Multivariable Thinking

Roxy Peck
Cal Poly, San Luis Obispo
2016 GAISE College Draft Report

- Six recommendations
- First recommendation with two new added areas of emphasis

1. Teach statistical thinking
   - Teach statistics as an investigative process of problem-solving and decision making.
   - Give students experience with multivariable thinking.

http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf
The Rationale

- From the 2016 Draft Report

“We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.”
But Wait!!!

- Isn’t this beyond what we do in the introductory course? Don’t students need more sophisticated methods to deal with multivariable situations???
- Well, not really. Much can be done with simple graphical methods and descriptive statistics.
- Students can explore multivariable data to come to understand that things are not always as they seem in a univariate or bivariate world. This is an important understanding.
- And it can be fun!
So what should we do?

- Provide students with opportunities to investigate and explore relationships among many variables.
- Recognize that in the intro stat course, can’t always cover this in depth, but the 2016 GAISE report says it is important to expose students to these ideas in the first course.
Where to Start??

The plan for today:
- Scottish Hill Racing Example
- Does Taking Your Time in College Pay Off? (Example from 2016 AP Statistics Exam)
- (if time) Discrimination in Services to the Disabled (Example from the Journal of Statistics Education)
- Confounding Examples
Scottish Hill Racing

- Running off road over upland country where the gradient climbed is a significant component of the difficulty.
- Date back to 10th century.
- Renewed popularity from 19th century forward.
- Mentioned in GAISE appendix as an example of a multi-variable data source.
- Extracted data on the races run between Jan. 1, 2016 and June 22, 2016.
- Variables:
  - Race name
  - Race distance (in km)
  - Race climb (in m)
  - 2016 fastest time for a male runner
  - 2016 fastest time for a female runner
What are some questions you might want to explore?

- How is time related to distance?
- How is time related to climb?
- Which of distance and climb is a better predictor of time?
- Could I predict time better using both distance and climb, or can I get predictions that would be just as good using just distance or just climb?
- What is a typical difference between best finish time for men and for women?
- Does the difference in finish time for men and women depend on the distance?
- Does the difference in finish time for men and women depend on climb?
- Which is a better predictor (distance or climb) of the difference in finish time for men and women?
How is time related to distance?

![Graph showing the relationship between distance and fastest time with data points for different genders.](image)
How is time related to climb?
Which of distance and climb is a better predictor of time?

**Fitted Line Plot**

2016 Fastest Time Male (min) = -8.222 + 5.913 Distance (km)

S 13.2857
R-Sq 86.2%
R-Sq(adj) 85.9%

2016 Fastest Time Male (min) = 6.420 + 0.08026 Climb (m)

S 12.8468
R-Sq 87.1%
R-Sq(adj) 86.8%
Could I predict time better using both distance and climb, or can I get predictions that would be just as good using just distance or just climb?

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.94893</td>
<td>96.32%</td>
<td>96.14%</td>
<td>95.79%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-6.43</td>
<td>2.16</td>
<td>-2.97</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Climb (m)</td>
<td>0.04552</td>
<td>0.00429</td>
<td>10.61</td>
<td>0.000</td>
<td>2.77</td>
</tr>
<tr>
<td>Distance (km)</td>
<td>3.218</td>
<td>0.318</td>
<td>10.13</td>
<td>0.000</td>
<td>2.77</td>
</tr>
</tbody>
</table>

Regression Equation

2016 Fastest Time Male (min) = -6.43 + 0.04552 Climb (m) + 3.218 Distance (km)
What is a typical difference between best finish time for men and for women?
Does the difference in finish time for men and women depend on the distance?
Does the difference in finish time for men and women depend on climb?
Does Taking Your Time in College Pay Off? (Example from 2016 AP Statistics Exam)

A newspaper in Germany reported that the more semesters need to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program. Information was collected on the number of semesters each person in the sample needed to complete the program and the starting salary, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.
a. Does the scatterplot support the newspaper report about number of semesters and starting salary?
Does Taking Your Time in College Pay Off? (Example from 2016 AP Statistics Exam)

- An independent researcher received the data from the newspaper and conducted a new analysis by separating the data into three groups based on the major of each person. A revised scatterplot identifying the major of each person is shown below.

c. Based on the people in the sample, describe the relationship between starting salary and number of semesters for the business majors.
Does Taking Your Time in College Pay Off? (Example from 2016 AP Statistics Exam)
Does Taking Your Time in College Pay Off?  
(Example from 2016 AP Statistics Exam)

e. Based on the analysis conducted by the independent researcher, how could the newspaper report be modified to give a better description of the relationship between the number of semesters and the starting salary for the people in the sample?
Does Taking Your Time in College Pay Off? (Example from 2016 AP Statistics Exam)

- Students were expected to explain that
  - Programs that require more semesters to complete tend to have higher starting salaries.

BUT

- Within a program, students who take a greater number of semesters to complete the program tend to have lower salaries.
Does Taking Your Time in College Pay Off? (Example from 2016 AP Statistics Exam)

- Example of Simpson’s Paradox
- Positive association for overall group and yet negative association within each group.
- Requires multivariable thinking—consider how a third variable can change understanding of the nature of the relationship between two variables.
- Based on real data from Teaching Statistics in School Mathematics—Challenges for Teaching and Teacher Education.
Discrimination in Services to the Disabled
(Example from the Journal of Statistics Education)

From JSE Volume 22, Number 1 (2014)
Simpson’s Paradox: A Data Set and Discrimination Case Study Exercise.

Investigates claim of discrimination in provision of services and support for individuals with developmental disabilities.
Discrimination in Services to the Disabled

- **Data Set**
  - Random sample of 1,000 clients of California Department of Developmental Services

- **Variables**
  - Annual expenditure for support to individual and family
  - Gender
  - Ethnicity
  - Age Cohort (based on amount of financial support typically required during a particular life phase)
    - (0 – 5 years, 6 – 12 years, 13 – 17 years, 18 – 21 years, 22 – 50 years, over 50 years)
Discrimination in Services to the Disabled

- Students explore
  - Discrimination by gender?
  - Discrimination by ethnicity?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Average of Expenditures ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>$18,130</td>
</tr>
<tr>
<td>Male</td>
<td>$18,001</td>
</tr>
<tr>
<td>All Consumers</td>
<td>$18,066</td>
</tr>
</tbody>
</table>

*Table 2. Average Expenditures by Gender*

<table>
<thead>
<tr>
<th>Ethnicity of Consumers</th>
<th>Average of Expenditures ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian</td>
<td>$36,438</td>
</tr>
<tr>
<td>Asian</td>
<td>$18,392</td>
</tr>
<tr>
<td>Black</td>
<td>$20,885</td>
</tr>
<tr>
<td>Hispanic</td>
<td>$11,066</td>
</tr>
<tr>
<td>Multi Race</td>
<td>$4,457</td>
</tr>
<tr>
<td>Native Hawaiian</td>
<td>$42,782</td>
</tr>
<tr>
<td>Other</td>
<td>$3,317</td>
</tr>
<tr>
<td>White non-Hispanic</td>
<td>$24,698</td>
</tr>
<tr>
<td>All Consumers</td>
<td>$18,066</td>
</tr>
</tbody>
</table>

*Table 1. Average Expenditures by Ethnicity*
Sample sizes were very small for most groups, so focus on comparing two groups—Hispanic and White non-Hispanic.

Evidence of discrimination? What else might explain difference in mean expenditure?

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Average of Expenditures ($)</th>
<th>% of Consumers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>$11,066</td>
<td>38%</td>
</tr>
<tr>
<td>White non-Hispanic</td>
<td>$24,698</td>
<td>40%</td>
</tr>
</tbody>
</table>

Table 5.1 Average Expenditures and # of Consumers by Ethnicity
Discrimination in Services to the Disabled

Considering Age Cohort

<table>
<thead>
<tr>
<th>Age Cohort</th>
<th>Hispanic (avg. of expenditures)</th>
<th>White non-Hispanic (avg. of expenditures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 5</td>
<td>$1,393</td>
<td>$1,367</td>
</tr>
<tr>
<td>6-12</td>
<td>$2,312</td>
<td>$2,052</td>
</tr>
<tr>
<td>13-17</td>
<td>$3,955</td>
<td>$3,904</td>
</tr>
<tr>
<td>18-21</td>
<td>$9,960</td>
<td>$10,133</td>
</tr>
<tr>
<td>22-50</td>
<td>$40,924</td>
<td>$40,188</td>
</tr>
<tr>
<td>51 +</td>
<td>$55,585</td>
<td>$52,670</td>
</tr>
<tr>
<td>All Consumers</td>
<td>$11,066</td>
<td>$24,698</td>
</tr>
</tbody>
</table>

Table 6. Average Expenditures by Ethnicity and Age Cohort
Confounding Variables
A Concern in Observational Studies and Poorly Designed Experiments

- There is a strong positive relationship between shoe size and reading ability for children
- There is a strong positive relationship between number of televisions per household and life expectancy for countries in the world
- There is a strong positive relationship between crime rate and ice cream sales
Different from Simpson’s Paradox

- The observed relationship between two variables can sometimes be explained by the fact that both are related to a third variable.
  - Shoe size and reading ability—age
  - Televisions and life expectancy—general standard of living
  - Crime rate and ice cream sales—temperature
- It is important to have students think about possible “third variables”.

AMATYC 2016
In Closing...

- Not suggesting that you need to do all of these activities in the intro course.
- Goal was to provide some examples that are accessible to students in the introductory course and that would be in the spirit of the GAISE recommendation to give students experience with multivariable thinking.
- Give it a try!
Thanks for attending this session.

Comments?
Questions?

Suggestions or questions at a later date:
rpeck@calpoly.edu