RSS International Conference 2025, Edinburgh

Some perspectives on models and their parametrisations

Heather Battey
Department of Mathematics, Imperial College London

September 3, 2025

Background to the first part

What is a treatment effect?

Two broad perspectives. One involves a statistical model for the outcome, the other usually does not.

Statistical models

Textbook definition:

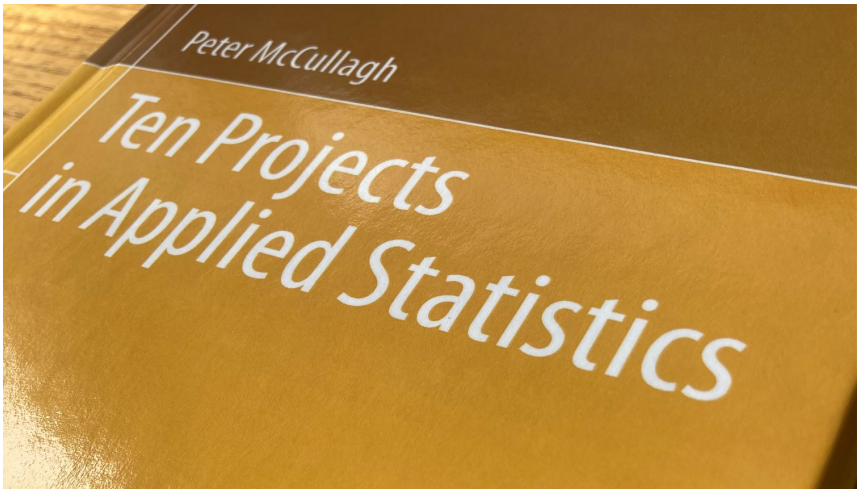
> Parametrised family of probability distributions $\{P_\theta : \theta \in \Theta\}$ to which the "true" distribution P_{θ^*} belongs. Data are realisations of random variables from P_{θ^*}.

Independence or dependence between random variables is part of the model.

Statistical models

McCullagh, P. (2002). What is a statistical model? *Ann. Statist.*, 30, 1225–1310.

- Textbook definition does not incorporate aspects usually invoked.
- Four "plainly absurd" examples are admissible under the textbook definition.
- Attempts to put scientific common sense on a mathematical footing.
- Main messages are conveyed in a less abstract way in Chapters 11 and 14 of ....


Peter McCullagh
Ten Projects
in Applied Statistics

What is a (sensible) statistical model?

Subjective claim: sensible statistical models are idealisations, stressing commonalities between individuals and attributing unexplained differences to randomness.

Purpose: insight/understanding; extrapolation of conclusions to new individuals.

Commonalities are encapsulated in stable parameters, e.g. treatment effects.

"individuals": more precisely, observational or experimental units.

Treatment effect: model-based definition

A treatment has an effect, not on the outcome, but on the probability distribution of the outcome.

McCullagh expresses this as a transformation induced by the action of a group on $\{P_\gamma : \gamma \in \Gamma\}$, the set of baseline distributions (prior to possible treatment).

Simplest models: group $\mathcal{G}$ acts on the baseline parameter space Γ

$$(g, \gamma) \to g\gamma \in \Gamma, \quad \gamma \in \Gamma, \ \ g \in \mathcal{G}.$$

A group is a set with a binary operation.

Group operation examples

A treatment has an effect on the probability distribution P_γ of the outcome.

In the simplest models, the group acts on the baseline parameter space Γ:

$$\mathcal{G} \times \Gamma \to \Gamma, \quad (g, \gamma) \to g\gamma \in \Gamma.$$

If $\Gamma = \mathbb{R}$, the group action is addition:

$$g = (\psi, +), \qquad g\gamma = \gamma + \psi, \quad \gamma \in \Gamma.$$

If $\Gamma = \mathbb{R}^+$, the group action is multiplication

$$g = (\psi, \times) \qquad g\gamma = \gamma\psi, \qquad \gamma \in \Gamma.$$

Treatment effect: model-based definition

Simplest setting: binary treatment; no interaction with intrinsic features.

Treatment effect in simplest models: treatment transforms P_γ to $P_{g\gamma}$.

Implicit: outcome distributions are compared for individuals who are comparable.

McCullagh (2022) gives a group-theoretic perspective on more elaborate models, including interaction etc.

Comparable individuals: general definition

Two, perhaps hypothetical, individuals i and i' are comparable if, for all events A in the sample space,

$$\begin{aligned}\mathsf{pr}_i(A \mid T = 0) &= \mathsf{pr}_{i'}(A \mid T = 0)\\ \mathsf{pr}_i(A \mid T = 1) &= \mathsf{pr}_{i'}(A \mid T = 1).\end{aligned}$$

Comparable individuals: parametrised model

Two, perhaps hypothetical, individuals i and i' are comparable if, for all events A in the sample space,

$$\begin{aligned}\mathsf{pr}_i(A \mid T = 0) &= \mathsf{pr}_{i'}(A \mid T = 0)\\ \mathsf{pr}_i(A \mid T = 1) &= \mathsf{pr}_{i'}(A \mid T = 1).\end{aligned}$$

Simplest parametrised models:

- individuals i and i' share the same baseline parameter γ_i.
- If both receive treatment, both are perturbed from baseline: $\gamma_i \mapsto g\gamma_i \in \Gamma$.

Comparable individuals: regression model

Two, perhaps hypothetical, individuals i and i' are comparable if, for all events A in the sample space,

$$\begin{aligned}\mathrm{pr}_i(A \mid T = 0) &= \mathrm{pr}_{i'}(A \mid T = 0)\\ \mathrm{pr}_i(A \mid T = 1) &= \mathrm{pr}_{i'}(A \mid T = 1).\end{aligned}$$

Simplest parametrised models:

- individuals i and i' share the same baseline parameter γ_i.

  If both receive treatment, both are perturbed from baseline: $\gamma_i \mapsto g\gamma_i \in \Gamma$.

Dependence on intrinsic features $w_i = w_{i'}$ corresponds to e.g.:

$$\begin{aligned}\gamma_i &= w_i^{\mathrm{T}}\beta, && \Gamma = \mathbb{R}\\ \gamma_i &= \exp(w_i^{\mathrm{T}}\beta), && \Gamma = \mathbb{R}^+.\end{aligned}$$

Treatment effect: model-free definitions

No outcome model in which to embed the treatment effect.

Many proposals, e.g. differences in expectations, differences in quantiles.

The definition of comparable individuals is the same:

$$\begin{aligned}\mathsf{pr}_i(A \mid T = 0) &= \mathsf{pr}_{i'}(A \mid T = 0)\\ \mathsf{pr}_i(A \mid T = 1) &= \mathsf{pr}_{i'}(A \mid T = 1).\end{aligned}$$

Not used in definitions based on counterfactuals...

Treatment effect: a model-free definition based on counterfactuals

For a binary treatment, a popular choice in terms of counterfactuals is:

$$\tau_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_i(Y_i(1) - Y_i(0)).$$

$Y_i(t)$: outcome on individual i for values $t \in \{0, 1\}$ of the treatment indicator.

For any i, a realisation of one of $Y_i(1)$ or $Y_i(0)$ is available, but not both.

Unobservable counterfactual

Two approaches:

- Impute the counterfactual $Y_i(t)$.
  Requires an outcome model. Defeats the object?
- (Rosenbaum & Rubin, 1983) Model and fit the process through which the counterfactuals becomes factual, then reweight actual observations.
  Exchanges assumptions on outcome distribution for other assumptions.

Two distinct considerations

- Security of causal claims/strength of uncheckable assumptions.
- The way the effect of (potential) causes is measured.

The two are decoupled under a fictitious idealisation for the counterfactuals.

Some prompts for reflection

Fictitious idealisation

Decouple any concerns over definition from concerns about counterfactuals.

Suppose, notionally, that the process through which the counterfactuals become factual is completely known, and that both outcomes are observable.

The counterfactual formulation then reduces to a matched pair problem: Each individual twinned with a perfect copy of itself.

Fictitious twin

From individual i and his or her fictitious twin, one of the two is randomised to treatment, the other to control. The corresponding outcomes are (Y_{i1}, Y_{i0}).

Question: If there is a (counterfactual) outcome model with a stable treatment parameter, is it desirable that the model-free treatment effect τ_n recovers it? Exactly? Qualitatively?

Tentative answer: otherwise, how can we be sure of its relevance in more difficult settings?

Example models giving compatible conclusions

Location model with unit treatment additivity, e.g.

$$Y_{i0} = \gamma_i + \varepsilon, \quad Y_{i1} = \gamma_i + \Delta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Linear regression if $\gamma_i = w_i^{\mathrm{T}} \beta$.

Model-based treatment effect is Δ
(additive on mean scale).

Model-free treatment effect: $\tau_n = \Delta$. ✓

Scale model with no shape parameters, e.g.

$$Y_{i0} \sim \mathrm{Exp}(\gamma_i), \quad Y_{i1} \sim \mathrm{Exp}(\gamma_i \theta).$$

Exponential regression if $\gamma_i = \exp(w_i^{\mathrm{T}} \beta)$.

Model-based treatment effect is θ.
(multiplicative on rate scale).

Model-free effect after log transformation:

$$\tau_n = -\log \theta. \quad ✓$$

Semi-compatible conclusions: Weibull example

Additional shape parameter:

$$Y_{i0} \sim \text{Weibull}(\alpha, \gamma_i), \quad Y_{i1} \sim \text{Weibull}(\alpha, \gamma_i \theta).$$

Model-based treatment effect is θ (multiplicative on rate scale).

Model-free treatment effect after log transformation:

$$\tau_n = -\frac{\log \theta}{\alpha}.$$

Only identifies θ up to sign of $\log \theta$. Stable over samples.

Incompatible conclusions: additive exponential example

$$Y_{i0} \sim \text{Exp}(\gamma_i), \quad Y_{i1} \sim \text{Exp}(\gamma_i + \Delta).$$

Model-based treatment effect is Δ (additive on rate scale).

Model-free treatment effect after log transformation:

$$\tau_n = \frac{1}{n} \sum_{i=1}^{n} \log\Big(\frac{\gamma_i + \Delta}{\gamma_i}\Big). \qquad \text{✗}$$

At $\Delta = 0$, $\tau_n = 0$, otherwise τ_n depends on the sample chosen.

Suppose $\Delta = 2$. If all γ_i are concentrated near zero, then τ_n is large; if all γ_i are large, τ_n is tiny.

Incompatible conclusions: binary example

$$Y_{i0} \sim \text{Bernoulli}\left(\frac{e^{\gamma_i}}{1+e^{\gamma_i}}\right), \quad Y_{i1} \sim \text{Bernoulli}\left(\frac{e^{\gamma_i+\Delta}}{1+e^{\gamma_i+\Delta}}\right).$$

Model-based treatment effect is Δ (additive on log-odds scale) or e^{Δ} (multiplicative on odds scale).

Model-free treatment effect:

$$\tau_n = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\gamma_i}(e^{\Delta}-1)}{(1+e^{\gamma_i})(e^{\gamma_i}e^{\Delta}+1)}. \qquad \text{✗}$$

At $\Delta = 0$, $\tau_n = 0$, otherwise τ_n depends on the sample chosen.

Rosemary Bailey's critique of a paper in Statistical Science

I have practised as a statistician for 40 years [. . .], in no case were the experimental units a random sample from a fixed finite population. They were convenient, and were deemed to be representative enough that results on them could be extrapolated to other units.

Bailey, R. A. (2017). Inference from randomized (factorial) experiments. *Statist. Sci.*, 32, 352–355.

Discussion of:

Ding, P. (2017). A paradox arising from randomization-based causal inference. *Statist. Sci.*, 32, 331–345.

Structure for compatibility

The model-based and model-free conclusions are compatible when:

(1) The group acts on the baseline distribution P_γ by transforming it to $P_{g\gamma}$;

(2) This group action can be described by the same group action on the observations;

and:

(3) There are no other nuisance parameters besides γ.

Otherwise the conclusions depend, in general, on the values of the other nuisance parameters or on the sample chosen via the baseline parameters.

Model-free definition of treatment effect

Arguments put forward in favour:

A.1 Allows the treatment to affect different individuals in different ways.

A.2 Does not necessarily require an outcome model.

A.3 Admits doubly-robust estimators.

Queries:

Q.1 If treatment effect varies by individual, should that not be explained, rather than absorbed in a composite?

Q.2 Stability? Interpretation?

Q.3 Only relevant if we accept the definition.

Model adequacy

Considerations regarding model adequacy

- The model is a provisional base; adequacy should ideally be assessed (sufficiency/co-sufficiency separation).
- Typically, more parameters $\simeq$ more flexibility. Careful construction of highly parametrised models sometimes allows elimination of nuisance parameters in the analysis (e.g. partial lik. for the PH model; stratum-specific nuisance parameters).
- When is standard likelihood-based inference for an interest parameter reliable in spite of arbitrary misspecification of a nuisance component?
- Several or many models may be compatible with the data.

A nuisance parameter is one needed to complete the model but of no direct subject-matter relevance.

Considerations regarding model adequacy

- The model is a provisional base; adequacy should ideally be assessed (sufficiency/co-sufficiency separation).
- Typically, more parameters $\simeq$ more flexibility. Careful construction of highly parametrised models sometimes allows elimination of nuisance parameters in the analysis (e.g. partial lik. for the PH model; stratum-specific nuisance parameters)
- When is standard likelihood-based inference for an interest parameter reliable in spite of arbitrary misspecification of a nuisance component?
- Several or many models may be compatible with the data.

A nuisance parameter is one needed to complete the model but of no direct subject-matter relevance.

Motivation

Aspects of core scientific interest are often convincingly specified in terms of a small number of interpretable parameters.

Other aspects might be chosen arbitrarily, e.g. on the basis of mathematical convenience.

Prospect of misspecification in the nuisance component is high.

A motivating example

Treatment effects and block effects

Matched pairs. One individual from each pair chosen at random to receive a treatment, the other is the control.

Outcomes (Y_{i1}, Y_{i0}) exponentially distributed of rates $\gamma_i\psi$ and γ_i/ψ.

Treatment param. ψ; n nuisance parameters $\gamma_1, \ldots, \gamma_n$ encapsulate arbitrary dependence on (perhaps unmeasured) covariates. Semipara.

Role of parametrisation.

Treating pair effects as fixed

If $\gamma_1, \ldots, \gamma_n$ are treated as fixed arbitrary constants they can be eliminated from the analysis by taking ratios $Z_i = Y_{i1}/Y_{i0}$.

Density function: $f_Z(z;\psi) = \dfrac{\psi^2}{(1+\psi^2 z)^2}$ no dependence on γ_i.

This is what I would do, but ...

Treating pair effects as random

An alternative approach treats pair-specific parameters as random.

Convenient choice of distribution in previous example, e.g. gamma.

A direct calculation shows that $\hat{\psi}$ is consistent in spite of arbitrary misspecification of the distribution for $(\gamma_i)_{i=1}^n$.

Questions

How sensitive is this conclusion to the formulation of the model?

- Can the response distribution be changed?
- Can the parametrisation be changed?
- Can the assumed distribution on $\gamma_1, \ldots, \gamma_n$ be changed?

Is there identifiable structure in the model that led to consistency of the MLE for the interest parameter?

Results for general models specialised to particular cases.

Misspecification: early history

Cox (1961, 1962)

Parametric model to be fitted: $m'(y;\theta)$, $y = (y_1, \ldots, y_n)$

True density $m(y)$.

MLE $\hat{\theta} \to_p \theta_m^0$, where θ_m^0 solves

$$\mathbb{E}_m[\nabla_\theta \log m'(Y;\theta)]_{\theta=\theta_m^0} = 0.$$

Equivalently θ_m^0 minimizes with respect to θ the Kullback-Leibler divergence

$$\int m(y) \log\left\{\frac{m(y)}{m'(y;\theta)}\right\} dy.$$

Asymptotic variance given by the "sandwich formula".

Tests of separate families (assumes one is correct).

The inferential target

Suppose that the inferential target has a stable interpretation in the true model and in the fitted model, e.g. treatment effect acts multiplicatively on the rate.

Other aspects are needed to complete the specification: these might be misspecified.

Formalisation

Battey, H. S. and Reid, N. (2024). On the role of parametrisation in models with a misspecified nuisance component. *Proc. Nat. Acad. Sci.*, 121 (36), e2402736121.

Misspecified nuisance component

True model contains true density function m and is parametrised in terms of an interest parameter β. True value β^*.

Assumed model m': same interpretable interest parameter β; misspecified in other ways; notional nuisance parameter α.

Log-lik for fitted model: $\ell(\beta, \alpha) = \log m'(y; \beta, \alpha)$; $(\beta, \alpha) \in \mathcal{B} \times \mathcal{A}$.
Maximisation: $(\hat{\beta}, \hat{\alpha}) \to_p (\beta_m^0, \alpha_m^0)$ solves

$$\mathbb{E}_m[\nabla_{(\beta,\alpha)} \ell(\beta_m^0, \alpha_m^0)] = 0.$$

Misspecification: no value of $\alpha \in \mathcal{A}$ gives back m.

The inferential target

Our inferential target is β^*, not (β_m^0, α_m^0) to which the classical literature applies.

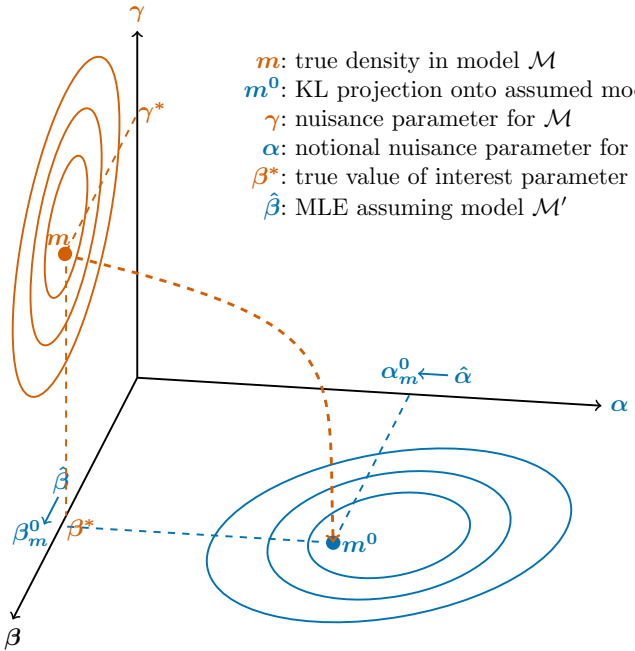
What structure implies that $\beta^* = \beta_m^0$ for any m?

Ideally would like to be able to check this without knowledge of m or α_m^0.

Orange: true model.

Blue: Assumed model.

Elliptical contours: contours of equal log-lik. under true and assumed models.



m: true density in model $\mathcal{M}$
m^0: KL projection onto assumed model $\mathcal{M}'$
γ: nuisance parameter for $\mathcal{M}$
α: notional nuisance parameter for $\mathcal{M}'$
β^*: true value of interest parameter
$\hat{\beta}$: MLE assuming model $\mathcal{M}'$

Parameter *m*-orthogonality

Definition. *Let* $\nabla^2_{\beta\alpha}\ell(\beta,\alpha)$ *denote the cross-partial derivative of the log-likelihood function for the assumed model. The parameter* β *is said to be m-orthogonal to the notional parameter* α *if* $\mathbb{E}_m[\nabla^2_{\beta\alpha}\ell(\beta,\alpha)] = 0$.

Notation. $\mathcal{B} \perp_m \mathcal{A}$: global *m*-orthogonality;
$\mathcal{B} \perp_m \alpha$ local *m*-orthogonality at α for any β;
$\beta \perp_m \mathcal{A}$ and local *m*-orthogonality at β for any α.

Geometrically. The stronger property $\nabla^2_{\beta\alpha}\ell(\beta,\alpha) = 0$ is an absence of torsion; the true model *m* trivially plays no role.

$\mathbb{E}_m[\nabla^2_{\beta\alpha}\ell(\beta,\alpha)] = 0 \Rightarrow$ any torsion is not systematic when data are generated from *m*.

Relevance of parameter m-orthogonality

Immediate: $\beta^* = \beta_m^0 \iff \underbrace{\mathbb{E}_m[\nabla_\beta \ell(\beta^*, \alpha_m^0)] = 0}_{C0}.$

But we don't know α_m^0 or m.

A first general result:
$\mathbb{E}_m[\nabla_\beta \ell(\beta^*, \alpha_m^0)] = 0$ is equivalent to
$\underbrace{(\mathbb{E}_m[\nabla_\beta \ell(\beta^*, \alpha)] = 0 \ \forall \alpha \in \mathcal{A})}_{C1.1}$ if and only if $\underbrace{\beta^* \perp_m \mathcal{A}}_{C1.2}$.

$C1.1$ and $C1.2$ imply $C0$ and therefore consistency of $\hat{\beta}$.
In some classes of models $C1.1$ & $C1.2$ can be guaranteed for any m through suitable parametrisation.

A weaker requirement

Suppose $C1.2$ fails, i.e. $\beta^* \not\perp_m \mathcal{A}$.

Second result: $(i^{\beta\beta} g_\beta + i^{\beta\alpha} g_\alpha = 0 \ \forall \alpha \in \mathcal{A}) \implies \beta^* = \beta^0_m$.
Scalar case: $(i_{\alpha\alpha} g_\beta + i_{\beta\alpha} g_\alpha = 0 \ \forall \alpha \in \mathcal{A}) \implies \beta^* = \beta^0_m$

where $i = i(\beta^*, \alpha) = \mathbb{E}_{\boldsymbol{m}}[-\nabla^2_{(\beta,\alpha)} \ell(\beta^*, \alpha)]$

$$\begin{pmatrix} i^{\beta\beta} & i^{\beta\alpha} \\ i^{\alpha\beta} & i^{\alpha\alpha} \end{pmatrix} = \begin{pmatrix} i_{\beta\beta} & i_{\beta\alpha} \\ i_{\alpha\beta} & i_{\alpha\alpha} \end{pmatrix}^{-1}.$$

$$\begin{aligned} g_\beta := g_\beta(\beta^*, \alpha) &= \mathbb{E}_{\boldsymbol{m}}[\nabla_\beta \ell(\beta^*, \alpha)], \\ g_\alpha := g_\alpha(\beta^*, \alpha) &= \mathbb{E}_{\boldsymbol{m}}[\nabla_\alpha \ell(\beta^*, \alpha)]. \end{aligned}$$

Examples

A class of natural examples involve misspecified random effects distributions.
The paper works through many examples.

A parameter cut trivially yields parameter *m*-orthogonality: these are relatively easy cases, although the conclusion may not have been obvious without having the structure made clear.

Examples of neither form...

Parameter cut: likelihood factorises as $L(\beta, \alpha) = L_1(\beta)L_2(\alpha)$.

Answers to earlier questions

How sensitive is the conclusion (consistency in exp. matched pair example with misspecified random effects) to the formulation of the model?

- Can the response distribution be changed? Yes.
- Can the parametrisation be changed? No.
- Can the assumed distribution on $\gamma_1, \ldots, \gamma_n$ be changed? Yes.

Is there identifiable structure in the model that led to consistency of the MLE for the interest parameter? Yes.

Evans and Didelez (2024) JRSSB discussion paper

Marginal structural model in a 'frugal parametrisation'.

Nuisance parameters enter through the propensity score.

E&D model has a parameter cut, implying parameter *m*-orthogonality when the propensity score is misspecified.

The proof of E&D's main theorem implicitly establishes the remaining condition. See HB discussion of E&D.

Considerations regarding model adequacy

- The model is a provisional base; adequacy should ideally be assessed (sufficiency/co-sufficiency separation).
- Typically, more parameters $\simeq$ more flexibility. Careful construction of highly parametrised models sometimes allows elimination of nuisance parameters in the analysis (e.g. partial lik. for the PH model; stratum-specific nuisance parameters)
- When is standard likelihood-based inference for an interest parameter reliable in spite of arbitrary misspecification of a nuisance component?
- Several or many models may be compatible with the data.

A nuisance parameter is one needed to complete the model but of no direct subject-matter relevance.

Confidence sets of regression models

Context

- Regression, broadly defined. Dimension $p \gg$ study individuals n.
- Aim: scientific understanding.
- In many genomics contexts, an assumption of sparsity is natural.
- Popular approaches based on penalised regression produce a single model.
- There are often several or many models that fit the data equivalently well.

Simple low-dimensional example

$$
\begin{aligned}
Y = X\beta + \varepsilon, \quad \varepsilon &= (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}, \quad \varepsilon_i \sim N(0,1), \quad n = 100, \\
\beta &= (1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p, \qquad\qquad p = 25.
\end{aligned}
$$

Rows of X drawn from $N_p(0, \Sigma)$: high correlation between first three variables.

Comprehensive model: $[p] := \{1, \ldots, p\}$; true model: $\mathcal{S} = \{1, 2\}$.

Lasso (with cross-validated tuning) selects a single model: $\{2\}$.

A likelihood-ratio test of each low-dimensional submodel $\mathcal{S}_m \subset \{1, \ldots, p\}$ against $[p]$ declared $\{2\}, \{1, 2\}$, $\{2, 3\}$ and $\{1, 2, 3\}$ as statistically indistinguishable from $[p]$.

Confidence set of models: $\mathcal{M} = \{\{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$

Operationalisation and theoretical insights in high dimensions

Cox, D. R. and Battey. H. S. (2017). Large numbers of explanatory variables: a semi-descriptive analysis. *Proc. Nat. Acad. Sci.*, 114 (32), 8592–8595

Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. Roy. Soc. Lond. A: Math. Phys. Sci.*, 474, 20170631.

Lewis, R. and Battey, H. S. (2025). Cox reduction and confidence sets of models: a theoretical elucidation. *Statist. Sci.*, to appear.

Battey, H. S., Rasines, D. G. and Tang, Y. (2025). Post-reduction inference for confidence sets of models. *arXiv: 2507.2507.10373*.

Usage of confidence sets of models

Confidence sets usually contain large numbers of models when $p \gg n$.
This is an honest reflection of the information in the data.

Any choice between statistically indistinguishable models requires either additional data or subject-matter expertise.

Compact messages can be extracted. In the example of Cox & Battey (2017):

- Two variables, v_1 and v_2, are present in 96% and 94% of models.
- In 78% of the models in which v_2 is absent, another variable, v_3, is present in its place.
- Only 1% of models include neither v_2 nor v_3.

The end

Thank you for your attention

Counterfactuals: Dawid (2000)/McCullagh (2022) critique

> *The good news is that the counterfactual extension provides an answer; the bad news is that every counterfactual extension provides a different answer. Moreover, there are infinitely many such extensions, all of which are indistinguishable on the basis of physical experiments.*
>
> (McCullagh, 2022, p. 244)

The objection is to modelling the "cross-world" between factual and counterfactual, for which no data are available because both are never simultaneously observed.