A Comment on “The Macroeconomics of Sticky Prices with Generalized Hazard Functions” by Alvarez et al.*

Anson T. Y. Ho[†] Kim P. Huynh[‡] Carson H. Rea[§]

Ottawa Replication Games on May 5, 2023

Abstract

We replicate the empirical results in Section 4 of Alvarez et al. (2022). First, we were able to reproduce the original authors’ major empirical results, but only after editing the program for it to run on our computing platform. There are small discrepancies in the empirical estimates, e.g. bootstrapped standard errors, that involve the use of simulations. Second, we replicated Alvarez et al.’s results by adopting the data cleaning criteria used by their original data source (Cavallo 2018) to evaluate its robustness to data handling procedures. We found noticeable changes in the empirical results that can have important implications on the effects of monetary policy. To conclude, we propose using Docker container to promote research reproducibility, and more attention is needed on data handling to improve the robustness of empirical research.

KEYWORDS: Sticky Price; Generalized Hazard Function
JEL CODES: E3, E5.

*We thank Abel Brodeur for his helpful comments. The views expressed in this article are those of the authors. The Bank of Canada is not responsible for any views expressed in the paper. All remaining errors are the authors’ responsibility.

[†]Ted Rogers School of Management, Toronto Metropolitan University, 55 Dundas Street West, Toronto ON, M5G 2C3, Canada. E-mail: atyho@torontomu.ca

[‡]Corresponding Author: Bank of Canada, 234 Wellington Ave., Ottawa ON, K1A 0G9, Canada. E-mail: khuynh@bankofcanada.ca.

[§]Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: chrea@emory.edu

1 Introduction

Alvarez et al. (2022) study firm's price-setting behavior in an economy with sticky prices. Through an analytical framework, they suggest that price-setting decisions can be described by a generalized hazard function, applicable to a large class of sticky-price models including those with random adjustment costs, information frictions, or exogenous random adjustment shocks. The model is fully identified using the observed distribution of price changes, and the cumulative impulse response of output to a monetary shock depends on the kurtosis of the distribution of price changes and the frequency of price adjustment.

Alvarez et al.'s framework is empirically applied to the U.S. economy using the price data from the Billion prices Project by Cavallo (2018). Specifically, the data set contains daily prices from March 2008 to August 2010, scraped from 4 of the largest retailers. It contains 49 categories, 172 products, and 28 million observations.

In this report, we investigate whether their empirical results are reproducible and examine their robustness to different data cleaning criteria. Alvarez et al. (2022) allows for longer tails in the distribution of price changes, when compared to the data processing criteria used in Cavallo (2018), the original data source. Our programs are available in our GitHub repository.[1]

To reproduce Alvarez et al.'s empirical results, we obtained their computation program and their data set from Alvarez et al. (2021). Programming edits and additional setups were needed before we could execute their computation program. Overall, we successfully reproduced Alvarez et al.'s major empirical estimates. There were small discrepancies in the bootstrapped standard errors. Also, unfortunately, we were not given sufficient information to use those empirical results to reproduce the underlying distributions and other empirical results reported in the original paper.

In our replication exercise, we dropped 5,538 observations (about 1.7%) from both

[1]URL: https://github.com/atyho/Ottawa-Replication-Games-2023

tails, instead of 87 (about 0.03%) in Alvarez et al. (2022). Our replication results show the expected effect of excluding outliers – a decrease in the variance and the kurtosis while the mean remains about the same. In particular, the kurtosis are reduced by 6% to 29%, depending on different product categories.

2 Reproducibility

We followed the data availability link on the Quarterly Journal of Economics website and downloaded the authors' replication program at the Harvard Dataverse (Alvarez et al. 2021). The data repository contains a Jupyter Notebook (.ipynb file) and a comma-separated-values data file. Unfortunately, it does not contain any instruction on how to execute the program, nor additional details on the programming language and the libraries being used.[2]

After studying the computation program, we recognized that it was written in Julia and some of the libraries used were not included in the default installation. Without further information on the version of Julia kernel and libraries being used in the original computation, we attempted to reproduce the results using the latest versions as of May, 2023.[3] We also needed to make minor edits to the program before it could be successfully executed. Our programs are available in our GitHub repository.

In general, we were able to reproduce Alvarez et al.'s major empirical results, except with minor discrepancies in the statistics involving simulations. Table 1 reports our reproduced summary statistics and main empirical estimates, corresponding to Table 1 in Alvarez et al. (2022). Our program reproduced the exact results as Alvarez et al.'s except the bootstrapped standard errors for the kurtosis for pooled observations and the kurtosis with unobserved heterogeneity, shown in parentheses in column 6 and 7, respectively. The discrepancies appeared in 5 out of the 14 estimates, and they are all

[2] Alvarez et al. (2021) does not include a README file or documentation other that some in-program comments.

[3] We used Julia LTS version 1.6.7 as our kernel. See our replication file for details on the required libraries and their versions.

less than 0.01. We examined the numerical results with more decimal places displayed, and verified that they were not due to numerical rounding.[4]

Given the empirical estimates, we did not have sufficient information in the authors' program to reproduce Figure 3 in Alvarez et al. (2022) for the estimated distribution and the implied cost functions (with unobserved heterogeneity). The figure was not generated by the program in the authors' data repository and the parametrization of the Gamma distribution was not specified in the paper. As such, we were not able to follow the authors' description on page 1015, second last paragraph, to reproduce the distribution of menu costs. Consequently, we could not verify the authors' claim that the mode, the median, and the mean of the menu cost is 6 basis points, 60 basis points and 80 basis points, respectively.

We also attempted to reproduce the empirical estimates reported in the Online Appendix and found similar discrepancies. Specifically, in the case that the price change distribution is the mixture of two Gamma distributions, our results in Table 2 shows that we were not able to reproduce some of the parameter and moment estimates. We believe the discrepancies in $\hat{\alpha}_{22}$ are due to the moments of the distributions being estimated via bootstrapping, so as the estimated standard deviation for the kurtosis using the first two price changes, reported in Table 3. For other discrepancies, it could be various reasons such as different numerical optimization techniques used in different package versions.

To summarize, the discrepancies in the empirical estimates are reported in Table 4. Our findings on the research reproducibility of Alvarez et al. (2022) are summarized in Table 5.

Computationally, we note that a random seed was set at the beginning of the program, so it is unlikely that the discrepancies were generated due to different random

[4] Although we were not able to execute the authors' original program to obtain more precise results, the kurtosis estimates reported in Table 3 (column 3) on Alvarez et al. (2022) online appendix page 24 contains 3 decimal places, which is also different from our reproduced estimates.

number sequences being used in bootstrapping. We further explore how the estimates may behave with different number of repetition in bootstrapping. Taking the standard errors for kurtosis in Table 1 as an example, we changed the number of repetitions from 20% to 500% of the original setup, i.e. 1,000 repetitions. The estimates are reported in Table 6. Our results show that the estimates also bounced around slightly to a similar extent as the discrepancies that we found in Table 4.

3 Robustness Analysis

We examine the robustness of the empirical results to different data cleaning criteria. This replication exercise is motivated by the important roles played by the variance and the kurtosis of price changes in the authors' theoretical framework.[5] A major analytical result (Section 5, page 1020) is that the cumulative impulse response (CIR) of output to a once-and-for-all monetary shock is proportional to the kurtosis. Also, technically, we noticed a difference in the authors' data cleaning criteria when compared to the one used by Cavallo (2018), for which the data set was collected. According to Section II C. paragraph 5 in Cavallo (2018),

> "There are a few price changes in each country that seem too large and are most likely the result of scraping mistakes. Although these are a negligible part of all observations, they can affect statistics related to the magnitude of price change. Consequently, all daily price changes that exceed 200% or -70% are excluded from all duration and size calculations."

On the other hand, Alvarez et al. (2022) removed log price changes greater than 1.5.[6] That translates into dropping price changes greater than 448.17% or -77.69%, which

[5] We follow the original authors' suggestion to call this section "Robustness Analysis" (see the original authors' response for details). According to Journal of Applied Econometrics (2024), replication in a narrow sense involve checking the submitted data against the primary sources for consistency and/or the validity of computations by carrying out the estimation (including standard errors) using other computer packages.

[6] Alvarez et al. (2022) footnote 19 on page 1014.

accounts for 87 out of 326,570 price changes in their sample.

We re-estimated the authors' model using a data exclusion criteria similar to Cavallo (2018), while following the same econometric method and optimization routines. Specifically, log price changes greater than 0.693 are dropped from the sample, which is equivalent to excluding price changes greater than 200% or -50%. Our criteria is not exactly the same as Cavallo (2018) as we opt to follow the authors' assumption of having a symmetric price change distribution.

Replication results are reported in Table 7. As a result of stricter data cleaning criteria, 5,538 out of the same 326,570 price changes were dropped. Compared to the results in Table 5, the effects of dropping tail observations are in the expected direction: (1) fewer products and price changes in each category, (2) expected price changes remained about the same but with smaller variation, and (3) the kurtosis estimates for the pooled sample and the one with unobserved heterogeneity both decreased.

Focusing on the (preferred) kurtosis with unobserved heterogeneity, the extent of the decrease varies across categories. On the lower end, it decreased by 5.89% (from 1.7 to 1.6) in categories 111 and 122. On the upper end, the kurtosis decreased by 28.57% (from 2.1 to 1.5) in category 1212.[7] These changes are noticeable and important as they directly translate into smaller CIR in an 1-to-1 scale.

Also note that the "Calvo-ness" measures in the last 2 columns, which show the fraction of price changes independent of the firm-specific price gap, also slightly decreased. These results follow our data processing that removing tail observations increased the empirical density of smaller price changes. In terms of economics, we dropped large price changes were likely to be triggered by substantial firm-level price deviation from the market.

[7]We thank the original authors for pointing out the error in percentage change in the previous version.

4 Conclusion

We encountered technical difficulties when we attempted to execute the authors' original computation program from their data repository. Our experience may potentially due to by using different versions of the libraries and/or a different operating environment. It sheds light on the potential benefits of using containers to enhance research reproducibility. If a container is used to specific the exact environment for replicating the authors' results, most of the programming issues we encountered could have been avoided. While containers are more commonly used in other disciplines in the scientific community (Boettiger 2015) and there have been advocates of its adoption in economics (AEA Data Editor 2021), its use is still rather uncommon and not supported by academic journals in economics. To illustrate the use of a container, as an example the replication program in our GitHub repository can be executed via a Docker container.

Unfortunately, without much information available from the data repository, we were unable to pin down the exact reason why the original program did not work for us. Although we were able to reproduce Alvarez et al.'s major empirical findings, we could not explain why and how we got the discrepancies in our reproduction exercise. Extra efforts were spent on investigating the program line-by-line to understand how it works, to identify variables/objects of interest, and to conduct the replication exercise.

Our replication exercise further suggests that more attention on data processing is needed in empirical applications. While our replication results are consistent with the theoretical model in Alvarez et al. (2022), they also serve as a note of caution for scholars, policymakers, and industry practitioners, who intend to apply the authors' important theoretical contribution to further empirical work. Future and ongoing empirical research using this framework is also discussed in Section 8 of Alvarez et al. (2022). Indeed, the rationale for their specific data cleaning criteria is not explained in neither Alvarez et al. (2022) nor Cavallo (2018). In generally, the appropriate rules for classifying, and dropping, outliers remain an open question. Sensitivity tests on data cleaning, in addition to that on statistical methods, are warranted for research work that depends on the use of higher statistical moments.

References

AEA Data Editor: 2021, Use of docker for reproducibility in economics, *Technical report*, Office of the AEA Data Editor.

URL: *https://aeadataeditor.github.io/posts/2021-11-16-docker*

Alvarez, F., Lippi, F. and Oskolkov, A.: 2021, Replication Data for: 'The Macroeconomics of Sticky Prices with Generalized Hazard Functions'.

URL: *https://doi.org/10.7910/DVN/IBM0IE*

Alvarez, F., Lippi, F. and Oskolkov, A.: 2022, The macroeconomics of sticky prices with generalized hazard functions, *The Quarterly Journal of Economics* **137**(2), 989–1038.

Boettiger, C.: 2015, An introduction to docker for reproducible research, *ACM SIGOPS Operating Systems Review* **49**(1), 71–79.

Cavallo, A.: 2018, Scraped data and sticky prices, *The Review of Economics and Statistics* **100**(1), 105–119.

Journal of Applied Econometrics: 2024, Contribute: Replication section, https://pericles.pericles-prod.literatumonline.com/page/journal/10991255/homepage/news.html#replication. Accessed: February 12, 2024.

5 Tables

Table 1: Reproduced Summary Statistics and Kurtosis Estimates

Category	Number products	Number p. changes	$\hat{\mathbb{E}}(\Delta p_{it})$	$\hat{\sigma}(\Delta p_{it})$	Kurtosis pooled	Kurtosis w/unobs. heterog.	C_{pooled}	C w/unobs. heterog.
111	3,439	75,144	0.002	0.34	3.4	1.7	0.07	0.06
					(0.16)	(0.07)		
119	3,228	56,898	0.002	0.33	3.8	2.0	0.10	0.07
					(0.10)	(0.05)		
1212	2,551	30,361	-0.001	0.25	3.5	2.1	0.07	0.06
					(0.28)	(0.17)		
122	1,405	27,563	0.002	0.34	3.0	1.7	0.10	0.07
					(0.09)	(0.06)		
118	1,390	30,492	0.003	0.31	3.6	2.0	0.06	0.06
					(0.23)	(0.12)		
117	1,154	21,123	0.007	0.31	3.5	2.0	0.08	0.06
					(0.14)	(0.05)		
561	1,032	17,782	0.002	0.26	3.3	1.8	0.05	0.04
					(0.22)	(0.13)		

Notes: Reproduced estimates. Bootstrapped standard errors are in parentheses. This table corresponds to Table 1 in Alvarez et al. (2022) page 1014, presented in the same format.

Table 2: Reproduced Moments Taken from the Data and the Estimated Parameters

Category	$\hat{\gamma}_{11}$	$\hat{\gamma}_{21}$	$\hat{\gamma}_{31}$	$\hat{\gamma}_{32}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1/\hat{\beta}_2$	$\hat{\omega}$	$\hat{\alpha}_{22}$
111	1.248	1.406	1.507	1.787	2.099	12.190	228.677	0.161	4.245
119	1.282	1.507	1.702	2.381	1.058	6.021	91.439	0.109	3.748
1212	1.242	1.476	1.786	2.930	0.599	3.873	73.414	0.000	4.151
122	1.243	1.397	1.508	1.903	1.848	9.779	173.048	0.131	4.449
118	1.289	1.539	1.777	2.552	3.123	9.836	0.628	0.580	3.606
117	1.281	1.511	1.721	2.484	0.967	5.442	84.154	0.089	3.802
561	1.216	1.394	1.586	2.271	0.998	5.783	103.470	0.031	4.784

Notes: Reproduced estimates. This table corresponds to Table 2 in Alvarez et al. (2022) Online Appendix page 23, presented in the same format.

Table 3: Reproduced Additional Statistics

Category	Skewness	Kurtosis	Kurtosis ($t = 1, 2$)	Implied Correlation	$CV(\Delta\hat{p}_{it})$
111	-0.121	1.656	1.426	0.44	1.555
		(0.068)	(0.062)		
119	0.011	1.955	1.288	0.339	1.683
		(0.049)	(0.042)		
1212	-0.020	2.052	1.710	0.284	1.589
		(0.167)	(0.183)		
122	-0.025	1.677	1.189	0.39	1.398
		(0.056)	(0.020)		
118	-0.012	2.044	1.663	0.295	1.62
		(0.117)	(0.149)		
117	-0.004	1.989	1.422	0.303	1.577
		(0.048)	(0.090)		
561	-0.006	1.778	1.403	0.374	1.524
		(0.133)	(0.065)		

Notes: Reproduced estimates. This table corresponds to Table 3 in Alvarez et al. (2022) Online Appendix page 24, presented in the same format. Bootstrapped standard errors are in parentheses. Kurtosis estimates in column 3 is the same kurtosis estimates for pooled observation in Table 1.

Table 4: Differences between Original and Reproduced Estimates

	Table 1		Table 2			Table 3		
Category	Kurtosis pooled std. dev.	Kurtosis w/unobs. heterog. std. dev.	$\hat{\gamma}_{32}$	$\hat{\alpha}_2$	$\hat{\alpha}_{22}$	Kurtosis	Kurtosis std. dev.	Kurtosis ($t = 1, 2$) std. dev.
111	0.00	0.00	0.000	0.000	0.003	0.000	-0.003	0.009
119	-0.01	0.00	0.000	-0.009	-0.001	0.000	0.001	0.000
1212	-0.01	-0.01	-0.007	0.000	0.000	-0.001	-0.005	0.003
122	0.00	-0.01	0.000	0.000	0.011	0.000	-0.005	-0.001
118	0.01	0.00	0.000	0.000	0.004	0.000	0.001	0.001
117	0.00	0.00	0.000	0.000	-0.001	0.000	-0.001	-0.001
561	0.00	0.00	0.000	0.000	-0.002	0.000	0.000	0.001

Notes: This table includes all discrepancies we found between those reported in Alvarez et al. (2022) and those we reproduced in Table 1, 2, and 3. Std. dev. refers to standard deviation estimated by bootstrapping.

Table 5: Summary on Research Reproducibility

Title	Location	Reproducible	Note
Table 1	Page 1014	Partial	Discrepancies in standard errors for kurtosis
Figure 3	Page 1015	No	Not included in the replication program
Table 2	Online Appendix page 23	Partial	Discrepancies in parameter estimates
Table 3	Online Appendix page 24	Partial	Discrepancies in standard errors for kurtosis
Figure 6	Online Appendix page 22	Yes	

Notes: Online Appendix refers to the online appendix available at the QJE web site as Supplementary data.

Table 6: Results with Different Number of Bootstrapping Repetitions

Std. dev. of kurtosis (pooled)						
Category	Alvarez et al. (2022)	20%	50%	100%	200%	500%
111	0.16	0.16	0.16	0.16	0.16	0.16
119	0.09	0.10	0.10	0.10	0.10	0.10
1212	0.27	0.27	0.28	0.28	0.27	0.27
122	0.09	0.09	0.09	0.09	0.09	0.09
118	0.24	0.23	0.23	0.23	0.23	0.23
117	0.14	0.13	0.14	0.14	0.13	0.14
561	0.22	0.20	0.22	0.22	0.21	0.21

Std. dev. of kurtosis w/unobs. heterog.						
Category	Alvarez et al. (2022)	20%	50%	100%	200%	500%
111	0.07	0.06	0.07	0.07	0.07	0.07
119	0.05	0.05	0.05	0.05	0.05	0.05
1212	0.16	0.16	0.17	0.17	0.17	0.16
122	0.05	0.05	0.05	0.06	0.05	0.05
118	0.12	0.11	0.11	0.12	0.11	0.11
117	0.05	0.05	0.05	0.05	0.05	0.05
561	0.13	0.12	0.13	0.13	0.13	0.13

Note: Standard errors for the kurtosis estimates are referring to those in Table 1. The authors' original computation program used 1,000 repetitions for their bootstrap estimations.

Table 7: Replication Results Using Cavallo's (2018) Data Cleaning Criteria

Category	Number products	Number p. changes	$\hat{\mathbb{E}}(\Delta p_{it})$	$\hat{\sigma}(\Delta p_{it})$	Kurtosis pooled	Kurtosis w/unobs. heterog.	C_{pooled}	C w/unobs. heterog.
111	3,411	72,691	0.002	0.31	2.4	1.6	0.07	0.06
					(0.02)	(0.02)		
119	3,210	54,736	0.002	0.29	2.7)	1.7	0.08	0.07
					(0.03)	(0.02)		
1212	2,536	30,169	0.000	0.24	2.6	1.5	0.07	0.06
					(0.04)	(0.03)		
122	1,399	26,579	0.002	0.31	2.3	1.6	0.09	0.07
					(0.03)	(0.02)		
118	1,367	29,741	0.003	0.28	2.4	1.7	0.06	0.05
					(0.04)	(0.03)		
117	1,153	20,627	0.007	0.28	2.7	1.7	0.08	0.06
					(0.05)	(0.02)		
561	1,025	17,663	0.002	0.25	2.6	1.5	0.05	0.04
					(0.06)	(0.04)		

Note: Price changes greater than 200% or smaller that -50% are dropped from the sample. Bootstrapped standard errors are in parentheses.