Model selection bias invalidates goodness of fit tests

When the best isn't good enough

Joshua Loftus and Weichi Yao

New York University

NYU

DENVER, COLORADO
JSM2019
JULY 27–AUGUST 1

1. Motivating examples

2. Selective unbiasedness

3. High-dimensional regression

4. Conclusion

Motivating examples

Regression example: F-tests (of unselected variables)

- Regression models $E[Y] = X_A\beta_A$ for some subset A of columns of a matrix X.
- With nested subsets $A \subsetneq A'$, we'll conduct an F-test and consider this as a goodness-of-fit test for the model with variables A.
- In `R` we just use the `anova` function with these two linear models.
- "Model specification test," "omitted variable bias"

The distribution of the F-statistic is derived, of course, under the assumption that A and A' have been chosen *a priori*. . .

Regression example: F-tests (of unselected variables)

- Regression models $E[Y] = X_A\beta_A$ for some subset A of columns of a matrix X.
- With nested subsets $A \subsetneq A'$, we'll conduct an F-test and consider this as a goodness-of-fit test for the model with variables A.
- In `R` we just use the `anova` function with these two linear models.
- "Model specification test," "omitted variable bias"

The distribution of the F-statistic is derived, of course, under the assumption that A and A' have been chosen *a priori*...

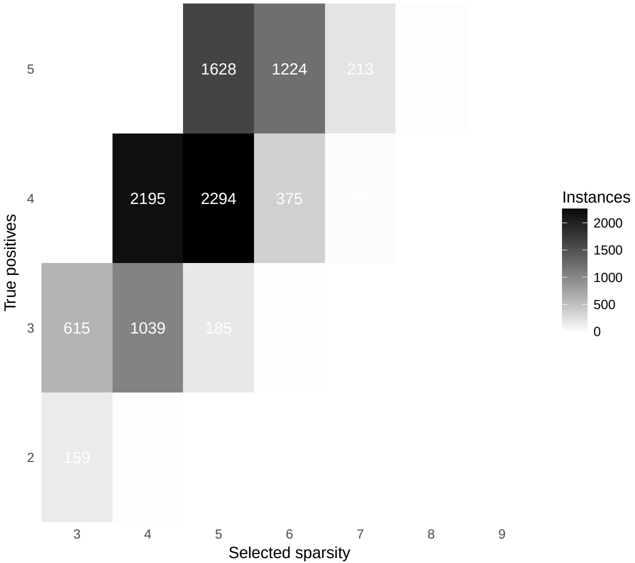
Model selection

When many predictors are available, we often use (automated) model selection procedures like forward stepwise or the LASSO (Tibshirani, 1996) to choose a "good" subset

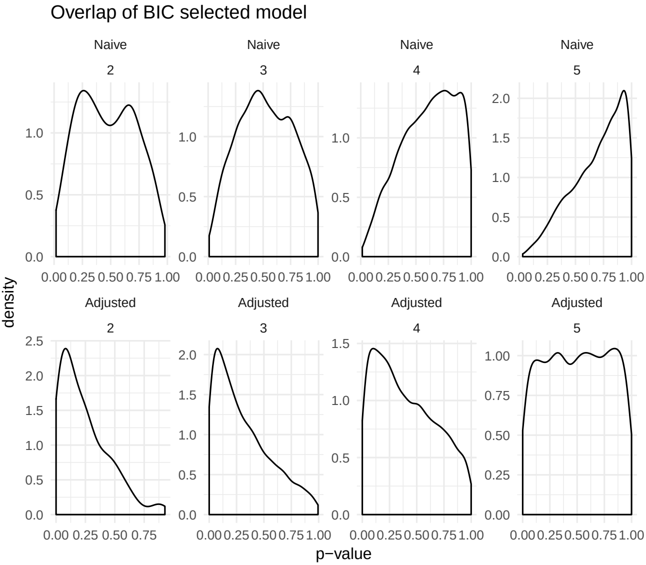
Regression variable selection

- For concreteness, consider selecting variables using forward stepwise with BIC, i.e. in `R` with `step(..., k = log(n))`.
- Simulation with $n = 100$ observations of $p = 10$ variables, the first two coefficients are larger than the next 3, and the last 5 are all 0.
- In this (low-dimensional) example, we'll take $A' = \{1, \ldots, 10\}$ for simplicity.
- Consider the F-test as a goodness-of-fit test for the selected $A \subset A'$, and compute both unadjusted (classical) and adjusted (selective) p-values.

Profile of model selection events



Distributions of p-values for full-model F-tests



Probability of rejection

Table 1: Probability of rejection at level 0.1, conditional on size of overlap

pvalue	overlap	Pr(reject)
Naive	2	0.056
Naive	3	0.032
Naive	4	0.017
Naive	5	0.005
Adjusted	2	0.328
Adjusted	3	0.251
Adjusted	4	0.156
Adjusted	5	0.101

Conditional power

Table 2: Probability of rejection at level 0.1, conditional on overlap less than 5

pvalue	Pr(reject)
Naive	0.022
Adjusted	0.186

Simple time series example: the AR(p) process

Autoregressive model set up

An AR(p) process $\{Y_t\}$ with mean μ satisfies

$$(Y_t - \mu) - \phi_1(Y_{t-1} - \mu) - ... - \phi_p(Y_{t-p} - \mu) = W_t$$

where $W_t \sim \mathcal{N}(0, \sigma^2)$. For simplicity, we consider the zero-mean autoregressive process with $\mu = 0$

$$Y_t - \phi_1 Y_{t-1} - ... - \phi_p Y_{t-p} = W_t,$$

where $W_t \sim \mathcal{N}(0, \sigma^2)$.

Example AR(p) paths

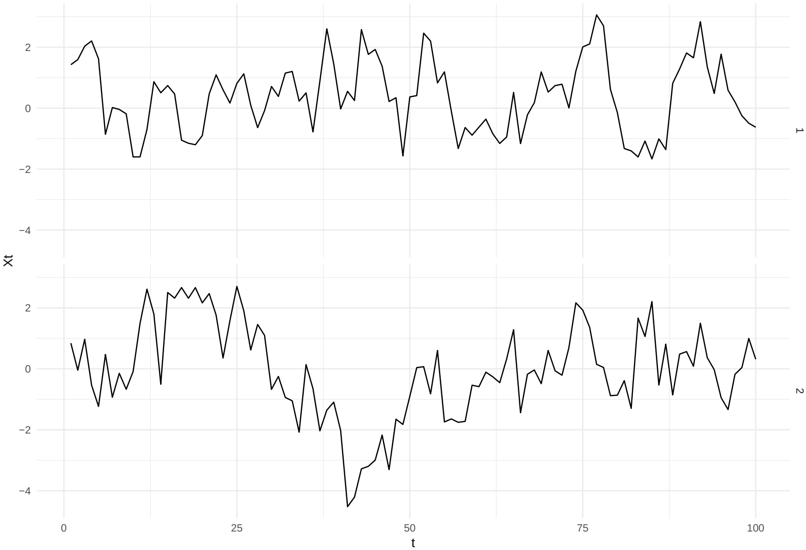


Figure 1: Top panel: AR(1), bottom panel: AR(2)

Goodness of fit test

Perhaps the most commonly used goodness of fit test in time series:

Ljung-Box test (Ljung & Box, 1978)

The Ljung-Box statistic with l lags is defined as

$$Q^{(l)}(\hat{r}) = n(n+2)\sum_{k=1}^{l}(n-k)^{-1}\hat{\mathrm{r}}(k)^2$$

where

$$\hat{\mathrm{r}}(k) = \frac{\sum_{t=k+1}^{n}(Y_t - \hat{Y}_t)(Y_{t-k} - \hat{Y}_{t-k})}{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}, \quad k = 1, ..., m$$

is the autocorrelation function.

If the data is truly AR(p) but we fit AR(q), with $q < p$, the LB test is powered to detect the residual^2 autocorrelation.

Observed LB vs null distribution, with selection

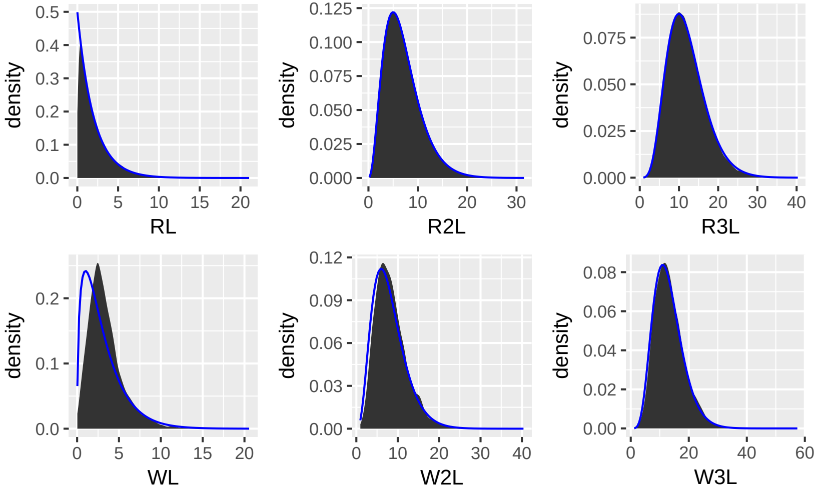


Figure 2: Truth is AR(3). Top row: when AIC_C selects $p = 3$. Bottom row: when $p = 2$ is selected. Distributions shown for LB with lags 5, 10, 15.

Model selection bias

- Model selection bias invalidates many hypothesis tests
- Previously, **significance** of selected variables: **anti-conservative**
- This work on **goodness of fit** tests: **conservative**
- *Conditional on selecting wrong model, goodness-of-fit tests have low power*

Intuitively obvious: use the data to pick the model, and then use the same data to answer “does this model fit the data?”

Next: one solution approach (used to compute adjusted p-values in regression example)

Selective unbiasedness

When the best isn’t good enough

- We’ve seen a few examples but it should be clear there are far more we haven’t mentioned
- Many combinations of model selection procedures and goodness of fit tests (or other diagnostics)
- In many such examples, such a test has power less than α conditional on selecting the wrong model

Pretty troubling! Goodness-of-fit tests worse at detecting the wrong model has been selected than just tossing an α-coin. . . What can we do?

Selective unbiasedness

Selective unbiasedness (Fithian, Sun, Taylor, 2014)

We say that a test is selectively unbiased if for any selected model *m* and alternative hypothesis $H_1(m)$,

$$P_{H_1(m)}(\text{reject } H_0 | M = m) \geq \alpha$$

We achieve this by using the conditional (truncated) distribution of the test statistic (if we can derive/compute it). *Same method used to control selective type 1 error will make (some) goodness-of-fit tests selectively unbiased*

Marginal screening example

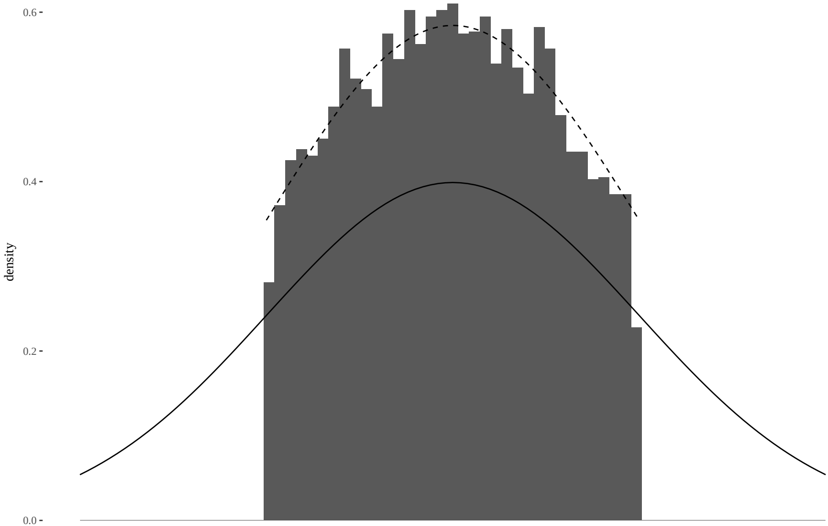
Consider a simpler problem of selecting marginal effects. From many independent effects, screen out those with small observed values, i.e. select those with large values (in this case bigger than 1)

```r
Z <- rnorm(10000)
unselected_Z <- data.frame(Z = Z[abs(Z) < 1])
```

These are generated under the global null, and $Z : |Z| > 1$ are selected. How can we test the "goodness of fit" of selecting these effects? Test based on unselected effects.

Conditional null distribution is truncated: $Z||Z| < 1$

```r
truncated_Z_pdf <- function(z) dnorm(z)/(pnorm(1)-pnorm(-1))
# plot code hidden
```



Interpretation/storytelling

- Keep model selection and inference compartmentalized
- The procedure/algorithm selects model m, this is evidence in favor of the model
- If model m contains variable j, this is evidence in favor of variable j
- We may still fail to reject the null hypothesis that $\beta_{m,j} = 0$, "the best model includes this variable but we cannot reject the null for it"
- Or we may reject the goodness of fit null: "this is the best model (in the class) but it still has specification problems"

"This is the best model we could find, and these inferences may be useful for understanding its limitations and directing future study"

Example: follow-up study after Benjamini-Hochberg

Idea: after using the BH(q) procedure to select a subset of hypotheses while controlling the false discovery rate, decide whether to conduct a follow-up study of the hypotheses that were *not* selected

Use the conditional distribution of the non-rejected $p_{(j)}$ to adjust e.g. Fisher's combination test

Interesting meta-analysis implications

Power of Fisher's combination test

Table 3: Probability of rejection at level 0.05 with different Beta$(1, \mu)$

μ	$p_{1\|0}$	unadjusted	adjusted
10	97	1.0000	1.0000
20	87	0.9614	1.0000
30	60	0.5091	0.9838
40	33	0.0900	0.8580
50	19	0.0040	0.6080
60	11	0.0000	0.3320
70	6	0.0000	0.2004
80	4	0.0000	0.1245
90	3	0.0000	0.0995

$p_{1|0}$ denotes the average proportion of true nonnulls that are not rejected

Conditioning on selection

- $M = M(y)$ model selection (e.g. marginal screening, forward stepwise)
- Observe $M = m$
- Test $H_0(m)$ with statistic T conditioned on $\{y : M(y) = m\}$

Selective type 1 error (Fithian, Sun, Taylor, 2014)

Reject $H_0(m)$ if $T > c_\alpha(m)$, where

$$P_{H_0(m)}(T > c_\alpha(m) \mid M = m) \leq \alpha$$

Conditioning on selection

- $M = M(y)$ model selection (e.g. marginal screening, forward stepwise)
- Observe $M = m$
- Test $H_0(m)$ with statistic T conditioned on $\{y : M(y) = m\}$

Selective type 1 error (Fithian, Sun, Taylor, 2014)

Reject $H_0(m)$ if $T > c_\alpha(m)$, where

$$P_{H_0(m)}(T > c_\alpha(m) \mid M = m) \leq \alpha$$

In practice we may also condition on other things (sufficient statistics for nuisance parameters)

Growing literature (including e.g. details of truncation regions) for various selection methods in various settings

Related literature

- Early works by Olshen (1973), Hurvich & Tsai (1990), Benjamini & Yekutieli (2005) used basic idea of conditioning to adjust for selection
- Methods controlling FDR or similar: "Screen & clean" Wasserman & Roeder (2009), "Stability selection" Meinshausen & Bühlmann (2010), an Empirical Bayes approach Efron (2011), "SLOPE" Bogdan et al. (2014), and "Knockoffs" Barber & Candès (2015)
- Controlling type 1 error: Debiasing methods Bühlmann (2013); Javanmard & Montanari (2014); Zhang & Zhang (2014), Causal inference for univariate treatment Belloni et al. (2014), PoSI: simultaneous for all submodels Berk et al. (2013), Impossibility results: Leeb & Pötscher (2005, 2006)
- Goodness of fit in high dimensions: Verzelen & Villers (2008), Nan & Yang (2014), Shah & Bühlmann (2017)

Conditional on selection literature

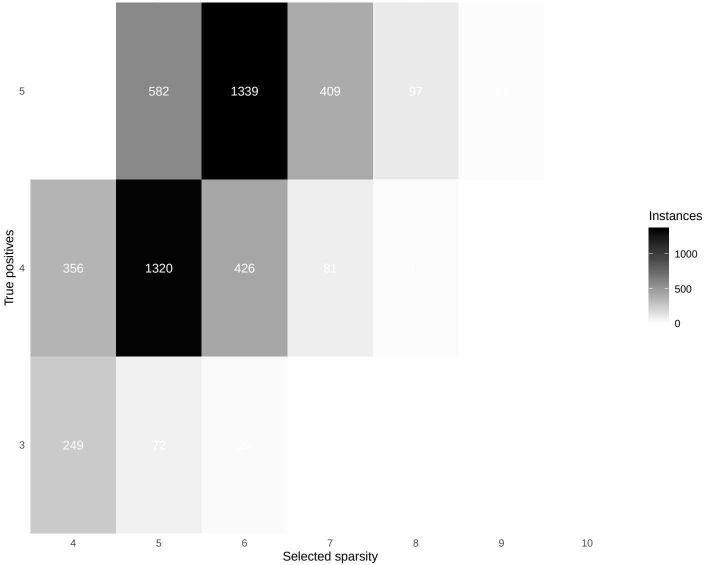
- Lasso, sequential Lockhart et al. (2014)
- General penalty, global null, geometry *Taylor et al. (2015)*; Azaïs et al. (2015)
- Forward stepwise, sequential *Loftus & Taylor (2014)*
- Matrices: PCA/CCA Choi et al. (2014)
- **Fixed λ Lasso / conditional Lee et al. (2015)**
- Framework, optimality Fithian et al. (2014)
- Forward stepwise and LAR Tibshirani et al. (2014)
- Unknown σ^2 *Tian et al. (2018)*; Gross et al. (2015)
- **Group selection / unknown σ^2 *Loftus & Taylor (2015)***
- Cross-validation Tian & Taylor (2015); *Loftus (2015)*; Markovic et al. (2017)

High-dimensional regression

Regression variable selection

- As before, but now $n = 100$ observations and $p = 200$ variables
- Can't use AIC/BIC, so we use RIC (Foster & George, 1994), i.e. in `R` with `step(..., k = 2log(p))`.
- Simulation with $n = 100$ observations of $p = 10$ variables, the first two coefficients are larger than the next 3, and the last 5 are all 0.
- Can't use full model as the alternative, so we'll find A' by letting forward stepwise add the next 5 variables

Profile of model selection events



Probability of rejection

Table 4: Probability of rejection at level 0.1

pvalue	overlap	Pr(reject)
Naive	5	$>$0.99
Naive	$<$5	$>$0.99
Adjusted	5	0.103
Adjusted	$<$5	0.134

Conclusion

The “quiet scandal” in statistics (Leo Breiman)

People use the data to make modeling choices

This can bias (all sorts of) inferences

Conditioning on those choices and using truncated distributions is a simple idea with wide applicability, including to goodness of fit tests

Thank you!

Preprint should be posted soon, more procedures added to `selectiveInference` package on CRAN as we progress.

joshualoftus.com