Занятие 0.

1. Сайт «Российского мониторинга экономического положения и здоровья населения НИУ ВШЭ» - Этот раздел для самостоятельного изучения ДО первого практического занятия!!!

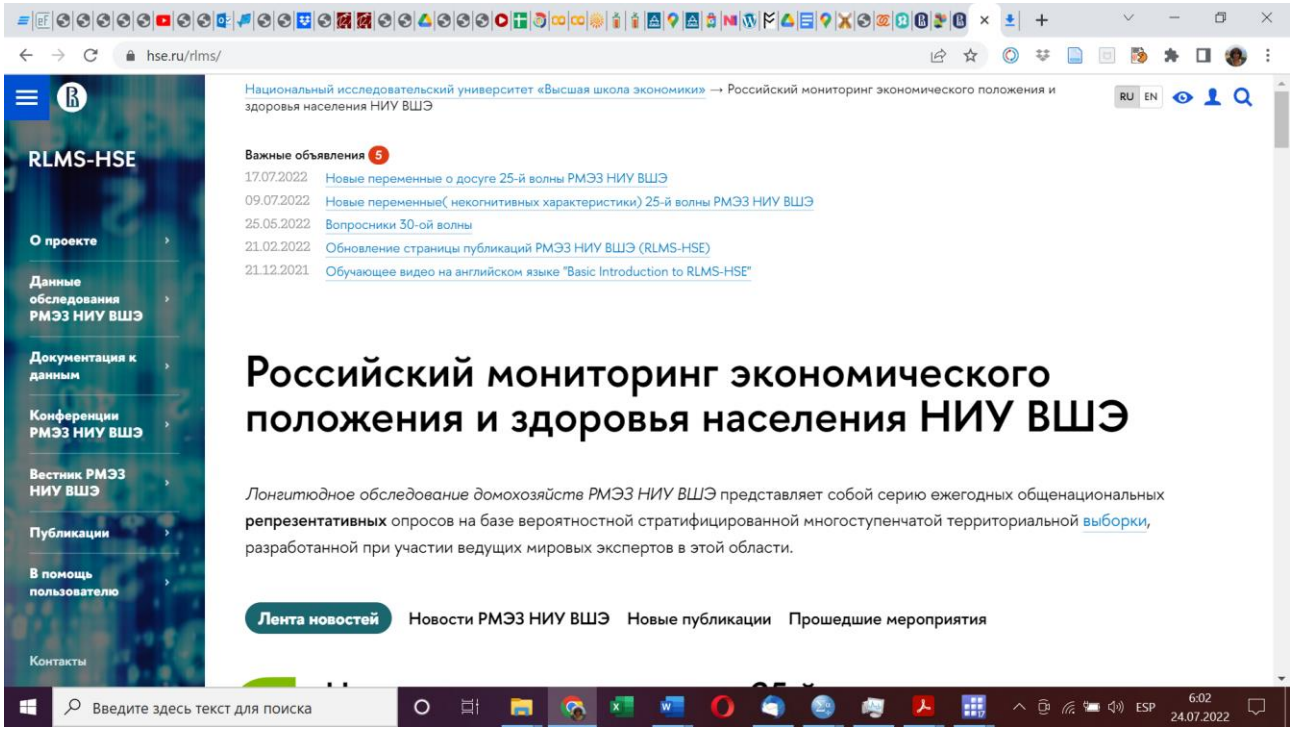
1.1. Основные разделы сайта.

В НИУ ВШЭ сбором, обработкой и первичным анализом данных РМЭЗ НИУ ВШЭ (RLMS HSE) занимается Центр лонгитюдных обследований (ЦЛО) при Институте социальной политики, заведующий – Козырева П.М. Сайт центра:

https://www.hse.ru/longitude/

Сайт «Российского мониторинга экономического положения и здоровья населения НИУ ВШЭ» (РМЭЗ НИУ ВШЭ или RLMS HSE)

https://www.hse.ru/rlms/



Рекомендую вам внимательно прочитать разделы сайта «О проекте» (описание проекта, модель выборки, график волн, и как ссылаться на данные при публикациях) и «Документация к данным» (Структура данных, Вопросники, Описание переменных, Идентификационные переменные, Кодификаторы).

В работе с данными также может помочь раздел «В помощь пользователю».

1.2. Вопросники и Codebooks

На странице https://www.hse.ru/rlms/question Вы можете скачать вопросники для каждой волны (индивидуальный взрослый – начиная с 14 лет; индивидуальный детский – заполняемый одним из родителей; домохозяйственный или семейный).

Каждый вопросник разбит на тематические блоки, названные буквами латинского алфавита. Внутри блока вопросы имеют внутреннюю нумерацию. Имейте, пожалуйста, ввиду, что номера вопросов в анкете и в файле могут различаться, так как в анкете нумерация делается «для респондента», а в файлах одним и тем же вопросам в разных волнах дается один и тот же номер (для возможности сопоставления и склеивания). Кроме того, данные некоторых вопросов за последние 3-4 года могут быть временно скрыты для их анализа разработчиками (они открываются через три года после первой публикации данных).

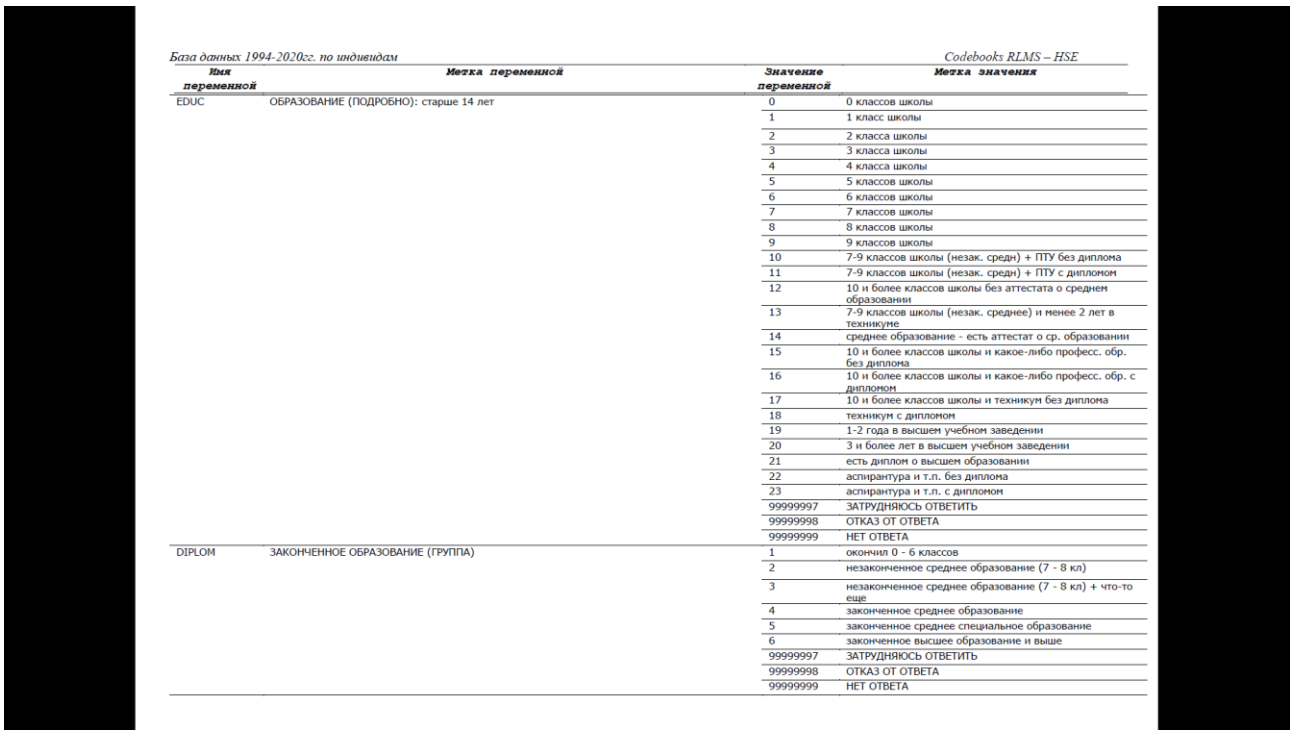
На странице https://www.hse.ru/rlms/code Вы можете скачать файлы с описанием переменных, представленных в базах данных. Так называемые Codebooks содержат информацию об имени переменных, их метках, значениях и метках значений. Файлы соответствуют файлам данных в форматах IBM SPSS Statistics и STATA. Это можно сделать для Объединенные базы данных (все волны представлены в одном файле), для отдельных волн, а также Единый коудбук в формате SPSS.

Имена всех переменных в любом файле сформированы по принципу: первая буква – номер волны (5- a, 6 – b, 7- c, 8 – d, 9- e, 10 – f, 11 – g, 12 – h, 13 – i, 14 – j, 15 - k и т.д.), вторая буква – номер раздела анкеты (a, b, c, d, e, f, h, i, j, l, m, n, o, k), цифры – номер вопроса в разделе. Так как вопросники изменялись и были добавлены новые вопросы, номера могут быть не только целыми (например, в разделе J индивидуальной анкеты есть вопрос 1: «Ваше основное занятие в настоящее время?», а также вопрос 1.1.1 : «Насколько Вы удовлетворены или не удовлетворены Вашей работой в целом?»). Имена переменных в фалах SPSS и STATA могут различаться в силу того, что в STATA в именах переменных запрещен символ точки; поэтому часто вместо точки в именах переменных формата SPSS используется нижнее подчеркивание.

ПРИМЕР:

Имя переменной, соответствующей вопросу 1.1.1 раздела J (Насколько Вы удовлетворены или не удовлетворены Вашей работой в целом?) волны 29 (2020 год) в формате SPSS выглядит как yj1.1.1, а в формате STATA - yj1_1_1. Первая буква имени – y – указывает на номер волны (29я). Этот же вопрос в объединенной базе данных (за все волны) имеет имя J1.1.1 в формате SPSS (первая буква – номер раздела, а не номер волны) и J1_1_1 в формате STATA.

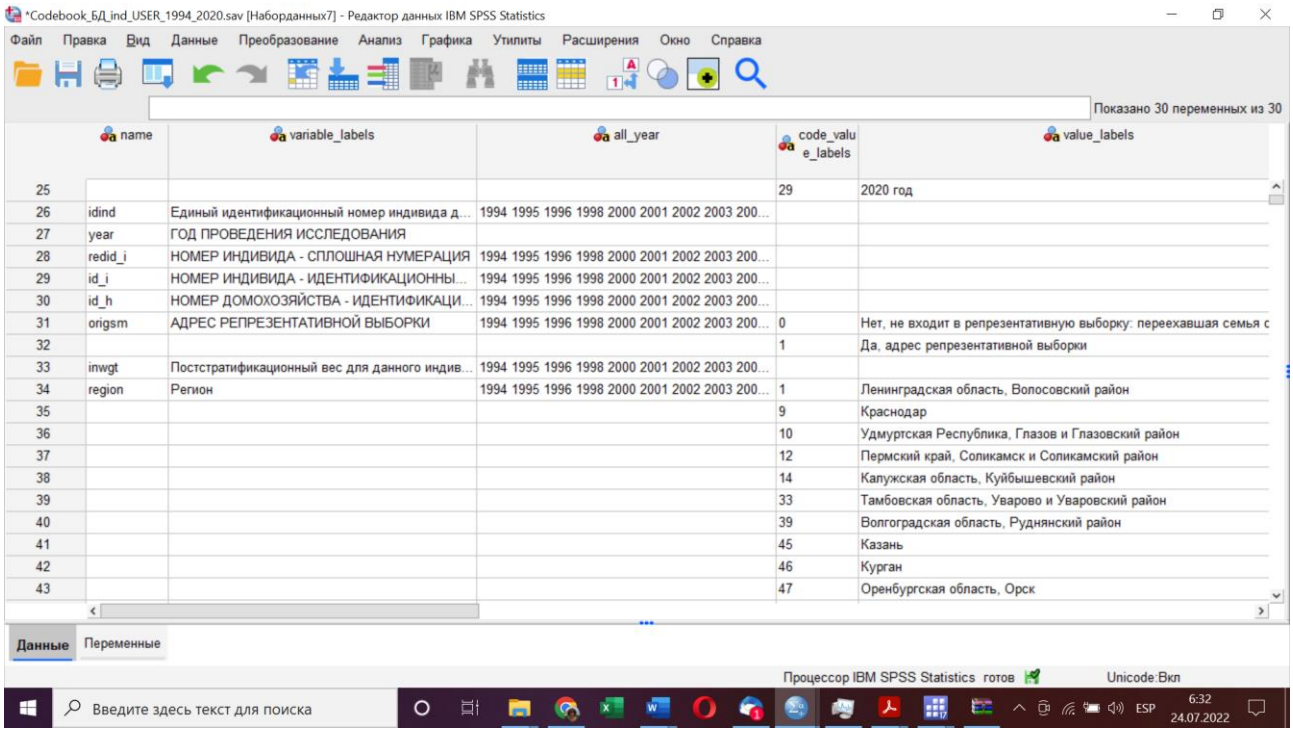
Codebook для файла отдельной волны и для файла объединенных баз данных выглядят примерно так (имена переменных в нем соответствуют формату SPSS). В Codebook для файла отдельной волны и для файла объединенных баз данных для одного и того же вопроса может различаться регистр. Это не важно для SPSS, но очень важно для STATA (в которой большие и малые буквы в именах – это разные символы).



База данных 1994-2020гг. по индивидам *Codebooks RLMS – HSE*

Имя переменной	Метка переменной	Значение переменной	Метка значения
EDUC	ОБРАЗОВАНИЕ (ПОДРОБНО): старше 14 лет	0	0 классов школы
		1	1 класс школы
		2	2 класса школы
		3	3 класса школы
		4	4 класса школы
		5	5 классов школы
		6	6 классов школы
		7	7 классов школы
		8	8 классов школы
		9	9 классов школы
		10	7-9 классов школы (незак. средн) + ПТУ без диплома
		11	7-9 классов школы (незак. средн) + ПТУ с дипломом
		12	10 и более классов школы без аттестата о среднем образовании
		13	7-9 классов школы (незак. среднее) и менее 2 лет в техникуме
		14	среднее образование - есть аттестат о ср. образовании
		15	10 и более классов школы и какое-либо професс. обр. без диплома
		16	10 и более классов школы и какое-либо професс. обр. с дипломом
		17	10 и более классов школы и техникум без диплома
		18	техникум с дипломом
		19	1-2 года в высшем учебном заведении
		20	3 и более лет в высшем учебном заведении
		21	есть диплом о высшем образовании
		22	аспирантура и т.п. без диплома
		23	аспирантура и т.п. с дипломом
		99999997	ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
		99999998	ОТКАЗ ОТ ОТВЕТА
		99999999	НЕТ ОТВЕТА
DIPLOM	ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА)	1	окончил 0 - 6 классов
		2	незаконченное среднее образование (7 - 8 кл)
		3	незаконченное среднее образование (7 - 8 кл) + что-то еще
		4	законченное среднее образование
		5	законченное среднее специальное образование
		6	законченное высшее образование и выше
		99999997	ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
		99999998	ОТКАЗ ОТ ОТВЕТА
		99999999	НЕТ ОТВЕТА

Единый коудбук в формате SPSS представляет собой файл с описанием всех переменных файлов данных по индивидам волн 5- 29 (с указанием волн). В этом файле вы можете увидеть, в каких волнах какие вопросы задавались. Следует иметь ввиду, что в некоторых вопросах с течением времени изменились варианты ответов. Например, в вопросе (x)j72.5a - «Вы учились или учитесь в институте, университете, академии?» в 1995-2005 гг. были варианты ответов «учился или учусь» (1) и «нет» (2), а начиная с 2006 г. – «учились» (1), «учитесь» (2), «нет» (3). Для сопоставимости данных разных волн, в волнах за 1995-2006 год ответы были перекодированы: «учился или учусь» (4) и «нет» (3), а в объединенном за все волны файле вы увидите все 4 варианта ответа: «учились» (1), «учитесь» (2), «нет» (3), «учился или учусь» (4). В анкеты прошлых лет были внесены изменения соответствующих значений.



1.3. Файлы сводных данных

На странице https://www.hse.ru/rlms/spss2 можно скачать файлы сводных данных, созданные аналитиками ЦЛО на базе основных данных, для облегчения работы с массивом РМЭЗ НИУ ВШЭ.

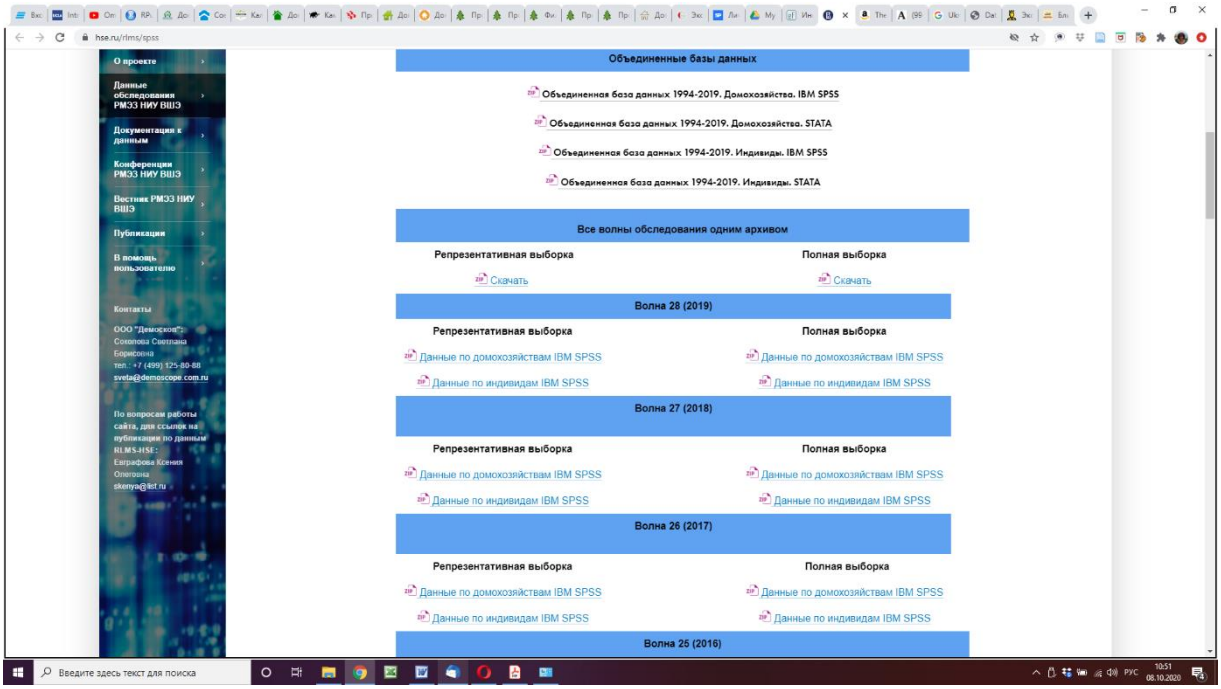
- Файл с перечнем всех респондентов, хотя бы раз опрошенных по индивидуальному вопроснику, с указанием для каждого респондента его уникального номера индивида и идентификационных номеров в каждой волне: Идентификационные номера индивидов 1994-2020 гг. (Для IBM SPSS)
- Файл с указанием уникального номера первоначально опрошенного домохозяйства для каждого домохозяйства, участвовавшего в РМЭЗ НИУ ВШЭ, и идентификационных номеров каждого домохозяйства в каждой волне: Идентификационные номера домохозяйств 1994-2019 гг. (Для IBM SPSS)
- Файл с указанием идентификационных номеров каждого родственника для каждого респондента в каждую волну, и пример использования этого файла: Идентификационные номера родственников 1994 - 20120гг. (Для IBM SPSS)
- Пример работы с файлом 'Идентификационные номера родственников' (SPS)
- Файл сконструированных переменных по доходам и расходам домохозяйства: Доходы и расходы: индексация. 1994-2020 гг. Для IBM SPSS Statistics; Доходы и расходы: индексация. 1994-2020 гг. Для STATA)

Пожалуйста, имейте ввиду, что любой файл формата IBM SPSS Вы легко можете

перевести в формат STATA, открыв этот файл в IBM SPSS Statistics и используя функцию «сохранить как» (выбрав формат STATA). О некоторых из этих файлов мы поговорим позже на наших семинарах.

1.4. Данные обследований и различия в файлах

На странице https://www.hse.ru/rlms/spss можно скачать данные РМЭЗ ВШЭ: данные за отдельные годы (панельная или репрезентативная выборка), данные за все годы одним архивом, или лонгитюдные данные (склеенные за все годы данные). Данные доступны в форматах SPSS или STATA (начиная с 25 волны, 2016 год, а также объединенные базы данных за 1994-2020 или 1994-2021, новые данные обычно появляются в сентябре-октябре следующего года).



Обратите внимание, что выборка RLMS-HSE представляет собой «повторяющуюся выборку» (repeated sample), с «разделяющейся панелью» (split-panel). Поэтому данные обследования представлены в двух типах файлов, различающихся по количеству наблюдений. В файлах под названием **«Репрезентативная выборка»** содержатся наблюдения по домохозяйствам/индивидам, которые в каждой отдельной волне репрезентируют население России. В файлах под названием **«Полная выборка»** содержатся наблюдения по всем домохозяйствам/индивидам, опрошенным в рамках данной волны. То есть, помимо наблюдений репрезентативной выборки, в этих файлах находятся данные по части «разделяющейся панели», которые не входят в состав репрезентативной выборки, но входят в общую панельную выборку.

Также напоминаем Вам, что при использовании данных RLMS-HSE в публичных целях **ссылка на источник** должна быть следующей:

«Российский мониторинг экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE)», проводимый Национальным исследовательским университетом "Высшая школа экономики" и ООО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии Федерального научно-исследовательского социологического центра РАН. (Сайты обследования RLMS-HSE: https://rlms-hse.cpc.unc.edu и http://www.hse.ru/rlms)».

Как вы видите, есть разные файлы, соответствующие данным по домохозяйствам (семейная анкета) и данным по индивидам. В файле по индивидам содержатся данные как по

детям до 13 лет, опрошенным по детской анкете, так и по взрослым (начиная с 14 лет), опрошенным по взрослой анкете.

Кроме того, в двух разных «колонках» содержатся данные «полной выборки» или «репрезентативной выборки». В чем разница? В «репрезентативной выборке» содержатся только те кейсы, которые входят в ежегодную репрезентативную выборку адресов РМЭЗ. В силу усыхания панельной выборки, каждый год из этой выборки выбывает некоторое количество адресов (например, люди переехали и т.д.), которые заменяются новыми из резервного списка. Для лучшего соответствия генеральной совокупности есть также возможность использовать взвешивание. Файлы для репрезентативных "непанельных" данных (где не отслеживаются переехавшие домохозяйства) можно получить отдельно, либо создать выбором из панельных данных по переменной "адрес в первоначальной выборке". Эти данные репрезентируют население России. Сравнительный непанельный анализ между волнами (cross-sectional analysis) должен осуществляться на массиве только этих семей и индивидов.

Кроме того, в каждой последующей волне исследователи также старались найти всех людей, участвовавших в исследовании ранее - и когда находили переехавших, то опрашивали их по их новым адресам (поиск осуществлялся в пределах только одного и того же населенного пункта). В «полной выборке», помимо адресов из репрезентативной выборки, есть также адреса семей, которые были в предыдущих волнах исследования, но переехали в пределах того же населенного пункта, разделились и т.д. (т.е. если хотя бы один член такого домохозяйства переехал). «Полные» данные содержат всех опрошенных индивидов (как проживающих по адресам выборки 1994 года, так и тех, кто хотя бы один раз ранее был опрошен по адресу выборки 1994 года, а в данной волне переехал на другой адрес и был опрошен по этому новому адресу). Это массивы данных для панельного анализа (панельные регрессии, многоуровневые регрессии, регрессии со смешанными эффектами) индивидов - взрослых и детей, и в более редких случаях – домохозяйств (причины того, что панельный анализ реже используется для домохозяйств, см. в лекциях). Полную выборку (лонгитюдные или панельные данные) необходимо использовать также в случаях, где необходимы данные с «лагом» (например, за будущую или предыдущую волну).

Поэтому «полная выборка», строго говоря, не репрезентативна. Но из нее всегда легко получить репрезентативную при помощи фильтра по специальной переменной, или используя взвешивание по специальной переменной. Я рекомендую всегда скачивать полную выборку, так как из нее легко сделать репрезентативную, но не наоборот.

Таким образом, вы можете скачать отдельные файлы за нужные вам годы, выбирая индивидуальные или домохозяйственные данные.

Файлы за отдельные годы

Раунд		Файлы (панельные)	
	Год	Анкета домохозяйства	Индивид. анкета
5	1994	r05hall41.sav	r05iall_42.sav
6	1995	r06hall41.sav	r06iall_42.sav
		….	
27	2018	r27hall41.sav r27hall41. dta	r27iall_42.sav r27iall_42. dta
28	2019	r28hall41.sav	r28iall_42.sav
29	2020	r29hall41.sav r29hall41.dta	r29iall_42.sav r29iall_42.dta
30	2021	…	…

Так как команда РМЭЗ постоянно совершенствует данные (даже за прошлые годы), у каждого файла есть «версия», которая содержится в цифре перед точкой в расширением (например, за 2020 год у индивидуального файла – версия 42, а у домохозяйстенного – 41).

Скачанные вами версии могут несколько отличаться от тех, которые мы будем использовать на занятиях (в зависимости от времени их скачивания), но это не критично.

Помимо того, что вы можете скачивать отельные файлы за нужные вам годы, вы можете скачать «Все волны обследования одним архивом (IBM SPSS Statistics)», то есть вы получите отдельные файлы за все годы обследования (по вашему выбору, индивидуальные или семейные). Это будет файл «Полная_выборка_10.10.2021.zip», вы скачаете одновременно файлы индивидов и домохозяйств.

Есть также еще одна возможность, скачать так называемые «Объединенные базы данных» - отдельно для индивидов или отдельно для домохозяйств, в формате либо SPSS, либо STATA.

RLMS_HH_1994_2021_rus_v4_dta.dta – данные по домохозяйствам

RLMS_IND_1994_2021_2022_08_21_1_v3_rus.dta - данные по индивидам.

Эти файлы содержат склеенные «по вертикали» данные за все годы, то есть сначала идут, например, все индивиды (или домохозяйства), опрошенные в 1994 году, потом в 1995, и т.д. вплоть до 2021. Понятно, что таким образом многие индивиды (и домохозяйства) принимают участие в опросе несколько раз. Это требует специальных методов расчета стандартной ошибки в анализе данных (например, в регрессиях, где используются данные за несколько лет, но не панельные регрессии).