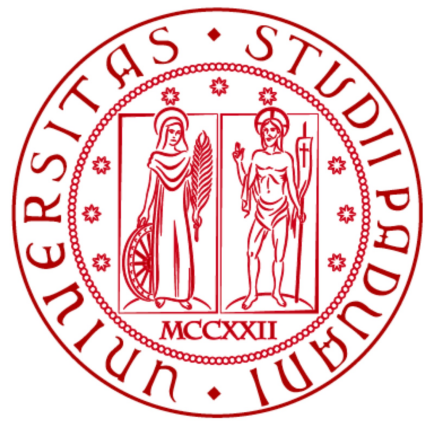
PIMA: An inferential framework for multiverse analysis

AIP – Sez. Sperimentale – Sept. 24, 2024

L. Finos, P. Girardi, A. Vesely,
D. Lakens, M. Pastore, A. Calcagnì,
G. Altoè

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 covariates/mediators
- possible interaction between X_1 or 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 covariates/mediators
- possible interaction between X_1 or 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 covariates/mediators
- possible interaction between X_1 or 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 the p-values and discuss'

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

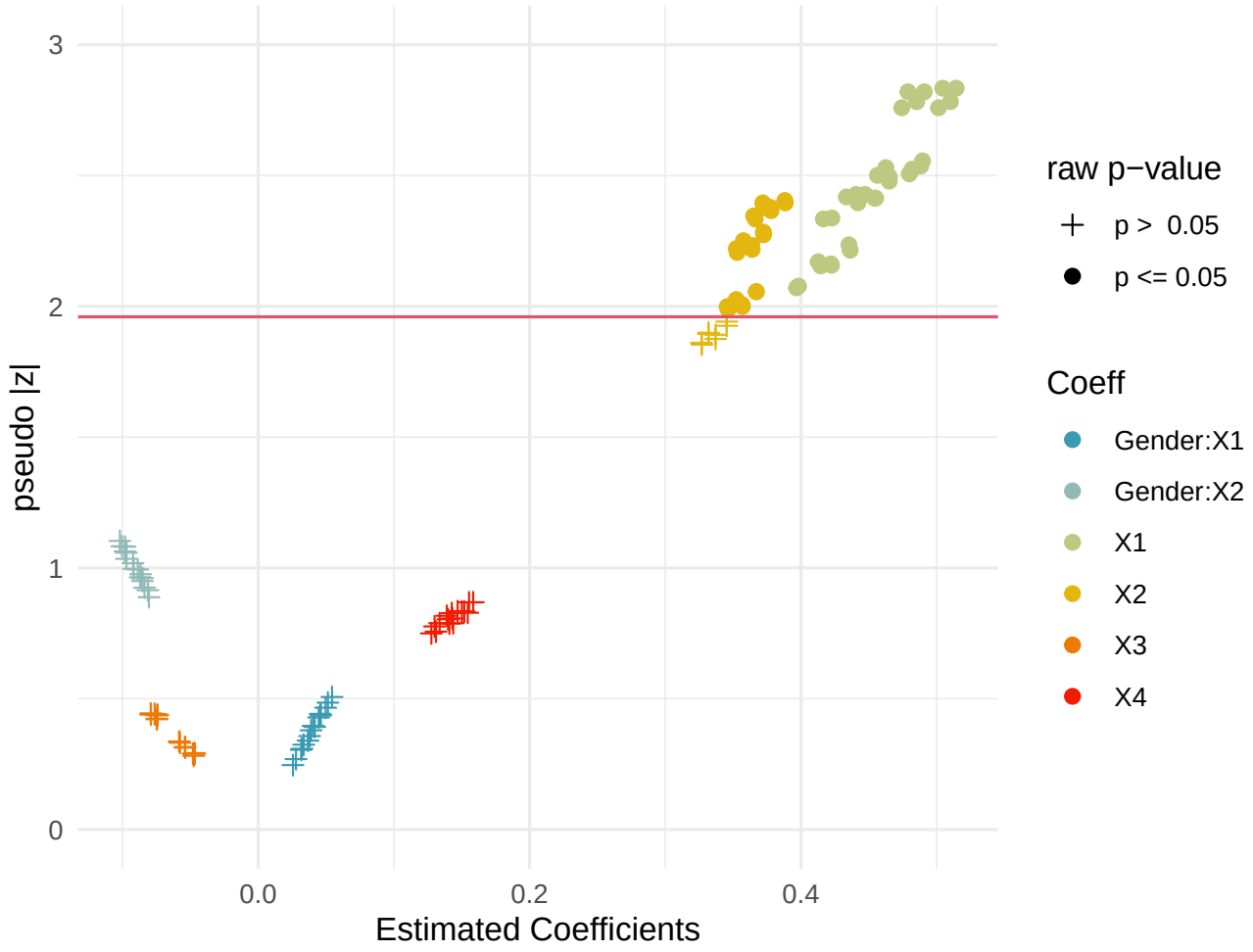
- robustness of results
- key choices that are most consequential in their fluctuation

Main tool: histogram of p-values

$\longrightarrow$ 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



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[2]

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)
Like Specification Curve[1], but done right

? Which models are significant?

! Adjusted p-values for each model (strong FWER control)
using the maxT algorithm → choose the model you like best!

[1] (not valid in GLM) Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208-1214.

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

Valid p-hacking via PIMA[2]

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)
Like Specification Curve[1], but done right

? Which models are significant?

! Adjusted p-values for each model (strong FWER control)
using the maxT algorithm → choose the model you like best!

[1](not valid in GLM) Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208-1214.

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

Valid p-hacking via PIMA[2]

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)
Like Specification Curve[1], but done right

? Which models are significant?

! Adjusted p-values for each model (strong FWER control)
using the maxT algorithm $\rightarrow$ choose the model you like best!

---

[1](not valid in GLM) Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208-1214.

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

PIMA

The models, the tested hypotheses

Consider K plausible general linear models (GLM):

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

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and 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 (GLM):

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

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and 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 (GLM):

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

- y_{ki}: response $\longrightarrow$ outlier deletion or leverage point removal
- x_{ki} and 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 (GLM):

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

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

Hypothesis testing

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

Sign flip score test (univariate)[1]

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

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

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

Under $H_0 : \beta = 0$: $T^{\text{obs}} \overset{d}{=} T^b$ 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 $f_{\beta,\boldsymbol{\gamma},x_i,\mathbf{z}_i}(y_i)$

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

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

Under $H_0 : \beta = 0$: $T^{\text{obs}} \overset{d}{=} T^b$ 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 $f_{\beta,\boldsymbol{\gamma},x_i,\mathbf{z}_i}(y_i)$

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

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.

Joint sign flip scores test

K models:

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

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

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: $(T_1^{obs}, \ldots, T_K^{obs})$

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

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

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

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

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

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

combined

---

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

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	$T^2 = \max\{T_k^2\}$

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

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

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

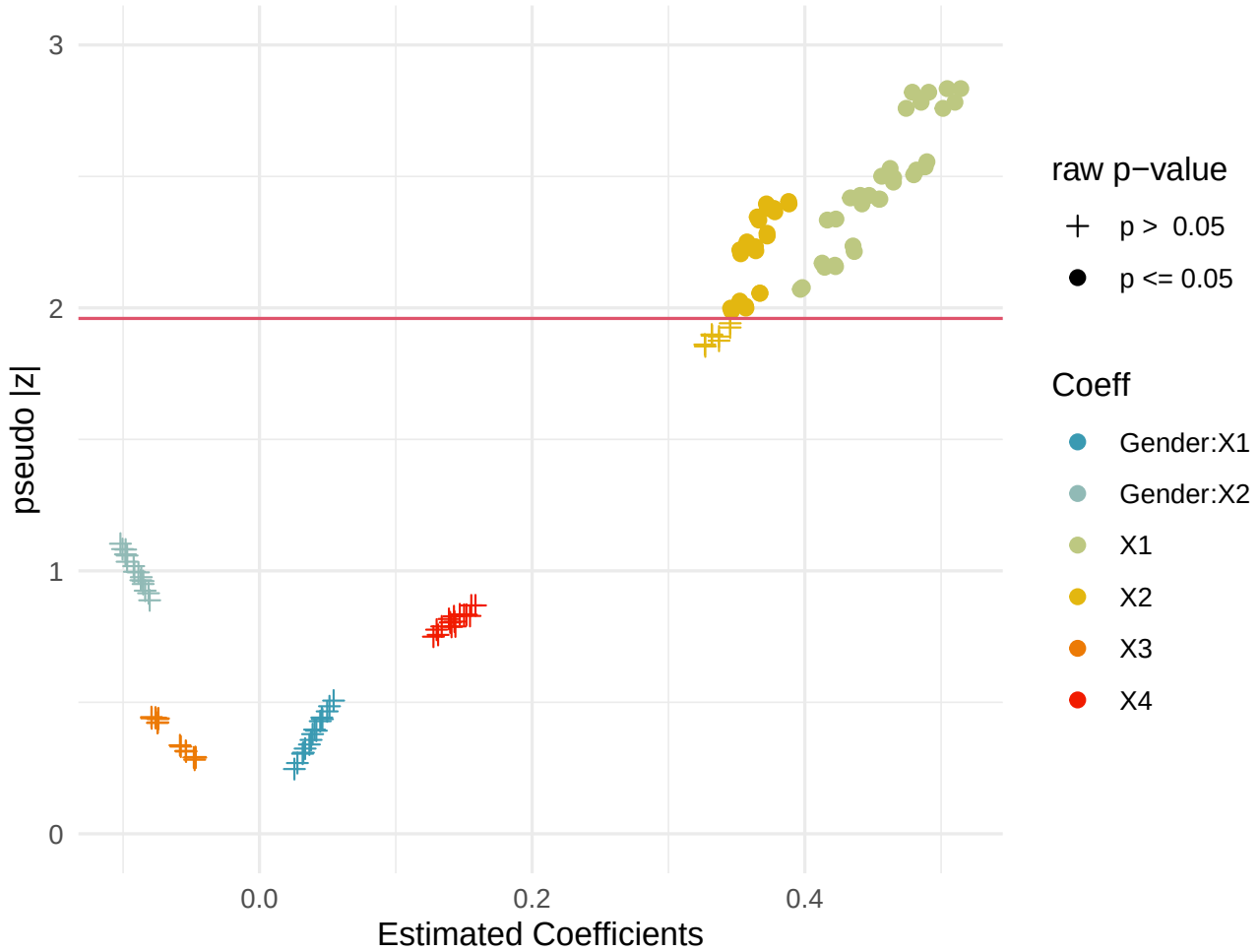
Properties

- 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. *ArXiv*, 2022.

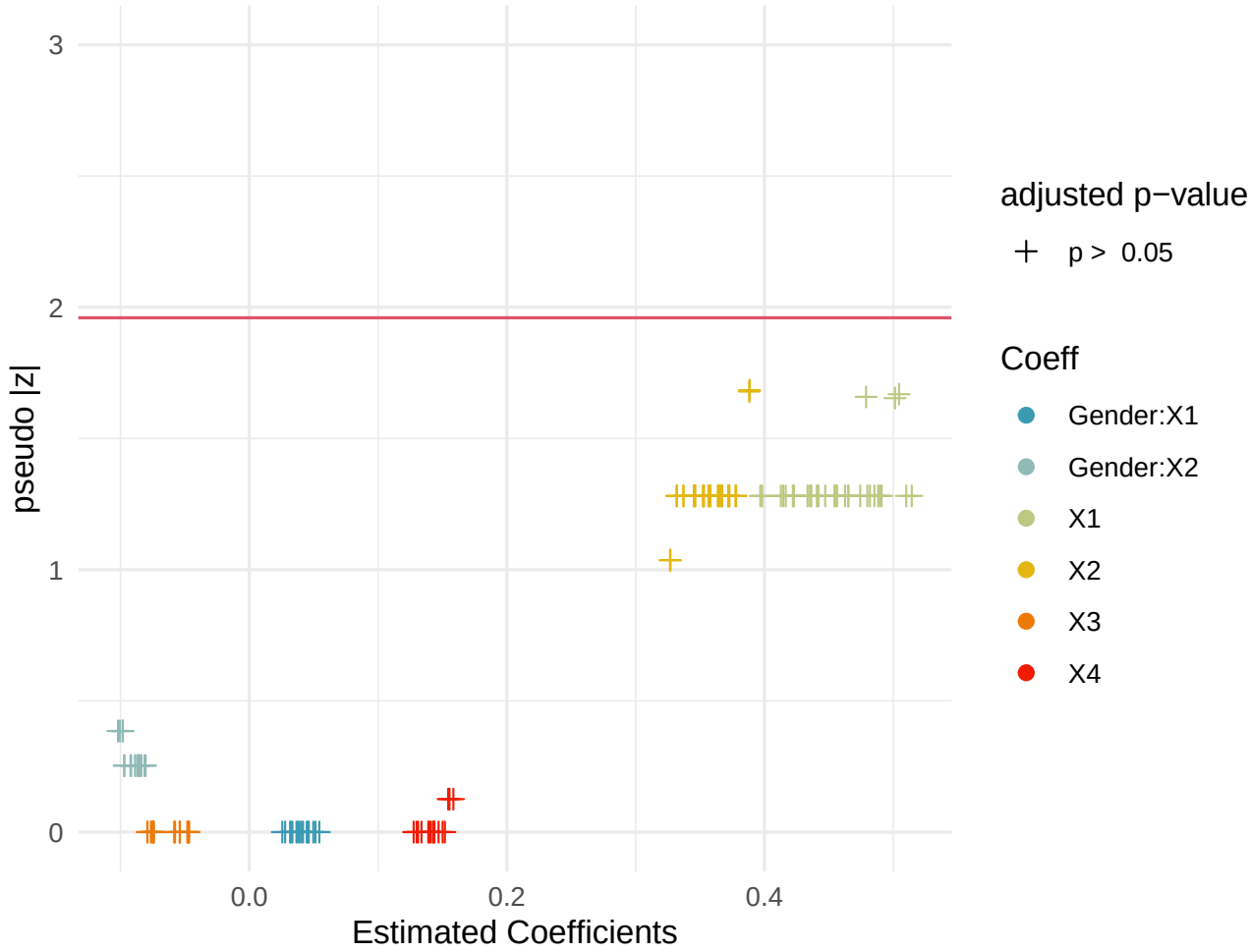
Results

Raw (unadjusted) p-values



Data are generated with no effects at all, these are ALL False Positives!

Adjusted p-values, strong FWER control

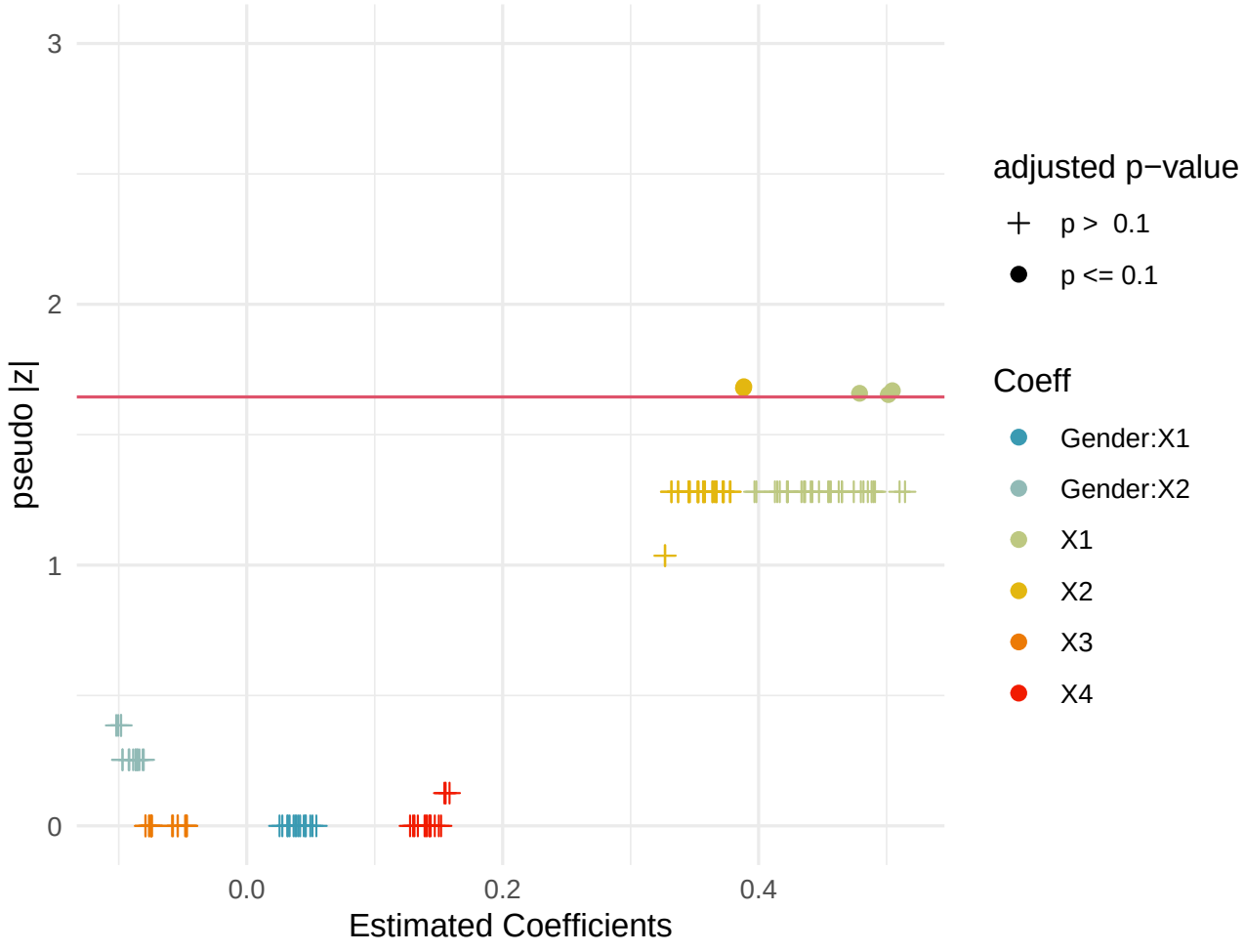


Global Null: p-value=0.089992 $\longrightarrow$ all null effects!

Conclusion

TakeHome Message

Assuming significance level 10% (instead of 5%)



TakeHome Message

Accounting for Selective Inference (i.e. Multiple Testing, adjusted p-values) is crucial

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

! Take the Global (i.e. max T) p-value: 0.089992
Yes, there is an overall effect (= at least one model)

? Which models are significant?

! There are 4 possible models:
Choose the model/story you like most!!

What is allowed and what is not

PIMA allows:

- any transformation of variables (predictors, responses)
- any GLM
- any outlier 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!

Sign flip score test

`github.com/livioivil/flipscores` and CRAN

- control of the type I error even for small sample size
- GLMs and any other model with score test
- robust to some model misspecifications

PIMA

`github.com/livioivil/jointest`

- inference framework for multiverse analysis
- model picking with adjusted p-values