Guidelines for running STATA codes to correct smoking prevalence for differential mortality

General information

This document describes the set of codes that we have written to adjust data on cohort-specific smoking prevalence over the life-course (estimated from retrospectively reported information) for differential mortality rates of smokers and non smokers. The codes apply the algorithm described in Christopoulou et al. (2011).

To describe the code, we use data from Australia. The input and output files will differ for each country. We reflect that fact in the file names we assign to both input and output files. In particular, the first two letters of every input or output file begins with the first two letters of the country's name. Here the prefix AU_ that appears at the beginning of each file name indicates that these programs use or produce data from Australia.

For each code-file, this document lists and describes the input data you will need (*Input*), the source of those data, and how you will need to organize the input data files. We also list and describe the set of output files that the code should produce (*Output*). Finally, we include some information specific to the code (*Notes*) and some tips that we hope will make it easier to adapt the code for each country (*Country-specific changes*).

Within the code-files we have embedded detailed comments to explain what the code does in each step. Please read the comments carefully because they also identify places in the code where the user needs to modify code in ways specific to a particular country. In most files, users only need to change lines that appear at the beginning of the code - the rest should run automatically. Again, please read these comments carefully.

We accompany the codes with sample input and output files (each file is limited to 10 observations). Hopefully these examples will provide users with a clear idea of how to format country-specific input files (e.g. long vs. wide shape of database, variable names, variable content etc.), and the form that output files should take.

Prefix A, B, C etc. in the names of the code files indicate the order in which one should run the codes.

Structure of codes:

1. **A_smok.do**

   Input: AU_retrosp.dta.
   Create this file with the retrospective smoking data from your country of interest. The file should contain the following variables: year of survey (this variable may take different values if you pool together data collected in different years), age at the survey year, sex, age started smoking, age quit smoking, unique person identifier, and sample or population weights. The file should have one and only one observation per surveyed individual.

   Output: AU_smoking.dta.
   The output file measures each person's smoking-status over the life-course.

   Country-specific changes: Change only the three first lines of code to:
   (i) specify the pathname for the input and output directories;
   (ii) specify the two letters representing the country name as local ***cn***.

2. **B_mort.do**

   Input: morticd7.dta; morticd8.dta; morticd9.dta; morticd10.dta.
   These files contain raw WHO mortality data that you can download at the following website: http://www.who.int/whosis/mort/download/en/index.html.

   Output: AU_mort.dta.
   The program produces harmonized mortality data by smoking-related cause, year, sex and age.

   Notes*:* Be very careful to check for country specific changes you might need to make in the code. In particular, check the WHO documentation that accompanies each country's mortality data for country specific differences. For example, the variable *list* may differ across countries (see documentation that accompanies the data in the WHO website). That documentation will help you identify where you need to modify the program in ways specific to your country.

   Note also that the WHO mortality data cover different time periods for different countries. For example, the WHO mortality data cover 1950-2004 for Canada; 1987-2000 for China; 1951-2005 for Spain; and 1978, 1979, 1981-1984, 1987 for Turkey.

   The code in B_mort.do is useful even when there are no WHO data for the country of interest **IF** you can get cause-specific mortality data from other (national) sources. That is what we did for the US for the years 1933-1949. We got those data from the US Vital Statistics. The WHO did not include for those years. Of course, one has to adjust the code accordingly.

When cause-specific mortality data are not available from the WHO for all years one needs (the WHO rarely includes data before 1950) and when no cause-specific mortality data are available from national sources, one can substitute data on overall mortality rates for the missing years. In this case (for these years) one has to input these data as described below in section 5.

Country-specific changes**:* In the WHO data there are two key variables, named ***list and ***cause*** respectively. The variable ***list*** identifies the ICD classification version and the variable ***cause*** identifies the respective causes of death. The variable ***list*** for Australia takes values: 07A, 08A, 09B, and 104. So, in B_mort.do we identify smoking-related causes of death by the ***cause*** values that correspond to ICD classifications 07A, 08A, 09B, and 104. The same ICD categories are available for other countries, such as Canada and Spain. Therefore, for these countries change only the first three lines of code to:

(i) specify the directory path;

(ii) specify the country-id in the WHO data as local *c* (for China, Canada, Spain, and Turkey, the relevant country-ids are listed in the code-file. For any other country you should refer to the WHO documentation);

(iii) specify two letters representing the country name as local *cn*.

Other countries have different data availability. For example, for Turkey the variable ***list*** takes values 08A and 08B. For China it takes value 09C. In these cases, the codes need to be adjusted further (e.g in command: *bys year sex age: egen pharynx7=total(deaths) if cause=="A044" & list=="07A"* change the number in the ***cause*** and ***list*** variables, as appropriate. To find these numbers, see the relevant tables in Part 2 of the file named *WHOdocumentation.doc*).

3. C_pop.do

Input: pop.dta.
This file contains raw population data by age, year, sex for the respective country. You can download the data at: http://www.who.int/whosis/mort/download/en/index.html.

Output: AUpop.dta; AUpop_bh.dta
These files will contain data as follows:
AU_pop.dta - Population data by 5-year age-category, year, sex
AU_pop_bh.dta - Population data by birth cohort, year, sex

Notes: These population data are needed in combination with the WHO mortality data in order to calculate death rates.
***Country-specific changes**:* Change only the first four lines of code to:

(i) specify the directory path;

(ii) specify the survey year as local *s* (if you combine together surveys from different years, specify latest survey year);
(iii) specify two letters representing the country name as local *cn;*
(iv) specify the country-id in the WHO data as local *c.*

4. D_sam.do

Input: CPS.dta; AU_mort.dta; AU_pop.dta
These data files contain the following information:
CPS.dta - data from the Cancer Prevention Study II. We include the CPS data in the homonymous folder that accompanies the codes;
AU_mort.dta - Harmonized mortality data by smoking-related cause, year, sex and age (produced by B_mort.do);
AU_pop.dta - Population data by 5-year age-category, year, sex (produced by C_pop.do).

Output: AUsam.dta
These data are estimates of smoking-attributable deaths by year, sex and birth cohort, calculated using the Peto et al. (1992) procedure.

Country-specific changes*:* Change only the first five lines of code to:
(i) specify the directory path for the CPS data,
(ii) specify pathnames for the other input files (may be same as (i));
(iii) specify the survey year as local *s* (if you combine together surveys from different years, specify latest survey year);
(iv) specify two letters representing the country name as local *cn;*
(v) specify the earliest year with available WHO cause-specific mortality data as local *m*.

5. E_adj.do

Input: AU_smoking.dta; AU_sam.dta; AU_pop_bh.dta; AU_mort_tot.dta
The above files contain data as follows:
AU_smoking.dta - Smoking prevalence by year, sex and birth cohort (produced by A_smok.do);
AU_sam.dta - Smoking-attributable deaths by year, sex and birth cohort, calculated using the Peto et al. procedure (produced by D_sam.do);
AU_pop_bh.dta - Population data by year, sex and birth cohort (produced by C_pop.do);
AU_mort_tot.dta - Overall (not cause-specific) death rates by cohort, sex, and year (to use when cause-specific mortality data are not available). Potential source is the Human Mortality Database (HMD): http://www.mortality.org/

Output: AU.dta
This file contains the adjusted smoking prevalence rates that have been corrected for differential mortality by the Harris (1983) formula with Peto et al. (1992) mortality inputs, and test-results of statistical significance of differential mortality bias, as described in Christopoulou et al. (2011).

Notes: Overall mortality rates are available in the HMD for specific countries and periods, e.g. Canada (1921-2006) and Spain (1908-2006). For other countries, such as Turkey and China, one should look for national sources.

Country-specific changes*:* Change the first five lines of code to:

(i) specify the directory path;
(ii) specify the survey year as local *s* (if you combine together surveys from different years, specify latest survey year);
(iii) specify two letters representing the country name as local *cn;*
(iv) specify the earliest year with available total mortality data as local *mt*;
(v) specify how stata should treat the weights that accompany the retrospective smoking data. (For info on weights type 'help weights' in STATA.)

6. F_figures.do

Input: AU.dta
These data measure smoking prevalence corrected for differential mortality and results from the tests of statistical significance (produced by E_adj.do).

Output: Corr_uncorr_AU.emf; Corr_chi2_AU.emf
This program produces two figures. They show:
Corr_uncorr_AU.emf - plots the adjusted and unadjusted smoking prevalence over the life-course for the three oldest generations.
Corr_chi2_AU.emf - presents test statistics from χ^2 tests of the statistical significance of the difference between the adjusted and unadjusted smoking prevalence rates at each age for each group. The figure includes the critical value of the test statistic for $p<.10$ and $p<.05$.

Notes: Run this code to see if your results make sense!

Country-specific changes*:* Change only the first three lines of code to:

(i) specify the directory path;
(ii) specify the survey year as local *s* (if you combine together surveys from different years, specify latest survey year);
(iii) specify two letters representing the country name as local *cn.*

References

Christopoulou, R., Han, J., Jaber, A., Lillard, D.R. 2011. Dying for a Smoke: How Much Does Differential Mortality of Smokers Affect Estimated Life-Course Smoking Prevalence? *Preventive Medicine*, 52(1), 66–70.

Harris, J.E., 1983. Cigarette smoking among successive birth cohorts of men and women in the United States during 1900–80. *J. Natl Cancer Inst*. 71, 473–479.

Peto, R., Lopez, A.D., Boreham, J., Thun, M., Heath, C. 1992. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet* 339, 1268–1278.