Generalized Roy Model

– Problem Set 4 –

In this problem set, we are going to make use of the open-source Python package `grmpy` to gain a practical understanding of the generalized Roy model and the economics behind it. We will employ `grmpy` for both simulation and estimation purposes and, in particular, estimate *marginal treatment effects* (MTE). To become familiar with the capabilities of `grmpy`, we will explore the relationship between college education and income using a toy data set in Task A. In Task B, we will then conduct a Monte Carlo analysis using our own simulated data. Auxiliary functions for Task B are provided in the file `auxiliary.py`.

Task A

Suppose we want to estimate the effect of post-secondary education on income. We use a toy model where log earnings (Y) is a function of experience (exp), experience squared ($expsq$), and mother's education ($momsEdu$). The decision to enroll in college (D) is modeled as a function of mother's education and distance to college ($distCol$).

(1) Open a `Jupyter Notebook` and import the data set `education-data-raw.pkl` — we recommend using the software presented in class (such as `pandas` and `seaborn`). Before analyzing the data and estimating the marginal treatment effect of college attendance, you need to clean the data set. Drop missing values and plot the wage distribution. *Are there any wage outliers?* If so, drop them. When you are done, save the data set as a new `pickle` file.

(2) Open the initialization file `education.grmpy.yml` and put the name of the cleaned data set you just saved in the blank after `["ESTIMATION"]["file"]:` (Hint: You do not need to do this in your Jupyter Notebook. You can use a simple Text Editor, such as `Sublime`, instead).

(3) You are now ready to estimate the marginal treatment effect of college attendance! `grmpy` offers two ways to do so. First, a parametric normal model. And second, the semiparametric method of *local instrumental variables* (LIV) (Heckman, Urzua, & Vytlacil, 2006). Using both models, estimate the MTE and plot it, while choosing different colors for the parametric and semiparametric plots, respectively. Note that for the semparametric model, the number of bootstrap iterations needs to be specified when plotting the MTE. Choose a suitable number.

Task B

To gain more insights into the economic implications behind `grmpy` and the realm of *marginal treatment effects*, you will now perform a simulated Monte Carlo analysis. Simulated data allows us to explore additional objects of interest, for which we are not able to obtain reliable information using empirical data sets. With simulated data, we have information on the whole range of potential outcomes for each individual and can bypass the evaluation problem. Hence, we can compute the joint distribution of potential outcomes. Moreover, we can directly construct conventional treatment effect parameters on the population level (e.g. see chapters 2.3, 2.4, and 3.3 in Heckman & Vytlacil (2007a)).

We will stick with our college example from above. For simplicity, however, assume now that log earnings (Y) only depend on mother's education ($momsEdu$), and the enrollment decision is solely driven by distance to college ($distCol$). To help you along with the analysis, we have set up some auxiliary functions in `auxiliary.py`.

(1) Simulate a data set based on the intialization file `sim.grmpy.yml`.

(2) Based on this simulated data set, compute the *average treatment effect* (ATE), the *average treatment effect on the treated* (TT), and the *average treatment effect on the untreated* (TUT) by hand (Note: there are no auxiliary functions required so far). *How does the ATE compare with the other treatment effect parameters? What does this imply for the marginal treatment effect?*

(3) Plot the distribution of college-related wage benefits ($Y_1 - Y_0$). *Hint:* You can verify your results in **2)** by plotting the effects along with the distribution of benefits.

We now introduce essential heterogeneity to our simulated sample.

(4) Simulate a new sample based on `sim-eh.grmpy.yml`. *How is essential heterogeneity reflected in the distribution of unobservables?* Compare the two simulated data sets.

(5) *Do the conventional treatment effect parameters differ now?* Construct them again as you did in **(2)**.

(6) Investigate the shape of the MTE with and without essential heterogeneity. *What does this imply for our toy model of the monetary effect of college attendance? Which individuals are more inclined to select into college?*

(7) Let us now begin with our Monte Carlo exploration! Simulate a new sample based on `mc.grmpy.yml`. Note that the correlation between U_1 and V (`["DIST"]["params"][2]`) is still zero, i.e. no essential heterogeneity is present so far. *But how do the ATE and TT diverge when the (absolute) correlation between U_1 and V is gradually increased?* Note that we consider the absolute value of the correlation coefficient. In the example here, the correlation becomes actually more negative.

(8) For three different estimation methods (assumption of random assignment, Ordinary Least Squares, `Instrumental Variables`) do the following:

a) Compute the ATE without essential heterogheneity using the data you simulated in **7)**.

b) Plot the estimated ATE in comparison with the underyling population ATE for incremental positive selection into treatment.

Note:

- Random assignment implies a simple comparison of average outcomes between treated and untreated individuals.
- For the OLS setup, choose a simple regression of Y on D plus a constant. The coefficient on D then captures the ATE.
- In your `IV/2SLS` setting, use the full model. *What is the endogeneous regressor? What is the instrument?*

Additionally, use `grmpy` to do step **b)** for both the paramteric and semiparametric MTE approach. Compare wih the estimation results above (for a discussion see chapter 3.1 in Heckman & Vytlacil (2007b)).

(9) Plot the joint distribution of potential outcomes.

References

grmpy. (2018). *grmpy: A Python package for the simulation and estimation of the generalized Roy model.* Retrieved from `http://doi.org/10.5281/zenodo.1162639`

Heckman, J. J., Urzua, S., & Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics*, *88*(3), 389–432.

Heckman, J. J., & Vytlacil, E. J. (2007a). Econometric evaluation of social programs, part I: Causal effects, structural models and econometric policy evaluation. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6B, pp. 4779–4874). Amsterdam, Netherlands: Elsevier Science.

Heckman, J. J., & Vytlacil, E. J. (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6B, pp. 4875–5144). Amsterdam, Netherlands: Elsevier Science.