UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

An Application of a Machine Learning Technique in Microeconomics: Using a Neural Network to Enhance Prediction in the Service of Estimation in the Context of the South African Child Support Grant

A Research Report submitted in partial fulfilment of the Degree of
Master of Commerce (Economics)
in the School of Economics and Finance
University of the Witwatersrand

Kyle Wootton

Student Number: 669201

Supervised by Dr Adeola Oyenubi

Word Count: 12 835

(Inclusive of all components)

Candidate's Declaration

i. I hereby submit my Research Report in partial fulfilment of the degree of Master of Commerce (Economics);

ii. I confirm neither the substance nor any part of my Research Report has been submitted in the past or is being or is to be submitted for a qualification at any other university;

iii. The information used in the research report has not been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.

Kyle Wootton

7 September 2021

Acknowledgements

First, to my supervisor Dr Adeola Oyenubi, thank you for your guidance, patience and support throughout this process. Your intellectual curiosity and willingness to challenge conventional approaches is inspiring. I have learnt so much about a method I started out knowing very little about. It has been a long road and I could not have done it without you. Second, to Dr Richard Klein, thank you for generously giving of your time to help a student from outside of your faculty construct a neural network. Finally, Mom and Dad, words cannot ever do justice to everything you have done for me. I am grateful beyond measure. This is for you both.

Abstract

In this study, my aim is to show how machine learning can be used for prediction in the service of estimation in the context of a microeconomic research question: namely, whether the South African Child Support Grant ('the grant') improves the nutrition of children who receive it. Specifically, I show how a fully connected artificial neural network can be used as a novel, and potentially superior, approach to constructing an input variable in microeconomic research where the input variable is the result of high-dimension prediction. The hypothesis is that, if a neural network can be used to improve predictive performance when constructing the input variable, the estimation step that relies on this input as a covariate should also be more accurate. The input variable in question is caregiver motivation which is a covariate used as part of the identification strategy when determining the impact of the grant on nutrition. Caregiver motivation is constructed as the standardised difference between predicted application delay and actual application delay. Actual application delay is the number of days between the child becoming eligible for the grant and receiving the grant. Predicted application delay is the expected number of days between a child becoming eligible for the grant and receiving the grant given a set of observable characteristics. When comparing eligible children who receive the grant to eligible children who do not, I find the motivation variable constructed using a neural network results in the grant having a statistically significant impact on child nutrition – a result consistent with theoretical expectations and qualitative empirical evidence. In contrast, the motivation variable constructed using ordinary least squares (OLS) regression finds no statistically significant improvement in child nutrition from the grant – a result contrary to theoretical expectations and qualitative empirical evidence. As a result, I argue the neural network is a better predictor of application delay for the grant than the OLS regression.

Table of Content

"My standard advice to graduate students [of economics] these days is go to the computer science department and take a class in machine learning" – Hal R. Varian (2014: 3)

1. Introduction

A poor workman always blames his tools, or so the old adage goes. But what then of the workman who fails to use the tools available to him in the first place? Machine learning tools borrowed from computer science have much to offer economists, yet economists have been slow to adopt these tools (Athey & Imbens, 2019: 1). While this is beginning to change, economists more often turn to machine learning as a tool to work with large datasets or to construct novel datasets (ibid.). This creates a risk that machine learning begins to be seen by economists as the exclusive domain of those in the field who work with large or novel datasets. A potential consequence is that other useful applications of machine learning may be overlooked; applications which could be of use to economists working with more conventional economic datasets.

Therefore, to provide an example of the role machine learning can play in economic research outside the realm of large or novel datasets, my aim in this study is to show how machine learning can be used for prediction in the service of estimation in the context of a microeconomic research question: namely, whether the South African Child Support Grant improves the nutrition of children who receive it. I show how a fully connected artificial neural network – a powerful machine learning prediction tool borrowed from computer science – can be used as a novel, and potentially superior, approach to constructing an input variable in microeconomic research where the input variable is the result of high-dimension prediction.

In this type of problem, prediction is required to produce an input variable which is then used as a covariate in an estimation problem. The hypothesis is that, if a neural network can be used to improve predictive performance when constructing the input variable, the estimation step that relies on this input as a covariate should also be more accurate.

The first point to address is why economists may benefit from considering additional techniques for prediction when prediction is required in the service of estimation. In econometrics, the conventional approach to this type of problem is to use ordinary least squares (OLS) regression to make the prediction that produces the input variable. If the functional form of this OLS regression can be correctly specified, then an additional machine learning technique like a neural network is not necessary. However, it can be both difficult and time consuming to correctly specify an OLS regression in the context of high-dimension prediction

– which refers to prediction that relies on many covariates or many complex interactions between covariates (Mullainathan & Spiess, 2017: 88).

It is when dealing with high-dimension prediction that neural networks have a potential advantage; a neural network is able to fit a complex functional form that does not have to be specified in advance (Varian, 2014: 5) (Mullainathan & Spiess, 2017: 88). It can uncover generalizable patterns in complex data that are not – and would likely be very difficult to be – specified in advance (Mullainathan & Spiess, 2017: 88). It can also fit complex but very flexible functional forms to the data without *necessarily* overfitting; and can find functions that work well out-of-sample (ibid.). In short, a neural network "…provides a powerful, flexible way of making quality predictions" (ibid., 98). Therefore, a neural network is more likely to outperform an OLS regression when the variable is predicted on many covariates or covariates with complex interactions that make specifying an appropriate functional form difficult.

With this in mind, the microeconomic research question I will use to show how a neural network offers a novel, and potentially superior, approach to high-dimension prediction in the service of estimation is whether the South African Child Support Grant (from here on 'the grant') improves the nutrition of children who receive it.

I begin by addressing why the grant may benefit from an additional technique for prediction. The grant was introduced in 1998 as an unconditional cash transfer aimed at supporting nutrition, health, and education among the poorest 30% of children (Agüero et al., 2007: 5). When the grant was introduced, it was made available to all children who met the eligibility criteria (ibid. 9). Therefore, whether an eligible child receives the grant depends on whether the child's caregiver applies for the grant on their behalf.

In addition to whether the eligible child receives the grant, it is important to consider how soon the child receives the grant after becoming eligible: children whose caregivers apply for the grant with little delay will receive the grant over a greater proportion of their lives than children whose caregivers take longer to apply for the grant. The earlier an eligible child receives the grant, the larger their treatment dose.

To account for this, the identification strategy used in prominent early research by Agüero et al. (2007) – and which is the approach followed in all subsequent research of which I am aware - is to construct a variable to control for unobserved differences between caregivers, which could cause certain caregivers to apply for the grant earlier than others (Coetzee, 2013: 433). This variable is referred to as 'caregiver motivation' (Agüero et al., 2007: 9-10). Conditional on caregiver motivation, "…the extent of [grant] treatment should be random

(related only to accidents of birth time and location) and hence orthogonal to the expected effect of the treatment" (Agüero et al, 2007: 11).

The caregiver motivation variable is constructed by standardising the difference between predicted application delay and actual application delay (ibid.). Predicted application delay is the number of days a caregiver is expected to wait before applying for the grant, based on a set of observed characteristics. Actual application delay is the number of days it took the caregiver to apply for the grant following the birth of the child. Actual delay is a function of both observed and unobserved characteristics. By standardising the difference between predicted and actual delay, the resulting caregiver motivation variable controls for latent family and caregiver characteristics (Agüero et al, 2007: 11).

Given the importance of the caregiver motivation variable to the identification strategy, it is worth interrogating the method used to predict delay. As far as I am aware, in all the grant literature where the caregiver motivation variable is included as a covariate, ordinary least squares regression is used to predict delay. By design, predicted application delay is the application delay expected based on observed characteristics, whereas actual application delay is a function of both observed and unobserved characteristics.

The difficulty this creates for measuring predictive performance is that predicted delay is not necessarily more accurate if it is closer to actual delay. So, predictive performance cannot be measured by comparing predicted values of delay to actual values of delay.

Despite the difficulty in measuring predictive performance, there is reason to think the OLS prediction of delay could be improved. A noteworthy finding in the grant literature is that, in the binary treatment case (i.e., when the treatment group of eligible children who are receiving the grant are compared to the control group of eligible children who are not receiving the grant), the grant is found to have no significant impact on nutrition (see, for example, Coetzee (2013) and Oyenubi (2018)).

This is despite there being both theoretical and qualitative empirical support for expecting to find a significant treatment effect in the binary treatment case. From a theoretical perspective, poor households in South Africa spend an average of a third of their income on food (Statistics South Africa, 2014: 19). Therefore, poor households who receive the grant are expected to spend more on food, improving nutrition compared to poor households who do not receive the grant (Oyenubi, 2018: 3). In a qualitative review, Patel and Plagerson (2016: 39) find the grant to be "…one of South Africa's most effective poverty reduction programmes".

And while Agüero et al (2007: 12) argue that the treatment effect would be dampened in the binary treatment case since a portion of the children in the treatment group would have

only just started receiving the grant, this does not account for why not even a small, statistically significant treatment effect can found. As noted in Oyenubi (2018: 3), one would still expect a statistically significant treatment effect in the binary case that captures the lower bound of the treatment effect.

A possible explanation for why no statistically significant treatment effect can be found in the binary treatment case is that delay predicted using an OLS regression is suboptimal. It may be that the functional form of the OLS regression used to predict delay is incorrectly specified and that the correct functional form is difficult to specify because of complex interactions between the variables.

While delay is predicted on six variables, I argue it is still effectively high-dimensional because the variables could enter linearly or logarithmically, and with various complex interactions. It is here where a neural network should be considered instead. Neural networks perform well in high dimension prediction problems (Mullainathan & Spiess: 2016: 101). If predicting application delay is effectively a high dimensional prediction problem and neural networks perform well in high dimensional prediction problems, a neural network may outperform an OLS regression.

The final point to address is how predictive performance will be measured when comparing the neural network prediction of delay to the OLS regression prediction of delay. As discussed above, the difficulty in measuring predictive performance is that predicted delay is not necessarily more accurate if it is closer to actual delay. So, predictive performance cannot be measured by comparing predicted values of delay to actual values of delay.

Instead, predictive performance must be measured indirectly. Given the strong theoretical and qualitative empirical justification for expecting a statistically significant lower bound treatment effect in the binary treatment case, a prediction of delay that results in an estimated treatment effect consistent with this expectation will be argued to be superior.

Turning to the results of the study: consistent with Coetzee (2013) and Oyenubi (2018), I find no significant improvement in nutrition from the grant in the binary treatment case when motivation is constructed using the simple OLS regression prediction of delay. However, when the motivation variable constructed from the neural network prediction of delay is used, I find a statistically significant improvement in nutrition among children who receive the grant in the binary treatment case. Since the neural network prediction of delay results in an outcome that is consistent with theoretical expectations and qualitative research findings, I argue it is superior to the OLS prediction of delay which produces a result contrary to theoretical expectations and qualitative research.

Finding a neural network produces a better prediction of delay compared to an OLS regression in the context of the grant is the primary contribution of this paper. It serves as an important example of how a machine learning tool can be of use to economists working with more conventional microeconomic questions and datasets.

Finally, using the generalised propensity score (GPS) approach of Hirano and Imbens (2004), I find children who receive the grant for a greater proportion of their lives have improved nutrition compared to the children who receive the grant for a smaller proportion of their lives. Given this outcome in the continuous case (where the dose effect of the grant is examined among children who are all receiving the grant), and the theoretical and qualitative empirical evidence in favour of the grant already mentioned, it would be surprising if no significant treatment effect was found in the binary treatment case. Therefore, since the neural network results in a small but statistically significant treatment effect in the binary treatment case, the outcome in this, the continuous treatment case supports the argument in the binary treatment case that the neural network prediction is superior to the OLS regression prediction.

2. Literature Review

For context, I start by discussing two areas where machine learning methods have been adopted more enthusiastically by economists: working with large datasets and generating novel datasets. Given that machine learning has been developed with large datasets in mind, it is perhaps unsurprising that economists are turning to machine learning methods to work with the vast amounts of data being collected on economic activity (Einav & Levin, 2014: 1-3). Technology is allowing data to be collected on a scale and on a range of activities that has not previously been possible (ibid.). Survey data that has been the workhorse of economic research is being replaced by rich, largescale administrative datasets - maintained by social security administrations and revenue services - with near-universal population coverage (ibid., 1-2). Also significant is the expansion of data collection on economic activity by private firms in almost every sector of the economy (Ibid., 3). Such data allows economists to begin looking into the 'black box' of markets and firms, and to gain new insights into economic behaviour. (ibid., 3-4).

While such rich data offers vast opportunities for new research, conventional econometric techniques are not always ideally suited to working with such large datasets (Varian, 2014: 3). Machine learning methods, on the other hand, are well suited to working with large datasets (ibid., 3-6). Therefore, economists are now more frequently turning to machine learning as a useful tool for working with increasingly large datasets (ibid.).

Another area where economists are turning to machine learning is to unlock data from novel sources (Mullainathan & Spiess, 2017: 99). For example, machine learning is being used to extract economically meaningful data from satellite images (ibid.). Lobell (2013) uses satellite images to estimate future harvest size while Michalopoulos and Papaioannou (2013) use data created using light pixel density from satellite images captured at night as a proxy for economic development; Donaldson and Storeygard (2016) provide a detailed overview of the growing body of research in economics using satellite images to generate economically meaningful data (Mullainathan & Spiess, 2017: 99). This is particularly useful in contexts where sources of more conventional economic data are either unavailable or unreliable (ibid.).

While the application of machine learning in economics to take advantage of large and novel datasets is important, machine learning has applications for economists that extend into other areas.

One such area is what Mullainathan & Spiess (2017: 100) call prediction in the service of estimation. An example of prediction in the service of estimation is linear instrumental variables (ibid.). Using a linear instrumental variable is a two-stage process. In the first stage, regress $x = \gamma' z + \delta$ on the instrument z. Then, in the second stage, regress $y = \beta' x + \epsilon$ on the fitted values $\hat{x}$ (ibid.). While the first stage is usually handled as an estimation problem using regression, it is actually a prediction problem. The coefficients in the first stage are a means to find the predicted values; but ultimately, only the predicted values $\hat{x}$ enter into the second stage. Since the first stage is a prediction problem, the goal is predictive performance rather than parameter estimation. This plays to the strength of machine learning. It is here, when constructing the instrument variable in stage one, that economists can look to machine learning to improve predictive performance. If machine learning can improve predictive performance when constructing the instrument variable in stage one, the estimation of the treatment effect in stage two will be improved.

To be clear, linear instrumental variables is one type of problem that fits under the umbrella of prediction in the service of estimation. However, the caregiver motivation variable constructed using predicted delay in the context of the Child Support Grant is not an instrumental variable. It is a covariate included to account for latent caregiver and household characteristics (Agüero et al, 2007:9-11). Using prediction to construct an input variable which is used as a covariate in an estimation problem is another type of problem that also fits under the umbrella of prediction in the service of estimation – and the type that will be the focus of this study.

3. Data

3.1 – Predicting Application Delay to Construct Caregiver Motivation

I use data from wave five of the National Income Dynamics Study (NIDS), a nationally representative panel dataset collected in 2017. In my sample to predict application delay, there are a total of 8399 children who are eligible to receive the grant. Of these, 7687 children receive the grant while 712 children do not receive the grant even though they are eligible. To be eligible to receive the grant, a child must be under the age of 18 and their primary caregiver must pass a means test (SASSA, 2016: 4). The grant is applied for on behalf of a child by their primary caregiver who need not be a parent (ibid.).

A child is age-eligible for the grant if they are younger than 18 years old. However, for the purposes of this study and of the NIDS survey, children are defined as those younger than 15 years old (Brophy et al., 2018: 18). Even though the grant is available to children up to the age of 18, I limit the sample to children younger than 15 in 2017. While the South African government's intention was to roll out the grant programme nationally in 1998, the reality is it took time for the information about the grant to reach all communities (Agüero et al, 2007: 11). It took approximately four years from the grant's introduction in 1998 for the average application delay to converge to its expected long-run level (ibid., 12-13). Therefore, by only including children younger than 15 in 2017, all the children in the sample are born after 2002; and so were born once the average expected application delay had stabilised at its long-run level. By excluding the initial rollout period, the sample is less likely to be affected by differences in the effectiveness of the rollout of the grant in different parts of the country.

In addition to being age eligible, a child's caregiver must also pass a means test before being allowed to receive the grant on a child's behalf. The grant is intended to be received by caregivers who have insufficient means to support themselves and the child or children in their care (South African Social Security Agency (SASSA), 2016: 7). The purpose of the means test is to ensure only those with insufficient means receive the grant by looking at the income of the primary caregiver applying to receive the grant on the child's behalf (ibid.). To pass the means test and be allowed to receive the grant on a child's behalf, a caregiver must have an annual income less than the threshold amount. The threshold amount for an unmarried primary caregiver is ten times the annual grant amount prevailing at the time of their application (ibid.). For a married primary caregiver, the threshold amount is doubled such that the combined annual income of the primary caregiver and their spouse must be less than 20 times the annual grant amount prevailing at the time of their application (ibid.). Since the grant amount increased

in April of 2017, the threshold amount used for the means test also changed in April of 2017. To account for this change, the threshold amount used for the means test depends on the date the caregiver was interviewed for the NIDS survey. Further detail on how I apply the means test to account for this change can be found in Appendix A.

To estimate each caregiver's income, I start with the income they report in the previous month. This includes income from both temporary and permanent sources and from formal and informal sources but excludes transfers from government like Unemployment Insurance Fund (UIF) payments (Coetzee, 2013: 450). I then annualise the income reported in the previous month to estimate the caregiver's annual income. It is the caregiver's annual income that is used to determine whether they pass the means test (SASSA, 2016: 4). My approach is consistent with the approach used by SASSA (ibid.). Appendix A provides a more detailed explanation of how caregiver income is calculated using the wave 5 NIDS data.

Since earnings may be seasonal, there is a chance my approach over- or under-estimates a caregiver's annual income. However, on balance, the risk this poses is low. First, suppose caregiver income is underestimated: the majority of caregivers who pass the means test have annual income that falls well below the income threshold amount. So even if caregiver income is underestimated, it would have to be underestimated by a large amount to make caregivers who are currently eligible, ineligible. For caregivers who are currently ineligible, correcting income that is underestimated has no impact on their ineligibility.

Now, suppose caregiver income is overestimated: caregivers whose income is currently below the threshold amount must necessarily remain below the threshold amount if their overestimated income were to be adjusted downwards. Furthermore, most ineligible caregivers who fail the means test have incomes far exceeding the threshold amount; their incomes would have to be overestimated by a large amount for them to go from being ineligible to eligible if their overestimated income were to be adjusted downwards. So, while seasonality means there is a possibility that incomes are over- or under-estimated, it would have to be by a large amount to affect whether or not most caregivers pass the means test.

The low probability of misestimating income notwithstanding, my approach is consistent with how the means test is applied by SASSA (2016). It should, therefore, lead to outcomes consistent with what would occur in reality when the means test is applied. Thus, my approach is an accurate reflection of the means test, even if the means test as applied by SASSA is also potentially susceptible to misestimation.

I now turn from how the sample is constructed to the choice of covariates for predicting application delay. Consistent with the approach used by Agüero et al. (2007) - and subsequently

by Coetzee (2013) – I include the two covariates *rural* and *child_age*: whether the household is located in a rural or an urban area is captured by the dummy variable, *rural*. Rural households may have to travel further to access social security agency offices to apply for the grant. Relative to urban households, the added burden and cost of having to travel further to apply for the grant may increase the application delay for caregivers in rural households. Application delay may, therefore, be greater for caregivers who live in rural households compared to caregivers who live in urban households. The NIDS defines households as urban or rural according to the classification in the 2011 census (Brophy et al, 2018: 76). The continuous variable *child_age* is measured in years and captures the age of the child at the time of the NIDS survey. A child's age relates to application delay in a straightforward way: the older the child, the greater their application delay can be, with the maximum application delay being the child's age. A child's application delay will be equal to their age if they were eligible to receive the grant at the time of their birth but are still not receiving it at the time of the NIDS survey.

Following the approach in Oyenubi (2018: 13-14), I also include a dummy for whether the primary caregiver is the child's parent, *pc_parent*. A primary caregiver who is not the child's parent must prove they are the primary caregiver by providing "an affidavit from a police official, a social worker's report, an affidavit from the biological parent or a letter from the principal of the school attended by the child" (SASSA, 2014). Acquiring the relevant documentation is likely to increase the application delay for primary caregivers who are not parents.

In addition to these three covariates, I include three further covariates when predicting application delay. As far as I am aware, this novel addition – that takes advantages of questions in the NIDS survey – has not been used in prior research. Nevertheless, there are theoretical justifications that warrant their inclusion. First, the dummy variable *caregiver_has_id* captures whether a caregiver has a South African identity document. A primary caregiver requires a South African identity document to apply for the grant (SASSA, 2016: 4). Second, the dummy variable *birth_cert* captures whether the primary caregiver has the child's birth certificate. A primary caregiver must also provide the child's birth certificate to apply for the grant. If a caregiver does not have a South African ID or the child's birth certificate, they are allowed to provide alternative documentation (ibid.). However, the process of finding out which alternative documentation is accepted and then acquiring the alternative documentation may increase application delay for these primary caregivers. To account for this, I include *caregiver_has_*id and *birth_*cert when predicting application delay. If a primary caregiver does not have the child's birth certificate, they may instead provide the child's road to health clinic

card (ibid.). Therefore, the third additional covariate I include is the dummy variable *clinic_card* which captures whether a primary caregiver has the child's clinic card. Among caregivers who do not have the child's birth certificate, those caregivers who have the child's clinic card are expected to have a shorter application delay than caregivers who do not have the child's clinic card since the clinic card is an accepted substitute (ibid.).

Having dealt with the covariates I use in predicting application delay, I now explain how actual application delay is calculated for the treatment group and the control group. For the most part, my approach follows that of Coetzee (2013), with one adjustment as suggested in Oyenubi (2018). For the treatment group – the children currently receiving the grant – actual delay is the number of days between their date of birth and the date they first received the grant. For the control group – the children eligible to receive but not currently receiving the grant – actual delay is calculated as the number of days between their date of birth and the date they were interviewed for the wave 5 NIDS survey. There is a small subgroup of this eligible but not receiving group who have previously but unsuccessfully applied for the grant. For these observations, actual delay is calculated as the number of days between the child's date of birth and the date on which the grant was first applied for on the child's behalf. This adjustment, as suggested by Oyenubi (2018: 12), is more appropriate when data on application date is available; it takes into account that these caregivers did make the effort to apply for the grant, even though they were not successful. By taking their attempted application into account when calculating actual delay in the control group, the guaranteed imbalance that results from equating actual delay to age may be mitigated (ibid.).

The implicit assumption I carry over from Coetzee (2013) when constructing actual delay for both the treatment and control group is that all eligible children have been eligible for the grant from their date of birth. I acknowledge this assumption need not necessarily hold true for all children in the sample. A caregiver may, for example, lose their job and be unable to find work again. As a result, a child who is currently eligible need not necessarily have been eligible since birth. The potential implication is actual delay may be inflated. In mitigation, Nonyana & Njuho (2018) find the unemployed poor in the South African labour market, and likely poor women in particular, are less mobile in their labour market state. Therefore, a primary caregiver who is currently unemployed or who is earning a basic subsistence income less than the threshold amount is unlikely to have been earning more than the threshold amount in the past. This supports the assumption that a child who is currently eligible has been eligible since birth. Given this support, there is sufficient justification to proceed with the assumption.

I do, however, concede that there is an opportunity for further empirical work in this area in the future to test the assumption.

Predicted delay and actual delay are then used as inputs to construct the caregiver motivation variable. Caregiver motivation is constructed by standardising the difference between predicted delay and actual delay. Caregivers whose actual delay is greater than their predicted delay are characterised as relatively less motivated while caregivers whose actual delay is less than their predicted delay are characterised as relatively more motivated.

While it is tempting to understand motivation as a caregiver characteristic, this would be a mistake. Motivation is the result of latent, unobserved characteristics that affect application delay; these characteristics may be caregiver characteristics but may also be household or other unobserved characteristics (Agüero et al., 2007: 9-11). For consistency with the prominent literature on the grant, I will continue to refer to the variable as caregiver motivation, but it really refers to latent characteristics that affect delay more generally.

3.2 – The Binary Treatment Case

In my sample for the binary treatment case, there are a total of 8399 children who are eligible to receive the grant. Of these, there are 7687 children who currently receive the grant. This receiving group will be referred to as the treatment group. There are another 712 children in the sample who are not receiving the grant even though they are eligible. This not-receiving group will be referred to as the control group. This suggests about 93% of the children who are eligible for the grant are receiving it – reaffirming the strong take-up of the grant reported in other research (Patel & Plagerson, 2016: 39).

Following Agüero et al. (2007), the outcome variable I use to measure child nutrition in both the binary and the continuous treatment case is the height-for-age z-score, *HAZ*. While Coetzee (2013) uses HAZ along with five other outcome variables that measure wellbeing more comprehensively, the aim of this paper is to demonstrate how a neural network can be used for prediction in the service of estimation. For this purpose, using *HAZ* as the one choice outcome variable that focuses particularly on nutrition is sufficient.

HAZ is the difference between a child's height and the median height of all children in the reference population, divided by the standard deviation of height for all children in the reference population (Coetzee, 2013: 430). For children under the age of five, the World Health Organisation (WHO) international child growth standards are used as the reference population (Brophy et al., 2018: 64). For children over the age of five, the WHO growth standards for school-aged children and adolescents are used as the reference population (ibid.).

The reliability of the *HAZ* variable depends on how accurately the child's height is measured. To help avoid measurement error, NIDS fieldworkers are instructed to take the child's height measurement twice and then a third time if the first two measures differ by more than one centimetre (ibid.). NIDS fieldworkers are also instructed to measure children in a particular position depending on their age which helps to minimise measurement error (ibid.).

HAZ is a good indicator of long-run nutritional attainment; it reflects the cumulative investment in a child's health and nutrition since birth (Coetzee, 2013: 430). It is a suitable outcome variable particularly because it measures the cumulative impact of the flow of funds from the grant on the long-term nutritional attainment of the child. A child whose height is less than the standard measure expected for their age cohort in healthy populations will have a negative *HAZ*. A negative *HAZ* indicates the child has likely received insufficient nutrition over an extended period; this may be the result of various long-term factors including "chronic insufficient protein and energy intake, frequent infection, sustained inappropriate feeding practices and poverty" (Agüero et al., 2007: 7). Children may be particularly vulnerable to nutrition shortfalls in the first three years of their life; growth failures due to nutrition shortfalls in this age range may not be reversible (ibid.). Therefore, delay in applying for the grant relative to the child's age is an important factor.

That being said, there is another strand of literature that finds children who have experienced nutrition shortfalls can still experience catch-up growth if nutrition is improved (Boersma & Wit, 1997 as cited in Oyenubi, 2018: 11). Catch-up growth is rapid linear growth that may allow a child to re-join their growth curve in spite of a period of insufficient nutrition (ibid.). Catch up growth means the potential for the grant to improve nutrition – and with it *HAZ* – still exists even if a child has not received the grant in their first three years of life.

For consistency, I use the same covariates in the binary treatment case as Coetzee (2013) and Oyenubi (2018) namely: the caregiver's employment status, *employed*; the caregiver's marital status, *married*; the caregiver's education (using dummy variables for the caregiver's highest education level: primary school, *edu_primary,* high school, *edu_high,* matric certificate, *edu_matric*, tertiary certificate, *edu_certificate*, and university degree *edu_degree*); the child's gender, child_male; the child's race, *child_black*; whether the child is underweight at birth, *underweight_at_birth*; whether the household has electricity, *electricity*; whether the household has piped water, *piped_water*; and whether the household has a flushing toilet, *flush_toilet*.

Comparing the characteristics of children in the treatment group and the control group, table 1 shows children in the treatment group are, on average, about five months older than

children in the control group. The only other significant difference between the treatment group and the control group is the proportion of children who have birth certificates: 99% in the treatment group but only 83% in the control group.

Turning from child characteristics to caregiver characteristics: table 1 shows primary caregivers of children in the treatment group and primary caregivers of children in the control group differ in several key areas. Most notable is the difference between the age of caregivers in the two groups and whether the caregiver is the child's parent. The average age of caregivers in the treatment group is approximately 38 years old, compared to an average age of approximately 45 years for caregivers in the control group. Moreover, 76% of caregivers in the treatment group are the child's parent compared to only 49% in the control group. This may help to explain why caregivers in the control group are not yet receiving the grant on behalf of the children in their care. It also supports the inclusion of these variables when predicting delay. Also noteworthy, and consistent with Coetzee (2013: 432), is that a much smaller proportion of caregivers in the control group are employed compared to caregivers in the treatment group.

Finally, looking at the household characteristics in table 1, a similar proportion of households in the treatment group and the control group have access to electricity, have access to a flush toilet, are located in a rural area and a have a household head who is male.

Where households in the treatment group and the control group differ is with respect to piped water, average total monthly expenditure per capita and average total monthly food expenditure per capita: households in the control group have better access to piped water and have greater total expenditure and food expenditure per capita. This may lend weight to the hypothesis that caregivers in the control group live in households that are, in some sense, better off.

Table 1

Descriptive Statistics

Variables	Treated					Not Treated
		Treatment Dose				
	All (0;100%]	Low (0;33%]	Medium (33;66%]	High (66;100%]		Eligible but not receiving
Child Characteristics						
n	**7687**	**454**	**542**	**6691**		**712**
child_age in years	**7.391**	**7.425**	**6.129**	**7.492**		**6.983**
	(4.021)	(4.33)	(4.467)	(3.943)		(4.332)
male	**0.493**	**0.522**	**0.52**	**0.489**		**0.521**
	(0.5)	(0.5)	(0.5)	(0.5)		(0.5)
black	**0.879**	**0.835**	**0.845**	**0.885**		**0.878**
	(0.326)	(0.372)	(0.362)	(0.319)		(0.328)
clinic card	**0.965**	**0.952**	**0.959**	**0.966**		**0.934**
	(0.184)	(0.215)	(0.198)	(0.181)		(0.248)
birth certificate	**0.991**	**0.991**	**0.987**	**0.991**		**0.829**
	(0.097)	(0.094)	(0.113)	(0.096)		(0.377)
HAZ	**-0.857**	**-0.889**	**-0.868**	**-0.854**		**-0.699**
	(1.218)	(1.329)	(1.382)	(1.196)		(1.354)
underweight at birth	**0.071**	**0.081**	**0.076**	**0.07**		**0.077**
	(0.257)	(0.274)	(0.265)	(0.255)		(0.267)
Caregiver Characteristics						
motivation (NN)	**0.178**	**-1.79**	**-0.693**	**0.382**		**-1.903**
	(0.762)	(1.164)	(0.69)	(0.414)		(1.462)
motivation (OLS)	**0.186**	**-1.801**	**-0.72**	**0.394**		**-1.919**
	(0.752)	(1.148)	(0.641)	(0.387)		(1.481)
actual_delay (days)	**343.534**	**2267.344**	**1115.386**	**150.476**		**2503.208**
	(701.267)	(1368.955)	(858.277)	(204.498)		(1603.36)
caregiver_age	**37.693**	**40.974**	**35.764**	**37.626**		**44.895**
(years)	(12.8)	(14.166)	(13.228)	(12.627)		(17.419)
married	**0.311**	**0.355**	**0.328**	**0.307**		**0.264**
	(0.463)	(0.479)	(0.47)	(0.461)		(0.441)
parent	**0.763**	**0.573**	**0.751**	**0.777**		**0.494**
	(0.425)	(0.495)	(0.433)	(0.417)		(0.5)
employed	**0.358**	**0.339**	**0.327**	**0.362**		**0.042**
	(0.479)	(0.474)	(0.469)	(0.481)		(0.201)
caregiver_has_id	**0.943**	**0.949**	**0.939**	**0.943**		**0.829**
	(0.231)	(0.22)	(0.239)	(0.231)		(0.377)
edu_primary	**0.193**	**0.262**	**0.183**	**0.181**		**0.263**
	(0.394)	(0.44)	(0.387)	(0.385)		(0.44)
edu_high	**0.45**	**0.351**	**0.465**	**0.465**		**0.323**
	(0.497)	(0.477)	(0.499)	(0.499)		(0.468)
edu_matric	**0.141**	**0.126**	**0.173**	**0.142**		**0.126**
	(0.349)	(0.332)	(0.379)	(0.349)		(0.331)
Household Characteristics						
electricity	**0.858**	**0.837**	**0.865**	**0.859**		**0.86**
	(0.349)	(0.37)	(0.342)	(0.348)		(0.348)
piped water	**0.299**	**0.308**	**0.382**	**0.292**		**0.346**
	(0.458)	(0.462)	(0.486)	(0.455)		(0.476)
flush toilet	**0.378**	**0.421**	**0.461**	**0.369**		**0.397**
	(0.485)	(0.494)	(0.499)	(0.483)		(0.49)
rural	**0.583**	**0.553**	**0.502**	**0.592**		**0.581**
	(0.493)	(0.498)	(0.5)	(0.491)		(0.494)
male headed	0.**277**	**0.304**	**0.321**	**0.272**		**0.26**
	(0.448)	(0.46)	(0.467)	(0.445)		(0.439)
total expenditure per	**772.833**	**794.695**	**957.624**	**756.38**		**1318.402**
capita	(1064.76)	(860.494)	(2225.953)	(921.199)		(3065.75)
food expenditure per	**232.467**	**239.132**	**248.464**	**230.719**		**289.511**
capita	(159.711)	(173.002)	(166.291)	(158.163)		(277.521)

Notes: Mean values are shown in bold while standard deviation values are shown in parentheses.

3.3 – The Continuous Treatment Case

The sample for the continuous case contains only the 7687 children who are currently receiving the grant. The continuous treatment case only looks at children who are currently receiving the grant and compares them based on the dose of the grant they have received. The treatment dose is the proportion of a child's life over which they receive the grant. Three dosage categories are defined to help simplify comparison. The 'low' dosage category includes children who have received the grant for up to 33% of their lives. The 'medium' dosage category includes children who have received the grant for between 33% and 66% of their lives. The 'high' dosage category includes children who have received the grant for more than 66% of their lives.

As can be seen in table 1, a large majority of children currently receiving the grant – approximately 87% – have received it for more than two thirds of their lives. Approximately 7% have received it for between two thirds and one third of their lives. Approximately 6% have received it for less than a third of their lives. This again reaffirms the success of the grant programme as a large majority of the children receiving the grant have received it for most of their lives. This also suggests that, if one were to find a treatment effect in the continuous treatment case, a smaller but still significant effect should be seen in the binary treatment case since the vast majority of grant recipients have received it for most of their lives.

The continuous treatment case uses the same covariates as the binary treatment case. However, an important difference between the binary and continuous treatment cases is that the motivation variable is not included in the continuous case. All of the children in the sample in the continuous case are currently receiving the grant and the dose is explicitly accounted for.

4. Methodology

4.1 – Using a Neural Network to Predict Application Delay

I propose using a neural network to predict application delay for the grant rather than the ordinary least square (OLS) regression approach used by Agüero et al. (2007) and Coetzee (2013). My focus here will be to explain (i) how a neural network is different to OLS regression, (ii) why a machine learning approach that relies on a neural network should be considered as an alternative to a conventional OLS regression approach for prediction in the service of estimation and (iii) how the neural network I use to predict application delay for the grant is specified in this study. I leave a more comprehensive theoretical explanation of neural networks to Appendix B.

4.1.1: How a Neural Network is different to an OLS Regression

Statistical modelling can be classified into two broad cultures: data modelling and algorithmic modelling (Breiman, 2001, as cited in Athey & Imbens, 2019: 1). Data modelling assumes the data are generated by a given stochastic model (Athey & Imbens, 2019: 1-2.). Ordinary least squares regression fits under the data modelling umbrella. In an OLS regression, the functional form of the model is specified in advance. On the other hand, algorithmic modelling treats the data mechanism as unknown (ibid.). Neural networks fit under the umbrella of algorithmic modelling (ibid.). The functional form of the neural network does not need to be specified in advance.

Algorithmic modelling can be used on large complex datasets but also, and crucially in support of the aim of this paper, "as a more accurate and informative alternative to data modelling [like OLS] on smaller datasets" (ibid.). The strength of machine learning methods like neural networks is they can fit very flexible functional forms without necessarily overfitting (Mullainathan & Spiess, 2017: 88, 99). This makes neural networks ideally suited to prediction problems.

Crucially, neural networks need a way to guard against overfitting. Take a trivial example where n linearly independent regressors are used to fit n observations: the regressors will fit the observations perfectly but out-of-sample performance will usually be poor (Varian, 2014: 6). In machine learning, this is referred to as the overfitting problem (ibid.). Poor out-of-sample performance is problematic because out-of-sample predictive performance is usually what we are trying to optimize. There tends to be a tradeoff between model complexity and out-of-sample predictive performance, so it is important to penalize models for excessive complexity. This is referred to in computer science as regularization (ibid., 7).

It is common practice to divide the data into separate sets for training and testing (which are usually combined) and validation (ibid.). The training data is used to estimate a model, the validation data is used to choose the model and the testing data is used to evaluate the performance of the chosen model (ibid.).

Next, by having an explicit numeric measure of model complexity, model complexity can be treated as a parameter that can be tuned to optimize out-of-sample predictive performance (ibid.). k-fold cross-validation is the standard approach used to choose a good value of the tuning parameter. It works as follows:

1. Divide the data into k equal subsets called folds and label them $s = 1, \dots, k$. Start with subset $s=1$.

2. Choose a value for the tuning parameter.
3. Fit the model using the $k-1$ subsets other than s.
4. Predict for subset s and measure the associated loss.
5. Stop if $s = k$, otherwise increment s by one and return to step two. (ibid.)

A common choice for k, which I use in this paper, is 10. After cross-validation, there are k values of the tuning parameter and the associated loss. These can then be used to choose an appropriate value for the tuning parameter. In this way, the test-train cycle and cross-validation balance in-sample goodness of fit and out-of-sample predictive performance by penalizing excessive complexity (ibid.). This is how neural networks can fit complex, flexible functional forms without necessarily overfitting.

4.1.2: Why a Neural Network is suited to prediction in the service of estimation in the case of the grant

The motivation variable relies on delay being predicted as accurately as possible based on observed characteristics. The six covariates I use to predict application delay likely interact in complex ways. Covariates may be included linearly or with higher orders and they may be interacted with one another. This means the problem is effectively high-dimensional (Mullainathan & Spiess, 2017: 101). Neural networks have two main advantages when dealing with high dimension prediction problems. First, neural networks allow for very flexible functional forms. Second, since a neural network does not require a functional form to be specified in advance, researchers avoid having to spend time testing different specifications.

4.1.3: How the Neural Network is specified in this study

Turning now to the neural network specification: the fully connected artificial neural network I use to predict application delay is built using Keras and TensorFlow. I specify two hidden layers with 10 neurons in each hidden layer. The network uses all the specified neurons. The activation function in the model is Tanh. The loss function in the model uses mean square error (MSE). The architecture of the network is determined using a manual search process with a validation set. The architecture selected is the simplest network that is still able to learn the trend as tested on a hold-out set.

Out of sample properties are optimised using 10-fold cross validation. The total sample of 8399 observations is randomly allocated between the folds.

4.2 – Using the GenMatch to Estimate the Binary Treatment Effect of the Grant

Once application delay has been predicted using the neural network and the OLS regression and the *motivation_NN* and *motivation_OLS* variables have been constructed, the next step is to estimate whether the grant has an effect on child nutrition in the binary treatment case. The binary treatment case compares children who are receiving the grant to children who are not receiving the grant even though they are eligible. First is the issue of matching: rather than using propensity score matching (PSM) as in Coetzee (2013), I follow Oyenubi (2018) in using a matching algorithm proposed by Diamond and Sekhon (2013) called Genetic Matching ('GenMatch').

4.2.1: Why Genetic Matching is preferred to Propensity Score Matching

The advantage of GenMatch is it weighs both the covariates and the propensity scores to achieve balance (ibid., 4). This is useful in the context of the grant: while the motivation variable is necessary as part of the identification strategy, it is – by construction – a strong predictor of treatment (ibid.). Including a variable that is a strong predictor of treatment may violate the common support assumption (ibid.). Thus, as found in Coetzee (2013) and Oyenubi (2018), it is difficult to find a propensity score specification that balances the distribution of covariates when the motivation variable is included (Oyenubi, 2018: 3). GenMatch, on the other hand, does not rely only on propensity scores. Therefore, GenMatch is better suited than propensity score matching in the context of the grant (ibid.).

4.2.2: How Genetic Matching Works

To perform multivariate matching, GenMatch uses "…an evolutionary search algorithm to determine the weight each covariate is given in order to satisfy the balancing condition in matched data. The algorithm allows the researcher to select a loss function to be optimised and the preferred measure of balance." (Oyenubi, 2018: 1). GenMatch can be thought of as a generalisation of the Mahalanobis metric that includes an additional weight matrix M; M is a $t\ x\ t$ positive definite weight matrix (Oyenubi, 2018: 6). With the distance $d(w_i, w_j)$ being used to match observations in the sample, the goal is to find the weight matrix M that optimises balance:

$$d(w_i, w_j) = \left\{(w_i - w_j)' \left(S^{-\frac{1}{2}}\right)' M\, S^{-\frac{1}{2}}(w_i - w_j)\right\}^{1/2} \quad (1)$$

where w_i is a vector of covariates for observation i and $S^{\frac{1}{2}}$ is the Cholesky decomposition of the variance covariance matrix of X (ibid.). This generalisation allows propensity scores to be included as one of the covariates (ibid.). On the one extreme, GenMatch

is equivalent to propensity score matching if propensity scores contain all the relevant information to balance the covariates. On the other extreme, GenMatch converges to Mahalanobis distance if propensity scores fail to achieve the optimal level of balance in the covariates (ibid.). The strength of GenMatch is between these extremes: GenMatch allows weight to be allocated to the propensity scores and all of the covariates (ibid.). Therefore, in addition to balancing the propensity scores, GenMatch is able to account for imbalances in individual covariates.

4.3 – Using Generalised Propensity Score to Estimate the Continuous Treatment Effect of the Grant

In this paper, the focus is the binary treatment case because the binary treatment case relies on the caregiver motivation variable. However, for completeness, I also present the continuous treatment case. The binary treatment case should provide the lower bound of the treatment effect; the continuous treatment case should affirm the direction (positive or negative) of the treatment effect. So even though the continuous treatment case does not rely on the caregiver motivation variable, finding a significant treatment effect in the continuous case is important to affirm the expected lower-bound treatment effect in the binary case.

To understand why the continuous treatment case is important, I start by explaining a shortcoming in the binary treatment case raised by Agüero et al. (2007). A child who has only received the grant for a very short period of time and a child who has received the grant for much longer – as a proportion of their lifetime - are both classified as treated. Yet, the grant is likely to have a greater treatment effect if the dose of the grant received by the child is larger. Here, treatment dose is the percentage of a child's life over which they have received the grant.

The continuous treatment case takes the potential dose effect into account by only looking at children who receive the grant and comparing the treatment effect of the grant based on the dose received. I follow Coetzee (2013: 435), in using the generalised propensity score (GPS) approach of Hirano and Imbens (2004). Since Coetzee (2013: 435) already provides a detailed explanation of how the GPS approach is applied in the context of the grant, I do not duplicate it here.

5. Results and Discussion

5.1 – Predicted Application Delay and Caregiver Motivation:

Comparing the neural network predictions in figure 1 to the OLS predictions in figure 2, it is clear the two approaches produce different predictions of delay. The OLS prediction produces delay values clustered strongly around linear bands. These linear bands are not as apparent in the neural network predictions where there appears to be more variation than the OLS prediction. This is likely the result of the neural network being able to accommodate a much more flexible functional form. Table 2 and 3 provide further details. The standard deviation of the neural network predictions is greater than the standard deviation of the OLS predictions. The neural network also results in more extreme outliers as can be seen by the very small minimum and very large maximum. However, turning to interquartile range, there is less dispersion in the middle 50% of the values for the neural network prediction than for the OLS prediction. Both the mean and the median predicted delay are also lower for the neural network. Therefore, the choice of prediction method matters for the outcome of predicted delay.

Figure 1: OLS prediction of delay and actual delay plotted again child age

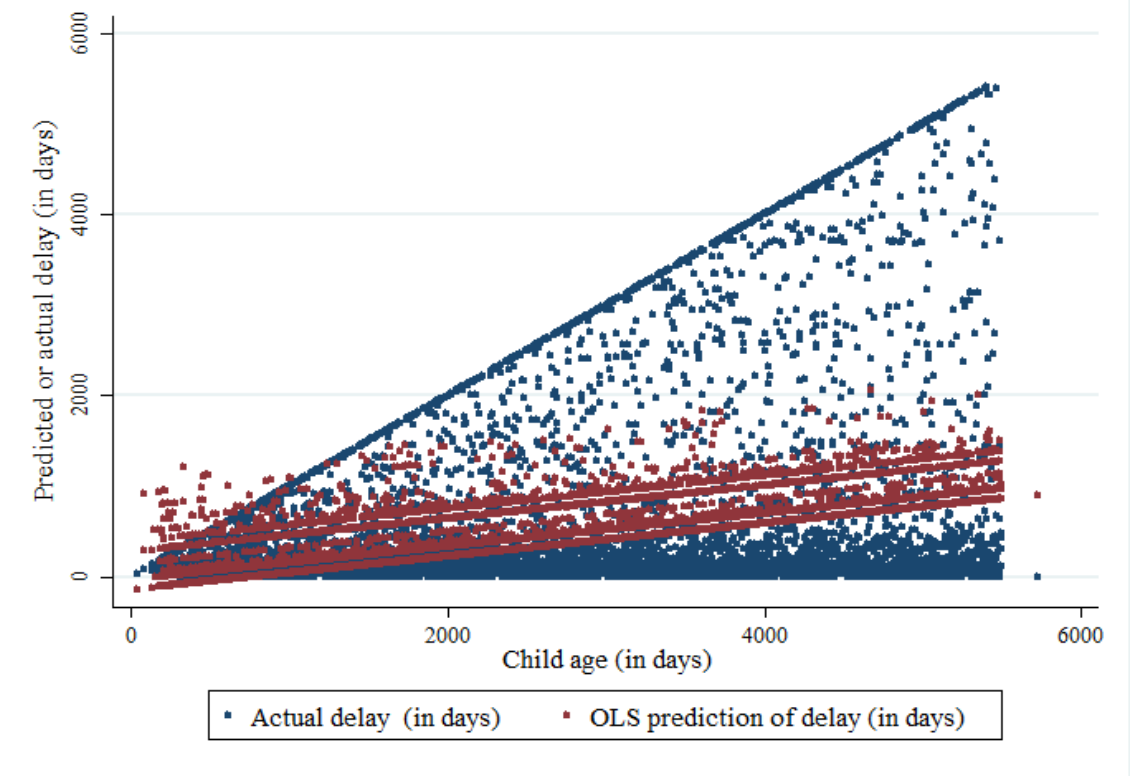


Figure 2: Neural network prediction of delay and actual delay plotted again child age

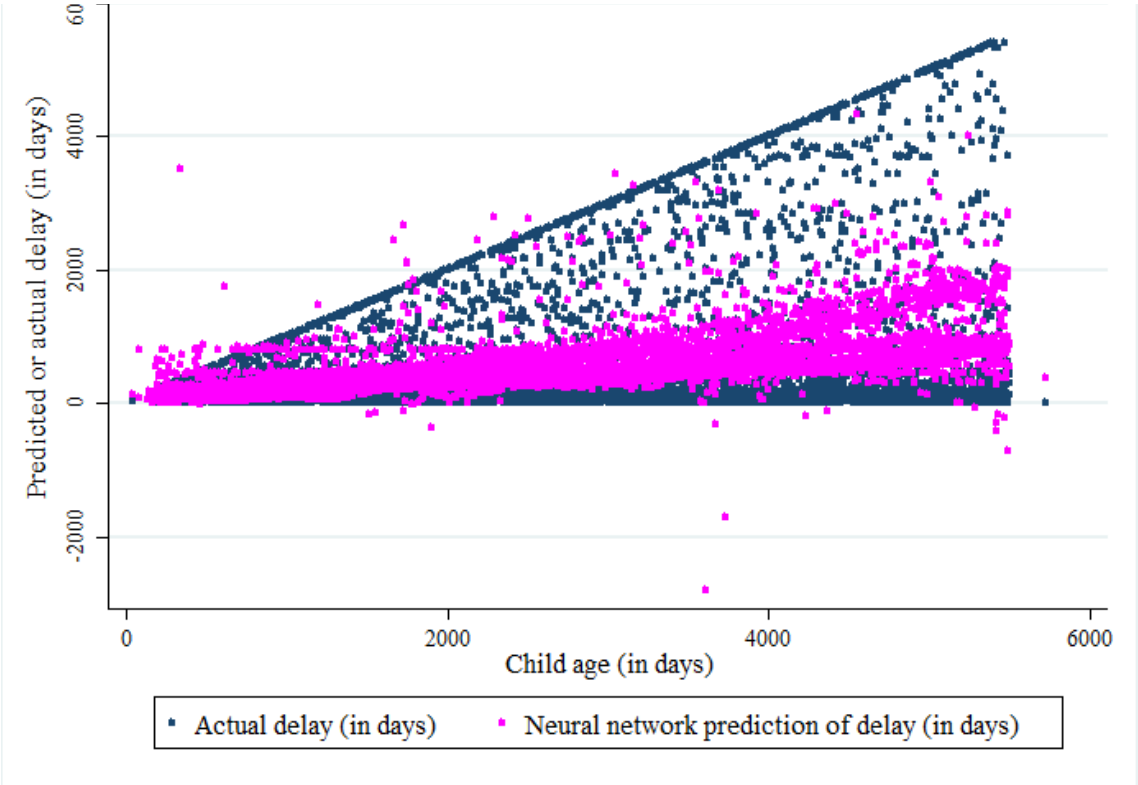


Table 2

Application Delay (in days) Predicted using OLS

Percentiles	*pdelay_OLS*	Smallest		
1%	-82.53	-134.46		
5%	7.38	-116.24		
10%	68.33	-108.95	**Observations**	8399
25%	242.97	-108.41		
50%	517.37		**Mean**	533.27
		Largest	**Std. Dev.**	356.48
75%	783.13	1860.58		
90%	1005.98	1938.03	**Variance**	127077.1
95%	1164.27	2011.517	**Skewness**	0.348
99%	1380.45	2058.23	**Kurtosis**	2.37

Notes: *pdelay_OLS* is application delay predicted using ordinary least squares regression.

Table 3

Application Delay (in days) Predicted using a Neural Network

Percentiles	*Pdelay_NN*	Smallest		
1%	49.68	-2795.05		
5%	114.35	-1712.44		
10%	159.07	-708.52	**Observations**	8399
25%	255.45	-413.55		
50%	430.95		**Mean**	529.55
		Largest	**Std. Dev.**	410.11
75%	703.89	3426.06		
90%	959.74	3512.19	**Variance**	168193.5
95%	1340.69	4015.21	**Skewness**	2.012
99%	1986.06	4323.36	**Kurtosis**	10.873

Notes: *pdelay_NN* is application delay predicted using a neural network.

As explained in section 3.1, the predictions of delay are used to construct caregiver motivation. It is important to examine the resulting motivation variables, as summarised in Table 4 and Table 5, because they represent the interaction between predicted delay and actual delay. Both motivation variables have similar mean, median and standard deviation values. As is the case with predicted delay, *Motivation_NN* has a smaller minimum and larger maximum than *Motivation_OLS*. *Motivation_NN* also has a larger interquartile range than *Motivation_OLS* (0.8 vs 0.7).

Looking at neural network and OLS predictions of delay, and the two resulting motivation variables, it is not possible to make a claim about which is superior. This is because the motivation variable represents latent characteristics. For now, it is sufficient to see that the two approaches do produce different results.

Table 4

Summary of Motivation_OLS

Percentiles	*Motivation_OLS*	Smallest		
1%	-4.06	-4.89		
5%	-2.42	-4.79		
10%	-1.22	-4.75	Observations	8399
25%	-0.12	-4.74		
50%	0.24		Mean	0.007
		Largest	Std. Dev.	1.023
75%	0.58	1.74		
90%	0.86	1.77	Variance	1.05
95%	1.02	1.96	Skewness	-2.18
99%	1.31	2.17	Kurtosis	8.22

Notes: *Motivation_OLS* is motivation constructed using delay predicted using ordinary least squares regression.

Table 5

Summary of Motivation_NN

Percentiles	*Motivation_NN*	Smallest		
1%	-3.83	-6.93		
5%	-2.44	-5.26		
10%	-1.25	-4.98	Observations	8399
25%	-0.32	-4.94		
50%	0.22		Mean	0.001
		Largest	Std. Dev.	1.024
75%	0.48	3.05		
90%	0.79	3.29	Variance	1.049
95%	1.02	3.44	Skewness	-2.05
99%	1.77	3.58	Kurtosis	8.32

Notes: *Motivation_NN* is motivation constructed using delay predicted using a neural network.

5.2 – The Binary Case: Treatment Effect of the Child Support Grant

The outcome of the binary treatment case is shown in Table 6. The *Motivation_NN* column shows the outcome in the binary treatment case when the neural network prediction of delay is used to construct the caregiver motivation variable (from here on 'the NN case'). The *Motivation_OLS* column shows the outcome in the binary treatment case when the OLS prediction of delay is used to construct the caregiver motivation variable (from here on 'the OLS case'). In the NN case, the grant has a strong positive treatment effect of 19.78% of standard deviation and the treatment effect is statistically significant at the 1% level. Conversely, in the OLS case, the grant has a weak negative treatment effect of -0.2% of standard deviation; the treatment effect is not statistically significant.

Table 6

Treatment Effect and Standard Error using GenMatch with Standardised Difference in Means as the Balance Measure

	Motivation_NN	*Motivation_OLS*
Estimate	0.19778	-0.0027175
Standard Error	0.022453	0.023968
t-stat	8.8087***	-0.11338
AI Standard Error	0.23465	0.13836
t-stat	0.84286	-0.019641

Notes: ***p<0.01, **p<0.05, *p<0.1
AI – Abadie and Imbens (2004) standard error that adjusts for matching.

The outcome in the NN case aligns with the theoretical expectation that children who receive the grant will have better nutritional outcomes than otherwise comparable children who do not receive the grant. The outcome in the OLS case is inconsistent with what is expected in theory and with qualitative empirical work on the grant.

It is possible the children in the treatment group who have only been receiving the grant for a small proportion of their lives may lessen the treatment effect in the binary treatment case. The binary treatment case may provide a lower bound for the treatment effect of the grant due to the dampening impact of children who have only received the grant for a short period of time. However, most children in the treatment group have received the grant for more than two thirds of their lives so there is reason to expect a positive and significant lower-bound treatment effect in the binary treatment case.

There is strong theoretical and qualitative empirical support for finding a positive and significant treatment effect in the binary treatment case. Therefore, *Motivation_NN* – which results in a positive and significant treatment effect – may be argued to be superior to *Motivation_OLS* – which does not – since it produces a result which is consistent with theoretical and qualitative empirical expectations.

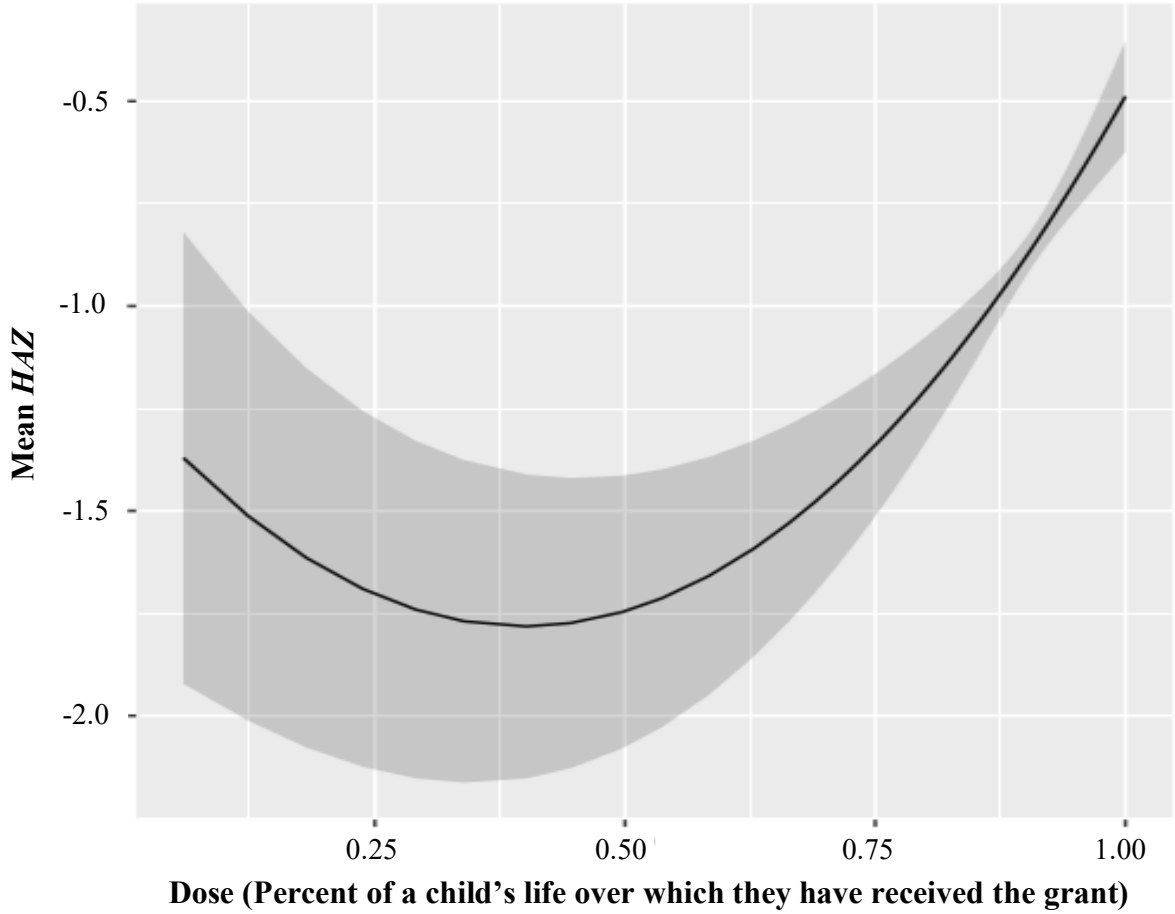
5.3 – The Continuous Case: Treatment Effect of the CSG

Finally, I present the results of the continuous treatment case. These results offer completeness: there should be a significant treatment effect in the continuous case given that the neural network prediction of delay results in a significant treatment effect in the binary treatment case – which should be the lower bound of the treatment effect.

The negative mean *HAZ* values seen on the vertical axis of Figure 3 are an indication that, on average, children who receive the grant have worse nutritional outcomes than comparably healthy children in their age cohort. This is unsurprising given the grant is targeted at the most vulnerable children (Agüero et al, 2007: 5) The negative mean *HAZ* across the sample shows, on average, children receiving the grant have not received consistently adequate nutrition. Crucially, however, the upward curve in the dose response function indicates mean *HAZ* improves as the dose of the grant increases: children who receive the grant over a greater proportion of their lives have, on average, better nutritional outcomes as measured by *HAZ*. This underscores how important the grant is for supporting nutrition among the vulnerable children who receive the grant.

In the context of this paper, the strong positive treatment effect of the grant in the continuous case is consistent with the positive and significant result using the neural network prediction in the binary treatment case. There would be an inconsistency if the binary treatment case (which should offer the lower bound of the treatment effect) found a positive and significant treatment effect, but the continuous treatment case did not. The presence of this inconsistency would call into question the finding that the neural network prediction of delay is superior. However, since this inconsistency has been avoided, my conclusion relating to the superiority of the neural network prediction in the binary treatment is not inconsistent.

Figure 3
Dose Response Function



6. Conclusion

Therefore, neural networks – which are powerful machine learning prediction tools borrowed from computer science – offer economists an alternative to OLS regression in the context of high dimension prediction in the service of estimation. While it would be too strong a claim to say neural networks are always superior to OLS regression, the example of the grant shows there are research contexts in microeconomics where using a neural network for prediction in the service of estimation may be advantageous. Neural networks are most likely to outperform OLS regression when the prediction problem is effectively high dimensional. By using a neural network to predict application delay, I found the grant has a positive and significant treatment effect in the binary case, a result consistent with both theoretical expectations, qualitative empirical research, and the finding in the continuous treatment case.

Looking to opportunities for further research, since the motivation variable constructed using a neural network has been shown to likely be superior in the context of the grant, there

is an opportunity to undertake a review of the impact of the grant on wellbeing, measured by the six outcome variables used in Coetzee (2013). This will give a more complete account of the impact of the grant on wellbeing measured by a broader set of outcome variables.

As conceded earlier, I assume a child who is currently eligible for the grant has always been eligible. There is an opportunity in future research to try to use the NIDS waves as a panel to examine whether this is a fair assumption to make. Additionally, Oyenubi (2018) shows how the entropy distance metric may be a superior balance measure compared to the standardized difference in means approach used in this study. Therefore, it may be worthwhile conducting further research using the entropy distance metric as the balance measure.

Ultimately, I hope to have shown in this study why machine learning methods such as neural networks should be given their own distinct place in the economist's toolbox, and not just because of their ability to work with large or novel datasets.

References

Agüero, J., Carter, M., & Woolard, I. (2007). The Impact of Unconditional Cash Transfers on Nutrition: The South African Child Support Grant, Working Paper 39, *International Poverty Centre*, United Nations Development Programme, pp. 1-27. https://ipcig.org/pub/IPCWorkingPaper39.pdf

Athey, S., & Imbens, G.W. (2019). Machine Leaning Methods Economists Should Know About. *Annual Review of Economics*, Vol. 11(1), pp. 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Brophy, T., Branson, N., Daniels, R.C., Leibbrandt, M., Mlatsheni, C., & Woolard, I. (2018). National Income Dynamics Study panel user manual. Release 2018. Version 1. Cape Town: *Southern Africa Labour and Development Research Unit*, pp. 1-82. http://www.nids.uct.ac.za/images/documents/20180831-NIDS-W5PanelUserManual-V1.0.pdf

Coetzee, M. (2013). Finding the Benefits: Estimating the Impact of the South African Child Support Grant. *South African Journal of Economics*, Vol. 81(3), pp. 427-450. https://doi.org/10.1111/j.1813-6982.2012.01338.x

Diamond, A & Sekhon, J.S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, Vol. 95(3), pp. 932-945. https://doi.org/10.1162/REST_a_00318

Donaldson, D. & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives*. Vol. 30(4), pp. 171-198. http://dx.doi.org/10.1257/jep.30.4.171

Einav, L. & Levin, J. (2014). Economics in the age of big data. *Science*, Vol. 346(6210), pp. 1-6. https://doi.org/10.1126/science.1243089

Hirano, K. & Imbens, G. (2004). The propensity score with continuous treatments. In A. Gelman and X. Meng (Eds.), *Applied Bayesian Modelling and Causal Inference from Incomplete-Data Perspectives*. Chichester: Wiley, pp. 73-83. https://doi.org/10.1002/0470090456.ch7

Lobell, D. B. (2013). The Use of Satellite Data for Crop Yield Gap Analysis. *Field Crops Research*. Vol. 143, pp. 56-64. https://doi.org/10.1016/j.fcr.2012.08.008

Michalopoulos, S. & Papaioannou, E. (2013). Pre-Colonial Ethnic Institutions and Contemporary African Development. *Econometrica*. Vol. 81(1), pp. 113-152. https://doi.org/10.3982/ECTA9613

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, Vol. *31*(2), pp. 87–106. https://doi.org/10.1257/jep.31.2.87

Nonyana, J. Z. & Njuho, P. M. (2018). Modelling the length of time spent in an unemployment state in South Africa. *South African Journal of Science*, Vol. 114(11-12), pp. 1-7. https://dx.doi.org/10.17159/sajs.2018/4313

Oyenubi, A. (2018). Optimizing Balance: The case of the South African Child Support Grant. Working Paper, pp. 1-31.

Patel, L. & Plagerson, S. (2016). The evolution of the Child Support Grant. *South African Child Gauge*. Cape Town. Children's Institute, University of Cape Town, pp. 39-43. http://www.ci.uct.ac.za/sites/default/files/image_tool/images/367/Child_Gauge/2006/Child_Gauge_2016-evolution_of_the_csg.pdf

Southern Africa Labour and Development Research Unit. (2018). National Income Dynamics Study 2017, Wave 5 [dataset]. Version 1.0.0 Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: *DataFirst* [distributor], 2018. https://doi.org/10.25828/fw3h-v708

South African Social Security Agency (SASSA). (2014). Child support grant. Pretoria: South African Social Security Agency. https://www.gov.za/services/child-care-social-benefits/child-support-grant

South African Social Security Agency (SASSA). (2016). You and Your Grant 2017/18. Pretoria: South African Social Security Agency, pp. 1-8. https://www.westerncape.gov.za/assets/departments/social-development/english_you_and_your_grants_2017-18_final.pdf

Statistics South Africa. (2014). Poverty Trends in South Africa. Pretoria: Stats SA, pp. 1-75. http://beta2.statssa.gov.za/publications/Report-03-10-06/Report-03-10-06March2014.pdf

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, Vol. *113*(27), pp. 7310-7315. https://doi.org/10.1073/pnas.1510479113

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, Vol. 28(2), pp. 3-28. https://doi.org/10.1257/jep.28.2.3

Zupan, J. (1994). Introduction to Artificial Neural Network Methods: What They Are and How to Use Them. *Acta Chimica Slovenica*, Vol. 41(3), pp. 327-352.

Appendix A: Caregiver Income and the Means Test

Here, I provides a more detailed explanation of how caregiver income is calculated and how I apply the means test to the NIDS data. The means test is used to determine if a child is eligible to receive the grant and depends on the income of their primary caregiver (SASSA, 2016: 4). If the calculated caregiver income is greater than the relevant threshold amount set by SASSA, the child or children associated with this caregiver are classified as ineligible to receive the grant. If the calculated caregiver income is less than the relevant threshold amount set by SASSA, the child or children associated with this caregiver are classified as eligible to receive the grant.

I calculate caregiver income as the sum of the following eight variables in the wave 5 NIDS data: (i) monthly take-home pay from primary and secondary jobs, *fwag* (ii) monthly take home pay from casual work, *cwag* (iii) monthly income form self-employment, *swag* (iv) income from a 13th cheque, *cheq* (v) income from profit share, *prof* (vi) monthly income from extra payments on a piece-rate basis, *extr* (vii) income from other bonuses from main job, *bonu* and, (viii) monthly income from rentals, *rnt*. Since these are monthly amounts, the total is then multiplied by 12 to get annual caregiver income.

The threshold amount for the means test is calculated as follows (SASSA, 2016: 7): the monthly grant amount is multiplied by 12 to get an annualised grant amount. The annualised grant amount is then multiplied by 10 to get the threshold amount. Since the grant amount changed on 1 April 2017, there are two possible threshold amounts. Table A.1 outlines how the two threshold amounts are calculated. Interviews for wave 5 of the NIDS survey were conducted between February and December 2017 (Brophy et al., 2018: 28). For those in the NIDS survey interviewed before April 2017, the threshold amount is R43200. For those in the NIDS survey interviewed after April 2017, the threshold amount is R45600. The wave 5 NIDS survey interview date is captured by the variable *intrv*.

Table A.1

How the 2017 Threshold Amounts used for the Means Test are Calculated

	Grant Amount (monthly)	Grant Amount (annualised)	Multiplier	**Threshold Amount**
Before 1 April 2017	R360	R4320	x 10	**R43200**
After 1 April 2017	R380	R4560	x 10	**R45600**

Appendix B: An Introduction to Neural Networks for Economists

To provide a more theoretical explanation of neural networks, I lean heavily on both Zupan (1994) and a presentation series by Sanderson (2017). For ease of understanding, a basic neural network with four layers and only one neuron in each layer is discussed first. The explanation is then extended to the case of multiple neurons. The simplified neural network with four layers and one neuron in each layer will be determined by three weights and three biases: six variables in total. Actual neural networks have many more. The goal here is to show how sensitive the cost function is to these six variables. From there, the aim is to show which adjustment to these terms is going to cause the most efficient decrease to the cost function.

Let y represent the value we want the last activation to be for a given training example (i.e., the desired output). For now, focus on the connection between the last two neurons. Let $a^{(L)}$ be the activation of the last neuron so the activation of the neuron before can be labelled as $a^{(L-1)}$: $a^{(L)} = \sigma(w^{(L)}a^{(L-1)} + b^{(L)})$. Note, these are not exponents but rather just a way of indexing. σ is a sigmoid function (but more generally, this is just some special non-linear function; it does not always have to be a sigmoid function; in this study, I use tanh). Therefore, the cost of this simple network for a single training example is given by $C_0(\dots) = (a^{(L)} - y)^2$. Let the weighted sum be represented by:

$$z^{(L)} = w^{(L)}a^{(L-1)} + b^{(L)} \text{ such that } a^{(L)} = \sigma(z^{(L)}) \tag{B.1}$$

The weight $w^{(L)}$, the previous activation $a^{(L-1)}$, and the bias $b^{(L)}$ are used together to compute $z^{(L)}$which then means we can compute $a^{(L)}$. The bias can be thought of as a hurdle that needs to be cleared before a neuron will be activated. $a^{(L-1)}$ is influenced by its own weight and bias and the activation of the neuron before it. The first goal is to find the sensitivity of the cost function to small changes in the weight ($\frac{\partial C_0}{\partial w^{(L)}}$). A small change in $w^{(L)}$ causes some change in $z^{(L)}$ which, in turn, causes some change to $a^{(L)}$ which directly influences the cost C_0.

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}}\frac{\partial a^{(L)}}{\partial z^{(L)}}\frac{\partial C_0}{\partial a^{(L)}} = a^{(L-1)}.\sigma'(z^{(L)}).2(a^{(L)} - y) \tag{B.2}$$

Notice this is just the chain rule. The above equation gives the sensitivity of C_0 to small changes in $w^{(L)}$.

Next, we compute the relevant derivatives:

$$\frac{\partial C_0}{\partial a^{(L)}} = 2(a^{(L)} - y) \tag{B.3}$$

The size of the derivative $\frac{\partial C_0}{\partial a^{(L)}}$ is proportional to the difference between the network's output and the desired output; if the difference is large, even slight changes to the activation stand to have a big impact on the final cost function.

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)}) \tag{B.4}$$

The sensitivity is just the derivative of whichever non-linearity is used.

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)} \tag{B.5}$$

The derivative $\frac{\partial z^{(L)}}{\partial w^{(L)}}$ shows that how much a small change to the weight $w^{(L)}$ influences the last layer activation $a^{(L)}$ depends on how strong the previous neuron $a^{(L-1)}$ is. This is the derivative with respect to weight only for the cost of a single training example (where a single training example would refer to the full set of input variables for one child in our sample). Since the full cost function involves averaging all of those costs across many training examples, its derivative requires averaging this expression that we just found ($\frac{\partial C_0}{\partial w^{(L)}}$) over all training examples.

$$\frac{\partial C}{\partial w^{(L)}} = \frac{1}{n}\sum_{k=0}^{n-1} \frac{\partial C_k}{\partial w^{(L)}} \tag{B.6}$$

This average is then just one component of the gradient vector which, itself, is built up from the partial derivatives of the cost function with respect to all the weights and biases.

$$\nabla C = \begin{bmatrix} \frac{\partial C}{\partial w^{(1)}} \\ \frac{\partial C}{\partial b^{(1)}} \\ \vdots \\ \frac{\partial C}{\partial w^{(L)}} \\ \frac{\partial C}{\partial b^{(L)}} \end{bmatrix} \tag{B.7}$$

The sensitivity of the cost to the bias is then:

$$\frac{\partial C_0}{\partial b^{(L)}} = \frac{\partial z^{(L)}}{\partial b^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}} = 1.\sigma'\left(z^{(L)}\right).2(a^{(L)} - y) \tag{B.8}$$

Notice we have just changed $\partial w^{(L)}$ for $\partial b^{(L)}$. This gives the sensitivity of C_0 to small changes in $b^{(L)}$ and $\frac{\partial z^{(L)}}{\partial b^{(L)}} = 1$.

The sensitivity of the cost function to the activation of the previous layer is given by:

$$\frac{\partial C_0}{\partial a^{(L-1)}} = \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}} = w^{(L)}.\sigma'(z^{(L)}).2(a^{(L)} - y) \tag{B.9}$$

$$\text{so } \frac{\partial z^{(L)}}{\partial a^{(L-1)}} = w^{(L)} \tag{B.10}$$

Again, even though we won't be able to directly influence that previous layer activation, it is important to keep track of. By iterating this chain rule idea backwards, we can see how sensitive the cost function is to previous weights and previous biases.

While this is a simplified example where each layer only has one neuron, the basic idea still holds true for more complicated neural networks. There will just be more indices to keep track of when there are multiple neurons in each layer. Rather than the activation of a given layer simply being $a^{(L)}$, it is now going to have a subscript indicating which neuron of layer L it is. Suppose we decide to use the letter k to index the layer (L-1) and the letter j to index the layer (L), we will have $a_k^{(L-1)}$ and $a_j^{(L)}$. For the cost, we again look at what the desired output (y) is. But now we add up the square of the difference between these (multiple) last layer activations and the desired output. Mathematically, $C_0 = \sum_{j=0}^{n_L-1}(a_j^{(L)} - y_j)^2$. Since there are many more weights, each one will have more indices to track its location. Let the weight of the edge connecting this k-th neuron to the j-th neuron be $w_{jk}^{(L)}$. As before, it is helpful to give a

name like z to the relevant weighted sum so that the activation of the last layer is just the special function, like sigmoid, applied to z:

$$a_j^{(L)} = \sigma(z_j^{(L)}) \text{ where } z_j^{(L)} = w_{jo}^{(L)} a_0^{(L-1)} + w_{j1}^{(L)} a_1^{(L-1)} + w_{j2}^{(L)} a_2^{(L-1)} + b_j^{(L)} \quad \text{(B.11)}$$

More generally, $z_j^{(L)} = \cdots + w_{jk}^{(L)} a_k^{(L-1)} + \cdots$ (B.12)

All of these are essentially the same equations as in the one-neuron-per-layer case. Again, the chain rule derivative expression describing how sensitive the cost is to a specific weight looks essentially the same as in the simpler case:

$$\frac{\partial C_0}{\partial w_{jk}^{(L)}} = \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}} \quad \text{(B.13)}$$

What does change here though is the derivative of cost with respect to one of the activations in the (L-1) layer. In this case, the difference is that each neuron in layer (L-1) influences the cost function through multiple paths. That is, on the one hand, it influences $a_0^{(L)}$ (which plays a role in the cost function) but it also influences $a_1^{(L)}$ (which also plays a role in the cost function). So, those influences must be added together such that:

$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{n_L - 1} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}} \quad \text{(B.14)}$$

Notice that this is the sum over layer L. Once the sensitivity of the cost function to the activation in the second last layer (L-1) is known, the same process is followed for all the weights and biases feeding into layer (L-1). This gives a clearer understanding of backpropagation which is the workhorse behind how neural networks learn. These chain rule expressions give the derivatives that determine each component in the gradient that helps minimise the cost of the network:

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = a_k^{(l-1)} \sigma'(z_j^{(l+1)}) \frac{\partial C}{\partial a_j^{(l)}} \quad \text{(B.15)}$$

where:

$$\frac{\partial C}{\partial a_j^{(l)}} = \sum_{j=0}^{n_{l+1}-1} w_{jk}^{(l+1)} \sigma'(z_j^{(l+1)}) \frac{\partial C}{\partial a_j^{(l+1)}} \text{ or } \frac{\partial C}{\partial a_j^{(l)}} = 2(a_j^{(L)} - y_j) \quad \text{(B.16)}$$

Appendix C: Additional Specifications and Results

In table C.1, I provide the treatment effect of the grant in the binary treatment case if GenMatch is specified with the default balance measure rather than standardised difference in means as used in Section 5.2. As can be seen in table C.1 below, the treatment effect is not significant regardless of whether the neural network or the OLS prediction of delay is used to construct motivation. This supports the use of the standardised difference in means approach.

Table C.1

Treatment Effect and Standard Error using GenMatch with the Default Balance Measure

	Neural Network	**OLS**
Estimate	0.014915	-0.0027175
Standard Error	0.023172	0.023968
t-stat	0.64367	-0.11338
AI Standard Error	0.13998	0.13836
t-stat	0.10656	-0.019641

Notes: ***p<0.01, **p<0.05, *p<0.1

AI – Abadie and Imbens (2004) standard error that adjusts for matching.