PIMA: Post-selection Inference in Multiverse Analysis

4M international Winter School, 20 Feb, 2025

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

University of Padova
livio.finos@unipd.it

The PIMA dream team

- F. Gambarota,
- G. Calignano,
- P. Girardi,
- A. Vesely,
- D. Lakens,
- M. Pastore,
- G. Altoè,
- L. Finos

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'

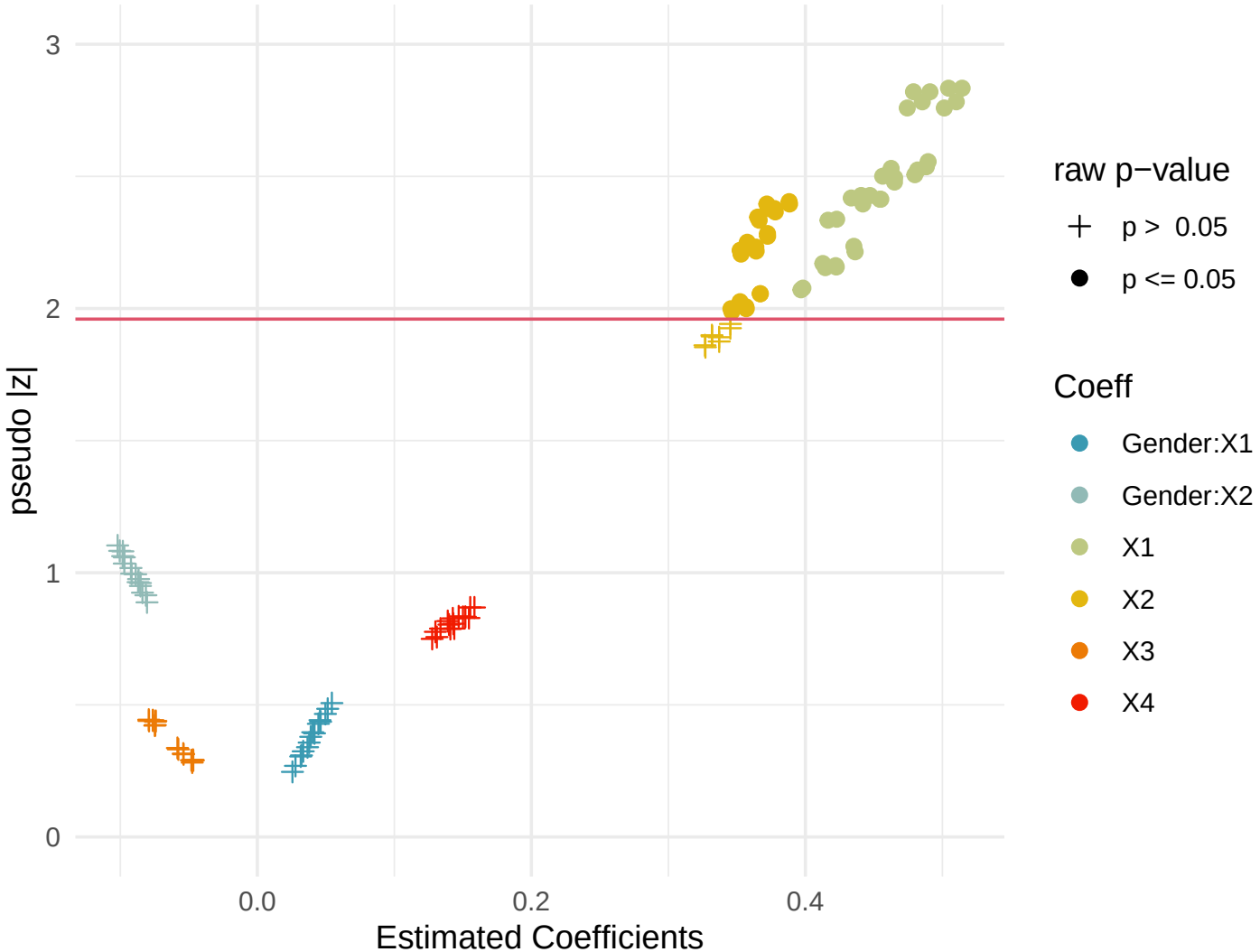
A philosophy of reporting the outcomes of many different analyses to explore:

- 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!

Inference in Multiverse Analysis (IMA)

Family-wise error rate (FWER)

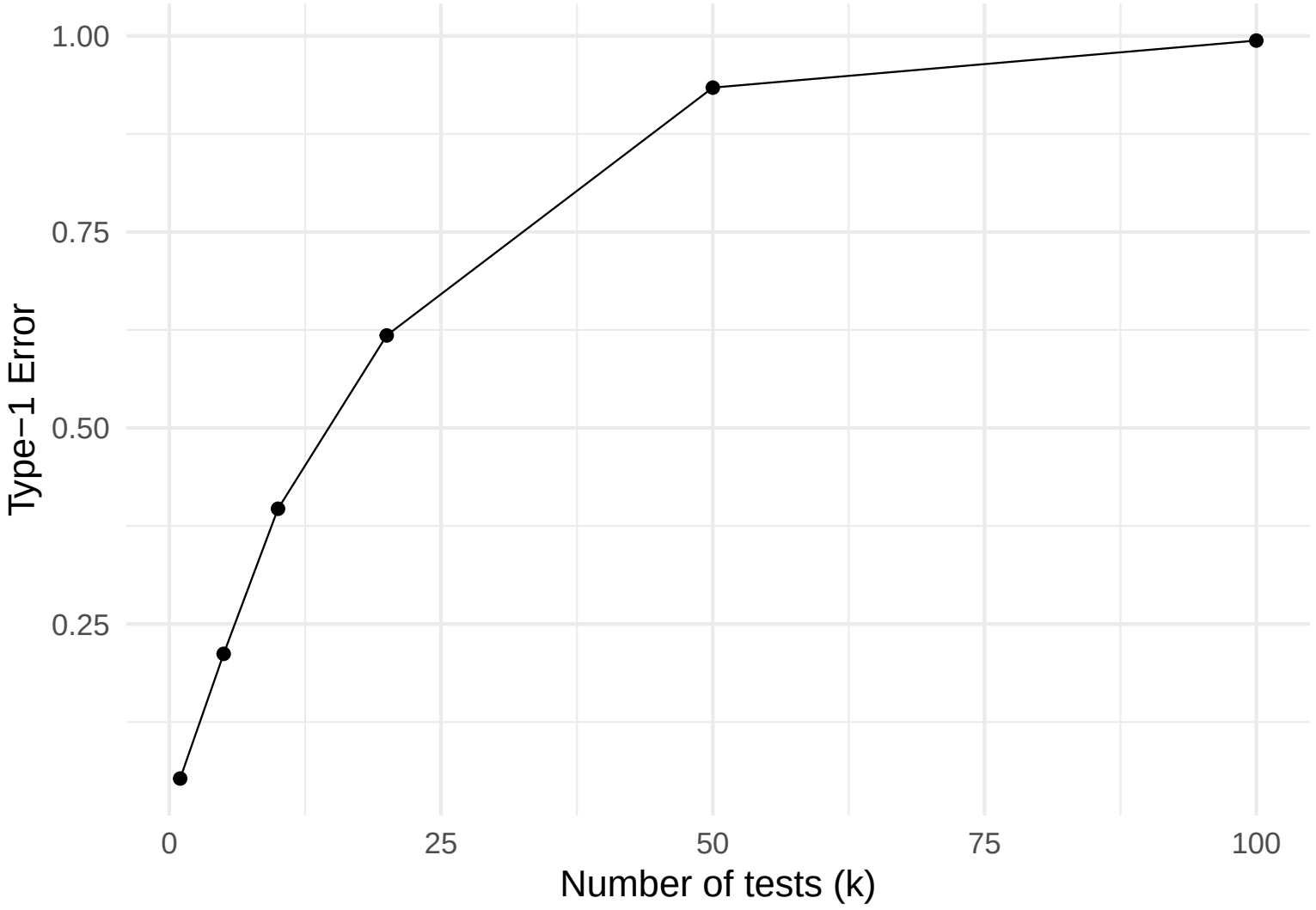
		H_0		
		False	**True**	**Tot**
Test	Rejected	**True Positive (S)**	**False Positive (V)**	R
	Not rejected	**False Negative (T)**	**True Negative (U)**	$m - R$
	Tot	m_1	m_0	m

The FWER is the probability of committing AT LEAST ONE type-1 error (i.e. false positive) thus $\Pr(V > 0)$. Controlling the FWER (whatever the methods) keep $\Pr(V > 0) \leq \alpha$.

There are different procedures for controlling the FWER, such as the Bonferroni or the Holm–Bonferroni method.

Why multiple testing issue?

Probability of at least one type 1 error as a function of number of (independent) tests



Power is nothing without control: Adjusting the p-values

The main problem is that the number of tests in a multiverse can be quite large.

As an example, we simulated a series of tests with different effect size to show the impact on the type-1 error rate and the power.

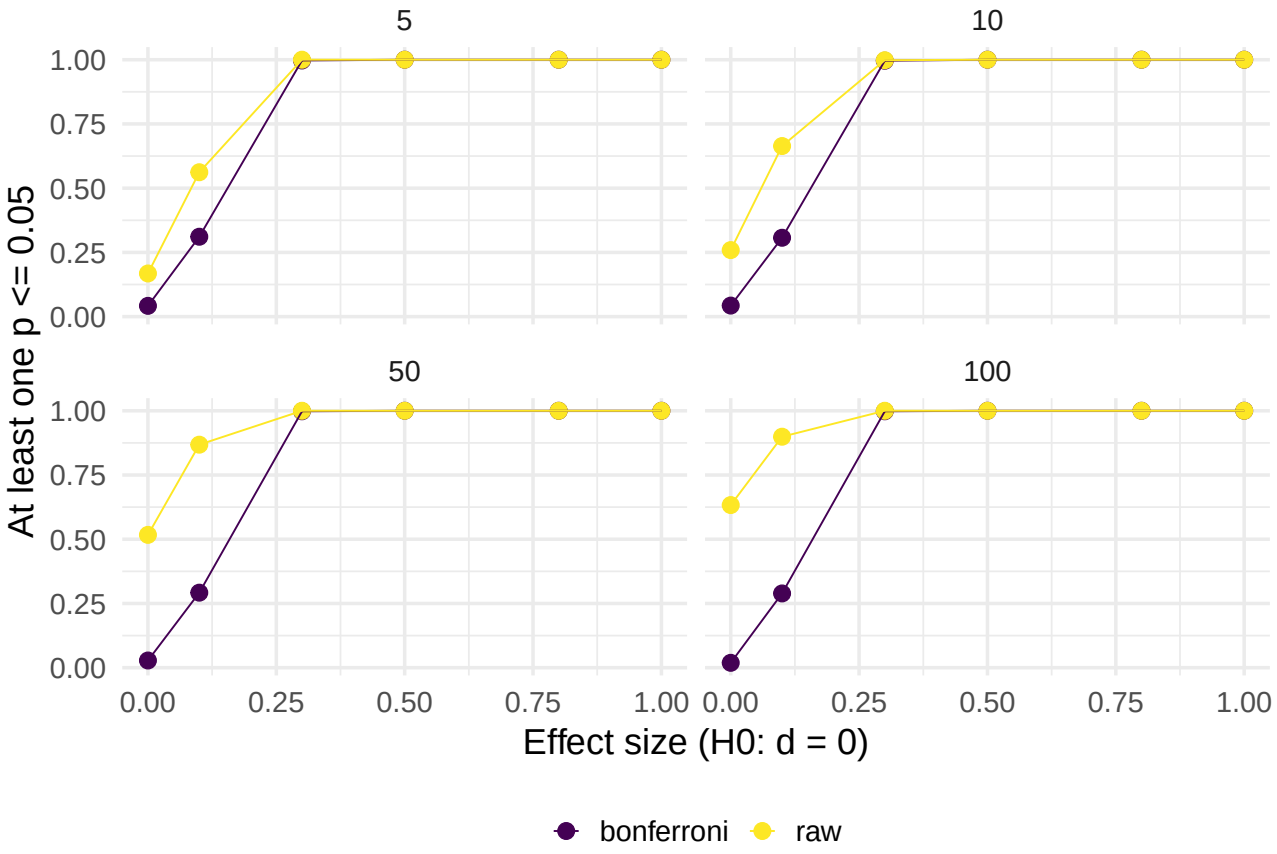
Power is nothing without control: Adjusting the p-values

The main problem is that the number of tests in a multiverse can be quite large.

As an example, we simulated a series of tests with different effect size to show the impact on the type-1 error rate and the power.

Adjusting the p-values

Using a standard method (e.g., Bonferroni multiplies each p-value by the number of tested hypos) clearly controls the type-1 error but reduces a lot the statistical power. At the same time, without correction the inflation is large.



Correlation between scenarios is (probably) large

The multiverse scenarios are computed on the same dataset thus the correlation between tests is probably medium-large. For example:

```
x <- runif(100, 5, 10)
y <- x * 0.1 + rnorm(100)

fit1 <- lm(y ~ x)
fit2 <- lm(y ~ cut(x, breaks = 2))
fit3 <- lm(y ~ log(x))
fit4 <- lm(y ~ poly(x, 2))

pp <- sapply(list(fit1, fit2, fit3, fit4), predict)
round(cor(pp), 2)
```

```
     [,1] [,2] [,3] [,4]
[1,] 1.00 0.88 1.00 0.99
[2,] 0.88 1.00 0.88 0.88
[3,] 1.00 0.88 1.00 1.00
[4,] 0.99 0.88 1.00 1.00
```

A more powerful correction method [1]

The Bonferroni and Holm methods are robust to any dependence structure, but the price it a reduced power.

The permutation-based methods (maxT, minP, etc.) take into account the correlation structure providing FWER control under H_0 but a more powerful test under H_1.

[1] maxT procedure Westfall & Stanley Young (1993)

Permutation testing in a nutshell

```
B <- 1e3 # number of permutations
tp <- matrix(NA, B,2)
tp[1,1] <- t.test(y ~ x1)$statistic
tp[1,2] <- t.test(y ~ x2)$statistic # first permutation always the observed dat

id=sample(30) # shuffling the group label
x1[id]
```

```
 [1] 1 1 1 1 1 0 1 1 1 1 0 0 0 0 0 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1
```

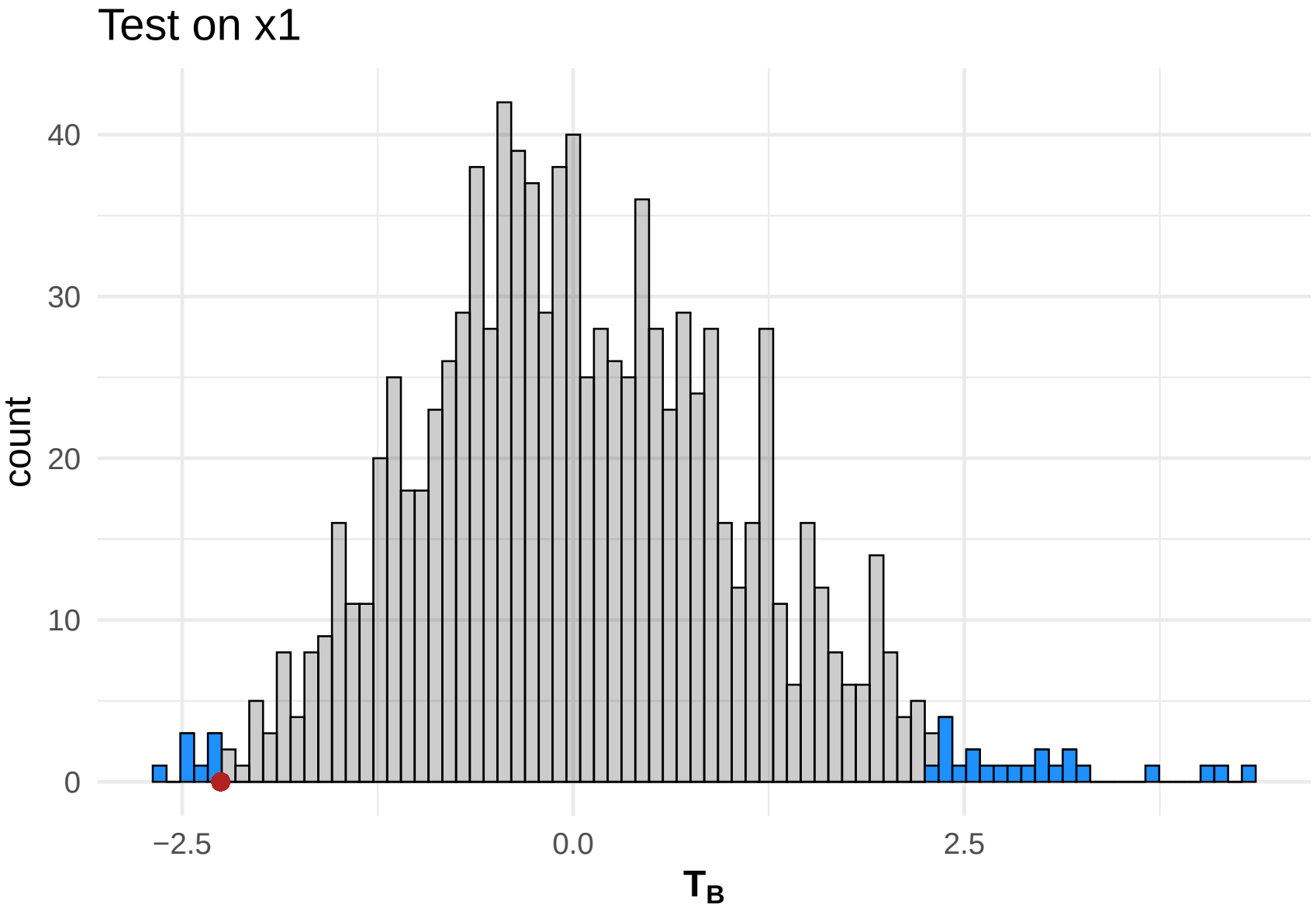
```
for(i in 2:B){
    id <- sample(30)
    tp[i,1] <- unname(t.test(y ~ x1[id])$statistic)
    tp[i,2] <- unname(t.test(y ~ x2[id])$statistic)
}

mean(abs(tp[,1]) >= abs(tp[1,1]))
```

```
[1] 0.03
```

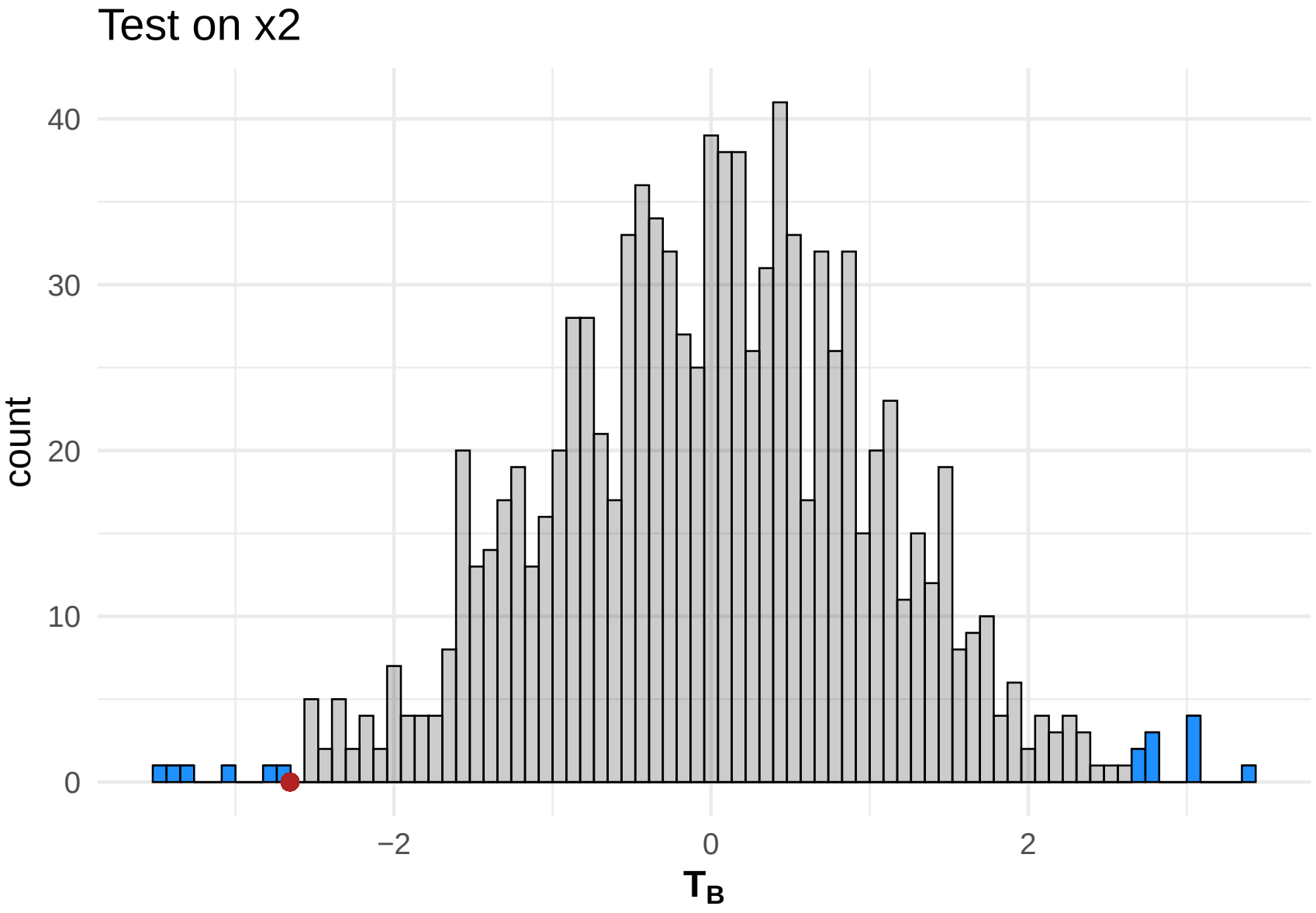
```
mean(abs(tp[,2]) >= abs(tp[1,2]))
```

```
[1] 0.016
```

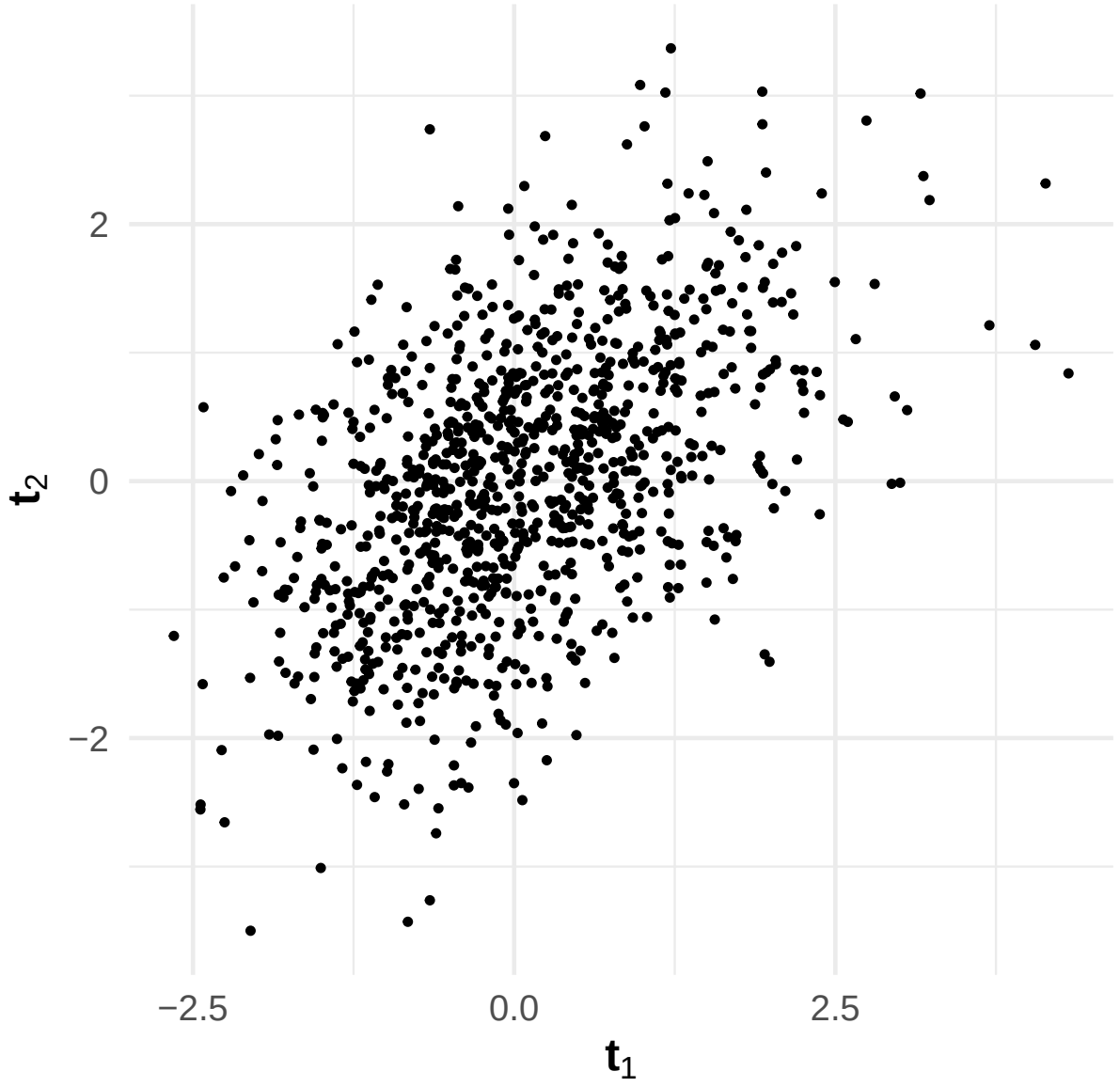
Permutation testing in a nutshell



Permutation testing in a nutshell



Permutation testing in a nutshell



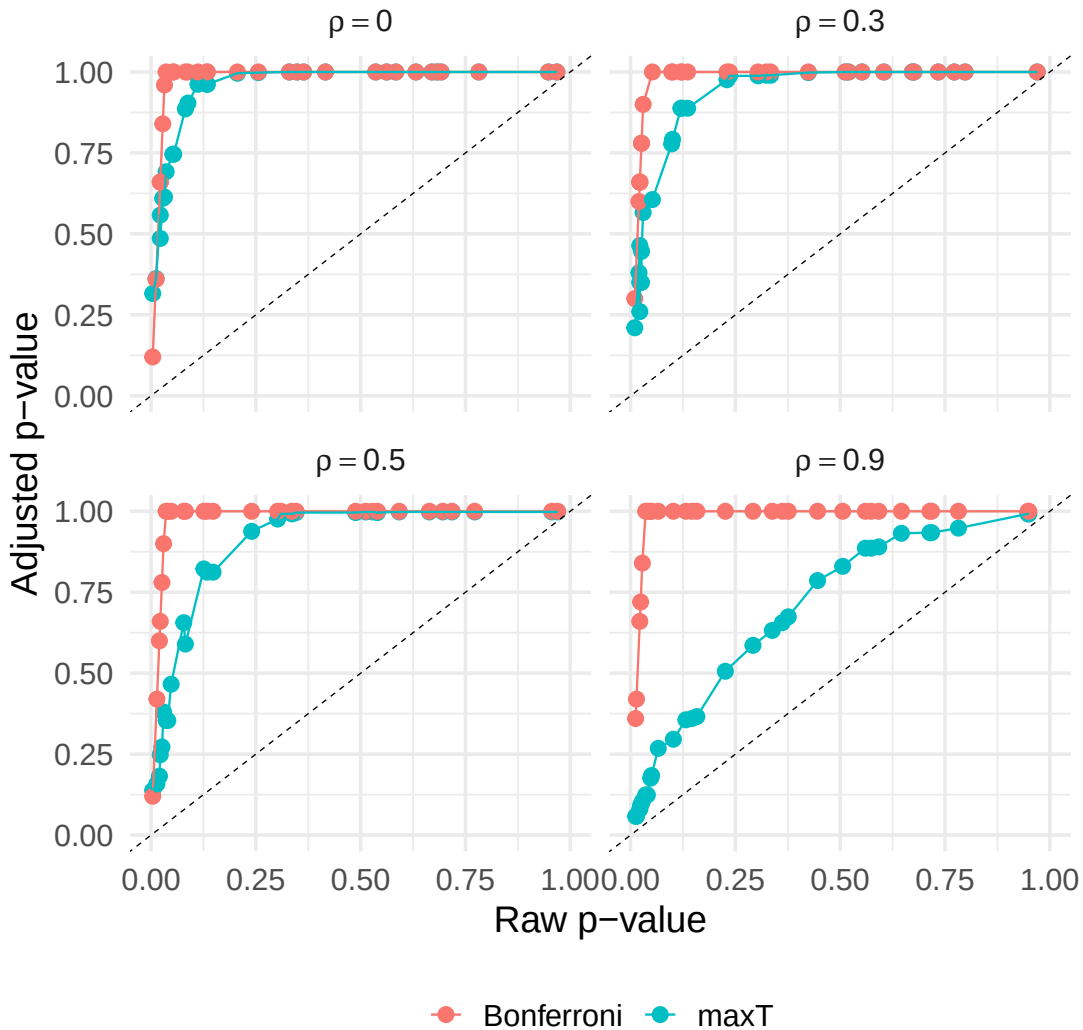
maxT procedure Westfall & Stanley Young (1993)

The maxT is a permutation-based method to control the FWER. With the method we can obtain:

- overall inference across M tests with *weak* control of FWER
- individually adjusted p-values for each test (i.e, *strong* FWER control)

maxT with correlated variables

Beyond the actual method and algorithm, the advantage of the maxT approach is taking into account the correlation between tests.



Inferential Methods

- Specification Curve (Simonsohn et al., 2020)
- Post-Selection Inference in Multiverse Analysis (PIMA; Girardi et al., 2024)

Specification Curve

The specification curve (Simonsohn et al., 2020) is the first attempt to build an inferential framework for multiverse analysis.

- provides only *weak* control of type-1 error
- is not directly applicable to GLMs (only standard linear models, see Girardi et al., 2024)
- is computationally expensive

Post-selection Inference in Multiverse Analysis (PIMA)

The PIMA recipe... (Girardi et al., 2024)

PIMA provides *weak* and *strong* type-1 error control with a powerful method based on permutations (maxT) and applicable to whatever GLM (Logistic, Poisson, etc.).

For constructing the inferential approach with PIMA we need:

- a flexible modelling framework: **Generalized Linear Models**
- a permutation-based inferential approach: **Flipscores**
- a permutation-based and powerful method for weak and strong FWER control: **maxT**

The core of PIMA, the flipscores method

- The formal part of the **flipscores** method is quite complex and beyond our scope and expertise. But a detailed description can be found in Hemerik et al. (2020) and Girardi et al. (2024).
- Essentially the **flipscores** method is an alternative way of doing inference for parameters of a GLM based on permutations.
- The idea is conceptually the same as the two-groups example, but can works for multiple regression models with covariates and interactions.

Intution of flipscores

This method can be extended to whatever GLM and to any number of predictors/confounders.

The actual permutation test is obtained flipping the sign of the scores/residuals thus obtaining the distribution under the null hypothesis of the test statistics.

Everthing is implemented into the `flipscores` package (Hemerik et al., 2020) and on CRAN https://cran.r-project.org/web/packages/flipscores/index.html and GitHub https://github.com/livioivil/flipscores .

`flipscores` package

With the `flipscores` function is very easy to fit a (generalized) linear model with permutations-based p-values.

```r
library(flipscores)
fit <- flipscores(Sepal.Length ~ Petal.Width + Species, data = iris)
summary(fit)
```

```
Call:
flipscores(formula = Sepal.Length ~ Petal.Width + Species, data = iris)

Coefficients:
                   Estimate     Score Std. Error   z value Part. Cor
(Intercept)         4.78044 160.25913   13.50622  11.86558     0.979
Petal.Width         0.91690   5.64500    1.27732   4.41941     0.365
Speciesversicolor  -0.06025  -0.26260    1.00098  -0.26234    -0.022
Speciesvirginica   -0.05009  -0.09030    0.64372  -0.14028    -0.012
                  Pr(>|z|)
(Intercept)         0.0002 ***
Petal.Width         0.0002 ***
Speciesversicolor   0.8104
Speciesvirginica    0.9016
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2313718)

    Null deviance: 102.17  on 149  degrees of freedom
```

Intuition of PIMA

The idea of PIMA is to extend the `flipscores` method to M models (where M is the number of scenarios) and perform inference at the multiverse level.

Using the maxT approach we can combine the M tests into a single test with weak control of FWER. The global null hypothesis is:

$$\mathcal{H} = \bigcap_{m=1}^{M} \mathcal{H}_m : \beta_m = 0 \text{ for all } m = 1, \ldots, M.$$

In addition, we can correct the indidual p-values with strong FWER control using the maxT method.

The `pima` package

We are implementing everything into the `pima` package that is under development. You are invited to try it, but please be patient, there could be bugs and breaking changes in the near future. Contact us for any issue!

You can explore the package here https://github.com/livioivil/pima. The package mainly depends on `jointest` that is the actual package for combining multiple (correlated) tests and correcting them.

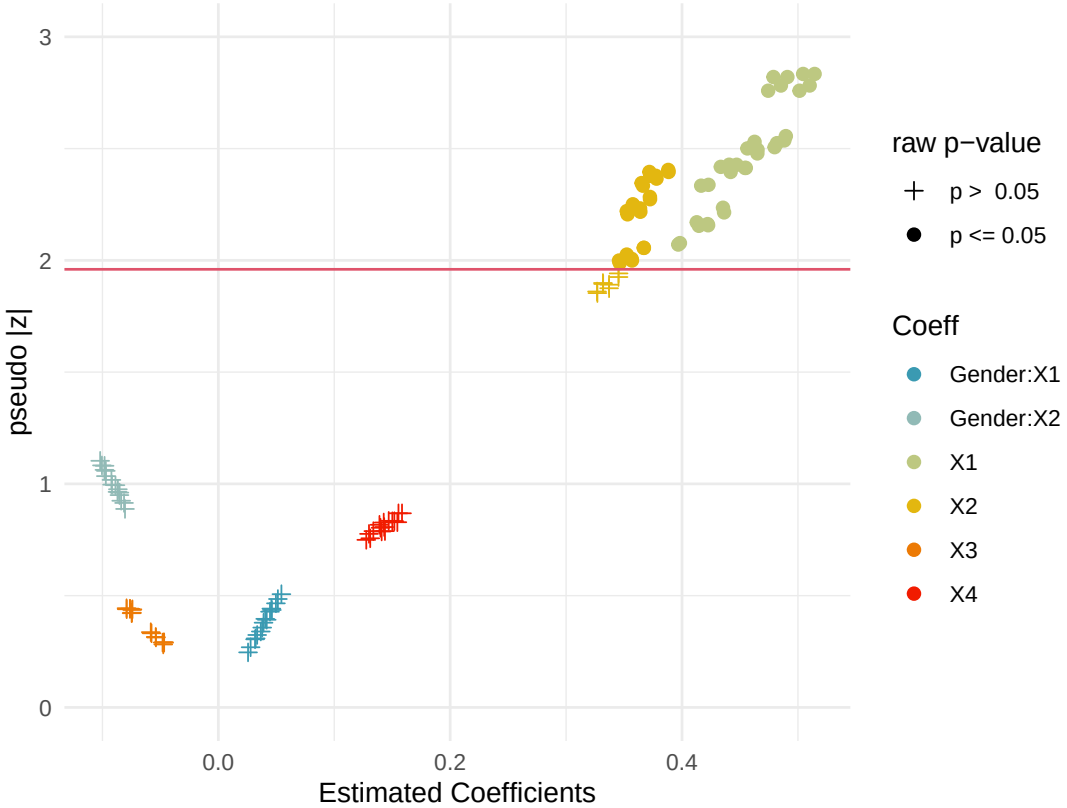
Properties

- 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.

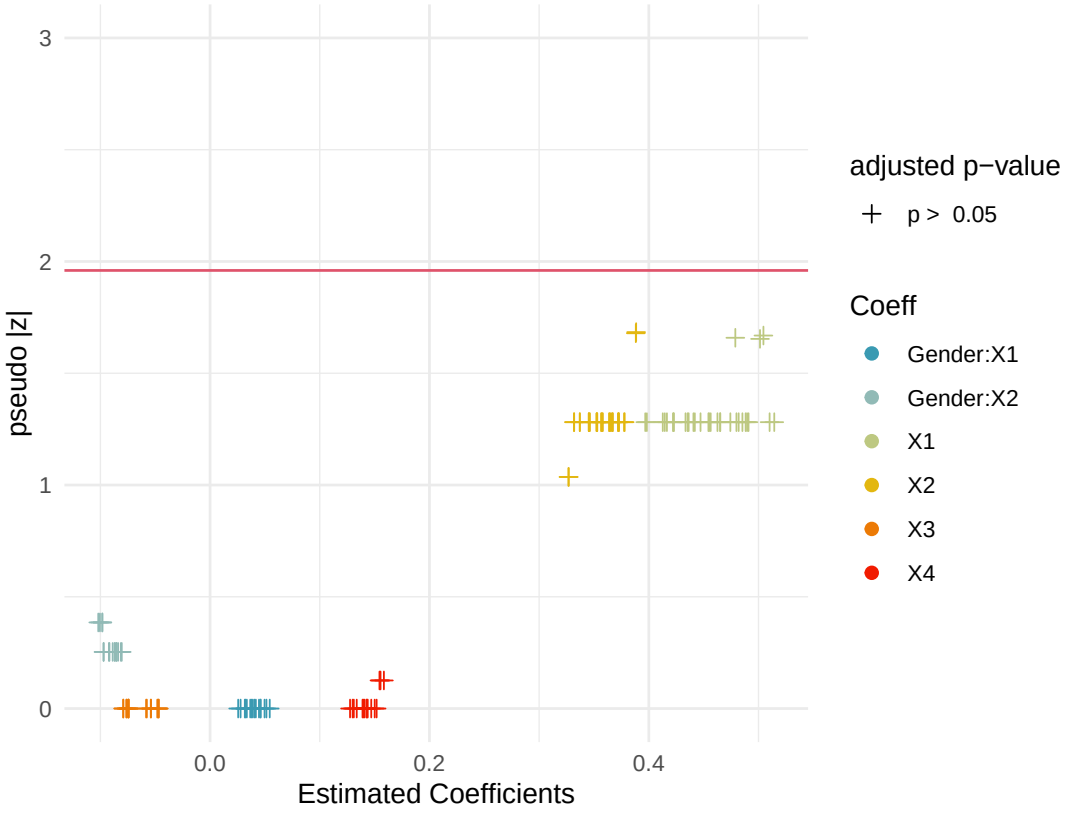
Results

Raw (unadjusted) p-values



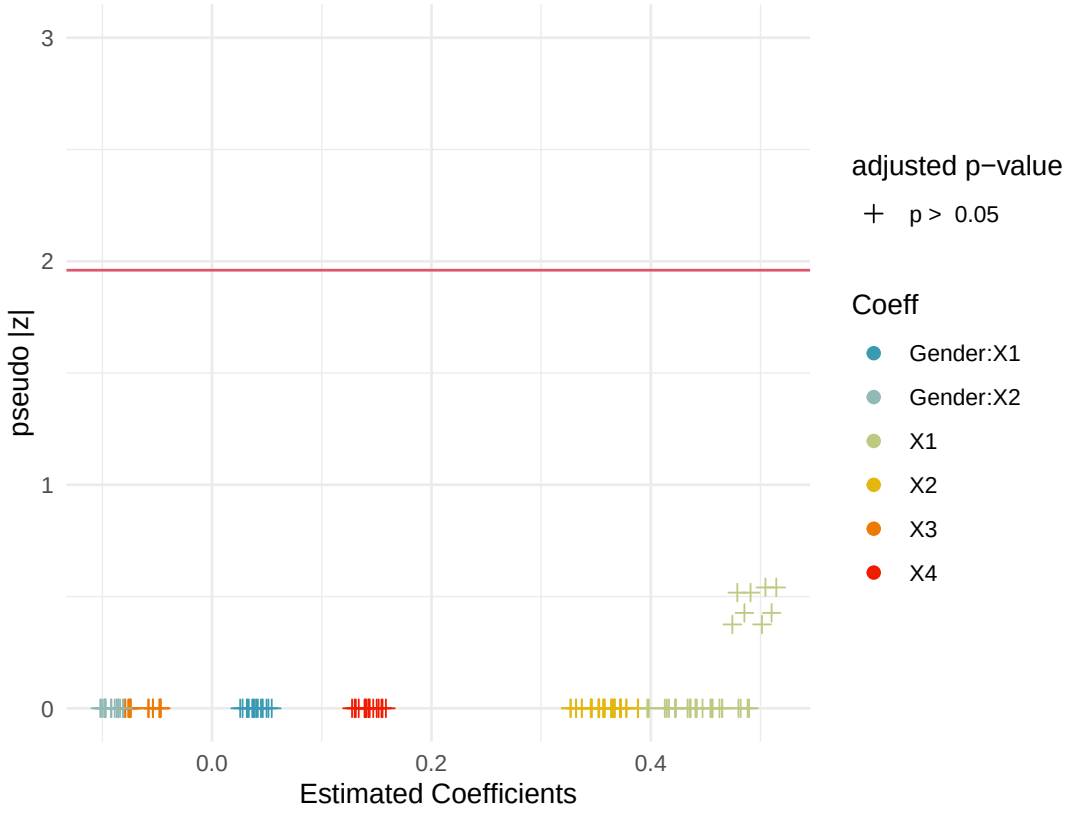
Data are generated with no effects at all, these are ALL False Positives!

Adjusted p-values, strong FWER control



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

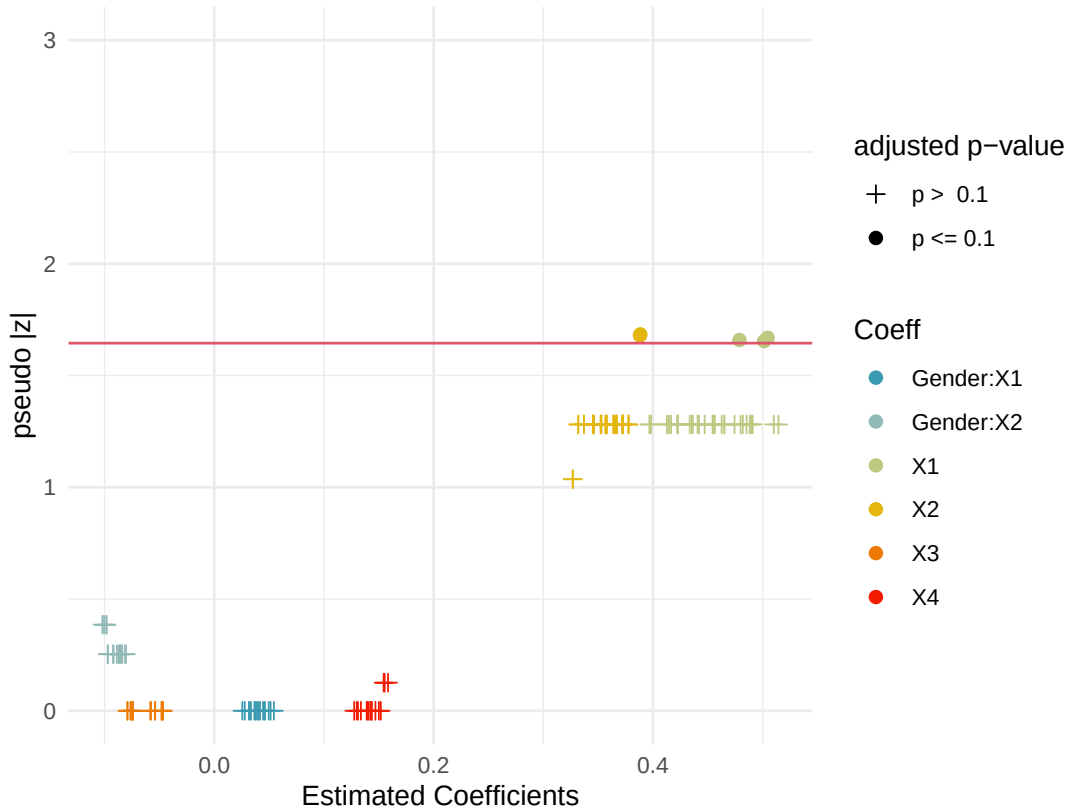
Bonferroni Adjusted p-values, strong FWER control



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

jointest

`github.com/livioivil/jointest`

- multivariate flipscores (joint distribution of test stats)

PIMA

`github.com/livioivil/pima`

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

References i

Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., & Finos, L. (2024). Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test. *Psychometrika*, *89*, 542–568. https://doi.org/10.1007/s11336-024-09973-6

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*, 1946–1978. https://doi.org/10.1002/sim.6082

References ii

Götz, M., Sarma, A., & O'Boyle, E. H. (2024). The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the r package multiverse. *International Journal of Psychology: Journal International de Psychologie*, *59*, 1003–1014. https://doi.org/10.1002/ijop.13229

Hemerik, J., Goeman, J. J., & Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *82*, 841–864. https://doi.org/10.1111/rssb.12369

References iii

Klau, S., Felix, Patel, C. J., Ioannidis, J. P. A., Boulesteix, A.-L., & Hoffmann, S. (2023). Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. *Meta-Psychology*, *7*. https://doi.org/10.15626/mp.2020.2556

McCaughey, N. J., Hill, T. G., & Mackinnon, S. P. (2022). The association of self-efficacy, anxiety sensitivity, and perfectionism with statistics and math anxiety. *Personality Science*, *3*. https://doi.org/10.5964/ps.7091

References iv

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*, 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Westfall, P. H., & Stanley Young, S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.