Economic and Social Data Service

Small area estimation using ESDS government surveys – An introductory guide

ESDS Government

Author: Alan Marshall

Version: 1.1
Date: July 2010

University of Essex
MANCHESTER 1824 The University of Manchester
JISC
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Contents

List of figures

1. Introduction

A key strength of the surveys supported by ESDS Government is the richness of data compared to aggregate data sources such as the Census. However, survey data is either not available for sub-national areas such as districts and wards or is based on such small sample sizes that estimates are not reliable. For many users of government surveys this represents a major weakness. Detailed information on the socio-economic characteristics of the population living in local areas is a valuable source to determine how resources should be allocated between areas, the nature of service provision within an area and to identify areas that would benefit most from specific policy interventions. Updateable subnational population information is an essential source to assess the impact of policies to tackle area based issues such as neighbourhood deprivation, crime and poor health.

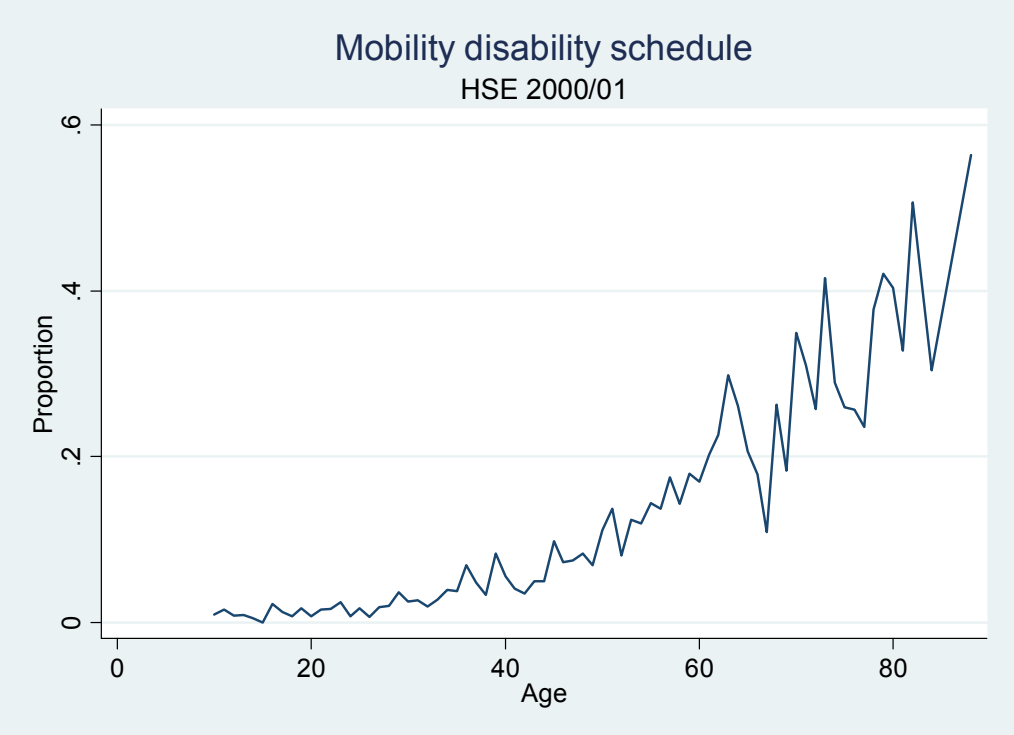
This guide addresses the weakness of survey data by introducing methods that can be used to generate local survey estimates by combining information from ESDS government surveys with other aggregate data that is reliably available for sub-national areas. The general principle of all the methods is that the aggregate data gives valuable local information that can be utilised to derive a survey estimate of a particular characteristic. For example, the census provides a range of socio-economic indicators for sub-national areas such as the age structure, levels of poor health/disability, tenure, employment status and the types of occupation undertaken by residents. This information can be used to generate local estimates of survey characteristics such as levels of smoking/drinking, specific health problems or disabilities, levels of charitable giving and the prevalence of obesity. For example, research suggests that people in the lower social classes are more likely to smoke than those in the higher social classes therefore we might expect higher levels of smoking in areas with high proportions of the population in the lower socioeconomic groups.

After this introduction the guide is divided into five sections. The first section is concerned with the geographical information that is available in UK surveys and includes three different types of information; direct boundaries, area classifications and primary sampling units. The availability of this geographical information is listed for each of the ESDS Government surveys. The second section describes the main sources of auxiliary (aggregate) information that can be combined with survey data in order to generate small area estimates. In the third section, the key small area estimation techniques are described and references to further information and examples of their use are given. The fourth section features a

practical case study that uses data from the Health Survey for England (HSE) and the Census to develop estimates of the population with a mobility disability in six UK districts.

Throughout this guide the words 'local' and 'small area' are regularly used and it is worth spending a little time describing their meaning. In this guide they refer to geographical areas (usually subnational) for which data is required but is either unavailable or estimates are unreliable as a result of small sample sizes. In some situations large areas may fall into the category of a 'small' or 'local' area according to this definition. If we examine the schedule[1] of age specific mobility disability rates for England (see figure 1), it is clear that there is a good deal of fluctuation from the general age pattern that stems from sampling variability. In these examples, the unreliability of the estimates means that techniques of 'small area' or 'local' estimation are valuable. Rao (2003) provides a comprehensive discussion of the definition of small area in relation to small area estimation see Rao (2003).

Figure 1: Mobility disability schedule (England- Males)



[1] A schedule is a curve of age specific rates for a characteristic that displays a strong age pattern

2. Geography in ESDS Government surveys

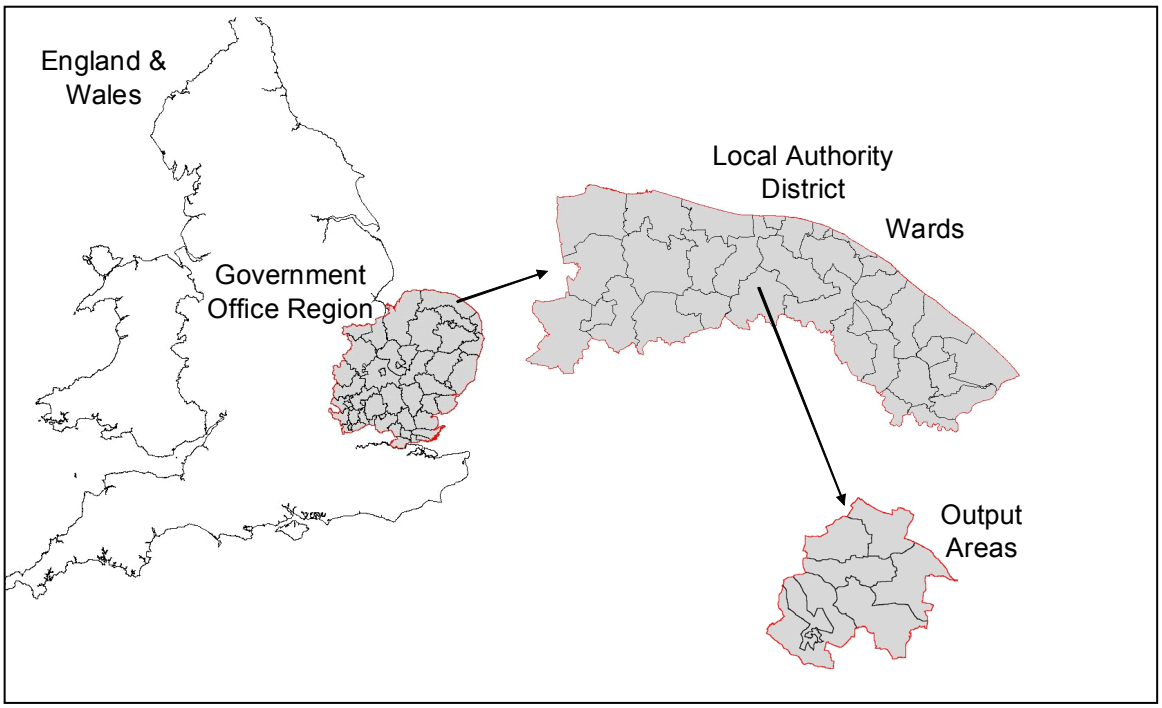
ESDS Government surveys include three types of information from which geographical information associated with the residence of a respondent can be determined. These are direct boundary information, area classifications and primary sampling units. The extent of geographical information that is included in a survey is important because it influences the nature of the model that is used to generate estimates for subnational areas.

2.1 Direct boundary information

A number of different types of boundaries are included in Government surveys. Some boundaries are common to many surveys (e.g. Government Office Regions) and others are survey-specific (e.g. British Crime Survey and Police Force Areas). The ONS have produced a Beginner's Guide to UK geography which gives a very informative introduction to the various geographies in the UK including downloadable maps. The guide is available at: http://www.statistics.gov.uk/geography/beginners_guide.asp

The administrative and census geographies in the UK feature in almost all the ESDS Government surveys. In England there are nine Government Office Regions which are subdivided into 376 Local authority districts, 8850 wards and 175,434 Output areas. Figure 2 divides the East of England GOR in its constituent local authority districts, wards and output areas.

Figure 2: Administrative and Census boundaries in the East of England



Source: Map provided by Dr Paul Norman

Maps of the Census and Administrative geographies can be downloaded from: http://www.statistics.gov.uk/geography/maps.asp and a full hierarchy including all the boundaries is given in figure 3.

Information on other geographies including Health, Postal, Electoral can be found on the ONS Beginner's guide to geography webpages:
(http://www.statistics.gov.uk/geography/beginners_guide.asp)

It is important to recognise that most boundaries have been subject to changes over time complicating time-series analysis of ESDS data. Norman (2003) considers strategies to achieve data compatibility when faced with such boundary changes.

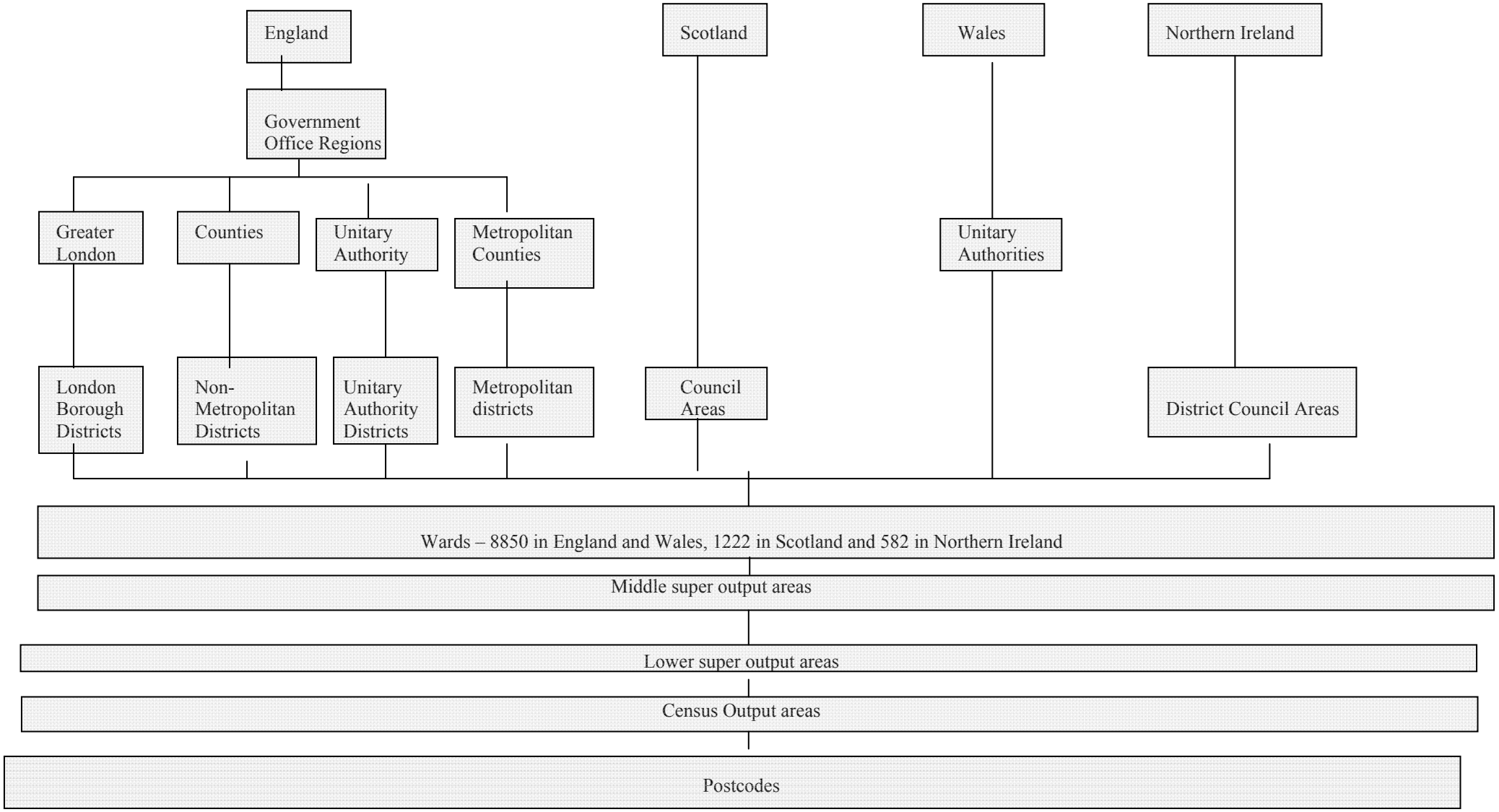
2.2 Area Classifications

Area classifications are regularly included in ESDS Government surveys and are increasingly used in a variety of research settings, such as the analysis of spatial patterns of migration (Duke-Williams 2008), participation in higher education (Singleton, Davidson-Burnett et al. 2007), crime (Bruton-Smith 2008), health outcomes, deprivation and other socioeconomic characteristics (Openshaw 1995). A number of area classifications have been developed, including the Indices of Multiple Deprivation (IMD) (DCLG 2007), MOSAIC (Experian 2009), ACORN (CACI 2009), the National Statistics 2001 Area Classification (ONS 2004) and the Output Area Classification (OAC) (Vickers,2006). These have been produced for varying purposes at differing geographical scales and time points.

The IMD 2007 is a multidimensional measure of deprivation that is available for Lower Super Output Areas and Districts in England. It uses data from the Census as well as more recent government administrative data on aspects such as health and benefits. The Department for Communities and Local Government (DCLG) commissioned the Social Disadvantage Research Centre (SDRC) at the University of Oxford to update the Indices of Deprivation 2004 (DCLG 2007).

MOSAIC and ACORN are two of the more established commercial area classifications and are produced by Experian and CACI respectively (CACI 2009; Experian 2009). Each classification uses both Census data and consumer information to generate a classification of all UK postcodes.

Figure 3: UK administrative and Census geographies



The National Statistics 2001 Area Classification (NSAC) uses a range of data from the Census 2001 which can be divided into six domains of demographic structure, household composition, housing, socio-economic character, employment and industry. All UK districts are divided into a three tier hierarchy of supergroups (8 clusters) groups (13 clusters) and subgroups (24 clusters) using the Ward's Clustering method followed by the k-means method. Methodological details are given elsewhere (ONS 2004).

At a finer geographical level than the NSAC, the Output Area classification (OAC) distills key results from the 2001 Census for the whole of the UK to indicate the character of local areas (census output areas). It was created in a collaboration between the Office for National Statistics (ONS) and the University of Leeds using the same well established methods as the NSAC classifications of local authorities. The OAC (like the NSAC) it is freely available from ONS.

2.3 Primary sampling units

Most of the ESDS Government surveys are multistage samples involving selection of a primary sampling unit (often a postcode or in a small number of cases a middle super output area or ward) and then selection of individuals from within the primary sampling unit (PSU). Some surveys include information on the primary sampling unit and, whilst this does not allow identification of the geographical location of the PSU it does allow analysts to determine whether individuals live with the same PSU. This information can be used to develop local estimates in a multilevel framework. For example, Twigg, Moon *et al*. (2000) develop estimates of smoking and drinking for wards in England using the Health Survey for England PSUs (postcodes). Section 4.2.4 provides more details on this study.

2.4 Geography in ESDS Government Surveys

Figure 4 indicates the areas for which data is available in the surveys supported by ESDS government. It also includes details of area classifications where these are available and whether or not the PSU is included in the deposited survey data. Such information is important in determining the nature of the model that is fitted. For example, the Annual Population Survey (APS) includes some information for local authority districts whilst the Health Survey for England (HSE) does not. Small area estimation techniques for local authority districts using the APS and HSE have different focuses. APS models would aim to improve the reliability of existing district survey estimates whilst HSE models would need to develop district estimates in the absence of any survey data at that level.

Figure 4: Geographical information in ESDS Government surveys

Survey	Sample size	Areas for which data is available	Geodemographic information	PSUs identified**
Annual Population Survey (April 2008-March 2009)	351,647 individuals	Government Office Regions (GORs); Unitary Authorities (England); Local Authority districts	None	Yes (can be derived)
British Crime Survey (2008/9)	48,136 individuals	Police force Area* Lower Super Output Area* Basic Command Unit*	Rural and Urban classification* Acorn classification* ONS classification (ward, district and output area)*	Yes
British Social Attitudes Survey 2008	4,468 individuals	Government Office Regions. Local Authority districts (prior to 2005)	None (some information on area type in previous surveys)	Yes
Continuous Household Survey (Northern Ireland) (2008/9	4,733 households	Northern Ireland	None	n/a (random sample of households)
Living Costs and Food Survey (previously Expenditure and Food Survey)	5,091 households in Great Britain, and 574 in Northern Ireland	Government Office Regions	Acorn classification Output area classification	No
Family Expenditure Survey	15,925 individuals	Government Office Regions	Area type (metropolitan, non-metropolitan and Greater London) Acorn classification	No
Family Resources Survey	24,977 households	Government Office Regions	Urban/rural classification of postcodes in Scotland ONS area classification Acorn classification (some years)	No
General Lifestyle Survey (previously General Household Survey)	20,503 individuals (8729 households)	Government Office Regions	Acorn classification (1996/7) Urban/rural classification (1982-1979)	Yes*
Health Survey for England	22,623 individuals	Government Office Regions Health Authorities	Ethnic mix (2000) ONS area classification (2006) Urban rural classification	Yes
Households below Average Income 2008/9	24,977	Government Office Regions	None	No
Labour Force Surveys	114,493	Unitary Authorities, Local Authorities	None	n/a (LFS is a random sample of households)
National Food Survey	6,700 households	Local Authorities	None	No
National Travel Survey	8,094 households (2008)	Government Office Regions	None	Yes

Survey	Sample size	Areas for which data is available	Geodemographic information	PSUs identified**
Northern Ireland Family Expenditure Survey	1223 individuals	Northern Ireland	Area type (metropolitan, non-metropolitan and Greater London)	No
Northern Ireland Labour Force Survey	8,842 individuals	Northern Ireland	No	n/a (NILFS is a random sample of households)
Northern Ireland Life and Times Survey	1,215 individuals	Northern Ireland	No	Na (random sample of addresses)
ONS Opinions Survey (formerly Omnibus Survey)	1,087 individuals	Government Office Regions	Acorn (June, July, September, October and November, 2004 and March 2005)	No
Scottish Crime and Justice Survey	16,003	Health Boards; Police Force Areas; Community Justice Authority Areas; National Criminal Justice Board Areas; Local Authority Areas	Scottish Index of Multiple Deprivation	No
Scottish Health Survey	8,215 individuals	Health Boards; Health Authority Regions/Districts	Urban/rural indicator (2008) Scottish Index of Multiple Deprivation (2003)	Yes in 2008
Scottish Social Attitudes	1,508 individuals	Local Authority Districts***	Urban Rural Classification	No
Survey of English Housing	38,105 individuals	Government Office Regions; Standard Statistical Regions; Local Authorities	Index of deprivation (2002/3 and 2001/2) Acorn classification (2001/2) Urban/rural indicator (1993/4 and 2007/8)	No
Time Use Survey	11,664 individuals	Government Office Regions	Population density Unemployment rate of postcode	No
Welsh Health Survey	15,966 individuals	None in 2008. Unitary Authority in 1998.	None	No

*available through special license only

** Information on primary sampling unit includes whether individuals/households are in the same postcode. The postcode itself is not given.

*** Ward and Mosaic variables are not included in the deposited dataset, but are available on request from the Scottish Centre for Social Research, subject to certain restrictions and conditions.

3. Auxiliary (locally available) data

All of the small area estimation techniques that are discussed in this guide combine survey data with other auxiliary (aggregate) information that is reliably available for local areas (based on large sample sizes). Three sources of information are described here; the Census, Administrative statistics and Mid-year estimates/ population projections.

3.1 Census of population

The census of population has been carried out since 1841, the most recent census was conducted on the 29th April 2001 with the next census day scheduled for the 27th May 2011. A key advantage of the census is that it enumerates all people and so is not prone to the sampling error that affects survey estimates particularly at neighbourhood level. The 2001 census form includes a wide range of questions covering various socio-demographic and health indicators.

Aggregate census data can be downloaded from a number of websites including Casweb, which is available to staff and students at UK higher and further education establishments, and Nomis, an ONS operated site, which is free for anyone to use. For more information on aggregate census output is provided by the Census Dissemination Unit.

Census data can also be accessed in microdata format through the Samples of Annonymised Records which are samples of individual records from the 2001 (and 1991) Censuses. There are five SAR files including the small area microdata which distinguishes local authority districts. This file is available under an End User License and includes information on 2.96 million individuals (5% sample).

3.2 Administrative statistics

Administrative statistics are another source of data that are often available for small areas and which can be combined with survey data to generate local survey estimates. For example, a wide range of data on benefits claimants collected by the Department for Work and Pensions are available from the Nomis website. Other administrative statistics that may be of use include vital statistics, GP records, hospital episode statistics, and eligibility to free school meals.

3.3 Mid-year estimates and population projections

One of the key disadvantages of the census is that it is only conducted once every ten years and so can become rather out of date. Mid-year estimates and population projections are produced by the National statistical offices in England, Scotland and Northern Ireland providing a more recent source of local information giving population counts for local authority districts that distinguish single year of age and sex.

Mid-year estimates and population projections are valuable for capturing the current and future population size and age structure, which can have implications for estimates of population characteristics that are strongly age related (such as the prevalence of ill health and disability). For example, if a population is projected to become more elderly then we can infer that the population with an illness will increase because the rates of illness tend to be greatest at the oldest ages. Mid-year estimates and population projections that distinguish marital status in addition to age and sex are also available for England and Wales and Scotland (but only at national level).

Figure 5: Mid-year population estimates and population projections

Title	Time points available*	Geography	Detail	Producer
Mid-year estimates England and Wales**	2002-2009	Local Authority districts	Quinary age groups and sex	ONS
Mid-year estimates Scotland	1982-2009	Council areas NHS board areas	Single year of age and sex	GROS
Mid-year estimates Northern Ireland	1991-2009	District council areas	Single year of age and sex	NISRA
Population projections – England**	2008-2033	Local Authority districts	Quinary age groups and sex	ONS
Population projections - Wales	2008-2033	Local Authority districts	Single year of age and sex	Stats Wales
Population projections - Scotland	2008-2033	Council areas	Single year of age and sex	GROS
Population projections – Northern Ireland	2008-2033	District council areas	Single year of age and sex	NISRA

*At time of writing (August 2010)

** rounded to the nearest 1000.

Although the ONS population projections and mid-year estimates are available for quinary age groups and are rounded to the nearest thousand it is possible to get unrounded data for single years of age by contacting ONS. It should be noted that ONS do not recommend reporting the unrounded data as it implies a false sense of accuracy in the figures they produce. ONS advise that these unrounded estimates should only be used for research purposes.

It is important when using mid-year estimates and population projections to consider the population enumerated. In general the total population is counted which creates a problem when combining these sources with survey data that usually exclude the institutional population. This discrepancy is particularly important at the oldest ages where the institutional population is largest. One approach to get around this issue is to use the census to calculate the proportion of the population who live in households in census year and then use this to adjust the population mid year estimate for a given year. The census table ST001 (Age by sex and type of resident) gives the institutional and household population totals by sex and single year of age can be used for this and is available from websites such as Casweb (academics) and Nomis.

4. Small area estimation techniques

This section gives a review of the methodologies that are used to derive local estimates of socio-demographic characteristics in the absence of reliable survey estimates and is divided into three parts based around broad methodological approaches. First, demographic models for the estimation of characteristics that are strongly linked to age are examined. Second, synthetic estimates that combine survey data with small area data from other sources in order to generate small area estimates are reviewed. Finally, microsimulation techniques that are concerned with the generation of large scale population micro-datasets for local areas (lists of individuals and their attributes) are discussed.

4.1 Demographic models

The basis of demographic models stems from the tendency for demographic rates, such as mortality, fertility and migration, to be strongly linked to age with the intensity of the event varying sharply across the age range (Preston, Heuveline et al. 2001). A key aspect of demographic research is to discover the regularities in such age schedules and a number of methodologies have been developed specifically for this purpose. (Coale and Trussell 1996). Where a socioeconomic characteristic is strongly linked to age then the techniques of demographic models offers an approach for the generation of small area estimates.

The two types of demographic models that are introduced here are curve fitting (also known in the literature as mathematical representation) and relational models. Each of these approaches aim to improve the reliability of sets of age-specific rates (or schedules) for a particular characteristic by smoothing the fluctuations in the observed curve of age-specific rates (for example see figure 6). The model schedules of rates can then be combined with local population counts distinguishing age and sex in order to generate neighbourhood counts of the population with the characteristic of interest.

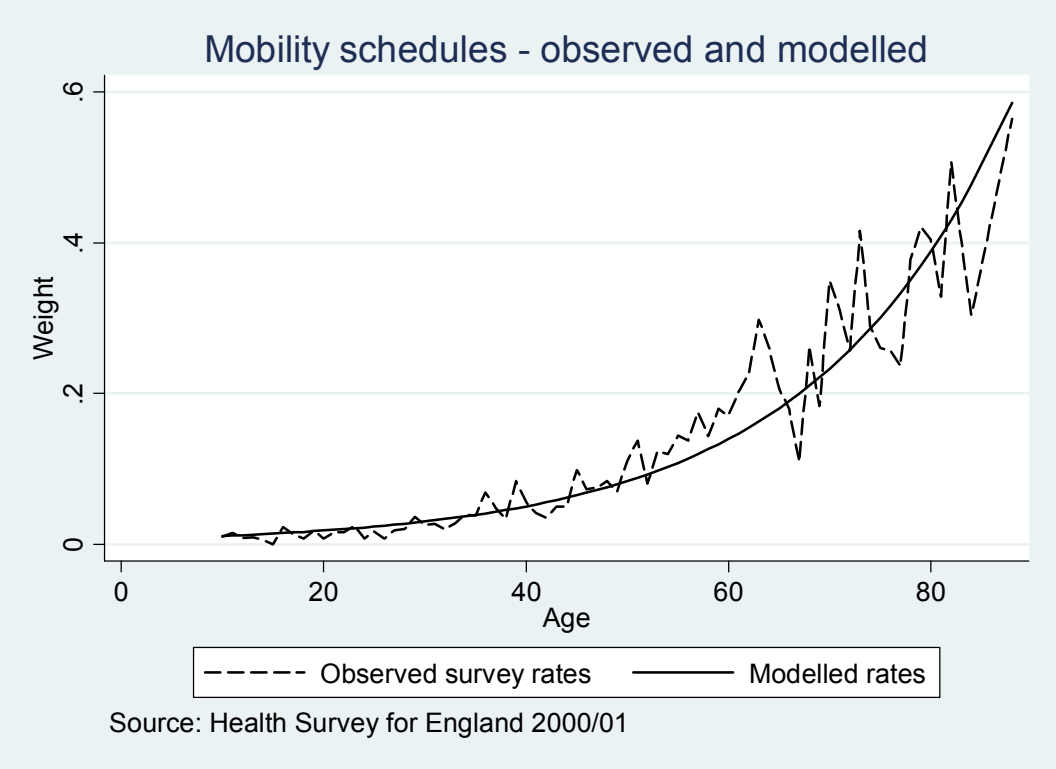
A review of the literature shows that curve fitting and relational models have been used to generate local schedules of disability rates, a characteristic that displays a similar age pattern to mortality and hence offers an opportunity to borrow approaches from the estimation of mortality schedules.

4.1.1 Curve fitting

Curve fitting involves fitting a function to represent the rates across the age profile (Coale and Trussell 1996). The statistical basis of this approach is usually non-linear regression

techniques (Congdon 1993). Non-linear regression models differ from linear regression both in the techniques that are used to fit the data and the interpretation of results. For further information on parameter estimation and interpretation of model statistics see Freund and Littell (2000). Michaud (1996) use an exponential curve to improve the reliability of age specific disability rates for Canadian territories and practical 1 of the case study demonstrates this approach using the data from the Health Survey for England. Figure 6 shows an exponential curve that is fitted to rates of mobility disability in England thus improving the reliability of the age-specific rates particularly at the oldest ages.

Figure 6: Mobility schedules: Observed survey rates and model rates derived from an exponential curve (Males - England)



4.1.2 Relational Models

Relational models comprise a (reliable) standard schedule of rates and a mathematical rule that maps the standard to another schedule in a population where information may be incomplete or unreliable (Preston, Heuveline et al. 2001).

The relational approach was originally developed by Brass (1971) for the modelling of mortality curves. The Brass model is based on a logit transformation of q(x) the probability of dying before age x.

$$Y(x)=\frac{1}{2}\text{Logit}[q(x)]=\frac{1}{2}\ln\left[\frac{q(x)}{1-q(x)}\right] \quad \textbf{1}$$

As q(x) varies between 0 and 1 then the logit of q(x) takes all values between -∞ to +∞. If we have the predicted value of the logit of q(x) denoted $\hat{Y}(x)$ then this can be converted to a predicted probability of dying by age x:

$$\hat{q}_x=\frac{\exp(2\hat{Y}(x))}{1+\exp(2\hat{Y}(x))} \quad \textbf{2}$$

The logit transformation of q(x) is valuable because the relationship between two logit mortality schedules turns out to be remarkably linear (Newall 1988). On the basis of this linear relationship, Brass proposed a simple relational formula to predict Y(x) from the logit of q(x) in the standard population $Y^s(x)$:

$$\hat{Y}(x)=\alpha+\beta * Y^s(x) \quad \textbf{3}$$

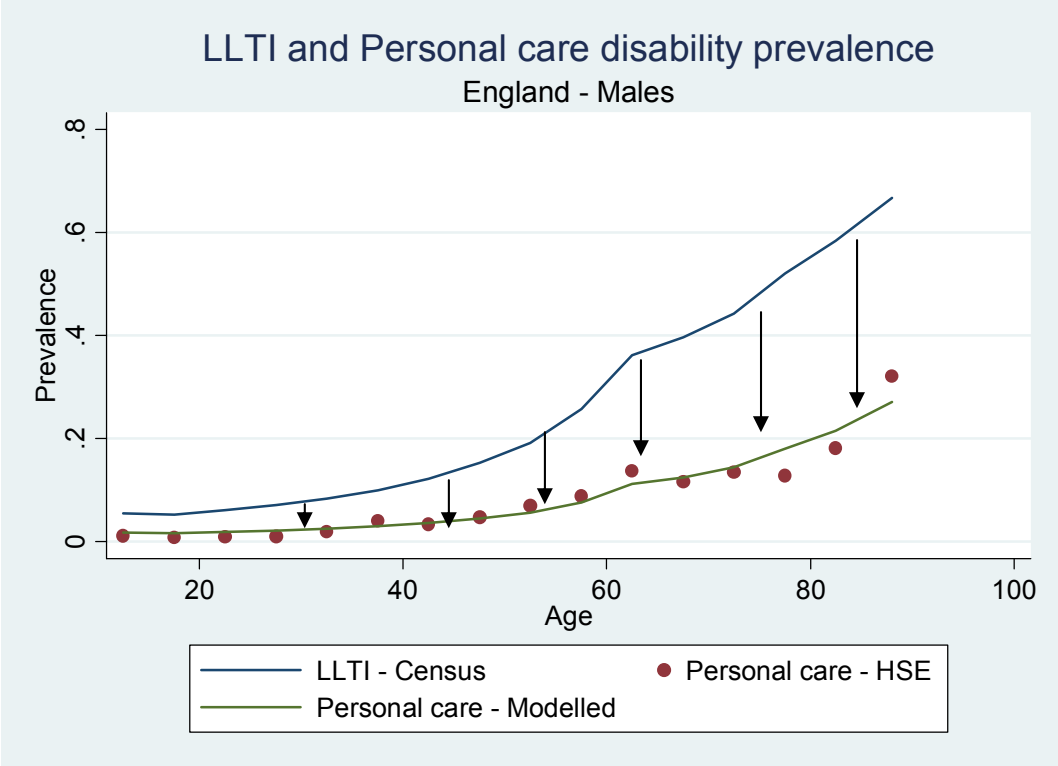
Altering α affects the level of mortality, with a value above 0 increasing the level of mortality compared with the standard schedule and a value less than 0 decreasing it. Altering values of β affects the relationship between childhood and adult mortality. The further β falls below 1 the lower is the predicted mortality at the younger ages (compared with the standard) and the higher is the mortality is at the higher ages (compared with the standard). The opposite is true the further β rises above 1. If α=0 and β=1 then the standard and predicted schedules are identical (Newall 1988; Congdon 1993; Coale and Trussell 1996).

Two features determine the success of the relational approach, these being the appropriateness of the standard and the relational rule (Preston, Heuveline et al. 2001). The relational approach can be used successfully with any standard, but it is most effective if the standard is close to the population being modelled (Keyfitz 1982).

Marshall (2009) extends the use of the Brass relational model to the estimation of disability schedules for districts in the UK using data from the Health Survey for England and the

Census. The general strategy is to use the Census LLTI schedule for England as the standard and employ the Brass relational model to relate this to schedules for various disability types as measured in the Health survey for England (see figure 7). The relational parameters from this model are stored and then used to adjust district LLTI curves and derive local schedules for different disability types in the absence of survey estimates. An example of the use of relational models in this way is given in practical 2.

Figure 7: Relational model: Personal care disability (Males - England)



An algebraic specification of the relational model described above is given below for the estimation on mobility disability:

Let:

p_{xsrd} = rate of disability at age x (x=10,11,….84,88) and sex s (s=1,2) in district r for disability type d.

l_{xsr} = prevalence of LLTI at age x and sex s in district r (Census 2001)

$p_{xs.d}$ = rate of disability at age x and sex s in England for disability type d (HSE00/01)

$l_{xs.}$ = prevalence of LLTI at age x and sex s in England (Census 2001)

Then the predicted prevalence of disability d for district r and age x is derived from:

$$\frac{1}{2}\log_e\left(\frac{\hat{p}_{xsrd}}{1-\hat{p}_{xsrd}}\right)=\hat{\alpha}+\hat{\beta}\left(\frac{1}{2}\log_e\left(\frac{l_{xsr}}{1-l_{xsr}}\right)\right) \quad \textbf{4}$$

Where $\hat{\alpha}$ and $\hat{\beta}$ are estimated from equation 5 (England level data):

$$\frac{1}{2}\log_e\left(\frac{p_{xs.d}}{1-p_{xs.d}}\right)=\alpha+\beta\left(\frac{1}{2}\log_e\left(\frac{l_{xs.}}{1-l_{xs.}}\right)\right)+e_x \quad \textbf{5}$$

4.2 Synthetic estimates

Synthetic estimation is a well established methodology that is used to generate local estimates by combining data that are reliably available for small areas with other data that are not (Charlton 1998). The approach uses a model-based technique to combine survey data containing a characteristic of interest with a set of associated covariate (or predictor) variables that are available for the small area (usually from the census or an administrative data source) in order to generate estimates of a characteristic of interest for each small area (Bajekal, Scholes et al. 2004).

There is a wide body of research on synthetic estimation and a number of different types of models have been developed. The models differ in terms of their complexity, data requirements and the levels (area/individual or both) at which that they operate. This section provides information on a selection of synthetic estimation techniques. For a more comprehensive review see Skinner (1993), Bajekal, Scholes et al. (2004) and Rao (2003).

4.2.1 Indirect standardisation

Indirect standardisation involves dividing the population into groups that are known to be associated with the characteristic of interest and to vary between small areas (Siegel 2002). Section 5. demonstrates how local estimates of disability can be generated using the product of national schedules of age-specific disability rates and local population estimates. The key advantage of this form of synthetic estimation is its ease of application. The census provides population group counts for small areas and surveys give national estimates of proportions with the characteristic of interest (in each population group) (Bajekal, Scholes et al. 2004).

The main disadvantage concerns the assumption that the national prevalence rates apply for all sub-national areas. This assumption means that if two areas have identical population characteristics then the indirect standardisation approach will generate identical estimates of disability.

Indirect standardisation can be extended to take into account contextual factors to some extent by calculating rates for different types of areas (according to a geodemographic area classification) and then applying these rates to the small areas in each area type (Bajekal, Scholes et al. 2004).

4.2.2 Individual-level synthetic regression estimation

Individual-level synthetic regression applies the indirect standardisation approach within a model framework (Bajekal, Scholes et al. 2004). In the first stage survey data is used to fit a regression model (usually logistic) that predicts the probability that an individual has a particular characteristic (e.g. a mobility disability) based on their characteristics (e.g. age, sex, social class). The probabilities associated with each combination of the explanatory characteristics are calculated and are then applied to the corresponding aggregate population groups in each small area to derive local estimates of the characteristic of interest (Skinner 1993).

Compared with indirect standardisation, the more formal modelling of individual synthetic regression estimation is valuable because it enables the success of explanatory variables to be formally assessed and for insignificant covariates to be removed from the model (Skinner 1993). This approach is still prone to the main weaknesses of indirect standardisation, namely that it does not take into account local area contextual factors or unmeasured variables that vary between areas.

Two further problems apply to both indirect estimation and individual synthetic regression models. First, each rely on the assumption that explanatory variables are measured in an identical way in the Census and survey data sources (Bajekal, Scholes et al. 2004). This assumption is often not met, or at least not across the whole age range. Second, both techniques require that cross tabulations of the covariates (or population groups) are available for small areas of interest. This limits the number of covariates that can be included

in models because extensive cross tabulations from sources such as the Census are often not available for confidentiality reasons (Skinner 1993; Twigg, Moon et al. 2000).

Charlton (1998) overcomes this problem to some extent by combining census microdata for the UK (1991 Sample of Anonymised Records (SARs)) with survey data to generate small area estimates of survey variables enabling the use of more extensive cross tabulations of census variables.

4.2.3 Area level synthetic regression

Area-level synthetic regression models use aggregate data only. First, area proportions with a characteristic of interest are modelled using the area proportions with each covariate characteristic as explanatory variables. This initial model predicts area variability in the characteristic of interest rather than individual variability as in the individual-level synthetic regression model. The parameter estimates from the initial model are then combined with small area covariate proportions (Census or administrative source) to generate estimates of the characteristic of interest in the absence of reliable direct estimates (Skinner 1993).

The area-level synthetic regression approach can be refined through the use of weights that take in to account the variability of the estimates of the characteristic of interest in the initial regression model. For example, areas that have larger sample sizes should be given more weight in the regression model than those with smaller sample sizes. Fay and Herriott (1979) develop a well known model based on this principle in order to generate small area estimates of income in the US.

The main advantage of area-level synthetic regression is that if the set of area covariates are available for all areas and if the relationship between them and the characteristic of interest is strong then good quality estimates can be produced easily. Unlike the individual-level synthetic regression models there is no restriction on the number of covariates that can be included (Bajekal, Scholes et al. 2004).

The main drawback of area-level synthetic regression is that it is difficult to disaggregate estimates for different population subgroups because the estimates apply to the whole social mix of adults living in an area. Fitting separate models or adding interaction terms are possible ways around this issue (Bajekal, Scholes et al. 2004).

Area-level synthetic regression has been used by the Small Area Estimation Programme team in the Office for National Statistics to estimate a number of characteristics (including three measures of poor health) at the ward level using data from the Family Resources Survey and the General Household Survey (Heady and Clarke 2003).

4.2.4 Synthetic regression models that combine individual and area covariates

Synthetic regression models that combine individual and area characteristics within a multilevel framework offer a method that includes both compositional and contextual factors that influence a characteristic of interest.

Twigg, Moon et al (2000) use a multilevel synthetic regression model to generate ward estimates of smoking and drinking behaviour using the Health Survey for England (HSE). They fit two multilevel models to predict the propensity of an individual to drink (excessively) or smoke using three levels of individual, postcode and health district. Although the HSE does not identify the postcode of respondents it does indicate whether respondents live in the same postcode and includes a geodemographic reference for each postcode. As the HSE does not include ward of residence and the Census does not include postcode detail the postcode and the ward assume analytical equivalence in the models. The model probabilities associated with particular types of individuals and areas are then applied to the relevant population totals in each ward to develop estimates of the number of smokers and excessive drinkers.

The main disadvantage of synthetic models that combine individual and area covariates is that calculation of confidence intervals for estimates is particularly complicated (Bajekal, Scholes et al. 2004). In addition, it is important to note that models require some survey information on the characteristic of interest for the small area (or at least a similar sized area) under investigation.

4.3 Microsimulation

Microsimulation involves generating a list of individuals from a micro-dataset (a source containing information on individuals) that matches the known aggregate data for these areas as closely as possible (Dorling, Rossiter et al. 2005; Ballas, Clarke et al. 2006). Microsimulation has become increasingly popular due to two factors; first, the increasing availability of data on individuals from sources such as the Sample of Anonymised Records (from the census) and major Government surveys; second, the increasing computer power

that overcomes the computing issues that hindered the microsimulation approach in the past (Williamson, Birkin et al. 1998; Ballas, Clarke et al. 2005). The utility of microsimulation for the generation of small area estimates of characteristics that are not recorded in the Census (such as disability) is a valuable use of this methodology (Charlton 1998; Williamson, Birkin et al. 1998).

Williamson (2002) discusses a number of techniques for estimating spatially detailed population microdata including stratified sampling (geodemographic profiling), data fusion, reweighting and synthetic reconstruction. A variant of the reweighting approach known as combinatorial optimisation emerges as a particularly promising approach with important advantages, linked to low data storage requirements and flexibility of application, compared with other techniques such as iterative proportional fitting and synthetic reconstruction (Voas and Williamson 2000; Williamson 2002). For more information on microsimulation see Marshall (2009) p103-105.

5. Case study - Estimating mobility disability using the Health survey for England

Introduction

This case study contains three practicals that each develop estimates of the number of people with a mobility disability for six districts in England (Barnet, Bury, Wakefield, South Bucks, Easington) using data from the Health Survey for England (2000 and 2001) and the Census. Each practical uses a different methodology, to create a set (or schedule) of age specific mobility disability rates which is then multiplied by district population data to generate estimates of the total population with a mobility disability.

Practical 1 uses the curve fitting approach (4.1.1) to improve the reliability of national and regional age-specific rates of mobility disability. Practical 2 uses relational models (4.1.2) to develop age-specific curves of mobility disability rates for each of the six case study districts estimates. Finally, practical 3 uses an individual level synthetic regression model (4.2.2) to develop estimates of the population with a mobility disability in all six districts.

In order to work through the casestudy you will first need to download the data and save it in on the c drive of your computer (note you can save the data elsewhere on your computer but if you choose to do this you will have to amend the stata scripts as these assume data is in c:\.

Data

If you open the folder 'ESDS mobility disability practical' you will find several Stata data files and a word file called metadata which contains more details on how this data was developed. The key data file is 'HSE data.dta' which contains data extracted from the Health Survey for England in 2000 and 2001 - each row denotes an individual (aged over 10). Figure 8 shows the variables in this dataset.

Figure 8: Variables in the 2000 to 2001 HSE data for small area estimation of disability file

Variable name	Description
Sex	1=male 2=female
Age	Single year of age
Gora	Government Office region (1=North East, 2=North West, 3=Yorkshire and Humberside, 4=West Midlands, 5=East Midlands, 6=East of England, 7=London, 8=South East, 9=South West)
Weight	Weights to account for disproportionate sampling of children and older people
Year	Survey year (2000 or 2001)
Disab	Overall disability - indicates whether a person has one of the 5 disabilities measured in the HSE (mobility, personal care, hearing, sight, communication) 1=has an disability 0=does not have a disability
Pcare	Personal care disability 1= has a personal care disability 0=does not have a personal care disability
Sight	Sight disability 1= has a sight disability 0=does not have a sight disability
Hear	Hearing disability 1= has a hearing disability 0=does not have a hearing disability
Mobility	Mobility disability 1= has a mobility disability 0=does not have a mobility disability
LLTI	Limiting long term illness (LLTI) 1= has an LLTI 2=does not have an LLTI
Agesq	Age squared=age*age
Agecub	Age cubed=age*age*age

There are three other data files which will be used:

1. *Practical 1 - task 5 - data.dta* – contains district population counts and model mobility disability rates (estimated during task 3 and 4 of practical 1) distinguishing single year of age and sex for the six case study districts

2. *Practical2 data.dta* – contains the data for practical 2 – aggregate schedules of LLTI (census) and mobility disability (HSE – calculated in practical 1) for England (Males and females)

3. *Population data practical 3.dta* – contains the population counts for the six case study districts (by age, sex and LLTI status) that are required to generate district estimates of the population with a mobility disability in practical 3

For more details on these datasets including how they were produced see the documentation for the small area estimation teaching datasets.

Practical structure and instructions

Each of the practicals in the case study are divided into tasks which cover a particular theme. Syntax is given allowing calculations to be carried out in Stata. Where syntax should be run in stata, the instructions and syntax appear in bold text. For example:

Calculate the proportion of people in England with a personal care disability by running the syntax below:

```
mean loco [pweight=weight]
```

Questions or additional tasks that you may choose to carry out appear in italics. For example,

Using the syntax above as a template calculate the mean disability prevalence for overall disability (disab).

In order to work through the practical you can either copy the syntax into a do file window and then run commands or you can open the appropriate do files (C:\ESDS SAE practical\Do files) which already contain all the syntax and run the commands from these files.

Practical 1: Indirect estimation and curve fitting

Introduction

In this practical you will:

- Calculate rates of age and sex specific curves (schedules) of disability rates
- Fit an exponential curve to these schedules of disability rates (at national and regional level)
- Estimate district populations with a mobility disability based on the district population structure.

Practical 1 - Task 1: Calculating disability prevalence rates

This task demonstrates how to calculate disability prevalence rates from a microdataset (which contains information on individuals).

First open Stata. Click on window and open a new do file editor window. This is where you will paste and run syntax. Note if you are using version of stata prior to version 11, you will have to set memory to 200 (set mem 200m) in order to open this data.

1.1.1 : Open the HSE 2000/01 dataset

```
clear

use "C:\ESDS SAE practical\Data\HSE data.dta"
```

Each of the disability variables (disab, pcare, mobility, sight, hear, mobility, llti) take the value of 1 if a person has a disability and 0 if they do not have a disability. We can calculate the proportion of people with a disability by calculating the mean of these variables. This makes intuitive sense if we consider a simple hypothetical example where 5 out of 10 people have a disability. In this example the mean of the disability variable is equal to 5/10=0.5:

Mean = Sum of disability variable/total number of people
Mean = (1+1+1+1+1+0+0+0+0+0)/10
Mean = 5/10 = 0.5.

When calculating means using the HSE dataset, we must remember to use the provided weights that take into account the disproportionate probabilities of selection amongst young people and the elderly.

1.1.2 : Calculate the proportion of people in England with a mobility disability by running the syntax below (note the confidence interval here does not account for the HSE's complex survey design):

```
mean mobility [pweight=weight]
```

The Stata output generated from this command should indicate that 13.8% (or 0.138) of people in England have a mobility disability.

1.1.3: Use the syntax below to calculate proportions for a particular population group (men aged 65).

```
mean mobility [pweight=weight] if sex==1&age==65
```

The Stata output generated from this command should indicate that 20.6% (or 0.206) of males aged 65 in England have a mobility disability.

Using the syntax above as a template calculate the mean disability prevalence for overall disability (disab), personal care disability (pcare), hearing disability (hear) and sight disability (sight). Fill in the figure 9 below:

Figure 9: Proportion of the population of England with a disability

Disability type	**Disability rate in England**	**Disability rate in England (Males aged 65)**
Overall disability		
Mobility (mobility) disability	0.138	0.206
Personal care disability		
Hearing		
Sight		

The table above suggests, as we might expect, that there is a strong relationship between disability and age, with rates increasing at the oldest ages. We can explore this by producing graphs of age-specific disability rates. Task 2 explains how to do this.

Practical 1 - Task 2: Generating graphs of age-specific disability schedules

This task develops a reduced data file in which each row contains a rate of disability for a single year of age. Single year rates are calculated for males and females so in total the reduced file produced here has 152 rows (Males – 10, 11,…….84, 88; Females – 10, 11,….84, 88).

The conventional way to calculate a new variable containing age and sex specific rates in Stata is to use the 'egen' command and the mean function. However, egen cannot be used with pweights (or the svy command) and so the unweighted rates that would result would be biased. In order to get around this issue two variables are created. The first variable (MO_num) contains a weighted count of the number of people with a mobility disability at each single year of age and sex. The second variable (MO_denom) contains a weighted count of the total number of people at each single year of age and for males and females. We can easily then calculate the age and sex specific (weighted) mobility disability rates by dividing MO_num by MO_denom.

Figure 10 gives a casestudy using a population of 6 people to show how MO_num and MO_denom are calculated and how appropriately weighted disability rates are then produced from these weighted counts.

Figure 10: Casestudy - calculation of weighted mobility disability prevalence rates

Person number	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Mobility (1=has disability, 0=no disability)	0	0	0	1	1	1	3
Weight	1	1	1	0.5	0.5	0.5	n/a
Weighted count of people with a mobility disability	0*1=0	0*1=0	0*1=0	1*0.5=0.5	1*0.5=0.5	1*0.5=0.5	MO_num =1.5
Weighted count of people	1*1=1	1*1=1	1*1=1	1*0.5=0.5	1*0.5=0.5	1*0.5=0.5	MO_denom= 4.5
Unweighted mobility disability prevalence = 3/6=0.5 **Weighted mobility disability prevalence = 1.5/4.5=0.33**							

1.2.1: Run the syntax below to produce a variable (count_w) from which we can derive a weighted count of population for each single year of age and for males and females:

```
gen count_w=1*weight
```

1.2.2: The syntax below creates a variable from which we can derive a weighted count of the population with a mobility disability distinguishing single year of age and sex (remember when mobility =1 a person has a mobility disability and when mobility=0 a person does not have a mobility disability):

.

```
gen mobility_w=mobility*weight
```

1.2.3: We can now calculate the MO_num and MO_denom variables:

```
sort sex age
by sex age: egen MO_num=total(mobility_w)
by sex age: egen MO_denom=total(count_w)
```

1.2.4: Next, calculate the age and sex specific mobility disability rates by dividing the number of people with a mobility disability by the total population:

```
gen MO_OBS_RT=MO_num/MO_denom
```

Browse the data editor then scroll through the data until you find a man aged 65. Confirm that the rate of mobility disability you estimate matches that from task 1.

Close the data browser

1.2.5: Now drop all values with duplicates values of age and sex and keep only the variables of age and sex and the variables that record the age specific mobility disability:

```
duplicates drop age sex, force

keep age sex MO_OBS_RT

sort sex age
```

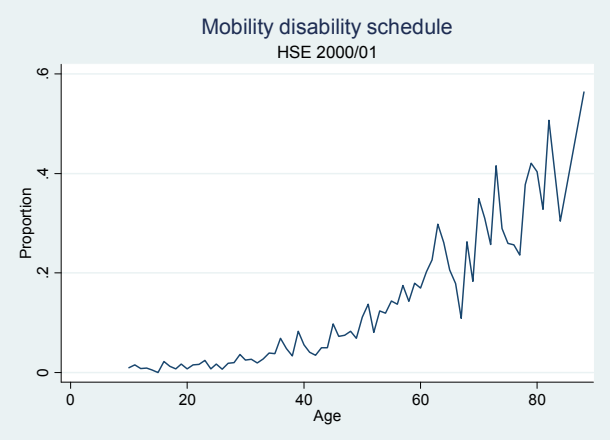
Browse the data editor to check you have 152 rows – one for each single year of age (male and females)

1.2.6: Create a graph of the mobility age specific mobility disability rates for men using the syntax below:

```
twoway (line MO_OBS_RT age if sex==1), ytitle(Proportion)
xtitle(Age) title(Mobility disability schedule) subtitle(HSE
2000/01)
```

Figure 11 shows the graph that the syntax above should have produced. The strong age relationship between age and mobility disability is clearly visible as are the fluctuations around this general age pattern that result from sampling error.

Figure 11: Mobility disability schedule (males)



1.2.7 Save the data you have created then clear the data out of stata's memory

```
save "C:\ESDS SAE practical\Saved practical data\Practical 1 - task
2.dta", replace

clear
```

Produce graphs for each of the other disability types. Can you create a graph with more than one disability type? (You will need to reopen the 2000 to 2001 HSE data for small area estimation of disability file first)

Practical 1 - Task 3: Curve fitting – fitting a function to disability schedules

You may have noticed that the graphs for each of disability types all follow a similar underlying pattern with low rates at the youngest ages that rise with age. Whilst the underlying pattern is clear there is also a fair degree of underlying variability about this pattern which stems from dealing with a sample rather than the total population.

One way to improve the reliability of the estimated rates is to fit a mathematical function to represent the underlying pattern as described in section. A function that has been used by statistics Canada (Michaud ??) to represent disability and which has been widely used to estimate mortality is the exponential function below:

$$D(x) = e^{a+bx} \qquad \mathbf{6}$$

Where:

$D(x)$ = the proportion of people with a disability at age x

This practical uses Stata's nl (non linear) command to fit a curve to the mobility schedule for males in England.

1.3.1: Re-open the data you saved at the end of the previous task (task 2)

```
clear

use "C:\ESDS SAE practical\Saved practical data\Practical 1 - task
2.dta"
```

1.3.2: Run the code below in Stata to fit the exponential function to mobility disability schedule for males.

```
nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1
```

Part of the output should include the estimates of the parameters of a and b in equation 6 (see figure 12):

Figure 12: Parameter estimates from the exponential model of the mobility schedule (Males)

LM_OBS_RT	Coef.	Std. Err.	T	P>t	[95% Conf.	Interval]
/a	-4.44	0.18	-25.12	<0.00	-4.79	-4.09
/b	0.04	0.00	18.53	<0.00	0.04	0.05

This output is telling us that the pattern of rates in the male mobility schedule (figure 11) can be best represented by our exponential function (equation 6) if a=-4.439 and b=0.04335. The table shows that both parameters make a significant contribution to the model. Other Stata output from the model shows the R^2 value is equal to 0.95, and, whilst there are some reservations about using the R^2 in a non-linear regression, this does provide some evidence that the model gives a good fit to the data.

1.3.3: Run the syntax below to create a new variable called 'pred_MO_M' which contains predict rates of mobility disability for males at each single year of age:

```
predict pred_MO_M if sex==1
```

1.3.4: Run the code below in Stata to fit the exponential function to mobility disability schedule for females.

```
nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2
```

1.3.5: Run the syntax below to create a new variable called 'pred_MO_F' which contains predict rates of mobility disability for females at each single year of age

```
predict pred_MO_F if sex==2
```

1.3.6: Generate a single variable 'pred_MO' containing the model mobility rates for males and females

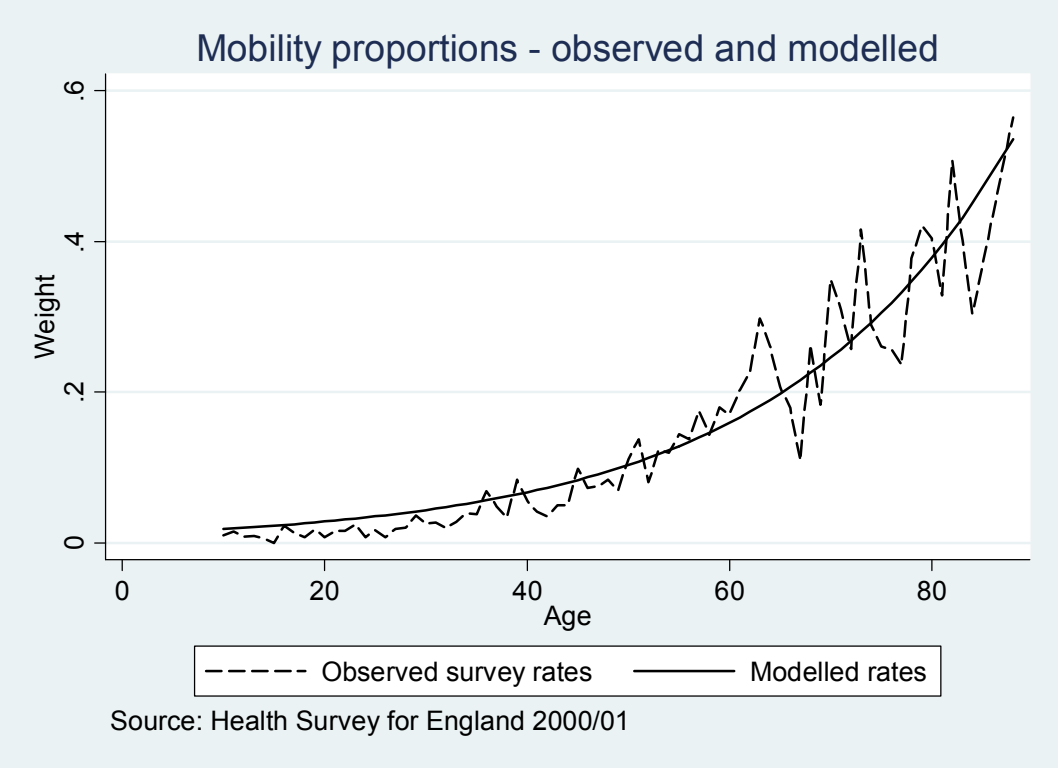
```
gen pred_MO=pred_MO_M if sex==1
replace pred_MO=pred_MO_F if sex==2
```

1.3.7: Graph the predicted (pred_MO) and observed (MO_RT_OBS) mobility disability rates for males by running the syntax below:

```
twoway (line MO_OBS_RT age if sex==1, lcolor(black) lpattern(dash))
(line pred_MO age if sex==1, lcolor(black)), ytitle(Weight)
xtitle(Age) title(Mobility proportions: Observed and modelled)
caption(Source: Health Survey for England 2000/01) legend(order(1
"Observed survey rates" 2 "Modelled rates"))
```

You should now have a graph that matches the one below:

Figure 13: Mobility schedules (males) observed and modelled



Examination of figure 13 (above) reveals evidence that the model is overestimating rates at the youngest ages. When we fitted our model Stata treated the rates at each age as being equally reliable. The graphs of observed mobility rates do not support this assumption; rates appear to fluctuate least (are most reliable) across the youngest ages and fluctuate the most (are least reliable) at the oldest ages. In order to account for the varying reliability of rates across the age range we can use age specific weights when fitting our exponential curve. These weights should be large at the younger ages (where rates appear most reliable – fluctuate least) and smallest at the oldest age (where rates appear least reliable and fluctuate most)

It is common to assume that the proportions/rates result from a binomial process where the variance associated with each estimated proportion can be derived using the formula below (Congdon 1993):

$$v_x(p) = \frac{p_x(1-p_x)}{N_x} \qquad \textbf{7}$$

Where p = proportion with a disability at age x and N_x equals the number of people sampled at age x.

The formula generates small variances (i.e. reliable estimates of proportions) where samples sizes (N_x) are large and where values of p_x are close to 1 or 0 (if an event occurs with probability 0 or 1 then there in no uncertainty in our sample measure of it). Large variances (i.e. unreliable estimates of proportions) result where sample sizes (N_x) are small or where p_x is close to 0.5. The inverse of the variance formula above gives a suitable age specific weight to apply when fitting our exponential curve:

$$w_x(p) = \frac{N_x}{p_x(1-p_x)} \qquad \textbf{8}$$

The formula for $w_x(p)$ is sensible because it results in larger weights where a proportion is reliable (the variance is low) and smaller weights where the proportion is unreliable (the variance is high). This becomes clear if we consider an example. Suppose that two rates p_1 and p_2 are associated with variances of v_1=0.1 and v_2=0.9 it is clear that p_2 is much less reliable than p_1 as its variance is larger. The weights associated with the inverse of the variances reflects this principle as a larger weight is associated with the most reliable proportion p_1:

w_1 = 1/0.1 = 10 and

w_2 =1/0.9 = 1.1

In order to calculate our weights we need to calculate values of N_x the number of observations at each age in the Health Survey for England. This requires us to clear our current data and reopen the HSE practical data.

1.3.8: Use the syntax below to clear any data in memory and reopen the HSE practical data:

```
clear

use"C:\ESDS SAE practical\Data\HSE data.dta"
```

1.3.9: Calculate age and sex specific disability proportions (MO_OBS_RT) as previously (see start of task 2).

```
gen count_w=1*weight
gen mobility_w=mobility*weight
sort sex age
by sex age: egen MO_num=total(mobility_w)
by sex age: egen MO_denom=total(count_w)
gen MO_OBS_RT=MO_num/MO_denom
```

1.3.10: Now run the syntax below to calculate the numbers of people (N_x in equation 8) at each age involved in the calculations of mobility disability rates:

```
egen mobilitycount=count(MO_OBS_RT), by (age sex)
```

1.3.11: Run the syntax below to calculate the age specific binomial weights (w_x in equation 8) associated with age and sex specific rates of mobility disability:

```
gen mobilityweight=mobilitycount/(MO_OBS_RT*(1-MO_OBS_RT))
```

1.3.12: Drop all duplicate values of age and sex so that we are left with a spreadsheet with one row for each single year of age:

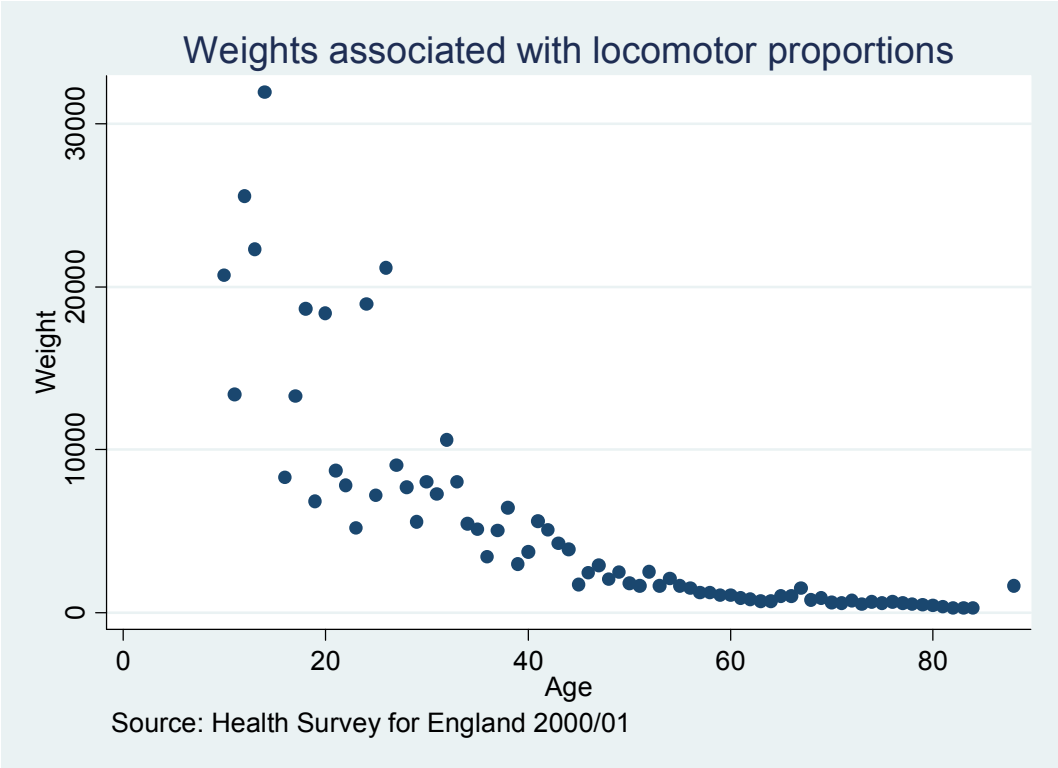
```
duplicates drop age sex, force
```

1.3.13: Use the syntax below to graph the age-specific weights for mobility disability (for males):

```
twoway (scatter mobilityweight age if sex==1), ytitle(Weight)
xtitle(Age) title(Weights associated with mobility proportions)
caption(Source: Health Survey for England 2000/01)
```

You should now have a graph matching figure 14. Examination of this graph shows that weights are highest at the youngest ages and lowest for the elderly. The weights therefore match our expectations of the reliability of proportions in figure 11 which fluctuates most at the oldest ages and least at the youngest.

Figure 14: Age specific weights for mobility proportions (males)



1.3.14: We can now refit the exponential model for the mobility schedule (males) but this time using our weights (mobilityweight). Note - Mobility weights are specified as analytic weights (aweight). Aweights are inversely proportional to the variance of an observed proportion.

```
nl (MO_OBS_RT=exp({a}+{b}*age)) [aweight=mobilityweight] if sex==1
```

1.3.15: Run the syntax below to generate predicted values of mobility rates for males:

```
predict pred_MO_ENG_M if sex==1
```

1.3.16: We can now refit the exponential model for the mobility schedule (females) but this time using our weights (mobilityweight).

```
nl (MO_OBS_RT=exp({a}+{b}*age)) [aweight=mobilityweight] if sex==2
```

1.3.17: Run the syntax below to generate predicted values of mobility rates for females:

```
predict pred_MO_ENG_F if sex==2
```

1.3.18: Generate a single variable 'pred_MO_ENG' containing the model mobility rates for males and females

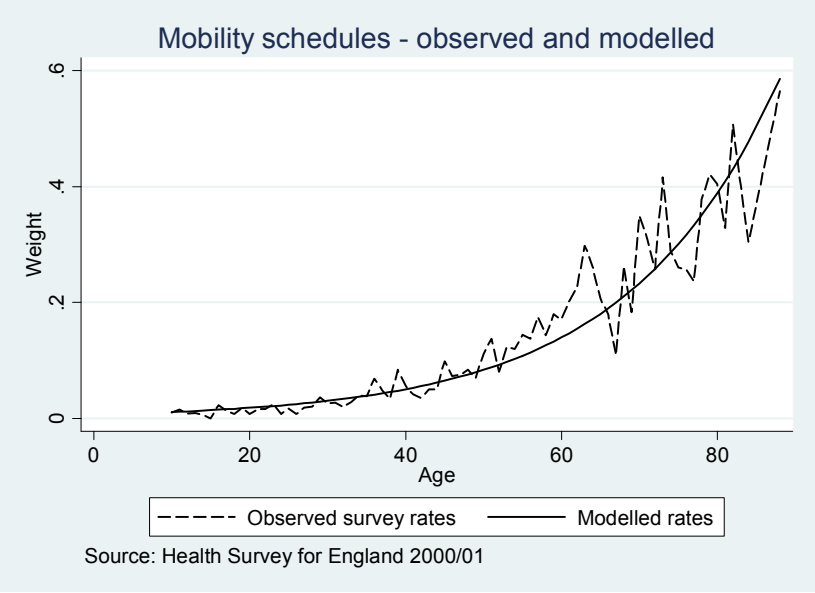
```
gen pred_MO_ENG=pred_MO_ENG_M if sex==1
replace pred_MO_ENG=pred_MO_ENG_F if sex==2
```

1.3.19: We can now produces graphs of the model and observed mobility schedules for males:

```
twoway (line MO_OBS_RT age if sex==1, lcolor(black) lpattern(dash))
(line pred_MO_ENG age if sex==1, lcolor(black)), ytitle(Weight)
xtitle(Age) title(Mobility schedules - Males) caption(Source: Health
Survey for England 2000/01) legend(order(1 "Observed survey rates" 2
"Modelled rates"))
```

You should now have the graph below. If you compare this with figure 13 you will notice that the model gives a better fit to the observed data particularly at the youngest ages where the model no longer overestimates the observed proportions.

Figure 15: Mobility proportions - observed and modelled



1.3.20: Sort data by sex then by age. Save your data

```
sort sex age
save"C:\ESDS SAE practical\Saved practical work\Practical 1 - task
3.dta", replace
```

Practical 1 - Task 4: Regional disability schedules

This task creates model schedules of mobility rates for the nine regions in England and involves two steps. The first step shows why we should be interested in calculating regional schedules by demonstrating that the differences in the probability of an individual having a mobility disability are significant across the nine Government Office Regions. The second step generates model mobility schedules for each of the nine Government Office Regions.

1.4.1: First clear the data in Stata's memory and open the HSE practical data:

```
Clear
use"C:\ESDS SAE practical\Data\HSE data.dta"
```

You may notice that 2,493 individuals take the value of -1 (item not applicable) for the Gora (Government Office Region) variable (run the command tab gora to observe this). These are the individuals living in institutions who were not asked about their region of residence. It is not possible to include the institutional population in the regional models without making an assumption about the regional distribution of this population group. For this reason the institutional population are dropped from the regional analysis undertaken here.

1.4.2: The following syntax drops individuals who have no GOR of residence attached to their data (institutional population). We can't include these cases in our regional analysis:

```
drop if gora==-1
```

Now fit the logistic regression model to predict the probability that a person has a mobility disability with explanatory variables of age, age squared and age cubed, sex and region:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{r=2}^{9} w_{ir}\delta_r \qquad \mathbf{9}$$

Where:

π_i = probability that a person has a particular disability type

x=age

w_{ir} =1 if person i lives in region r and 0 otherwise

Age squared and age cubed are included in this model to reflect the quadratic increase in disability prevalence with age.

1.4.3: The syntax below fits the logistic regression model for males defined above:

```
xi: logit mobility age agesq agecub i.gora if sex==1, or
```

After running the model above, part of the output that is returned should include figure 16

Figure 16: Parameter statistics for the regional mobility disability logistic regression model

Parameters	Odds ratio	Std. Err.	T	P>t	[95% Conf.	Interval]
Age	1.08	0.04	2.03	0.04	1.00	1.16
age2	1.00	0.00	0.03	0.98	1.00	1.00
age3	1.00	0.00	-0.30	0.76	1.00	1.00
North West	0.73	0.10	-2.30	0.02	0.56	0.96
Yorkshire and Humberside	0.69	0.10	-2.60	0.01	0.52	0.91
West Midlands	0.69	0.10	-2.57	0.01	0.53	0.92
East Midlands	0.65	0.10	-2.91	<0.0000	0.49	0.87
East of England	0.39	0.06	-6.31	<0.0000	0.29	0.52
London	0.58	0.09	-3.70	<0.0000	0.44	0.78
South East	0.37	0.05	-6.94	<0.0000	0.28	0.49
South West	0.43	0.06	-5.66	<0.0000	0.32	0.58

*Note the standard errors here do not take into account the complex survey design of the HSE

Figure 16 provides strong evidence that region of residence influences the probability of having a mobility disability for males. The reference category is the North East region and we can see from the odds ratio that living in most regions (with the exception of Yorkshire and Humberside and the North West) is associated with a significant reduction in the probability of having a mobility disability.

Fit the same logistic regression model for females. Is there also evidence of regional differences in the probability of having a mobility disability (after accounting for age and sex)?

As there appear to be differences between the regions in terms of the probability of an individual having a mobility disability it is sensible to model the mobility schedule in each region.

1.4.4: Calculate age and sex specific mobility rates in each region in a similar way to that described at the start of task 2 (the only difference being the inclusion of government office region (gora), along with age and sex, in the 'by' command

```
gen count_w=1*weight
gen mobility_w=mobility*weight
sort sex age gora
by sex age gora: egen MO_num=total(mobility_w)
by sex age gora: egen MO_denom=total(count_w)
gen MO_OBS_RT=MO_num/MO_denom
```

1.4.5: Generate binomial weights in the same way as in task 3 using the syntax below (note: these weights are calculated using national (England) age-specific proportions and counts. Weights for regions tend to be unreliable due to the smaller samples that contribute to age specific disability rates at regional level. The national weights are appropriate because they follow an age pattern that is broadly consistent with the regional pattern.

First age/sex specific counts (England)

```
egen mobilitycount=count(MO_OBS_RT), by (age sex)
```

Then age/sex specific disability rates (England)

```
sort sex age
by sex age: egen MO_num_Eng=total(mobility_w)
by sex age: egen MO_denom_Eng=total(count_w)
gen MO_OBS_RT_Eng=MO_num_Eng/MO_denom_Eng
```

Finally calculate the binomial weights:

```
gen mobilityweight=mobilitycount/(MO_OBS_RT_Eng *(1- MO_OBS_RT_Eng))
```

1.4.6: We now need to reduce the dataset so that it comprises a rate of mobility disability for each single year of age, for males and females and for each of the nine regions. This is achieved by dropping duplicate values of age sex and region:

```
duplicates drop age sex gora, force
```

1.4.7: Keep the relevant variables:

```
keep age sex gora MO_OBS_RT mobilityweight
```

1.4.8: Sort the data:

```
sort gora sex age
```

1.4.9: The next step involves fitting an exponential curve to the mobility curve and generating predicted values region by region for males:

```
nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==1 [aweight=mobilityweight]
predict pred_MO1_M
```

```
nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==2 [aweight=mobilityweight]
predict pred_MO2_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==3 [aweight=mobilityweight]
predict pred_MO3_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==4 [aweight=mobilityweight]
predict pred_MO4_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==5 [aweight=mobilityweight]
predict pred_MO5_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==6 [aweight=mobilityweight]
predict pred_MO6_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==7 [aweight=mobilityweight]
predict pred_MO7_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==8 [aweight=mobilityweight]
predict pred_MO8_M

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==1&gora==9 [aweight=mobilityweight]
predict pred_MO9_M
```

1.4.10: The syntax below fits an exponential curve to the female mobility curve for each of the nine English regions and generates variable giving predicted values for each region:

```
nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==1 [aweight=mobilityweight]
predict pred_MO1_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==2 [aweight=mobilityweight]
predict pred_MO2_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==3 [aweight=mobilityweight]
predict pred_MO3_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==4 [aweight=mobilityweight]
predict pred_MO4_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==5 [aweight=mobilityweight]
predict pred_MO5_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==6 [aweight=mobilityweight]
predict pred_MO6_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==7 [aweight=mobilityweight]
predict pred_MO7_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==8 [aweight=mobilityweight]
predict pred_MO8_F

nl (MO_OBS_RT=exp({a}+{b}*age)) if sex==2&gora==9 [aweight=mobilityweight]
predict pred_MO9_F
```

1.4.11: The syntax above created eighteen variables containing predicted values from our exponential curve models for males and females in each of the nine regions. A

single variable of predicted mobility rates (pred_LM) can be generated from these nine variables:

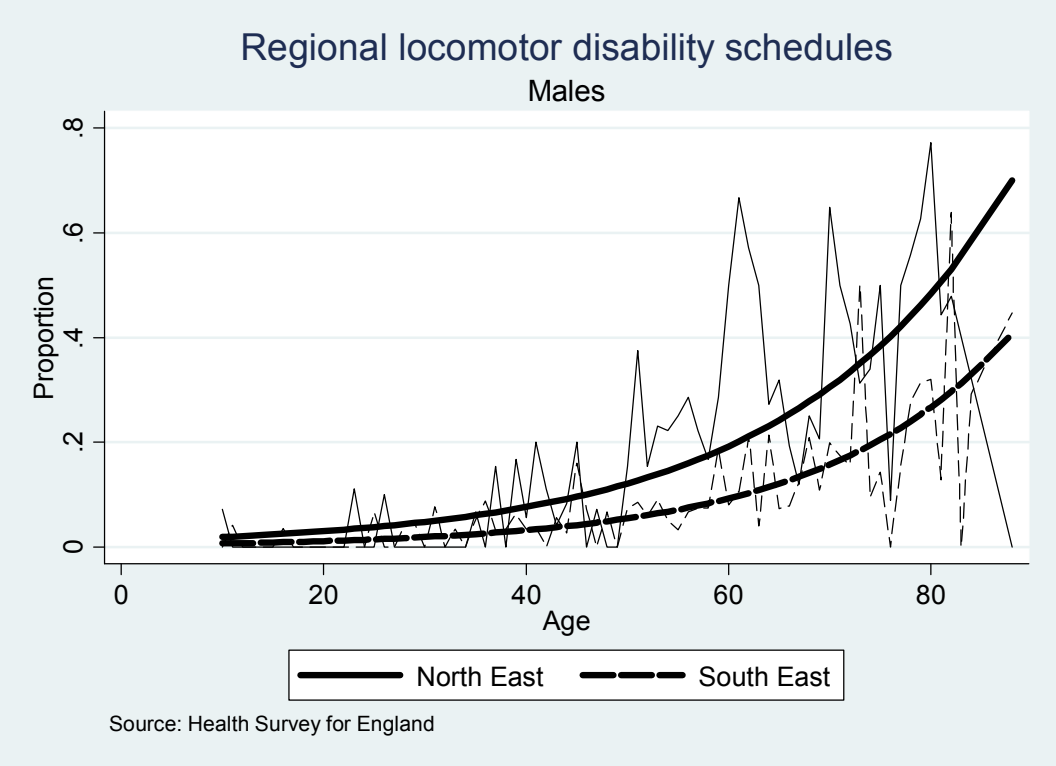
```
gen pred_MO=.

replace pred_MO=pred_MO1_M if gora==1&sex==1
replace pred_MO=pred_MO2_M if gora==2&sex==1
replace pred_MO=pred_MO3_M if gora==3&sex==1
replace pred_MO=pred_MO4_M if gora==4&sex==1
replace pred_MO=pred_MO5_M if gora==5&sex==1
replace pred_MO=pred_MO6_M if gora==6&sex==1
replace pred_MO=pred_MO7_M if gora==7&sex==1
replace pred_MO=pred_MO8_M if gora==8&sex==1
replace pred_MO=pred_MO9_M if gora==9&sex==1


replace pred_MO=pred_MO1_F if gora==1&sex==2
replace pred_MO=pred_MO2_F if gora==2&sex==2
replace pred_MO=pred_MO3_F if gora==3&sex==2
replace pred_MO=pred_MO4_F if gora==4&sex==2
replace pred_MO=pred_MO5_F if gora==5&sex==2
replace pred_MO=pred_MO6_F if gora==6&sex==2
replace pred_MO=pred_MO7_F if gora==7&sex==2
replace pred_MO=pred_MO8_F if gora==8&sex==2
replace pred_MO=pred_MO9_F if gora==9&sex==2
```

1.4.12: The syntax below creates a graph of observed and modelled mobility schedules for the North East and the South East.

```
twoway (line pred_MO age if gora==1&sex==1, lwidth(thick) lcolor(black))
(line pred_MO age if gora==8, lwidth(thick) lcolor(black)
lpattern(longdash)) (line MO_OBS_RT age if gora==1&sex==1, lcolor(black)
lwidth(thin)) (line MO_OBS_RT age if gora==8&sex==1, lwidth(thin)
lcolor(black) lpattern(dash)) if sex==1, ytitle(Proportion) xtitle(Age)
title(Regional mobility disability schedules) subtitle(Males) note(Source:
Health Survey for England) legend(order(1 "North East" 2 "South East"))
```

You should now have a graph that matches figure 17 (below). It is clear from this graph that the Health Survey for England provides strong evidence to suggest that rates of mobility disability are much higher in the North East than the South East.

Figure 17: Mobility disability schedules - North East and South East



1.4.13: Drop redundant variables:

```
drop pred_MO1_M - pred_MO9_F
```

1.4.14: Sort then save your data file

```
sort gora sex age
save"C:\ESDS SAE practical\Saved practical data\Practical 1 - task
4.dta", replace
```

Try developing your own syntax to generate model schedules for overall disability (disab) by fitting an exponential curve to regional data.

Practical 1 - Task 5: Generating district estimates of the numbers of people with mobility disabilities

In the previous tasks of practical 1 you have fitted models that smooth the observed rates of mobility disability across the age range. Model schedules of age specific rates were developed for England (task 3) and then for each of the nine regions (task 4). It is possible to use the modelled rates in combination with data on population totals in a particular district to estimate the numbers of people who have a mobility disability in that district. In this task this is achieved in two ways:

Method 1 – Mobility schedules for England

1. Multiply model rates of mobility disability for England at each single year of age (for males and females) by the district population count for males and females at each single year of age.
2. Sum the district estimates of the numbers of males and females with a mobility disability across all ages.

Method 2 – Mobility schedules for each region

1. Multiply model rates of mobility disability for the region (within which a district is contained) at each single year of age (and for males and females) by the district population count for males and females at each single year of age.
2. Sum the district estimates of the number of males and females with a mobility disability across all ages.

Methods 1 and 2 can be expressed algebraically as below:

$$L_{.d} = \sum_{x=10}^{88} N_{xd}\, p_{x.} \qquad \textbf{10}$$

$$L_{.d} = \sum_{x=10}^{88} N_{xd}\, p_{xr} \qquad \textbf{11}$$

Where:

$L_{..d}$ = number of people with a mobility disability in district d across all ages (10+) and for males and females

N_{xsd} = number of people at age x (x=10,11,12,….84,88), sex s (s=1,2) and district d

$p_{xs.}$ = mobility disability rate at age x and sex s for England

p_{xsr} =mobility disability rate at age x and sex s for region r (r=1,….,9)

This final task of practical 1 develops estimates of numbers of people with a mobility disability for six local authority districts (Easington, Bury, Wakefield, Barnet, South Bucks and Stroud) in 2001 (using census data). This is achieved using a prepared data file containing population counts (by age and sex) and the model rates developed in previous tasks.

1.5.1: Open the file for task 5

```
use "C:\ESDS SAE practical\Data\Practical 1 - task 5 - data.dta",
clear
```

The data file you have opened contains a row for each single year of age (10, 11,….84,88) for males and females in each of the six districts. Figure 18 shows the variables that are included in this dataset:

Figure 18: Variables in the task 5 data file

Variable name	Description
Zonecode	ONS code for local authority district
Zonename	Name of local authority district
Gora	Government office region of local authority district
Sex	Sex (1=male, 2=female)
Age	Age
Pop_2001	District population count in 2001 (Census)
Pop_2021	Projected district population count in 2021 (2006 based population projections
Pred_MO	Regional predicted regional mobility rates (from practical 4)
Pred_MO_ENG	England predicted mobility rates (from practical 3

1.5.2: View the data file to confirm the explanation of the data file above

```
Browse
```

1.5.3: Now generate counts of people with disability (at each age and by sex) by multiplying the regional rates of mobility disability by the population counts in 2001.

```
gen mo_pop_Reg_01=pop_2001 * pred_MO
```

1.5.4: Now generate counts of people with disability (at each age and by sex) by multiplying the England rates of mobility disability by the population counts in 2001.

```
gen mo_pop_Eng_01=pop_2001 * pred_MO_ENG
```

1.5.5: Browse the data viewer to confirm that the new variables, containing estimated counts of mobility disability in 2001 have been created.

```
browse
```

Compare counts of mobility disability in Easington generated using England rates (mo_pop_Eng) and regional rates (mo_pop_Reg)? Which set of estimates are larger and why do you think this is the case (hint: look at figure 17)?

1.5.6: Now, sort the data by zonecode then calculate total counts of mobility disability for each of the districts in 2001.

```
sort zonecode
```

1.5.7: First for the estimates based on regional schedules.

```
by zonecode: egen mo_tot_01_reg= sum(mo_pop_Reg_01)
```

1.5.8: Then for the estimates based on England schedules

```
by zonecode: egen mo_tot_01_eng= sum(mo_pop_Eng_01)
```

1.5.9: Generate a total population count

```
by zonecode: egen pop_tot_01= sum(pop_2001)
```

1.5.10: Drop duplicate values of age and sex so that we are left with a count of people with a mobility disability in each district

```
duplicates drop zonecode, force
```

1.5.11: Finally, generate % of population with mobility disability

```
gen percent_eng=(mo_tot_01_eng/pop_tot_01)*100
gen percent_reg=(mo_tot_01_reg/pop_tot_01)*100
```

If you view the data browser you should observe the district estimates of populations with mobility disability that are shown in figure 19 for 2001. The figure clearly shows that the importance of taking into account regional variability from the national age specific rates. For example, in South Bucks using the South East rates to generate estimate leads to 1,742 fewer cases of mobility disability than if the lower rates observed in England are used.

Figure 19: District estimates of the population with a locomotor disability (2001)

District	2001 estimate using England rates		2001 estimates using regional rates	
	Number	%	Number	%
Barnet	28,881	10.4	28,881	10.6
Bury	16,804	10.7	16,804	12.9
Wakefield	29,833	10.8	29,833	11.8
South Bucks	6,525	12.0	6,525	8.8
Easington	9,113	11.1	9,113	14.3
Stroud	11,479	12.1	11,479	9.4

Whilst the regional estimates represent an improvement on those generated using the National (England rates) they do not incorporate potential variability in levels of disability within a region. Practical 2 addresses this weakness through the use of relational models to generate district schedules of mobility disability.

Try repeating the above task to produce estimates of the population with a mobility disability in 2021. Use the same disability rates but different population totals.

Practical 2 – Relational models

Introduction

The previous practical used information on the age/sex population structure of an area (estimated from census/population projections) and smoothed age and sex specific rates of mobility disability for English regions (estimated from the Health Survey for England) to estimate district populations with a mobility disability. The key drawback of this approach is that levels of disability vary within a region and so simply applying the regional rate will not give accurate results for sub-regional areas. This practical uses relational models to overcome this weakness generating model mobility disability schedules for each district (rather than for each region as in the previous practical).

The general strategy is to use a relational model to quantify the relationship between the Census LLTI schedule for England and the mobility disability schedules for England as measured in the Health Survey for England. It turns out that after we take a logit transform of the LLTI and mobility rates the relationship between the two logit schedules is remarkably linear and so can be described using two parameters (intercept and slope). The relational parameters from this model are stored and then used to adjust district LLTI curves (which are available from the Census) allowing us to derive local schedules for different disability types in the absence of survey estimates. Under this method if two districts have different levels of LLTI they will also have different levels of mobility disability. An algebraic specification of the model is given below:

Let:

p_{xsrd} = rate of disability at age x (x=10,11,….84,88) and sex s (s=1,2) in district r for disability type d.

l_{xsr} = prevalence of LLTI at age x and sex s in district r (Census 2001)

$p_{xs.d}$ = rate of disability at age x and sex s in England for disability type d (HSE00/01)

$l_{xs.}$ = prevalence of LLTI at age x and sex s in England (Census 2001)

Then the predicted prevalence of disability d for district r and age x is derived from:

$$\frac{1}{2}\log_e\left(\frac{\hat{p}_{xsrd}}{1-\hat{p}_{xsrd}}\right)=\hat{\alpha}+\hat{\beta}\left(\frac{1}{2}\log_e\left(\frac{l_{xsr}}{1-l_{xsr}}\right)\right) \quad \textbf{12}$$

Where $\hat{\alpha}$ and $\hat{\beta}$ are estimated from equation 5 (England level data):

$$\frac{1}{2}\log_e\left(\frac{p_{xs.d}}{1-p_{xs.d}}\right)=\alpha+\beta\left(\frac{1}{2}\log_e\left(\frac{l_{xs.}}{1-l_{xs.}}\right)\right)+e_x \quad \textbf{13}$$

In this practical you will learn to:

1. Apply a logit transform to schedules of LLTI and mobility disability rates
2. Fit a Brass relational model (2 parameters) to capture the relationship between the logit schedule of LLTI rates and the logit schedule of mobility disability rates for England
3. Use the parameter estimates from the national relational model to adjust district logit LLTI schedules and derive model rates of mobility disability schedules for each district
4. Estimate district populations with a mobility disability using model rates of mobility disability and the district population age structure.

This practical uses the dataset 'Practical 2.dta' which contains a row for each single year of age (10,11, 12,…..83, 84, 88) for males and females in England and each of the six casestudy districts. For more details on the production of this dataset see the documentation for the small area estimation teaching datasets.

The variables that are included in this dataset are shown in figure 20.

Figure 20: Variables in the practical 2 task 1 dataset

Variable	Description
Zonecode	ONS code of local authority district
Zonename	Name of local authority district
Gora	Name of government office Region
Sex	Sex (1=male, 2=female)
age	Age (10,11,……83,84,88)
llti_2001	Age and sex specific LLTI rates for England
pop_2001	Population counts (by single year of age and sex) in 2001 (census)
pop_2021	Population counts (by single year of age and sex) in 2021 (2006 based population projections)

MO_OBS_RT	Age and sex specific rates of mobility disability for England
D_OBS_RT	Age and sex specific rates of overall disability for England
PC_OBS_RT	Age and sex specific rates of personal care disability for England
HR_OBS_RT	Age and sex specific rates of hearing disability for England
ST_OBS_RT	Age and sex specific rates of sight disability for England
mobilitycount	The number of HSE observations (individuals) at each single year of age that contribute towards the calculations for the mobility rates for males and females (England). This variable is used (along with MO_OBS_RT) to calculate analytic weights (aweights) for use when fitting the relational models
Disabweight	Relational model weights for overall disability
Pcareweight	Relational model weights for personal care disability
Hearweight	Relational model weights for hearing disability
Sightweight	Relational model weights for sight disability

Practical 2 - Task 1: Fitting a relational model

In task 1 we will fit a relational model to link the census LLTI schedule for England (llti_2001_ENG) to the HSE mobility disability schedule for England (MO_OBS_RT)

2.1.1: First open the 'Practical 2 task 1' dataset and keep only the England data - the first stage of the relational approach is at national level

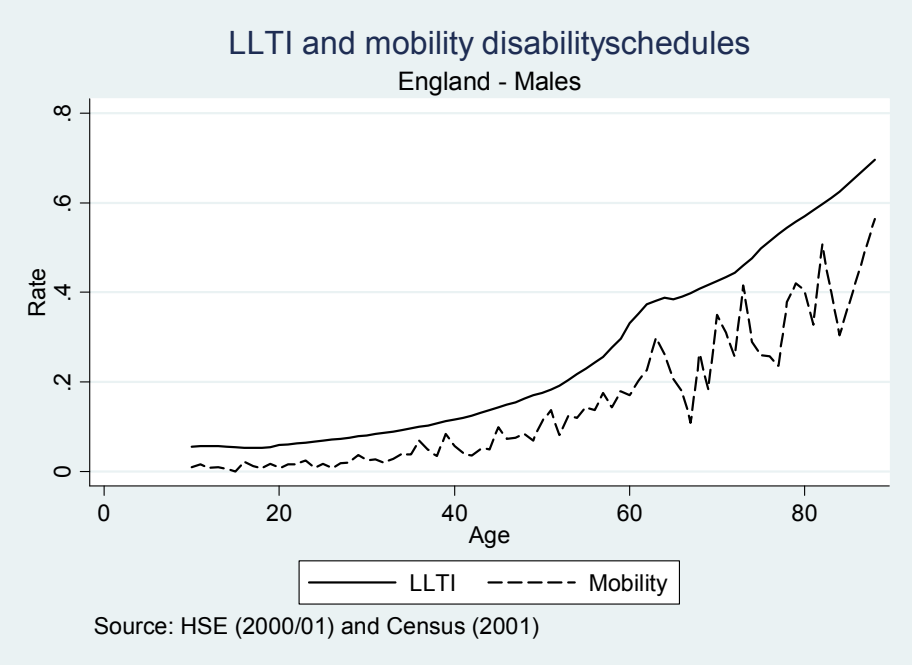
```
use "C:\ESDS SAE practical\Data\Practical 2 data.dta", clear
keep if zonename=="ENGLAND"
```

2.1.2: In order to observe the similar pattern of the LLTI and mobility disability (and the suitability of the relational approach described above) we can plot the two schedules together on the same graph:

```
twoway (line llti_2001 age, lcolor(black) lpattern(solid)) (line
MO_OBS_RT age, lcolor(black) lpattern(dash)) if sex==1, ytitle(Rate)
xtitle(Age) title(LLTI and mobility disability schedules)
subtitle(England - Males) caption(Source: HSE (2000/01) and Census
(2001)) legend(order(1 "LLTI" 2 "Mobility"))
```

Running the syntax above should give the graph in figure 22. The LLTI and mobility disability schedules follow a very similar age patterns - if we were to move the LLTI curve downwards then the adjusted line would give a good fit to the curve of mobility disability rates.

Figure 21: LLTI and mobility disability schedules (England - Males)



In the next part of task 1 we will calculate logit LLTI (l) and mobility (m) rates for each year of age (x) and sex (s). This involves taking the logit of the LLTI and mobility disability rates for England as below (note: multiplying the logit transform by a half is a convention of the relational modelling approach developed by Brass – the same model rates are derived with or without the multiplier):

Let:

$l_{xs.}$ = LLTI rate at age x (x=10,11,12,…84,88) and for sex s (s=1 (male), 2 (female) for England

$m_{xs.}$ =Mobility disability rate at age x and sex s for England

$$\frac{1}{2}\left(Logit\left(l_{xs}\right)\right)=\frac{1}{2}\log_e\left(\frac{l_{xs.}}{1-l_{xs.}}\right) \quad \textbf{14}$$

$$\frac{1}{2}\left(Logit\left(m_{xs}\right)\right)=\frac{1}{2}\log_e\left(\frac{m_{xs.}}{1-m_{xs.}}\right) \quad \textbf{15}$$

2.1.3: The syntax below creates two variables containing the logit LLTI and logit mobility schedules:

```
gen logit_LLTI=0.5*ln(llti_2001/(1-llti_2001))
gen logit_mob=0.5*ln( MO_OBS_RT/(1- MO_OBS_RT))
```

If we produce a scatterplot comparing the two logit schedules we can see the linear nature of the relationship between the two schedules. This is useful because it means that we can represent the relationship between them using just two parameters that quantify the slope and intercept of the linear relationship.

2.1.4: Use the syntax below to produce a scatterplot of logit_LLTI and logit_mob

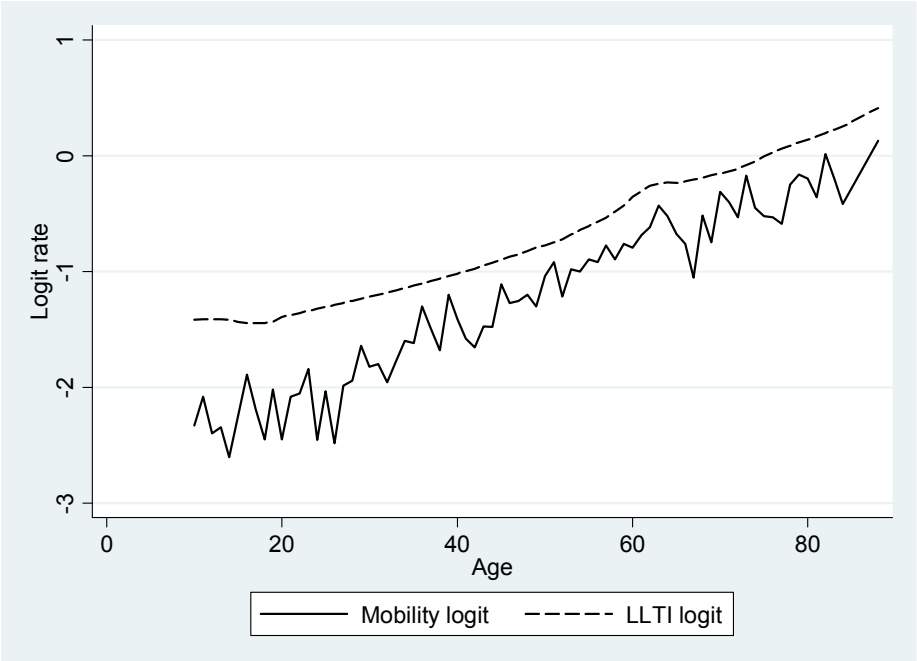
```
twoway (scatter logit_mob logit_LLTI) if sex==1, ytitle(Logit of
mobility rate) xtitle(Logit of LLTI rate) yline(0) xline(0)
caption(Source: census 2001 and HSE 2000/01)
```

Figure 22: Scatterplot of the relationship between the logit LLTI schedule and the logit mobility disability schedule (Males – England)

2.1.5 An alternative way to view the relationship is to produce line graphs of the logit schedules against age using the syntax below:

```
twoway (line logit_mob age, lcolor(black)) (line logit_LLTI age,
lcolor(black) lpattern(dash)) if sex==1, ytitle(Logit rate)
xtitle(Age) legend(order(1 "Mobility logit" 2 "LLTI logit"))
```

Figure 23: Logit LLTI and logit mobility disability schedules (Males -England)



Before we fit a relational model to express the relationship between the two logit schedules in figure 23 it is important that we first calculate weights that play a similar role to those defined in practical 1 – task 3. A model that uses the logit of a proportion as the dependent variable should use weights at each age x (w_x) based on equation 7-1 (Congdon 1993):

$$w_{xs} = m_{xs.}(1 - m_{xs.})N_{xs} \qquad \textbf{16}$$

Where:

$m_{xs.}$ = the rate of mobility disability at age x and sex s in England

N_x = the HSE sample size at age x and sex s in England

The result of weights specification above is that proportions that are close to 0 or 1 or that are based on small samples are given less weight during the model fitting process. Conversely, proportions that are near 0.5 or that are based on larger sample sizes are given greater weight.

2.1.6 The following syntax calculates the age and sex specific weights (w_{xs}):

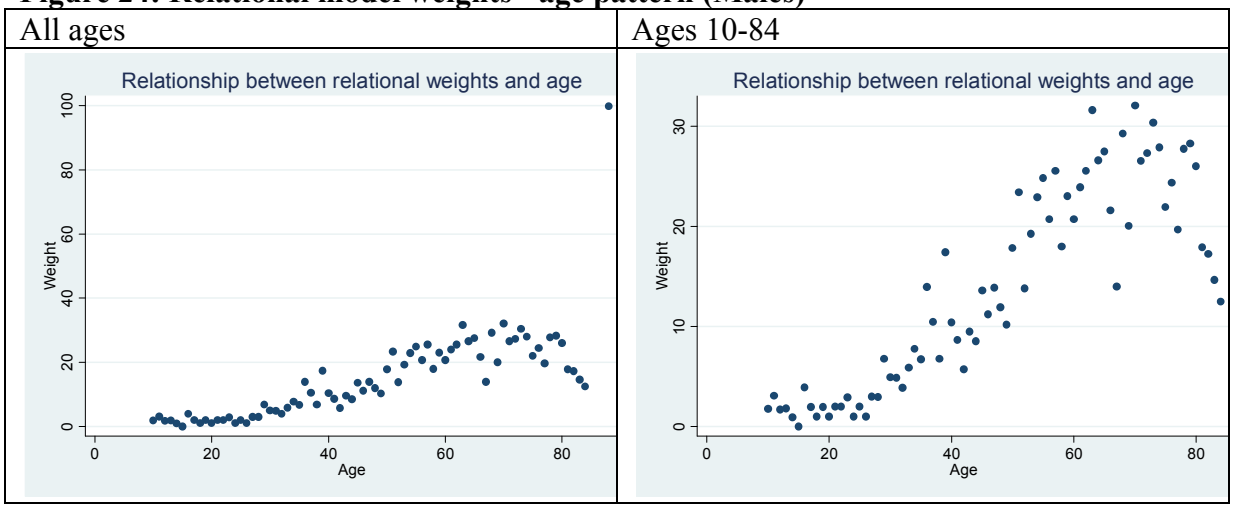
```
gen rel_weights=MO_OBS_RT *(1- MO_OBS_RT )* mobilitycount
```

2.1.7: We can produce a scatterplot of the weight age pattern and another scatterplot that excludes the oldest age group using the syntax below:

```
twoway (scatter rel_weights age) if sex==1, ytitle(Weight)
xtitle(Age) title(Relationship between relational weights and age)

twoway (scatter rel_weights age) if sex==1&age<87, ytitle(Weight)
xtitle(Age) title(Relationship between relational weights and age)
```

Running syntax line 2.1.7 should give the graphs displayed in Figure 25. The smallest weights are found at the youngest ages and examination of the observed fluctuations in the logit locomotor disability schedule suggests this is least reliable part of this schedule (see figure 24). Weights drop at the very oldest ages because sample sizes become markedly smaller after the age of 70. There is a very large weight for the largest age (88) as this includes all people who are 85 or older and as such contains a large sample compared to the sample sizes for single years

Figure 24: Relational model weights - age pattern (Males)



We can now fit the relational model (see below) which describes the (linear) relationship between the logit mobility disability schedule ($m_{xs.}$) (dependant variable) and the logit LLTI schedule ($l_{xs.}$) (explanatory variable):

$$\frac{1}{2}\log_e\left(\frac{m_{xs.}}{1-m_{xs.}}\right)=\alpha+\beta\left(\frac{1}{2}\log_e\left(\frac{l_{xs.}}{1-l_{xs.}}\right)\right)+e_x \qquad \textbf{17}$$

2.1.7: The syntax below is used to fit the model for males and females separately and then produces model logit mobility schedules for males (logit_mob_mod_M) and females (logit_mob_mod_M):

```
regress logit_mob logit_LLTI if sex==1 [aweight=rel_weights]
predict logit_mob_mod_M if sex==1

regress logit_mob logit_LLTI if sex==2 [aweight=rel_weights]
predict logit_mob_mod_F if sex==2
```

2.1.8: We can then create a single variable containing the predicted model logit mobility schedules for both males and females:

```
gen logit_mob_mod=logit_mob_mod_M if sex==1
replace logit_mob_mod=logit_mob_mod_F if sex==2
```

2.1.9: The syntax below converts the model logit mobility schedule to a schedule of mobility rates

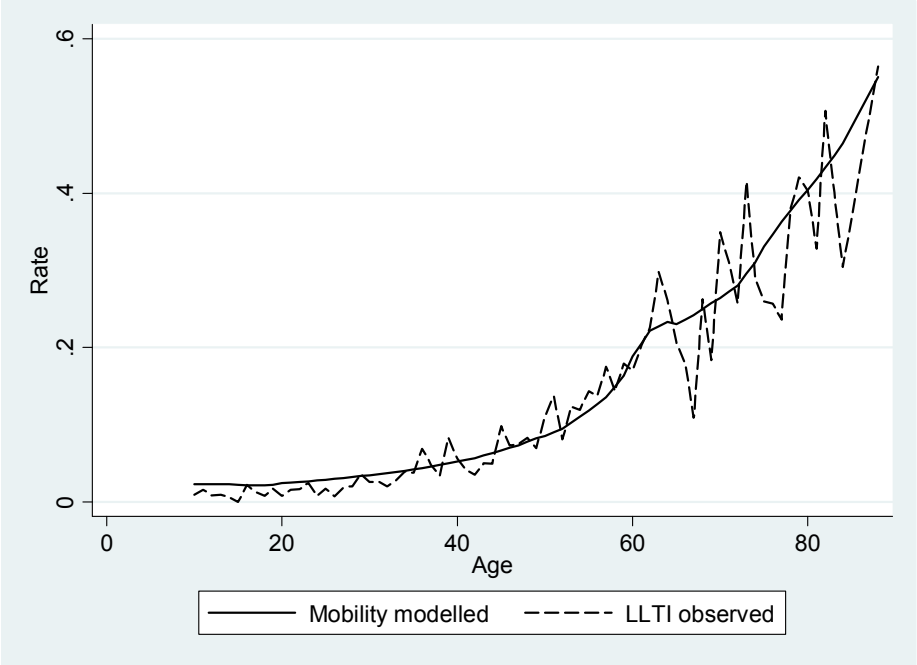
```
gen mob_mod=(exp(2*logit_mob_mod))/(1+exp(2*logit_mob_mod))
```

2.1.10 Finally, we can graph the modelled and observed mobility disability schedules to confirm that relational models used in this way give a good fit.

```
twoway (line mob_mod age, lcolor(black)) (line MO_OBS_RT age,
lcolor(black) lpattern(dash)) if sex==1, ytitle(Rate) xtitle(Age)
legend(order(1 "Modelled" 2 "Observed"))
```

After running the syntax above you should have produced an equivalent graph to figure 26.

Figure 25: Mobility schedules - observed and modelled



Part of the output from the relational models you fitted (see output associated with syntax line 2.1.7) should have included estimates of the α (_cons) and β (logit_LLTI) parameters (see figures 27 and 28). the α parameter represents the intercept and the β parameter the slope of the linear relationship between the logit LLTI and logit mobility disability schedules. For both males and females α is approximately equal to -0.3 and β is approximately equal to 1; this tells us that we need to move the logit llti schedule

downwards (as α is negative) but don't really need to change its slope in order to reproduce the logit mobility disability schedule.

Look at figure 23 (which plots the relationship between the logit LLTI and logit mobility disability schedules) to confirm that these parameter estimates seem reasonable.

Figure 26: Relational model - Parameter estimates (males)

	Coef.	Std. Err.	T	P>t	[95% Conf.	Interval]
β	1.08	0.04	27.51	<0.0000	1.00	1.16
α	-0.35	0.02	-15.38	<0.0000	-0.40	-0.31

Figure 27: Relational model - Parameter estimates (females)

	Coef.	Std. Err.	T	P>t	[95% Conf.	Interval]
β	1.09	0.02	47.39	<0.0000	1.05	1.14
α	-0.30	0.01	-20.50	<0.0000	-0.33	-0.27

The parameter estimates above are important as we shall use them to adjust district LLTI schedules in the following task

Practical 2 - Task 2: Generating district mobility disability schedules

This task uses the relational parameter estimates from the previous task (see figures 27 and 28) along with district LLTI schedules to generate district mobility disability schedules. A new dataset 'practical 2 – task 2' is used for this task which contains a row for each single year of age (10,11, 12,…..83, 84, 88) for males and females and for each of the six casestudy districts (Barnet, Bury, Easington, South Bucks, Stroud, Wakefield). The variables that are included in this dataset are shown in figure 21.

2.2.1: The first step if to open the practical 2 – task 2 dataset and add two new variables containing the α (intercept) and β (slope) estimates for males and females:

```
clear

use "C:\ESDS SAE practical\Data\Practical 2 data.dta", clear

gen a=-0.35111117 if sex==1
replace a=-0.2969268 if sex==2

gen b=1.080914 if sex==1
replace b=1.093486 if sex==2
```

2.2.2: Now we must calculate logit LLTI schedules for each district:

```
gen logit_LLTI=0.5*ln(llti_2001/(1-llti_2001))
```

2.2.3: The syntax below adjusts the district logit LLTI curves using the parameters α (intercept) and β (slope)

```
gen mob_logit_mod=a+b*logit_LLTI
```

2.2.4: The following syntax converts the model logit mobility schedule (derived above) to a schedule of mobility rates

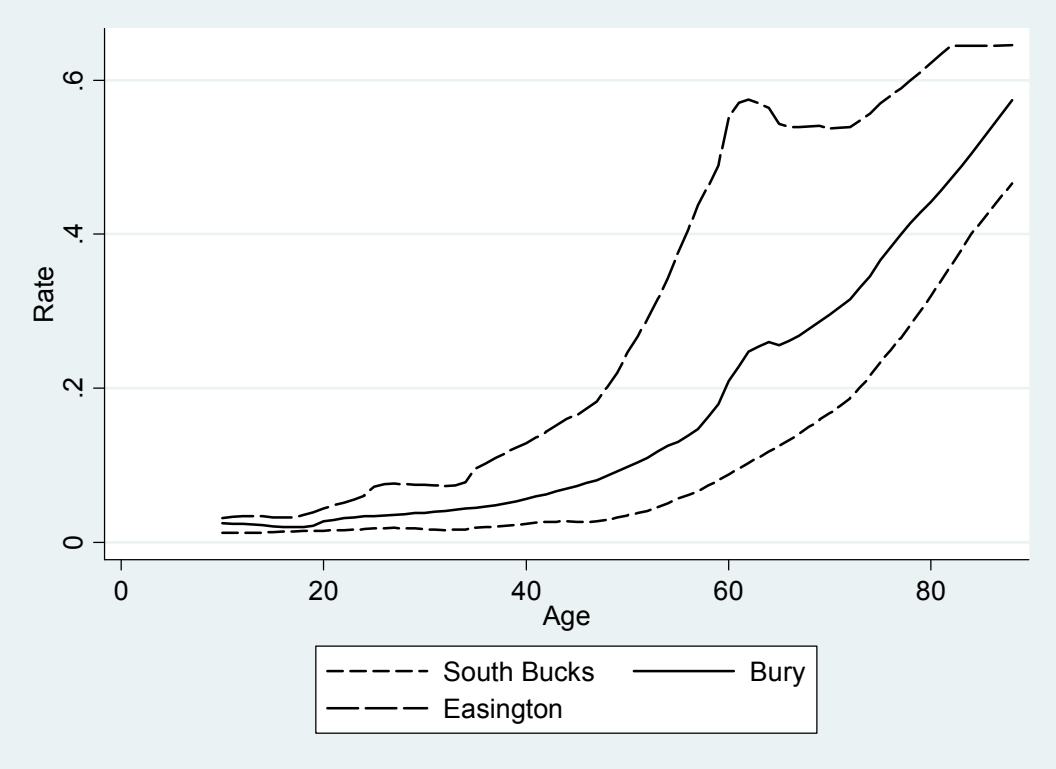
```
gen mob_modelled=(exp(2*mob_logit_mod))/(1+exp(2*mob_logit_mod))
```

2.2.5: Finally, we can produce a graph comparing model mobility disability schedules for a selection of districts

```
twoway (line mob_modelled age if zonename=="South Bucks",
lcolor(black) lpattern(dash)) (line mob_modelled age if
zonename=="Bury", lcolor(black) lpattern(solid)) (line mob_modelled
age if zonename=="Easington", lcolor(black) lpattern(longdash)) if
sex==1, ytitle(Rate) xtitle(Age) legend(order(1 "South Bucks" 2
"Bury" 3 "Easington"))
```

Figure 29 shows the graph comparing model mobility disability schedules for South Bucks, Bury and Easington. Clearly the local information we have on LLTI and the correlations we observe between LLTI and mobility disability imply a great deal of variation in levels of mobility disability between these districts. This variability would not be picked up in the estimates derived in practical 1.

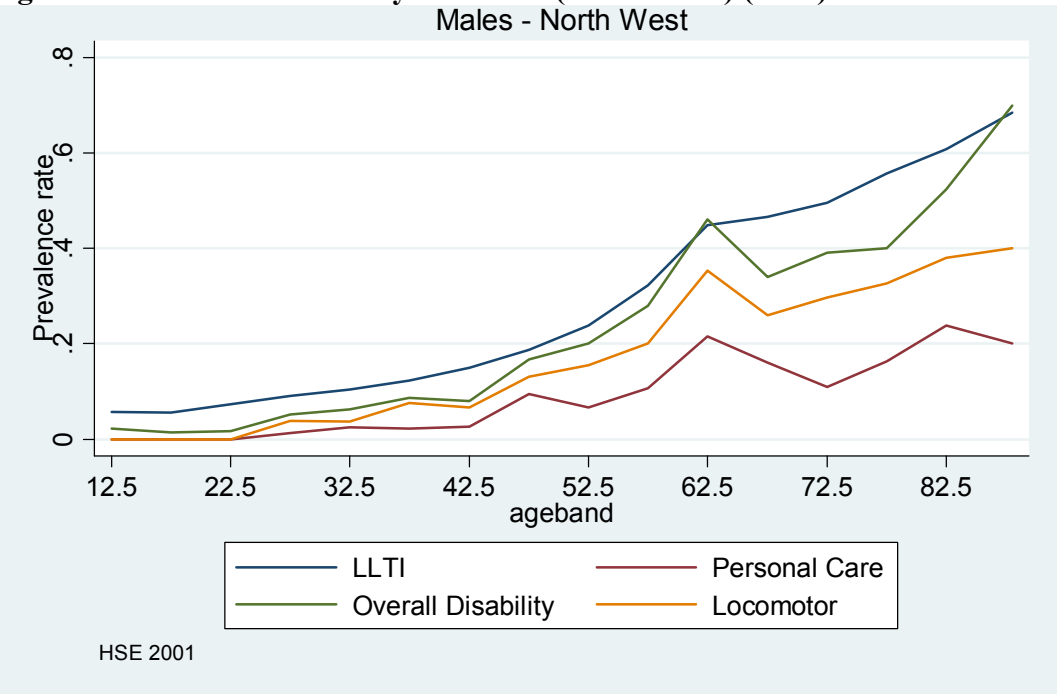
Figure 28: Model mobility disability schedules: South Bucks, Bury and Easington (males)



The districts of Bury and Easington display a kink in the LLTI schedule around retirement age which is preserved in the model mobility disability schedules. The LLTI retirement kink is a feature noted by other researchers. For example, Bellaby (2006) finds a tailing off in the increase in LLTI after retirement ages, particularly for those in manual occupations, and a similar result from clinical assessments of health using standardised methods (e.g. forced expiratory volume (FEV1), blood pressure (allowing for control by medication), and body mass index). Figure 30, shows evidence of a retirement kink for different disability types in

the North West region providing evidence to support its transfer to different disability types in the relational estimation approach.

Figure 29: LLTI and disability schedules (North West) (2001)



2.2.6: Now, sort the data and then save the model rates you have created:

```
sort sex zonecode age

save "C:\ESDS SAE practical\Saved practical work\Practical 2 - task 2.dta", replace
```

Practical 2 - Task 3: Generating district estimates of the numbers of people with mobility disabilities

This task multiplies the model age and sex specific rates of mobility disability (calculated in the previous task) by the appropriate population counts in each district to generate local estimates of the numbers of people with a mobility disability.

2.3.1: First open the data file you saved at the end of the last task:

```
use "C:\ESDS SAE practical\Data\Practical 2 - task 2.dta", clear
```

2.3.2: Now generate a new variable containing counts of people with a mobility disability in 2001 for each district, single year of age and sex

```
gen mob_pop_2001= pop_2001 * mob_modelled
```

2.3.3: Now calculate total counts of mobility disability for each of the districts in 2001 and 2021.

```
sort zonecode
by zonecode: egen mo_tot_01_rel= sum(mob_pop_2001)
```

2.3.4: Generate a variable containing counts of the total population in each district in 2001

```
by zonecode: egen pop_tot_01= sum(pop_2001)
```

2.3.5: Drop the duplicate values of age and sex so that we are left with a count of people with a mobility disability in each district. Keep required variables only

```
duplicates drop zonecode, force
keep zonecode zonename mo_tot_01_rel pop_tot_01
```

2.3.6: Generate a variable containing the % of people with a mobility disability in each district

```
gen prev_01=(mo_tot_01_rel/pop_tot_01)*100
```

If you browse the data editor you will be able to view the information in Figure 31 which compares the district estimates of populations with mobility disabilities in 2001 using regional rates (from the previous practical) and relational model rates (that take into account likely regional variability in levels of mobility disability). The table shows that in many districts taking into account information on the level of LLTI can have a large impact on our

estimate of the population with a mobility disability compared to estimates derived from regional rates of mobility disability.

Figure 30: District estimates of the population with a mobility disability (2001) - Regional and relational models

District	2001 Estimate using regional rates		2001 Estimates using relational model rates	
	Number	%	Number	%
Barnet	29,593	10.6	26,455	9.5
Bury	20,305	12.9	20,391	13.0
Wakefield	32,788	11.8	44,397	16.0
South Bucks	4,786	8.8	4,612	8.5
Easington	11,819	14.4	19,670	23.9
Stroud	8,957	9.4	9,931	10.4

2.3.7: Finally save the data:

```
save "C:\ESDS SAE practical\Saved practical work\Practical 2 - task 3.dta", replace
```

Try repeating the above task to produce estimates of the population with a mobility disability in 2021.

Try repeating tasks 1 and 2 for hearing disability and if you have time other disability types. Note the weights are already calculated for you (hearweight, disabweight,pcareweight, sightweight) . For the disability types of hearing, sight and personal care these weights are based on quinary age groups rather than single years of age. Single year weights were found to be too unstable for these less common disabilities.

Practical 3 – Individual level synthetic regression

Introduction

This final practical generates district estimates of populations with a mobility disability using the individual synthetic regression technique (see 4.2.2 for further details on this approach). Individual level synthetic regression models are a valuable addition to those introduced previously (curve fitting and relational models) because such models are suitable for the estimation of characteristics that may not display a strong age pattern of rates.

In this practical you will learn how to:

1. Fit a logistic regression model to predict the probability of a person having a mobility disability on the basis of covariates of age, age squared and age cubed, sex, LLTI and LLTI*age (interaction between LLTI and age).
2. Use the parameter estimates from this model to calculate the model probabilities of having a mobility disability at each single year of age for males and females and for those with and without an LLTI.
3. Generate district estimates of the population with a mobility disability by multiplying the model probabilities by the appropriate population counts (derived from the Census).

The individual synthetic regression model is defined more formally below:

The logistic synthetic regression model (fitted separately for males and females) is defined below:

Let:

M_{xr} =number of people with a mobility disability at age x in district r

$p_{x.}$ =rate of mobility disability at age x (x=10,11,….84,88) in England

π_i =probability of having a mobility disability for individual i

j=0 (no LLTI) or 1 (has an LLTI)

z_{ij} =1 if j=1 (individual i has an LLTI) and 0 otherwise

N_{xrj} =Number of people at age x in district r and LLTI group j (j=0,1)

$N_{xr.}$ = Number of people at age x in district r

$p_{x.j}$ =probability of having a mobility disability at age x and LLTI group j in England

Then:

$$M_{xr} = \sum_{j=0}^{1} \hat{p}_{x.j} N_{xrj} \qquad \textbf{18}$$

Where:

$$\hat{p}_{x.j} = \frac{\exp\left(\hat{\beta}_0 + \sum_{1}^{3}\left(\hat{\alpha}_t x_i^t\right)_i + \hat{\beta}_1 z_{ij} + \hat{\delta}_1 z_{ij} x_i\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_{1}^{3}\left(\hat{\alpha}_t x_i^t\right) + \hat{\beta}_1 z_{ij} + \hat{\delta}_1 z_{ij} x_i\right)} \qquad \textbf{19}$$

and $\hat{\alpha}_t$, $\hat{\beta}_1$, $\hat{\delta}_1$ are calculated from the maximum likelihood regression model below:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{1}^{3}\left(\alpha_t x_i^t\right) + \beta_1 z_{ij} + \delta_1 z_{ij} x_i \qquad \textbf{20}$$

The individual level synthetic regression model makes an assumption that is not necessary for relational models (or the curve fitting approach). The model assumes that the covariates in the HSE and the census are measured identically. This turns out not to be the case. Figure 32 demonstrates that census rates of LLTI tend to be lower that the HSE 95% confidence intervals for LLTI under the age of 60 and almost above them at the oldest ages. The focus on health in the HSE is thought to lead to higher estimates of LLTI rates compared to the census (Bajekal, Harries et al. 2003).

Census LLTI schedules were adjusted using differentials between HSE and census LLTI prevalence rates below the age of 60. For more details see Marshall (2009) (p169). Figure 32 shows the unadjusted census LLTI rates along with the HSE rates whilst figure 33 shows the adjusted LLTI rates. The same adjustment was applied to the LLTI schedules in each of the districts in this practical.

Figure 31: Age specific rates of LLTI in the census (2001) and the Health Survey for England (2001)

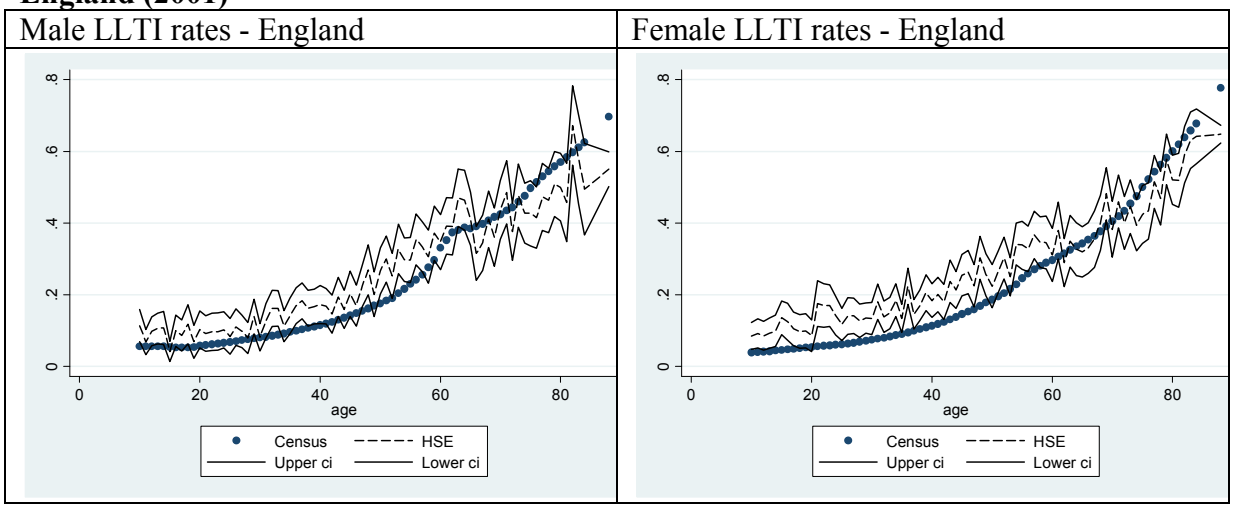
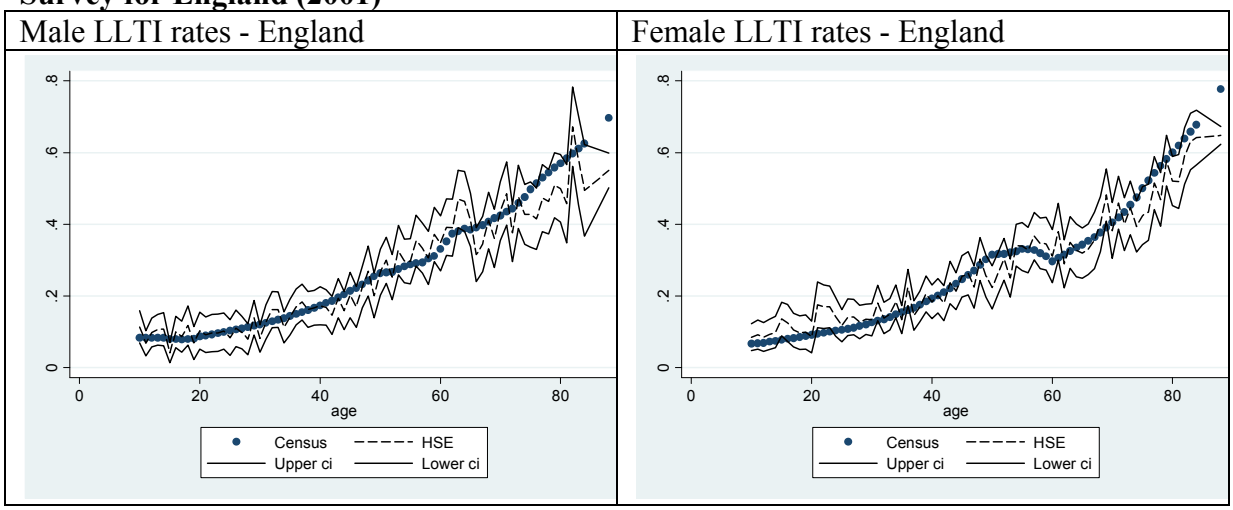


Figure 32: Age specific rates of LLTI in the census (2001 - adjusted) and the Health Survey for England (2001)



The non-linear nature of the logit transformation of the dependent variable in a logistic regression model requires that all combinations of the explanatory variables must exist as a cross-tabulation in a census table for each district (Bajekal, Scholes et al. 2004). Variables of age, LLTI and sex, which are known to be correlated with mobility disability (and whose distribution varies across districts) were chosen on this basis. It is possible to develop more detailed census cross-tabulations, allowing inclusion of more explanatory variables, by combining data from the census and Sample of Anonymised Records (Simpson and Tranmer 2005; Charlton 1998) or through the use of the Controlled Access Micro data Samples. These options are not pursued here.

Practical 3 - Task 1: Generating model probabilities from a logistic regression model

There are two stages to this first task. First, a logistic regression model (equation 16) is fitted to data from the HSE00/01 to predict the probability of an individual having a mobility disability based upon their age, sex and whether or not they have an LLTI. The explanatory variables include age (single year) sex (male or female), age squared and age cubed (in recognition that the rise in mobility disability rates with age might not be linear), LLTI (has an LLTI or not) and an interaction between age and LLTI (reflecting the possibility that the impact of having an LLTI on the probability of having a mobility disability might vary according to age).

Second, model probabilities are generated for each single year of age for males and females and for the LLTI population and non-LLTI population. 304 probabilities are generated: 2 sex groups* 2 llti groups* 76 age groups=304).

3.1.1: First open the HSE practical data

```
clear



use "C:\ESDS SAE practical\Data\HSE data.dta", clear
```

3.1.2: Drop missing values of LLTI (the LLTI question was not asked in proxy interviews in 2000)

```
drop if llti==.
```

3.1.3: Now fit the logistic regression model for males and generate predicted probabilities of mobility disability for each of the males in the HSE

```
xi: logit mobility i.llti*age agesq agecub if sex==1
[pweight=weight]

predict MO_RT_M if sex==1
```

3.1.4: Now fit the logistic regression model for females and generate predicted probabilities of mobility disability for each of the females in the HSE

```
xi: logit mobility i.llti*age agesq agecub if sex==2
[pweight=weight]

predict MO_RT_F if sex==2
```

3.1.5: The following syntax creates a single variable (LM_RT) containing the predicted probabilities for males and females

```
gen MO_RT=MO_RT_M if sex==1
replace MO_RT=MO_RT_F if sex==2
```

The model we fitted for males and female (see syntax lines 3.1.3 and 3.1.4) predicted the log odds of having a mobility disability on the basis of the following explanatory variables: age, age^2, age^3, LLTI and LLTI*age (interaction between age and LLTI). From this model we calculated model probability of having a mobility disability for each possible combination of the age and LLTI variables (i.e. has LLTI – 10,11,12,….84,88 and No LLTI – 10,11,12,….,84,88). If we drop duplicate values in terms of age and sex we will be left with a single row corresponding to each of the probabilities generated by the model.

3.1.6: Now drop the duplicates in terms of sex age and LLTI leaving a probability of mobility disability for each age/sex/llti combination

```
duplicates drop sex age llti, force
```

3.1.7: Sort the data and keep only the required variables

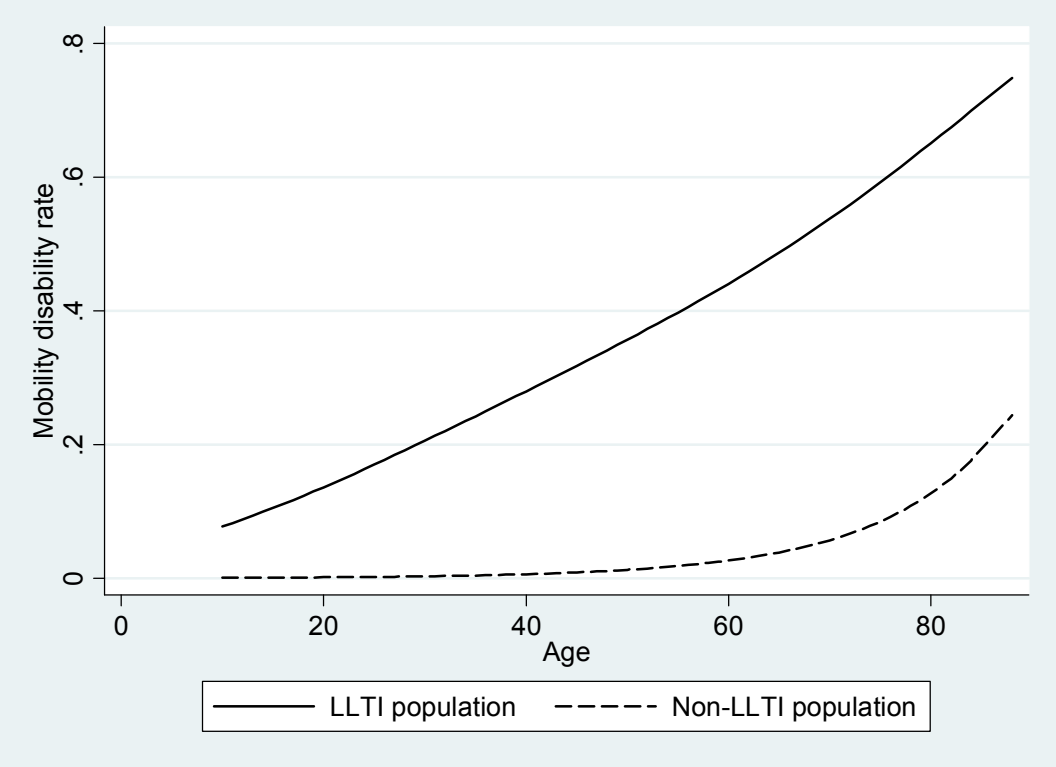
```
sort llti sex age

keep sex age llti MO_RT

browse
```

3.1.8: Produce a graph of model mobility disability rates for males in the LLTI and non-LLTI populations

```
twoway (line MO_RT age if llti==1, lcolor(black) lpattern(solid))
(line MO_RT age if llti==0, lcolor(black) lpattern(dash)) if sex==1,
ytitle(Mobility disability rate) xtitle(Age) legend(order(1 "LLTI
population" 2 "Non-LLTI population"))
```

Running the syntax in line 3.1.8 should generate the graph in figure 34. As would be expected the population with an LLTI are always more likely to have a mobility disability than the non-LLTI population. The probability of having a mobility disability increases with age for the both the LLTI and non-LLTI populations.

Figure 33: Mobility disability model schedules for LLTI and non -LLTI population



3.1.9: Sort then save your data

```
sort sex llti age

save "C:\ESDS SAE practical\Saved practical work\Practical 3 - task
1.dta"
```

Practical 3 - Task 2: Generating district estimates of the population with a mobility disability

3.2.1 First, open the population data file for task 3 and then browse its structure

```
use "C:\ESDS SAE practical\Data\Population data - practical 3.dta",
clear

browse
```

You should notice from the data browser that the population data for practical 3 contains a population count for each district distinguishing each combination of age sex and LLTI. Through the multiplication of these population counts by the model probabilities we developed in the previous task we can derive estimates of population counts with a mobility disability.

3.2.2 The syntax below merges the population data with the model probabilities saved at the end of task 1:

```
sort sex llti age

merge sex llti age using "C:\ESDS SAE practical\Data\Practical 3 -
task 1.dta"
```

Browse the data file to confirm the merge has gone to plan.

3.2.3: Now multiply the model probabilities by the census population counts:

```
gen MO_POP= MO_RT * pop
```

3.2.4: Now calculate total counts of mobility disability for each of the districts in 2001:

```
sort zonecode
by zonecode: egen Mob_Tot= sum(MO_POP)
```

3.2.5: Generate counts of total population for each of the districts in 2001

```
by zonecode: egen Pop_Tot= sum(pop)
```

3.2.6: Drop the duplicate values of age, sex and LLTI so that we are left with a count of people with a mobility disability in each district. Keep required variables only

```
duplicates drop zonecode, force
```

```
keep zonecode zonename Mob_Tot Pop_Tot
```

3.2.7: Calculate the % of people with a mobility disability in each district

```
gen prev=(Mob_Tot/Pop_Tot)*100
```

Browse the data editor and confirm the results for the individual level synthetic regression model in figure 35.

Figure 34: Mobility disability estimates: Relational model and Individual level synthetic regression model

District	2001 Estimate - relational models		2001 Estimates using individual level synthetic regression model	
	Number	%	Number	%
Barnet	26,455	9.5	27,995	10.0
Bury	20,391	13.0	19,276	12.2
Wakefield	44,397	16.0	38,626	14.0
South Bucks	4,612	8.5	5,415	10.0
Easington	19,670	23.9	14,937	18.1
Stroud	9,931	10.4	10,781	11.3

Try repeating tasks 1 and 2 of practical 3 for hearing disability

Case study discussion

In the previous case studies we used the following four techniques to develop schedules of mobility disability rates which were then used to create district estimates of the population with a mobility disability:

1. Curve fitting using England mobility rates
2. Curve fitting using regional mobility rates
3. Relational models
4. Individual level synthetic regression

The percentages of people with a mobility disability in each of the six districts under each of the 4 models are shown in figure 36. The variability in the percentages estimated using England rates stems purely from the differences in age structure across the six districts. Stroud has the most elderly population and so it has the highest percentage of mobility disability, whilst Barnet has the lowest estimate because it has the youngest population structure. Estimates derived in this way are useful to give an indication of the size of the population with a disability that might result from an area's age structure. The POPPI (http://www.poppi.org.uk/) and PANSI (http://www.pansi.org.uk/) website allow users to generate estimates of disability and other health conditions in this way.

A key weakness of the estimates derived from England rates is that it is well known that levels of poor health and disability vary across the UK after accounting for population age structure. The estimates using regional rates take this into account to some extent. South Bucks and Stroud have considerably lower estimated percentages with a mobility disability (compared to estimates using England rates) reflecting the lower level of mobility disability in the South East and South West regions that we observe in the Health Survey for England.

The weakness of using regional rates of mobility disability is that it makes the unrealistic assumption that all districts within a region experience exactly the same level of mobility disability. The relational and individual level synthetic regression models attempt to model this variability using the level of LLTI (census) within a district as a proxy for level of mobility disability. Both models predict higher percentages of mobility disability in Easington and Wakefield than are estimated using England and regional rates. This result

makes intuitive sense as these districts are known to have particularly poor health as can be seen by their high Standardised Illness Ratios (SIRs[2]). Conversely, in the healthy districts (with SIRs below 1 - South Bucks and Stroud) the estimates of the percentage of people with a mobility disability are lower that the corresponding estimate from regional and England rates.

The differences between the estimates from relational and synthetic regression models occur most prominently in districts with high or low SIRs. In districts with a high SIR the estimates of the population with a mobility disability are higher than for the synthetic regression model. Similarly, where a district has a low SIR the estimate of the mobility disability prevalence rate is higher under a relational model compared to a synthetic regression model.

Figure 35: Model percentage of people with a mobility disability in six English districts under 4 different modelling approaches

	Model % of population with a mobility disability				SIR[2]
	England rates (%)	Regional rates	Relational	Synthetic regression	
Barnet	10.4	10.6	9.5	10.0	0.83
Bury	10.7	12.9	12.9	12.2	1.05
Wakefield	10.8	11.8	15.9	14.0	1.22
South Bucks	12.0	8.8	8.5	10.0	0.66
Easington	11.1	14.4	23.6	18.1	1.63
Stroud	12.1	9.4	10.4	11.3	0.79

It is difficult to choose between the relational and synthetic regression models largely because we have no district estimates of mobility disability with which to make a comparison. Marshall (2009) (see chapter 7) compares the performance of relational and individual level synthetic regression models for the estimation of several types of disability in the HSE for regions (where data does exist for comparison). Although a more complex relational model is needed for some disability types there is no evidence to separate the different approaches on the basis of the regional data.

[2] SIRs are calculated for each district in 2001 using the direct standardisation procedure. See Newall (1988) for more details. All SIRs are relative to the UK in 2001, a value over 1 indicates higher levels of LLTI in a district relative to the UK (2001) whilst an SIR below 1 indicates lower levels of LLTI relative to the UK (2001).

Given the similarity of the models in terms of closeness of their fit to the observed data, a key additional requirement of the selected model is parsimony in terms of parameters, assumptions and data requirements. A parsimonious model is less likely to be affected by error resulting from the influence of missing data, false assumptions or inaccurate estimation of parameters.

The individual level synthetic regression model relies on an assumption that census and HSE measures of explanatory variables are identical. It has already been shown that this equivalence is not exact with LLTI rates from the census falling outside HSE confidence intervals. The adjustment of districts census rates to match HSE estimates is based on a comparison of England as a whole and assumes that the same relationships are found at sub-national areas. The relational approach does not rely on the assumption of equivalence of census and HSE measures and so is not prone to the error that might result from any deviation from this assumption.

Relational models offer a more parsimonious approach than the individual synthetic regression model in terms of parameters minimising the error that might arise during this process. The relational models fitted here uses four parameters (2 for males and 2 for females) compared to 12 parameters in the synthetic regression model.

The synthetic regression model requires information on whether HSE respondents have an LLTI which is not needed in the relational models. 5% of cases (1,288) in the combined HSE 2000/2001 sample have missing values for the LLTI variable. This occurs predominantly in the care home sample in situations where a proxy interview was undertaken (the LLTI question not being included in proxy interviews in 2000). Missing data is an issue because it can lead to model error as those who are excluded tend to have different characteristics to the participants (Henry 1990). One approach around this would be to assume the presence of an LLTI for all those in care homes who had information entered by a proxy. However this represents another assumption that is not required in a relational modelling approach.

The methodological advantages of relational models discussed above certainly point in favour of using relational models in this instance if one model were to be picked. However, for the estimation of other characteristics that might not be strongly related to age the synthetic regression model is more appropriate. It is important to remember that estimates

from all the other models developed here are very helpful in building confidence in the final selected approach. Taking an average of the estimates from relational and synthetic regression models is another approach that could be followed. It is always helpful to discuss local estimates with local experts and policy makers to assess whether they are reasonable.

Model extensions

Research suggests that the simple Brass relational model is not always flexible enough for estimation of disability where the disability age pattern deviates from the LLTI schedule. In these situations it is possible to generate a better set of estimates using a more complex relational model based on that developed by Ewbank *et al* (1983). For more details see Marshall (2009).

The relational and individual synthetic regression models assume that national relationships apply for districts. For example, the relational models fitted here assume that the adjustments made to LLTI curves in order to derive a mobility disability schedule are the same in each district as for England. This may not be the case; there could be area specific factors that might result in a different local relationship between LLTI and mobility disability compared to that at play nationally. In such cases two districts may have exactly the same population composition (in terms of the population characteristics included in our models) but the levels of mobility disability may vary. An interesting piece of work that illustrates the importance of area specific or contextual factors in determining the levels of health in an area is the comparison of premature mortality in Middlesbrough and Sunderland (Phillimore and Morris, 1995). This study shows that despite having very similar socio-economic characteristics, levels of premature mortality were consistently and markedly higher in Middlesbrough in the early 1980s.

One way to get around this issue is to include area covariates in the small area estimation models. Twigg, Moon *et al.* (2000) employ such an approach to estimate levels of smoking and drinking using data from the Health Survey for England and the Census (see section 4.2.4). The Health survey for England includes the ONS area classification (as well as Acorn classifications) that categorises the district a person according to 6 area types of Inner London, Mining and Industrial, Urban, Mature, Prosperous and Rural. Research shows that a surprisingly large proportion of the variability in district levels of LLTI can be explained by the use of such area classifications (see Marshall (2009) – p260-262). Inclusion of area

classification variables in each of the models fitted here (curve fitting, relational models, synthetic regression) offers a way to improve estimates by accounting for contextual factors that might influence levels of mobility disability.

A weakness of the individual level synthetic regression model as fitted in this practical is that the number of explanatory variables is somewhat restricted. This is because the method requires a local population crosstabulation for all the explanatory variables that are included in the model. In practical 3, LLTI age and sex were included and a census table including a crosstabulation of population counts distinguishing age, sex and LLTI status was used to derive estimates of mobility disability for each district. Adding further explanatory variables such as occupation, tenure or marital status might our local estimates but as there isn't a census crosstabulation of age, sex, LLTI and an additional variable it isn't immediately possible. Charlton (1998) and Simpson (2005) describe approaches to overcome this problem by combining census microdata for the UK with survey data to generate more detailed cross tabulations of census variables than exists in the available census output.

6. References

Bajekal, M., Harries, T., Breman, R. and Woodfield, K. (2003). Review of disability estimates and definitions. London, HMSO.

Bajekal, M., Scholes, S., Pickering, K. and Purdon, S. (2004). Synthetic estimation of healthy lifestyle indicators: Stage 1 report. NatCen, London

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B. and Rossiter, D. (2005). "SimBritain: A Spatial Microsimulation Approach to Population Dynamics." Population, Space and Place **11**(1): 13-34.

Ballas, D., Clarke, G., Rigby, J. and Wheeler, B. (2006). "Using geographical information systems and spatial microsimulation for the analysis of health inequalities." Health Informatics Journal **12**(1): 65-79.

Bellaby, P. (2006). "Can they carry on working? Later retirement, health and social inequality in an aging population." International Journal of Health Services **36**(1): 1-23.

Brass, W. (1971). Biological aspects of demography. On the scale of mortality. Brass, W. London, Taylor Francis: 69-110.

Bruton-Smith, I. (2008). The use of the ACORN classifications and the Index of Deprivation in British Crime Survey analysis. Geodemographics in action seminar, Royal Statistical Society, London. Available at: http://areaclassification.org.uk/2008/09/29/geodemographics-in-action-seminar-15-september/

CACI. (2009). "New ACORN classification map." Retrieved 2/10/09, from http://www.caci.co.uk/acorn/acornmap.asp

Charlton, J. (1998). "Use of the Census Samples of Anonymised Records (SARs) and survey data in combination to obtain estimates at local authority level." Environment and Planning A **30**(5): 775-784.

Coale, A. and Trussell, J. (1996). "The development and use of demographic models." Population Studies **50**(3): 469-84.

Congdon, P. (1993). "Statistical graduation in local demographic analysis and projection." Journal of the Royal Statistical Society. Series A. Statistics in society **156**(2): 237-70.

Dorling, D., Rossiter, D., Thomas, B. and Clarke, G. (2005). Geography matters: simulating the local impacts of national social policies. York, Joseph Rowntree Foundation.

Duke-Williams, O. (2008). Using OAC for analysis of the 2001 Census interaction data including migration Geodemographics in action: seminar, Royal Statistical Society, London. Available at: http://areaclassification.org.uk/2008/09/29/geodemographics-in-action-seminar-15-september/

Ewbank, D. C., Gomez de Leon, J. and Stoto, M. (1983). "A reducible four-parameter system of model life tables." Population Studies **37**(1): 105-127.

Experian. (2009). "Mosaic UK: The consumer classification for the UK." Retrieved 2/10/09, from www.experian.co.uk/www/pages/what_we_offer/products/mosaic_uk.html.

Heady, P. and Clarke, P. (2003). Model-based small area estimation series No 2. Small area estimation project report. London, Office for National Statistics.

Henry, G. (1990). Practical Sampling. London, Sage.

Keyfitz, N. (1982). "Choice of function for mortality analysis: Effective forecasting depends on a minimum parameter representation." Theoretical Population Biology **21**(3): 329-352.

Langford, I. (1994). "Using Empirical Bayes estimates in the geographical analysis of risk." Area **26**(2): 142-149.

Marshall, A. (2009) Developing a methodology for the estimation and projection of limiting long term illness and disability, PhD Thesis, School of Social Sciences, University of Manchester. Available at http://www.ccsr.ac.uk/staff/am.htm

Michaud, J. (1996). Projections of persons with disabilities (limited at work/perception), Canada, Provinces and Territories, 1993-2016. S. Canada, Minister of Industry.

Newall, C. (1988). Methods and Models in Demography. New York, The Guildford Press.

Norman, P. (2003). "Achieving data compatibility over space and time: creating consistent geographical zones." International journal of population geography 9(5): 365-386.

Openshaw, S. (1995). "Geodemographic segmentation systems for screening health data." Journal of Epidemiological Community Health **49**(2): 34-38.

Orange, S. (2008). Using OAC as a tool to help inform health insight planning in Yorkshire and the Humber Geodemographics in action: seminar, Royal Statistical Society, London. Available at: http://areaclassification.org.uk/2008/09/29/geodemographics-in-action-seminar-15-september/

Phillimore, P. and Morris, D. (1991) Discrepant legacies: premature mortality in two industrial towns. Social Science and Medicine, 33(2):139-152.

Powell, G. (2008). Realising the Power of OAC using ONS' Social Surveys. Geodemographics in action: seminar, Royal Statistical Society, London. Available at: http://areaclassification.org.uk/2008/09/29/geodemographics-in-action-seminar-15-september/

Preston, S., Heuveline, P. and Guillot, M. (2001). Demography: Mesuring and modelling population processes. Oxford, Blackwell.

Rao, J. (2003). Small area estimation. New Jersey, John Wiley & Sons Inc.

Siegel, J. (2002). Applied Demography. London, Academic Press.

Simpson, L. and M. Tranmer (2005). "Combining sample and census data in small area estimates: iterative proportional fitting with standard software." The Professional geographer **57**(2): 222-234.

Singleton, A., Davidson-Burnett, G. and Longley, P. (2007). University Market Area Analysis for Widening Participation. Centre for Education in the Built Environment Working Paper Series. London

Skinner, C. (1993). The Use of Synthetic Estimation Techniques to Produce Small Area Estimates. New Methodology Series NM18. OPCS. London.

Twigg, L., Moon, G. and Jones, K. (2000). "Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators." Social Science & Medicine **50**(7-8): 1109-1120.

Vickers, D. (2006). Multi-Level Integrated Classifications Based on the 2001 Census. School of Geography. Leeds, The University of Leeds. Available at: http://www.geog.leeds.ac.uk/people/old/d.vickers/thesis.html

Voas, D. and Williamson, P. (2000). "An evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata." International Journal of Population Geography **6**(5): 349-366.

Williamson, P. (1996). Community care policies for the elderly, 1981 and 1991: a microsimulation approach. Microsimulation for Urban and Regional Policy Analysis. G. Clarke. (Ed). London, Pion**:** 64-87.

Williamson, P. (2002). Synthetic Microdata. The Census Data System. P. Rees, D. Martin and P. Williamson. Chicester, John Wiley**:** 231-242.

Williamson, P., Birkin, M. and Rees, P. (1998). "The estimation of population microdata by using data from small area statistics and samples of anonymised records." Environment and Planning A **30**(5): 785-816.