INTERMEDIATE QUANTITATIVE METHODS

Avoiding Pitfalls in Data Analysis

Per Engzell (Nuffield College)

Hilary term 2018

Sociology Department, University of Oxford

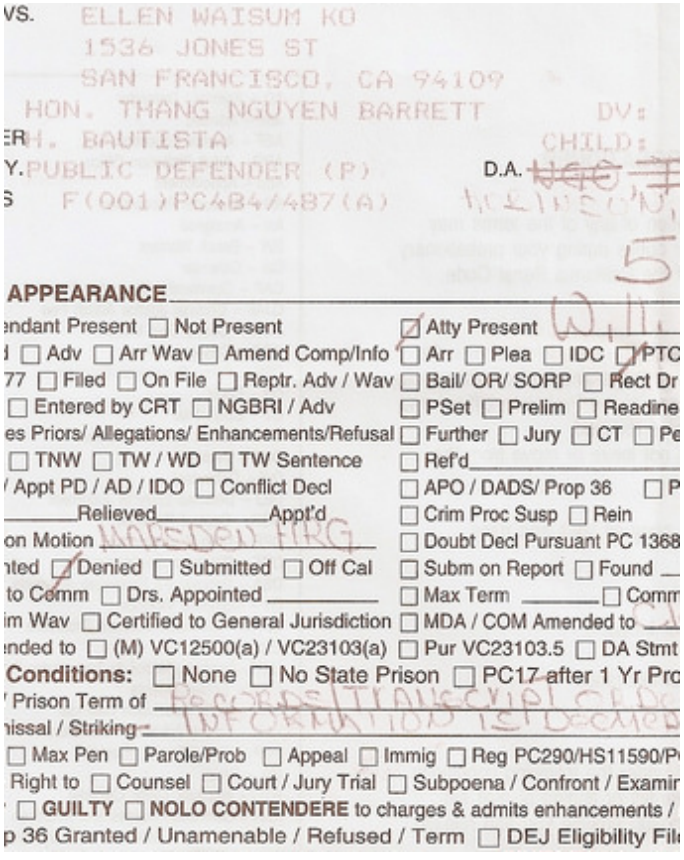
PRELIMINARIES

PRELIMINARIES

- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- If at first you succeed, try again
- Documentation, documentation

- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- If at first you succeed, try again
- Documentation, documentation

How did we get from this ...


VS. ELLEN WAISUM KO
1536 JONES ST
SAN FRANCISCO, CA 94109
HON. THANG NGUYEN BARRETT DV:
ER H. BAUTISTA CHILD:
Y. PUBLIC DEFENDER (P) D.A.
S F(001)PC484/487(A)

APPEARANCE

ndant Present ☐ Not Present ☑ Atty Present
☐ Adv ☐ Arr Wav ☐ Amend Comp/Info ☐ Arr ☐ Plea ☐ IDC ☐ PTC
77 ☐ Filed ☐ On File ☐ Reptr. Adv / Wav ☐ Bail/ OR/ SORP ☐ Rect Dr
☐ Entered by CRT ☐ NGBRI / Adv ☐ PSet ☐ Prelim ☐ Readine
es Priors/ Allegations/ Enhancements/Refusal ☐ Further ☐ Jury ☐ CT ☐ Pe
☐ TNW ☐ TW / WD ☐ TW Sentence ☐ Ref'd
/ Appt PD / AD / IDO ☐ Conflict Decl ☐ APO / DADS/ Prop 36 ☐ P
Relieved Appt'd ☐ Crim Proc Susp ☐ Rein
on Motion ☐ Doubt Decl Pursuant PC 1368
nted ☑ Denied ☐ Submitted ☐ Off Cal ☐ Subm on Report ☐ Found
to Comm ☐ Drs. Appointed ☐ Max Term ☐ Comm
im Wav ☐ Certified to General Jurisdiction ☐ MDA / COM Amended to
nded to ☐ (M) VC12500(a) / VC23103(a) ☐ Pur VC23103.5 ☐ DA Stmt
Conditions: ☐ None ☐ No State Prison ☐ PC17 after 1 Yr Pro
Prison Term of
issal / Striking
☐ Max Pen ☐ Parole/Prob ☐ Appeal ☐ Immig ☐ Reg PC290/HS11590/P
Right to ☐ Counsel ☐ Court / Jury Trial ☐ Subpoena / Confront / Examir
☐ **GUILTY** ☐ **NOLO CONTENDERE** to charges & admits enhancements /
p 36 Granted / Unamenable / Refused / Term ☐ DEJ Eligibility Fil

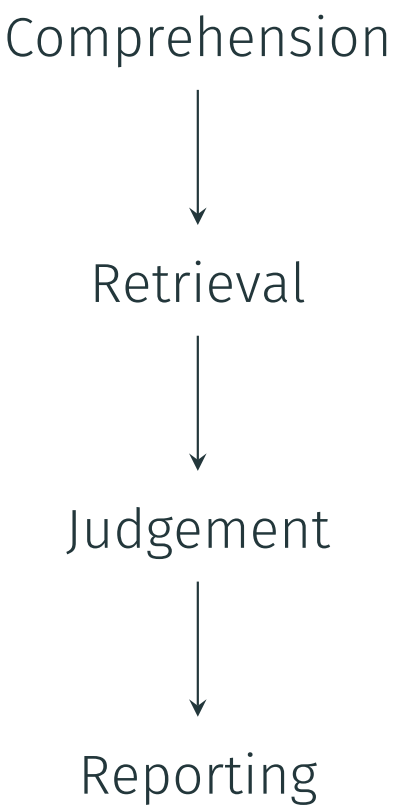
... to this?

	id	v3	v4	v5
1	2630	22	3	1930
2	8590	23	2	1745
3	7523	23	4	1700
4	8114	19	1	1730
5	9036	24	6	1000
6	2270	23	6	1010
7	8290	16	4	1000
8	1738	20	2	1600
9	7498	17	4	1130
10	11240	20	4	1845
11	3995	0	0	9900
12	1538	19	5	1715
13	5960	17	1	1300
14	11556	22	3	1755
15	7090	21	2	1800
16	10896	24	4	1930
17	3467	24	3	1800
18	7615	0	0	9900
19	4111	20	1	1700

Photo credit: `http://www.flickr.com/photos/elleko/7539221216` (CC-BY-2.0)

KNOW YOUR DATA

Your variables do not equal the concepts you want to measure!



To provide data, every respondent has to

- interpret the question (ideally in the same way as each other and the researcher)
- gather the information from memory
- make a judgement about how it corresponds to given alternatives
- formulate and communicate an answer

In each of these steps, mistakes and random variability are inevitable. Additional errors arise from vague response options, coarseness of categories, coding errors etc.

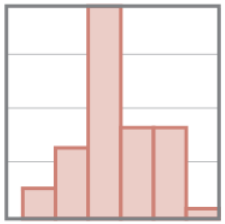
Adapted from: Robert M. Groves, et al. Survey Methodology (Wiley, 2011).

- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- Documentation, documentation
- If at first you succeed, try again

One Variable

sysuse auto, clear

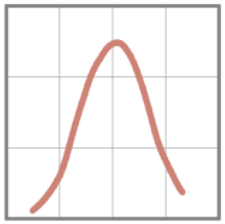
Continuous



histogram mpg, **width**(5) **freq** **kdensity** **kdenopts**(**bwidth**(5))

histogram

bin(#) • width(#) • density • fraction • frequency • percent • addlabels
addlabopts(<options>) • normal • normopts(<options>) • kdensity
kdenopts(<options>)



kdensity mpg, **bwidth**(3)

smoothed histogram

bwidth • kernel(<options>
normal • normopts(<line options>)

← main plot-specific options; see help for complete set

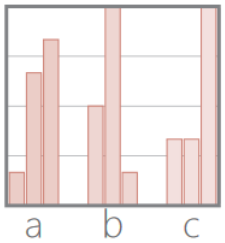
Discrete

graph bar (count), **over**(foreign, **gap**(*0.5)) **intensity**(*0.5)

bar plot

graph hbar draws horizontal bar charts

(asis) • (percent) • (count) • over(<variable>, <options: gap(*#) •
relabel • descending • reverse>) • cw •missing • nofill • allcategories •
percentages • stack • bargap(#) • intensity(*#) • yalternate • xalternate



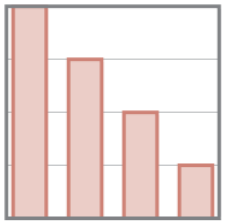
graph bar (percent), **over**(rep78) **over**(foreign)

grouped bar plot

graph hbar ...

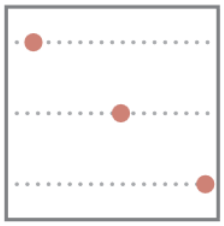
(asis) • (percent) • (count) • over(<variable>, <options: gap(*#) •
relabel • descending • reverse>) • cw •missing • nofill • allcategories •
percentages • stack • bargap(#) • intensity(*#) • yalternate • xalternate

Discrete X, Continuous Y



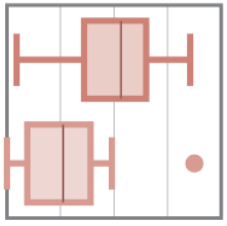
graph bar (median) price, **over**(foreign) graph hbar ...

bar plot (asis) • (percent) • (count) • (stat: mean median sum min max ...)
over(<variable>, <options: gap(*#) • relabel • descending • reverse
sort(<variable>)>) • cw • missing • nofill • allcategories • percentages
stack • bargap(#) • intensity(*#) • yalternate • xalternate



graph dot (mean) length headroom, **over**(foreign) **m**(1, ms(S))

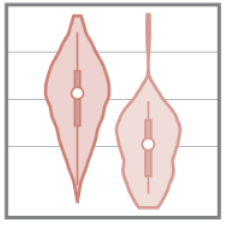
dot plot (asis) • (percent) • (count) • (stat: mean median sum min max ...)
over(<variable>, <options: gap(*#) • relabel • descending • reverse
sort(<variable>)>) • cw • missing • nofill • allcategories • percentages
linegap(#) • marker(#, <options>) • linetype(dot | line | rectangle)
dots(<options>) • lines(<options>) • rectangles(<options>) • rwidth



graph hbox mpg, **over**(rep78, descending) **by**(foreign) **missing**

box plot graph box draws vertical boxplots

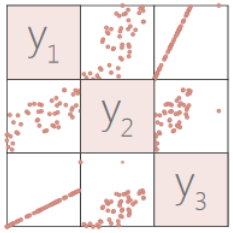
over(<variable>, <options: total • gap(*#) • relabel • descending • reverse
sort(<variable>)>) • missing • allcategories • intensity(*#) • boxgap(#)
medtype(line | line | marker) • medline(<options>) • medmarker(<options>)



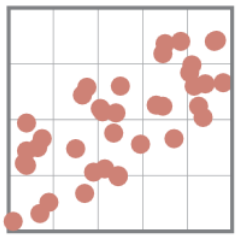
vioplot price, **over**(foreign) ssc install vioplot

violin plot over(<variable>, <options: total • missing>)>) • nofill •
vertical • horizontal • obs • kernel(<options>) • bwidth(#) •
barwidth(#) • dscale(#) • ygap(#) • ogap(#) • density(<options>)
bar(<options>) • median(<options>) • obsopts(<options>)

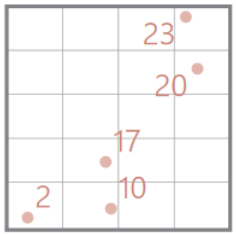
Two+ Continuous Variables



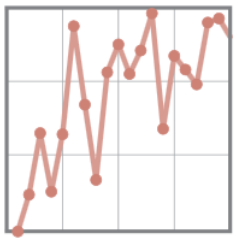
graph matrix mpg price weight**, half**
scatter plot of each combination of variables
half • jitter(#) • jitterseed(#)
diagonal • [aweights(<variable>)]



twoway scatter mpg weight**, jitter(7)**
scatter plot
jitter(#) • jitterseed(#) • sort • cmissing(yes | no)
connect(<options>) • [aweight(<variable>)]

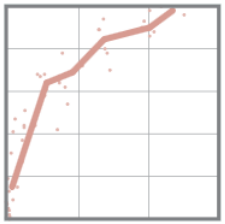


twoway scatter mpg weight**, mlabel(**mpg**)**
scatter plot with labelled values
jitter(#) • jitterseed(#) • sort • cmissing(yes | no)
connect(<options>) • [aweight(<variable>)]



twoway connected mpg price**, sort(**price**)**
scatter plot with connected lines and symbols
jitter(#) • jitterseed(#) • sort
connect(<options>) • cmissing(yes | no)
see also **line**

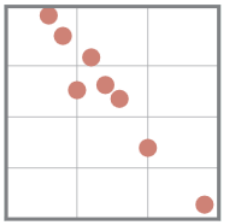
Summary Plots



twoway mband mpg weight **|| scatter** mpg weight

plot median of the y values

bands(#)

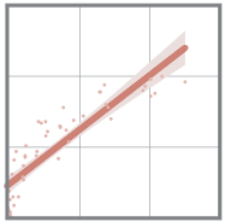


binscatter weight mpg, **line(**none**)** ssc install binscatter

plot a single value (mean or median) for each x value

medians • nquantiles(#) • discrete • controls(<variables>) • linetype(lfit | qfit | connect | none) • aweight[<variable>]

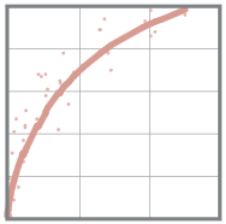
Fitting Results



twoway lfitci mpg weight **|| scatter** mpg weight

calculate and plot linear fit to data with confidence intervals

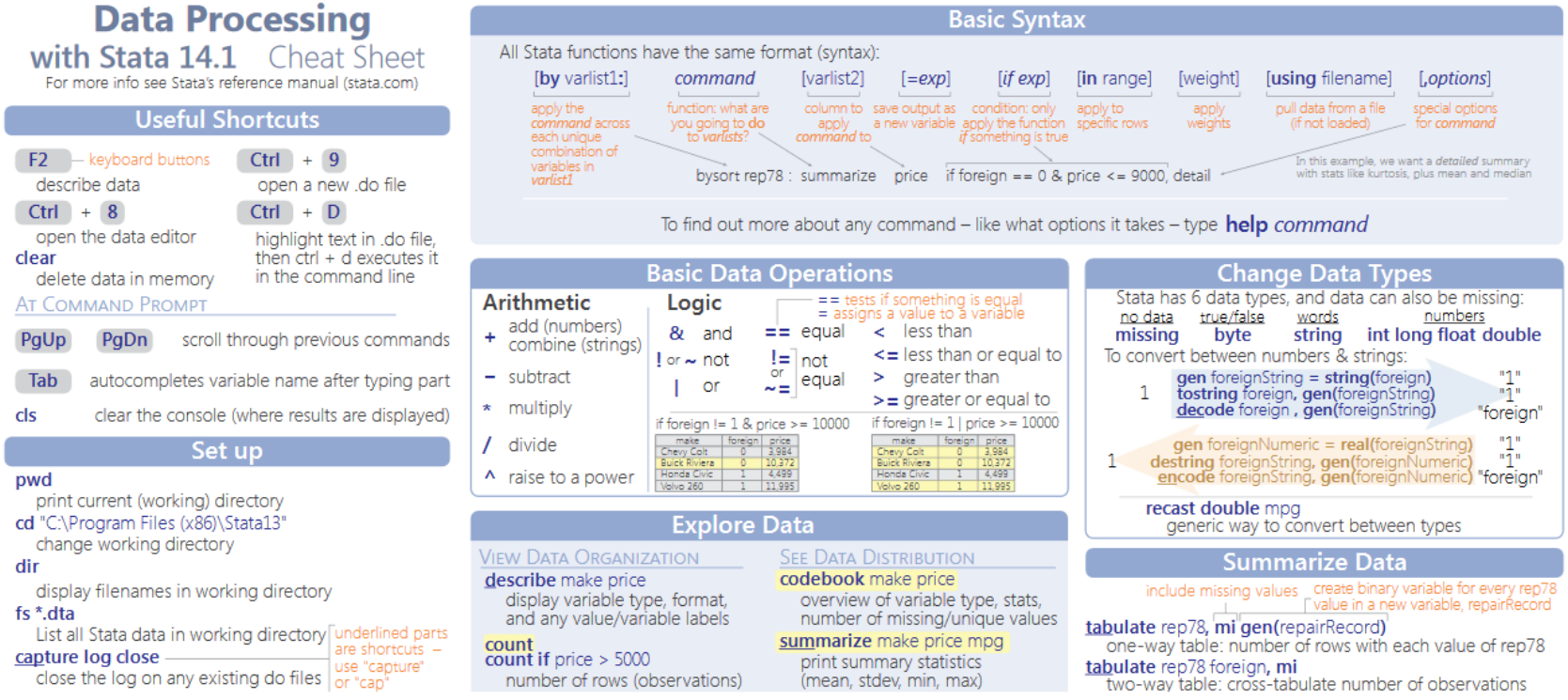
level(#) • stdp • stdf • nofit • fitplot(<plottype>) • ciplot(<plottype>) • range(# #) • n(#) • atobs • estopts(<options>) • predopts(<options>)



twoway lowess mpg weight **|| scatter** mpg weight

calculate and plot lowess smoothing

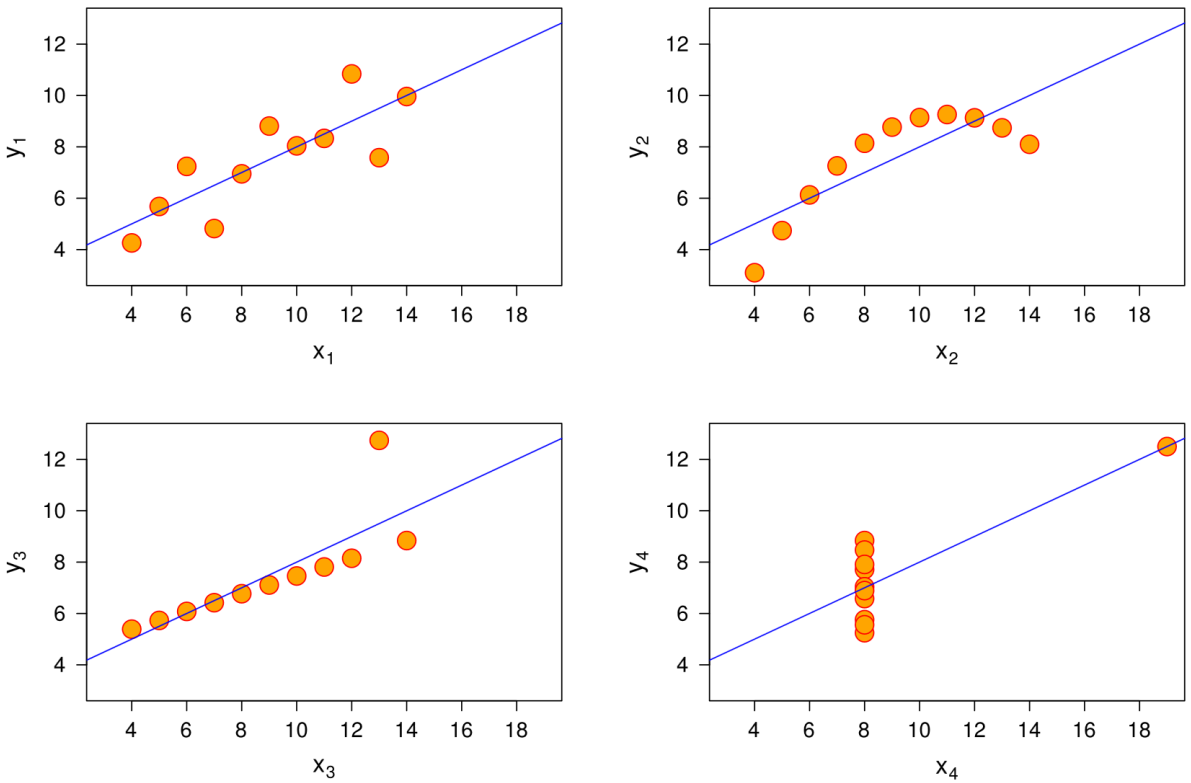
bwidth(#) • mean • noweight • logit • adjust



More excellent cheat sheets like these at
`http://geocenter.github.io/StataTraining/`

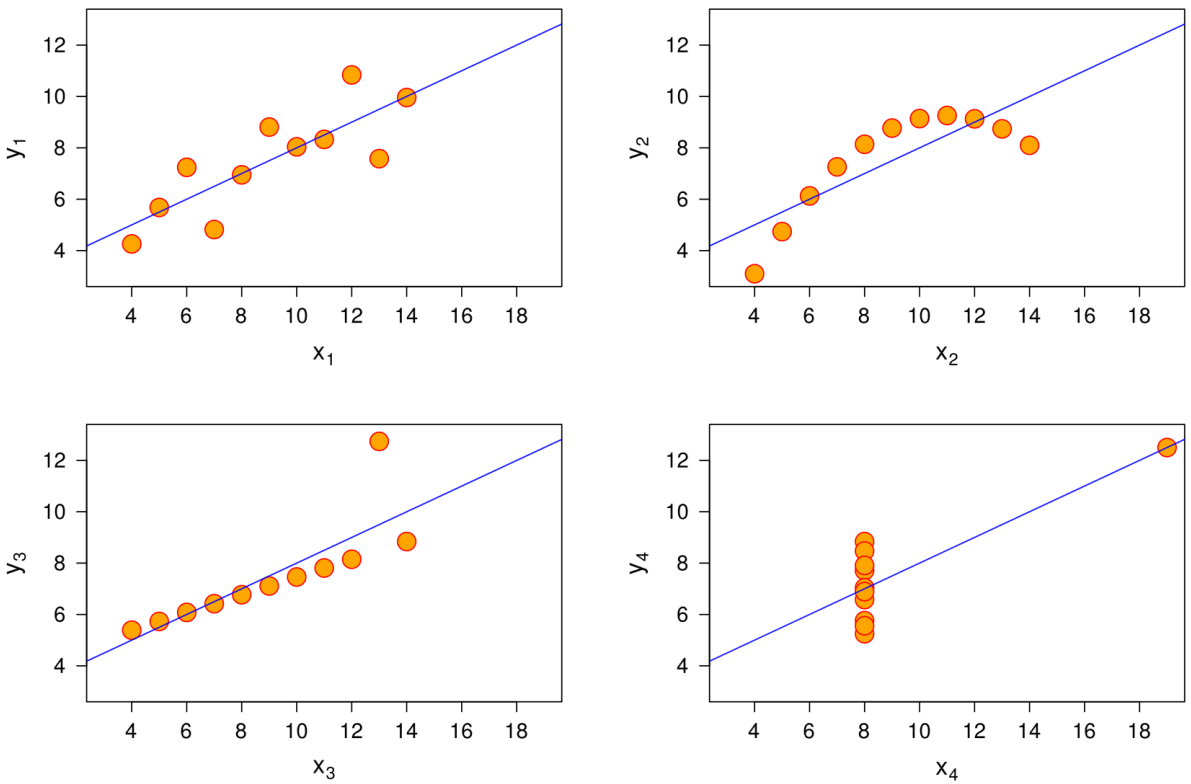
Also, always remember to use Stata help files, e.g. typing
`<help twoway>` into the command line will get you a long way.

What do these distributions have in common?



Credit: `https://commons.wikimedia.org/wiki/File:Anscombe.svg` (CC-BY-SA-3.0)

What do these distributions have in common?



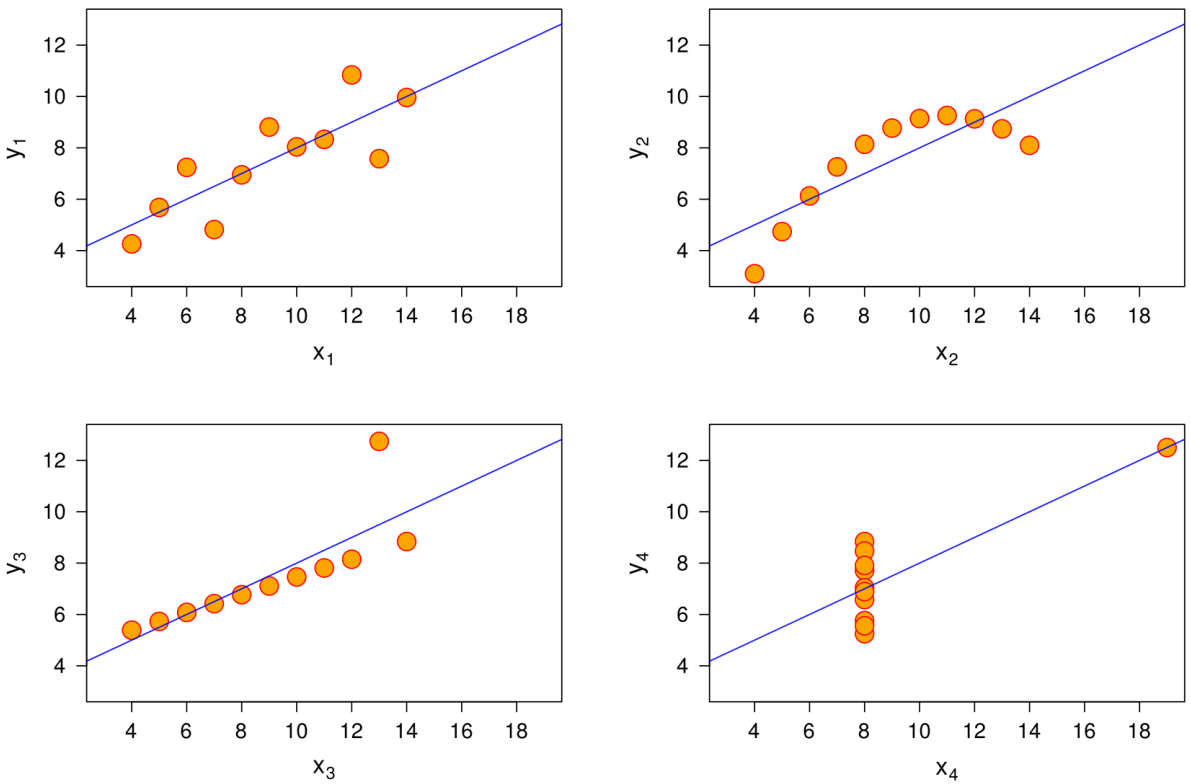
Mean of x	9.0
Variance of x	11.0
Std dev. of x	3.32
Mean of y	7.5
Variance of y	4.12
Std dev. of y	2.03
Covariance	5.493
Correlation	0.816

Regression line

$$y = 3 + 0.5x$$

Credit: `https://commons.wikimedia.org/wiki/File:Anscombe.svg` (CC-BY-SA-3.0)

What do these distributions have in common?



Mean of x	9.0
Variance of x	11.0
Std dev. of x	3.32
Mean of y	7.5
Variance of y	4.12
Std dev. of y	2.03
Covariance	5.493
Correlation	0.816

Regression line

$$y = 3 + 0.5x$$

Credit: `https://commons.wikimedia.org/wiki/File:Anscombe.svg` (CC-BY-SA-3.0)

That said, you will be surprised at how often the linear model gives a reasonable first approximation!

- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- Documentation, documentation
- If at first you succeed, try again

SIMPLICITY

”Can we just do this by OLS instead?”

Regardless of scale level, whenever relationships are approximately linear, ordinary least squares is always* a justifiable option

Even with nonlinear relationships, ordinary least squares has an intuitive interpretation as the sample-weighted average slope of the true regression function across the range of the variable

A special case of this is the linear probability model (LPM) which tells us the average marginal effect (i.e., the expected percentage point difference) for the probability of the outcome

Heteroskedasticity can be handled by robust standard errors (in Stata: **regress depvar indepvar, robust**)

*Except maybe if you are writing an exam on nonlinear models

For more on these points, see Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly Harmless Econometrics. Princeton Univ Press, 2008.

OVERFITTING

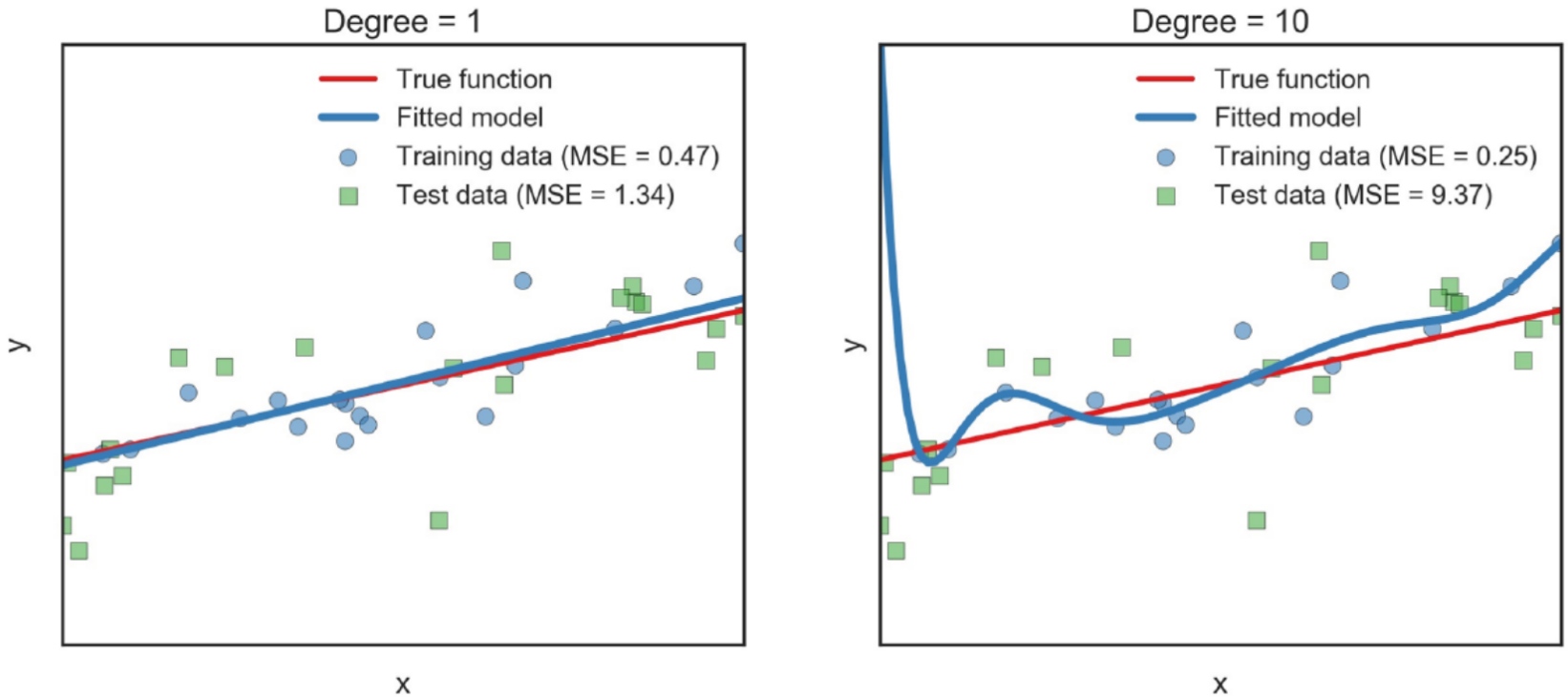


Fig. 1. Training and test error produced by fitting either a linear regression (left) or a 10th-order polynomial regression (right) when the true relationship in the population (red line) is linear. In both cases, the test data (green) deviate more from the model's predictions (blue line) than the training data (blue). However, the flexibility of the 10th-order polynomial model facilitates much greater overfitting, resulting in lower training error but much higher test error than the linear model. MSE = mean squared error.

Source: Yarkoni, Tal, and Jacob Westfall. "Choosing prediction over explanation in psychology: Lessons from machine learning." Perspectives on Psychological Science 12, 6 (2017): 1100-1122.

- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- Documentation, documentation
- If at first you succeed, try again

Often you will want to repeat the same task for a number of variables, values, commands, etc. For example, you may want to:

- Recode the same values in several different variables
- Run the same model with different outcomes/controls
- Do something separately by subgroups in your data

It is possible to do so by copy/pasting and substituting the relevant arguments in each new line. However, doing so tedious, verbose and error-prone.

This is where loops come in handy! Stata has a number of different looping commands (**foreach, forvalues, while**) but **foreach** is the most versatile.

Anatomy of a Loop

see also **while**

Stata has three options for repeating commands over lists or values: **foreach**, **forvalues**, and **while**. Though each has a different first line, the syntax is consistent:

Credit: `http://geocenter.github.io/StataTraining/` (CC-BY-4.0)

Two great user written Stata commands for publication-quality output

- **outreg2** takes your regression output and exports it to a formatted table in your preferred file format (e.g., MS Word)
- **coefplot** takes your regression output and plots point estimates and confidence intervals as a Stata graph

- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- Documentation, documentation
- If at first you succeed, try again

DOCUMENTATION

Well documented and human readable code allows you to

- Ensure replicability
- Search for and identify mistakes
- Redo an analysis with minor changes
- Refashion code for another purpose
- Bring on collaborators
- Pick up your own past projects

Last point is important: the person you are most likely to share your data and code with is your future self!

Readable code is not all about commenting but also about using understandable variable names, structure and spacing of type, etc. (Don't just document your code – code your documentation!)

Poor code

```
use "C:\Users\peen5137\Box Sync\Books in the home\corr.dta", clear
ren agr k
replace item = item + ((coef-1)*5)

egen median = median(k), by(item)
egen upq = pctile(k), p(75) by(item)
egen loq = pctile(k), p(25) by(item)
egen iqr = iqr(k), by(item)
egen upper = max(min(k, upq + 1.5 * iqr)), by(item)
egen lower = min(max(k, loq - 1.5 * iqr)), by(item)

cap drop jitter
set seed 3753
gen jitter = item - .5/2 + runiform()/2

twoway rbar med upq item if item==1, pstyle(p1) blc(gs10) bfc(gs14) barw(0.75)
 || rbar med loq item if item==1, pstyle(p1) blc(gs10) bfc(gs14) barw(0.75) //
 || rbar med upq item if item==2, pstyle(p1) blc(gs10) bfc(white) barw(0.75) /
 || rbar med loq item if item==2, pstyle(p1) blc(gs10) bfc(white) barw(0.75) /
 || rbar med upq item if item==3, pstyle(p1) blc(gs10) bfc(gs7) barw(0.75) ///
 || rbar med loq item if item==3, pstyle(p1) blc(gs10) bfc(gs7) barw(0.75) ///
 || rbar med upq item if item==4, pstyle(p1) blc(gs10) bfc(gs11) barw(0.75) //
 || rbar med loq item if item==4, pstyle(p1) blc(gs10) bfc(gs11) barw(0.75) //
 || rbar med upq item if item==6, pstyle(p1) blc(gs10) bfc(gs14) barw(0.75) //
 || rbar med loq item if item==6, pstyle(p1) blc(gs10) bfc(gs14) barw(0.75) //
 || rbar med upq item if item==7, pstyle(p1) blc(gs10) bfc(white) barw(0.75) /
 || rbar med loq item if item==7, pstyle(p1) blc(gs10) bfc(white) barw(0.75) /
 || rbar med upq item if item==8, pstyle(p1) blc(gs10) bfc(gs7) barw(0.75) ///
 || rbar med loq item if item==8, pstyle(p1) blc(gs10) bfc(gs7) barw(0.75) ///
 || rbar med upq item if item==9, pstyle(p1) blc(gs10) bfc(gs11) barw(0.75) //
 || rbar med loq item if item==9, pstyle(p1) blc(gs10) bfc(gs11) barw(0.75) //
 || rspike upq upper item, pstyle(p1) lc(gs10) ///
 || rspike loq lower item, pstyle(p1) lc(gs10) ///
 || rcap upper upper item, pstyle(p1) lc(gs10) msize(*2) ///
 || rcap lower lower item, pstyle(p1) lc(gs10) msize(*2) ///
 || scatter k jitter, ms(Oh) fxsize(105) fysize(100)  ///
 || pci .66 .5 .66 4.5, lp(dash) lw(medthin) || pci .53 .5 .53 4.5, lp(dash) l
    legend(order(1 "Parent Books" "vs. Parent" "Child Books" 3 "Student Books"
```

- No explanatory remarks
- Lines extend beyond margin
- Noninformative variable names
- Lots of repetition
- No indentation
- No working folder

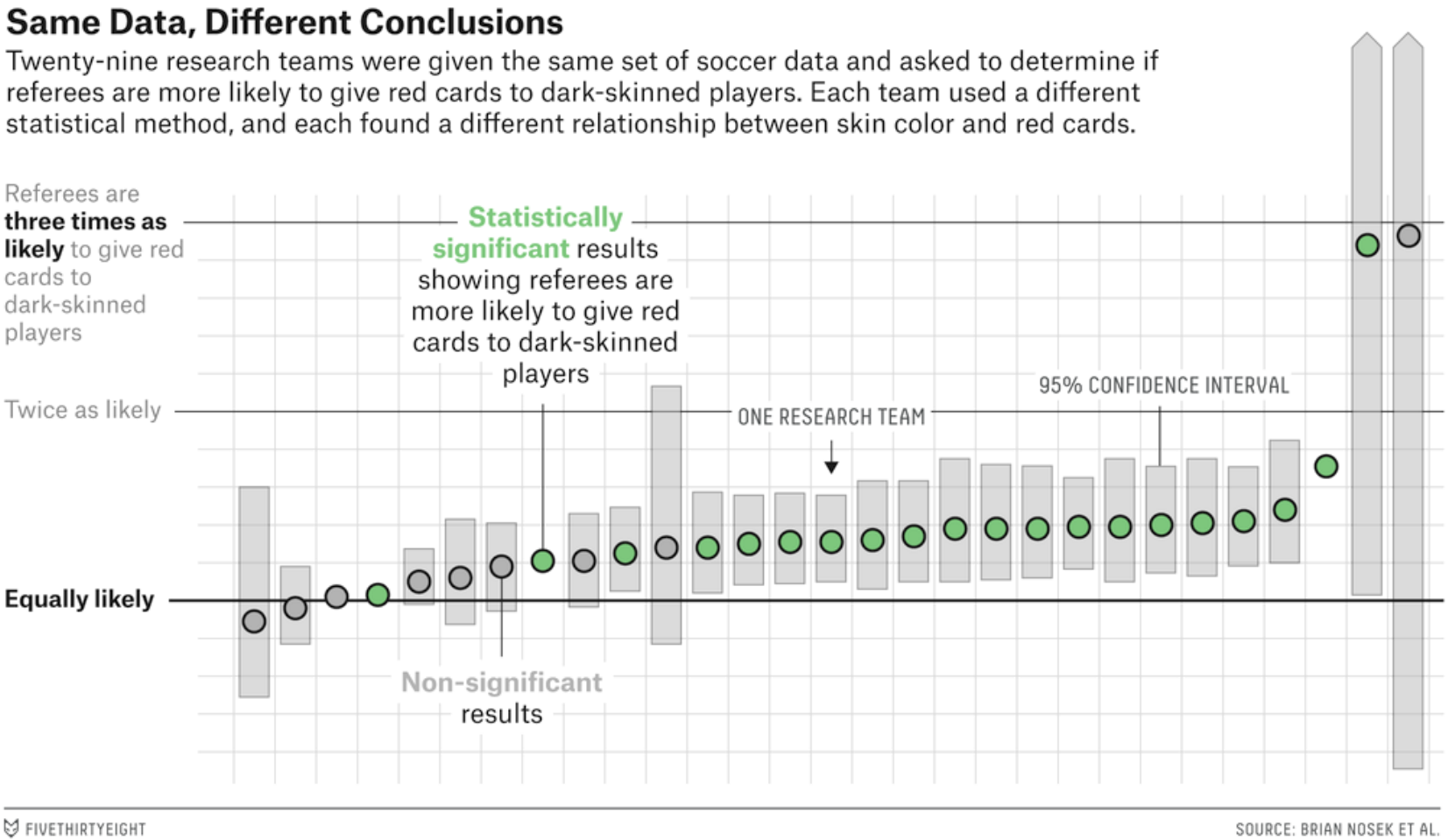
Better code

```
**************************************
*** Figure 3: Differences by gender ***
**************************************

use "booksdata.dta", clear
matrix drop _all
scalar drop _all
set more off

* Create matrix to store results *
set matsize 800
mat r = [ .,.,. ]

* Ordered logit of each measure regressed on gender *
distinct cntid
scalar npirls = r(ndistinct)
levelsof cntid, local(cntid)

foreach c of local cntid {
    di ""
    di "cntid: `c'"
foreach var in bookspa chbkspa booksch educpa occupa {
    qui ologit `var' female if cntid == `c', robust cluster(cluster)
    mat r = [r \ `c', exp(_b[female]), e(p)]
    }
    }

* Gender analyses of PISA 2012 data
* This requires first downloading the student file "INT_STU12_DEC03.zip"
* from http://pisa2012.acer.edu.au/downloads/ and converting to Stata format
cd "$ddir/PISA 2012/"
use "INT_STU12_DEC03.dta", clear
set more off

* Create/rename variables *
gen female = (ST04Q01==1)
ren ST28Q01 booksch
egen cluster = group(CNT SCHOOLID)
```

- Clearly commented
- Lines stay within margin
- Memorable variable names
- Loops do away with repetition
- Uses indentation (see loops)
- Uses working folder

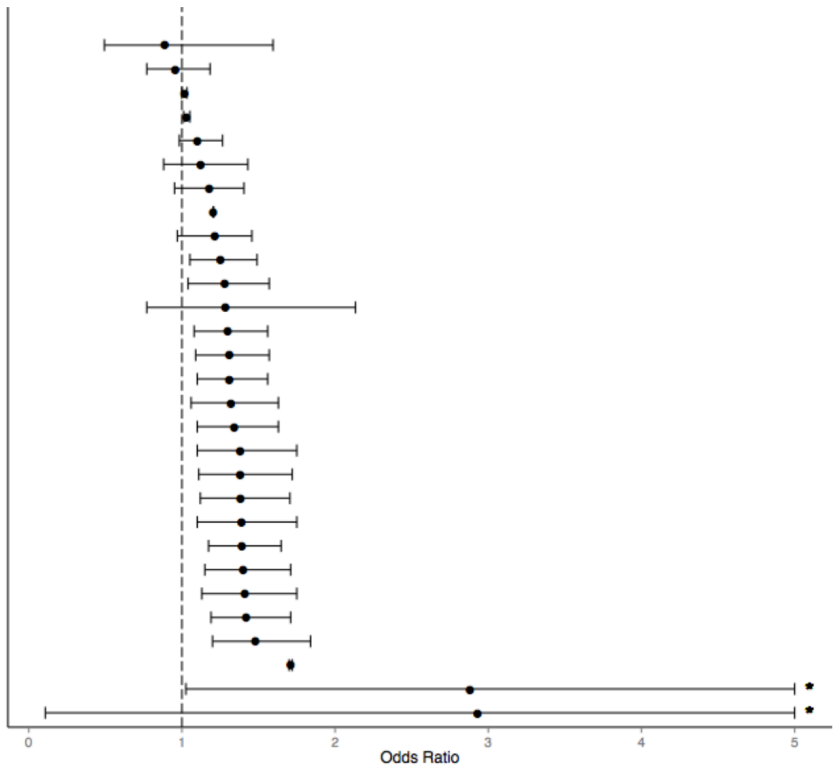
- Know your data
- Visualise early, visualise often
- Do the simple before the complex
- Automate routine tasks
- Documentation, documentation
- If at first you succeed, try again

Silberzahn, et al. (2017, in press) “Many analysts, one dataset: Making transparent how variations in analytical choices affect results”.



Silberzahn, et al. (2017, in press) “Many analysts, one dataset: Making transparent how variations in analytical choices affect results”.

Team	Analytic Approach	OR
12	Zero–inflated Poisson regression	0.89
17	Bayesian logistic regression	0.96
15	Hierarchical log–linear modeling	1.02
10	Multilevel regression and logistic regression	1.03
18	Hierarchical Bayes model	1.10
31	Logistic regression	1.12
1	Ordinary least squares with robust standard errors, logistic regression	1.18
4	Spearman correlation	1.21
14	Weighted least squares regression with referee fixed–effects and clustered standard errors	1.21
11	Multiple linear regression	1.25
30	Clustered robust binomial logistic regression	1.28
6	Linear Probability Model	1.28
26	Three–level hierarchical generalized linear modeling with Poisson sampling	1.30
3	Multilevel Binomial Logistic Regression using bayesian inference	1.31
23	Mixed model logistic regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear probability model, logistic regression	1.34
5	Generalized linear mixed models	1.38
24	Multilevel logistic regression	1.38
28	Mixed effects logistic regression	1.38
32	Generalized linear models for binary data	1.39
8	Negative binomial regression with a log link analysis	1.39
20	Cross–classified multilevel negative binomial model	1.40
13	Poisson Multi–level modeling	1.41
25	Multilevel logistic binomial regression	1.42
9	Generalized linear mixed effects models with a logit link function	1.48
7	Dirichlet process Bayesian clustering	1.71
21	Tobit regression	2.88
27	Poisson regression	2.93



Silberzahn, et al. (2017, in press) "Many analysts, one dataset: Making transparent how variations in analytical choices affect results".

Team	Analytic Approach	N covariates	Treatment of Non-Independence	Distribution	Reported Effect Size				Odds Ratio (OR)		
					Unit	Size	95% CI		OR	95% CI	
10	Multilevel regression and logistic regression	3	Variance component	Linear	R	0.01	0.00	0.01	1.03	1.01	1.05
1	Ordinary least squares with robust standard errors, logistic regression	7	Clustered SE	Linear	OR	1.18	0.95	1.41	1.18	0.95	1.41
4	Spearman correlation	3	None	Linear	D	0.10	0.10	0.10	1.21	1.20	1.21
14	Weighted least squares regression with referee fixed-effects and clustered SE	6	Clustered SE	Linear	OR	1.21	0.97	1.46	1.21	0.97	1.46
11	Multiple linear regression	4	None	Linear	D	0.12	0.03	0.22	1.25	1.05	1.49
6	Linear Probability Model	6	Clustered SE	Linear	OR	1.28	0.77	2.13	1.28	0.77	2.13
17	Bayesian logistic regression	2	Variance component	Logistic	OR	0.96	0.77	1.18	0.96	0.77	1.18
15	Hierarchical log-linear modeling	1	None	Logistic	OR	1.02	1.00	1.03	1.02	1.00	1.03
31	Logistic regression	6	Clustered SE	Logistic	OR	1.12	0.88	1.43	1.12	0.88	1.43
30	Clustered robust binomial logistic regression	3	Clustered SE	Logistic	OR	1.28	1.04	1.57	1.28	1.04	1.57
3	Multilevel Binomial Logistic Regression using Bayesian inference	2	Variance component	Logistic	OR	1.31	1.09	1.57	1.31	1.09	1.57
23	Mixed model logistic regression	2	Variance component	Logistic	OR	1.31	1.10	1.56	1.31	1.10	1.56
2	Linear probability model, logistic regression	6	Clustered SE	Logistic	OR	1.34	1.10	1.63	1.34	1.10	1.63
5	Generalized linear mixed models	0	Variance component	Logistic	OR	1.38	1.10	1.75	1.38	1.10	1.75
24	Multilevel logistic regression	3	Variance component	Logistic	OR	1.38	1.11	1.72	1.38	1.11	1.72
28	Mixed effects logistic regression	2	Variance component	Logistic	OR	1.38	1.12	1.71	1.38	1.12	1.71
32	Generalized linear models for binary data	1	Clustered SE	Logistic	OR	1.39	1.10	1.75	1.39	1.10	1.75
8	Negative binomial regression with a log link analysis	0	None	Logistic	OR	1.39	1.17	1.65	1.39	1.17	1.65
25	Multilevel logistic binomial regression	4	Variance component	Logistic	OR	1.42	1.19	1.71	1.42	1.19	1.71
9	Generalized linear mixed effects models with a logit link function	2	Variance component	Logistic	OR	1.48	1.20	1.84	1.48	1.20	1.84
7	Dirichlet process Bayesian clustering	0	None	Miscellaneous	OR	1.71	1.70	1.72	1.71	1.70	1.72
21	Tobit regression	4	Clustered SE	Miscellaneous	R	0.28	0.01	0.56	2.88	1.03	11.47
12	Zero-inflated Poisson regression	2	Fixed effect	Poisson	IRR	0.89	0.49	1.60	0.89	0.49	1.60
26	Three-level hierarchical generalized linear modeling with Poisson sampling	6	Variance component	Poisson	IRR	1.30	1.08	1.56	1.30	1.08	1.56
16	Hierarchical Poisson Regression	2	Variance component	Poisson	IRR	1.32	1.06	1.63	1.32	1.06	1.63
20	Cross-classified multilevel negative binomial model	1	Variance component	Poisson	IRR	1.40	1.15	1.71	1.40	1.15	1.71
13	Poisson Multi-level modeling	1	Variance component	Poisson	IRR	1.41	1.13	1.75	1.41	1.13	1.75
27	Poisson regression	1	None	Poisson	IRR	2.93	0.11	78.66	2.93	0.11	78.66
32	Generalized linear models for binary data	1	Clustered SE	Logistic	OR	1.39	1.10	1.75	1.39	1.10	1.75

MODEL FLEXIBILITY

Lenz and Sahn (2017) “Achieving statistical significance with covariates”.

- Use replication data from American Journal of Political Science (AJPS) 2013-2015 (N=63 articles)
- Find that in 40% of studies, statistical significance depended on covariate adjustment
- Statistical significance shifted not primarily due to increased precision, but increased effect sizes
- Articles where this is the case were less likely to report a baseline specification

Lenz and Sahn (2017) “Achieving statistical significance with covariates”.

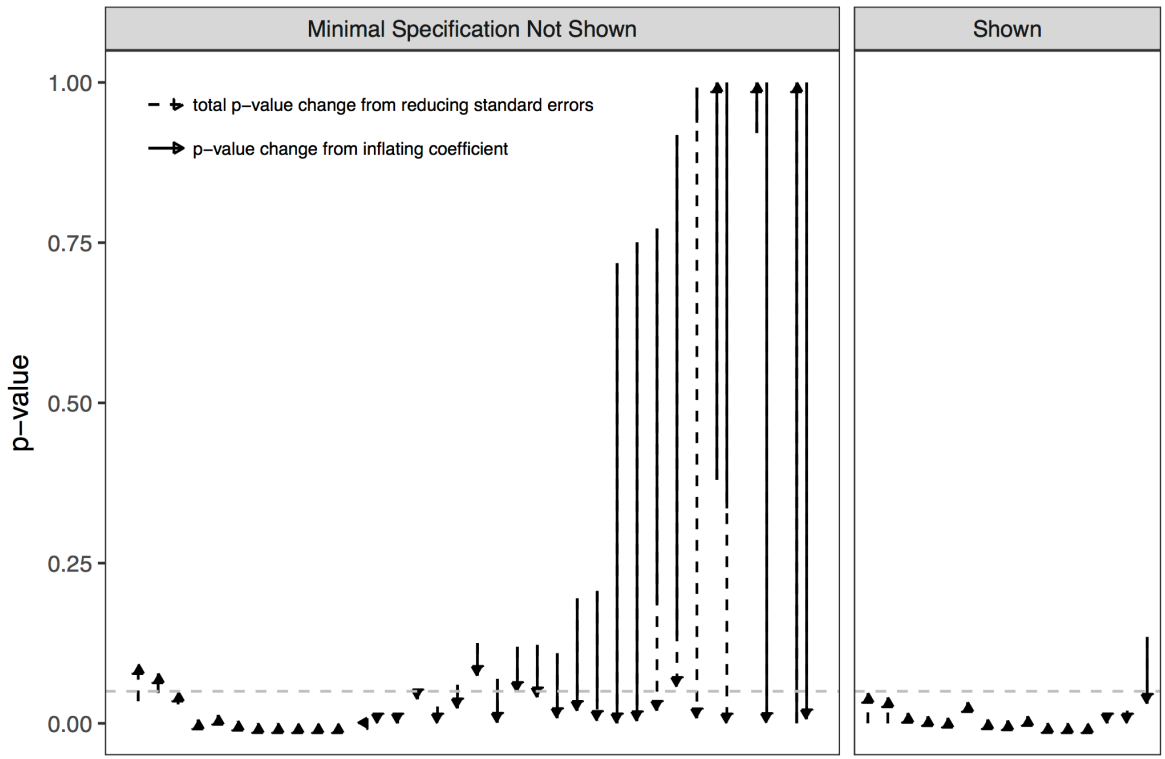


Figure 2: P-Value Changes in Observational Studies. For each article, the arrows show the total p-value changes from the minimal to the full specification. The solid part of the arrows shows the p-value changes from only coefficient estimate changes, while the dotted part shows p-value changes from standard error changes. The figure shows that, when articles failed to present a minimal specification, they often achieve statistical significance (lower their p-values) by including covariates.

QUESTIONABLE RESEARCH PRACTICES (QRP:S)

- Failing to report all of a study's dependent measures
- Deciding whether to collect more data after looking at results
- Failing to report all of a study's conditions
- Stopping collecting data earlier than planned because one found the result that one had been looking for
- Downward rounding of p-values (e.g., reporting p=.054 as "p<.05")
- In a paper, selectively reporting studies that "worked"
- Deciding whether to exclude data after looking at impact on results
- Reporting an unexpected finding as having been predicted from start
- Claiming results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)
- Falsifying data

Source: John et al. (2012). "Measuring the prevalence of questionable research practices with incentives for truth telling". Psychological science,

NONREPRODUCIBLE RESEARCH

- Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance: Carney, Cuddy & Yap – Psychological Science, 2010
- Big and Tall Parents/Violent Men/Engineers Have More Sons, Nurses/Beautiful Parents Have More Daughters: Kanazawa (various studies)
- Changes in Women's Choice of Dress Across the Ovulatory Cycle: Durante et al – Personality and Soc Psych, 2008
- Voting and the Ovulatory Cycle: Durante et al – Psychological Science, 2013
- Female Hurricanes Deadlier Than Male Hurricanes: Jung et al – PNAS, 2014
- First Class Seating on Airplanes Predicts Air Rage: Decelles & Norton – PNAS, 2016
- Spread of Obesity in a Social Network: Christakis & Fowler – N Engl J Med, 2007
- Labor Market Returns to Childhood Interventions: Heckman et al (various studies)
- Elderly Word Priming and Walking Speed: Bargh et al (various studies)
- Extrasensory Perception: Daryl Bem (various studies)

Questions?