Bangor Business School

Working Paper

BBSWP/19/14

Mitigating misleading implications for policy: Treatment of outliers in a difference-in-differences framework

By

Philip Molyneux[a] Alessio Reghezza[b*1] Jonathan Williams[b]

[a] University of Sharjah, College of Business Administration, P. O. Box 27272, Sharjah, UAE

[b] Bangor University, Bangor Business School, Hen Goleg, College Road, LL57 2DG, Bangor, UK

May, 2019

Bangor Business School
Bangor University
Hen Goleg
College Road
Bangor
Gwynedd LL57 2DG
United Kingdom
Tel: +44 (0) 1248 382277

E-mail: business@bangor.ac.uk

[1] * Corresponding author. Email address: elx64e@bangor.ac.uk

Abstract

Applications of the difference-in-differences estimator in economics, banking and finance, and management commonly treat outliers using the winsorize method. However, failure to winsorize outliers in both the treatment and controls groups introduces volatility in estimated coefficients, significance levels, and standard errors. A faulty process can lead to an exogenous event realising a significant effect that proper process would fail to detect. In demonstration, we randomly generate placebo interventions in bank-level data and discuss how to detect and limit the problem.

JEL: C00, C10, C13, C18, C19, C80

Keywords: Winsor, Difference-in-Differences, Outliers, Financial Data, Research Design

1. Introduction

The practice of 'winsorizing' (or 'winsorization') is a valid and popular tool for researchers needing to deal with outliers in a distribution of data.[2] Named after 20th century biostatistician Charles Winsor, winsorization replaces extreme value/s (or 'outliers') with the value of the highest data point not considered an outlier; winsorizing transforms data to limit the effects of outliers rather than removing observations.[3] An 80% winsorized mean averages data below 10% and above 90%; retention of data is a benefit for significance testing purposes as opposed to reducing sample size. Notwithstanding valid reasons to contain outliers, Leone et al (2014) caution against losing important information inherent in extreme observations.

The difference-in-differences (DD) framework creates additional considerations. Applications of DD investigate difference in trends between two groups; a treatment (group affected) and control (group unaffected) following an intervention and/or exogenous shock. We must recognise that outlying values might be different for treatment and control groups. This implies we should winsorize data separately for each group; otherwise, we are applying values from one group to the other, which may cause serious estimation problems. Reading a random sample of 50 papers that apply DD suggests the problem is widespread. We examine values of maximum and minimum descriptive statistics to detect the anomaly; equal values of maximum and minimum in treatment and control groups before and after intervention suggests incorrect application of winsor.[4] Avoidance of detailed summary statistics in DD papers applying winsor is a signal of potential problems.[5]

Using bank-level data, we construct DD placebo interventions to demonstrate how severe the problem is. The estimations uncover large variation in estimated coefficients, significance, and standard errors on applying different winsor techniques. Our results speak to the DD literature;

---

[2] Using a sample of top finance journal papers (JF, JFE, RFS, JFQA) over the period 2008-2012, Adams et al. (2017) show that winsorizing covers 49% of the outliers' mitigation methods.
[3] Trimming and dropping remove data completely from the sample.
[4] Section 2 will provide a detailed explanation of how to detect the problem.
[5] DD Winsor papers tend to display the first and last percentiles in descriptive statistics rather than minimum and maximum values. To gain further insight into this problem, we randomly select 50 DD Winsor and 50 Winsor-no-DD papers. 26 DD Winsor papers present just mean and standard deviation in comparison with 7 Winsor-no-DD. 11 DD-Winsor papers display first and last percentiles in comparison with 5 Winsor-no-DD. Finally, 13 DD-Winsor papers show minimum and maximum values in comparison with 37 Winsor-no-DD. These results concur with Adams et al. (2017) and Leone et al. (2015) who find papers avoid mentioning preferred methods to mitigate outliers and related information.

we identify ramifications arising from a technical issue in DD applications, which could lead to faulty policy recommendations if unchecked (Bertrand et al., 2004; Donald and Lang, 2007; Imbens and Wooldridge, 2009; Roberts and Whited, 2012). The problem we detect is little discussed in the wider academic literature. By using DD placebo regressions, we quantify the extent of mismeasurement due to improper treatment of outliers. Lastly, we offer suggestions on how to detect and tackle the problem in future research.

In what follows section 1 explains the problem and how to detect it. Section 2 introduces data and methodology. Section 4 presents results and Section 5 concludes.

2. Detecting the problem

We use a stylised normal density function to show the erroneous application of Winsor. Figure 1 shows two distributions of data divided by the control (0 on top of the figure) and the treatment (1 on top of the figure). In the treatment group, we arbitrarily introduce an outlier to skew the distribution to the left to justify winsorizing.[6] We apply two strategies: normal winsorization and winsorization by group. Figure 2 illustrates the outcome of applying winsor without distinguishing between two groups. Winsor replaces the outlier with the smallest value in the two groups (zero in this case). However, zero is a value belonging only to the control group; zero appears an outlier following winsorization.[7] Winsorizing by group and treating outliers separately mitigates this problem (see figure 3).

[6] Examining the most suitable technique (winsorizing, trimming or dropping, for example) in this circumstance is beyond the scope of this paper.

[7] Section 3 will discuss problems related to the estimation.

Figure 1. Normal density function (no-Winsor)

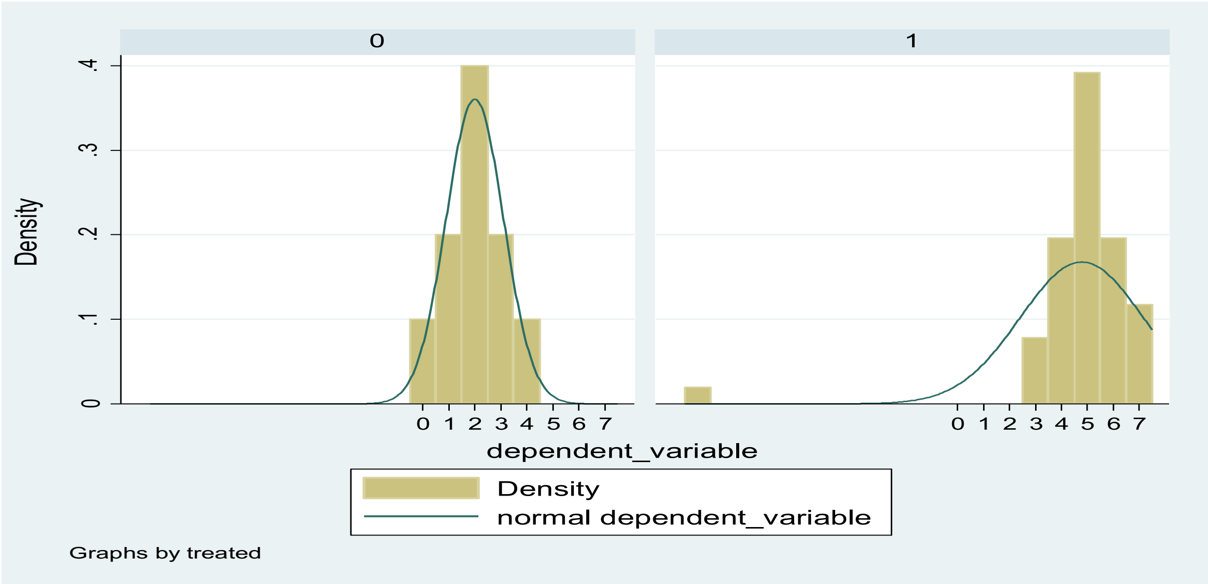


Figure 2. Normal Density Function (Winsor)

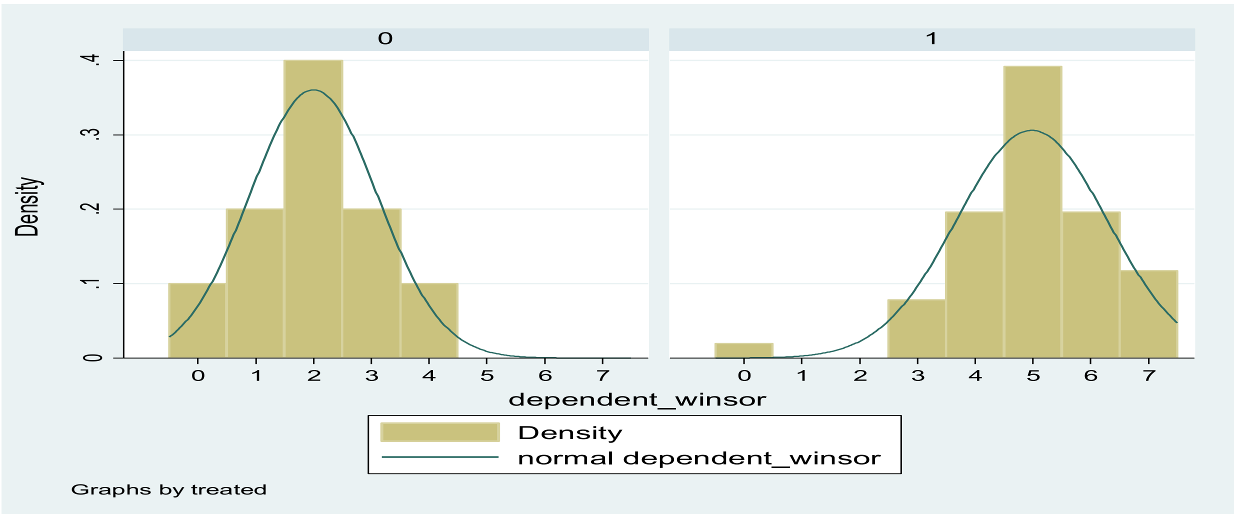
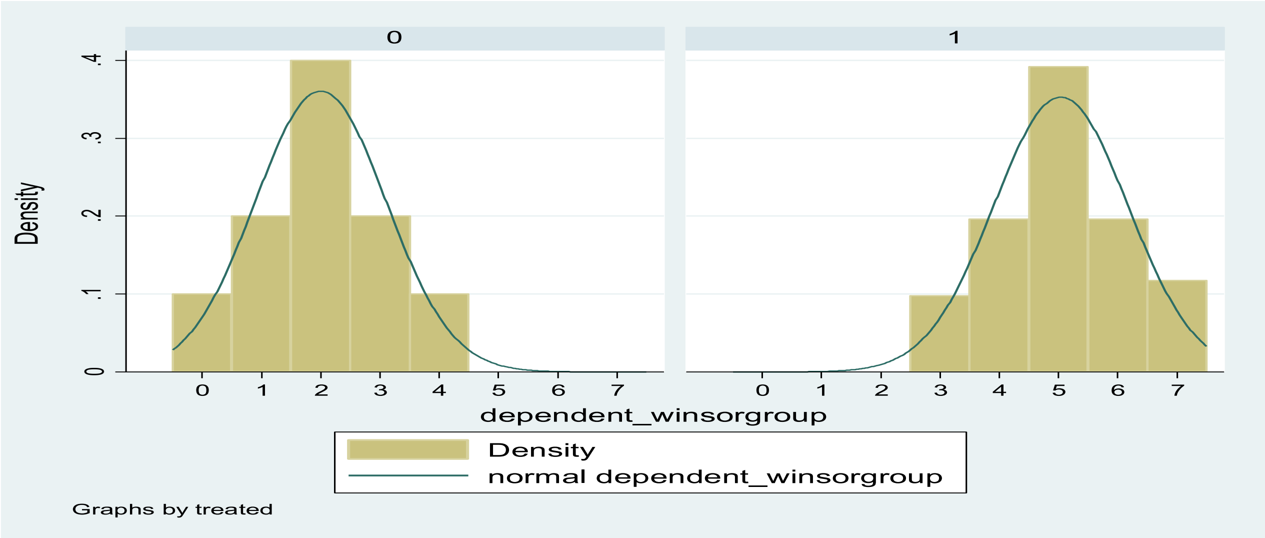


Figure 3. Normal Density Function (Winsor by group)



Picture 1 shows how we detect the winsor problem if authors provide descriptive statistics for treatment and control groups. The authors state that they winsorize the variable *Daily efficiency* at 2 standard deviations from the mean. The picture shows *Min* and *Max* are equal for both treatment and control groups prior and after the intervention, which signals application of winsorization without differentiating between groups. Improper treatment of outlying values of *Daily efficiency* can lead to serious estimation problems.

Picture 1. Winsor Problem Detection in Min and Max Descriptive Statistics

	Pre-award					Award period				
Variable	Obs	Mean	Sd	Min	Max	Obs	Mean	Sd	Min	Max
					Treatment					
Tardy	7,126	0.04	0.20	0.0	1.0	5,742	0.03	0.16	0.0	1.0
Minutes late	7,126	−1.73	5.37	−59.7	54.0	5,737	−2.19	5.10	−49.0	58.2
Total absences	7,846	0.04	0.19	0.0	1.0	6,318	0.05	0.21	0.0	1.0
Single absences	7,846	0.01	0.07	0.0	1.0	6,318	0.01	0.11	0.0	1.0
Daily efficiency	6,576	125.60	32.40	54.5	218.0	5,669	124.70	36.20	54.5	218.0
Late	7,126	0.15	0.35	0.0	1.0	5,742	0.10	0.29	0.0	1.0
Monthly absences	7,861	0.79	1.65	0.0	8.0	6,339	1.03	1.80	0.0	10.0
Total hrs worked	7,861	8.25	1.64	0.0	23.7	6,339	8.17	1.69	0.0	13.1
Tenure	7,861	3,373	2,681	176	9,262	6,339	3,186	2,708	176	9,262
Age	7,861	43.51	10.05	21.0	62.0	6,339	43.63	9.84	21.0	62.0
Male	7,861	0.43	0.50	0.0	1.0	6,339	0.43	0.49	0.0	1.0
Base salary	7,861	25,695	845	19,240	28,600	6,339	25,495	1,363	19,240	28,600
					Control					
Tardy	23,600	0.03	0.17	0.0	1.0	14,081	0.02	0.15	0.0	1.00
Minutes late	23,600	−2.12	4.73	−59.4	59.0	14,081	−1.88	5.07	−59.0	59.0
Total absences	27,449	0.05	0.22	0.0	1.0	16,300	0.05	0.22	0.0	1.00
Single absences	27,449	0.02	0.13	0.0	1.0	16,300	0.02	0.12	0.0	1.00
Daily efficiency	22,320	120.30	34.70	54.5	218.0	14,148	125.30	35.60	54.5	218.00
Late	23,600	0.17	0.38	0.0	1.0	14,081	0.14	0.34	0.0	1.00
Monthly absences	27,684	1.06	1.76	0.0	15.0	16,474	1.13	1.78	0.0	12.00
Total hrs worked	27,684	8.01	1.38	0.0	44.0	16,474	8.09	1.25	0.0	16.37
Tenure	27,684	1,720	1,920	4.0	8,566	16,474	1,754	1,992	60.0	8,566
Age	27,684	39.36	12.88	18.0	69.0	16,474	39.87	12.79	19.0	69.00
Male	27,684	0.34	0.47	0.0	1.0	16,474	0.32	0.47	0.0	1.00
Base salary	27,522	19,850	4,899	8,320	48,526	16,474	19,734	5,167	8,320	48,526

Source: Gubler, T., Larkin, I., Pierce, L., 2016. Motivational spillovers from awards: Crowding out in a multitasking environment. Organization Science 27(2): 286-303 (page 292). https://doi.org/10.1287/orsc.2016.1047

3. Methodology and Data

3.1 Methodology

We use random samples of bank-level data and randomly generate placebo interventions to capture the effect of applying different winsor techniques. Equation (1) shows a standard DD framework:

$$Y_{it} = \alpha + \beta_1 Treated_i + \beta_2 Post_t + \beta_3 (Treated_i * Post_t) + \varphi_t + \gamma_i + \varepsilon_{it} \qquad (1)$$

Where Y_{it} is the dependent variable[8] of bank i at time t. $Treated_i$ is a dummy variable taking the value of 1 if bank i is affected by an intervention/shock and 0 otherwise[9]. $Post_t$ is a dummy variable equal to 1 following intervention/shock and 0 before[10]; β_3 represents the average difference in the dependent variable between banks in the treatment and control groups prior to and after the intervention/shock. In common with DD applications in banking, γ_i and φ_t capture bank and year fixed effects, respectively, and limit potential for bias in estimates of $\beta_{3.}$[11]

Our estimations of equation (1) apply three treatments to outliers. First, we exclude treatment of outliers. Next, we use Winsor indiscriminately on both groups (the method we regard as technically incorrect). Lastly, we apply Winsor to treated and untreated banks (the 'best alternative'). In all applications, we winsorize the dependent variable at 1% and 99%.[12]

3.2 Data

Our sample includes 16,675 financial institutions from 2008 to 2015; Orbis Bank Focus is the source of our bank balance sheet data.

Table 1 presents descriptive statistics from the three treatments. The first and fourth rows show statistics without treatment of outliers, absent Winsor; the standard deviation of E/TA (equity-to-total assets) is significantly larger than treated data for both groups due to the number of outliers in the sample. The second and fifth rows show statistics for indiscriminate use of winsor, E/TA Winsor; group minimum and maximum are equal and standard deviation comparable across groups. The third and sixth rows show application of winsor by group. Now, minimum and maximum values and standard deviations differ between groups. E/TA Winsor

[8] Initially, we employ the ratio of equity-to-total assets as an illustration. This ratio is a measure of bank solvency, which is an important indicator of the financial resilience of banks and the wider financial sector. Hence, providing erroneous results due to improper mitigation of how to treat outliers can have strong repercussions for interpretation of policy relevant empirical studies. For robustness, and later in the paper, we select different bank performance indicators; the ratios of net interest income-to-average earnings, securities-to-total assets, non-interest income-to-gross revenue, liquid assets-to-deposits and short-term borrowing, and deposits-to-total liabilities. The decision to use these variables is not casual; we choose variables that exhibit a substantial number of outliers in the dataset. In addition, we also apply alternative intervention windows.

[9] The sample banks come from OECD countries. Treated banks (dummy equal to 1) are European, untreated (dummy equal to 0) are non-European.

[10] In this case, the dummy variable equals 1 after 2014 and 0 otherwise (intervention window 2012-2016). We use different estimation windows to confirm the validity of our results.

[11] Standard errors are robust and clustered at bank level.

[12] Researchers apply different Winsor levels, such as, 10% and 90% or 5% and 95%. Most common are 1% and 99%. Of the 50 papers we examine, 39 use 1% and 99% levels of Winsorization.

presents two problems. The first concerns the mean. E/TA Winsor replaces outliers with equal values in both treatment and control groups. For the treatment group, the mean of E/TA Winsor is larger in comparison to no treatment (absent Winsor) and winorizing by group. In contrast, the mean of the control group of E/TA Winsor is lower than the other two means. Second, the standard deviations of E/TA Winsor show less variation, which renders more homogenous the two groups.

Table 1. Descriptive Statistics: Control and Treatment groups

	Obs.	Mean	Std. Dev.	Min	Max
Treatment					
E/TA absent Winsor	17182	12.94	18.45	-967.21	100
E/TA Winsor	17182	14.03	8.51	4.52	30.95
E/TA Winsor by group	17182	10.14	5.15	4.07	21.76
Control					
E/TA no Winsor	9654	18.55	29.66	-969.91	100
E/TA Winsor	9654	10.96	7.07	4.52	30.95
E/TA Winsor by group	9654	16.55	13.91	5.47	51.27

4. Results

Table 2 shows results from estimating equation (1) and applying each Winsor technique. All estimations specify bank and year fixed effects. Column 1 of table 1 shows the result without Winsor. The coefficient of interest, *Treatment (E/TA),* is statistically insignificant with large standard error. The coefficient is insignificant at conventional levels when we winsorize by group (see column 3). However, the coefficient's significance changes if we fail to differentiate between treatment and control groups; it is statistically significant at the one percent level and the magnitude of standard error smaller (see column 2). In a DD framework, the inference would be that the effect of the treatment on the treated group led to a significant difference in E/TA in the control group after intervention. In terms of our application, the effect of treatment realises improvement in bank solvency, of interest to bank management, bank regulators and policymakers, but the implication is misleading because of improper use of winsorization.

Table 2. Winsorization Techniques and DD Methodology

	No-Winsor (1)	Winsor (2)	Winsor by group (3)
Treated	-5.670***	-3.085***	-6.626***
	(0.606)	(0.196)	(0.279)
Period	0.385	0.203***	0.252**
	(0.353)	(0.0620)	(0.0989)
Treatment (E/TA)	0.129	0.190***	0.147
	(0.365)	(0.0719)	(0.103)
No. of Banks	7467	7467	7467
No. of Observations	26836	26836	26836

To ensure sample selection does not affect the results, we re-estimate equation (1) on multiple random samples and different intervention windows. Table 3 shows the Treatment (E/TA) coefficient rejection rate of the three Winsor techniques for alternative intervention windows (2008-2012, 2012-2016)[13], 106 random samples and 6 variables. When winsorization is applied to both groups the rejection rate, namely rejecting the null hypothesis that β_3 is equal to zero in favour of the alternative that β_3 differs from zero, is the highest supporting the validity of our previous results.[14]

[13] Intervention is set in 2010 and 2014 for the two windows.
[14] We consider as rejected under a 10% p-value significant level.

Table 3. Rejection Rate Winsor Techniques

Placebo windows	No. of Samples	No. of Variables	No-Winsor Rejection Rate Treatment (E/TA) (1)	Winsor Rejection Rate Treatment (E/TA) (2)	Winsor by Groups Rejection Rate Treatment (E/TA) (3)
2012-2016	57	6	21%	75%	36%
2008-2012	49	6	18%	69%	37%
	53	6	20%	72%	37%

5. Conclusion

Much of the economics, management, banking and finance literature that applies difference-in-differences use winsorized data to deal with outliers. However, the practice of winsorizing data, and replacing outliers with values equal in both treatment and control groups, can produce misleading results, and by extension, faulty inferences for policy. This paper demonstrates the effects of improper use of winsorization. Our recommendation is to always apply winsorization separately for treatment and control groups in a DD framework.

References

Adams, J.C., Hayunga, D.K., Mansi, S., Reeb, D.M., 2017. Identifying outliers in finance. Available at SSRN: https://ssrn.com/abstract=2986928

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust difference-in-differences estimates? *Quarterly Journal of Economics* 119, 249-275.

Donald, S., Lang, K., 2007. Inference with difference-in-differences and other panel data. *Review of Economics and Statistics* 89, 221-233.

Gubler, T., Larkin, I., Pierce, L., 2016. Motivational spillovers from awards: Crowding out in a multitasking environment. *Organization Science* 27(2): 286-303 (page 293). https://doi.org/10.1287/orsc.2016.1047

Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47, 5-86.

Leone, J.A., Minutti-Meza, M., Wasley, C.E. 2014. Influential observations and inference in accounting research. *Simon Business School Working Paper No. FR 14-06.* Available at SSRN: https://ssrn.com/abstract=2407967 or http://dx.doi.org/10.2139/ssrn.2407967.

Roberts, M.R., Whited, T.M., 2012. Endogeneity in empirical corporate finance. *Handbook of the Economics of Finance*, Editors G.M. Constantinides, M. Harris and R. M. Stulz, Volume 2 Part A, Chapter 7, 493-572 (NY: Elsevier).