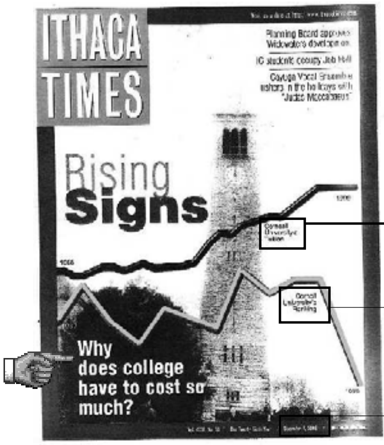
Smoking, Stephen Jay Gould and Convertibles:
Statistical Graphics for Data Presentation and Analysis

Thomas E. Love, Ph.D.
TEL3@po.cwru.edu & www.chrp.org

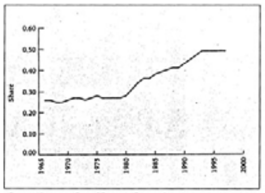
January 25, 2002



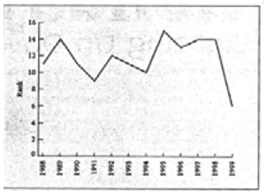
Cornell University's Tuition

Cornell University's Ranking

7 December 2000



By the Numbers: Over 35 years, Cornell's tuition has taken an increasingly larger share of its median student family income.



Pecking Order: Over 12 years, Cornell's ranking in *US News & World Report* has risen and fallen erratically.

By the Numbers:

Over 35 years, Cornell's Tuition has taken an increasingly larger share of its median student family income.

Pecking Order:

Over 12 years, Cornell's ranking in US News & World Report has risen and fallen erratically.

Outline

- Philosophy
- Aim of good graphics
- "Plotting" – smoking and Y-Y
- Good advice and MLB finances
- S. J. Gould and "Goosing"
- Showing Balance
- Some Bad Ideas
- Some Good Ideas
  - Transformations
  - Graphical arrays

Make the data stand out. Avoid clutter.

Data Analysis is like Doing Experiments

- Discovery is usually more exciting and important than confirmation.
- Interaction, feedback, and trial and error are all critically important.
- Better to start trying to obtain and explain specific findings rather than figure everything out at once.
- Insight more important than objectivity.

Aim of Good Data Graphics

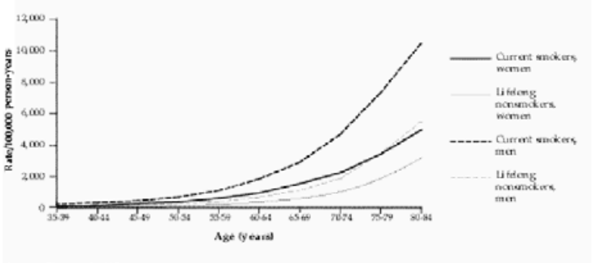
- The aim of good data graphics is to display data accurately and clearly.
- A good graph is quiet and lets the data tell their story clearly and completely.
- Graphs are best when they "force us to see what we never expected."

Smoking, Women and "Plots"

Thanks to Howard Wainer and the CDC web site

Visual clarity must be preserved under reduction and reproduction

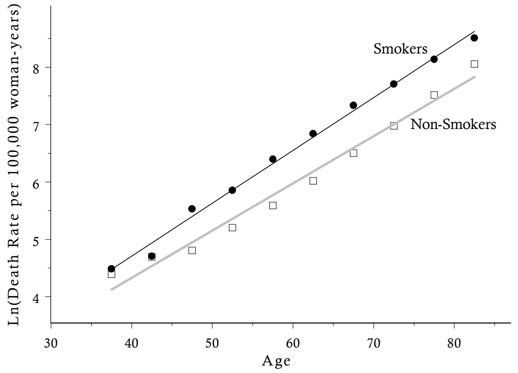
Figure 3.1. All-cause death rates for current smokers and lifelong nonsmokers, by age and gender, Cancer Prevention Study II, 1982–1988



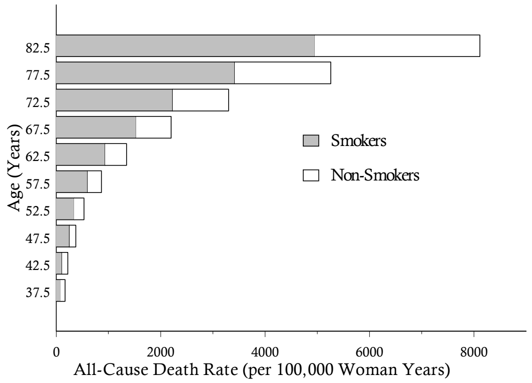
Sources: Thun et al. 1997a,c.

www.cdc.gov/tobacco/sgr/sgr_forwomen/pdfs/chp3.pdf

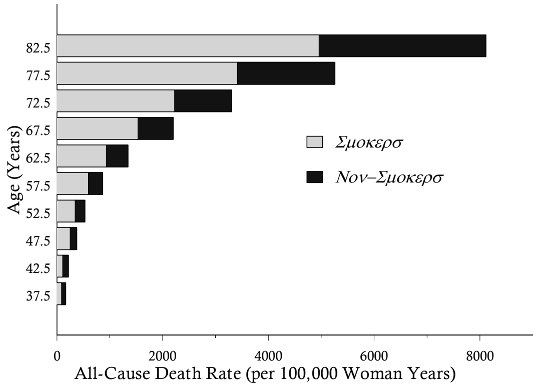
All-Cause Death Rate (Log Scale) Plotted Against Age
Women in Cancer Prevention Study II, 1982-1988



Smoking and Death Rates Shown by Age



Smoking and Death Rates Shown by Age



Surgeon General Reports
Aging is the Primary Cause of Death

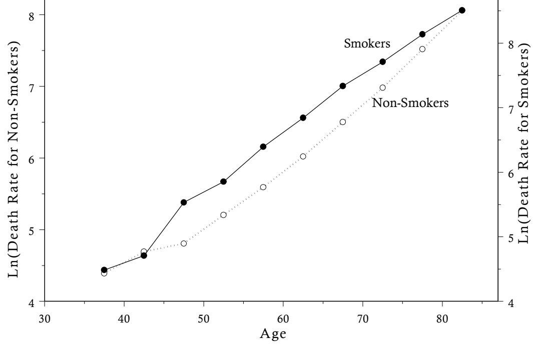
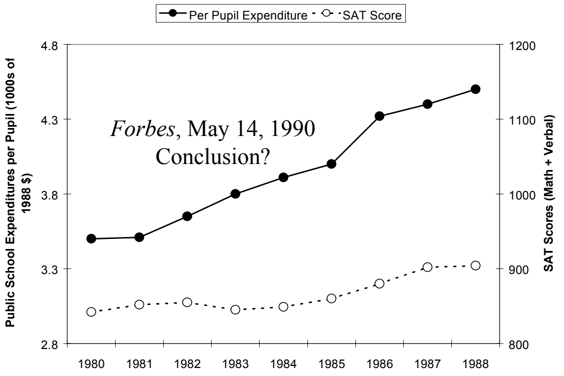


Figure 1: Expenditures vs. SAT Scores



Forbes, May 14, 1990
Conclusion?

Figure 2: SAT Scores and funds for education rise together

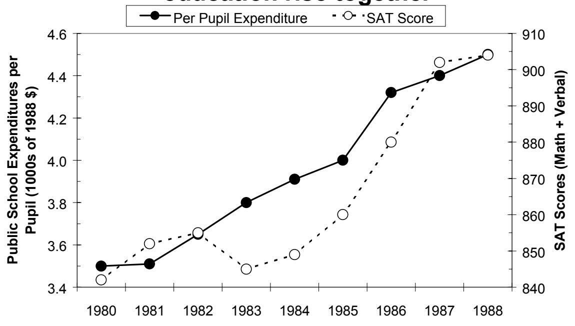
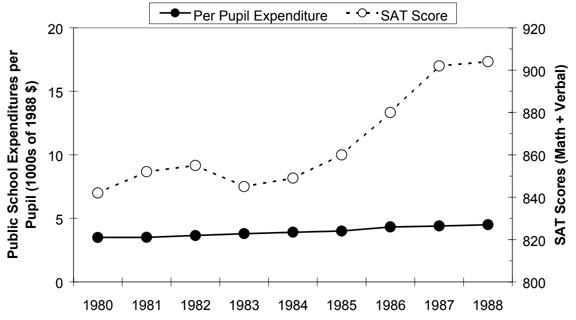


Figure 3: SAT scores soar despite sluggish funding of education



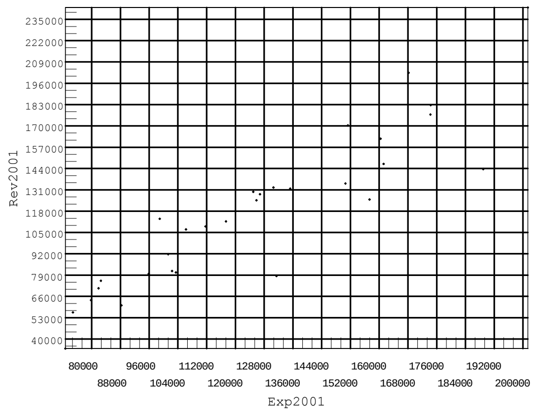
Graphical Displays Should…

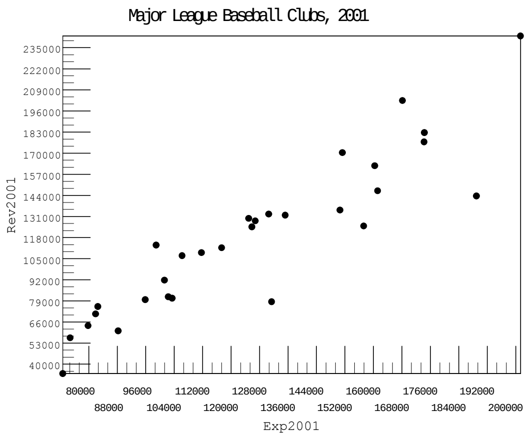
- Show the data.
- Induce the viewer to think about the substance of the data.
- Avoid distorting what the data have to say.
- Serve a clear purpose.

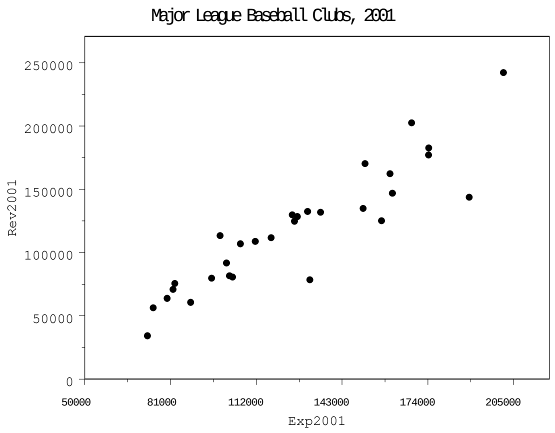
MAJOR LEAGUE BASEBALL 2001 ECONOMIC FORECASTS

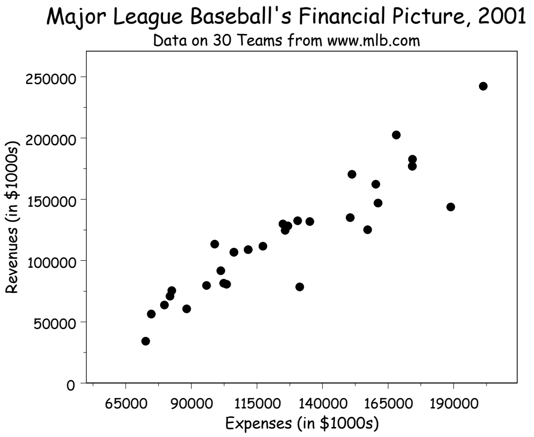
www.mlb.com/mlb/hearings/downloads/overview.pdf

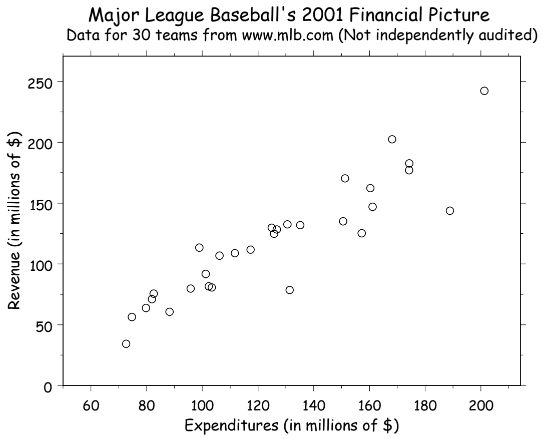
Major League Baseball Clubs, 2001

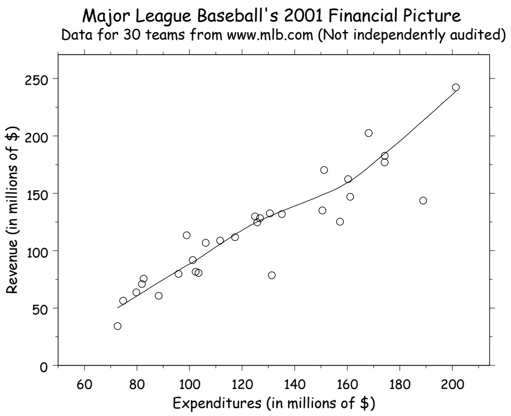


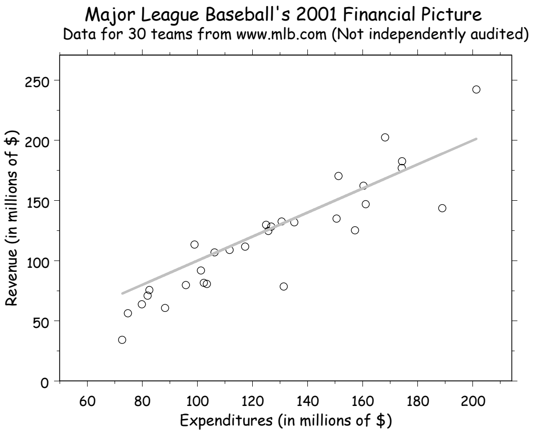

Major League Baseball Clubs, 2001
Rev2001
235000
222000
209000
196000
183000
170000
157000
144000
131000
118000
105000
92000
79000
66000
53000
40000
80000
88000
96000
104000
112000
120000
128000
136000
144000
152000
160000
168000
176000
184000
192000
200000
Exp2001

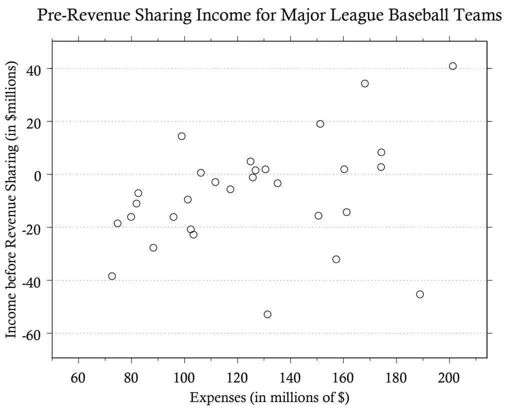

Major League Baseball Clubs, 2001
Rev2001
250000
200000
150000
100000
50000
0
50000
81000
112000
143000
174000
205000
Exp2001


Major League Baseball's Financial Picture, 2001
Data on 30 Teams from www.mlb.com
Revenues (in $1000s)
250000
200000
150000
100000
50000
0
65000
90000
115000
140000
165000
190000
Expenses (in $1000s)

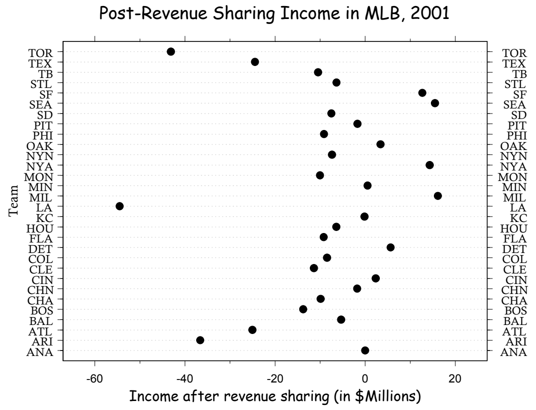

Major League Baseball's 2001 Financial Picture
Data for 30 teams from www.mlb.com (Not independently audited)
Revenue (in millions of $)
250
200
150
100
50
0
60
80
100
120
140
160
180
200
Expenditures (in millions of $)


Major League Baseball's 2001 Financial Picture
Data for 30 teams from www.mlb.com (Not independently audited)
Revenue (in millions of $)
250
200
150
100
50
0
60
80
100
120
140
160
180
200
Expenditures (in millions of $)

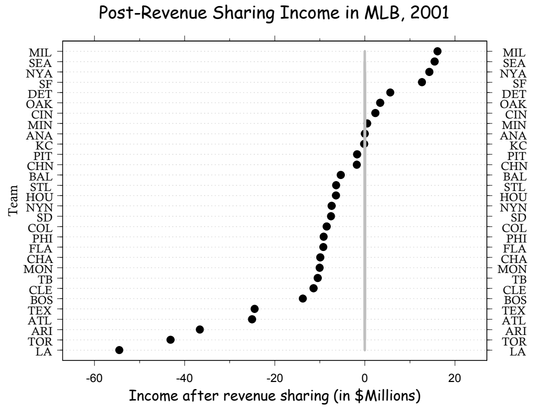

Major League Baseball's 2001 Financial Picture
Data for 30 teams from www.mlb.com (Not independently audited)
Revenue (in millions of $)
250
200
150
100
50
0
60
80
100
120
140
160
180
200
Expenditures (in millions of $)



Pre-Revenue Sharing Income for Major League Baseball Teams



Post-Revenue Sharing Income in MLB, 2001



Post-Revenue Sharing Income in MLB, 2001

Make the data stand out.

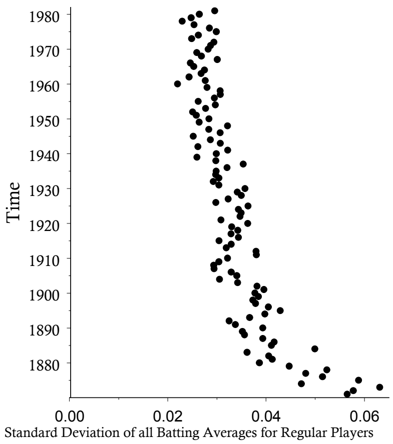
- Don't let anything obscure the data.
  - Use visually prominent graphical elements to show the data.
  - Make the data rectangle slightly smaller than the scale-line rectangle.
  - Tick marks should point outward. Do not overdo the tick marks or grid lines.
  - Overlapping plotting symbols must be visually distinguishable.

Make the data stand out.

- Do not clutter the interior of the scale-line rectangle.
  - Use a reference line when there is an important value that must be seen across the entire graph, but don't let the line interfere with the data.
  - Do not allow data labels to interfere with the data or clutter the graph.
  - Put keys outside the scale-line rectangle, and put notes in the caption or text.

STEPHEN JAY GOULD AND PLOT "GOOSING"

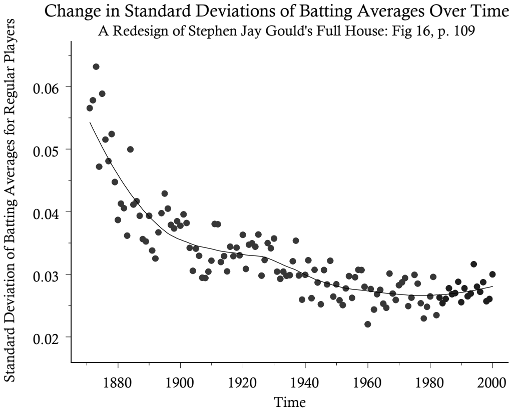
SJ Gould Full House (1996), figure 16, page 109



Redrawn
to match
Gould, S.J.
Full House
(1996),
p.109,
fig 16

Change in Standard Deviations of Batting Averages Over Time

A Redesign of Stephen Jay Gould's Full House: Fig 16, p. 109

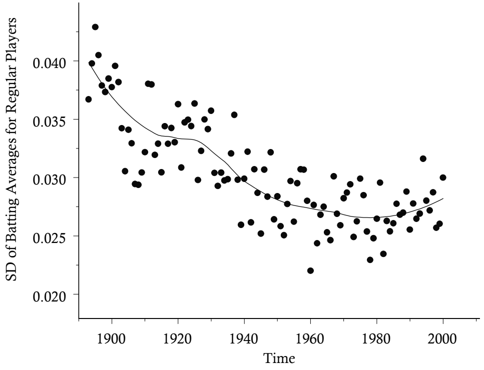


LOWESS Smoothing

- LOWESS (Loess) = Locally Weighted Regression Smoothing
- Purpose – summarize the middle of the distribution of Y for a given X.
  - Produces smoothed Y values at any point on the X scale, then connects the smooths with line segments.
  - Smoothing parameter α requires judgment, or can automate this.

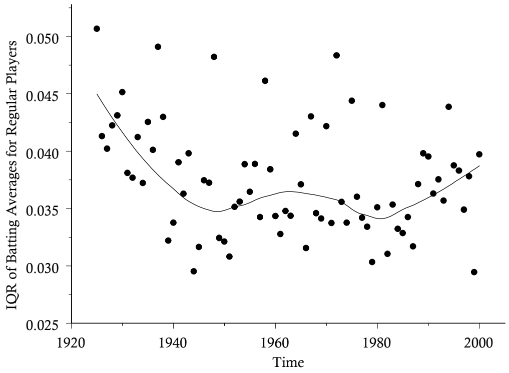
Standard Deviations of Batting Averages Since 1893 (60 ft 6 in)

Prompted by SJ Gould's Full House, Fig 16, p. 109 with Lowess Smooth



InterQuartile Ranges of Batting Averages Over The Last 75 Years

Prompted by Stephen Jay Gould's Full House: Fig 16, p. 109, With Lowess Smooth



Checking for Covariate Balance

Checking for Covariate Balance: Large Tabular Presentations

Variable	(Rx) RP %	(Ctrl) RT %	Unadjusted Wald F (p)	Wald F (p) **adj. for PS**
Incontinent	3	8	12.2 (<.001)	0.09 (.76)
Impotent	21	38	36.1 (<.001)	0.85 (.36)
CHF	5	8	3.8 (.05)	0.20 (.66)
Lung Dx	7	12	5.7 (.02)	0.02 (.89)
Hypertens.	41	45	1.2 (.28)	0 (.98)
Angina	9	18	17.0 (<.001)	0.25 (.62)

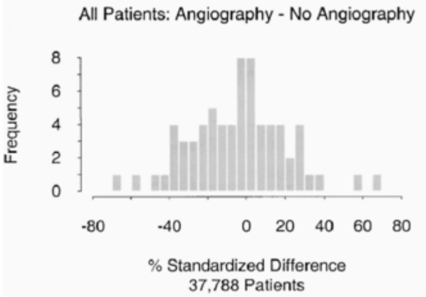
Adapted from Table 1 in Potosky et al. (2000) p. 1585

Does Matching By Propensity Scores Help Reduce Selection Bias?

Standardized Differences are an Appropriate Summary Statistic to Use in Assessing Covariate Balance

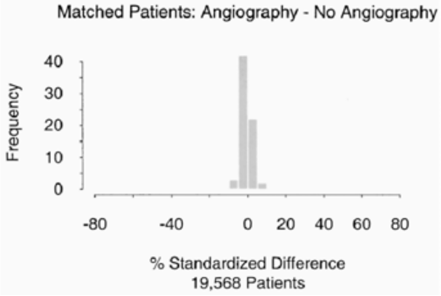
$$d = \frac{100\left(\bar{x}_{Treatment} - \bar{x}_{Control}\right)}{\sqrt{\frac{\left(s^2_{Treatment} + s^2_{Control}\right)}{2}}}$$

Standardized Differences (%) in Covariate Means: Before Matching

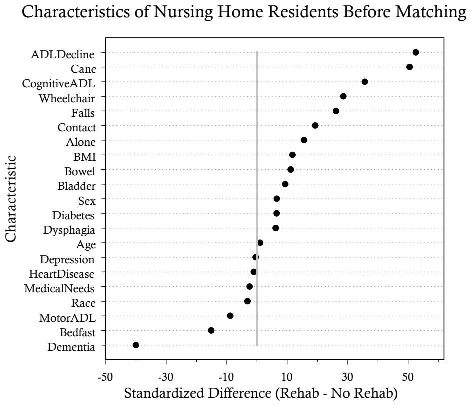


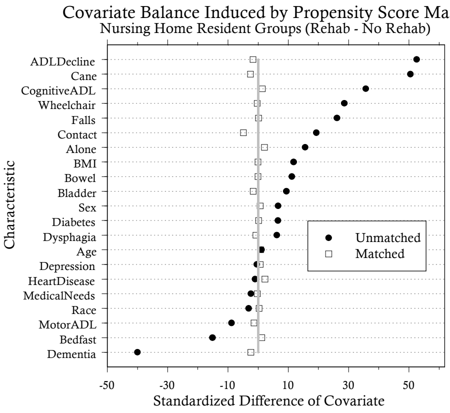
Normand et al. (2001) p. 395

Standardized Differences (%) in Covariate Means: After Matching



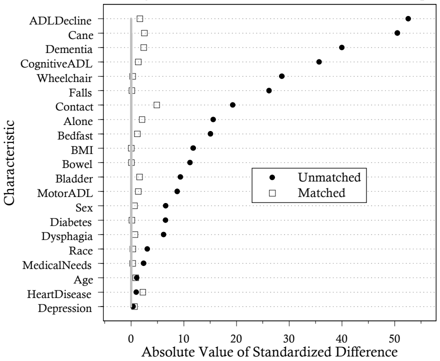
Normand et al. (2001) p. 395



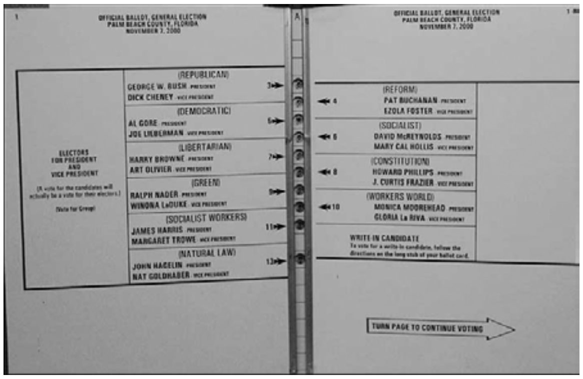


Covariate Balance Induced by Propensity Score Matching

Nursing Home Residents (Rehab vs. No Rehab Groups)



To Transformations



OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS
FOR PRESIDENT
AND
VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3→

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5→

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7→

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9→

(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11→

(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13→

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

←4
(REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

←6
(SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

←8
(CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

←10
(WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

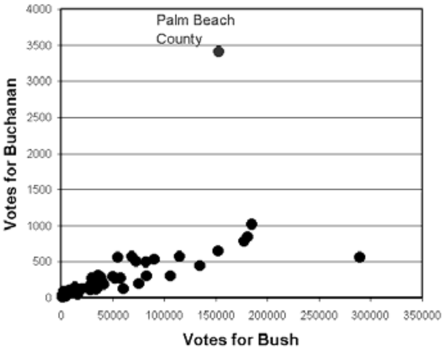
WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

TURN PAGE TO CONTINUE VOTING

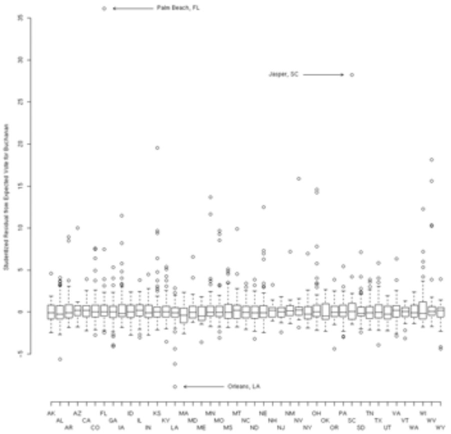
http://madison.hss.cmu.edu/

http://www.sun-sentinel.com/graphics/news/ballot.htm

Presidential Election Results for Florida, by County

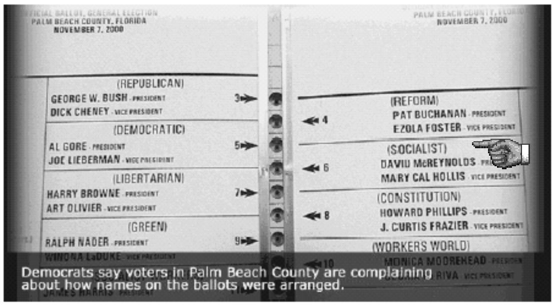


Prof. Greg Adams (Carnegie Mellon Univ.) Prof. Chris Fastnow (Chatham College)



McReynolds Effect?

Sample ballot



OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3→

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5→

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7→

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9→

PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

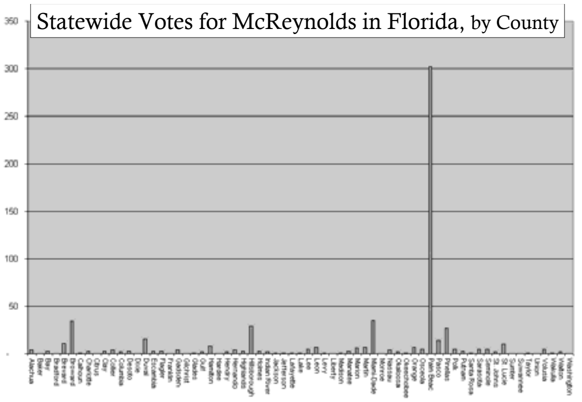
←4
(REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

←6
(SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

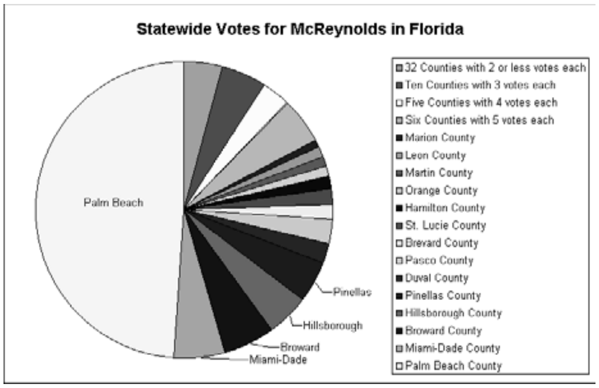
←8
(CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

←10
(WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

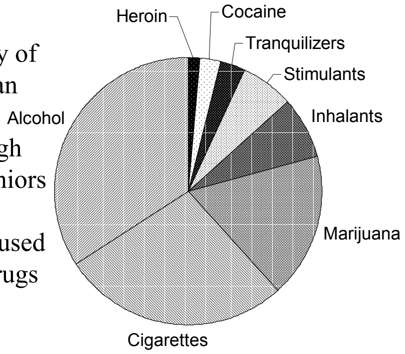
Democrats say voters in Palm Beach County are complaining about how names on the ballots were arranged.



http://www.bestbookmarks.com/election/

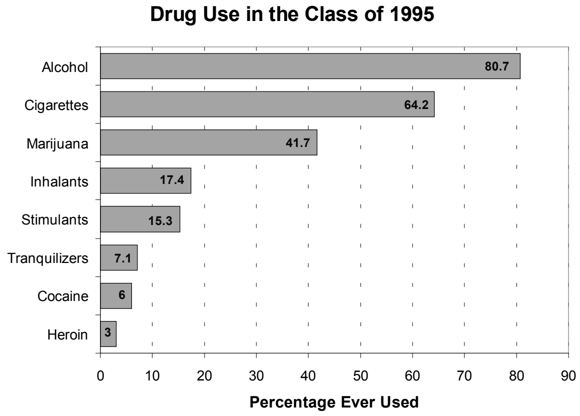


Drug Use in the Class of 1995

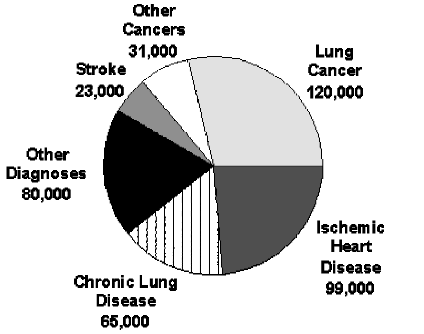


University of Michigan

% of High School Seniors who have ever used various drugs



Deaths Attributable to Cigarette Smoking — United States, 1990

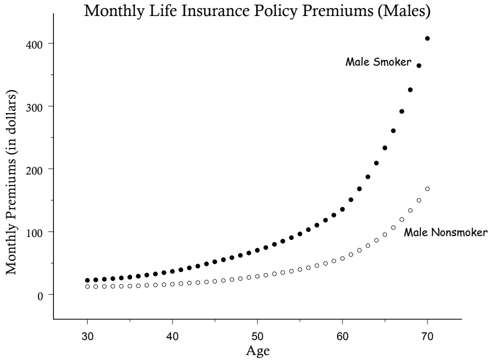


Source: CDC SAMMEC, MMWR 1993; 42:645-9.

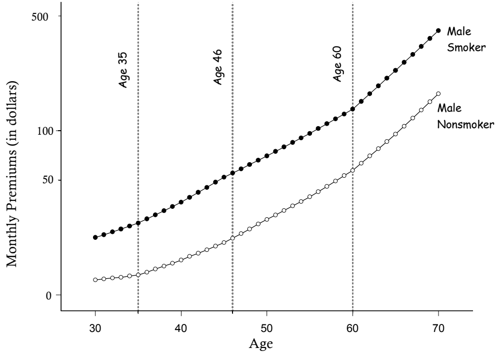
To Automation and ARC

THE POWER OF TRANSFORMATIONS

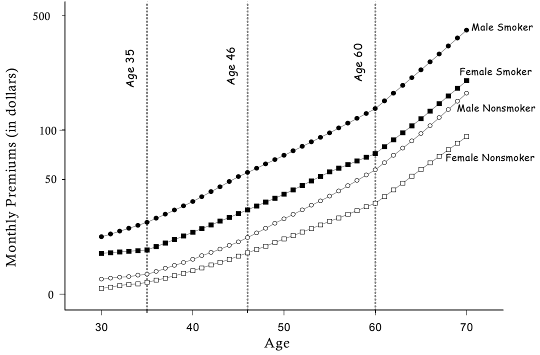
Jackson National's 10-year Level-Term $100,000 Policy
Monthly Life Insurance Policy Premiums (Males)



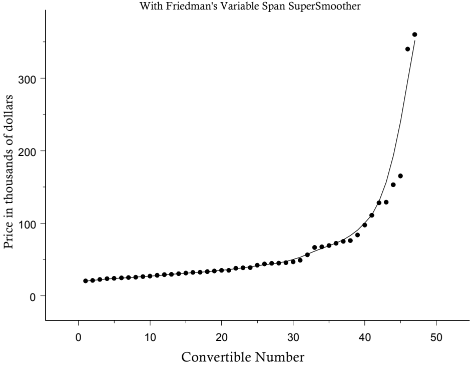
Insurance Premiums for Males as a function of age
and smoking classification (spaced logarithmically)



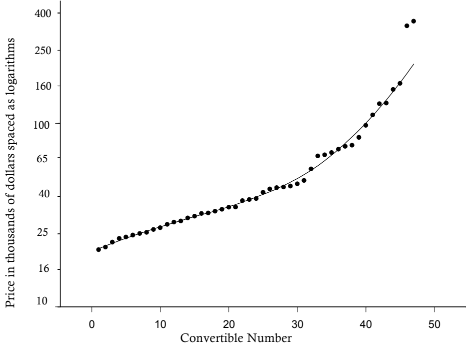
Insurance Premiums for $100,000 policies by gender,
age and smoking classification (spaced logarithmically)



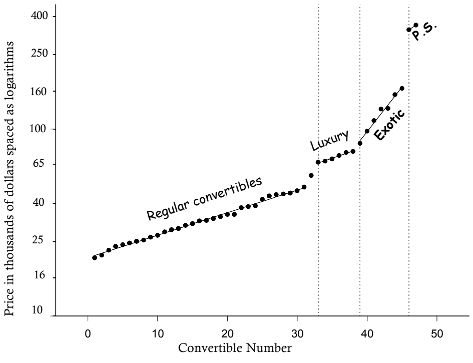
Forty-seven 2001 convertibles as a function of their price
With Friedman's Variable Span SuperSmoother



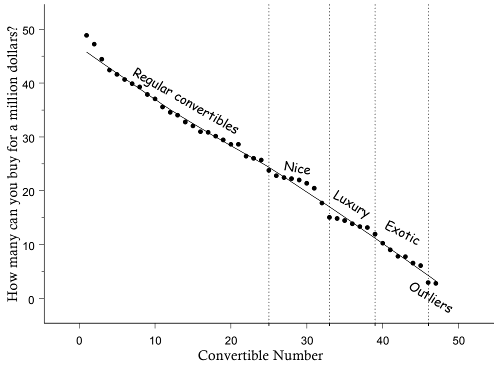
MSRP for 2001 Convertibles (with Lowess Smooth)
Taking logs doesn't always linearize



Taking Logs Emphasizes Linear Price Tiers of 2001 Convertibles
Based on Wainer, H. (2001) Chance, 14.3, p. 44



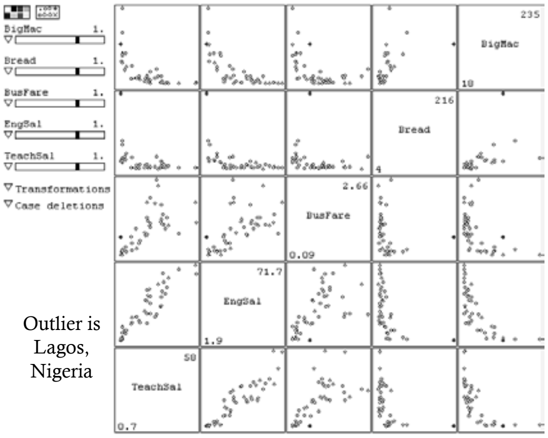
An Inverse Transformation provides an understandable metric, yields a good linear fit throughout and suggests other groupings



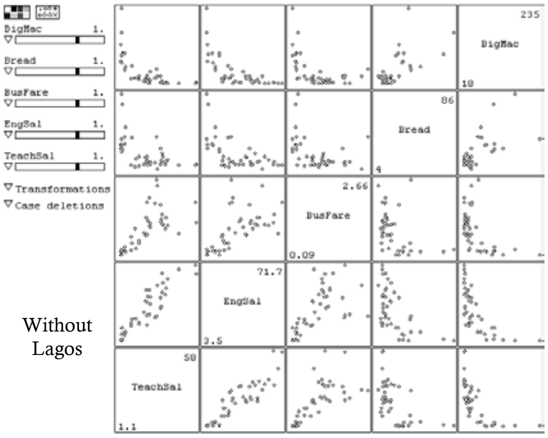
To More Graphical Arrays

Transformation, Scatterplot Matrices & Automation

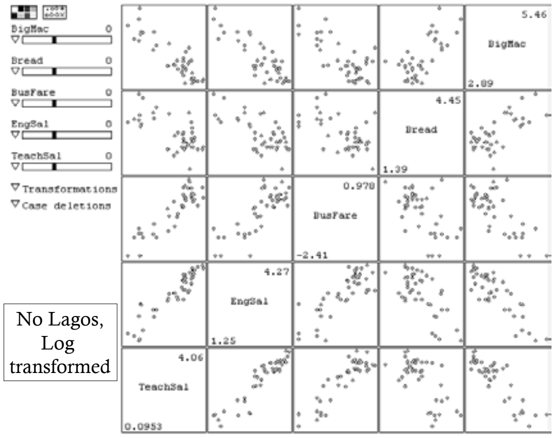
http://www.stat.umn.edu/arc/



Outlier is Lagos, Nigeria

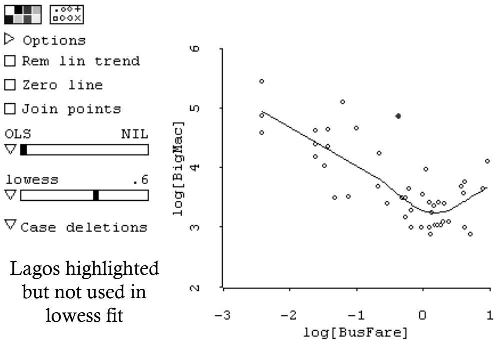


Without Lagos

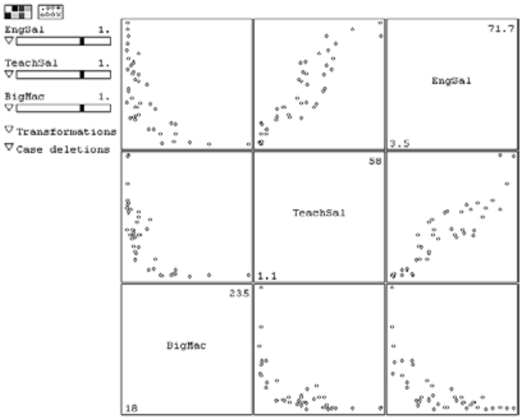


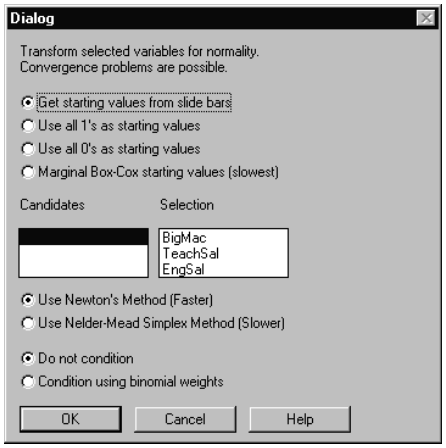
No Lagos, Log transformed

Zooming in on one plot



Lagos highlighted but not used in lowess fit




Dialog

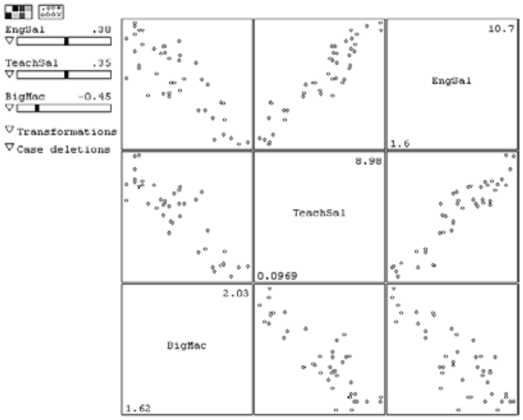
Transform selected variables for normality.
Convergence problems are possible.

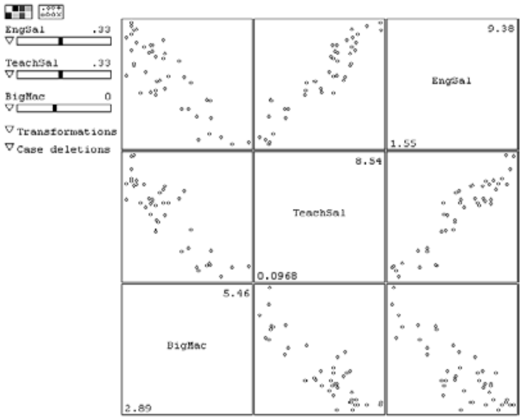
● Get starting values from slide bars
○ Use all 1's as starting values
○ Use all 0's as starting values
○ Marginal Box-Cox starting values (slowest)

Candidates	Selection
	BigMac TeachSal EngSal

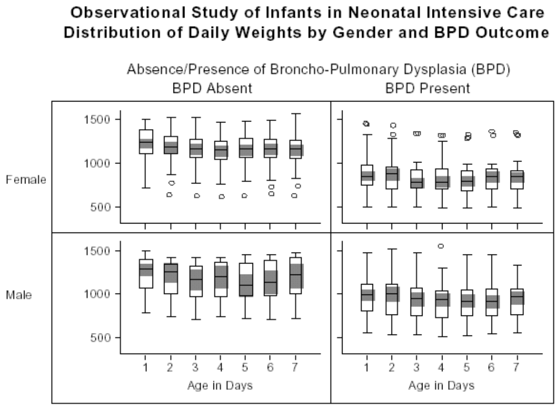
● Use Newton's Method (Faster)
○ Use Nelder-Mead Simplex Method (Slower)

● Do not condition
○ Condition using binomial weights





More Graphical Arrays

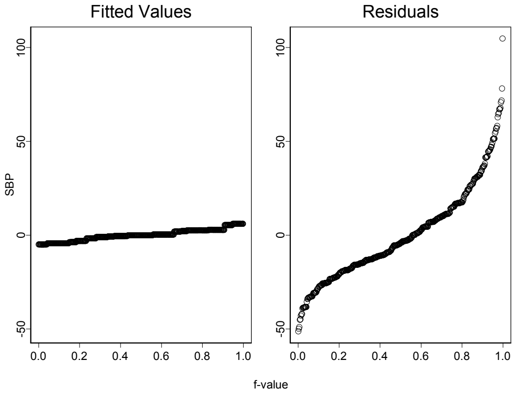

Observational Study of Infants in Neonatal Intensive Care
Distribution of Daily Weights by Gender and BPD Outcome

http://www.belmont.com/belweb2/software/cg/cg_examp/boxplot.pdf

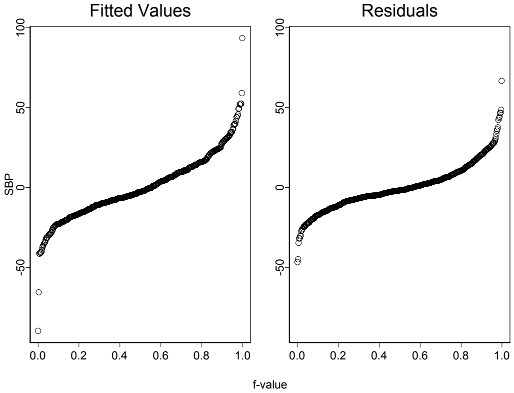
Residual-Fit Spread Plots

- Fitted (predicted) values and residuals each have a distribution.
- An r-f spread plot compares the spreads of the residuals and the fitted values.
- Graphical analog to R^2 statistic.

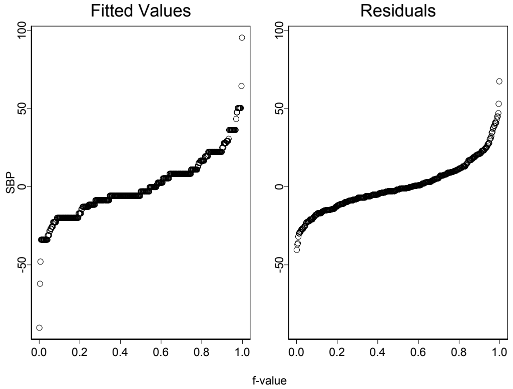
Dominican HTN Predicting SBP from Gender and Village

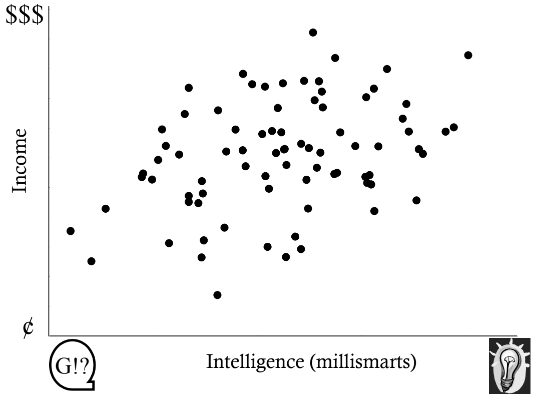


Regression Model for Dominican Hypertension Study Data



Regression model SBP = 14.54 + 1.40 DBP


Fitted Values
Residuals
SBP
f-value


$$$
Income
¢
G!?
Intelligence (millismarts)