THE PROBLEM OF MULTIPLE COMPARISONS

A ROLE-PLAYING ACTIVITY

June 12th 2024
eCOTS 2024

Jonathan Wells
wellsjon@grinnell.edu
Grinnell College | Department of Statistics

Grinnell College

MULTIPLE VARIABLES, MULTIPLE COMPARISONS

- 2015 revised GAISE report encourages instructors to incorporate multivariate thinking in classes.
- However, in exploring multivariate models, students also have potential to commit errors based on the **problem of multiple comparisons (PMC):**
  - Performing multiple significance tests inflates Type I error rate, compared to nominal significance level for each
  - Identifying variables of interest based on observing small p-values also inflates Type I error rate
- P-hacking or data dredging involves errors based on the problem of multiple comparisons, and is one cause of the **replication crisis**
  - Many published results using statistical methods in social and natural sciences cannot be replicated

TEACHING THE PROBLEM OF MULTIPLE COMPARISONS

- The problem of multiple comparisons can be challenging for students to understand
  - The distinction between reporting a p-value for a specific coefficient in a multivariate regression model and finding a variable that has a small p-value is subtle
  - Without experience with probability, students may struggle to understand why the problem of multiple comparisons occur
- Survey of several intro- and intermediate statistics texts reveals several discuss problem of multiple comparisons
- However, homework exercises often do not provide opportunity to make errors based on multiple comparisons
  - Exercises may ask whether a specific variable in multivariate model is significant (no problem of multiple comparisons)
  - Or may explicitly instruct to use specific method to correct for multiple comparisons
- The problem of multiple comparisons is a decision-making error, rather than an implementation error

ACTIVITY: GOALS

- Produce a 50-minute classroom activity on the problem of multiple comparisons that…
  - Creates meaningful, memorable and intuitive experience that reflects authentic structure of the problem of multiple comparisons
  - Facilitates class discussion of the problem of multiple comparisons and motivate potential remedies
  - Provides reference example for later comparison of ANOVA F-tests and t-tests for individual slope coefficients
  - Actively engages students in learning process

ACTIVITY: TEACHING CONTEXT

- Activity was developed for and implemented in a classroom with the following attributes:
  - **Class level:** Statistics 2 on multivariate statistical modeling (Multilinear regression, ANOVA, Logistic Regression)
  - **Class size:** 15 – 20 students
  - **Institutional type:** selective small liberal arts college
- The activity assumes students are familiar with the following:
  - Hypothesis testing and interpreting p-values
  - Creating and interpreting scatterplots and computing correlation
  - Building and interpreting multilinear models
- Activity is most effective if it precedes discussion of ANOVA F-tests for multilinear regression
- Modifications for other classroom structure, level and pre-requisites are possible

ACTIVITY: STUDENT INSTRUCTIONS

- Students arranged into groups of 3 – 4 and asked to imagine they are hired as a team of data scientists for a biotech company
  - Each team's goal is to research a potentially life-saving pharmaceutical intervention
  - Teams have opportunity to provide a name, geographic location and quirk for their company
- Teams provided data for 25 observations on 21 quantitative variables (1 response and 20 covariates)
  - Each team has difference subset of data drawn from the same population
  - Variable names have been removed, and all variables have been standardized with mean 0 and variance 1
- Teams tasked to identify the covariate most strongly associated with the response
  - Teams are given 15 minutes to perform statistical analysis of data and 5 minutes to summarize findings
- Teams that correctly identify most important covariate will receive a significant pay bonus

ACTIVITY: WORK PERIOD

- During 20-minute work period, students given access to data .csv and use R software to explore and analyze
- Teams are prompted to consider the following:
  - Scatterplots and correlations between response and individual covariates
  - Simple linear models and multilinear models
  - Diagnostic plots and model conditions
- Instructor is available to answer technical and implementation questions, and will prompt groups that are stuck, but won't provide substantive modeling advice

ACTIVITY: RESULTS, REVEAL AND DISCUSSION

- After 20 minutes, each team announces to the class which variable identified and reports the observed statistics and p-values supporting their claim
- Invariably, nearly all groups identify a different variable, with correlations above 0.4 and p-values less than 0.1
- Anticipation (and perhaps some confusion) builds as instructor prepares to announce results:
- *None of the variables*, in fact, are intrinsically related to the response
  - All were generated by instructor independently of each other and of the response; any observed correlations in the data set were due to chance alone
- Following the reveal, the instructor provides opportunity for class discussion
- During class or as an after-class reflection, students invited to consider possible remedies for problem.

ACTIVITY: DEBRIEF AND DISCUSSION

- Instructor then invites class discussion about the activity. Some prompts include:
  - What does a p-value for a specific hypothesis test represent? Both informal and formal interpretations
  - Why did teams report a significant p-value for a hypothesis test, if none of the variables were related to the response?
  - Why did teams each identify a different variable?
  - In what ways were teams incentivized to seek the variable with smallest p-value?
  - How might results differ if each group was asked to determine whether a specific variable was associated with response?
- Students may have questions about data structure. Instructor can share code for generating data and can also generate new data set with comparable results

ACTIVITY: REFLECTION AND REMEDIES

- During class discussion or as after-class reflection, students invited to consider possible remedies for problem.
- Analytic remedies:
  - Using lower significance level (Bonferroni method)
  - Modify test statistic (Tukey Honestly Significant Difference Test)
  - Control family-wise error rate (Fisher's Least Significant Difference)
- Community-oriented remedies:
  - Share initial hypotheses and provide preliminary results
  - Collaborate and share data, using methods of meta-analysis to synthesis results
  - Change incentive structures for research to avoid quest for smallest p-value
  - Raise awareness about problem of multiple comparisons

NOTE ON ROLE-PLAYING ELEMENTS

- At the outset of activity, students are asked to imagine they have been hired as data scientists for a company, and are invited to create a name, location and quirk for their company
- These role-playing elements are essential for success of activity:
  - Enhances student buy-in and provides ready-made context
  - Evokes real-world application and structure for statistical investigations
  - Provides incentive structure for task, without offering real (and hence problematic) payouts like extra credit, money, or gifts
  - Allows students to dissociate from the role, which can be important given the duplicity involved in the activity

Thank you!

Live Q&A

Wednesday June 12th, 2:40 pm - 3:35 pm ET

Contact Info:

Jonathan Wells
wellsjon@grinnell.edu