PIMA: An inferential framework for multiverse analysis

Girardi, P. et al. (2024) Psychometrika.
https://doi.org/10.1007/s11336-024-09973-6

L. Finos

9 Dec., 2024

University of Padova
livio.finos@unipd.it

A leading example

In real data analysis, researchers face many choices:

- variable transformation (log, sqrt, splines, etc.)
- inclusion of covariates and interactions
- outlier deletion
- ...

Example

- one over 4 possible predictors X_1, X_2, X_3, X_4
- *gender* + (a subset of) other 4 covariates/mediators
- possible interaction between X_1/X_2 and *gender*

$\longrightarrow$ We easily get lost in the forest of possible models!

A leading example

In real data analysis, researchers face many choices:

- variable transformation (log, sqrt, splines, etc.)
- inclusion of covariates and interactions
- outlier deletion
- ...

Example

- one over 4 possible predictors X_1, X_2, X_3, X_4
- *gender* + (a subset of) other 4 covariates/mediators
- possible interaction between X_1/X_2 and *gender*

$\longrightarrow$ We easily get lost in the forest of possible models!

A leading example

In real data analysis, researchers face many choices:

- variable transformation (log, sqrt, splines, etc.)
- inclusion of covariates and interactions
- outlier deletion
- ...

Example

- one over 4 possible predictors X_1, X_2, X_3, X_4
- *gender* + (a subset of) other 4 covariates/mediators
- possible interaction between X_1/X_2 and *gender*

$\longrightarrow$ We easily get lost in the forest of possible models!

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.

Multiverse analysis[1] solves the problem!

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

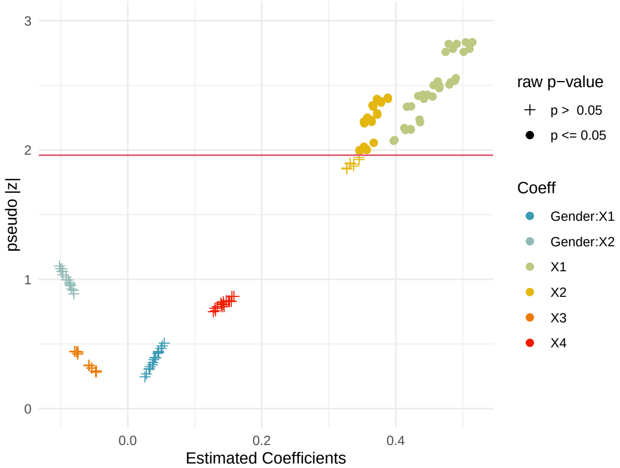
A philosophy of reporting the outcomes of many different analyses to explore:

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

Main tool: histogram of p-values
$\longrightarrow$ discussed in terms of % of significant p-values

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

Results: p-values in the example



$$\text{pseudo } |z| = \text{qnorm}(1 - p/2)$$

Multiverse analysis solves the problem! Really?

Ok, let's go multiverse!
43% of the tested coefficients have $p \leq 0.05$.
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.
Make it formal and get p-values that account for this multiplicity!

Multiverse analysis solves the problem! Really?

Ok, let's go multiverse!
43% of the tested coefficients have $p \leq 0.05$.
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.
Make it formal and get p-values that account for this multiplicity!

Multiverse analysis solves the problem! Really?

Ok, let's go multiverse!
43% of the tested coefficients have $p \leq 0.05$.
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.
Make it formal and get p-values that account for this multiplicity!

Valid p-hacking via PIMA[1]

PIMA constructs permutation-based test statistics/p-values, combining information from all plausible models

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

- **?** Which models are significant?
- **!** Adjusted p-values for each model (strong FWER control)
  using the maxT algorithm → choose the model you like best!

- **?** How many models are significant? (How many for a given predictor/transformation/model-choice)
- **!** Confidence interval for the proportion (TDP) via closed testing
  using pARI, SumSome.. or NOTIP!

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

PIMA

The models, the tested hypotheses

Consider K plausible general linear models (GLM):

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

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

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

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

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

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

Hypothesis testing

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

The models, the tested hypotheses

Consider K plausible general linear models (GLM):

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

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

Hypothesis testing

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

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

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.

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

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

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

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.

Sign flip score test (univariate+multivariate)

Two refinements:

- effective score (more powerful) [1]
- standardized effective score
  ('almost' exact type I error in finite sample)[2]

Extension to Multivariate responses

- Fit a model for each response (each model possbily with different predictors and/or responses), joint distribution is dealt simply[3]

[1] Hemerik, Goeman and Finos (2020) JRSS-B

[2] De Santis et al. Inference in generalized linear models with robustness to misspecified variances. *ArXiv*, 2024.

[3] De Santis, Goeman, Davenport, Hemerik, Finos (2024) Permutation-based multiple testing when fitting many generalized linear models, arXiv

Sign flip score test (univariate+multivariate)

Two refinements:

- effective score (more powerful) [1]
- standardized effective score
  ('almost' exact type I error in finite sample)[2]

Extension to Multivariate responses

- Fit a model for each response (each model possbily with different predictors and/or responses), joint distribution is dealt simply[3]

[1] Hemerik, Goeman and Finos (2020) JRSS-B

[2] De Santis et al. Inference in generalized linear models with robustness to misspecified variances. *ArXiv*, 2024.

[3] De Santis, Goeman, Davenport, Hemerik, Finos (2024) Permutation-based multiple testing when fitting many generalized linear models, arXiv

Joint sign flip scores test

K models:

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

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

obtained by jointly flipping the signs of $\pm(\nu_{1i}, \ldots, \nu_{Ki})$

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

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

A multiverse p-value is obtained combining the single tests (e.g., $T^b = max\{T_1^b, \ldots, T_K^b\}$)

Joint sign flip scores test

K models:

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

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

obtained by jointly flipping the signs of $\pm(\nu_{1i}, \ldots, \nu_{Ki})$

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

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

A multiverse p-value is obtained combining the single tests (e.g., $T^b = max\{T_1^b, \ldots, T_K^b\})$

Joint sign flips of the score contributions

$$\begin{array}{lccccc}
 & +\nu_{11} & +\nu_{12} & \dots & +\nu_{1K} & \\
 & +\nu_{21} & +\nu_{22} & \dots & +\nu_{2K} & \\
 & \vdots & \vdots & & \vdots & \\
 & +\nu_{n1} & +\nu_{n2} & \dots & +\nu_{nK} & \\
 & & & & & \text{combined} \\
\hline
\text{obs} & T_1^{\text{obs}} & T_2^{\text{obs}} & \dots & T_K^{\text{obs}} & T^{\text{obs}} = \max\{T_k^{\text{obs}}\}
\end{array}$$

Joint sign flips of the score contributions

$$
\begin{array}{lccccc}
 & -\nu_{11} & -\nu_{12} & \dots & -\nu_{1K} & \\
 & +\nu_{21} & +\nu_{22} & \dots & +\nu_{2K} & \\
 & \vdots & \vdots & & \vdots & \\
 & -\nu_{n1} & -\nu_{n2} & \dots & -\nu_{nK} & \\
 & & & & & \text{combined} \\
\hline
\text{obs} & T_1^{\text{obs}} & T_2^{\text{obs}} & \dots & T_K^{\text{obs}} & T^{\text{obs}} = \max\{T_k^{\text{obs}}\} \\
\text{perm}(2) & T_1^2 & T_2^2 & \dots & T_K^2 & T^2 = \max\{T_k^2\}
\end{array}
$$

Joint sign flips of the score contributions

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

$$\begin{array}{cccccc} & & & & & \text{combined} \\ \hline \text{obs} & T_1^{\text{obs}} & T_2^{\text{obs}} & \dots & T_K^{\text{obs}} & T^{\text{obs}} = \max\{T_k^{\text{obs}}\} \\ \text{perm}(2) & T_1^2 & T_2^2 & \dots & T_K^2 & T^2 = \max\{T_k^2\} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \text{perm}(B) & T_1^B & T_2^B & \dots & T_K^B & T^B = \max\{T_k^B\} \end{array}$$

Properties

- Can be used whenever we can write a score test (GLMs and much more)
- Asymptotically exact (exact, in practice[1])
- Very robust to model - 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. *ArXiv*, 2024.

Simulation Study

Specification Curve, a good competitor?

Simonsohn, Simmons and Nelson (2020) in *Nature Human Behaviour*

- **First Paper** with inference in Multiverse!
- it proposes a solution via Bootstrap.
- Computationally very intensive: refit the multiverse $\times$ bootstrap
- Asymptotically ok in LM, but Very problematic in GLM
- It provides only the overall combination
  (i.e. no model selection, Weak FWER control)

- we don't discuss the alternative solution which is restricted to orthogonal designs and it has low power.

Simulation setting 1/2

Unobserved variable U

- Real: $g(\mu) = U\beta + Z\gamma + \gamma_0$
- $(U, Z) \sim$ Multivariate Normal, $\rho_{U,Z} = 0.6$.

Observed variables X_k (proxy of U):

- Fitted: $g(\mu) = X_k\beta_k + Z\gamma + \gamma_0$
- $(X_k, U) \sim$ Multivariate Normal, $\rho_{X_k,U} = 0.85$.

Multiverse analysis with five models:

- $H_0 : \ \beta_k = 0, \ k = 1, \dots, 5$

(5000 MC)

Simulation setting 2/2

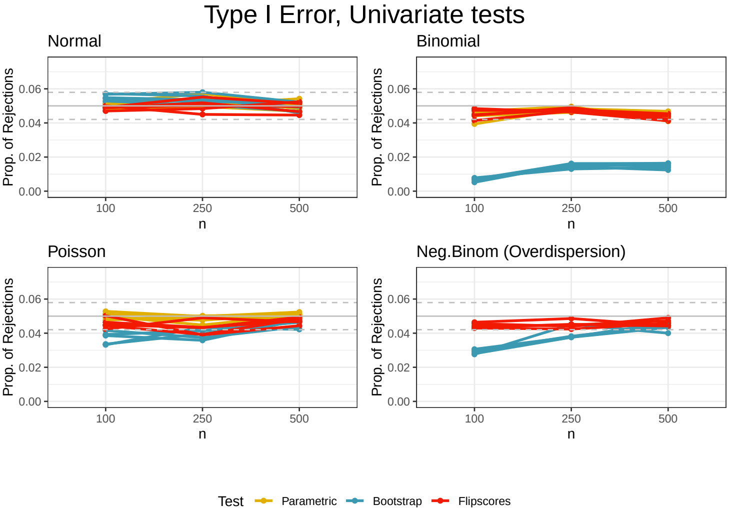
Scenarios,

1. LM with homoschedastic Gaussian errors:
2. Binomial logit-link model:
3. Poisson log-link model:
4. *Overdispersion*
   Real: Negative Binomial log-link model,
   Fitted: while Poisson log-link model.

Methods:

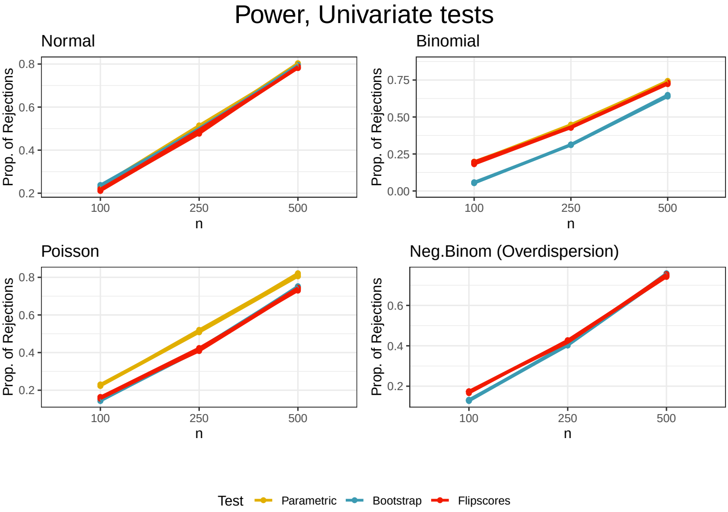
- Flipscores,
- Bootstrap (Simonsohn et al, 2020),
- Parametric test (t-test in LM, Wald-test in others GLMs).

Simulation: H_0, univariate



Prop. of Rejections for Parametric test in Neg Binom setting ranges between 0.154 and 0.170

Simulation: H_1, univariate



Prop. of Rejections for Parametric test in Neg Binom setting is not displayed since it does not control the type I error

Simulation: multivariate

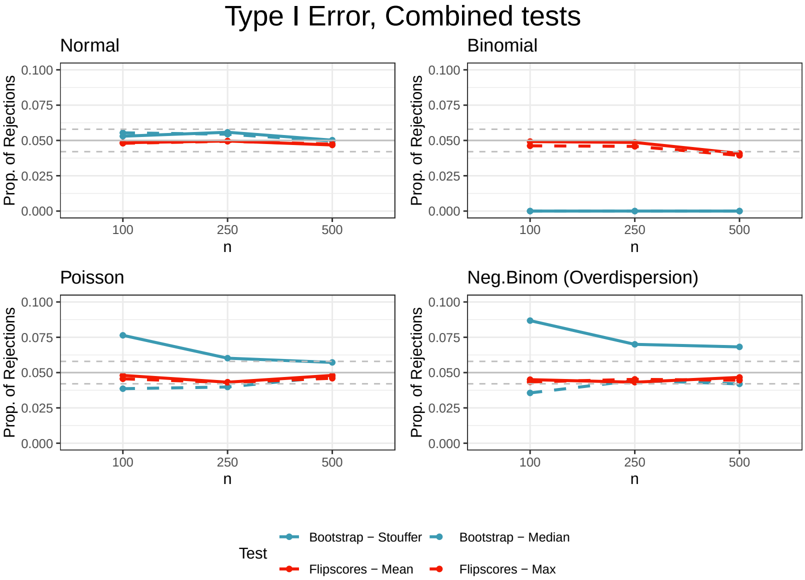
In order to ensure (strong) FWER control with any multiple testing procedure we must ensure control of the Type I error control of the combinated (i.e. multivariate) test of any of subsets of tested hypothesis (by Closed Testing principle).

Combining Methods:

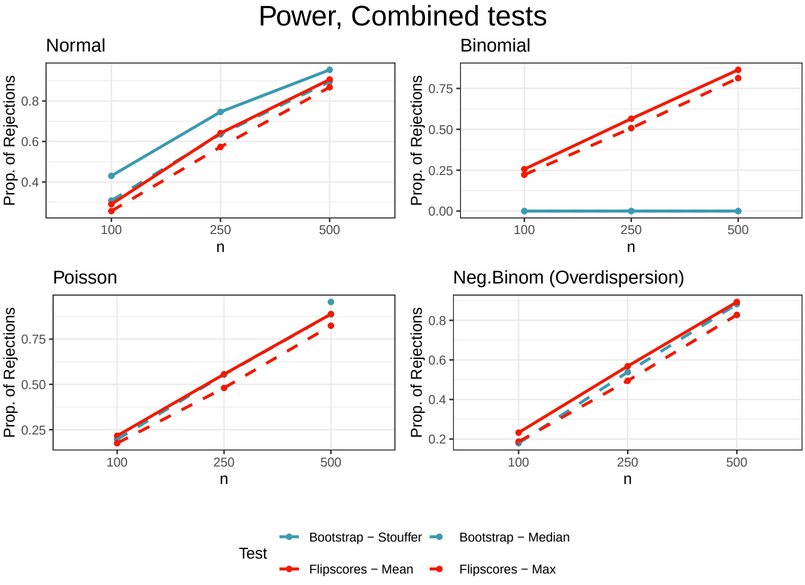
- **Flipscores**:
  - *Mean* of the test statistics,
  - *Max* of the test statistics,
- **Bootstrap**:
  - *Stouffer/Liptak* (Sum of the z-tranfomed p-values),
  - *Median* of the test statistics,
- Parametric: Bonferroni. Not shown because extremely conservative and under-powered).

Sims: combine the 5 tests (i.e. weak FWER)

Simulation: H_0, multivariate

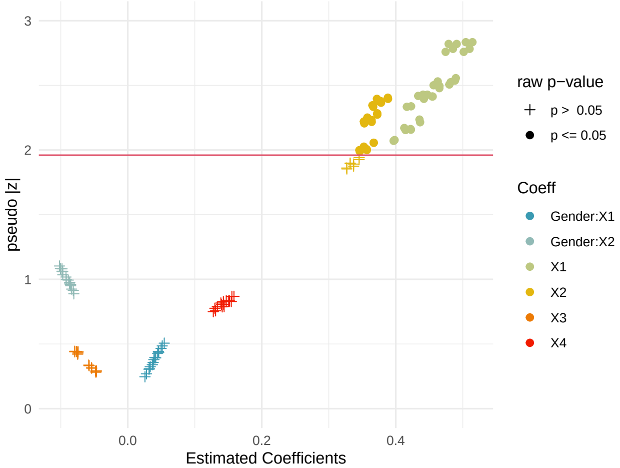


Simulation: H_1, multivariate



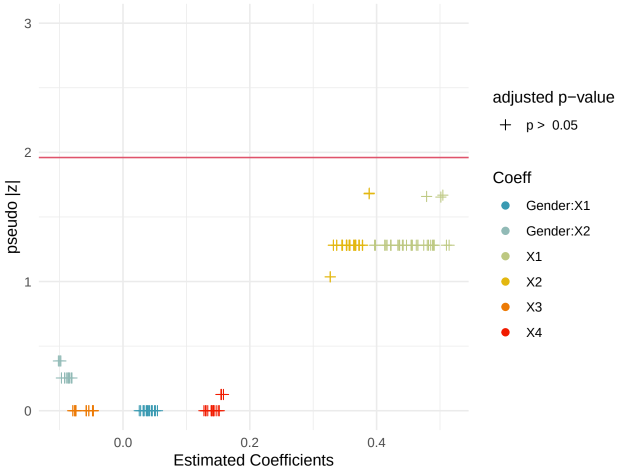
Results

Raw (unadjusted) p-values



Data were generated with no effects → all false positives!

Adjusted p-values, strong FWER control



Global p-value $\approx 0.09 \rightarrow$ all null effects

Take-home message

Accounting for selective inference (multiple testing, adjusted p-values) is crucial

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

! Take the global p-value

? How many models are significant? (How many for a given predictor/transformation/model-choice)

! Confidence interval for the proportion (TDP) via closed testing

? Which models are significant?

! Take the adjusted p-values and choose the model/story you like most

What is allowed and what is not

PIMA allows:

- any GLMs (and Cox models comming soon)
- any transformation of variables (predictors, responses)
- any outlier/leverage deletion method

BUT all the above models must be

- planned in advance
- valid (at least the right link)

There is no free lunch

Enjoy p-hacking, it is now valid!

`flipscores`: `github.com/livioivil/flipscores` and CRAN

- Sign flip score test: **GLMs** and any other model with score
- robust to some model misspecifications

`jointest`: `github.com/livioivil/jointest`

- inference framework for **multivariate** inference with `flipscores` (and more)
- FWER and (address to) TDP control

`pima`: `github.com/livioivil/pima`

- inference framework for **multiverse** analysis
- model picking with adjusted p-values
- see **vignettes** there