# ECO 344 - Applied Econometrics I
## *Chapter 7*

Muhammad Salman Khalid

School of Economics & Social Sciences

November 2, 2026



# Outline


1. Introduction
2. Dummy Variable as an Independent Variable
3. Using Multiple Dummy Variables as Independent Variables
4. Interaction Terms for Dummy Variables
5. Binary Variable as a Dependent Variable
6. Policy Analysis and Program Evaluation
7. Discrete Dependent Variable
8. Summary

# Introduction

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.
- Similarly, some variables are represented by numbers but are not numerical in nature. Cardinal vs Ordinal.

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.
- Similarly, some variables are represented by numbers but are not numerical in nature. Cardinal vs Ordinal.
- However, we need to convert categories and ordinal variables into numbers to include them in our analysis. We can include such information using:

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.
- Similarly, some variables are represented by numbers but are not numerical in nature. Cardinal vs Ordinal.
- However, we need to convert categories and ordinal variables into numbers to include them in our analysis. We can include such information using:
  - Dummy variables - They can take only two values (1 and 0).

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.
- Similarly, some variables are represented by numbers but are not numerical in nature. Cardinal vs Ordinal.
- However, we need to convert categories and ordinal variables into numbers to include them in our analysis. We can include such information using:
  - Dummy variables - They can take only two values (1 and 0).
  - Categorical variables - They can take multiple values but are neither cardinal nor ordinal in nature.

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.
- Similarly, some variables are represented by numbers but are not numerical in nature. Cardinal vs Ordinal.
- However, we need to convert categories and ordinal variables into numbers to include them in our analysis. We can include such information using:
  - Dummy variables - They can take only two values (1 and 0).
  - Categorical variables - They can take multiple values but are neither cardinal nor ordinal in nature.
  - Ordinal variables - They can take multiple values that are ordered but not cardinal.

# Qualitative Information

- Qualitative information is information that cannot be described naturally in numbers. For example: gender, race, religion etc.
- Have you used qualitative information in any analysis using numbers?
- What does it mean when we say that it cannot be represented by numbers? Number vs Category.
- Similarly, some variables are represented by numbers but are not numerical in nature. Cardinal vs Ordinal.
- However, we need to convert categories and ordinal variables into numbers to include them in our analysis. We can include such information using:
  - Dummy variables - They can take only two values (1 and 0).
  - Categorical variables - They can take multiple values but are neither cardinal nor ordinal in nature.
  - Ordinal variables - They can take multiple values that are ordered but not cardinal.
- Be careful while including such variables in the analysis and while interpreting them.

# Dummy Variable as an Independent Variable

# Dummy Variable as an Independent Variable

- A dummy variable takes a value of either 1 or 0. For example: owning a laptop is a dummy variable which takes the value 1 if the participant owns a laptop and 0 otherwise.

# Dummy Variable as an Independent Variable

- A dummy variable takes a value of either 1 or 0. For example: owning a laptop is a dummy variable which takes the value 1 if the participant owns a laptop and 0 otherwise.
- Similarly, a categorical variable with only two categories can easily be converted into a dummy variable.

# Dummy Variable as an Independent Variable

- A dummy variable takes a value of either 1 or 0. For example: owning a laptop is a dummy variable which takes the value 1 if the participant owns a laptop and 0 otherwise.
- Similarly, a categorical variable with only two categories can easily be converted into a dummy variable.
    - For example: assume gender has two categories, male & female. We can convert it into a dummy variable.

# Dummy Variable as an Independent Variable

- A dummy variable takes a value of either 1 or 0. For example: owning a laptop is a dummy variable which takes the value 1 if the participant owns a laptop and 0 otherwise.
- Similarly, a categorical variable with only two categories can easily be converted into a dummy variable.
  - For example: assume gender has two categories, male & female. We can convert it into a dummy variable.
  - The variable "male" takes the value 1 if the participant is male and 0 if the participant is female.

# Dummy Variable as an Independent Variable

- A dummy variable takes a value of either 1 or 0. For example: owning a laptop is a dummy variable which takes the value 1 if the participant owns a laptop and 0 otherwise.
- Similarly, a categorical variable with only two categories can easily be converted into a dummy variable.
  - For example: assume gender has two categories, male & female. We can convert it into a dummy variable.
  - The variable "male" takes the value 1 if the participant is male and 0 if the participant is female.
  - Similarly, the variable "female" takes the value 1 if the participant is female and 0 if the participant is male.

# Dummy Variable as an Independent Variable

- A dummy variable takes a value of either 1 or 0. For example: owning a laptop is a dummy variable which takes the value 1 if the participant owns a laptop and 0 otherwise.
- Similarly, a categorical variable with only two categories can easily be converted into a dummy variable.
  - For example: assume gender has two categories, male & female. We can convert it into a dummy variable.
  - The variable "male" takes the value 1 if the participant is male and 0 if the participant is female.
  - Similarly, the variable "female" takes the value 1 if the participant is female and 0 if the participant is male.
- We can include the dummy variable as an independent variable in the regression. Let us see what this does to our model.

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$
  - In case of female, $female = 1$. Therefore, $wage = (\beta_0 + \delta_f) + \beta_1 education + \mu$

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$
  - In case of female, $female = 1$. Therefore, $wage = (\beta_0 + \delta_f) + \beta_1 education + \mu$
  - Holding education fixed, $\delta_f = E[wage|female, education] - E[wage|male, education]$

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$
  - In case of female, $female = 1$. Therefore, $wage = (\beta_0 + \delta_f) + \beta_1 education + \mu$
  - Holding education fixed, $\delta_f = E[wage|female, education] - E[wage|male, education]$
- So, the coefficient of a dummy variable is the difference in the average outcome between the group coded 1 and the base group (coded 0), keeping the other variables constant.

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$
  - In case of female, $female = 1$. Therefore, $wage = (\beta_0 + \delta_f) + \beta_1 education + \mu$
  - Holding education fixed, $\delta_f = E[wage|female, education] - E[wage|male, education]$
- So, the coefficient of a dummy variable is the difference in the average outcome between the group coded 1 and the base group (coded 0), keeping the other variables constant.
- In the example above, $\delta_f$ is the difference in the average wage of a female and a male with the same education.

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$
  - In case of female, $female = 1$. Therefore, $wage = (\beta_0 + \delta_f) + \beta_1 education + \mu$
  - Holding education fixed, $\delta_f = E[wage|female, education] - E[wage|male, education]$
- So, the coefficient of a dummy variable is the difference in the average outcome between the group coded 1 and the base group (coded 0), keeping the other variables constant.
- In the example above, $\delta_f$ is the difference in the average wage of a female and a male with the same education.
- If $\delta_f < 0$, females earn less than comparable males.

# Dummy Variable as an Independent Variable

- Consider the following model:

$$wage = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- What would be the interpretation of $\delta_f$ in our model?
- To clarify the interpretation of $\delta_f$, let us consider male and female separately.
  - In case of male, $female = 0$. Therefore, $wage = \beta_0 + \beta_1 education + \mu$
  - In case of female, $female = 1$. Therefore, $wage = (\beta_0 + \delta_f) + \beta_1 education + \mu$
  - Holding education fixed, $\delta_f = E[wage|female, education] - E[wage|male, education]$
- So, the coefficient of a dummy variable is the difference in the average outcome between the group coded 1 and the base group (coded 0), keeping the other variables constant.
- In the example above, $\delta_f$ is the difference in the average wage of a female and a male with the same education.
- If $\delta_f < 0$, females earn less than comparable males.
- Under what condition can we say that there is no gap between male and female wages? ($\delta_f = 0$)

# Dummy Variable: Intercept Change

- Can we call this an intercept change?

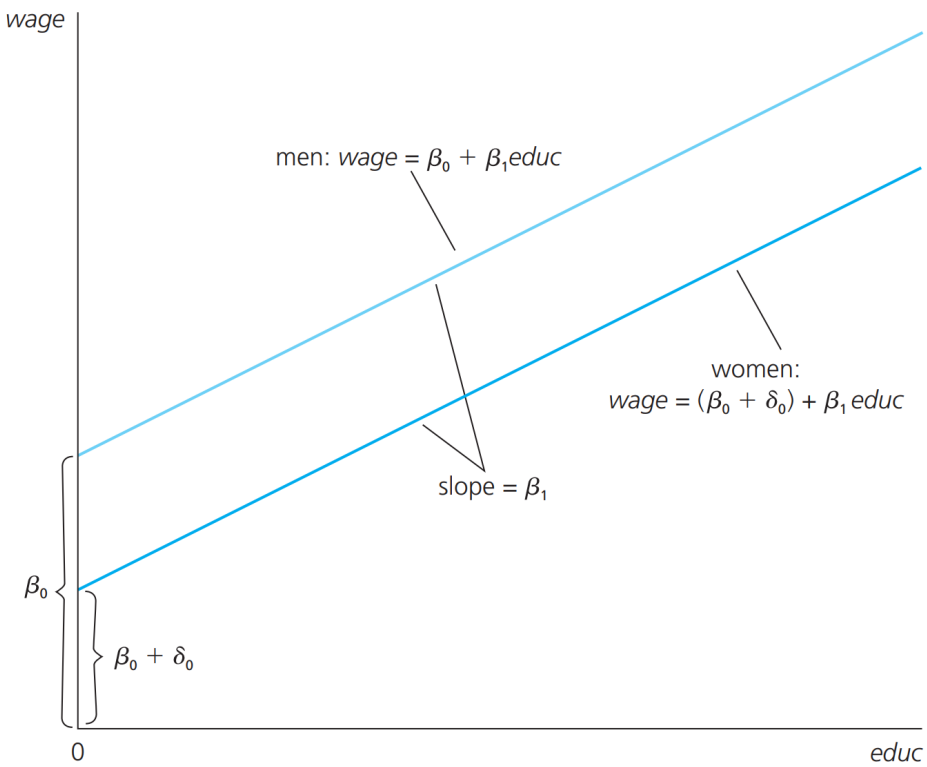
# Dummy Variable: Intercept Change

- Can we call this an intercept change?
- The intercept of male is $\beta_0$ while the intercept of female is $\beta_0 + \delta_f$. The slope of education is the same for both groups.

# Dummy Variable: Intercept Change

- Can we call this an intercept change?
- The intercept of male is $\beta_0$ while the intercept of female is $\beta_0 + \delta_f$. The slope of education is the same for both groups.

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



# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?

# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?
- **NO**, because of perfect multicollinearity as $male + female = 1$, which is the constant. This is the Dummy Variable Trap.

# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?
- **NO**, because of perfect multicollinearity as $male + female = 1$, which is the constant. This is the Dummy Variable Trap.
- The only way to add both variables is to run the regression without an intercept.

# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?
- **NO**, because of perfect multicollinearity as $male + female = 1$, which is the constant. This is the Dummy Variable Trap.
- The only way to add both variables is to run the regression without an intercept.
  - The benefit is that the coefficients are directly the intercepts of male and female.

# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?
- **NO**, because of perfect multicollinearity as $male + female = 1$, which is the constant. This is the Dummy Variable Trap.
- The only way to add both variables is to run the regression without an intercept.
  - The benefit is that the coefficients are directly the intercepts of male and female.
  - However, there is no agreed upon R-Squared calculation for a regression without an intercept, and testing the difference between groups becomes harder.

# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?
- **NO**, because of perfect multicollinearity as $male + female = 1$, which is the constant. This is the Dummy Variable Trap.
- The only way to add both variables is to run the regression without an intercept.
  - The benefit is that the coefficients are directly the intercepts of male and female.
  - However, there is no agreed upon R-Squared calculation for a regression without an intercept, and testing the difference between groups becomes harder.
- Generally, the coefficient of a dummy variable tells us the difference in $y$ as compared to the base group (the group where the dummy variable is zero).

# Dummy Variable Trap

- Can we add both the male and the female dummy variable in the regression?
- **NO**, because of perfect multicollinearity as $male + female = 1$, which is the constant. This is the Dummy Variable Trap.
- The only way to add both variables is to run the regression without an intercept.
  - The benefit is that the coefficients are directly the intercepts of male and female.
  - However, there is no agreed upon R-Squared calculation for a regression without an intercept, and testing the difference between groups becomes harder.
- Generally, the coefficient of a dummy variable tells us the difference in $y$ as compared to the base group (the group where the dummy variable is zero).
- The choice of base group does not change the fit of the model. It only changes how the coefficients are read.

# Is the Gap Discrimination?

- Now, can we say based on the model discussed before that $\delta_f$ really measures discrimination against females?

# Is the Gap Discrimination?

- Now, can we say based on the model discussed before that $\delta_f$ really measures discrimination against females?
- Think about the ceteris paribus interpretation of $\delta_f$ and the model together.

# Is the Gap Discrimination?

- Now, can we say based on the model discussed before that $\delta_f$ really measures discrimination against females?
- Think about the ceteris paribus interpretation of $\delta_f$ and the model together.
- $\delta_f$ is the gap that remains after controlling for the variables in the model. It is consistent with discrimination.

# Is the Gap Discrimination?

- Now, can we say based on the model discussed before that $\delta_f$ really measures discrimination against females?
- Think about the ceteris paribus interpretation of $\delta_f$ and the model together.
- $\delta_f$ is the gap that remains after controlling for the variables in the model. It is consistent with discrimination.
- But it also captures any productivity differences between males and females that we have not controlled for, such as experience, tenure or occupation.

# Is the Gap Discrimination?

- Now, can we say based on the model discussed before that $\delta_f$ really measures discrimination against females?
- Think about the ceteris paribus interpretation of $\delta_f$ and the model together.
- $\delta_f$ is the gap that remains after controlling for the variables in the model. It is consistent with discrimination.
- But it also captures any productivity differences between males and females that we have not controlled for, such as experience, tenure or occupation.
- So the dummy gives us a conditional gap. A causal claim about discrimination needs more than adding a dummy to the regression.

# Independent Dummy Variable with Dependent Log Variable

- How do we interpret the coefficient of a dummy variable when the dependent variable is $\log(y)$?

$$\log(wage) = \beta_0 + \delta_f female + \beta_1 education + \mu$$

# Independent Dummy Variable with Dependent Log Variable

- How do we interpret the coefficient of a dummy variable when the dependent variable is $\log(y)$?

$$\log(wage) = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- The coefficient has an interpretation in terms of percentage.

# Independent Dummy Variable with Dependent Log Variable

- How do we interpret the coefficient of a dummy variable when the dependent variable is $\log(y)$?

$$\log(wage) = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- The coefficient has an interpretation in terms of percentage.
- The approximate difference is $100 \times \delta_f$ percent. A female earns about $100 \times \delta_f$ percent less than a male (if $\delta_f < 0$) keeping education constant.

# Independent Dummy Variable with Dependent Log Variable

- How do we interpret the coefficient of a dummy variable when the dependent variable is $\log(y)$?

$$\log(wage) = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- The coefficient has an interpretation in terms of percentage.
- The approximate difference is $100 \times \delta_f$ percent. A female earns about $100 \times \delta_f$ percent less than a male (if $\delta_f < 0$) keeping education constant.
- This approximation is good only when $\delta_f$ is small. When the coefficient is large, the log approximation is poor.

# Independent Dummy Variable with Dependent Log Variable

- How do we interpret the coefficient of a dummy variable when the dependent variable is $\log(y)$?

$$\log(wage) = \beta_0 + \delta_f female + \beta_1 education + \mu$$

- The coefficient has an interpretation in terms of percentage.
- The approximate difference is $100 \times \delta_f$ percent. A female earns about $100 \times \delta_f$ percent less than a male (if $\delta_f < 0$) keeping education constant.
- This approximation is good only when $\delta_f$ is small. When the coefficient is large, the log approximation is poor.
- The exact percentage difference between the group coded 1 and the base group is $100 \times (e^{\delta_f} - 1)\%$.

# Example 1: Hourly Wage Equation (Example 7.1)

- Estimating the model we have been discussing, with experience and tenure added as controls, gets us the following sample regression function.

$$\widehat{wage} = -1.57 - 1.81 female + 0.572 education + 0.025 exper + 0.141 tenure$$

# Example 1: Hourly Wage Equation (Example 7.1)

- Estimating the model we have been discussing, with experience and tenure added as controls, gets us the following sample regression function.

$$\widehat{wage} = -1.57 - 1.81 female + 0.572 education + 0.025 exper + 0.141 tenure$$

- What is the interpretation of the coefficient of female?

# Example 1: Hourly Wage Equation (Example 7.1)

- Estimating the model we have been discussing, with experience and tenure added as controls, gets us the following sample regression function.

$$\widehat{wage} = -1.57 - 1.81 female + 0.572 education + 0.025 exper + 0.141 tenure$$

- What is the interpretation of the coefficient of female?
- On average, a female earns 1.81 USD per hour less than a male with the same education, experience and tenure.

# Example 1: Hourly Wage Equation (Example 7.1)

- Estimating the model we have been discussing, with experience and tenure added as controls, gets us the following sample regression function.

$$\widehat{wage} = -1.57 - 1.81 female + 0.572 education + 0.025 exper + 0.141 tenure$$

- What is the interpretation of the coefficient of female?
- On average, a female earns 1.81 USD per hour less than a male with the same education, experience and tenure.
- Without any controls the raw gap in the same data is larger. Controlling for the other variables shrinks it, but does not remove it.

# Example 1: Hourly Wage Equation (Example 7.1)

- Estimating the model we have been discussing, with experience and tenure added as controls, gets us the following sample regression function.

$$\widehat{wage} = -1.57 - 1.81 female + 0.572 education + 0.025 exper + 0.141 tenure$$

- What is the interpretation of the coefficient of female?
- On average, a female earns 1.81 USD per hour less than a male with the same education, experience and tenure.
- Without any controls the raw gap in the same data is larger. Controlling for the other variables shrinks it, but does not remove it.
- Is this an indication of discrimination? Remember the previous slide.

# Example 2: Computer Ownership and College GPA (Example 7.2)

- We are interested in estimating the effect of owning a computer on college GPA. We run the following model:

$$colGPA = \beta_0 + \beta_1 PC + \beta_2 hsGPA + \beta_3 ACT + \mu$$

# Example 2: Computer Ownership and College GPA (Example 7.2)

- We are interested in estimating the effect of owning a computer on college GPA. We run the following model:

$$colGPA = \beta_0 + \beta_1 PC + \beta_2 hsGPA + \beta_3 ACT + \mu$$

- We get the following estimation:

$$\widehat{colGPA} = 1.26 + 0.157 PC + 0.447 hsGPA + 0.0087 ACT$$

# Example 2: Computer Ownership and College GPA (Example 7.2)

- We are interested in estimating the effect of owning a computer on college GPA. We run the following model:

$$colGPA = \beta_0 + \beta_1 PC + \beta_2 hsGPA + \beta_3 ACT + \mu$$

- We get the following estimation:

$$\widehat{colGPA} = 1.26 + 0.157 PC + 0.447 hsGPA + 0.0087 ACT$$

- What is the interpretation of the coefficient of PC?

# Example 2: Computer Ownership and College GPA (Example 7.2)

- We are interested in estimating the effect of owning a computer on college GPA. We run the following model:

$$colGPA = \beta_0 + \beta_1 PC + \beta_2 hsGPA + \beta_3 ACT + \mu$$

- We get the following estimation:

$$\widehat{colGPA} = 1.26 + 0.157 PC + 0.447 hsGPA + 0.0087 ACT$$

- What is the interpretation of the coefficient of PC?
- Students who own a computer score 0.157 points higher in college GPA than students who do not own a computer, keeping high school GPA and ACT score constant.

# Example 2: Computer Ownership and College GPA (Example 7.2)

- We are interested in estimating the effect of owning a computer on college GPA. We run the following model:

$$colGPA = \beta_0 + \beta_1 PC + \beta_2 hsGPA + \beta_3 ACT + \mu$$

- We get the following estimation:

$$\widehat{colGPA} = 1.26 + 0.157 PC + 0.447 hsGPA + 0.0087 ACT$$

- What is the interpretation of the coefficient of PC?
- Students who own a computer score 0.157 points higher in college GPA than students who do not own a computer, keeping high school GPA and ACT score constant.
- Is this a causal effect of the computer? What might be missing from the model?

# Example 3: Log Wage and Gender (Example 7.4)

- Let us consider the following regression for the wage gap between females and males.

$$\log(wage) = \beta_0 + \delta_1 female + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \mu$$

# Example 3: Log Wage and Gender (Example 7.4)

- Let us consider the following regression for the wage gap between females and males.

$$\log(wage) = \beta_0 + \delta_1 female + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \mu$$

- The estimated regression function is:

$$\widehat{\log(wage)} = 0.417 - 0.297 female + 0.080 educ + 0.029 exper - 0.00058 exper^2 + 0.032 tenure - 0.00059 tenure^2$$

# Example 3: Log Wage and Gender (Example 7.4)

- Let us consider the following regression for the wage gap between females and males.

$$\log(wage) = \beta_0 + \delta_1 female + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \mu$$

- The estimated regression function is:

$$\widehat{\log(wage)} = 0.417 - 0.297 female + 0.080 educ + 0.029 exper - 0.00058 exper^2 + 0.032 tenure - 0.00059 tenure^2$$

- What is the interpretation of the coefficient of female?

# Example 3: Log Wage and Gender (Example 7.4)

- Let us consider the following regression for the wage gap between females and males.

$$\log(wage) = \beta_0 + \delta_1 female + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \mu$$

- The estimated regression function is:

$$\widehat{\log(wage)} = 0.417 - 0.297 female + 0.080 educ + 0.029 exper - 0.00058 exper^2 + 0.032 tenure - 0.00059 tenure^2$$

- What is the interpretation of the coefficient of female?
- Approximately, a female earns 29.7 percent less than a male with the same education, experience and tenure.

# Example 3: Log Wage and Gender (Example 7.4)

- Let us consider the following regression for the wage gap between females and males.

$$\log(wage) = \beta_0 + \delta_1 female + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \mu$$

- The estimated regression function is:

$$\widehat{\log(wage)} = 0.417 - 0.297 female + 0.080 educ + 0.029 exper - 0.00058 exper^2 + 0.032 tenure - 0.00059 tenure^2$$

- What is the interpretation of the coefficient of female?
- Approximately, a female earns 29.7 percent less than a male with the same education, experience and tenure.
- The exact difference is $100(e^{-0.297} - 1) = -25.7$. So a female earns about 25.7 percent less than a comparable male.

# Using Multiple Dummy Variables as Independent Variables

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?
- Suppose we have the categorical variable education with the categories 1: Less than primary/Illiterate, 2: Primary Education, 3: Matriculation, 4: Finished College, 5: Undergraduate, 6: Master's, 7: Doctorate, 8: Post-Doctorate, in the regression below:

$$wage = \beta_0 + \beta_1 education + \mu$$

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?
- Suppose we have the categorical variable education with the categories 1: Less than primary/Illiterate, 2: Primary Education, 3: Matriculation, 4: Finished College, 5: Undergraduate, 6: Master's, 7: Doctorate, 8: Post-Doctorate, in the regression below:

$$wage = \beta_0 + \beta_1 education + \mu$$

- What is the interpretation of $\beta_1$ in this case?

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?
- Suppose we have the categorical variable education with the categories 1: Less than primary/Illiterate, 2: Primary Education, 3: Matriculation, 4: Finished College, 5: Undergraduate, 6: Master's, 7: Doctorate, 8: Post-Doctorate, in the regression below:

$$wage = \beta_0 + \beta_1 education + \mu$$

- What is the interpretation of $\beta_1$ in this case?
- Does the interpretation feel strange? It assumes that every step up the code has the same effect on wage.

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?
- Suppose we have the categorical variable education with the categories 1: Less than primary/Illiterate, 2: Primary Education, 3: Matriculation, 4: Finished College, 5: Undergraduate, 6: Master's, 7: Doctorate, 8: Post-Doctorate, in the regression below:

$$wage = \beta_0 + \beta_1 education + \mu$$

- What is the interpretation of $\beta_1$ in this case?
- Does the interpretation feel strange? It assumes that every step up the code has the same effect on wage.
- The solution is to create a dummy variable for each category and include them in the regression.

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?
- Suppose we have the categorical variable education with the categories 1: Less than primary/Illiterate, 2: Primary Education, 3: Matriculation, 4: Finished College, 5: Undergraduate, 6: Master's, 7: Doctorate, 8: Post-Doctorate, in the regression below:

$$wage = \beta_0 + \beta_1 education + \mu$$

- What is the interpretation of $\beta_1$ in this case?
- Does the interpretation feel strange? It assumes that every step up the code has the same effect on wage.
- The solution is to create a dummy variable for each category and include them in the regression.
- The model could be:

$$\begin{aligned} wage = \beta_0 &+ \beta_1 illit + \beta_2 prim + \beta_3 matric + \beta_4 college \\ &+ \beta_5 undergrad + \beta_6 master + \beta_7 phd + \beta_8 postdoc + \mu \end{aligned}$$

# Adding Categorical Independent Variable

- We have seen how we can incorporate a dummy variable in the regression. What about a categorical variable with more than two categories?
- Suppose we have the categorical variable education with the categories 1: Less than primary/Illiterate, 2: Primary Education, 3: Matriculation, 4: Finished College, 5: Undergraduate, 6: Master's, 7: Doctorate, 8: Post-Doctorate, in the regression below:

$$wage = \beta_0 + \beta_1 education + \mu$$

- What is the interpretation of $\beta_1$ in this case?
- Does the interpretation feel strange? It assumes that every step up the code has the same effect on wage.
- The solution is to create a dummy variable for each category and include them in the regression.
- The model could be:

$$\begin{aligned} wage = \beta_0 &+ \beta_1 illit + \beta_2 prim + \beta_3 matric + \beta_4 college \\ &+ \beta_5 undergrad + \beta_6 master + \beta_7 phd + \beta_8 postdoc + \mu \end{aligned}$$

- What is wrong with the above model?

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.
- The correct model would be:

$$\begin{aligned} wage = \beta_0 + \beta_1 prim + \beta_2 matric + \beta_3 college + \beta_4 undergrad \\ + \beta_5 master + \beta_6 phd + \beta_7 postdoc + \mu \end{aligned}$$

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.
- The correct model would be:

$$\begin{aligned} wage = \beta_0 + \beta_1 prim + \beta_2 matric + \beta_3 college + \beta_4 undergrad \\ + \beta_5 master + \beta_6 phd + \beta_7 postdoc + \mu \end{aligned}$$

- The dummy variable we drop acts as the base group or comparison group.

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.
- The correct model would be:

$$\begin{aligned} wage = \beta_0 + \beta_1 prim + \beta_2 matric + \beta_3 college + \beta_4 undergrad \\ + \beta_5 master + \beta_6 phd + \beta_7 postdoc + \mu \end{aligned}$$

- The dummy variable we drop acts as the base group or comparison group.
- What is the interpretation of $\beta_5$?

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.
- The correct model would be:

$$\begin{aligned} wage = \beta_0 + \beta_1 prim + \beta_2 matric + \beta_3 college + \beta_4 undergrad \\ + \beta_5 master + \beta_6 phd + \beta_7 postdoc + \mu \end{aligned}$$

- The dummy variable we drop acts as the base group or comparison group.
- What is the interpretation of $\beta_5$?
- An individual with a master's degree earns $\beta_5$ USD per hour more **than an individual with less than primary education**, keeping everything else constant.

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.
- The correct model would be:

$$\begin{aligned} wage = \beta_0 + \beta_1 prim + \beta_2 matric + \beta_3 college + \beta_4 undergrad \\ + \beta_5 master + \beta_6 phd + \beta_7 postdoc + \mu \end{aligned}$$

- The dummy variable we drop acts as the base group or comparison group.
- What is the interpretation of $\beta_5$?
- An individual with a master's degree earns $\beta_5$ USD per hour more **than an individual with less than primary education**, keeping everything else constant.
- All the discussion regarding a log dependent variable applies here as well.

# Adding Categorical Independent Variable

- **Dummy Variable Trap** - the nine terms (intercept plus eight dummies) are perfectly collinear. What is the solution? Drop one of the dummy variables.
- The correct model would be:

$$\begin{aligned} wage = \beta_0 + \beta_1 prim + \beta_2 matric + \beta_3 college + \beta_4 undergrad \\ + \beta_5 master + \beta_6 phd + \beta_7 postdoc + \mu \end{aligned}$$

- The dummy variable we drop acts as the base group or comparison group.
- What is the interpretation of $\beta_5$?
- An individual with a master's degree earns $\beta_5$ USD per hour more **than an individual with less than primary education**, keeping everything else constant.
- All the discussion regarding a log dependent variable applies here as well.
- To compare two included groups, for example master's and undergraduate, look at $\beta_5 - \beta_4$. A t-test on this difference needs the standard error of the difference, or we can simply change the base group.

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.
- Consider the following model, which captures the effect of rating on the interest rate.

$$interest = \beta_0 + \beta_1 CR + \mu$$

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.
- Consider the following model, which captures the effect of rating on the interest rate.

$$interest = \beta_0 + \beta_1 CR + \mu$$

- What is the interpretation of $\beta_1$?

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.
- Consider the following model, which captures the effect of rating on the interest rate.

$$interest = \beta_0 + \beta_1 CR + \mu$$

- What is the interpretation of $\beta_1$?
- We can say that $\beta_1$ is the change in the interest rate from a 1 unit increase in rating.

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.
- Consider the following model, which captures the effect of rating on the interest rate.

$$interest = \beta_0 + \beta_1 CR + \mu$$

- What is the interpretation of $\beta_1$?
- We can say that $\beta_1$ is the change in the interest rate from a 1 unit increase in rating.
- The interpretation looks fine, but it is not right. It forces the move from Poor to Fair to have the same effect as the move from Good to Excellent.

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.
- Consider the following model, which captures the effect of rating on the interest rate.

$$interest = \beta_0 + \beta_1 CR + \mu$$

- What is the interpretation of $\beta_1$?
- We can say that $\beta_1$ is the change in the interest rate from a 1 unit increase in rating.
- The interpretation looks fine, but it is not right. It forces the move from Poor to Fair to have the same effect as the move from Good to Excellent.
- The solution is to include a dummy variable for each category, leaving one as the base group.

$$interest = \beta_0 + \beta_1 Fair + \beta_2 Good + \beta_3 Excellent + \mu$$

# Adding Ordinal Independent Variable

- An ordinal variable preserves the order of categories. For example: credit rating of a bond.
- Let us assume the credit rating of a bond $CR$ is coded 4: Excellent, 3: Good, 2: Fair, 1: Poor.
- Consider the following model, which captures the effect of rating on the interest rate.

$$interest = \beta_0 + \beta_1 CR + \mu$$

- What is the interpretation of $\beta_1$?
- We can say that $\beta_1$ is the change in the interest rate from a 1 unit increase in rating.
- The interpretation looks fine, but it is not right. It forces the move from Poor to Fair to have the same effect as the move from Good to Excellent.
- The solution is to include a dummy variable for each category, leaving one as the base group.

$$interest = \beta_0 + \beta_1 Fair + \beta_2 Good + \beta_3 Excellent + \mu$$

- Here the base group is Poor, and $\beta_3$ is the difference in the interest rate between an Excellent and a Poor rating.

# Adjusting Categories

- So far, we have seen that categories can be added to the regression through individual dummy variables.

# Adjusting Categories

- So far, we have seen that categories can be added to the regression through individual dummy variables.
- However, what if we have a large number of categories? For example: the name of the political party.

# Adjusting Categories

- So far, we have seen that categories can be added to the regression through individual dummy variables.
- However, what if we have a large number of categories? For example: the name of the political party.
- We need to add $n-1$ dummy variables to capture the effect of the political party on our outcome. That $n-1$ could be a very large number.

# Adjusting Categories

- So far, we have seen that categories can be added to the regression through individual dummy variables.
- However, what if we have a large number of categories? For example: the name of the political party.
- We need to add $n-1$ dummy variables to capture the effect of the political party on our outcome. That $n-1$ could be a very large number.
- The solution is to re-categorize into a few dominant parties and put all the rest into "others".

# Adjusting Categories

- So far, we have seen that categories can be added to the regression through individual dummy variables.
- However, what if we have a large number of categories? For example: the name of the political party.
- We need to add $n-1$ dummy variables to capture the effect of the political party on our outcome. That $n-1$ could be a very large number.
- The solution is to re-categorize into a few dominant parties and put all the rest into "others".
- This reduces the number of dummy variables that need to be included in the regression.

# Adjusting Categories

- So far, we have seen that categories can be added to the regression through individual dummy variables.
- However, what if we have a large number of categories? For example: the name of the political party.
- We need to add $n-1$ dummy variables to capture the effect of the political party on our outcome. That $n-1$ could be a very large number.
- The solution is to re-categorize into a few dominant parties and put all the rest into "others".
- This reduces the number of dummy variables that need to be included in the regression.
- In this example, we drop the dummy of "others" so that it becomes the base group. The coefficient of each party is then read against all the small parties.

# Interaction Terms for Dummy Variables

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.
- We can use the same concept with dummy variables. There are two types of interactions that we can implement in a model involving dummy variables.

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.
- We can use the same concept with dummy variables. There are two types of interactions that we can implement in a model involving dummy variables.
  1. Interaction between dummy variables - creates subcategories.

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.
- We can use the same concept with dummy variables. There are two types of interactions that we can implement in a model involving dummy variables.
  1. Interaction between dummy variables - creates subcategories.
  2. Interaction of a dummy variable and a numerical variable - slope change among different groups.

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.
- We can use the same concept with dummy variables. There are two types of interactions that we can implement in a model involving dummy variables.
  1. Interaction between dummy variables - creates subcategories.
  2. Interaction of a dummy variable and a numerical variable - slope change among different groups.
- Consider the following model:

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \mu$$

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.
- We can use the same concept with dummy variables. There are two types of interactions that we can implement in a model involving dummy variables.
  1. Interaction between dummy variables - creates subcategories.
  2. Interaction of a dummy variable and a numerical variable - slope change among different groups.
- Consider the following model:

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \mu$$

- What do $\beta_1$ and $\beta_2$ represent?

# Dummy Interaction Terms

- We have seen that we can add interaction terms when we think that the effect of one independent variable depends on another.
- We can use the same concept with dummy variables. There are two types of interactions that we can implement in a model involving dummy variables.
  1. Interaction between dummy variables - creates subcategories.
  2. Interaction of a dummy variable and a numerical variable - slope change among different groups.
- Consider the following model:

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \mu$$

- What do $\beta_1$ and $\beta_2$ represent?
- This model forces the effect of marriage to be the same for males and females. What if we want to see the effect of being a married female?

# Interaction Between Dummy Variables

- To capture the effect of being a married female, let us consider the following model:

$$wage = \beta_0 + \beta_1 marriedfemale + \mu$$

where *marriedfemale* is a dummy variable taking the value 1 when the respondent is female and married.

# Interaction Between Dummy Variables

- To capture the effect of being a married female, let us consider the following model:

$$wage = \beta_0 + \beta_1 marriedfemale + \mu$$

  where *marriedfemale* is a dummy variable taking the value 1 when the respondent is female and married.
- What is wrong with the model? Think about the interpretation of $\beta_1$.

# Interaction Between Dummy Variables

- To capture the effect of being a married female, let us consider the following model:

$$wage = \beta_0 + \beta_1 marriedfemale + \mu$$

  where *marriedfemale* is a dummy variable taking the value 1 when the respondent is female and married.
- What is wrong with the model? Think about the interpretation of $\beta_1$.
- The base group now mixes unmarried males, married males and unmarried females. So $\beta_1$ compares married females with an average of three different groups.

# Interaction Between Dummy Variables

- To capture the effect of being a married female, let us consider the following model:

$$wage = \beta_0 + \beta_1 marriedfemale + \mu$$

  where *marriedfemale* is a dummy variable taking the value 1 when the respondent is female and married.
- What is wrong with the model? Think about the interpretation of $\beta_1$.
- The base group now mixes unmarried males, married males and unmarried females. So $\beta_1$ compares married females with an average of three different groups.
- The correct model would be:

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female \times married + \mu$$

# Interaction Between Dummy Variables

- To capture the effect of being a married female, let us consider the following model:

$$wage = \beta_0 + \beta_1 marriedfemale + \mu$$

  where *marriedfemale* is a dummy variable taking the value 1 when the respondent is female and married.
- What is wrong with the model? Think about the interpretation of $\beta_1$.
- The base group now mixes unmarried males, married males and unmarried females. So $\beta_1$ compares married females with an average of three different groups.
- The correct model would be:

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female \times married + \mu$$

- Now the base group is clear: unmarried males. Each of the four groups has its own average wage. How? Write the possibilities on paper.

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.
- When female $= 0$ & married $= 1$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.
- When female $= 0$ & married $= 1$
  - $\widehat{wage} = 5.16 + 2.82 = 7.98$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.
- When female $= 0$ & married $= 1$
  - $\widehat{wage} = 5.16 + 2.82 = 7.98$
  - So $\beta_2$ is the extra/lower wage of a married male relative to an unmarried male.

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.
- When female $= 0$ & married $= 1$
  - $\widehat{wage} = 5.16 + 2.82 = 7.98$
  - So $\beta_2$ is the extra/lower wage of a married male relative to an unmarried male.
- When female $= 1$ & married $= 1$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.
- When female $= 0$ & married $= 1$
  - $\widehat{wage} = 5.16 + 2.82 = 7.98$
  - So $\beta_2$ is the extra/lower wage of a married male relative to an unmarried male.
- When female $= 1$ & married $= 1$
  - $\widehat{wage} = 5.16 - 0.56 + 2.82 - 2.86 = 4.56$

# Example: Wage, Gender and Marital Status

- The estimated model of wage on female, married and their interaction is:

$$\widehat{wage} = 5.16 - 0.56\, female + 2.82\, married - 2.86\, female \times married$$

- When female $= 0$ & married $= 0$
  - $\widehat{wage} = 5.16$
  - So $\beta_0$ is the average wage of unmarried males.
- When female $= 1$ & married $= 0$
  - $\widehat{wage} = 5.16 - 0.56 = 4.60$
  - So $\beta_1$ is the extra/lower wage of an unmarried female relative to an unmarried male.
- When female $= 0$ & married $= 1$
  - $\widehat{wage} = 5.16 + 2.82 = 7.98$
  - So $\beta_2$ is the extra/lower wage of a married male relative to an unmarried male.
- When female $= 1$ & married $= 1$
  - $\widehat{wage} = 5.16 - 0.56 + 2.82 - 2.86 = 4.56$
  - So $\beta_3$ is the extra/lower wage of a married female beyond the two individual effects.

# Allowing for Different Slope

- So far, we have seen how belonging to one category can have an extra/lower effect on the outcome variable.

# Allowing for Different Slope

- So far, we have seen how belonging to one category can have an extra/lower effect on the outcome variable.
- What if we want to check whether belonging to a category changes the effect of an independent variable on the outcome?

# Allowing for Different Slope

- So far, we have seen how belonging to one category can have an extra/lower effect on the outcome variable.
- What if we want to check whether belonging to a category changes the effect of an independent variable on the outcome?
- For example: does an extra year of schooling have a different effect on the wages of females as compared to males?

# Allowing for Different Slope

- So far, we have seen how belonging to one category can have an extra/lower effect on the outcome variable.
- What if we want to check whether belonging to a category changes the effect of an independent variable on the outcome?
- For example: does an extra year of schooling have a different effect on the wages of females as compared to males?
- As our question suggests that the effect of education depends on gender, we can include the interaction term of gender and education.

# Allowing for Different Slope

- So far, we have seen how belonging to one category can have an extra/lower effect on the outcome variable.
- What if we want to check whether belonging to a category changes the effect of an independent variable on the outcome?
- For example: does an extra year of schooling have a different effect on the wages of females as compared to males?
- As our question suggests that the effect of education depends on gender, we can include the interaction term of gender and education.
- This allows a different slope between education and wage for males and females.

# Allowing for Different Slope

- So far, we have seen how belonging to one category can have an extra/lower effect on the outcome variable.
- What if we want to check whether belonging to a category changes the effect of an independent variable on the outcome?
- For example: does an extra year of schooling have a different effect on the wages of females as compared to males?
- As our question suggests that the effect of education depends on gender, we can include the interaction term of gender and education.
- This allows a different slope between education and wage for males and females.
- It is not necessary that the slope is different for males and females. How would that show up in our estimation?

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.
  - $\beta_0$ is the average wage of a male with zero education.

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.
  - $\beta_0$ is the average wage of a male with zero education.

- When female $= 1$

$$wage = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) education + \mu$$

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.
  - $\beta_0$ is the average wage of a male with zero education.

- When female $= 1$

$$wage = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) education + \mu$$

  - Therefore, $\beta_1 + \delta_1$ is the effect of education on the wages of females.

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.
  - $\beta_0$ is the average wage of a male with zero education.

- When female $= 1$

$$wage = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) education + \mu$$

  - Therefore, $\beta_1 + \delta_1$ is the effect of education on the wages of females.
  - $\delta_1$ is the extra/lower effect of education that a female gets as compared to a male.

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.
  - $\beta_0$ is the average wage of a male with zero education.

- When female $= 1$

$$wage = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) education + \mu$$

  - Therefore, $\beta_1 + \delta_1$ is the effect of education on the wages of females.
  - $\delta_1$ is the extra/lower effect of education that a female gets as compared to a male.
  - Similarly, $\beta_0 + \delta_0$ is the average wage of a female when education is zero.

# Allowing for Different Slope

- Consider the following model:

$$wage = \beta_0 + \delta_0 female + \beta_1 education + \delta_1 female \times education + \mu$$

- When female $= 0$

$$wage = \beta_0 + \beta_1 education + \mu$$

  - Therefore, $\beta_1$ is the effect of education on the wages of males.
  - $\beta_0$ is the average wage of a male with zero education.

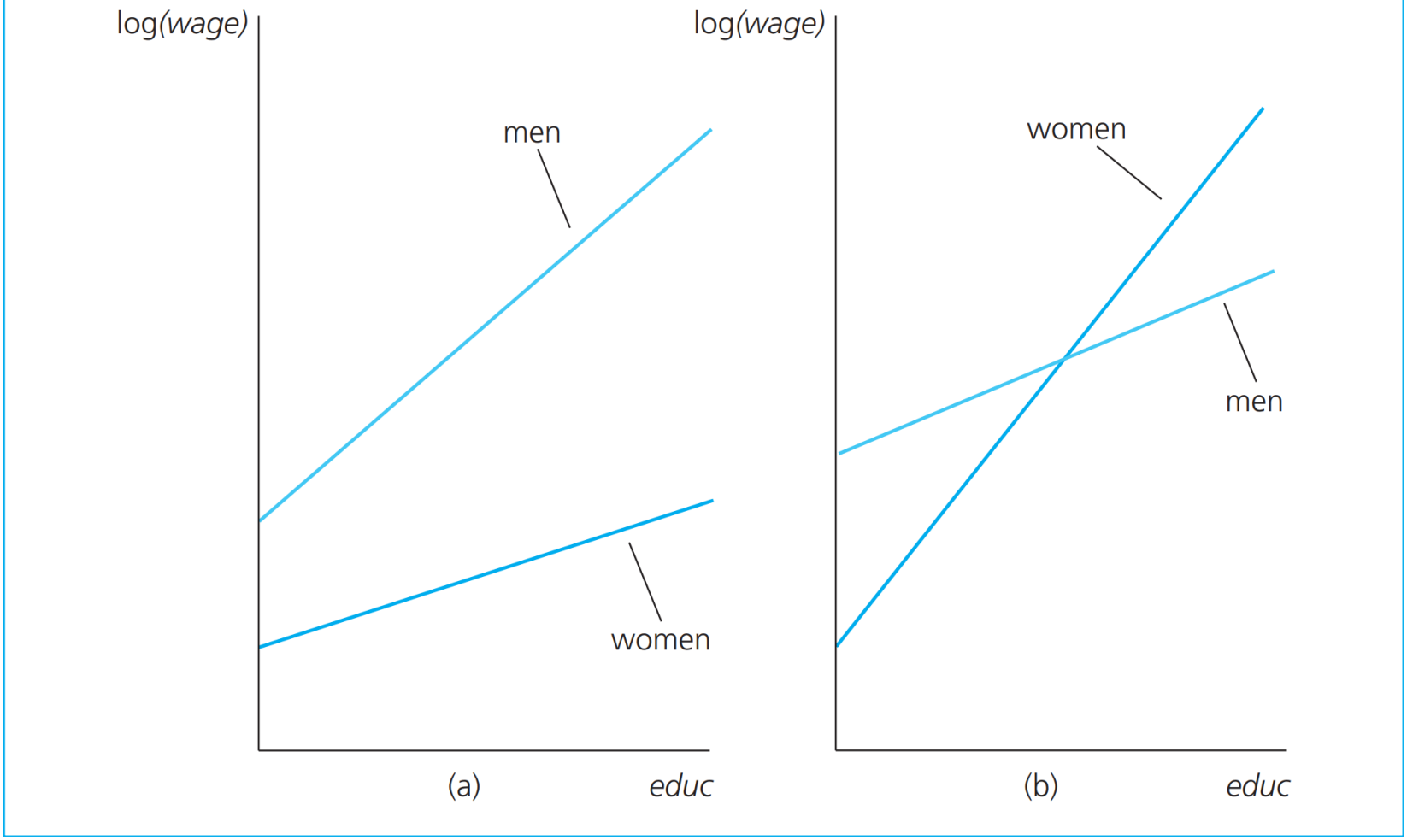
- When female $= 1$

$$wage = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) education + \mu$$

  - Therefore, $\beta_1 + \delta_1$ is the effect of education on the wages of females.
  - $\delta_1$ is the extra/lower effect of education that a female gets as compared to a male.
  - Similarly, $\beta_0 + \delta_0$ is the average wage of a female when education is zero.
  - $\delta_0$ is the extra/lower wage a female gets with no education as compared to a male with no education.

# Allowing for Different Slope

**FIGURE 7.2 Graphs of equation (7.16): (a) $\delta_0 < 0, \delta_1 < 0$; (b) $\delta_0 < 0, \delta_1 > 0$.**



# Example: Different Returns to Education

- Allowing the return to education to differ by gender gives:

$$\widehat{\log(wage)} = 0.389 - 0.227\, female + 0.082\, educ - 0.0056\, female \times educ + \ldots$$

# Example: Different Returns to Education

- Allowing the return to education to differ by gender gives:

$$\widehat{\log(wage)} = 0.389 - 0.227\, female + 0.082\, educ - 0.0056\, female \times educ + \ldots$$

- The return to education is about 8.2 percent per year for males and $8.2 - 0.56 = 7.6$ percent for females.

# Example: Different Returns to Education

- Allowing the return to education to differ by gender gives:

$$\widehat{\log(wage)} = 0.389 - 0.227\, female + 0.082\, educ - 0.0056\, female \times educ + \ldots$$

- The return to education is about 8.2 percent per year for males and $8.2 - 0.56 = 7.6$ percent for females.
- The interaction coefficient is small and statistically insignificant, so we cannot reject that the returns are equal.

# Example: Different Returns to Education

- Allowing the return to education to differ by gender gives:

$$\widehat{\log(wage)} = 0.389 - 0.227\, female + 0.082\, educ - 0.0056\, female \times educ + \ldots$$

- The return to education is about 8.2 percent per year for males and $8.2 - 0.56 = 7.6$ percent for females.
- The interaction coefficient is small and statistically insignificant, so we cannot reject that the returns are equal.
- The female coefficient $-0.227$ is the gap at $educ = 0$. Nobody in the sample has zero education, so this number has no practical meaning on its own.

# Example: Different Returns to Education

- Allowing the return to education to differ by gender gives:

$$\widehat{\log(wage)} = 0.389 - 0.227\, female + 0.082\, educ - 0.0056\, female \times educ + \ldots$$

- The return to education is about 8.2 percent per year for males and $8.2 - 0.56 = 7.6$ percent for females.
- The interaction coefficient is small and statistically insignificant, so we cannot reject that the returns are equal.
- The female coefficient $-0.227$ is the gap at $educ = 0$. Nobody in the sample has zero education, so this number has no practical meaning on its own.
- A better question is the gap at a meaningful education level, for example 12 years: $-0.227 - 0.0056 \times 12 \approx -0.29$.

# Testing for Difference in Groups

- We have studied that there could be intercept and slope differences among groups.

# Testing for Difference in Groups

- We have studied that there could be intercept and slope differences among groups.
- How do we check whether two groups differ in terms of intercept and slopes at the same time?

# Testing for Difference in Groups

- We have studied that there could be intercept and slope differences among groups.
- How do we check whether two groups differ in terms of intercept and slopes at the same time?
- This is a joint hypothesis, so it calls for an F-test.

# Testing for Difference in Groups

- We have studied that there could be intercept and slope differences among groups.
- How do we check whether two groups differ in terms of intercept and slopes at the same time?
- This is a joint hypothesis, so it calls for an F-test.
- Consider the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$$

# Testing for Difference in Groups

- We have studied that there could be intercept and slope differences among groups.
- How do we check whether two groups differ in terms of intercept and slopes at the same time?
- This is a joint hypothesis, so it calls for an F-test.
- Consider the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$$

- We want to test whether the estimated relationship is different for males and females.

# Testing for Difference in Groups

- We have studied that there could be intercept and slope differences among groups.
- How do we check whether two groups differ in terms of intercept and slopes at the same time?
- This is a joint hypothesis, so it calls for an F-test.
- Consider the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$$

- We want to test whether the estimated relationship is different for males and females.
- Let us conduct the F-test for this.

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.
2. What would be the Unrestricted Model?

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.
2. What would be the Unrestricted Model?
   - $wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + \beta_2 exper + \delta_2 female \times exper + \beta_3 tenure + \delta_3 female \times tenure + \mu$

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.
2. What would be the Unrestricted Model?
   - $wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + \beta_2 exper + \delta_2 female \times exper + \beta_3 tenure + \delta_3 female \times tenure + \mu$
   - What will be the hypothesis?

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.
2. What would be the Unrestricted Model?
   - $wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + \beta_2 exper + \delta_2 female \times exper + \beta_3 tenure + \delta_3 female \times tenure + \mu$
   - What will be the hypothesis?
   - $H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$ and $H_A$ : Not($H_0$)

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.
2. What would be the Unrestricted Model?
   - $wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + \beta_2 exper + \delta_2 female \times exper + \beta_3 tenure + \delta_3 female \times tenure + \mu$
   - What will be the hypothesis?
   - $H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$ and $H_A$ : Not$(H_0)$
3. The F-statistic can be calculated as

$$F = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{df_{ur}}} = \frac{\frac{SSR_r - SSR_{ur}}{k+1}}{\frac{SSR_{ur}}{n-2(k+1)}}$$

where $q = k + 1$ is the number of restrictions and $k$ is the number of regressors.

# F-Test for Difference in Group

1. $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$
   - Is this the Unrestricted Model?
   - No, it is the Restricted Model. It forces both groups to share the same coefficients.
2. What would be the Unrestricted Model?
   - $wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + \beta_2 exper + \delta_2 female \times exper + \beta_3 tenure + \delta_3 female \times tenure + \mu$
   - What will be the hypothesis?
   - $H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$ and $H_A$ : Not$(H_0)$
3. The F-statistic can be calculated as

$$F = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{df_{ur}}} = \frac{\frac{SSR_r - SSR_{ur}}{k+1}}{\frac{SSR_{ur}}{n-2(k+1)}}$$

   where $q = k + 1$ is the number of restrictions and $k$ is the number of regressors.
4. If $F > F_{k+1,\, n-2k-2}$ (critical value), we reject $H_0$ and say that both groups have different regression functions. Otherwise, we conclude there is no evidence of a difference between the regression functions of males and females.

# Chow Test for Difference in Groups

- The Chow test is the same F-test that we conducted before, with a different approach to calculating $SSR_{ur}$.

# Chow Test for Difference in Groups

- The Chow test is the same F-test that we conducted before, with a different approach to calculating $SSR_{ur}$.
- When we have a large number of independent variables, it gets cumbersome to add all the dummy variables and interaction terms to get the unrestricted model.

# Chow Test for Difference in Groups

- The Chow test is the same F-test that we conducted before, with a different approach to calculating $SSR_{ur}$.
- When we have a large number of independent variables, it gets cumbersome to add all the dummy variables and interaction terms to get the unrestricted model.
- Under the Chow test, we run the same regression separately for each group and get $SSR_1$ and $SSR_2$.

# Chow Test for Difference in Groups

- The Chow test is the same F-test that we conducted before, with a different approach to calculating $SSR_{ur}$.
- When we have a large number of independent variables, it gets cumbersome to add all the dummy variables and interaction terms to get the unrestricted model.
- Under the Chow test, we run the same regression separately for each group and get $SSR_1$ and $SSR_2$.
- The SSR of the unrestricted model is $SSR_{ur} = SSR_1 + SSR_2$ and $n = n_1 + n_2$. The restricted $SSR_r$ comes from the pooled regression.

# Chow Test for Difference in Groups

- The Chow test is the same F-test that we conducted before, with a different approach to calculating $SSR_{ur}$.
- When we have a large number of independent variables, it gets cumbersome to add all the dummy variables and interaction terms to get the unrestricted model.
- Under the Chow test, we run the same regression separately for each group and get $SSR_1$ and $SSR_2$.
- The SSR of the unrestricted model is $SSR_{ur} = SSR_1 + SSR_2$ and $n = n_1 + n_2$. The restricted $SSR_r$ comes from the pooled regression.
- The Chow statistic is:

$$Chow = \frac{\frac{SSR_r - (SSR_1 + SSR_2)}{k+1}}{\frac{SSR_1 + SSR_2}{n_1 + n_2 - 2(k+1)}} \sim F_{k+1,\, n-2k-2}$$

# Chow Test for Difference in Groups

- The Chow test is the same F-test that we conducted before, with a different approach to calculating $SSR_{ur}$.
- When we have a large number of independent variables, it gets cumbersome to add all the dummy variables and interaction terms to get the unrestricted model.
- Under the Chow test, we run the same regression separately for each group and get $SSR_1$ and $SSR_2$.
- The SSR of the unrestricted model is $SSR_{ur} = SSR_1 + SSR_2$ and $n = n_1 + n_2$. The restricted $SSR_r$ comes from the pooled regression.
- The Chow statistic is:

$$Chow = \frac{\frac{SSR_r - (SSR_1 + SSR_2)}{k+1}}{\frac{SSR_1 + SSR_2}{n_1 + n_2 - 2(k+1)}} \sim F_{k+1,\, n-2k-2}$$

- Caution: this form of the test assumes the error variance is the same in both groups. If it is not, use the dummy and interaction version with heteroskedasticity robust standard errors.

# Binary Variable as a Dependent Variable

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?
- Consider the following model:

$$inlf = \beta_0 + \beta_1 education + \beta_2 kidslt6 + \mu$$

where *inlf* is 1 if a woman is in the labor force and 0 if she is not, and *kidslt6* is the number of children under six years of age.

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?
- Consider the following model:

$$inlf = \beta_0 + \beta_1 education + \beta_2 kidslt6 + \mu$$

where *inlf* is 1 if a woman is in the labor force and 0 if she is not, and *kidslt6* is the number of children under six years of age.

- What will be the interpretation of $\beta_1$ and $\beta_2$?

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?
- Consider the following model:

$$inlf = \beta_0 + \beta_1 education + \beta_2 kidslt6 + \mu$$

  where *inlf* is 1 if a woman is in the labor force and 0 if she is not, and *kidslt6* is the number of children under six years of age.
- What will be the interpretation of $\beta_1$ and $\beta_2$?
- Is it that an extra year of schooling leads to a $\beta_1$ unit increase in labor force?

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?
- Consider the following model:

$$inlf = \beta_0 + \beta_1 education + \beta_2 kidslt6 + \mu$$

  where *inlf* is 1 if a woman is in the labor force and 0 if she is not, and *kidslt6* is the number of children under six years of age.
- What will be the interpretation of $\beta_1$ and $\beta_2$?
- Is it that an extra year of schooling leads to a $\beta_1$ unit increase in labor force?
- Whenever the dependent variable is a dummy, $E[y|x] = P(y = 1|x)$, so the interpretation is in terms of probability.

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?
- Consider the following model:

$$inlf = \beta_0 + \beta_1 education + \beta_2 kidslt6 + \mu$$

  where *inlf* is 1 if a woman is in the labor force and 0 if she is not, and *kidslt6* is the number of children under six years of age.
- What will be the interpretation of $\beta_1$ and $\beta_2$?
- Is it that an extra year of schooling leads to a $\beta_1$ unit increase in labor force?
- Whenever the dependent variable is a dummy, $E[y|x] = P(y = 1|x)$, so the interpretation is in terms of probability.
- Therefore, an extra year of schooling changes the probability of working by $\beta_1$ (in percentage points when multiplied by 100).

# Binary Variable as a Dependent Variable

- We have seen how a dummy independent variable changes the interpretation. What about using a dummy variable as the dependent variable?
- Consider the following model:

$$inlf = \beta_0 + \beta_1 education + \beta_2 kidslt6 + \mu$$

  where *inlf* is 1 if a woman is in the labor force and 0 if she is not, and *kidslt6* is the number of children under six years of age.
- What will be the interpretation of $\beta_1$ and $\beta_2$?
- Is it that an extra year of schooling leads to a $\beta_1$ unit increase in labor force?
- Whenever the dependent variable is a dummy, $E[y|x] = P(y = 1|x)$, so the interpretation is in terms of probability.
- Therefore, an extra year of schooling changes the probability of working by $\beta_1$ (in percentage points when multiplied by 100).
- Similarly, each additional child under six changes the probability of being in the labor force by $\beta_2$, keeping all other factors constant.

# Linear Probability Model

- OLS regression with a dummy variable as the dependent variable is known as the **Linear Probability Model (LPM)**.

# Linear Probability Model

- OLS regression with a dummy variable as the dependent variable is known as the **Linear Probability Model (LPM)**.
- The predicted value of such a model tells us the probability of the outcome being 1, given the explanatory variables.

# Linear Probability Model

- OLS regression with a dummy variable as the dependent variable is known as the **Linear Probability Model (LPM)**.
- The predicted value of such a model tells us the probability of the outcome being 1, given the explanatory variables.
- Each $\beta_j$ is the change in the probability of success when $x_j$ increases by one unit, and this change is constant across all values of $x_j$.

# Linear Probability Model

- OLS regression with a dummy variable as the dependent variable is known as the **Linear Probability Model (LPM)**.
- The predicted value of such a model tells us the probability of the outcome being 1, given the explanatory variables.
- Each $\beta_j$ is the change in the probability of success when $x_j$ increases by one unit, and this change is constant across all values of $x_j$.
- The LPM is easy to use and interpret, but it has some problems.

# Example: Women's Labor Force Participation

- A linear probability model for women's labor force participation (753 women) gives:

$$\widehat{inlf} = 0.586 - 0.0034\, nwifeinc + 0.038\, educ + 0.039\, exper - 0.0006\, exper^2 - 0.016\, age - 0.262\, kidslt6 + 0.013\, kidsge6$$

# Example: Women's Labor Force Participation

- A linear probability model for women's labor force participation (753 women) gives:

$$\widehat{inlf} = 0.586 - 0.0034\, nwifeinc + 0.038\, educ + 0.039\, exper - 0.0006\, exper^2 - 0.016\, age - 0.262\, kidslt6 + 0.013\, kidsge6$$

- An extra year of education raises the probability of working by 0.038, that is 3.8 percentage points, keeping the other factors constant.

# Example: Women's Labor Force Participation

- A linear probability model for women's labor force participation (753 women) gives:

$$\widehat{inlf} = 0.586 - 0.0034\, nwifeinc + 0.038\, educ + 0.039\, exper - 0.0006\, exper^2 - 0.016\, age - 0.262\, kidslt6 + 0.013\, kidsge6$$

- An extra year of education raises the probability of working by 0.038, that is 3.8 percentage points, keeping the other factors constant.
- Each additional child under six lowers the probability of working by 0.262, that is 26.2 percentage points.

# Example: Women's Labor Force Participation

- A linear probability model for women's labor force participation (753 women) gives:

$$\widehat{inlf} = 0.586 - 0.0034\, nwifeinc + 0.038\, educ + 0.039\, exper - 0.0006\, exper^2 - 0.016\, age - 0.262\, kidslt6 + 0.013\, kidsge6$$

- An extra year of education raises the probability of working by 0.038, that is 3.8 percentage points, keeping the other factors constant.
- Each additional child under six lowers the probability of working by 0.262, that is 26.2 percentage points.
- What happens to the predicted probability for a woman with a very low income, a lot of experience and no young children? Could it exceed 1?

# Problems with the Linear Probability Model

- The LPM has three main limitations.

# Problems with the Linear Probability Model

- The LPM has three main limitations.
  1. It does not restrict the predicted probability to lie between 0 and 1. Fitted values can be negative or above 1.

# Problems with the Linear Probability Model

- The LPM has three main limitations.
  1. It does not restrict the predicted probability to lie between 0 and 1. Fitted values can be negative or above 1.
  2. The effect of $x_j$ is constant, so the probability can change linearly forever, which cannot be true for a probability.

# Problems with the Linear Probability Model

- The LPM has three main limitations.
  1. It does not restrict the predicted probability to lie between 0 and 1. Fitted values can be negative or above 1.
  2. The effect of $x_j$ is constant, so the probability can change linearly forever, which cannot be true for a probability.
  3. It violates the homoskedasticity assumption (MLR.5), because

$$Var[y|x] = p(x)(1 - p(x))$$

depends on $x$.

# Problems with the Linear Probability Model

- The LPM has three main limitations.
  1. It does not restrict the predicted probability to lie between 0 and 1. Fitted values can be negative or above 1.
  2. The effect of $x_j$ is constant, so the probability can change linearly forever, which cannot be true for a probability.
  3. It violates the homoskedasticity assumption (MLR.5), because

$$Var[y|x] = p(x)(1 - p(x))$$

  depends on $x$.
- Heteroskedasticity does not make OLS biased, but the usual standard errors and test statistics are not valid. We use heteroskedasticity robust standard errors.

# Problems with the Linear Probability Model

- The LPM has three main limitations.
  1. It does not restrict the predicted probability to lie between 0 and 1. Fitted values can be negative or above 1.
  2. The effect of $x_j$ is constant, so the probability can change linearly forever, which cannot be true for a probability.
  3. It violates the homoskedasticity assumption (MLR.5), because

$$Var[y|x] = p(x)(1 - p(x))$$

     depends on $x$.
- Heteroskedasticity does not make OLS biased, but the usual standard errors and test statistics are not valid. We use heteroskedasticity robust standard errors.
- If we want to classify each observation as 1 or 0, we also need a rule (a probability cutoff) to turn the predicted probability into a prediction.

# Problems with the Linear Probability Model

- The LPM has three main limitations.
  1. It does not restrict the predicted probability to lie between 0 and 1. Fitted values can be negative or above 1.
  2. The effect of $x_j$ is constant, so the probability can change linearly forever, which cannot be true for a probability.
  3. It violates the homoskedasticity assumption (MLR.5), because

$$Var[y|x] = p(x)(1 - p(x))$$

  depends on $x$.
- Heteroskedasticity does not make OLS biased, but the usual standard errors and test statistics are not valid. We use heteroskedasticity robust standard errors.
- If we want to classify each observation as 1 or 0, we also need a rule (a probability cutoff) to turn the predicted probability into a prediction.
- Logit and probit models fix the first two problems. We will not cover them here.

# Policy Analysis and Program Evaluation

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).
- Consider the following model:

$$outcome = \beta_0 + \delta_0 program + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).
- Consider the following model:

$$outcome = \beta_0 + \delta_0 program + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- $\delta_0$ is the average difference in the outcome between participants and non-participants with the same values of the control variables.

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).
- Consider the following model:

$$outcome = \beta_0 + \delta_0 program + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- $\delta_0$ is the average difference in the outcome between participants and non-participants with the same values of the control variables.
- Under what condition is $\delta_0$ the causal effect of the program?

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).
- Consider the following model:

$$outcome = \beta_0 + \delta_0 program + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- $\delta_0$ is the average difference in the outcome between participants and non-participants with the same values of the control variables.
- Under what condition is $\delta_0$ the causal effect of the program?
- Only if participation is unrelated to the unobserved factors in $\mu$ once the controls are included.

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).
- Consider the following model:

$$outcome = \beta_0 + \delta_0 program + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- $\delta_0$ is the average difference in the outcome between participants and non-participants with the same values of the control variables.
- Under what condition is $\delta_0$ the causal effect of the program?
- Only if participation is unrelated to the unobserved factors in $\mu$ once the controls are included.
- If participants selected themselves into the program, for example the more motivated workers, then $\delta_0$ mixes up the program effect and self-selection.

# Policy Analysis and Program Evaluation

- Dummy variables are also the natural way to evaluate a program or a policy. We create a dummy that equals 1 for participants (the treatment group) and 0 for non-participants (the control group).
- Consider the following model:

$$outcome = \beta_0 + \delta_0 program + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- $\delta_0$ is the average difference in the outcome between participants and non-participants with the same values of the control variables.
- Under what condition is $\delta_0$ the causal effect of the program?
- Only if participation is unrelated to the unobserved factors in $\mu$ once the controls are included.
- If participants selected themselves into the program, for example the more motivated workers, then $\delta_0$ mixes up the program effect and self-selection.
- This is why randomized assignment is the best case, and why we add controls when assignment is not random.

# Discrete Dependent Variable

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?
- Similarly, given the values of the explanatory variables, suppose the predicted value is 2.7. Is it reasonable?

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?
- Similarly, given the values of the explanatory variables, suppose the predicted value is 2.7. Is it reasonable?
- In these cases, we convert the interpretation into averages or the relevant rate.

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?
- Similarly, given the values of the explanatory variables, suppose the predicted value is 2.7. Is it reasonable?
- In these cases, we convert the interpretation into averages or the relevant rate.
  - So, we can say that with an extra year of schooling the average number of children decreases by 0.3 as compared to a comparable individual. OR

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?
- Similarly, given the values of the explanatory variables, suppose the predicted value is 2.7. Is it reasonable?
- In these cases, we convert the interpretation into averages or the relevant rate.
  - So, we can say that with an extra year of schooling the average number of children decreases by 0.3 as compared to a comparable individual. OR
  - On average, 3 extra years of schooling lead to 1 fewer child. OR

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?
- Similarly, given the values of the explanatory variables, suppose the predicted value is 2.7. Is it reasonable?
- In these cases, we convert the interpretation into averages or the relevant rate.
  - So, we can say that with an extra year of schooling the average number of children decreases by 0.3 as compared to a comparable individual. OR
  - On average, 3 extra years of schooling lead to 1 fewer child. OR
  - With an extra year of schooling, the fertility rate decreases by 0.3.

# Discrete Dependent Variable

- Although discrete variables are not qualitative variables, using them as a dependent variable has implications for interpretation.
- Consider the following model, where the dependent variable is the number of children.

$$numchild = \beta_0 + \beta_1 education + \beta_2 income + \mu$$

- What is the interpretation of $\beta_1$ if $\beta_1 = -0.3$?
- Is it that an extra year of schooling will decrease the number of children by 0.3?
- Similarly, given the values of the explanatory variables, suppose the predicted value is 2.7. Is it reasonable?
- In these cases, we convert the interpretation into averages or the relevant rate.
  - So, we can say that with an extra year of schooling the average number of children decreases by 0.3 as compared to a comparable individual. OR
  - On average, 3 extra years of schooling lead to 1 fewer child. OR
  - With an extra year of schooling, the fertility rate decreases by 0.3.
  - The same discussion applies to the predicted values. A prediction of 2.7 is an expected number of children, not a count we will ever observe for one person.

# Summary

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.
- With $g$ categories, include $g - 1$ dummies to avoid the dummy variable trap. The omitted category is the base group.

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.
- With $g$ categories, include $g-1$ dummies to avoid the dummy variable trap. The omitted category is the base group.
- With a log dependent variable, $100 \times \delta$ is an approximation, and $100(e^{\delta}-1)$ is the exact percentage difference.

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.
- With $g$ categories, include $g-1$ dummies to avoid the dummy variable trap. The omitted category is the base group.
- With a log dependent variable, $100 \times \delta$ is an approximation, and $100(e^{\delta}-1)$ is the exact percentage difference.
- Interaction of two dummies creates subgroups. Interaction of a dummy and a numerical variable changes the slope.

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.
- With $g$ categories, include $g-1$ dummies to avoid the dummy variable trap. The omitted category is the base group.
- With a log dependent variable, $100 \times \delta$ is an approximation, and $100(e^{\delta}-1)$ is the exact percentage difference.
- Interaction of two dummies creates subgroups. Interaction of a dummy and a numerical variable changes the slope.
- To test whether two groups follow the same regression, use an F-test on all the group dummy and interaction terms, or equivalently the Chow test.

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.
- With $g$ categories, include $g-1$ dummies to avoid the dummy variable trap. The omitted category is the base group.
- With a log dependent variable, $100 \times \delta$ is an approximation, and $100(e^{\delta}-1)$ is the exact percentage difference.
- Interaction of two dummies creates subgroups. Interaction of a dummy and a numerical variable changes the slope.
- To test whether two groups follow the same regression, use an F-test on all the group dummy and interaction terms, or equivalently the Chow test.
- With a binary dependent variable, OLS gives the Linear Probability Model. Read coefficients as changes in probability and use robust standard errors.

# Summary

- A dummy variable shifts the intercept. Its coefficient is the difference from the base group, holding the other variables constant.
- With $g$ categories, include $g-1$ dummies to avoid the dummy variable trap. The omitted category is the base group.
- With a log dependent variable, $100 \times \delta$ is an approximation, and $100(e^{\delta}-1)$ is the exact percentage difference.
- Interaction of two dummies creates subgroups. Interaction of a dummy and a numerical variable changes the slope.
- To test whether two groups follow the same regression, use an F-test on all the group dummy and interaction terms, or equivalently the Chow test.
- With a binary dependent variable, OLS gives the Linear Probability Model. Read coefficients as changes in probability and use robust standard errors.
- A program dummy gives the causal effect only if participation is unrelated to the unobserved factors.