Agony and ecstasy: teaching a computationally intensive introductory statistics course using Stata

Nicholas J. Horton

Smith College

August 13, 2007, NASUG 2007

nhorton@email.smith.edu
http://www.math.smith.edu/∼nhorton/agony.pdf

Introduction

- Introduction and motivation
- Modern approaches to teaching intro stats
- Teaching using Stata
- Conclusions and discussion

Background on Smith

- private selective women's liberal arts college in Northampton, Massachusetts
- n=2,800 undergraduate students
- most classes typically small
- focus on opportunities for student research (summer, thesis, special studies)
- 7 intro stats courses offered on campus, 4 within Mathematics and Statistics

Background on Smith

- MTH107: Statistical Thinking (no prereq)
- MTH190: Statistical Methods for Undergraduate Research (pre-calc prereq) [shared with Psychology]
- MTH245: Introduction to Probability and Statistics (calc or discrete prereq)
- MTH241: Probability and Statistics for Engineers (calc III and CS prereq)
- plus Sociology, Economics and Government 100 level courses

Background on Smith

- MTH107: Statistical Thinking (no prereq)
- MTH190: Statistical Methods for Undergraduate Research (pre-calc prereq) [shared with Psychology]
- **MTH245: Introduction to Probability and Statistics (calc or discrete prereq)**
- **MTH241: Probability and Statistics for Engineers (calc III and CS prereq)**
- plus Sociology, Economics and Government 100 level courses

Caveats

- drawing with broad brushstrokes today
- multiple courses allow us to stratify our intro stats offerings
- Smith has relatively small class sizes (20-25 for computer classes; 50-60 for lecture, with 15-20 in a lab)
- lots of people are doing similar things
- relatively novice Stata user (2 years as primary analysis environment)
- not an expert Stata programmer (successfully completed NC151, but still clumsy)
- much of the agony may be fixed in Stata 10

Modern approaches to teaching intro stats

- Statistical education reform
- Importance of multiple regression
- Simulations and activities

Statistical education reform

- Fostered by Cobb (1992) Focus Group on Statistics Education (MAA)
- Addressed shortcomings of traditional statistics education
- Widely adopted tenets
- Much anecdotal evidence of success, some more rigorous benefits shown
- Codified by GAISE (Guidelines for Assessment and Instruction for Statistics Education) project

GAISE College report

1. Emphasize statistical literacy and develop statistical thinking;

GAISE College report

1. Emphasize statistical literacy and develop statistical thinking;
2. Use real data;

GAISE College report

1. Emphasize statistical literacy and develop statistical thinking;
2. Use real data;
3. Stress conceptual understanding rather than mere knowledge of procedures;

GAISE College report

1. Emphasize statistical literacy and develop statistical thinking;
2. Use real data;
3. Stress conceptual understanding rather than mere knowledge of procedures;
4. Foster active learning in the classroom;

GAISE College report

1. Emphasize statistical literacy and develop statistical thinking;
2. Use real data;
3. Stress conceptual understanding rather than mere knowledge of procedures;
4. Foster active learning in the classroom;
5. Use technology for developing conceptual understanding and analyzing data;

GAISE College report

1. Emphasize statistical literacy and develop statistical thinking;
2. Use real data;
3. Stress conceptual understanding rather than mere knowledge of procedures;
4. Foster active learning in the classroom;
5. Use technology for developing conceptual understanding and analyzing data;
6. Use assessments to improve and evaluate student learning

Key role of multiple regression

- Main question: is X associated with Y? What role does Z (possible confounder) play?
- Big conceptual idea to communicate
- Recent study in *NEJM* found that more than half of all original articles used multiple regression (Horton and Switzer, 2005)
- excellent topic to address at length in intro stats (particularly if a terminal course!)
- facilitated by modern texts (e.g. Moore and McCabe's IPS)

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2
- Turn to sampling, probability, and inference

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2
- Turn to sampling, probability, and inference
- Cover only one form of testing (one-sample t?) and confidence intervals (two sample difference in proportions?) rather than traditional laundry list

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2
- Turn to sampling, probability, and inference
- Cover only one form of testing (one-sample t?) and confidence intervals (two sample difference in proportions?) rather than traditional laundry list
- Return to inference for ordinary least squares linear regression

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2
- Turn to sampling, probability, and inference
- Cover only one form of testing (one-sample t?) and confidence intervals (two sample difference in proportions?) rather than traditional laundry list
- Return to inference for ordinary least squares linear regression
- Extend to multiple regression

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2
- Turn to sampling, probability, and inference
- Cover only one form of testing (one-sample t?) and confidence intervals (two sample difference in proportions?) rather than traditional laundry list
- Return to inference for ordinary least squares linear regression
- Extend to multiple regression
- Cover other topics (dramatic pruning needed!) during last 3 weeks

Teaching multiple regression in intro stats

- Review ordinary least squares linear regression (as descriptive technique) in week 2
- Turn to sampling, probability, and inference
- Cover only one form of testing (one-sample t?) and confidence intervals (two sample difference in proportions?) rather than traditional laundry list
- Return to inference for ordinary least squares linear regression
- Extend to multiple regression
- Cover other topics (dramatic pruning needed!) during last 3 weeks
- Students work on projects involving analysis using multiple regression with three variables then present results in poster session

Importance of simulations and activities

- hands on activities help to fix concepts
- computer labs reinforce key ideas
- chunk of class/lab as group work, not lecture

Teaching using Stata

- Analysis
- Lab activities
- Simulations and empirical problem-solving

Analysis and smart calculator (Ecstasy)

- simple to learn
- incredibly powerful
- syntax logical and easy to communicate
- natural interface to Microsoft Word to write up results
- ‘use’ command is incredible
- ‘display’ is a great calculator
- facilitates lookup of critical values to complement tables in book

Addresses GAISE recommendations

- Emphasize statistical literacy and develop statistical thinking;
- Use real data;
- Stress conceptual understanding rather than mere knowledge of procedures;
- Foster active learning in the classroom;
- Use technology for developing conceptual understanding and analyzing data;

Lab activities (Ecstasy)

- provide opportunities for students to analyze real data as part of an extended analysis
- UCLA labs (Gould) adapted to Smith
- 1-2 page writeup (plus graphs) turned in
- approximately 10 due during semester, graded pass/fail
- lab manual provides a useful Stata reference

Addresses GAISE recommendations

- Emphasize statistical literacy and develop statistical thinking;
- Use real data;
- Stress conceptual understanding rather than mere knowledge of procedures;
- Foster active learning in the classroom;
- Use technology for developing conceptual understanding and analyzing data;

Simulations (Agony and ecstasy)

- use of Stata as a toolbox for experimentation
- activities typically requiring only a few dozen lines of code
- facilitate explorations of statistical concepts and experimentation
- 6-8 throughout semester (in addition to 4-5 hands-on activities)
- Two examples: snowstorms and the CLT

Example: Snowstorm insurance

An insurance company sells snowstorm insurance. For each snowstorm that closes business the insurance company pays $10,000. But the coverage does not include the first snowstorm of the year. Assume that the number of snowstorms (X) is a Poisson random variable with rate parameter 1.5 storms per year.

What is the expected return from the policy?

Example: Snowstorm insurance

An insurance company sells snowstorm insurance. For each snowstorm that closes business the insurance company pays $10,000. But the coverage does not include the first snowstorm of the year. Assume that the number of snowstorms (X) is a Poisson random variable with rate parameter 1.5 storms per year.

What is the expected return from the policy?

Exp. return = \$15,000 - \$10,000*$P(X > 0) = \$7,231.02$

Empirical solution in Stata (1/2)

```
clear
set obs 5000
gen x = uniform( )
gen u = x
local lambda = 1.5
local cp=0
local n =0
while ‘cp’<=.99999 {
  local p = exp(-1*‘lambda’)*(‘lambda’^‘n’)/exp(lnfact(‘n’
  local cp= ‘cp’+‘p’
  local pcp = ‘cp’-‘p’
  quietly recode x (‘pcp’/‘cp’ = ‘n’)
  local n = ‘n’ + 1
}
```

Empirical solution in Stata (2/2)

```
generate y=x-1 if x>0
replace y=0 if x==0
gen return = y*10000
```

```
. sum return

  Variable |     Obs      Mean     Std.Dev.    Min      Max
-----------+-------------------------------------------------
    return |    5000      7274     9901.946      0    70000
```

Postmortem

- lack of easy way to generate Poisson rv is painful
- other packages (particularly R) have excellent support for sampling from distributions
- other components of the simulation extremely straightforward
- students can use their empirical solution to check their analytic solution
- repeated runs show sampling variability of the simulation

Example: CLT in Stata (1/3)

```
/* Generate samples from exponential rv
   and capture summary statistics
/* assumes "rnd" package from STB-41 is installed */

clear
local numsim 500
local sampsize 10

set obs ‘numsim’
generate obsmean=.
generate obsmax=.
generate obsmedian=.
```

Example: CLT in Stata (2/3)

```
local i 0
while ‘i’ < ‘numsim’ {
  local i=‘i’+1
  preserve
  clear

  quietly rndexp ‘sampsize’ 2
  collapse (mean) obsmean = xe (max) obsmax = xe (median)
  scalar obsmean=sum(obsmean)
  scalar obsmax=sum(obsmax)
  scalar obsmedian=sum(obsmedian)
```

Example: CLT in Stata (3/3)

```
  restore
  quietly replace obsmean=scalar(obsmean) in ‘i’
  quietly replace obsmax=scalar(obsmax) in ‘i’
  quietly replace obsmedian=scalar(obsmedian) in ‘i’
  _dots ‘i’ 0
}

sum
```

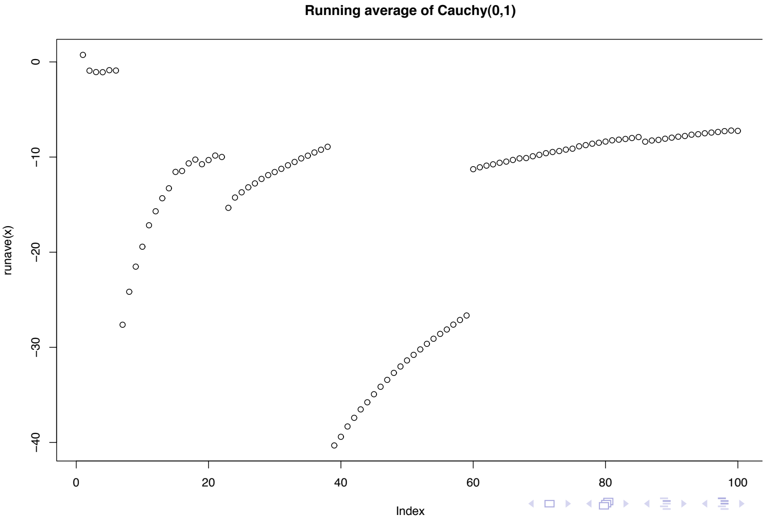
Postmortem

- very clunky
- quietly, scalar, restore, _dots, local, preserve, collapse all confusing
- students get lost in the syntax
- but workable if I provide the base code, and only ask them to tweak it (e.g. different sample sizes, different distributions)

Running average (in R)

```
runave <- function(x) {
        n <- length(x)
        ret <- rep(0,n)
        for (i in 1:n) {
                ret[i] <- mean(x[1:i])
        }
        ret
}
x <- rcauchy(100)
plot(runave(x))
title("Running average of Cauchy(0,1)")
```

Running average



Postmortem

- Stata superb environment for programming by experts to add new bullet-proofed functionality

Postmortem

- Stata superb environment for programming by experts to add new bullet-proofed functionality
- less accessible to the type of ‘sandbox’ work that I describe

Postmortem

- Stata superb environment for programming by experts to add new bullet-proofed functionality
- less accessible to the type of ‘sandbox’ work that I describe
- is it possible to support both?

Postmortem

- Stata superb environment for programming by experts to add new bullet-proofed functionality
- less accessible to the type of 'sandbox' work that I describe
- is it possible to support both?
- nonetheless provides an excellent technology for my intro stat courses

Conclusions

Cobb (1992) noted:

> *of the usual supposed facts about the beginning course, neither its content, nor its organization, nor its mode of delivery is essential for effective learning about statistics. We are actually much freer than we often think to rebuild our curriculum from the ground up.*

Conclusions

David Moore, quoted in Cobb (1992) stated:

> *If I use regression to give students the experience they need and you use time series forecasting, that's fine. What matters most is the experience with practical reasoning about data.*

Conclusions

- students need to analyze data using some sort of technology to effectively learn how to use and interpret statistics
- regression key component to consider in intro stats
- computer exercises and activities can (and should!) be integrated by providing students with working code and instructions for how to modify it (though code is sometimes clunky)
- Stata provides a workable environment for the entire package (as opposed to using applets or multiple systems)

Resources and references

CAUSE http://www.causeweb.org

Cobb, G. *Teaching Statistics*, in Heeding the Call for Change: Suggestions for Curricular Action, ed. L. Steen, MAA Notes No. 22, Washington: Mathematical Association of America, pp. 3-43 (1992).

Cobb, G. http://www.amstat.org/publications/jse/v1n1/cobb.html

First course http://www.amstat.org/publications/jse/v10n2/garfield.html

GAISE http://www.amstat.org/education/gaise

Resources and references

Garfield, J. http://www.education.umn.edu/EdPsych/Projects/Impact.html

NEJM multiple regression Horton and Switzer 2005, 2007, http://www.math.smith.edu/~nhorton/doctor.pdf

R for Math Stats TAS (2003), http://www.math.smith.edu/~nhorton/R

Discussion and questions

Agony and ecstasy: teaching a computationally intensive introductory statistics course using Stata

Nicholas J. Horton

Smith College

August 13, 2007, NASUG 2007

nhorton@email.smith.edu
http://www.math.smith.edu/∼nhorton/agony.pdf