



THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Regression Lab

Lecture 20

STA 371G

1. Sit with the other members of your project group—today you'll work with them on a model-building workshop.
2. Log in to Learning Catalytics and click on the Model Building Workshop module.
3. Everyone in your group should enter your group name and number as the team name, e.g., "Group 15".

Reminders

- Midterm 2 is this Thursday/Friday
- Make sure to sign up for a time slot if you haven't already
- You can bring 2 pages of notes to the test

Model building workshop

- Today you'll work through a set of interesting questions about predicting income using data from the General Social Survey (GSS)
- The GSS is an annual survey of attitudes and behaviors that has been conducted since the 1970s
- Much of the work is in getting the data into a form that you can work with and dealing with data issues—this will be good practice for your project!

Loading the data

- Load the GSS data into a variable
- Rename the column to simpler names that are easy to remember and type
- **Very important:** save all of your cleaning steps into a script so you can easily reproduce them later!

Cleaning data pass 1

- Replace missing data codes with NA
- Convert numbers that were read in as character fields (strings) back to numbers
- Transform values to something more meaningful where appropriate

Model building pass 1

- Select a meaningful subset of the data
- Conduct initial model building

Cleaning data pass 2

- Use boxplots to think about which levels are meaningful
- Transform character fields (strings) to levels

Model building pass 2

- Collapse levels where appropriate
- Build another interesting model

Model selection

- Use automated tools (stepwise and best-subsets regression) to help narrow down the possible models
- Examine and remove incomplete rows
- Use the output of automated tools to decide what to explore further