Занятие 3. Перекрестное использование данных лонгитюдных файлов.

Общая рекомендация ко всем выполняемым вами заданиям:

1) Сохраняйте исходные файлы под новым именем, чтобы работать с ними.

2) СРАЗУ открывайте и сохраняйте файл аутпута. Первая команда в нем должна быть вида:

***Фамилия – номер семинара – номер задания**

3) Сохраняйте проделанную вами работу в виде кода, используя «сохранение» правильных команд в STATA (или функцию “paste” SPSS). В этом случае вы сможете дома повторить все сделанное вами в классе. Кроме того, рекомендуется прикладывать программу к вашим исследованиям.

4) В качестве отчета за семинар нужно предъявить созданные файлы данных, файл аутпута с вашей фамилией, и файл с кодом.

Исходные файлы.

Склеенные данные за 5-30 волны

hh_5_30_main.dta

ind_5_30_main.dta

ind_relatives_id_i_idind_USER_RLMS-HSE_HH_1994_2021_rus_DTA.dta

fam_all_5_30_members_year_birth.dta

11. Агрегирование данных по индивидам для лонгитюдных данных.

11.1. Предварительно, до занятия, я загрузила индивидуальный лонгитюдный файл за 1994-2021 годы с именем: **RLMS_IND_1994_2021_2022_08_21_1_v3_rus.dta**

```
use
"C:\Dropbox\work\Data_work\RLMS_panel\R30_2021\панель\RLMS_IND_1994_2021_2022_08_21_1_v3_rus.dta", clear
```

И оставила только нужные мне переменные, сохранив под новым названием.

```
keep id_w idind id_i id_h origsm inwgt region status adult child marst occup08 diplom h5 h6 born_m age i4 j1 j322 j323m j324 j325m j325y
save "C:\RLMS_work\seminar_3\data\ind_5_30_main.dta"
```

В случае, если переменные у вас названы заглавными буквами (это может быть в некоторых версиях), может потребоваться переименовать переменные для сопоставимости снтаксисов:

```
rename ( ID_W ID_I ID_H OCCUP08 H5 H6 I4 J1 J322 J323M J324 J325M J325Y) ( id_w id_i id_h occup08 h5 h6 i4 j1 j322 j323m j324 j325m j325y)
```

Вы можете этот файл **ind_5_30_main.dta** скачать из нашего урока.

ЗАДАЧА: на основе индивидуального лонгитюдного файла создать файл с агрегированными по домохозяйству в каждой волне данными (например, доля занятых).

11.2. Откройте программу STATA. Если она у вас открыта, тогда закройте файл аутпута, и откройте новый для заданий 11-13.

Начнем с создания файла результатов.

```
log using "C:\RLMS_work\seminar_3\data\Seminar_3_задания 11_13.smcl"
```

У кого STATA версии 14 и ниже, выполним команду (чтобы результаты выводились не построчно, а сразу).

set more off

Теперь загрузите файл данных:

use "C:\RLMS_work\seminar_3\data\ind_5_30_main.dta", clear

11.3. Давайте посмотрим на переменные.
Обратите внимание на разницу в именах переменных по сравнению с файлом одной волны:
1) В начале имени переменной нет буквы, соответствующей году опроса (что понятно).
2) В некоторых версиях фалов имена переменных могут быть написаны заглавными буквами, поэтому часть кода для одной волны может не подойти (тогда надо менять имена). Вообще-то это большая проблема для склеивания данных «по вертикали», если вы к имеющейся лонгитюдной базе приклеиваете новую волну самостоятельно. Но сейчас эта проблема решена.

Посмотрите также пожалуйста, что в файле есть переменная **id_w**, отвечающая за номер волны (год) и переменная идентификатора домохозяйства **id_h** – ОНА ПОДХОДИТ ТОЛЬКО ДЛЯ СКЛЕИВАНИЯ ДАННЫХ В ОДНОЙ И ТОЙ ЖЕ ВОЛНЕ (без ЛАГА, то есть данных прошлых лет)!!!

ОБЯЗАТЕЛЬНО сохраните файл под новым именем.

save "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta"

11.4. Сделаем некоторые преобразования переменных:

11.4.1. *миссинги для переменных пол (**h5**), год рождения (**h6**), возраст (**age**) и занятость (**j1**)

recode marst h5 h6 age j1 (99999997 =.a) (99999998 =.b) (99999999=.c)

```
(marst: 770 changes made)
(h5: 0 changes made)
(h6: 0 changes made)
(age: 19 changes made)
(j1: 237 changes made)
```

11.4.2. *переменную занятость (**j1**) перекодируем в переменную «есть работа» (**employed**), как мы это делали для файла одного года

recode j1 (1/4 =1) (5=0) (99999997 / 99999999 = .), gen (employed)
label variable employed "есть работа"
label define YES 1 "Да" 0 "Нет"
label values employed YES

11.4.3. *создадим группы возраста

generate age_group = age, after(age)
recode age_group (0/2 = 1) (3/5 =2) (6/15=3) (16/17=4) (18/64=5) (65/max=6)
label variable age_group "группы возраста"
label define age_group 1 "0 – 2" 2 "3-5" 3 "6-15" 4 "16-17" 5 "18- 64" 6 ">=65", replace
label values age_group age_group

и посмотрим распределение получившихся переменных

tab employed, missing

```
       есть |
      работ |
          а |      Freq.     Percent        Cum.
------------+-----------------------------------
        Нет |    154,935       38.05       38.05
         Да |    181,766       44.64       82.69
          . |     70,253       17.25       99.94
         .a |        163        0.04       99.98
         .b |         59        0.01      100.00
         .c |         15        0.00      100.00
------------+-----------------------------------
      Total |    407,191      100.00
```

tab age_group, missing

```
      групп |
          ы |
      возра |
        ста |      Freq.     Percent        Cum.
------------+-----------------------------------
      0 - 2 |     14,116        3.47        3.47
        3-5 |     15,198        3.73        7.20
       6-15 |     51,030       12.53       19.73
      16-17 |     10,369        2.55       22.28
     18- 64 |    259,649       63.77       86.04
       >=65 |     56,810       13.95      100.00
         .a |         19        0.00      100.00
------------+-----------------------------------
      Total |    407,191      100.00
```

11.4.4. создадим переменную «состоит в браке, включая неформальный» - для этого *посмотрим распределение по годам исходной переменной

codebook marst

```
marst
СЕМЕЙНОЕ ПОЛОЖЕНИЕ В ТЕКУЩЕЙ ВОЛНЕ
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------

                  type:  numeric (long)
                 label:  marst

                 range:  [1,7]                        units:  1
           unique values:  7                    missing .:  70,281/407,191
       unique mv codes:  4                      missing .*:  770/407,191

            tabulation:  Freq.   Numeric  Label
                        65,190         1  Никогда в браке не состояли
                       154,814         2  Состоите в зарегистрированном
                                          браке
                        32,581         3  Живете вместе, но не
                                          зарегистрированы
                        26,476         4  Разведены и в браке не состоите
                        39,659         5  Вдовец (вдова)
                         1,256         6  ОФИЦИАЛЬНО ЗАРЕГИСТРИРОВАНЫ, НО
                                          ВМЕСТЕ НЕ Ж
                        16,164         7  Состоите в браке
                        70,281         .
                           202        .a
                           124        .b
                           444        .c
```

Вы видите, что 70,281– миссинги, это дети. Отберем только тех, кто во «взрослой» анкете и посмотрим распределение переменной по годам опроса

tabulate id_w marst if adult ==1

НОМЕР ВОЛНЫ	СЕМЕЙНОЕ ПОЛОЖЕНИЕ В ТЕКУЩЕЙ ВОЛНЕ							
	Никогда в	Состоите	Живете вм	Разведены	Вдовец (в	ОФИЦИАЛЬН	Состоите	Total
1994 год	1,183	0	0	686	1,017	0	5,642	8,528
1995 год	1,416	0	0	629	1,024	0	5,315	8,384
1996 год	1,456	0	0	631	1,007	0	5,207	8,301
1998 год	1,599	4,868	588	582	1,040	0	0	8,677
2000 год	1,781	4,899	691	606	1,087	0	0	9,064
2001 год	2,009	5,257	862	725	1,224	0	0	10,077
2002 год	2,159	5,270	972	777	1,276	37	0	10,491
2003 год	2,167	5,272	1,076	828	1,264	0	0	10,607
2004 год	2,251	5,277	1,043	818	1,260	0	0	10,649
2005 год	2,205	5,120	980	804	1,214	0	0	10,323
2006 год	2,655	6,028	1,221	1,016	1,445	99	0	12,464
2007 год	2,575	5,973	1,249	939	1,442	93	0	12,271
2008 год	2,479	5,807	1,136	944	1,391	93	0	11,850
2009 год	2,385	5,771	1,354	898	1,368	25	0	11,801
2010 год	3,481	8,896	2,010	1,300	1,997	111	0	17,795
2011 год	3,570	9,074	2,090	1,409	2,059	90	0	18,292
2012 год	3,487	9,278	2,202	1,466	2,135	104	0	18,672
2013 год	3,287	8,888	2,136	1,434	2,121	86	0	17,952
2014 год	2,776	7,527	1,809	1,213	1,761	67	0	15,153
2015 год	2,795	7,529	1,729	1,184	1,810	64	0	15,111
2016 год	2,774	7,683	1,684	1,247	1,815	86	0	15,289
2017 год	2,937	7,679	1,754	1,254	1,781	57	0	15,462
2018 год	2,864	7,375	1,578	1,263	1,784	68	0	14,932
2019 год	2,940	7,284	1,527	1,279	1,768	52	0	14,850
2020 год	2,967	7,074	1,439	1,256	1,785	61	0	14,582
2021 год	2,992	6,985	1,451	1,288	1,784	63	0	14,563
Total	65,190	154,814	32,581	26,476	39,659	1,256	16,164	336,140

Из этих данных видно, что в 1994-96 гг. в вопросе не было различия между официальным и неофициальным браком (код 7), а с 1998 года из разделили (коды 2 и 3). Кроме того, только с 2006 года добавили опцию «ОФИЦИАЛЬНО ЗАРЕГИСТРИРОВАНЫ, НО ВМЕСТЕ НЕ ЖИВЕТЕ».
Будем также считать не живущих вместе не состоящими в браке.
Создадим переменную «в браке (включая неофициальный)» (**married**)

recode marst (1=0) (2/3=1) (4/6=0) (7=1) (99999997 / 99999999 = .) if adult ==1, gen (married)
label variable married "в браке (включая неофициальный)"
label values married YES

11.5. *Посмотрим распределения новых переменных (**age_group, employed, married**) в репрезентативной выборке, а также с учетом того, что две последние определены только для взрослых

tabulate age_group if origsm ==1, missing

группы возраста	Freq.	Percent	Cum.
0 - 2	8,632	2.86	2.86
3-5	9,582	3.17	6.03
6-15	35,641	11.80	17.83
16-17	7,832	2.59	20.42

```
     18- 64 |    191,484       63.39       83.81
       >=65 |     48,906       16.19      100.00
         .a |          7        0.00      100.00
------------+-----------------------------------
      Total |    302,084      100.00
```

tabulate employed if (origsm ==1 & adult==1) , missing

```
       есть |
      работ |
          а |      Freq.     Percent        Cum.
------------+-----------------------------------
        Нет |    124,696       48.79       48.79
         Да |    130,682       51.13       99.92
         .a |        146        0.06       99.98
         .b |         40        0.02      100.00
         .c |         12        0.00      100.00
------------+-----------------------------------
      Total |    255,576      100.00
```

tabulate married if (origsm ==1 & adult==1) , missing

```
          в |
      браке |
     (включ |
         ая |
     неофи  |
     циаль  |
       ный) |      Freq.     Percent        Cum.
------------+-----------------------------------
        Нет |    105,564       41.30       41.30
         Да |    149,311       58.42       99.73
          . |         17        0.01       99.73
         .a |        147        0.06       99.79
         .b |        106        0.04       99.83
         .c |        431        0.17      100.00
------------+-----------------------------------
      Total |    255,576      100.00
```

11.6. *Создадим дамми возраста

quietly tabulate age_group, generate(age_)

СОХРАНИТЕ ВАШ ФАЙЛ ДАННЫХ!!!

save "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", replace

11.7. *Создадим агрегированный файл по ключу «идентификатор домохозяйства».
*Посчитаем среднее по занятым (**employed**) (реально это среди взрослых) и суммируем по остальным переменным (**married** и **age_1 age_2 age_3 age_4 age_5 age_6**):

collapse (mean) employed (sum) married age_1 age_2 age_3 age_4 age_5 age_6 , by(id_w id_h)

При этом сразу получается новый файл, где кейсом является домохозяйство. Отсортируем и сохраним его:

sort id_w id_h

save "C:\RLMS_work\seminar_3\data\h_5_30_aggr.dta"

11.8. Посмотрим характеристики переменной **id_h**

summarize id_h

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        id_h |    151,868     1834824     3561536       1001   2.41e+07
```

В этом файле 151,868 кейсов (домохозяйств за все годы).

12. Приклеивание к семейному файлу агрегированных данных по индивидам для лонгитюдных данных.

Предварительно до занятия, я загрузила семейный лонгитюдный файл за 1994-2021 годы **RLMS_HH_1994_2021_2022_09_06_rus_DTA.dta**

И оставила только нужные мне переменные, сохранив под новым названием.

keep id_w id_h hhwgt origsam nfm f14
save "C:\RLMS_work\seminar_3\data\hh_5_30_main.dta"

(В некоторых версиях файла может понадобиться переименовать переменные, если верхний регистр)

rename (ID_W ID_H F14) (id_w id_h f14)

Вы можете этот файл **hh_5_30_main.dta** скачать из нашего урока.

ЗАДАЧИ: 1) создать переменную дефлированного душевого дохода; 2) к семейному файлу приклеить переменные из ранее созданного агрегированного файла

12.1. *Откроем этот файл домохозяйств.

use "C:\RLMS_work\seminar_3\data\hh_5_30_main.dta", clear

*И сохраним под новым именем:

save "C:\RLMS_work\seminar_3\data\hh_5_30_main_s3.dta"

12.2. *Перекодируем миссинги для переменной суммарного дохода д\х **F14** (аналогично файлу за одну волну)

recode f14 (99999997 / 99999999 =.)

*Рассчитаем НОМИНАЛЬНЫЙ душевой доход делением на количество членов семьи **nfm**

gen INCOME_PC_N = f14 / nfm if nfm>0
label variable INCOME_PC_N "Душевой доход номинальный"

```
(6,095 missing values generated)
```

12.3. *Посмотрим описательные характеристики переменных

sum nfm f14 INCOME_PC_N

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         nfm |    151,906    2.760029    1.460367          1         16
         f14 |    145,811    97358.67    472050.1          0   3.75e+07
   INCOME_PC_N |    145,811    37044.08    180143.1          0   1.85e+07
```

Всего 151,906 кейсов, в переменных **f14** и **INCOME_PC_N** есть пропущенные значения.
Вы видите, что в файле домохозяйств на 38 кейсов больше (151,906 – 151,868), чем в агрегированном файле (так как есть д\х, в которых не опрошен ни один человек).
Максимальные значения суммарного и номинального душевого дохода очень велики, так как в 1994-1996 гг. цены и доходы в абсолютном выражении были очень большими.

12.4. Создадим переменную «дефлятор приведения к 2021 году» (**deflat_30**), так как каждый год цены растут (инфляция), и, кроме того, за 1994-1996 гг. значения нужно разделить на 1000, так как в 1998 г. была деноминация. Используем цепные среднероссийские индексы потребительских цен (ИПЦ) по данным Росстата. Я взяла ИПЦ «декабрь к декабрю предыдущего года», так как данные собирают в октябре-декабре.
https://rosstat.gov.ru/statistics/price#
Для 2021 года (30 волна) дефлятор равен 1, для 2020 (29 волна) – самому ИПЦ, для 2019 года (28 волна) – произведению ИПЦ за два года, и т.д.
Для первых трех волн дефлятор дополнительно делим на 1000 в силу деноминации в 1998 году.
(код на новой странице)

gen deflat_30=1 if id_w == 30
replace deflat_30= 1.0839 if id_w == 29
replace deflat_30= 1.0839 * 1.0491 if id_w == 28
replace deflat_30= 1.0839 * 1.0491 * 1.03 if id_w == 27
replace deflat_30= 1.0839 * 1.0491 * 1.03 * 1.043 if id_w == 26
replace deflat_30= 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 if id_w == 25
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 if id_w == 24
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291 if id_w == 23
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135 if id_w == 22
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647 if id_w == 21
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 if id_w == 20
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 if id_w == 19
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 * 1.088 if id_w == 18
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 if id_w == 17
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 if id_w == 16
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 if id_w == 15
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09 if id_w == 14
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109 if id_w == 13
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117 if id_w == 12
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12 if id_w == 11
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12*1.151 if id_w == 10
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12*1.151*1.186 if id_w == 9
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12*1.151*1.186*1.202*1.365 if id_w == 8
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12*1.151*1.186*1.202*1.365*1.844*1.11/1000 if id_w == 7
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12*1.151*1.186*1.202*1.365*1.844*1.11*1.218/1000 if id_w == 6
replace deflat_30 = 1.0839 * 1.0491 * 1.03 * 1.043 * 1.0251 * 1.0539 * 1.1291*1.1135*1.0647*1.0657 * 1.061 *1.088 * 1.088 * 1.133 * 1.119 * 1.09*1.109*1.117*1.12*1.151*1.186*1.202*1.365*1.844*1.11*1.218*2.3/1000 if id_w == 5
label variable deflat_30 "дефлятор прив. цен к 2021 по среднерос. ИПЦ (умножать на него)"

12.5. *Рассчитаем душевой доход реальный (дефлированный) **INCOME_PC**, умножив номинальный доход на дефлятор.

gen INCOME_PC = INCOME_PC_N* deflat_30
label variable INCOME_PC "Душевой доход дефлированный"

*Рассчитаем средний душевой доход (и номинальный и реальный) по годам с весом по домохозяйствам (для репрезентативности)

tabstat INCOME_PC INCOME_PC_N [aw = hhwgt], stat (mean) by (id_w)

```
Summary statistics: Mean
Group variable: ID_W (Волна (год проведения исследования))

Summary statistics: mean
  by categories of: id_w (Волна (год проведения исследования))

    id_w |  INCOME~C  INCOME~N
---------+--------------------
1994 год |  10941.54   187923.8
1995 год |   8862.105  350080.6
1996 год |   8899.748  428209.4
1998 год |   7531.021  741.6794
2000 год |   7767.512  1255.109
2001 год |   9862.835  1890.105
2002 год |  10721.03   2364.809
2003 год |  11825.65   2921.477
2004 год |  12861.35   3549.091
2005 год |  15271.01   4673.369
2006 год |  17776.74   5929.812
2007 год |  18295.02   6828.915
2008 год |  21884.1    9255.023
2009 год |  22195.65   10212.81
2010 год |  23986.67   12008.16
2011 год |  24395.37   12957.74
2012 год |  25723.72   14560.98
2013 год |  26469.05   15952.27
2014 год |  25392.35   17040.3
2015 год |  25287.04   19160.41
2016 год |  24131.02   19270.01
2017 год |  25626.11   20977.56
2018 год |  25089.9    21421.78
2019 год |  25253.38   22208.21
2020 год |  24412.75   22523.06
2021 год |  26146.92   26146.92
---------+--------------------
   Total |  19932.51   43276.51
```

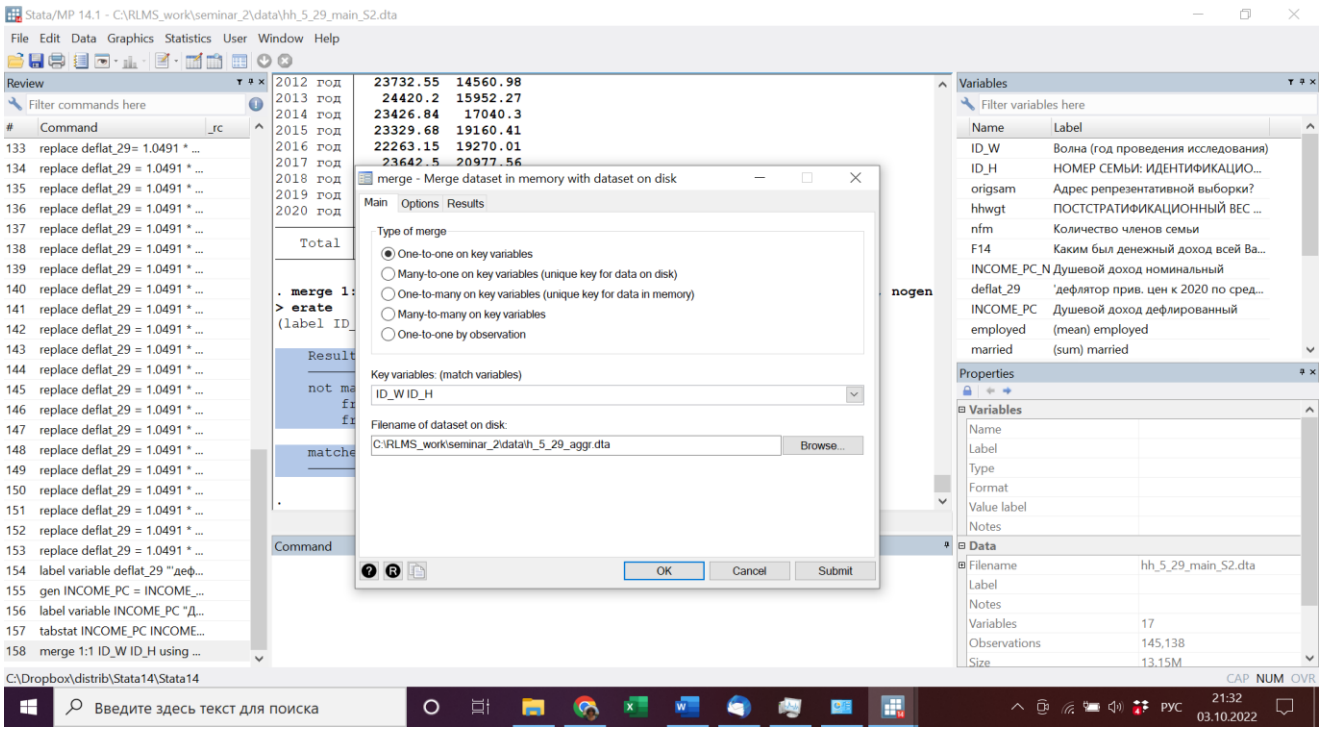
Вы видите, что реальные душевые доходы были минимальными в 1998 году и максимальными в 2013 году (2021 близок к 2013).

12.6. *Приклеим агрегированные ранее данные («один к одному»)

merge 1:1 id_w id_h using "C:\RLMS_work\seminar_3\data\h_5_30_aggr.dta", nogenerate

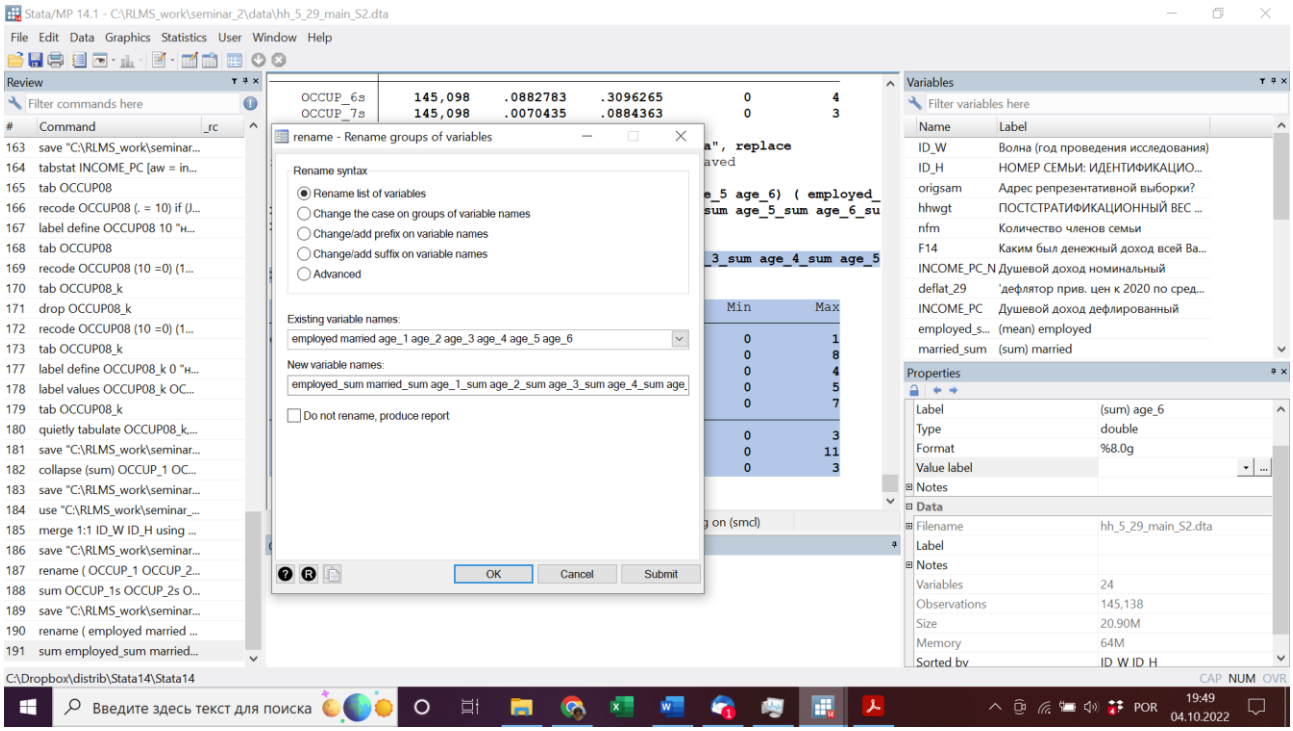
```
    Result                           # of obs.
    -----------------------------------------
    not matched                             38
        from master                         38
        from using                           0

    matched                           151,868
    -----------------------------------------
```

Как и ожидалось, для 38 кейсов не нашлось агрегированных данных.
Обязательно посмотрите на данные в таблице (браузер данных).



12.7. *Переименуйте приклеенные переменные (для того, чтобы различать с индивидуальными переменными с такими же именами) и их лейблы

rename (employed married age_1 age_2 age_3 age_4 age_5 age_6) (employed_mean married_sum age_1_sum age_2_sum age_3_sum age_4_sum age_5_sum age_6_sum)



label variable age_1_sum " (sum) age_1 0-2"
label variable age_2_sum " (sum) age_2 3-5"
label variable age_3_sum " (sum) age_3 6-15"
label variable age_4_sum " (sum) age_4 16-17"
label variable age_5_sum " (sum) age_5 18-64"
label variable age_6_sum " (sum) age_6 >=65"

СОХРАНИТЕ ФАЙЛ ДАННЫХ!!!

save "C:\RLMS_work\seminar_3\data\hh_5_30_main_s3.dta", replace

12.8. Самостоятельное задание в классе.
- Посмотрите распределение или описательные характеристики приклеенных переменных.

13. Приклеивание семейных данных к индивидуальным в лонгитюдном файле.

ЗАДАЧА: приклеить переменные суммарного семейного дохода, душевого дохода и количества человек в семье к индивидуальному лонгитюдному файлу.

13.1. *Откроем снова преобразованный индивидуальный файл

use "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", clear

*и приклеим к нему данные из семейного файла (преобразованного), указав нужные переменные

merge m:1 id_w id_h using "C:\RLMS_work\seminar_3\data\hh_5_30_main_s3.dta", keepusing(INCOME_PC INCOME_PC_N f14 nfm)

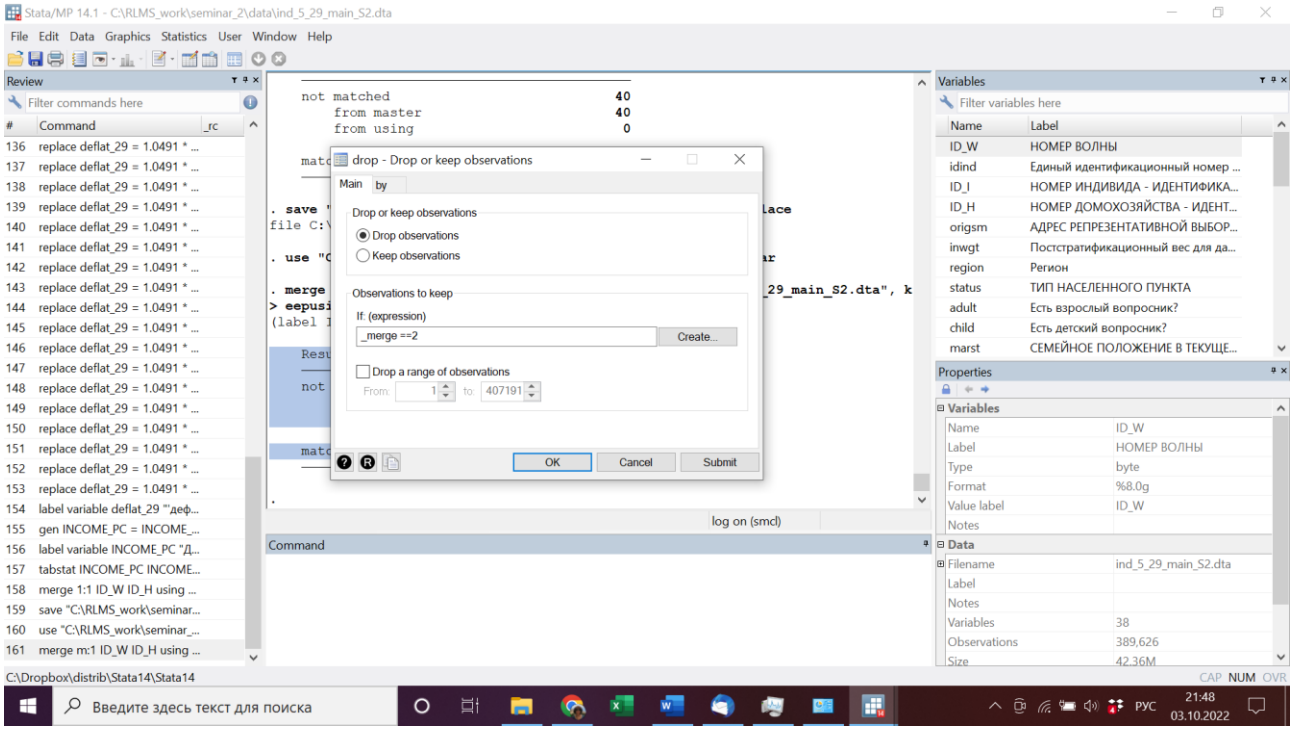
```
(label id_w already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             38
        from master                          0  (_merge==1)
        from using                          38  (_merge==2)

    matched                            407,191  (_merge==3)
    -----------------------------------------
```

Те же 38 кейсов добавились к нашей базе индивидов. Их нужно удалить, так как им не соответствуют заполненные анкеты индивидов.

drop if _merge ==2
drop _merge

СОХРАНИТЕ ФАЙЛ ДАННЫХ!!!

save "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", replace



13.2. И также посмотрим распределение среднего реального дохода по годам, но веса – индивидуальные

tabstat INCOME_PC [aw = inwgt], stat (mean) by (id_w)

```
Summary for variables: INCOME_PC
     by categories of: id_w (НОМЕР ВОЛНЫ)

    id_w |      mean
---------+----------
1994 год | 10327.52
1995 год |  8186.163
1996 год |  8343.832
1998 год |  9741.812
2000 год |  7448.683
2001 год |  9621.384
2002 год |  10248.98
2003 год |  11324.26
2004 год |  12350.98
2005 год |  14549.11
2006 год |  17436.86
2007 год |  17697.71
2008 год |  21402.75
2009 год |  20447.31
2010 год |  21809.1
2011 год |  22434.89
2012 год |  23952.36
2013 год |  24656.96
2014 год |  23586.03
2015 год |  23569.76
2016 год |  22277.37
2017 год |  23996.92
2018 год |  23546.04
2019 год |  23573.56
2020 год |  22859.92
2021 год |  24087.19
---------+----------
   Total |  18583.87
--------------------
```

13.3. Сравните средний по годам реальный доход, рассчитанный по семьям, и по индивидам. Почему есть разница?

13.4. Сохраните do-файл (файл кода) с названием примерно «семинар 3 часть 1 фамилия». Закройте лог-файл (аутпута).
Оба эти файла – ваш отчет за первую половину семинара 3.

13.5. Самостоятельное задание (домашнее).
В файле отдельного аутпута для самостоятельного задания наберите команду:
***Фамилия – номер семинара – номер задания**

- Посмотрите распределение переменной **occup08.**
- Сделайте преобразования переменной **occup08** аналогично тем, которые мы делали для файла одного года на прошлом занятии:
1) перекодируйте в переменной **occup08** миссинг (.) в 10, если переменная **J1** равна 5 (человек не работает) и человек взрослый (переменная **adult** равна 1).
2) добавьте метку значений 10 «не работает». Посмотрите распределение.
3) перекодируйте в новую переменную **occup08_k**, создав градации:
1 – не работает (код 10 переменной **occup08**);
2 – законодатели, руководители и т.д. (код 1 переменной **occup08**);
3 – специалисты (код 2 переменной **occup08**);
4 – служащие (коды 3, 4 и 5 переменной **occup08**);
5 – рабочие (коды 6, 7 и 8 переменной **occup08**);
6 – чернорабочие (код 9 переменной **occup08**);
7 – военные (код 0 переменной **occup08**).
Отсутствие ответа (99999997, 99999998, 99999999) перекодировать в миссинг (.).
4) Сделайте новые лейблы для градаций этой переменной.
5) Посмотрите распределение этой новой переменной.
6) Создайте из нее дамми: **occup_1 occup_2 occup_3 occup_4 occup_5 occup_6 occup_7**
7) Сохраните ваш файл
9) аггрегируйте по году и идентификатору семьи, создав новый файл с количеством человек в каждой категории занятости (на основе созданных вами дамми)
10) откройте файл для домохозяйств и приклейте к нему агрегированный файл
11) переименуйте эти новые переменные (имена новых переменных: **occup_1sum occup_2sum occup_3sum occup_4sum occup_5sum occup_6sum occup_7sum**) и посмотрите их характеристики коммандой **sum**
12) сохраните файл.
- Скопируйте результаты и команды из окна аутпута, и вставьте их в текстовый файл с вашими ответами на задания этого семинара. Один текстовый файл для всех выполненных заданий. Назовите файл вашей ФИО и номер группы, укажите номер семинара.

14. Слияние индивидуальных данных членов одного домохозяйства (семьи) в базе данных РМЭЗ НИУ ВШЭ. Работа с файлом родственников.

14.1. Мы будем работать сначала с файлом родственных связей
ind_relatives_id_i_idind_USER_RLMS-HSE_HH_1994_2021_rus.sav
Он предварительно был преобразован к формату STATA:
ind_relatives_id_i_idind_USER_RLMS-HSE_HH_1994_2021_rus_DTA.dta

Если у вас остался открыт файл аутпута, пожалуйста, закройте его и откройте новый файл:

log using "C:\RLMS_work\seminar_3\data\Seminar_3_задания 14_16.smcl"

Наберите в нем такую команду, заменив Вашу фамилию:

***Фамилия - семинар 3 – задания 14-16**
set more off

14.2. Откроем этот файл родственных связей:
use "C:\RLMS_work\seminar_3\data\ind_relatives_id_i_idind_USER_RLMS-HSE_HH_1994_2021_rus_DTA.dta", clear

В этом файле каждому человеку в данной волне сопоставляется идентификатор его родственников.
И сохраним с новым именем, чтобы не «испортить» исходный файл:
save "C:\RLMS_work\seminar_3\data\ind_relatives_5_30.dta"
!!!! ВАЖНО!!!!
В данном файле в качестве идентификаторов этих родственников используются идентификаторы индивидов в данной волне (**id_i**), а не универсальный идентификатор **idind.**
Поэтому использовать этот файл непосредственно для склеивания информации между разными волнами НЕЛЬЗЯ. В исходной версии переменная **idind** есть только для самого респондента.
МОЖНО: заменить для каждого родственника **id_i** на его **idind**, создав предварительно файл соответствия этих идентификаторов в каждой волне.

14.3. Посмотрите, как устроен файл и имена переменных. Сначала идут переменные по типам родственников без различий по полу, затем родственники-мужчины, затем родственницы-женщины.

* Теперь создадим несколько полезных переменных о наличии в семье для каждого индивида его\ее:

* наличие супруга (супруги)
recode r01a1 (.=0 "нет") (1 / max =1 "да") , into (spouse_yes)
label variable spouse_yes "есть супруг(а) "

* наличие родителей, отчима или мачехи
recode r02m1 (.=0 "нет") (1 / max =1 "да") , into (father_yes)
recode r03m1 (.=0 "нет") (1 / max =1 "да") , into (father_l_yes)
label variable father_yes "есть отец в семье"
label variable father_l_yes "есть отчим в семье"

recode r02f1 (.=0 "нет") (1 / max =1 "да"), into (mother_yes)
recode r03f1 (.=0 "нет") (1 / max =1 "да"), into (mother_l_yes)

label variable mother_yes "есть мать в семье"
label variable mother_l_yes "есть мачеха в семье"

14.4. * Посчитаем количество родственников человека, живущих с ним в одном домохозяйстве (команда **egen newv4 = rownonmiss(v1 v2 v3)** подсчитывает количество непропущенных значений в перечисленных переменных; в случае отсутствия таковых, будет значение «ноль»). Обратите внимание, что мы не узнаем, есть ли у человека такие родственники, живущие отдельно.

* количество дедушек
egen gr_father = rownonmiss(r08m1 r08m2)
label variable gr_father "количество дедушек в семье"

* количество бабушек
egen gr_mother = rownonmiss(r08f1 r08f2)
label variable gr_mother " количество бабушек в семье"

* количество прабабушек и прадедушек (вместе)
egen grgr_parents = rownonmiss(r17a1 r17a2)
label variable grgr_parents "количество прадедушек и\или прабабушек в семье"

*количество братьев и\или сестер
egen siblings = rownonmiss(r06a1 r06a2 r06a3 r06a4 r06a5 r06a6 r06a7 r06a8 r07a1 r07a2 r07a3 r07a4 r07a5 r07a6 r07a7)
label variable siblings "количество братьев и сестер в семье"

*количество двоюродных братьев и\или сестер
egen cousins = rownonmiss(r16a1 r16a2 r16a3 r16a4 r16a5)
label variable cousins "количество кузенов\кузин в семье"

Сохраните файл
save "C:\RLMS_work\seminar_3\data\ind_relatives_5_30.dta", replace

14.5. Самостоятельное задание (домашнее).

В файле отдельного аутпута для самостоятельного задания наберите команду:
***Фамилия – номер семинара – номер задания**

- Посчитайте количество у человека в семье: тещ\свекровей (переменная **lawmother**); тестей\свекров (**lawfather**); внуков\внучек (**gr_chilgren**). Создайте соответствующие лейблы.
- Посмотрите распределения (или описательные характеристики) получившихся новых переменных (начиная от **spouse_yes** до **gr_chilgren**).
- Сохраните файл.
- Скопируйте результаты и команды из окна аутпута, и вставьте их в текстовый файл с вашими ответами на задания этого семинара. Один текстовый файл для всех выполненных заданий. Назовите файл вашей ФИО и номер группы, укажите номер семинара.

14.6. Так как имена переменных однотипные и не говорят об их сущности, создадим новые переменные для каждого из нужных родственников. ВАЖНО!!! Так как идентификаторы **id_i**, которые используются для установления соответствий между родственниками, очень большие «числа» (до 8-10 знаков), они в исходном файле «родственников» имеют тип «double» (число «двойной точности»). И также в файлах данных по индивидам.

Если создать новую переменную другого типа, в результате склейки будет много ошибок (в одном из вариантов работы у меня приклеилось всего около 60 тыс.кейсов вместо около примерно 160 тысяч). Поэтому при создании новых переменных сразу указываем тип «double».

* **idi** супруга \или супруги
generate double idi_spouse= r01a1
label variable idi_spouse "ID_I супруга\супруги"

* **idi** отца (отчима), при этом меняем значение на **id_i** отчима, если значение переменной «миссинг»
generate double idi_father=r02m1
replace idi_father=r03m1 if idi_father == .

Обратите внимание, эти замены – это количество отчимов у всех респондентов за все годы.
```
(12,096 real changes made)
```

* **idi** матери (мачехи), при этом меняем значение на **id_i** мачехи, если значение переменной «миссинг»
generate double idi_mother=r02f1
replace idi_mother=r03f1 if idi_mother== .
label variable idi_father "ID_I отца или отчима"
label variable idi_mother "ID_I матери или мачехи"

Обратите внимание, эти замены – это количество мачех у всех респондентов за все годы – оно в десять раз меньше количества отчимов.
```
(1,209 real changes made)
```

* **idi** родных и неродных детей
generate double idi_lawchild1 = r05a1
label variable idi_lawchild1 "ID_I 1 неродн.ребенка"
generate double idi_lawchild2 = r05a2
label variable idi_lawchild2 "ID_I 2 неродн.ребенка"
generate double idi_lawchild3 = r05a3
label variable idi_lawchild3 "ID_I 3 неродн.ребенка"
generate double idi_lawchild4 = r05a4
label variable idi_lawchild4 "ID_I 4 неродн.ребенка"
generate double idi_lawchild5 = r05a5
label variable idi_lawchild5 "ID_I 5 неродн.ребенка"
generate double idi_lawchild6 = r05a6
label variable idi_lawchild6 "ID_I 6 неродн.ребенка"
generate double idi_child2 = r04a2
label variable idi_child2 "ID_I 2 ребенка"
generate double idi_child3 = r04a3
label variable idi_child3 "ID_I 3 ребенка"
generate double idi_child4 = r04a4
label variable idi_child4 "ID_I 4 ребенка"
generate double idi_child5 = r04a5
label variable idi_child5 "ID_I 5 ребенка"
generate double idi_child6 = r04a6
label variable idi_child6 "ID_I 6 ребенка"
generate double idi_child7 = r04a7
label variable idi_child7 "ID_I 7 ребенка"

generate double idi_child8 = r04a8
label variable idi_child8 "ID_I 8 ребенка"
generate double idi_child9 = r04a9
label variable idi_child9 "ID_I 9 ребенка"
generate double idi_child1 = r04a1
label variable idi_child1 "ID_I 1 ребенка"

Сохраните файл
save "C:\RLMS_work\seminar_3\data\ind_relatives_5_30.dta", replace

14.7. * Отсортируем файл по переменным номер волны и идентификатора человека, сохраним файл под ДРУГИМ именем, сохраним нужные переменные, и еще раз сохраним файл (будьте внимательны, если вы просто сохраните этот файл, вы потеряете почти все исходные переменные). Операция необратима!

sort id_w idind
save "C:\RLMS_work\seminar_3\data\ind_relatives_5_30_short.dta"

keep id_w idind id_i idi_spouse idi_father idi_mother idi_child1 idi_child2 idi_child3 idi_child4 idi_child5 idi_child6 idi_child7 idi_child8 idi_child9 idi_lawchild1 idi_lawchild2 idi_lawchild4 idi_lawchild5 idi_lawchild6 spouse_yes father_yes father_l_yes mother_yes mother_l_yes gr_father gr_mother grgr_parents siblings

save "C:\RLMS_work\seminar_3\data\ind_relatives_5_30_short.dta", replace

15. Формирование вспомогательных файлов с данными по родственникам.

15.1. Откроем исходный лонгитюдный индивидуальный файл.

use "C:\RLMS_work\seminar_3\data\ind_5_30_main.dta", clear

(Он может быть получен из лонгитюдного файла, скачанного с сайта, с сохранением нужных переменных:
id_w idind id_i id_h origsm inwgt region status adult child marst occup08 diplom h5 h6 born_m age i4 j1 j322 j323m j324 j325m j325y
и удалением миссингов 99999997-99999999).

* Обратите внимание на различия идентификаторов **idind** и **id_i**!!!
idind - идентификатор для ЛЮБОЙ волны
id_i - идентификатор ВНУТРИ ТОЛЬКО ОДНОЙ волны
(формируется на основе номера места жительства, номера семьи и номера члена семьи)
в разных волнах у одного и того же человека этот номер не совпадает, его нельзя использовать для склеивания данных в разных волнах

15.2. Сохраним несколько вспомогательных файлов

*сохраним этот же файл как "вспомогательный" для приклеивания данных родственников

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_relatives.dta"

(он теперь открыт в программе)

* Удалим ненужные переменные и снова сохраним файл

drop id_h origsm inwgt region status adult child marst j322 j323m j324 j325m j325y
save "C:\RLMS_work\seminar_3\data\ind_5_30_for_relatives.dta", replace

* Используем этот файл для сохранения данных по супругу

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_spouse.dta"

* Используем этот файл для сохранения данных по отцу (отчиму)

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_father.dta"

* Используем этот файл для сохранения данных по матери (мачехе)

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_mother.dta"

* Используем этот файл для сохранения данных по первому ребенку

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_child_1.dta"

15.3.
В программе открыт файл (если нет, откройте) **ind_5_30_for_child_1.dta**
* Оставим только нужные переменные в файле с данными по первому ребенку

drop occup08 diplom i4 j1

*сохраним файл

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_child_1.dta", replace

* переименуем переменные

rename (idind id_i h5 h6 born_m age) (ch1_idind idi_child1 ch1_h5 ch1_h6 ch1_born_m ch1_age)

* можно также изменить лейблы переменных, добавив в начале CH1_
* сохраним файл

label variable ch1_idind "CH1_Единый идентификационный номер индивида дл"
label variable idi_child1 "CH1_НОМЕР ИНДИВИДА - ИДЕНТИФИКАЦИОННЫЙ"
label variable ch1_h5 "CH1_Пол респондента"
label variable ch1_h6 "CH1_Год рождения респондента (=J69.9C)"
label variable ch1_age "CH1_Количество полных лет"
label variable ch1_born_m "CH1_Месяц рождения ребенка"

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_child_1.dta", replace

15.4. Откроем файл для данных по супругу

use "C:\RLMS_work\seminar_3\data\ind_5_30_for_spouse.dta", clear

*удалим ненужные переменные

drop born_m

* переименуем переменные

rename (idind id_i occup08 diplom h5 h6 age i4 j1) (s_idind idi_spouse s_occup08 s_diplom s_h5 s_h6 s_age s_i4 s_j1)

* можно также изменить лейблы переменных, добавив в начале S_
* сохраним файл

label variable s_idind "S_Единый идентификационный номер индивида дл"
label variable idi_spouse "S_НОМЕР ИНДИВИДА - ИДЕНТИФИКАЦИОННЫЙ"
label variable s_occup08 "S_ПРОФЕССИОНАЛЬНАЯ ГРУППА - по коду J2COD08"
label variable s_diplom "S_ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА)"
label variable s_h5 "S_Пол респондента"
label variable s_h6 "S_Год рождения респондента (=J69.9C)"
label variable s_age "S_Количество полных лет"
label variable s_i4 "S_Кем Вы себя считаете по национальности? Я им"
label variable s_j1 "S_Ваше основное занятие в настоящее время"

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_spouse.dta", replace

15.5. * Откроем файл для данных по матери

use "C:\RLMS_work\seminar_3\data\ind_5_30_for_mother.dta", clear

*удалим ненужные переменные
drop h5 h6 born_m i4

* переименуем переменные

rename (idind id_i occup08 diplom age j1) (m_idind idi_mother m_occup08 m_diplom m_age m_j1)

* можно также изменить лейблы переменных, добавив в начале M_
* сохраним файл

label variable m_idind "M_Единый идентификационный номер индивида дл"
label variable idi_mother "M_НОМЕР ИНДИВИДА - ИДЕНТИФИКАЦИОННЫЙ"
label variable m_occup08 "M_ПРОФЕССИОНАЛЬНАЯ ГРУППА - по коду J2COD08"
label variable m_diplom "M_ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА)"
label variable m_age "M_Количество полных лет"
label variable m_j1 "M_Ваше основное занятие в настоящее время"

save "C:\RLMS_work\seminar_3\data\ind_5_30_for_mother.dta", replace

16. Приклеивание к рабочему файлу вспомогательных файлов с данными по родственникам.

16.1. теперь снова откроем «рабочий» индивидуальный файл

use "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", clear

*посмотрим, сколько в нем кейсов (без взвешивания, лонгитюдная выборка)

tab id_w

```
      НОМЕР |
      ВОЛНЫ |      Freq.     Percent        Cum.
------------+-----------------------------------
   1994 год |     11,289        2.77        2.77
   1995 год |     10,666        2.62        5.39
   1996 год |     10,464        2.57        7.96
   1998 год |     10,675        2.62       10.58
   2000 год |     10,976        2.70       13.28
   2001 год |     12,121        2.98       16.26
   2002 год |     12,523        3.08       19.33
   2003 год |     12,656        3.11       22.44
   2004 год |     12,641        3.10       25.54
   2005 год |     12,237        3.01       28.55
   2006 год |     14,689        3.61       32.16
   2007 год |     14,505        3.56       35.72
   2008 год |     14,026        3.44       39.16
   2009 год |     13,991        3.44       42.60
   2010 год |     21,343        5.24       47.84
   2011 год |     21,991        5.40       53.24
   2012 год |     22,534        5.53       58.78
   2013 год |     21,753        5.34       64.12
   2014 год |     18,372        4.51       68.63
   2015 год |     18,429        4.53       73.16
   2016 год |     18,756        4.61       77.76
   2017 год |     18,954        4.65       82.42
   2018 год |     18,234        4.48       86.89
   2019 год |     18,060        4.44       91.33
   2020 год |     17,701        4.35       95.68
   2021 год |     17,605        4.32      100.00
------------+-----------------------------------
      Total |    407,191      100.00
```

В файле «родственников» у нас было 419,273 кейсов (в нем содержатся вообще все люди из опрошенных домохозяйств); в рабочем файле индивидов - 407,191 (кто ответил на анкету). Эта разница образуется за счет того, что некоторые члены домохозяйства не были опрошены. Но их очень мало – за 27 лет это всего чуть больше 12 тыс.человек, или менее 2% выборки.

* приклеим (один к одному) к файлу идентификаторы родственников из «короткого» файла по ключевым переменным **id_w id_i** (так как **idind** в нем пока нет для 30й волны).

merge 1:1 id_w id_i using "C:\RLMS_work\seminar_3\data\ind_relatives_5_30_short.dta"

```
    Result                           # of obs.
    -----------------------------------------
    not matched                         12,082
        from master                          0  (_merge==1)
        from using                      12,082  (_merge==2)

    matched                            407,191  (_merge==3)
    -----------------------------------------
```

У нас приклеились «лишние» 12082 кейса (для которых нет индивидуальных анкет). Удалим эти кейсы.

drop if _merge == 2

```
(12,082 observations deleted)
```

* Распределения по раундам получаются идентичные.

tab id_w

В этом аутпуте мы тоже видим, что ко всем кейсам были приклеены переменные из файла родственников. Поэтому можем сохранить файл. И удалим переменную-индикатор.

drop _merge

save "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", replace

16.2. *Приклеим к нашему рабочему файлу переменные из файла «первый ребенок» (для примера), «многие к одному», по переменным-ключам **id_w idi_child1**.

Проверим, у скольких респондентов есть информация об идентификаторе первого ребенка. *Для этого перекодируем идентификатор в проверочную дихотомическую переменную и посмотрим ее распределение

recode idi_child1 (.=0) (1/max=1), into (check)
tab check

```
  RECODE of |
 idi_child1 |
    (ID_I 1 |
      ребен |
        ка) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |    241,104       59.21       59.21
          1 |    166,087       40.79      100.00
------------+-----------------------------------
      Total |    407,191      100.00
```

Такая информация есть у 166087 респондентов. Приклеим теперь данные о первом ребенке:

merge m:1 id_w idi_child1 using
"C:\RLMS_work\seminar_3\data\ind_5_30_for_child_1.dta"

```
    Result                           # of obs.
    -----------------------------------------
    not matched                       543,881
        from master                   244,938  (_merge==1)
        from using                    298,943  (_merge==2)

    matched                           162,253  (_merge==3)
    -----------------------------------------
```

Обратите внимание, информация о первом ребенке «приклеилась» к 162253 респондентам. То есть почти по 4м тысячам детей информация отсутствует (на них не были заполнены анкеты). Всего 244938 респондентов не имеют первого ребенка, или о нем отсутствует информация.
Но к нашему файлу также приклеились 298943 кейса из «внешнего» файла – это те респонденты, которые «не являются детьми». Нам нужно их удалить.

drop if _merge == 2

```
(298,943 observations deleted)
```

*Посмотрите распределение вспомогательной переменной. У нас по-прежнему 407,191 кейсов.

tab _merge

```
                _merge |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
        master only (1) |    244,938       60.15       60.15
            matched (3) |    162,253       39.85      100.00
-----------------------+-----------------------------------
                 Total |    407,191      100.00
```

drop _merge check
Сохраним файл.

*альтернативная команда, сохраняющая только кейсы в мастер-файле:
***merge m:1 id_w idi_child1 using "C:\RLMS_work\seminar_3\data\ind_5_30_for_child_1.dta", assert(master) keep(master)**

16.3. Если вы посмотрите на таблицу данных, вы увидите, что сейчас данные стали отсортированы по переменным-ключам **id_w idi_child1.** Но вы увидите также «пропуски» в данных о первом ребенке: это происходит потому, что на этих детей не была заполнена анкета. Но мы можем восстановить данные о годе рождения и о поле из файла домохозяйств.

* Проверим, для всех ли детей имеются данные о годе рождения. Для этого создадим новую вспомогательную переменную, которая равна 1, если идентификатор ребенка не миссинг, но пропущена информация о годе рождения ребенка

gen check_child1=1 if (idi_child1 != . & ch1_h6 == .)

```
(403,357 missing values generated)
```

tab check_child1, missing

```
check_child |
          1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      3,834        0.94        0.94
          . |    403,357       99.06      100.00
------------+-----------------------------------
      Total |    407,191      100.00
```

Для 3834 кейсов у нас нет данных о годе рождения
Это происходит потому, что для этих кейсов не была заполнена анкета в соответствующем году. Это можно исправить, если учитывать данные из файла домохозяйства - там для каждого члена семьи указывается год рождения, независимо от того, заполнена ли на него анкета. На его основе можно сделать файл со всеми членами семьи и их годом рождения.
fam_all_5_30_members_year_birth.dta

*Откроем такой файл с данными о всех членах семьи.

use "C:\RLMS_work\seminar_3\data\fam_all_5_30_members_year_birth.dta", clear

*и сохраним файл под новым именем

save "C:\RLMS_work\seminar_3\data\IND_5_30_for_child_1_YB.dta"

* сохраним в новом файле нужные переменные

keep id_w idind id_i bn_4 bn_5

*переименуем переменные

rename (idind id_i bn_4 bn_5) (child1_idind idi_child1 ch1_gender ch1_year_birth)

label variable child1_idind "CH1_индивидуальный номер "
label variable ch1_gender "CH1_Какого пола этот человек?"
label variable ch1_year_birth "CH1_В каком году (он/она) родились?"

*сохраним файл

save "C:\RLMS_work\seminar_3\data\IND_5_30_for_child_1_YB.dta", replace

* теперь снова откроем рабочий индивидуальный файл

use "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", clear

* Теперь приклеим год рождения и пол по первому ребенку из этого нового файла

merge m:1 id_w idi_child1 using
"C:\RLMS_work\seminar_3\data\IND_5_30_for_child_1_YB.dta"

```
Result                           # of obs.
    -----------------------------------------
    not matched                       549,079
        from master                   241,104  (_merge==1)
        from using                    307,975  (_merge==2)

    matched                           166,087  (_merge==3)
```

* удалим «лишние» кейсы, приклеенные из внешнего файла

drop if _merge == 2
```
(307,975 observations deleted)
```

drop _merge

Сохраним файл.

* Проверим, для всех ли детей теперь имеются данные о годе рождения

gen check2_child1=1 if (idi_child1 != . & ch1_year_birth == .)
```
(407,191     missing      values generated)
```
Действительно, теперь для всех детей есть информация о годе рождения и поле (у всех кесов в этой переменной – пропущенные значения).

drop check2_child1

Сохраните файл
save "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", replace

*альтернативная команда, сохраняющая только кейсы в мастер-файле:

***merge m:1 id_w idi_child1 using "C:\RLMS_work\seminar_3\data\IND_5_30_for_child_1_YB.dta", assert(master) keep(master)**

16.4. * Приклеивание данных по супругу\супруге.

Проверим, у скольких респондентов есть информация об идентификаторе супруга. *Для этого перекодируем идентификатор в проверочную дихотомическую переменную и посмотрим ее распределение

recode idi_spouse (.=0) (1/max=1), into (check)
tab check

```
  RECODE of |
idi_spouse |
     (ID_I |
     супру |
    га\суп |
     руги) |      Freq.     Percent        Cum.
-----------+-----------------------------------
          0 |    207,470       50.95       50.95
          1 |    199,721       49.05      100.00
-----------+-----------------------------------
      Total |    407,191      100.00
```

Такая информация есть у 199721 респондентов. Приклеим теперь данные по супругу\супруге

merge m:1 id_w idi_spouse using "C:\RLMS_work\seminar_3\data\ind_5_30_for_spouse.dta"

```
    Result                           # of obs.
    -----------------------------------------
    not matched                       426,557
        from master                   213,277  (_merge==1)
        from using                    213,280  (_merge==2)

    matched                           193,914  (_merge==3)
    -----------------------------------------
```

Снова мы видим, что не по всем супругам есть информация – это те люди, которые не заполнили анкету. По ним также можно приклеить год рождения на основании семейного файла, но обычно нас интересуют другие переменные по супругу\супруге (работает ли, заработная плата и т.д.).

* удалим «лишние» кейсы, приклеенные из внешнего файла

drop if _merge == 2
```
(213,280 observations deleted)
```

drop _merge check

Сохраним файл.

save "C:\RLMS_work\seminar_3\data\ind_5_30_main_s3.dta", replace

16.5. * Сравним национальности супругов, для этого перекодируем соответствующие переменные в другие с укрупненными градациями. Перед этим уберем миссинги.

recode i4 s_i4 (99999997 / 99999999 = .)

```
(i4: 7411 changes made)
(s_i4: 1304 changes made)
```

recode i4 (1 =1) (50=1) (70=1) (85=1) (99=1) (102=1) (105=1) (106=1) (112=1) (113=1) (114=1) (115=1) (120=1) (121=1) (124=1) (129=1) (130=1) (131=1) (136=1) (137=1) (138=1) (148=1) (151=1) (152=1) (156=1) (159=1) (174=1) (2=2) (4 = 2) (13 = 2) (127=2) (132=2) (141 =2) (18=3) (22=3) (27=3) (28=3) (29=3) (32=3) (33=3) (34=3) (36=3) (48=3) (56=3) (65=3) (76=3) (82=3) (81=3) (98=3) (95=3) (103=3) (123=3) (166=3) (11=4) (12=4) (14=4) (16=4) (35=4) (39=4) (75=4) (17=4) (9=5) (21=5) (158=5) (170=5) (172= 5) (3=6) (5=6) (6=6) (7=6) (8=6) (10=6) (20=6) (24=6) (25=6) (26=6) (31=6) (37=6) (41=6) (47=6) (51=6) (52=6) (67=6) (68=6) (70=6) (75=6) (86=6) (90=6) (96=6) (107=6) (110=6) (111=6) (119=6) (144=6) (145=6) (155=6) (161=6) (167=6) (171=6) (175=6) (176=6) (19=7) (30=7) (38=7) (40=7) (42=7) (43=7) (49=7) (58=7) (60=7) (64=7) (72=7) (77=7) (80=7) (81=7) (83=3) (87=7) (88=7) (93=7) (95=7) (100=7) (118=7) (123=7) (134=7) (142=7) (154=7) (155=7) (158=7) (164=7) (.=.) (else = 7), into (i4_k)

recode s_i4 (1 =1) (50=1) (70=1) (85=1) (99=1) (102=1) (105=1) (106=1) (112=1) (113=1) (114=1) (115=1) (120=1) (121=1) (124=1) (129=1) (130=1) (131=1) (136=1) (137=1) (138=1) (148=1) (151=1) (152=1) (156=1) (159=1) (174=1) (2=2) (4 = 2) (13 = 2) (127=2) (132=2) (141 =2) (18=3) (22=3) (27=3) (28=3) (29=3) (32=3) (33=3) (34=3) (36=3) (48=3) (56=3) (65=3) (76=3) (82=3) (81=3) (98=3) (95=3) (103=3) (123=3) (166=3) (11=4) (12=4) (14=4) (16=4) (35=4) (39=4) (75=4) (17=4) (9=5) (21=5) (158=5) (170=5) (172= 5) (3=6) (5=6) (6=6) (7=6) (8=6) (10=6) (20=6) (24=6) (25=6) (26=6) (31=6) (37=6) (41=6) (47=6) (51=6) (52=6) (67=6) (68=6) (70=6) (75=6) (86=6) (90=6) (96=6) (107=6) (110=6) (111=6) (119=6) (144=6) (145=6) (155=6) (161=6) (167=6) (171=6) (175=6) (176=6) (19=7) (30=7) (38=7) (40=7) (42=7) (43=7) (49=7) (58=7) (60=7) (64=7) (72=7) (77=7) (80=7) (81=7) (83=3) (87=7) (88=7) (93=7) (95=7) (100=7) (118=7) (123=7) (134=7) (142=7) (154=7) (155=7) (158=7) (164=7) (.=.) (else = 7), into (s_i4_k)

label variable i4_k "группы национальностей кратко"
label variable s_i4_k "S_группы национальностей кратко"
label define ethnicity 1 "русские, смешанные русские" 2 "украинцы, беларусы, молдаване" 3 "народы Сев.Кавказа" 4 "народы Поволжья и Севера" 5 "татары, башкиры" 6 "прочие европейские" 7 "прочие неевропейские"
label values i4_k ethnicity
label values s_i4_k ethnicity

Сохраните файл

*Посмотрим распределение

tab i4_k

```
          RECODE of i4 (Кем Вы |
                 себя считаете |
                           по |
                 национальности? |
                           Я им) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
   русские, смешанные русские |    283,853       86.14       86.14
украинцы, беларусы, молдаване |      5,857        1.78       87.92
            народы Сев.Кавказа |     13,202        4.01       91.92
     народы Поволжья и Севера |     10,992        3.34       95.26
               татары, башкиры |      8,831        2.68       97.94
            прочие европейские |      5,407        1.64       99.58
          прочие неевропейские |      1,386        0.42      100.00
--------------------------------+-----------------------------------
                          Total |    329,528      100.00
```

*Сделаем кросс-таблицу национальностей мужа и жены (отобрав для i4_k только мужчин, чтобы не было повторов)

*Сравним национальности супругов

tab i4_k s_i4_k if h5 == 1

```
   RECODE of i4 (Кем |
             Вы себя |
          считаете по |              RECODE of s_i4 (S_Кем Вы себя считаете по
           национальн |                        национальности? Я им)
          ости? Я им) | русские,   украинцы,  народы Се  народы По  татары, б  прочие ев |     Total
----------------------+----------------------------------------------------------------------+----------
русские, смешанные ру |    77,392      1,156        114      1,363        725        486 |    81,381
украинцы, беларусы, м |     1,649        338          2         85         29         28 |     2,133
   народы Сев.Кавказа |       425          1      3,288          0          5         22 |     3,744
народы Поволжья и Сев |       978         36          0      2,153         69          3 |     3,244
      татары, башкиры |     1,083         15          7         51      1,606          4 |     2,768
   прочие европейские |     1,056         43         25         32         48        810 |     2,014
 прочие неевропейские |       284          0          1         15         18          1 |       520
----------------------+----------------------------------------------------------------------+----------
                Total |    82,867      1,589      3,437      3,699      2,500      1,354 |    95,804
```

```
                      | RECODE of
                      |    s_i4
                      |   (S_Кем
                      |     Вы
                      |    себя
                      |   счита
                      |    ете
                      |     по
    RECODE of i4 (Кем |   нацио
             Вы себя |   нальн
          считаете по |   ости?
           национальн |   Я им)
          ости? Я им) | прочие не |     Total
----------------------+-----------+----------
русские, смешанные ру |       145 |    81,381
украинцы, беларусы, м |         2 |     2,133
   народы Сев.Кавказа |         3 |     3,744
народы Поволжья и Сев |         5 |     3,244
      татары, башкиры |         2 |     2,768
   прочие европейские |         0 |     2,014
 прочие неевропейские |       201 |       520
----------------------+-----------+----------
                Total |       358 |    95,804
```

Как интерпретировать эту таблицу с точки зрения теории брака Г.Беккера?

16.6. *Приклеим данные по матери

merge m:1 id_w idi_mother using "C:\RLMS_work\seminar_3\data\ind_5_30_for_mother.dta"

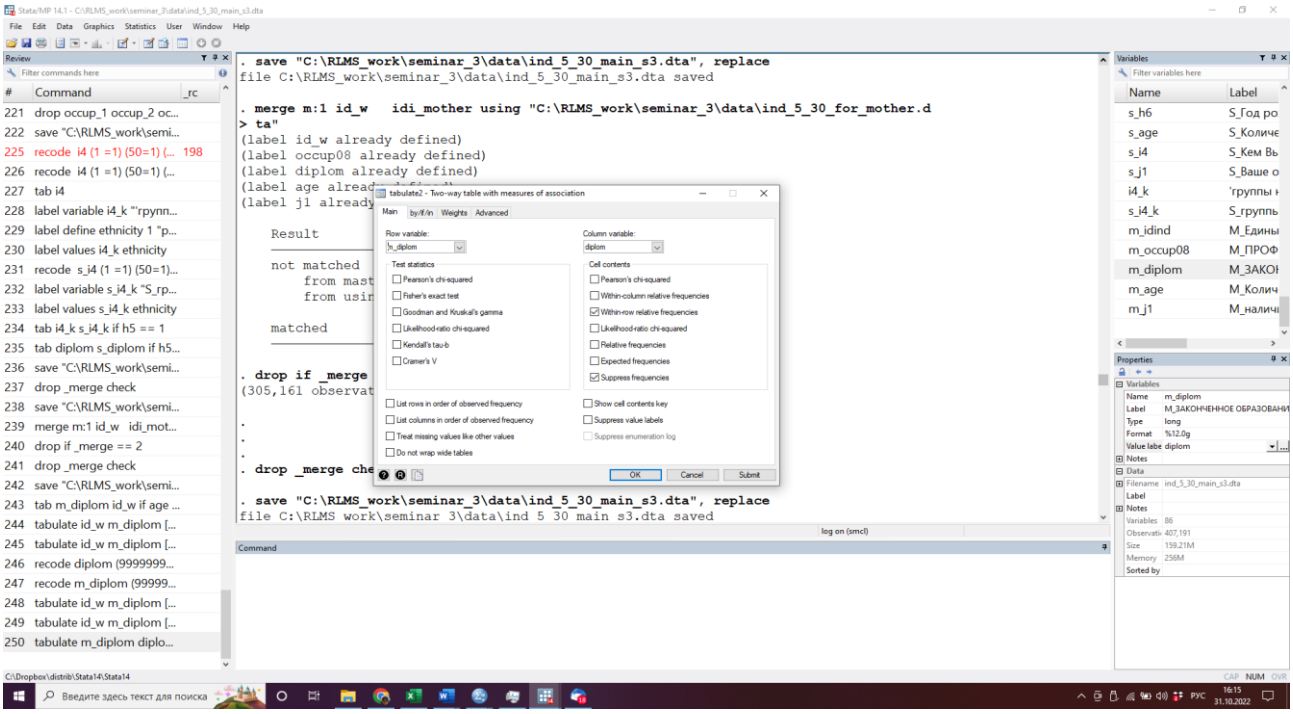
```
    Result                           # of obs.
    -----------------------------------------
    not matched                        567,043
        from master                    261,882  (_merge==1)
        from using                     305,161  (_merge==2)

    matched                            145,309  (_merge==3)
    -----------------------------------------
```

drop if _merge == 2

```
(305,161 observations deleted)
```

drop _merge

*Посмотрим, как изменялось образование матерей за все время обследования. Взвешенные данные, годы по строкам, горизонтальный %.



*Перед этим перекодируем «9999999» в миссинг.

recode diplom (99999997/99999999 = .)

```
(diplom: 588 changes made)
```

recode s_diplom (99999997 / 99999999 =.)

```
(s_diplom: 272 changes made)
```

recode m_diplom (99999997 / 99999999 =.)

```
(m_diplom: 184 changes made)
```

tabulate id_w m_diplom [aweight = inwgt] if age <= 15, nofreq row

НОМЕР ВОЛНЫ	M_ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА) окончил 0-6 кл	незаконч ср.О.(7-8)	незаконч сред +	законч. среднее	законченн ср.спец.	законченн высшее	Total
1994 год	0.23	4.79	3.60	40.97	30.74	19.67	100.00
1995 год	0.25	3.33	2.86	45.07	30.58	17.91	100.00
1996 год	0.14	3.27	2.89	44.94	30.55	18.22	100.00
1998 год	0.16	3.35	2.79	43.81	30.62	19.28	100.00
2000 год	0.34	2.66	3.03	41.63	31.37	20.96	100.00
2001 год	0.12	1.97	3.65	44.00	29.52	20.74	100.00
2002 год	0.27	2.48	3.63	40.24	31.57	21.81	100.00
2003 год	0.07	3.01	3.65	39.78	29.85	23.64	100.00
2004 год	0.67	2.54	4.81	39.64	29.86	22.47	100.00
2005 год	0.70	3.46	5.23	38.94	28.08	23.60	100.00
2006 год	0.00	3.40	8.39	33.29	29.72	25.21	100.00
2007 год	0.07	2.89	8.69	32.74	27.94	27.67	100.00
2008 год	0.00	3.02	8.44	30.98	29.67	27.89	100.00
2009 год	0.00	2.76	9.40	29.98	27.13	30.72	100.00
2010 год	0.10	1.92	9.82	30.95	26.22	30.98	100.00
2011 год	0.32	1.76	9.02	30.59	25.02	33.29	100.00
2012 год	0.37	1.88	14.78	26.39	24.25	32.34	100.00
2013 год	0.34	1.37	10.11	29.21	24.31	34.66	100.00
2014 год	0.51	1.30	10.46	29.14	23.04	35.56	100.00
2015 год	0.10	1.20	10.33	27.48	23.72	37.17	100.00
2016 год	0.47	1.19	9.80	27.10	21.32	40.13	100.00
2017 год	0.49	0.99	8.91	28.27	21.13	40.21	100.00
2018 год	0.44	0.63	9.51	26.99	22.08	40.35	100.00
2019 год	0.44	1.20	9.32	27.48	21.15	40.41	100.00
2020 год	0.39	1.44	10.16	25.85	22.05	40.12	100.00
2021 год	0.34	1.12	9.64	26.23	22.30	40.37	100.00
Total	0.27	2.29	7.37	34.17	26.76	29.14	100.00

Проинтерпретируйте результаты.
Теперь сравним образование матерей и их детей в возрасте от 25 лет и старше (разумеется, это касается только тех «детей», которые продолжают жить с родителями); ограничение по возрасту – так как к этому возрасту большинство уже получают профессиональное образование. Образование матери – по строкам, детей – в колонки, горизонтальный %.

tabulate m_diplom diplom [aweight = inwgt] if age >=25, nofreq row

M_ЗАКОНЧЕНН ОЕ ОБРАЗОВАНИ Е (ГРУППА)	ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА) окончил 0 0-6 кл	незакончен. ср.О.(7-8)	Незаконч. сред +	Закончен. среднее	Закончен. ср.спец.	Закончен. высшее	Total
окончил 0 - 6 классов	4.74	10.49	6.13	48.72	20.42	9.50	100.00
незак. среднее 7-8 кл	0.82	7.43	9.23	45.41	23.38	13.73	100.00
незак. Сред. + что-то	0.12	4.53	23.74	34.44	21.56	15.62	100.00
законч. среднее общее	0.44	2.19	10.25	44.64	20.00	22.49	100.00
законч. среднее спец.	0.26	1.26	6.63	27.17	27.47	37.21	100.00
законч. высшее обр.	0.30	0.51	2.39	15.31	20.18	61.30	100.00
Total	0.86	3.29	8.04	35.04	22.52	30.25	100.00

Проинтерпретируйте результаты.

16.7. Сохраните do-файл (файл кода) под другим названием (семинар 3 часть 2 фамилия). Закройте лог-файл (аутпута).
Оба эти файла – ваш отчет за вторую половину семинара 3.

16.8. Самостоятельное задание (домашнее)

В файле отдельного аутпута для самостоятельного задания наберите команду:
***Фамилия – номер семинара – номер задания**

- Аналогично преобразованиям в файле матери, сделайте преобразования в файле отца **ind_5_30_for_father.dta**
- приклейте данные по отцу к основному рабочему файлу
- посмотрите, как изменялось образование отцов детей до 16 лет (от 0 до 15) за все время обследования. Взвешенные данные, годы по строкам, горидонтальный %. Перед этим надо перекодировать в переменной «образование отца» 99999997/9999999 в миссинг.
- Сравните образование отцов и их детей в возрасте от 25 лет и старше. Образование отца – по строкам, детей – в колонки, горизонтальный %
- Команды и результаты из окна аутупта скопируйте и вставьте в текстовый файл вашего отчета за задания к семинару.