

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

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Outline

1 What is Policy Relevant Analysis?

- Defining Terms

- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?

- Barriers to Data Analysis

- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting Started

4 More Advanced Functions

- Analysis

- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses

- Coordinating and Social Coding

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?

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

Options?

Option

- Contract with university faculty
- Consult existing literature

Options?

Option

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

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.

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.

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.

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



Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Not Enough Jen's



How do we do more?

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

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

Policy relevant research is...

- Fast

Policy Relevant Research

Policy relevant research is...

- Fast
- Narrowly focused

Policy Relevant Research

Policy relevant research is...

- Fast
- Narrowly focused
- Approximate

Policy Relevant Research

Policy relevant research is...

- Fast
- Narrowly focused
- Approximate
- Fast

Traditional research is...

- Long term

Policy Relevant Research

Policy relevant research is...

- Fast
- Narrowly focused
- Approximate
- Fast

Traditional research is...

- Long term
- Branches to new ideas and new paths

Policy Relevant Research

Policy relevant research is...

- Fast
- Narrowly focused
- Approximate
- Fast

Traditional research is...

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

Policy Relevant Research

Policy relevant 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 Relevant 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

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

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

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
- Data visualization to demonstrate problem scope for grants and press materials

How?

Policy Relevant Analysis is FAST



Policy Relevant Analysis is FOCUSED



Policy Relevant Analysis is APPROXIMATE



Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

• What can R do?

- R Examples
- Getting Started

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

• Why is Data Analysis So Important?

- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

• What can R do?

- R Examples
- Getting Started

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

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 can tell us how choices in policy will actually affect the population
- Lessons learned can help us build simulations to weigh policy outcomes before making decisions—decision support

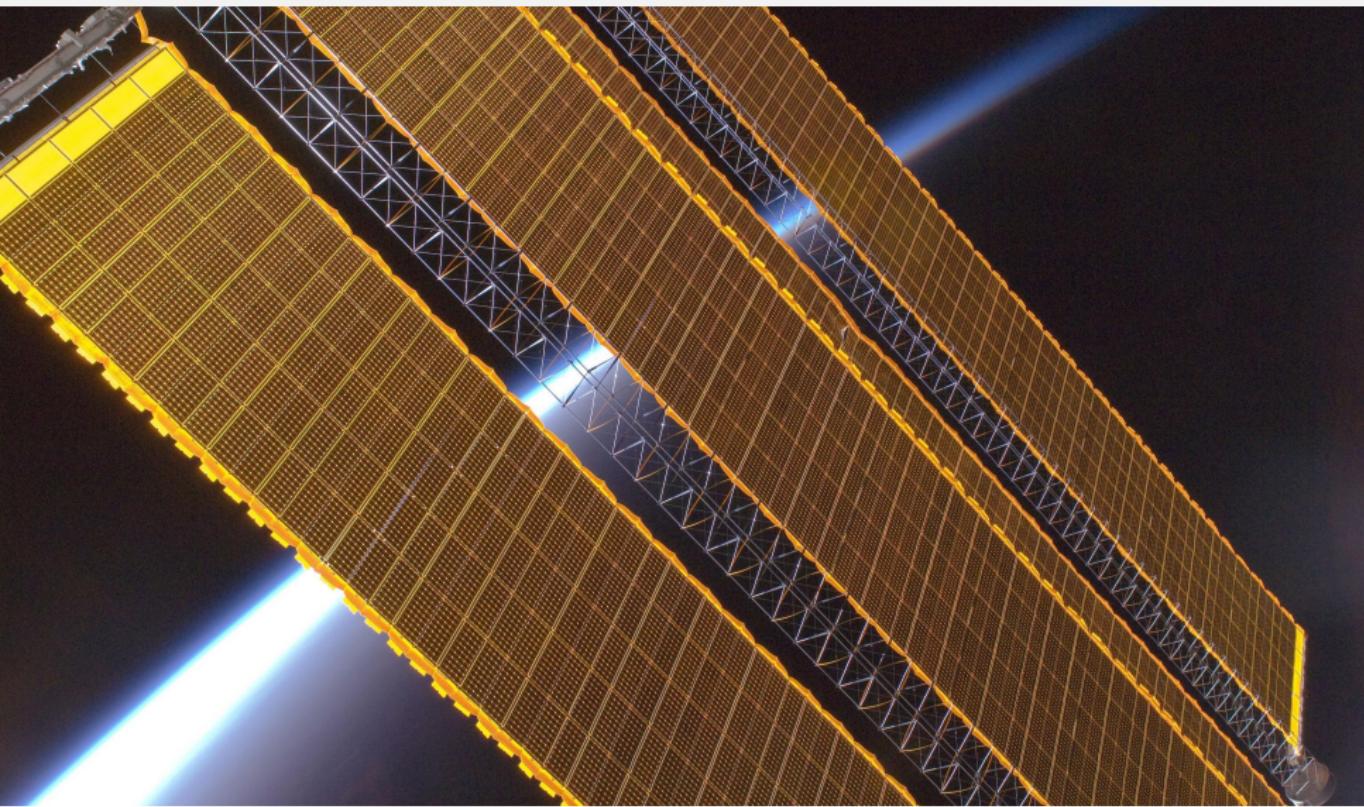
Data is like ore



Analysis concentrates its value



And it can be used to produce something



Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?

• Barriers to Data Analysis

- What is the solution?

3 Introduction to R

- What is R?

• What can R do?

- R Examples
- Getting Started

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Institutional Frustrations

We just need to get our jobs done. We need to do them efficiently, but also transparently and in a reproducible manner. This is currently costly in time, money, and management resources.



Institutional Frustrations

We just need to get our jobs done. We need to do them efficiently, but also transparently and in a reproducible manner. This is currently costly in time, money, and management resources.

- Acquiring proprietary tools from vendors takes agreements, legal documents, and lag time

Institutional Frustrations

We just need to get our jobs done. We need to do them efficiently, but also transparently and in a reproducible manner. This is currently costly in time, money, and management resources.

- 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 are often done in proprietary tool sets, poorly documented, and unable to be reproduced with updated data later



What about analyses we do?

Analyses we do complete are not used enough to drive decisions.

Incoherence

age draw further modifications of the unit printed on page N. — at the left a



FIG. 1.

Analyses Don't Get Used

Often when we do an in house analysis it does not get used or only gets used once.

- In house analysis often relies on the expertise of one or two staff who are obligated elsewhere.

Analyses Don't Get Used

Often when we do an in house analysis it does not get used or only gets used once.

- 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.

Analyses Don't Get Used

Often when we do an in house analysis it does not get used or only gets used once.

- 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.

Analyses Don't Get Used

Often when we do an in house analysis it does not get used or only gets used once.

- 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

Analyses Don't Get Used

Often when we do an in house analysis it does not get used or only gets used once.

- 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

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



UNSERE GÄRTEN

GOLF KILLS

www.wien.at



Gartenbezirk 7



Wiesen



Mundenh.



Mundenh.



LICHTTELEFON
797 75-8033



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.

Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- **What is the solution?**

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Open Source Tools



R

```

66 vars<-c("linkdensity", "sd_degreedist", "apl", "diameter", "mean_diameter", "sd_diameter", "mean.cons_omnivory", "skew.cons_trophiclevel",
67 for(o.var in vars){
68   newrow<-new.omnivary.row(c(o.var, alpha=alpha)
69   slopes<-rbind(slopes, newrow)
70 }
71
72
73 #write.table(slopes, file="removal_group.slopes.csv", sep=",", row.names=F)
74 write.table(slopes, file="removal_sp.slopes.csv", sep=",", row.names=F)
75
76
77 dfitfb<-rep(NA, 6)
78
79 vars<-c("linkdensity", "sd_degreedist", "apl", "diameter", "mean_diameter", "sd_diameter", "mean.cons_omnivory", "skew.cons_tropiclevel",
80 for(o.var in vars){
81   dfd<-df[,df[[o.var]] >= 1]^o.dfcchange)
82   lmd<-lm(df[[o.var]] ~ I((o.var)^o.dfcchange))
83   o.dfc<-subset(removal_data, removal_datayear==2009)
84   lmd<-lm(df[[o.var]] ~ I((o.var)^o.dfcchange))
85   o.dfc<-subset(removal_data, removal_datayear==2008)
86   lmd<-lm(df[[o.var]] ~ I((o.var)^o.dfcchange))
87
88 difs<-slope_compare(lmd, lmd2, tailed=1)
89 names(diffbt)<-c("var", names(difs))
90 diffbt<-rbind(difftb, c(o.var, difs))
91
92 difftb<-difftb[1,]
93 write.table(difftb, file="removal_slope_change.csv", sep=",", row.names=F)
94
95

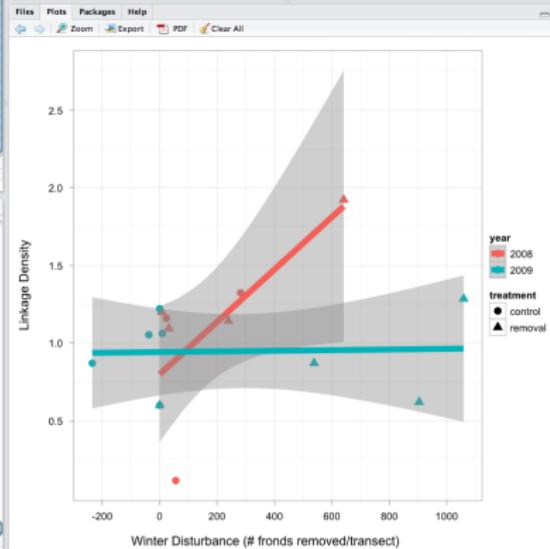
```

```

93 #graphing
94 library(ggplot2)
95 #tiff("removal_slopes.tiff", units="px", width=2000, height=1400, res=300)
96
97 gplot<-c1change, richness, data-removal_data, colour-year, shape-treatment, group-year, xlab="\\nWinter Disturbance (# fronds removed/transc
98 scale_lineetype("Year")-scale.colour,grey("Year"), start=0, end=0.5)-stat_smooth(method="lm", size=2, fill="grey55")-
99 scale.colour,grey("Year", start=0, end=0.5)-scale.colour,brewer("Year", palette="Set1")
100
101 gplot
102 #dev.off()
103
104 gplot<-c1change, linkdensity, data-removal_data, colour-year, shape-treatment, group-year, xlab="\\nWinter Disturbance (# Fronds removed/transc
105 gplot<-c1change, sd_degreeist, data-removal_data, colour-year, shape-treatment, group-year, size=5)+theme_bw() + stat_smooth(method="lm")
106 gplot
107
```

```
#dev.off()
plot(-1*tchange, linkdensity, data=remove_data, colour=year, shape=treatment, group=year, xlab="v/Hinter Disturbance (# fronds  
-1) over time", ylab="Link density", xres=100, col=1:100, bty="l", main="Link density vs. v/Hinter Disturbance (# fronds  
-1) for each year", font.main=1, font.lab=2, font.axis=2, font.ticks=2, font.legend=1, legend="treatment", lwd=2)
```

```
[1] "linkdensity"
[2] "sd_degreestd"
[3] "opl"
[4] "diameter"
[5] "mean_diameter"
[6] "sd_diameter"
[7] "mean_cons_omnivory"
[8] "skew_cons_trophiclevel"
[9] "mean_cons_trophiclevel"
[10] "power_exp"
[11] "prey_richness"
[12] "consumer_richness"
[13] "algae_richness"
[14] "sessile_invert_richness"
[15] "mobile_species_richness"
[16] "proportion_fruit" function "slope_compare"
> opl <- tchamp(linkdensity, datremoval_dots, colour=year, shape=treatment, group=year, xlab="v0#Inter Disturbance (# fronds removed/seedling)", ylab="Linkage Density", size=15) tchamp.hm(qplot.size=15) tchamp.smooth(method="lo", size=4)
```



R As the Solution

Objections to Data Analysis

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

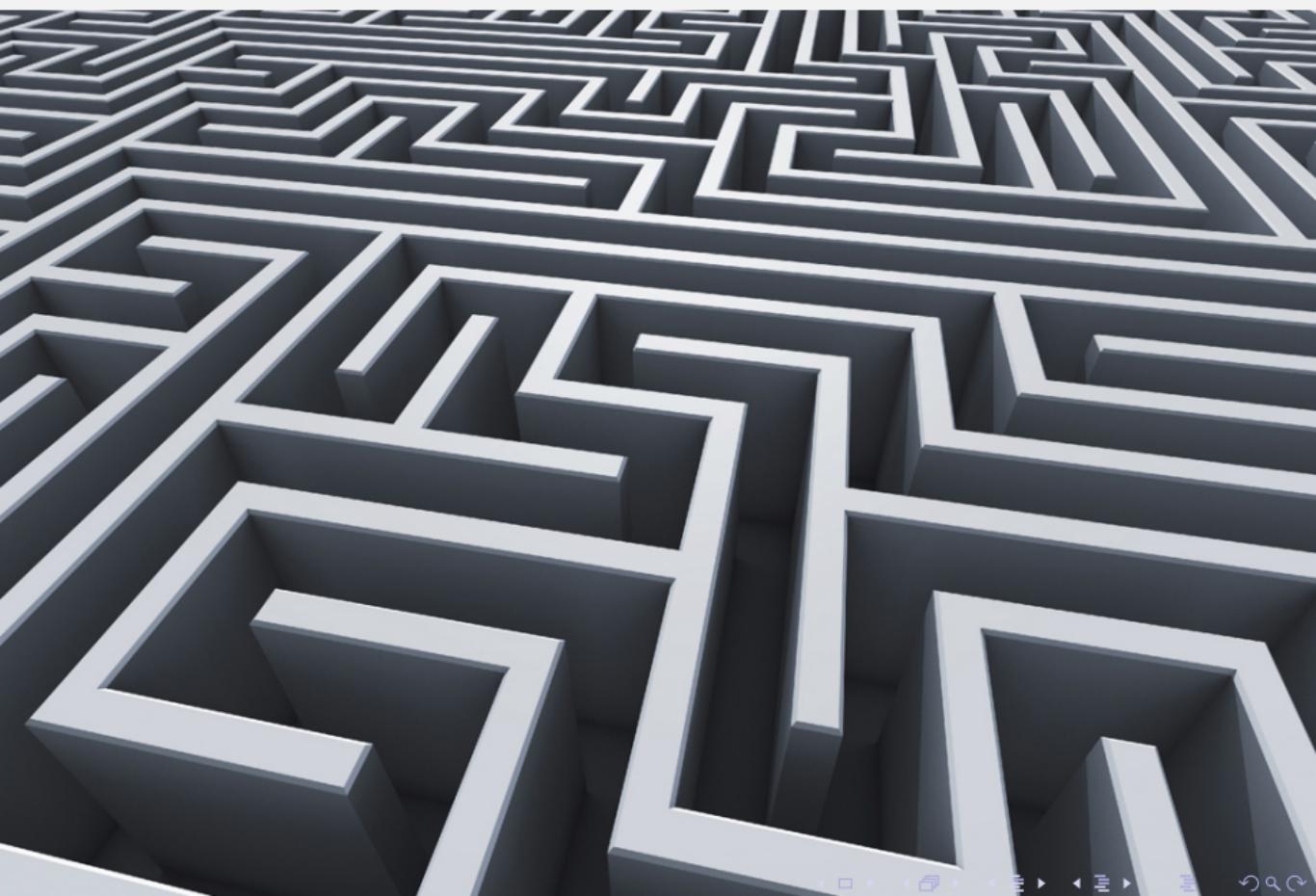
The R Solution

- 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.

The Truth

- The short answer is no.
- R has a high startup cost, but we are working to bring that down.

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.

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

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?
- R Examples
- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

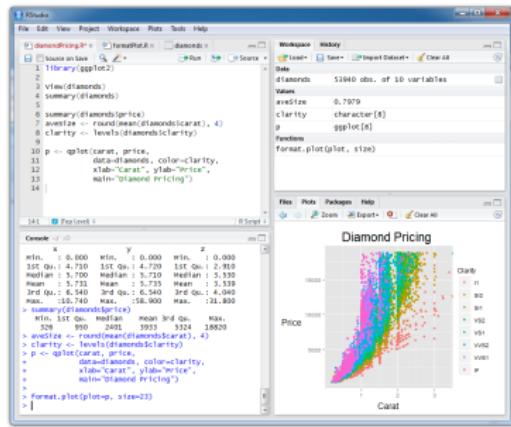
- R is an Open Source (and freely available) environment for statistical computing and graphics
- Available for Windows, Mac OS X, and Linux
- R is being actively developed with two major releases per year and dozens of releases of add on packages
- R can be extended with 'packages' that contain data, code, and documentation to add new functionality

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 wildly by both professional and academic developers

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

How does it help?

- R can be built to do what you want without waiting for a vendor to upgrade software

How does it help?

- R can be built to do what you want without waiting for a vendor to upgrade software
- All add-ons and functionality to R are free and shared immediately

How does it help?

- R can be built to do what you want without waiting for a vendor to upgrade software
- All add-ons and functionality to R are free and shared immediately
- Student Growth Percentiles (SGP) package is available for growth modeling now

How does it help?

- R can be built to do what you want without waiting for a vendor to upgrade software
- All add-ons and functionality to R are free and shared immediately
- Student Growth Percentiles (SGP) package is available for growth modeling now
- It standardizes processes and makes them more transparent

Standardization is Good

- By standardizing common data tasks staff are freed up to do other tasks

Standardization is Good

- By standardizing common data tasks staff are freed up to do other tasks
- Wisconsin is using R to calculate AYP directly from LDS dataset

Standardization is Good

- By standardizing common data tasks staff are freed up to do other tasks
- Wisconsin is using R to calculate AYP directly from LDS dataset
- One script, one run, all reports and error checks run—saving weeks of work

Standardization is Good

- By standardizing common data tasks staff are freed up to do other tasks
- Wisconsin is using R to calculate AYP directly from LDS dataset
- 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 turnover

Standardization is Good

- By standardizing common data tasks staff are freed up to do other tasks
- Wisconsin is using R to calculate AYP directly from LDS dataset
- 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 turnover
- Other reports and data analyses are being standardized—quality checks on LDS data, etc.

Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

• What can R do?

- R Examples
- Getting Started

4 More Advanced Functions

- Analysis
- Linear Model Example

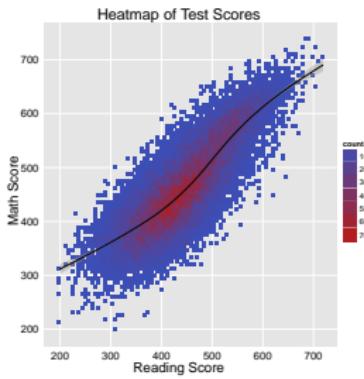
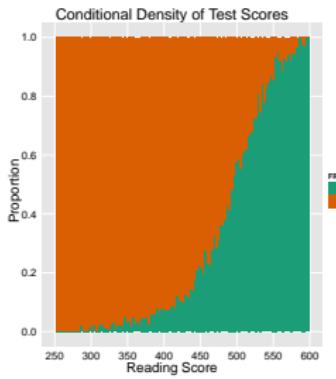
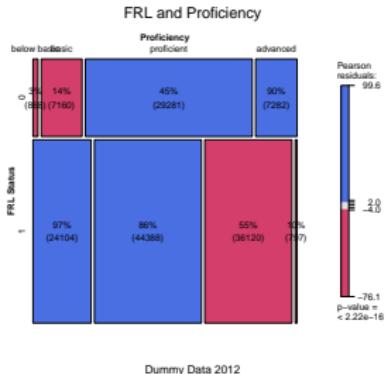
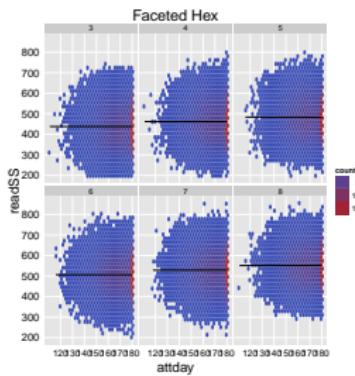
5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Visualization

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

Examples of R Figures



Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

• R Examples

- Getting StaRted

4 More Advanced Functions

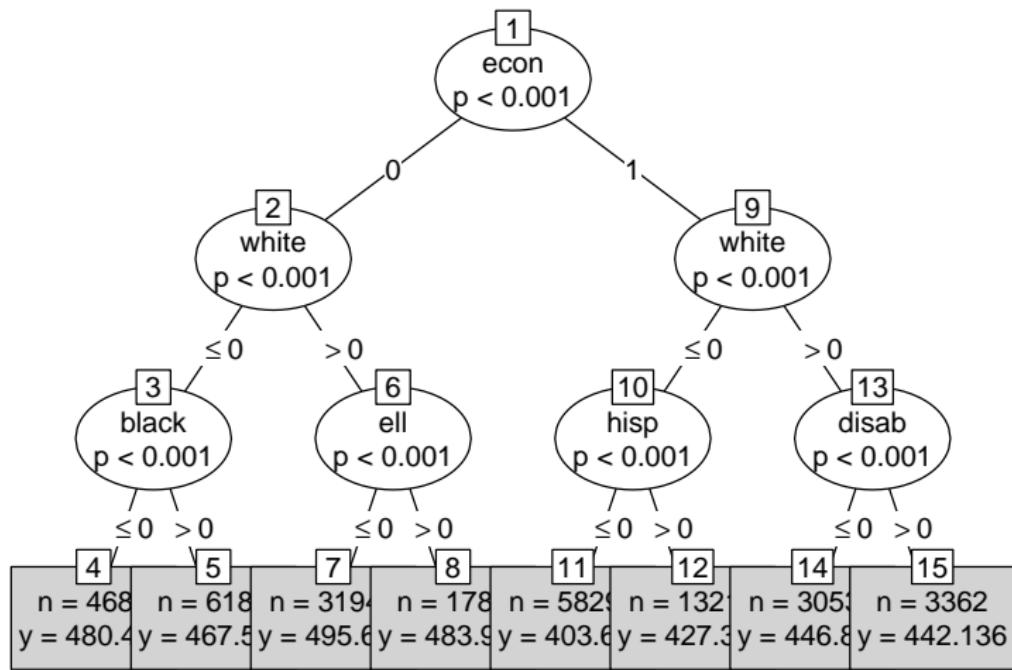
- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Inference Trees

Splitting Categorical Data



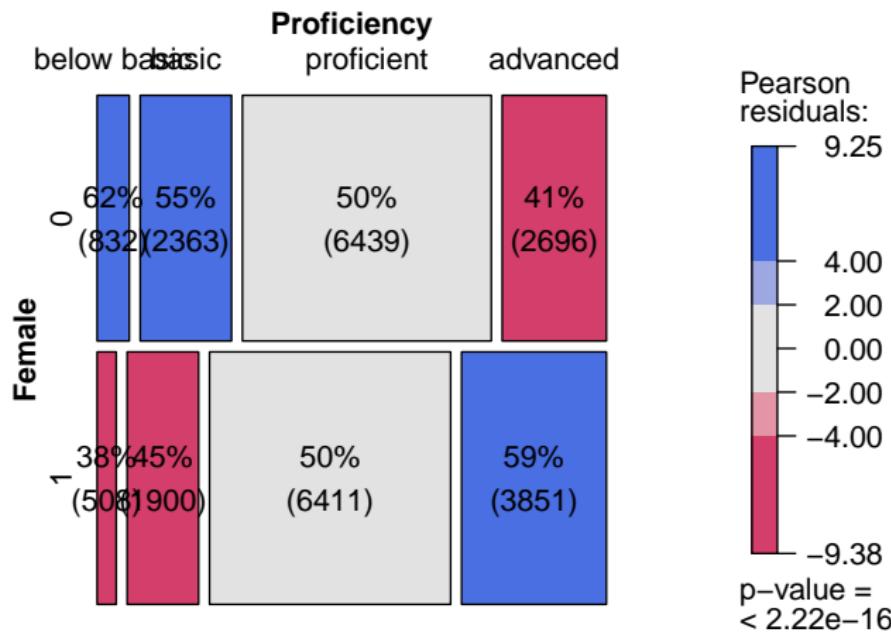
Easy Code for Plot

Code to make this plot:

```
> z1<-ctree(readSS~black+hisp+asian+indian+white+ell+disab+econ  
+           +attday,data=subset(student_long,year=='2000',grade=5),  
+           controls=bonf)  
> plot(z1,type='simple',main="Splitting Categorical Data")
```

Visual Crosstabs

Gender and Proficiency

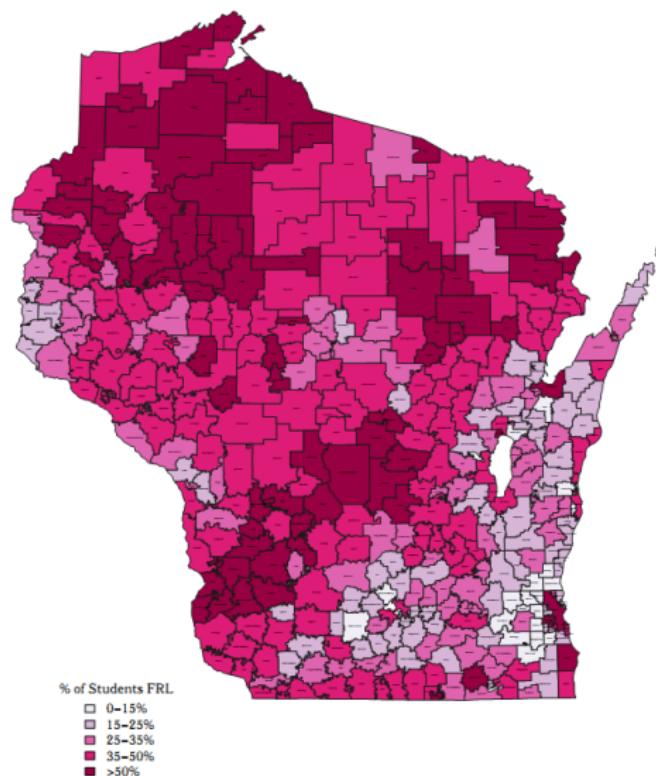


R Code

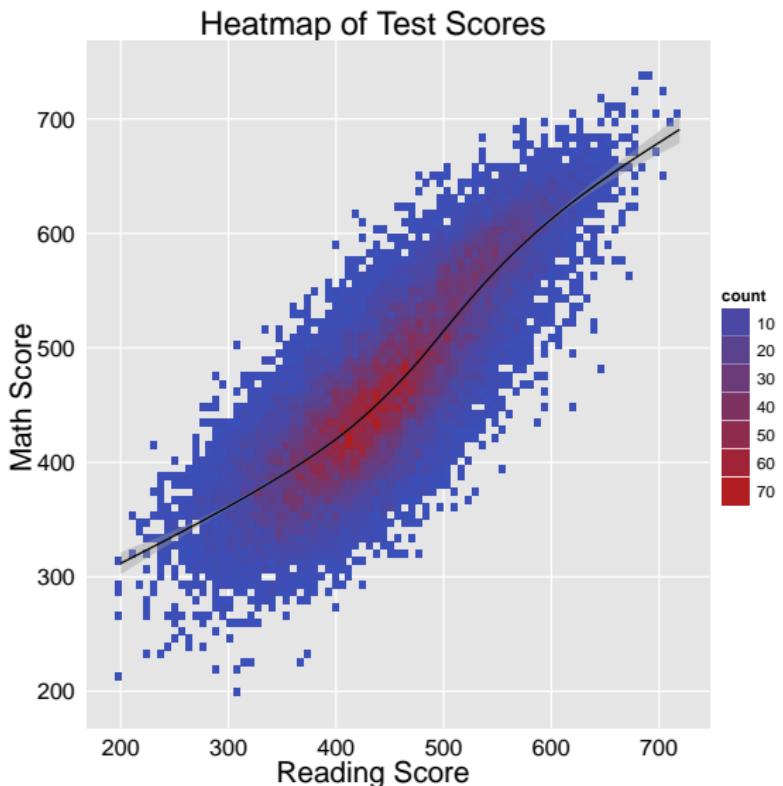
Code for this plot that uses 'mosaictabs' function from the LDS_TOOLS package

```
> 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 2012')
```

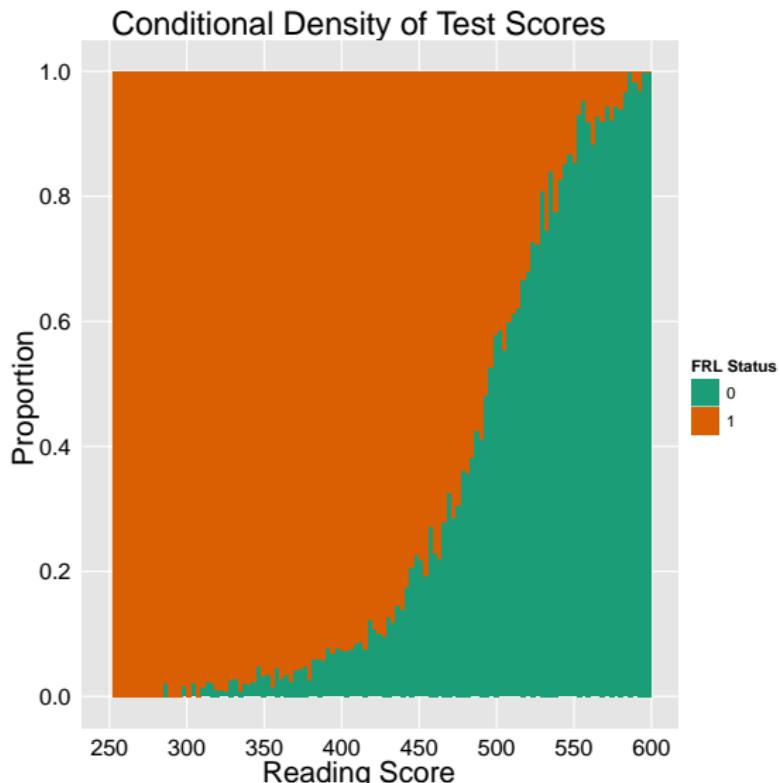
Maps



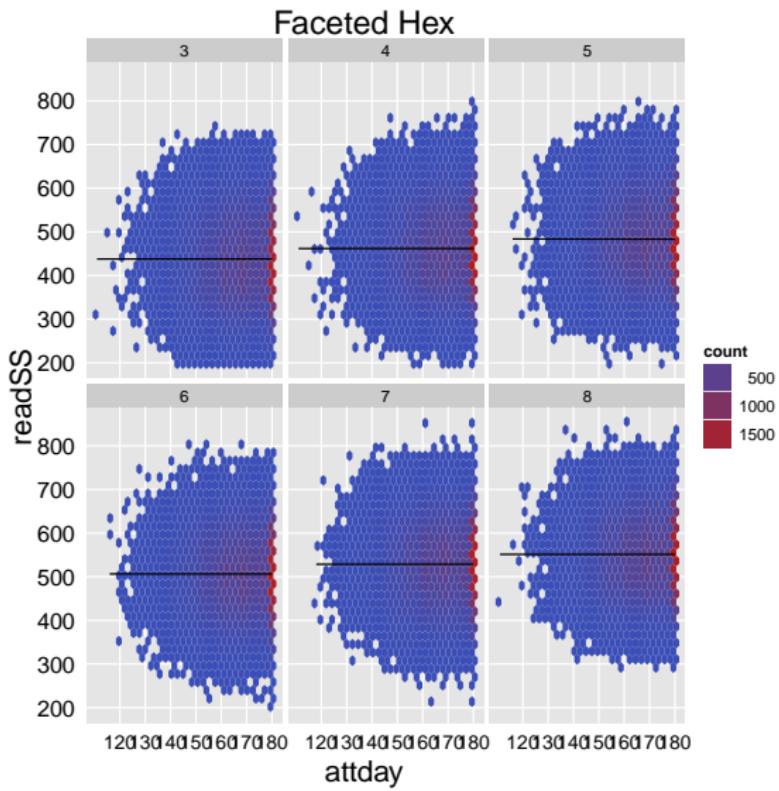
Heatmap



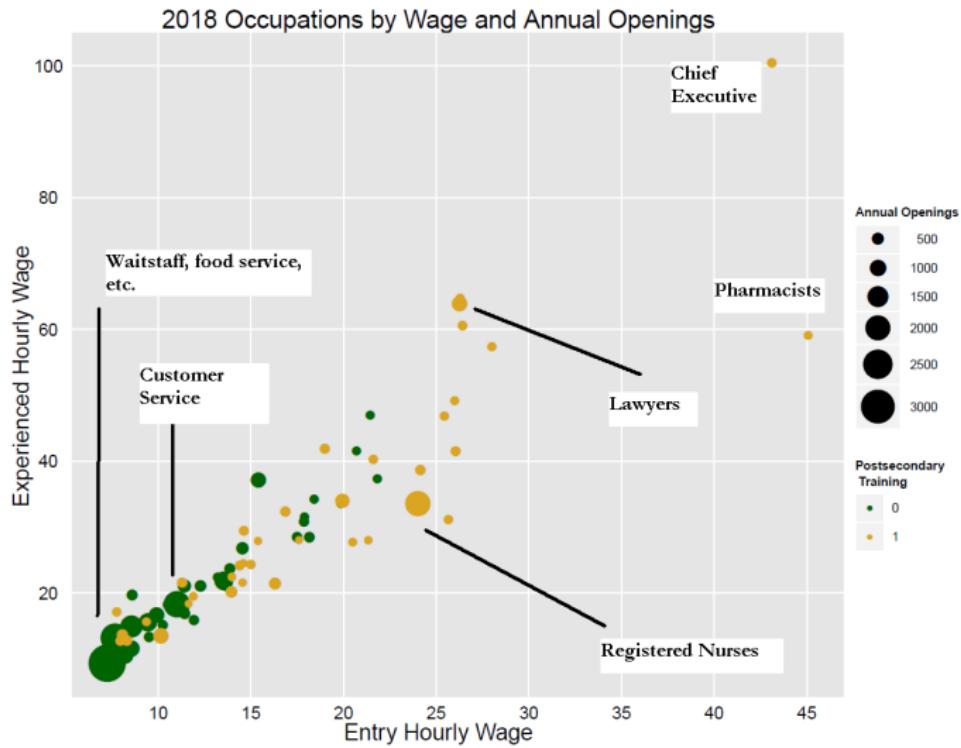
Conditional Density



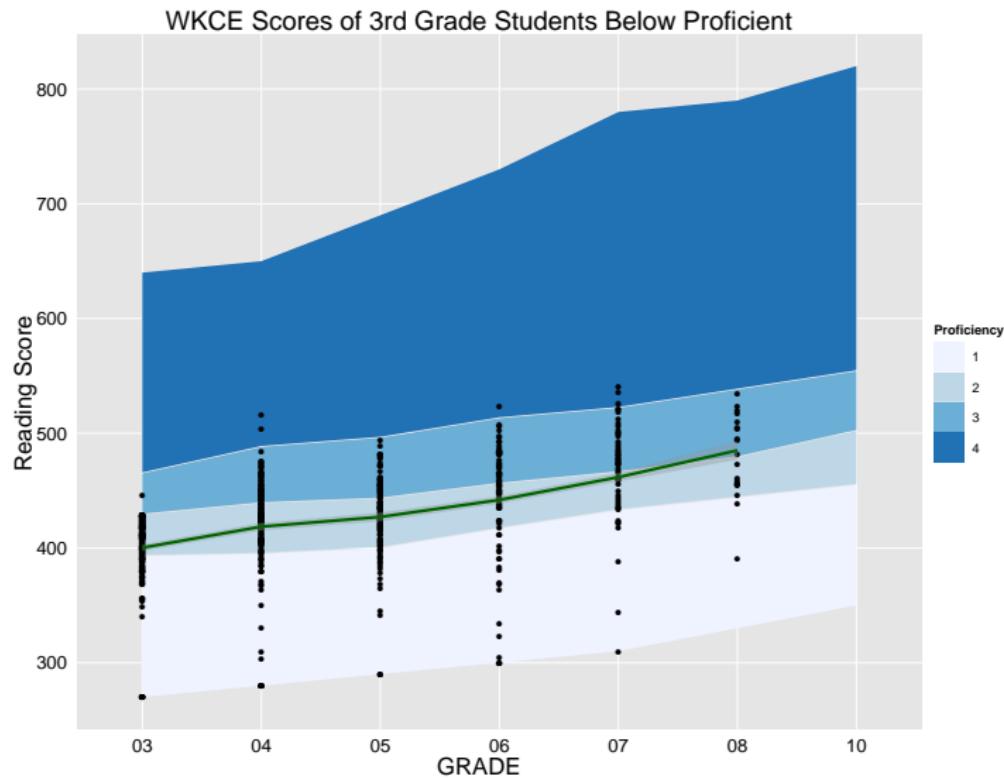
Faceted Hex



Polished Scatter

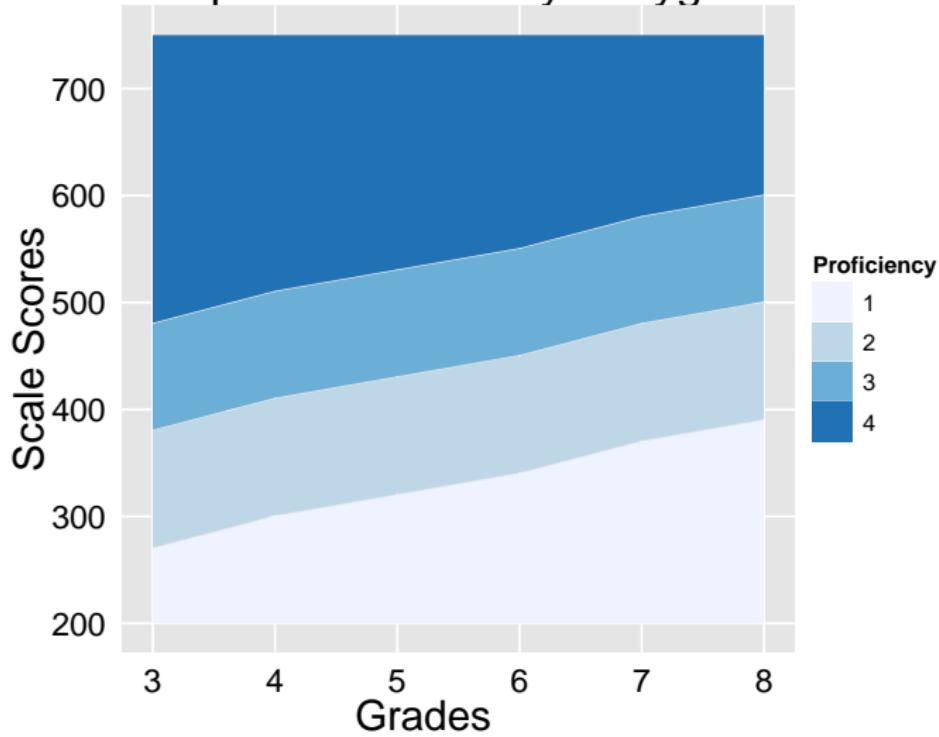


Longitudinal Data



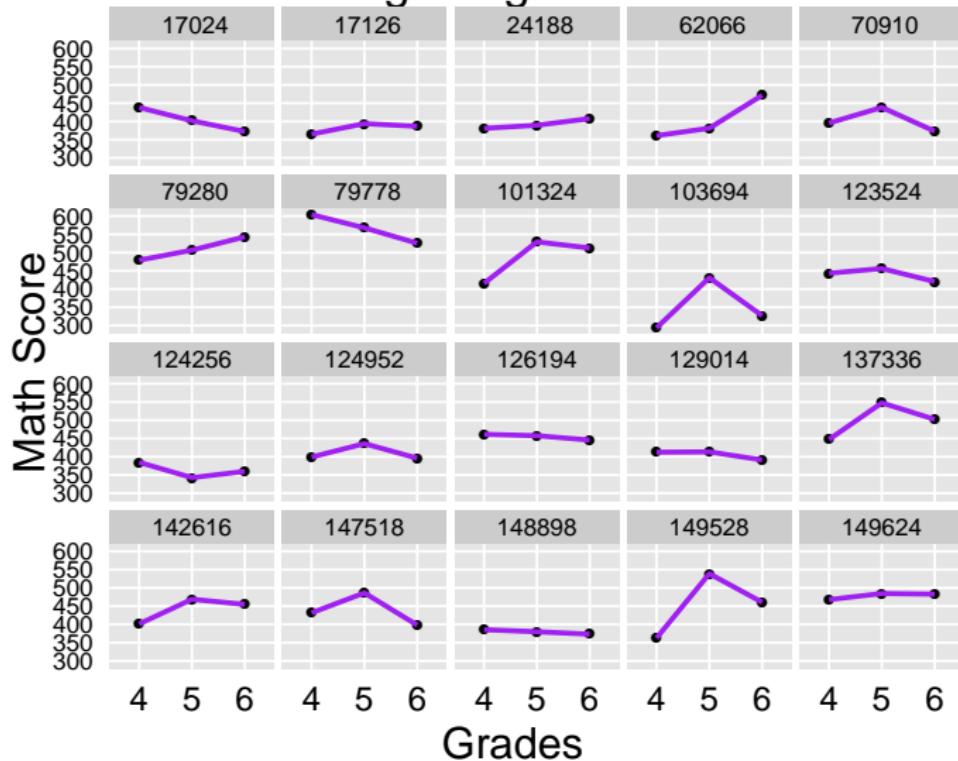
Proficiency Polygon

Example of Proficiency Polygon

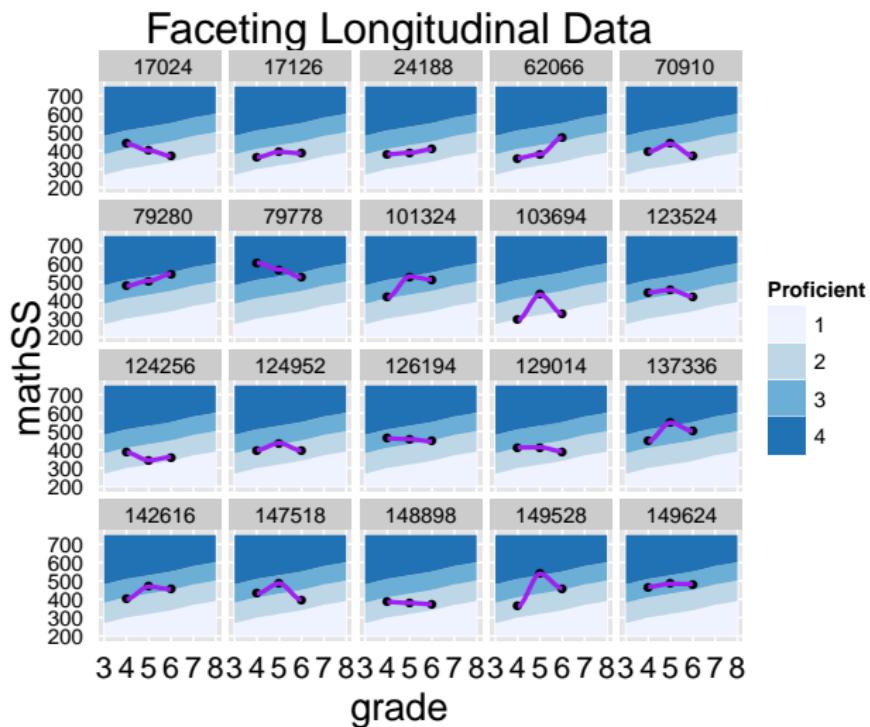


Individual Growth Trajectories

Faceting Longitudinal Data

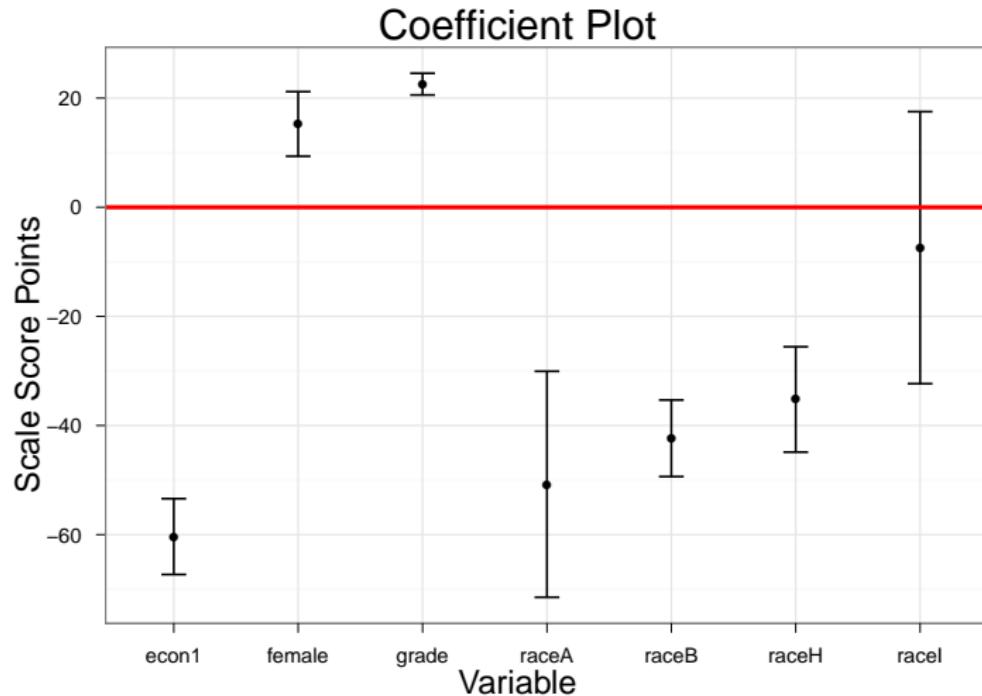


Individual Growth Trajectories

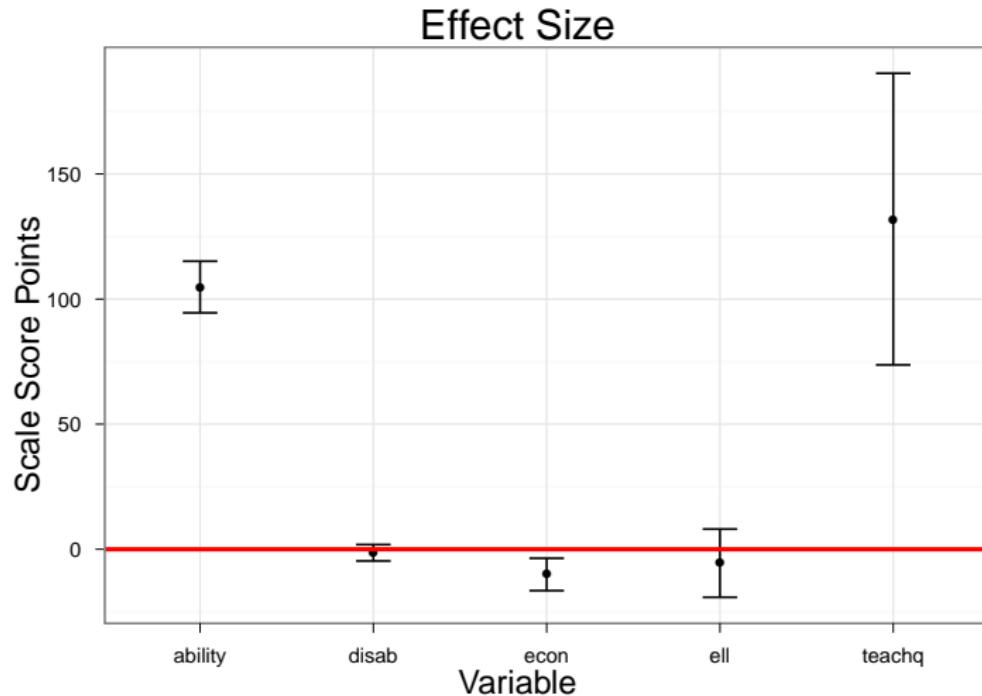


```
| null device  
|       1
```

Communicate Statistical Models



Communicate Statistical Models II



Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

• Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

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

Simple R Operations

```
> with(student_long,mean(readSS[year=='2001' & grade==4]))
```

```
[1] 444.7696
```

```
> with(student_long,median(readSS[year=='2001' & grade==4]))
```

```
[1] 442.1458
```

```
> with(student_long,max(readSS[year=='2001' & grade==4]))
```

```
[1] 719.052
```

```
> with(student_long,min(readSS[year=='2001' & grade==4]))
```

```
[1] 200
```

```
> with(student_long,summary(readSS[year=='2001' & grade==4]))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
200.0	393.3	442.1	444.8	495.2	719.1



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

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?
- R Examples
- Getting Started

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

• What can R do?

- R Examples
- Getting Started

4 More Advanced Functions

• Analysis

- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

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.

ANOVA

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

```
> novat1<-aov(readSS~female*race*econ,data=novaset)
> summary(novat1)
```

	Df	Sum Sq	Mean Sq	F value
female	1	2176464	2176464	766.610
race	4	33929701	8482425	2987.743
econ	1	20785160	20785160	7321.103
female:race	4	50009	12502	4.404
female:econ	1	12835	12835	4.521
race:econ	4	266621	66655	23.478
female:race:econ	4	18117	4529	1.595
Residuals	24980	70920091	2839	
		Pr(>F)		
female		< 2e-16 ***		
race		< 2e-16 ***		
econ		< 2e-16 ***		
female:race		0.00147 **		
female:econ		0.03349 *		

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	766.61	0.0000
race	4	33929701.29	8482425.32	2987.74	0.0000
econ	1	20785160.43	20785160.43	7321.10	0.0000
female:race	4	50008.56	12502.14	4.40	0.0015
female:econ	1	12835.21	12835.21	4.52	0.0335
race:econ	4	266620.58	66655.15	23.48	0.0000
female:race:econ	4	18116.94	4529.24	1.60	0.1725
Residuals	24980	70920091.02	2839.07		

Outline

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting Started

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

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 = stu
```

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
raceI	2.48529	2.15941	1.151
raceA	0.45644	1.38578	0.329

A simple OLS Model II

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
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

A simple OLS Model III

	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 ***
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 ***

A simple OLS Model IV

```
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
```

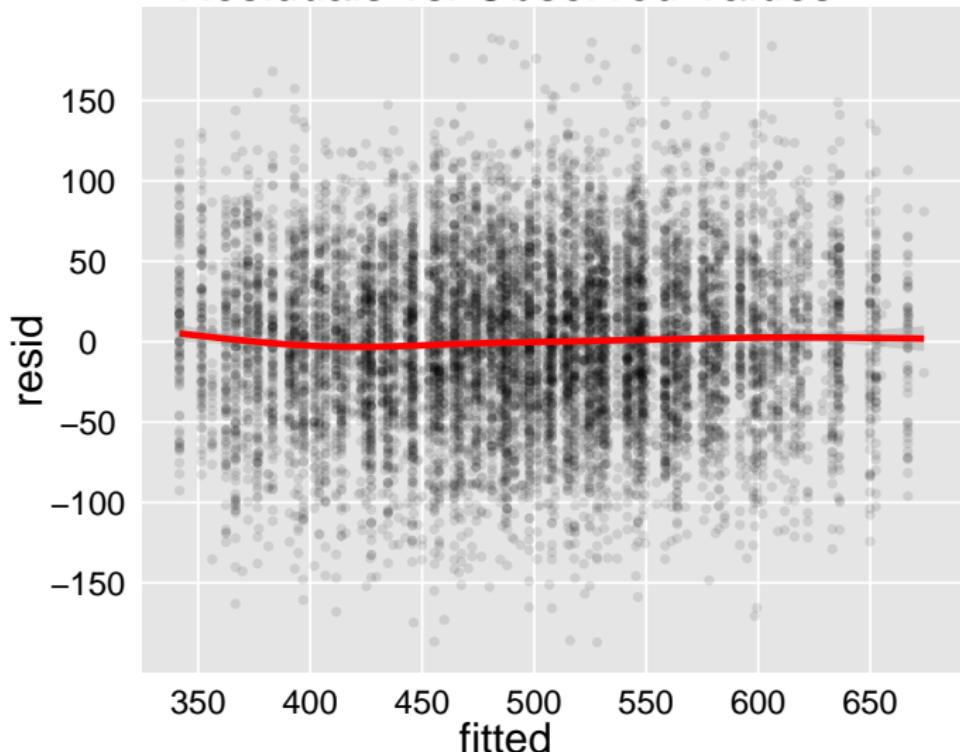
Residual standard error: 51.53 on 449975 degrees of freedom

Multiple R-squared: 0.6742, Adjusted R-squared: 0.6742

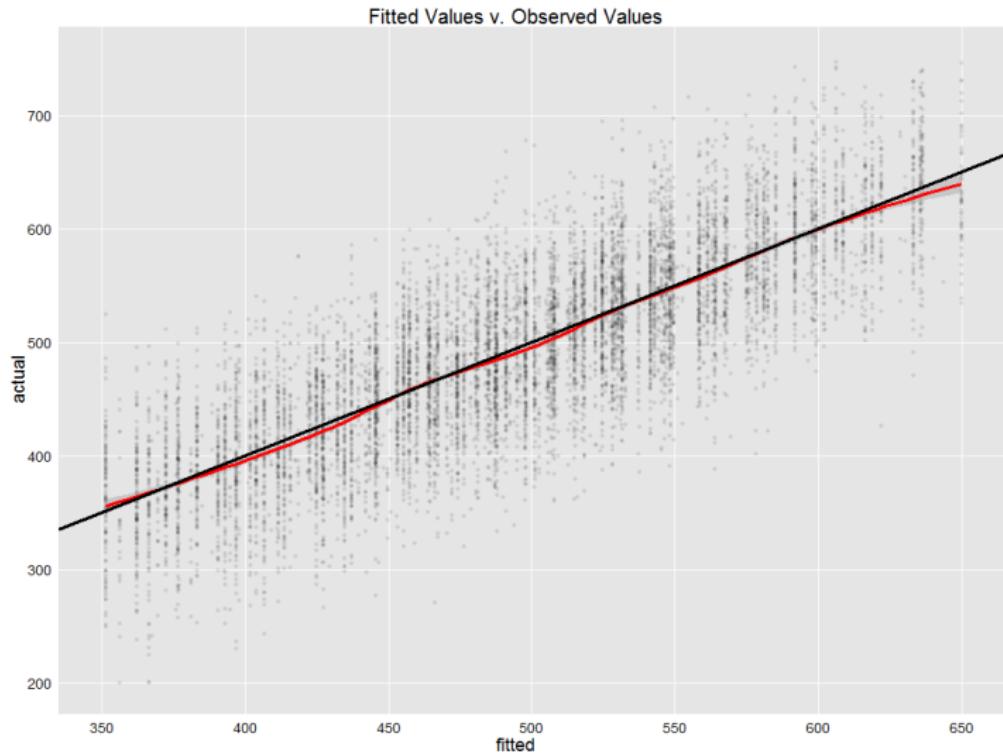
F-statistic: 3.88e+04 on 24 and 449975 DF, p-value: < 2.2e-16

Model Evaluation

Residuals vs. Observed Values

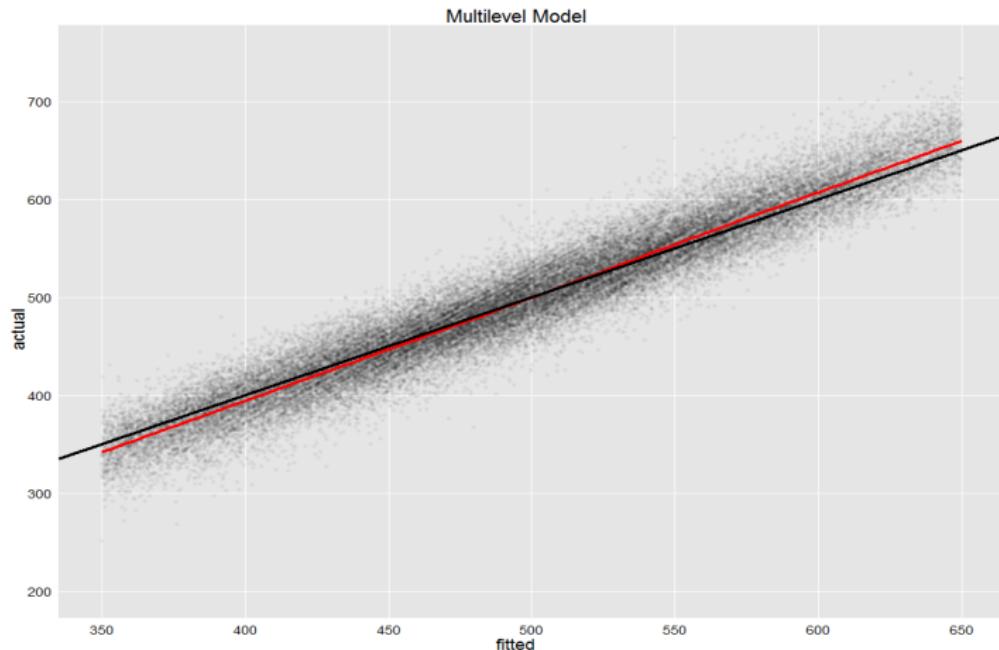


Model Evaluation Part 2



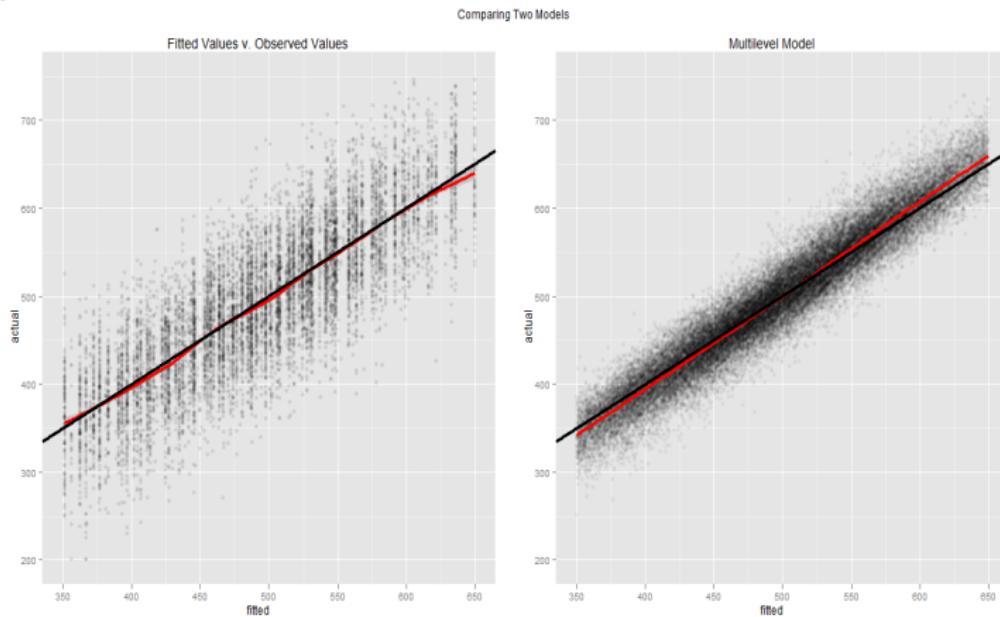
Better Fitting

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

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

• What can R do?

• R Examples

• Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

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

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

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

1 What is Policy Relevant Analysis?

- Defining Terms
- The Problem

2 Extracting Meaning from Data

- Why is Data Analysis So Important?
- Barriers to Data Analysis
- What is the solution?

3 Introduction to R

- What is R?

- What can R do?

- R Examples

- Getting StaRted

4 More Advanced Functions

- Analysis
- Linear Model Example

5 Putting It Together to Collaborate

- Same Data, Similar Analyses
- Coordinating and Social Coding

GitHub

GitHub provides an excellent 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

Example

This entire presentation was created with R, L^AT_EX and is available online at GitHub to be edited, modified, and reused.

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/>)