Selection of Variables and Functional Forms in Multivariable Analysis: Current Issues and Future Directions

Frank E Harrell Jr

Department of Biostatistics
Vanderbilt University School of Medicine

STRATOS Banff Alberta 2016-07-04

Functional Forms

- Fractional polynomials, regression splines, penalized splines, smoothing splines, nonparametric smoothers, etc.
- Advantages of regression splines
  - Shape flexibility; allows sharp changes and long flat segments
  - Can pre-set d.f.
  - Simple progression of form as knots added
  - Origin-free when knots are selected from marginal distribution of X, i.e., $f(X + c)$ is a horizontal shift from $f(X)$
  - Allows $X \leq 0$
  - Extension to tensor splines for interaction modeling
  - Easy to penalize some types of complexity

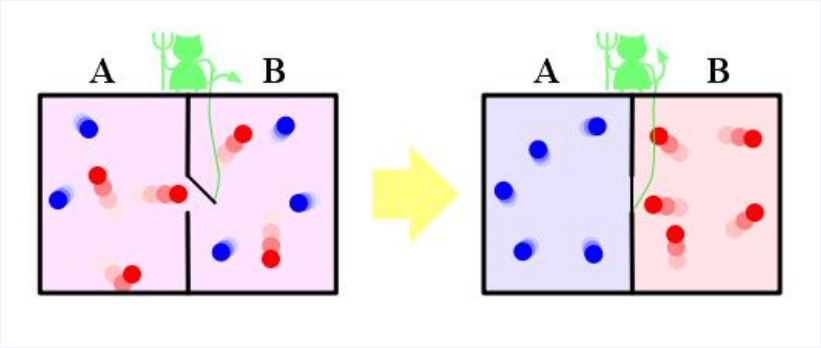
Setting

- Exploratory vs. formal/predictive analysis
- Too many variables, too few subjects
- Many investigators and some statisticians feel that variable selection helps

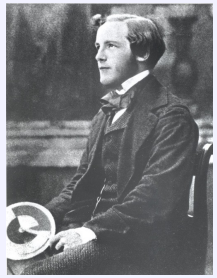
Variable Selection

- Is misleading to the researcher
- Often arbitrary and unstable
- Attempts to obtain clear result in the presence of competition among predictors; result is clear only because it is misleading
- Reduce the number of variables to collect in the future
- Some of the information in the data is spent on variable selection instead of using all information for estimation

Maxwell's Demon



James Clerk Maxwell



Maxwell imagines one container divided into two parts, A and B. Both parts are filled with the same gas at equal temperatures and placed next to each other. Observing the molecules on both sides, an imaginary demon guards a trapdoor between the two parts. When a faster-than-average molecule from A flies towards the trapdoor, the demon opens it, and the molecule will fly from A to B. Likewise, when a slower-than-average molecule from B flies towards the trapdoor, the demon will let it pass from B to A. The average speed of the molecules in B will have increased while in A they will have slowed down on average. Since average molecular speed corresponds to temperature, the temperature decreases in A and increases in B, contrary to the second law of thermodynamics.

Szilárd pointed out that a real-life Maxwell's demon would need to have some means of measuring molecular speed, and that the act of acquiring information would require an expenditure of energy. Since the demon and the gas are interacting, we must consider the total entropy of the gas and the demon combined. The expenditure of energy by the demon will cause an increase in the entropy of the demon, which will be larger than the lowering of the entropy of the gas.

commons.wikimedia.org/wiki/File:YoungJamesClerkMaxwell.jpg

en.wikipedia.org/wiki/Maxwell's_demon

Variable Selection, *continued*

- Example

Method	Apparent Rank Correlation of Predicted vs. Observed	Over-Optimism	Bias-Corrected Correlation
Full Model	0.50	0.06	0.44
Stepwise Model	0.47	0.05	0.42

- Model *specification* is preferred to model *selection*
- Information content of the data is usually insufficient for reliable variable selection
- Bootstrap or repeating the selection on a new sample expose the difficulty of the task

Stepwise Methods Can't Select the Right Variables

8 candidate predictors, 4 have $\beta \neq 0$, Gaussian errors
Stepwise selection using AIC

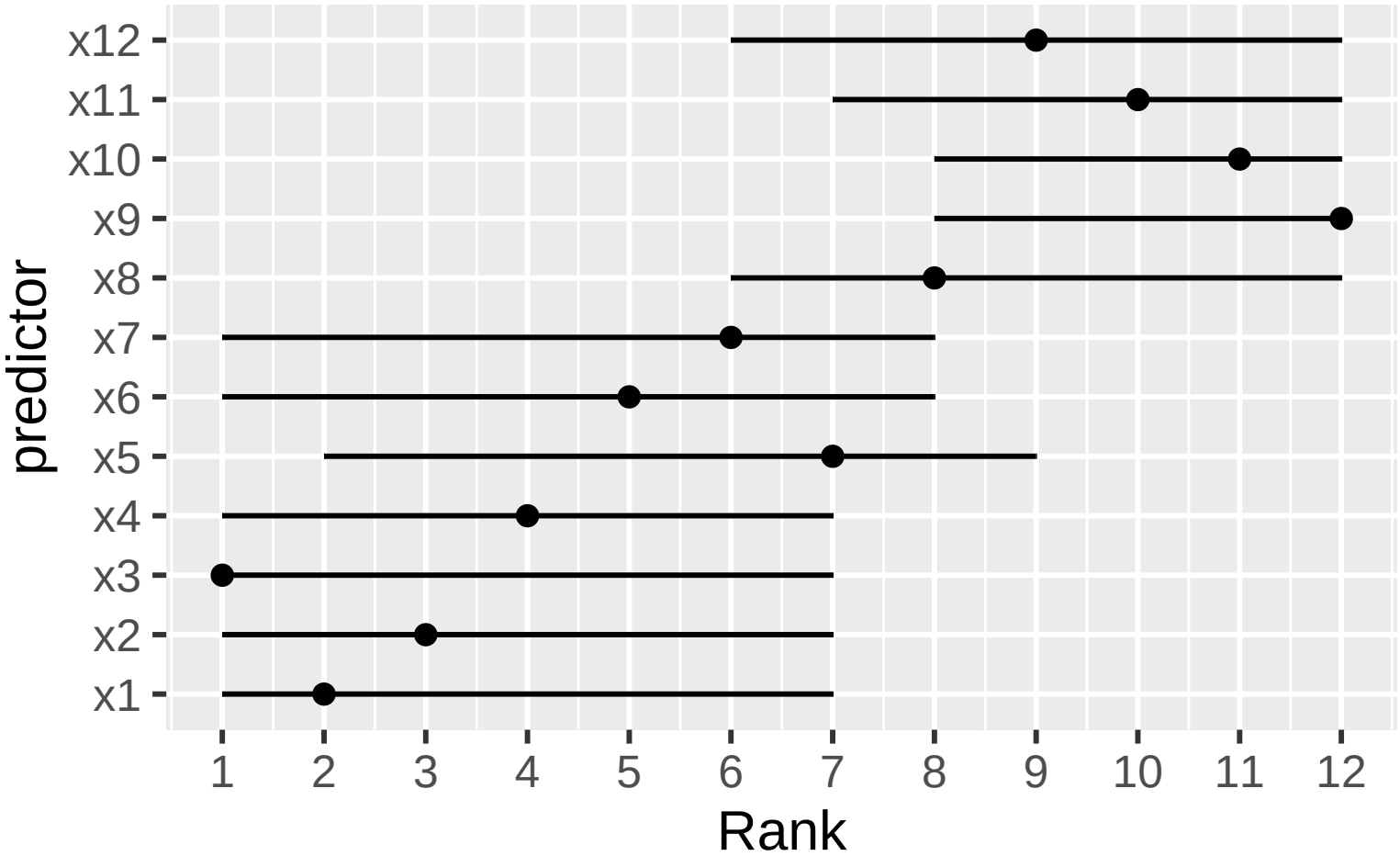
n	Correct Model	gdf	$\hat{\sigma}^2/\sigma^2$
20	0	8.1	0.70
40	0	7.3	0.87
150	0.13	9.1	0.97
300	0.38	9.3	0.98
2000	0.52	6.3	1.00

Generalized d.f.: gdf : $\frac{\text{SSE}}{n-gdf-1}$ unbiased for σ^2

Harrell 2015, Ye 1998, Freedman, Pee, and Midthune 1992

Bootstrapping Importance Ranks of Predictors

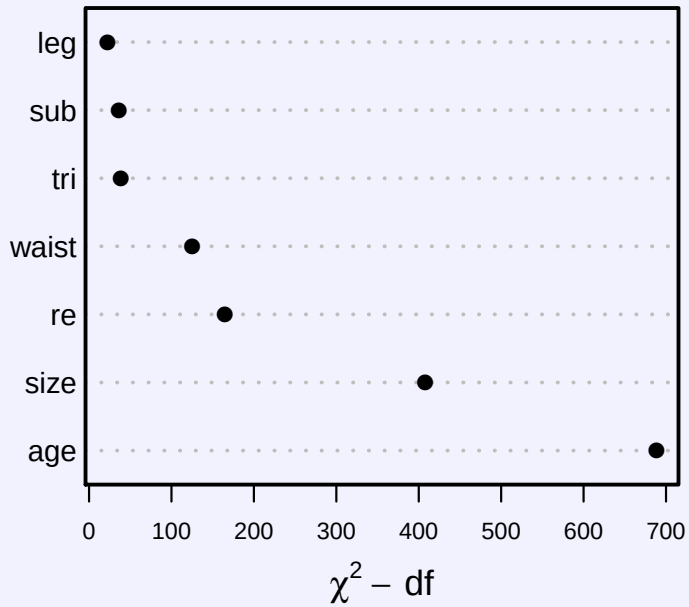
$n = 300$, 12 predictors, $\beta_i = i, \sigma = 9$; rank partial χ^2



Harrell 2015

Pooled Tests Instead of Variable Selection

- Example: semiparametric ordinal model for predicting glycohemoglobin in NHANES cohort
- There are better body size measures for predicting preclinical diabetes than height & weight
- Body size measures compete
- Compute "chunk test"



Prediction vs. Understanding

- Many good approaches exist for developing accurate predictive models, e.g.
  - Data reduction, e.g., variable clustering; summarizing clusters with PC_1
  - Penalized MLE with large numbers of predictors and nonlinear terms
  - Machine learning algorithms
- Predictive discrimination is best when the model is not decoded/simplified
- **Prediction Uncertainty Principle**: a prediction tool can be optimally predictive or parsimonious but not both
- Related to LJ Savage's anti-parsimony principle

Aside: *lasso* and Related Methods

- Trade one problem (high $\frac{p}{n}$) for another
- Assuming all effects are linear
- Scaling problems make it difficult to know how to include nonlinear terms

Advantages of Model Pre-Specification

- Apparent d.f. = actual d.f.
- Accurate type I error
- Accurate confidence intervals
- Regression splines: pre-specify number & location of knots
- Most "found" interactions turn out to be false

Perspective of a Bayesian Modeler

- Everything is pre-specified but general enough to allow flexibility
- Prior distributions for regression coefficients generally centered at zero to bring skepticism/shrinkage
- Gaussian prior $\equiv$ quadratic penalty (ridge regression)
- Laplace (double exponential) prior $\equiv$ lasso/variable selection
- Terms are not just in or out of the model; they can be half–in
- Natural for interaction terms
- Bayesian posterior provides formal inference
- With frequentist penalized MLE formal inference not fully developed

One Future Direction: Interaction Modeling

- Precision medicine, aka personalized medicine, aka heterogeneity of treatment effect
  (New name scheduled for 2017)
- Most clinicians attempt to use subgroup analysis; this is doomed
- Needs to be well planned, formal interaction analysis
- Interaction effects have lower precision than main effects, and more co-linearity problems
- Main effects need to be liberally and nonlinearly accounted for before considering interaction effects
- Dimensionality and identifiability problem
  - Does the drug work more on older patients, or patients with more disease?

Reduced Rank Interaction Modeling

- Main effects fully expanded as usual
- PC_1, PC_2, ..., PC_k for interaction terms of dimension $K > k$
- k chosen to be the maximum the effective sample size will fully support
- Attempt to interpret linear combination of PCs in terms of constituent variables
- Need to develop more sensible restrictions than PCs, e.g., penalty function
  - E.g. penalize interact effects for poorly distributed predictors
  - Penalize for co-linear predictors
- Bayesian approaches

References

Freedman, L. S., D. Pee, and D. N. Midthune (1992). "The Problem of Underestimating the Residual Error Variance in Forward Stepwise Regression". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 41.4, pp. 405–412 (cit. on p. 7).

Harrell, F. E. (2015). *Regression Modeling Strategies, with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Second edition. New York: Springer. isbn: 978-3-319-19424-0 (cit. on pp. 7–9).

Ye, J. (1998). "On measuring and correcting the effects of data mining and model selection". In: *J Am Stat Assoc* 93, pp. 120–131 (cit. on p. 7).