Influence of COVID-19 on disability and mortality: a multiverse analysis with post-selection inference

EAPS Health, Morbidity and Mortality Working Group
11 Sept. 2025

Anna Vesely and Rossella Miglio

University of Bologna - anna.vesely2@unibo.it

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAI lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006450001

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Motivation

In real data analysis, researchers face many choices:

- variable transformation (log, sqrt, splines, etc.)
- inclusion of covariates and interactions
- outlier and/or leverage point removal
- ...

Often these decisions

- are arbitrary
- are based on subjective beliefs
- have equally justifiable alternatives

This range of choices can be abused $\longrightarrow$ replicability crisis

Motivation

In real data analysis, researchers face many choices:

- variable transformation (log, sqrt, splines, etc.)
- inclusion of covariates and interactions
- outlier and/or leverage point removal
- ...

Often these decisions

- are arbitrary
- are based on subjective beliefs
- have equally justifiable alternatives

This range of choices can be abused $\longrightarrow$ replicability crisis

Motivation

In real data analysis, researchers face many choices:

- variable transformation (log, sqrt, splines, etc.)
- inclusion of covariates and interactions
- outlier and/or leverage point removal
- ...

Often these decisions

- are arbitrary
- are based on subjective beliefs
- have equally justifiable alternatives

This range of choices can be abused $\longrightarrow$ replicability crisis

p-hacking and the replicability crisis

p-hacking (data snooping or data dredging)

Performing many statistical tests on the same data and only reporting those that give significant results

Consequences

Dramatically increases and understates the risk of false positives

This is a main reason of the replicability crisis in psychology, neuroscience, biology, economics, etc.[1]

[1] Ioannidis. Why most published research findings are false. *PLoS Med.*, 2005.

Influence of COVID-19 on disability and mortality

Survey of Health, Ageing, and Retirement in Europe (SHARE)

SHARE[1] is a longitudinal study of adults 50+ and household members

Does SARS-CoV-2 infection increase the likelihood of health deterioration?

- Wave 8 (2019/20) → select healthy subjects
- Corona-specific supplementary survey (June/August 2020) → register COVID-19-related events
- Wave 9 (2021/22) → register declining health

[1] Börsch-Supan et al. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *Int. J. Epidemiol.*, 2013.

Survey of Health, Ageing, and Retirement in Europe (SHARE)

SHARE[1] is a longitudinal study of adults 50+ and household members

Does SARS-CoV-2 infection increase the likelihood of health deterioration?

- Wave 8 (2019/20) → select healthy subjects
- Corona-specific supplementary survey (June/August 2020) → register COVID-19-related events
- Wave 9 (2021/22) → register declining health

[1] Börsch-Supan et al. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *Int. J. Epidemiol.*, 2013.

Variables of interest

Predictor of interest: indicator X of COVID-19-related events

- self-reported symptoms
- positive test
- hospitalization

Response variable: indicator Y of disability onset or mortality

- self-reported Global Activity Limitation Index (GALI)
- death

Logit model and hypothesis testing

$$y_i \sim \text{Bernoulli}(p_i), \qquad g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta x_i + \boldsymbol{\gamma} \mathbf{z}_i$$

$$H_0 : \beta = 0 \;\text{ vs }\; H_1 : \beta \neq 0$$

Nuisance covariates:

- gender
- age
- area
- presence of a partner
- education level
- financial stress
- pre-existing chronic illnesses

Model specifications

How to specify nuisance covariates?

- age: linear/quadratic/splines/4 classes
- area: 3/5 classes
- partner: any/cohabiting
- education level: binary/3 classes
- financial stress: binary/4 classes
- chronic illnesses: number/1+/2+

Any interaction with gender?

- predictor: excluded/included and tested
- other covariates: always included

$\longrightarrow$ 384 possible models, 576 statistical tests

Model specifications

How to specify nuisance covariates?

- age: linear/quadratic/splines/4 classes
- area: 3/5 classes
- partner: any/cohabiting
- education level: binary/3 classes
- financial stress: binary/4 classes
- chronic illnesses: number/1+/2+

Any interaction with gender?

- predictor: excluded/included and tested
- other covariates: always included

⟶ 384 possible models, 576 statistical tests

Model specifications

How to specify nuisance covariates?

- age: linear/quadratic/splines/4 classes
- area: 3/5 classes
- partner: any/cohabiting
- education level: binary/3 classes
- financial stress: binary/4 classes
- chronic illnesses: number/1+/2+

Any interaction with gender?

- predictor: excluded/included <u>and tested</u>
- other covariates: always included

$\longrightarrow$ 384 possible models, 576 statistical tests

Post-selection inference in multiverse analysis

Multiverse analysis[1]

'Don't hide what you tried, report all the p-values and discuss'

A philosophy of reporting the outcomes of many different analyses to explore:

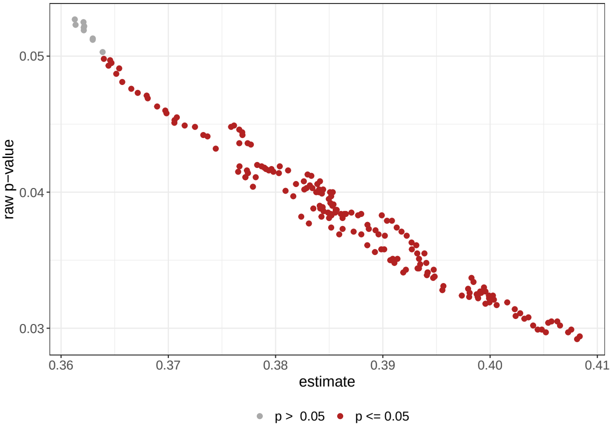
- robustness of results
- key choices that are most consequential in their fluctuation

Main tool: histogram of p-values, discussed in terms of % of significant p-values

[1]Steegen et al. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.*, 2016.

Results

- With `X:gender` interaction → no significant effect
- Without interaction → significant effects of COVID-19 in $183/192 = 95.3\%$ models



Multiverse analysis solves the problem! Really?

Quite a strong evidence, isn't it?

No! We don't get any inferential clue from it

Multiverse analysis is important to make data analysis transparent, but a formal inferential approach is missing

p-hacking is an informal selective inference problem
Let's make it formal and get p-values that account for this multiplicity!

Multiverse analysis solves the problem! Really?

Quite a strong evidence, isn't it?

No! We don't get any inferential clue from it

Multiverse analysis is important to make data analysis transparent, but a formal inferential approach is missing

p-hacking is an informal selective inference problem
Let's make it formal and get p-values that account for this multiplicity!

Multiverse analysis solves the problem! Really?

Quite a strong evidence, isn't it?

No! We don't get any inferential clue from it

Multiverse analysis is important to make data analysis transparent, but a formal inferential approach is missing

p-hacking is an informal selective inference problem
Let's make it formal and get p-values that account for this multiplicity!

Post-selection Inference in Multiverse Analysis (PIMA)

PIMA[1] combines information from all specifications to construct permutation-based test statistics/p-values

? Is there any non-null effect among the tested models?

! Global p-value (weak FWER control)

Similar to Specification Curve[2], but valid for all GLMs

? Which models are significant?

! Adjusted p-values for each model (strong FWER control)

⟶ choose the model you like best!

[1]Girardi et al. Post-selection Inference in Multiverse Analysis (PIMA): An inferential framework based on the sign flipping score test. *Psychometrika*, 2024.

[2]Simonsohn et al. Specification curve analysis. *Nat. Hum. Behav*, 2020.

Post-selection Inference in Multiverse Analysis (PIMA)

PIMA[1] combines information from all specifications to construct permutation-based test statistics/p-values

- **?** Is there any non-null effect among the tested models?
- **!** Global p-value (weak FWER control)
  Similar to Specification Curve[2], but valid for all GLMs

- **?** Which models are significant?
- **!** Adjusted p-values for each model (strong FWER control)
  $\longrightarrow$ choose the model you like best!

[1] Girardi et al. Post-selection Inference in Multiverse Analysis (PIMA): An inferential framework based on the sign flipping score test. *Psychometrika*, 2024.

[2] Simonsohn et al. Specification curve analysis. *Nat. Hum. Behav*, 2020.

The models, the tested hypotheses

Consider K plausible general linear models (GLMs):

$$g_k(\mathbb{E}(y_{ki})) = \beta_k x_{ki} + \gamma_k \mathbf{z}_{ki} \qquad (i = 1, \ldots, n)$$

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and $\mathbf{z}_{ki}$: transformed predictors $\longrightarrow$ selection, combination and transformation

Hypothesis testing

$$\text{Model } k\text{: } H_{0k} : \beta_k = 0, \qquad \text{Global null: } H_0 : \bigcap_{k=1}^{K} H_{0k}$$

The models, the tested hypotheses

Consider K plausible general linear models (GLMs):

$$g_k(\mathbb{E}(y_{ki})) = \beta_k x_{ki} + \gamma_k \mathbf{z}_{ki} \qquad (i = 1, \ldots, n)$$

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and $\mathbf{z}_{ki}$: transformed predictors $\longrightarrow$ selection, combination and transformation

Hypothesis testing

Model k: $H_{0k} : \beta_k = 0$, Global null: $H_0 : \bigcap_{k=1}^{K} H_{0k}$

The models, the tested hypotheses

Consider K plausible general linear models (GLMs):

$$g_k(\mathbb{E}(y_{ki})) = \beta_k x_{ki} + \gamma_k \mathbf{z}_{ki} \qquad (i = 1, \ldots, n)$$

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and $\mathbf{z}_{ki}$: transformed predictors $\longrightarrow$ selection, combination and transformation

Hypothesis testing

$$\text{Model } k\text{: } H_{0k} : \beta_k = 0, \qquad \text{Global null: } H_0 : \bigcap_{k=1}^{K} H_{0k}$$

The models, the tested hypotheses

Consider K plausible general linear models (GLMs):

$$g_k(\mathbb{E}(y_{ki})) = \beta_k x_{ki} + \gamma_k \mathbf{z}_{ki} \qquad (i = 1, \ldots, n)$$

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and $\mathbf{z}_{ki}$: transformed predictors $\longrightarrow$ selection, combination and transformation

Hypothesis testing

$$\text{Model } k\text{: } H_{0k} : \beta_k = 0, \qquad \text{Global null: } H_0 : \bigcap_{k=1}^{K} H_{0k}$$

Properties of PIMA

- Can be used whenever we can write a score test (GLMs and much more)
- Asymptotically exact (exact, in practice[1])
- Very robust to variance misspecification, if the link function is correctly specified
- Can be extended to the case of multiple parameters of interest

[1] De Santis et al. Inference in generalized linear models with robustness to misspecified variances. *JASA*, 2025.

But... Multiverse is a slippery floor

Multiverse doesn't solve the problem of validity of the assumptions: if the model is wrong, a significant p-value doesn't mean anything

For instance, if the true model is
`Y ∼ X + gender + age + X:gender`
then a model without the interaction `X:gender` is wrong

Indeed, residuals are not independent/normal/etc., and the test on X may fail to control the type I error

Think before testing!

But... Multiverse is a slippery floor

Multiverse doesn't solve the problem of validity of the assumptions: if the model is wrong, a significant p-value doesn't mean anything

For instance, if the true model is
`Y ~ X + gender + age + X:gender`
then a model without the interaction `X:gender` is wrong

Indeed, residuals are not independent/normal/etc., and the test on X may fail to control the type I error

Think before testing!

But... Multiverse is a slippery floor

Multiverse doesn't solve the problem of validity of the assumptions: if the model is wrong, a significant p-value doesn't mean anything

For instance, if the true model is
`Y ~ X + gender + age + X:gender`
then a model without the interaction `X:gender` is wrong

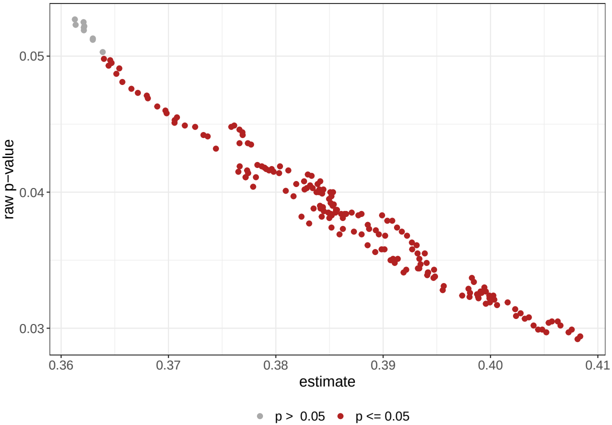
Indeed, residuals are not independent/normal/etc., and the test on X may fail to control the type I error

Think before testing!

Results

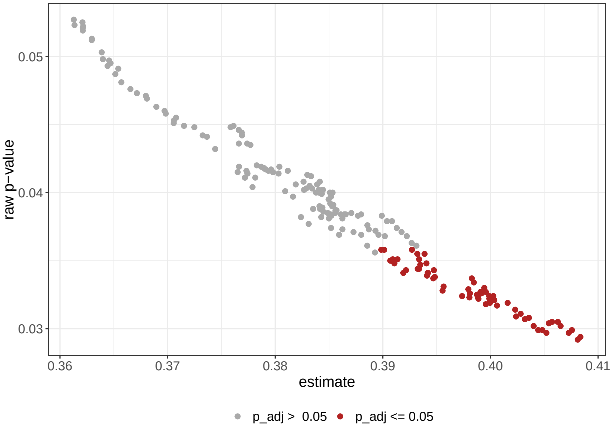
Raw (unadjusted) p-values

- With `X:gender` interaction → no significant effect
- Without interaction → significant effects of COVID-19 in $183/192 = 95.3\%$ models



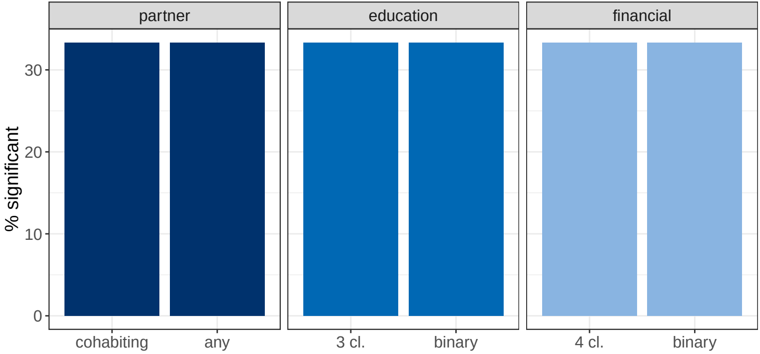
Adjusted p-values, strong FWER control

- With `X:gender` interaction → no significant effect
- Without interaction → significant effects of COVID-19 in $64/192 = 33.3\%$ models



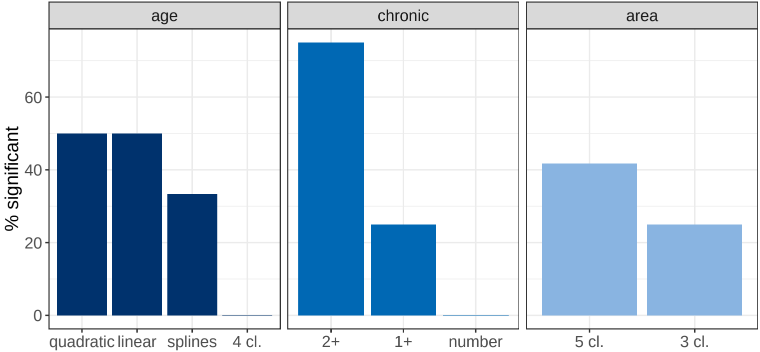
Some data-analytic choices do not affect results...

Models without `X:gender` interaction:



... while others are influential

Models without `X:gender` interaction:



Influential data-analytic choices

- Interaction `X:gender` - inclusion leads to less precise estimates and higher standard errors
- Age - non-linear effects may not be captured by grouping
- Chronic illnesses - differences reflect variations in the type of illnesses, with some being more common and milder (hypertension, high colesterol)
- Geographical level

Conclusion

- Significant effects arise only from specific modeling choices
- Some results may be due to inadequate modeling, which can induce spurious correlations
- A multiverse approach provides deeper insights into the analysis
- Every significant test must be evaluated with care

Main reference: Vesely and Miglio. Influence of COVID-19 on disability and mortality: a multiverse analysis with post-selection inference. *Statistics for Innovation IV*, 2025.

Appendix

Nuisance covariates: geographical location

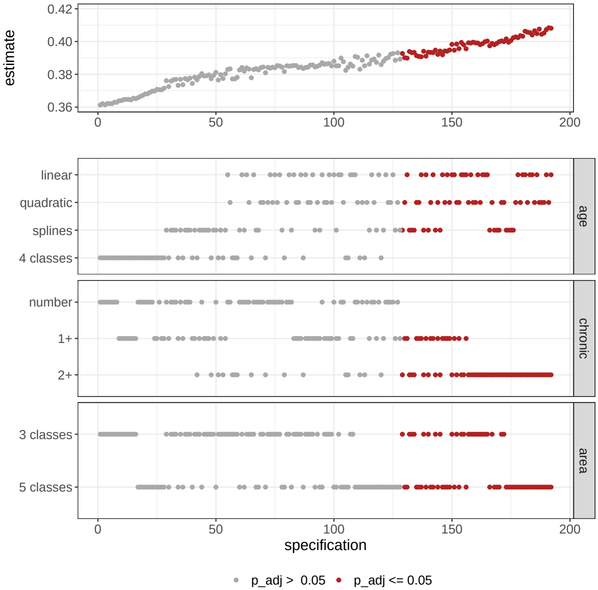
Areas

- West - Germany, Austria, Switzerland, France, Belgium, Netherlands, Luxembourg
- South - Spain, Italy, Greece, Malta
- East - Slovenia, Czech Republic, Slovakia, Poland, Hungary
- North - Sweden, Denmark, Finland
- Baltic/Balkan - Croatia, Romania, Bulgaria, Estonia, Lithuania, Latvia, Cyprus

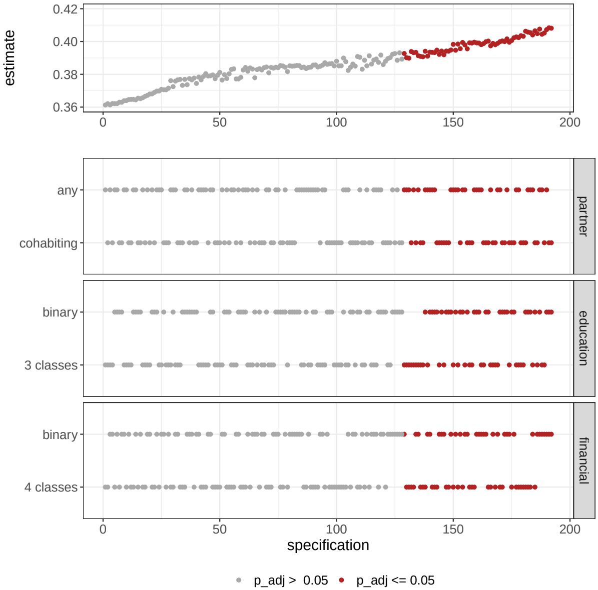
Macroareas

- Bismarck: payroll-based social health insurance - West, East
- Beveridge: tax-funded national health service - North, South
- hybrid/transitioning - Baltic/Balkan

Results: influential data-analytic choices



Results: non-influential data-analytic choices



Basis of PIMA: sign flip score test (univariate)[1]

Single model: n independent observations with density $f_{\beta,\gamma,x_i,\mathbf{z}_i}(y_i)$

Score test: $T^1 = T^{\text{obs}} = \sum_{i=1}^{n} \nu_i, \qquad \nu_i = \frac{\partial}{\partial \beta} \log f_{\beta,\gamma,x_i,\mathbf{z}_i}(y_i) \mid_{\hat{\gamma},\beta=0}$

Random sign flips: $T^b = \sum_{i=1}^{n} \pm \nu_i \qquad (b = 2, \ldots, B)$

Under $H_0 : \beta = 0$: $T^{\text{obs}} \overset{d}{=} T^b$ asymptotically

$$\text{p-value} = \frac{\#_b(T^b \geq T^{\text{obs}})}{B}$$

[1] Hemerik et al. Robust testing in generalized linear models by sign flipping score contributions. *JRSS-B*, 2020.

Basis of PIMA: sign flip score test (univariate)[1]

Single model: n independent observations with density $f_{\beta,\gamma,x_i,\mathbf{z}_i}(y_i)$

Score test: $$T^1 = T^{\text{obs}} = \sum_{i=1}^{n} \nu_i, \qquad \nu_i = \frac{\partial}{\partial \beta} \log f_{\beta,\gamma,x_i,\mathbf{z}_i}(y_i) \mid_{\hat{\gamma},\beta=0}$$

Random sign flips: $$T^b = \sum_{i=1}^{n} \pm \nu_i \qquad (b = 2, \ldots, B)$$

Under $H_0 : \beta = 0$: $T^{\text{obs}} \overset{d}{=} T^b$ asymptotically

$$\text{p-value} = \frac{\#_b(T^b \geq T^{\text{obs}})}{B}$$

[1]Hemerik et al. Robust testing in generalized linear models by sign flipping score contributions. *JRSS-B*, 2020.

Basis of PIMA: sign flip score test (univariate)[1]

Single model: n independent observations with density $f_{\beta,\gamma,x_i,\mathbf{z}_i}(y_i)$

Score test: $T^1 = T^{\text{obs}} = \sum_{i=1}^{n} \nu_i, \qquad \nu_i = \frac{\partial}{\partial \beta} \log f_{\beta,\gamma,x_i,\mathbf{z}_i}(y_i) \mid_{\hat{\gamma},\beta=0}$

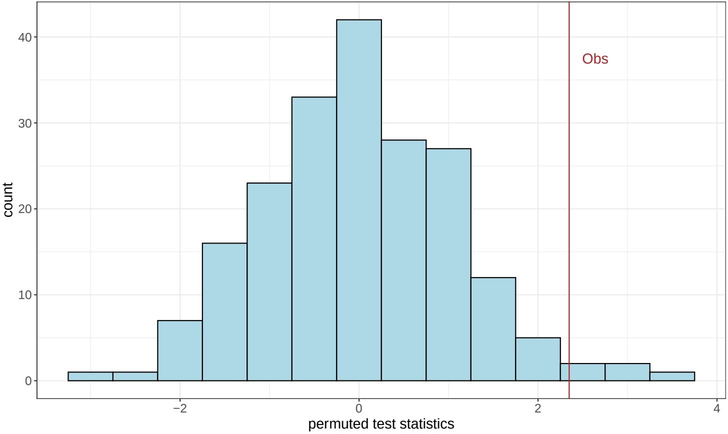
Random sign flips: $T^b = \sum_{i=1}^{n} \pm \nu_i \qquad (b = 2, \ldots, B)$

Under $H_0 : \beta = 0$: $T^{\text{obs}} \overset{d}{=} T^b$ asymptotically

$$\text{p-value} = \frac{\#_b(T^b \geq T^{\text{obs}})}{B}$$

[1] Hemerik et al. Robust testing in generalized linear models by sign flipping score contributions. *JRSS-B*, 2020.

Permutation test



$$\text{p-value} = \frac{\#_b(T^b \geq T^{\text{obs}})}{B}$$

Joint sign flip scores tests (multivariate)

K models:

K score test statistics: $(T_1^{\text{obs}}, \dots, T_K^{\text{obs}})$

Random sign flips: $(T_1^b, \dots, T_K^b)$ $(b = 2, \dots, B)$

obtained by jointly flipping the signs of the K-variate contributions

$$\pm(\nu_{1i}, \dots, \nu_{Ki})$$

$\longrightarrow$ each observation i is subject to the same sign flips in all K models

Joint sign flip scores tests (multivariate)

$$\text{sign flips}\quad \overset{\text{models}}{\begin{matrix} T_1^{\text{obs}} & \cdots & T_K^{\text{obs}} \\ T_1^2 & \cdots & T_K^2 \\ \vdots & & \vdots \\ T_1^B & \cdots & T_K^B \end{matrix}} \quad \xrightarrow{\psi} \quad \overset{\text{joint}}{\begin{matrix} T^{\text{obs}} \\ T^2 \\ \vdots \\ T^B \end{matrix}}$$

ψ: suitable combining function, such as the (weighted) mean and the maximum

Under $H_0 : \beta_1 = \ldots = \beta_K = 0$:

$$(T_1^{\text{obs}}, \ldots, T_K^{\text{obs}}) \overset{d}{=} (T_1^b, \ldots, T_K^b) \text{ asymptotically}$$

$$\implies T^{\text{obs}} \overset{d}{=} T^b \text{ asymptotically}$$

Joint sign flip scores tests (multivariate)

$$
\begin{array}{c|ccc|c|c}
 & & \text{models} & & & \text{joint} \\
\hline
\text{sign flips} & T_1^{\text{obs}} & \dots & T_K^{\text{obs}} & & T^{\text{obs}} \\
 & T_1^2 & \dots & T_K^2 & \xrightarrow{\psi} & T^2 \\
 & \vdots & & \vdots & & \vdots \\
 & T_1^B & \dots & T_K^B & & T^B
\end{array}
$$

ψ: suitable combining function, such as the (weighted) mean and the maximum

Under $H_0 : \beta_1 = \ldots = \beta_K = 0$:

$$(T_1^{\text{obs}}, \ldots, T_K^{\text{obs}}) \overset{d}{=} (T_1^b, \ldots, T_K^b) \text{ asymptotically}$$

$$\implies T^{\text{obs}} \overset{d}{=} T^b \text{ asymptotically}$$

Joint sign flips of the score contributions

$$\begin{array}{llcl} +\nu_{11} & +\nu_{12} & \dots & +\nu_{1K} \\ +\nu_{21} & +\nu_{22} & \dots & +\nu_{2K} \\ \vdots & \vdots & & \vdots \\ +\nu_{n1} & +\nu_{n2} & \dots & +\nu_{nK} \end{array}$$

					combined
obs	T_1^{obs}	T_2^{obs}	$\dots$	T_K^{obs}	$T^{\text{obs}} = \max\{T_k^{\text{obs}}\}$

Joint sign flips of the score contributions

$$\begin{array}{cccc} -\nu_{11} & -\nu_{12} & \cdots & -\nu_{1K} \\ +\nu_{21} & +\nu_{22} & \cdots & +\nu_{2K} \\ \vdots & \vdots & & \vdots \\ -\nu_{n1} & -\nu_{n2} & \cdots & -\nu_{nK} \end{array}$$

					combined
obs	T_1^{obs}	T_2^{obs}	$\dots$	T_K^{obs}	$T^{\text{obs}} = \max\{T_k^{\text{obs}}\}$
perm(2)	T_1^2	T_2^2	$\dots$	T_K^2	$T^2 = \max\{T_k^2\}$

Joint sign flips of the score contributions

$$\begin{array}{cccc} +\nu_{11} & +\nu_{12} & \dots & +\nu_{1K} \\ -\nu_{21} & -\nu_{22} & \dots & -\nu_{2K} \\ \vdots & \vdots & & \vdots \\ +\nu_{n1} & +\nu_{n2} & \dots & +\nu_{nK} \end{array}$$

					combined
obs	T_1^{obs}	T_2^{obs}	$\dots$	T_K^{obs}	$T^{\text{obs}} = \max\{T_k^{\text{obs}}\}$
perm(2)	T_1^2	T_2^2	$\dots$	T_K^2	$T^2 = \max\{T_k^2\}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
perm(B)	T_1^B	T_2^B	$\dots$	T_K^B	$T^B = \max\{T_k^B\}$

Post-hoc inference

? Is there any non-null effect among the tested models?

! Global p-value defined from $(T^{\text{obs}}, \ldots, T^B)$

? Which models are significant?

! Adjusted p-values for each model using the maxT

Post-hoc inference

? Is there any non-null effect among the tested models?

! Global p-value defined from $(T^{\text{obs}}, \dots, T^B)$

? Which models are significant?

! Adjusted p-values for each model using the maxT