Multivariable thinking in algebra-based second courses

eCOTS workshop

Wednesday May 23, 2018; 11:00-12:45PMET

Overview

- Goals: Discussing with you about
  - Goals for a second algebra-based second course in statistics
  - National trends
  - Conceptual teaching strategies/examples with applets you can use in your class
  - Discussing outstanding questions/debates/assessment results

Who are the presenters?

- Beth Chance, Cal Poly
- Karen McGaughey, Cal Poly
- Nathan Tintle, Dordt College

Who you are

- <Results from survey>

- Why attending? Goals of attending?
- Goals of a second course?
- Experience teaching a second course?
- Current second course curriculum?

Overview of session

- 11:00 - 11:15 – Getting to know the audience and presenters (Lead presenter: Nathan Tintle)
- 11:15 – 11:40 – Big picture opportunities for a second course in statistics (Lead presenter: Nathan Tintle)
- 11:40 – 12:00 – Example #1 – Randomized complete block design (Lead presenter: Karen McGaughey
- 12:00 – 12:20 – Examples #2 and #3 – Interaction simulation and Multiple regression visualization (Lead presenter: Beth Chance)
- 12:20 – 12:40 – Assessment, Technology and outstanding questions (Lead presenter: Karen and Beth)
- 12:40 – 12:45 – Next steps (Lead presenter: Nathan Tintle)

Big picture

- Why an algebra-based second course?
  - Lots of students in the first course
  - Historically more calculus, linear algebra, etc. before a second course
  - Multivariable thinking (GAISE)
  - More of what's done in practice
  - Increasing statistics in K-12
  - Alternative entry point to minor/major

Big picture

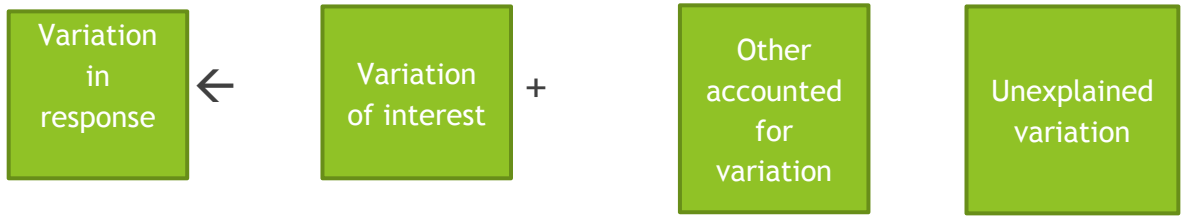
- Goals of a second course
  - Explore multivariable statistical thinking in the context of general modelling framework with
    - A single response variable of any type
    - One or more explanatory variables of any type with additive and interacting relationships
  - How can we make this more conceptual? How can we present students content which is more conceptual?

Content Strategy #1. Focus on explained variation

- Explained variation to drive content; intuitive examples; compelling visualization
- Variable relationship model (next slide)
- Regularly ask students to reflect on three sources of variation in response variable
  - Source(s) of variation of interest?
  - Additional sources of variation being controlled for by design or analysis?
  - How much and possible sources of variation left unexplained?
- At the end of the analysis, what next? What would help explain additional variation (Design? Analysis?)

Content Strategy #1. Focus on explained variation

- Variable relationship model
- Visual



- Symbolic $Y = X_1 + X_2 + \ldots X_n + Z_1 + Z_2 + \ldots + Z_m +$ ERROR
- Verbal

  Variation in the response is explained by variation in sources of interest, variation in other sources we have accounted (or can account for), and unexplained variation

Content Strategy #2. Focus on multivariable thinking

- Potential impact of confounding variables
- Visualizing adjusted vs. unadjusted associations
  - Subtracting off of effects
  - Implications of choice of design
- Patterns in the residuals and how to explain more of the variation
- Additive vs. interaction models
- Use of simulation

Pedagogical strategy #1. Integration of exposition, examples, and explorations

- Multiple paths through materials
- Motivated by context

Pedagogical strategy #2. Easy to use technology

- Finding a bridge between ‘locking into’ a specific software package
  - Giving translatable skills, but also ensuring some proficiency
- Starting with pedagogically focused applets to focus on key conceptual ideas

Pedagogical strategy #3. Real data from genuine studies

- Probably easier for a second course, but still find studies that are accessible, interesting, published research from a variety of fields

Pedagogical strategy #4. Flexible content ordering

- Purposefully developing materials that allow flexibility in ordering (or even choice within a class) so can be maximally impactful for students in the course

Pedagogical strategy #5. Reinforcing key principles

- Strike a balance of new material and review/discussion/ reinforcement of key objectives from the first course
  - Overarching process of statistical inference including looking back and looking ahead
  - Logic and scope of inference and connections to design and analysis strategies

Q+A on overarching themes, goals, and strategies

Example 2: Best ad?

- Does the type of claim or the size of the image impact consumers rating of a product?

		Type of imagery	
		Verbal	Visual
Type of claim	Abstract ("great taste")	2.636	2.909
	Concrete ("won 5 out of 5 taste tests")	2.955	5.545

Example: Best ad?

Sample data: (Response, EV1, EV2)

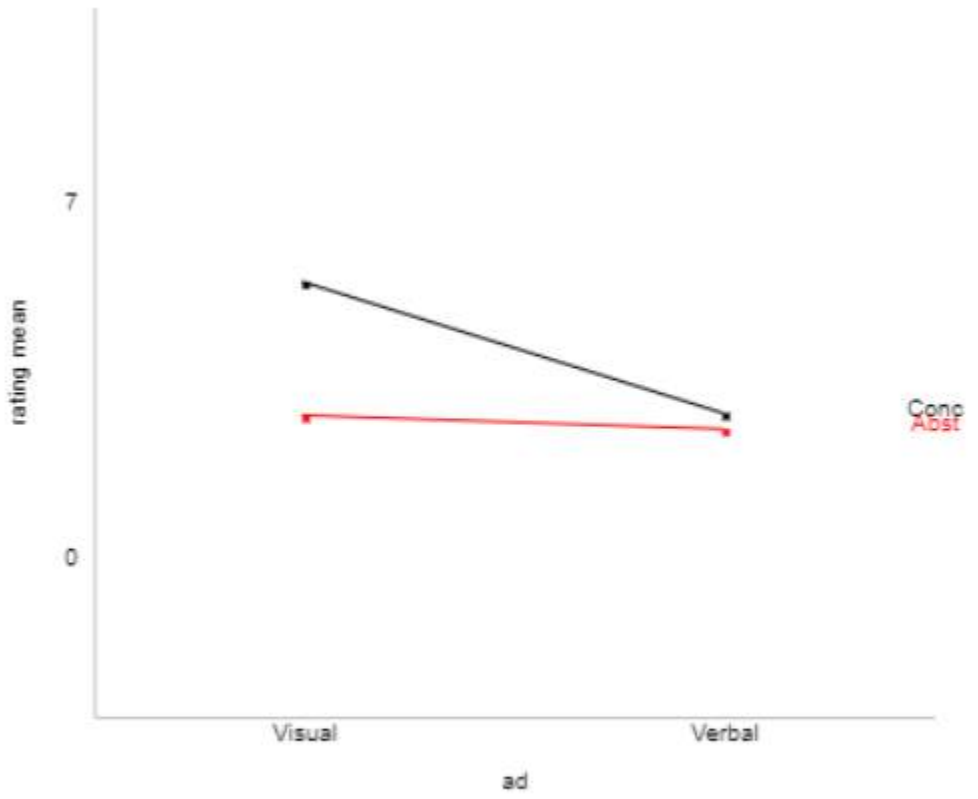
rating	claim	ad
0	Concrete	Visual
5	Concrete	Visual
4	Concrete	Visual
3	Concrete	Visual
6	Concrete	Visual
8	Concrete	Visual
5	Concrete	Visual
6	Concrete	Visual

Use Data | Clear | Top/Bottom

Statistic: Diff in Diffs ▼

Show Means:

Show ANOVA table:



Observed Diff in Diff = 2.32

Simulation?

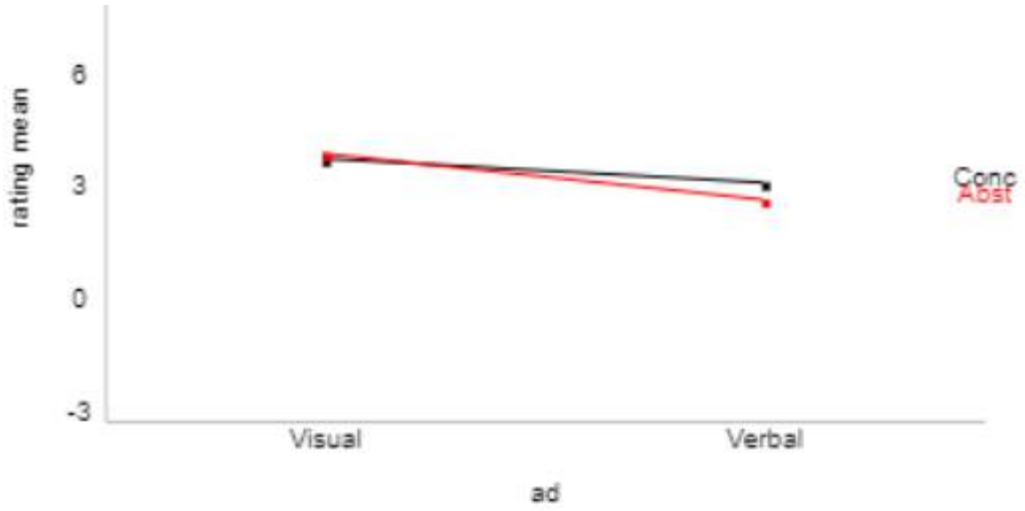
Original data

Sample data: (Response, EV1, EV2)

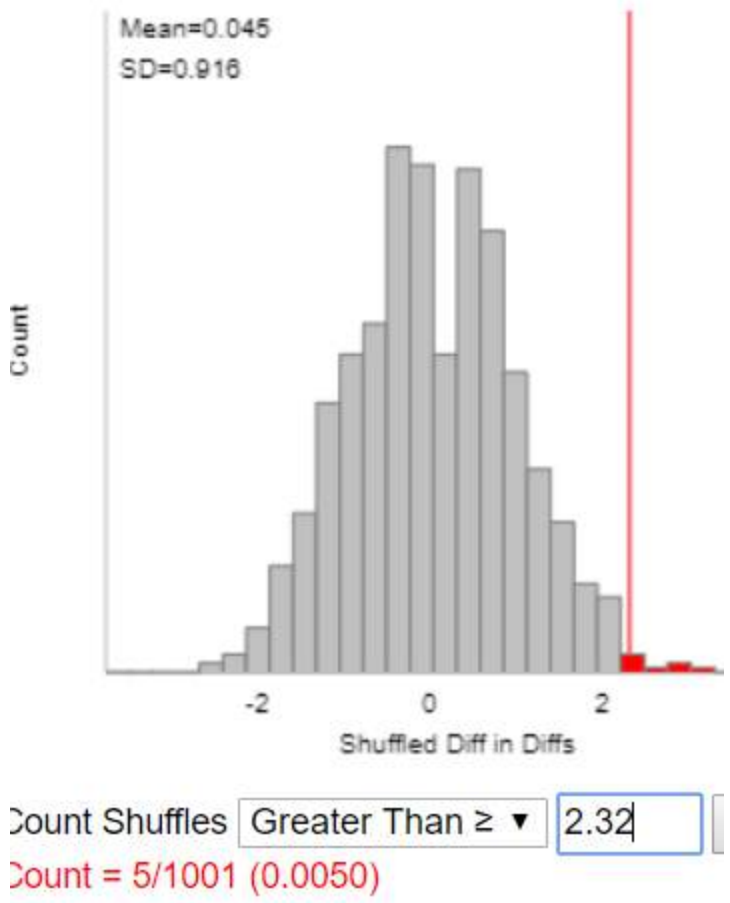
rating	claim	ad
0	Concrete	Visual
5	Concrete	Visual
4	Concrete	Visual
3	Concrete	Visual
6	Concrete	Visual
8	Concrete	Visual
5	Concrete	Visual
6	Concrete	Visual

Shuffled data

rating	claim	ad
6	Concrete	Visual
3	Concrete	Visual
-2	Concrete	Visual
1	Concrete	Visual
6	Concrete	Visual
3	Concrete	Visual
6	Concrete	Visual



Simulation?



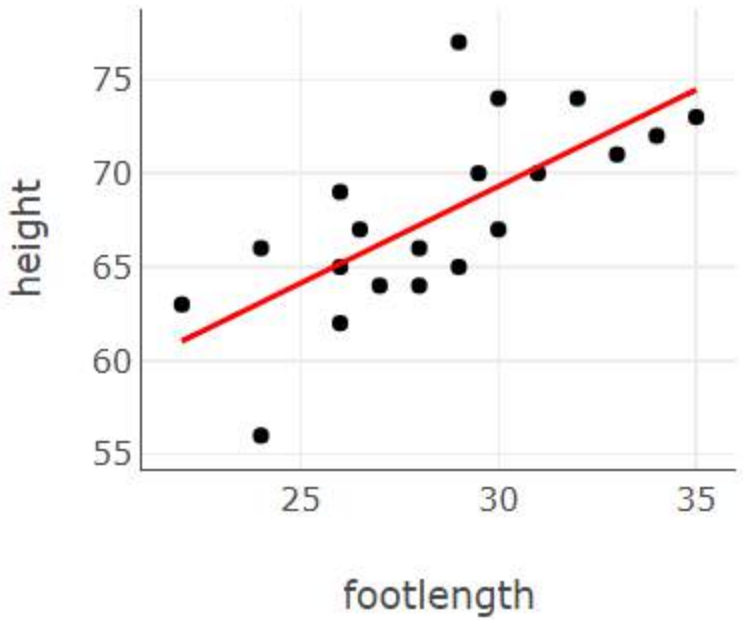
Key Ideas

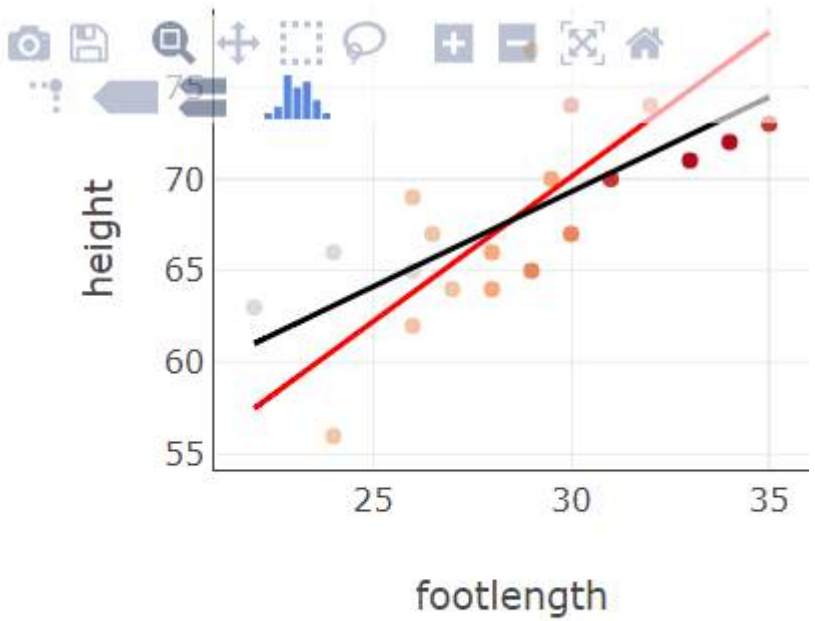
- Shuffling the response maintains the same x-variable structure
- Can focus on a more intuitive statistics and get to the “punchline” sooner

Example 3: Multiple regression

- New applet under development
  - http://www.rossmanchance.com/applets/multreg/multreg7.html
- Auto mpg dataset
  - https://archive.ics.uci.edu/ml/datasets/Auto+MPG
  - Divide weight by 1000

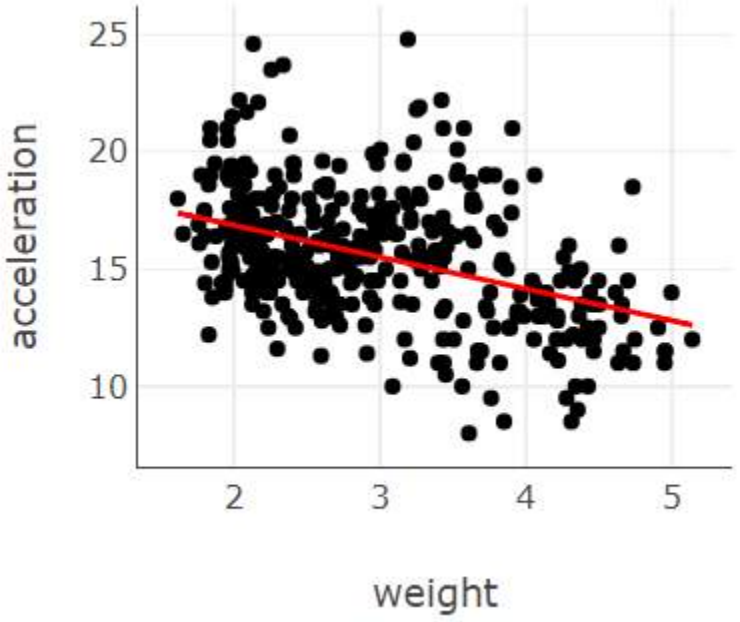
Example: Multiple Regression

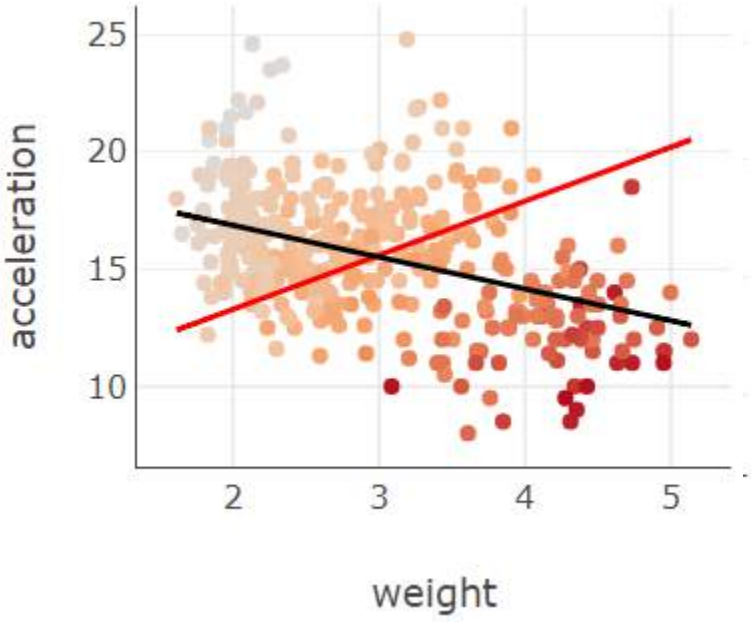




Show Regression Line: ☑
unadjusted footlength slope = 1.03
adjusted footlength slope =1.58

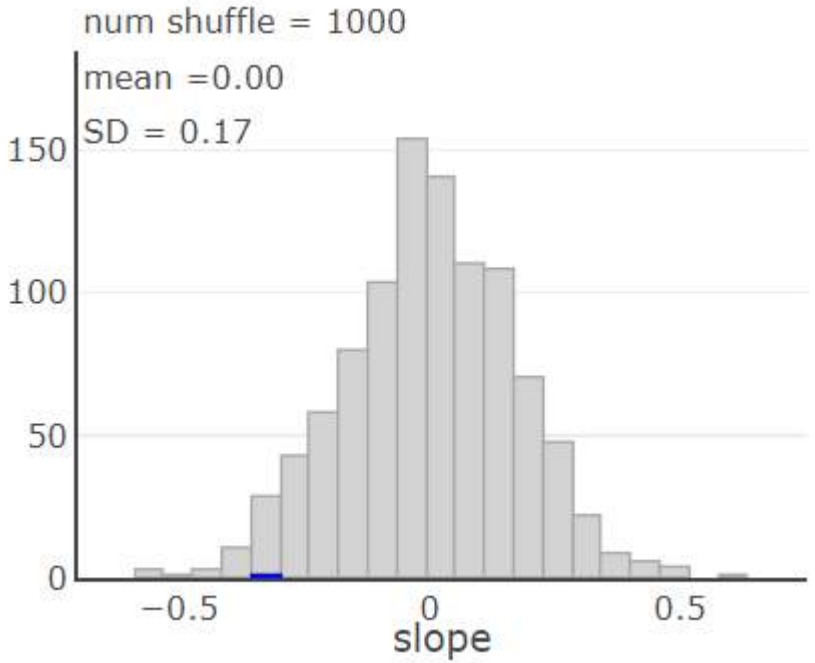
Example: Multiple Regression



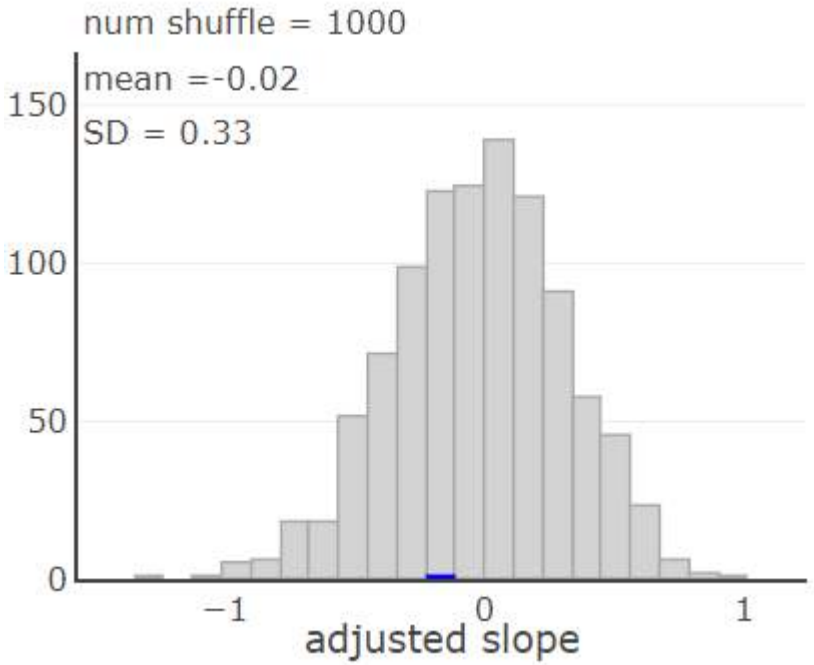


Example: Multiple Regression

Unadjusted slopes

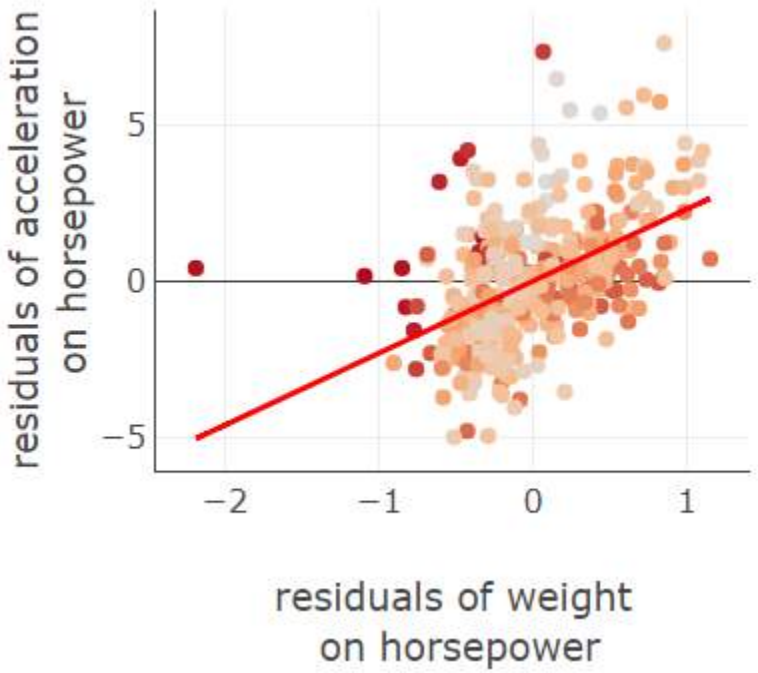


Adjusted slopes



Key Ideas

- Helps student to visualize the adjusted association
- Can see impact of variation inflation from correlated variables
- Quick/easy way to produce added variable plots



Assessment

- Fall class test (Cal Poly)
- Winter/spring class testing (Cal Poly, Dordt, Hope, Georgia, Nebraska)
  - Undergraduate and graduate students
  - Accelerated course
- Pre/post multiple choice questions
- Student feedback
- Final exam questions

Example MC questions

It is suspected that the antioxidants in fruits and vegetables prevent various types of cancers. Suppose a study asks 1000 male volunteers to keep track of the amount of fruit and vegetables they eat over several years. Then the researchers compare the proportions who get colon cancer between the men who ate fruits and vegetables regularly and those who did not eat fruits and vegetables regularly. Suppose the researchers find a statistically significantly smaller proportion with colon cancer among those who eat fruits and vegetables regularly. Which of the following is the best explanation of a potential confounding variable in this study?

	Winter (n = 139)
○ There shouldn't be any confounding variables because the result was statistically significant.	3%
○ The sample sizes between the two groups are probably not the same, so sample size is a potential confounding variable.	8%
○ Only males were involved in this study, so gender is a potential confounding variable.	21.6%
○ Males who eat fruits and vegetables regularly may be more genetically predisposed to not get colon cancer, so genetic predisposition to get colon cancer is a potential confounding variable.	38.9%
○ Colon cancer is more dependent on environmental factors than on diet, so environmental factors are a potential confounding variable.	28.8%

Example MC questions (*REGRESS)

Reconsider the previous question. A second analysis also included minutes of exercise per week (see output below). What do you conclude from the second analysis?

- ○ There must have been an error in the second analysis because the coefficient and p-value of BMI has changed
- ○ There is an association between BMI and amount of exercise
- ○ BMI is not related to pulse rate
- ○ Increasing exercise by one minute per week lowers someone's pulse rate by 1.72 bpm

Pre	Post
59.4%	70-90%
33.3%	

Second analysis

	Coeff	SE	p-value
Intercept	58.2	21.8	0.007
BMI	0.15	0.005	0.021
Exercise	-1.72	0.54	0.002

First analysis

	Coeff	SE	p-value
Intercept	55.3	42.5	0.103
BMI	0.3	0.16	0.036

Example final exam question

- A study is going to be carried out to investigate the impact of taking notes with pen and paper versus using a laptop on success of college students. The study will be carried out in a large general education lecture class, with anywhere from 100-200 students. Success will be measured using the final exam score in the course.

(c) For the observational study, what is the best way to address the issue that GPA could be a confounding variable?

(d) For the experimental study, what are two ways to address the concern that GPA could be a confounding variable?

Example final exam question

- A study was carried out to investigate the impacts of alcohol consumption on inflammation in the body, as measured by the level of C-reactive protein (CRP) in the blood. The table shows the mean and standard deviation of CRP levels for each type of drinker.

Type of Drinker	n	mean CRP (mg/dL)	sd CRP (mg/dL)
Non-drinker	500	3.0	1.5
Light	1120	4.5	1.6
Moderate	910	7.0	2.1
Heavy	496	10.0	2.8

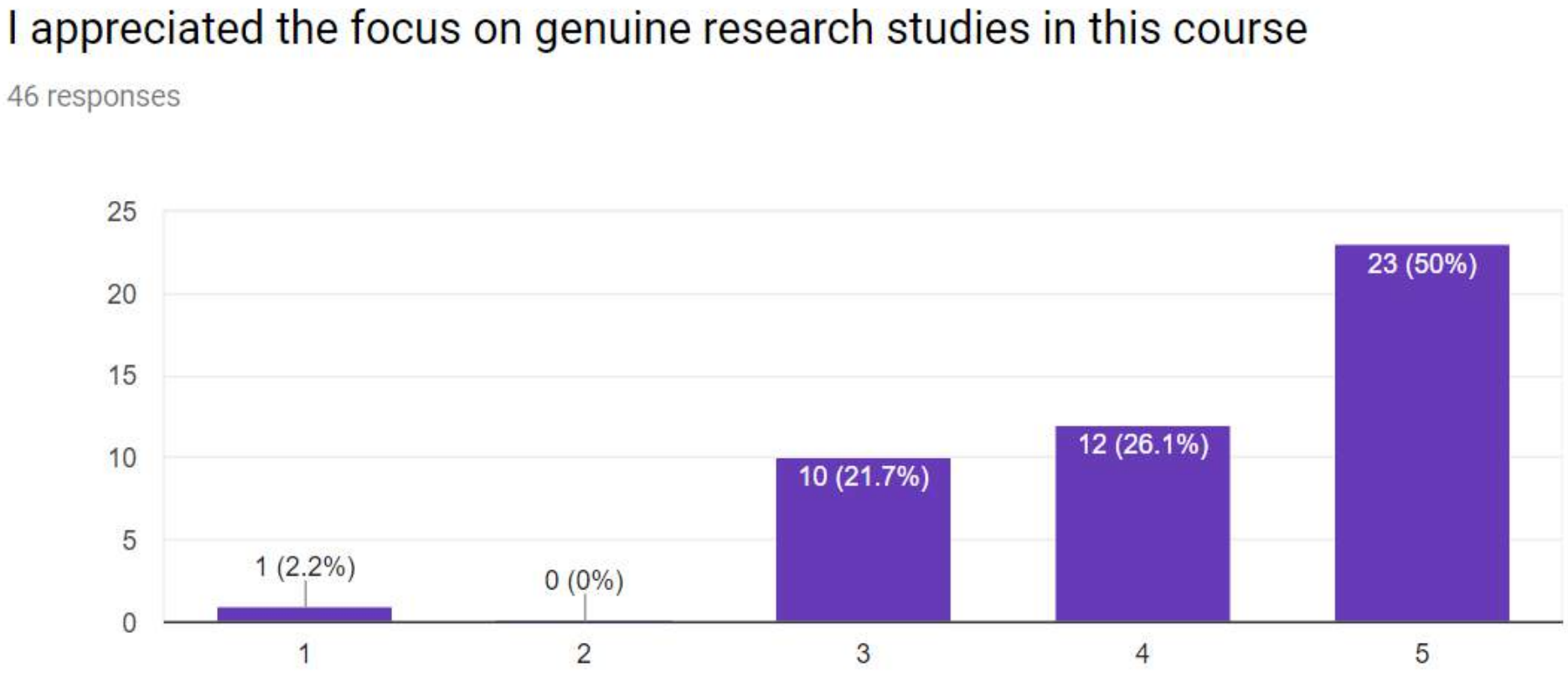
- Because people who tend to engage in one unhealthy practice tend to engage in others, participants were also categorized based on how much time they spent exercising per week: Minimal (0 to 60 minutes/week), Moderate (60 to 120 minutes/week), High (120+ minutes/week). Suppose we fit a model to explain variation in CRP from type of drinker and level of exercise, and we used that model to estimate the mean CRP for each type of drinker. How do you expect these exercise-adjusted means to compare to those in the table above?

Lessons learned

- Getting good information on what students do/do not know coming in
- The “catch up” on simulation-based inference is not a hurdle
- Improvement on most assessment questions

  (60% $\rightarrow$ 90%; 10% $\rightarrow$ 30%)
- Still struggle with interaction vs. independence/confounding/ collinearity

Student feedback (fall/spring)



Open ended responses

- Favorite part of course
  - looking at data with multiple variables and how the impacts differed when certain variables/interactions were included and not included
- Best feature
  - loved how we had an interactive day where we got to work on our own and ask questions when needed. However, this lab part of class would not have been beneficial if we did not get the class time the next day to discuss everything.
  - the course allows us to get practice doing statistical analyses ourselves, and be able to discuss them later.

Future work

- Maintain focus on visualization, multivariable thinking, better understanding of future literature in the field
- Improve linkages of topics/focus on variation explained throughout the course
- Considerations
  - More data science type skills?
  - More applets/role of technology?

Next steps

- Pilot assessment instrument
- Try some materials
  - Contact us:
    bchance@calpoly.edu
    kmcgaughey@calpoly.edu
    nathan.tintle@dordt.edu
- NSF grant DUE-1612201
- Acknowledgments: Soma Roy, Todd Swanson, Jill VanderStoep; numerous class testers and students