
# Scottish Longitudinal Study (SLS)
# Research Working Paper Series

## Research working paper 10

# Estimating an occupational based wage in the census: A mixed model approach to generate empirical bayes estimates

**Chris Dibben**
**Tom Clemens**

Department of Geography and Geoscience, University of St Andrews, UK
Email: tc245@st-andrews.ac.uk

**SLS Office**
**National Records of Scotland**
**Ladywell House**
**Ladywell Road**
**Edinburgh**
**EH12 7TF**
**Tel 0131 314 4210**

**Put online: 9 October 2012**

# 1. Introduction

Knowing the income, wealth and financial circumstances of individuals and groups is often central to a number of key topics including studies of migration, housing, demography and health. Income, for example, is often considered one of the key social determinants of health (Galobardes, Shaw et al. 2006) and has been linked with mental health (Mullis 1992; Lundberg and Fritzell 1994; Lynch, Kaplan et al. 1997; Benzeval and Judge 2001), mortality (Menchik 1993; Wolfson, Rowe et al. 1993; Duncan 1996; McDonough, Duncan et al. 1997) and self-assessed health (Frijters, Haisken-DeNew et al. 2005; Jones and Wildman 2008). However, despite this research imperative for income information, it is not included in the UK national census with national income figures instead based on smaller scale social surveys and government databases. This is unfortunate as the census has a number of advantages for social research including complete population coverage and information on a wide selection of different and important social characteristics.

Commonly, a lack of income information in the census is mitigated analytically through the use of proxy measures. Area based deprivation scores are the most common of these in which various domains of deprivation, obtained from the census, are aggregated to different spatial scales to give an indication of the characteristics of an area. Furthermore, other measures of socio-economic position such as occupational social class and education are often used to proxy some of the effect of material and financial circumstances. Though both measure independent components of socio-economic confounding (SEP) they may not capture entirely the health effects of income (Galobardes, Shaw et al. 2006). An alternative approach is to produce an estimated synthetic measure using regression modelling. However, whilst such a technique has been utilised for the production of aggregated area estimates of income (Williamson and Voas 2000), the potential for estimates generated at the individual scale, within the UK census, have yet to be investigated. The aim of this working paper is to investigate the usefulness of multilevel models based on Standard Occupation Classification (SOC) groups to estimate an occupational based wage measure in the census.

# 2. Methods

### 2.1 Data

Individual wage was obtained from the Labour Force Survey (LFS). The LFS is a population representative sample of individuals over a large number of years including both census years in the SLS (1991 and 2001). In addition it also contains demographic information including occupational type from the SOC which can also be found in the census. This study used data from all quarters in the years 2001-2005 and 2007-2010. Data for 2006 was excluded from the analysis and was used to validate the final models and to compare the accuracy of the estimates. For the remainder of the paper 2006 data shall be referred to as validation data and the original data as the master data.

Both the master and validation data samples were restricted to individuals of working age which was 16-64 for men and 16-60 for women. The LFS is conducted quarterly

with each quarter consisting of 53,000 UK households. The study design of the LFS incorporates a longitudinal element as each quarter contains five waves of data with each wave containing approximately 11,000 households. This design means that households are interviewed five times in consecutive quarters of the survey and thus each successive quarter contains an 80% sample overlap with the previous quarter. Therefore, when pooling data from many years of the survey it is necessary to exclude data from waves 2 to 5 to avoid analysing households more than once (LFS 2007). Finally, individuals missing information for wage, SOC, age and sex were also excluded leaving a final sample size of 232,108 in the original data from which our models were estimated. The validation sample with the same sample restrictions contained 27560 cases.

### 2.2 Standard Occupational Classification (SOC)

Our estimates were based upon the standard occupational classification variable which was originally developed by the Office of National Statistics in 1990 and has since been revised in 2000. The revised version was designed to better capture variations in managerial occupations and the rise of the IT industry for example. The revised version was used in this study. The structure of SOC is hierarchical and is designed to capture information on an individual's position within the hierarchy of their particular sector. The SOC contains four levels defined as follows; unit groups (n = 353) nested in minor groups (n = 81) nested in sub-major groups (n = 25) nested in major groups (n = 9). Each descending level provides increasingly detailed descriptions of occupational type with the main purpose being to capture the kind of work performed and the competence required to complete tasks and duties. At the coarsest level (major groups) occupations are categorised into the following broad categories; Managers and Senior Officials, Professional Occupations, Associate professional and technical occupations, Administrative and secretarial occupations, Skilled trades occupations, Personal service occupations, Sales and customer service occupations, Process, plant and machine operatives and Elementary occupations (ONS 2000).

### 2.4 Deriving a measure of occupational wage

Our wage measure was derived from a variable recording 'gross weekly income in main job'. This variable was adjusted in a number of way; firstly to account for inflation, secondly to remove outliers in the wage distribution and finally to log transform the remaining values. We used the consumer price index (CPI) of inflation which measures annual price increases of the basket of essential goods and services. The CPI rather than the Retail Price Index (RPI) was used as it is the more recent measure (developed in 1996) and is also the measure that is used by the government as the official measure of inflation (Pike, Marks et al. 2008). The CPI figures were used to calculate compound inflation i.e. a cumulative adjustment based on the previous years adjusted figures and were used to adjust salaries to match those in 2006 since this was the year from which data would be used to validate the models. Thus, adjusting wage figures from 2003 for example, we adjust upwards for 2004 inflation, and then adjust these figures upwards for 2005 and so on up to 2006.

The second adjustment removed outliers from the right tail of the wage distribution. These cases will have a disproportionate impact on mean wage and are, by definition,

unusual. The extent to which a mean, with these values included, is representative of the average wage for that whole SOC grouping is therefore questionable (ie at another 'sampled' time point they might well not be in the group and therefore the mean wage would be quite different). The removal of these cases will have introduced error into the final estimates for the excluded individuals. However, as the unit of research interest is usually the individual and this procedure affects very few individuals, then the error for most analyses will be negligible. Furthermore, the purpose of these estimates is not to exactly reproduce any individual's wage level but more to capture the broader underlying patterns of wage levels across occupational groups. Exclusion of 'outlying' wage values is therefore appropriate in this case. Trimming of these cases was achieved by removing a proportion of the right tail of the wage distribution for each minor category of SOC. The optimum cut-off was determined through sensitivity analysis comparing the skewness of the trimmed distributions at 95%, 99% and 100% cut-off values. A 95% cut-off reduced the skewness of the overall wage distribution by 97% and this was the approach used. In order to compare the effectiveness of the models appropriately we applied the same 95% cut-off to the validation data. Finally, we took the natural log of the wage values in order to account for the general right tail shape of wage distributions and fitted models to this transformed variable.

### 2.5 Modelling approach and validation

We estimated prediction models from the master dataset using a number of modelling approaches. We began by estimating the geometric mean and variance of the wage distribution in order to provide a baseline comparison. Next, we fitted four mixed models with different random effects. For the purposes of this study level one shall refer to the individual, level two to the unit group level of SOC and level three to the minor group level of SOC. In the first model we fitted a two level mixed model with level two consisting of random intercepts only for each of the 81 SOC minor groups. The equation for this model was as follows:

$$\log(\text{wage}_{ij}) = \beta_1 + \beta_2 \text{age}_{2ij} + \beta_3 \text{sex}_{3ij} + \zeta_j + \varepsilon_{ij} \qquad (1)$$

Where $\log(\text{wage}_{ij})$ is the log transformed weekly wage for the $i^{th}$ individual in the $j^{th}$ SOC minor group, $\beta_1$ is the grand intercept, $\beta_2 \text{age}_{2ij}$ and $\beta_3 \text{sex}_{3ij}$ are the fixed age and sex coefficients respectively for the $i^{th}$ individual in the $j^{th}$ SOC minor group, $\zeta_j$ is the random intercept for the $j^{th}$ SOC minor group and $\varepsilon_{ij}$ is the random error term corresponding to the deviation of the $i^{th}$ individuals wage from $\zeta_j$. Thus this model adds a random correction term to the fixed effects which adjusts for the variation in wage across SOC minor groups. In the third model, we added an interaction in the random effects between age and the level two SOC minor groups to allow the slopes and the intercepts of the random effects to vary. This model therefore allowed for the differential effect of age across different occupational groups. The equation for this model is given by:

$$\log(\text{wage}_{ij}) = \beta_1 + \beta_2 \text{age}_{2ij} + \beta_3 \text{sex}_{3ij} + \zeta_{1j} + \zeta_{2j} \text{age}_{ij} + \varepsilon_{ij} \qquad (2)$$

Thus, $\zeta_{2j} \text{age}_{ij}$ is a term for the age of the $i^{th}$ individual in the $j^{th}$ SOC minor group. In this model the error term comprises the deviation of the $i^{th}$ individuals wage from a SOC minor specific regression line with an age specific slope. Thus, this model

simply adds a random effect, $\zeta_{2j}age_{ij}$, allowing age to vary across SOC minor groups. The final two models add an additional level to the models corresponding to SOC unit.

$$\log(wage_{ij}) = \beta_1 + \beta_2 age_{2ikj} + \beta_3 sex_{3ikj} + \zeta_{kj} + \zeta_k + \varepsilon_{ikj} \qquad (3)$$

$$\log(wage_{ij}) = \beta_1 + \beta_2 age_{2ikj} + \beta_3 sex_{3ikj} + \zeta_{1kj} + \zeta_{2kj} age_{ikj} + \zeta_{1k} + \zeta_{2k} age_{ikj} + \varepsilon_{ikj} \qquad (4)$$

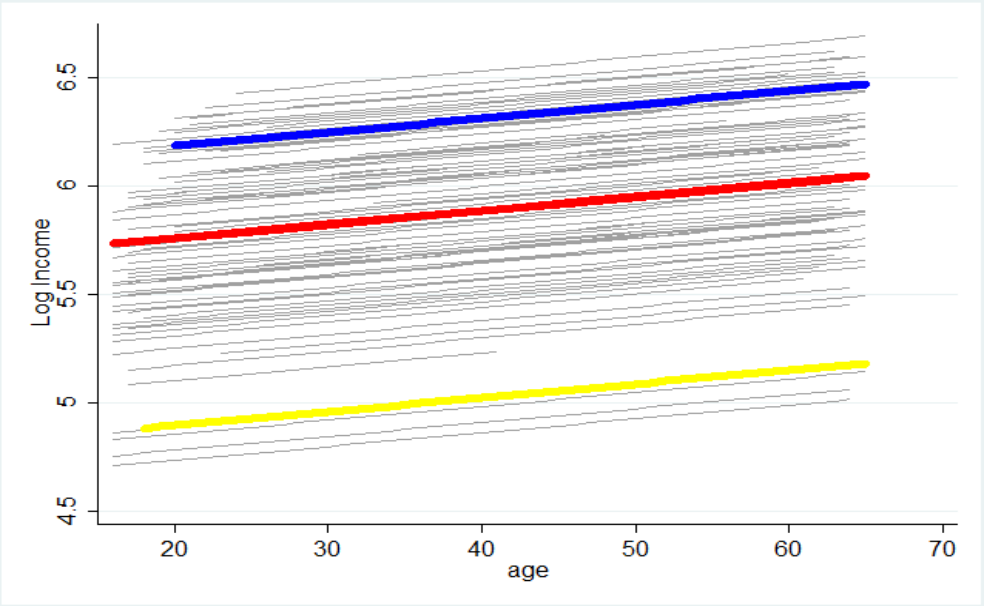
Equation three includes random intercept terms for unit and minor SOC levels while equation four additionally includes random slopes for age at both levels. Thus SOC unit groups are nested within SOC minor group, which themselves are allowed to have varying intercepts and slopes for age around the slope corresponding to the higher level SOC minor category.

The fit of each these models was assessed within the original data by examining the variance of the residuals across models. This was appropriate as each of the models was fitted to the same sample. All of the models were fitted using STATA version 11 and the mixed models were estimated using the xtmixed command. Once the models were fitted we extracted the fixed effect coefficients for age, sex and the intercept and the random effects parameters. Because the random elements of the mixed models are not estimated directly the predicted values were calculated using a best linear unbiased estimator (BLUP) - an empirical bayes estimator. This estimator has the helpful characteristic of not over-fitting to the observed data, unlike say the directly derived estimate. Where there are few cases in a particular grouping the estimate is 'shrunk' towards the mean of the higher level. The degree of shrinkage is inversely proportional to the variance and standard error of the values being estimated. These shrunk parameter estimates, together with the fixed coefficients were attached to the validation data and used to construct estimates of log wage according to equations one to four. The predicted values in the validation data were exponentiated to return them to the standard scale. The accuracy of the predictions was evaluated across models by calculating the average distance of the predicted wage from the actual wage in the validation data.

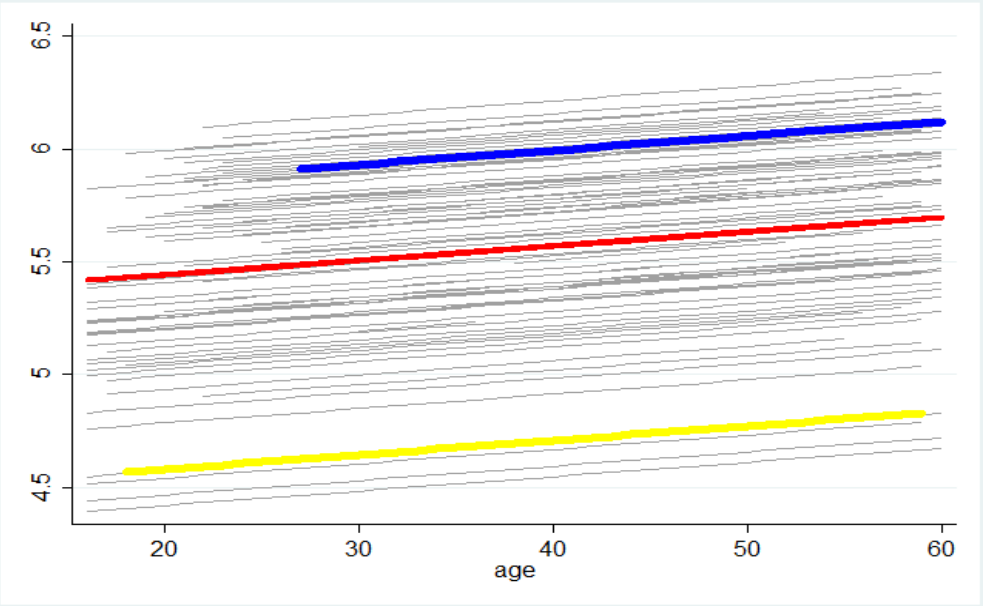
## 3. Results

The variability of the random effects used to predict the log of occupational wage are displayed in panels one and two with line graphs displaying individual regression lines in grey for each SOC category. Each graph also displays the fixed effect regression line from each model in red. The results for men and women are presented separately in order to compare gender effects and to allow the graphical representation of the random effects. Each graph effectively illustrates the variability of mean occupational wage within SOC level relative to the grand mean represented by the fixed effect line. To aid interpretation and to provide a comparison of high and low earning occupations, the random effects for 'corporate managers and senior officials' and 'elementary security officials' are highlighted in blue and yellow respectively.

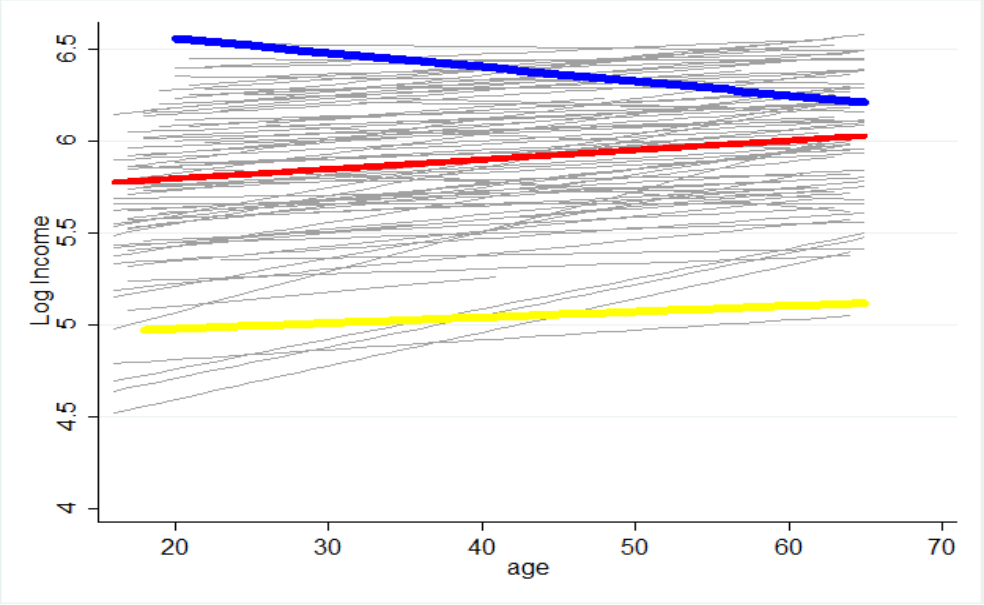
Panel one: Regression lines from mixed models with two levels predicting occupational wage. Fixed effects lines in red and SOC minor random effects in grey. For comparison, blue and yellow lines represent SOC minor groups 111 (Corporate managers and senior officials) and 924 (Elementary security occupations) respectively.



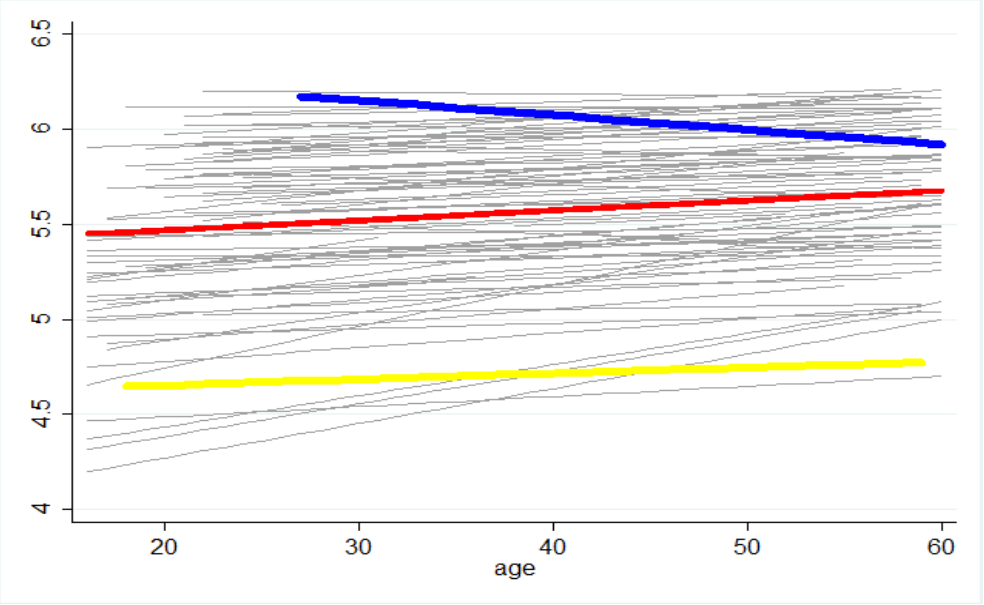

A) Random intercept model (men)




B) Random intercept model (women)




C) Random intercept with random age slopes (men)



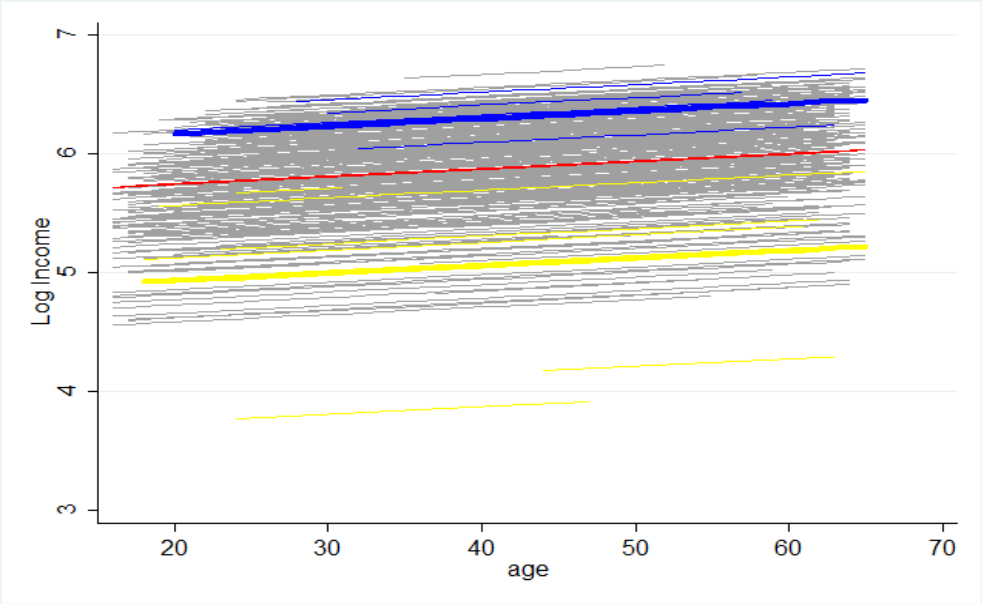

D) Random intercept with random age slopes (women)

The graphs in panel one display effects from the two level models with figures A (males) and B (females) showing results from the random intercept model (equation 1) and figures C and D showing results from the random intercept and age slope model (equation 2). Figures A and B demonstrate significant variation in occupational mean wage across SOC minor categories and also demonstrate a substantial differential in wage between men and women which can be examined by comparing the fixed effect lines. This figure equates to men earning approximately £80 more than women. Comparing the yellow and blue lines indicates significant difference in wage between these occupations of approximately £400. Furthermore, the slope of the lines indicates that mean wage increases with age with figures C and D illustrating that this age effect also varies considerably between occupational groups. Interestingly, the blue line in figures C and D indicates that age has an inverse relationship with log wage in these particular groups.

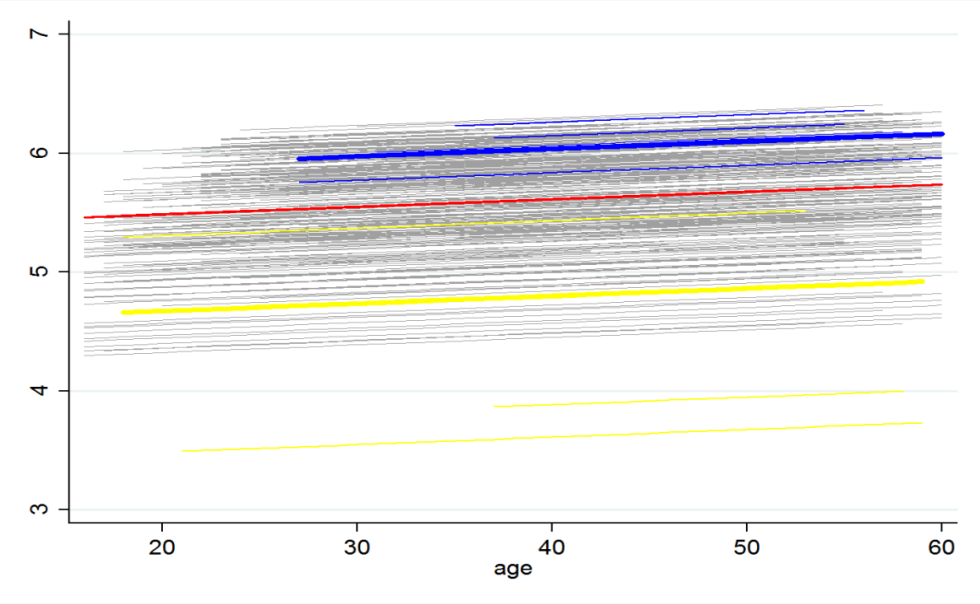
Panel two displays results from the models which include information from level two of the SOC structure (equations three and four). Again, regression lines from the higher and lower earning occupations outlined above are illustrated. The thinner yellow and blue lines indicate random effects at level three relative to the corresponding higher level regression lines which are comprised of the level two portions of equation three and four. The spread of the lines in the graphs illustrates significant variation in mean wage across unit occupation groups which is evident from the variation in the intercepts of each of the regression lines. Similarly, the regression lines from the random slopes model indicates that the effect of age for wage also varies significantly between level 3 occupation groups. Furthermore, the blue level three lines show a positive effect for age with the exception of senior officials in local government and special interest organisations.

Table two displays summary statistics of the prediction errors of the estimates that were calculated from equations one to four in the validation data. In a distribution of mean £358, the results suggest that the models ‘explain’ over 50% of the variance in wage which corresponds to an improved accuracy of around £75 per person on average from a simple geometric mean measure of wage. For the mixed models, the addition of level three random effects had the greatest relative improvement in accuracy. Comparing the intercept only models, adding a random intercept for level two improved the accuracy by an average of around £5 per individual with a similar level of improvement in the intercept and slope models. The addition of varying age slopes at level three however was less marked with smaller improvements in accuracy when comparing the two level models. In the three level models accuracy was improved by around a £1 per person on average with the addition of varying age slopes.

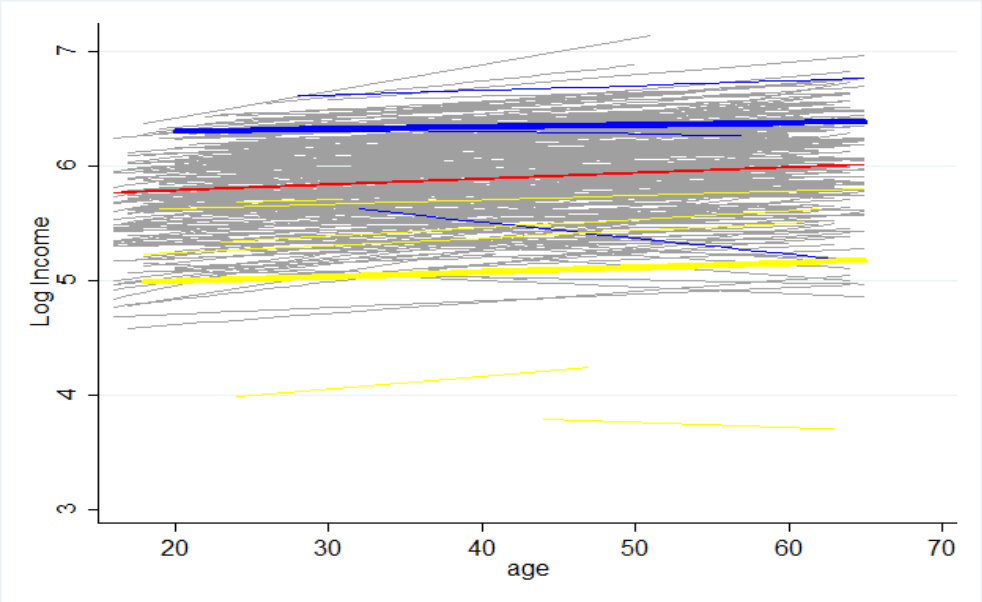
Panel two: Regression lines from mixed models with three levels predicting occupational wage. Fixed effects lines in red and SOC minor random effects in grey. For comparison, thicker blue and yellow lines represent SOC minor groups 111 (Corporate managers and senior officials) and 924 (Elementary security occupations) respectively and thinner lines represent the associated lower level occupation subgroups.



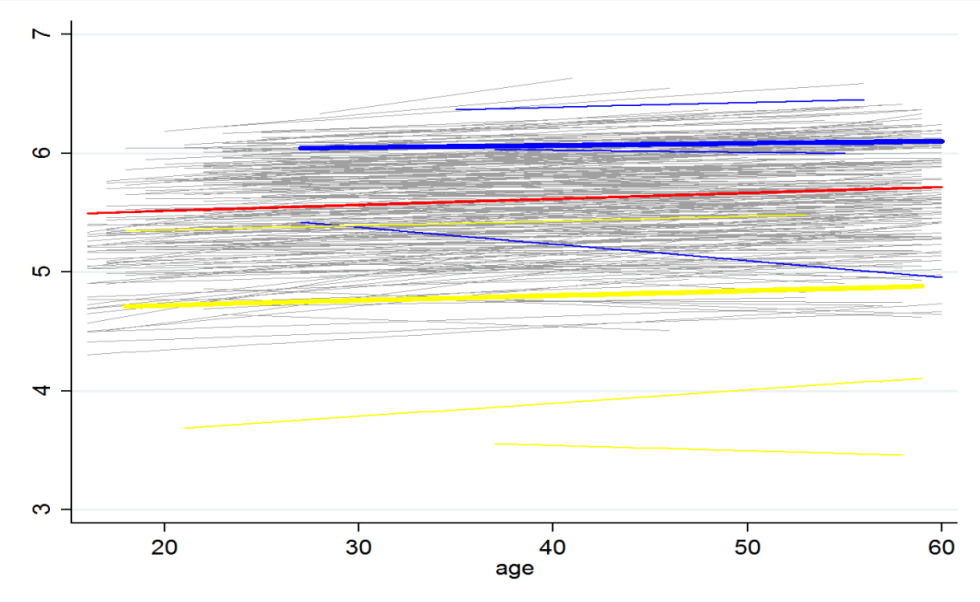

A) Random intercept model (men)




B) Random intercept model (women)




C) Random intercept with random age slopes (men)




D) Random intercept with random age slopes (women)

**Table 2: Variance, percentage reduction in variance and standard deviation of the wage predictions in the validation data**

| | Validation data generated from model parameters and coefficients | | |
|---|---|---|---|
| **Model** | **Variance of prediction errors** | **Standard deviation of prediction errors** | **% reduction in variance of prediction errors** |
| **Grand Sample Geometric Mean Wage** | 43815.65 | 209.3219 | 0% |
| **Geometric Mean wage within SOC Unit Group** | 43665.67 | 208.9633 | 0.34% |
| **2 level intercept** | 22730.3 | 150.7657 | 48.12% |
| **2 level intercept and age slopes** | 22602.38 | 150.3409 | 48.41% |
| **3 level intercept** | 21192.92 | 145.5779 | 51.63% |
| **3 level intercept and age slopes** | 20987.17 | 144.8695 | 52.10% |
| **N** | 27560 | | |

# 4. Discussion

This paper has outlined and explored a mixed model approach to the estimation of occupation based wage for the SLS based on the detailed classifications of occupation in the 2000 standard occupational classification. It has used wage data from the labour force survey to estimate and compare a number of models predicting log weekly gross wage. Four models were estimated and then assessed for predictive power using a separate year of the labour force survey. These models estimated fixed coefficients that captured the overall relationship between age and sex across all SOC groups together with random components that captured the differences in these fixed effects between nested groups within the SOC structure.

There were a number of reasons for choosing a mixed model approach rather than for example a standard fixed effects model. Firstly, it was important to utilise as much of the detail of the SOC structure as possible including the lowest level unit categories to maximise the available information included in the SOC variable. In a fixed effects model this would require the estimation of 353 fixed parameters for each of the categories. In those categories containing fewer cases these effects would be subject to large standard errors and therefore the derived estimates would be a poor estimate of the actual, unmeasured, average SOC group wage level or over fitted to the particular sample used in the model estimation. In the random effects framework, estimation of random parameters can account for this problem by adjusting or 'shrinking' the estimates towards a higher level parameter value depending on the precision of the estimate and the variance between the parameters of interest. Though this procedure does introduce bias into the estimate it reduces the chances of

capturing random noise within the models through overfitting to the sample used in the modelling.

Our findings illustrate the importance of utilising the more detailed lower levels of the SOC. Incorporating level two information into our models improved their accuracy by, on average, around £5 per person. In terms of the overall variability inherent in the wage distribution (sample average of £358) this appears relatively small yet any degree of extra precision is likely to reduce misclassification. No model could plausibly capture all of the determinants of individual wage because it is determined by a large number of factors or characteristics, some random, that we could never hope to capture in a survey. Of greater importance in statistical terms and certainly in the context of regression adjustment is the approximate ordering of individuals by wage. If this is achieved, then the effect estimate of wage can be examined or its confounding effects controlled for irrespective of the magnitude of the value of their estimated wage. From this perspective, even small increases in precision are a worthwhile exercise.

A significant limitation with the approach that we have described which also applies more generally when using wage to proxy wealth is that our estimates are unable to account for levels of wealth or capital such as house value, material possessions and savings for example. We are also unable to take account of income from other sources such as state benefits, tax credits or other sources that are separate from the working wage. For the latter, however, it would be fairly straightforward to estimate state benefits for the unemployed or workless for example based on current welfare benefit levels.

In summary, we are confident that we have been able to produce a robust estimate of occupational wage. A STATA module for replicating these estimates has been written and can be made available to SLS users. Future work will thus look to produce these estimates within the SLS and determine ‘face validity’ through examining its effects on various relevant outcomes as well as extending the estimates to the 1991 census. Furthermore, we will also utilise household composition information from both the 1991 and 2001 censuses in the SLS to produce an equivalised variable which will reflect differences in the size of the household over which the estimated wage may be spread. This can be achieved by following the method advocated by the Organisation for Economic Co-operation and Development in which the total household income is divided by a weight which is derived from the numbers of dependant individuals in the household. The utility of the resulting equivalised household wage is that it is likely to be a more accurate measure of material circumstances because it allows for the fact that the material wealth of a household is likely shared around the household.

## References


Benzeval, M. and K. Judge (2001). "Income and health: the time dimension." Social Science & Medicine **52**(9): 1371-1390.

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

Frijters, P., J. Haisken-DeNew, et al. (2005). "The causal effect of income on health: Evidence from German reunification." Journal of Health Economics **24**(5): 997-1017.

Galobardes, B., M. Shaw, et al. (2006). Indicators of socioeconomic position. Methods in social epidemiology. J. M. Oakes and J. S. Kaufman. San Francisco, Jossey-Bass Inc Pub. **7:** 47-85.

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

LFS (2007). Labour Force Survey: Background and Methodology. L. F. Survey.

Lundberg, O. and J. Fritzell (1994). "Income distribution, income change and health: on the importance of absolute and relative income for health status in Sweden." WHO regional publications. European series **54**: 37.

Lynch, J., G. 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.

McDonough, P., G. Duncan, et al. (1997). "Income dynamics and adult mortality in the United States, 1972 through 1989." American Journal of Public Health **87**(9): 1476.

Menchik, P. (1993). "Economic status as a determinant of mortality among black and white older men: does poverty kill?" Population Studies **47**(3): 427-436.

Mullis, R. (1992). "Measures of economic well-being as predictors of psychological well-being." Social Indicators Research **26**(2): 119-135.

ONS (2000). Standard Occupational Classification 2000 (SOC2000): Summary of Structure. http://www.ons.gov.uk/about-statistics/classifications/archived/SOC2000/summary-of-structure.pdf. O. f. N. Statistics.

Pike, R., C. Marks, et al. (2008). "Measuring UK inflation." Data and support: 18.

Williamson, P. and D. Voas (2000). "Income imputation for small areas: Interim progress report." CCSR.

Wolfson, M., G. Rowe, et al. (1993). "Career earnings and death: a longitudinal analysis of older Canadian men." Journal of gerontology **48**(4): S167.