Variable Selection Uncertainty in Linear Models

Aldo Solari

Joint work with Ningning Xu and Jelle Goeman

Padua, June 16, 2017

1. **Variable Selection Algorithms**

2. **Variable Selection Uncertainty**

3. **Predictive modeling**

4. **Explanatory modeling**

5. **Discussion**

Outline

1. **Variable Selection Algorithms**
2. Variable Selection Uncertainty
3. Predictive modeling
4. Explanatory modeling
5. Discussion

Classic example

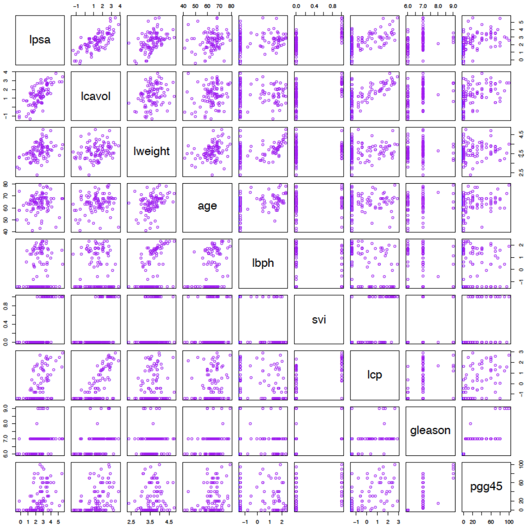


FIGURE 1.1. *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* `svi` *and* `gleason`*, are categorical.*

Prostate cancer data

Response

`lpsa` for $n = 67$ subjects

Predictors

1, `lcavol`, `lweight`, `age`, `lbph`, `svi`, `lcp`, `gleason`, `pgg45`

Variable selection problem

Select the "optimal" subset of predictors

Number of possible subsets

$2^p = 256$

Selection algorithms

	C_P	BIC	LASSO	FS
`lcavol`	•	•	•	•
`lweight`	•	•	•	•
`age`	•			
`lbph`	•		•	
`svi`	•		•	•
`lcp`	•			
`gleason`				
`pgg45`	•			

C_P	best subsets selection with min C_P/AIC
BIC	best subsets selection with min BIC
LASSO	10-fold CV with 1-SE rule (Hastie et al. 2009)
FS	forward stop rule on LAR path at 10% FDR (G'Sell et al. 2016)

Post-selection inference

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.32592    0.77998  -0.418   0.6775    
lcavol       0.50552    0.09256   5.461 8.85e-07 ***
lweight      0.53883    0.22071   2.441   0.0175 *  
lbph         0.14001    0.07041   1.988   0.0512 .  
svi          0.67185    0.27323   2.459   0.0167 *  

Residual standard error: 0.7275 on 62 degrees of freedom
Multiple R-squared:  0.6592,Adjusted R-squared:  0.6372 
F-statistic: 29.98 on 4 and 62 DF,  p-value: 6.911e-14
```

Naïve interpretation of results

Variables in the selected model

Important

Importance quantified by p-values/confidence intervals for the coefficients, which are calculated without taking into account the selection

Variables not in the selected model

Not important

Interpreted as if they had coefficients $= 0$

A "quiet scandal"

Breiman (1992) referred to this naïve post-selection inference as a "quiet scandal" in the statistical community

Two problems

Post-selection inference

Significance of each variable within the selected model, taking into account the selection

- PoSI (Berk et al., 2013)

Comparison of models

Many-to-one comparisons with the full model

- Adequate models (Mallows, 1973; Spjøtvoll, 1977)
- Primitive models (Cox and Snell, 1974)

Selection algorithm = point estimation

$y \sim 1 + x_1 + x_2 + x_3$

$\widehat{\text{Best}}$

$\boxed{y \sim 1 + x_1 + x_2}$

$y \sim 1 + x_1 + x_3$

$y \sim 1 + x_2 + x_3$

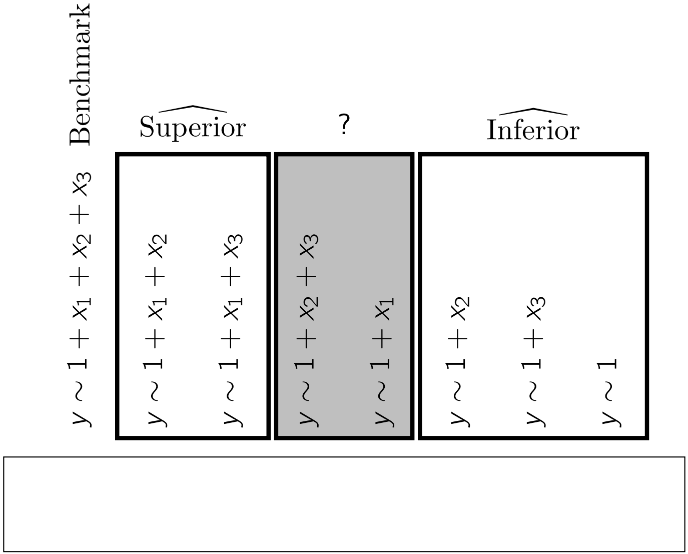
$y \sim 1 + x_1$

$y \sim 1 + x_2$

$y \sim 1 + x_3$

$y \sim 1$

Confidence set = uncertainty



Outline

1. Variable Selection Algorithms
2. **Variable Selection Uncertainty**
3. Predictive modeling
4. Explanatory modeling
5. Discussion

Linear model

$$y \sim \mathcal{N}(\mu, \sigma^2 I_n)$$

- $y \in \mathbb{R}^n$: random response
- $\mu \in \mathbb{R}^n$, $\sigma^2 > 0$: unknown parameters

Design matrix

- $X \in \mathbb{R}^{n \times p}$: design matrix
- $\mathrm{range}(X)$: column space of X

First-order misspecification

- $\mu \in \mathrm{range}(X)$ iff $\mu = X\beta$: correct (unbiased) full model
- $\mu \notin \mathrm{range}(X)$: first-order misspecification
- "If all models are wrong, the practical question is how wrong do they have to be to not be useful" (Box and Draper, 1986)

Models

- $M \subseteq F = \{1, \ldots, p\}$ with $\# M = m$
- $X_M \in \mathbb{R}^{n \times m}$: model M design matrix
- Orthogonal projector onto $\mathrm{range}(X_M)$

$$P_M = X_M (X_M^\top X_M)^{-1} X_M^\top$$

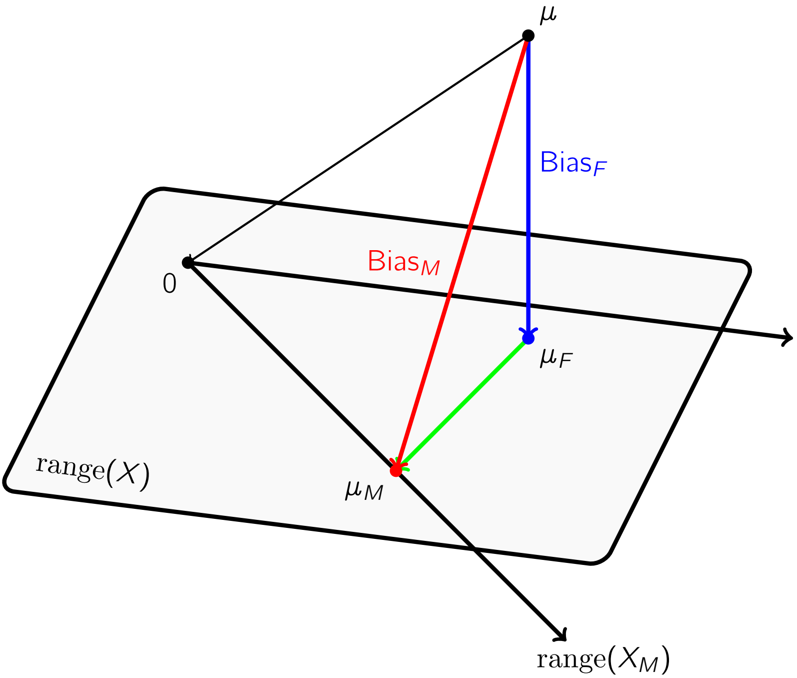
- Model M estimator

$$\hat{\mu}_M \sim \mathcal{N}(\mu_M, \sigma^2 P_M)$$

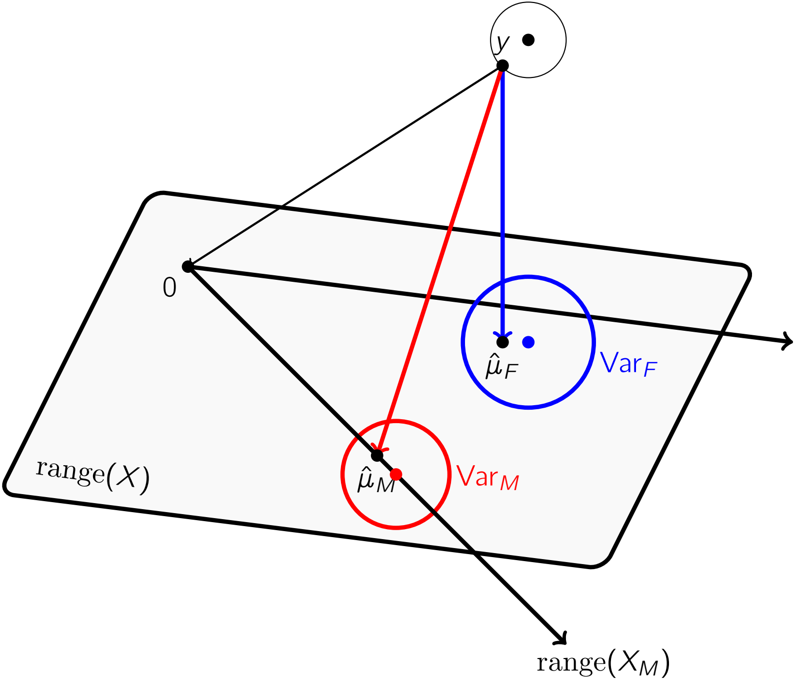
  with $\hat{\mu}_M = P_M y$ and $\mu_M = P_M \mu$
- Candidate models

$$\mathcal{M} = \{M : M \subseteq F\}$$

  with $\# \mathcal{M} = 2^p$



μ
Bias_F
Bias_M
0
μ_F
μ_M
range(X)
range(X_M)


y
0
$\hat{\mu}_F$
Var_F
$\hat{\mu}_M$
Var_M
$\mathrm{range}(X)$
$\mathrm{range}(X_M)$

To explain or to predict?

Explanatory modeling

- Obtain the most accurate representation of the underlying theory
- Avoid/minimize Bias
- Omitted-variable bias compromises interpretation

Predictive modeling

- Generate good predictions of new y
- minimize $\text{Bias}^2 + \text{Variance}$
- A biased model can predict better than an unbiased one

Outline

1. Variable Selection Algorithms
2. Variable Selection Uncertainty
3. **Predictive modeling**
4. Explanatory modeling
5. Discussion

Inferior and superior models

Mean Squared Error

$$\frac{\mathrm{MSE}_M}{\sigma^2} = \lambda_M + m$$

where $\lambda_M = \dfrac{\|\mu_M - \mu\|^2}{\sigma^2}$

Relative efficiency

$$\mathrm{e}_M = \frac{\mathrm{MSE}_M}{\mathrm{MSE}_F} = \frac{\lambda_M + m}{\lambda_F + p} > 1 \quad \text{iff} \quad \lambda_M^F = \frac{\|\mu_M - \mu_F\|^2}{\sigma^2} > p - m$$

where $\lambda_M = \lambda_M^F + \lambda_F$ (Pythagoras' theorem)

Inferior and superior models

- $\mathcal{I} = \{M \in \mathcal{M} : \lambda_M^F > p - m\}$
- $\mathcal{S} = \{M \in \mathcal{M} : \lambda_M^F \leq p - m\}$

Hypothesis testing

One true hypothesis

$M \in \mathcal{I} : \lambda_M^F > p - m$ or $M \in \mathcal{S} : \lambda_M^F \leq p - m$

Testing for superiority

- Null $M \in \mathcal{I}$ against alternative $M \in \mathcal{S}$
- If $M \in \mathcal{I}$ rejected at level α, then $M \in \hat{\mathcal{S}}_\alpha$

Testing for inferiority

- Null $M \in \mathcal{S}$ against alternative $M \in \mathcal{I}$
- If $M \in \mathcal{S}$ rejected at level α, then $M \in \hat{\mathcal{I}}_\alpha$

Uncertainty

If both nulls $M \in \mathcal{S}$ and $M \in \mathcal{I}$ not rejected at α, then $M \in \hat{\mathcal{U}}_\alpha$

Confidence sets

$1 - \alpha$ **confidence of no type I errors**

$$\mathrm{P}(\{\hat{\mathcal{S}}_\alpha \cap \mathcal{I} = \emptyset\} \cap \{\hat{\mathcal{I}}_\alpha \cap \mathcal{S} = \emptyset\}) \geq 1 - \alpha$$

Familywise error control

The probability of at least one type I error in testing the family of 2^{p+1} null hypotheses $\{(M \in \mathcal{I}, M \in \mathcal{S}), M \in \mathcal{M}\}$ should be at most α

Uncertainty set

$$\hat{\mathcal{U}}_\alpha = \mathcal{M} \setminus (\hat{\mathcal{S}}_\alpha \cup \hat{\mathcal{I}}_\alpha)$$

Correct full model assumption

Assumption $*$

$$* : \lambda_F = 0 \text{ iff } \mu \in \mathrm{range}(X)$$

$$\mathrm{e}^*_M = \frac{\lambda^F_M + m}{p} > 1 \quad \text{iff} \quad \lambda^F_M = \frac{\|\mu_M - \mu_F\|^2}{\sigma^2} > p - m$$

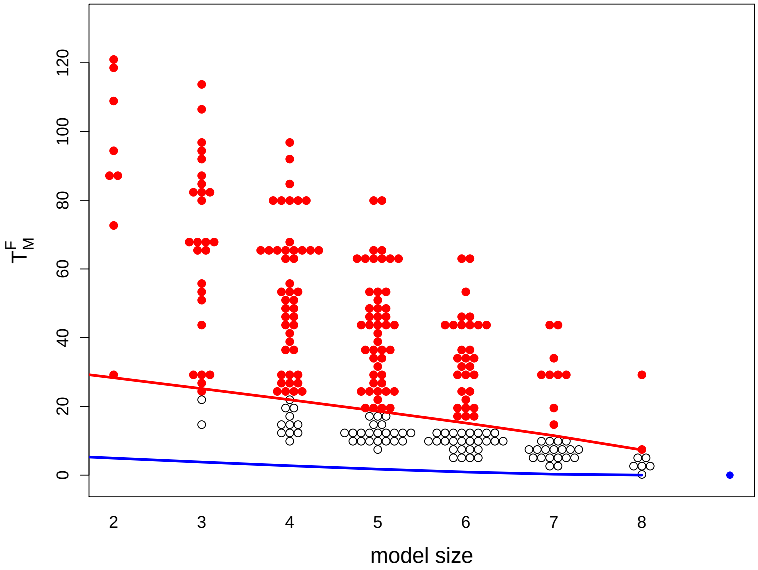
$\mathcal{F}$ test statistic

$$T^F_M = \frac{\|\hat{\mu}_M - \hat{\mu}_F\|^2}{\hat{\sigma}^2_F} \overset{*}{\sim} (p-m)\mathcal{F}'_{p-m,n-p}(\lambda^F_M)$$

where $\hat{\sigma}^2_F = \dfrac{\|\hat{\mu}_F - y\|^2}{n-p}$ is the full model estimator of σ^2

Reject $M \in \mathcal{S} : \lambda^F_M \leq p - m$ if $T^F_M > (p-m){f'}^{1-\alpha}_{p-m,n-p}(p-m)$

Reject $M \in \mathcal{I} : \lambda^F_M > p - m$ if $T^F_M < (p-m){f'}^{\alpha}_{p-m,n-p}(p-m)$



0
20
40
60
80
100
120
T_M^F
2
3
4
5
6
7
8
model size

Assumption $*$ - free

$\mathcal{F}$ test statistic

$$T_M^F = \frac{\|\hat{\mu}_M - \hat{\mu}_F\|^2}{\hat{\sigma}_F^2} \sim (p-m)\mathcal{F}''_{p-m,n-p}(\lambda_M^F, \lambda_F)$$

Stochastic order

$$? \overset{\text{st}}{\leq} \mathcal{F}''_{p-m,n-p}(\lambda_M^F, \lambda_F) \overset{\text{st}}{\leq} \mathcal{F}'_{p-m,n-p}(\lambda_M^F)$$

Conservative testing for inferiority

Reject $M \in \mathcal{S} : \lambda_M^F \leq p-m$ if $T_M^F > (p-m){f'}^{1-\alpha}_{p-m,n-p}(p-m)$

Scheffé's method

Maximum test statistic

$$T_{\varnothing}^{F} = \max_{M \in \mathcal{M}} T_M^F \sim p\mathcal{F}'_{p,n-p}(\lambda_{\varnothing}^F, \lambda_F)$$

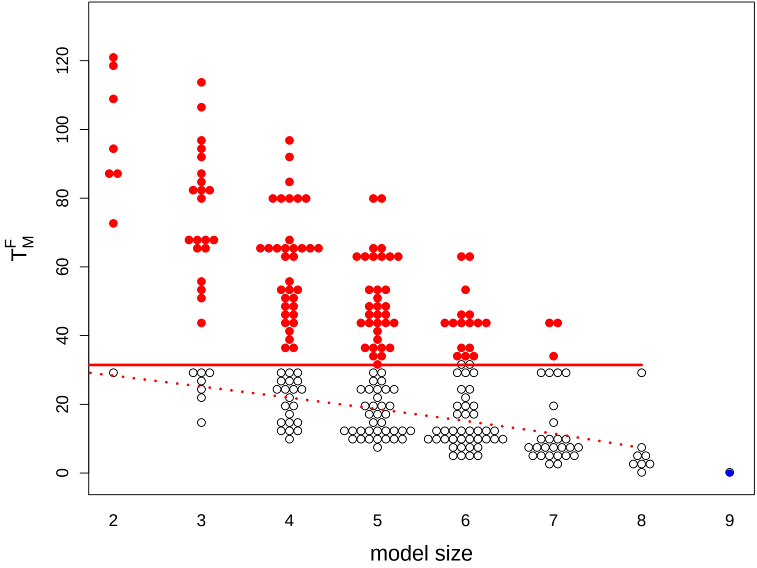
Simultaneous testing for inferiority

Reject $M \in \mathcal{S} : \lambda_M^F \leq p - m$ if $T_M^F > p f'^{1-\alpha}_{p,n-p}(p)$

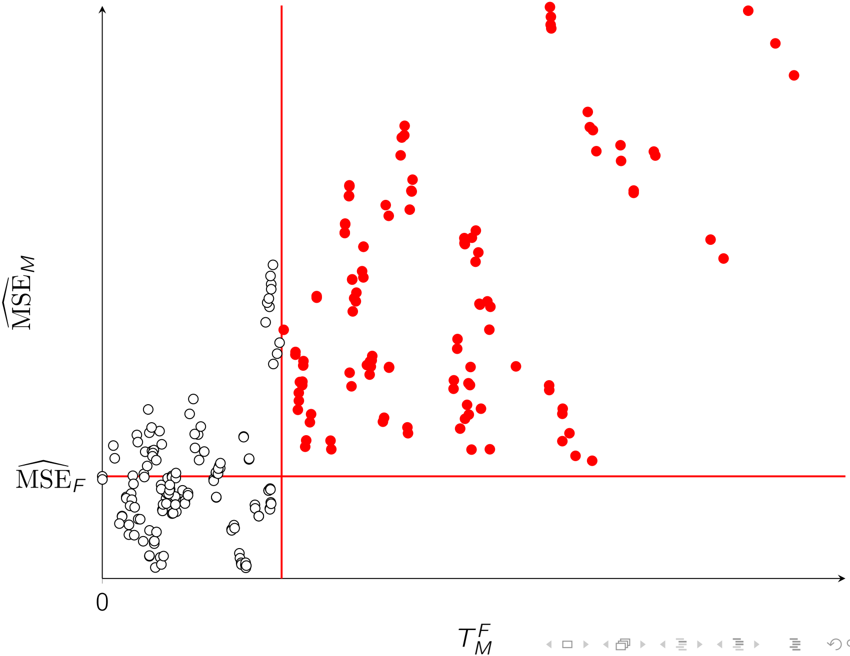
$1 - \alpha$ **confidence of no type I errors**

$$\mathrm{P}(\hat{\mathcal{I}}_\alpha \cap \mathcal{S}) \geq 1 - \alpha$$

Inferior models: $u_{5\%} = 54.1\ \%$



Predictions



Outline

1. Variable Selection Algorithms
2. Variable Selection Uncertainty
3. Predictive modeling
4. **Explanatory modeling**
5. Discussion

Explanatory modeling

Correct (unbiased) models

$\mathcal{C} = \{M \in \mathcal{M} : \lambda_M = 0\}$

Wrong (biased) models

$\mathcal{W} = \{M \in \mathcal{M} : \lambda_M > 0\}$

Testing for correctness?

Null $M \in \mathcal{W}$ against point alternative $M \in \mathcal{C}$ implies α power

Confidence about wrong models only

$\mathrm{P}(\hat{\mathcal{W}}_\alpha \cap \mathcal{C} = \emptyset) \geq 1 - \alpha$

More power

$\mathcal{C} \subseteq \mathcal{S}$

$\mathcal{W} \supseteq \mathcal{I}$ implies a more powerful confidence set $\hat{\mathcal{W}}_\alpha \supseteq \hat{\mathcal{I}}_\alpha$

Adequate Models

Adequate models

$\mathcal{A} = \{M \in \mathcal{M} : \ \lambda_M^F = 0\}$

Non-adequate models

$\mathcal{B} = \{M \in \mathcal{M} : \ \lambda_M^F > 0\}$

Relationships

$\mathcal{C} \subseteq \mathcal{A} \subseteq \mathcal{S} \quad \text{and} \quad \mathcal{I} \subseteq \mathcal{B} \subseteq \mathcal{W}$

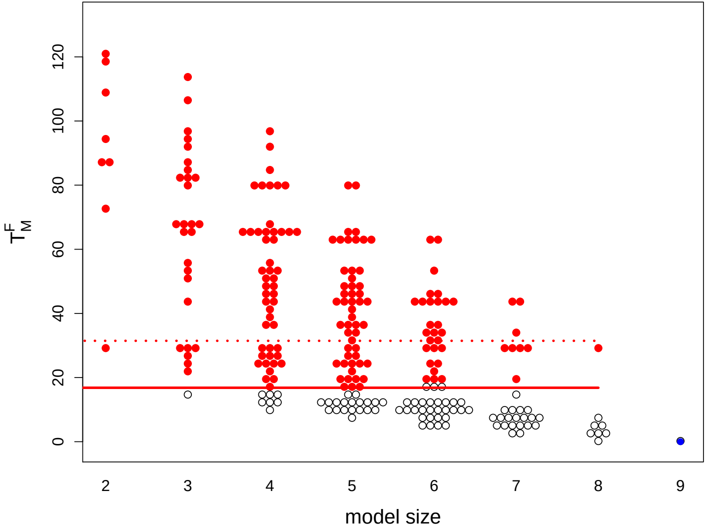
Mallows (1973)

Assumption $*$ and Scheffé's method

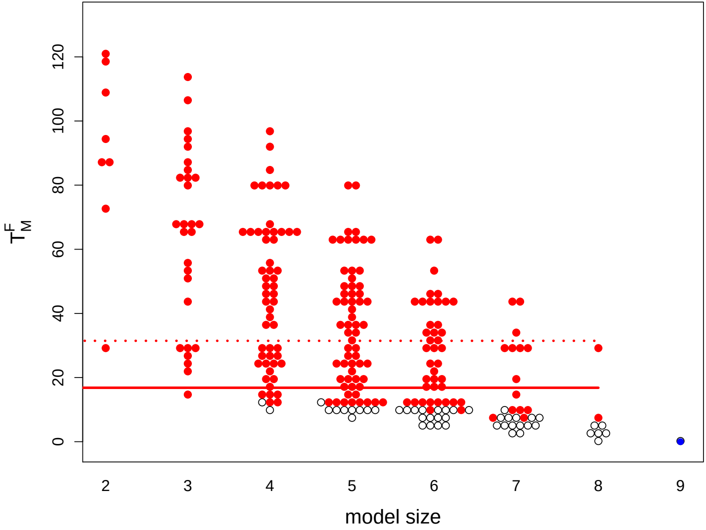
Spjøtvoll (1977)

Assumption $*$ and closed testing method

Scheffé's method: $u_{5\%} = 32.5$ %



Closed testing method: $u_{5\%} = 18.8$ %



Summary

Training set: $n = 67$

Inference	*Method*	*Size*	*Uncertain*	$u_{5\%}$
Inferior	Scheffé	117	138	54.1 %
Non-adequate	Scheffé	172	83	32.5 %
Non-adequate	Closed testing	207	48	18.8 %

Training + test: $n = 97$

Inference	*Method*	*Size*	*Uncertain*	$u_{5\%}$
Inferior	Scheffé	136	119	46.6 %
Non-adequate	Scheffé	179	76	29.8 %
Non-adequate	Closed testing	223	32	12.5 %

Outline

1. Variable Selection Algorithms
2. Variable Selection Uncertainty
3. Predictive modeling
4. Explanatory modeling
5. **Discussion**

Alternative assumptions

Known σ^2 (Tibshirani et al. 2016)

- $T_M = \|\hat{\mu}_M - y\|^2/\sigma^2 \sim \chi^2_{n-m}(\lambda_M)$
- $T_M^F = \|\hat{\mu}_M - \hat{\mu}_F\|^2/\sigma^2 \sim \chi^2_{p-m}(\lambda_M^F)$

Estimator $\hat{\sigma}^2$ (Berk et al. 2011)

- $\hat{\sigma}^2$ with $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$, $\hat{\sigma}^2 \perp y$ and g degrees of freedom
- $T_M = \|\hat{\mu}_M - y\|^2/\hat{\sigma}^2 \sim (n-m)\mathcal{F}'_{n-m,g}(\lambda_M)$
- $T_M^F = \|\hat{\mu}_M - \hat{\mu}_F\|^2/\hat{\sigma}^2 \sim (p-m)\mathcal{F}'_{p-m,g}(\lambda_M^F)$

Alternative null hypotheses

Forward stepwise testing (G'Sell et al. 2016)

- Nested models: $\varnothing \subset M_1 \subset \ldots \subset M_s \subset \ldots \subset M_{p-1} \subset F$
- Complete null : $\lambda_{M_s} = 0$ for $s = 1, 2, \ldots, p$
- Incremental null : $\lambda_{M_{s-1}} = \lambda_{M_s}$ for $s = 1, 2, \ldots, p$

PoSI (Berk et al. 2011)

- $\lambda_{M\setminus\{j\}} = \lambda_M$ for every $j \in M$ and $M \in \mathcal{M}$
- Power improvement with closed testing is possible

Discussion

NP-hard approach

Shortcuts needed to reduce exponential complexity

High-dimensional data

$p \gg n$: change of paradigm?

All comparisons

Ranking of models e.g. $\mathrm{rank}(F) \in [1, \ldots, 139]$

Measuring uncertainty

is a statistician's task.

Bibliography

Berk, Brown, Buja, Zhang, and Zhao (2013)
Valid post-selection inference
Annals of Statistics, 41:802–837

Mallows (1973)
Some comments on C_P
Technometrics, 15:661-765

Spjøtvoll (1977)
Alternatives to plotting C_P in multiple regression
Biometrika, 64:1–8