The Journal of Open Source Software

pydynpd: A Python package for dynamic panel model

Dazhong Wu[1¶], **Jian Hua**[1], **and Feng Xu**[1]

1 Department of Business Management, School of Business and Public Administration, University of the District of Columbia, USA ¶ Corresponding author

DOI: 10.21105/joss.04416

Software

- Review
- Repository
- Archive

Editor: Chris Hartgerink

Reviewers:

- @Athene-ai
- @mhu48

Submitted: 09 April 2022
Published: 07 March 2023

Summary

We present pydynpd, a Python package which implements all the features in dynamic panel model with GMM (general method of moments). These features include: (1) difference and system GMM, (2) one-step, two-step, and iterative estimators, (3) robust standard errors including the one suggested by (Windmeijer, 2005), (4) Hansen over-identification test, (5) Arellano-Bond test for autocorrelation, (6) time dummies, (7) allows users to collapse instruments to reduce instrument proliferation issue, and (8) a simple grammar for model specification. As far as we know, pydynpd is the first Python package that allows researchers to estimate dynamic panel model.

What distinguishes pydynpd from any other dynamic panel model packages is its innovative feature: the capability to search for models based on users' request, rather than just run the model specified by users as other packages do. To the best of our knowledge, there is no other econometric software/package that offers this feature, let alone dynamic panel model packages.

Statement of need

Over the past decade, dynamic panel model has become increasingly popular in empirical studies. For example, researchers use dynamic panel model to study the environmental impacts of climate change (Phillips, 2020) and COVID-19 (Anser et al., 2020; Oehmke et al., 2021). This is because many aspects of our social and natural systems are inherently dynamic, and the GMM methods proposed by Arellano & Bond (1991) and Blundell & Bond (1998) allow us to model the dynamics that traditional static panel models are not able to capture. Correspondingly, the growing popularity of dynamic panel model will stimulate demand for the related packages in open source programs such as R, Python, and Julia,

Statement of field

So far, there are several related packages in Stata and R. Stata is a commercial software, while existing R packages have some issues. For example, in our benchmark test R package panelvar (Sigmund & Ferstl, 2021) is more than 100 times slower than Stata package xtabond2 (Roodman, 2009). On the other hand, R package plm (Croissant & Millo, 2008) is fast enough, but it has calculation issue for system GMM. A third R package, pdynmc, crashed or refused to work several times in our tests. Due to these reasons, R packages above are far less popular than xtabond2, according to citations they have received.

Moreover, there is no Python or Julia package yet to estimate dynamic panel model due to the complexity involved in implementation. Our package contributes to the open source community because (1) it implements all of the major features in the associated commercial packages in Stata, (2) its innovative feature (as mentioned above) will stimulate similar or even

more revolutionary features in the empirical computing community, and (3) though Python is interpreted, our package is almost as fast as xtabond2 which was compiled as shown in figure below. This package will increase the usability of open source software in estimating dynamic panel models, because for a package to be attractive, it must be both accurate and fast. Moreover, unlike existing R packages which rely heavily on R-specific components (that is a main reason they are not fast), our code uses components common to any programming language, making it easy to translate to R or Julia.

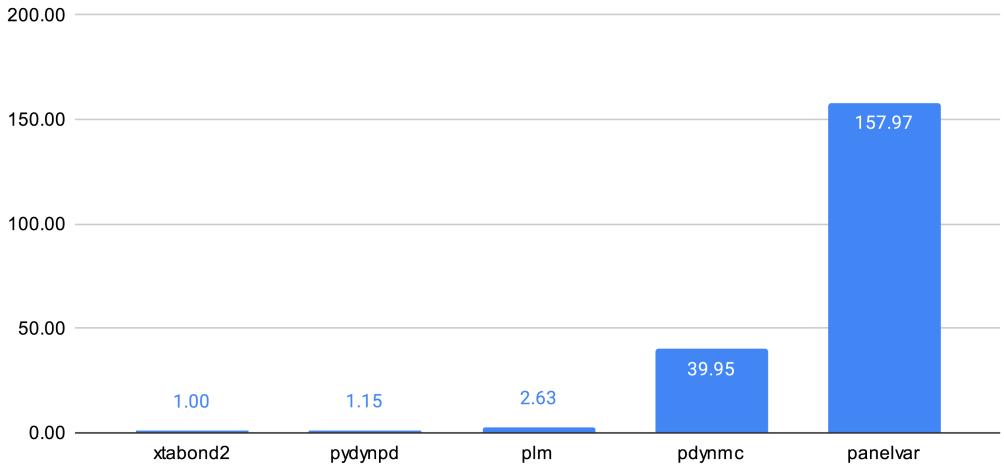

Figure 1: Running time (relative to the fastest).

The pydynpd package

pydynpd is able to estimate the most complicated linear dynamic panel models:

$$y_{it} = \sum_{j=1}^{p} \alpha_j y_{i,t-j} + \sum_{k=1}^{m} \sum_{j=0}^{q_k} \beta_{jk} r_{i,t-j}^{(k)} + \delta d_{i,t} + \gamma s_{i,t} + u_i + \epsilon_{it}$$

In the model above, $y_{i,t-j}$ $(j = 1, 2, \ldots, p)$ denotes a group of p lagged dependent variables. $r_{i,t-j}^{(k)}$ represents a group of m endogenous variables other than lagged y. d_{it} is a vector of predetermined variables which may potentially correlate with past errors, s_{it} is a vector of exogenous variables, and u_i represents fixed effect. As lagged dependent variables such as $y_{i,t-1}$ are included as regressors, the popular techniques in static panel models no longer produce consistent results. Researchers have developed many methods to estimate dynamic panel models. Essentially there are two types of GMM estimates, difference GMM and system GMM. Just like other R and Stata packages, pydynpd fully implements these two methods.

Due to space limit, we focus here on general discussion of the package. A detailed statistical/technique description of our package is available on GitHub.

For illustration purpose, consider the following equation:

$$y_{it} = \sum_{j=1}^{\text{p}} \alpha_j y_{i,t-j} + \sum_{j=1}^{\text{q}_k} \beta_j r_{i,t-j} + \delta d_{i,t} + \gamma_{i,t} + u_i + \epsilon_{it}$$

The equation above is related to a group/family of models with different combinations of p and q_k values. Unless existing economic theory indicates exactly what model to choose,

researchers need to guess and try the values of p and q_k as highlighted in equation above. For example, if $p = 2$ and $q_k = 1$, then a specific model is formed:

$$y_{it} = \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-2} + \beta_j r_{i,t-j} + \delta d_{i,t} + \gamma_{i,t} + u_i + \epsilon_{it}$$

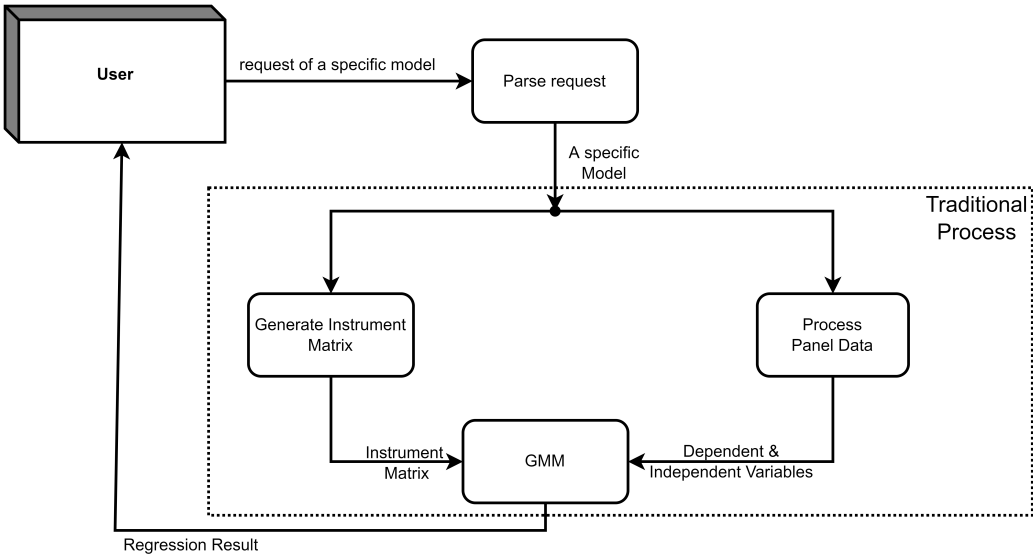

Figure 2: How alternative packages work.

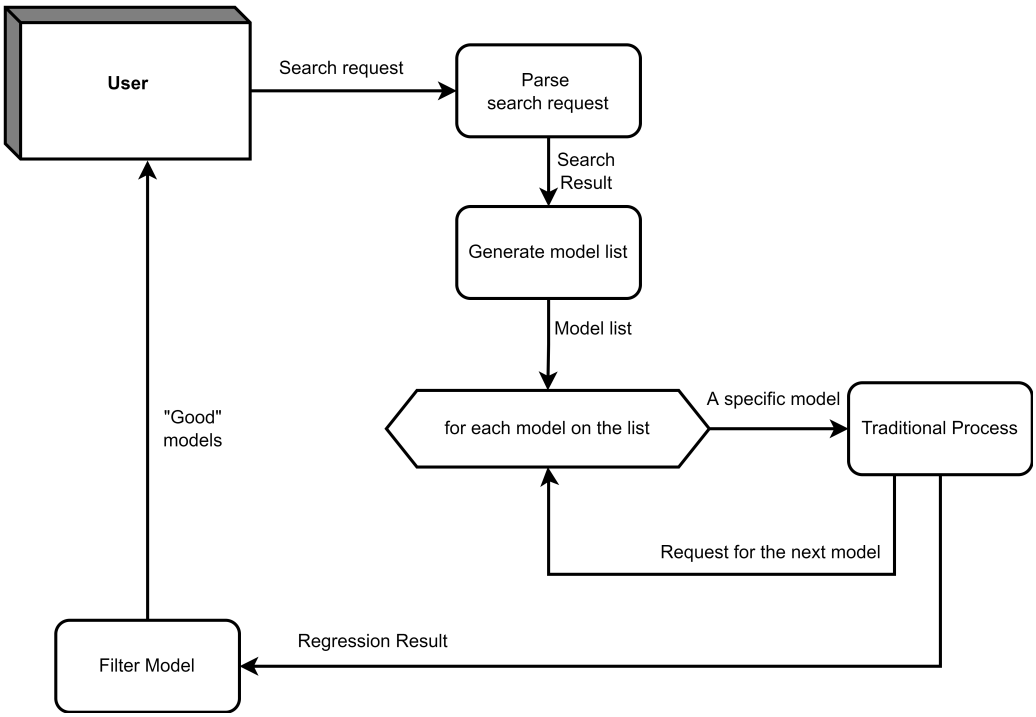

Figure 3: The automatic mode in pydnynpd.

Figure 2 shows how other packages work: a user needs to choose a specific model, then based on that particular model the system generates the corresponding instrument matrix and panel

data with dependent/independent variables so that the GMM process can produce regression results. An innovative feature of pydynpd is that it can also run in its "automatic" mode in which it doesn't require users to choose a particular model. Instead, users may let pydynpd search for the lags (e.g., p and q_k) so that the corresponding models satisfy certain standards. In other words, users may use pydynpd to estimate the following model with question markers indicating values not determined yet:

$$y_{it} = \sum_{j=1}^{?} \alpha_j y_{i,t-j} + \sum_{j=1}^{?} \beta_j r_{i,t-j} + \delta d_{i,t} + \gamma_{i,t} + u_i + \epsilon_{it}$$

Figure 3 shows how pydynpd's automatic mode works: a user indicates what values pydynpd needs to search for (e.g., the question marks in equation above), and then pydynpd tries all possible models, and returns "good" models that pass dynamic models' specification tests (e.g., Hansen overidentification test and AR(2) test). Note that processes included in the dotted box in Figure 2 is represented as a black-box process named "traditional process" in Figure 3.

References

Anser, M. K., Yousaf, Z., Khan, M. A., Nassani, A. A., Alotaibi, S. M., Abro, M. M. Q., Vo, X. V., & Zaman, K. (2020). Does communicable diseases (including COVID-19) may increase global poverty risk? A cloud on the horizon. *Environmental Research*, *187*, 109668. https://doi.org/10.1016/j.envres.2020.109668

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277–297. https://doi.org/10.2307/2297968

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115–143. https://doi.org/10.1016/S0304-4076(98)00009-8

Croissant, Y., & Millo, G. (2008). Panel data econometrics in r: The plm package. *Journal of Statistical Software*, *27*(2). https://doi.org/10.18637/jss.v027.i02

Oehmke, T. B., Post, L. A., Moss, C. B., Issa, T. Z., Boctor, M. J., Welch, S. B., & Oehmke, J. F. (2021). Dynamic panel data modeling and surveillance of COVID-19 in metropolitan areas in the united states: Longitudinal trend analysis. *Journal of Medical Internet Research*, *23*(2), e26081. https://doi.org/10.2196/26081

Phillips, P. C. B. (2020). Dynamic panel modeling of climate change. *Econometrics*, *8*(3). https://doi.org/10.3390/econometrics8030030

Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in stata. *The Stata Journal*, *9*(1), 86–136. https://doi.org/10.1177/1536867X0900900106

Sigmund, M., & Ferstl, R. (2021). Panel vector autoregression in r with the package panelvar. *The Quarterly Review of Economics and Finance*, *80*, 693–720. https://doi.org/10.1016/j.qref.2019.01.001

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, *126*(1), 25–51. https://doi.org/10.1016/j.jeconom.2004.02.005