Determining the effect of change in income on self-rated health using regression models – ordered, linear, fixed or random?

Fiona Imlach Gunasekara*, Kristie Carter*, Tony Blakely*, I-Ming Liu**

*Health Inequalities Research Programme, SoFIE-Health
Department of Public Health
University of Otago
Wellington, New Zealand

*School of Mathematics, Statistics and Operations Research
Victoria University
Wellington, New Zealand

Working Paper to be presented at the British Household Panel Survey Conference
July 9-11th 2009, University of Essex, Colchester

Abstract

Many observational studies have shown an association between income and health. However, much less research has assessed how changes in income influence health over time. This could help illuminate the issue of whether income is causally related to health, if longitudinal survey data were analysed with methods to account for bias from both measured and unmeasured time-invariant confounders. The aims of this paper are to discuss econometric and biostatistical regression approaches to such analyses using panel data (i.e., fixed and random effects, and ordinal versus continuous specification of the five-level self-rated health (outcome) variable).

We used data from 16,365 adults from the first four waves (2002-2005) of the New Zealand fixed household panel Survey of Family, Income and Employment (SoFIE). The outcome was annual self-rated health (SRH; five responses: excellent, very good, good, fair, poor),and the main exposure variable was annual household equivalised income. Potential confounding variables included age, sex, ethnicity, education, employment status, marital status, family structure and area deprivation level. Random and fixed effects linear and proportional odds models were fitted using SAS 8.2.

Theoretically, we argue that the best analytical approach is to treat the self-rated heath outcome as ordinal (not continuous), and with a fixed effects methodology. A fixed effects proportional odds logistic regression model in SAS is possible using the NLMIXED procedure to compute a hybrid model. This model produces fixed effects estimates for time-varying covariates. The NLMIXED procedure can also be used to estimate the random effects proportional odds model, allowing the inclusion of time-invariant covariates.

Empirically, the random effects proportional odds models found a very small but significant positive effect of increasing income on the odds of a higher/better level of SRH. This diminished and became non-significant in the (theoretically preferred) fixed effects/hybrid proportional odds models. Qualitatively similar results were found in the linear random and fixed effects models.

Substantively, changes in income in our study have at most a small effect on SRH over the short-term at least. Econometric and biostatistical approaches can be brought together and usefully applied to panel data to answer questions of interest to both the health and economic fields.

Keywords: income, self-rated health, categorical data analysis, longitudinal survey

Corresponding author: Fiona Imlach Gunasekara, Department of Public Health University of Otago, PO Box 7343, Wellington South, phone +64 4 385 5541, fax +64 4 389 5319, email Fiona.gunasekara@otago.ac.nz

Acknowledgments: Many thanks are due to Ken Richardson for assistance and advice with the analyses.

Disclaimer: Access to the data used in this presentation was provided by Statistics New Zealand under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975. The results presented are the work of the authors, not Statistics New Zealand. The authors take full responsibility for the results and Statistics New Zealand will not be held accountable for any error or inaccurate findings within this paper.

Funding: This work was funded by the New Zealand Health Research Council through a PhD scholarship. Data collection and initial processing of the Survey of Family Income and Employment (SoFIE) has been the responsibility of Statistics New Zealand, funded by the New Zealand Government. An extension of the survey, SoFIE-Health, has been funded by the New Zealand Health Research Council.

Conflicts of interest: We declare that no conflicts of interest exist in the presentation of these data.

Introduction

Income is related to health. When those who are in poverty are compared to those who are wealthy, we find higher rates of disease, ill health and mortality, both across and within nations (Feinstein 1993; Lynch, Kaplan et al. 1997; Lin, Rogot et al. 2003). The link between income and a range of health outcomes appears intuitively obvious and has been the subject of so many investigations it hardly merits another. Or does it? Many studies of the relationship between income and health, with a few notable exceptions (e.g. (McDonough and Berglund 2003; Contoyannis, Jones et al. 2004; Jones and Wildman 2008)) are cross sectional (e.g. (Martikainen, Adda et al. 2003; Molarius, Berglund et al. 2007) or, if longitudinal, use only one income measure (e.g. (Salas 2002; Buckley, Denton et al. 2004)). The problem with cross sectional surveys, or cross sectional analyses of longitudinal data, is that this does not shed light on the causal nature of the relationship between income and health. The first issue to overcome, in determining whether income is truly related to health, is to determine the direction of the relationship. Lower income could lead to poorer health (social causation) or poorer health could lead to lower incomes (health selection). The social causation hypothesis appears to be the dominant pathway for mental health (Chandola, Bartley et al. 2003; Orpana, Lemyre et al. 2009) but for other health outcomes, the direction of effect could go either way.

An advance on cross sectional surveys, which allows for temporality and which effect precedes the other, is the longitudinal survey. However, if only one measure of income is included in an analysis of longitudinal data, this can introduce more problems of interpretation. There are two main issues to consider. First, is that income is not a static variable – income varies widely over time, at an individual and household level (Duncan 1996; Krieger, Williams et al. 1997). Longitudinal and cross sectional analyses of income mobility give different results (Jenkins 2000). A single income estimate in a longitudinal analysis of data introduces a substantial amount of measurement error, because the analyst is assuming that the income estimate for each individual is fixed when in reality it may change and fluctuate significantly. The single income measurement is an inaccurate reflection of long term income. Second, and even more important, is that any estimate of average or long term income is going to be a biased estimate. This is because at an average level, the association of income with health could be confounded by any number of other factors, some of which can be easily measured such as education, labour force activity, wealth, occupation, age and sex; but others which cannot be so easily captured in a general survey. These unmeasured factors include cognitive ability, personality, social standing and social capital, beliefs, upbringing, genetic characteristics and childhood and life experiences. Any association of 'average' income with health may be contaminated by any of these factors that have not been controlled for in the analysis – the differences between individuals. The differences that are not measured, but which we know are there, cause the bias that is most concerning.

These problems, of having a bidirectional association between the outcome and exposure or reverse causation (health selection), and of unmeasured confounders (or unobserved heterogeneity) are two examples of endogeneity (Wooldridge 2002). The first, reverse causation, may be taken into account using structural equation or simultaneous equation modelling. The second problem, of unobserved heterogeneity, is amenable to a modelling solution if we reflect that the income *changes* over time of individuals, not the average income, is free of bias from the time invariant differences that exist between or across individuals. This is because if only changes that occur in the income level of each individual are the focus of analysis, not differences in income levels between individuals, then each individual effectively acts as their own control, in terms of fixed individual characteristics that might bias the relationship between income and health (Allison 2005). If one person is much more optimistic and outgoing than another, and this leads to both their income and health being better than average, this will not cause bias in an analysis that only compares changes occurring within each individual. The extrovert optimist is being compared to herself; the introvert pessimist to himself and the differences averaged over the sample (Allison 2005). Therefore, the unmeasured factors described above are “controlled” for in the analysis. The methodology that will compute changes over time and control for fixed or time-invariant characteristics of people that would otherwise cause bias is the fixed effects family of models.

Fixed effects models are standard econometric analyses but suffer from several limiting features. Their main attraction is also their main drawback – they remove the nuisance of time-invariant confounders from the model, but with the unmeasured confounders also goes the possibility of estimating parameters for any *measured* time-invariant variables and confounders. This means that if the estimation of time-invariant variables such as sex, ethnicity, education and birthplace is important, then the fixed effects model will not be acceptable. Second, fixed effects models only work if enough change occurs in the exposure and outcome variables. If few individuals experience change, most of the data cannot be used for the analysis.

An alternative method to fixed effects is the random effects model, which allows estimation of the effects of time-invariant variables as well as time varying variables (Andress and Brockel 2007). These are typically more statistically efficient than the equivalent fixed effects model, as they use not only within-individual but also between-individual variation in the analysis (Allison 2005). However, if the between-individual variation in income, for example, is correlated with unmeasured confounders or the individual differences that are not accounted for in the model, then the estimate for income given by the random effects model will be biased.

The outcome variable for the fixed effects model must be one that changes and can be measured repeatedly over time. Self-reported health (SRH) is generally acknowledged to be a valid and useful proxy for ‘true’ health status, as seen by its ability to predict subsequent mortality (Idler and Benyamini 1997; Benyamini and Idler 1999; Burstrom and Fredlund 2001; Singh-Manoux, Dugravot et al. 2007). However, reservations have been raised about its subjective nature (Currie and Madrian 1999) and how it is possible that different subgroups of people report their health in different ways to others (van Doorslaer and Gerdtham 2003; Lindeboom and van Doorslaer 2004; Johnston, Propper et al. 2007). Putting aside these reservations, SRH as an outcome variable poses further challenges. Participants are asked to rate their health using a question in the form of: ‘Would you say your health is excellent, very good, good, fair or poor?’ This gives five possible responses and a decision has to be made whether to treat this variable as interval, assuming that the differences between adjacent categories are equal and consistent (e.g. a change from poor to fair health is equivalent to a change from very good to excellent health), or ordinal, which only assumes that there is a rank order from poor to excellent health, but the distance between the categories is unknown or not necessarily the same across different parts of the ranking. If SRH is treated as interval (1, 2, 3, 4, 5), then ordinary least squares generalised linear models can be used, which have advantages in terms of quicker computational and programming times, easier model testing and more familiar interpretations. Although SRH is an ordinal variable, it is not uncommon for SRH as an outcome to be treated as interval for the convenience of being able to use a linear model(Singh-Manoux, Martikainen et al. 2006) or as a preliminary model which is subsequently compared to a more complex type of model(Ferrer-i-Carbonell and Frijters 2004; Jones and Wildman 2008). However, treating SRH as ordinal is more ‘faithful’ to the nature of SRH measurement, makes fewer assumptions, and is probably preferable, even though categorical models are more complex and time intensive to compute. The third option, of dichotomising SRH into two categories and using logistic models, is not preferred as this will result in a loss of data and may even give different results depending on where the cut point is chosen (Altman and Royston 2006).

This paper has three objectives. The first is to estimate the effect of changes in income on SRH using four waves of data from the Survey of Income, Family and Employment (SoFIE), which is the first longitudinal survey of its kind in New Zealand. This is a preliminary result, to be extended when the survey is completed and eight waves of data are available. The second objective is to compare the results of the fixed and random effects models and see how much effect unobserved heterogeneity has on the estimates. If estimates are severely biased, then the fixed effects model, even with its limitations, will be the preferred model. The third objective is to compare the results of the categorical and linear models and determine whether treating SRH as ordinal or linear would have changed the conclusions about the effect of income on SRH.

Data and Methods

Sample

Origin of the survey

The data used in the analyses in this paper come from the Survey of Family, Income and Employment (SoFIE). SoFIE is a longitudinal household survey that is run by Statistics New Zealand, which started in October 2002 and is expected to continue for eight years, after which it will terminate. It began with funding from the Foundation for Research, Science and Technology (a government organisation) for a feasibility study for a longitudinal survey in New Zealand that would investigate the dynamics of income, labour force status and families. This was driven by the fact that similar surveys existed in many other countries (e.g. British Household Panel Survey in the UK, Panel Study of Income Dynamics in the US, Household Income and Labour Dynamics of Australia Survey in Australia, German Socioeconomic Panel in Germany, Survey of Labour and Income Dynamics in Canada, etc), providing valuable research information to assist in policy decisions, and no comparable data were available in New Zealand.

Sample Population and Selection

SoFIE is a fixed panel survey, so the same people who initially entered the survey are followed up over time and new entrants to the survey are not sought. The target population is ‘the usually resident New Zealand population living in permanent, private dwellings’(Carter, Cronin et al. 2009). By ‘usually resident’, this excludes those who are visiting New Zealand and resident for less than twelve months, and by ‘private, permanent dwellings’, this excludes those who live in institutions, boarding houses, hostels and caravans.

The sampling frame used by Statistics New Zealand divides the North Island, South Island and Waiheke Island into around 19,000 geographical areas known as Primary Sampling Units (PSUs). Each PSU includes on average 70 dwellings, but may vary in magnitude from 30 to 260 dwellings (Carter, Cronin et al. 2009).

The random sample for SoFIE was selected by a three-stage stratified cluster approach (Carter, Cronin et al. 2009). The first stage involved allocating PSUs to strata according to such factors as region, urban or rural, high or low Māori (indigenous New Zealanders) population density and socio-economic features derived from the latest census. Second, a sample of PSUs was selected from each stratum by systematic sampling, to give an independent sample of PSUs . Third, a systematic random sample was taken of the permanent private dwellings within each selected PSU, resulting in a sample of 15,000 households. From these, came a final sample of 11,500 households and over 22,000 adults, which included all eligible residents aged 15 or over of the selected dwellings who agreed to participate in the survey, giving an initial response rate of 77% (Carter, Cronin et al. 2009).

The original participants are known as original sample members (OSMs) and consist of the core SoFIE sample that was included in wave one. Each OSM is interviewed annually over the eight waves, even if they move into another household. The only other way to become an adult OSM is to be the child of an OSM at wave one. Once the children of an OSM reach the age of 15, they are also interviewed as adult OSMs, with a full range of questions. This distinction is important as after wave one, households may undergo change so that OSMs may be now living with people who are not OSMs ('non-OSMs'). While some information from these non-OSMs is essential, for example to understand household compositional changes, non-OSMs are only asked SoFIE questions while they are living with an OSM and are not followed up if they leave the OSM household. If an OSM could not be found or refused to be interviewed for two or more years, they were no longer followed up.

Data collection

Data collection for SoFIE is annual and is undertaken by face-to-face, computer-assisted interviewing in the homes of respondents. There are two main questionnaires – the household questionnaire, which is answered by one adult OSM in the household, and the personal questionnaire, which is asked of all adult OSMs of the household. The household questionnaire contains questions about family connections and the household composition, and standard of living, indicated by such things as the number of bedrooms, whether the house is rented or owned, number of vehicles owned, types of appliances owned, etc. The core personal questionnaire includes questions on demographics, income, education, labour force history, current labour force involvement and family. Every two years (Waves 2, 4, 6 and 8) information on assets and liabilities is collected to monitor net worth and savings. A health module, collecting information on health related quality of life, psychological distress, co-morbidities (e.g. stroke, diabetes, injury), lifestyle factors (smoking and alcohol consumption) perceived stress and primary care usage was developed as part of the Health Inequalities Research Programme, University of Otago, Wellington, and funded by the Health Research Council and is asked in waves three, five and seven of the survey.

There are two types of data that are collected in SoFIE – point-in-time data, which is information relevant to one time point, usually the interview date, such as self-rated health; and spell data, which is information about what happened over a period of time, such as length of employment at a particular job. The variables used in these analyses may be taken from the point-in-time or spell data.

Interviews take place continuously throughout the year by eighty to ninety interviewers. The first wave of data collection took place from 1st October 2002 to 30th September 2003, with interviews conducted evenly throughout the 12 months. This means that within any one wave of data, information from respondents has been collected throughout that calendar year, so data refers to different reference periods, because respondents are asked to recall information from the previous 12 months from the date of their interview.

Variables

This section describes the variables used in the analyses, beginning with the dependent (outcome) variable self-rated health, the main exposure variable of interest, income, then describing the time-invariant and time-varying variables. This includes how the variables are collected in SoFIE and how these are grouped or coded, if necessary. Note that some variables correspond directly to a question asked in the survey but others are derived variables, which are created by aggregating two or more questions asked within the questionnaire. Annual household income is a derived variable, for example, which is created by adding together the income received from all sources of every member of the household in question within the annual reference period (Statistics New Zealand 2005). This may end up being derived from a number of separate questions.

Most of the variables discussed below are categorical and for the purposes of the regression analysis, one category was chosen as the comparison or reference group to which the other categories can be compared. For consistency in interpretation, the reference group has been made the highest socioeconomic group for the socioeconomic position variables, and the group expected to have the best outcome for the demographic variables (e.g. ‘Married’ for marital status, ‘European/Other’ for ethnicity).

Outcome - Self-rated health

Self-rated health is asked of all adult OSMs at each interview, as: ‘In general would you say your health is excellent, very good, good, fair or poor?’ ‘Don’t know’ and ‘refused’ are two other optional responses. However, as only approximately ten people did not give a response to the SRH question in the data analysis population these were dropped from the analysis. SRH suffers from ‘ceiling’ effects, as the majority of New Zealanders rate themselves as in ‘Very Good’ or ‘Excellent’ health , which does not leave much room to improve. Relatively few (<2% from the New Zealand Health Survey 2006/07) are in the ‘Poor’ health category (Gerritsen, Stefanogiannis et al. 2008). However, the distribution of changes in SRH over time is much less skewed, with 75% of people reporting a change in SRH at some point of the four waves of SoFIE. Out of this 75%, 20% report an increase in SRH at one or more waves, 14% report a decrease in SRH at one or more waves and 41% experience both an increase and decrease.

The SRH question in SoFIE does not include a time frame, unlike in the British Household Panel Survey, where SRH over the last 12 months is explicitly asked about. Depending on how the participants choose to interpret the question, SRH in SoFIE implicitly measures health at the time of the interview. In contrast, the income question asked at the same wave measures income from the past 12 months. This creates a natural time-lag between income and SRH for these analyses.

SRH is coded as Excellent=5, Very good=4, Good=3, Fair=2, Poor=1. Therefore a positive estimate for the continuous variable income means an increase in income is associated with an improvement in SRH.

Exposure - Income

As the main exposure variable, information about income is asked of all adults in SoFIE at every wave, so that both individual and household income can be derived. Income is gross and covers the period from 12 months before the household enumeration date (the annual reference period). For example, if the household wave one interview was on 1st October 2002, the income information for wave one would be from 1st October 2001-30th September 2002. Total individual or personal income includes earnings from employment and self-employment, government benefits, private pensions and superannuation, investments and interest from bank accounts, and any other regular or one-off payments.

These analyses use annual household income, which is derived from personal income data, and is the total of income received by all adult (aged 15 years and over) individuals within a household over the annual reference period. Household income is also equivalised to account for household composition using the Jensen Index, which was created specifically for New Zealand (Jensen 1988), which works by accounting for the number and of both children and adults in the household. The equivalised annual household income is also adjusted for inflation using the Consumer Price Index (CPI) (Statistics New Zealand 2006) to the baseline income quarter 1st October 2001 to 31st December 2001 (Carter, Hayward et al. 2008).

Income polynomials in health economic analyses are common, either the addition of a quadratic term or the transformation into log(income) because the relationship between income and health is non-linear and log(income) has a linear association with health (Hauck and Rice 2004; Jones and Wildman 2005). However, in the SoFIE dataset, a scatterplot of income and SRH shows the relationship between the two to be basically linear. Various income specifications were tested in different models. Adding an income quadratic to linear models of income and SRH made no improvement to the residuals of the random effects linear model and worsened the residuals of the fixed effects model. Income as a continuous variable was used in the final models as a transformation of income was not thought to add any advantage.

The distribution of household annual income within the SoFIE dataset is highly skewed, with a small number of people with very high incomes and also some with negative incomes. These extreme high and low values resulted in convergence problems with some of the analyses and so a restricted income variable was used which included only those people who were within the 1st and 99th percentiles of income. This was in one sense a practical solution to model convergence errors but also excluded those people with negative and very low incomes who are not typical of the population we are trying to model, which is those people who receive and live off an income which is able to support them. Those with negative and zero income must have means of sustenance that are not evident from the data and do not have the usual characteristics of those with low income. This includes those who are self-employed, others who may have other means of support which they have not declared and the relatively large number of zero incomes may conceal some data error that is best to remove. Excluding some genuinely poor people from these few extreme values is unlikely to affect the results, but including these values, which will include people who are not really poor in amongst the low income people, may have an adverse effect on the results. Restricting the dataset in this way excluded approximately 1,310 observations (from a total of 65,610) from the analysis over four waves.

Non-response is more common with income questions than with many other questions, due to the sensitivity respondents may feel about disclosing their income.

Missing data and item non-response for income were imputed by Statistics NZ as so many separate items are used to derive the overall income variable.

Income is mean centred to improve interpretation, as is common practice in social science (Frees 2004). Mean centring ensures the interpretation of the intercept is not the mean response when income is zero, but the mean response for someone with average income. The mean used is that of the sample over the four waves ($55,000). Income is also scaled by a factor of $10,000 to bring the estimates closer to 1.

Time-Invariant Confounders

Age is treated as a time-invariant variable, although it changes at each wave, it changes at the same rate as the survey progresses (yearly) so it is not treated as time-varying. Age (derived from Date of Birth) is used as the age of the respondent at the household enumeration date. A non-linear relationship with age was tested by including higher order polynomial terms for age in models of income and health. The addition of the age quadratic made no difference to the overall age-SRH relationship, so this was dropped from the model.

Sex is a time-invariant covariate. Respondent's sex (male/female) is identified during their initial interview. 'Male' is used as the reference group.

Ethnicity is treated as a time-invariant variable, although it is acknowledged that ethnicity may change over time (Carter, Hayward et al. 2009). Ethnicity is collected once in the household questionnaire but is asked of each respondent during the personal questionnaire at every interview. Ethnicity is self-identified using a question similar to the NZ census and respondents can choose as many categories of ethnic group as they please. The 'New Zealand European' category is used as the reference group. Although ethnicity is assumed to be a static variable, there is some flux from waves one to four. To deal with this, the ethnicity most frequently chosen by each participant over the four waves was determined and used as the chosen ethnicity for the participants. For those who chose more than one ethnicity the same number of times, the usual prioritisation system used in health analyses was applied. The ethnic groups with the worst health outcomes were given higher priority so that Māori have highest priority, then Pacific peoples, then Asian peoples, and lastly Pakeha/New Zealand European. For example, if a participant reported that they were Pacific and European at all four waves, then they were classified as Pacific.

Education is treated as stable over the four waves. SoFIE collects information on the types and level of qualifications that respondents have attained and also at each wave, respondents are asked if they have undertaken any further study towards a qualification during the time covered by the survey questionnaire. The highest attained qualification across the four waves is used as the education level for each respondent. In this analysis, education is aggregated into four categories: 'No Qualification', 'School Qualification', 'Post-school Vocational Qualification' and 'Degree or Higher'. The reference category is 'Degree or Higher'.

In SoFIE, all adults in the household are interviewed about their assets and liabilities in waves two and four, with information collected on types of assets and their estimated worth. Net worth or total wealth is calculated by subtracting the total value of all liabilities from the total value of all assets for individuals and couples (Carter, Hayward et al. 2008). For respondents in a couple, individual wealth is calculated by dividing the couple wealth by two; for respondents not in a couple, individual wealth is used. Wealth at wave two is included in the model as a continuous variable. This is mean centred at a mean wealth of $150,000 and scaled by a factor of $100,000 to aid interpretation of the estimate.

Time-Varying Confounders

Labour force status is asked at each interview, and captures the respondent's employment status at the household enumeration date and employment history over the past year (spell data). The labour force status at the household enumeration date was used as the labour force activity variable. There are two possible categories: 'Employed' and 'Not Employed' which includes those who are seeking work and not seeking work. The reference category is 'Employed'.

Marital status is included as a time-varying variable, as marital status may change over time. The main marital status variable in SoFIE includes only legal (and not social) marital status, giving the options of 'Never married', 'Separated/divorced/widowed', and 'Married', which includes legal spouse/other partnership but not de facto couples. The reference group for this variable is 'Married'.

Family structure is a time-varying variable, as families may break up and reconstitute between waves. Statistics New Zealand collects information on 'family nuclei' which are defined as 'Couple only', 'Couple with dependent and/or adult child(ren)', 'Sole parent with dependent and/or adult child(ren)' and 'Not in a family nucleus' (e.g. single people living alone, people in flatting type situations, etc). A 'couple' includes legally married, de facto, and same-sex partnerships (Statistics New Zealand 2005). 'Couple with children' is used as the reference group.

New Zealand Deprivation 2001 is a time-varying variable reflecting area-level rather than individual deprivation. The New Zealand Deprivation index (NZDep2001) is calculated for each household, according to the decile of the dwelling location. The NZDep2001 is a deprivation scale from 1 to 10 that divides New Zealand into tenths, where 1 represents the areas with the least deprived scores and 10 the areas with the most deprived scores (Salmond and Crampton 2002). For example, a value of 10 indicates that the area is in the most deprived 10 percent of areas in New Zealand, according the NZDep2001 scores. The scales are derived from a number of variables that measure deprivation taken from Census data, such as overcrowding, income, education, access to telephone and unemployment. It is important to note that NZDep2001 measures deprivation at an area not an individual level, so a household in a high deprivation area may not necessary suffer any indicators of deprivation themselves. In these analyses, NZDep2001 is aggregated to five levels, with NZDep1 corresponding to the two least deprived areas and NZDep5 to the two most deprived areas. NZDep1 is the reference category.

The models also include dummy variables to account for the effect of time on the outcome SRH – one for each year, with the reference being the final year (wave four).

Data analysis population

The models presented are using a balanced panel of OSMs, including those adult OSMs (aged 15 and over) who responded at wave one and on whom there was complete information also in all subsequent four waves (N=16,415). The advantage of using a balanced panel is that missing data can be ignored, making computation of the models easier. The disadvantages are a loss of efficiency because of a reduced sample size (which is not so problematic for SoFIE, given the large sample size of the survey to begin with) and possibility of introducing bias, if data are not missing at random due to attrition (Vandenbroucke, von Elm et al. 2007). Figure 1 shows the attrition rate from wave one to wave four in total OSMs was 13% overall and the number of adult OSMs started at 22,165 in wave one and fell to 17,785 by wave four (20% attrition). Note that child OSMs (aged less than 15 years) can subsequently be counted as an adult OSM if they turn 15 between interview dates. Figure 2 shows the numbers of the final analysis datasets, starting with a balanced panel of 16,415 individuals then converted to observations (four per individual) with some further limits due to missing data and the truncation of the income variable. Figure 2 also details are given of missing values of the covariates used in the models.. Although the observations which are missing are not used in the analyses, the individuals with the missing data are kept in the models estimated by maximum likelihood (the linear random effects and the categorical models) under the assumption that the data are missing at random.

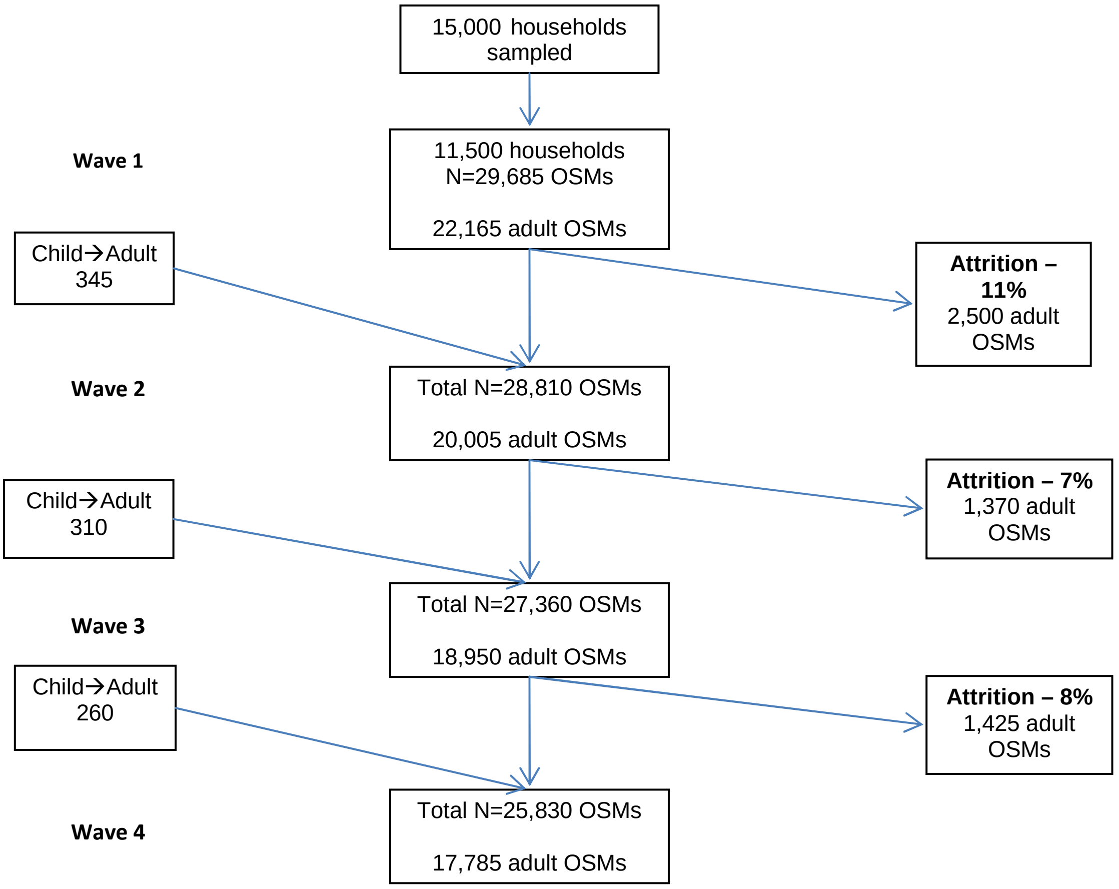
The random effects linear model, estimated using PROC MIXED, was not modelled using the full dataset due to limited computer memory capacity. Therefore, for these models only, a random sample of 40,000 observations on 10,169 individuals was taken and used as the data analysis population.

Analysis

All analyses in this paper were conducted using SAS version 8.2 in the Statistics New Zealand Data Laboratory in Wellington, New Zealand, under strict confidentiality conditions. At the current time, only the first four waves of data are available for analysis, from SoFIE data release 5. Results from regression models are given as sample estimates and are not weighted to the total New Zealand population.

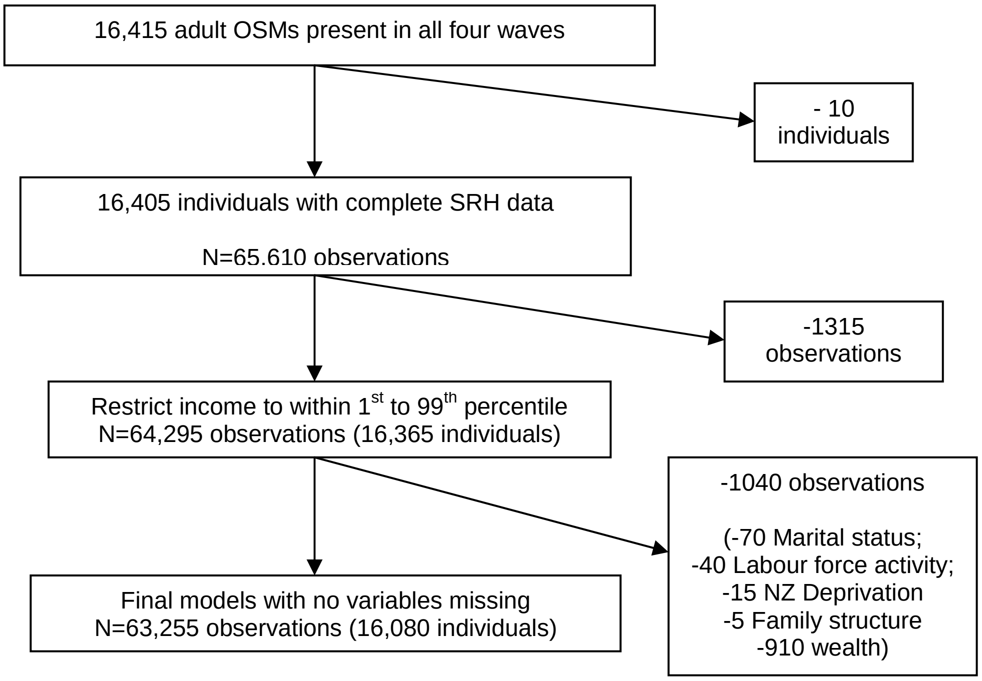
Four types of model are presented in this paper: two linear models which treat the SRH outcome as an interval variable, the fixed effects and random effects linear models and two categorical models (fixed effects and random effects), which treat SRH as an ordinal variable. The categorical models are ordered cumulative logit models of the proportional odds type, the first being a simple random effects model and the second a hybrid model with estimates that can be interpreted as if from a fixed effects model.

Figure 1: Attrition in SoFIE over waves one to four



*Numbers in this figure have been rounded to base 5 due to confidentiality reasons.

Figure 2: Flowchart of restrictions to balanced panel of analysis dataset


16,415 adult OSMs present in all four waves
- 10 individuals
16,405 individuals with complete SRH data
N=65,610 observations
-1315 observations
Restrict income to within 1st to 99th percentile
N=64,295 observations (16,365 individuals)
-1040 observations
(-70 Marital status;
-40 Labour force activity;
-15 NZ Deprivation
-5 Family structure
-910 wealth)
Final models with no variables missing
N=63,255 observations (16,080 individuals)

*Numbers in this figure have been rounded to base 5 due to confidentiality reasons.

Linear random effects models

In longitudinal data analysis, random effects models (REMs) are used as a way to model differences within (the same) individuals over time, and between (separate) individuals over time. To discuss the model, we begin with a linear regression equation representing longitudinal data. Take the following equation:

$$E(Y_{it}) = \mu_t + \beta X_{it} + \gamma Z_i + \alpha_i + \varepsilon_{it}$$

where *Y* is the outcome variable, *X* is a column vector of time-varying independent variables with co-efficient vector *β, μ_t* is the intercept which may vary over time, Z is a column vector of time-invariant independent variables with co-efficient vector γ, '*i*' refers to different individuals (individual *i*=1....n, where n is sample size) and '*t*' to different points in time (*t* = 1...T, where T is number of data collection periods). All factors of the equation can vary over time (except Z_i and α_i) and between individuals.

In longitudinal data, the error term is conceived of as having a time-invariant (α_i) and time-varying (ε_{it}) component (Verbeek 2004). These two error components are assumed to be uncorrelated with each other and with the independent variables in each time period (Wooldridge 2006). The time-invariant term, α_i is also known as unobserved heterogeneity, or the unknown and unchanging characteristics of individuals that can influence the independent and dependent variables. Unobserved heterogeneity, α_i, can be thought of as capturing all the time-invariant unknown confounders (or unknown or mismeasured covariates) that are associated with one or more independent variables in the model and impact on the outcome. The remaining part of the error term, ε_{it}, is then assumed to be the truly 'random' error varying over time. This part of the error term will contain any time-varying confounding, measurement error and random error. The problem, and potential, of longitudinal data, is how to handle the time-invariant error term α_i. Different models make different assumptions about how this is associated with other variables.

If the α_i parameter is included in the model, this can be treated as a fixed parameter (the fixed effects model) or a random vector (the random effects model). In a random effects model, the α_i component, representing the stable characteristics of individuals, is treated as a random variable, normally distributed and assumed to be independent of all other variables in the model. Random effects models (REMs) measure α_i and estimate how much individual differences influence the data. If the variance of the random effect α_i is significant, this indicates there is a lot of variability between individuals and individual differences account for much of the change in the outcome.

Random effects models assume that the exposure variables in the model are not correlated with α_i or the error term, including unobserved variables, omitted variables and unmeasured or unknown confounders. However, if α_i is correlated with the exposure variables then estimates for these may be biased, e.g. income levels vary by (unmeasured) ability which affects health.

The linear mixed (random effects) models were fitted using PROC MIXED and restricted maximum likelihood.

Linear fixed effects models

Fixed effects models (FEMs) are a type of regression model that control for time invariant variables, those that have and have not been measured. The model works by only computing the change in the dependent variable over time for each individual ('within individual change') using each individual as their own control, ignoring changes that happen between different individuals ('between individual change'), which may be biased (Allison 2005). FEMs allow for correlation between the unobserved and observed variables, whereas random effects models assume that there is no correlation between the unobserved and observed variables, which may not be an appropriate or correct assumption in many cases.

In a fixed effects model, α_i is the constant varying over individuals, treated as N fixed, unknown parameters (Verbeek 2004). The error term ε_{it} is assumed to be uncorrelated with the independent variables.

Theoretically, FEMs work by including dummy variables for each individual so graphically, would have N-1 parallel regression lines, where N is the number of individuals in the study. But calculating the model in this way is difficult for large samples and the dummy variables usually do not add useful information. A more parsimonious way to calculate the model is to subtract the mean (over time) for all time-varying variables for each individual.

Starting with : $Y_{it} = \mu_t + \beta X_{it} + \gamma Z_i + \alpha_i + \varepsilon_{it}$

Taking the individual mean of the time-varying variables:
$\overline{Y}_i = \overline{\mu}_i + \beta \overline{X}_i + \gamma Z_i + \alpha_i + \overline{\varepsilon}_i$

Then subtracting the mean from the time-varying variables to give:
$Y_{it} - \overline{Y}_t = (\mu_t - \overline{\mu}_i) + \beta(X_{it} - \overline{X}_i) + (\varepsilon_{it} - \overline{\varepsilon}_i)$

The vector of time-invariant variables, γZ_i, and α_i which also does not vary over time, are eliminated in this step. The estimate for β from this method is the 'within' estimator or the fixed effects estimator, as it looks at how changes in the time-varying exposure variable cause the outcome to vary around the mean within the individual (Verbeek 2004). This estimator does not contain 'group' level effects or the differences between individuals, only using the within individual variation to estimate β. The FEM thus controls for all time invariant predictors and confounders, whether they are measured or not, including the unchanging characteristics of people that can introduce bias into the model.

There is one major caution with using FEMs. The between-individual variation is not used because it is assumed to be correlated to some extent with α_i, but relying solely on the within-individual variation to estimate the model may lead to biased results in some cases. This can occur if the variation over time in the variables is less reliable than variation at each cross-section, for example due to significant measurement error in the independent variables or a lack of variation over time.

The linear fixed effects models were fitted using PROC GLM and putting the longitudinal subject identifier in the ABSORB statement, which eliminates the individual heterogeneity (α_i term) and performs the regression on mean-adjusted values of the outcome and time-varying variables. This gives the same result as treating α_i as a fixed term and calculating dummy variables (Allison 2005) but has the advantage of decreasing the computation time and allowing a larger sample size to be used in the analysis.

Proportional odds (categorical) random effects models

The proportional odds models, estimated using the NLMIXED procedure, are nonlinear models with the dependent variable, SRH, treated as an ordinal variable. This model allows for unobserved individual heterogeneity by including individuals as a random effect – random variables with a normal distribution and a common variance (Allison 2005), accounting for the clustering that occurs because repeated measurements that are made on the same individuals over time are more similar than those measurements made on different people. However, the model also assumes that these random effects (which account for individual variation) are not correlated with covariates included in the model. Therefore, the estimates of covariates may be biased if there is correlation between the individual effects and the covariates, as discussed above in the section on linear random effects models. REMs have been shown to be sensitive to bias because of misspecification in the context of longitudinal categorical data (Heagerty and Kurland 2001).

Since there are five categories of SRH, four intercepts are output in the model. This is because in this model, the cumulative probability of the outcome SRH is modelled:

> *Logit[P(Y≤j)],* where *j* is the number of outcome categories (with possible responses of one to five, reflecting the range from poor to excellent health).

A cumulative logit model with a five-category outcome will estimate four cumulative logits but the final logit is not included as this must be equal to one (Agresti 2007). The cumulative probability of 'good' versus 'less good' health is the probability modelled in these models, with the first cumulative logit modelling the probability of excellent health (health=5) versus the probability of any other health response; the second cumulative logit is the probability of excellent or very good health versus good, fair or poor health; as so on. Thus the estimates of covariates can be interpreted as the odds of having better SRH, at any point of the SRH scale, for each unit increase in the covariate and holding all others fixed (Agresti 2007). The important assumption of the proportional odds model that allows its presentation like that of a binary logistic model (apart from the *j-1* intercepts) is that the estimated effect of each covariate is the same over all categories of the outcome. If this were not true, then a separate parameter for each cumulative logit would have to be estimated. For example, if the effect of income was different in people with fair health from people in excellent and good health, then the proportional odds model would not be the best fitting model, as it assumes that a change in income would have the same effect over all categories of SRH. This does not mean that the probabilities of different health states do not vary by income level. For example, at an income of $10,000 the probability of excellent health may be low and the probability of poor health high, but the effect of a change in income remains the same for people of both poor and excellent health (assuming the proportional odds model is valid).

In these models, intercept 1 corresponds to the first estimated cumulative probability where health = 5 (excellent); intercept 2 gives the cumulative probability for where health is ≥ 4 (excellent or very good); and so on (Agresti 2007). These intercepts may be used to calculate the probabilities of the outcome (cumulative or category – i.e. the probability that health is ≥ 4 (excellent or very good); or the probability of health being only '4' or very good) but are not generally of interest by themselves.

PROC NLMIXED models in SAS were fitted by maximum likelihood, using Gauss-Hermite quadrature approximation and the Newton-Raphson technique to maximise parameter estimates. In order to help models to converge, parameters from the equivalent marginal (population averaged) proportional odds models, estimated using PROC GENMOD, were used as starting values.

Proportional odds (categorical) fixed effects models

An extension of the simple random effects proportional odds model is the hybrid model. This uses the same estimation techniques as described above but the model gives a within person estimate that corresponds to the result that be estimated from a fixed effects model (Allison 2005). This is done by calculating a within-individual mean (by taking the individual-specific mean from each individual's response) and an average estimate, corresponding to the between person variation (simply the individual-specific average), for all time varying covariates in the model. The individual mean centred part of the covariate is only capturing change within individuals over time and is not influenced by differences between individuals, at least in terms of the fixed characteristics of individuals, as each individual acts as their own control in this analysis. This individual mean centred estimate, for the time-varying covariates, should be free from bias from time-invariant unobserved heterogeneity and will be equivalent to an estimate from a conditional fixed effects analysis (Allison 2005). If the mean centred and the average components of the time-varying covariates are significantly different (tested using the CONTRAST statement in PROC NLMIXED), the simple random effects proportional odds model is likely to be biased. If the average or between-person variation components of the hybrid model are correlated with unmeasured confounders or unobserved heterogeneity, the coefficients of these variables are not reliable, being biased by the unobserved differences that exist between individuals. If these components are biased, then the estimates from the simple random effects proportional odds model will also be biased, as this model uses information from between-individual variation in its estimation procedure.

The hybrid models allow estimation of time-invariant variables; however, the estimates for these variables will only be free from bias if the assumption holds that there is no correlation between these variables and the error term, including α_i.

Results

The results are given in two sections. The first presents the proportional odds (categorical) models that treat SRH as an ordinal variable and contrast the random and fixed effects methodologies. The second contrasts these with the linear random and fixed effects models, which treat SRH as an interval variable.

Categorical models

Random effects proportional odds models

Two models are shown – the crude model, which gives the results for income, and the full model, which adds in all the other covariates. The results of the crude analysis are given in Table 1. In this model, income is the main exposure variable included, along with time to account for the variation of SRH over time. The estimates given are presented as odds ratios by taking the exponential of the *β* estimate. This means that for any level of SRH, an increase in household annual income of $10,000 would increase the odds of reporting a higher level of health by 1.088 ($e^{0.0841}$). The income odds ratio has a 95% confidence interval of 1.080-1.095, which is significant as it does not include the null but is nevertheless very close to zero. The first two years, compared to the final year, have estimates which give odds ratios of around 1.5, with fairly narrow confidence intervals, indicating that SRH decreases over time as estimates for Time 1 and 2 are significantly larger than the reference (time 4). This is often the case in longitudinal studies, because highest rates of attrition occur in the early stages and people in poorer states of health are more likely to attrit or be lost to the study (e.g. from death, moving into institutional care). This effect has ceased by the third year.

Table 1: Crude random effects proportional odds model with income as main covariate, SRH outcome

Income only model, N = 64,295 observations					
Variables	**Estimate**	**Standard Error**	**Odds Ratio**	**Odds Ratio 95% Confidence Interval**	
Household annual income	0.0841	0.0038	1.088	1.080	1.095
Baseline covariates					
Time 1	0.4258	0.0237	1.531	1.484	1.577
Time 2	0.3838	0.0236	1.468	1.422	1.514
Time 3	0.0371	0.0232	1.038	0.992	1.083
Time 4 (Reference)	0				
Intercept 1	-1.4463	0.0280			
Intercept 2	1.2749	0.0281			
Intercept 3	4.0014	0.0349			
Intercept 4	6.5491	0.0513			
Random effect					
α_i	2.5612	0.0227			

AIC= 143386

Table 2: Final random effects proportional odds model with income, time-invariant and time-varying covariates, SRH outcome

	Final model, N = 63,255 observations				
Variables	**Estimate**	**Standard Error**	**Odds Ratio**	**Odds Ratio 95% Confidence interval**	
Household annual income	0.0474	0.0039	1.042	1.035	1.050
Time-invariant confounders					
Age	-0.0529	0.0016	0.948	0.945	0.952
Sex – Female	0.1286	0.0407	1.137	1.057	1.217
Sex – Male (Reference)	0				
Ethnicity – Asian	-0.4187	0.0921	0.658	0.477	0.838
Ethnicity – Pacific	-0.1029	0.1084	0.902	0.690	1.115
Ethnicity – Maori	-0.3069	0.0725	0.736	0.594	0.878
Ethnicity – European/Other (Reference)	0				
Education – No qualification	-1.1634	0.0708	0.312	0.174	0.451
Education – School qualification	-0.6339	0.0666	0.531	0.400	0.661
Education – Post school qualification	-0.5779	0.0629	0.561	0.438	0.684
Education - Degree or Higher (Reference)	0				
Wealth	0.0747	0.0075	1.078	1.063	1.092
Time-varying confounders					
LFS Not employed	-0.5402	0.0330	0.583	0.518	0.647
LFS Employed (Reference)	0				
Marital status Not married	-0.1913	0.0514	0.826	0.725	0.927
Marital status DWS*	-0.1376	0.0555	0.871	0.762	0.980
Marital status Married (Reference)	0				
Family structure Couple	-0.064	0.0405	0.938	0.859	1.017
Family structure Sole parent	-0.1669	0.0612	0.846	0.726	0.966
Family structure Not in family	-0.1177	0.0501	0.889	0.791	0.987
Family structure Couple with children (Reference)	0				
NZ Dep 1	-0.6498	0.0559	0.522	0.413	0.632
NZ Dep 2	-0.5355	0.0511	0.585	0.485	0.686
NZ Dep 3	-0.3389	0.0519	0.713	0.611	0.814
NZ Dep 4	-0.2995	0.0503	0.741	0.643	0.840
NZ Dep 5 (Reference)	0				
Baseline covariates					
Time 1	0.4336	0.0240	1.543	1.496	1.58978
Time 2	0.3896	0.0237	1.476	1.430	1.5229
Time 3	0.03838	0.0234	1.039	0.993	1.085

Time 4 (Reference)	0				
Intercept 1	-0.1481	0.0702			
Intercept 2	2.5744	0.0714			
Intercept 3	5.3162	0.0754			
Intercept 4	7.8708	0.0850			
Random effect					
α_i	2.2492	0.0208			

AIC= 137754

The standard deviation of the random effect, which measures the variability of the individual subjects, is high, suggesting that the differences between individuals have an important impact on SRH.

The standard deviation of the random effect, which measures the variability of the individual subjects, is high, suggesting that the differences between individuals have an important impact on SRH.

Fixed effects hybrid proportional odds models

The results of the crude analysis are given in Table 3. In this model, income is the main exposure variable included, with estimates given for mean-centred income (the ‘fixed effects’ or ‘within-individual’ estimate) and average income (the ‘between-individual’ estimate). Time is included as a baseline factor to account for the potential variation of SRH over time. In this model, for any level of SRH, an increase in household annual income of $10,000 would increase the odds of reporting a higher level of health by 1.007 ($e^{0.0071}$). The 95% confidence interval for the odds ratio for this ‘within-individual’ income estimate is 0.998 to 1.016, which includes the null. This is even without accounting for confounding and bias from other factors, which is tested when more covariates are added to the model.
includes all covariates, both time-invariant and time-varying. The main exposure variable, income is still significant but is diminished with the addition of the other variables, from 0.0841 to 0.0474 – a reduction of over 40% from the initial crude model. The final effect of an increase of income of $10,000 on the odds of improving one’s level of SRH is 1.042 (4.2%).

Of the other covariates in the model, education, labour force status and NZ Deprivation index have strong influences on changes in SRH. Being out of employment compared to being employed reduces the odds of having better SRH by 0.583 and living in the most deprived neighbourhood compared to living in the least deprived neighbourhood reduces those odds by 0.522, with a clear declining gradient from higher to lower deprivation status. Education has the largest effect with those having no qualifications compared to at least a degree having 0.312 odds of reporting a higher level of SRH, at any given category of SRH. Having a school or post school qualification also diminishes the odds of reporting a higher level of SRH, compared to higher education, but only by about half as much as having no qualification at all.

However, as already discussed, the problem with these random effects proportional odds results is the possibility that they may be biased, as the individual differences, included in these models as the random effect, may be correlated with covariates. The next section presents fixed effects hybrid proportional odds models that control for this type of bias.

Fixed effects hybrid proportional odds models

The results of the crude analysis are given in Table 3. In this model, income is the main exposure variable included, with estimates given for mean-centred income (the ‘fixed effects’ or ‘within-individual’ estimate) and average income (the ‘between-

individual' estimate). Time is included as a baseline factor to account for the potential variation of SRH over time. In this model, for any level of SRH, an increase in household annual income of $10,000 would increase the odds of reporting a higher level of health by 1.007 ($e^{0.0071}$). The 95% confidence interval for the odds ratio for this 'within-individual' income estimate is 0.998 to 1.016, which includes the null. This is even without accounting for confounding and bias from other factors, which is tested when more covariates are added to the model.

Table 3: Crude hybrid fixed effects proportional odds model, with income as main covariate, SRH outcome

Income only model, N = 64,295 observations					
Variables	**Estimate**	**Standard Error**	**Odds Ratio**	**Odds Ratio 95% Confidence Interval**	
Household annual income – mean centred (FE)	0.0071	0.0046	1.007	0.998	1.016
Household annual income – average	0.2423	0.0066	1.274	1.258	1.290
Baseline covariates					
Time 1	0.3979	0.0237	1.489	1.421	1.559
Time 2	0.3697	0.0236	1.447	1.382	1.516
Time 3	0.0330	0.0233	1.034	0.988	1.082
Time 4 (Reference)	0				
Intercept 1	-1.4467	0.0278			
Intercept 2	1.2855	0.0279			
Intercept 3	4.0251	0.0348			
Intercept 4	6.5815	0.0512			
Random effect					
α_i	2.5190	0.0221			

AIC = 142542

Table 4: Test of fixed and random effects for initial hybrid proportional odds model

Contrasts	**F Value**	**P Value**
Test of mean centred versus average income (fixed versus random)	863.38	<.0001

In contrast to the fixed effects income estimate, the average income estimate is much more substantial, with an odds ratio of 1.274, and a confidence interval of 1.258-1.290. The estimate for the average income is more than 30 times larger than the mean centred income estimate. The results of the CONTRAST test comparing the mean centred and average income estimates are presented in Table 4. This confirms what is evident from the great disparity between these estimates – that they are significantly different. It is the between-person differences in income that account for the apparent effect of income on level of SRH. An increase in average income of $10,000, appears to increase the odds of having a better level of SRH by 1.274, but this mean income estimate is seriously biased by unmeasured confounders. It is the within-person estimate that gives the unbiased figure – the change in income that occurs with individuals acting as their own controls, so that unmeasured confounding (by time-invariant factors) is not an issue. And this gives the result that an increase in individual income of $10,000 will not result in a significant impact on SRH.

The final hybrid model (Table 5) shows that the estimates for income diminish further with the inclusion of other covariates. The final estimate for mean-centred income is 0.0060 – giving an odds ratio of 1.006 and a 95% confidence interval of 0.997 to 1.015. This compares to the odds ratio of 1.042 from the final random effects proportional odds model.

Of the other time-varying covariates in the model, only labour force status has a fixed effects (mean-centred) estimate with a 95% confidence interval that does not include

Table 5: Final hybrid fixed effects proportional odds model with income, time invariant and time varying covariates, SRH outcome

	Final model, N = 63,255 observations				
Variables	**Estimate**	**Standard Error**	**Odds Ratio**	**Odds Ratio 95% Confidence interval**	
Household annual income – mean centred (FE)	0.0056	0.0046	1.006	0.997	1.015
Household annual income – average	0.0997	0.0073	1.105	1.091	1.119

Time-invariant confounders					
Age	-0.0438	0.0019	0.957	0.953	0.961
Sex – Female	0.2324	0.0411	1.262	1.181	1.342
Sex – Male (Reference)	0				
Ethnicity – Asian	-0.8526	0.0718	0.426	0.286	0.567
Ethnicity – Pacific	-0.4312	0.0670	0.650	0.518	0.781
Ethnicity – Maori	-0.4165	0.0632	0.659	0.536	0.783
Ethnicity – European/Other (Reference)	0				
Education – No qualification	-0.8452	0.0714	0.429	0.373	0.494
Education – School qualification	-0.4360	0.0668	0.647	0.567	0.737
Education – Post school qualification	-0.3932	0.0628	0.675	0.597	0.763
Education - Degree or Higher (Reference)	0				
Wealth	0.0366	0.0076	1.037	1.022	1.052
Time-varying confounders					
LFS Not employed – mean centred	-0.1298	0.0409	0.878	0.798	0.958
LFS Employed – mean centred (Reference)	0				
LFS Not employed – average	-1.1105	0.0572	0.329	0.217	0.441
LFS Employed – average (Reference)	0				
Marital status Not married – mean centred	0.0476	0.0798	1.049	0.892	1.205
Marital status DWS* – mean centred	-0.0875	0.0848	0.916	0.750	1.082
Marital status Married – mean centred (Reference)	0				
Marital status Not married - average	-0.0818	0.0766	0.921	0.771	1.072
Marital status DWS* - average	-0.0434	0.0829	0.958	0.795	1.120
Marital status Married – average (Reference)	0				
Family structure Couple - mean centred	0.0037	0.0553	1.004	0.895	1.112
Family structure Sole parent - mean centred	0.0701	0.0802	1.073	0.915	1.230
Family structure Not in family - mean centred	-0.0049	0.0634	0.995	0.871	1.119
Family structure Couple with children - mean centred (Reference)	0				
Family structure Couple - average	-0.1491	0.0597	0.861	0.745	0.978
Family structure Sole parent – average	-0.2839	0.0967	0.753	0.563	0.942
Family structure Not in family – average	-0.2000	0.0877	0.819	0.647	0.991
Family structure Couple with children - average (Reference)	0				
NZ Dep 1 – mean centred	-0.0904	0.0839	0.914	0.749	1.078
NZ Dep 2 – mean centred	-0.0785	0.0765	0.924	0.775	1.074
NZ Dep 3 – mean centred	-0.0175	0.0767	0.983	0.832	1.133
NZ Dep 4 – mean centred	-0.0951	0.0754	0.909	0.761	1.057
NZ Dep 5 – mean centred (Reference)	0				
NZ Dep 1 – average	-0.9120	0.0750	0.402	0.255	0.549
NZ Dep 2 – average	-0.7658	0.0689	0.465	0.330	0.600
NZ Dep 3 – average	-0.4817	0.0706	0.618	0.479	0.756
NZ Dep 4 – average	-0.4063	0.0670	0.666	0.535	0.797
NZ Dep 5 – average (Reference)	0				
Baseline covariates					
Time 1	0.3992	0.0241	1.491	1.443	1.538
Time 2	0.3744	0.0238	1.454	1.407	1.501
Time 3	0.0315	0.0234	1.032	0.986	1.078
Time 4 (Reference)	0				
Intercept 1	-0.0706	0.0782			
Intercept 2	2.6584	0.0793			

Intercept 3	5.4126	0.0832			
Intercept 4	7.9798	0.0922			
Random effect					
α_i	2.2253	0.0203			

*DWS=Divorced, widowed or separated; AIC = 137259

the null (0.878, 0.798-0.958). This can be interpreted as an unemployed person has a reduced odds of 0.878 of reporting a better SRH compared to someone in employment, at any level of SRH. This is higher than the odds ratio from the random effects model of 0.583. NZ Deprivation index, which had large estimates in the final simple random effects model, has much smaller estimates in the fixed effects model which, like income, appear be confounded by between-person differences.

The time-invariant estimates in this model are fairly similar to those from the final simple random effects model. The estimates for education in the hybrid model are diminished compared to the simple random effects model but the pattern of fewer qualifications reducing the odds of better health remains the same. Given that it is difficult to know whether these variables are confounded or not, it is impossible to guess the ‘true’ effect of the time-invariant covariates on health. It seems likely, given past evidence and the effects seen here, that education is particularly important, and that age, ethnicity and sex have a role to play.

Results of the tests comparing the mean-centred and average estimates of the time-varying covariates are in Table 6. These show that most of the estimates are significantly different, except for marital status and several family status categories. For those that are significant, this indicates that the average estimates are not reliable and the estimates that depend on variation within individuals over time, which control for all time-invariant confounding, should be preferred (Allison 2005). The estimates that are not significantly different can be taken to indicate that the average estimate is not overly biased. It seems that wealth accounts for much of the confounding between marital status, family structure and SRH, as a model without wealth found fewer significant contrasts between these variables.

Table 6: Tests of fixed and random effects for final hybrid proportional odds model

Contrasts	**F Value**	**P Value**
Test of mean centred versus average income	119.78	<.0001
Test of mean centred versus average labour force status	195.00	<.0001
Test of mean centred versus average family status - couple	3.52	0.0605
Test of mean centred versus average family status – sole parent	7.95	0.0048
Test of mean centred versus average family status – not in family	3.24	0.0717
Test of mean centred versus average marital status – not married	1.37	0.2423
Test of mean centred versus average marital status – divorced, widowed or separated	0.14	0.7099
Test of mean centred versus average NZ Dep – 1	53.29	<.0001
Test of mean centred versus average NZ Dep – 2	44.58	<.0001
Test of mean centred versus average NZ Dep – 3	19.81	<.0001
Test of mean centred versus average NZ Dep – 4	9.50	0.0021

Linear models

Random and fixed effects linear models are presented in this section, for the purpose of seeing whether these simpler models, that assume that SRH is an interval variable, will give similar results to the categorical models, that have the more rigorous assumption that SRH is ordered in nature. Although the parameter estimates from the linear models cannot be directly compared to the categorical proportional odds models, the changes between the crude and full models, the changes between the random and fixed effects models and the measures of error around the estimates can be compared.

Random effects linear models

Table 7: Crude random effects linear model, income only, outcome SRH

Income only model, N = 40,000 observations				
Variables	**Estimate**	**Standard Error**	**95% Confidence Interval**	
Household annual income	0.0260	0.0014	0.0233	0.0287
Baseline covariates				
Time 1	0.1279	0.0090	0.1103	0.1455
Time 2	0.1137	0.0089	0.0963	0.1311
Time 3	0.0110	0.0089	-0.0064	0.0284
Time 4 (Reference)	0			
Intercept	3.8584	0.0101		
Random effects				
α_i (individual variation)	0.3978			
□ (residual)	0.6191			

AIC=96615.3

The random effects linear models were estimated on a subsample of the analysis population due to computer memory limitations. The crude model (Table 7) contains the main exposure variable, household annual income, and time. The estimates for income can be interpreted as an increase of $10,000 household annual income will

Table 8: Final random effects linear model, income and time-invariant and time-varying covariates, outcome SRH

	Final model, N = 39,940 observations			
Variables	**Estimate**	**Standard Error**	**95% Confidence interval**	
Household annual income	0.0124	0.0014	0.0096	0.0153
Time-invariant confounders				
Age	-0.0162	0.0006	-0.0174	-0.0150
Sex – Female	0.0537	0.0219	0.0108	0.0967
Sex – Male (Reference)	0			
Ethnicity – Asian	-0.1467	0.0351	-0.2154	-0.0780
Ethnicity – Pacific	-0.0625	0.0406	-0.1422	0.0171
Ethnicity – Maori	-0.0963	0.0272	-0.1497	-0.0429
Ethnicity – European/Other (Reference)	0			
Education – No qualification	-0.3351	0.0268	-0.3876	-0.2826
Education – School qualification	-0.1642	0.0247	-0.2126	-0.1158
Education – Post school qualification	-0.1260	0.0238	-0.1726	-0.0794
Education - Degree or Higher (Reference)	0			
Wealth	0.0251	0.0029	0.0194	0.0308
Time-varying confounders				
LFS Not employed	-0.1711	0.0117	-0.1941	-0.1481
LFS Employed (Reference)	0			
Marital status Not married	-0.0802	0.0193	-0.1181	-0.0423
Marital status DWS*	-0.0618	0.0207	-0.1023	-0.0212
Marital status Married (Reference)	0			
Family structure Couple	-0.0057	0.0154	-0.0359	0.0245
Family structure Sole parent	-0.0368	0.0216	-0.0791	0.0056
Family structure Not in family	-0.0223	0.0192	-0.0599	0.0154
Family structure Couple with children (Reference)	0			

NZ Dep 1	-0.2041	0.0210	-0.2452	-0.1630
NZ Dep 2	-0.1627	0.0191	-0.2001	-0.1253
NZ Dep 3	-0.0991	0.0194	-0.1371	-0.0611
NZ Dep 4	-0.0848	0.0187	-0.1215	-0.0481
NZ Dep 5 (Reference)	0			
Baseline covariates				
Time 1	0.1327	0.0090	0.1150	0.1504
Time 2	0.1159	0.0090	0.0983	0.1335
Time 3	0.0127	0.0090	-0.0049	0.0302
Time 4 (Reference)	0			
Intercept	4.2188	0.0313		
Random effect				
α_i (individual variation)	0.4764	0.0082		
□ (residual)	0.3976	0.0033		

AIC= 94334.8

increase the probability of better SRH by 2.6%. Similar to the categorical simple random effects models, this shows a very small effect of increased income on the probability of improving SRH, with a 95% confidence interval that does not cross the null.

The final random effects linear model (Table 8) includes the time-invariant and time-varying covariates and, as in the categorical random effects model, results in a decrease in income estimate, of approximately 50%. The final estimate for income is very small – an increase in income of $10,000 might increase the probability of improving SRH by 1.2% overall. The 95% confidence interval of 0.0096-0.0153 barely escapes the null.

The comparison between these models and the categorical models show more likenesses than differences. Labour force status, New Zealand Deprivation Index and education are three variables that have the strongest effects on SRH, with reasonably narrow confidence intervals. These linear random effects models gave results that were in good concordance with the categorical, simple random effects proportional odds models, despite the linear model using a subsample of the full balanced panel.

Fixed effects linear models

The crude fixed effects model used the same dataset as the crude fixed effects hybrid proportional odds model as there were no computational limitations with this model.

Table 9: Fixed effects linear model, income only, outcome SRH

	Income only model, N = 64,295 observations			
Variables	**Estimate**	**Standard Error**	**95% Confidence interval**	
Household annual income	0.0022	0.0013	-0.0004	0.0048
Time 1	0.1223	0.0071	0.1084	0.1361
Time 2	0.1105	0.0071	0.0967	0.1244
Time 3 (ref Time 4)	0.0114	0.0070	-0.0024	0.0252
R^2	0.715267			

The estimate for income from the fixed effects linear model (Table 9), as in the hybrid fixed effects proportional odds model, is very small with a 95% confidence interval that includes the null. Both fixed effects models give a much smaller estimate of income than the respective random effects model estimate – by a factor of six to eight. The final random effects linear model estimate for income was 0.0124 – the fixed effects linear estimate is 0.0022. (In a similar vein, the final random effects proportional odds model gave an estimate for income of 0.0474 but the final hybrid proportional odds model reduced that to 0.0056.) The models are both telling the same story – the random effects estimates are biased upwards, regardless of whether SRH is treated as a continuous or ordinal variable. The fixed effects models, taking out the between individual variability, diminishes further the small effect of increasing income on SRH. When only changes in income that occur within the same individuals are modelled, increases in income do not have any significant impact on the probability of increasing the level of SRH.

The full linear fixed effects model (Table 10) includes time varying covariates and these can be compared to the final fixed effects categorical models. The results are consistent. Only labour force status stands out as having a modest effect with a confidence interval that does not encompass the null, so those who are not employed compared to employed persons have a 5% lower probability of increasing SRH. The additional variables reduce the income estimate by a small amount.

Table 10: Fixed effects linear model, income and time-varying covariates, outcome SRH

	Final model, N = 64,165 observations			
Variables	**Estimate**	**Standard Error**	**95% Confidence interval**	
Household annual income	0.0019	0.0013	-0.0007	0.0045
Labour force status – Not employed (ref Employed)	-0.0548	0.0120	-0.0783	-0.0312
Marital status – Not married	0.0026	0.0231	-0.0427	0.0478
Marital status – Divorced, widowed or separated (ref Married)	-0.0325	0.0253	-0.0820	0.0169
Family structure – Couple	0.0007	0.0161	-0.0309	0.0322
Family structure – Sole parent	0.0098	0.0235	-0.0362	0.0558
Family structure – Not in family (ref Couple with children)	-0.0080	0.0183	-0.0439	0.0280
NZ Dep 1	-0.0215	0.0244	-0.0693	0.0262
NZ Dep 2	-0.0231	0.0220	-0.0661	0.0200
NZ Dep 3	-0.0114	0.0220	-0.0546	0.0317
NZ Dep 4 (ref NZ Dep 5)	-0.0279	0.0215	-0.0700	0.0143
Time 1	0.1237	0.0072	0.1097	0.1378
Time 2	0.1119	0.0071	0.0980	0.1258
Time 3 (ref Time 4)	0.0119	0.0071	-0.0019	0.0258
R^2	0.715524			

The linear models emerge to be consistent with the categorical models – the same conclusions would be reached from these models concerning the main exposure of income and other key variables, especially labour force status. In light of this, it would seem acceptable to use linear models in some analyses of SRH, at least for the exploratory stages. A seminal methodological paper in the ‘ordinal outcome’ field (Ferrer-i-Carbonell and Frijters 2004), tested whether treating happiness as ordinal or linear made a substantial difference and concluded that it did not – that controlling for the unobserved heterogeneity was far more important.

Discussion

This paper had three objectives:
1. to use four waves of data from the longitudinal New Zealand panel study SoFIE to see if changes in income lead to changes in SRH

2. to compare the results of fixed and random effects models to understand the importance of unobserved heterogeneity on these results and in these data
3. to compare linear and categorical models to see if treating SRH as ordinal or interval changes the conclusions reached about the income-health relationship.

To see how these objectives have been met, Table 11 compares the various estimates for income from the linear and categorical random and fixed effects models. This shows clearly that the random effects models over-estimate and give biased estimates for income compared to the fixed effects models. Unobserved heterogeneity is an important source of bias in such models and must be controlled for to get results free of confounding from both measured and unmeasured time-invariant factors. It also shows that the linear and categorical models would lead to similar conclusions about the relationship between income and SRH - that there is at best a very small positive effect of income on SRH, which reduces when more covariates are added to the model and is largely explained by bias from unobserved heterogeneity. Even if this were a real effect apart from confounding variables and measurement error, it is so small it is likely to be inconsequential compared to other factors.

Table 11: Comparison of results for income from categorical and linear models

Model	Income estimate	Odds ratio	95% Confidence Interval	
Random effects POM[1], income only	0.0841	1.088	1.080	1.095
Random effects POM, final	0.0474	1.042	1.035	1.050
Fixed effects hybrid POM, income only (FE estimate)	0.0071	1.007	0.998	1.016
Fixed effects hybrid POM, final	0.0056	1.006	0.997	1.015
Random effects linear model, income only	0.0260	-	0.0233	0.0287
Random effects linear model, final	0.0124	-	0.0096	0.0153
Fixed effects linear, income only	0.0022	-	-0.0004	0.0048
Fixed effects linear model, final	0.0019	-	-0.0007	0.0045

[1] POM=Proportional odds model

There are many extensions to this analysis that could be done in the future. It may be that the causal pathway or lag between income and health is longer than one year, which requires more waves of data and inclusion of time lags. With more waves of data, the problem of endogeneity from reverse causation may also be addressed using time lagged structural equation models. Perhaps it takes more years of data to find a meaningful effect of changing income on SRH such as that found in the British Household Panel Survey (Contoyannis, Jones et al. 2004; Jones and Wildman 2008). Or perhaps the New Zealand population is different to the UK, and income changes are of trivial importance to SRH, compared to other things, which is why country-specific surveys are necessary for policy makers. It is hoped that SoFIE will help to answer many more questions, both in New Zealand and in comparison with other longitudinal data around the world.

References

Agresti, A. (2007). An introduction to categorical data analysis. Hoboken, New Jersey, John Wiley and Sons, Inc.

Allison, P. D. (2005). Fixed effects regression analysis for longitudinal data using SAS. Cary, North Carolina, SAS Institute Inc.

Altman, D. G. and P. Royston (2006). "The cost of dichotomising continuous variables." BMJ **332**(7549): 1080-.

Andress, H. J. and M. Brockel (2007). "Income and life satisfaction after marital disruption in Germany." Journal of Marriage and the Family **69**(2): 500-512.

Benyamini, Y. and E. L. Idler (1999). "Community Studies Reporting Association between Self-Rated Health and Mortality: Additional Studies, 1995 to 1998." Research on Aging **21**(3): 392-401.

Buckley, N. J., F. T. Denton, et al. (2004). "The transition from good to poor health: an econometric study of the older population." Journal of Health Economics **23**(5): 1013-1034.

Burstrom, B. and P. Fredlund (2001). "Self rated health: Is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes?" Journal of Epidemiology and Community Health **55**(11): 836.

Carter, K., M. Hayward, et al. (2009). "How much and for who does self-identified ethnicity change over time in New Zealand? Answers from a longitudinal study." Social Policy Journal of New Zealand **In Press**.

Carter, K., M. Hayward, et al. (2008). SoFIE-Health Baseline Report: Study Design and Associations of Social Factors and Health in Waves 1 to 3. SoFIE-Health Report 2. Wellington, University of Otago.

Carter, K. N., M. Cronin, et al. (2009). "Cohort profile: Survey of Families, Income and Employment (SoFIE) and Health Extension (SoFIE-Health)." Int. J. Epidemiol. **10.1093/ije/dyp215** (http://ije.oxfordjournals.org/cgi/content/full/dyp215v1).

Chandola, T., M. Bartley, et al. (2003). "Health selection in the Whitehall II study, UK." Social Science & Medicine **56**(10): 2059-72.

Contoyannis, P., A. M. Jones, et al. (2004). "The Dynamics of Health in the British Household Panel Survey." Journal of Applied Econometrics **19**(4): 473-503.

Currie, J. and B. C. Madrian (1999). Health, health insurance, and the labor market. Handbook of Labor Economics, vol 3. O. Ashenfelter and D. Card. Amsterdam, Elsevier.

Duncan, G. J. (1996). "Income dynamics and health." International Journal of Health Services **26**(3): 419-44.

Feinstein, J. S. (1993). "The Relationship between Socioeconomic Status and Health: A Review of the Literature." The Milbank Quarterly **71**(2): 279-322.

Ferrer-i-Carbonell, A. and P. Frijters (2004). "How important is methodology for the estimates of the determinants of happiness?" Economic Journal **114**(497): 641-659.

Frees, E. W. (2004). Longitudinal and panel data. Cambridge, Cambridge University Press.

Gerritsen, S., N. Stefanogiannis, et al. (2008). A portrait of health: key results from the 2006/07 New Zealand Health Survey. Wellington, Ministry of Health.

Hauck, K. and N. Rice (2004). "A longitudinal analysis of mental health mobility in Britain." Health Economics **13**(10): 981-1001.

Heagerty, P. J. and B. F. Kurland (2001). "Misspecified maximum likelihood estimates and generalised linear mixed models." Biometrika **88**(4): 973-985.

Idler, E. L. and Y. Benyamini (1997). "Self-rated health and mortality: a review of twenty-seven community studies." J Health Soc Behav **38**: 21-37.

Jenkins, S. P. (2000). "Modelling household income dynamics." Journal of Population Economics **13**: 529-67.

Jensen, J. (1988). Income Equivalences and the Estimation of Family Expenditure on Children. Wellington, Department of Social Welfare (unpublished).

Johnston, D. W., C. Propper, et al. (2007). Comparing subjective and objective measures of health: evidence from hypertension for the income/health gradient. Discussion Paper No. 2737. Bonn, Institution for the Study of Labor (IZA).

Jones, A. M. and J. Wildman (2005). Disentangling the relationship between health and income. HEDG Working Papers, 05/07. York, Health, Econometrics and Data Group: University of York.

Jones, A. M. and J. Wildman (2008). "Health, income and relative deprivation: Evidence from the BHPS." Journal of Health Economics **27**(2): 308-324.

Krieger, N., D. R. Williams, et al. (1997). "Measuring social class in US public health research: concepts, methodologies, and guidelines." Annual Reviews of Public Health **18**: 341-78.

Lin, C. C., E. Rogot, et al. (2003). "A further study of life expectancy by socioeconomic factors in the National Longitudinal Mortality Study." Ethnicity & Disease **13**(2): 240-247.

Lindeboom, M. and E. van Doorslaer (2004). "Cut-point shift and index shift in self-reported health." Journal of Health Economics **23**(6): 1083–1099.

Lynch, J. W., G. A. Kaplan, et al. (1997). "Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning." New England Journal of Medicine **337**(26): 1889-95.

Martikainen, P., J. Adda, et al. (2003). "Effects of income and wealth on GHQ depression and poor self rated health in white collar women and men in the Whitehall II study." Journal of Epidemiology and Community Health **57**(9): 718-23.

McDonough, P. and P. Berglund (2003). "Histories of poverty and self-rated health trajectories." Journal of Health & Social Behavior **44**(2): 198-214.

Molarius, A., K. Berglund, et al. (2007). "Socioeconomic conditions, lifestyle factors, and self-rated health among men and women in Sweden." European Journal of Public Health **17**(2): 125-34.

Orpana, H. M., L. Lemyre, et al. (2009). "Income and psychological distress: The role of the social environment." Health Reports **20**(1): 21-28.

Salas, C. (2002). "On the empirical association between poor health and low socioeconomic status at old age." Health Economics **11**(3): 207-220.

Salmond, C. and P. Crampton (2002). NZDep2001 Index of Deprivation. Wellington, Department of Public Health, University of Otago.

Singh-Manoux, A., A. Dugravot, et al. (2007). "The association between self-rated health and mortality in different socioeconomic groups in the GAZEL cohort study." Int. J. Epidemiol. **36**(6): 1222-1228.

Singh-Manoux, A., P. Martikainen, et al. (2006). "What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies." J Epidemiol Community Health **60**(4): 364-372.

Statistics New Zealand (2005). Survey of Family, Income and Employment Dynamics (Year ended 30 September 2003) - Reference Report. Wellington, Statistics New Zealand.

Statistics New Zealand (2005). Survey of Family, Income and Employment Dynamics (Year ended 30 September 2003) - Reference Report. . Wellington, Statistics New Zealand.

Statistics New Zealand (2006). Consumers Price Index Review. Wellington, Statistics New Zealand.

van Doorslaer, E. and U. G. Gerdtham (2003). "Does inequality in self-assessed health predict inequality in survival by income? Evidence from Swedish data." Soc Sci Med **57**(9): 1621-9.

Vandenbroucke, J. P., E. von Elm, et al. (2007). "Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration." PLOS Medicine **4**(10): e297.

Verbeek, M. (2004). A guide to modern econometrics. Chichester, John Wiley and Sons.

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge, Massachusetts, MIT Press.