ln K$ and $\ln L$, the intercept will give us an estimate of $\ln A$ and the coefficients will give us estimates of $\alpha$ and $\beta$

# Polynomials and Multivariate Data

- Polynomials offer a very flexible way to fit nonlinear trends
- Recall the example of income and age (the U-shaped curve meant we should use a quadratic in age):

$$\ln wage_i = \beta_1 + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 edu_i + \varepsilon_i$$

- If we think that there is a nonlinear relationship between $y$ and a particular regressor $x_j$, we should consider including a polynomial in $x_j$ in our regression $(x_j, x_j^2, x_j^3, ...)$

# Dummy Variables and Multivariate Data

- We may want to use dummy variables to include categorical data in our regressions
- Recall that a dummy variable is either zero or one depending on the value of a particular categorical variable (eg. male equals one, female equals zero)
- When we considered categorical variables with more than two values, we split the values into two groups so that we could use a binary dummy variable
- If we are willing to use several regressors, we have another option available to us: multiple dummy variables

# Using Multiple Dummy Variables

- Suppose we have a categorical variable for education (*edu*) that can take on any of the following values: some high school, high school graduate, some college, college graduate
- To include this variable in our regression, we can use several dummy variables
- Each dummy variable still needs to be either zero or one, for example the dummy variable for 'some high school' would be defined as:

$$d_{somehs} = 1 \text{ if } edu = \text{“some HS”}, 0 \text{ otherwise}$$

- We could define a dummy variable this way for each educational cateogory: $d_{somehs}, d_{hsgrad}, d_{somecol}, d_{colgrad}$

# Using Multiple Dummy Variables

| edu | d(somehs) | d(hsgrad) | d(somecol) | d(colgrad) |
|---|---|---|---|---|
| some college | 0 | 0 | 1 | 0 |
| high school graduate | 0 | 1 | 0 | 0 |
| college graduate | 0 | 0 | 0 | 1 |
| high school graduate | 0 | 1 | 0 | 0 |
| some high school | 1 | 0 | 0 | 0 |
| some college | 0 | 0 | 1 | 0 |
| college graduate | 0 | 0 | 0 | 1 |

# Using Multiple Dummy Variables

- Including these dummy variables as regressors will allow us to estimate average differences in the dependent variable across different groups
- For example, suppose we regress books read per year on our education dummies and age:

$$books = b_1 + b_2 age + b_3 d_{somehs} + b_4 d_{hsgrad} + b_5 d_{somecol}$$

- Notice that I didn't include $d_{colgrad}$, it is very important to exclude one of the dummy variables
- Why?

# Using Multiple Dummy Variables

- Suppose we had included $d_{colgrad}$:

$$books = b_1 + b_2 age + b_3 d_{somehs} + b_4 d_{hsgrad} +$$

$$b_5 d_{somecol} + b_6 d_{colgrad}$$

- But $d_{colgrad} = 1 - d_{somehs} - d_{hsgrad} - d_{somecol}$, so the above equation can be rewritten as:

$$books = (b_1 + b_6) + b_2 age + (b_3 - b_6) d_{somehs}$$

$$+(b_4 - b_6) d_{hsgrad} + (b_5 - b_6) d_{somecol}$$

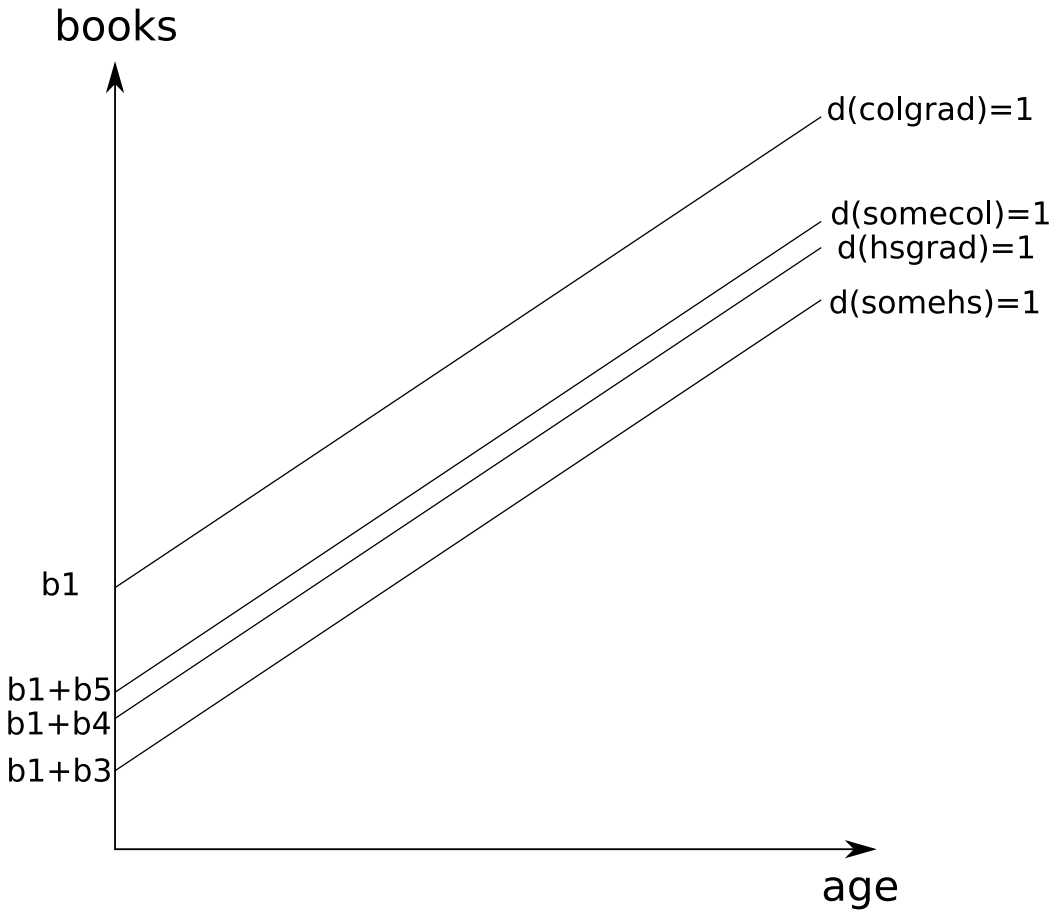
- There won't be a unique set of coefficients for the regression equation (it's like the multicollinearity problem we talked about before)
- To get around this problem, we need to drop one of the dummy variables

# Using Multiple Dummy Variables

$$books = b_1 + b_2 age + b_3 d_{somehs} + b_4 d_{hsgrad} + b_5 d_{somecol}$$

- The coefficient in front of each dummy variable can be interpreted as the difference between the average outcome for the dummy variable group and the average outcome for the omitted group
- For example, $b_3$ would be the difference between the average number of books read by high school dropouts and college graduates
- We can also compare coefficients to each other, $b_4 - b_3$ is the difference in the average number of books read by high school grads and high school dropouts

# Using Multiple Dummy Variables



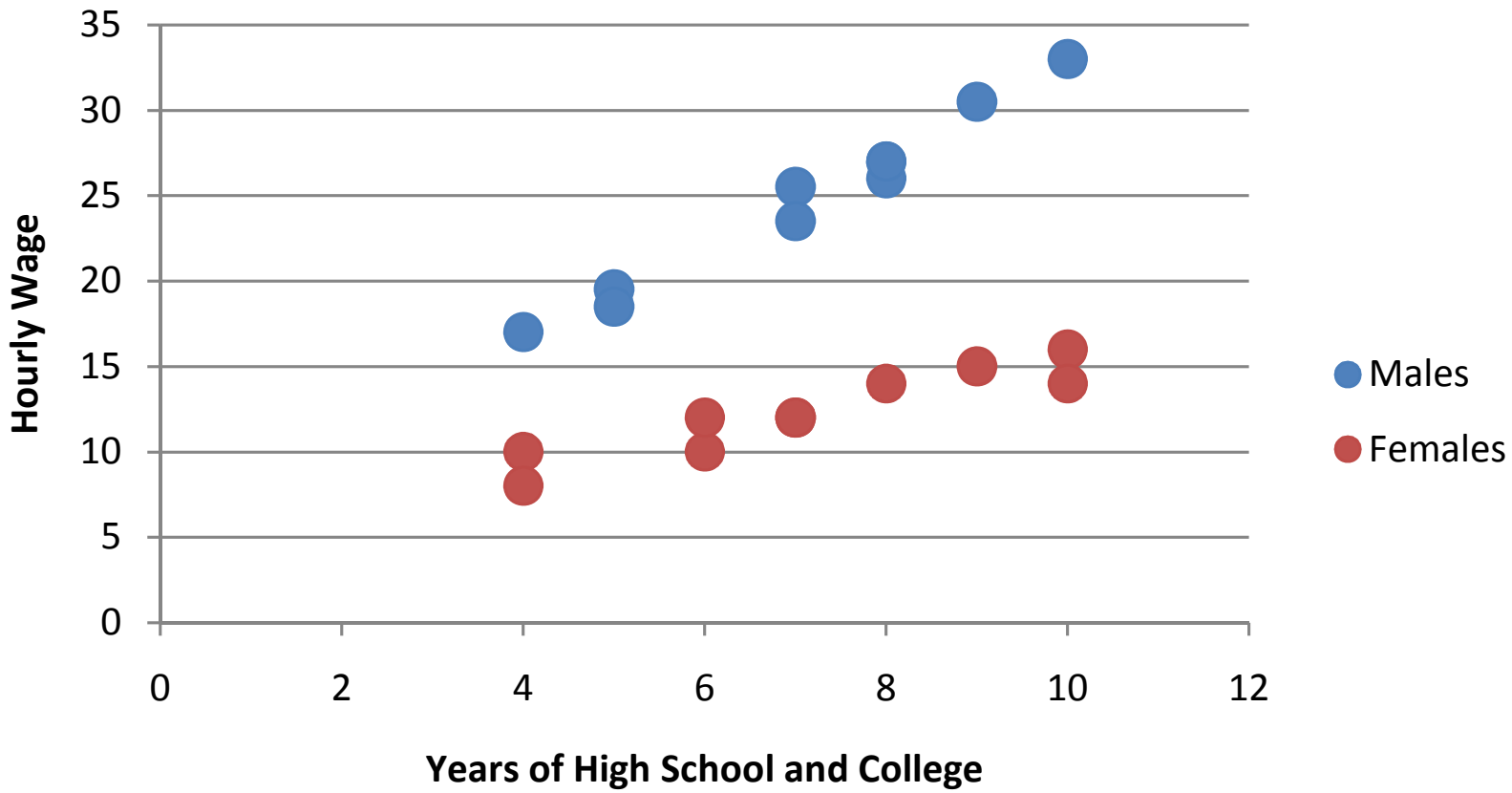
# Dummy Variables and Interaction Terms

- What if the difference between two groups isn't quite this simple?
- What if the coefficients on certain variables differ by group?
- We can model a situation like this by including *interaction terms* in our regression
- An interaction term is when we multiply one variable by another and include the result as a regressor
- So if we thought being a college grad both shifted the line for books as a function of age and changed its slope, we would include $d_{colgrad}$ in our regression and $d_{colgrad}$ times *age* in our regression

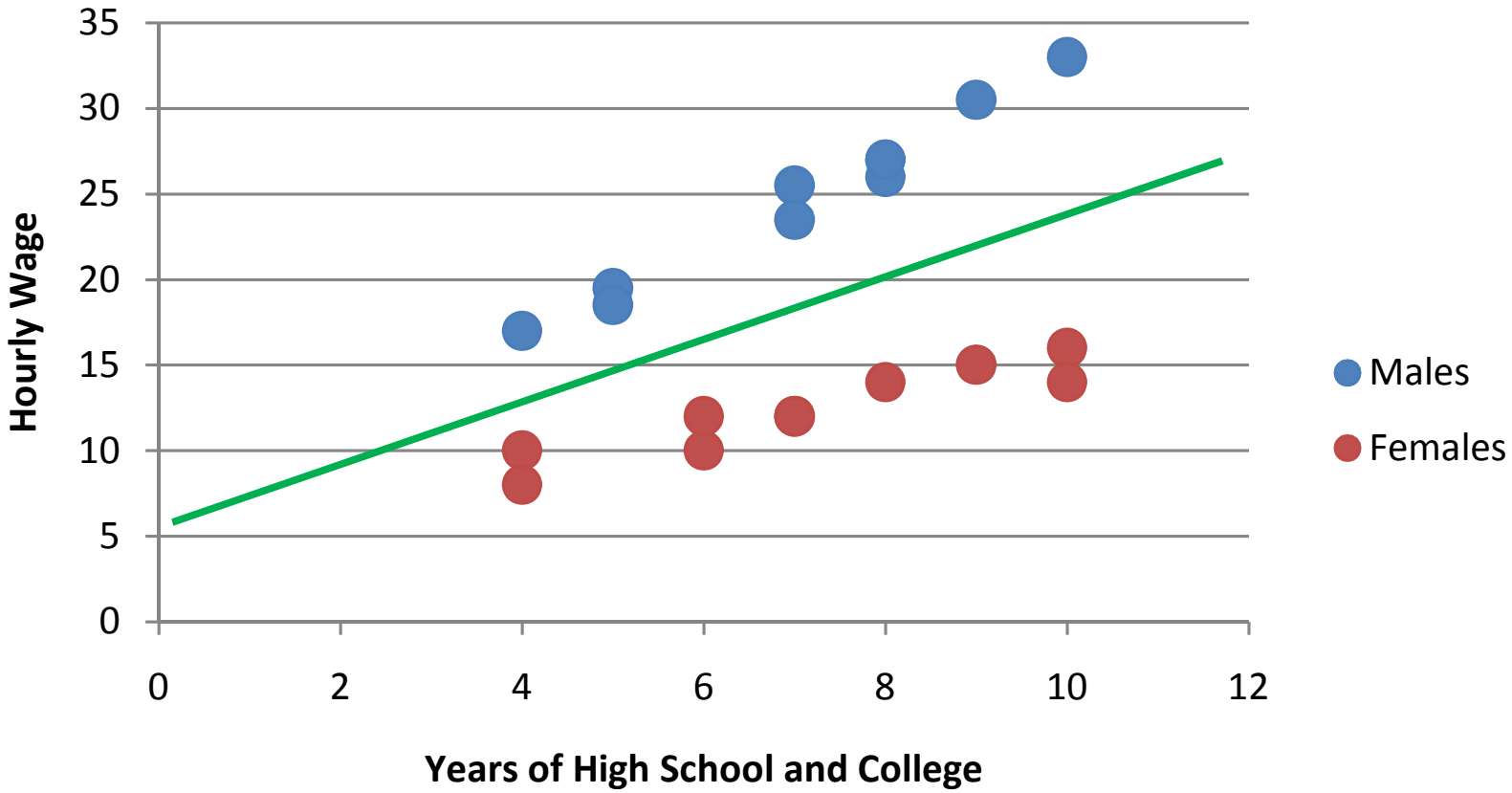
# Dummy Variables and Interaction Terms

- For an example of interaction terms, let's think about wage discrimination
- We may be concerned that women are payed less than men on average
- We could test for this by including a gender dummy variable in a wage regression
- However, we may also be concerned that women also get a smaller return to education than men
- To test for this, we would also need to include an interaction term between the gender dummy and education in our wage regression

# Dummy Variables and Interaction Terms

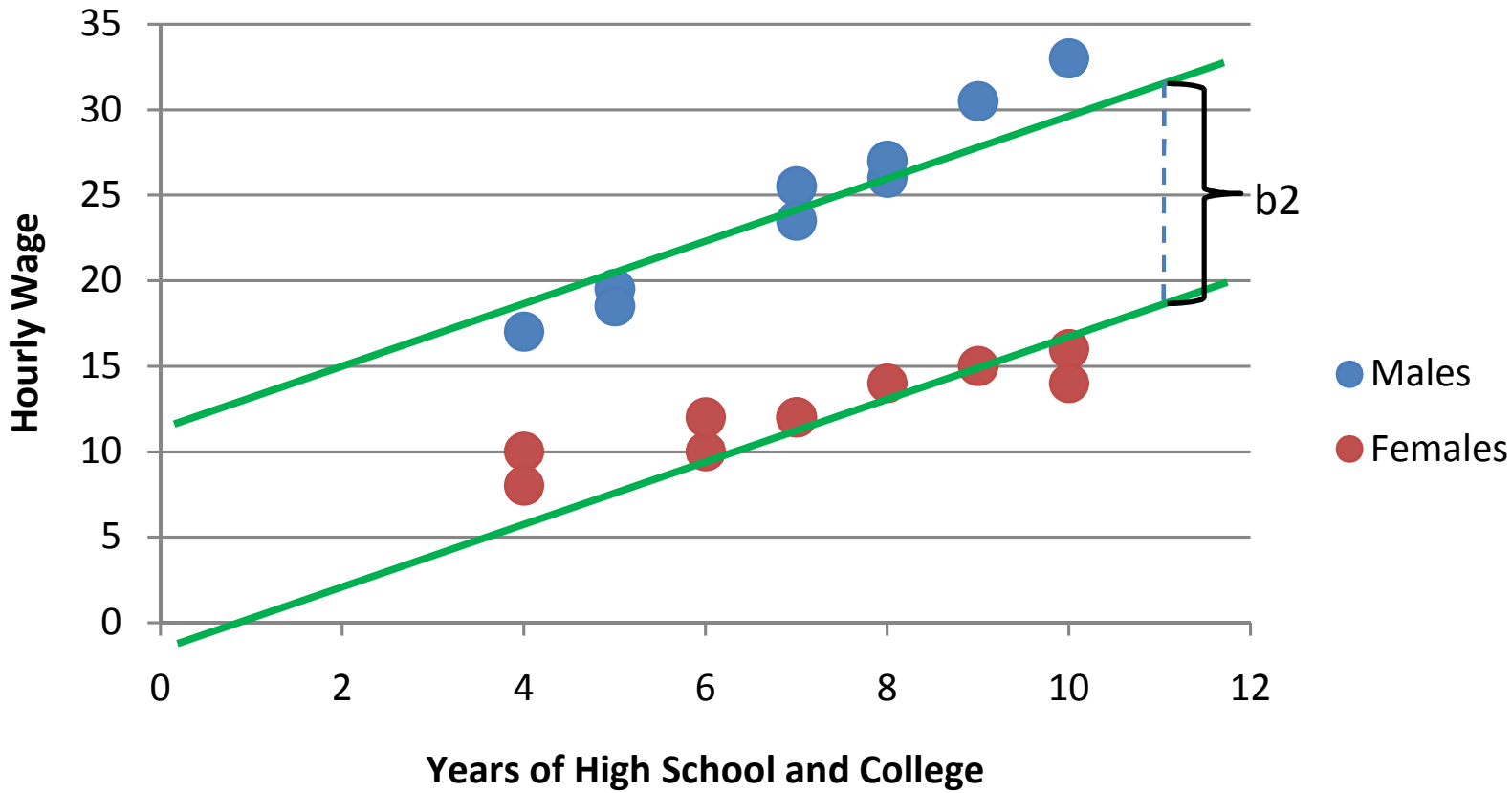


# Dummy Variables and Interaction Terms



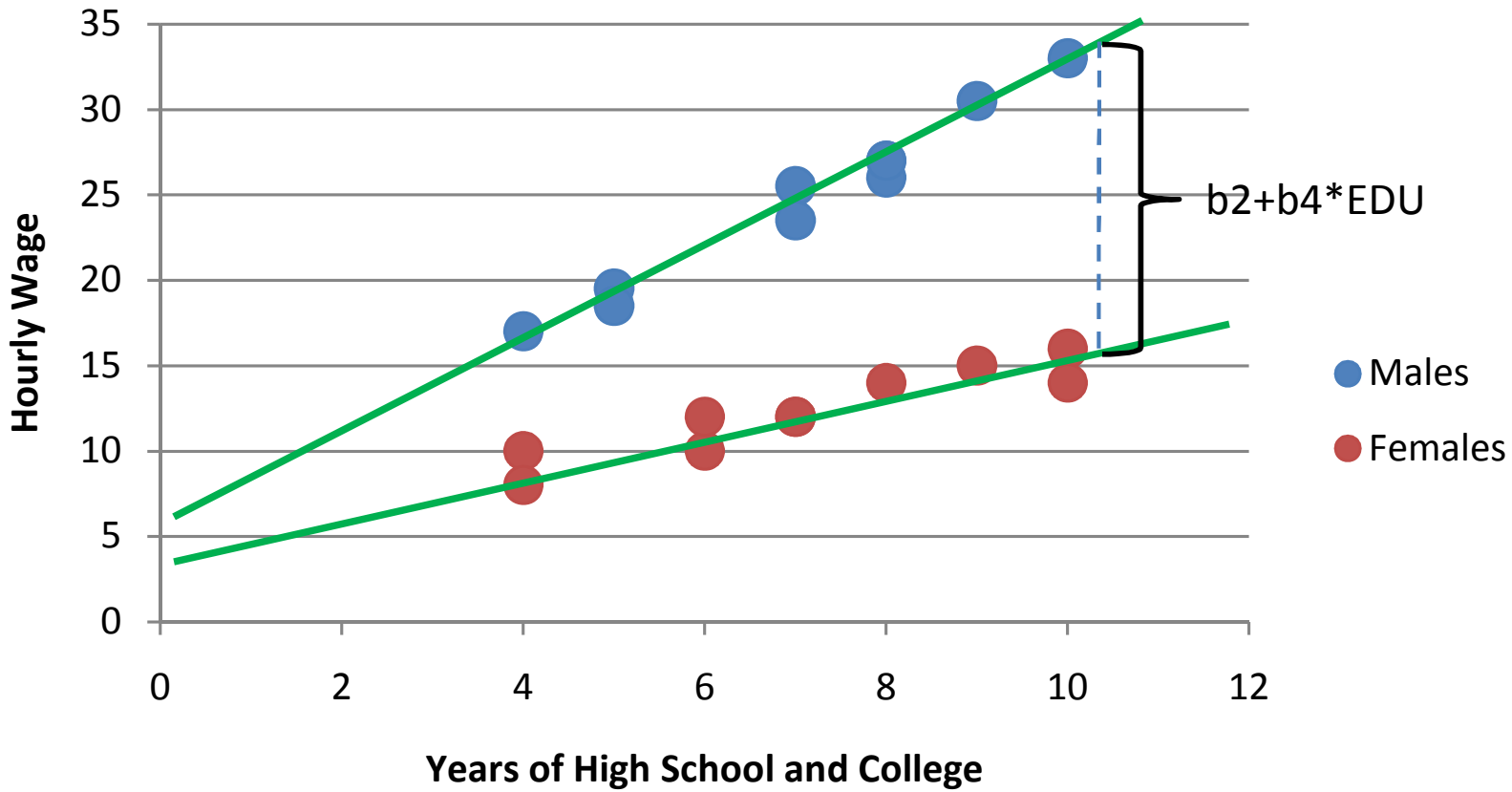

$$W = b_1 + b_2 \cdot EDU$$

# Dummy Variables and Interaction Terms




$$W = b_1 + b_2 \cdot MALE + b_3 \cdot EDU$$

# Dummy Variables and Interaction Terms




$$W = b_1 + b_2 \cdot MALE + b_3 \cdot EDU + b_4 \cdot MALE \cdot EDU$$

# Dummy Variables and Interaction Terms

- How do we interpret the coefficients?

$$W = b_1 + b_2 \cdot MALE + b_3 \cdot EDU + b_4 \cdot MALE \cdot EDU$$

- The difference in the predicted wage for males and females will depend on level of education:

$$\widehat{W_{male}} = (b_1 + b_2) + (b_3 + b_4)EDU$$

$$\widehat{W_{female}} = b_1 + b_3 EDU$$

$$\widehat{W_{male}} - \widehat{W_{female}} = b_2 + b_4 \cdot EDU$$

# Dummy Variables and Interaction Terms

- How do we interpret the coefficients?

$$W = b_1 + b_2 \cdot MALE + b_3 \cdot EDU + b_4 \cdot MALE \cdot EDU$$

- $b_2$: predicted difference in wages between a man and a woman who each have zero years of education
- $b_2 + b_4 \cdot EDU$: predicted difference in wages between a man and woman who each have $EDU$ years of education
- $b_4$: additional return to a year of education for males relative to females

# Interpreting Coefficients with Dummy Variables

- Suppose we regressed log wages on age, education, a dummy equal to one for males and and interaction between that dummy and education and got the following (standard errors are in parentheses):

$$\underset{}{\ln \text{wage}} = \underset{(.80)}{5} + \underset{(.001)}{0.03}\,\text{age} + \underset{(.002)}{0.08}\,\text{edu} + \underset{(.08)}{0.04}\,\text{male} + \underset{(.03)}{0.01}\,\text{male*edu}$$

- Would we conclude that there is gender discrimination in wages?
- The predicted difference in male and female log wages is $0.04 + 0.01 \cdot edu$
- This is positive at all education levels and increasing as education increases
- However, neither the coefficient on *male* or on $male \cdot edu$ is statistically significant

# Interpreting Coefficients with Dummy Variables

| ln wage | = | 5 | + | 0.03 | age | + | 0.08 | edu | + | 0.04 | male | + | 0.01 | male*edu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | (.80) | | (.001) | | | (.002) | | | (.08) | | | (.03) | |

- There is another thing to be careful about when interpreting the $male \cdot edu$ coefficient
- It's lack of significance does not mean that education has no effect on wage for males
- It means that education has no significant additional effect for males beyond the effect common to males and females
- One final caution: even if the coefficients were significant we don't have information on why male wages differ from female wages