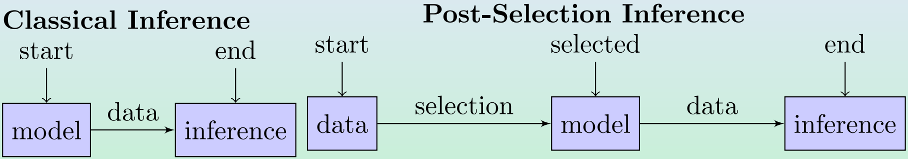


Classical Inference

start

end

model

data

inference

Post-Selection Inference

start

selected

end

data

selection

model

data

inference

Post-Selection Inference

Todd Kuffner
Washington University in St. Louis

WHOA-PSI 2016

The problem of post-selection inference

Classical inference assumes the model is chosen independently of the data.

The problem of post-selection inference

Classical inference assumes the model is chosen independently of the data.

Using the data to select the model introduces additional uncertainty

$\Rightarrow$ invalidates classical inference

The problem of post-selection inference

Classical inference assumes the model is chosen independently of the data.

Using the data to select the model introduces additional uncertainty

⇒ invalidates classical inference

Do you believe me?

Example 1: Forward Stepwise Regression

R. Lockhart, J. Taylor, Ryan Tibshirani, Rob Tibshirani (2014), 'A significance test for the lasso', *Annals of Statistics*.

Classical inference for linear regression: two fixed, nested models

Model A variable indices $M \subset \{1, \ldots, p\}$

Model B variable indices $M \cup \{j\}$

Goal: test significance of jth predictor in Model B

Compute drop in RSS from regression on $M \cup \{j\}$ and M

$$R_j = (\mathrm{RSS}_M - \mathrm{RSS}_{M \cup \{j\}})/\sigma^2 \quad \text{versus} \quad \underbrace{\chi^2_1}_{\text{for } \sigma^2 \text{ known}}$$

Post-selection inference: first use selection procedure, then do inference

- want to do the same test as above for Models A and B which are not fixed, but rather outputs of selection procedure

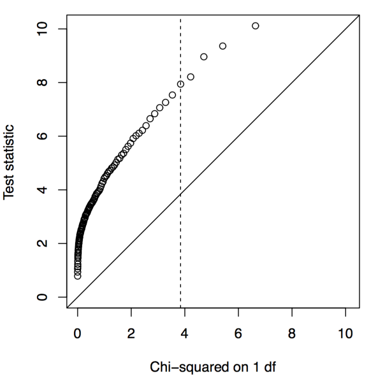
e.g. forward stepwise

- start with empty model $M = \emptyset$
- enter predictors one at a time: choose predictor j giving largest drop in RSS
- FS chooses j at each step to to maximize $R_j = (\mathrm{RSS}_M - \mathrm{RSS}_{M \cup \{j\}})/\sigma^2$
- each $R_j \sim \chi^2_1$ (under null)

$\Rightarrow$ max possible R_j stochastically larger than χ^2_1 under null

Illustration

Compare quantiles of R_1 in forward stepwise regression, i.e. chi-square for <u>first</u> predictor to enter versus those of χ^2_1 variable, when $\beta_k = 0 \ \forall k = 1, \ldots, p$.



$n = 100$, $p = 10$ (orthogonal); all true coefficients are zero; 1000 simulations of statistic R_1, versus χ^2_1 distribution; dotted line is 0.95 quantile of χ^2_1

At 0.05 level, using χ^2_1 quantile (3.84) has *actual* type I error probability of 0.39

Example 2: *File Drawer Effect* (Fithian, 2015)

Observe $Y_1, \ldots, Y_n$ independently $\sim \mathcal{N}(\mu_i, 1)$

Suppose you focus on 'apparently' large effects, $|Y_i| > 1$:

$$\hat{I} = \{i : |Y_i| > 1\}$$

Goal: test $H_{0,i} : \mu_i = 0$ for each $i \in \hat{I}$ at level 0.05.

- Usual approach: reject $H_{0,i}$ when $|Y_i| > 1.96$

Not Valid Due to Selection *Why?*

Seems counterintuitive: probability of falsely rejecting a given $H_{0,i}$ is *still* α, since most of the time $H_{0,i}$ is not tested at all.

Problem: for those hypotheses selected for testing, type I error rate is possibly much higher than α

Proof of concept

- let n_0 be # of true null effects
- assume $n_0 \to \infty$ as $n \to \infty$

Long-run fraction of errors among the true nulls we test:

$$\begin{aligned}
\frac{\text{\# false rejections}}{\text{\# true nulls selected}} &= \frac{\frac{1}{n_0}\sum_{i:H_{0,i}\ \text{true}} 1\{i \in \hat{I},\ \text{reject } H_{0,i}\}}{\frac{1}{n_0}\sum_{i:H_{0,i}\ \text{true}} 1\{i \in \hat{I}\}} \\
&\to \frac{\mathbb{P}_{H_{0,i}}(i \in \hat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \hat{I})} \\
&= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})
\end{aligned}$$

For nominal test, this is $\Phi(-1.96)/\Phi(-1) \approx 0.16$

Other examples

1. Box-Cox transformation (Chen, Lockhart & Stephens, 2002); Even using the data to find suitable transformations causes problems.
2. Marginal screening (Lee & Taylor, 2014): screen variables by largest correlation with response
3. Selection of ‘events’ in high-energy physics
4. Choosing low p-values
5. Even graphical displays of data to look for patterns!

Are Bayesians immune?

Dawid (1994): selection should have no effect

> *Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data.*

Yekutieli (2012, 'Adjusted Bayesian inference for selected parameters,' *JRSSB*): (More from Jonathan Taylor and Snigdha Panigrahi)

Actually, selection can affect Bayesian inference *... but since Professor Yekutieli is here, I will let him explain more about that!*

Bayesian inference for parameters selected after viewing the data is a 'truncated' data problem.

Why should the sciences care?

We hate false discovery as much as you do.

- control of FDR, FCR, FWER are key desiderata in PSI
- applications are endless: need to formalize the informal ‘data snooping’ (adaptive selection) process to properly account for uncertainty

To make things extra complicated...

- select minimum signal threshold, then do inference for selected signals
- selection of ‘events’
- data transformations based on data snooping

Broad (*crude*) Classification of PSI

1. **data splitting** (Cox, Wasserman) and **data carving** (Fithian)

   idea: the source of the problem is using the same data for selection and inference; solution: use some data for selection, the rest for inference

2. **high-dimensional inference** (the Swiss, signal processing, machine learning, econometrics; the old-school bootstrappers):

   idea: ignore selection, view as single procedure followed by interval correction; not *really* PSI?

3. **simultaneous inference** (Benjamini, Yekutieli, Heller, Wharton)

   idea: control FDR for all models ever under consideration by selection procedure; solution: fix the confidence intervals

4. **selective inference** (Benjamini, Yekutieli, Stanford)

   idea: inference for selected hypotheses

Warning/Disclosure

In this tutorial I only talk about simultaneous and selective inference. This doesn't mean I think the other approaches are any less valuable, but rather that selective and simultaneous inference are (I think) less familiar to the typical statistician

Point of Contention: Full model vs. Submodel (Berk et al., 2013, Appendix)

Suppose we have

$$\text{Full model:} \quad Y_i = \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i$$

Full model viewpoint: full observed set of p covariates *defines* the data generating process for the response.
⇒ coefficients have *fixed meaning* as *full model parameters*

Crucially, this means that the **targets** of estimation and testing are fixed.

Full Model Viewpoint continued

Estimation of full model parameters by penalized methods results in a *shrinkage* estimator for the full model parameters (Econometric-speak: ‘preliminary test estimators’

The submodel resulting from variable selection is only a *summary* or *compression* of the data: not a legitimate competitor model worthy of study on its own

In the full model world, a variable selection procedure leads to biases which can adversely affect subsequent inference.

- could omit partially collinear covariates ⇒ causing correlation with unobserved error among the selected covariates ⇒ biased estimators

The Dark World of the Full Model

Leeb & Pötscher (2006-2008): ‘impossible’ to consistently estimate sampling distributions of post-selection coefficient estimators

Submodel Viewpoint

Suppose we apply selection procedure, result is

$$\text{Selected/sub- model:} \quad Y_i = \sum_{k \in \hat{M}} \beta_{ik} x_{ik} + \gamma_i$$

with $\hat{M} \subseteq \{1, \ldots, p\}$.

Parameter spaces are not the same; should we do inference about full model parameters or submodel parameters?

Submodel Viewpoint

The parameters corresponding to any subset of covariates selected by a variable selection procedure are specific to the submodel.

Thus the ***targets*** are submodel parameters, which are in general different for different models and, crucially, are not conceptually the same as the full model parameters.

Viewed this way, the submodel estimates are not targeting the full model parameters, and the bias criticisms arising in the full model view are not present.
Berk et al. (2013): submodel interpretation of parameters is more consistent with statistical theory, and admits the interpretation of submodels as linear approximations to the truth.

Some notation

Variable selection procedure *followed by* ordinary least squares estimation: choose some subset $M \subset \{1, \ldots, p\}$ using, say, forward stepwise (FS) or least angle regression (LAR), and seek the unique linear combination of predictors belonging to M such that the expected error is minimized,

$$\beta^M \equiv \arg\min_{b^M} \mathbb{E}\|Y - X_M b^M\|^2 = X_M^+ \mu, \tag{1}$$

- X_M the matrix composed of columns X_j for $j \in M$,
- β^M is the coefficient vector for model M and $X_M^+ \equiv (X_M^T X_M)^{-1} X_M^T$ is the Moore-Penrose pseudo-inverse of X_M.

Clearly, for two different models $M_1 \neq M_2$, the estimation targets, say $\beta_j^{M_1}$ and $\beta_j^{M_2}$, **are generally not the same**.

- Interpretation: linear regression coefficients are understood as changes in the mean of the response *controlling for other predictors.*
- Thus the coefficient of a predictor is not directly comparable across models; this gives a simple intuition for adopting the submodel viewpoint.

Comment from Lee, Sun Sun Taylor (2015)

That the target β^M will change depending on which model is selected *does not* mean the parameters themselves must be regarded as random.

Rather the randomness of which model is selected only implies that the set of parameters to be considered is random, but within any particular submodel these parameter values may be regarded as fixed.

Next Point of Departure: Which frequency properties do we want for P-S interval estimates?

A priori there are 2^p possible models, and with one parameter corresponding to each coefficient in each of these models, there are $p2^{p-1}$ well-defined population parameters:

$$\{\beta_j^M : M \subset \{1, \dots, p\}, j \in M\}.$$

Inference only takes place for the parameters $\beta_j^{\hat{M}}$ in the selected model $\hat{M}$.

As argued in a series of papers since 2005, Benjamini & Yekutieli, along with Hochberg, Heller and others, demonstrated that the adaptive selection of which parameters to consider may lead to undesirable frequency properties.

Suppose we seek a confidence interval $C_j^{\hat{M}}$ for the parameter $\beta_j^{\hat{M}}$. In the classical setting, we would desire

$$\mathbb{P}(\beta_j^{\hat{M}} \in C_j^{\hat{M}}) \geq 1 - \alpha,$$

but β_j^M is not defined when $j \notin M$, and thus this criterion is ill-posed (LSST, 2015); BBBZZ (2013) suggest two alternate frequency properties to consider:

Alternative 1

We only form a confidence interval for β_j^M if model M is selected ($\hat{M} = M$), and thus we should condition on this selection event and require that the confidence interval satisfy a *conditional coverage* property

$$\mathbb{P}(\beta_j^M \in C_j^M | \hat{M} = M) \geq 1 - \alpha. \tag{2}$$

This has the benefit of never comparing coefficients across different models.

- Fithian, Sun, Taylor (2015) motivate this criterion by *data splitting* as in Cox (1975), where half the data is used for selection and the remainder for inference.
- It is argued in FST (2015) that accepting data splitting as a valid approach leads to consideration of frequency properties *conditional on the model selected*, including conditional coverage and *selective type I error*.

Data splitting is common in practice, but is cautioned against in the post-selection inference literature, as it necessarily reduces the available sample information for both selection *and* inference, and is not always applicable (e.g. in time series models).

Alternative 2

One could also consider events defined simultaneously over all $j \in \hat{M}$. The proposal in BBBZZ (2013) is to control the familywise error rate (FWER),

$$\text{FWER} \equiv \mathbb{P}(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} \text{ for any } j \in \hat{M}).$$

This controls the probability of making *any* error. A less demanding criterion would be to control the expected proportion of errors. The false coverage-statement rate (FCR) was proposed in Benjamini & Yekutieli (2005) as

$$\text{FCR} \equiv \mathbb{E}\left[\frac{\left|\{j \in \hat{M} : \ \beta_j^{\hat{M}} \notin C_j^{\hat{M}}\}\right|}{\left|\hat{M}\right|}; \ \left|\hat{M}\right| > 0\right],$$

where $|\cdot|$ measures cardinality of a set and the error is set to zero when zero variables are selected $(\left|\hat{M}\right| = 0)$.

Continued

Lemma (LSST (2015))

Consider a family of intervals $\{C_j^{\hat{M}}\}_{j \in \hat{M}}$ *each having conditional coverage* $1-\alpha$,

$$\mathbb{P}(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} \mid \hat{M} = M) \leq \alpha, \quad \forall M, \, \forall j \in M.$$

Then $FCR \leq \alpha$.

This says that conditional coverage (2) implies FCR control. A more general result can be found in FST (2015). Thus it is sensible to control conditional coverage error.

Selective vs Simultaneous

Controlling the conditional coverage is an example of a *selective inference* procedure, whereas controlling the FCR is a type of *simultaneous inference*.

The simultaneous inference considered in BBBZZ (2013), termed 'post-selection inference' (PoSI) constructs simultaneous $1 - \alpha$ confidence intervals based on least-squares estimates for the parameters of *all* linear regression models that were ever considered.

Thus, regardless of how the model is chosen, one can control, at level α, the *overall* probability of constructing any incorrect (non-covering) interval.

Selective inference differs from simultaneous inference, in which all models considered by the variable selection procedure are considered relevant for inference.

In the *selective inference* paradigm, inference is conditional on the selection event, where the selection event means one particular outcome of the variable selection procedure, which may be interpreted as the **selection of which hypotheses to test**.

More on Selective Inference

Selective inference is actually a broader concept than suggested by the conditional coverage statement above and, in fact, there is a notion of unconditional inference within the selective inference paradigm (Tibshirani, Rinaldo, Tibshirani, Wasserman).

Many of the newer papers in this area can be classified as *selective pivotal inference*

More on Selective Inference

The selection of a model is a random event.

- helpful toy example: the set of selected variables in regression is a random set; hypotheses are only tested for selected variables, thus the hypotheses are random
- to condition on selection event, need to characterize this event in a manner suitable to uncertainty quantification

e.g. Lasso and forward stepwise partition $\mathbb{R}^n$ into convex polyhedra: if $y \in \text{ConvPoly}_m$, then model m is selected