Regression Analysis in Stata

Hsueh-Sheng Wu
CFDR Workshop Series
October 3, 2022
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 the most common statistical technique that sociologists use to answer research questions. Regression analysis is further extended into other advanced statistical techniques such as structural equation modeling and hierarchical linear models.

• 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 outcomes.

• Stata regression commands have options to provide better estimates of regression coefficients by taking account 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 conducted 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 to test regression coefficients or examine marginal effects to answer their research questions.

• The goal of this workshop to demonstrate how Stata can be used to conduct regression analysis and answer research questions.
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 support the hypothesis

Analysis:
- decide on the statistical technique,
- specify models, and
- decide if hypothesis is rejected

Data:
- Target population,
- Sample, and
- Variables

Research Questions:
- Formulate
- Testable
- Hypotheses
Steps of Conducting Regression Analysis

1. **Research Question**
   - Testable Research Hypothesis
   - SEM, HLM

2. **Number of Dependent Variables**
   - >1
   - SEM, HLM

3. **Measurement of the Dependent Variable and specify models**
   - OLS regression
   - Logistic/Probit Regression
   - Ordered Logistic Regression
   - Multinomial Logistic Regression
   - Poisson Regression
   - Negative Binomial Regression

4. **Characteristics of Variable, Sample, and Data**
   - Each respondent carries different weight on the findings
   - Respondents are not independent from each other
   - Only interested in a sub-set of sample

5. **Post-estimation Analysis**
   - Testing the equality of regression coefficients
   - Testing the total effect of a variable
Research Questions and Hypotheses

1. Regression

\[ X_1 \rightarrow b_1 \rightarrow Y \]
\[ X_2 \rightarrow b_2 \rightarrow Y \]

2. Regression with a two-way interaction term

\[ X_1 \rightarrow b_1 \rightarrow Y \]
\[ Z \rightarrow b_2 \rightarrow Y \]
\[ Z_1 \times X \rightarrow b_3 \rightarrow Y \]
3. Regression with a three-way interaction

\[ Y = b_1 X_1 + b_2 Z + b_3 W + b_4 X_1 Z + b_5 X_1 W + b_6 X_1 Z W \]
Table 1. Research Question, Null Hypothesis, Statistical Evidence, and Analysis

<table>
<thead>
<tr>
<th>#</th>
<th>Research Question</th>
<th>Null Hypothesis</th>
<th>Statistical Evidence</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>With X1 in the model, is X2 an important predictor of Y?</td>
<td>b2 = 0</td>
<td>b2 is significantly different from 0</td>
<td>Regression or post-estimation commands</td>
</tr>
<tr>
<td>2</td>
<td>Do X1 and X2 have significant, but different relations with Y?</td>
<td>b1 = b2</td>
<td>The differences between b1 and b2 is significantly different from 0</td>
<td>Regression and post-estimation commands</td>
</tr>
<tr>
<td>3</td>
<td>Do the effects of X1 and X2 cancel each out?</td>
<td>b1 = -b2 or b1 + b2 = 0</td>
<td>the sum of b1 and b2 is significantly different from 0</td>
<td>Regression and post-estimation commands</td>
</tr>
<tr>
<td>4</td>
<td>Does the relation between X1 and Y change with the levels of Z?</td>
<td>b3 = 0</td>
<td>b3 is significantly different from 0</td>
<td>Regression or post-estimation commands</td>
</tr>
<tr>
<td>5</td>
<td>When a regression model has an interaction term, what is the total effect of X1?</td>
<td>b1 + b3 = 0 (X1 is involved in a two-way interaction); b1 + b4 + b5 + b6 = 0 (X1 is involved in two- and three-way interactions)</td>
<td>The sum of b1, b4, b5, and b6 is significantly different from 0</td>
<td>Post-estimation commands</td>
</tr>
</tbody>
</table>
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)

• 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 when the data were collected with a complex survey design and the -svy- and -subpop- options are not used
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
  `poisson depvar indepvars [if] [in] [weight] [, options ]`
  `nbreg depvar indepvars [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 indepvars [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 indepvars [if] [in] [, nbreg options]
Post-estimation Commands

• Post-estimation commands are used after the regression model has 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

<table>
<thead>
<tr>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>byte</td>
<td>%9.0g</td>
<td>sex</td>
<td>1=male, 2=female</td>
</tr>
<tr>
<td>race</td>
<td>byte</td>
<td>%9.0g</td>
<td>race</td>
<td>1=white, 2=black, 3=other</td>
</tr>
<tr>
<td>height</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>height (in.)</td>
</tr>
<tr>
<td>weight</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>weight (lbs.)</td>
</tr>
<tr>
<td>sampwgt</td>
<td>double</td>
<td>%9.0g</td>
<td></td>
<td>sampling weight</td>
</tr>
<tr>
<td>state</td>
<td>byte</td>
<td>%9.0g</td>
<td></td>
<td>State ID (strata)</td>
</tr>
<tr>
<td>county</td>
<td>byte</td>
<td>%9.0g</td>
<td></td>
<td>County ID (PSU)</td>
</tr>
<tr>
<td>school</td>
<td>byte</td>
<td>%9.0g</td>
<td></td>
<td>School ID (SSU)</td>
</tr>
<tr>
<td>id</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>Person ID</td>
</tr>
<tr>
<td>ncounties</td>
<td>byte</td>
<td>%9.0g</td>
<td></td>
<td>Stage 1 FPC</td>
</tr>
<tr>
<td>nschools</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>Stage 2 FPC</td>
</tr>
</tbody>
</table>

Sorted by: state  county  school

• 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.
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