Module 1

Day 4: Introduction to Regression Analysis in R

Recap last week

- facet_wrap()
- wtd.mean(), wtd.median()
- stargazer()
- scale_color_manual(), theme(), theme_
- unique()
- as.Date(), scale_x_date()

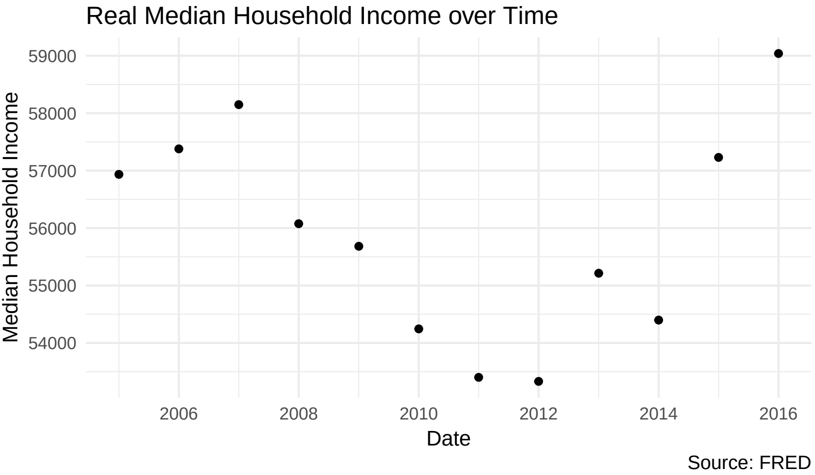
Organize your folders

- You should have one folder for this class
- Inside that folder there should be one folder for each lecture
- Make a folder for your project with sub folders - data, plots, liturature

Recap last week

- Read in the data file "fred_median_income.csv"
- Convert the data column from a character to a Date
- Filter to data post 2005
- Make a scatter plot with the date on the x axis and median_hh_income on the y axis
- Be sure to label your chart appropriately
- Use the scale_x_dates() function to label every 2 years

Recap last week



What do we mean when we say "real" income?

lag() and lead()

What do the dplyr functions lag() and lead() do?

lag() and lead()

What do the dplyr functions lag() and lead() do?

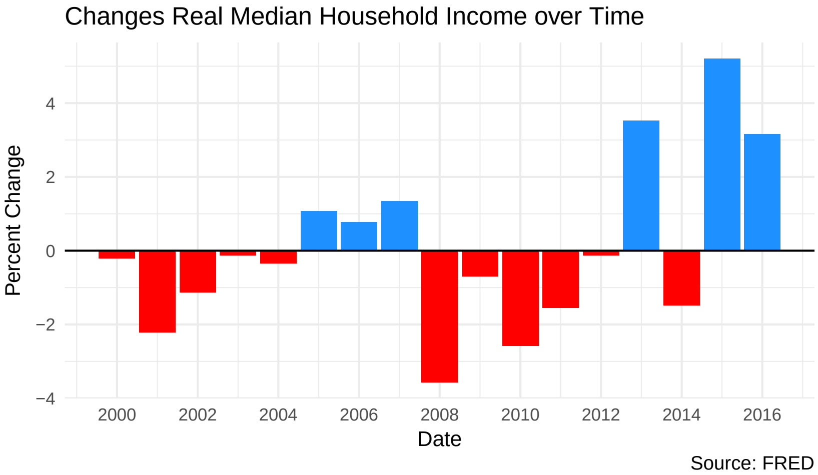
- Find the "next" or "previous" values in a vector.
- Useful for comparing values ahead of or behind the current values.

How could we use this function to calculate the percent change in median household income?

Shrinking Real Incomes

- Use the lag() function to calculate the percent change in household income
- Add a new column called shrunk that "Yes" if the percent change is less than zero and "No" otherwise
- Make a coloumn chart (geom_col) of the percent change in median household income since 2000
- Use the variable shrunk as the color axis
- Use geom_hline() to add a horizontal line at y = 0 to highlight years where real income shrunk
- Turn off the color axis by using guides(color = FALSE)
- Use scale_color_manual() so that when income shrinks the point is red and if it grows it's blue

Shrinking Real Incomes



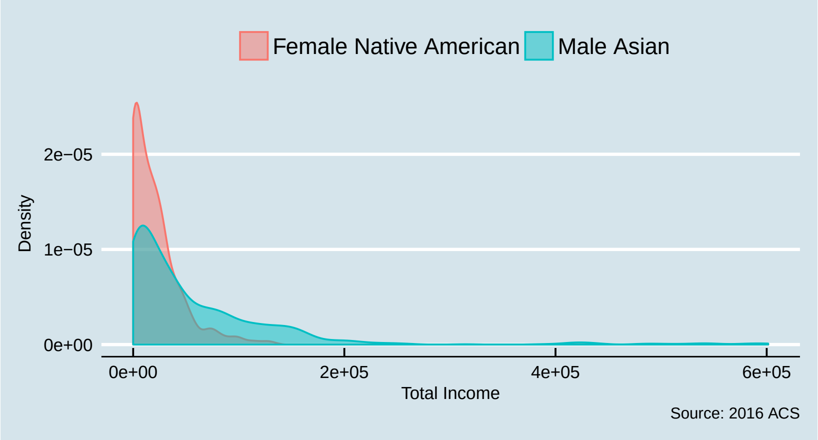
Recap last week

- Read in the data file "acs_2016_sample.csv"
- Apply our standard transformations
- Create a new data frame called acs_mean_median that is the weighted mean and median income by race and sex
- Add a column that is the difference between the mean and median income
- Which group has the biggest difference? Which the smallest?

Recap last week

- Going back to the whole sample filter to only Native American Females and Asian Males
- Use the paste function to create a new variable called group that is paste(sex, race)
- Make a density plot of the income of the two groups
- Set the alpha parameter to 0.5

Recap last week



What does it mean for the distribution if there is a big difference between the mean and median?

Regression Analysis

We have spent the past few weeks

- Learning some R
- Uncovering relationships between characteristics and income

Time to formalize our understanding

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

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

What relationships have we already uncovered in the data?

Prepping the data

Create a new data frame acs_2016_cleaned which is acs_2016_transformed filtered to people:

- Between 18 and 65
- In the workforce
- With a total wage $<=$ 1,000,000
- Worked more than 0 hours a week
- Worked more than 0 weeks
- Add a column for hourly wage

What fraction of the original observations do we have?

Preping the data

Select the columns

- wage_income, age, hrs_worked , weeks_worked, and hourly_wage

Make a stargazer summary table

Table 1: Summary Statistics

Statistic	Mean	St. Dev.	Min	Max
wage_income	46,831.870	60,970.610	0	665,000
age	41.660	13.588	18	65
hrs_worked	38.843	12.626	1	99
weeks_worked	47.403	10.510	13	52
hourly_wage	23.988	83.714	0.000	9,246.154

OLS Regression

- Let's write down a baseline model of an individual's hourly wage as a function of their age.

$$\text{Hourly Wage}_i = \beta_0 + \beta_1 \text{Age}_i$$

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

OLS Regression

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

OLS Regression

- Some example code:

```
# 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 ages

```
baseline_model <- lm(hourly_wage ~ age, acs_2016_cleaned)
```

- What are the results?
- How can we interpret the result?
- How much more per hour is a 40 year old expected to earn than a 20 year old?
- What is the structure of the model object?

OLS Regression

How do we add weights?

```
baseline_model <- lm(hourly_wage ~ age, weights = weight,
                     acs_2016_cleaned)
```

stargazer()

- Once again, we can use stargazer to look at the results

```
stargazer(baseline_model, title = "Baseline Model",
          header = F, dep.var.caption = "",
          omit.stat = c("ser", "f"),
          no.space = T)
```

- Lots of options to customize your stargazer table — read more here

Simple regression results

Table 2: Baseline Model

	hourly_wage
age	0.39***
	(0.04)
Constant	7.02***
	(1.79)
Observations	19,929
R^2	0.004
Adjusted R^2	0.004
Note:	*p<0.1; **p<0.05; ***p<0.01

Interpreting the results

- Is the effect of age on income per hour significant
  - Statistically?
  - Economically?
- What is the marginal effect that one year of age has on how much you earn per hour?
- Is this a good model? Why or why not

The Broom Package

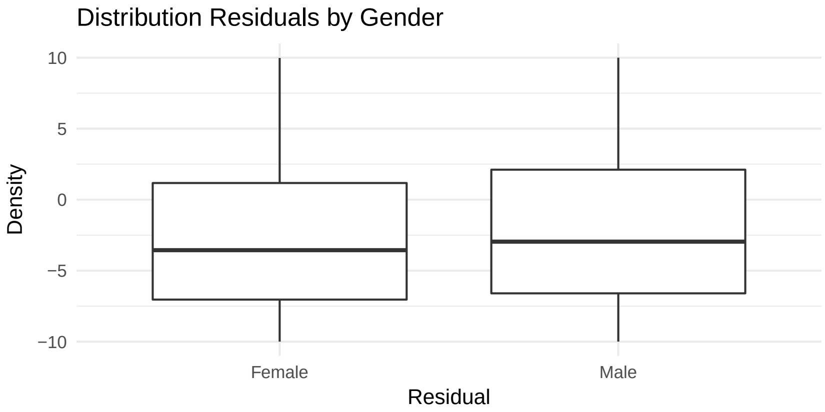
- Model results are messy and hard to work with by themselves in R
- The broom package is there to help!
- The broom package can turn these messy and unfamiliar model objects into good old data frames.
- The three main functions of the broom package are
  - tidy() - for creating a data frame of component 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?



Dummy Variables

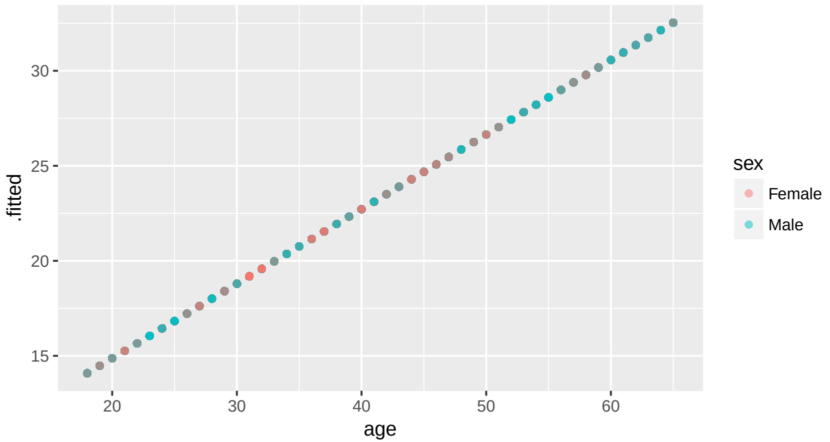
- Use categorical variables in regressions — but first must be transformed to dummy variables
- Dummy variable — any variable that takes on a value of 0 or 1 to indicate whether an observation fits into a particular category.
- For example, in our data:

$$sexMale = \begin{cases} 1 & \text{for} \quad \text{male} \\ 0 & \text{for} \quad \text{female} \end{cases}$$

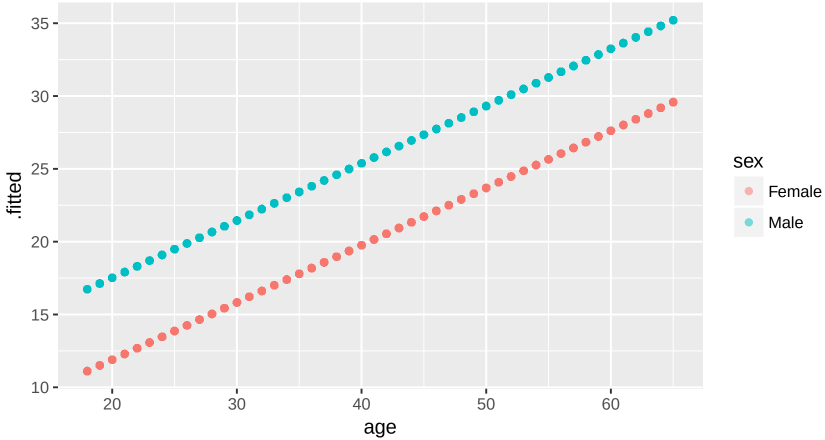
Dummy Variables

- But why do we use dummy variables?
- Recall from you econometrics class that dummy variables allow effects of different levels of a category to vary
  - The difference between no high school diploma and a high school diploma is different than a bachelor's degree and a PhD
- WE NEED TO EXCLUDE ONE DUMMY VARIABLE FROM THE REGRESSION
  - Called the base group
  - Cannot run a regression with all of the dummy variables in the model
- including dummy variables changes the interpretation of our β_0 coefficient

Dummy Variables



Dummy Variables



Improving our model

- Let's run the regression described by

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i$$

- How do the two models compare?

Improving our model

- Let's update the code for our model:

```
improved_model <- lm(hourly_wage ~ age + sex,
                     weights = weight,
                     acs_2016_cleaned)
```

- It's real easy to compare multiple models with stargazer()

```
stargazer(baseline_model, improved_model,
          title = "Model Comparison",
          header = F, dep.var.caption = "",
          omit.stat = c("ser","f"),
          no.space = T)
```

Regression Results

Table 3: Model Comparison

	hourly_wage	
	(1)	(2)
age	0.392***	0.393***
	(0.042)	(0.042)
sexMale		5.623***
		(1.132)
Constant	7.019***	4.036**
	(1.792)	(1.889)
Observations	19,929	19,929
R^2	0.004	0.006
Adjusted R^2	0.004	0.005
Note:	*p<0.1; **p<0.05; ***p<0.01	

Interpreting the results

- Luckily, our interpretation of age is unchanged!
- What is the omitted group?
- $\beta_0 =$ wage of a worker that is 0 years old
- $\beta_2 =$ "bonus" for being a man

R^2

- What is R^2

R^2

- A statistical measure of how close the data are to the regression line.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

What is the range of values R^2 can have?

Adjusted R^2

- Important for models with multiple variables
- Similar to R^2, except there is an *adjustment* for adding additional terms
- A way of testing weather the added variables are actually helping your model

Some more ways of understanding R Squared

Comparing the two models

- Are these coefficients significant:
  - Statistically?
  - Economically?
- Are the coeficients different?
- Now that we know about adjusted R^2, which of the two models is better (marginally)?

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
- 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