**Instructions for Exercise 7: Health by Social Variables**

**Revised September 2020**

**Preliminary Notes**

* The training exercise materials are available here: <https://www.york.ac.uk/che/research/equity/handbook/>
* There are two relevant spreadsheets: “Ex 7 – student”, and “Ex 7 – solution”. Please open the student version and go through the worksheets filling in the relevant cells yourself. You can look at the solution file if you get stuck.
* When the spreadsheet is first opened a ‘Security Warning’ may be displayed below the menu bar.  Select ‘Enable this content’.
* This exercise was produced for the Handbook of Distributional Cost-Effectiveness Analysis by James Love-Koh and edited by Richard Cookson, with help from James Lomas.

**Introduction**

The objective of this exercise to construct an *ex ante* baseline distribution of health-adjusted life expectancy (HALE) for the English population. The distribution is made up of ten subgroups – five socioeconomic and two geographical groups – and is parameterised using real-world data.

The principal stages of the exercises are:

* Construct life tables for each subgroup to predict HALE at birth, using Sullivan’s method to adjust life expectancy estimates;
* Arrange distributions of HALE and calculate some basic inequality measures.

## Step-by-step guide

Open the file ‘*Ex 7 - student.xlsx’* and select the <Inputs> worksheet. The cells that require completing in this exercise are shaded yellow. All other cells are complete and ready for use in estimating the baseline distribution of health.

The sheet contains all the necessary inputs for the exercise: health-related quality of life (HRQL) weights, mortality rates and population estimates. Mortality rates and population estimates are taken from external sources. In order to populate the life tables, each input is specified by age as well as region and socioeconomic status.

### Health-related quality of life weights

Take some time to familiarise yourself with how the HRQL weights have been calculated within the spreadsheet. A regression analysis was conducted on health survey data to generate predicted values of health-related quality of life. The predicted scores are obtained through the following equation:

$$\hat{HRQL}=α+\sum\_{i=1}^{14}β\_{i}age\_{i}+\sum\_{j=1}^{4}γ\_{j}SES\_{j}+δNorth $$

where $α$ is the constant term, $β\_{i}$, $γ\_{j}$ and $δ$ are the coefficients for a series of age, socioeconomic and region dummy variables. These take on either a value of 0 or 1 depending on subgroup membership. When all the variables are set to zero, the predicted weight is the constant term only and represents the ‘reference’ group: individuals aged 15 to 19, residing in the south and in the lowest socioeconomic group (SES1).

The cells containing the constant term and the age-group, socioeconomic and region coefficients are found on the <Regression> worksheet and have been named. These can be verified in the Excel ‘Name Manager’ (in the formula tab), and the names are used in formulae instead of the cell reference.

### Constructing life tables

Select the <Life Tables> worksheet. This worksheet contains the life tables for each socioeconomic group for England and the North and South of England separately. They are calculated from data and estimates from the <Inputs> worksheet. All of them have been completed except for one. Our objective is to complete the first life table on the sheet for the lowest socioeconomic group for the south of England.

The first two columns of the life table are provided: $n\_{x}$ is the number of years in each interval and $a\_{x}$ is the proportion of the interval survived by those who die, set at a default value of 0.5. Begin by importing the relevant mortality rates $M\_{x}$ (column E) and HRQL weights $u\_{x}$ (column K) from the <Inputs> worksheet. The remaining columns perform a set of sequential mathematical operations to generate HALE:

1. The probability of dying $q\_{x}$ differs from the mortality rate, as it adjusts the mid-year population to account for the fact that those who die live only half a year. The correction equates to: $q\_{x}=\frac{n\_{x}M\_{x}}{1+(1-a\_{x})n\_{x}M\_{x}}$

The probability of dying in the last interval is always equal to 1.

1. The surviving cohort $I\_{x}$is the number alive in the previous interval multiplied by 1 minus the probability of dying during the previous interval. The cohort size is set in the first interval (cell G21); in this exercise we set it to 1.
2. The number of life years lived in the interval $L\_{x}$ reflects the years lived by both those who survive and those who die. Those who survive live $n\_{x}$ years, while those that die live for only $a\_{x}n\_{x}$ years.

*Hint: The number dying in the interval can be calculated as the difference between those alive in the current and subsequent interval. In the last interval all individuals die, meaning that the value for* $L\_{x}$ *is given by the formula* $n\_{x}a\_{x}I\_{x}$*.*

1. $T\_{x}$ is the sum of the life years lived in the current interval and all future intervals. In the last interval, this reduces to $L\_{x}$.
2. Life expectancy $e\_{x}$for each interval the average future number of life years experienced by the surviving cohort: $\frac{T\_{x}}{I\_{x}}$

Take a moment to review the past 5 steps and how your calculations relate to the terms defined in cells B5:C17.

Next, these life expectancy estimates are adjusted for HRQL using Sullivan’s method. The number of health-adjusted life years in each interval $Y\_{x}$ is obtained by multiplying $L\_{x}$ by the respective HRQL weight, $u\_{x}$. Then, repeat steps iv and v using $Y\_{x}$ in place of $L\_{x}$ to derive estimates of HALE.

### Ordering distributions of health

Open the worksheet <Dist SES N&S>. The main table (cells A8:K26) amasses all of the HALE estimates from the life tables for the five socioeconomic groups of the South and North of England. Underneath this table (in rows 28 to 104), ordered distributions and their associated inequalities are summarised again using the slope and relative indices of inequality, as well as the Atkinson and Kolm indices of health-related social welfare. A similar arrangement is found in the worksheet <Dist SES Only>, which looks at socioeconomic inequality only.

We will now create a distribution of health for the ten regional and socioeconomic subgroups in the <Dist SES N&S> worksheet. Our three tasks here are to rank the subgroups in terms of their health, use the rankings to create an ordered multivariate distribution and then calculate some inequality measures.

Firstly, note that the population statistics are already completed using the estimates from the <Inputs> worksheet. The respective population fraction of each group is provided in cells D34:D43.

1. Rank the subgroups in terms of their HALE at birth in cells B34:B43. The RANK function can be invoked here on cells A34:A43 (the help file can be viewed to understand the function arguments). As the lowest HALE value should return a rank of 1, the ‘order’ argument in the function should be set to 1.
2. We can use these ranks to create an ordered distribution of HALE in cells F47:F56. For this we need to familiarise ourselves the VLOOKUP function, which can be used to lookup a particular rank value (from cells A47:A56), match it to the same rank in the unordered table and automatically return the value of a separate column (for example, the population or HALE value). Again, the Excel help file for the VLOOKUP function can be read for a full description.
3. Next complete the fractional ranks, which will be used to calculate some inequality indices. The fractional rank is the cumulative mid-point of each group within the population. Calculate the fractional rank for each group.

*Hint: For group 1, this is simply half the population fraction. For group 2, it is the fraction of group 1 plus half the fraction of group 2.*

*Summarising health inequality*

The multivariate distribution of HALE across both region and socioeconomic groups is now complete and ready for analysis. We can summarise the inequality using four different metrics:

1. The absolute gap is the difference between the top and bottom groups: $G\_{A}=H\_{T}-H\_{B}$
2. The relative gap is the ratio of these minus one: $G\_{R}=\frac{H\_{T}}{H\_{B}}-1$
3. The slope index of inequality is the coefficient of a simple (one variable) regression of HALE ($H$) on fractional rank ($r$). The formula for this coefficient can be derived mathematically as: $SII=\frac{Cov(H,r)}{Var(r)}$. The relevant Excel functions are COVARIANCE.P and VAR.P.
4. The relative index of inequality transforms the slope index on to a relative scale by dividing through by mean HALE: $RII=\frac{SII}{\overline{H}}$

*Hint: Mean HALE is calculated as a weighted average of HALE according to the size of the subgroup. The SUMPRODUCT function can simplify this calculation considerably by multiplying the respective values of two equally sized arrays.*

Below these inequality measures there are tables calculating Atkinson and Kolm indices, which estimate health-related social welfare. These are a type of measure that evaluates both changes to average health and changes to the distribution of health simultaneously. These are the subject of Chapter 11 and are included here for completeness. Note that the ‘equally distributed equivalent’ lifetime health (in cells B83 and B102) is lower than average HALE – this is because the measures reflect social preferences (captured by the inequality aversion parameters) that inequalities are bad.

The tables for the bivariate distribution (worksheet <Dist SES Only> cells A13:E19), i.e. the distribution of HALE across socioeconomic groups, have also been completed. Take a moment to look over them and validate the calculations you have just made.

Summary graphs of the distributions we have calculated are presented in the <Summary> worksheet. The bar graph used for Distribution 2a, which plots multivariate inequality, does not accurately reflect the unequal size of the subgroups in the population. Since creating bar plots with variable bin width is troublesome in Excel, HALE predictions are presented over a series of 1,000 values using the subgroup population size. These produce Distribution 2b.