Pacini, D. (2018). On the Size Distortion of a Test for Equality between the ATE and FE Estimands. *Journal of Econometric Methods*, 8(1), [20180011]. <https://doi.org/10.1515/jem-2018-0011>

Publisher's PDF, also known as Version of record

Link to published version (if available):
[10.1515/jem-2018-0011](https://doi.org/10.1515/jem-2018-0011)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via DeGruyter at <https://www.degruyter.com/view/journals/jem/8/1/article-20180011.xml> . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

David Pacini¹

On the Size Distortion of a Test for Equality between the ATE and FE Estimands

¹ Department of Economics, University of Bristol, Bristol, UK, E-mail: David.Pacini@bristol.ac.uk

Abstract:

This note investigates the numerical performance of an existing asymptotic test for the null hypothesis of equality between the average treatment effect (ATE) and the group fixed-effect (FE) estimands based on the standardized difference between ATE and FE estimators. It shows that this test has a size distortion. This distortion has implications to empirical economic research. It can lead to erroneously confirm the relevance of heterogeneous responses to policy interventions.

Keywords: difference test, grouped fixed effects, size distortion

JEL classification: C21, C18, C52

DOI: 10.1515/jem-2018-0011

1 The T-Equality Test

To confirm the relevance of heterogeneous responses to a policy intervention, Gibbons, Suarez Serrato, and Urbancic (2018) introduce an asymptotic statistical test based on comparing two estimators of the ATE. Let us call this test, which is described below, the T-Equality test. This note documents that the T-Equality test can incorrectly reject the null hypothesis of interest more often than the pre-specified level chosen by the test's user. Being aware of this distortion is important because, in practice, it can lead to erroneously reject homogeneity in behavior.

The size distortion is more likely to be a problem when the variance of the treatment is similar, although not necessarily identical, across group. To see this, consider

$$y_i = x_i\beta_g + z_i'\gamma + \epsilon_i, \quad (1)$$

where y_i is the outcome for individual i , x_i is a covariate of interest representing the exposure to a policy intervention (the treatment), $z_i := (z_{1i}, z_{2i})'$ is a list of control covariates, and ϵ_i is a disturbance term satisfying $E(\epsilon_i|x_i, z_i) = 0$. z_{2i} is a list of $G - 1$ dummy variables indicating group membership and coding a discrete variable g_i with support $1, \dots, G$. The coefficients β_g and γ are unknown. β_g are group-specific for each of the $g = 1, \dots, G$ known groups. Data are available from an i.i.d. sample $\{y_i, x_i, z_i\}_{i=1}^N$. The parameter of interest is the ATE $\beta^{ATE} := \sum_{g=1}^G \pi_g \beta_g$, where π_g is the population frequency of group g .

The T-Equality test compares either the regression-weighted estimator $\hat{b}^{RWE}$ or the interaction-weighted estimator $\hat{b}^{IWE}$ of β^{ATE} with the group fixed-effects estimator $\hat{b}^{FE}$.¹ Let β^{FE} denote the limit in probability of $\hat{b}^{FE}$. Let call β^{FE} the FE estimand. The test statistic is

$$T_E := \frac{(\hat{b}^{ATE} - \hat{b}^{FE})^2}{\hat{v}\hat{a}r(\hat{b}^{ATE} - \hat{b}^{FE})},$$

where $\hat{b}^{ATE}$ is either $\hat{b}^{RWE}$ or $\hat{b}^{IWE}$ and the denominator is the estimated variance of $\hat{b}^{ATE} - \hat{b}^{FE}$. The critical value comes from a chi-squared distribution with one degree of freedom [see Gibbons, Suarez Serrato, and Urbancic (2018); Proposition 5]. The T-Equality test tests the null hypothesis $H_0 : \beta^{FE} = \beta^{ATE}$ against $H_A : \beta^{FE} \neq \beta^{ATE}$.

The intuition behind the size distortion documented in the next section is the potential presence of a division by zero (or, more precisely, an indeterminate form) not being ruled out by the T-Equality test under the null hypothesis. To build this intuition, start by comparing the definitions of the estimators

¹David Pacini is the corresponding author.

©2019 Walter de Gruyter GmbH, Berlin/Boston.

$$\hat{b}^{RWE} := \left(\sum_{i=1}^N \tilde{x}_i \hat{w}_i^2 \tilde{x}_i' \right)^{-1} \sum_{i=1}^N \tilde{x}_i \hat{w}_i^2 \tilde{y}_i \quad \text{and} \quad \hat{b}^{FE} := \left(\sum_{i=1}^N \tilde{x}_i \tilde{x}_i' \right)^{-1} \sum_{i=1}^N \tilde{x}_i \tilde{y}_i,$$

where $\tilde{y}_i$ (res. $\tilde{x}_i$) is the residual in the linear projection of y_i (x_i) on z_i . The only difference is the presence of the estimator $\hat{w}_i$ of the inverse of the square root of the conditional variance of $\tilde{x}_i$ given g_i . If $\hat{w}_i$ was the same for all i , $\hat{b}^{RWE}$ and $\hat{b}^{FE}$ would coincide and the dispersion of the difference between $\hat{b}^{RWE}$ and $\hat{b}^{FE}$ would be zero. We now have a link between the dispersion of $\hat{b}^{RWE} - \hat{b}^{FE}$ and the variance of the treatment across groups. Turn now to the definition of the T_E test statistic. For large samples, this test statistic behaves like the square of the ratio of a normal random variable and its variance. Under the null hypothesis, the T-equality test approximates this behavior by a chi-squared distribution. This approximation is valid except when the denominator in the ratio is zero, which occurs when the asymptotic dispersion of $\hat{b}^{RWE} - \hat{b}^{FE}$ is zero.² Using now the link established above, one should intuit that, to avoid using a test statistic behaving like a ratio with a zero in the denominator, the variance of the treatment has to be sufficiently different across groups.

One potential example where the size distortion is likely to be a problem is in the re-analysis by [Gibbons, Suarez Serrato, and Urbancic 2018, Table 6 (b), (d), (f)] of the application by Banerjee and Iyer (2005). In this re-analysis, the group variables are year dummies. The treatment (an historical measure of non-landlord control) is time-invariant. Hence, its variance is likely to differ little across years in this application.

2 Main Result: The T-Equality Test has Size Distortions

The numerical performance of the T-Equality test has not yet been explored in the literature. The following simulation exercises show that the T-Equality test can incorrectly reject the null hypothesis more often than the level chosen by the test's user.

The data generating process (dgp) is an extension of the one in Gibbons, Suarez Serrato, and Urbancic (2018). The extension is motivated by the need to show that the T-Equality has size distortions even when the variance of the treatment is not identical across groups. One thousand datasets ($M = 1000$) with one thousand replications ($N=1000$) are generated according to (1), where:

- ϵ_i follows a normal distribution with mean 0 and variance 10 (denoted $\epsilon_i \sim \mathcal{N}(0, 10)$).
- $z_i = (z_{1i}, z_{2i})'$, with $z_{1i} \sim \mathcal{N}(0, 2)$ and z_{2i} a list of five dummy variables coding the realizations of g_i , which follows a categorical distribution with support $\{1, 2, \dots, 5\}$. Each point of support have equal probability $\pi_g = 1/5$.
- $\gamma = (0.75, 1, 1, 1, 1)$.
- $x_i|z_i \sim \mathcal{N}(\alpha z_{1i}, \sigma(g_i, B))$ with $\alpha = 0.3$ and

$$\sigma(g_i, B) = \begin{cases} 58.33 - B \times 57.33 & \text{if } g_i = 1 \\ 15.03 - B \times 14.03 & \text{if } g_i = 2 \\ 7.39 - B \times 6.39 & \text{if } g_i = 3 \\ 4.57 - B \times 3.57 & \text{if } g_i = 4 \\ 2.18 - B \times 1.18 & \text{if } g_i = 5 \end{cases}$$

- $\beta_g = 3.5$ for all g .

In this dgp, $H_0 : \beta^{FE} = \beta^{ATE}$ is true for any value of $B \in [0, 1]$. Hence, the rejection frequencies from the simulations approximate the Type I error of the test. When $B = 0$ (benchmark exercise), $\sigma(g_i, 0)$ is the same as in Gibbons, Suarez Serrato, and Urbancic 2018 and the conditional variance $V(\tilde{x}_i|g_i)$ of $\tilde{x}_i$ given g_i is different from the unconditional variance $V(\tilde{x}_i)$ of $\tilde{x}_i$. Otherwise, $\sigma(g_i, B)$ is different. When $B = 1$, $\sigma(g_i, 1) = 1$ and $V(\tilde{x}_i|g_i) = V(\tilde{x}_i)$. Given the discussion in the previous section, the T-Equality test could loose control of the Type I error as B moves from 0 to 1.

Figure 1A plots the rejection frequencies for the T-Equality test based on $\hat{b}^{RWE}$ and the heteroscedastic-robust variance estimator at different values of B . When B is such that $V(\tilde{x}_i|g_i)$ and $V(\tilde{x}_i)$ are sufficiently different, the T-Equality test has rejection frequencies (represented by the dots) below the nominal level (indicated by the dotted line). A size distortion (the positive difference between the rejection frequency and the nominal level) appears as $V(\tilde{x}_i|g_i)$ and $V(\tilde{x}_i)$ approach each other.

Figure 1: Monte Carlo Exercises – Size Distortions for the 5% T-Equality Test.
(A) Rejection Frequencies, (B) QQ Plot ($B = 0.8$).

The distortion appears before $V(\tilde{x}_i|g_i)$ becomes equal to $V(\tilde{x}_i)$ – i.e. when the variance of the treatment is similar but not necessarily identical across groups. Figure 1B plots the quantiles of a chi-square random variable with one degree of freedom (the test's critical values) against the quantiles of the T_E statistic (the target) when $V(\tilde{x}_i|g_i)$ is still different from $V(\tilde{x}_i)$. If the critical values were close to the target, the points would lie close to the diagonal line.

To check the robustness of the result, Table 1 reports the empirical rejection probabilities for nominal levels equal to 10%, 5% and 1%. Four variants of the test, resulting from combining $\hat{b}^{IWE}$ and $\hat{b}^{RWE}$ with homoscedastic-only or heteroscedastic-robust variance estimators, are considered. For two versions (rows "rwe/hetero" and "iwe/homo") the rejection frequencies are above the pre-specified nominal levels. For the other two versions, the rejection frequencies are generally below the pre-specified nominal levels.

Table 1: Monte Carlo Exercises – Size Distortions for the T-Equality Test at $B = 1$.

Nominal	$N = 1000 \ M = 1000$			$N = 4000 \ M = 1000$			$N = 4000 \ M = 4000$		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
rwe / hetero	41%	33%	22%	32%	25%	13%	35%	27%	14%
rwe / homo	7%	6%	4%	0%	0%	0%	1%	0%	0%
iwe / hetero	3%	2%	1%	3%	2%	1%	3%	2%	1%
iwe / homo	15%	10%	6%	16%	12%	7%	15%	11%	7%

N stands for the sample size and M for the number of Monte Carlo replications.

Table 1 also reports the rejection frequencies when, first, the sample size is fourfold and, then, the number of simulations is fourfold. The rejection frequencies in these exercises remain well above or below the pre-specified significance level without a clear pattern.

3 Discussion

A size distortion in the T-Equality test, like the one documented in this note, can lead to erroneously confirm the relevance of heterogeneous responses when estimating the average effect of a policy intervention. To provide a theoretical justification for this distortion, it suffices to notice the existence of a dgp for which the FE and ATE estimands are the same and the denominator in T-Equality test statistic converges fast enough to zero. The dgp described in Section 2 is a case in point. The size distortion is more likely to appear when the variance of the treatment does not vary enough across groups. The problem with the T-Equality test, which is not specific to the dgp in Section 2, is that it does not specify how much variation should be enough to guarantee size control.

To rectify the size distortions, one could consider testing $H_0 : \beta^{FE} = \beta^{ATE}, V(\tilde{x}_i|g_i = g) \neq V(\tilde{x}_i) \forall g$ against $H_A : \beta^{FE} \neq \beta^{ATE}, V(\tilde{x}_i|g_i = g) \neq V(\tilde{x}_i) \forall g$. Now, the same nuisance parameter $V(\tilde{x}_i|g_i = g) \neq V(\tilde{x}_i)$ is under the null and the alternative hypotheses, which are both composite. Elliot, Muller, and Watson (2015) describe a general approach to construct a nearly optimal test for this type of testing problems. Tailoring this general approach to testing the difference between the ATE and FE estimands is out of the scope of this note.

Acknowledgement

I would like to thank Helene Turon, Chris Muris, and Yanos Zylberberg for useful suggestions.

Notes

- 1 Relevant definitions are provided below. To distinguish $\hat{b}^{FE}$ from the individual dummy variable least-squares estimator – also called the “fixed-effects” estimator – we call $\hat{b}^{FE}$ the “group fixed-effects” estimator.
- 2 Without loss of generality, set $\hat{w}_i := (z'_{2i} \tilde{\sigma} / \sum_{i=1}^N \tilde{x}_i^2)^{-1/2}$ where $\tilde{\sigma} := (\sum_{i=1}^N z_{2i} z'_{2i})^{-1} \sum_{i=1}^N z_{2i} \tilde{x}_i^2$. Let $A_N := [N^{1/2}(\hat{b}^{RWE} - \hat{b}^{FE})]^2$. One can write $T_E = A_N / a\hat{var}(\hat{b}^{RWE} - \hat{b}^{FE})$, where $a\hat{var}()$ is an estimator of the asymptotic variance of $\hat{b}^{RWE} - \hat{b}^{FE}$. In particular, for the heteroscedastic-robust asymptotic variance estimator, $a\hat{var}(\hat{b}^{RWE} - \hat{b}^{FE})$ converges in probability to $E(w_i^2 \tilde{x}_i^2)^{-2} E(w_i^2 \tilde{x}_i^2 \epsilon_i^2) + E(\tilde{x}_i^2)^{-2} E(\tilde{x}_i^2 \epsilon_i^2) - 2E(w_i^2 \tilde{x}_i^2)^{-1} E(\tilde{x}_i w_i \epsilon_i^2 \tilde{x}_i) E(\tilde{x}_i^2)^{-1}$, where $w_i := V(\tilde{x}_i | g_i) / V(\tilde{x}_i)$ is the limit in probability of $\hat{w}_i$. Under $V(\tilde{x}_i | g_i) = V(\tilde{x}_i)$, one has $w_i = 1$ and the limit of $a\hat{var}(\hat{b}^{RWE} - \hat{b}^{FE})$ becomes zero. Then, under the null hypothesis, both A_N and $a\hat{var}(\hat{b}^{RWE} - \hat{b}^{FE})$ can converge in probability to zero. If $a\hat{var}(\hat{b}^{RWE} - \hat{b}^{FE})$ converges in probability faster than A_N , the T-Equality test in the limit will overreject. Otherwise, it will underreject. The same prediction follows if one uses the homoscedastic-only asymptotic variance estimator. In such a case, $a\hat{var}(\hat{b}^{RWE} - \hat{b}^{FE})$ converges in probability to $\sigma_\epsilon^2 [E(w_i^2 \tilde{x}_i^2)^{-1} + E(\tilde{x}_i^2)^{-1} - 2E(w_i^2 \tilde{x}_i^2)^{-1} E(\tilde{x}_i w_i \tilde{x}_i) E(\tilde{x}_i^2)^{-1}]$, where σ_ϵ^2 is the variance of ϵ_i . The Monte Carlo exercises in Section 2 bear out these theoretical predictions.

References

- Banerjee, A., and L. Iyer. 2005. “History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India.” *American Economic Review* 95 (4): 1190–1213.
- Elliot, G., U. Muller, and M. Watson. 2015. “Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis.” *Econometrica* 83 (2): 771–811.
- Gibbons, C., J. Suarez Serrato, and M. Urbancic. 2018. “Broken or Fixed Effects?” *Journal of Econometric Methods*. DOI: 10.1515/jem-2017-0002.