A Discussion of 3 Talks from the 2014 Joint Statistical Meetings

Eric Slud, UMCP Math Dept. & Census CSRM

Stat Seminar, Oct. 30, 2014

Three talks/topics

I. *Fisher Lecture* by Stephen Stigler (U. Chic.),

"The Seven Pillars of Statistical Wisdom"

II. Invited talk: Andreas Buja (U. Penn, Wharton),

"Valid Post-Selection Inference"

III. Contributed: Qixuan Chen (Columbia Biostat.),

"Bayesian post-stratification models using multilevel penalized spline regression"

The Seven Pillars, annotated

reconstructed from notes & blog of R. Wicklin, SAS

(1) Aggregation – summary, descriptive parameters

(2) Law of diminishing information – the '$\sqrt{n}$ rule'

(3) Likelihood – conditional prob's, multiple quest.'s

(4) Intercomparisons – contrasts, k-sample tests

(5) Regression & multivariate analysis – catchall

(6) Design – experimental design, randomization

(7) Models and Residuals – model-criticism,
including statistical adequacy & goodness of fit

Stigler's Lecture on What is Statistics, cont'd

Important categories (for me) to find in the list:

- idea of formal hypothesis tests, incl. specification of **alternatives** to a statistical hypothesis ? **(7)**
- **decision theory (3)**
- **data-representation (5)**
- **nonparametrics** ? **orthogonal decomposition** ?

Techniques & algorithms not highlighted on the list
Simulation, MCMC, Bootstrap or Bayes

Large Some General Themes from the Meetings

- Big data – Genomics, finance, brain imaging, etc.
- 'Reproducible research' – reporting of inferences
- Bayes methods – continuing computational trend

A. Buja talk on 'PoSI Project'

(joint with R. Berk, L. Brown, K. Zhang, L. Zhao)

based on 2013 Annals of Stat paper

- Asymptotics to account for degrees of freedom in model choice (including variable selection, transformation of variables, etc.) with Scheffé method of multiple comparisons as special case

Following slides taken from the talk also appear in a long talk-pdf on Buja's web-site

"PoSI" — Valid Post-Selection Inference

Andreas Buja

joint work with

Richard Berk, Lawrence Brown, Kai Zhang, Linda Zhao

Department of Statistics, The Wharton School
University of Pennsylvania
Philadelphia, USA

UF 2014/01/18

Larger Problem: Non-Reproducible Empirical Findings

- Indicators of a problem
  (from: Berger, 2012, "Reproducibility of Science: P-values and Multiplicity")
  - Bayer Healthcare reviewed 67 in-house attempts at replicating findings in published research:
    $< 1/4$ were viewed as replicated.
  - Arrowsmith (2011, Nat. Rev. Drug Discovery 10):
    Increasing failure rate in Phase II drug trials
  - Ioannidis (2005, PLOS Medicine):
    "Why Most Published Research Findings Are False"
  - Simmons, Nelson, Simonsohn (2011, Psychol.Sci):
    "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant,"

- Many potential causes – two major ones:
  - publication bias: "file drawer problem" (Rosenthal 1979)
  - statistical biases: "researcher degrees of freedom" (SNS 2011)

Statistical Biases – one among several

- Hypothesis: A statistical bias is due to
  an absence of accounting for model/variable selection.
- Model selection is done on several levels:
  - formal selection: AIC, BIC, Lasso, ...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: "The effect size is too small in relation to the cost of data collection to warrant inclusion of this predictor."
- Suspicions:
  - All three modes of model selection may be used in much empirical research.
  - Ironically, the most thorough and competent data analysts may also be the ones who produce the most spurious findings.
  - If we develop valid post-selection inference for "adaptive Lasso", say, it won't solve the problem because few empirical researchers would commit themselves a priori to one formal selection method and nothing else.
    $\Rightarrow$ "Meta-Selection Problem"

The Problem of Post-Selection Inference

How can Variable Selection invalidate Conventional Inference?

- Conventional inference after variable selection ignores the fact that the model was obtained through a stochastic selection process.
- Stochastic variable selection distorts sampling distributions of the post-selection parameter estimates: Most selection procedures search for strong, hence highly significant looking predictors.

Evidence from a Simulation

Generate Y from the following linear model:

$$Y = \beta x + \sum_{j=1}^{10} \gamma_j z_j + \epsilon,$$

where $p = 11$, $N = 250$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ iid.

▸ More Details

- For simplicity: "Protect" x and select only among z_1,...,z_{10}; interest is in inference for β.
- Model selection: All-subset search with BIC among z_1,...,z_{10}; always including x.
- Proper coverage of a 95% CI on the slope β of x under the chosen model requires that the t-statistic is about $\mathcal{N}(0, 1)$ distributed.

Background to Q. Chen Talk

Definitions related to (Bayesian) survey sampling

Survey weights w_i are estimator coefficients, modifying $1/\pi_i$ where probabilities $\pi_i = P(i \in S)$. Often related to demographic or geographic predictor variables for response.

Post-stratification standard method of applying weights to units based on known aggregate-population totals; usually cell-based, weight-factor $r \cdot p_h/r_h$ in cell h based on known population proportion p_h and sample responder-proportion r_h/r.

Related (frequentist) term: calibration, (generalized) raking

Place of models in surveys

Little (2001): Bayesian poststratification model.

For unit-level discrete observations $z_i \in \{1, \ldots, H\}$,

$$y_i \sim \mathcal{N}(\mu_h, \sigma_h^2) \quad \text{given} \quad z_i = h$$

where μ_h may themselves be modelled further, as mean plus *random effects* or with discrete main-effect terms α_k, β_l for $h = (k, l)$ and interactions within random effects

$$\mu_h = \alpha_k + \beta_l + \gamma_{kl} \ , \quad \gamma_{kl} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\gamma^2)$$

Keywords: small area estimation, borrowing strength

Spline survey models in terms of weights

Zheng & Little (2003), PPS sampling ($\pi_i \propto$ scalar x_i), derives HT estimator $\sum_i y_i/\pi_i$ from model $y_i = \beta\pi_i + \pi_i e_i$, proposes

$$y_i = f(\pi_i, \beta) + \epsilon_i \ , \quad \mathcal{N}(0, \pi_i^{2q}\sigma^2)$$

Fitting basis-expanded P-splines with a sum of coefficients-squared roughness penalty is equivalent to treating highest-order coefficients as random effects, (design-consistently for $q = 1/2$), in

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{k=1}^{m} \gamma_k (\pi_i - \kappa_k)_+^p + \pi_i^q e_i$$

where κ_k are knots, τ is a tuning parameter

$$e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \ \gamma_k \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$$

New elements in Q. Chen talk (co-authors incl. A. Gelman)

$$y_i \sim \mathcal{N}(\mu_h, \sigma_h^2) \quad \text{for} \quad z_i = h$$

$$\mu_h \sim \mathcal{N}(g_1(x_h), \tau_1^2) \ , \quad \sigma_h^2 \sim \mathcal{N}(g_2(x_h), \tau_2^2)$$

where $x_h = \pi_h, 1/\pi_h$ or $-\log(\pi_h)$

$g_j(x) =$ spline of degree p_j with

iid $\mathcal{N}(0, \phi_j^2)$ coefficients of degree-p_j terms

Bayesian model fitting using MCMC, noninformative priors for lower-degree spline coeff's and ϕ_j

References

Buja, A., Brown, L. et al. (2013) Ann. Stat.
Valid Post-Selection Inference

Little, R. (2001, JASA) *Post-stratification* ...

Zheng, H. & Little, R. (2003) *Jour. Official Stat.*