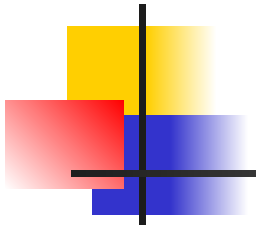
Statistics and Quantitative Analysis U4320

Segment 12:

Extension of Multiple Regression Analysis

Prof. Sharyn O'Halloran



Key Points

- **Dummy Variables**
  - **Shift in the intercept**
- **Difference of Means Test**
  - **Are two groups significantly different**
- **Interactive Terms**
  - **Shift in the Slope**

I. Introduction to Dummy Variables

- A. Definition:

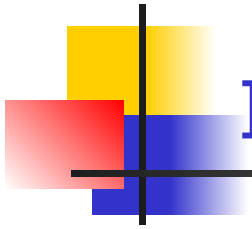
  Any dichotomous variable that is coded zero-one

  - Examples:

    - Gender: 1 for female
      O for male

    - Employment: 1 if employed
      0 if unemployed

    - Religion: 1 if Catholic
      0 otherwise



I. Introduction to Dummy Variables

- B. Example
  - Suppose that a certain drug is suspected of raising blood pressure.
  - One way to test this hypothesis is to conduct a controlled experiment.
  - For example, suppose we randomly sample 10 women, 6 take the drug, 4 do not take the drug.

I. Introduction to Dummy Variables (cont.)

- B. Example
  - 1. Variables
    - We represent this information by a dummy variable.
      - Y = Level of Blood Pressure
      - D = number of doses of this drug daily
        - D = 1 if she took the drug
        - D = 0 if she did not (i.e., was a control)

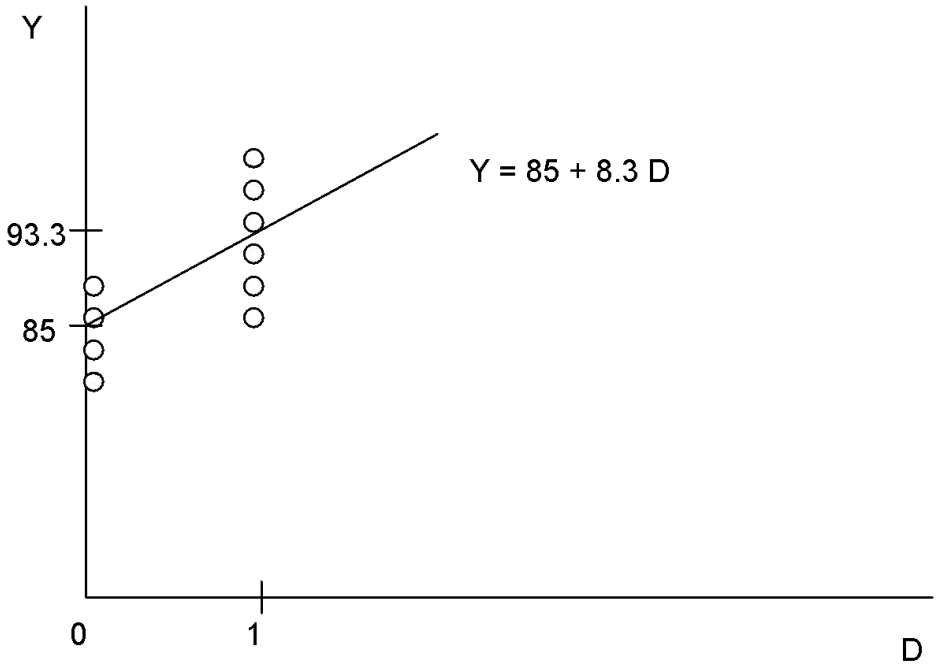
I. Introduction to Dummy Variables (cont.)

- B. Example
  - 2. Data

OBS	PRESSURE	DRUG
1	85	0
2	95	1
3	90	1
4	75	0
5	100	1
6	90	0
7	90	0
8	90	1
9	100	1
10	85	1

I. Introduction to Dummy Variables (cont.)

- B. Example
  - 3. Graph



I. Introduction to Dummy Variables (cont.)

- B. Example
  - 4. Results
    - There are four observations when D = 0. The mean of these observations is 85.
    - There are six observations when D = 1. The mean of these observations is 93.3.

I. Introduction to Dummy Variables (cont.)

- B. Example
  - 4. Results (cont.)
    - So how do we get a regression line from these data?
      - To draw the regression line, we just connect the means of each group!
      - This minimizes the sum of the squared distances from the data points to the line.

I. Introduction to Dummy Variables (cont.)

- B. Example
  - 4. Results (cont.)
    - The equation would be:

$$Y = 85 + 8.3\,Drug.$$
$$(3.22)^{*} \quad (4.16)$$

I. Introduction to Dummy Variables (cont.)

- B. Example
  - 5. Interpretation
    - The intercept is the average blood pressure level for those who did not take the drug
      - The group of all people who did not take the drug is called the reference group, or the control group.
      - The intercept is the average of the dependent variable when the dummy variable is 0.

I. Introduction to Dummy Variables (cont.)

- B. Example
  - 5. Interpretation (cont.)
    - The coefficient on the Dummy variable states by how much an individual's blood pressure increases, on average, when given the drug.

I. Introduction to Dummy Variables (cont.)

- C. Difference of Means Test
  - 1. Slope
    - What is the difference between the means when the sample is divided into those who took the drug and those who did not?
      - As the dummy variable goes from 0 to 1, the value of Y rises by 8.33 points.
        - The difference between the means is 8.33.

I. Introduction to Dummy Variables (cont.)

- C. Difference of Means Test
  - 2. Significant Difference
    - Now, how can we tell if this slope is significantly different from zero?
      - a. Hypothesis

$$H_0: \beta = 0$$
$$H_a: \beta > 0$$

I. Introduction to Dummy Variables (cont.)

- C. Difference of Means Test
  - 2. Significant Difference (cont.)
    - b. t-statistic
    - c. Accept of Reject
      - the variable is significant at the 5% level.

$n=10; d.f. = 8; \alpha = .05$

$t=8.3/4.16=2.00$

I. Introduction to Dummy Variables (cont.)

- C. Difference of Means Test
  - 3. Interpretation
    - What does it mean that the slope is significant?
    - People who took the drug did have a significantly higher blood pressure than those who didn't.

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - Does education vary according to a respondent's religion?
  - I investigated whether Catholics had more or less education than the general population.
- 1. Hypothesis

$$H_0: \beta = 0; \quad H_a: \beta \neq 0$$

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis (cont.)
    - 1. Commands
      - I made my own variable called Catholic.
      - I recoded it and gave it value labels, so that all Catholics were coded 1 and all others got a 0.
      - I then made a variable for education, called Smarts.
      - Finally, I specified my regression.

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 2. Regression Results

Variable	B	SE B	Beta	T	Sig T
CATHOLIC	.314567	.178289	.047002	1.764	.0779
(Constant)	12.755382	.093351		136.639	.0000

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 2. Regression Results (cont.)
      - What was the average education level of non-Catholics in the population?
      - What's the average education level of Catholics?

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 3. Significance
      - Is the coefficient significant at the 5% level?

$$\frac{3.145}{0.178} = 1.764$$

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 4. Difference of Means Results
      - For comparison, I did a difference of means test on the same data.
        - **Is There A Significant Difference Between Catholics**

        And The Rest Of The Population?

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 4. Difference of Means Results (cont.)
      - **T-TEST /GROUPS CATHOLIC (0,1) /VARIABLES SMARTS.**

      - Independent samples of CATHOLIC
      - Group 1: CATHOLIC EQ .00 Group 2: CATHOLIC EQ 1.00
      - t-test for: SMARTS

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 4. Difference of Means Results (cont.)

	Number of Cases	Mean	Standard Deviation	Standard Error
Group 1	1022	127554	3.031	.095
Group 2	386	130699	2.855	.145

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 4. Difference of Means Results (cont.)

		Pooled Variance Estimate			Separate Variance Estimate		
F Value	2-Tail Prob	t Value	Degrees of Freedom	2-Tail Prob	t Value	Degrees of Freedom	2-Tail Prob
1.13	.165	-1.76	1406	.078	-1.81	732.50	.070

I. Introduction to Dummy Variables (cont.)

- D. SPSS Example 1:
  - 1. Hypothesis
    - 5. Interpretation
      - The results suggest that there is no significant difference between Catholics and the rest of the population.

II. More than One Dummy Variable

- A. Sample Equation
  - If we had many different drugs that we thought might cause high blood pressure, then we could write:

$$Y = b_0 + b_1 D_1 + b_2 D_2 + b_3 D_3 + \ldots$$

  - Then b_1, b_2, b_3, the slope coefficient on the dummy variables, are the difference between the blood pressure of someone taking that drug and a control group which takes none of the drugs.

II. More than One Dummy Variable (cont.)

- B. 1. Example: More than One Dummy Variable
  - Say that we rerun the regression from before, but we allow religion to be either Protestant, Catholic, Jewish, or Other.
  - We create separate variables for each category.

II. More than One Dummy Variable (cont.)

- B. 1. Example: More than One Dummy Variable (cont.)

Variables:

Catholic	1 if yes 0 otherwise
Protestant	1 if yes 0 otherwise
Jewish	1 if yes 0 otherwise
Other	1 if other than catholic, Protestant, etc. 0 otherwise

II. More than One Dummy Variable (cont.)

- 2. Question:
  - Can we run a regression that looks like this:

$Y = b_0 + b_1$ Catholic $+ b_2$ Protestant, $+ b_3$ Jewish $+ b_4$ Other ?

II. More than One Dummy Variable (cont.)

- 3. Model Specification
  - Let's have Catholic be our base group.
  - Then the correct equation is:

$Y = b_0 + b_1$ Protestant $+ b_2$Jewish $+ b_3$Other.

II. More than One Dummy Variable (cont.)

- C. Difference of Means
  - 1. Results

Variable	B	SE B	Beta	T	Sig T
OTHER	1.654190	.564760	.078695	2.929	.0035
JEWISH	2.773802	.539575	.138466	5.141	.0000
PROTSTNT	-.476816	.176751	-.074344	-2.698	.0071
(Constant)	13.069948	.149293		87.546	.0000

II. More than One Dummy Variable (cont.)

- C. Difference of Means
  - 2. Interpretation
    - The intercept is the average education of the reference group.
      - It is **13.06**. This agrees with our earlier results when we just did Catholic as our single category.
    - Slopes is the difference in years of education from the base case.

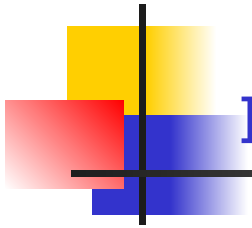
II. More than One Dummy Variable (cont.)

- C. Difference of Means
  - 3. Differences between other groups
    - We can also use these results to tell the difference between any two groups.
    - For instance, the education difference between **Jewish and Other** is:

2.77- 1.64=1.13.

II. More than One Dummy Variable (cont.)

- C. Difference of Means
  - 3. Differences between other groups (cont.)
    - However, to tell if this difference is significant, we'd have to re-run the regression, using either Jewish or Other as the reference group.



III. Dummy Variables and Confounding Variables

- We may think that our results are confounded by other factors.

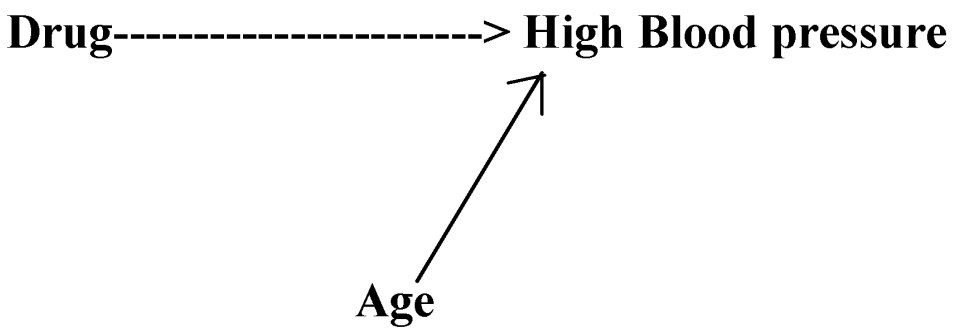
A. Parallel Lines for Two Categories

1. Blood Pressure Example

- What we want to do is compare the blood pressure of those women who take the drug against those who do not.
- In testing the effect of the drug, however, we will also want to control for age.

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 2. Path Diagram



  - 3. Data

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 3. Data

OBS	PRESSURE	AGE	DRUG
1	85	30	0
2	95	40	1
3	90	40	1
4	75	20	0
5	100	60	1
6	90	40	0
7	90	50	0
8	90	30	1
9	100	60	1
10	85	30	1

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 4. Results
    - a. Regression
      - Suppose after we run a regression line then the results look like this
      - Blood Pressure = $b_0 + b_1$AGE + b_2DRUG

VARIABLE	COEFFICIENT	ST. DEV.	T-STATISTIC
CONSTANT	69.535	2.905	23.93
AGE	0.44186	0.07301	6.05
DRUG	4.651	1.885	2.47

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 4. Results
    - a. Regression (cont.)
      - Our regression equation is

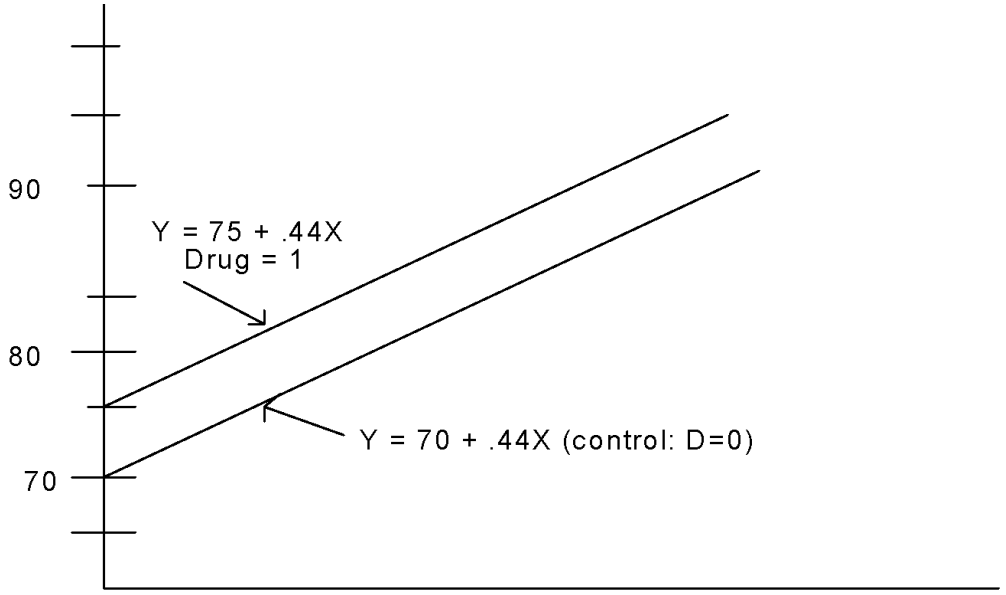
$$Y = 69.5 + 0.44(\text{Age}) + 4.65(\text{Drugs})$$

$$Y = 69.5 + 0.44(\text{AGE}) \qquad D = 0$$

$$Y = 69.5 + 0.44(\text{AGE}) + 4.65 \qquad D = 1$$

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 4. Results
    - b. Graph
      - What is the regression line then, for those women who take the drug?



III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 4. Results
    - c. Interpretation
      - How do we interpret the coefficient?
        - The coefficient on Drugs is the change in Blood pressure that accompanies a unit change in Drugs, while age remains constant.

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - 4. Results
    - c. Interpretation
      - How do we interpret the coefficient? (cont.)
        - There is an increase in blood pressure of 5 units as we go from a woman without the drug (D=0) to a woman of the same age with the drug (D=1).

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - **5. Construct a 95% confidence interval**
    - a. State hypothesis

$$H_o: \beta = 0$$

$$H_a: \beta \neq 0$$

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - **5. Construct a 95% confidence interval**
    - b. Confidence Interval

$$\beta = b \_ t_{.025} \frac{s}{\sqrt{\square x^2}}.$$

$$\beta = b \_ t_{.025} * SE$$

$$d.f. = n\text{-}k1 = 10\text{-}2\text{-}1 = 7$$

$$\beta = 4.7 \_ 2.36 * (1.88)$$

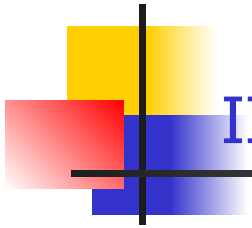
$$4.7 \_ 4.4$$

III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - **5. Construct a 95% confidence interval**
    - c. Accept or Reject
      - So can we reject the null hypothesis that the difference is 0?

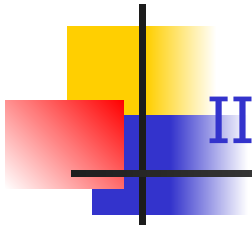
III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - **5. Construct a 95% confidence interval**
    - d. Interpretation
      - This suggests that on average women taking the drug have significantly higher blood pressure than those who do not.



III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - **5. Construct a 95% confidence interval**
    - e. Note
      - We should note, however, that this does not suggest that the drug **causes** high blood pressure.



III. Dummy Variables and Confounding Variables (cont.)

- A. Parallel Lines for Two Categories
  - **5. Construct a 95% confidence interval**
    - e. Note (cont.)
      - Other confounding factors, such as stress levels, eating and exercise habits, may also influence blood pressure.

III. Dummy Variables and Confounding Variables (cont.)

- B. Parallel Lines for Several Categories
  - 1. Data
    - What if we are now testing two drugs, A and B, against a control C, with a sample of 30 patients.
    - In measuring drugs, we use two dummy variables.
      - D_A = 1 if drug A given; 0 otherwise
      - D_B = 1 if drug B given ; 0 otherwise

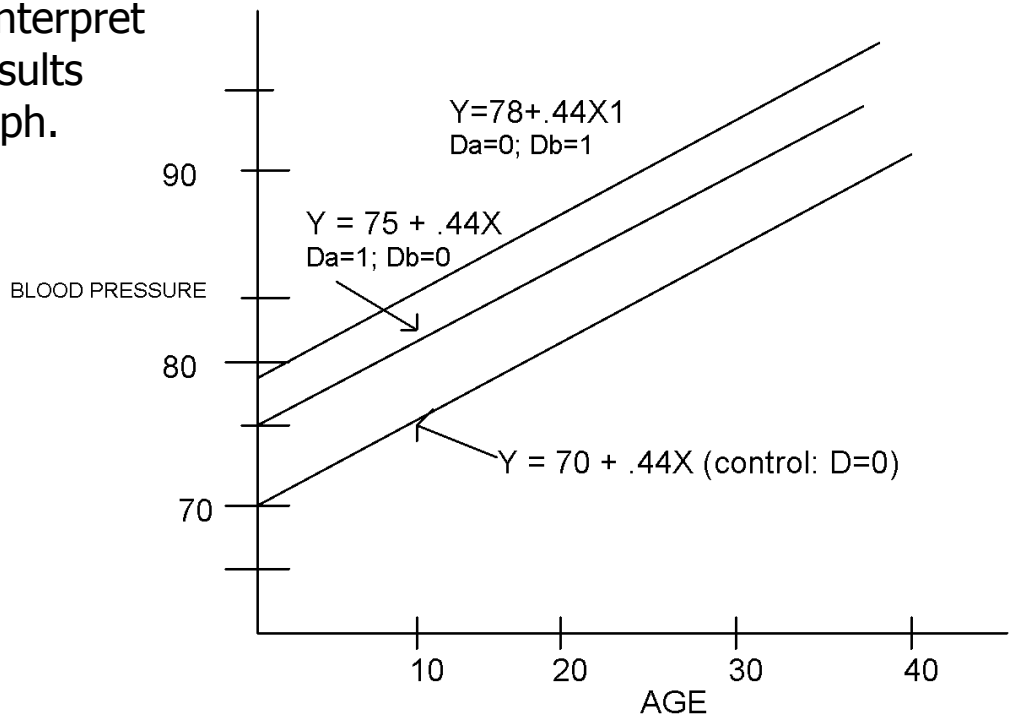
III. Dummy Variables and Confounding Variables (cont.)

- B. Parallel Lines for Several Categories
  - 2. Estimates

    - Suppose the regression line turned out to be:
      - $Y = 70 + 5D_A + 8D_B + .44AGE + ...$
    - From this regression, we can easily write down three separate equations for the three groups:
      - For the control group, set $D_A = 0$, $D_B = 0$: $Y = 70 + .44(AGE)$
      - For the group on Drug A, set $D_A = 1$, $D_B = 0$: $Y = 75 + .44(AGE)$
      - For the group on Drug B, set $D_A = 0$, $D_B = 1$: $Y = 78 + .44(AGE)$

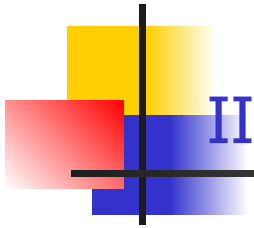
III. Dummy Variables and Confounding Variables (cont.)

- B. Parallel Lines for Several Categories
  - 3. Graph
    - The easiest way to interpret these results by a graph.



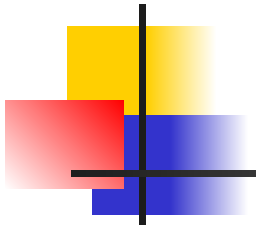
III. Dummy Variables and Confounding Variables (cont.)

- B. Parallel Lines for Several Categories
  - 4. Interpretation
    - The results show that the group on drug A exceeds the controls by 5 units-- the coefficient on D_A.
    - Similarly, the group on Drug B exceeds the controls by 8 units-- the coefficient of D_B.



III. Dummy Variables and Confounding Variables (cont.)

- B. Parallel Lines for Several Categories
  - NOTE
    - Again, in designing our models we need to make sure that we identify our equation. That is, we need to have a least one category left as a reference group.



IV. Different Slopes as Well as Intercepts: Interactive Terms

- A. Interactive Terms
  - What if the relation between age and the drug is not additive but multiplicative?
    - That is, not only does the intercept change, but so does the slope of the regression line.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

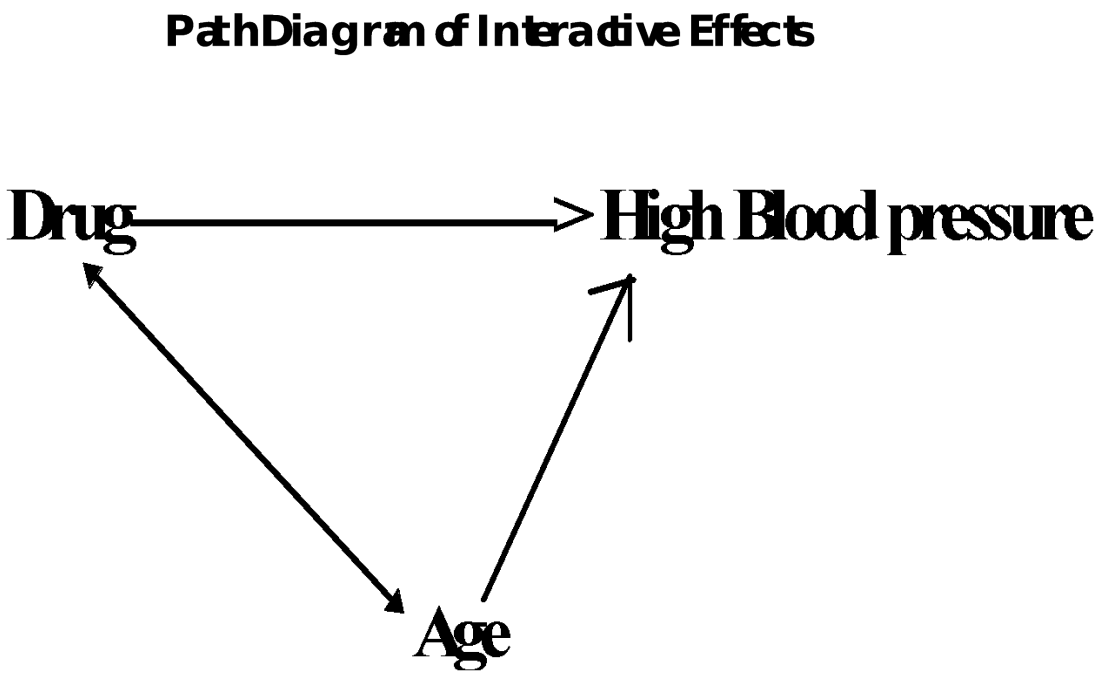
- A. Interactive Terms
  - 1. Definition
    - What if the drug had a greater impact on your blood pressure as a person gets older.
      - We can investigate this possibility by adding a multiplicative term to the regression.
        - Y = c + Drug + Age +Age*Drug .....
  - The term Age*Drug is called an <u>interactive term</u>.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 1. Definition(cont.)
    - Interactive terms captures the possibility that the effect of one independent variable might vary with the level of another independent variable.
      - In this case, the effect of the drug on your blood pressure depends on your age.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 1. Definition (cont.)



IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 1. Definition (cont.)
    - 2. Data

OBS	PRESSURE	AGE	DRUG	Age*Drug
1	85	30	0	0
2	95	40	1	40
3	90	40	1	40
4	75	20	0	0
5	100	60	1	60
6	90	40	0	0
7	90	50	0	0
8	90	30	1	30
9	100	60	1	60
10	85	30	1	30

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 2. Results
    - Suppose that when we run a regression, we get the following results:
    - Again we set D = 0 for the control group and D = 1 for those taking the drug.

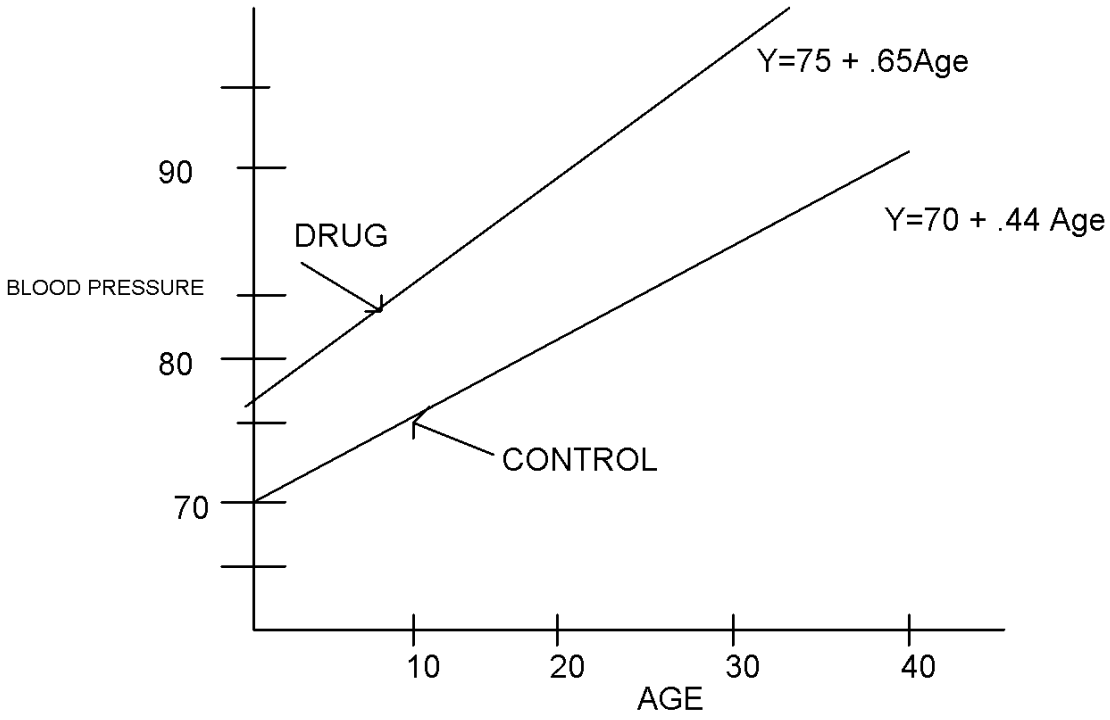
      Y = 70 + 5(Drug) + .44(Age) + .21(Drug*Age)
    - We obtain two separate equations for the two groups:

      set D = 0: Y = 70 + .44Age

      Set D = 1: Y = 75 + .65Age

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 3. Graph



IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 4. Interpretation
    - Note that for those taking the drug not only does the intercept increase (that is, the average level of blood pressure), but so does the slope.
    - Interpretation of an interactive term -- The effect of one independent variable (DRUG) depends on the level of another independent variable (AGE).

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- A. Interactive Terms
  - 1. Definition
  - 2. Results
  - 3. Graph
  - 4. Interpretation (cont.)
    - The results here suggest that for people not taking the drug, each additional year adds .44 units to blood pressure.
    - For people taking the drug, each additional year increases blood pressure by .65 units.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 1. Setup

PARTYID = C + MONEY + MALE + MONEY*MALE (MANBUCKS)

- What is the interpretation of this interactive variable between Money and Gender?
  - As a male's income increases are they more likely to identify themselves as a Republican than are women?

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 2. Commands
    - The next two pages of your handout give the results.
    - First, I made variables for everything that I want to analyze.
    - MYPARTY is party ID.
    - MONEY is income.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 2. Commands (cont.)
    - MALE is whether or not you're a male.
    - And MANBUCKS is the interactive term for MALE and MONEY.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 3. Results
    - 1. No Interactive Effect
      - Notice that both MALE and MONEY are significant determinants of Party ID.

Equation Number 1 Dependent Variable.. MYPARTY

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
MALE	.391620	.113794	.093874	3.441	.0006
MONEY	.046016	.010763	.116615	4.275	.0000
(Constant)	2.112151	.155388		13593	.0000

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 3. Results
    - 2. Interactive Effect
      - Now we put in the interactive term. Is it significant?

Equation Number 2 Dependent Variable.. MYPARTY

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
MANBUCKS	5.009256E-04	.022138	.001915	.023	.9820
MONEY	.045823	.013720	.116128	3.340	.0009
MALE	.384671	.327505	.092209	1.175	.2404
(Constant)	2.114600	.189432		11.163	.0000

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 3. Results
    - 3. Presentation of Results

Model Variables	Model II	Model III
Constant	2.11** (0.15)	2.11 (0.19)**
MALE	0.39** (0.11)	0.38 (0.33)**
MONEY	0.046** (0.01)	0.046 (0.014)
MANBUCKS		0.0005 (0.02)

Standard errors in parentheses; Significant at $\alpha < .05$

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 3. Results
    - 4. Interpretation
      - So we conclude that as men and women' s income increases, they become Republican at the same rate.
      - Men and women are equally as likely to become Democrats as they get poor.

IV. Different Slopes as Well as Intercepts: Interactive Terms (cont.)

- B. Example 3:
  - 3. Results
    - 5. Notice
      - Notice that the t-statistics for the other two variables went haywire when we added the interactive term into the equation.
      - This is an important general point; you can use the t-value for the interactive term, but the other t-values become meaningless.

VII. Term Paper

- 1. Clearly state your hypothesis.
  - Use a path diagram to present the causal relation.
  - Use the correlations to help you determine what causes what.
  - State the alternative hypothesis.
- 2. Present descriptive statistics.

VII. Term Paper (cont.)

- 3. Estimate your model.
  - You can do simple regression, or include interactive terms, or do path analysis, or use dummy variables; whatever is appropriate to your hypothesis.
- 4. Present your results.
- 5. Interpret your results.

VII. Term Paper (cont.)

- 6. Draw out the policy implications of your analysis.

- 7. The paper should begin with a brief which states the basic project and your main findings.