

Using GitHub and RStudio to Facilitate Authentic Learning Experiences in a Regression Analysis Course

Maria Tackett
Duke University

WSDS 2019



bit.ly/wsd2019-regression

Introduction

Assignments

Challenges
+
Learning
Outcomes



Introduction

Assignments

Challenges
+
Learning
Outcomes

Motivation

“To be prepared for statistics and data science careers, students need facility with professional statistical analysis software, the ability to access and wrangle data in various ways, and the ability to perform algorithmic problem solving.”

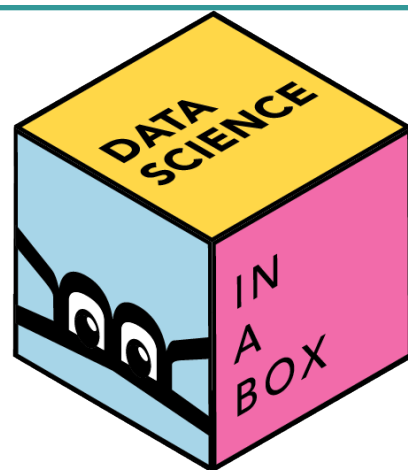
2014 ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science

Motivation

“To be prepared for statistics and data science careers, students need facility with professional statistical analysis software, the ability to access and wrangle data in various ways, and the ability to perform algorithmic problem solving.”

2014 ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science

Innovations in
intro statistics
and data science
courses

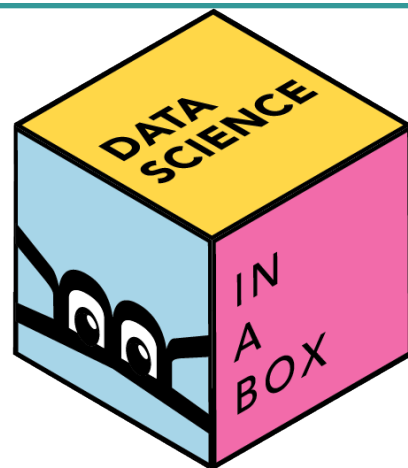


Motivation

“To be prepared for statistics and data science careers, students need facility with professional statistical analysis software, the ability to access and wrangle data in various ways, and the ability to perform algorithmic problem solving.”

2014 ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science

Innovations in
intro statistics
and data science
courses



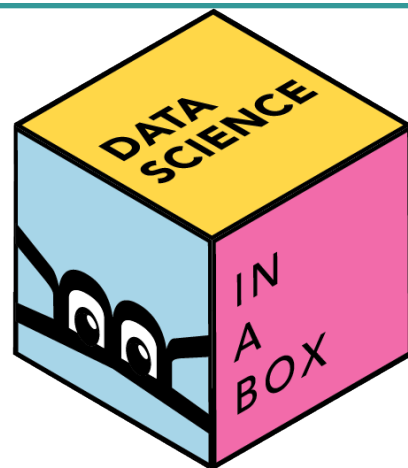
Goal: Create learning experiences to continue cultivating these skills in intermediate courses

Motivation

*“To be prepared for statistics and data science careers, **students need facility with professional statistical analysis software**, the ability to access and wrangle data in various ways, and the ability to perform algorithmic problem solving.”*

2014 ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science

Innovations in
intro statistics
and data science
courses



Goal: Create learning experiences to continue cultivating these skills in intermediate courses

The Course

75 students who have taken introductory statistics, data science, or probability course



Applied course focusing on linear regression, analysis of variance and logistic regression

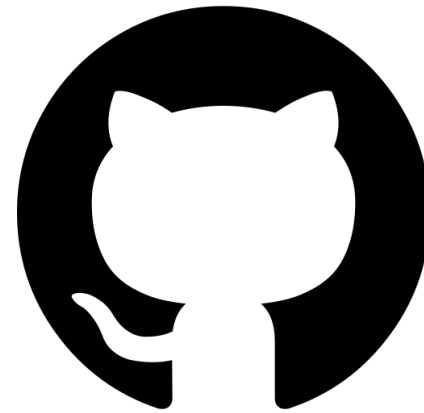
In-class activities + computing labs + homework assignments

Computing Tools



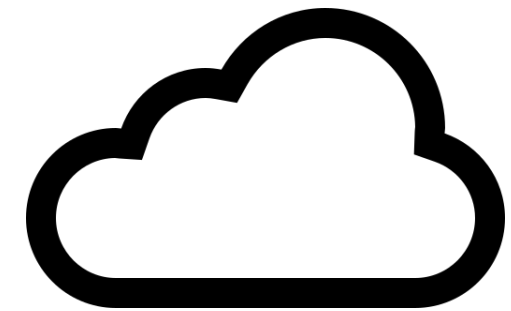
R + R Markdown
for analysis and
write up

Run Git
commands
through Git pane



Assign and submit
assignments

Platform for
collaboration on
group
assignments



RStudio Cloud

Packages installed
+ Git configured
on Day 1

Introduction

Assignments

Challenges
+
Learning
Outcomes

In-Class Activities

1. Create RStudio Cloud project with R Markdown template and data.
2. Students make copy of RStudio Cloud project and answer questions by writing short lines of code and interpreting the results.
3. Display a student's RStudio Cloud project on the classroom screen as the student (or group) shares their work with the class.

```
STA 210: Regression Analysis (Fall 2019) / Cats!

File Edit Code View Plots Session Build Debug Profile Tools Help

cats.Rmd x
Go to file/function Addins

78
79 1. Fill in the code below to fit the regression model describing the relationship between a
80   cat's body weight and heart weight. Display the output.
81 ```{r model, eval = TRUE}
82 bwt_hwt_model <- lm(Hwt ~ Bwt, data = cats)
83
84 tidy(bwt_hwt_model, conf.int = TRUE, level = 0.95) %>%
85   kable(digits = 3)
86 ```
87
88 2. Write the model equation.
89
90 3. Before doing any statistical inference or interpretation, let's check the model assumptions.
91   Fill in the code below to calculate the residuals.
92 ```{r resid, eval = FALSE}
93 cats <- cats %>%
94   mutate(resid = residuals(____))
95 ```
96
97 4. Use the code below to plot the residuals versus the predictor variable. Include an
98   informative title and informative labels for the x and y axes.
99 ```{r resid-v-pred, eval = FALSE}
100 ggplot(data = cats, aes(x = Bwt, y = resid)) +
101   geom_point() +
102   geom_hline(yintercept = 0, color = "red") +
103   labs(x = "____", y = "____",
104        title = "____")
105 ```
106

87:1 # Part 2: Fit the Model & Check Assumptions R Markdown
```

Computing Assignments

The data for this part of the lab is the `Hitters` dataset in the `ISLR` package. Your goal is to fit a regression model that uses the performance statistics of baseball players to predictor their salary. There are 19 potential predictor variables, so you will use the `regsubsets` function to conduct forward selection to choose a final model.

Exercise 7. Read through the data dictionary for the `Hitters` dataset. You can access it by typing `?Hitters` in the console. What is the difference between the variables `HmRun` and `CHmRun`?

Exercise 8. Some observations have missing values for `Salary`. Filter the data, so only observations that have values for `Salary` are included. You will use this filtered data for the remainder of the lab.

Exercise 9. Fill in the code below to conduct forward selection and save the results in an object called `sel_summary` (selection summary).

```
regfit_forward <- regsubsets(_____, _____, method="forward", nvmax = 19)
sel_summary <- summary(_____)
```

The `nvmax` option indicates the maximum-sized variable subsets to consider in the model selection.

Homework Assignments

Part 2: Data Analysis

The *Data Analysis* section of homework contains open-ended data analysis questions. Your response should be neatly organized and read as a complete narrative. This means that in addition to addressing the question there should also be exploratory data analysis and an analysis of the model assumptions. In short, these questions should be treated as “mini-projects”.

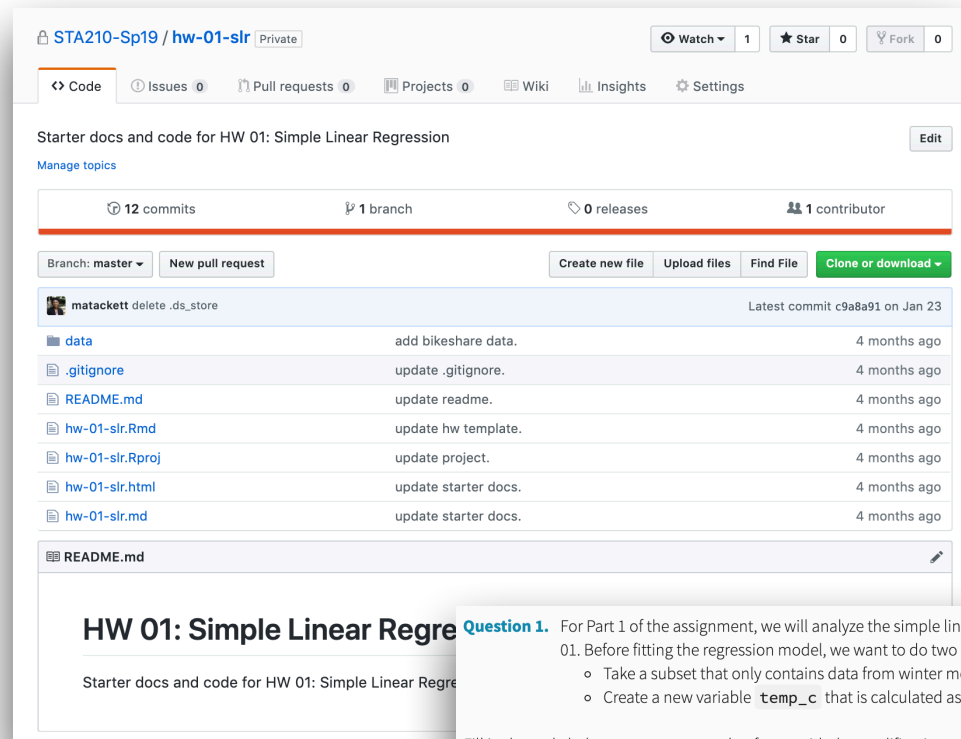
Question 8. For this portion, you will use the `housing` data you started analyzing in [Lab 04](#). Use the code below to load the data and prepare the data.

```
houses <- read_csv("data/KingCountyHouses.csv")
houses <- houses %>%
  filter(bedrooms <= 5 ) %>%
  mutate(floorsCat = as.factor(floors),
         sqftCent = sqft - mean(sqft),
         bedroomsCent = bedrooms - mean(bedrooms),
         bathroomsCent = bathrooms - mean(bathrooms),
         logprice = log(price))
```

Fit a regression model with `logprice` as the response and `floorsCat`, `sqftCent`, `bedroomsCent`, `bathroomsCent`, and `waterfront` as predictor variables. In your analysis, include the following:

- Briefly explain why we should use the log-transformed version of `price` instead of the original version of the variable.
- Describe the relationship between a house’s price and square footage (holding all else constant), including the appropriate confidence intervals.
- Describe how the expected price differs based on the number of floors in the house (holding all else constant). Include discussion about whether or not the differences are statistically significant.

Assignment Workflow



Question 1. For Part 1 of the assignment, we will analyze the simple linear regression model you created in Lab 01. Before fitting the regression model, we want to do two things to prepare the data:

- Take a subset that only contains data from winter months.
- Create a new variable `temp_c` that is calculated as `temp * 41`.

Fill in the code below to create a new dataframe with the modifications stated above and assign the dataframe to `winter_data`.

```
_____ <- bikeshare %>%  
  filter(_____) %>%  
  mutate(_____)
```

Question 2. Fit a regression model with `temp_c` as the predictor variable and `count` as the response and assign the results to `winter_model`. Use the code below to display the model coefficients along with the test statistics and confidence intervals for the coefficients.

```
winter_model %>%  
  tidy(conf.int=TRUE)
```

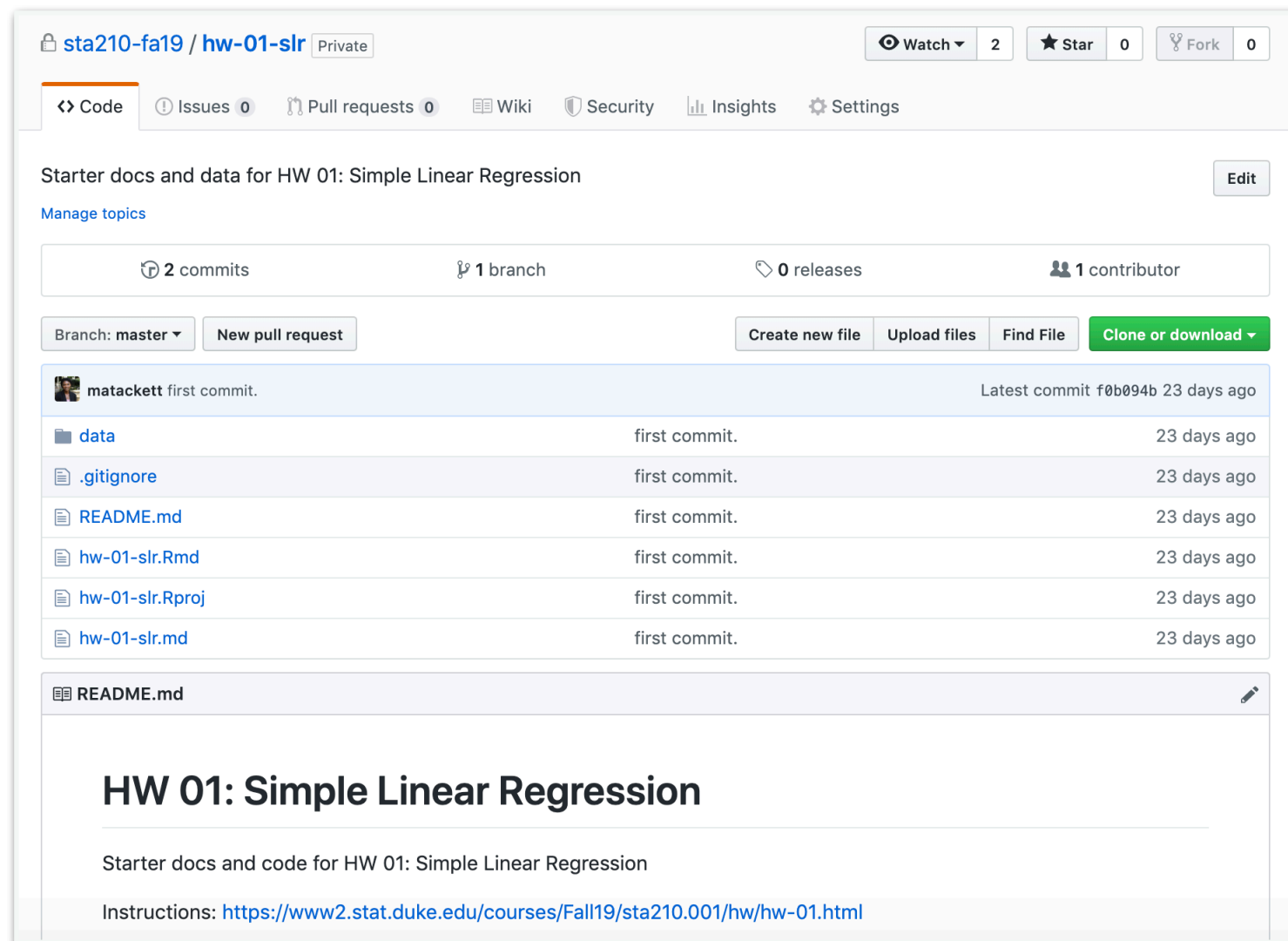


1. Create starter repo in GitHub.
2. Make a copy of the starter repo for each student (or team) using **ghclass**, an R package for managing the GitHub organization for a course.
3. Students clone repo into RStudio, write code and narrative in R Markdown, and push their work to GitHub.
4. Clone repos and grade the completed Markdown files in Gradescope.

ghclass: <https://rundel.github.io/ghclass/articles/ghclass.html>
Gradescope: <https://www.gradescope.com/>

Starter Repo

Create starter repo as a R project and push to GitHub.
Make copy of repo for each student (or team).

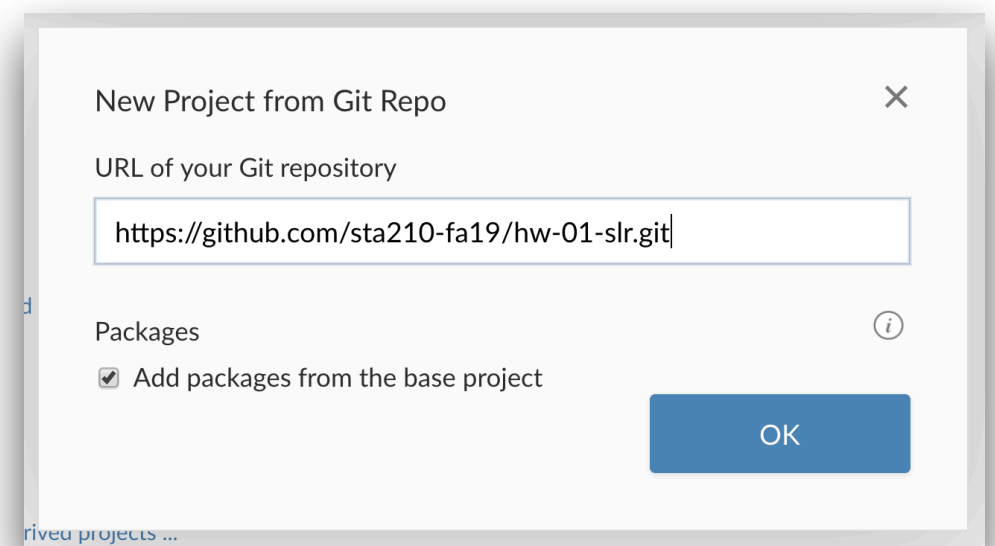
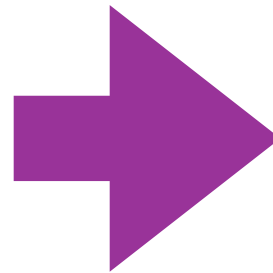
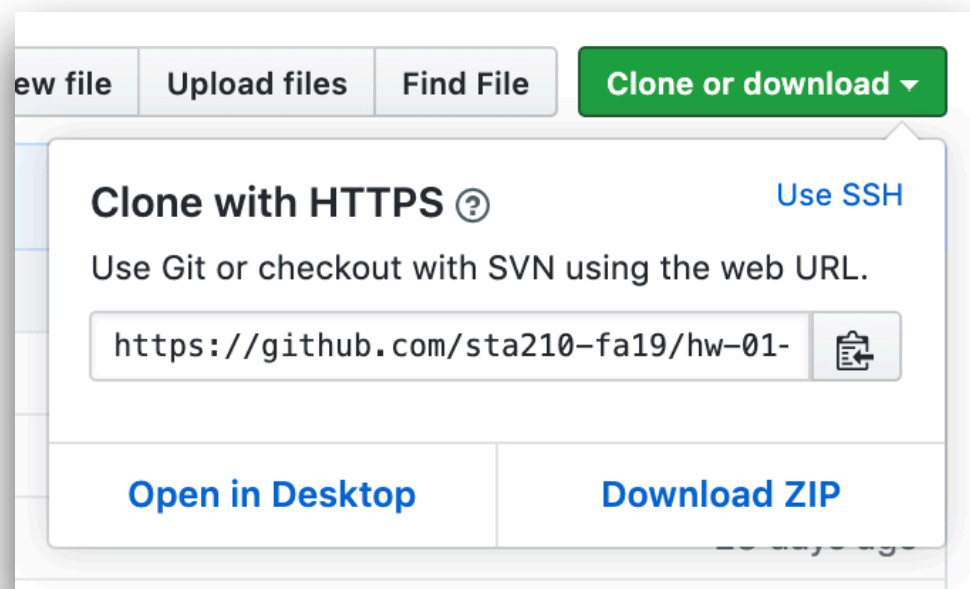


Starter repo includes...

- R Markdown template for student responses
- README with link to assignment instructions
- Folder containing the datasets

Start New Project

Students start a new project in RStudio Cloud from their GitHub repo



Work on Assignment

Students write code and narrative in R Markdown and push work to GitHub periodically as they work on the assignment.

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar below the menu contains icons for file operations and a search bar. The main editor window shows an R Markdown document titled 'hw-01-slr.Rmd'. The document content includes a YAML header with title, author, date, and output settings, followed by R code chunks for loading packages (tidyverse, readr, broom, knitr) and data (bikeshare.csv). The narrative section, titled 'Part 1: Computations & Concepts', describes filtering the data for the winter season. The right sidebar contains the Environment, History, Connections, and Git panels. The Files panel at the bottom right shows a list of files in the project directory, including .gitignore, .Rhistory, data, hw-01-slr.md, hw-01-slr.Rmd, hw-01-slr.Rproj, and README.md. The Console panel at the bottom left is currently empty.

```
1 title: "HW 01 - Simple Linear Regression"
2 author: "Your Name"
3 date: "`r Sys.Date()`"
4 output: github_document
5 ---
6
7
8 ### Load packages & Data
9
10 ```{r load-packages, message=FALSE}
11 library(tidyverse)
12 library(readr)
13 library(broom)
14 library(knitr)
15 ```
16
17 ```{r load-data, message=FALSE}
18 bikeshare <- read_csv("data/bikeshare.csv")
19 ```
20
21 ## Part 1: Computations & Concepts
22
23 ### Question 1
24
25 We begin by filtering the data, so we only have the dates during the winter season.
26
27 ```{r}
28 winter_data <- bikeshare %>%
29   filter(season == 1)
30 ```
```

Name	Size	Modified
..		
.gitignore	50 B	Oct 4, 2019, 11:23 AM
.Rhistory	0 B	Oct 4, 2019, 11:23 AM
data		
hw-01-slr.md	1.2 KB	Oct 4, 2019, 11:26 AM
hw-01-slr.Rmd	1.4 KB	Oct 4, 2019, 11:27 AM
hw-01-slr.Rproj	205 B	Oct 4, 2019, 11:23 AM
README.md	175 B	Oct 4, 2019, 11:23 AM

Grading

Clone the student repos and upload to Gradescope for grading.

The screenshot displays the Gradescope interface for grading a student submission. The main area shows the question content, which includes R code and a table of model results. The right sidebar shows the grading rubric with five items, each with a score and a description. The bottom bar shows the submission status and navigation controls.

Question 2

```
winter_model <- lm(count ~ temp_c, data = winter_data)
tidy(winter_model) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-111.038	238.312	-0.466	0.642
temp_c	222.416	18.459	12.049	0.000

2: Question 2

75 OF 75 GRADED

TOTAL POINTS
2.0 / 3.0 pts

[Rubric Settings](#)
[Collapse View](#)

- +3.0**
Correct model code output with confidence intervals and test statistics.
- +2.0**
Correct model code and output.
- +1.0**
Displayed confidence intervals.
- +0.0**
Exercise incorrect or missing.
- +0.0**
Needs review!

[+ Add Rubric Item](#) [Import...](#)

Submission: 1 of 75

[Previous Ungraded](#) [Previous](#) [Next](#) [Next Ungraded](#)

Introduction

Assignments

Challenges
+
Learning
Outcomes

Challenges

Challenges

Challenges

Different computing backgrounds

- Begin semester with short in-class exercises to expose students to RStudio and GitHub early-on
- Have students work in groups that include diverse computing experiences

Challenges

Different computing backgrounds

- Begin semester with short in-class exercises to expose students to RStudio and GitHub early-on
- Have students work in groups that include diverse computing experiences

Learning curve for Git/GitHub

- Use Git through Git pane in RStudio
- Only focus on basic Git functions (push, pull, commit)

Challenges

Different computing backgrounds

- Begin semester with short in-class exercises to expose students to RStudio and GitHub early-on
- Have students work in groups that include diverse computing experiences

Learning curve for Git/GitHub

- Use Git through Git pane in RStudio
- Only focus on basic Git functions (push, pull, commit)

Handling large classes

- RStudio Cloud (or other server) to reduce computing differences
- **ghclass** package for easy course management

Learning Outcomes

Learning Outcomes

Learning Outcomes

Iterative Workflow

- Students "submit" work frequently by pushing it to GitHub
- Easier for students to incorporate feedback, make changes and include new ideas along the way

Learning Outcomes

Iterative Workflow

- Students "submit" work frequently by pushing it to GitHub
- Easier for students to incorporate feedback, make changes and include new ideas along the way

Professional Computing Skills

- R and GitHub are popular tools in industry
- Students gain computing skills that are applicable for internships, jobs, and higher-level classes

Learning Outcomes

Iterative Workflow

- Students "submit" work frequently by pushing it to GitHub
- Easier for students to incorporate feedback, make changes and include new ideas along the way

Professional Computing Skills

- R and GitHub are popular tools in industry
- Students gain computing skills that are applicable for internships, jobs, and higher-level classes

Professional Collaboration

- GitHub makes it easier to truly collaborate on group assignments
- Students gain skills for collaboration and sharing work in the workplace

Thank You!



maria.tackett@duke.edu



[@MT_statistics](https://twitter.com/MT_statistics)



bit.ly/wsds2019-regression



www.introregression.org

Maria Tackett
Duke University

References

- Çetinkaya-Rudel, M., & Rundel, C. (2017). Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum. *The American Statistician*, 72, 58-65.
- Cloud and GitHub icons from www.flaticon.com
- Course Management with ghclass. Retrieved from <https://rundel.github.io/ghclass/>
- Curriculum Guidelines for Undergraduate Programs in Statistical Science. Retrieved from <https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf>
- Data Science in a Box. Retrieved from <https://datasciencebox.org/>