Policy Relevant Visualization and Analysis of LDS Data With Open Source Tools

Jared Knowles

Policy Research Advisor
Wisconsin Department of Public Instruction

January 20, 2012 / MIS Conference

Outline

Outline

Outline

Defining Terms

Policy relevant analysis is answering questions that inform policy in a timely fashion and presenting results in an accessible and engaging fashion.

The Big Questions

- States and LEAs have an abundance of data, but how do we extract meaning from it?
- Can we do data analysis fast enough to inform decisions and improve outcomes?
- Can we produce analyses that are approachable to policy makers and the public so that they galvanize change?
- Can we do these things in a time of reduced staffing, decreased budgets, and a shortage of time?
- Can we do analyses under constraints that remain rigorous and valid and provide accurate information?

Example

The state chief school officer asks:

"Do our state bilingual-bicultural programs provide any benefit to our students? Should we focus on ESL more or keep our BLBC programs?"

Options?

Option

- Contract with university faculty
- Consult existing literature
- Call the REL

Caveats

- This will take months. Budget proposal is due in three weeks. Costly.
- No studies in our state. State legislators not impressed.
- Study will take months.

Or Ask Jen

Not Enough Jen's

Outline

How do we do more?

What tools exist to help us turn data into usable information that informs decisions?

Policy Research

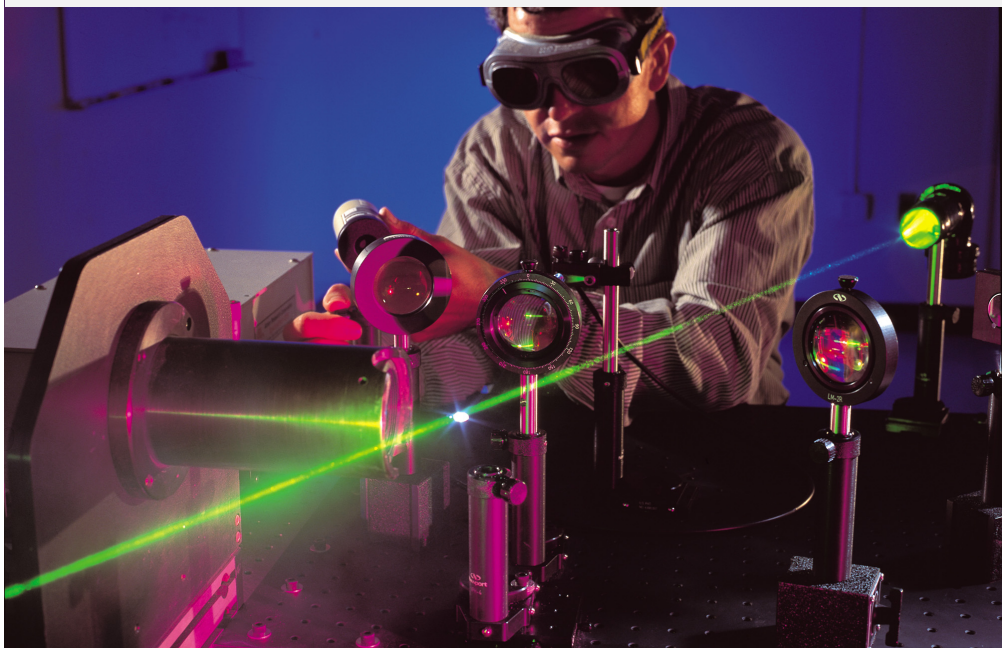
Summary

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

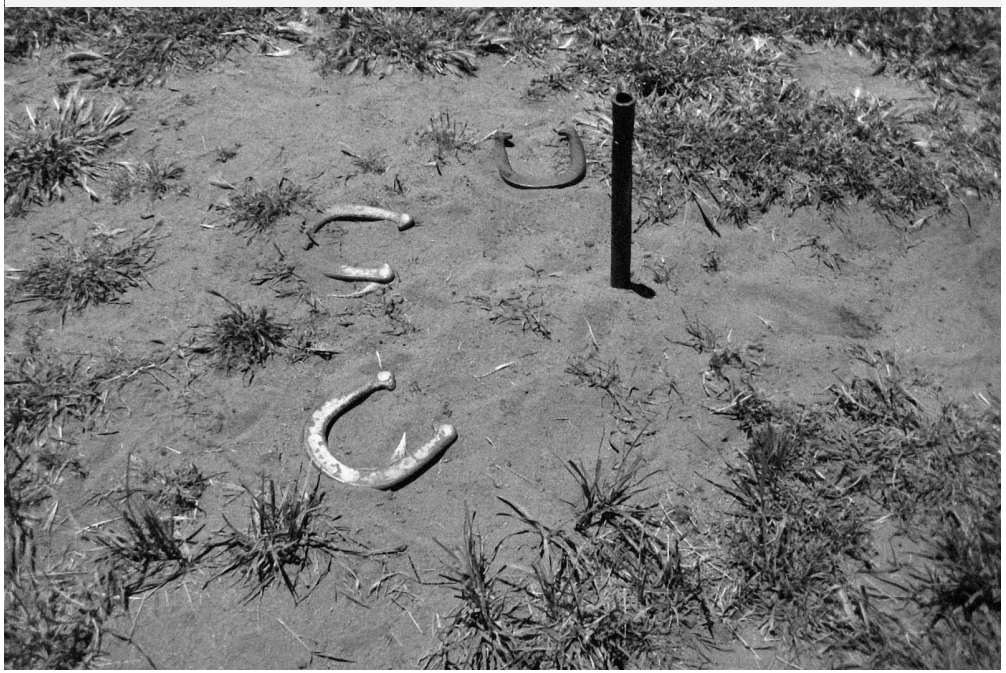
John Tukey

Policy Relevant Research is FAST

Policy Relevant Research is FOCUSED



Policy Relevant Research is APPROXIMATE



Policy Research

Policy research is...

- Fast
- Narrowly focused
- Approximate
- Fast

Traditional research is...

- Long term
- Branches to new ideas and new paths
- Focused on precision
- Peer-reviewed

Policy Research Is Not...

- A visualization and summary statistics
- Descriptive statistics organized with some text
- Ignorant of policy limitations and context

But it also is not:

- Focused on purely causal relationships
- Overly concerned with precision of estimates
- Irrefutable

Examples from One State

In Wisconsin we have done a few analyses that have helped us make decisions.

- An analysis of the effectiveness of bilingual-bicultural programs
- An analysis of reading performance and markers of struggling literacy
- An analysis of concentration in dropouts
- An anaysis of feeder patterns for teacher supply from teacher prep programs

Outline

Outline

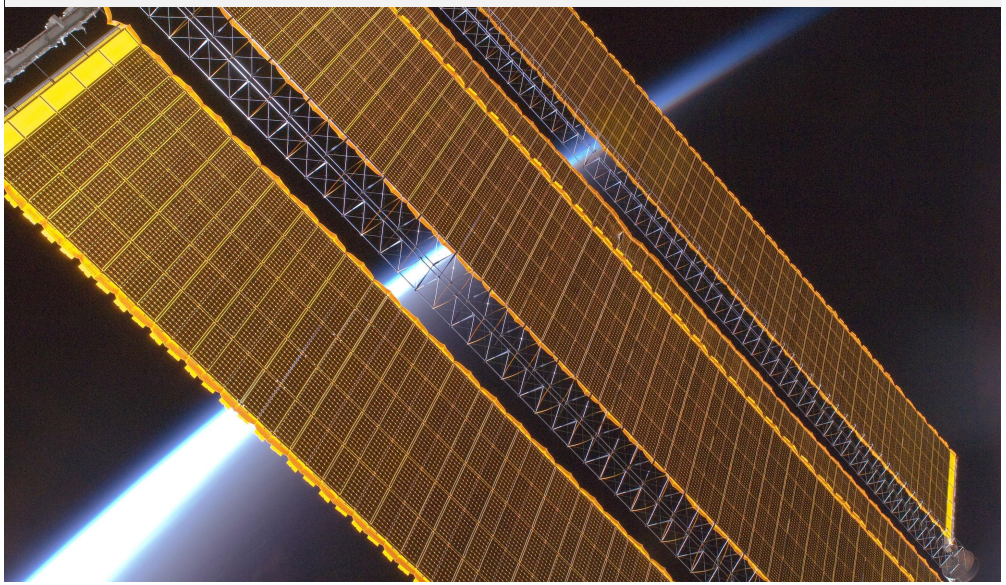
Gathering More Data

- States and districts collect hundreds of attributes about millions of students
- Data is collected before children reach school age and after they have moved to a college or a career
- Patterns may inform how choices in policy will affect the population
- Lessons learned can help us build simulations to weigh policy outcomes before making decisions–decision support

Data is like ore

Analysis concentrates its value

And it can be used to produce something



How?

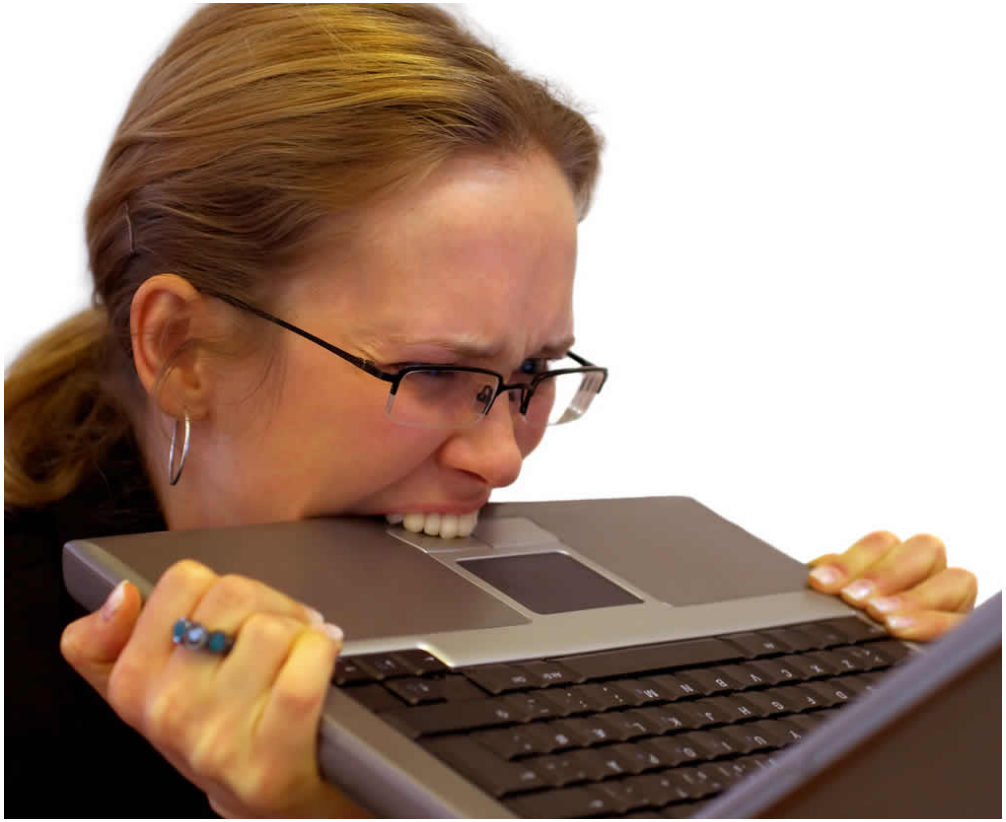
Outline

Institutional Frustrations

We just need to get our jobs done. We need to do them:

- efficiently
- transparently
- and in a reproducible manner

This is currently **costly** in time, money, and management resources.



Institutional Frustrations

A number of barriers make using data difficult and supporting decisions on short timelines hard:

- Acquiring proprietary tools from vendors takes agreements, legal documents, and lag time
- Sharing data with external researchers requires legal agreements, levels of management approval, and planning time to specify narrow scope

Institutional Frustrations

Analyses that get done are not part of cohesive strategy, but done ad hoc.

- Done using proprietary tool sets
- Poorly documented
- Unable to be reproduced with updated data later

Knowledge is lost, not accumulated.

Analyses Don't Get Used

Often when we do an in house analysis it does not get used to support decisions.

- In house analysis often relies on the expertise of one or two staff who are obligated elsewhere.
- Analysis are done using ad-hoc tools distributed among expertise of individual staff with no comprehensive standard.
- The information we have is dependent on individual staff and the analysis projects they undertake and their tenure supporting these efforts.
- Staff turnover threatens to change the information available to make decisions as knowledge leaves the agency, breaking continuity with previous reports

Incoherence

Or we do the wrong analysis

Sometimes analyses are done at the whim of an analyst or two and not tied to the needs of decision makers or stakeholders

Irrelevant

Outline

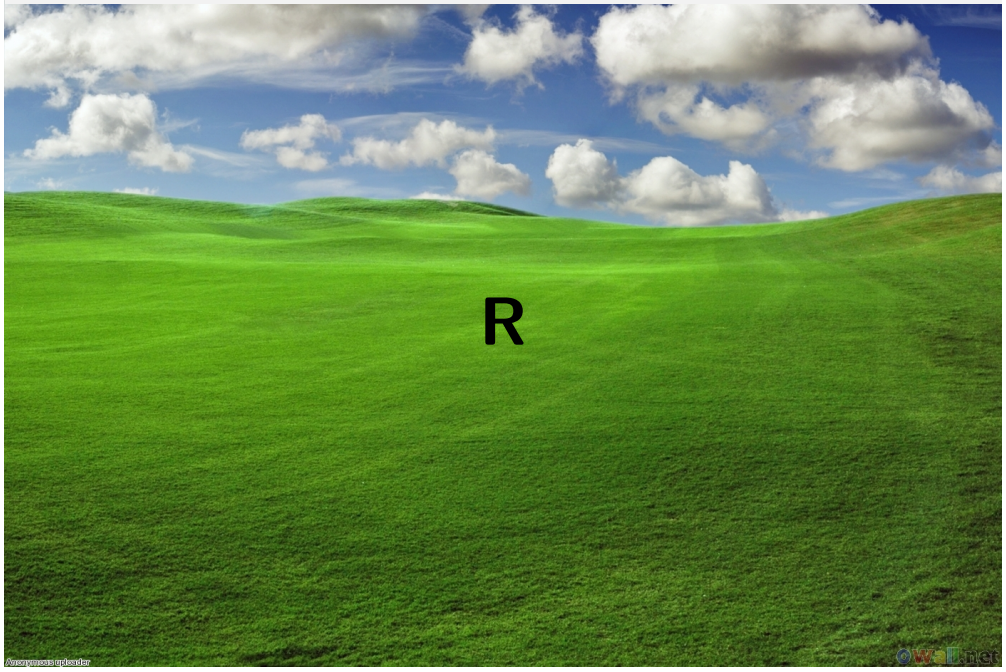
Example

The state chief school officer asks:
"Do our state bilingual-bicultural programs provide any benefit to our students? Should we focus on ESL more or keep our BLBC programs?"

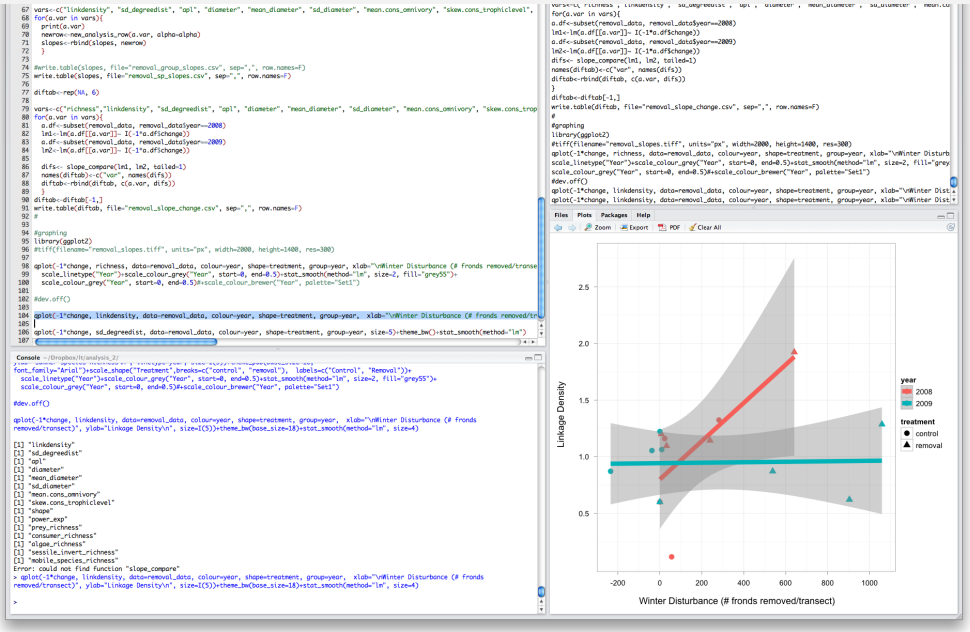
Instead we answer:

- Our ELL students are doing better than last year.
- National research is inconclusive on these programs.
- Our data shows participation is up in BLBC programs.
- We found a researcher to help with this in six months.

Open Source Tools



R



R As Part of the Solution

Objections to Data Analysis

- Costly
- Slow and Time Consuming
- Technical and complex
- Opaque and not actionable

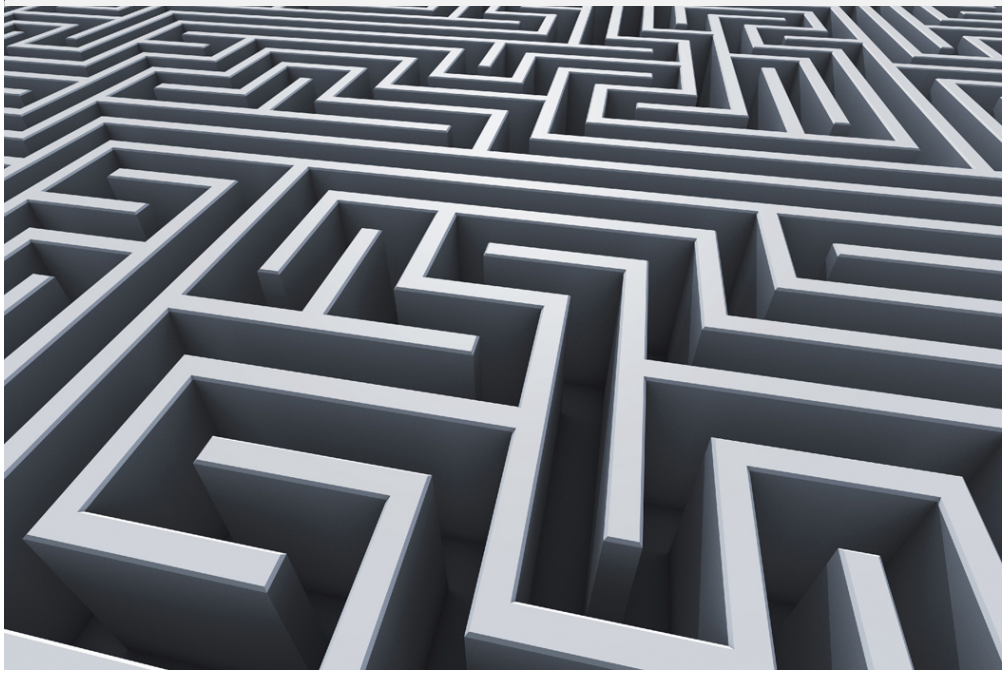
The R Solution

- R is free and open source
- R allows reproducible and sharable analysis across researchers
- R can be scripted to do common tasks
- R is a lingua franca that standardizes common tasks

Caveats

But wait...? Isn't R?

Confusing?



Full of Bugs?

Inefficient?

The Truth

- The short answer is no.
- R has a high startup cost, but we are working to bring that down.
- R has some quirks, but all software does.
- And, R can be amazingly more efficient through collaboration and sharing of code and tools.

Outline

Outline

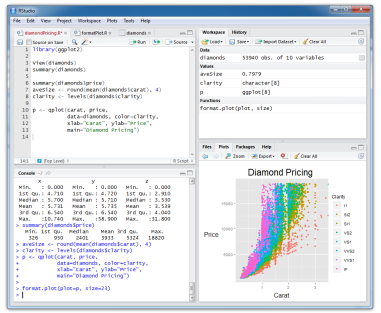
- R is an Open Source (and freely available) environment for statistical computing and graphics
- Available for Windows, Mac OS X, and Linux
- R is being actively developed with two major releases per year
- R can be extended with 'packages' that contain data, code, and documentation to add new functionality
  - 'SGP package' to calculate Student Growth Percentiles
  - 'googleVis' package to build Google Visualizations easily out of data
  - 'ggplot2' package to build attractive custom data visualizations

More Support for R

- R is a common tool among data experts at major universities
- No need to go through procurement, R can be installed in any environment on any machine and used with no licensing or agreements needed
- R source code is very readable to increase transparency of analyses
- R code is easily borrowed from and shared with others
- R is incredibly flexible and can be adapted to specific local needs
- R is under active development, improving greatly, and supported strongly by both professional and academic developers
- R is top of the line best in class statistics software maintained by professional statisticians and supported by an active user community

Using R

- R can be used with an excellent Integrated Development Environment
- RStudio makes many of the basic tasks in R much easier like
  - Importing data
  - Previewing plots
  - Version control
  - Collaboration
- Greatly increases ease of use



Pros and Cons of R

Pros of **R**

- Open source and freely available on all platforms
- Scripting for reproducible and transparent analyses
- Extensible to fit skills, needs, and cutting edge techniques
- Excellent graphical and output capabilities

Cons

- Steep learning curve and command line interface
- Requires specific inputs to get desired results
- Unforgiving of misspecification of inputs
- Data input can be tricky at first

R in the SEA

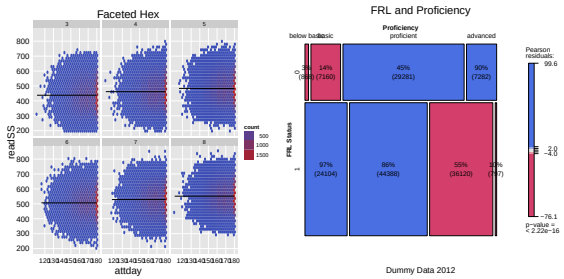
- By standardizing common data tasks staff are freed up to do other tasks
- Wisconsin is using R to calculate AYP directly from the LDS
- One script, one run, all reports and error checks run–saving weeks of work
- Easy to understand how calculation is done protecting against discontinuity if staff turnsover
- Other reports and data analyses are being standardized–quality checks on LDS data, etc.
- More time to do policy research, more transparent data analyses that can be reproduced, accumulation of a knowledge base

Outline

Visualization

One of the major strengths of R is its ability to create informative and compelling visualizations of data.

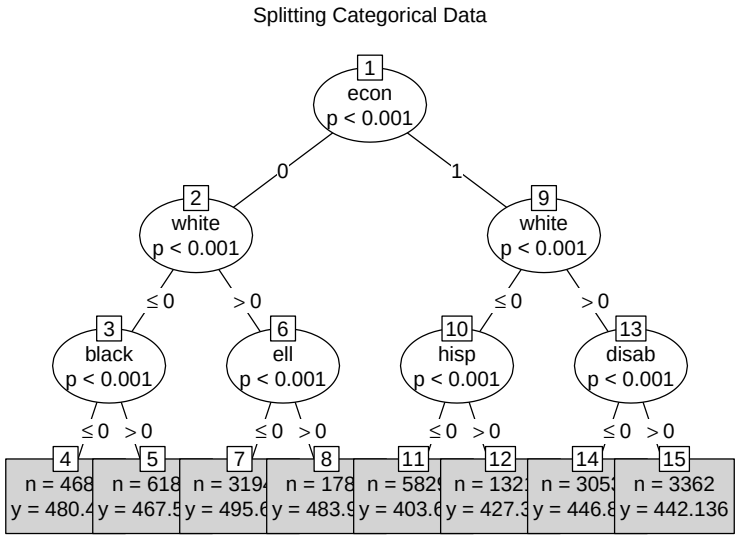
Examples of R Figures



Dummy Data 2012

Outline

Inference Trees

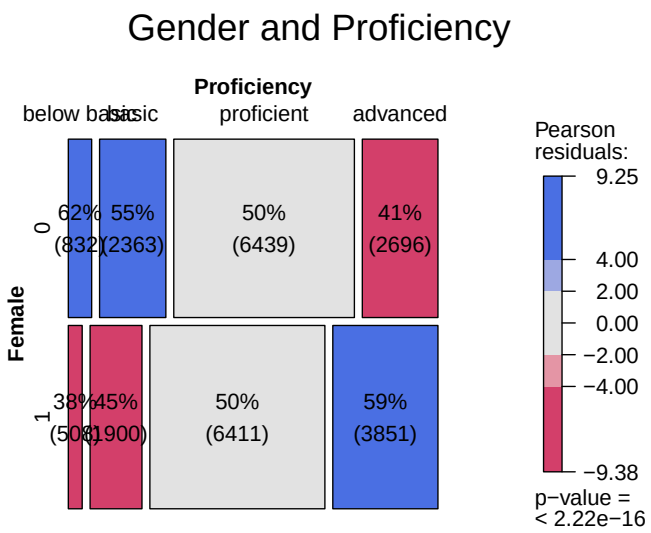


Easy Code for Plot

Code to make this plot:

```
> z1<-ctree(readSS~black+hisp+asian+indian+white+ell+disab+econ
+           +attday,data=subset(student_long,year=='2000',grade=
+           controls=bonf)
> plot(z1,type='simple',main="Splitting Categorical Data")
```

Visual Crosstabs



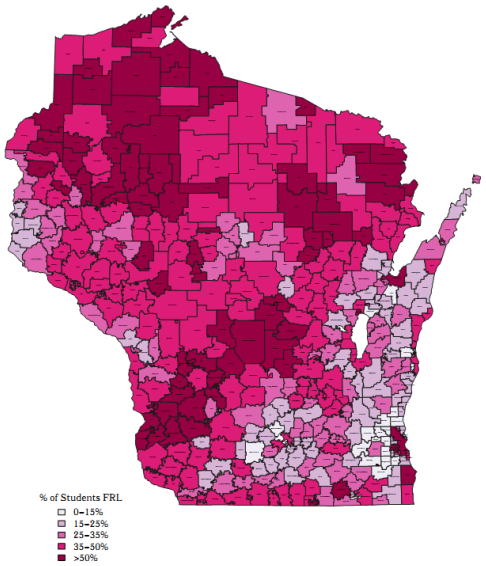
R Code

Code for this plot that uses 'mosaictabs' function from the LDS_TOOLS package

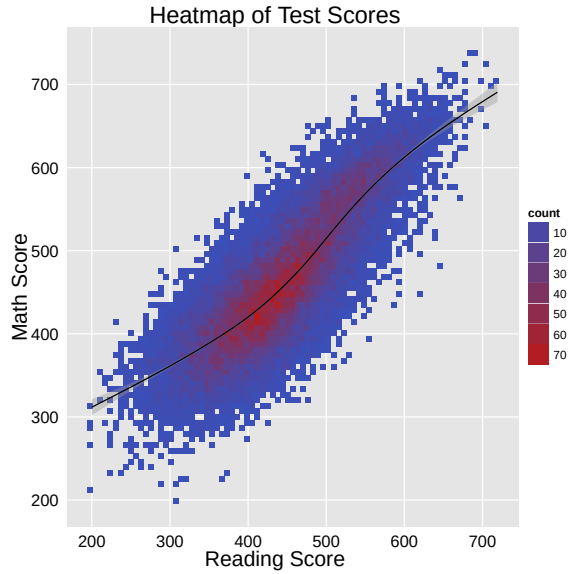
```
> plotsub<-subset(student_long,year=='2001' & grade==6)
> varnames<-c('Female','Proficiency')
> mosaictabs.label(plotsub,plotsub$female,plotsub$proflvl,
+                  varnames,'Gender and Proficiency','Dummy Data
```


Dummy Data 2012

Maps

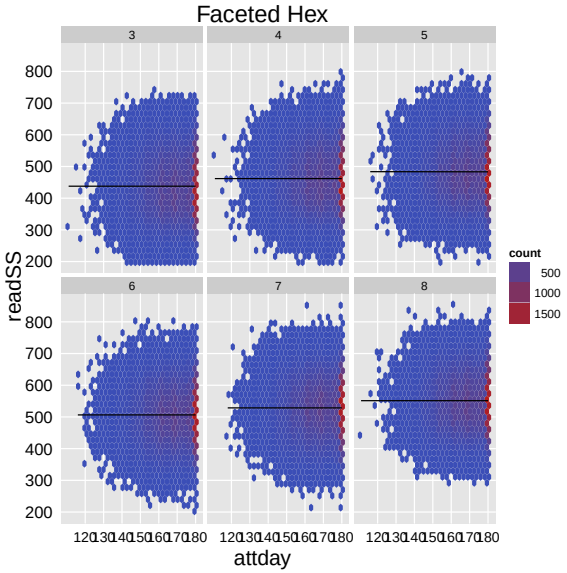


Heatmap

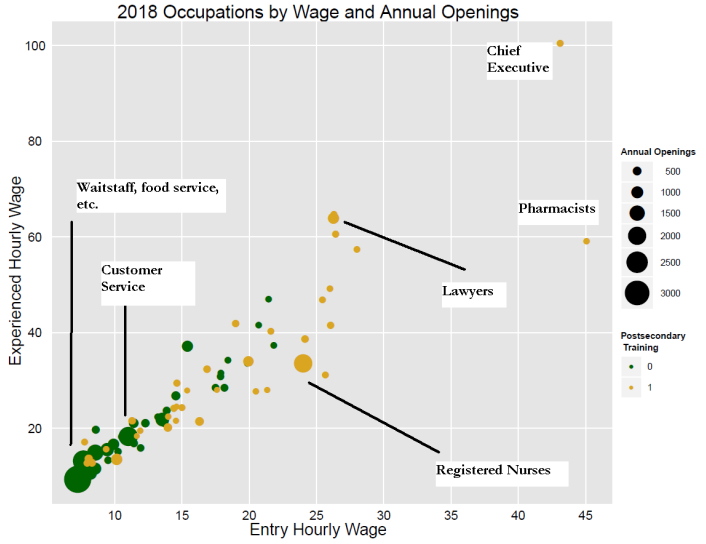


Conditional Density

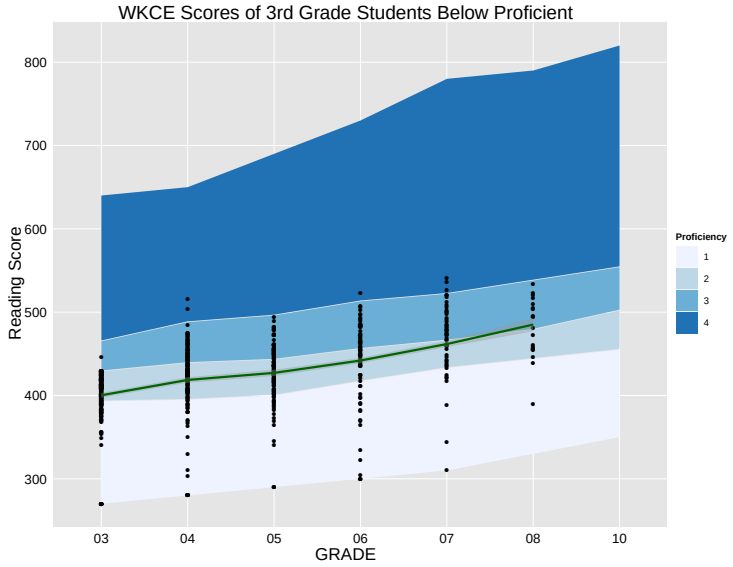
Faceted Hex



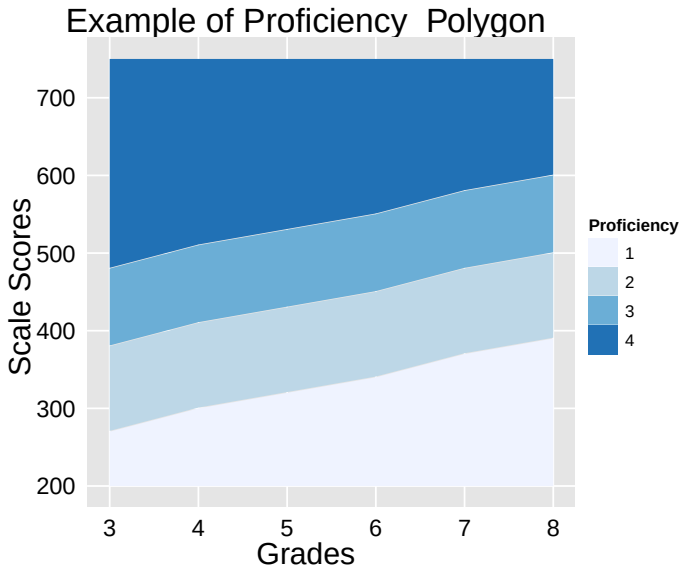
Polished Scatter



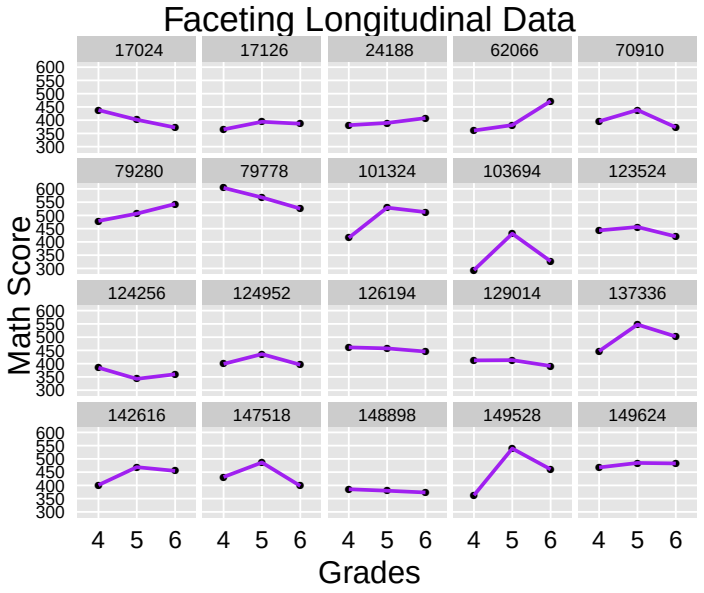
Longitudinal Data



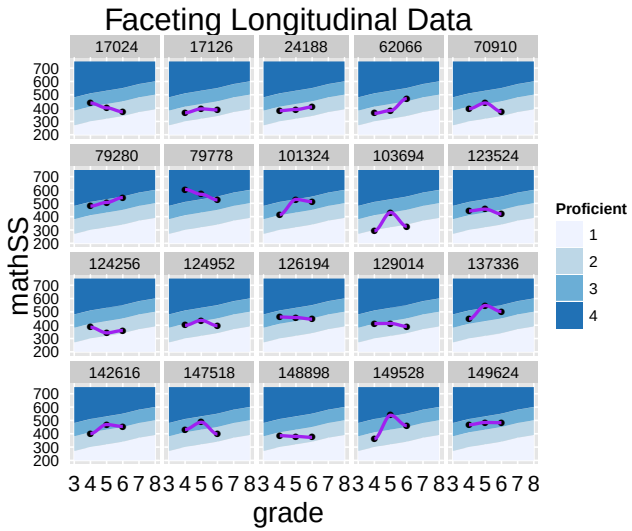
Proficiency Polygon



Individual Growth Trajectories

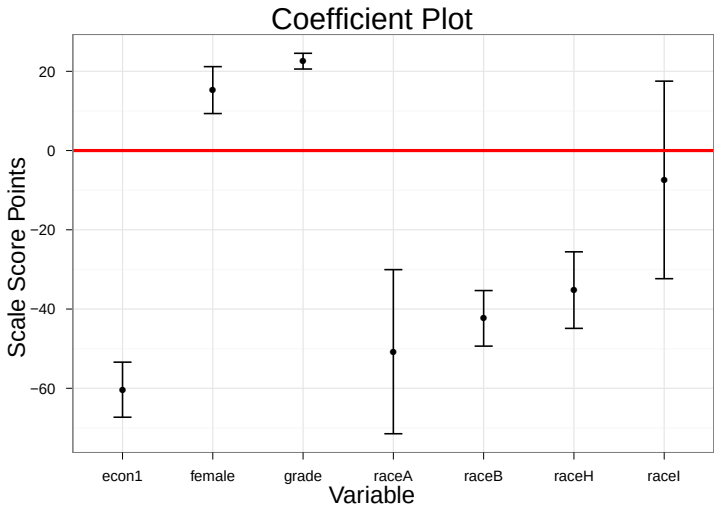


Individual Growth Trajectories

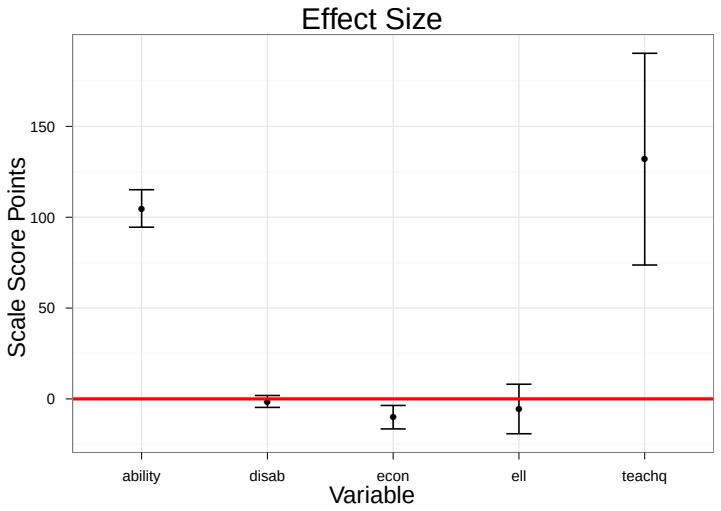



```
null device
          1
```

Communicate Statistical Models



Communicate Statistical Models II



Outline

The Command Line

- R can be tricky because it uses command lines.
- This is powerful, but requires a learning curve.
- Some simple calculations can give a feel for how R works

```
> 2+2
```

```
[1] 4
```

```
> 7*4
```

```
[1] 28
```

```
> exp(3)
```

```
[1] 20.08554
```

```
> pi
```

```
[1] 3.141593
```

Deconstruct R Commands

```
> summary(student_long[,28:30])
```

```
     readSS          mathSS
 Min.   :200.0   Min.   :200.0
 1st Qu.:430.9   1st Qu.:420.8
 Median :494.4   Median :481.2
 Mean   :494.9   Mean   :483.7
 3rd Qu.:558.4   3rd Qu.:543.8
 Max.   :850.4   Max.   :857.5
        proflvl
 below basic: 35927
 basic      : 85983
 proficient :198393
 advanced   :129697
```

- **summary** is the function
- **student_ long** is the data object

Crosstabs

Let's test for balance among some categories of students

```
> with(subset(student_long,year=='2001'
+             & grade==3),table(female,race))
```

```
      race
female    W    B    H    I    A
     0 5570 5081 1365  107  224
     1 5653 5252 1394  116  238
```

```
> #As proportions
> with(subset(student_long,year=='2001'
+             & grade==3),round(prop.table
+             (table(female,race))*100),4)
```

```
      race
female  W  B  H  I  A
     0 22 20  5  0  1
     1 23 21  6  0  1
```


Crosstabs

We can even output the results of R commands into a print-ready format. As we have below.

	W	B	H	I	A
0	22.00	20.00	5.00	0.00	1.00
1	23.00	21.00	6.00	0.00	1.00

Outline

Doing More than the Basics

- R can routinize basic functions like tables, crosstabs, and visualization of data
- R can also be extended to do more advanced analyses like multilevel modeling, spatial error modeling, Bayesian data analysis, forecasting, and simulation
- R can do advanced graphical functions as well
- R can even be expanded to incorporate additional programming languages like Python, C++, and Java

The downside of this is that these functions can have a steep learning curve.

Outline

Outline

ANOVA

We can also do statistical tests using both Bayesian and Frequentist methods.

```
> nova1<-aov(readSS~female*race+econ,data=novaset)
> summary(nova1)
```

```
               Df   Sum Sq  Mean Sq  F value
female          1  2176464  2176464  763.682
race            4 33929701  8482425 2976.331
econ            1 20785160 20785160 7293.140
female:race     4    50009    12502    4.387
Residuals   24989 71217664     2850
               Pr(>F)
female       < 2e-16 ***
race         < 2e-16 ***
econ         < 2e-16 ***
female:race 0.00152 **
Residuals
---
```

Pretty Output

We can also do print-ready model outputs with R's extensible formatting

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female	1	2176463.53	2176463.53	763.68	0.0000
race	4	33929701.29	8482425.32	2976.33	0.0000
econ	1	20785160.43	20785160.43	7293.14	0.0000
female:race	4	50008.56	12502.14	4.39	0.0015
Residuals	24989	71217663.75	2849.96		

Outline

A simple OLS Model I

```
> mod1<-lm(readSS~female*race*econ+grade*year,data=student_long)
> summary(mod1)
```

```
Call:
lm(formula = readSS ~ female * race * econ + grade * year, data = s

Residuals:
     Min       1Q   Median       3Q      Max
-284.566  -34.219    0.104   34.321  246.441

Coefficients:
                 Estimate Std. Error  t value
(Intercept)      383.11107    0.49828  768.865
female            14.24756    0.32404   43.969
raceB            -37.21069    0.57997  -64.160
raceH            -41.01409    0.78264  -52.405
```


A simple OLS Model II

```
raceI                9.48529    2.15941    1.151
raceA                0.45644    1.38578    0.329
econ1              -54.71664    0.32522 -168.246
grade               20.81004    0.07791  267.111
year2001            -3.48643    0.63452   -5.495
year2002           134.55881    0.63452  212.065
female:raceB        -0.88573    0.82287   -1.076
female:raceH         4.46915    1.09121    4.096
female:raceI         4.72824    2.88579    1.638
female:raceA        -0.04431    1.93005   -0.023
female:econ1        -3.22112    0.45766   -7.038
raceB:econ1        -11.97227    0.64901  -18.447
raceH:econ1          6.46426    0.89487    7.224
raceI:econ1        -25.56913    2.62396   -9.744
raceA:econ1        -23.59874    1.72464  -13.683
grade:year2001       9.50620    0.11018   86.281
grade:year2002      -3.95024    0.11018  -35.853
```

A simple OLS Model III

```
female:raceB:econ1  -0.29569    0.91935   -0.322
female:raceH:econ1  -2.03755    1.25022   -1.630
female:raceI:econ1   1.70605    3.57366    0.477
female:raceA:econ1  -2.64903    2.39991   -1.104
                   Pr(>|t|)
(Intercept)         < 2e-16 ***
female              < 2e-16 ***
raceB               < 2e-16 ***
raceH               < 2e-16 ***
raceI                 0.250
raceA                 0.742
econ1               < 2e-16 ***
grade               < 2e-16 ***
year2001           3.92e-08 ***
year2002            < 2e-16 ***
female:raceB          0.282
female:raceH       4.21e-05 ***
```


A simple OLS Model IV

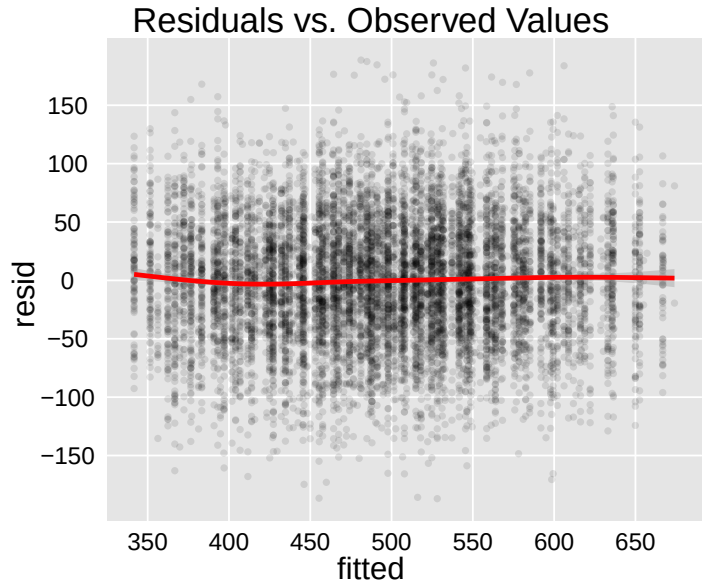
```
female:raceI          0.101
female:raceA          0.982
female:econ1       1.95e-12 ***
raceB:econ1         < 2e-16 ***
raceH:econ1        5.07e-13 ***
raceI:econ1         < 2e-16 ***
raceA:econ1         < 2e-16 ***
grade:year2001      < 2e-16 ***
grade:year2002      < 2e-16 ***
female:raceB:econ1    0.748
female:raceH:econ1    0.103
female:raceI:econ1    0.633
female:raceA:econ1    0.270
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.53 on 449975 degrees of freedom
```

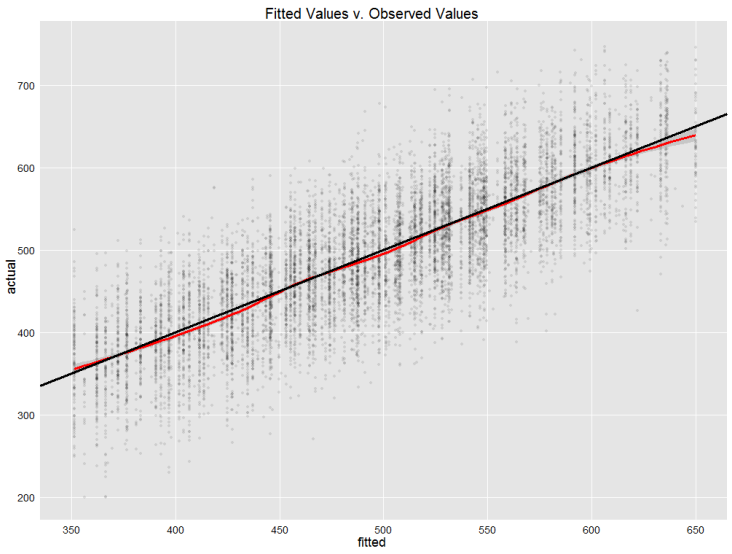

A simple OLS Model V

```
Multiple R-squared: 0.6742,  Adjusted R-squared: 0.6742
F-statistic: 3.88e+04 on 24 and 449975 DF,  p-value: < 2.2e-16
```


Model Evaluation

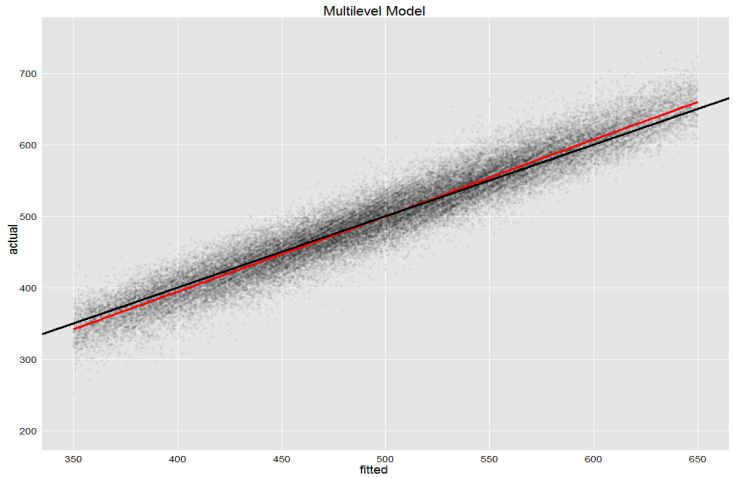


Model Evaluation Part 2



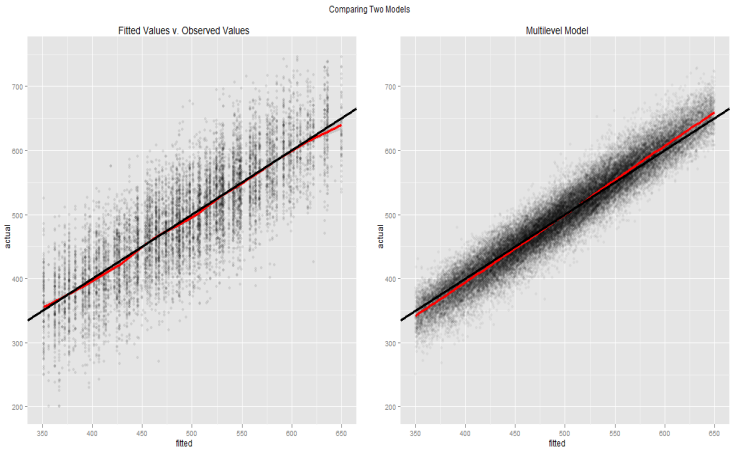
Better Fitting

Using advanced techniques we can greatly improve our model fit over the OLS model.



Compare OLS and Mixed Effects

A simple mixed-effects model estimated with an R package can outperform the simple OLS without much additional effort



Outline

Some stuff

The advantage of R comes when we specialize to develop advanced tools that are compatible across our very similar datasets.

Collaborative Development

Outline

Leverage Similar Data

Why can we do this?

- We all share similar data with similar attributes and similar reporting needs
- Standardizing on our analysis language makes applying analysis from one place to another easy due to data similarities
- **Build it. Share it. Use it.**

Outline

GitHub

GitHub provides an excllent way to do this.

LDS_TOOLS

https://github.com/jknowles/LDS_TOOLS

- Uses the **'git'** version control system to track collaborative coding on the same source document
- Free and open source coding environment that plays well with RStudio
- No need to contribute, provides easy way to access work of others

Open Source Analysis Code

- LDS_TOOLS is a fledgling effort to open source many of the graphics and analyses from earlier in this presentation
- Make R code and LaTeX code available to be applied to other SEA and LEA data
- Packaged with a dummy dataset representing common educational data attributes for testing and sharing
- Share visualization techniques, statistical models, and even full reports
- Plug and play–change the variables to match your data and produce the same visualization, report, analysis

LDS_TOOLS Available Now

You can visit online at: https://github.com/jknowles/LDS_TOOLS

Available Tools:

- Improved mosaic plots
- Gantt charts for project planning
- Proficiency polygons for assessments
- Convenience functions for education data

Planned Tools:

- Easy to make maps
- Data mining and statistical modeling routines
- Pre-built data quality reports
- Summary reports for NSC, Assessment, and Enrollment data

Who can get involved?

Current staff can use these tools, and even contribute to them

- Analysts with a couple of days to learn basic R skills
- Anyone with programming skills
- University partners
- The R Community

Outline

Questions?

Questions?

Contact Information

For more information:

Jared Knowles
Wisconsin Department of Public Instruction
jared.knowles@dpi.wi.gov
https://github.com/jknowles/LDS_TOOLS

Example

This entire presentation was created with R, LaTeX and is available online at GitHub to be edited, modified, and reused.
https://github.com/jknowles/mis-presentation

Tools

- **R** (http://cran.r-project.org/)
  - An open source statistics package that is freely available for all platforms.
- **RStudio** (http://www.rstudio.org/)
  - An enhanced front-end for R. An Integrated Development Environment (IDE) for statistical programming.
- **Quantum GIS** (http://www.qgis.org/)
  - A GIS package that provides most of the functionality of ArcGIS but is freely available.
- **GeoDa** (http://geodacenter.asu.edu/)
  - A geo-spatial statistics package for analyzing clustering and spatial correlation of datasets.
- **LaTeX** (http://www.latex-project.org/)
  - A typesetting and document building tool that integrates with R.

Tutorials

- R Reference (http://www.statmethods.net/)
- First R Commands to Learn (https://github.com/hadley/devtools/wiki/vocabulary)
- Beginning with LaTeX (http://en.wikibooks.org/wiki/LaTeX)
- Quantum GIS Guide (http://qgis.org/en/documentation/manuals.html)
- R Graph Gallery (http://addictedtor.free.fr/graphiques/)