Stata Tips and Tricks: Data Analysis

Hsueh-Sheng Wu

CFDR Workshop Series

June 5, 2023

BGSU Center for Family and Demographic Research

Outline

- The objective of statistical analysis
- Three types of research questions
- Model building for various research questions
- Techniques for testing relations among variables
- Techniques for correcting methodological issues
- Procedures for conducting analysis
- A Stata example
- Conclusions

The Objective of Statistical Analysis

- The purpose of statistical analysis is to answer research questions satisfactorily by applying statistical techniques and procedures to data and extracting valid evidence.

- To meet this objective, scholars must understand:
  - What are their research questions?
  - What are their hypotheses?
  - What evidence is required to support or refute study hypotheses?
  - What statistical procedures are available to obtain such evidence?
  - What methodological issues may arise as a result of sampling design, data, and variable construction, and what strategies can be used to minimize or mitigate these issues?

- Important principles of statistical analysis
  - Study design always focuses on providing the most valid and reliable evidence to test research hypotheses and answer research questions.
  - Data analyses comprise techniques addressing theoretical or methodological issues, and these techniques do not always work with each other. Thus, it is critical for researchers to select adequate techniques and plan on how to apply these techniques before undertaking the actual data analysis.

Three Types of Research Questions

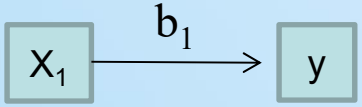
- Social science researchers generally examine three types of research questions:
  - What variables predict the outcome variable?
  - What variables mediate the relations between the independent variables and the outcome variable?
  - What variables moderate the relations between the independent variables and the outcome variables?

- Different types of evidence are needed to address these questions

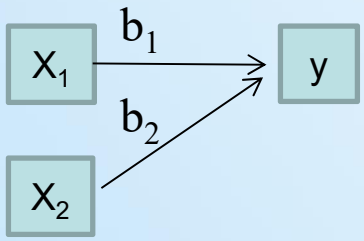
Model Building for Various Research Questions

- Model Building for Identifying if X_2 is a Significant Predicator of the Outcome Variable Y

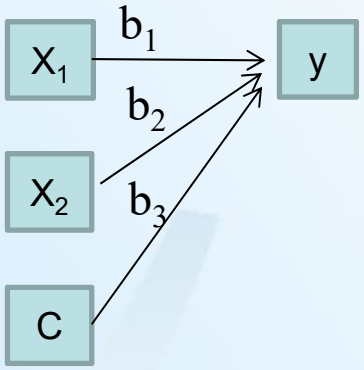
1.1 Test if X_1 predicts Y



1.2 Test if X_2 predicts Y when X_1 is already in the model
1.3 Test if b_1 equals b_2



1.4. Test if X_1 and X_2 still predict Y when C is in the model
1.5 Test if b_1 equals b_2 when C is in the model



Model Building for Various Research Questions (Cont.)

- Model Building for Identifying A Significant Variable that Mediates the Relation between the Independent Variable X and the Outcome Variable Y

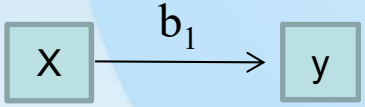
2.1 Test if X predicts Y

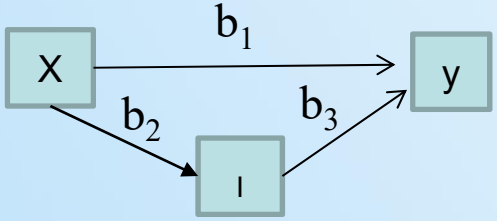
2.2. Test if the direct path linking X to Y is significant
2.3 Test if the indirect path linking X to I and then to Y is significant
2.4 Test which path has more significant impact on Y

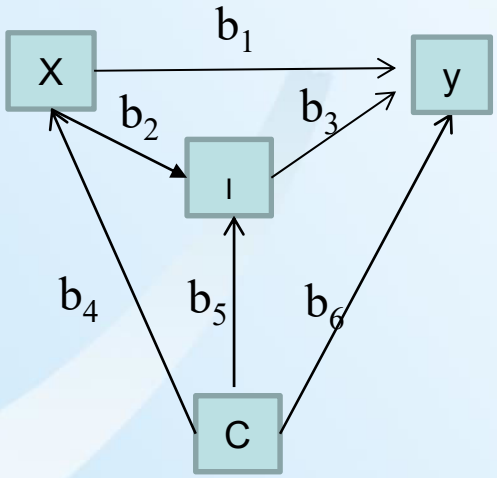
2.5. Test if the direct path linking X to Y is significant when C is already in the model.

2.6 Test if the indirect path linking X to I and then to Y is when C is already in the model.

2.7 Test which path has more significant impact on Y when C is already in the model.







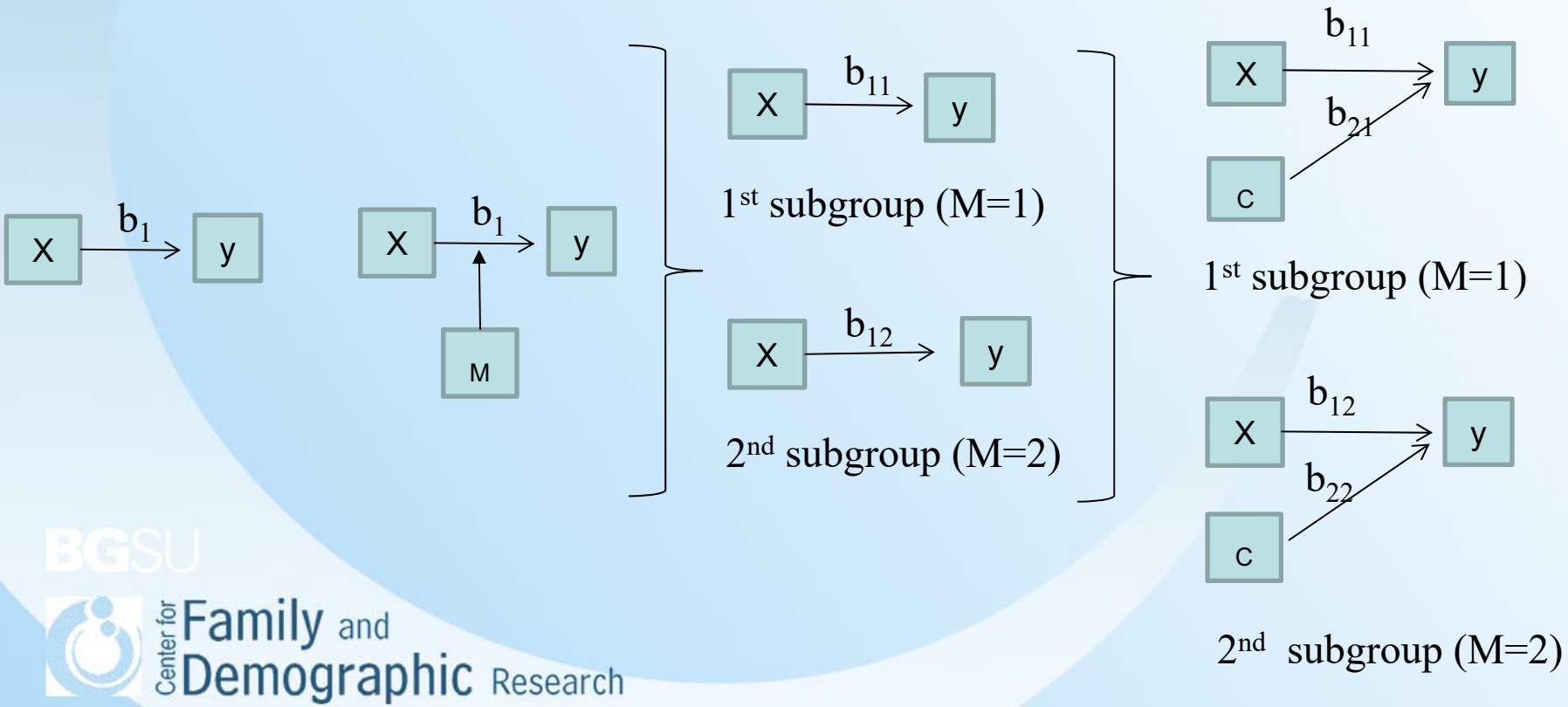
Model Building for Various Research Questions (Cont.)

- Two Types of Model Building for Identifying a Significant Variable that Moderates the relation between the Independent Variable X and the Outcome Variable Y

3.1.1 Test if X predicts Y

3.1.2 Test if M modifies the relation between X and Y, which is equivalent to examining if b_{11} and b_{12} are the same

3.1.3. Test if M still modifies the relation between X and Y (if $b_{11} = b_{12}$), when C is controlled for within the 1st and 2nd subgroups.

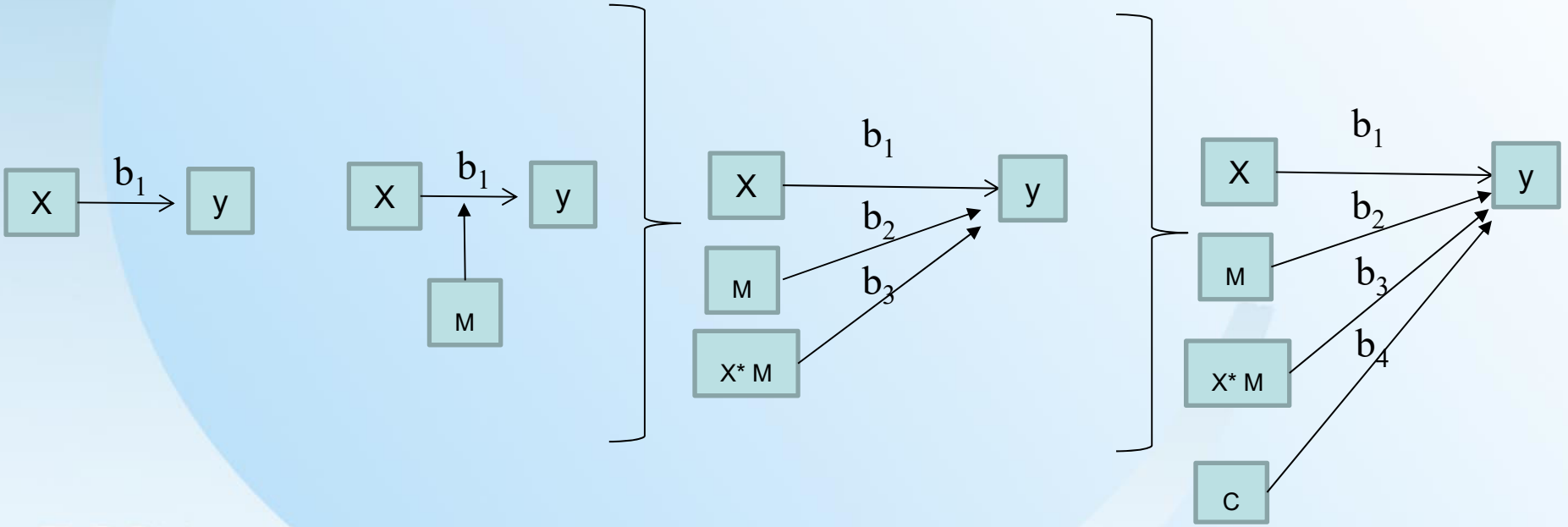


Model Building for Various Research Questions (Cont.)

3.2.1. Test if X predicts Y

3.2.2 Test if M modifies the relation between X and Y, which is equivalent to examining if the total effect of X on Y differ significantly, depending on the levels of M (i.e., (b_1*X) vs (b_1*X + b_3*X*M)

3.3.3. Test if M still modifies the relation between X and Y (i.e., (b_1*X) vs (b_1*X + b_3*X*M) , when C is controlled for within the 1st and 2nd subgroups.



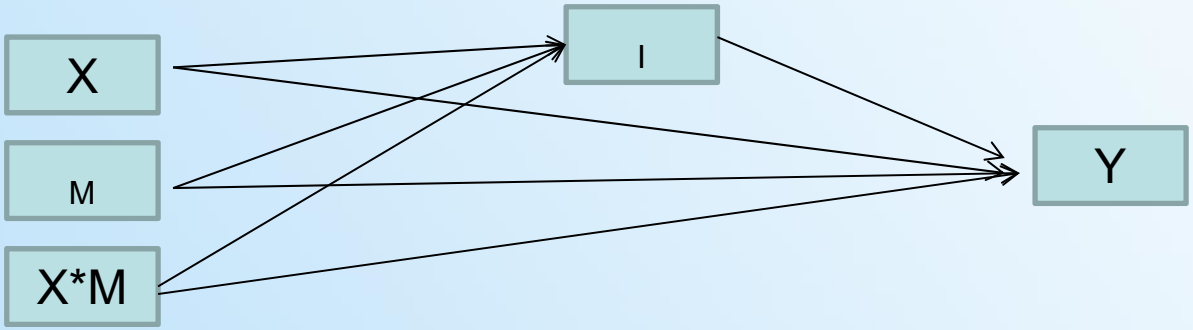
The whole sample

The whole sample

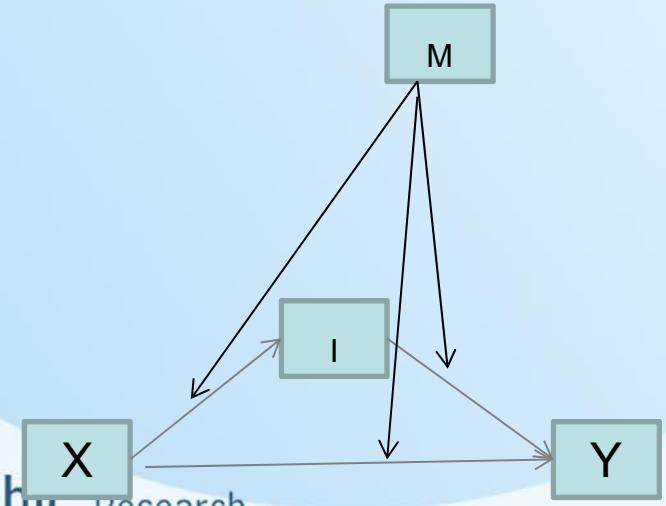
Model Building for Various Research Questions (Cont.)

Extension of these Research Questions

- Mediated Moderation Effects



- Moderated Mediation Model



Techniques for Testing Relations among Variables

Sociologists frequently employ regression-based techniques to test theoretical hypotheses, including:

- Single-level analysis with one dependent variable: OLS regression, Logit/probit regression, ordered logit regression, multinomial logistic regression, passion regression, negative binomial regression, discrete time event history analysis
- Single-level analysis with two or more dependent variables: path analysis, structural equation modeling
- Single-level analysis results from different groups: structural equation modeling and single-level analyses described above
- Multiple-level analysis with one or more dependent variables at each level: structural equation modeling and hierarchical linear model

Techniques for Correcting Methodological Issues

- While establishing the theoretical relation between independent variable and dependent variables is at the heart of data analysis, biased results could occur if methodological concerns related with sampling design, missing data, and arbitrary variable operationalizations are not handled appropriately.
- Respondents in a survey may have different weights because their chances of getting selected into the sample and continuing to participate in the survey differ. To address this issue, use the Stata survey command -svy-.
- Some survey questions may be refused by respondents. The Stata multiple imputation command -mi- can be used to make the most of the available data and boost statistical power.
- Researchers may sometimes need to arbitrarily operationalize a variable, such as determining the cut-off points on an income scale in order to categorize respondents as low, middle, or high income. Then, sensitivity testing with various cut-off values should be carried out.

Procedure of Conducting Analysis

The following procedure suggests how to perform data analysis.

- Clearly define research topics and hypotheses.
- Determine the evidence required to test these hypotheses.
- Determine the statistical procedures to employ based on the number and measurement level of dependent variables, as well as whether single- or multi-level analyses are required.
- Examine the sampling strategy, data, and variable constructs to determine what methodological difficulties must be addressed when undertaking data analysis.
- Determine the sequence of procedures to be employed, for example.
  - Step 1: To lessen the impact of missing data on analysis findings, use the -mi- command.
  - Step 2: Use the -SVY- command to specify that the data is collected using a complicated survey design and that the analysis findings will be weighted.
  - Step 3: Provide descriptive statistics and bivariate variable relationships.
  - Step 4: Perform a series of regression analyses to put the theoretical questions to the test.
  - Step 5: Examine the results and determine whether the model needs to be modified.
  - Step 6: If the model has to be modified, repeat steps 3, 4,5 until satisfactory evidence is obtained..
  - Step 7: if there is no need to adjust the model and an independent or outcome variable is arbitrarily formed, conduct a sensitivity test
  - Step 8: Compare results of Step 6 and 7 and decide whether results support or reject research hypotheses.

A Stata Example

- See the command and log files

Conclusions

- The goal of data analysis is to obtain valid evidence to support or refute research hypotheses and subsequently to answer research questions.
- Data analysis techniques address two issues: estimating the theoretical relations between variables and reducing biases caused by the sample, missing data, and arbitrary operationalization of variables.
- When developing models for research topics, it is critical to consider what regression coefficients are required, how they will be compared to one another, and which procedures offer such coefficients.
- Before conducting data analysis, it is important to select adequate techniques and think through the sequences of statistical techniques to be used. In general, techniques for dealing with missing data and sampling weights should be used first, followed by regression-based analysis, and last, sensitivity testing. This is especially important because some analytic techniques do not work with techniques dealing with methodological issues.
- Feel free to contact me (wuh@bgsu.edu) if you have any queries regarding using Stats to analyze data.