Learning Statistics with R, from the Ground Up

Xiaofei (Susan) Wang

Lecturer in Statistics
Department of Mathematics and Statistics
Amherst College

JSM 2015

Overview

- Briefly discuss R's role in introductory statistics.
- Consider common objectives of the second course and discuss where R fits in.
- Introduce 2 examples that align with those objectives.

Introductory Statistics Core Objectives

GAISE 2012:

> *The desired result of all introductory statistics courses is to produce statistically educated students, which means that students should develop* ***statistical literacy*** *and the ability to* ***think statistically****.*

Introductory Statistics Core Objectives (ct'd)

An interpretation – students should:

- understand basic inferential techniques
- apply basic inferential methods to data
- critically evaluate statistical approaches

R should play a role in all of these goals.

Understanding and Applying

Data: Heights of active men ($n = 247$) from Open Intro (Diez, et al. 2012)

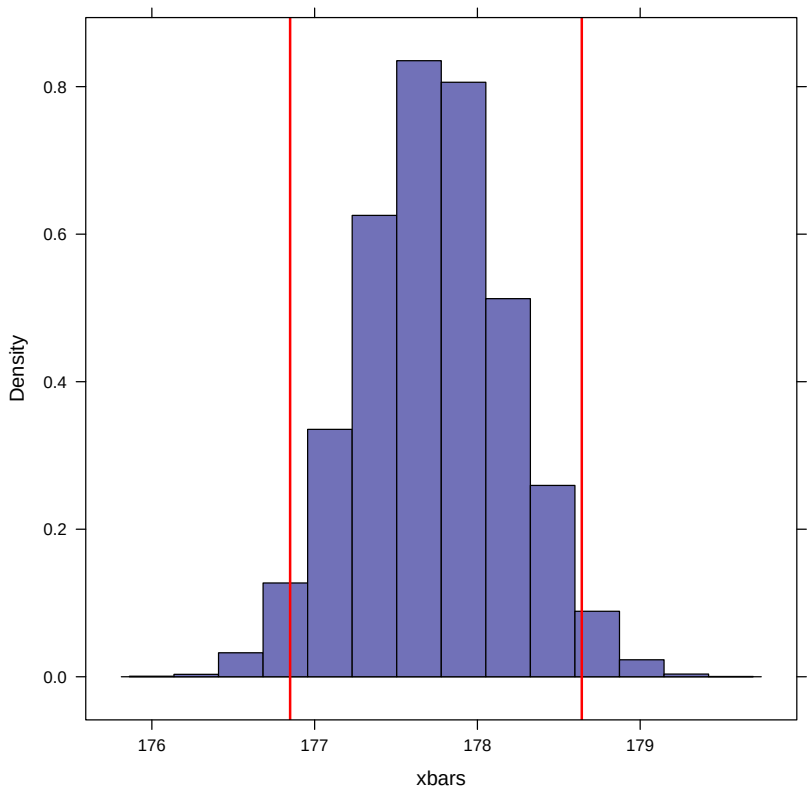
Q: Construct and interpret a 95% confidence interval for the mean height (cm) of active men.

t-interval

```
confint(t.test(~hgt, data = bdims))

## mean of x     lower     upper
##    177.75    176.85    178.65
##     level
##      0.95
```

bootstrap confidence interval



Example of R Script at Intro-Level

Open Intro Lab #7 (Diez, et al. 2012) via `mosaic` package (Pruim, et al. 2015):

```
load(url("http://www.openintro.org/stat/data/mlb11.RData"))
xyplot(runs ~ at_bats, data=mlb11)
cor(runs ~ at_bats, data=mlb11)
m1 <- lm(runs ~ at_bats, data=mlb11)
xyplot(resid(m1) ~ fitted(m1))
```

Example of R Script at Intro-Level

Open Intro Lab #7 (Diez, et al. 2012) via `mosaic` package (Pruim, et al. 2015):

```
load(url("http://www.openintro.org/stat/data/mlb11.RData"))
xyplot(runs ~ at_bats, data=mlb11)
cor(runs ~ at_bats, data=mlb11)
m1 <- lm(runs ~ at_bats, data=mlb11)
xyplot(resid(m1) ~ fitted(m1))
```

- plug n' play data
- simple syntax
- straight-forward application of textbook methods
- basic model assessment

Transitioning to the Second Course

Curriculum Guidelines for Undergraduate Programs in Statistical Science (2014):

> *We concur and recommend that a focus on data be a major component of introductory and advanced statistics courses and that students work with* ***authentic data*** *throughout the curriculum.*

What are authentic data?

Transitioning to the Second Course

Curriculum Guidelines for Undergraduate Programs in Statistical Science (2014):

> *We concur and recommend that a focus on data be a major component of introductory and advanced statistics courses and that students work with* ***authentic data*** *throughout the curriculum.*

What are authentic data?

authentic data

even more authentic data

Transitioning to the Second Course

Nolan & Temple Lang (2010):

> *Successful statisticians must be facile with the computer, for they are expected to be able to access data from various sources, apply the latest statistical methodologies, and communicate their findings to others in novel ways and via new media.*

Possible Second Course Objectives

- Develop data manipulation skills to facilitate working with more authentic data.

Possible Second Course Objectives

- Develop data manipulation skills to facilitate working with more authentic data.
- Garner experience with data problems that are **not** direct applications of textbook methods.

Possible Second Course Objectives

- Develop data manipulation skills to facilitate working with more authentic data.
- Garner experience with data problems that are **not** direct applications of textbook methods.
- Develop flexibility in computing skills for communicating results in variety of ways.

Possible Second Course Objectives

- Develop data manipulation skills to facilitate working with more authentic data.
- Garner experience with data problems that are **not** direct applications of textbook methods.
- Develop flexibility in computing skills for communicating results in variety of ways.

Now, on to some examples of activities that are aligned with these objectives...

Example 1: College Basketball

Adapted from Jay Emerson's data analysis course activity.

FINAL 2006-07 COLLEGE BB SCORES & POINTSPREADS

The Logs contain the following information: KEY: H-Home Game. V-Away Game. N-Neutral Site Game. Number to right of H or V indicates number of overtimes. HN or VN means a game was played at a site not the home court, but favoring the team designated HN. W-Won vs. Pointspread. L-Lost vs. Pointspread. N-No Decision. CT-Conference Tournament. NC-NCAA Tournament. NI-NIT. Date of game is indicated in first column. Example: 11/30 means the game was played on Nov. 30. Middle number in column represents consensus line. '-indicates 1/2-point. P-Pick Game. NL-No Line.

```
AIR FORCE
(SUR: 26-9  PSR: 12-16-1)
11/10  Ark.-Pine Bluff                 81-45   H
11/14  Long Beach St.       L  -8      69-68   N
11/15  Stanford             W  +4'     79-45   V
11/18  Colorado             W  -6'     84-46   V
11/20  Duke                 L  +5'     56-71   N
11/21  Texas Tech           W  -5'     67-53   N
11/22  Radford                         83-59   H
11/29  Wake Forest          W  -11     94-58   HN
12/2   IPFW                            78-66   H
```

Raw

```
##       team1          team2 bookiespread score1
## 1 AIR FORCE Long Beach St.          -8.0     69
## 2 AIR FORCE       Stanford           4.5     79
## 3 AIR FORCE       Colorado          -6.5     84
## 4 AIR FORCE           Duke           5.5     56
## 5 AIR FORCE     Texas Tech          -5.5     67
## 6 AIR FORCE    Wake Forest         -11.0     94
##   score2 gamespread site spresult  date
## 1     68         -1    N        L 11/14
## 2     45        -34    V        W 11/15
## 3     46        -38    V        W 11/18
## 4     71         15    N        L 11/20
## 5     53        -14    N        W 11/21
## 6     58        -36    H        W 11/29
```

Scraped

Example 1: College Basketball

How do we get from the **raw data** to the **scraped data**?

- reading raw text: `scan()` or `readLines()`
- text manipulation: `gsub()`, `substring()`, `grep()`, `strsplit()`
- type conversion: `as.numeric()`, `factor()`, `data.frame()`
- others: `if()`, `for()`, `match()`

Example 1: College Basketball

```
x <- scan("http://www.goldsheet.com/historic/cbblog06.html",
          what="", sep="\n")
# x <- scan("cbblog06.html", what="", sep="\n")

########################################################

# Use R's regular expressions to strip out all the nasty HTML:
y <- gsub("<[^<>]*>", "", x)
y <- gsub("&amp;", "&", y)

# Quick and dirty: limit our attention to lines that matter:
z <- y[36:6955]

# Get the column of spread results:
result <- substring(z, 29, 29)

# Get and process the column of point spreads:
spread <- substring(z, 32, 35)
spread <- gsub("P", "0", spread)
spread <- as.numeric(gsub("'", ".5", spread))

# Get and process the scores:
scores <- substring(z, 36, 44)
```

That's a lot of work!

Example 1: College Basketball

A Motivating Question

Bookies know the game inside and out... or do they? Goldsheet.com tracks the bookies' pointspreads and actual scores in all college basketball games going way back. Using data from the 2006 to 2007 college basketball season, explore the relationship between the pointspreads and the actual outcomes.

Example 1: College Basketball

A Motivating Question

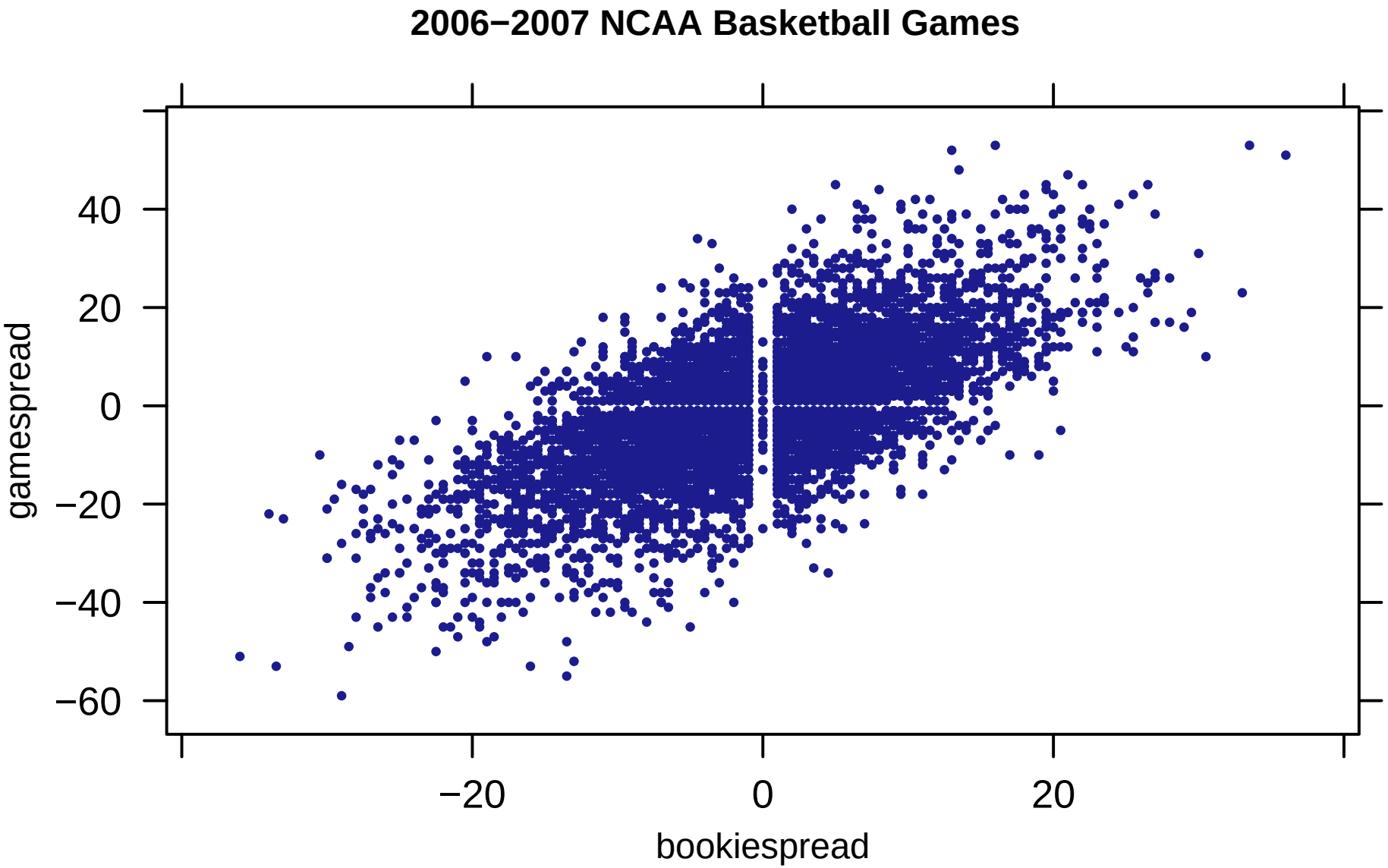
Bookies know the game inside and out... or do they? Goldsheet.com tracks the bookies' pointspreads and actual scores in all college basketball games going way back. Using data from the 2006 to 2007 college basketball season, explore the relationship between the pointspreads and the actual outcomes.

Depending on available time, audience, course focus, adjust the amount of scraping the students need to do.

Example 1: College Basketball

We scraped the data... now what?

Example 1: College Basketball



Example 1: College Basketball

Most games are actually recorded twice!

```
library(dplyr)
```

```
max(bb$bookiespread)
```

```
## [1] 36
```

```
min(bb$bookiespread)
```

```
## [1] -36
```

```
filter(bb, abs(bookiespread) == 36)
```

```
##       team1     team2 bookiespread score1 score2 gamespread site spresult
## 1 DARTMOUTH    Kansas           36     32     83         51    V        L
## 2    KANSAS Dartmouth          -36     83     32        -51    H        W
##    date
## 1 11/28
## 2 11/28
```

Example 1: College Basketball

The ramifications...

- Full dataset (with duplicates):

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.058      0.144    -0.4     0.69
## bookiespread      1.050      0.015    69.2     0.00
```

- Cleaned dataset (duplicates removed):

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.14       0.20    0.72     0.47
## bookiespread       1.07       0.02   52.61     0.00
```

Example 1: College Basketball

The ramifications...

- Full dataset (with duplicates):

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.058      0.144    -0.4     0.69
## bookiespread      1.050      0.015    69.2     0.00
```

- Cleaned dataset (duplicates removed):

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.14       0.20    0.72     0.47
## bookiespread       1.07       0.02   52.61     0.00
```

Coefficients are similar, standard errors off by a factor of $\sqrt{2}$!

Example 2: Olympic Diving

The dataset: Dives and dive scores, by diver and judge, from all events in the 2000 Olympics.

```
##   event round    diver dcountry rank divenum difficulty score
## 1 M3mSB Final XIONG Ni      CHN    1       1        3.1   8.0
## 2 M3mSB Final XIONG Ni      CHN    1       1        3.1   9.0
## 3 M3mSB Final XIONG Ni      CHN    1       1        3.1   8.5
## 4 M3mSB Final XIONG Ni      CHN    1       1        3.1   8.5
## 5 M3mSB Final XIONG Ni      CHN    1       1        3.1   8.5
## 6 M3mSB Final XIONG Ni      CHN    1       1        3.1   8.5
##                      judge jcountry
## 1 RUIZ-PEDREGUERA Rolando       CUB
## 2              GEAR Dennis      NZL
## 3            BOYS Beverley      CAN
## 4            JOHNSON Bente      NOR
## 5          BOUSSARD Michel      FRA
## 6           CALDERON Felix      PUR
```

Example 2: Olympic Diving

Possible tasks:

- Explore whether or not judges exhibit nationalistic bias (Emerson & Meredith, 2010).

Example 2: Olympic Diving

Possible tasks:

- Explore whether or not judges exhibit nationalistic bias (Emerson & Meredith, 2010).
- Create an interactive tool, like Shiny, to display the dataset (and possibly results of some analysis).

Example 2: Olympic Diving

Possible tasks:

- Explore whether or not judges exhibit nationalistic bias (Emerson & Meredith, 2010).
- Create an interactive tool, like Shiny, to display the dataset (and possibly results of some analysis). ✓

Example 2: Olympic Diving

Olympic Diving 2000 | Overall | By Diver | By Judge

Score by Difficulty

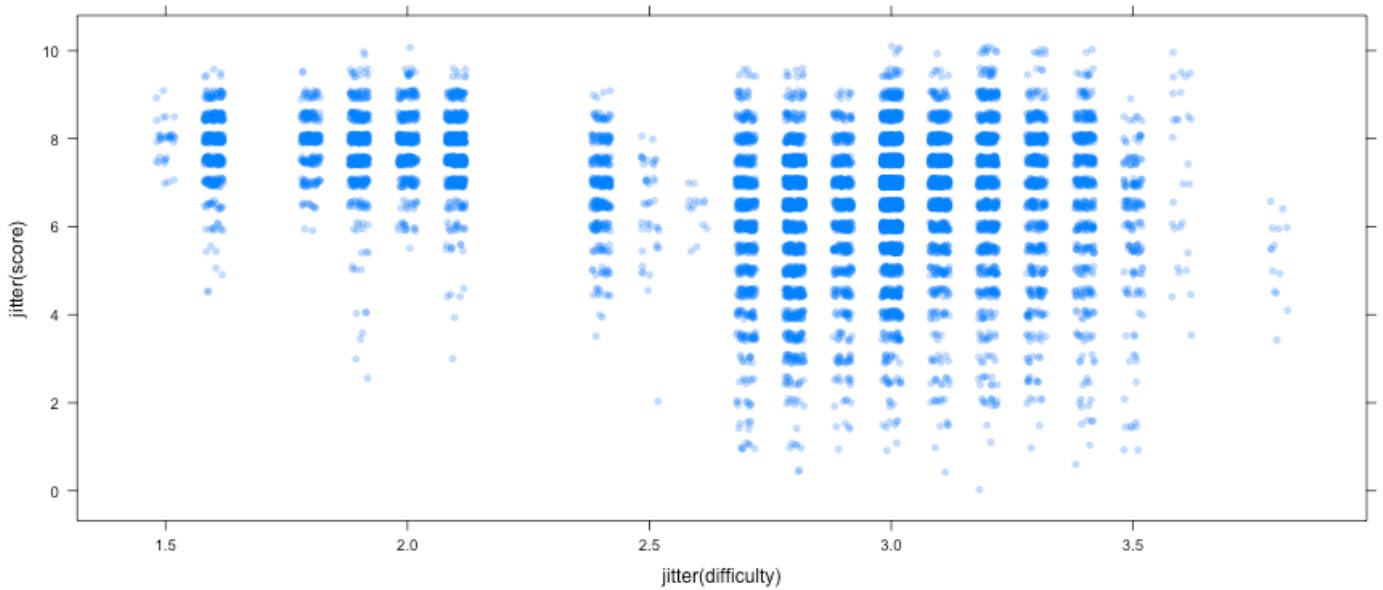
Event:

- M3mSB
- M10mPF
- W3mSB
- W10mPF

Diver Rank:

1 — 49

1 6 11 16 21 26 31 36 41 46 49



Score by Event

Example 2: Olympic Diving

```
```{r, echo=FALSE}
navbarPage("Olympic Diving 2000",
  tabPanel("Overall",
    h3("Score by Difficulty"),
    inputPanel(
      checkboxGroupInput("events", "Event:",
                 events, selected=events),
      sliderInput("rank", "Diver Rank:",
                  min=1, max=49, value=c(1,49))
    ),
    renderPlot({
      xyplot(jitter(score) ~ jitter(difficulty),
             data=filter(data, event %in% input$events,
                               rank >= input$rank[1],
                               rank <= input$rank[2]),
             alpha=0.3, pch=16)
    }),
    h3("Score by Event"),
#     renderPlot({
#
#     }),
```

code template
```

Example 2: Olympic Diving

Provide a template that:

- can be compiled to show a basic, working app
- gives examples of plots (in R code) and some formatted text

Ask students to:

Example 2: Olympic Diving

Provide a template that:

- can be compiled to show a basic, working app
- gives examples of plots (in R code) and some formatted text

Ask students to:

- generate code for the remaining plots
- carry out analyses and integrate their results with the app

Example 2: Olympic Diving

Provide a template that:

- can be compiled to show a basic, working app
- gives examples of plots (in R code) and some formatted text

Ask students to:

- generate code for the remaining plots
- carry out analyses and integrate their results with the app

Students love the appeal of an interactive tool, and the ability to display their statistical prowess on the web!

Example 2: Olympic Diving

Provide a template that:

- can be compiled to show a basic, working app
- gives examples of plots (in R code) and some formatted text

Ask students to:

- generate code for the remaining plots
- carry out analyses and integrate their results with the app

Students love the appeal of an interactive tool, and the ability to display their statistical prowess on the web!

Bonus: Students learn to independently troubleshoot broken code.

Shiny: Student final project examples

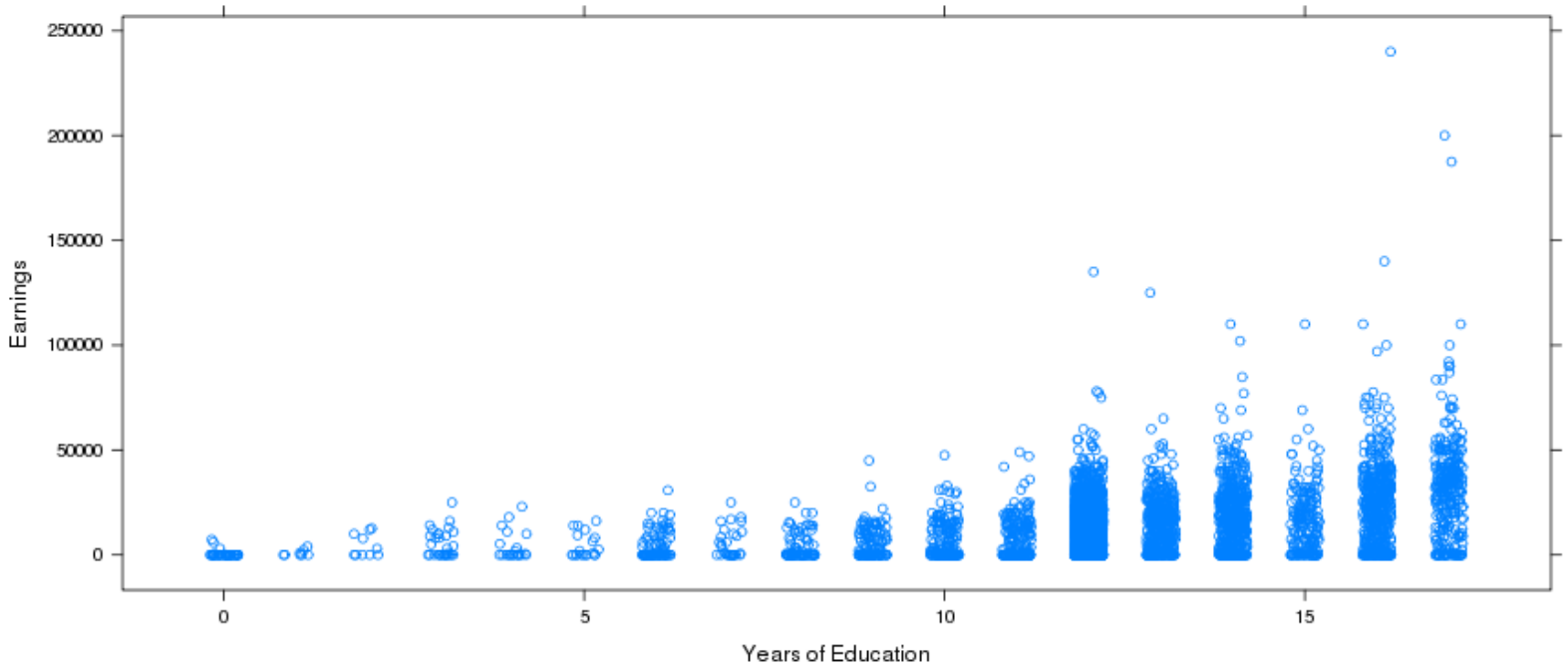
Predicting Economic Mobility Among Low Income Families

by Paul Gramieri, Connor Haley, and Lara Min

Tabs | Data Introduction | Our Analyses ▾ | Low-Income Bracket Indicator Exploration | Marital Status Exploration

Education Analysis

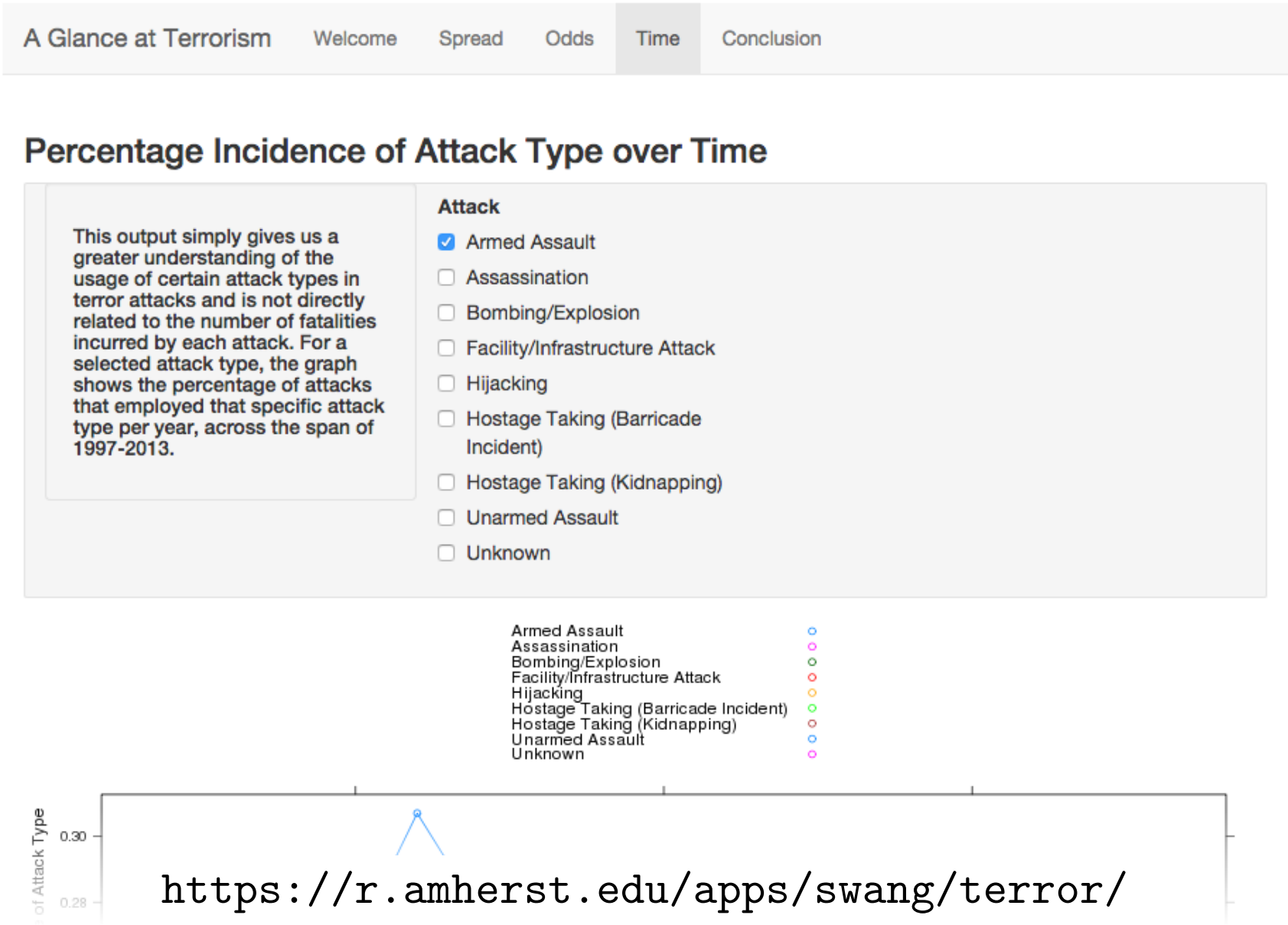
Before diving into this dataset, we predicted that a subject's earnings was positively correlated with his or her years of education.



The above scatterplot of Earnings vs. Years of Education confirms this. However, looking at the differences in mean earnings of education

https://r.amherst.edu/apps/swang/lowincome/

Shiny: Student final project examples



https://r.amherst.edu/apps/swang/terror/

Final Remarks

- Statistics-related jobs in industry require strong computational skills.

Final Remarks

- Statistics-related jobs in industry require strong computational skills.
- R should be used in introductory statistics courses to complement core learning outcomes.

Final Remarks

- Statistics-related jobs in industry require strong computational skills.
- R should be used in introductory statistics courses to complement core learning outcomes.
- Varying amounts of R can be injected into lab activities in second courses and beyond to provide more authentic data experiences.

Final Remarks

- Statistics-related jobs in industry require strong computational skills.
- R should be used in introductory statistics courses to complement core learning outcomes.
- Varying amounts of R can be injected into lab activities in second courses and beyond to provide more authentic data experiences.

Thank you!

References

- Chang, W., et al. (2015). shiny: Web Application Framework for R. R package version 0.12.1. `http://CRAN.R-project.org/package=shiny`
- Curriculum Guidelines for Undergraduate Programs in Statistical Science. (2014). `http://www.amstat.org/education/pdfs/guidelines2014-11-15.pdf`
- Diez, D. M., Barr, C. D., Cetinkaya-Rundel, M. (2012), *OpenIntro Statistics* (2nd ed.).
- Emerson, J. W. and Meredith, S. Nationalistic judging bias in the 2000 Olympic diving competition. *Math Horizons*. (2010). `http://www.stat.yale.edu/~jay/EmersonMaterials/MathHorizons.pdf`
- GAISE College Report. (2012). `http://www.amstat.org/education/gaise`
- Nolan, D., and Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64(2).
- Pruim, R., Kaplan, D., & Horton, N. (2015). mosaic: Project MOSAIC statistics and mathematics teaching utilities. R package version 0.10.0. `http://CRAN.R-project.org/package=mosaic`