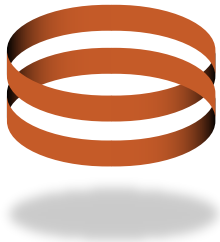
Regression in Stata

Ista Zahn

Harvard MIT Data Center

April 12 2013



The Institute
for Quantitative Social Science
at Harvard University

Outline

1. Introduction
2. Univariate regression
3. Multiple regression
4. Interactions and Categorical IV
5. Exporting and saving results
6. Wrap-up

Topic

1 Introduction

2 Univariate regression

3 Multiple regression

4 Interactions and Categorical IV

5 Exporting and saving results

6 Wrap-up

Documents for today

USERNAME: dataclass PASSWORD: dataclass

- Find class materials at: Scratch > StataStatistics
- FIRST THING: copy this folder to your desktop!

Organization

- Please feel free to ask questions at any point if they are relevant to the current topic (or if you are lost!)
- There will be a Q&A after class for more specific, personalized questions
- Collaboration with your neighbors is encouraged
- If you are using a laptop, you will need to adjust paths accordingly
- Make comments in your Do-file rather than on hand-outs
- Save on flash drive or email to yourself

Today's Dataset

- We have data on a variety of variables for all 50 states
- Population, density, energy use, voting tendencies, graduation rates, income, etc.
- We're going to be predicting SAT scores
- Univariate Regression: SAT scores and Education Expenditures
- Does the amount of money spent on education affect the mean SAT score in a state?
- Dependent variable: csat
- Independent variable: expense

Opening Files in Stata

- Look at bottom left hand corner of Stata screen – This is the directory Stata is currently reading from
- Files are located in the StataDatMan folder
- Start by changing directory and loading the data

```
// change directory
cd "C:/Users/dataclass/Desktop/StataStatistics"
// use dir to see what is in the directory:
dir
// tell Stata to use the states data set
use states.dta
```

Steps for Running Regression

1. Examine descriptive statistics
2. Look at relationship graphically and test correlation(s)
3. Run and interpret regression
4. Test regression assumptions

Topic

1 Introduction

2 Univariate regression

3 Multiple regression

4 Interactions and Categorical IV

5 Exporting and saving results

6 Wrap-up

Univariate Regression: Preliminaries

- We want to predict csat scores from expense
- First, let's look at some descriptives

```
// generate summary statistics for csat and expense
sum csat expense
// look at codebok
codebook csat expense
```

Univariate Regression

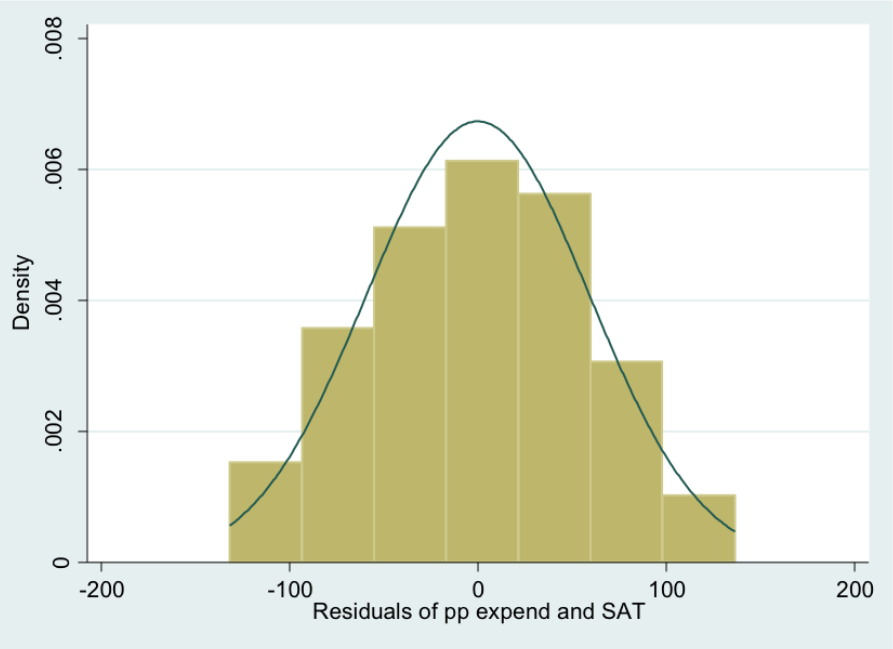
- Look at scatterplots, compute correlation matrix, and regress SAT scores on expenditures

```
// graph expense by csat
twoway scatter expense csat

// correlate csat and expense
pwcorr csat expense, star(.05)

// run the regression
regress csat expense
```

Postestimation Commands–Predicted and Residual Values

- Modeling functions in Stata usually save the results so you can do further computations with them–see `help regress` for the list of saved values
- We can use postestimation commands to do computations on the results
- For example, histogram of the residuals can be informative

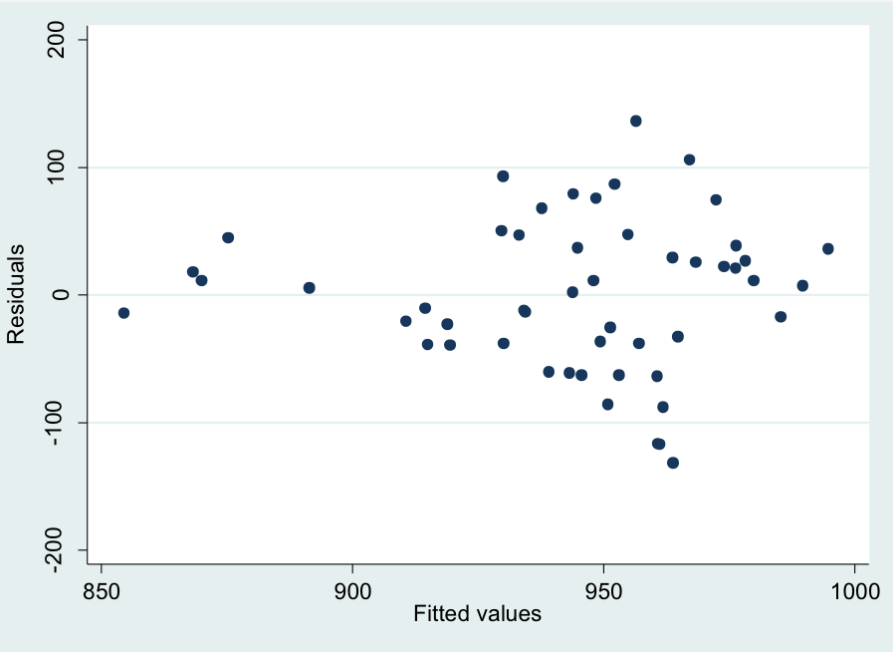
```
// graph the residual values of csat
predict resid, residual
histogram resid, normal
```



Postestimation Commands: rvfplot

- We could do this manually with the predicted and residual values we saved earlier, or we can have Stata do it for us with the `rvfplot` postestimation command:

`rvfplot`



Topic

1 Introduction

2 Univariate regression

3 Multiple regression

4 Interactions and Categorical IV

5 Exporting and saving results

6 Wrap-up

Multiple Regression

- Just keep adding predictors
- Let's try adding some predictors to the model of SAT scores

income % students taking SATs

percent % adults with HS diploma (high)

Multiple Regression Preliminaries

- As before, start with descriptive statistics and correlations

```
// descriptive statistics
sum income percent high

// generate correlation matrix
pwcorr csat expense income percent high

// regress csat on exense, income, percent, and high\
regress csat expense income percent high
```

Exercise 1: Multiple Regression

Open the datafile, states.dta.

1. Select a few variables to use in a multiple regression of your own. Before running the regression, examine descriptive of the variables and generate a few scatterplots.
2. Run your regression
3. Examine the plausibility of the assumptions of normality and homogeneity

Topic

1 Introduction

2 Univariate regression

3 Multiple regression

4 Interactions and Categorical IV

5 Exporting and saving results

6 Wrap-up

Interactions

- What if we wanted to test an interaction between percent & high?
- Stata uses the # sign to represent interactions–see `help fvvarlist` for details

```
// use the # sign to represent interactions
regress csat percent high c.percent # c.high
// same as . regress csat c.percent##high
```

- Alternatively we can use the ## operator to automatically include the lower-order terms:

```
// use the # sign to represent interactions
regress csat percent high c.percent # c.high
// same as . regress csat c.percent##high
```

Categorical Predictors

- For categorical variables, we first need to dummy code
- Use region as example
  - Option 1: create dummy codes before fitting regression model

```
// create region dummy codes using tab
tab region, gen(region) // could also use gen / replace

//regress csat on region
regress csat region1 region2 region3
```

Categorical Predictors

- For categorical variables, we first need to dummy code using `i.` notation–see `help fvvarlist` for details
- Use region as example:

```
// regress csat on region using fvvarlist syntax
// see help fvvarlist for details
regress csat i.region
```

- You can change the reference level using `ib#` notation:

```
// regress csat on region using fvvarlist syntax
// see help fvvarlist for details
regress csat ib4.region
```

Exercise 2: Regression, Categorical Predictors, & Interactions

Open the datafile, states.dta.

1. Add on to the regression equation that you created in exercise 1 by generating an interaction term and testing the interaction.
2. Try adding a categorical variable to your regression. You could use region or high25, or generate a new categorical variable from one of the continuous variables in the data set.

Topic

1 Introduction

2 Univariate regression

3 Multiple regression

4 Interactions and Categorical IV

5 Exporting and saving results

6 Wrap-up

Saving and exporting regression tables

- Usually when we're running regression, we'll be testing multiple models at a time
- Can be difficult to compare results
- Stata offers several user-friendly options for storing and viewing regression output from multiple models
- First, download the necessary packages:

```
* install outreg2 package
findit outreg2
```

Saving and replaying

- You can store regression model results in Stata

```
// fit two regression models and store the results
regress csat expense income percent high
estimates store Model1
regress csat expense income percent high i.region
estimates store Model2
```

Saving and replaying

- Stored models can be recalled

```
// Display Model1
estimates replay Model1
```

Saving and replaying

- Stored models can be compared

```
// Compare Model1 and Model2 coefficients
estimates table Model1 Model2
```

Exporting into Excel

- Avoid human error when transferring coefficients into tables
- Excel can be used to format publication-ready tables

```
outreg2 [Model1 Model2] using csatprediction.xls, replace
```

Topic

1 Introduction

2 Univariate regression

3 Multiple regression

4 Interactions and Categorical IV

5 Exporting and saving results

6 Wrap-up

Help Us Make This Workshop Better

- Please take a moment to fill out a very short feedback form
- These workshops exist for you–tell us what you need!
- `ttp://tinyurl.com/StataRegressionFeedback`

Additional resources

- training and consulting
  - IQSS workshops:
    `http://projects.iq.harvard.edu/rtc/filter_by/workshops`
  - IQSS statistical consulting: `http://rtc.iq.harvard.edu`
- Stata resources
  - UCLA website: `http://www.ats.ucla.edu/stat/Stata/`
  - Great for self-study
  - Links to resources
- Stata website: `http://www.stata.com/help.cgi?contents`
- Email list: `http://www.stata.com/statalist/`