Regression Analysis in Stata

Hsueh-Sheng Wu

CFDR Workshop Series

February 18, 2019

BGSU

Center for Family and Demographic Research

Overview

- Introduction to regression
- Venn diagram of question, data, and regression analysis
- Steps of conducting regression analysis
- Research questions and hypotheses
- Attributes of variables, samples, and data
- Specify regression models
- Post-estimation commands
- Stata examples
- Conclusions

Introduction to Regression

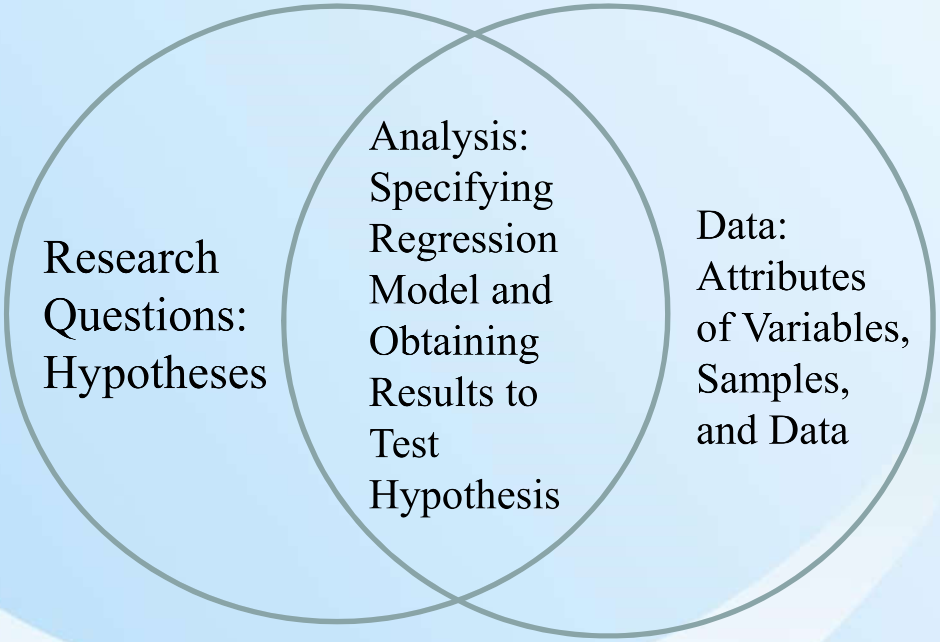
- Regression analysis is probably the most common statistical technique that sociologists use to answer a research question
- Regression analysis assumes a linear relation between the predictor and the outcome variable. Since the outcome variables may follow different distributions, Stata has commands for conducting regression analysis for each of these outcome variables
- Stata regression commands have many options. These options are used to account for special features of the model and overcome particular problems related with how sample is selected, how to adjust the estimate of variance of the regression coefficient when respondents are not independent from each other, whether the analysis is done for a subset of observations, and so on

Introduction to Regression (Cont.)

- After fitting a regression model, researchers may need to use post-estimation commands of testing regression coefficients or examining marginal effects to answer their research questions
- The goal of this workshop to explain how conducting a regression analysis and answering a research question is linked together

Venn Diagram of Question, Data, and Regression Analysis

- Regression analysis lies in the overlapping areas of research question and data
- The goal for researchers conducting regression analyses is to consider both research questions and attributes of data to obtain most valid findings to reject or accept the hypothesis

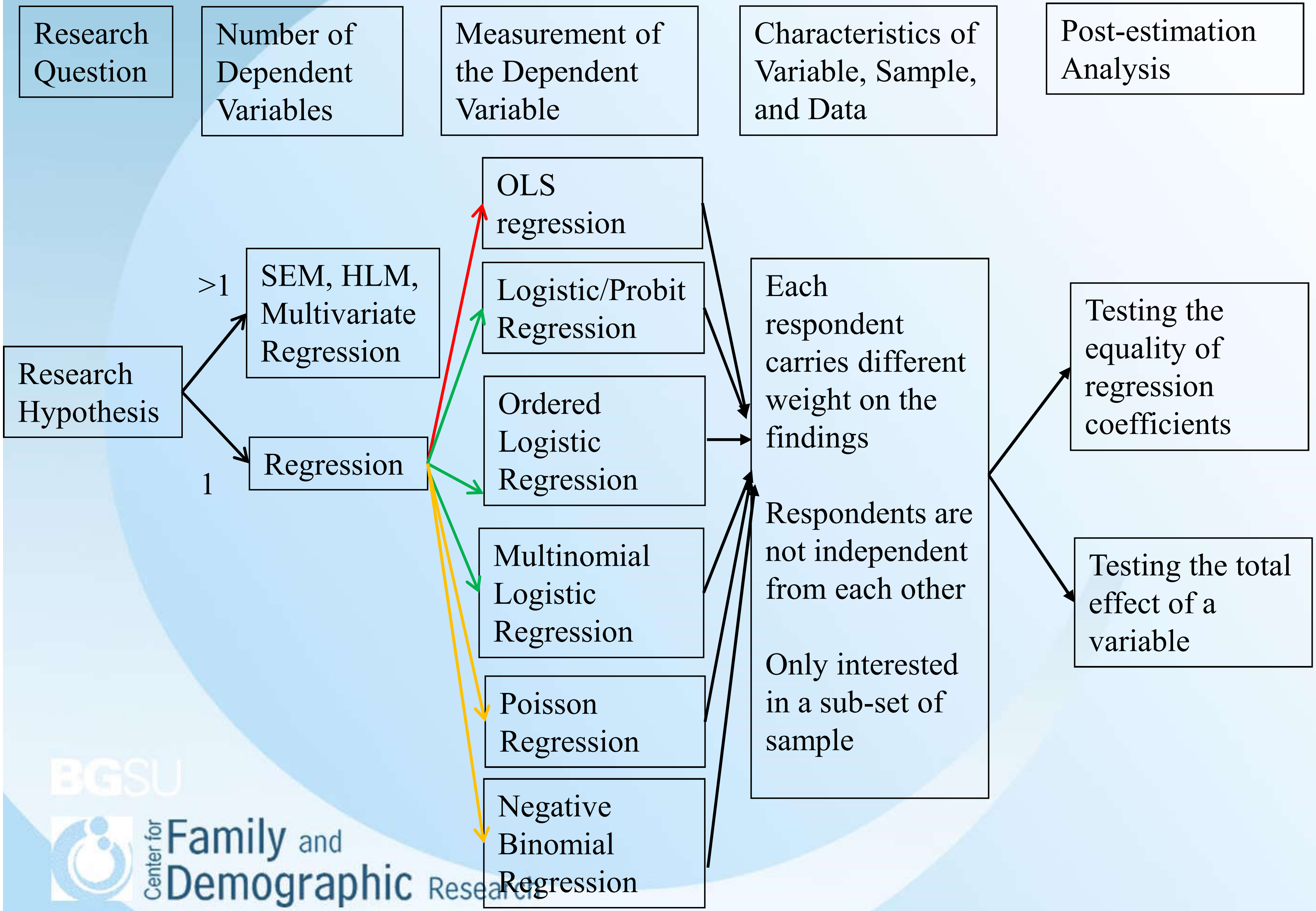

Research Questions: Hypotheses

Analysis: Specifying Regression Model and Obtaining Results to Test Hypothesis

Data: Attributes of Variables, Samples, and Data

BGSU Center for Family and Demographic Research

Steps of Conducting Regression Analysis



Research Question

Number of Dependent Variables

Measurement of the Dependent Variable

Characteristics of Variable, Sample, and Data

Post-estimation Analysis

Research Hypothesis

>1

SEM, HLM, Multivariate Regression

1

Regression

OLS regression

Logistic/Probit Regression

Ordered Logistic Regression

Multinomial Logistic Regression

Poisson Regression

Negative Binomial Regression

Each respondent carries different weight on the findings

Respondents are not independent from each other

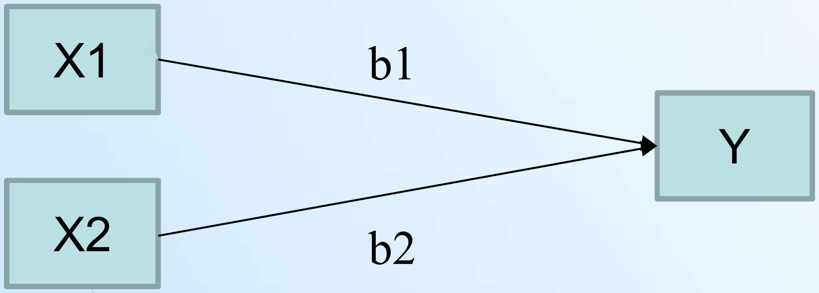
Only interested in a sub-set of sample

Testing the equality of regression coefficients

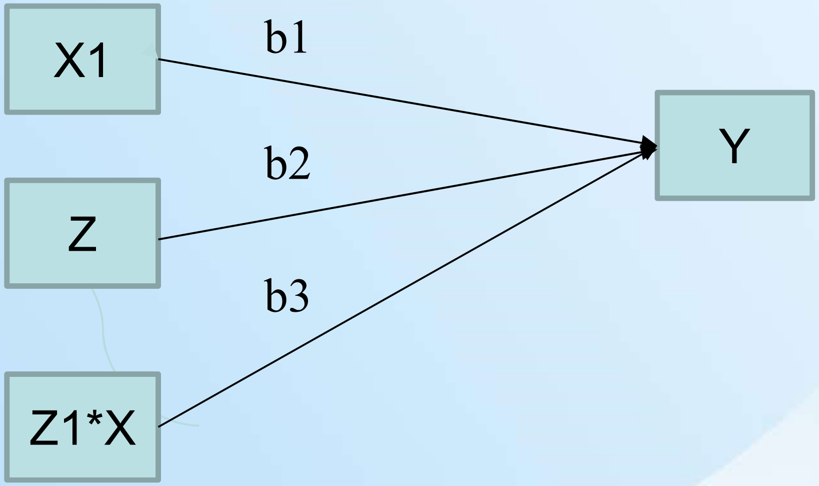
Testing the total effect of a variable

Research Questions and Hypotheses

1. Regression



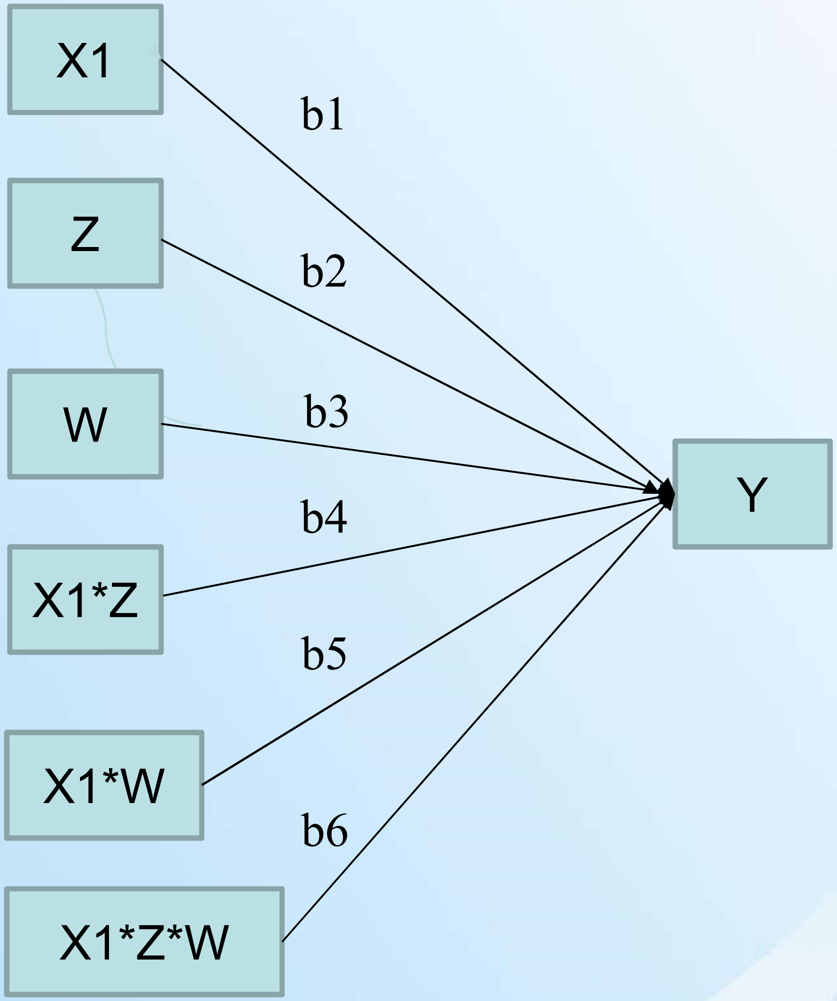
2. Regression with a two-way interaction term



BGSU Center for Family and Demographic Research

Research Questions and Hypotheses (Cont.)

3. Regression with a three-way interaction



Research Questions and Hypotheses (Cont.)

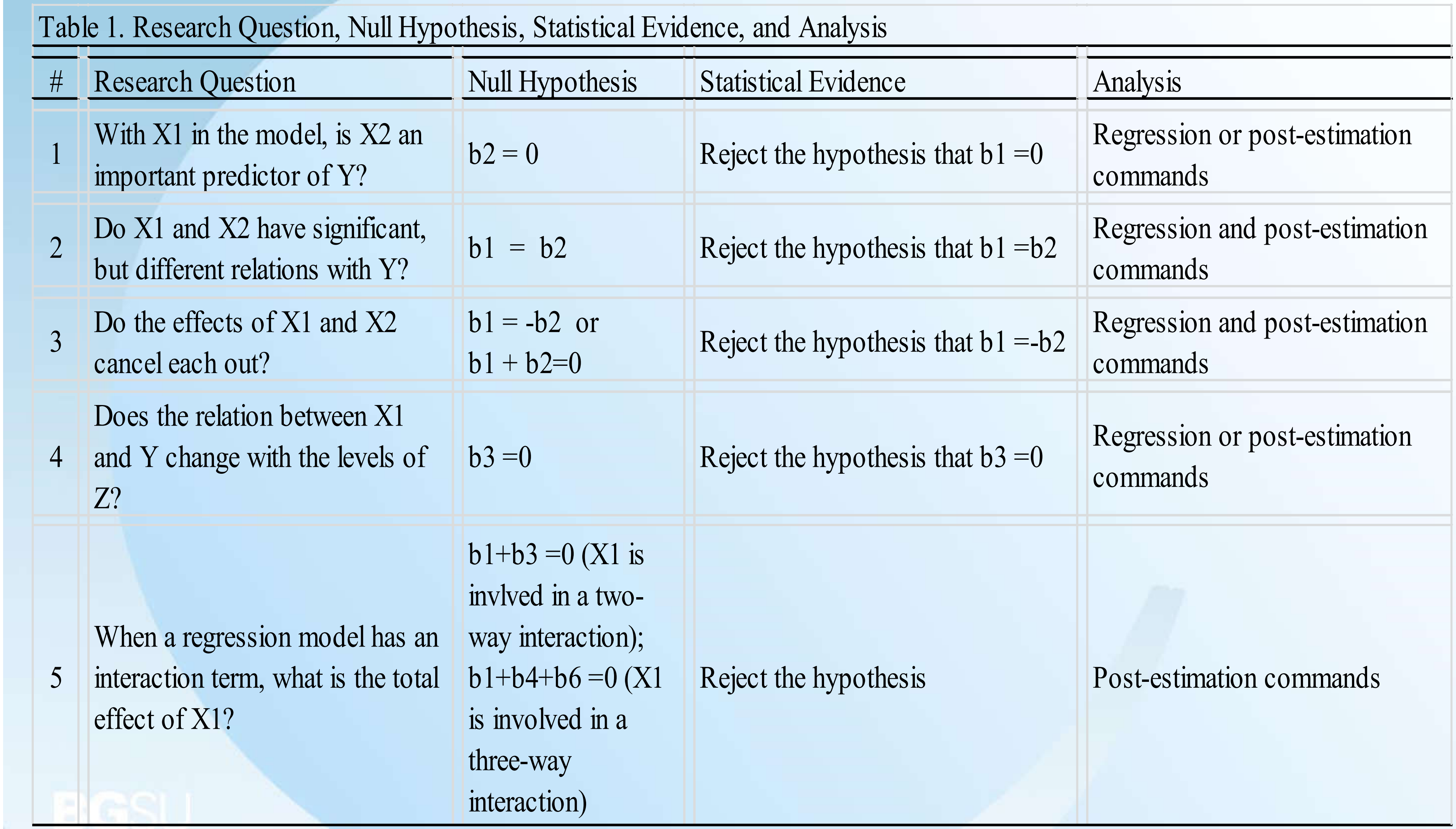


Table 1. Research Question, Null Hypothesis, Statistical Evidence, and Analysis

#	Research Question	Null Hypothesis	Statistical Evidence	Analysis
1	With X1 in the model, is X2 an important predictor of Y?	b2 = 0	Reject the hypothesis that b1 =0	Regression or post-estimation commands
2	Do X1 and X2 have significant, but different relations with Y?	b1 = b2	Reject the hypothesis that b1 =b2	Regression and post-estimation commands
3	Do the effects of X1 and X2 cancel each out?	b1 = -b2 or b1 + b2=0	Reject the hypothesis that b1 =-b2	Regression and post-estimation commands
4	Does the relation between X1 and Y change with the levels of Z?	b3 =0	Reject the hypothesis that b3 =0	Regression or post-estimation commands
5	When a regression model has an interaction term, what is the total effect of X1?	b1+b3 =0 (X1 is invlved in a two-way interaction); b1+b4+b6 =0 (X1 is involved in a three-way interaction)	Reject the hypothesis	Post-estimation commands

Attributes of Variables, Samples, and Data

- The number of dependent variables and/or the nested data structure determine the number of regression equations needed (e.g. OLS regression vs. SEM, HLM, and Multivariate Regression)
- The measurement level of dependent variable (regression vs. logistic regression)
- If the respondents were selected with unequal probabilities, the results need to be weighted using the -svy- command or -pweight- command
- If some respondents are not independent from each other, it can be dealt with using the robust option or choose a method that takes into account the dependence of the observations
- Analyzing a subpopulation may create an inaccurate estimate of variance if the data were collected with a complex survey design and the -subpop- option is not used

Center for Family and Demographic Research

Specify Regression Models

The measurement level of the dependent variable determines the type of regression model used:

Data collected without a complex survey design

Continuous dependent variable (e.g., income)

regress *depvar indepvars* [*if]* [*in]* [*weight*] [, *options*]

Binary, ordered, and nominal dependent variable

logit *depvar indepvars* [*if]* [*in]* [*weight*] [, *options*]

ologit *depvar indepvars* [*if]* [*in]* [*weight*] [, *options*]

mlogit *depvar indepvars* [*if]* [*in]* [*weight*] [, *options*]

Count variable

possion *depvar indepvars* [*if]* [*in]* [*weight*] [, *options*]

nbreg *depvar* [*if]* [*in]* [*weight*] [,*nbreg options*]

Specify Regression Models (Cont.)

Regression using data collected with a single-stage survey design

```
svyset [psu] [weight] [, design_options options]
```

Continuous dependent variable (e.g., income)

```
    svy: regress depvar indepvars  [if]  [in] [, options ]
```

Binary, ordered, and nominal dependent variable

```
    svy: logit depvar indepvars    [if]  [in]  [, options ]
    svy: ologit depvar indepvars  [if]  [in]  [, options ]
    svy: mlogit depvar indepvars [if]  [in]  [, options ]
```

Count variable:

```
    svy: possion depvar indepvars    [if]  [in]  [, options ]
    svy: nbreg depvar                        [if]  [in]  [, nbreg options]
```

Specify Regression Models (Cont.)

Regression using data collected with a single-stage survey design and analyze only a sub-sample

Continuous dependent variable (e.g., income)

svy, subpop(indicator): regress *depvar indepvars* [*if*] [*in*] [, *options*]

Binary, ordered, and nominal dependent variable

svy, subpop(indicator): logit *depvar indepvars* [*if*] [*in*] [, *options*]
svy, subpop(indicator): ologit *depvar indepvars* [*if*] [*in*] [, *options*]
svy, subpop(indicator): mlogit *depvar indepvars* [*if*] [*in*] [, *options*]

Count variable:

svy, subpop(indicator): possion *depvar indepvars* [*if*] [*in*] [, *options*]
svy, subpop(indicator): nbreg *depvar* [*if*] [*in*] [, *nbreg options*]

Post-estimation Commands

- Post-estimation commands are used after the regression model had been fitted

- Post-estimation commands allow researchers to test the equality and linear combination of regression coefficients

- Post-estimation commands are very useful when the regression models involve interaction terms and/or categorical dependent variables

- Two most commonly used post-estimation commands are -test- and -margins-

Sample Stata Code

- Descriptions of the variables

```
variable name   type    format     label      variable label
----------------------------------------------------------------------------------
-
stratid         byte    %9.0g                 stratum identifier, 1-32
psuid           byte    %9.0g                 primary sampling unit, 1 or 2
finalwgt        long    %9.0g                 sampling weight (except lead)
company_id      float   %9.0g                 company ID
sex             byte    %9.0g      sex        1=male, 2=female
race            byte    %9.0g      race       1=white, 2=black, 3=other
age             byte    %9.0g                 age in years
illness         float   %9.0g                 how many illnesses do you have?
hlthstat        byte    %9.0g                 1=excellent,..., 5=poor
```

- The sample Stata codes are in the accompanying handouts.

Conclusions

- An accurate application of regression analysis requires a clear specification of research hypothesis, choosing the correct regression model and options, and using a suitable test for the hypothesis
- Research hypotheses determine what regression coefficients will be tested in the end
- The number and measurement level of the dependent variables decide the specification of the regression model and analysis
- Depending on whether the equality, linear combination, or the total effect of variables is tested, different post-estimation commands will be used

BGSU Center for Family and Demographic Research

Conclusions (Cont.)

- The Sample Stata code can be used for dependent variables that are categorical or counts
- When your research question involves more than one dependent variable, it is likely your research question is not one listed in Table 1. If you are not sure what research hypothesis will be tested and/or how to specify the regression model, please stop by my office and we can discuss it

Additional Information

1. Estimation and post-estimation commands: https://www.stata.com/manuals13/u20.pdf
2. svy postestimation: https://www.stata.com/manuals13/svysvypostestimation.pdf#svysvypostestimation
3. Test linear hypotheses after estimation: https://www.stata.com/manuals13/rtest.pdf
4. Nonlinear combinations of estimators: https://www.stata.com/manuals13/rnlcom.pdf
5. Marginal means, predictive margins, and marginal effects: https://www.stata.com/manuals13/rmargins.pdf
6. Stata survey data: https://www.stata.com/manuals13/svy.pdf

BGSU

Center for Family and Demographic Research