Teaching the Tidyverse in the Second Semester, Undergraduate Statistics Course
Kelly McConville, Swarthmore College, U.S.A.

Background

**Goal:** Modernize my second semester, undergraduate statistics course. Want course to satisfy two popular but conflicting ideas:
- Teach the entire data analysis workflow, of which modeling is only one step.
- Teach a more diverse set of models, especially statistical learning techniques.

**Problem:** How do I find time to teach more of the data analysis workflow and to cover new modeling techniques?

**Proposed Solution:**
- Streamline the process of teaching the data analysis workflow using the Tidyverse.
- Shorten the discussion of specific regression models.
- Use freed up class time to cover predictive modeling techniques.

**Examples:** In this poster, I present example activities which:
- Use Tidyverse packages.
- Emphasize the importance of the Data Wrangling and the Exploration and Visualization steps.
- Reflect an iterative approach to the data analysis workflow.
- Include statistical learning methods.
- Follow a reproducible workflow.

Case Study 1: Are volcanic eruptions increasing?

**Question Formulation:**
- After learning simple linear regression, the students can frame this problem as:
  - Is there a positive, linear relationship between time and number of eruptions?

**Data Acquisition:**
- Data file from the Smithsonian Institution’s Global Volcanism Program website.
  
  ```
  read_csv()
  ```

  Eruptions
  ```
  # A tibble: 118 × 3
  #> Volcano_Number Volcano_Name Eruption_Number
  #> <int> <chr> <int>
  #>  1 1     Aira 2
  #>  2 2     Kamchatka 1
  #>  3 3     Ksitarau 2
  ```

  **Data Wrangling:**
  - Filter by date and confirmed eruptions.
  - Group by start year.
  - Record year, number of eruptions, and average size of eruptions.

  ```
  dataset %>%
  filter(Years) %>%
  group_by(start.Year) %>%
  summarize(mean.size = mean(size))
  ```

**Exploration and Visualization:**
- Students construct graphics to explore multivariate relationships.

**Modeling and Inference:**
- Consider full two-way interaction model with 1,030 potential variables.
- Use cross-validation to select hyperparameters.

**Communicating Findings:**
- Students write up their work using RMarkdown.
- Students also use this data to construct interactive maps of the world’s volcanoes using shiny and leaflet.

Case Study 2: Build a model for household income

**Question Formulation:**
- When covering model selection techniques, the students complete the following task:
  - Build a model for income. Conduct model selection to determine an appropriate set of predictors.

**Data Acquisition:**
- Data from the US Bureau of Labor Statistics Consumer Expenditure Survey.
  - Two files from the fourth quarter of 2015:
    - Household data
    - Data on each individual
  - Merge the principal earner’s information into the household dataset.
  - Resulting in 2,469 households.

**Data Wrangling:**
- Build a model for income. Conduct model selection to determine an appropriate set of predictors.
- Use freed up class time to cover predictive modeling techniques.

**Exploration and Visualization:**
- Students construct graphics to explore multivariate relationships.

**Modeling and Inference:**
- Fit an elastic net model.
- Use cross-validation to select hyperparameters.

**Communicating Findings:**
- Resulting model contains 163 variables.

Conclusions

**Students** get a lot of satisfaction out of making impressive plots with ggplot2 and polished reports with RMarkdown.
- This provides motivation to improve their skills and to overcome errors.
- Students struggle with data wrangling. My suggestions are:
  - Make LOTS of pictures.
  - Use the pipes to breakdown each step.
  - Stress the importance of the wrangling step to the entire workflow.
  - Must drop some topics.
- With freely available or “found” data, it is so important to emphasize the potential pitfalls of generalizing results.

Acknowledgments

I would like to thank the Smithsonian Institution and the US Bureau of Labor Statistics for providing public use datasets. My classes have also greatly benefited from the RStudio Server.

References


Acknowledgments