I’m interested in building a Python package which will be helpful for easily using econometrics tools. Currently, most economists conduct their analysis using packages in either Stata or R. Unfortunately, Stata is behind a paywall and both of them, while great, are languages meant only for data analysis (rather than a ‘traditional’ programming language). This naturally means both of these softwares have limitations; for example, economists still do all their web-scraping and a bunch of machine learning methods via Python. My project aims to bridge this issue by building a package in Python which can serve as an alternative to these software’s.

To see if I’m successful, I hope to test my package by seeing if you can type in Stata commands and successfully get the code to implement that command in Python. I’d love to implement every single Stata command, but to be honest Stata has many niche but helpful commands for me to implement them all in 1 year. Realistically, I’ll be happy if I can build a package which implements all the basic econometrics tools (i.e., all the tools that will be covered in a class like econ 140) in Python. As a reach goal, I’d love to implement all the commands covered in the first page of the Stata cheatsheet, but I understand that goal may be tough.

It is hard to categorize what data techniques exactly I will use because I’m not sure exactly what it’ll take to accomplish my goal, but I will be creating a package to do data processing so I’m sure I’ll end up using some data techniques along the way.

For dataset ideas, I could just look at popular datasets for teaching econometrics with Stata and reimplement their analysis in Python. I will likely use multiple datasets, and at the end, it’d be interesting to do a comparison of processing times with my package vs in Stata.

I would really appreciate it if Dr. Eric Van Dusen could be my advisor. Dr. Van Dusen has been really pioneering classes that aim to combine both economics and data science in recent years. He is very knowledgeable about both fields and will have very relevant advice for my thesis. He is also among the very few faculty who are well connected in both the economics and data science departments, so he will have relevant advice for me from both departments.

I aim to get a PhD in economics and have worked as a research assistant in economics for the past two years, which includes me writing a thesis in econ 191: research methods in economics last semester. While I’m not sure which dataset I will be using exactly, I am familiar with econometric research tools in general.

I already mentioned I seek a PhD in economics, so this research is directly relevant to my intended career. My (soon-to-be) domain emphasis is applied mathematics. I chose this domain emphasis because economics graduate school requires plenty of mathematics classes in order to succeed, and so it was a natural fit. While I’m not sure right now how my domain emphasis will explicitly tie in to my project, I’m sure it’ll find a way somehow – after all, you do need a strong mathematical basis to understand a lot of the statistical tests.

I think there are two major goals my project needs to achieve to ensure I adequately cover human contexts and ethics. For one, I need to ensure the outputs of my packages are correct. If my package has an incorrect output in even one case, that would compromise the integrity of my data and if a policy decision is based off that, it could potentially have disastrous consequences. Secondly, I need to ensure I word the outputs of my package very thoughtfully to ensure I don’t make any claims which cannot be 100% backed up by statistics. For example, if I’m reporting a confidence interval, I cannot say there is a 95% chance the true population parameter is in the interval. I need to be more careful and instead say if many, many confidence intervals were made, the true population parameter would be in 95% of them. These seemingly small statistical nuances can have a big impact on how the results are interpreted, which could then have large unintended consequences on public policy.