**Part 3: Technical Report**

**David D. Bui**

**Grand Canyon University**

**DSC-540**

**Professor Aiman Darwiche**

**November 3rd, 2021**

**Part 2 Link for Python Code:**

**https://github.com/DouglasBui/GCU/tree/main/DSC-540/Assignment1**

Overviewing the following articles, “The Seven Tools of Causal Inferences, with Reflections on Machine Learning” and “Using Machine Learning to Translate Applicant Work History into Predictors of Performance and Turnover”, we can gather an in-depth understanding of an effective machine learning process that is applicable to fields of business, finance, social sciences, medicine, and economics. There is an intricate relationship weaving between the models described within these two articles that amplify one another and allowing for researchers to pull further understanding from data with great efficiency.

The first article imparts the importance of a well-articulated question when drafting models of statistical measure. It explains a three-tier hierarchy that depicts the range and extend certain questions can derive information depending on the way they are structured. It then places this hierarchy amid the structural causal model. This SCM relies on 7 tools that are implemented on the database that allow for a reasonable claim for causation. A claim that can face a discerning level of human criticism and understanding.

The second article involves a self-supervising machine learning model that can distinguish optimal applicants for job employment by expanding the range of variables and applying machine learning models. This exploit attempts to capture hidden information that is not quite easily quantified and at the same time allow for discernment in comparison. An example of this that they provided is the difference between an individual with 5 years of experience in childcare vs someone with 3 years of experience in corporate training. The analysis of these new variables involves reading into and discovering higher potential performance and turnover based off this extended work history. It is affirmed that machine learning has the capability of understanding even text line questions that involve reason for leaving previous employment and applicant behavior from tenure and work history. After examining the variables in play, it fixates on which model would best support the data. When evaluating the hiring process, it is important to note that the process does not involve a random selection. Which causes an issue when using traditional OLS regressions, since OLS estimates would suffer from sample selection bias. With this observation they surmised the Heckman Selection Correction to be the most suitable method to apply before performing Heckman regression.

The Heckman selection correction is a statistical technique that involves correcting the bias of when non-random sampling is performed. The staging of this process is first based on economic theory of probability of working. The Heckman regression model proceeds in two stages.

Stage 1:

*Note: Z represents the explanatory variables, are the unknown parameters, and D is part of a binary probability.*

The next stage involves a transformation of the predicted probabilities of explanatory variables. Baseline error terms and jointly normal assumptions are then evoked that allows for the transformation. The formula that is depicted uses the wage equation as an example.

Stage 2: Wage Equation = →

*Note: represents wage, ρ is the correlation between unobserved determinants of propensity to work, , is the inverse Mills ratio.*

Within the stage 2 process the equation shows how sample selection is a related form of omitted-variable bias. This process isn’t illustrated in detail inside of the article and just merely stated without explanation or presence of assumption, thus I thought it important to expand upon.

To add to the subject matter of learning algorithms, the back-propagating neural network is a unique approach to handling regression problems and could be an alternative to Heckman regression, while still allowing for a three-tier hierarchy and utilization of tools. This process orchestrates a layered system that allows for a complex and intricate webbing of connections of measured weight and bias. The network leads into a conclusion or even filtering process for optimization of desired grouping. It does require a non-random sampling technique to be employed properly beforehand though. It also needs to have a understanding of assumptions that make up each grouping or for example job positions. The application of this method would allow companies to multiple optimal job positions, while the Heckman model focused on a singular path. This network holds a greater value in quick utility and greater comparison for diversity and filtration.

**References**

Bastos, F. de S., Barreto-Souza, W., & Genton, M. G. (2020). *A Generalized Heckman Model With Varying Sample Selection Bias and Dispersion Parameters*. <https://doi-org.lopes.idm.oclc.org/10.5705/ss.202021.0068>

Olugbenga Omotayo Alabi, Ayoola Olugbenga Oladele, & Mohammed Bello Usman. (2021). Determinants of Agricultural Loan Decision Making Process for Rice (Oryza Sativa) Farmers in Abuja, Nigeria. Applications of Heckman Two-Stage Model and Factor Analysis. *Journal of Agribusiness and Rural Development*, *59*(1). <https://doi-org.lopes.idm.oclc.org/10.17306/J.JARD.2021.01381>

Huang, J.-C., Ko, K.-M., Shu, M.-H., & Hsu, B.-M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Computing & Applications*, *32*(10), 5461–5469. <https://doi-org.lopes.idm.oclc.org/10.1007/s00521-019-04644-5>