COS 4807 Assignment 1

Adriaan Louw (53031377)

June 2, 2019

Contents

List of Figures

List of Tables

1 Abstract

hello

2 Introduction

Human beings have always had a fascination with longevity. Myths like the fountain of youth or the Holy Grail are a testiment to this fact. Today, longevity and causes of mortality are studies by profesionals like Demographers and Actuaries. Trying to determine why some people or group have long lives.

This study investigates the use of Machine Learning techniques in studying the determinants of life expectancy for coutries. Indicators, shown to have some form of correlation with life expectancy, will be selected. Their relationship with life expectancy will be investigated using various techniques from Machine Learning. This analysis will seek out to prove the appropriateness of using these machine learning algorithms for use in research to find the exact correlation between indicators and life expectancy life expectancy. In the hope that the causes of long life expectancies in certain countries can be better understood. This study does not aim to prove causation between the indicators chosen and life expectancy, but rather the usfullness of machine learning algorithms as a tool.

Machine learning is used in medicine Chen & Asch (2017).

life expectancy vs mortality rate?

Machine learning techniques can find relationships in the data that regression analysis cannot Chen & Asch (2017)

cohort life expectancy vs period life expectancy (https://ourworldindata.org/life-expectancy-how-is-it-calculated-and-how-should-it-be-interpreted)

Rajkomar et al. (2018) Google uses machine learning to predict in hospital medical events for patients.

Human attempts to mathematically predict life expectance is not a new endeavour. Gompertz (1825) introduced an equation to predict life expectancy, which was modified in Makeham (1860) to create the famous Gompertz−Makeham law.

3 Literature Review

Forecasting Mortality in Developed Countries Tabeau 2001

3.1 ?Grossman?

2017 determinants of health: an economic perspective ???? 1972 The Demand for Health: A Theoretical and Empirical Investigation,

Grossman (2000)

3.2 Determinants of life expectancy

3.2.1 Income

The relationship beteen income and life expectancy has been given a lot of attention in academic circles (Preston 1975, Hu et al. 2015, Chetty et al. 2016, Oeppen 2019).

Preston (1975) was the first to show the relationship between life expectancy and per capita income. His original curve can be seen in Figure 1. As we can see from Figure 1, for low income countries, life expectancy increases rapidly with per capita income. Whereas in high income countries a small increase in per capita income does not have a large effect on life expectancy.

This relationship has also been shown in more recent studies (Chetty et al. 2016, Oeppen 2019). Even though Shkolnikov et al. (2019) found that in Russia the Preston curve is not an

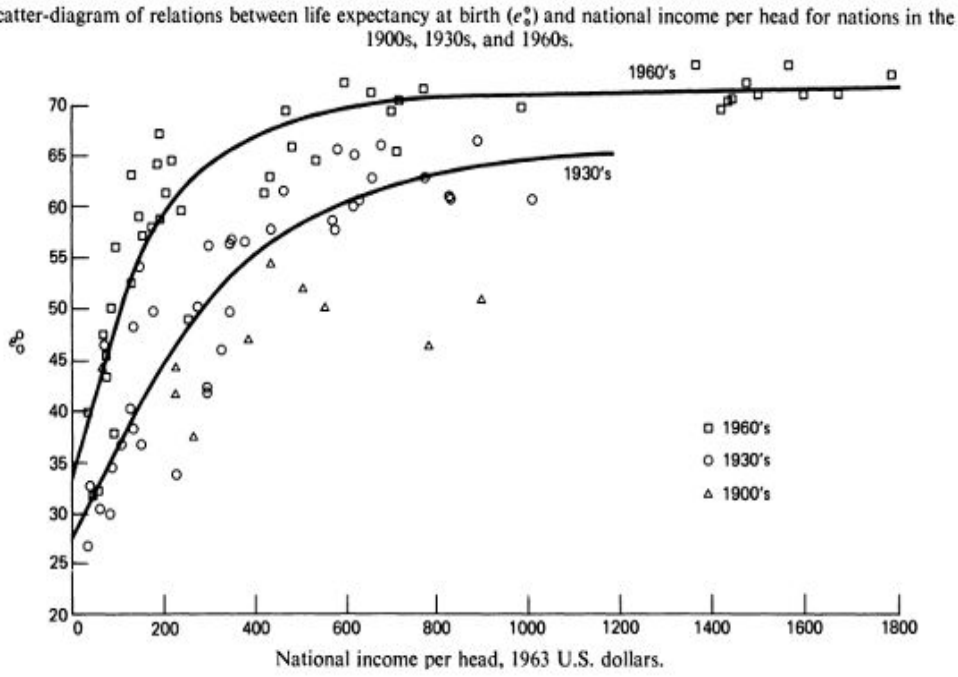

Figure 1: The original Preston curve from Preston (1975)

accurate predictor of life expectancy. They found that the actual life expectancy should be "substantially higher" when comparing to the Preston curve predicted value.

Studies in first world countries involving mortality rather that life expectancy have also found a relationship with income level (Blakely et al. 2004, Kalwij et al. 2013, von Gaudecker & Scholz 2007).

Just 16% of the increase in life expectancy between 1930s and 1960s could be explained by rising income levels Preston (1975). Which seems to indicate that a coutries life expectancy is dependant on more than income levels.

Kalwij (2014)

Oeppen (2019) Very Good!!

Preston (1975) is a seminal work according to Oeppen (2019)

inequality Hu et al. (2015)

Chetty et al. (2016) in the US

income inequality does not affect health of a a country Jason Beckfield (2004)

Tarkiainen et al. (2012) (To be downloaded)

3.2.2 Education atainment

Kaplan et al. (2015) investigated the relationship between educational atainment and life expectancy in eight states in the United States. They found that even when controlling for variables like income, race, sex and common medical issues like cardiovascular disease, the relationship between educational antainment and life expectance remains statistically significant.

Luy et al. (2019) studied 3 developed nations, namely the United States, Italy and Denmark. They have also found a strong correlation between education levels and longevity.

But what is the nature of this correlation? According to Deary & Gottfredson (2004) Intelligence Quotient or IQ could explain the association. While Hayward et al. (2015) does not believe in a "causal relationship" but rather that it depends on factors like "time, place, and the social environment".

In an attempt to find a causal relationship between education and life expectancy, van

Kippersluis et al. (2011) investigated the result of the Netherlands increasing the mandatory number of years a child had to attend school to 7 years. It was 6 years previously. van Kippersluis et al. (2011) found a decrease in mortality of 3% for 81 year old males who had the additional year of schooling.

This relationship appears strongest in more developed countries where the life expectancy is already above 60 years (Bulled & Sosis 2010). In these countries, any educational investment leads to greater compensation for the learner than they would get in a less developed country Bulled & Sosis (2010), Handwerker (1986). In addition, Kabir (2008) also studied this relationship, among others, with regards to developing countries and did not find a correlation.

The question remains, which educational indicators should be used when investicating the relationship between education and life expectancy?

Various educational indicators have been used in the literature for comparing to life expectancy. One approach is to use the International Classification of Education (ISCED) system (UNESCO Institute for Statistics 2012). The ISCED 2011 standard consists of 9 levels ranging from ISCED level 0 (Early childhood education) to ISCED level 8 (Doctoral or equivalent level).

Luy et al. (2019) used the United Nations ISCED-97 (consisting of 7 levels) scale to break education attainment down into 3 levels namely Low (None to Lower Secondary), Medium (Upper secondary) and High (Tertiary education). In van Kippersluis et al. (2011) the Dutch SOI system (Standaard Onderwijs Indeling). Which according to van Kippersluis et al. (2011) is similar to the ISCED system. While in Deboosere et al. (2009) educational attainment was broken into 5 levels also ranging from no education to Tertiarty education.

Kaplan et al. (2015) broke educational attainment into 4 levels ranging from less than high school to college graduate.

In the study Bulled & Sosis (2010), the relationship between educational invesments and fertility against life expectancy, over 193 countries, was investigated. They used adult literacy and the enrolment ratios for primary,secondary and tertiary schooling.

For more see Montez & Friedman (2015) Much information!!!!

helping individuals to mobilise health resources Elo & Preston (1996) from Deboosere et al. (2009)

Study in Belgium Deboosere et al. (2009)

Inverse relationship Hoque et al. (2019)

netherlands van Kippersluis et al. (2011)

van Baal et al. (2016)

3.2.3 Per capita spending on health

Healthcare spending and life expectancy in the Unites States, between 1960 and 2000, was compared in Cutler et al. (2006). They found that the increased spending on health per capita, controlling for inflation, is positively correlated to US life expectancy for the time period in question.

Most Eastern European countries, who have joined the European Union, have seen an increase in healthcare spending. This has generally been accompanied by an increase in life expectancy (Jakovljevic et al. 2016). This has to be seen in the light of the so called "Russian Mortality Crisis" where former Soviet Union countries faced a sudden drop in life expectancy after the fall of the Berlin wall (Brainerd & Cutler 2005). Jakovljevic et al. (2016) found that the best metric to use when comparing health spending of coutries to be their total per capita health spending in US dollars.

The same relationship was found in Canada. When spending on healthcare is decreased, life expectancy follows (Crémieux et al. 1999).

It is well known that life expectancy in Sub-Saharan Africa is low. Here spending on health care can also be correlated to increases in life expectancy. Even though poor governance can undo some of the effects of increased spending (**?**).

A countries per capita healthcare is not neccesarily in proportion to its per capita income. In 2005 the United States spent 50% more on healthcare per capita than its income per capita would suggest (Anderson & Frogner 2008).

Shaw et al. (2005) showed that pharmaceutical expenditures shows a positive correlation with life expectancy in OECD countries.

medical spending Cutler et al. (2006)

?Grossman? 2017 determinants of health: an economic perspective ???? 1972 The Demand for Health: A Theoretical and Empirical Investigation,

Grossman (2000)

3.2.4 Access to safe drinking water

3.2.5 Infant mortality

Centers for Disease Control & Prevention (1999)

3.2.6 Turmoil

(Low et al. 2008) p211

3.2.7 The gender gap

Rochelle et al. (2015)

3.2.8 Unemployment

unemployment Bonamore et al. (2015) Roelfs et al. (2011) Roelfs et al. (2015)

non linear Bonamore et al. (2015)

3.3 Cross country studies

Bulled & Sosis (2010) Pearsons r and multivariate regression.Bulled & Sosis (2010) aims to show correlation not prediction.

Shaw et al. (2005) assumes a linear model and uses regression.

Kabir (2008) investigated how well the life expectancy of 93 developing countries were predicted by indicators like income,education and fertility (among others). It classied a countries life expectancy into 3 categories. Then used a probit model where the input variables have a linear relationship. Multiple Ordinary Least Squares Regression was then applied to study inicators' influences.

The study Hu et al. (2015), also used a linear regression model of GDP per capita, Gini indeces, ect with respect to life expectancy. The intention of the study was to link income inequality to mortality rates and life expectancy.

4 Methodology/Procedure

There are many studies that attempt to extrapolate future life expectancy for countries based on current data. This includes studies for high income countries (Kontis et al. 2017) and low income countries ????cite.

This study will attempt to create a model that can predict life expectancy for a country based on various socio-economic conditions in the country.

The philosophical standpoint of this study is Positivism. By using the scientific method, this study will comprise of an experiment to inductively determine whether machine learning techniques can provide more accurate life expectancy models than those created using regression. This cross-sectional study will use life expectancy indicators, shown from the literature, to have some correlation to life expectancy.

segment data into groups where each group has the same amount of data points???

Unlike Shaw et al. (2005), this study will not take into account the age distribution of each country.

As for HDI from Bulled & Sosis (2010) Adult literacy rate

primary secondary and tertiary enrolment ratios

GDP per Capita (Purchasing power parity)

"National datasets must be regarded with some level of caution as data gaps and issues of inconsistency and incoherence remain owing to differences in the effectiveness of infrastructure, political agendas, and additional factors, such as internal conflicts" Bulled & Sosis (2010)

The impact of finishing secondary school is different before vs after the seconf world war Deboosere et al. (2009)

4.1 Choice of dataset

4.2 Choise of algorithms

4.2.1 Algorithms Chosen

Regression Linear Regression is a popular technique, used to find relationships in data. As the name suggests Linear Regression assumes a linear relationship between the input variables and the result (Murphy 2012). This might not be the case for the target function. The target function could be any potential function. In the case of life expectancy modelling, we know that according to the Preston curve (Section 3.2.1) the relationship between income and life expectancy is not linear. Thus using Linear Regression should return a sub-optimal result. The same logic applies to Logistic Regression. It assumes a linear relationhip between inputs. The difference is that this linear sum is passed through the sigmoid function (Murphy 2012). This alsom makes it inappropriate for non-linear target functions. In this study Linear and Logistic Regression will be used as a baseline for comparison on the dataset.

Ridge Regression

k-Nearest Neighbour The k-Nearest Neighbour algorithm (kNN) is an instance based form of machine learning. It uses the classification of those datapoints closest to the datapoint to be classied to determine its classification. The kN-algorithm allows for non-linear problems spaces to be classified, because it does not make an assumption on the nature of the problem space. Additionally, how the algorithm determined its output value is transparent and can be used to study how various components affects the end result. In this study the standard kNN-algoritm will be altered to accomodate a real valued output and not just a class classification. This will be accomplised by taking the mean life expectancy for all the datapoints determined to be closest to the target point. Care will have to be taken to reduce the number of features the data, because this algorithm is sensitive to the so-called "curse of dimentionality" (Mitchell 1997).

using PCA to reduce dimentions

compare different distance metrics

can be used to determine missing data

Support Vector Machines Smola & SCHöLKOPF (2004) introduction

Bayes

Priciple Component Analysis (PCA) Duda et al. (2001)

Kernel Methods Alpaydin (2010)

4.2.2 Ignored Algorithms

Neural Networks Even though Neural networks are capable of representing non-linear hypothesis spaces (Mitchell 1997), theyare not appropriate for this study for a couple of reasons. Firsly, the datasets that are available are not large enough. Neural networks typically require thousands if not tens of thousands of datapoints. Secondly the amount of processing power and processing time required, will not be available to this study. Thirdly, the results of nural networks are hard to interperet. How the Neural Network came to its conclusion is not clear to the researcher. Which makes it unsuitable as a tool to study the relationship between life expectancy and its various indicators.

Decision Trees Traceability and understandability are some of the hallmarks of Decision Trees. These algorithms are well suited problem spaces where the target function and the input attributes are discrete values. It is possible to approximate continuous input attributes by making a branch in the tree when a value is smaller or greter than some value or is between some value. For functions where input attributes span over large ranges, this leads to very large and sub-optimum trees (Mitchell 1997). The problem of determining life expectancy from socio-economic indices has a continuous target function output and continuous input attributes. Therefor, Decision Trees will be excluded from this study.

4.3 Cross-validation

By using stratified k-fold cross validation, this study will aim to reduce the impact of the relative small dataset that will be analysed. This form of cross validation will ensure that when the validation set is chosen, no important datapoints are ignored for training. The data will be broken down randomly into k subsets of equal size. Each data subset will also contain equal amounts of datapoints with low and high life expectancies, so that no dataset is completely towards one end of the data range. One data subset is chosen to be the validation subset and the remaining $k-1$ subsets are combined into the training set. The model is then trained on the training dataset andt its performance is measured against the validation subset. This is done k times in order for each subset to be the validation subset. For each of the training runs the mean of the error will be calculated (Mitchell 1997, Murphy 2012). The value of k will be dependant on the final dataset.

References

Alpaydin, E. (2010), *Introduction to Machine Learning*, second edi edn, MIT Press, London, England.
URL: *https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf*

Anderson, G. F. & Frogner, B. K. (2008), 'Health spending in OECD countries: Obtaining value per dollar', *Health Affairs* **27**(6), 1718–1727.

Blakely, T., Kawachi, I., Atkinson, J. & Fawcett, J. (2004), 'Income and mortality: The shape of the association and confounding New Zealand Census - Mortality study, 1981-1999', *International Journal of Epidemiology* **33**(4), 874–883.

Bonamore, G., Carmignani, F. & Colombo, E. (2015), 'Addressing the unemployment-mortality conundrum: Non-linearity is the answer', *Social Science and Medicine* **126**, 67–72.

Brainerd, E. & Cutler, D. M. (2005), Autopsy on an Empire: Understanding Mortality in Russia and the Former Soviet Union, Technical Report 1.
URL: *http://pubs.aeaweb.org/doi/10.1257/0895330053147921*

Bulled, N. L. & Sosis, R. (2010), 'Examining the Relationship between Life Expectancy, Reproduction, and Educational Attainment', *Human Nature* **21**(3), 269–289.

Centers for Disease Control & Prevention (1999), Achievements in Public Health, 1900-1999 Healthier: Healthier Mothers and Babies, Technical report.

Chen, J. H. & Asch, S. M. (2017), 'Machine Learning and Prediction in Medicine Beyond the Peak of Inflated Expectations', *New England Journal of Medicine* **376**(26), 2507–2509.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A. & Cutler, D. (2016), 'The association between income and life expectancy in the United States, 2001-2014', *JAMA - Journal of the American Medical Association* **315**(16), 1750–1766.

Crémieux, P.-Y., Ouellette, P. & Pilon, C. (1999), 'Health care spending as determinants of health outcomes', *Health Economics* **8**(7), 627–639.

Cutler, D. M., Rosen, A. B. & Vijan, S. (2006), 'The Value of Medical Spending in the United States, 19602000'.

Deary, I. J. & Gottfredson, L. S. (2004), 'Intelligence Predicts Health and Longevity, but Why?', *Current Directions in Psychological Science* **13**(1), 1–4.

Deboosere, P., Gadeyne, S. & Van Oyen, H. (2009), 'The 1991-2004 evolution in life expectancy by educational level in Belgium based on linked census and Population register data', *European Journal of Population* **25**(2), 175–196.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), 'Pattern Classification'.

Elo, I. T. & Preston, S. H. (1996), 'Educational differentials in mortality: United States, 1979-85', *Social Science and Medicine* **42**(1), 47–57.

Gompertz, B. (1825), 'XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c', *Philosophical Transactions of the Royal Society of London* **115**, 513–583.

Grossman, M. (2000), THE HUMAN CAPITAL MODEL, *in* 'Handbook of Health Economics, Volume 1', Vol. 1, pp. 348–407.

Handwerker, W. P. (1986), 'The Modern Demographic Transition: An Analysis of Subsistence Choices and Reproductive Consequences', *American Anthropologist* **88**(2), 400–417.

Hayward, M. D., Hummer, R. A. & Sasson, I. (2015), 'Trends and Group Differences in the Association between Educational Attainment and U.S. Adult Mortality: Implications for Understanding Education's Causal Influence *', *Soc Sci Med* **127**, 8–18.

Hoque, M. M., King, E. M., Montenegro, C. E. & Orazem, P. F. (2019), 'Revisiting the relationship between longevity and lifetime education: global evidence from 919 surveys', *Journal of Population Economics* **32**(2), 551–589.

Hu, Y., van Lenthe, F. J. & Mackenbach, J. P. (2015), 'Income inequality, life expectancy and cause-specific mortality in 43 European countries, 19872008: a fixed effects study', *European Journal of Epidemiology* **30**(8), 615–625.

Jakovljevic, M. B., Vukovic, M. & Fontanesi, J. (2016), 'Life expectancy and health expenditure evolution in Eastern EuropeDiD and DEA analysis', *Expert Review of Pharmacoeconomics and Outcomes Research* **16**(4), 537–546.

Jason Beckfield (2004), 'Does Income Inequality Harm Health? New Cross-National Evidence', *Journal of Health and Social Behavior* **45**(3), 231–248.

Kabir, M. (2008), 'Determinants of Life Expectancy in Developing Countries', *The Journal of Developing Areas* **41**(2), 185–204.

Kalwij, A. S. (2014), 'An empirical analysis of the importance of controlling for unobserved heterogeneity when estimating the income-mortality gradient', *Demographic Research* **31**(1), 913–940.

Kalwij, A. S., Alessie, R. J. M., Knoef, M. G., Kalwij, A. S., Alessie, R. J. M. & Knoef, M. G. (2013), 'The Association Between Individual Income and Remaining Life Expectancy at the Age of 65 in the Netherlands', *Demography* **50**(1), 181–206.

Kaplan, R. M., Howard, V. J., Safford, M. M. & Howard, G. (2015), 'Educational attainment and longevity: Results from the REGARDS U.S. national cohort study of blacks and whites', *Annals of Epidemiology* **25**(5), 323–328.

Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K. & Ezzati, M. (2017), 'Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble', *The Lancet* **389**(10076), 1323–1335.

Low, B. S., Hazel, A., Parker, N. & Welch, K. B. (2008), 'Influences on Women's Reproductive Lives', *Cross-Cultural Research* **42**(3), 201–219.

Luy, M., Zannella, M., Wegner-Siegmundt, C., Minagawa, Y., Lutz, W. & Caselli, G. (2019), 'The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA', *Genus* **75**(1), 11.

Makeham, W. M. (1860), 'On the Law of Mortality and Construction of Annuity Tables', *The Assurance Magazine and Journal of the Institute of Actuaries* **8**(06), 301–310.

Mitchell, T. M. (1997), *Machine Learning*, McGraw-Hill international editions - computer science series, McGraw-Hill.

Montez, J. K. & Friedman, E. M. (2015), 'Educational attainment and adult health: Under what conditions is the association causal?', *Social Science and Medicine* **127**(1), 1–7.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*.

Oeppen, J. (2019), Life Expectancy Convergence Among Nations Since 1820: Separating the Effects of Technology and Income, *in* T. Bengtsson & N. Keilman, eds, 'Old and New Perspectives on Mortality Forecasting', Springer International Publishing, Cham, pp. 197–219.

Preston, S. H. (1975), 'The Changing Relation between Mortality and level of Economic Development', *Population Studies* **29**(2), 231–248.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Liu, P. J., Liu, X., Sun, M., Sundberg, P., Yee, H., Zhang, K., Duggan, G. E., Flores, G., Hardt, M., Irvine, J., Le, Q., Litsch, K., Marcus, J., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S. L., Chou, K., Pearson, M., Madabushi, S., Shah, N. H., Butte, A. J., Howell, M., Cui, C., Corrado, G. & Dean, J. (2018), 'Scalable and accurate deep learning for electronic health records', *Digital Medicine* **1**, 18.

Rochelle, T. L., Yeung, D. K., Bond, M. H. & Li, L. M. W. (2015), 'Predictors of the gender gap in life expectancy across 54 nations', *Psychology, Health and Medicine* **20**(2), 129–138.

Roelfs, D. J., Shor, E., Blank, A. & Schwartz, J. E. (2015), 'Misery loves company? A meta-regression examining aggregate unemployment rates and the unemployment-mortality association', *Annals of Epidemiology* **25**(5), 312–322.

Roelfs, D. J., Shor, E., Davidson, K. W. & Schwartz, J. E. (2011), 'Losing life and livelihood: A systematic review and meta-analysis of unemployment and all-cause mortality', *Social Science and Medicine* **72**(6), 840–854.

Shaw, J. W., Horrace, W. C. & Vogel, R. J. (2005), The Determinants of Life Expectancy: An Analysis of the OECD, Technical Report 4.

Shkolnikov, V. M., Andreev, E. M., Tursun-zade, R. & Leon, D. A. (2019), 'Patterns in the relationship between life expectancy and gross domestic product in Russia in 200515: a cross-sectional analysis', *The Lancet Public Health* **4**(4), e181–e188.

Smola, A. J. & SCHöLKOPF, B. (2004), 'A tutorial on support vector regression', *Statistics and Computing* **14**(3), 199–222.

Tarkiainen, L., Martikainen, P., Laaksonen, M. & Valkonen, T. (2012), 'Trends in life expectancy by income from 1988 to 2007: decomposition by age and cause of death', *Journal of Epidemiology and Community Health* **66**(7), 573 LP – 578.

UNESCO Institute for Statistics (2012), *International Standard Classification of Education ISCED 2011*.

van Baal, P., Peters, F., Mackenbach, J. & Nusselder, W. (2016), 'Forecasting differences in life expectancy by education', *Population Studies* **70**(2), 201–216.

van Kippersluis, H., O'Donnell, O. & van Doorslaer, E. (2011), 'Long Run Returns to Education: Does Schooling Lead to an Extended Old Age?', *Journal of Human Resources* **46**(4), 695–721.

von Gaudecker, H. M. & Scholz, R. D. (2007), 'Differential mortality by lifetime earnings in Germany', *Demographic Research* **17**, 83–108.

Appendices