Discussion Paper Series

University of Tokyo
Institute of Social Science
Panel Survey

東京大学社会科学研究所　パネル調査プロジェクト
ディスカッションペーパーシリーズ

東大社研・若年壮年パネル調査における横断ウェイトの作成

Constructing cross-sectional sample weights

in the Japanese Life Course Panel Survey Youth and Middle-aged

小川和孝
（東北大学）

Katsunori OGAWA

August 2020

No.122

東京大学社会科学研究所
INSTITUTE OF SOCIAL SCIENCE　UNIVERSITY OF TOKYO

東京大学社会科学研究所パネル調査プロジェクト
ディスカッションペーパーシリーズ No.122
2020 年 8 月

東大社研・若年壮年パネル調査における横断ウェイトの作成

小川和孝（東北大学）

要約

本論文の目的は、東京大学社会科学研究所が実施する「若年・壮年パネル調査」（以下、JLPS-YM）における Wave1 調査の横断ウェイトについて検討することである。ウェイトの作成にあたっては、性別年齢階級、学歴、就業状態、配偶状態、居住地域を使用し、これらの周辺分布が母集団を表すとみなしたベンチマーク統計の分布に一致するように、raking と呼ばれる手法によって反復調整を行った。結果として、これらの変数に関しては適切な補正を行うウェイトが作成できた。しかし、ウェイトの作成に使用をしていない労働時間と年間世帯収入に注目した場合には、適切に分布を補正できているとは必ずしも言えないことが示された。

謝辞

本研究は、科学研究費補助金基盤研究（S）（18103003, 22223005）、特別推進研究事業（25000001, 18H05204）の助成を受けたものである。東京大学社会科学研究所パネル調査の実施にあたっては、社会科学研究所研究資金、株式会社アウトソーシングからの奨学寄付金を受けた。パネル調査データの使用にあたっては社会科学研究所パネル調査運営委員会の許可を受けた。

1. 問題設定

本論文の目的は、東京大学社会科学研究所が実施する「若年・壮年パネル調査」（以下、JLPS-YM）における Wave1 調査の横断ウェイトについて検討することである。既存の様々なパネル調査と同様に、同一対象を追跡するという特性上、JLPS-YM にも脱落によるサンプルサイズの減少が生じている。もし脱落が特定の社会経済的属性に関して不均等に起きる場合に、誤った推論を起こしてしまう危険性がある。また、パネル調査に関しては Wave1 時点での非回答と、その後の脱落を区別して補正の議論を行う必要がある。

本論文の以下では、標本調査おける社会調査における一般的な事後層化ウェイト作成の方法や、既存の社会調査におけるウェイトの作成状況についても触れつつ、JLPS-YM の Wave1 調査における横断的ウェイトの作成を検討する。

2. 分析の焦点化

2.1 標本調査における事後層化ウェイト

標本調査において、回答拒否などの理由によって母集団との構成比率が異なる場合の補正の一つとして、事後層化（poststratification）によるウェイトが知られている（Cochran 1977; Holt and Smith 1979）。事後層化の主な手法の一つは、センサスなどによって母集団の下位集団（性別・年代など）の情報が利用可能な場合に、これに標本の（同時）分布が一致するようにウェイトを作成するというものである。

表 1　仮想的な事例（年齢を考慮した事後層化）

層j	母集団P_j	標本p_j	ウェイトw_j
20 代	0.13	0.08	1.625
30 代	0.15	0.11	1.364
40 代	0.16	0.14	1.143
50 代	0.18	0.20	0.9
60 代	0.20	0.25	0.8
70 代	0.18	0.22	0.818
計	1.00	1.00	

母集団および標本における層（グループ）jの比率を、それぞれP_jおよびp_jとするとき、それぞれの層に対して割り当てられるウェイトは、$w_j = P_j/p_j$となる。たとえば、表 1 の

ような仮想的な事例を考える。20 代〜70 代までの異なる年代の人々に関して、母集団と標本では構成比率が異なっているとする。具体的には、社会調査ではしばしば見られるように、若年者の標本比率が母集団比率にくらべて過小となっている。

上記の仮想データにおいては、若年者には 1 より大きいウェイトが、高齢者には 1 より小さいウェイトが与えられる。たとえば 20 代の回答者に対しては、$w_j = 1.625$倍に重みづけをして集計が行われる。

事後層化ウェイトを作成する際には各層内において、それぞれの個人が等確率で標本に含まれうるという仮定が満たされることが望ましい（Gelman and Carlin 2002）。このためには、なるべく層化に使用する変数を増やした方がよい。しかし、他方で層化変数を多くするほど、各層内の標本サイズが小さくなり、ゼロに近くなる層が多くなってしまう（標本サイズがゼロの層に対してはウェイトの計算・適用ができなくなる）。

このため実際には、事後層化ウェイトを作成する際には比較的少数の重要と判断される変数を選択するのが一般的である。たとえば、基本的な社会人口学的変数である、性別・年齢・人種・地域などである。

2.2 既存の社会調査における事後層化ウェイト

これまでの日本の社会調査においても、これを補正するためのウェイトが構築されてきている。たとえば、「日本版総合社会調査」（JGSS）では、2000 年から 2005 年調査までは、「地域ブロック（6 区分）×市郡（2 区分）×男女（2 区分）」を考慮した、計 144 区分がウェイトの算出に用いられている。しかし、平成の大合併のために多くの郡が市になったことにより、市郡の区別を行うことの意味が低下したとともに、回答者が少ない区分においてウェイトの値が安定しない問題があったという。そのため、JGSS-2006 以降では「男女（2 区分）×年齢階級（7 区分）」の 14 区分が用いられるようになっている（松井 2006）。また、このウェイトを用いることで、回収率の低い男性・若年層により大きいウェイトが置かれることになるという。

「社会階層と社会移動調査」（SSM 調査）の 2005 年調査では、2005 年国勢調査に基づき、「地域（6 区分）×市郡（2 区分）×男女（2 区分）×年齢（5 区分）」の計 120 区分がウェイト算出に用いられている。

クロスセクション調査では適切な変数が選択されていれば事後層化ウェイトは 1 つあれば十分であるのに対して、同一対象を追跡するパネル調査においては、第 2 波以降の調査における脱落を補正するための縦断ウェイトを作成が必要になることがある。さらにパネル調査においては、初期サンプルの脱落によるサンプルサイズの減少を補うために追加調査が行われることがある。追加サンプルと、初期サンプルを同時に集計するためにはさら

に、統合ウェイトと呼ばれるものが作成されることがある。石井・野崎（2014）は「慶應義塾家計パネル調査」（KHPS）・「日本家計パネル調査」（JHPS）における横断・縦断・統合ウェイトの作成を行っている。

3. JLPS-YM のデザイン

JLPS-YM では当初、20～34 歳（若年調査）と 35～40 歳（壮年調査）の人々を母集団とし、地域・都市規模・性別・年齢によって層化した上でサンプルが抽出されている。Wave1 調査は 2007 年 1 月から 4 月にかけて行われた。Wave1 調査は 2007 年 1 月から 4 月にかけて行われており、有効アタック数に対する回収率は 36.4%である（石田 2015）。

三輪（2008）によると、JLPS-YM の Wave1 調査サンプルには就業者が過大に含まれており、同時期の「労働力調査」の値をおよそ 7 ポイント上回っている。ただし調査は事前に継続的に行わることに同意してもらった人々を対象としていることにくわえ、JLPS が対象とする年齢層における回収率の一般的な低さなども考慮する必要があるとされている。また、類似した設計のパネル調査とくらべても、JLPS-YM の偏りが特段大きいわけではないとも述べられている。

4. 横断ウェイト作成の手続き

4.1 ウェイト作成における一般的手続き

Watson（2012）は事後層化ウェイト作成に関して、以下のような一般的な手続きを挙げている。第一に、母集団の中で対象とするサンプル単位を特定する。第二に、選択確率の逆数として初期ウェイトを算出する。第三に、回答傾向が同質的なグループを特定するか、あるいは回答確率をモデル化することで非回答の調整をする。そして第四に、ウェイト付けされた推定値が（センサス統計などから得られる）既知のベンチマークに一致するように、ウェイトを調整する（calibration）。

calibration の方法としてはそれぞれの変数の組み合わせに関するサンプルのセル比率が、ベンチマーク統計におけるそれらと一致するように調整するのが、単純な事後層化の方法である。しかし、ベンチマーク統計においてすべての変数の組み合わせに関するクロス表を得るのが難しいことや、サンプルにおいてゼロセルが多くなることから、この方法は望ましくない。これに代わる手法として raking、すなわちウェイトに使用するそれぞれの変数の周辺確率が、ベンチマーク統計におけるそれぞれの変数の周辺確率に一致するように反復調整する方法が用いられている。

4.2 JLPS-YM における変数選択

石井・野崎（2014）は KHPS/JHPS において次のように第 1 波横断ウェイトを作成している。

まず KHPS/JHPS は層化二段無作為抽出が行われていることから、初期ウェイトの作成は省略している。ウェイト作成に使用している変数は、学歴 5 区分（中学卒／高校卒／短大・高専卒／大学・大学院卒／在学中）、就業状態 6 区分（主に仕事／通学のかたわら仕事／家事のかたわら仕事／休職中／求職中／通学・家事）、年齢階級（5 歳刻み）、地域 8 区分である。就業状態と配偶状態には国勢調査、学歴には就業構造基本調査、年齢階級には人口推計、居住地域には KHPS/JHPS を使用している。ただし、ベンチマークとなる統計は毎年収集されているとは限らないため、その場合にはもっとも近い年次のベンチマーク統計が使用されている。

この、石井・野崎（2014）の方法を参考にして、JLPS-YM の Wave1 横断ウェイト作成には表 2 のように変数とベンチマーク統計の選択をした。

表 2　使用する変数とベンチマーク統計

変数	ベンチマーク統計	カテゴリー
性別・年齢階級	人口推計（2006年）	男性20～24歳／男性25～29歳／男性30～34歳／男性35～40歳／女性20～24歳／女性25～29歳／女性30～34歳／女性35～40歳
学歴	就業構造基本調査（2007年）	中学／高校／短大・高専・専門／大学・大学院／在学中
就業状態	就業構造基本調査（2007年）	仕事／通学／家事／その他
配偶状態	国勢調査（2005年）	既婚／未婚／離別・死別
居住地域	人口推計（2006年）	北海道／東北／関東／中部／関西／中国／四国／九州・沖縄

性別・年齢階級には 2006 年人口推計を使用し、「男性 20～24 歳／男性 25～29 歳／男性 30～34 歳／男性 35～40 歳／女性 20～24 歳／女性 25～29 歳／女性 30～34 歳／女性 35～40 歳」の 8 区分とした。学歴には 2007 年就業構造基本調査を使用し、「中学／高校／短大・高専・専門／大学・大学院／在学中」の 5 区分とした[1]。就業状態には 2007 年就業構造基本調査を使用し、「仕事／通学／家事／その他」の 4 区分とした。配偶状態には 2005 年国勢調査を使用し、「既婚／未婚／離別・死別」の 4 区分とした。居住地域には 2006 年人口推計を使用し、「北海道／東北／関東／中部／関西／中国／四国／九州・沖縄」の 8 区分とした。ベンチマーク統計は JLPS-YM にあわせるために 20～39 歳の年齢階級の値

[1] 就業構造基本調査は学歴に関しては、最終卒業学校について尋ねており、中退の場合はその前の学校となる。JLPS-YM でも中退は同様の扱いとした。

を使用している。

5. JLPS-YM における横断ウェイトの作成

表 3　サンプルとベンチマーク統計のずれ（%）

	JLPS-YM	ベンチマーク	差
性別年齢階級（N=4800）			
男性・20～24歳	9.85	10.50	-0.65
男性・25～29歳	10.44	11.39	-0.95
男性・30～34歳	14.98	13.76	1.22
男性・35～40歳	14.00	15.22	-1.22
女性・20～24歳	10.42	9.94	0.48
女性・25～29歳	10.98	10.99	-0.01
女性・30～34歳	13.48	13.35	0.13
女性・35～40歳	15.85	14.86	0.99
学歴（N=4730）			
中学	3.95	5.08	-1.13
高校	28.48	34.45	-5.97
短大・高専・専門	30.53	26.71	3.82
大学・大学院	28.25	25.51	2.74
在学中	8.79	8.25	0.54
就業状態（N=4784）			
仕事をしている	82.17	77.26	4.91
仕事をしていない・通学	3.01	5.48	-2.47
仕事をしていない・家事	11.85	12.38	-0.53
仕事をしていない・その他	2.97	4.88	-1.91
配偶状態（N=4800）			
既婚	46.98	44.82	2.16
未婚	49.63	51.93	-2.30
離死別	3.40	3.25	0.15
地域ブロック（N=4800）			
北海道	4.08	4.43	-0.35
東北	6.63	7.56	-0.93
関東	36.04	32.51	3.53
中部	19.56	16.98	2.58
関西	15.54	17.73	-2.19
中国	5.85	6.01	-0.16
四国	2.79	3.21	-0.42
九州・沖縄	9.50	11.58	-2.08

表 3 は横断ウェイト作成に使用する変数に関して、サンプルとベンチマーク統計における分布を示したものである。

性別年齢階級については、ベンチマーク統計と比較して、35～40 歳男性の回答率がやや高く、他の年齢層の男性の回答率がやや低い。学歴に関しては JLPS-YM では、中学と高校の比率が低く、他方で短大・高専・専門と大学・大学院の比率が高くなっており、高学歴者に比率がやや偏っている。就業状態に関しては、回答時点で仕事をしている人々の比率がベンチマーク統計にくらべて高くなっている。配偶状態では、未婚者にくらべて既婚者のサンプル比率がベンチマーク統計にくらべてやや高い。地域ブロックに関して JLPS-YM では、関東と中部の回答比率がやや高く、関西と九州・沖縄ではやや低くなっている。

次に上述した raking の手法を用いて、それぞれの変数の周辺確率が、ベンチマーク統計の周辺確率に一致するように反復調整を行った。これには Stata の ipfweight コマンドを使用した（Bergmann 2011）。欠損値については除外せず、それぞれのステップにおいて重みが 1 となるように設定した。また各ステップにおいて許容度は定めず、反復回数は 10 回とした。

以上の手続きによってウェイト変数が計算された。その要約統計を示したのが表 4 である。サンプルの中では最大でおよそ 3.67 倍の重みを与えられ、最小ではおよそ 0.49 倍の重みが与えられることになる。

表 4　ウェイト変数の要約

平均	1.001
標準偏差	0.305
最大値	3.665
最小値	0.488
N	4800

表 5 には、ウェイト後の JLPS-YM の各変数の分布を示し、ベンチマーク統計のそれぞれの値と比較した。結果として、すべての変数・カテゴリーにおいて、小数点以下 2 桁で値が一致しており、これらの選択した変数の周辺分布については母集団における分布にうまく調整されたと言える。

表 5　ウェイト後の各変数の分布（%）

	JLPS-YM（ウェイト後）	ベンチマーク
性別年齢階級		
男性・20～24歳	10.50	10.50
男性・25～29歳	11.39	11.39
男性・30～34歳	13.76	13.76
男性・35～40歳	15.22	15.22
女性・20～24歳	9.94	9.94
女性・25～29歳	10.99	10.99
女性・30～34歳	13.35	13.35
女性・35～40歳	14.86	14.86
学歴		
中学	5.08	5.08
高校	34.45	34.45
短大・高専・専門	26.71	26.71
大学・大学院	25.51	25.51
在学中	8.25	8.25
就業状態		
仕事をしている	77.26	77.26
仕事をしていない・通学	5.48	5.48
仕事をしていない・家事	12.38	12.38
仕事をしていない・その他	4.88	4.88
配偶状態		
既婚	44.82	44.82
未婚	51.93	51.93
離死別	3.25	3.25
地域ブロック		
北海道	4.43	4.43
東北	7.56	7.56
関東	32.51	32.51
中部	16.98	16.98
関西	17.73	17.73
中国	6.01	6.01
四国	3.21	3.21
九州・沖縄	11.58	11.58

6. 横断ウェイトのパフォーマンス評価

横断ウェイトのパフォーマンスを評価するために、ウェイト作成に使用していない変数について、ウェイト適用前後の分布を検討する。ここでは、週労働時間と年間世帯収入に注目する。ベンチマーク統計として、週労働時間には 2007 年労働力調査（1～3 月期）を使用し、年間世帯収入には 2007 年国民生活基礎調査を使用する。

表 6　週労働時間のウェイト適用前後の比較（%）

週労働時間	JLPS-YM ウェイト前	JLPS-YM ウェイト後	労働力調査
1～14時間	5.5	4.5	4.0
15～34時間	15.2	14.4	14.2
35～42時間	24.9	25.4	30.8
43～48時間	16.2	17.1	20.4
49～59時間	17.6	17.3	17.5
60時間以上	20.6	21.3	13.1

表 6 には週労働時間に関して、JLPS-YM のウェイト前、JLPS-YM のウェイト後、労働力調査の 3 つの分布を示した。三輪（2008）でも指摘されているように、JLPS-YM のウェイト前の値を見ると、労働力調査にくらべて週 1～14 時間および週 60 時間以上という短時間・長時間労働の両極でより比率が高くなっている。

ウェイトを適用することで、週 1～14 時間のカテゴリーでは 5.5→4.0%と、労働力調査の 4.0%により近い値となった。しかし、週 60 時間以上のカテゴリーでは、20.6→21.3%と、逆に労働力調査と乖離する方向の値を示した。

図１　年間世帯収入のウェイト適用前後の比較（%）

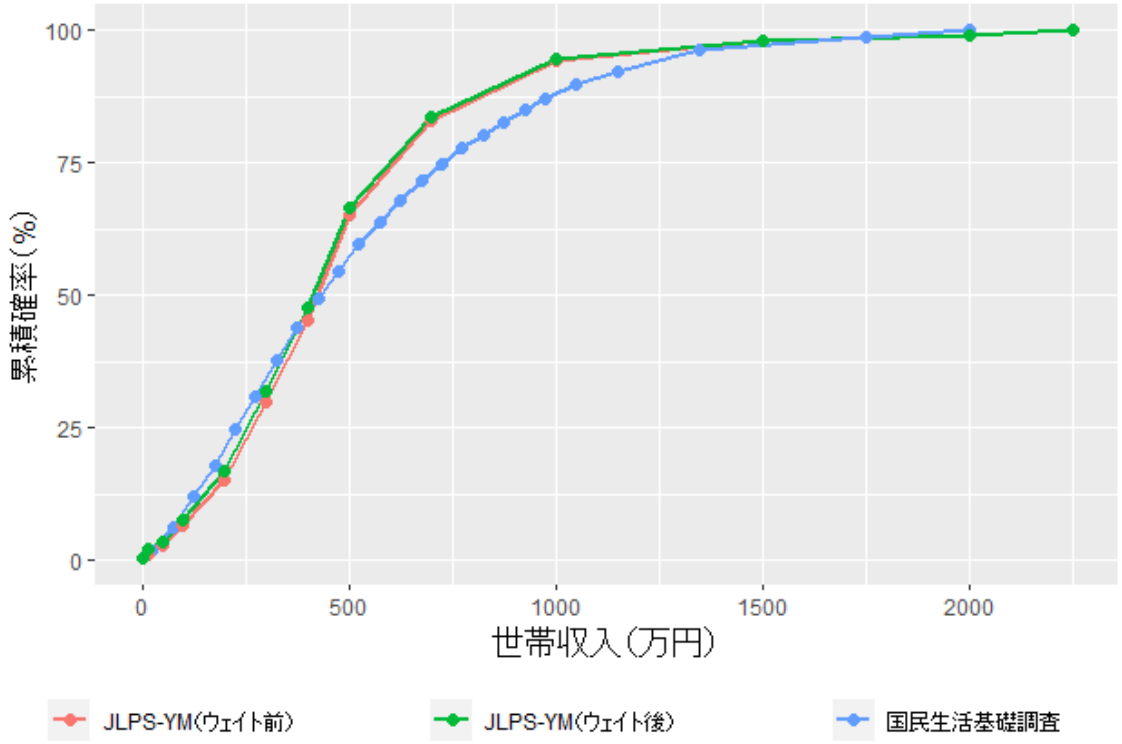

図 1 は年間世帯収入について、JLPS-YM のウェイト前、JLPS-YM のウェイト後、就業構造基本調査の 3 つの分布を示した。縦軸は累積確率である。なお、国民生活基礎調査は世帯調査であるため、JLPS-YM にはさらに同居人数の逆確率をウェイトとして乗じて集計した。ただし国民生活基礎調査はすべての世帯を対象としているのに対して、JLPS-YM は 20～40 歳の人々を対象としているため、厳密な結果の比較はできず、あくまで参考である。

年間世帯収入で 500 万円を境として、JLPS-YM は国民生活基礎調査の累積確率を上回り、収入がより高い方向に偏っていることがわかる。これはウェイトの適用後にも分布はほとんど変化しないことが示されている。

7. まとめと結論

本論文では標本調査における事後層化ウェイトの手法を用いて、JLPS-YM の Wave1 調査における横断ウェイトの作成について検討してきた。考慮した変数（性別年齢階級、学歴、就業状態、配偶状態、居住地域）については、周辺分布がベンチマーク統計に一致するように横断ウェイトが作成できた。しかしウェイトによって、就労時間や世帯所得とい

った変数の分布は必ずしも改善されないという限界もわかった。これより、ウェイト作成に使用した変数以外による回答拒否や欠損のメカニズムがより複雑であることが示唆される。こうした限界はあるにせよ、横断ウェイトの１つを示すことで、今後 JLPS-YM を用いた分析の結果が頑健であるかどうかなどの検証可能性をより高めることに貢献できたのではないかと考えられる。また今後は縦断ウェイト、および追加サンプルとの統合ウェイトの作成も課題である。

引用文献

Bergmann, Michael. 2011. "IPFWEIGHT: Stata Module to Create Adjustment Weights for Surveys." Statistical Software Components S457353, Boston College Department of Economics.

Gelman, Andrew and John B. Carlin. 2002. "Poststratification and Weighting Adjustments." Pp. 289-302 In *Survey Nonresponse*, edited by R. M. Groves et al. New York: Wiley.

Holt, D. and T. M. F. Smith. 1979. “Post Stratification.” *Journal of the Royal Statistical Society. Series A*. 142(1): 33-46.

石井加代子・野崎華世，2014，「『慶應義塾家計パネル調査（KHPS）』と『日本家計パネル調査（JHPS）』における Cross-sectional/Longitudinal ウェイトおよびパネル統合ウェイトの作成」『三田商学研究』57(4): 123-45.

松井博，2006，「ウエイトの改定について」『日本版 General Social Surveys 基礎集計表・コードブック JGSS-2006』29-30.

三輪哲，2008，「働き方とライフスタイルの変化に関する全国調査 2007 における標本特性と欠票についての基礎分析」『東京大学社会科学研究所パネル調査プロジェクト ディスカッションペーパーシリーズ』No.10.

Watson, Nicole. 2012. “Longitudinal and Cross-sectional Weighting Methodology for the HILDA Survey.” *HILDA Project Technical Paper Series* No.2/12, Melbourne Institute of Applied Economic and Social Research.

東京大学社会科学研究所パネル調査プロジェクトについて

労働市場の構造変動、急激な少子高齢化、グローバル化の進展などにともない、日本社会における就業、結婚、家族、教育、意識、ライフスタイルのあり方は大きく変化を遂げようとしている。これからの日本社会がどのような方向に進むのかを考える上で、現在生じている変化がどのような原因によるものなのか、あるいはどこが変化してどこが変化していないのかを明確にすることはきわめて重要である。

本プロジェクトは、こうした問題をパネル調査の手法を用いることによって、実証的に解明することを研究課題とするものである。このため社会科学研究所では、若年パネル調査、壮年パネル調査、高卒パネル調査、中学生親子パネル調査の４つのパネル調査を実施している。

本プロジェクトの推進にあたり、以下の資金提供を受けた。記して感謝したい。

文部科学省・独立行政法人日本学術振興会科学研究費補助金
基盤研究 S：2006 年度～2009 年度、2010 年度～2014 年度　基盤研究 C：2013 年度～2016 年度　特別推進研究：2015 年度～2017 年度　若手研究 A：2015 年度～2018 年度
基盤研究 B：2016 年度～2020 年度　特別推進研究：2018 年度～2024 年度

厚生労働科学研究費補助金
政策科学推進研究：2004 年度～2006 年度

奨学寄付金
株式会社アウトソーシング（代表取締役社長・土井春彦、本社・静岡市）：2006 年度～2008 年度

東京大学社会科学研究所パネル調査プロジェクト
ディスカッションペーパーシリーズについて

東京大学社会科学研究所パネル調査プロジェクトディスカッションペーパーシリーズは、東京大学社会科学研究所におけるパネル調査プロジェクト関連の研究成果を、速報性を重視し暫定的にまとめたものである。

東京大学社会科学研究所 パネル調査プロジェクト
https://csrda.iss.u-tokyo.ac.jp/panel/