“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”)

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

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"

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,"

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:

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.

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:

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, ...

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, ...

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."

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:

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.

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.

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?

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.

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.

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.
- Some forms of the problem has been known for decades:

  Koopmans (1949); Buehler and Fedderson (1963); Brown (1967); and Olshen (1973); Sen (1979); Sen and Saleh (1987); Dijkstra and Veldkamp (1988); Arabatzis et al. (1989); Hurvich and Tsai (1990); Regal and Hook (1991); Pötscher (1991); Chiou and Han (1995a,b); Giles (1992); Giles and Srivastava (1993); Kabaila (1998); Brockwell and Gordeon (2001); Leeb and Pötscher (2003; 2005; 2006a; 2006b; 2008a; 2008b); Kabaila (2005); Kabaila and Leeb (2006): Berk, Brown and Zhao (2009); Kabaila (2009).

Example: Length of Criminal Sentence

Question: What covariates predict length of a criminal sentence best?

A small empirical study:

- $N = 250$ observations.

Example: Length of Criminal Sentence

Question: What covariates predict length of a criminal sentence best?

A small empirical study:

- $N = 250$ observations.
- Response: `log-length of sentences`

Example: Length of Criminal Sentence

Question: What covariates predict length of a criminal sentence best?

A small empirical study:

- $N = 250$ observations.
- Response: `log-length of sentences`
- $p = 11$ covariates (predictors, explanatory variables):
  - `race`
  - `gender`
  - `initial age`
  - `marital status`
  - `employment status`
  - `seriousness of crime`
  - `psychological problems`
  - `education`
  - `drug related`
  - `alcohol usage`
  - `prior record`

Example: Length of Criminal Sentence

Question: What covariates predict length of a criminal sentence best?

A small empirical study:

- $N = 250$ observations.
- Response: `log-length of sentences`
- $p = 11$ covariates (predictors, explanatory variables):
  - `race`
  - `gender`
  - `initial age`
  - `marital status`
  - `employment status`
  - `seriousness of crime`
  - `psychological problems`
  - `education`
  - `drug related`
  - `alcohol usage`
  - `prior record`
- What variables should be included?

Example: Length of Criminal Sentence (contd.)

- All-subset search with BIC chooses a model $\hat{\mathrm{M}}$ with seven variables:
  - `initial age`
  - `gender`
  - `employment status`
  - `seriousness of crime`
  - `drugs related`
  - `alcohol usage`
  - `prior records`

Example: Length of Criminal Sentence (contd.)

- All-subset search with BIC chooses a model $\hat{\mathrm{M}}$ with seven variables:
  - `initial age`
  - `gender`
  - `employment status`
  - `seriousness of crime`
  - `drugs related`
  - `alcohol usage`
  - `prior records`
- t-statistics of selected covariates, in descending order:
  - $|t_{\texttt{alcohol}}| = 3.95$;
  - $|t_{\texttt{prior records}}| = 3.59$;
  - $|t_{\texttt{seriousness}}| = 3.57$;
  - $|t_{\texttt{drugs}}| = 3.31$;
  - $|t_{\texttt{employment}}| = 3.04$;
  - $|t_{\texttt{initial age}}| = 2.56$;
  - $|t_{\texttt{gender}}| = 2.33$.

Example: Length of Criminal Sentence (contd.)

- All-subset search with BIC chooses a model $\hat{\mathrm{M}}$ with seven variables:
  - `initial age`
  - `gender`
  - `employment status`
  - `seriousness of crime`
  - `drugs related`
  - `alcohol usage`
  - `prior records`
- t-statistics of selected covariates, in descending order:
  - $|t_{\texttt{alcohol}}| = 3.95$;
  - $|t_{\texttt{prior records}}| = 3.59$;
  - $|t_{\texttt{seriousness}}| = 3.57$;
  - $|t_{\texttt{drugs}}| = 3.31$;
  - $|t_{\texttt{employment}}| = 3.04$;
  - $|t_{\texttt{initial age}}| = 2.56$;
  - $|t_{\texttt{gender}}| = 2.33$.
- Can we use the cutoff $t_{.975,250-8} = 1.97$?

Linear Model Inference and Variable Selection

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ = fixed design matrix, $N \times p$, $N > p$, full rank.
- $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$

In textbooks:

1. Variables selected
2. Data seen
3. Inference produced

In common practice:

1. Data seen
2. Variables selected
3. Inference produced

Linear Model Inference and Variable Selection

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ = fixed design matrix, $N \times p$, $N > p$, full rank.
- $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$

In textbooks:

1. Variables selected
2. Data seen
3. Inference produced

In common practice:

1. Data seen
2. Variables selected
3. Inference produced

Is this inference valid?

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

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.

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.

Evidence from a Simulation

Generate Y from the following linear model:

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

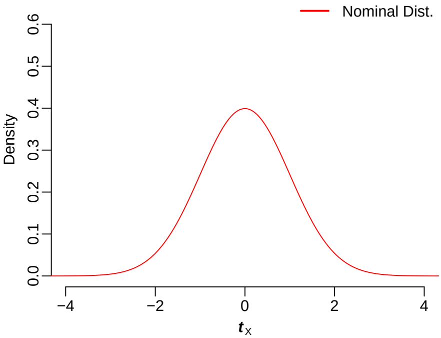
where $p = 11$, $N = 250$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ iid.

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.

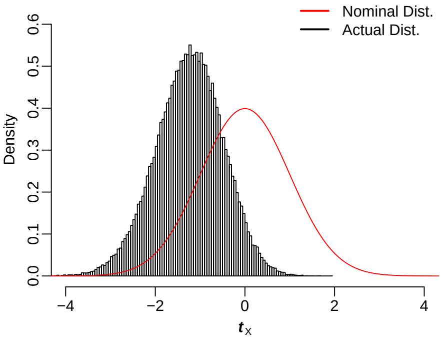
Evidence from a Simulation (contd.)

Marginal Distribution of Post-Selection t-statistics:



Evidence from a Simulation (contd.)

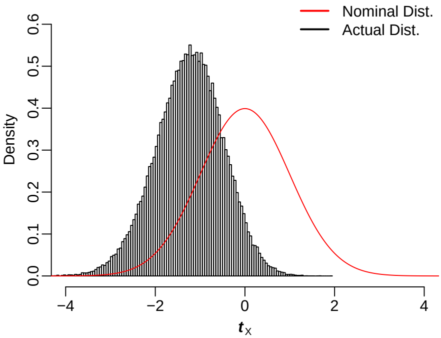
Marginal Distribution of Post-Selection *t*-statistics:



- The overall coverage probability of the conventional post-selection CI is $83.5\% < 95\%$.

Evidence from a Simulation (contd.)

Marginal Distribution of Post-Selection *t*-statistics:



- The overall coverage probability of the conventional post-selection CI is $83.5\% < 95\%$.
- For $p = 30$, the coverage probability can be as low as 39%.

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.
- Simultaneity is across all submodels and all slopes in them.

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.
- Simultaneity is across all submodels and all slopes in them.
- As a result, we obtain

valid PoSI for all variable selection procedures!

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.
- Simultaneity is across all submodels and all slopes in them.
- As a result, we obtain

  valid PoSI for all variable selection procedures!

- But first we need some preliminaries on

  Targets of Inference and Inference in Wrong Models

PoSI — What's in a word?

http://www.thefreedictionary.com/posies

Translations

posy [ˈpəʊzɪ] *N* → ramillete *m*

posy [ˈpəʊzi] *n* → petit bouquet *m*

posy

n → Sträußchen *nt*

posy [ˈpəʊzɪ] *n* → mazzolino (di fiori)

posy

n **posy** [ˈpəuzi]

a small bunch of flowers *a posy of primroses.* ruiker باقة китка kytička lille buket der Strauß μπουκετάκι ramillete (väike) lillekimp دسته گل kukkavihko petit bouquet (de fleurs) זר פְּרָחִים पुष्पगुच्छ kitica cvijeća kis csokor seikat bunga blómvöndur mazzolino 花束 꽃다발 puokštelė puķu pušķītis sejambak bunga boeket liten bukett, blomst bukiecik ramalhete buchețel букет(ик) цветов kytička šopek buketić [] bukett ช่อดอกไม้ küçük çiçek demeti 花束 букетик квітів گلدستہ چھوٹا chùm hoa nhỏ 花束

Submodels — Notation, Parameters, Assumptions

Submodels — Notation, Parameters, Assumptions

- Denote a submodel by the integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$ for the predictors:

$$\mathbf{X}_{\mathrm{M}} = \left(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}\right) \in \mathrm{I\!R}^{N \times m}.$$

Submodels — Notation, Parameters, Assumptions

- Denote a submodel by the integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$ for the predictors:

$$\mathbf{X}_{\mathrm{M}} = (\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}) \in \mathbb{R}^{N \times m}.$$

- The LS estimators in the submodel M are

$$\hat{\beta}_{\mathrm{M}} = (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \mathbf{Y} \in \mathbb{R}^m$$

Submodels — Notation, Parameters, Assumptions

- Denote a submodel by the integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$ for the predictors:

$$\mathbf{X}_{\mathrm{M}} = \left(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}\right) \in \mathbb{R}^{N \times m}.$$

- The LS estimators in the submodel M are

$$\hat{\beta}_{\mathrm{M}} = \left(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}}\right)^{-1} \mathbf{X}_{\mathrm{M}}^T \mathbf{Y} \in \mathbb{R}^m$$

- What does $\hat{\beta}_{\mathrm{M}}$ estimate, **not** assuming the truth of M?

Submodels — Notation, Parameters, Assumptions

- Denote a submodel by the integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$ for the predictors:

$$\mathbf{X}_{\mathrm{M}} = (\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}) \in \mathbb{R}^{N \times m}.$$

- The LS estimators in the submodel M are

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \mathbf{Y} \in \mathbb{R}^m$$

- What does $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ estimate, **not** assuming the truth of M?
  A: Its expectation — i.e., we ask for unbiasedness.

$$\boldsymbol{\mu} := \mathbf{E}[\mathbf{Y}] \in \mathbb{R}^N \quad \textbf{arbitrary!!}$$

$$\boldsymbol{\beta}_{\mathrm{M}} := \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] = (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \boldsymbol{\mu}$$

- Once again: We do not assume that the submodel is correct,
  i.e., we allow $\boldsymbol{\mu} \neq \mathbf{X}_{\mathrm{M}} \boldsymbol{\beta}_{\mathrm{M}}$! But $\mathbf{X}_{\mathrm{M}} \boldsymbol{\beta}_{\mathrm{M}}$ is the best approximation to $\boldsymbol{\mu}$.

Submodels versus the Full Model — Confusions

- Abbreviate $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{\mathrm{M}_F}$ and $\boldsymbol{\beta} := \boldsymbol{\beta}_{\mathrm{M}_F}$, $\mathrm{M}_F = \{1, 2, ..., p\}$ = full model.
  Questions:
  - How do submodel estimates $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ relate to full-model estimates $\hat{\boldsymbol{\beta}}$?
  - How do submodel parameters $\boldsymbol{\beta}_{\mathrm{M}}$ relate to full-model parameters $\boldsymbol{\beta}$?

  Is $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ a subset of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$ a subset of $\boldsymbol{\beta}$?

Submodels versus the Full Model — Confusions

- Abbreviate $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{\mathrm{M}_F}$ and $\boldsymbol{\beta} := \boldsymbol{\beta}_{\mathrm{M}_F}$, $\mathrm{M}_F = \{1, 2, ..., p\}$ = full model.
  Questions:
  - How do submodel estimates $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ relate to full-model estimates $\hat{\boldsymbol{\beta}}$?
  - How do submodel parameters $\boldsymbol{\beta}_{\mathrm{M}}$ relate to full-model parameters $\boldsymbol{\beta}$?

  Is $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ a subset of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$ a subset of $\boldsymbol{\beta}$?
- Answer: Unless **X** is an orthogonal design,
  - $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ is not a subset of $\hat{\boldsymbol{\beta}}$ and
  - $\boldsymbol{\beta}_{\mathrm{M}}$ is not a subset of $\boldsymbol{\beta}$.

Submodels versus the Full Model — Confusions

- Abbreviate $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{\mathrm{M}_F}$ and $\boldsymbol{\beta} := \boldsymbol{\beta}_{\mathrm{M}_F}$, $\mathrm{M}_F = \{1, 2, ..., p\}$ = full model.
  Questions:
  - How do submodel estimates $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ relate to full-model estimates $\hat{\boldsymbol{\beta}}$?
  - How do submodel parameters $\boldsymbol{\beta}_{\mathrm{M}}$ relate to full-model parameters $\boldsymbol{\beta}$?

  Is $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ a subset of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$ a subset of $\boldsymbol{\beta}$?
- Answer: Unless **X** is an orthogonal design,
  - $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ is not a subset of $\hat{\boldsymbol{\beta}}$ and
  - $\boldsymbol{\beta}_{\mathrm{M}}$ is not a subset of $\boldsymbol{\beta}$.
- Reason: Slopes, both estimates and parameters, depend on what the other predictors are — in value and in meaning.

Submodels versus the Full Model — Confusions

- Abbreviate $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{\mathrm{M}_F}$ and $\boldsymbol{\beta} := \boldsymbol{\beta}_{\mathrm{M}_F}$, $\mathrm{M}_F = \{1, 2, ..., p\}$ = full model.
  Questions:
  - How do submodel estimates $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ relate to full-model estimates $\hat{\boldsymbol{\beta}}$?
  - How do submodel parameters $\boldsymbol{\beta}_{\mathrm{M}}$ relate to full-model parameters $\boldsymbol{\beta}$?

  Is $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ a subset of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$ a subset of $\boldsymbol{\beta}$?
- Answer: Unless **X** is an orthogonal design,
  - $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ is not a subset of $\hat{\boldsymbol{\beta}}$ and
  - $\boldsymbol{\beta}_{\mathrm{M}}$ is not a subset of $\boldsymbol{\beta}$.
- Reason: Slopes, both estimates and parameters, depend on what the other predictors are — in value and in meaning.
- Message: $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ does not estimate full-model parameters!
  (Exception: The full model is causal or "data generating" ...
  Submodel estimates suffer then from "omitted variables bias.")

‘Slopes depend on ...’ — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: “`LoP`” = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: `Age`
  - Predictor 2: `Income`

'Slopes depend on ...' — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: "`LoP`" = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: `Age`
  - Predictor 2: `Income`
- Expectation: Younger customers have higher `LoP`, that is, $\beta_{\texttt{Age}} < 0$.

'Slopes depend on ...' — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: "`LoP`" = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: `Age`
  - Predictor 2: `Income`
- Expectation: Younger customers have higher `LoP`, that is, $\beta_{\texttt{Age}} < 0$.
- Outcome of the analysis:

‘Slopes depend on ...’ — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: “`LoP`” = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: `Age`
  - Predictor 2: `Income`
- Expectation: Younger customers have higher `LoP`, that is, $\beta_{\texttt{Age}} < 0$.
- Outcome of the analysis:
  - Against expectations, a regression of `LoP` on `Age` alone indicates that older customers have higher `LoP`: $\beta_{\texttt{Age}} > 0$

‘Slopes depend on ...’ — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: “`LoP`” = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: `Age`
  - Predictor 2: `Income`
- Expectation: Younger customers have higher `LoP`, that is, $\beta_{\texttt{Age}} < 0$.
- Outcome of the analysis:
  - Against expectations, a regression of `LoP` on `Age` alone indicates that older customers have higher `LoP`: $\beta_{\texttt{Age}} > 0$
  - But a regression of `LoP` on `Age` and `Income` indicates that, adjusted for `Income`, younger customers have higher `LoP`: $\beta_{\texttt{Age}\bullet\texttt{Income}} < 0$

'Slopes depend on ...' — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: "LoP" = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: Age
  - Predictor 2: Income
- Expectation: Younger customers have higher LoP, that is, $\beta_{\texttt{Age}} < 0$.
- Outcome of the analysis:
  - Against expectations, a regression of LoP on Age alone indicates that older customers have higher LoP: $\beta_{\texttt{Age}} > 0$
  - But a regression of LoP on Age and Income indicates that, adjusted for Income, younger customers have higher LoP: $\beta_{\texttt{Age}\bullet\texttt{Income}} < 0$
  - Enabling factor: (partial) collinearity between Age and Income.

'Slopes depend on ...' — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: "`LoP`" = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: `Age`
  - Predictor 2: `Income`
- Expectation: Younger customers have higher `LoP`, that is, $\beta_{\texttt{Age}} < 0$.
- Outcome of the analysis:
  - Against expectations, a regression of `LoP` on `Age` alone indicates that older customers have higher `LoP`: $\beta_{\texttt{Age}} > 0$
  - But a regression of `LoP` on `Age` and `Income` indicates that, adjusted for `Income`, younger customers have higher `LoP`: $\beta_{\texttt{Age}\bullet\texttt{Income}} < 0$
  - Enabling factor: (partial) collinearity between `Age` and `Income`.
- A case of Simpson's paradox: $\beta_{\texttt{Age}} > 0 > \beta_{\texttt{Age}\bullet\texttt{Income}}$.
  - The marginal and the `Income`-adjusted slope have very different values and different meanings.

Adjustment, Estimates, Parameters, t-Statistics

- Notation and facts for the components of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$, assuming $j \in \mathrm{M}$:
  - Let $\mathbf{X}_{j\bullet\mathrm{M}}$ be the predictor $\mathbf{X}_j$ adjusted for the other predictors in M:

$$\mathbf{X}_{j\bullet\mathrm{M}} \;:=\; \left(\mathbf{I} - \mathbf{H}_{\mathrm{M}\smallsetminus\{j\}}\right)\mathbf{X}_j \;\;\perp\;\; \mathbf{X}_k \;\forall k \in \mathrm{M}\smallsetminus\{j\}.$$

Adjustment, Estimates, Parameters, *t*-Statistics

- Notation and facts for the components of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$, assuming $j \in \mathrm{M}$:
  - Let $\mathbf{X}_{j\bullet\mathrm{M}}$ be the predictor $\mathbf{X}_j$ adjusted for the other predictors in M:

$$\mathbf{X}_{j\bullet\mathrm{M}} \;:=\; \left(\mathbf{I} - \mathbf{H}_{\mathrm{M}\setminus\{j\}}\right)\mathbf{X}_j \;\;\perp\;\; \mathbf{X}_k \;\forall k \in \mathrm{M}\setminus\{j\}.$$

  - Let $\hat{\beta}_{j\bullet\mathrm{M}}$ be the slope estimate and $\beta_{j\bullet\mathrm{M}}$ be the parameter for $\mathbf{X}_j$ in M:

$$\hat{\beta}_{j\bullet\mathrm{M}} \;:=\; \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T\mathbf{Y}}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|^2}\,, \qquad \beta_{j\bullet\mathrm{M}} \;:=\; \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T\mathbf{E}[\mathbf{Y}]}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|^2}.$$

Adjustment, Estimates, Parameters, *t*-Statistics

- Notation and facts for the components of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$, assuming $j \in \mathrm{M}$:
  - Let $\mathbf{X}_{j\bullet\mathrm{M}}$ be the predictor $\mathbf{X}_j$ adjusted for the other predictors in M:

$$\mathbf{X}_{j\bullet\mathrm{M}} \;:=\; \left(\mathbf{I} - \mathbf{H}_{\mathrm{M}\setminus\{j\}}\right)\mathbf{X}_j \;\;\perp\;\; \mathbf{X}_k \;\forall k \in \mathrm{M}\setminus\{j\}.$$

  - Let $\hat{\beta}_{j\bullet\mathrm{M}}$ be the slope estimate and $\beta_{j\bullet\mathrm{M}}$ be the parameter for $\mathbf{X}_j$ in M:

$$\hat{\beta}_{j\bullet\mathrm{M}} \;:=\; \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T\mathbf{Y}}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|^2}\,, \qquad \beta_{j\bullet\mathrm{M}} \;:=\; \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T\mathbf{E}[\mathbf{Y}]}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|^2}.$$

  - Let $t_{j\bullet\mathrm{M}}$ be the *t*-statistic for $\hat{\beta}_{j\bullet\mathrm{M}}$ and $\beta_{j\bullet\mathrm{M}}$:

$$t_{j\bullet\mathrm{M}} \;:=\; \frac{\hat{\beta}_{j\bullet\mathrm{M}} - \beta_{j\bullet\mathrm{M}}}{\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|} \;=\; \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|\;\hat{\sigma}}.$$

Parameters One More Time

Parameters One More Time

- Once more: If the predictors are partly collinear (non-orthogonal) then

$$\boxed{\mathrm{M} \neq \mathrm{M}' \quad \Rightarrow \quad \beta_{j\bullet\mathrm{M}} \neq \beta_{j\bullet\mathrm{M}'}}$$ in value and in meaning.

Motto: A difference in adjustment implies a difference in parameters.

Parameters One More Time

- Once more: If the predictors are partly collinear (non-orthogonal) then

  $$\boxed{\mathrm{M} \neq \mathrm{M}' \quad \Rightarrow \quad \beta_{j\bullet\mathrm{M}} \neq \beta_{j\bullet\mathrm{M}'}}$$ in value and in meaning.

  Motto: A difference in adjustment implies a difference in parameters.

- It follows that there are up to $p\,2^{p-1}$ different parameters $\beta_{j\bullet\mathrm{M}}$!

Parameters One More Time

- Once more: If the predictors are partly collinear (non-orthogonal) then

$$\boxed{\mathrm{M} \neq \mathrm{M}' \quad \Rightarrow \quad \beta_{j\bullet\mathrm{M}} \neq \beta_{j\bullet\mathrm{M}'}}$$ in value and in meaning.

  Motto: A difference in adjustment implies a difference in parameters.

- It follows that there are up to $p\,2^{p-1}$ different parameters $\beta_{j\bullet\mathrm{M}}$!

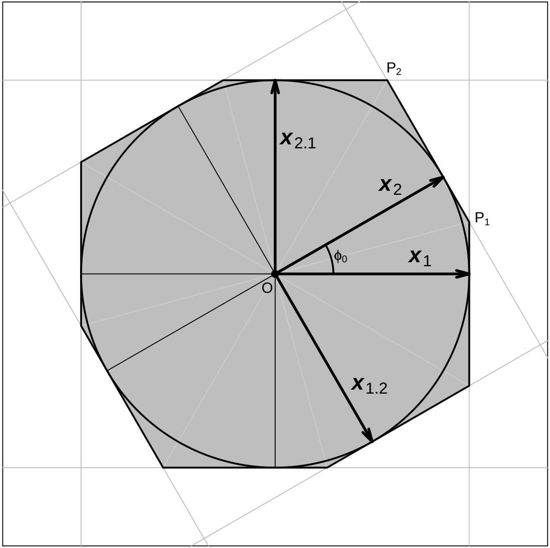
- However, they are intrinsically p-dimensional:

$$\boldsymbol{\beta}_{\mathrm{M}} \;=\; (\mathbf{X}_{\mathrm{M}}^{T}\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^{T}\,\mathbf{X}\,\boldsymbol{\beta}$$

  where $\mathbf{X}$ and $\boldsymbol{\beta}$ are from the full model.

- Hence each $\beta_{j\bullet\mathrm{M}}$ is a lin. comb. of the full model parameters $\beta_1, ..., \beta_p$.

Geometry of Adjustment



Column space of **X** for $p = 2$ predictors, partly collinear

Error Estimates $\hat{\sigma}^2$

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}}\hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M; (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model;

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j \bullet \mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}} \hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M; (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model;

$\Longrightarrow$ $t_{j \bullet \mathrm{M}}$ will have a t-distribution with the same dfs $\forall \mathrm{M}, \ \forall j \in \mathrm{M}$.

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}}\hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M;
    (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model;

$\Longrightarrow$ $t_{j\bullet\mathrm{M}}$ will have a t-distribution with the same dfs $\forall\mathrm{M}, \ \forall j \in \mathrm{M}$.

- What if even the full model is 1st order wrong?

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j \bullet \mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}} \hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M; (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model;

$\Longrightarrow$ $t_{j \bullet \mathrm{M}}$ will have a t-distribution with the same dfs $\forall \mathrm{M},\ \forall j \in \mathrm{M}$.

- What if even the full model is 1st order wrong?
  Answer: $\hat{\sigma}^2_{Full}$ will be inflated and inference will be conservative.
  But better estimates are available if ...

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j \bullet \mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}} \hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M; (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model;

  $\Longrightarrow$ $t_{j \bullet \mathrm{M}}$ will have a t-distribution with the same dfs $\forall \mathrm{M}, \; \forall j \in \mathrm{M}$.

- What if even the full model is 1st order wrong?
  Answer: $\hat{\sigma}^2_{Full}$ will be inflated and inference will be conservative.
  But better estimates are available if ...
  - exact replicates exist: use $\hat{\sigma}^2$ from the 1-way ANOVA of replicates;
  - a larger than the full model can be assumed 1st order correct: use $\hat{\sigma}^2_{Large}$;
  - a previous dataset provided a valid estimate: use $\hat{\sigma}^2_{previous}$;
  - nonparametric estimates are available: use $\hat{\sigma}^2_{nonpar}$ (Hall and Carroll 1989).

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}}\hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M;
    (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model;

$\Longrightarrow$ $t_{j\bullet\mathrm{M}}$ will have a t-distribution with the same dfs $\forall \mathrm{M}, \ \forall j \in \mathrm{M}$.

- What if even the full model is 1st order wrong?
  Answer: $\hat{\sigma}^2_{Full}$ will be inflated and inference will be conservative.
  But better estimates are available if ...
  - exact replicates exist: use $\hat{\sigma}^2$ from the 1-way ANOVA of replicates;
  - a larger than the full model can be assumed 1st order correct: use $\hat{\sigma}^2_{Large}$;
  - a previous dataset provided a valid estimate: use $\hat{\sigma}^2_{previous}$;
  - nonparametric estimates are available: use $\hat{\sigma}^2_{nonpar}$ (Hall and Carroll 1989).

PS: In the fashionable $p > N$ literature, what is their $\hat{\sigma}^2$?

Statistical Inference under First Order Incorrectness

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If r = dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

$$\mathrm{CI}_{j\bullet\mathrm{M}}(K) \;:=\; [\,\hat{\beta}_{j\bullet\mathrm{M}} \pm K\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|\,]$$

  satisfy each $\qquad \mathbf{P}[\,\beta_{j\bullet\mathrm{M}} \in \mathrm{CI}_{j\bullet\mathrm{M}}(K)\,] \;=\; 1-\alpha.$

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If $r =$ dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

$$\mathrm{CI}_{j\bullet\mathrm{M}}(K) \;:=\; [\,\hat{\beta}_{j\bullet\mathrm{M}} \pm K\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|\,]$$

  satisfy each $\quad \mathbf{P}[\,\beta_{j\bullet\mathrm{M}} \in \mathrm{CI}_{j\bullet\mathrm{M}}(K)\,] \;=\; 1-\alpha.$

- Achieved so far:

$$\mathbf{Y} \;=\; \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$$

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If $r =$ dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

$$\mathrm{CI}_{j\bullet\mathrm{M}}(K) := [\,\hat{\beta}_{j\bullet\mathrm{M}} \pm K\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|\,]$$

  satisfy each $\mathbf{P}[\,\beta_{j\bullet\mathrm{M}} \in \mathrm{CI}_{j\bullet\mathrm{M}}(K)\,] = 1-\alpha.$

- Achieved so far:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$$

  - No assumption is made that the submodels are 1st order correct;

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If $r =$ dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

$$\mathrm{CI}_{j\bullet\mathrm{M}}(K) \;:=\; [\,\hat{\beta}_{j\bullet\mathrm{M}} \pm K\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|\,]$$

  satisfy each $\qquad \mathbf{P}[\,\beta_{j\bullet\mathrm{M}} \in \mathrm{CI}_{j\bullet\mathrm{M}}(K)\,] \;=\; 1-\alpha.$

- Achieved so far:

$$\mathbf{Y} \;=\; \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$$

  - No assumption is made that the submodels are 1st order correct;
  - Even the full model may be 1st order incorrect
    if a valid $\hat{\sigma}^2$ is otherwise available;

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If $r =$ dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

$$\mathrm{CI}_{j\bullet\mathrm{M}}(K) \;:=\; [\,\hat{\beta}_{j\bullet\mathrm{M}} \pm K\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|\,]$$

  satisfy each $\quad \mathbf{P}[\,\beta_{j\bullet\mathrm{M}} \in \mathrm{CI}_{j\bullet\mathrm{M}}(K)\,] \;=\; 1-\alpha.$

- Achieved so far:

$$\mathbf{Y} \;=\; \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$$

  - No assumption is made that the submodels are 1st order correct;
  - Even the full model may be 1st order incorrect if a valid $\hat{\sigma}^2$ is otherwise available;
  - A single error estimate opens up the possibility of simultaneous inference across submodels.

Variable Selection

Variable Selection

- What is a variable selection procedure?

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\;\mathrm{I\!R}^N \to \mathcal{P}(\{1, ...p\})$

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathrm{I\!R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathrm{I\!R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Facing up to post-selection inference: Confusers!

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y}), \quad \mathrm{I\!R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathrm{I\!R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on $\mathbf{X}$ does not invalidate inference (example: deselect predictors based on VIF, $\mathbf{H}$, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, its components $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ for $j \in \hat{\mathrm{M}}(\mathbf{Y})$.

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, its components $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ for $j \in \hat{\mathrm{M}}(\mathbf{Y})$.
  - The target of inference is random.

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on $\mathbf{X}$ does not invalidate inference (example: deselect predictors based on VIF, $\mathbf{H}$, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, its components $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ for $j \in \hat{\mathrm{M}}(\mathbf{Y})$.
  - The target of inference is random.
  - The target of inference has a random dimension: $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{|\hat{\mathrm{M}}(\mathbf{Y})|}$

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, its components $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ for $j \in \hat{\mathrm{M}}(\mathbf{Y})$.
  - The target of inference is random.
  - The target of inference has a random dimension: $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{|\hat{\mathrm{M}}(\mathbf{Y})|}$
  - Conditional on $j \in \hat{\mathrm{M}}$, the target component $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ has random meanings.

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y}), \quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, its components $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ for $j \in \hat{\mathrm{M}}(\mathbf{Y})$.
  - The target of inference is random.
  - The target of inference has a random dimension: $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{|\hat{\mathrm{M}}(\mathbf{Y})|}$
  - Conditional on $j \in \hat{\mathrm{M}}$, the target component $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ has random meanings.
  - When $j \notin \hat{\mathrm{M}}$ both $\beta_{j\bullet\hat{\mathrm{M}}}$ and $\hat{\beta}_{j\bullet\hat{\mathrm{M}}}$ are undefined.

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y}), \quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, its components $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ for $j \in \hat{\mathrm{M}}(\mathbf{Y})$.
  - The target of inference is random.
  - The target of inference has a random dimension: $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{|\hat{\mathrm{M}}(\mathbf{Y})|}$
  - Conditional on $j \in \hat{\mathrm{M}}$, the target component $\beta_{j\bullet\hat{\mathrm{M}}(\mathbf{Y})}$ has random meanings.
  - When $j \notin \hat{\mathrm{M}}$ both $\beta_{j\bullet\hat{\mathrm{M}}}$ and $\hat{\beta}_{j\bullet\hat{\mathrm{M}}}$ are undefined.
  - Hence the coverage probability $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ is undefined.

Universal Post-Selection Inference

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\;\&\;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\quad(\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\;\&\;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\qquad (\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$
  - $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \mid j \in \hat{\mathrm{M}}\,]$ $\qquad (\mathbf{P}[\, j \in \hat{\mathrm{M}}\,] \;=\; ???)$

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\;\&\;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\qquad(\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$
  - $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \,\big|\, j \in \hat{\mathrm{M}}\,]$ $\qquad(\mathbf{P}[\, j \in \hat{\mathrm{M}}\,] = ???)$
  - $\mathbf{P}[\, \forall j \in \hat{\mathrm{M}} : \; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$

  All are meaningful; the last will be our choice.

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\;\&\;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\qquad (\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$
  - $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \,\big|\, j \in \hat{\mathrm{M}}\,]$ $\qquad (\mathbf{P}[\, j \in \hat{\mathrm{M}}\,] \;=\; ???)$
  - $\mathbf{P}[\, \forall j \in \hat{\mathrm{M}} :\;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$

  All are meaningful; the last will be our choice.

- Overcoming the next difficulty:
  - Problem: None of the above coverage probabilities are known or can be estimated for most selection procedures $\hat{\mathrm{M}}$.

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\;\&\;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\qquad (\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$
  - $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \,\big|\, j \in \hat{\mathrm{M}}\,]$ $\qquad (\mathbf{P}[\, j \in \hat{\mathrm{M}}\,] = ???)$
  - $\mathbf{P}[\, \forall j \in \hat{\mathrm{M}} : \;\beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$

  All are meaningful; the last will be our choice.

- Overcoming the next difficulty:
  - Problem: None of the above coverage probabilities are known or can be estimated for most selection procedures $\hat{\mathrm{M}}$.
  - Solution: Ask for more!

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\; \& \;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\qquad (\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$
  - $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \,|\, j \in \hat{\mathrm{M}}\,]$ $\qquad (\mathbf{P}[\, j \in \hat{\mathrm{M}}\,] \; = ???)$
  - $\mathbf{P}[\, \forall j \in \hat{\mathrm{M}} : \;\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$

  All are meaningful; the last will be our choice.

- Overcoming the next difficulty:
  - Problem: None of the above coverage probabilities are known or can be estimated for most selection procedures $\hat{\mathrm{M}}$.
  - Solution: Ask for more!
    Universal Post-Selection Inference for all selection procedures is doable.

Reduction to Simultaneous Inference

Reduction to Simultaneous Inference

Lemma

For any variable selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$*, we have the following "significant triviality bound":*

$$\max_{j \in \hat{\mathrm{M}}} |t_{j \bullet \hat{\mathrm{M}}}| \;\leq\; \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}| \qquad \forall\, \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^N.$$

Reduction to Simultaneous Inference

Lemma

For any variable selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$*, we have the following "significant triviality bound":*

$$\max_{j \in \hat{\mathrm{M}}} |t_{j \bullet \hat{\mathrm{M}}}| \;\leq\; \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}| \qquad \forall\, \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^N .$$

Theorem

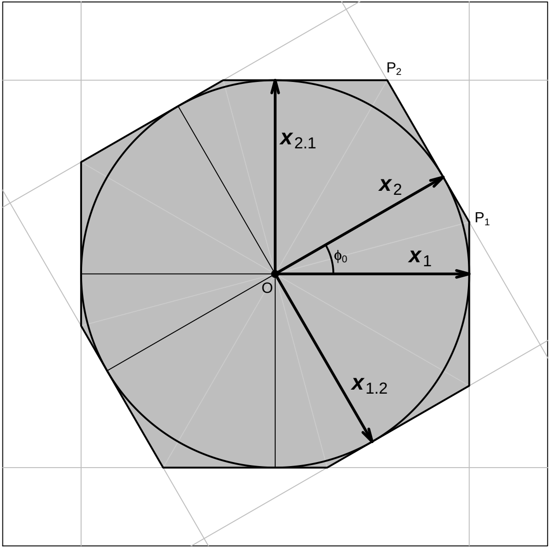
Let K *be the* $1-\alpha$ *quantile of the "max-max-*$|t|$*" statistic of the lemma:*

$$\mathbf{P}\Big[\, \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}| \leq K \,\Big] \stackrel{(\geq)}{=} 1 - \alpha .$$

Then we have the following universal PoSI guarantee:

$$\mathbf{P}\Big[\, \beta_{j \bullet \hat{\mathrm{M}}} \in CI_{j \bullet \hat{\mathrm{M}}}(K) \;\; \forall j \in \hat{\mathrm{M}} \,\Big] \;\geq\; 1 - \alpha \qquad \forall \hat{\mathrm{M}}.$$

PoSI Geometry — Simultaneity



PoSI polytope
= intersection
of all t-bands.

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?
- Consider $\hat{\mathrm{M}}$ defined as follows:

$$\hat{\mathrm{M}} := \mathrm{argmax}_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}|$$

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?
- Consider $\hat{\mathrm{M}}$ defined as follows:

$$\hat{\mathrm{M}} \;:=\; \mathrm{argmax}_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}|$$

A polite name: "Single Predictor Adjusted Regression" $=:$ SPAR

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?
- Consider $\hat{\mathrm{M}}$ defined as follows:

$$\hat{\mathrm{M}} \;:=\; \mathrm{argmax}_{\mathrm{M}} \; \max_{j \in \mathrm{M}} \; |t_{j \bullet \mathrm{M}}|$$

A polite name: "Single Predictor Adjusted Regression" $=:$ SPAR

A crude name: "Significance Hunting"

a special case of "p-hacking" (Simmons, Nelson, Simonsohn 2011)

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?

- Consider $\hat{\mathrm{M}}$ defined as follows:

$$\hat{\mathrm{M}} \;:=\; \mathrm{argmax}_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}|$$

  A polite name: "Single Predictor Adjusted Regression" $=:$ SPAR

  A crude name: "Significance Hunting"

  a special case of "p-hacking" (Simmons, Nelson, Simonsohn 2011)

- SPAR requires the full PoSI protection — by construction!

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?
- Consider $\hat{\mathrm{M}}$ defined as follows:

$$\hat{\mathrm{M}} \;:=\; \mathrm{argmax}_{\mathrm{M}} \; \max_{j \in \mathrm{M}} \; |t_{j \bullet \mathrm{M}}|$$

A polite name: "Single Predictor Adjusted Regression" $=:$ SPAR
A crude name: "Significance Hunting"
a special case of "p-hacking" (Simmons, Nelson, Simonsohn 2011)

- SPAR requires the full PoSI protection — by construction!
- How realistic is SPAR in describing real data analysts behaviors?
  - It ignores the goodness of fit of the selected model.
  - It looks for the minimal achievable p-value / strongest "effect".

Computing PoSI

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\|t_{j\bullet\mathrm{M}}\|$	1	4	12	32	80	192	448	1, 024	2, 304	5, 120
p	11	12	13	14	15	16	17	18	19	20
$\#\|t_{j\bullet\mathrm{M}}\|$	11, 264	24, 576	53, 248	114, 688	245, 760	524, 288	1, 114, 112	2, 359, 296	4, 980, 736	10, 485, 760

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	1	4	12	32	80	192	448	1,024	2,304	5,120
p	11	12	13	14	15	16	17	18	19	20
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	11,264	24,576	53,248	114,688	245,760	524,288	1,114,112	2,359,296	4,980,736	10,485,760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \overset{<}{\approx} 20$.

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\vert t_{j\bullet\mathrm{M}}\vert$	1	4	12	32	80	192	448	1, 024	2, 304	5, 120
p	11	12	13	14	15	16	17	18	19	20
$\#\vert t_{j\bullet\mathrm{M}}\vert$	11, 264	24, 576	53, 248	114, 688	245, 760	524, 288	1, 114, 112	2, 359, 296	4, 980, 736	10, 485, 760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \overset{<}{\approx} 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\|t_{j\bullet\mathrm{M}}\|$	1	4	12	32	80	192	448	1,024	2,304	5,120
p	11	12	13	14	15	16	17	18	19	20
$\#\|t_{j\bullet\mathrm{M}}\|$	11,264	24,576	53,248	114,688	245,760	524,288	1,114,112	2,359,296	4,980,736	10,485,760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \overset{<}{\approx} 20$.
- Computations are specific to a design **X**: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$
- Computations depend only on the inner product matrix $\mathbf{X}^T\mathbf{X}$.

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	1	4	12	32	80	192	448	1,024	2,304	5,120
p	11	12	13	14	15	16	17	18	19	20
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	11,264	24,576	53,248	114,688	245,760	524,288	1,114,112	2,359,296	4,980,736	10,485,760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \overset{<}{\approx} 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$
- Computations depend only on the inner product matrix $\mathbf{X}^T\mathbf{X}$.
  $\Rightarrow$ The limiting factor is p (N may only matter for $\hat{\sigma}^2$).

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	1	4	12	32	80	192	448	1,024	2,304	5,120
p	11	12	13	14	15	16	17	18	19	20
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	11,264	24,576	53,248	114,688	245,760	524,288	1,114,112	2,359,296	4,980,736	10,485,760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \lessapprox 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$
- Computations depend only on the inner product matrix $\mathbf{X}^T\mathbf{X}$.
  $\Rightarrow$ The limiting factor is p (N may only matter for $\hat{\sigma}^2$).
- One Monte Carlo computation is good for any α and any error df.

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	1	4	12	32	80	192	448	1,024	2,304	5,120
p	11	12	13	14	15	16	17	18	19	20
$\#\lvert t_{j\bullet\mathrm{M}}\rvert$	11,264	24,576	53,248	114,688	245,760	524,288	1,114,112	2,359,296	4,980,736	10,485,760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \overset{<}{\approx} 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$
- Computations depend only on the inner product matrix $\mathbf{X}^T\mathbf{X}$.
  $\Rightarrow$ The limiting factor is p (N may only matter for $\hat{\sigma}^2$).
- One Monte Carlo computation is good for any α and any error df.
- Computations of universal upper bounds:

$$K_{\mathrm{univ}}(p, \alpha, df) \;\geq\; K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df) \;\; \forall \mathbf{X}_{\ldots \times p}.$$

Scheffé Protection Yields Valid PoSI

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \; \hat{\sigma}} \;\sim\; \sqrt{p \, F_{p,df}}.$$

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \, \hat{\sigma}} \sim \sqrt{p\, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear “contrasts”.

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \, \hat{\sigma}} \sim \sqrt{p \, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear “contrasts”.
  - The Scheffé constant is $K_{\mathrm{Sch}} = K_{\mathrm{Sch}}(p, \alpha, df) = \sqrt{p \, F_{p,df;1-\alpha}}$.

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \; \hat{\sigma}} \quad \sim \quad \sqrt{p \, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear “contrasts”.
  - The Scheffé constant is $K_{\mathrm{Sch}} = K_{\mathrm{Sch}}(p, \alpha, df) = \sqrt{p \, F_{p,df;1-\alpha}}$.

- Compare: **PoSI Simultaneous Inference** is based on the statistic

$$\max_{\mathrm{M}} \max_{j \in \mathrm{M}} \frac{|\mathbf{X}_{j \bullet \mathrm{M}}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{X}_{j \bullet \mathrm{M}}\| \; \hat{\sigma}}$$

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \; \hat{\sigma}} \quad \sim \quad \sqrt{p \, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear “contrasts”.
  - The Scheffé constant is $K_{\mathrm{Sch}} = K_{\mathrm{Sch}}(p, \alpha, df) = \sqrt{p \, F_{p,df;1-\alpha}}$.

- Compare: **PoSI Simultaneous Inference** is based on the statistic

$$\max_{\mathrm{M}} \max_{j \in \mathrm{M}} \frac{|\mathbf{X}_{j \bullet \mathrm{M}}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{X}_{j \bullet \mathrm{M}}\| \; \hat{\sigma}}$$

  The PoSI contrasts are a subset of the Scheffé contrasts, hence:

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \; \hat{\sigma}} \quad \sim \quad \sqrt{p \, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear "contrasts".
  - The Scheffé constant is $K_{\mathrm{Sch}} = K_{\mathrm{Sch}}(p, \alpha, df) = \sqrt{p \, F_{p,df;1-\alpha}}$.

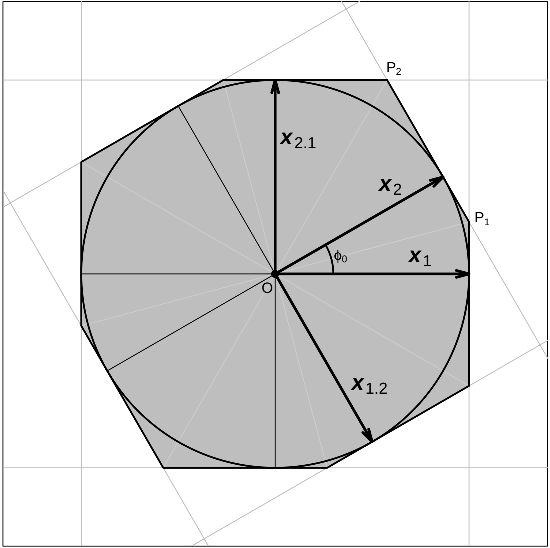
- Compare: **PoSI Simultaneous Inference** is based on the statistic

$$\max_{\mathrm{M}} \max_{j \in \mathrm{M}} \frac{|\mathbf{X}_{j\bullet \mathrm{M}}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{X}_{j\bullet \mathrm{M}}\| \; \hat{\sigma}}$$

  The PoSI contrasts are a subset of the Scheffé contrasts, hence:

  - Scheffé statistic $\geq$ PoSI statistic
  - $K_{\mathrm{Sch}} \geq K_{\mathrm{PoSI}}$
  - Scheffé yields universally valid conservative PoSI.

The Scheffé Ball and the PoSI Polytope



Circle = Scheffé Ball

The PoSI polytope is tangent to the ball.

Orthogonal Designs have the Smallest PoSI Constants

Orthogonal Designs have the Smallest PoSI Constants

- In orthogonal designs there is no adjustment:

$$\mathbf{X}_{j\bullet \mathrm{M}}^{\mathrm{orth}} = \mathbf{X}_{j}^{\mathrm{orth}} \quad \forall \mathrm{M}, j \,(\ni \mathrm{M})$$

Orthogonal Designs have the Smallest PoSI Constants

- In orthogonal designs there is no adjustment:

$$\mathbf{X}_{j\bullet \mathrm{M}}^{\mathrm{orth}} = \mathbf{X}_{j}^{\mathrm{orth}} \quad \forall\, \mathrm{M},\, j\,(\ni \mathrm{M})$$

- The PoSI statistic simplifies to $\max_{j=1\ldots p} |t_{j\bullet\{j\}}|$,
  hence the PoSI guarantee reduces to

  simultaneity for p orthogonal contrasts.

Orthogonal Designs have the Smallest PoSI Constants

- In orthogonal designs there is no adjustment:

$$\mathbf{X}_{j\bullet \mathrm{M}}^{\mathrm{orth}} = \mathbf{X}_{j}^{\mathrm{orth}} \quad \forall \mathrm{M}, j\,(\ni \mathrm{M})$$

- The PoSI statistic simplifies to $\max_{j=1\ldots p} |t_{j\bullet\{j\}}|$,
  hence the PoSI guarantee reduces to

  simultaneity for p orthogonal contrasts.

- The PoSI constant for orthogonal designs is uniformly smallest:

$$K_{\mathrm{orth}}(p, \alpha, df) \leq K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df) \quad \forall p, \alpha, df, \mathbf{X}_{\ldots\times p}$$

PoSI Asymptotics

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots \times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots \times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:

$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:

$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$

  - This represents an upper bound on the PoSI rate.

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:

$$K_{\mathrm{Sch}}(p, \alpha) \; = \; \sqrt{\chi^2_{p;1-\alpha}} \; \sim \; \sqrt{p}.$$

  - This represents an upper bound on the PoSI rate.
  - We know a sharper rate bound to be $0.866\ldots\sqrt{p}$.

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:
$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$
  - This represents an upper bound on the PoSI rate.
  - We know a sharper rate bound to be $0.866...\sqrt{p}$.
  - We know of design sequences that reach $0.78...\sqrt{p}$.

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:

$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$

  - This represents an upper bound on the PoSI rate.
  - We know a sharper rate bound to be $0.866...\sqrt{p}$.
  - We know of design sequences that reach $0.78...\sqrt{p}$.
- The lowest rate is achieved by orthogonal designs with a rate

$$K_{\mathrm{orth}}(p, \alpha) \;\sim\; \sqrt{2 \log p}.$$

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:

$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$

  - This represents an upper bound on the PoSI rate.
  - We know a sharper rate bound to be $0.866...\sqrt{p}$.
  - We know of design sequences that reach $0.78...\sqrt{p}$.
- The lowest rate is achieved by orthogonal designs with a rate

$$K_{\mathrm{orth}}(p, \alpha) \;\sim\; \sqrt{2 \log p}.$$

- Hence there is a wide range of rates for the PoSI constants:

$$\sqrt{2 \log p} \;\lessapprox\; K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha) \;\lessapprox\; \sqrt{p}$$

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots \times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots \times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:
$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$
  - This represents an upper bound on the PoSI rate.
  - We know a sharper rate bound to be $0.866\ldots\sqrt{p}$.
  - We know of design sequences that reach $0.78\ldots\sqrt{p}$.
- The lowest rate is achieved by orthogonal designs with a rate
$$K_{\mathrm{orth}}(p, \alpha) \;\sim\; \sqrt{2 \log p}.$$
- Hence there is a wide range of rates for the PoSI constants:
$$\sqrt{2 \log p} \;\lessapprox\; K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots \times p}, \alpha) \;\lessapprox\; \sqrt{p}$$
- Under all circumstances, K should not be $t_{df;1-\alpha/2} = \cancel{O(1)}$.

Worst Case PoSI Asymptotics — Some Details

- Comments on upper bound $K_{\mathrm{PoSI}} \lesssim 0.866...\sqrt{p}$:

Worst Case PoSI Asymptotics — Some Details

- Comments on upper bound $K_{\mathrm{PoSI}} \lesssim 0.866...\sqrt{p}$:
  - Ignores the PoSI structure, i.e., the many orthogonalities from adjustment.

Worst Case PoSI Asymptotics — Some Details

- Comments on upper bound $K_{\mathrm{PoSI}} \lesssim 0.866...\sqrt{p}$:
  - Ignores the PoSI structure, i.e., the many orthogonalities from adjustment.
  - Based purely on the growth rate: $|\{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}| = p\,2^{p-1} \sim 2^p$

Worst Case PoSI Asymptotics — Some Details

- Comments on upper bound $K_{\mathrm{PoSI}} \lesssim 0.866...\sqrt{p}$:
  - Ignores the PoSI structure, i.e., the many orthogonalities from adjustment.
  - Based purely on the growth rate: $\quad |\{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}| \;=\; p\,2^{p-1} \;\sim\; 2^p$
  - Bound is achieved by random selection of $\sim 2^p$ random vectors:

$$\mathbf{X}_1, ..., \mathbf{X}_{p2^{p-1}} \sim U(S^{p-1}) \text{ i.i.d.} \quad \text{versus} \quad \{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}$$

Worst Case PoSI Asymptotics — Some Details

- Comments on upper bound $K_{\mathrm{PoSI}} \lesssim 0.866...\sqrt{p}$:
  - Ignores the PoSI structure, i.e., the many orthogonalities from adjustment.
  - Based purely on the growth rate: $|\{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}| = p\,2^{p-1} \sim 2^p$
  - Bound is achieved by random selection of $\sim 2^p$ random vectors:

    $$\mathbf{X}_1, ..., \mathbf{X}_{p2^{p-1}} \sim U(S^{p-1}) \text{ i.i.d. } \quad \text{versus} \quad \{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}$$
  - Reduction to radial problem: $\max_{j,\mathbf{M}:\, j\in\mathbf{M}} \langle \mathbf{U}, \mathbf{X}_{j\cdot\mathbf{M}} \rangle \quad (\mathbf{U} \sim U(S^{p-1}))$.

    $\Rightarrow$ Wyner's (1967) bounds on sphere packing apply.

Worst Case PoSI Asymptotics — Some Details

- Comments on upper bound $K_{\mathrm{PoSI}} \overset{<}{\sim} 0.866...\sqrt{p}$:
  - Ignores the PoSI structure, i.e., the many orthogonalities from adjustment.
  - Based purely on the growth rate: $|\{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}| = p\,2^{p-1} \sim 2^p$
  - Bound is achieved by random selection of $\sim 2^p$ random vectors:

    $$\mathbf{X}_1, ..., \mathbf{X}_{p2^{p-1}} \sim U(S^{p-1}) \text{ i.i.d.} \quad \text{versus} \quad \{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}$$
  - Reduction to radial problem: $\max_{j,\mathbf{M}:\, j\in\mathbf{M}} \langle \mathbf{U}, \mathbf{X}_{j\cdot\mathbf{M}} \rangle \quad (\mathbf{U} \sim U(S^{p-1}))$.

    $\Rightarrow$ Wyner's (1967) bounds on sphere packing apply.

- Comments on lower bound $K_{\mathrm{PoSI}} \overset{>}{\sim} 0.78\,\sqrt{p}$:
  - Best lower bound known to date is found by construction of an example.

$$\left|\begin{array}{cccccc} 1 & 0 & 0 & 0 & \ldots & c \\ 0 & 1 & 0 & 0 & \ldots & c \\ 0 & 0 & 1 & 0 & \ldots & c \\ 0 & 0 & 0 & 1 & \ldots & c \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \sqrt{1-(p-1)c^2} \end{array}\right|$$

  - This is not be the ultimate worst case yet.

Example: Length of Criminal Sentence (contd.)

- Reminder: t-statistics of selected covariates, in descending order:
  - $|t_{\texttt{alcohol}}| = 3.95$;
  - $|t_{\texttt{prior records}}| = 3.59$;
  - $|t_{\texttt{seriousness}}| = 3.57$;
  - $|t_{\texttt{drugs}}| = 3.31$;
  - $|t_{\texttt{employment}}| = 3.04$;
  - $|t_{\texttt{initial age}}| = 2.56$;
  - $|t_{\texttt{gender}}| = 2.33$.

Example: Length of Criminal Sentence (contd.)

- Reminder: t-statistics of selected covariates, in descending order:
  - $|t_{\texttt{alcohol}}| = 3.95$;
  - $|t_{\texttt{prior records}}| = 3.59$;
  - $|t_{\texttt{seriousness}}| = 3.57$;
  - $|t_{\texttt{drugs}}| = 3.31$;
  - $|t_{\texttt{employment}}| = 3.04$;
  - $|t_{\texttt{initial age}}| = 2.56$;
  - $|t_{\texttt{gender}}| = 2.33$.
- The PoSI constant is $K_{\mathrm{PoSI}} \approx 3.1$, hence we would claim significance for the four variables on the left.
- For comparison, the Scheffé constant is $K_{\mathrm{Sch}} \approx 4.5$, leaving us with no significant predictors at all.
- Similarly, Bonferroni with $\alpha/(p\,2^{p-1})$ yields $K_{\mathrm{Bonf}} \approx 4.7$.

Conclusions

- Valid universal post-selection inference is possible.

Conclusions

- Valid universal post-selection inference is possible.
- Necessary buy-ins:
  - Each submodel has its own slope parameters.
  - Use one $\hat{\sigma}$ you believe in for all $t_{j\bullet\mathrm{M}}$.
  - Valid inference for "wrong" models is meaningful.

Conclusions

- Valid universal post-selection inference is possible.
- Necessary buy-ins:
  - Each submodel has its own slope parameters.
  - Use one $\hat{\sigma}$ you believe in for all $t_{j\bullet\mathrm{M}}$.
  - Valid inference for “wrong” models is meaningful.
- PoSI is not procedure-specific, hence is conservative. However:
  - PoSI is valid even for selection that is informal and post-hoc.
  - PoSI is necessary for selection based on “significance hunting”.

Conclusions

- Valid universal post-selection inference is possible.
- Necessary buy-ins:
  - Each submodel has its own slope parameters.
  - Use one $\hat{\sigma}$ you believe in for all $t_{j\bullet\mathrm{M}}$.
  - Valid inference for "wrong" models is meaningful.
- PoSI is not procedure-specific, hence is conservative. However:
  - PoSI is valid even for selection that is informal and post-hoc.
  - PoSI is necessary for selection based on "significance hunting".
- Asymptotics in p suggests strong dependence of K_{PoSI} on design $\mathbf{X}$.

Conclusions

- Valid universal post-selection inference is possible.
- Necessary buy-ins:
  - Each submodel has its own slope parameters.
  - Use one $\hat{\sigma}$ you believe in for all $t_{j \bullet \mathrm{M}}$.
  - Valid inference for “wrong” models is meaningful.
- PoSI is not procedure-specific, hence is conservative. However:
  - PoSI is valid even for selection that is informal and post-hoc.
  - PoSI is necessary for selection based on “significance hunting”.
- Asymptotics in p suggests strong dependence of K_{PoSI} on design $\mathbf{X}$.
- Challenges:
  - PoSI under heteroskedasticity, random $\mathbf{X}$, general misspecification, ...
  - Understanding the design geometry that drives $K_{\mathrm{PoSI}}(\mathbf{X})$.
  - Computations of K_{PoSI} for large p.

Conclusions

- Valid universal post-selection inference is possible.
- Necessary buy-ins:
  - Each submodel has its own slope parameters.
  - Use one $\hat{\sigma}$ you believe in for all $t_{j\bullet \mathrm{M}}$.
  - Valid inference for "wrong" models is meaningful.
- PoSI is not procedure-specific, hence is conservative. However:
  - PoSI is valid even for selection that is informal and post-hoc.
  - PoSI is necessary for selection based on "significance hunting".
- Asymptotics in p suggests strong dependence of K_{PoSI} on design $\mathbf{X}$.
- Challenges:
  - PoSI under heteroskedasticity, random $\mathbf{X}$, general misspecification, ...
  - Understanding the design geometry that drives $K_{\mathrm{PoSI}}(\mathbf{X})$.
  - Computations of K_{PoSI} for large p.

THANK YOU!

Details of Simulation

- Number of simulated datasets: $100,000$.
- $\beta = 0$, $\gamma_j = 4, \forall j = 1, \ldots, 10$.
- Vectors $\mathbf{x}$ and $\mathbf{z}_j$'s are standardized to have zero mean and unit variance.
  - The correlation between $\mathbf{x}$ and each $\mathbf{z}_j$ is 0.7, $\forall j = 1, \ldots, 10$.
  - The correlation between $\mathbf{z}_{j_1}$ and $\mathbf{z}_{j_2}$ is 0.5, $\forall j_1, j_2 = 1, \ldots, 10$.

Back

PoSI1 — SPAR1: Focus on One Predictor of Interest

- Sometimes there is one focal predictor of interest, $\mathbf{X}_p$.

PoSI1 — SPAR1: Focus on One Predictor of Interest

- Sometimes there is one focal predictor of interest, $\mathbf{X}_p$.
- Inference is desired only for $\beta_{p\bullet\mathrm{M}}$ (test statistics: $t_{p\bullet\mathrm{M}}$).

PoSI1 — SPAR1: Focus on One Predictor of Interest

- Sometimes there is one focal predictor of interest, $\mathbf{X}_p$.
- Inference is desired only for $\beta_{p \bullet \mathrm{M}}$ (test statistics: $t_{p \bullet \mathrm{M}}$).
- Search only models M that contain p: $p \in \mathrm{M}$ $(\# = 2^{p-1})$.
  Purpose: Boost the statistical significance of $\hat{\beta}_{p \bullet \mathrm{M}}$.

PoSI1 — SPAR1: Focus on One Predictor of Interest

- Sometimes there is one focal predictor of interest, $\mathbf{X}_p$.
- Inference is desired only for $\beta_{p \bullet \mathrm{M}}$ (test statistics: $t_{p \bullet \mathrm{M}}$).
- Search only models M that contain p: $p \in \mathrm{M}$ $(\# = 2^{p-1})$.
  Purpose: Boost the statistical significance of $\hat{\beta}_{p \bullet \mathrm{M}}$.
- PoSI1 produces a constant K_{PoSI1} whose intervals $\mathrm{CI}_{p \bullet \mathrm{M}}(K_{\mathrm{PoSI1}})$ are valid after any variable selection procedure $\hat{\mathrm{M}}$ that is subject to $p \in \hat{\mathrm{M}}$.

PoSI1 — SPAR1: Focus on One Predictor of Interest

- Sometimes there is one focal predictor of interest, $\mathbf{X}_p$.
- Inference is desired only for $\beta_{p \bullet \mathrm{M}}$ (test statistics: $t_{p \bullet \mathrm{M}}$).
- Search only models M that contain p: $p \in \mathrm{M}$ $(\# = 2^{p-1})$.
  Purpose: Boost the statistical significance of $\hat{\beta}_{p \bullet \mathrm{M}}$.
- PoSI1 produces a constant K_{PoSI1} whose intervals $\mathrm{CI}_{p \bullet \mathrm{M}}(K_{\mathrm{PoSI1}})$ are valid after any variable selection procedure $\hat{\mathrm{M}}$ that is subject to $p \in \hat{\mathrm{M}}$.
- A selection procedure that requires the full PoSI1 protection:
  SPAR1, defined by $\hat{\mathrm{M}} := \mathrm{argmax}_{\mathrm{M} \ni p} |t_{p \bullet \mathrm{M}}|$.

PoSI1 — SPAR1: Focus on One Predictor of Interest

- Sometimes there is one focal predictor of interest, $\mathbf{X}_p$.
- Inference is desired only for $\beta_{p\bullet\mathrm{M}}$ (test statistics: $t_{p\bullet\mathrm{M}}$).
- Search only models M that contain p: $p \in \mathrm{M}$ $(\# = 2^{p-1})$.
  Purpose: Boost the statistical significance of $\hat{\beta}_{p\bullet\mathrm{M}}$.
- PoSI1 produces a constant K_{PoSI1} whose intervals $\mathrm{CI}_{p\bullet\mathrm{M}}(K_{\mathrm{PoSI1}})$ are valid after any variable selection procedure $\hat{\mathrm{M}}$ that is subject to $p \in \hat{\mathrm{M}}$.
- A selection procedure that requires the full PoSI1 protection:
  SPAR1, defined by $\hat{\mathrm{M}} := \mathrm{argmax}_{\mathrm{M}\ni p} |t_{p\bullet\mathrm{M}}|$.

Conclusions:

- PoSI1 is more appropriate for some situations than full PoSI.
- Trivially, $K_{PoSI1} < K_{PoSI}$, but sometimes not by much!

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
  - $\mathcal{M}_{\mathrm{all}} = \{\mathrm{M} : \mathrm{M} \subset \{1, 2, ..., p\},\ 0 < |\mathrm{M}| \le \min(n, p),\ \mathrm{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\}.$
- Useful sub-universes:

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
  - $\mathcal{M}_{\mathrm{all}} = \{\mathrm{M} : \mathrm{M} \subset \{1, 2, ..., p\},\ 0 < |\mathrm{M}| \leq \min(n, p),\ \mathrm{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\}$.
- Useful sub-universes:
  - Protect one or more predictors, as in PoSI1: $\mathcal{M} = \{\mathrm{M} : p \in \mathrm{M}\}$.

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
  - $\mathcal{M}_{\text{all}} = \{\mathrm{M} : \mathrm{M} \subset \{1, 2, ..., p\},\ 0 < |\mathrm{M}| \leq \min(n, p),\ \mathrm{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\}$.
- Useful sub-universes:
  - Protect one or more predictors, as in PoSI1: $\mathcal{M} = \{\mathrm{M} : p \in \mathrm{M}\}$.
  - Sparsity, i.e., submodels of size m' or less: $\mathcal{M} = \{\mathrm{M} : |\mathrm{M}| \leq m'\}$.

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
  - $\mathcal{M}_{\mathrm{all}} = \{\mathrm{M} : \mathrm{M} \subset \{1, 2, ..., p\},\ 0 < |\mathrm{M}| \leq \min(n, p),\ \mathrm{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\}$.
- Useful sub-universes:
  - Protect one or more predictors, as in PoSI1: $\mathcal{M} = \{\mathrm{M} : p \in \mathrm{M}\}$.
  - Sparsity, i.e., submodels of size m' or less: $\mathcal{M} = \{\mathrm{M} : |\mathrm{M}| \leq m'\}$.
  - Richness, i.e., drop fewer than m' predictors from the full model: $\mathcal{M} = \{\mathrm{M} : |\mathrm{M}| \geq p - m'\}$.

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
  - $\mathcal{M}_{\text{all}} = \{\mathrm{M} : \mathrm{M} \subset \{1, 2, ..., p\},\ 0 < |\mathrm{M}| \le \min(n, p),\ \mathrm{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\}$.
- Useful sub-universes:
  - Protect one or more predictors, as in PoSI1: $\mathcal{M} = \{\mathrm{M} : p \in \mathrm{M}\}$.
  - Sparsity, i.e., submodels of size m' or less: $\mathcal{M} = \{\mathrm{M} : |\mathrm{M}| \le m'\}$.
  - Richness, i.e., drop fewer than m' predictors from the full model: $\mathcal{M} = \{\mathrm{M} : |\mathrm{M}| \ge p - m'\}$.
  - Nested sets of models, as in polynomial regression, AR models, ANOVA.

PoSI Significance: Strong Error Control

For each $j \in \mathrm{M}$, consider the t-test statistic

$$t_{0,j\bullet\mathrm{M}} = \frac{\hat{\beta}_{j\bullet\mathrm{M}} - 0}{\hat{\sigma}_{\bullet}\left(\left(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}}\right)^{-1}_{jj}\right)^{\frac{1}{2}}}.$$

Theorem

Let H_1 be the random set of true alternatives in $\hat{\mathrm{M}}$,
and $\hat{H}_1$ the random set of rejections in $\hat{\mathrm{M}}$:

$$\hat{H}_1 = \{(j, \hat{\mathrm{M}}) : j \in \hat{\mathrm{M}}, |t_{0,j\bullet\hat{\mathrm{M}}}| > K\} \quad \textit{and} \quad H_1 = \{(j, \hat{\mathrm{M}}) : j \in \hat{\mathrm{M}}, \beta_{j\bullet\hat{\mathrm{M}}} \neq 0\}.$$

Then

$$\mathrm{P}\left(\hat{H}_1 \subset H_1\right) \geq 1 - \alpha.$$

If we repeat the sampling process many times, the probability that all PoSI rejections are correct is at least $1 - \alpha$, no matter how the model is selected.

Back