# No free lunch in inference

Ben Bolker
McMaster University

31 January 2022

# acknowledgements

money: NSERC

ideas: Jonathan Dushoff, Marm Kilpatrick, Brian McGill, Daniel Park, Daniel Turek

# no free lunch

- **Conjecture**: Data-driven model tuning can increase the *accuracy* of a point estimate, but cannot decrease its *uncertainty* (without further strong assumptions)
- *accuracy*: e.g. mean squared error
- *uncertainty*: e.g. width of the confidence interval

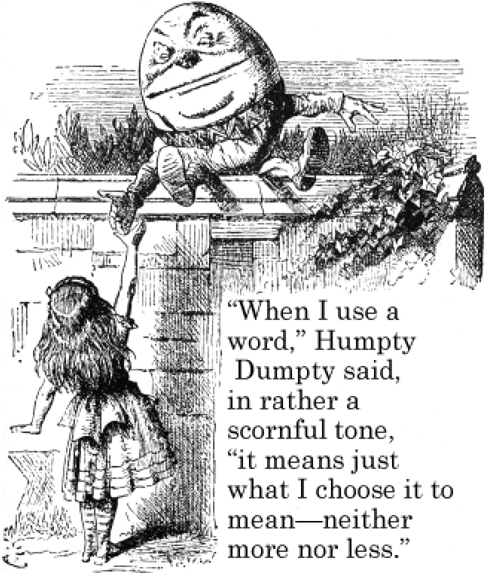
# scope of this talk

- full-rank ($p < n$), **non-sparse** problems
- science oriented (ecology/evolution)

# what are we doing when we do statistics?

- exploration
  - look for interesting patterns
  - confirm with followup observations
- prediction
  - best guess at future outcomes under specified conditions
- inference
  - estimate effects of processes *and their uncertainty*

# terminology



"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less."

# what do I mean by inference?

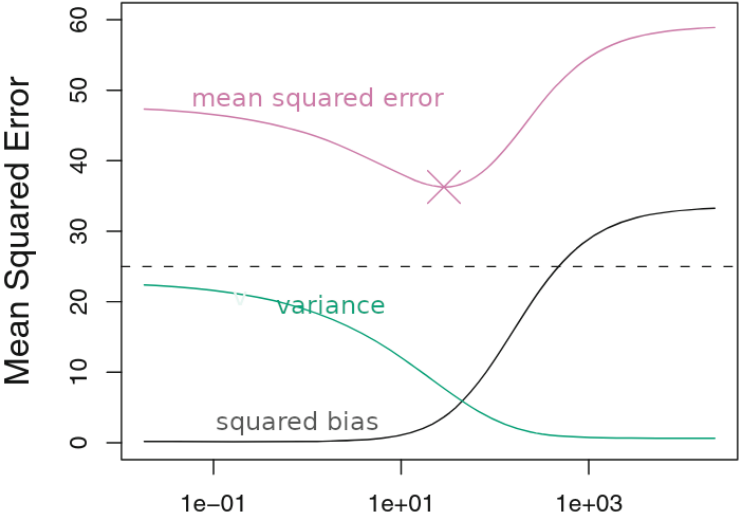
- **not** concerned with *formal* causal inference
- evaluation of uncertainty
- what would we expect to see in future data?
  - *coefficients*: uncertainty around the effect of a change in the predictors
  - *predictions*: uncertainty around the value observed for specified predictor values
  - *p-values*: uncertainty around a counterfactual null

# data-driven model tuning for point estimates

- avoid omitting potentially important predictors
- avoid overfitting
- ≈ optimize bias-variance tradeoff

Rackham 1918
Wikipedia

# e.g. bias-variance tradeoff in ridge regression



James et al.
ISLR fig 6.5

$\lambda$ (ridge penalty)

# data-driven tuning

- stepwise/subset regression, ridge/lasso/elastic net, random forests, boosting . . .
- need to choose *appropriate* model complexity
  - model size (selection) or complexity (shrinkage)
  - estimate out-of-sample error without (explicit) cross-validation (AIC, Cp, BIC, out-of-bag error, . . . )
  - or explicit cross-validation
- may need to tune model **hyperparameters** (e.g. via cross-validation)

# but what about uncertainty?

- statistical learning strongly focused on *prediction*
- **but** appropriate decisions require uncertainty quantification!
  - well appreciated in clinical trials
  - underappreciated in modern data science?

# how do we assess uncertainty quantification?

- false positive/type 1 error rate
- **coverage**: does an $x\%$ confidence interval include the true value $x\%$ of the time?
- (mentioned $0 \times$ in James et al. (2013), $1 \times$ in Hastie et al. (2009))

# coverage example (Li et al. 2018)

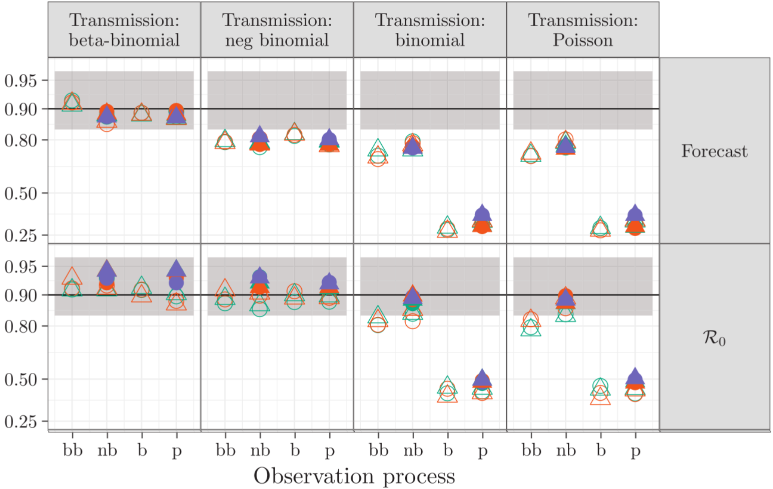


Figure 1: coverage plot for 90% intervals

# why coverage is better than type 1 error (mini-rant)

- type 1 error focuses on rejecting null hypotheses
- NH ($\beta = 0$) never(?) true in applied problems outside physics
- coverage reduces to type-1 error *if* $\beta = 0$
- type 1 assessment encourages unrealistic simulation setups

# naive selection methods

Post-selection inference that *ignores the selection process* is always be overoptimistic

- Altman et al. (1989)
  *Any form of data-dependent variable selection is likely to lead to overoptimistic goodness of fit; we expect a worse fit to a new set of data; bootstrapping with stepwise variable selection gave similar individual predictions but larger confidence intervals for estimated survival probabilities.*

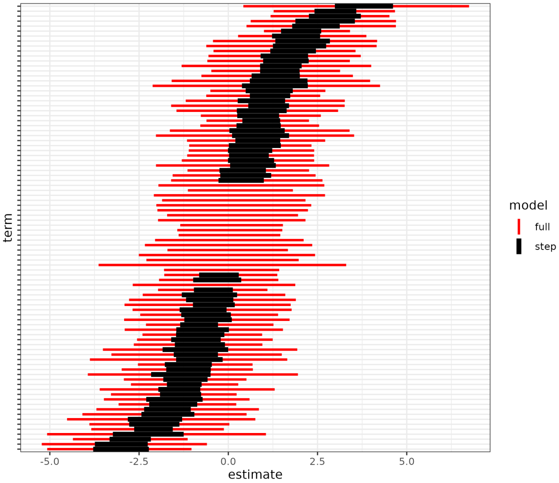
See also Harrell et al. (1996).

## for example . . .

AIC-stepwise regression, simulated data

`lm(y ~ .) + step()`

$n = 100$, $p = 90$, $\beta \sim U(-1, 1)$, $\sigma_r^2 = 5$; 90% CIs



cf. Freedman et al. (1983)

# coverage results (90% CIs)

| model | n | n_ok | prop | lwr | upr |
|---|---:|---:|---:|---:|---:|
| full | 18000 | 16284 | 0.905 | 0.900 | 0.909 |
| step | 14008 | 4220 | 0.301 | 0.294 | 0.309 |

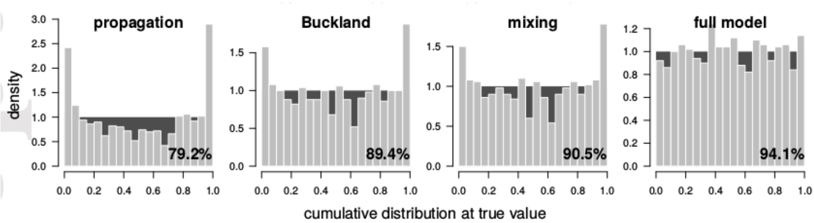
# what about something smarter, e.g. ridge regressoin?

- Obenchain (1977): properly constructed CI width $\geq$ full least-squares CI
- other methods may give OK results (e.g. Crivelli et al. (1995), Vinod (1987), Efron et al. (2020))
  - often involve *additional assumptions* - e.g. on the distribution of parameters
  - bootstrapping methods must incorporate the *full tuning process*

# multimodel averaging (MMA)

- various methods for constructing MMA CIs (Burnham et al. 2002; Fletcher et al. 2012; Kabaila et al. 2016)
- MMA CIs are generally **too narrow** (Turek 2013; Kabaila et al. 2016; Dormann et al. 2018) but cf. Burnham et al. (2002)

# MMA results (Dormann et al. (2018), Figure 5)



# What about post-selection inference?

- Lots of exciting work
- focused on high dimensions, depends on strong assumptions
  - sparsity
  - coefficient gap (minimum size of smallest $|\beta|$)
  - asymptopia
- e.g. Dezeure et al. (2015):
  *When the truth (or the linear approximation of the true model) is nonsparse, the methods are expected to break down . . .*
- See Cosma Shalizi's notes at
  {http://bactra.org/notebooks/post-model-selection-inference.html}

# paying for lunch in other ways

- go Bayesian!
  - Bayesian CIs are well-calibrated *by definition*, conditional on the model and the priors . . .
    (Gelman et al. 1995; Cook et al. 2006; Talts et al. 2020)
- pseudo-Bayesian assumptions about effect size distributions

# conclusions: what should you do?

- for **inference**:
    - use the full model
    - *a priori* model reduction (Harrell 2001)
- for **prediction**:
    - use CIs from shrinkage estimates with caution
    - use non-neutral, informative Bayesian priors? (Crome et al. 1996)

# there ain't no such thing as a free lunch . . .

# (blank)

# what are multifactorial systems?

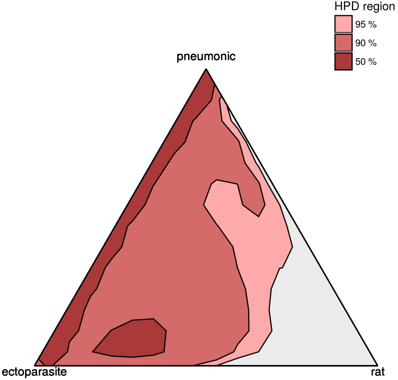
- many processes contribute to pattern
- quantify *how* each process affects the system, rather than testing *whether* we can detect its impact
- related:
  - Chamberlin's method of multiple working hypotheses (Raup et al. 1995)
  - tapering effect sizes (Burnham et al. 2002)

# conceptual problem: discretization

- model selection, or evidential statistics (Taper et al. 2016), focus on differentiating **discrete** hypotheses/models
- submodels are always straw men
- expand models to cover the whole space

# conceptual problem: discretization

- model selection, or evidential statistics (Taper et al. 2016), focus on differentiating **discrete** hypotheses/models
- submodels are always straw men
- expand models to cover the whole space



{Estimated contribution of plague transmission modes in Eyam 1665}

# References I

Altman, DG et al. 1989.. *Statistics in Medicine* 8 (7): 771–783. doi:10.1002/sim.4780080702. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780080702.

Burnham, KP et al. 2002. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. Springer.

Cook, SR et al. 2006.. *Journal of Computational and Graphical Statistics* 15 (3) (September): 675–692. doi:10.1198/106186006X136976. http://www.tandfonline.com/doi/abs/10.1198/106186006X136976.

Crivelli, A et al. 1995.. *Communications in Statistics - Simulation and Computation* 24 (3) (January): 631–652. doi:10.1080/03610919508813264. https://doi.org/10.1080/03610919508813264.

Crome, FHJ et al. 1996.. *Ecological Applications* 6: 1104–1123.

Dezeure, R et al. 2015.. *Statistical Science* 30 (4) (November): 533–558. doi:10.1214/15-STS527. https://projecteuclid.org/euclid.ss/1449670857.

Dormann, CF et al. 2018.. *Ecological Monographs*. doi:10.1002/ecm.1309. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1309.

Efron, B et al. 2020.. *Journal of Computational and Graphical Statistics* 29 (3) (July): 608–619. doi:10.1080/10618600.2020.1714633. https://doi.org/10.1080/10618600.2020.1714633.

Fletcher, D et al. 2012.. *Journal of agricultural, biological, and environmental statistics* 17 (1): 38–51.

Freedman, DA et al. 1983.. *The American Statistician* 37 (2) (May): 152–155. doi:10.1080/00031305.1983.10482729. https://www.tandfonline.com/doi/abs/10.1080/00031305.1983.10482729.

Gelman, A et al. 1995.. *Sociological Methodology* 25: 165–173. doi:10.2307/271064. http://www.jstor.org/stable/271064.

Harrell, F. 2001. *Regression Modeling Strategies*. Springer.

# References II

Harrell, F et al. 1996.. *Stata FAQ*. https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/.

Hastie, T et al. 2009. *The elements of statistical learning data mining, inference, and prediction*. New York: Springer. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=437866.

James, G et al. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

Kabaila, P et al. 2016.. *Scandinavian Journal of Statistics* 43 (1): 35–48. doi:10.1111/sjos.12163. http://onlinelibrary.wiley.com/doi/10.1111/sjos.12163/abstract.

Li, M et al. 2018.. *Statistical Methods in Medical Research* 27 (7): 1956–1967. doi:10.1177/0962280217747054.

Obenchain, R. 1977.. *Technometrics* 19 (4): 429–439.

Raup, DC et al. 1995.. *The Journal of Geology* 103 (3): 349–354. http://www.jstor.org/stable/30071227.

Talts, S et al. 2020.. *arXiv:1804.06788 [stat]* (October). http://arxiv.org/abs/1804.06788.

Taper, ML et al. 2016.. *Population Ecology* 58 (1) (January): 9–29. doi:10.1007/s10144-015-0533-y. http://link.springer.com/10.1007/s10144-015-0533-y.

Turek, DB. 2013.. PhD thesis, University of Otago. https://www.otago.ourarchive.ac.nz/bitstream/handle/10523/3923/TurekDanielB2013PhD.pdf.

Vinod, HD. 1987.. In *Time Series and Econometric Modelling*, ed by. Ian B. MacNeill et al., 279–300. Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-4790-0_19. http://link.springer.com/10.1007/978-94-009-4790-0_19.