Expository Data Analysis in R

Regressions

Day 1

When to use Regression Analysis in Economics

When to use Regression Analysis in Economics

- Trying to identify causation
- Correlation vs. causation
  - Height vs. Weight
  - Get taller gain weight!
  - Spurious correlations

Regression Analysis More Formally Defined

- Regression analysis is used to describe the relationship between:
  - A single response variable Y and
  - One or more predictor variables X_1, X_2, X_3, ..., X_n
- What conditions must the response variable meet for OLS?

Regression Analysis More Formally Defined

- Regression analysis is used to describe the relationship between:
  - A single response variable Y and
  - One or more predictor variables X_1, X_2, X_3, ..., X_n
- What conditions must the response variable meet for OLS?
  - Continuous! but ... (sometimes economists cheat)
- What conditions must the predictor variables meet?

Regression Analysis More Formally Defined

- Regression analysis is used to describe the relationship between:
  - A single response variable Y and
  - One or more predictor variables X_1, X_2, X_3, ..., X_n
- What conditions must the response variable meet for OLS?
  - Continuous! but ... (sometimes economists cheat)
- What conditions must the predictor variables meet?
  - None! These variables can be continuous, discrete, or categorical

Steps to take before you put your data into a regression

Steps to take before you put your data into a regression

- Check for:
  - Missing values
  - Outliers
  - Asymmetric distributions
  - Clustering of values
  - Unexpected patterns
- Numerical Summaries
  - Mean, min, max, variance, etc.
  - Correlations
- Graphical Summaries
  - Scatter plots
  - Histograms
  - Box plots

ACS/Census Data from IPUMS

- IPUMS is a great resource!
- Let's check out how you can create a sample of data to download

ACS/Census Data

- Please change the code in your regressions_lecture.R file so that you can read in the IPUMS data file a with this lecture.
- Also be sure to load the appropriate packages (dplyr, ggplot2)
- What variables do we have in our data? What are the variable classes in the data?
- Run the summary function on the dataset. What do we learn?
- Check out the code book. What variables are we going to have to re-code?

ACS/Census Data

The American Community Survey... is a survey!

Survey Data

Why do we weight survey data?

Survey Data

Why do we weight survey data? To make statistics computed from the data more representative of the population.

- Design Weight - compensate for over- or under-sampling of specific cases

Example?

- Post-Stratification or Non-response Weight - compensate for that fact that persons with certain characteristics are not as likely to respond to the survey.

Example?

Survey Data

- Weights primarily adjust means and proportions.
- May adversely affect inferential data and standard errors.
- Weights almost always increase the standard errors of your estimates.
- Introduce instability into your data.
- Very large weights (or very small ones) can also introduce instabilities (fivethirtyeight).

Recoding Variables

- How should we re-code the variable EDUC to transform the education variable from a categorical variable to a continuous variable?

Recoding Variables

- NA - N/A or no schooling
- 5 - Nursery school to grade 4
- 9 - Grade 5, 6, 7, or 8
- 10 - Grade 9
- 11 - Grade 10
- 12 - Grade 11
- 13 - Grade 12
- 14 - 1 year of college
- 15 - 2 years of college
- 16 - 3 years of college
- 17 - 4 years of college
- 18 - 5+ years of college

Recoding Variables

- For ease of plotting let's also re-code the SEX, RACE, and HISPAN variables
- SEX
  - 1 - Male
  - 2 - Female
- RACE
  - 1 - White
  - 2 - Black
  - 3 - American Indian or Alaska Native
  - (4,5,6) - Asian or Pacific Islander
  - 7 - Other
  - What is a problem with how I am handling this variable?
- HISPAN
  - 0 - Not Hispanic
  - (1,2,3,4) - Hispanic

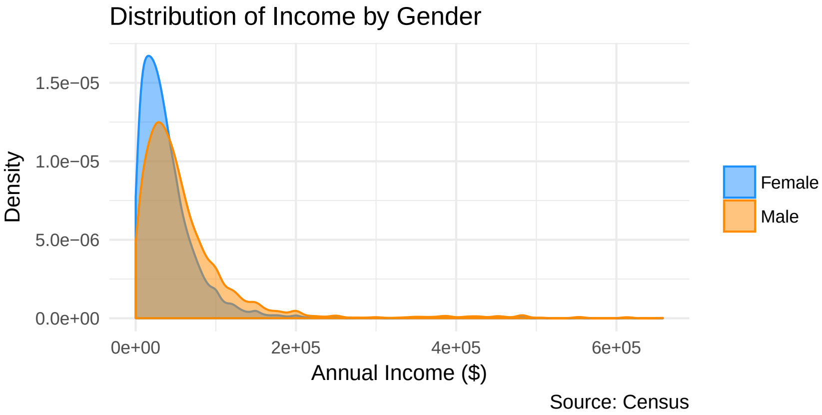
Filtering Variables

Filter out all individuals:

- That are missing data for new_educ
- Younger than 18 or older than 65
- Not in the workforce
- Missing data for OCC
- With a negative salary

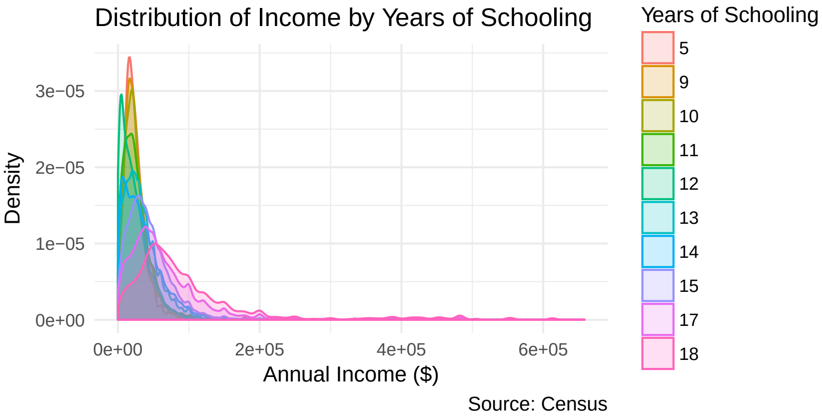
Plotting the Data

- Please make a density plot of wages by gender. It should look something like:



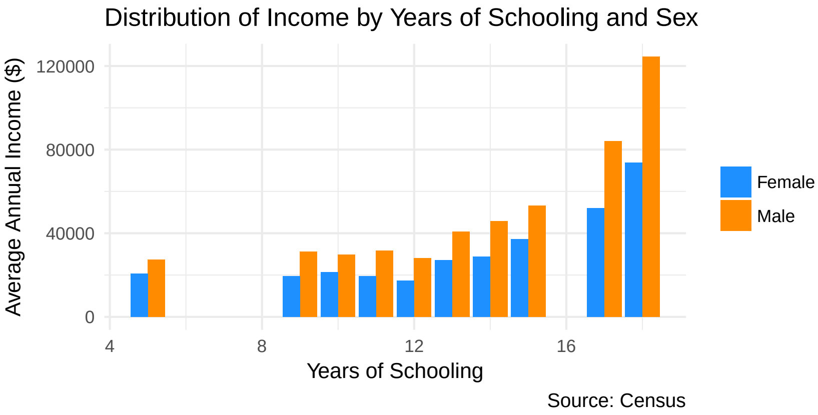
Plotting the Data

- Please plot the distribution of wages by years of schooling. It should look something like:



Plotting the Data

- Lets make a barchart of average income vs. years of schooling by gender:



OLS Regression

- Let's write down a baseline model of an individual's Salary as a function of the years of education.

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Years of Education}$$

- What do you think? What variables might be missing?

OLS Regression

- How do we run a OLS regression in R? With the lm() function.
- What are the arguments to the lm() function?

OLS Regression

- Some example code:

```r
# run a multiple linear regression
my_model <- lm(y ~ x1 + x2 + x3, data = mydata)

#show results
summary(my_model)
```

OLS Regression

- Try it out! Run a simple regression of salary on years of education.
- What are the results?
- How do we add weights?
- How can we interpret the result?
- What is the structure of the model object?

The Broom Package

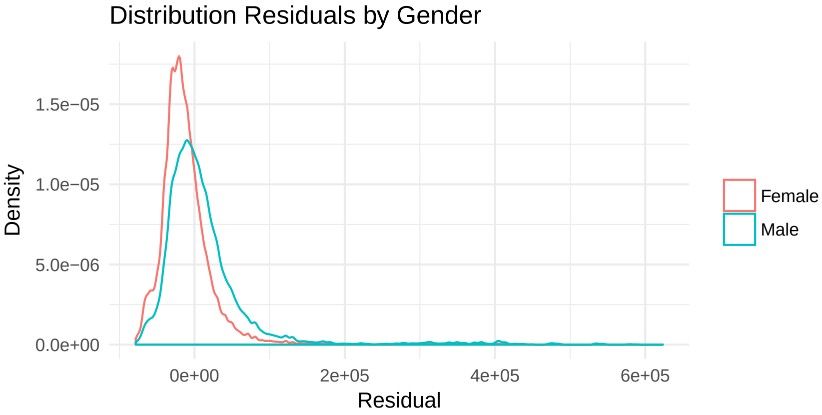
- Model results are messy and hard to work with by themselves in R
- The broom package is there to help!
- The broom package can turn these messy and unfamiliar model objects into good old data frames.
- The three main functions of the broom package are
  - tidy() - for creating a data frame of component statistics
  - augment() - for observation level statistics (like fitted values and residuals)
  - glance()- for model level statistics (like R-squared etc.)

The Broom Package

- Let's try it out!
- tidy, augment, and glance at the results of the baseline model
- How can we use the augment function to keep all of our original columns?

Improving our model

- Let's make a plot of the distribution of residuals by gender.
- What do we learn from this chart?



Improving our model

- Let's run the regression described by

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Years of Education}_i + \beta_2 \text{Gender}_i$$

- How do the two models compare?

Improving our model

- What if we think that the difference of an additional year of education on salary differs by gender?
- How does this change our model?
- How can we calculate interaction terms using lm()?

Presenting Regression Results

- The stargazer package is designed to beautify the results of a regression in R.
- Let's install the package and run stargazer() on the baseline model.
- The output of the function is the code to create a beautiful latex table.
- We do not expect you to use latex for this class.
- You can plug the latex into http://quicklatex.com/ to create a nice image of the table.

Put my models to shame

- Pair up!
- Take 15 - 20 mins to improve on the models we have done so far.
- I want to see plots that explain why you are adding in variables or interaction terms
- I want to see beautiful regression output tables
- I want you to spend 5 minutes writing up a post on piazza that includes a graph, a table, and a brief explanation of your model