The Journal of Open Source Software

IDCeMPy: Python Package for Inflated Discrete Choice Models

Nguyen K. Huynh[1], **Sergio Béjar**[2], **Vineeta Yadav**[1], **and Bumba Mukherjee**[1]

1 Dept. of Political Science, Pennsylvania State University **2** Dept. of Political Science, San Jose State University

DOI: 10.21105/joss.03322

Software

- Review
- Repository
- Archive

Editor: Gabriela Alessio Robles

Reviewers:

- @cmaimone
- @jungtaekkim
- @tmickleydoyle

Submitted: 19 February 2021
Published: 20 July 2021

Summary

Scholars and Data Scientists often use discrete choice models to evaluate ordered dependent variables using the ordered probit model and unordered polytomous outcome measures via the multinomial logit (MNL) estimator (Greene, 2002; Richards & Bonnet, 2018; Sarrias, 2016). These models, however, cannot account for the possibility that in many ordered and unordered polytomous choice outcomes, a disproportionate share of observations —stemming from two distinct data generating processes (d.g.p's)— fall into a single category which is thus "inflated." For instance, ordered outcome measures of self-reported smoking behavior that range from 0 for "no smoking" to 3 for "smoking 20 cigarettes or more daily" contain excessive observations in the zero (no smoking) category that includes individuals who never smoke cigarettes and those who smoked previously but temporarily stop smoking because of an increase in cigarette costs (Greene et al., 2015; Harris & Zhao, 2007). The "indifference" middle-category in ordered measures of immigration attitudes is inflated since it includes respondents who are genuinely indifferent about immigration and those who select "indifference" because of social desirability reasons (Bagozzi & Mukherjee, 2012; Brown et al., 2020). The baseline category of unordered polytomous variables of Presidential vote choice is also often inflated as it includes non-voters who abstain from voting owing to temporary factors and routine non-voters who are disengaged from the political process (Bagozzi & Marchetti, 2017; Campbell & Monson, 2008). Inflated discrete choice models have been developed to address such category inflation in ordered and unordered polytomous outcome variables as failing to do so leads to model misspecification and incorrect inferences (Bagozzi & Mukherjee, 2012; Brown et al., 2020; Harris & Zhao, 2007).

`IDCeMPy` is an open-source Python package that enables researchers to fit three distinct sets of discrete choice models used by Data Scientists, Economists, Engineers, Political Scientists, and Public Health researchers: the Zero-Inflated Ordered Probit (ZiOP) model without and with correlated errors (ZiOPC model), Middle-Inflated Ordered Probit (MiOP) model without and with correlated errors (MiOPC), and Generalized-Inflated Multinomial Logit (GiMNL) models. Functions that fit the ZiOP(C) model in `IDCeMPy` evaluate zero-inflated ordered dependent variables that result from two d.g.p's, while functions that fit the MiOP(C) models account for inflated middle-category ordered outcomes that emerge from distinct d.g.p's. The functions in `IDCeMPy` that fit GiMNL models account for the large share and heterogeneous mixture of observations in the baseline and other lower outcome categories in unordered polytomous dependent variables. The primary location for the description of the functions that fit the models listed above is available at the IDCeMPy package's documentation website.

State of the Field

Software packages and code are available for estimating standard (non-inflated) discrete choice models. In the R environment, the packages `MASS` (Venables & Ripley, 2002) and `micEcon` (Henningsen, 2014) fit binary and discrete choice models. The package `Rchoice` (Sarrias, 2016) allows researchers to estimate binary and ordered probit and logit models as well as the Poisson model by employing various optimization routines. The proprietary LIMDEP package NLOGIT (Greene, 2002) fits conventional binary and ordered discrete choice models but is neither open-sourced nor freely available. The R packages `mlogit` (Croissant, 2012) and `mnlogit` (Hasan et al., 2016) provide tools for working with conventional MNL models, while `gmnl` (Sarrias et al., 2017) and `PReMiuM` (Liverani et al., 2015) estimate MNL models that incorporate unit-specific heterogeneity. There exists proprietary LIMDEP software and R code —but not an R package— that fit few inflated ordered probit and MNL models (Bagozzi & Marchetti, 2017; Bagozzi & Mukherjee, 2012; Harris & Zhao, 2007). Outside R, the Python package `biogeme` (Bierlaire, 2016) fits mixed logit and MNL models. Further, Dale & Sirchenko (2021)'s ZiOP STATA command (but not package) fits the Zero-Inflated Ordered Probit without correlated errors. Xia et al. (2019)'s `gidm` STATA command fits discrete choice models without correlated errors for inflated zero and other lower-category discrete outcomes.

The R or LIMDEP software, along with the STATA commands listed above, are undoubtedly helpful, however to our knowledge, there are no R or Python packages to fit a variety of statistical models that account for the excessive (i.e., "inflated") share of observations in the baseline, and other higher categories of ordered and unordered polytomous dependent variables, which are commonly analyzed across the natural and social sciences. As discussed below, our Python package `IDCeMPy` thus fills an important lacuna by providing an array of functions that fit a substantial range of inflated discrete choice models applicable across various disciplines.

Statement of Need

Although our `IDCeMPy` package also fits standard discrete choice models, what makes it unique is that unlike existing software, it offers functions to fit and assess the performance of both Zero-Inflated and Middle-Inflated Ordered Probit (OP) models without and with correlated errors as well as a set of Generalized-Inflated MNL models. The models included in `IDCeMPy` account for the excessive proportion of observations in any given ordered or unordered outcome category by combining a single binary probit or logit split-stage equation with either an ordered probit outcome stage (for the Zero and Middle-Inflated OP models) or an MNL outcome-stage equation. Users can treat the error terms from the two equations in the Zero and Middle-Inflated OP models as independent or correlated in the package's estimation routines. `IDCeMPy` also provides functions to assess each included model's goodness-of-fit via the AIC statistics, extract the covariates' marginal effects from each model, and conduct Vuong tests for comparing the performance between the standard and inflated discrete choice models.

The functions in `IDCeMPy` use quasi-Newton optimization methods such as the Broyden-Fletcher-Goldfarb-Shanno algorithm for Maximum-Likelihood-Estimation (MLE), which facilitates convergence and estimation speed. Another feature is that the coefficients, standard errors, and confidence intervals obtained for each model estimated in `IDCeMPy` are in `pandas.DataFrame` (McKinney, 2010) format and are stored as class attribute `.coefs`. This allows for easy export to csv or excel, which makes it easier for users to perform diagnostic tests and extract marginal effects. `IDCeMPy` is thus essential as it provides a much-needed unified software package to fit statistical models to account for category inflation in several ordered and unordered outcome variables used across fields as diverse as Economics, Engi-

neering, Marketing, Political Science, Public Health, Sociology, and Transportation research. Users can employ the wide range of statistical models in `IDCeMPy` to assess:

- Zero-inflation in self-reported smoking behavior (Harris & Zhao, 2007), demand for health treatment (Greene et al., 2015), and accident injury-severity (Fountas et al., 2018).
- Middle-category inflation in ordered measures of monetary policy (Brown et al., 2020) and European Union (EU) membership attitudes (Elgün & Tillman, 2007).
- Inflated unordered polytomous outcomes such as transportation choice, environmental policy and consumer demand (Richards & Bonnet, 2018), and Presidential vote choice (Campbell & Monson, 2008).

Functionality and Applications

`IDCeMPy` contains the functions listed below to estimate via MLE the following inflated discrete choice models listed earlier:

- `opmod`; `iopmod`; `iopcmod`: Fits the ordered probit model, the Zero-Inflated (ZIOP) and Middle-Inflated ordered probit (MIOP) models without correlated errors, and the ZIOPC and MIOPC models that incorporate correlated errors.
- `opresults`; `iopresults`; `iopcresults`: Presents covariate estimates, Variance-Covariance (VCV) matrix, Log-Likelihood, and AIC statistics of the object models.
- `iopfit`; `iopcfit`: Computes fitted probabilities from each estimated model's objects.
- `vuong_opiop`; `vuong_opiopc`: Calculates Vuong test statistic for comparing the performance of the OP with the ZiOP(C) and MiOP(C) models.
- `split_effects`; `ordered_effects`: Estimates marginal effects of covariates in the split-stage and outcome-stage respectively.
- `mnlmod`; `gimnlmod`: Fits MNL model and Generalized-Inflated MNL models.
- `mnlresults`; `gimnlresults`; `vuong_gimnl`: Presents covariate estimates, VCV matrix, Log-Likelihood, and AIC statistics of `mnlmod`; `gimnlmod`. Vuong test statistic for comparing MNL to GIMNL models obtained from `vuong_gimnl`.

Details about the functionality summarized above are available at the package's documentation website, which is open-source and hosted by ReadTheDocs. The features of the functions in `IDCeMPy` that fit the,

(i) ZiOP(C) models are presented using the ordered self-reported tobacco consumption dependent variable from the 2018 National Youth Tobacco Dataset,

(ii) MiOP(C) models are illustrated using the ordered EU support outcome variable from Elgün & Tillman (2007).

(iii) GiMNL models are evaluated using the unordered polytomous Presidential vote choice dependent variable from Campbell & Monson (2008).

Availability and Installation

IDCeMPy is an Open-source software made available under the GNU General Public License. It can be installed from PyPI or from its GitHub repository.

References

Bagozzi, B. E., & Marchetti, K. (2017). Distinguishing occasional abstention from routine indifference in models of vote choice. *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2015.42

Bagozzi, B. E., & Mukherjee, B. (2012). A mixture model for middle category inflation in ordered survey responses. *Political Analysis*, 369–386. https://doi.org/10.1093/pan/mps020

Bierlaire, M. (2016). PythonBiogeme: A short introduction. Report TRANSP-OR 160706, series on biogeme. *Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Switzerland*.

Brown, S., Harris, M. N., & Spencer, C. (2020). Modelling category inflation with multiple inflation processes: Estimation, specification and testing 1. *Oxford Bulletin of Economics and Statistics*, *82*(6), 1342–1361. https://doi.org/10.1111/obes.12366

Brown, S., Harris, M. N., & Spencer, C. (2020). Modelling category inflation with multiple inflation processes: Estimation, specification and testing 1. *Oxford Bulletin of Economics and Statistics*, *82*(6), 1342–1361. https://doi.org/10.1111/obes.12366

Campbell, D. E., & Monson, J. Q. (2008). The religion card: Gay marriage and the 2004 presidential election. *Public Opinion Quarterly*, *72*(3), 399–419. https://doi.org/10.1093/poq/nfn032

Croissant, Y. (2012). Mlogit: Multinomial logit models. *R Package Version 0.2-2*. https://cran.r-project.org/package=mlogitl

Dale, D., & Sirchenko, A. (2021). Estimation of nested and zero-inflated ordered probit models. *The Stata Journal*, *21*(1), 3–38. https://doi.org/10.1177/1536867X211000002

Elgün, Ö., & Tillman, E. R. (2007). Exposure to european union policies and support for membership in the candidate countries. *Political Research Quarterly*, *60*(3), 391–400. https://doi.org/10.1177/1065912907305684

Fountas, G., Anastasopoulos, P. C., & Abdel-Aty, M. (2018). Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. *Analytic Methods in Accident Research*, *18*, 57–68. https://doi.org/10.1016/j.amar.2018.04.003

Greene, W. H. (2002). *NLOGIT: Version 3.0; reference guide*. Econometric Software, Incorporated.

Greene, W. H., Harris, M. N., & Hollingsworth, B. (2015). Inflated responses in measures of self-assessed health. *American Journal of Health Economics*, *1*(4), 461–493. https://doi.org/10.1162/ajhe_a_00026

Harris, M. N., & Zhao, X. (2007). A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics*, *141*(2), 1073–1099. https://doi.org/10.1016/j.jeconom.2007.01.002

Hasan, A., Zhiyu, W., & Mahani, A. S. (2016). Fast estimation of multinomial logit models: R package mnlogit. *Journal of Statistical Software*, *75*(3), 1–24. https://doi.org/10.18637/jss.v075.i03

Henningsen, A. (2014). micEcon: Microeconomic analysis and modelling. *R Package Version 0.6-12*. https://cran.r-project.org/package=micEcon

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., & Richardson, S. (2015). PReMiuM: An R package for profile regression mixture models using dirichlet processes. *Journal of Statistical Software*, *64*(7), 1. https://doi.org/10.18637/jss.v064.i07

McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Richards, T. J., & Bonnet, C. (2018). New empirical models in consumer demand. In *The routledge handbook of agricultural economics* (pp. 488–511). Routledge. https://doi.org/10.4324/9781315623351-27

Richards, T. J., & Bonnet, C. (2018). New empirical models in consumer demand. In *The routledge handbook of agricultural economics* (pp. 488–511). Routledge. https://doi.org/10.4324/9781315623351-27

Sarrias, M. (2016). Discrete choice models with random parameters in R: The Rchoice package. *Journal of Statistical Software*, *74*(10), 1–31. https://doi.org/10.18637/jss.v074.i10

Sarrias, M., Daziano, R., & others. (2017). Multinomial logit models with continuous and discrete individual heterogeneity in R: The gmnl package. *Journal of Statistical Software*, *79*(2), 1–46. https://doi.org/10.18637/jss.v079.i02

Venables, W., & Ripley, B. (2002). Random and mixed effects. In *Modern applied statistics with s* (pp. 271–300). Springer. https://doi.org/10.1007/978-0-387-21706-2

Xia, Y., Zhou, Y., & Cai, T. (2019). Gidm: A command for generalized inflated discrete models. *The Stata Journal*, *19*(3), 698–718. https://doi.org/10.1177/1536867X19874246