Post-selection Inference in Multiverse Analysis (PIMA):

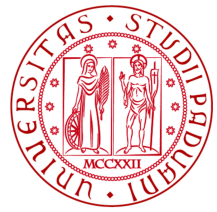
an inferential framework based on the sign flipping score test

Livio Finos

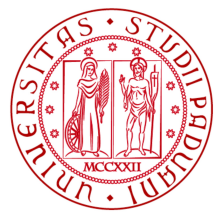
Joint work with P. Girardi, A. Vesely

G. Altoè, A. Calcagnì, M. Pastore, D. Lakens

Psicostat meeting – 24 May, 2024



Outline



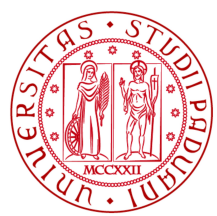
Toy example

Response variable Y: quantitative variable: e.g. behavioural measure, opinion (scale) etc

Predictors

- A few demographic confounders (e.g. *Gender*, *Age*, etc)
- 4 potential covariates/mediators C_1, C_2, C_3, C_4: it is interesting to see which subset of them is more relevant
- MAIN INTEREST: 4 possibly related traits (scales?). X_1, X_2, X_3, X_4 tested separately
- Interaction of X_1 and X_2 with *Gender* is plausible by literature review

Q: Is Y explained by X_1, X_2, X_3, or X_4 after accounting for (a subset of) C_1, C_2, C_3, C_4 and the other confounders?



Many possible Multiple Linear Models

- Should I use X_1, X_2, X_3, or X_4 in my model? (4 options)
- If X_1 or X_2, should I add the interaction with *Gender*? (+2 more options)
- Which subset of covariates C_1, C_2, C_3, C_4 should I use? (2^4 subsets)

E.g.

`Y ~ X2 + X2:Gender + C1 + C3 + C4 + Gender + Other`

or

`Y ~ X4 + C2 + C4 + Gender + Other Confounders`

We easily get lost in the forest of $(4 + 2) * 2^4 = 96$ models!

Furthermore, in some model we test for X_1 (or X_2) and $X_1 : Gender$ (or $X_2 : Gender$); **there are** 128 **tests altogether!**

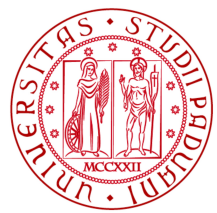
p-hacking and replicability crisis

p-hacking (i.e. Data snooping or Data dredging):
performing many statistical tests on the data and only reporting those that come back with 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, Management, etc

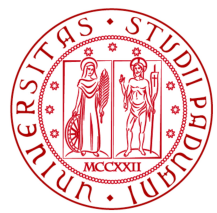
One for all: Ioannidis (2005) Why Most Published Research Findings Are False, Plos Medicine, 13,000 citations today



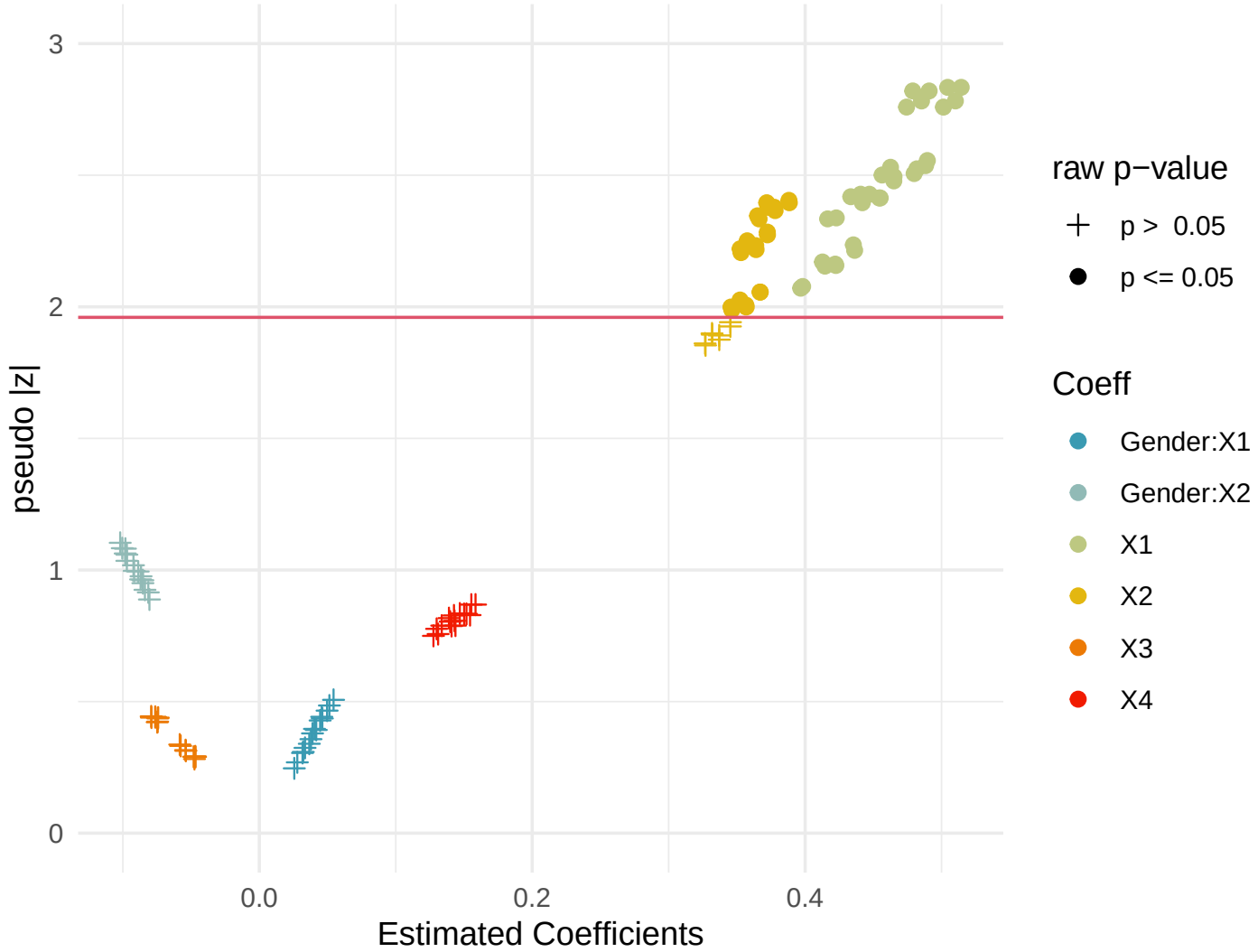
The Multiverse analysis solves the problem!

- Philosophy of statistical reporting the outcomes of many different statistical analyses showing how robust findings are (Dragicevic et al., 2019)
- Multiverse Analysis displays robustness of a finding across different options for all steps in data processing (Steegen et al., 2016).

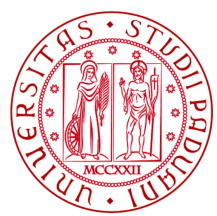
Multiverse made simple: don't hide what you've tried, report all the p-values and discuss them...



Summary of the results



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



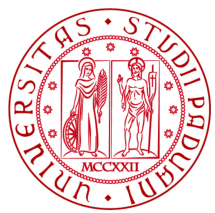
The Multiverse analysis solves the problem! Really?

Ok, I'et's go Multiverse!
I've got 43% coefficients with $p \leq 0.05$ (i.e. 58 over 128)
Quite a strong evidence to support our hypothesis! Isn't it?

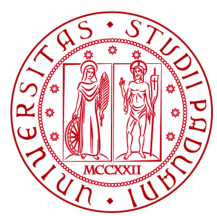
NO! We don't get any inferential clue from it.

Multiverse is important to make data analysis transparent,
BUT **a formal inferential approach is NEEDED**

p-hacking is an informal **Selective Inference** problem. Make it formal and get p-values that accounts for this multiplicity!



Outline



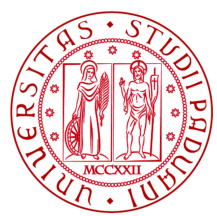
Valid p-hacking via sign-flip score test!

There is a lack of a general and valid inferential framework for multiverse analysis

The specification curve (Simonsohn et all, 2020, Nature Human Behaviour) is the only inferential method but is limited to standard linear models

Girardi et al. (2024) Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test, Psychometrika, 1-27

It uses a multivariate extension of the sign-flip score test (Hemerik, Goeman & Finos, 2020)



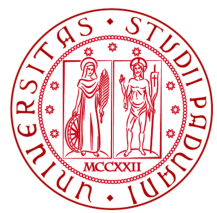
Valid p-hacking via sign-flip score test!

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

! **Ensemble Inference**: combining the info from all models in a single p-value

? **Which models** are significant?

! **model-picking**: choose the model you better like while accounting for Selective Inference![1]



[1] FamilyWise Error Rate control

The models, the tested hypotheses

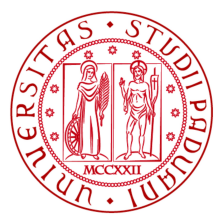
K models, for each model a General Linear Model (GLM):

$$g_k(E(y_{ki})) = \gamma_{k0} + \gamma_{k1} z_{ki} + \beta_k x_{ki}, \ i = 1, \ldots, n$$

- $\forall k = 1, \ldots, K$ models:
  y_{ki}, z_{ki}, x_{ki}: transformed y_i, z_i, x_i,
  g_k: link function for model k
- nuisances: γ_{k0}, γ_{k1}
- tested: $H_{0k} : \beta_k = 0$

We want to test:

$$H_0 : \cap_{k=1}^{K} H_{0k} : \ \beta_k = 0 \ \forall k = 1, \ldots, K$$



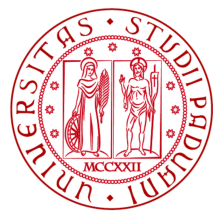
Sign Flip Score Test[2] (univariate)

In a nutshell:
(n independent observations with density $f_{\beta,\gamma,x_i,z_i}(y_i)$)

- Score test:

$$T = \sum_{i=1}^{n} \nu_i = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log f_{\beta,\gamma,x_i,z_i}(y_i) \mid_{\hat{\gamma},\beta=0}$$

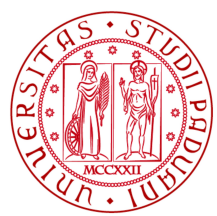
- $T^{*b} = \sum_{i=1}^{n} \pm \nu_i$
- Under H_0:
  . $E(T) = E(T^{*b}) = 0$,
  . $T^{*b} \stackrel{d}{=} T$, asymptotically normal (CLT)



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

Joint Sign Flip Scores Test

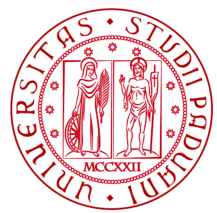
- Instead of 'only' one model, in multiverse we have K of them, i.e. K score statistic $(T_1, \ldots, T_K)'$
- k-variate score contributions: $(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})', \; i = 1, \ldots, n$
- jointly flip the sign of all K contributions: $\pm(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})$
- $T_k^{*b} = \sum_i \pm\nu_{ik}, \; k = 1, \ldots, K$
- under H_0 : $(T_1^{*b}, \ldots, T_K^{*b}) \stackrel{d}{=} (T_1, \ldots, T_K)$ – jointly, approximated
- combine the K test stats in a single test, e.g. $\max_k T_k$
- Multiverse p-value: $\#_b(\max_k T_k^{*b} \geq \max_k T_k)/(B+1)$



Joint Sign Flip Scores Test

n scores' contribution (observations),
K tests (models) for $H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_K = 0$

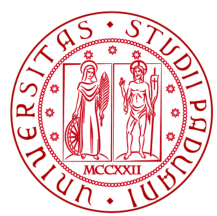
$+\nu_{11}$	$+\nu_{12}$	$\ldots$	$+\nu_{1K}$	
$+\nu_{21}$	$+\nu_{22}$	$\ldots$	$+\nu_{2K}$	
$\ldots$	$\ldots$	$\ldots$	$\ldots$	
$+\nu_{n1}$	$+\nu_{n2}$	$\ldots$	$+\nu_{nK}$	Combined
$\mathbf{S}_1^O$	$\mathbf{S}_2^O$	$\ldots$	$\mathbf{S}_K^O$	$max_k \mathbf{S}_k^O$



Joint Sign Flip Scores Test

n scores' contribution (observations),
K tests (models) for $H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_K = 0$

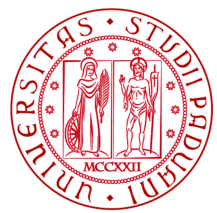
$-\nu_{11}$	$-\nu_{12}$	$\ldots$	$-\nu_{1K}$	
$+\nu_{21}$	$+\nu_{22}$	$\ldots$	$+\nu_{2K}$	
$\ldots$	$\ldots$	$\ldots$	$\ldots$	
$-\nu_{n1}$	$-\nu_{n2}$	$\ldots$	$-\nu_{nK}$	Combined
$\mathbf{S}_1^O$	$\mathbf{S}_2^O$	$\ldots$	$\mathbf{S}_K^O$	$max_j \mathbf{S}_j^O$
$\mathbf{S}_1{}^{*1}$	$\mathbf{S}_2{}^{*1}$	$\ldots$	$\mathbf{S}_K{}^{*1}$	$max_k \mathbf{S}_k{}^{*1}$



Joint Sign Flip Scores Test

n scores' contribution (observations),
K tests (models) for $H_0 : \beta_1 = 0,\ \beta_2 = 0, \ldots,\ \beta_K = 0$

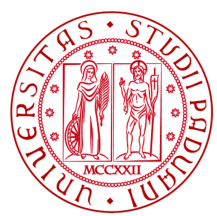
$+\nu_{11}$	$+\nu_{12}$	$\ldots$	$+\nu_{1K}$	
$-\nu_{21}$	$-\nu_{22}$	$\ldots$	$-\nu_{2K}$	
$\ldots$	$\ldots$	$\ldots$	$\ldots$	
$+\nu_{n1}$	$+\nu_{n2}$	$\ldots$	$+\nu_{nK}$	Combined
$\mathbf{S}_1^O$	$\mathbf{S}_2^O$	$\ldots$	$\mathbf{S}_K^O$	$max_j \mathbf{S}_j^O$
$\mathbf{S}_1{}^{*1}$	$\mathbf{S}_2{}^{*1}$	$\ldots$	$\mathbf{S}_K{}^{*1}$	$max_k \mathbf{S}_k{}^{*1}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathbf{S}_1{}^{*B}$	$\mathbf{S}_2{}^{*B}$	$\ldots$	$\mathbf{S}_K{}^{*B}$	$max_k \mathbf{S}_k{}^{*B}$



Joint Sign Flip Scores Test

n scores' contribution (observations),
K tests (models) for $H_0 : \beta_1 = 0,\ \beta_2 = 0, \ldots,\ \beta_K = 0$

$+\nu_{11}$	$+\nu_{12}$	$\ldots$	$+\nu_{1K}$	
$-\nu_{21}$	$-\nu_{22}$	$\ldots$	$-\nu_{2K}$	
$\ldots$	$\ldots$	$\ldots$	$\ldots$	
$+\nu_{n1}$	$+\nu_{n2}$	$\ldots$	$+\nu_{nK}$	Combined
$\mathbf{S}_1^O$	$\mathbf{S}_2^O$	$\ldots$	$\mathbf{S}_K^O$	$max_k \mathbf{S}_k^O$
$\mathbf{S}_1{}^{*1}$	$\mathbf{S}_2{}^{*1}$	$\ldots$	$\mathbf{S}_K{}^{*1}$	$max_k \mathbf{S}_k{}^{*1}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathbf{S}_1{}^{*B}$	$\mathbf{S}_2{}^{*B}$	$\ldots$	$\mathbf{S}_K{}^{*B}$	$max_k \mathbf{S}_k{}^{*B}$
p-values	$\frac{\#_b(S_k^{*b} \geq S_k^O)}{B+1}$			$\frac{\#_b(\max_k S_k^{*b} \geq \max_k S_k^O)}{B+1}$



Joint Sign Flip Scores in a drop

The estimate of a coefficient β in (G)LM can be written as the sum of n contributions: $T = \sum_{i=1}^{n} \nu_i$

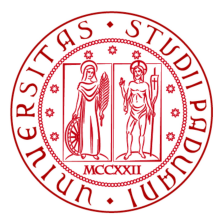
Each contribution ν_i has mean 0 when $H_0 : \beta = 0$

We can flip the signs of ν_is hence creating new pseudo-scores $T = \sum_{i=1}^{n} \pm\nu_i$ (under H_0)

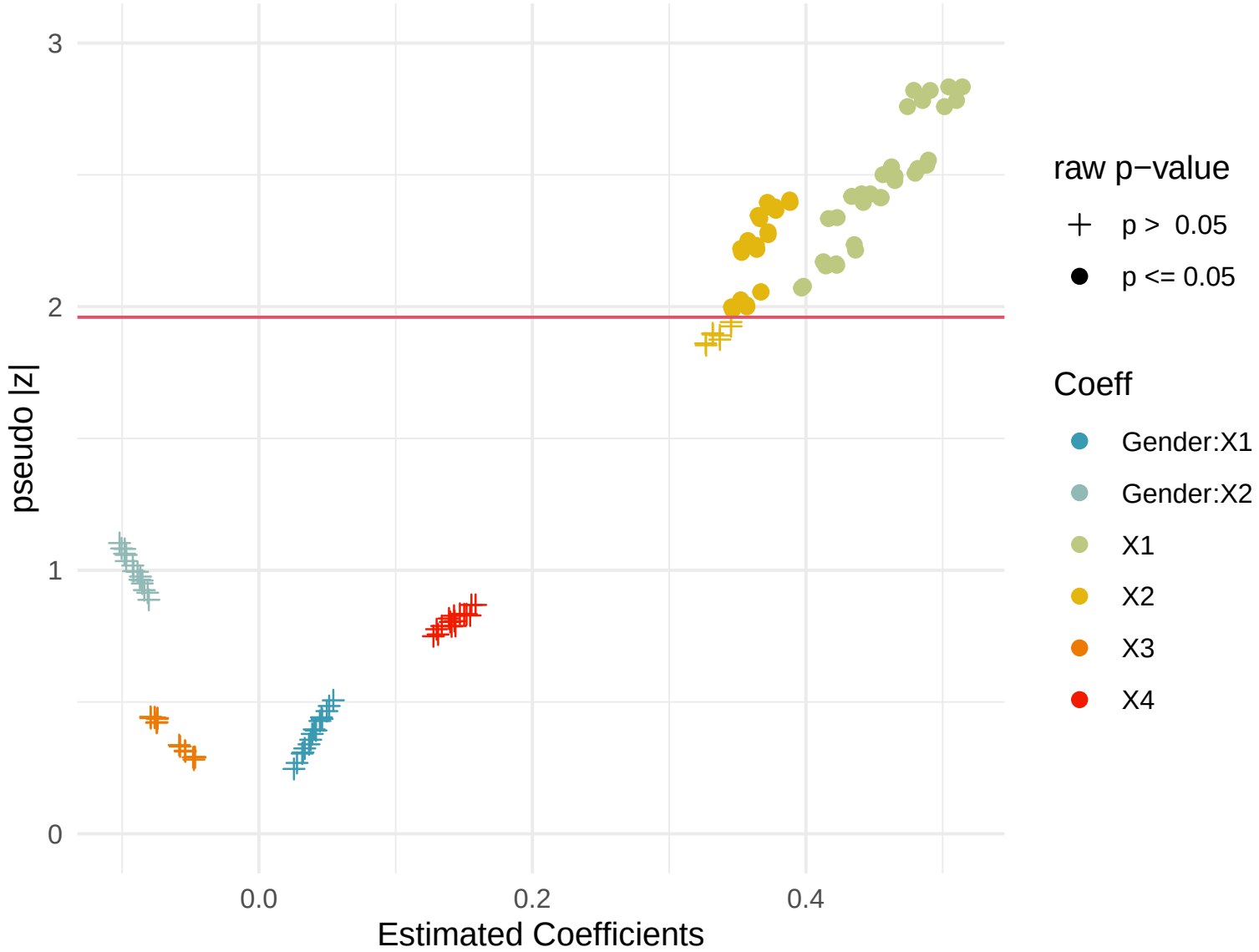
Properties

- you can use it whenever you can write a score test (i.e. lm, glm and much more)
- asymptotically exact (exact, in practice)
- very robust to variance misspecification
  (OK if link function is OK)
- the resampling approach easily accounts for the (very strong) dependence among tests, i.e. powerful approach.

Outline

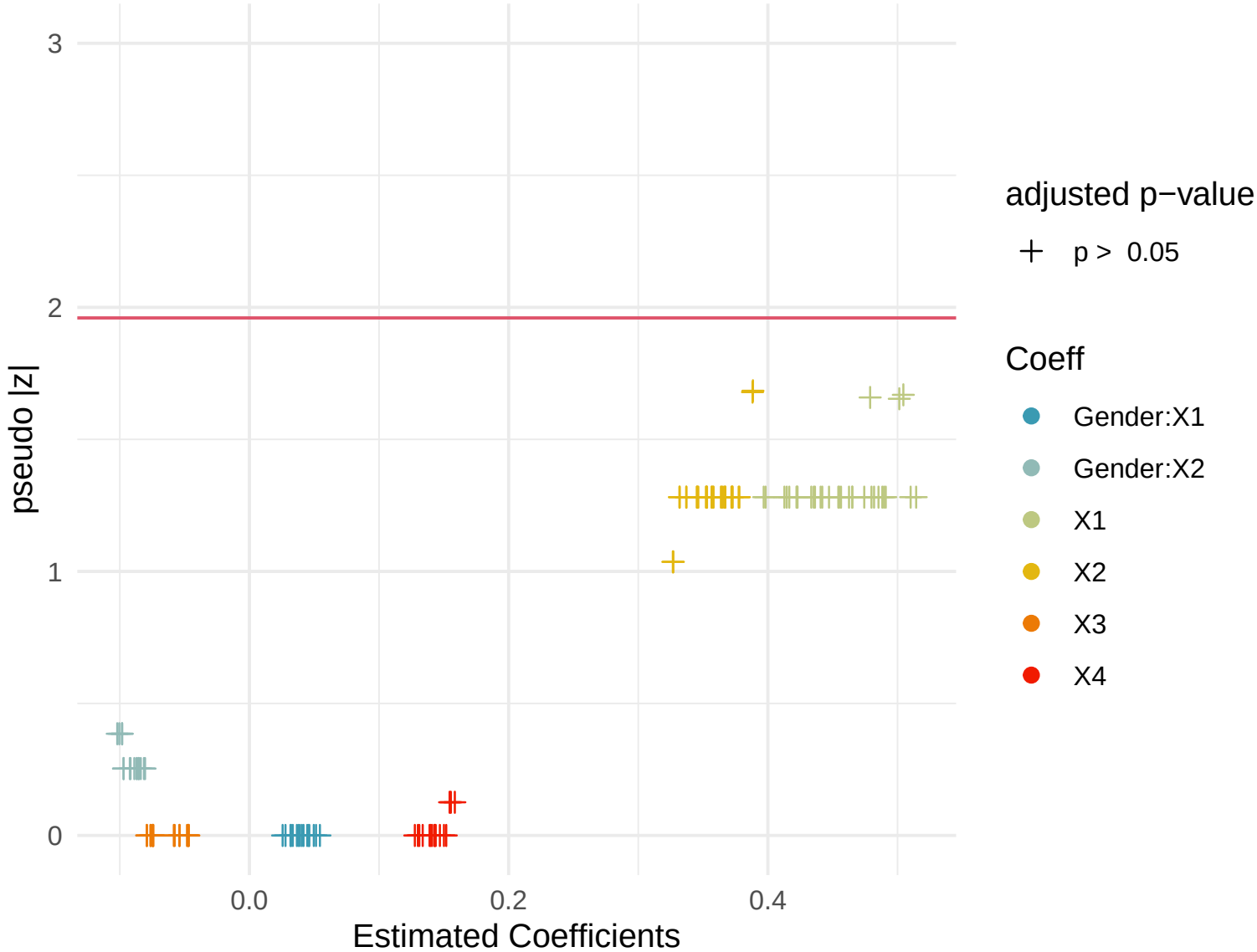


Raw (unadjusted) p-values

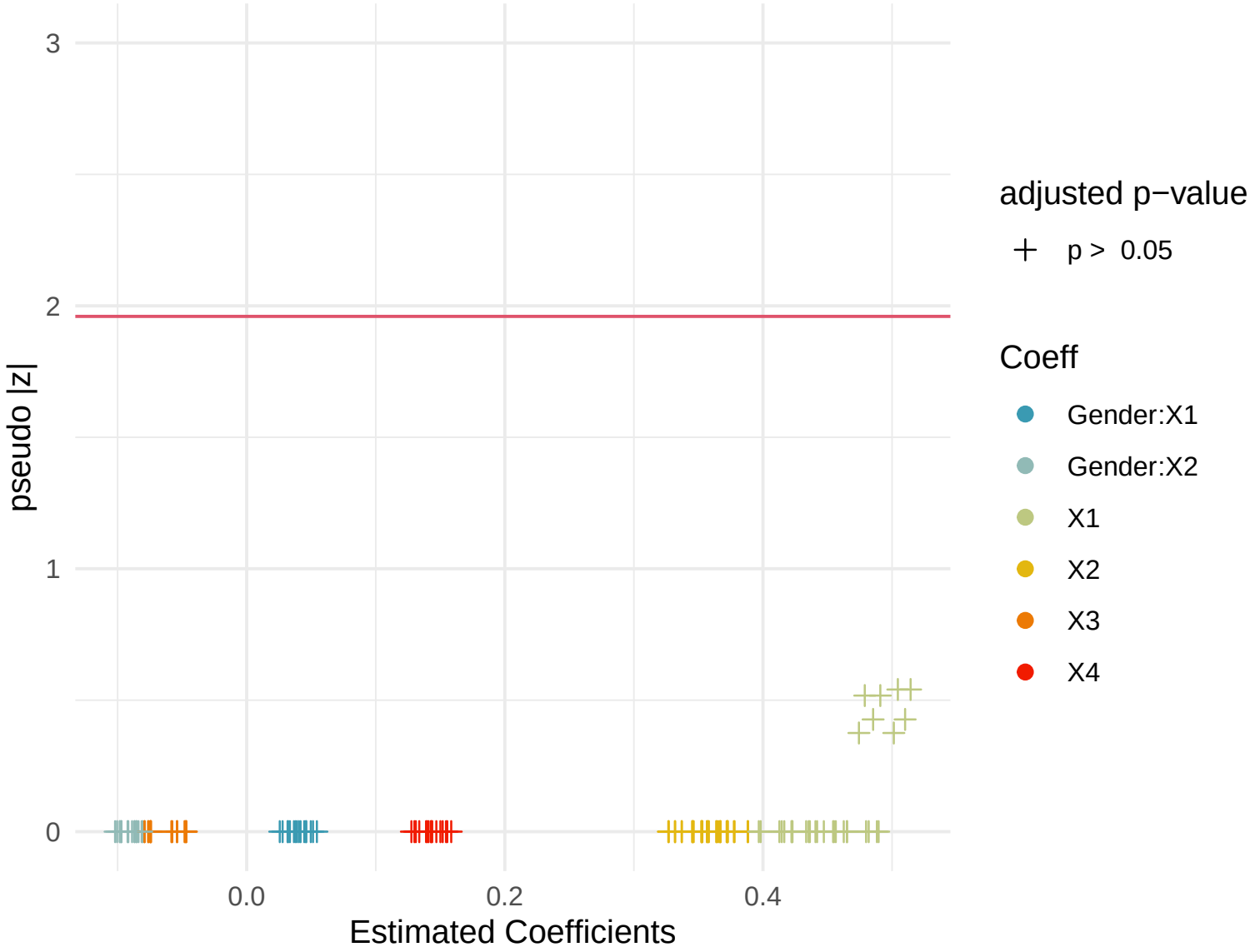


Which coefficients are non-null? Adjusted p-values

NONE OF THEM!

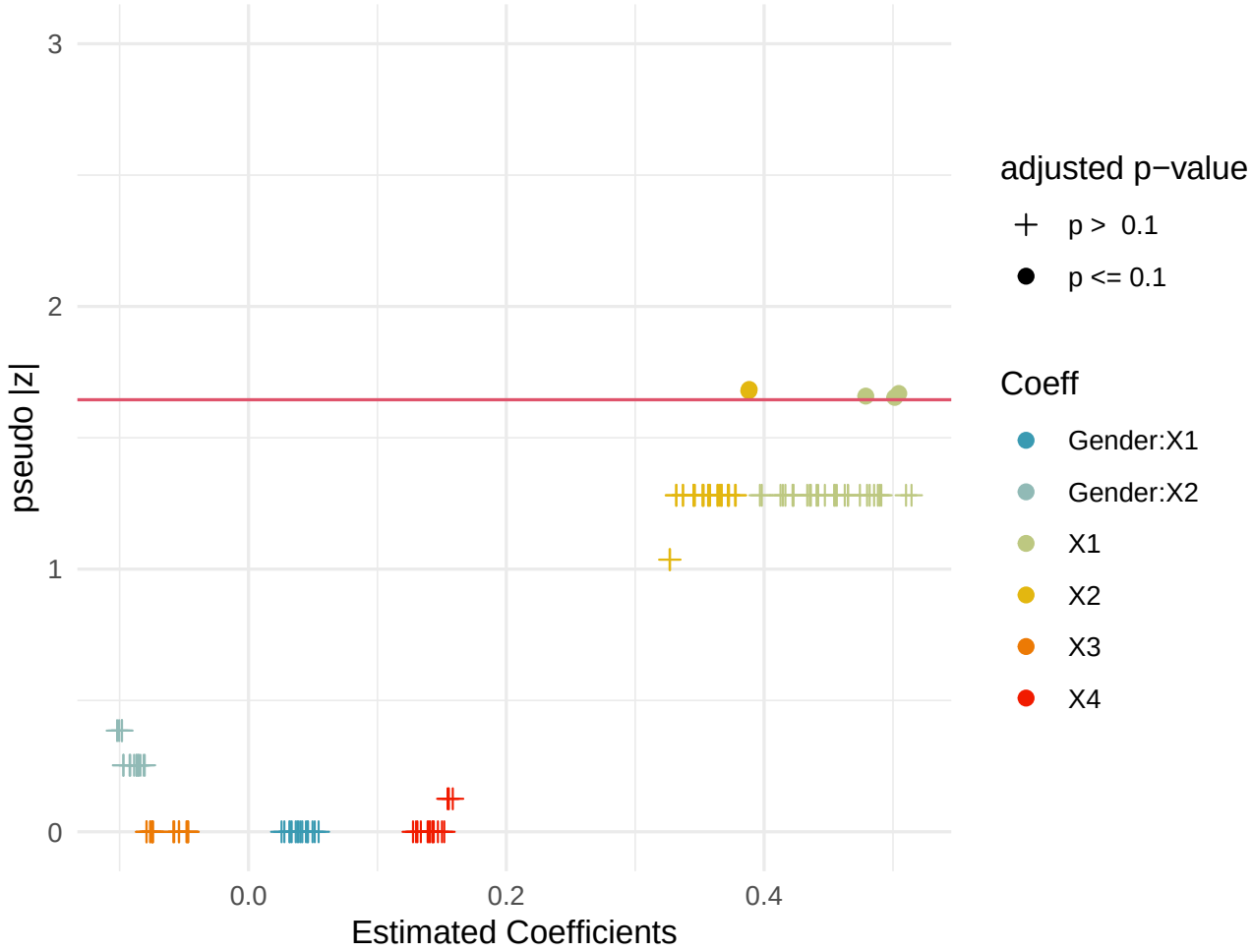


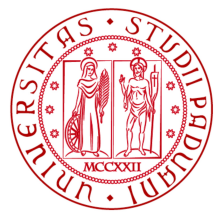
Bonferroni-Holm Adjusted p-values (FWER)



TakeHome Message 1:
Pick the model, choose the story to tell :)

Assuming significance level 10% (instead of 5%)



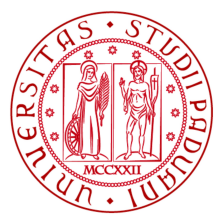


TakeHome Message 1:
Pick the model, choose the story to tell :)

4 selected models, all plausible:

- `Y ~ X1 + C1 + Gender + Other Confounders`
- `Y ~ X1 + C4 + Gender + Other Confounders`
- `Y ~ X2 + X2:Gender + C1 + C3 + C4 + Gender + Other Confounders`
- `Y ~ X2 + X2:Gender + C1 + C4 + Gender + Other Confounders`

Which one do you like most?



TakeHome Message 2: Multiverse is a slippery floor

Multiverse does not solve the problem of validity of the assumptions: If the model is wrong a significant p-value does not mean anything!

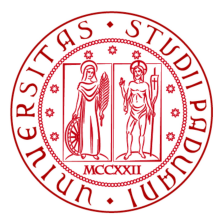
E.g. If the true model is

`Y ~ X2 + X2:Gender + C1 + C3 + C4 + Gender + Other`

the model without interaction term `X2:Gender` is wrong!

(Residuals are not independent, not normal etc, the test on `X2` may fail to control the false positive)

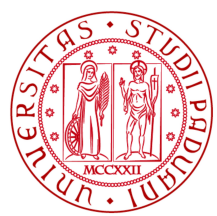
Think before testing! (Altoè, 2001)



What is allowed and what is not

PIMA approach allows:

- Any variable transformation (predictors, responses)
- Any GLM model (e.g. log-normal, Poisson, negative Binomial)
- Any outlier deletion methods



What is allowed and what is not

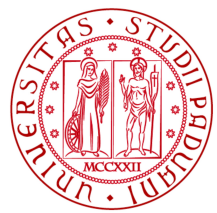
PIMA approach allows:

- Any variable transformation (predictors, responses)
- Any GLM model (e.g. log-normal, Poisson, negative Binomial)
- Any outlier deletion methods

BUT all the above models

- MUST be planned IN ADVANCE
- MUST be valid
  (at least the right link, variance is not a problem)

There is no free lunch



Take Home message

Testing coefficients in a GLM:
`flipscores`: github.com/livioivil/flipscores (and CRAN)

- Control of the Type I Error
  Sims: good control even for tiny sample size (e.g. n=20)
- (not only LM) GLM and any other model with score test
- Robust to model miss-specification (i.e. heteroscedasticity)

PIMA approach (i.e. combine the test of `flipscores`):
`jointest`: github.com/livioivil/jointest

- Ensemble Inference (and Post-hoc) made easy
- Model picking (with adjusted significant p-value)

Enjoy p-hacking, it is now valid!