# Effective communication as a learning objective in an intermediate statistics course

Teaching and Evaluating Communication at Scale Workshop

January 10, 2024

Maria Tackett
Duke University

🔗 bit.ly/effective-communication-2024

Background
+
Motivation

Activities
+
Assessment

Challenges
+
Lessons Learned

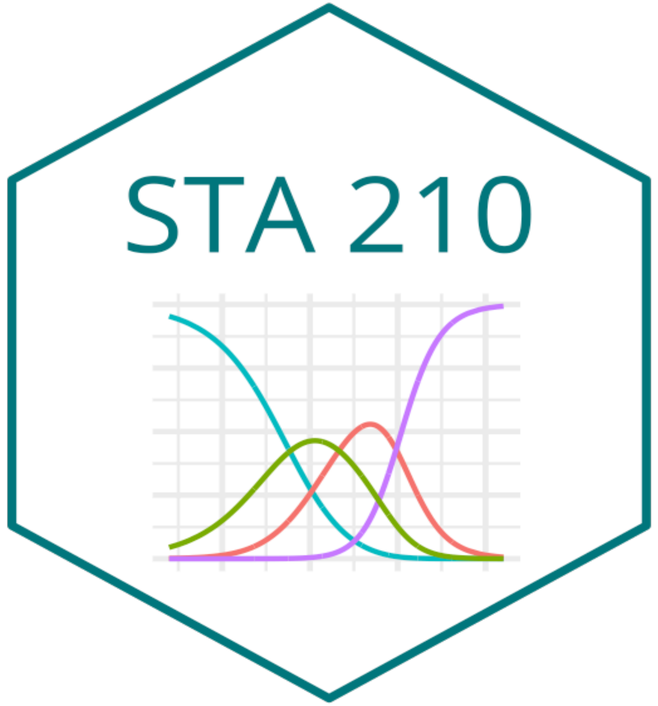
Background
+
Motivation

Activities
+
Assessment

Challenges
+
Lessons Learned

# Courses I teach


Introduction to Data Science




Regression Analysis


Generalized Linear Models

# STA 210: Regression Analysis

A course primarily on linear and logistic regression with a focus on application


**Students:** ~100 who have taken introductory statistics, data science, or probability course

**Class Meetings:** 2 lectures with in-class activities and 1 lab

**Teaching team:** instructor, undergraduate and graduate teaching assistants

**Assessments:** labs, homework, exams, final group project

# Course computing toolkit


Write reproducible analysis reports using Quarto


Implement version control and collaboration


Submit assignments and receive feedback through online rubric

# 2014 ASA Undergraduate Curriculum Guidelines

*"Effective statisticians at any level need to master an integrated combination of skills built upon statistical theory, statistical application, data management, computation, mathematics, and* ***communication*** *…****These skills need to be introduced, supported, and reinforced throughout a student's academic program****…Such scaffolded exposure helps students connect statistical concepts and theory to practice."* (pg. 9)

ASA Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines for Undergraduate Programs in Statistical Science*

# 2014 ASA Undergraduate Curriculum Guidelines

*"Graduates should be expected to **write clearly**, speak fluently, and **construct effective visual displays and compelling written summaries**…They should be able to **communicate complex statistical methods in basic terms** to managers and other audiences and **visualize results in an accessible manner**. Undergraduate majors in statistics often will be hired into analyst positions, where they need to be able to **understand and communicate statistical findings**." (pg. 10)*

ASA Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines for Undergraduate Programs in Statistical Science*

# Challenges for students

- Moving from interpretations to big picture conclusions
- Understanding why precise language matters
- Writing interpretations in a meaningful way
- Determining how much detail to include in reports
- Choosing visualizations that add value
- Effectively presenting results to a non-technical audience

# STA 210 learning objectives

By the end of the semester, you will be able to…

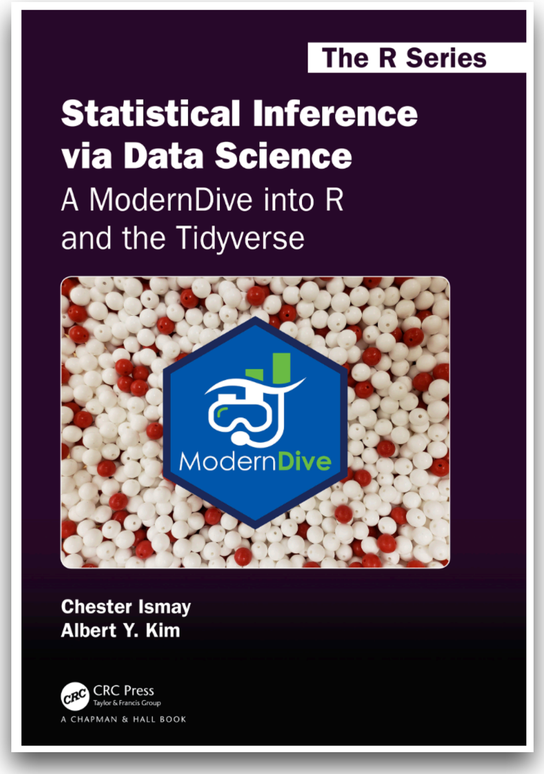
- analyze real-world data to answer questions about multivariable relationships.
- use R to fit and evaluate linear and logistic regression models.
- assess whether a proposed model is appropriate and describe its limitations.
- use Quarto to write reproducible reports and GitHub for version control and collaboration.
- effectively communicate statistical results through writing and oral presentations.

Background
+
Motivation

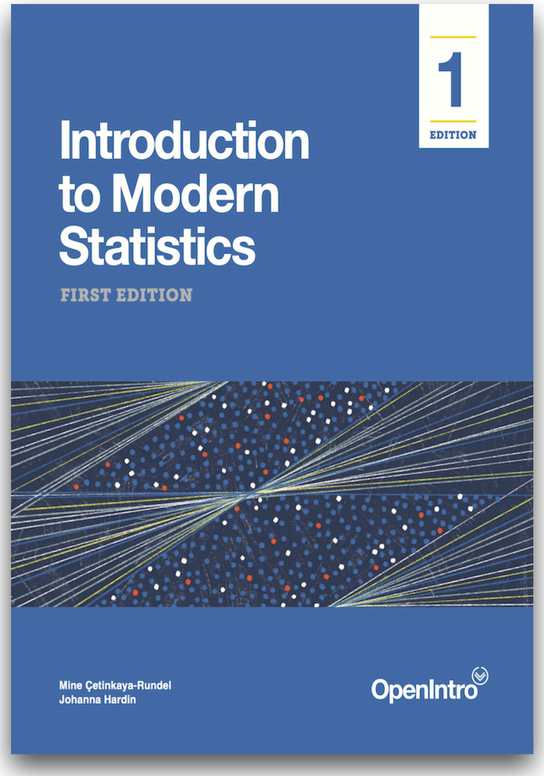
Activities
+
Assessment

Challenges
+
Lessons Learned

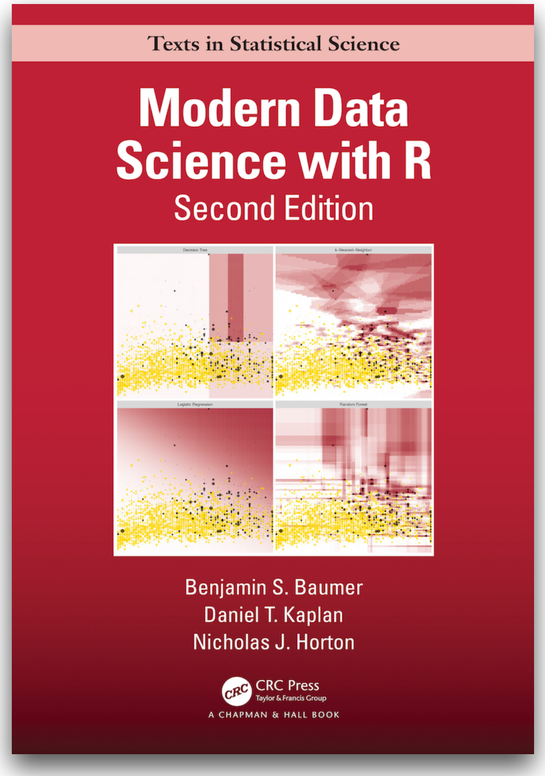
# Build on skills from introductory course



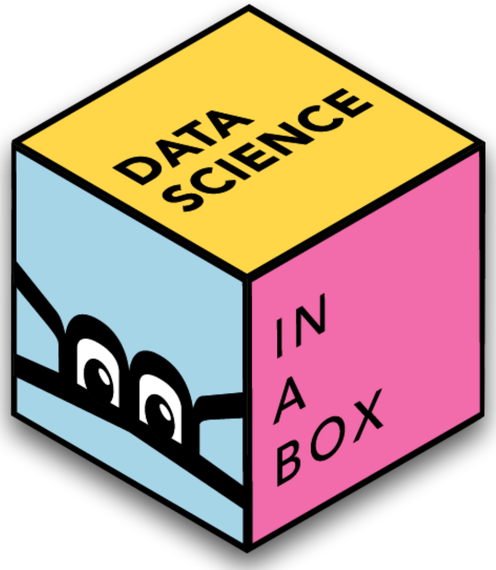

[Modern Dive]



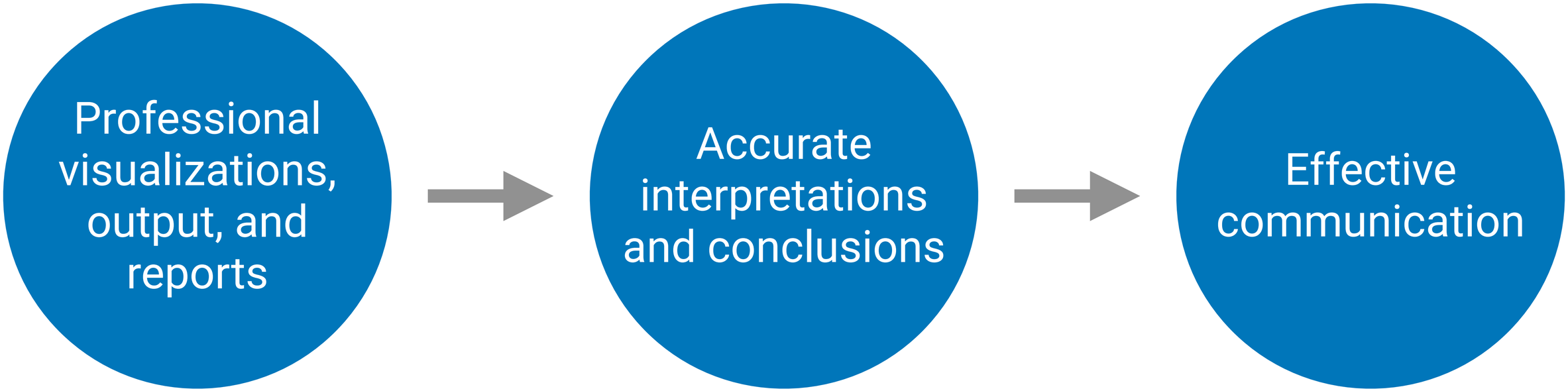

[Introduction to Modern Statistics]




[Modern Data Science with R]

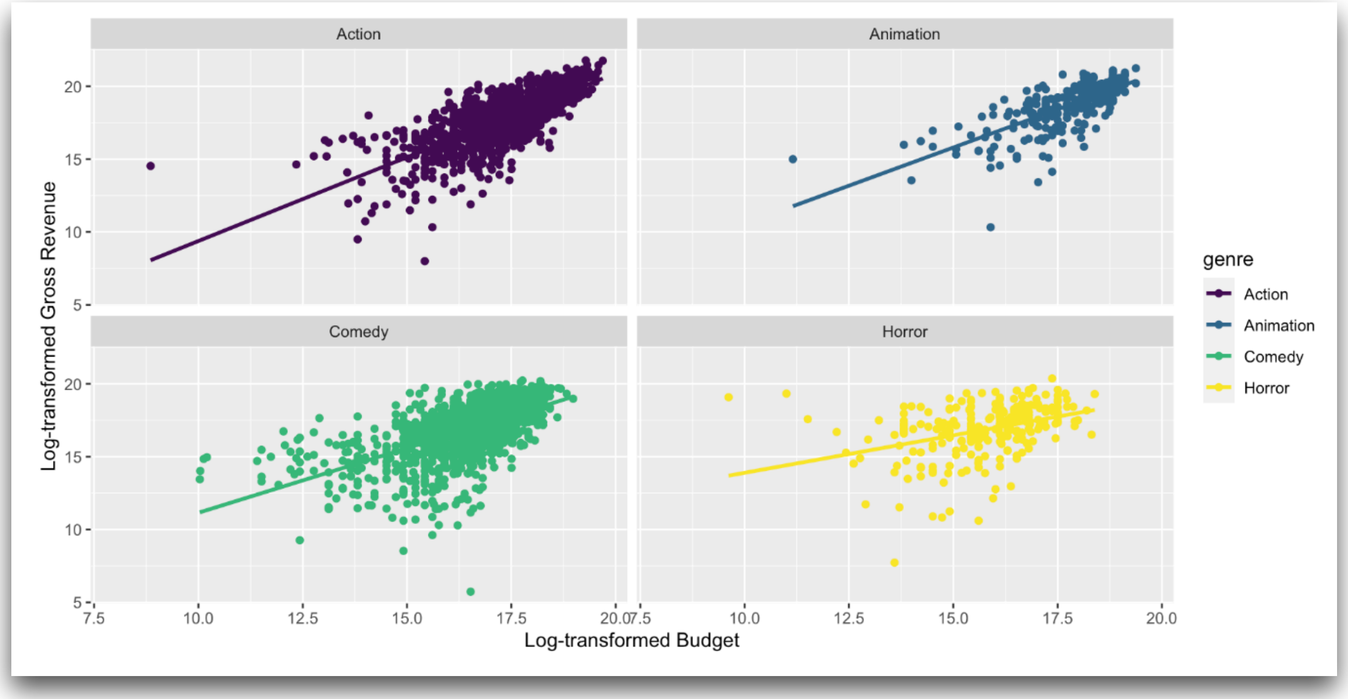



[Data Science in a Box]


Professional visualizations, output, and reports
Accurate interpretations and conclusions
Effective communication

# First day of class

In-class exercise exploring relationship between movie budgets and revenues



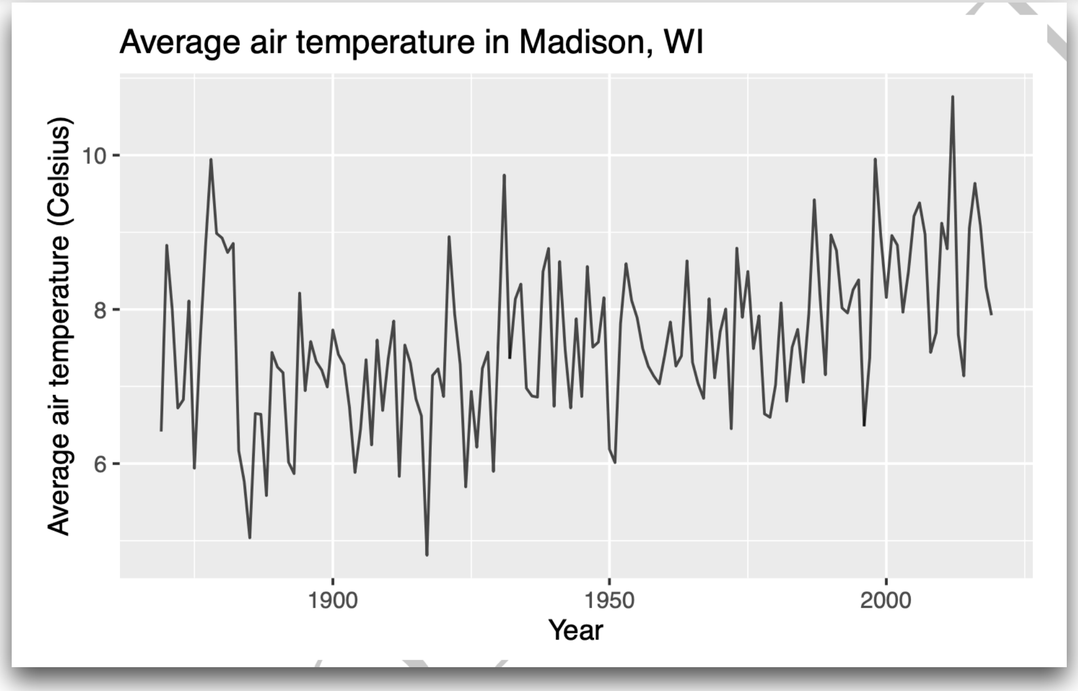

- General observations from the plot
- Observations about the relationship between the variables and differences by genre
- Discussion about advantages / disadvantages of visualizing log-transformed values

# Assessing visualizations early in semester

## Exercise 6

Create a visualization of `avg_air_temp` versus `year`. Include informative axis labels and an informative title on the visualization.

- Use the visualization to write two observations about the trend of average air temperature over time.
- Based on the visualizations of average ice duration and average air temperature over time, would you expect the linear model describing the association between average ice duration and average air temperature to have a positive or negative slope? Briefly explain.




**Total Points**

**5.0 / 5.0 pts**

Rubric Settings

Collapse View

1 **-0.0**

Full Credit:
1. Appropriate plot with informative axis labels and title
2. Correct variables on the x and y axis
3. Two correct and relevant observations about the trend of air temperature over time
4. Conclusion that there would be a negative slope
5. Correct reasoning for why there would be a negative slope

2 **-1.0**

One criteria incorrect or missing

3 **-2.0**

Two or three criteria incorrect or missing

4 **-4.0**

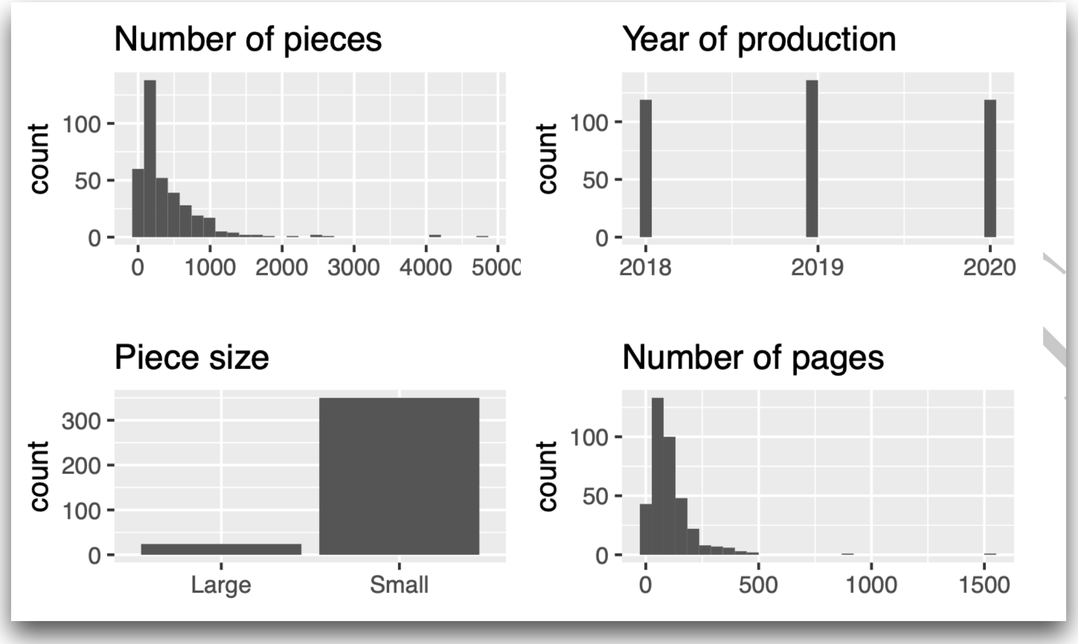
Four or five criteria incorrect or missing

sta210-fa22.netlify.app/labs/lab-02

# Assessing visualizations later in semester

## Exercise 2

Visualize the distributions of the predictor variables `Pieces`, `Size`, `Year`, and `Pages`. Neatly arrange the plots using the patchwork package.




**Total Points**

**4.0 / 4.0 pts**

Rubric Settings

Collapse View

1 **-0.0**

Full credit. Answer includes:
1) Appropriate visualizations of the specified variables with informative labels
2) Arranges the plots using patchwork.

2 **-2.0**

Criteria 1 incorrect or missing

3 **-2.0**

Criteria 2 incorrect or missing

4 **-4.0**

Blank

sta210-fa22.netlify.app/hw/hw-03

# Assessing regression output and interpretations

## Output

- ✓ Accuracy
- ✓ Neatly formatted table
- ✓ Reasonable number of digits

## Interpretations

- ✓ Accuracy
- ✓ Non-causal language
- ✓ Written in context
  - Use variable names
  - Use units

# ⚠️ Avoid points for individual components

**Question 4.** Fit the regression model. *(You can use the R functions to fit the model; yo*
*to calculate the slope and intercept by hand).*

- Display the model output, including the 95% confidence interval f
  slope and intercept.
- Write the regression equation using mathematical notation.

sta210-sp21.netlify.app/hw/hw-01

**Total Points**

**3.0 / 5.0 pts**

Rubric Settings

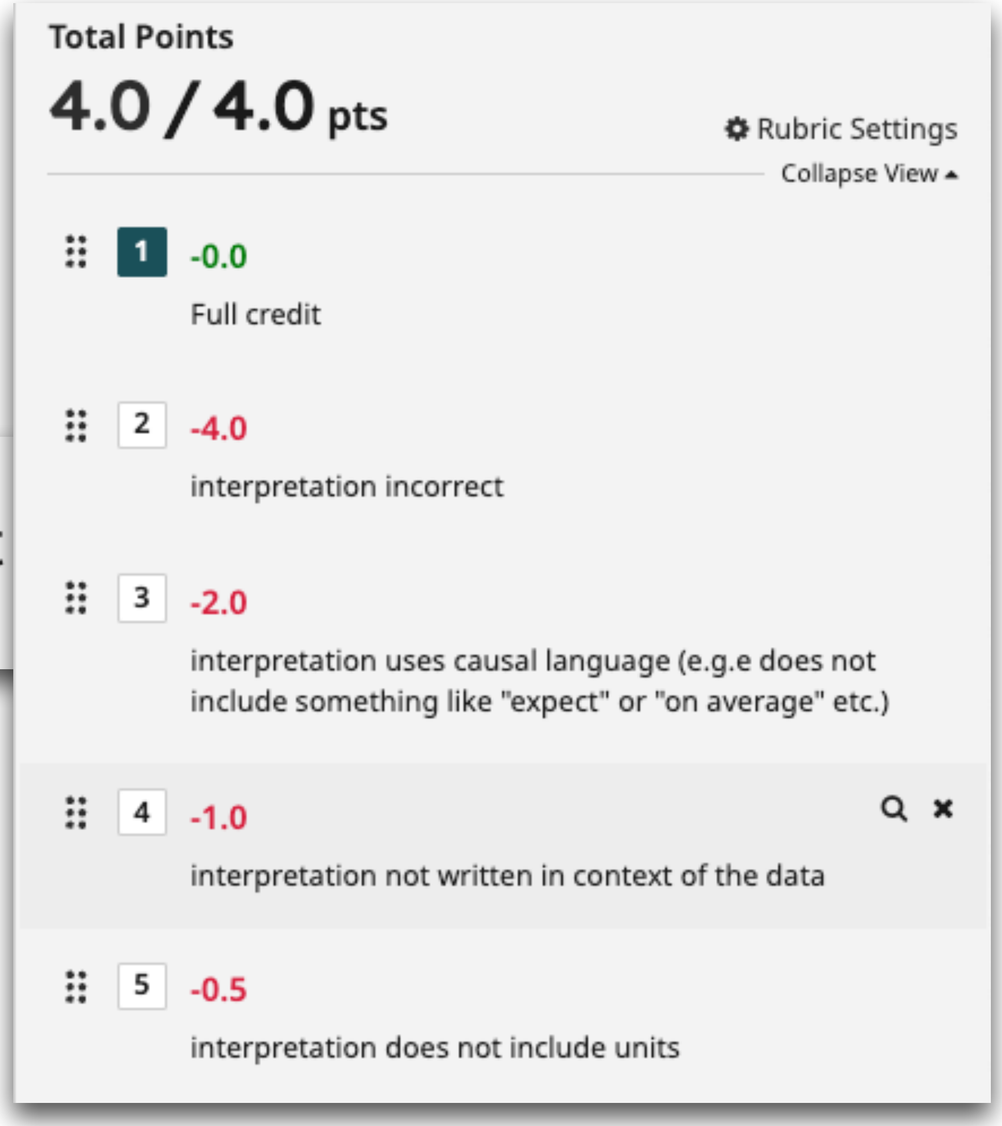
Collapse View

| | | |
|---|---|---|
| 1 | -0.0 | Full credit |
| 2 | -2.0 | no model output displayed |
| 3 | -1.0 | model output does not include 95% confidence interval |
| 4 | -2.0 | equation not consistent with output |
| 5 | -2.0 | equation not written using LaTex |
| 6 | -1.0 | equation does not use variable names |
| 7 | -1.0 | no "hat" above predicted values |
| 8 | -1.0 | included error term in the equation |
| 9 | -1.0 | Filpped response and predictor |

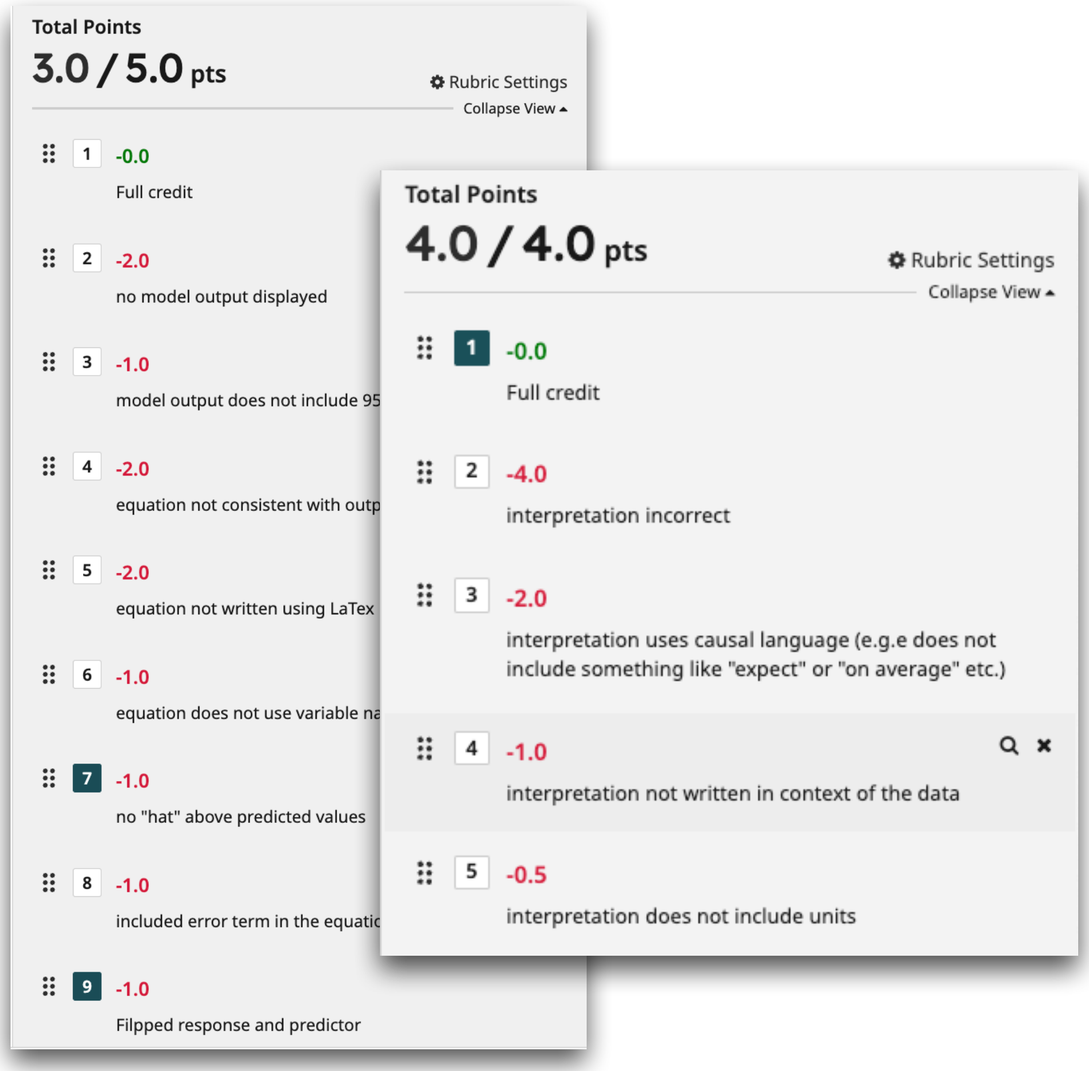
# ⚠️ Avoid points for individual components

**Question 6.** Interpret the slope in the context

sta210-sp21.netlify.app/hw/hw-01



# ⚠ Avoid points for individual components



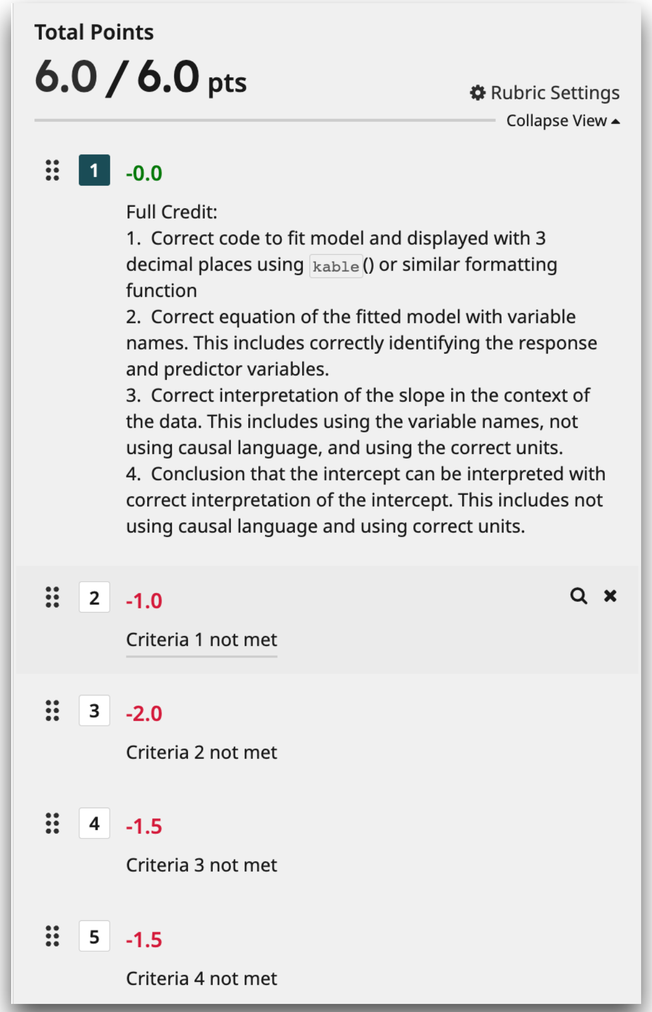

- Difficult for students (and me!) to understand point allocation for each component
- Encourages a “check the box” approach over understanding

# Assessing output and interpretations

## Exercise 8

Fit and display the output of the regression model corresponding to the statistical model in the previous exercise. **As in the previous exercise, only observations from Lake Monona should be included in the analysis.** Use the `tidy` and `kable` functions to neatly display the model output using three decimal places.

- Write the equation of the fitted model. Use mathematical notation and variable names (`avg_air_temp` and `avg_ice_duration`) in the equation.
- Interpret the slope in the context of the data.
- Does the intercept have a meaningful interpretation in this context? If so, interpret the intercept in the context of the data. Otherwise, explain why not.




sta210-fa23.netlify.app/labs/lab-02

# Assessing output and interpretations

## Exercise 8

Fit and display the output of the regression model corresponding to the statistical model in the previo… includ… three …

- W… (a…
- In…
- D… in…

**Exercise 8**

```
monona <- ice_air_joined |>
  filter(lakeid == "Lake Monona")

ice_fit <- linear_reg() |>
  set_engine("lm") |>
  fit(avg_ice_duration ~ avg_air_temp, data = monona)

tidy(ice_fit) |>
  kable(digits=3)
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 149.177 | 10.987 | 13.578 | 0 |
| avg_air_temp | -6.248 | 1.433 | -4.361 | 0 |

$$\widehat{avg\_ice\_duration} = 149.177 - 6.248 \times avg\_air\_temp$$

For each one degree Celsius increase in the average air temperature, the average ice duration at Lake Monona is expected to decrease by 6.248 days, on average.

Yes. There are values close to 0 degrees Celsius, so it is reasonable to interpret the intercept. The average ice duration for years with average air temperature of 0 degrees Celsius is expected to be about 149.177 days.

**Total Points**

**6.0 / 6.0 pts**

Rubric Settings

Collapse View

1 **-0.0**

Full Credit:
1. Correct code to fit model and displayed with 3 decimal places using `kable()` or similar formatting function
2. Correct equation of the fitted model with variable names. This includes correctly identifying the response and predictor variables.
3. Correct interpretation of the slope in the context of the data. This includes using the variable names, not using causal language, and using the correct units.
4. Conclusion that the intercept can be interpreted with correct interpretation of the intercept. This includes not using causal language and using correct units.

2 **-1.0**

Criteria 1 not met

3 **-2.0**

Criteria 2 not met

4 **-1.5**

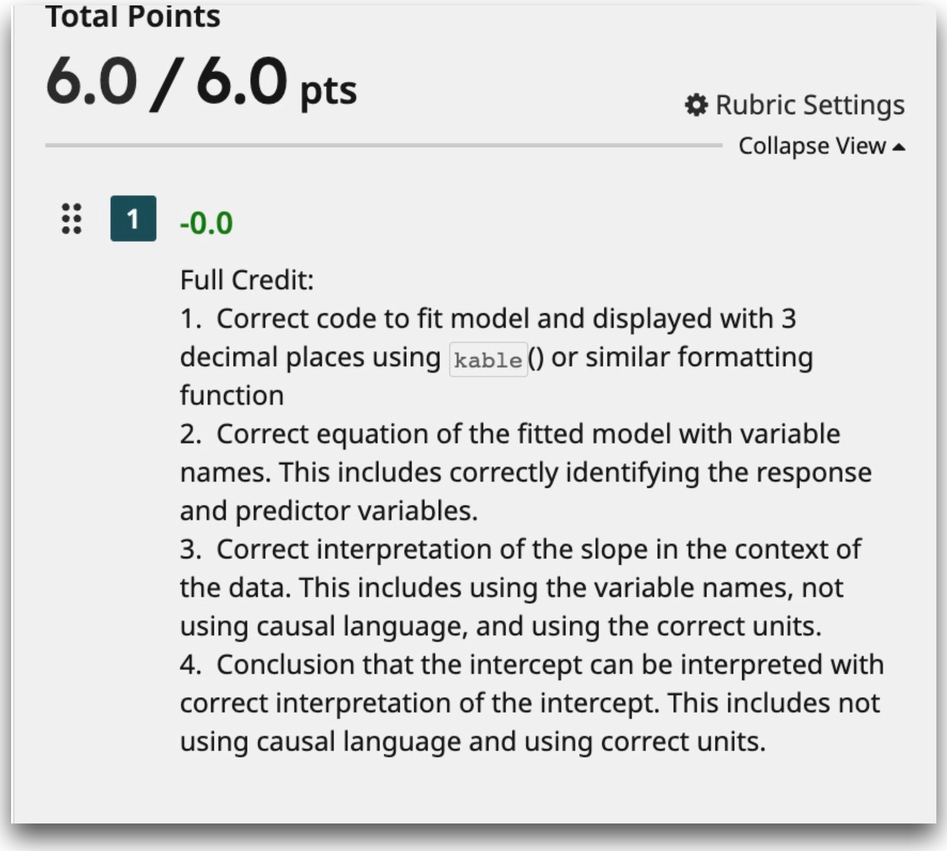
Criteria 3 not met

5 **-1.5**

Criteria 4 not met

sta210-fa23.netlify.app/labs/lab-02

# ✅ Holistic grading for interpretations




- Helps emphasize how all components connect
- Encourages focus on understanding rather than “check the box” approach
- Easier to grade (kind of)
- Most effective with written feedback in addition to rubric items

# “What’s the ‘so what’ ?”

- Goal is for students to get beyond interpretations and communicate what they learn from the analysis
- Create assignments in which students analyze data, then …
  - Evaluate if the conclusions they draw from the analysis are consistent with previously written conclusions
  - Write one paragraph summarizing the results
- Assess analysis and summary separately to more easily identify student misunderstanding

# “What’s the ‘so what’ ?”

Students analyze data  from the FiveThirtyEight article “Why Many Americans Don’t Vote” to examine relationship between demographics, political party, and odds of being a “frequent voter”

## Exercise 5

Let’s start by fitting a model using the demographic factors - `ppage`, `educ`, `race`, `gender`, `income_cat` - to predict the odds a person is a frequent voter.

- Split the data into training (75%) and testing sets (25%). Use a seed of `29`.
- Fit the model on the training data. Display the model using 3 digits.
- Consider the relationship between `ppage` and one’s voting behavior. Interpret the coefficient of `ppage` in the context of the data in terms of the **odds** a person is a frequent voter.

sta210-fa23.netlify.app/hw/hw-04

# “What’s the ‘so what’ ?”

Students analyze data  from the FiveThirtyEight article “Why Many Americans Don’t Vote” to examine relationship between demographics, political party, and odds of being a “frequent voter”

## Exercise 7

Display the model chosen from the previous exercise using 3 digits.

Then use the model selected to write a short paragraph (2 - 5 sentences) describing the effect (or lack of effect) of political party on the odds a person is a frequent voter. The paragraph should include an indication of which levels (if any) are statistically significant along with specifics about the differences in the odds between the political parties, as appropriate.

sta210-fa23.netlify.app/hw/hw-04

# “What’s the ‘so what’ ?”

Students analyze data  from the FiveThirtyEight article “Why Many Americans Don’t Vote” to examine relationship between demographics, political party, and odds of being a “frequent voter”

## Exercise 8

In the article, the authors write

> *“Nonvoters were more likely to have lower incomes; to be young; to have lower levels of education; and to say they don’t belong to either political party, which are all traits that square with what we know about people less likely to engage with the political system.”*

Consider the model you selected in Exercise 6. Is it consistent with this statement? Briefly explain why or why not.

sta210-fa23.netlify.app/hw/hw-04

# Putting it all together

- Final project in groups of 3 - 4 students
- *“The goal of the final project is for you to use regression analysis to analyze a data set of your own choosing.”*
- Project components
  - Proposal
  - Draft report + peer review
  - Project meetings (optional)
  - **Round 1 submission (optional)**
  - Presentation + presentation comments
  - Written report
  - Reproducibility and repo organization

sta210-fa23.netlify.app/project-instructions

# Project: Round 1 submission

*The Round 1 submission is an opportunity to receive detailed feedback on your analysis and written report before the final submission.*

*Therefore, to make the feedback most useful,* ***you must submit a complete written report to receive feedback.***

*You will also be notified of the grade you would receive at that point.*

*You will have the option to keep the grade (and thus you don't need to turn in an updated report) or resubmit the written report by the final submission deadline to receive a new grade.*

sta210-fa23.netlify.app/project-instructions

# Project: Round 1 submission

Use same rubric they see for final submission along with written feedback

| | Excellent | Strong | Satisfactory | Needs improvement | Incomplete / missing |
|---|---|---|---|---|---|
| **Introduction** | | | | | |
| **Data** | | | | | |
| **Method** | | | | | |
| **Results** | | | | | |
| **Discussion and conclusion** | | | | | |
| **Organization and formatting** | | | | | |

sta210-fa23.netlify.app/project-instructions

# Project: Round 1 submission

🙂 Pros

- Gives students an opportunity to improve work resulting in better final projects
- Provides opportunity for more in-depth conversation about the analysis
- Encourages focus on feedback rather than letter grade

😕 Cons

- Time consuming
- Some reports still in very rough draft form

Background
+
Motivation

Activities
+
Assessment

Challenges
+
Lessons Learned

# Challenges

Student buy-in

- Not immediately clear why writing is important in a regression course
- **Idea**: Assign an article/ podcast/video about communicating statistical results

# Statistics experience

Student reflection on how the Stats + Stories podcast episode COVID by Numbers connects to STA 210 (emphasis mine)

*"The podcast likes to reinforce the idea of being precise while conveying statistics, as well as not making statements that could potentially be wrong.* ***Dr. Spiegelhalter explains how statistics is effectively conveyed and understood if described precisely.*** *For example, the false positive rate can mean many different things due to the ambiguity of the denominator.* ***This explains why in class we typically like to use a lot of words in order to analyze our results so that readers won't misunderstand the results and draw false conclusions.***"

sta210-fa23.netlify.app/stats-experience

# Challenges

## Student buy-in

- Not immediately clear why writing is important in a STEM course
- **Idea**: Assign an article/ podcast/video about communicating statistical results

## Consistent grading

- Most grading done by teaching assistants who are new to grading and providing feedback
- **Idea**: Utilize TA meetings to discuss grading questions and do short grading exercises

# Additional information

- American Statistical Association “Curriculum Guidelines for Undergraduate Programs in Statistical Science”: amstat.org/docs/default-source/amstat-documents/edu-guidelines2014-11-15.pdf
- STA 210: Regression Analysis Fall 2023 course website: sta210-fa23.netlify.app
- Tackett, M. (2023). Three Principles for Modernizing an Undergraduate Regression Analysis Course. *Journal of Statistics and Data Science Education*, 31(2), 116-127. doi.org/10.1080/26939169.2023.2165989

# Thank you!

maria.tackett@duke.edu

🔗 bit.ly/effective-communication-2024