INTERMEDIATE QUANTITATIVE METHODS

Avoiding Pitfalls in Regression Analysis

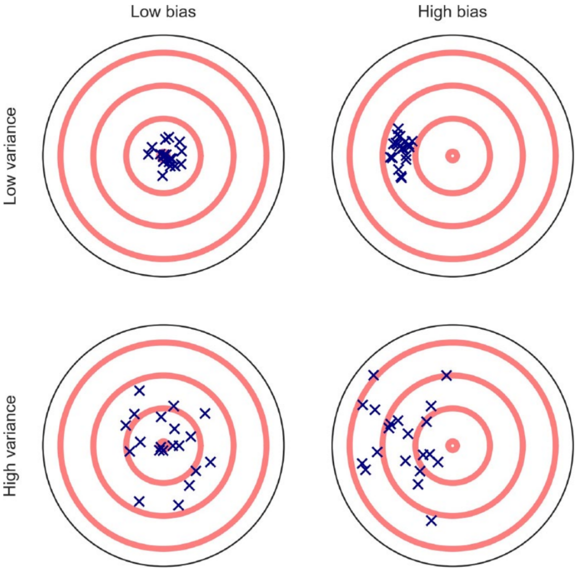
Per Engzell (Nuffield College)

Hilary term 2018

Sociology Department, University of Oxford

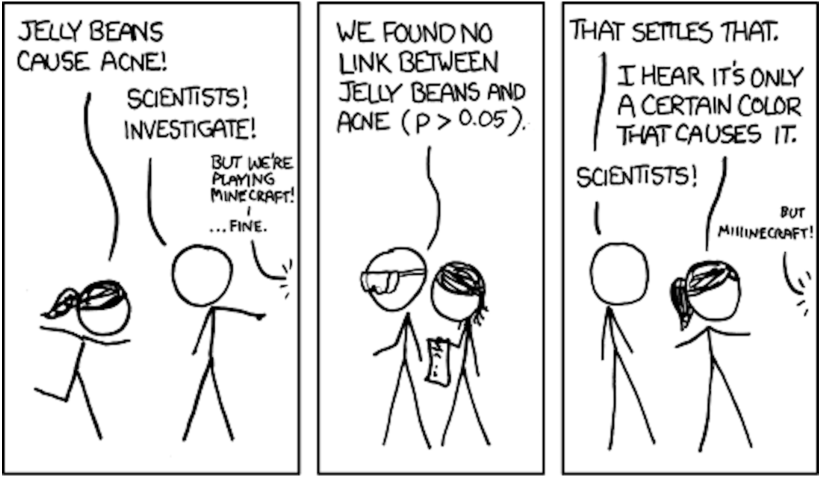
MULTIPLE COMPARISONS

BIAS VS VARIANCE



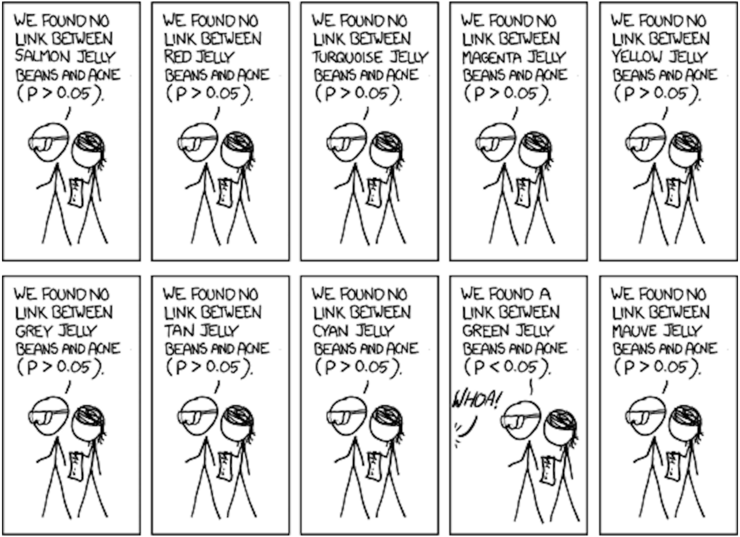
Source: Yarkoni, Tal, and Jacob Westfall. ”Choosing prediction over explanation in psychology: Lessons from machine learning.” Perspectives on Psychological Science 12, 6 (2017): 1100-1122.

MULTIPLE COMPARISONS

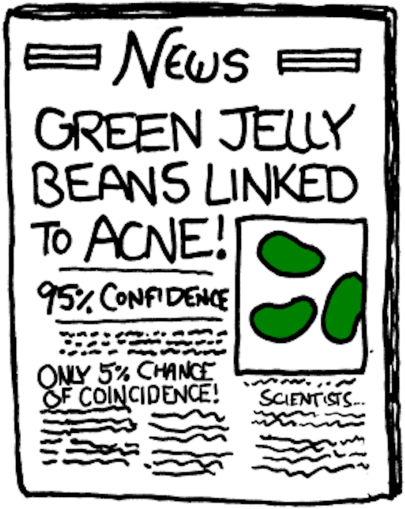

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

Credit: `https://xkcd.com/882/`

MULTIPLE COMPARISONS


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

Credit: `https://xkcd.com/882/`


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Credit: `https://xkcd.com/882/`

MULTIPLE COMPARISONS

This cartoon illustrates the multiple comparison problem

- The probability of attaining a false positive increases with the number of hypotheses tested

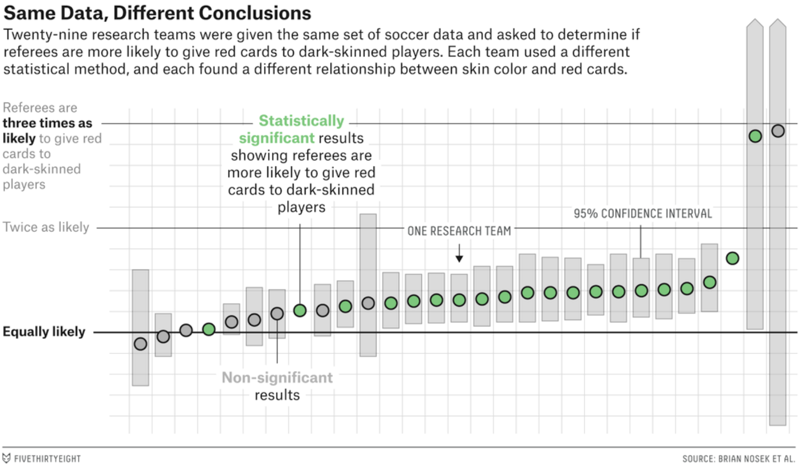
The same problem occurs when a large number of different models can potentially be used to test the same hypothesis

- Should you include control variables? If so, which?
- Should you allow for nonlinearities? If so, how?
- Should you model interactions? If so, which?
- Should you exclude outliers? If so, how defined?

The number of possible models increases exponentially with each choice: $2 \times 2 \times 2 \times 2 = 16$ and so on

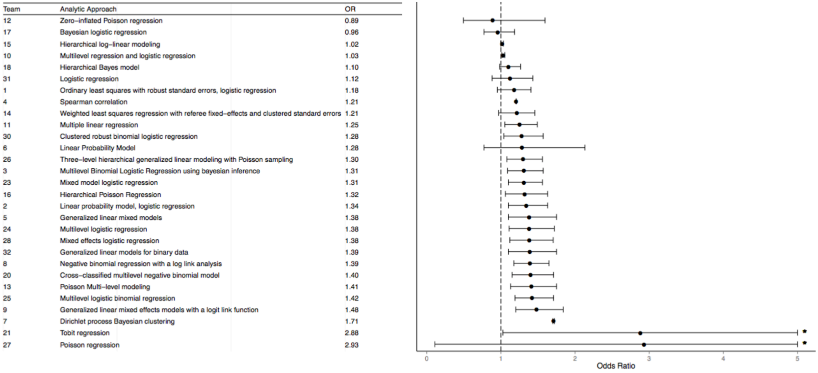
MODEL DEPENDENCE

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



MODEL DEPENDENCE

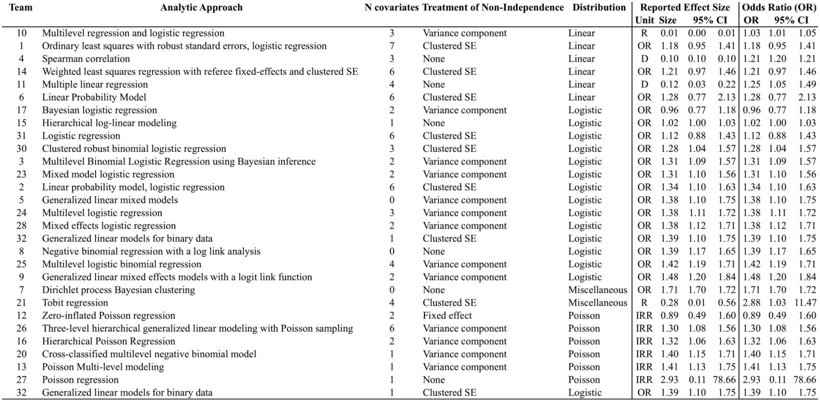
Silberzahn, et al. (2017, in press) “Many analysts, one dataset: Making transparent how variations in analytical choices affect results”.



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

MODEL DEPENDENCE

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

CHERRY PICKING

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 control variables
- 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 without controls

CHERRY PICKING

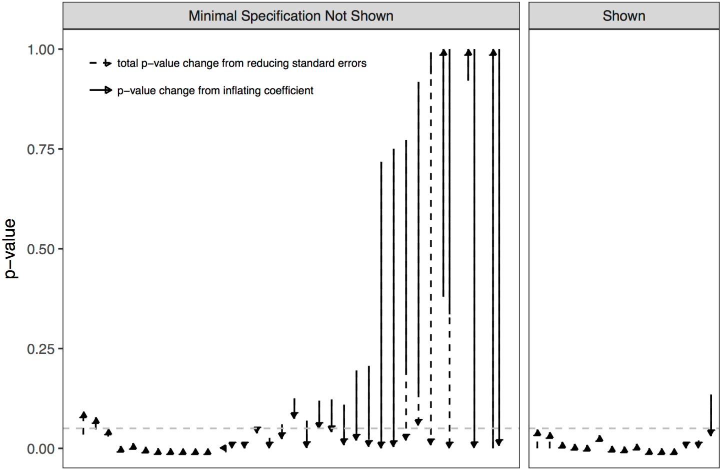


Figure 2: P-Value Changes in Observational Studies. For each article, the arrows show the total p-value changes from the minimal to the full specification. The solid part of the arrows shows the p-value changes from only coefficient estimate changes, while the dotted part shows p-value changes from standard 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.

SOLUTIONS

What to do? For the standard multiple comparison problem with several hypotheses there are procedures to correct your p-values

Bonferroni correction

- Very simple: with k hypotheses, divide alpha by k *or* multiply p-value by k (it’s the same thing)
- Keeps constant probability of finding at least one false positive (family-wise error rate)

Other tests (e.g., Benjamini-Hochberg procedure) are less penalising and instead control the expected proportion of false positives among significant results (false discovery rate)

For testing many group-level interactions, multilevel models are helpful. This is because they use ”shrinkage” to pull subgroup estimates closer to the pooled mean

AUTOMATION

The foregoing are specific solutions to specific problems. More generally, you want to make sure that your results are robust, that is, do not depend on a single specification

The solution is not to estimate fewer models but to estimate more of them – and get a sense of which results withstand scrutiny

Possible to do so by copy/pasting and substituting the relevant arguments anew in your code each time. However, doing so is 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

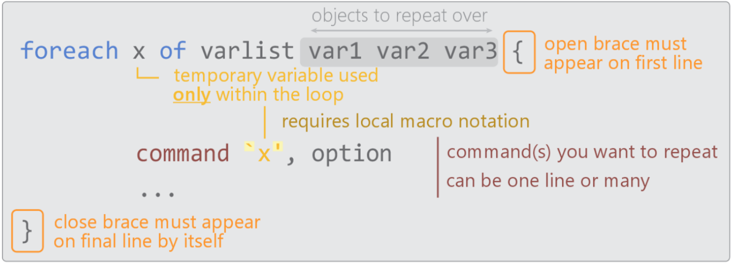
Two great user written Stata commands for publication-quality output are outreg2 and coefplot

AUTOMATION

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:



```
foreach x of varlist var1 var2 var3 {
        command `x', option
        ...
}
```

objects to repeat over

open brace must appear on first line

temporary variable used only within the loop

requires local macro notation

command(s) you want to repeat can be one line or many

close brace must appear on final line by itself

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

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!)

WHAT HAVE WE LEARNED?

Be smart about statistical problems and threats to valid inference

Avoid unwarranted causal claims – description is valuable enough

Look at coefficient sizes and not just statistical significance

Avoid cherry-picking specifications that tell the story you want to tell

Learn to code well, and make sure that your results are robust across more than one specification

Methods like IV can be useful, but always be careful about the assumptions you impose

WHAT HAVE WE LEARNED?

“Pseudo-facts have a way of inducing pseudo-problems, which cannot be solved because matters are not as they purport to be.”

- Robert K. Merton, “Notes on Problem-Finding in Sociology,” Sociology Today, Robert K. Merton, et al. (eds.). Basic Books (New York, 1959), p. xv.

“Before you come up with some smart explanation of how the pig got into the tree, just be sure that it is the pig that is in the tree.”

- William H. Sewell, quoted by John H. Goldthorpe, Sociology as a Population Science. Cambridge University Press (Cambridge, 2015), p. 87.

FURTHER READING

On measurement error/misclassification

- Bound, J., Brown, C., Mathiowetz, N. (2001). Measurement error in survey data. In Handbook of econometrics (Vol. 5, pp. 3705-3843). Elsevier.
- Hyslop, D. R., Imbens, G. W. (2001). Bias from classical and other forms of measurement error. Journal of Business Economic Statistics, 19(4), 475-481.

On survivorship bias

- Elwert, F., Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. Annual Review of Sociology, 40, 31-53.
- Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. Advances in Methods and Practices in Psychological Science.
- Wainer, H., Palmer, S., Bradlow, E. T. (1998). A selection of selection anomalies. Chance, 11(2).

FURTHER READING

On regression to the mean

- Jerrim, J., Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(4), 887-906.
- Kahneman, D. (2011). Thinking, Fast and Slow, Ch. 17. Penguin books.

On common support

- Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. Sociology of education, 341-374.
- Morgan, S. L., Winship, C. (2007). Counterfactuals and Causal Analysis: Methods and principles for social research, Ch. 4. Cambridge University Press.

FURTHER READING

On multiple comparisons

- Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention. Journal of the American statistical Association, 103(484), 1481-1495.
- Gelman, A. (2015). Working through some issues. Significance, 12(3), 33-35. www.stat.columbia.edu/~gelman/research/published/psych_crisis_minipaper4.pdf
- Gelman, A., Loken, E. (2016). The statistical crisis in science. American Scientist, 102. www.stat.columbia.edu/~gelman/research/published/ForkingPaths.pdf

On instrumental variables

- Angrist, J. D., Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion, Ch 4. Princeton University Press.
- Morgan, S. L., Winship, C. (2007). Counterfactuals and Causal Analysis: Methods and principles for social research, Ch. 7. Cambridge University Press.

Questions?