DSA3361 - Individual Report

A0219739N - Le Van Minh

November 18, 2022

1 Overview and Peer Evaluation

1.1 Project Development

The project started with many difficulties stemming from the lack of information and direction. Back then we have little knowledge of regression, so deciding whether to choose simulation or regression was uneasy. Simulation seemed straightforward, easy to assign tasks, and even to fake data, but model-building could be tedious. Regression, on the other hand, was more difficult, but also more promising. We decided on regression for the wide range of possibilities it offered.

Many good sample projects were provided. The datasets were good, with many potentials, and guidance was given. However, we didn't want to rely on well-known datasets because we could gain little real experience from them. We wanted to challenge ourselves and went on to find an interesting problem to tackle.

We were looking on Kaggle for some highly-rated datasets and projects and found a post about MPI - Multidimensional Poverty Index. It was an active thread with many recommendations, and a provoking topic to pursue. We started working on the problem statement and quickly realised that the dataset is not enough to conduct a proper study. We needed more data, more variables, and more insights. To make sure we would not spend our effort pointlessly, I conducted some preliminary analysis before getting deeper into the study.

After the first few analysis we found out that the relationship between variables was not straightforward, MPI and expenditure had a delayed effect of approximately 6 years. That means we could only use 20% of the dataset, which was not a very big set. Furthermore, accounting for the delay will cost us about thrice as much time, and the return was not worth it. These findings had proven that MPI was not a mature topic for study, hence the limited publications we could find. We needed to change direction.

We decided to go switch the scope of the project to something broader. Traditional poverty indication was a good candidate for its proximity with the original plan, and the abundant guidance it could provide. We had more data, more variables, studies and other resources. Alongside them, there was a gold mine: missing values.

The problem that missing values created presented us with an unforeseen opportunity: autonomous work. By tackling the issue with a different approach, each member can build their model, and be independent of each other. Because everyone had to work on their models, free-riding on others' labour is also harder. By switching the topic of the project, we solved two major problems: limited resources and task assignments.

At the end of the project, our models were diverse in terms of power and interpretation. Some variables were agreed to be good predictors of poverty, but some had more arguable results. The most powerful models disagreed on many aspects, proving that how we approach missing values can drastically affect the conclusion of the study. This is a major outcome we got from this project, aside from the main problem statement. It also posed a question to me personally: How much did these decisions affect our understanding of the world?

1.2 Peer Evaluation

Callista is a hard-working and conscientious contributor. She started on her part early on and found many helpful information that contributed to others' work, pushing the project forwards. In discussions, she asks questions and pays attention, clarifies the scope of the projects and proves to be a good facilitator for our meetings/consultations. Her strongest point is knowing the right questions to ask, and it helps herself along with other teammates very much. She understands her work well and how to move forwards.

Imtiyaaz is not a strong technical, but a good teamworker. He knows how to cooperate with others and call for help when needed. He always makes sure he understand the assignment, and does his part well. He is a strong choice in any collaborative project.

Isaac is very hard to work with because he rarely speaks his thought. When facing difficulty, he would not ask for help or guidance. Consquently, he made very slow progress and his final results were hard to draw meaningful interpretation. We could not include his portion in the final conclusion and interpretation, as his model was incomplete and thorough procedure was not carried out. It is a disappointment we could not observe the proper outcome of this approach.

2 Technical Aspect

The project is where I could learn how to tackle problems in real-life datasets that the lectures were too limited to address. Missing values, normality, delay effects, etc. are some topics I encountered during the development of the projects. Thanks to my background as a programmer, I could navigate through technical issues quite comfortably and make faster progress, exploring new possibilities. Many discoveries were not used in the final project, however, I believe they can be good resources for subsequent studies and real-life scenarios.

Boxcox is a transformation used to make more normally distributed residuals linear regressions. I learnt how to use this technique when we were working on the MPI dataset. I realised that this assumption is quite hard to achieve especially in a complicated dataset like that, and appropriate transformation is not always obvious. Its performance was good for the MPI set but fell when we increased the number of variables and used the poverty headcount set. Although it was not included in the project for its relatively insignificant impact, it remains an important asset for future uses.

Imputation is one of our approaches to tackling missing values. It was not a hard concept to grasp or an algorithmic complication. However, using the imputed data cautiously and following the practice to minimise bias is hard work. I am not sure if I have done it right, because the data is partially artificial. However, I think it is an important idea to consider in the future and learning to use it effectively is very helpful.

Panel data analysis is another unused content. This technique is way outside of the class scope and needs a lot of time for practising to master. I have read multiple articles and lectures and figured this is the most suitable technique for our project. Combining this with imputation is one of the planned models, but it was cancelled due to time constraints. This allows us to account for entity-to-entity and time-to-time different in a way that pure linear regression cannot. It is a shame we could not include this in our project, as it was widely used in the research I studied.

They are more most noticeable contents. There are more I would like to cover, such as step-wise AIC, Goldfeld-Quandt test, ggplot2, etc. However, they are not as significant as those listed above.