Technical report: Logistic regression and latent class analysis of the poorest personal well-being using the Annual Population Survey (2014 to 2016)

Description of the statistical methods and techniques which underpin the article: Understanding well-being inequalities: Who has the poorest personal well-being?

Contact:
Edward Pyle
qualityoflife@ons.gov.uk
+44 (0)1633 582486

Release date:
11 July 2018

Next release:
To be announced

Table of contents

1. Introduction

2. The Annual Population Survey, 2014 to 2016

3. Logistic regression

4. Latent class analysis

1 . Introduction

This technical report accompanies Understanding well-being inequalities: Who has the poorest personal well-being? an exploration of factors associated with the lowest reports of personal well-being. Using three years of data from the Annual Population Survey (APS) (January 2014 to December 2016), the characteristics and circumstances of people with poorest personal well-being were compared against others who reported higher personal well-being. Following this are more in-depth analyses which explore the nature of associations between these factors and personal well-being.

The research took an iterative approach, involving descriptive analysis followed by logistic regression and latent class analysis (LCA). The logistic regression isolates single factors that impact on the odds of reporting the lowest personal well-being levels. The LCA identifies combinations of factors that frequently occur together among those with poorest personal well-being. Logistic regression and LCA are complementary techniques, appreciating that while single factors affect personal well-being, in practice, combinations of influential factors tend to go together. This methodology paper describes how these techniques were applied.

2 . The Annual Population Survey, 2014 to 2016

The three-year Annual Population Survey (APS) dataset has a sample size of 543,298 respondents of which 284,456 were aged 16 years and over and eligible to be asked personal well-being questions. Of these, 280,003 (over 98%) answered all four personal well-being questions and were included for analysis.

Both logistic regression and latent class analysis (LCA) cannot be applied to missing data. With more variables included in a model, there is a greater likelihood that a case will contain missing data and so be excluded from analysis. We found that better LCA data was produced when applied to fewer variables when compared to the logistic regression. As a result, 192,567 cases were included in the logistic regression model and 227,139 in the LCA.

There are four personal well-being questions:

1. Overall, how satisfied are you with your life nowadays?
2. Overall, to what extent do you feel the things you do in your life are worthwhile?
3. Overall, how happy did you feel yesterday?
4. Overall, how anxious did you feel yesterday?

The responses to all four personal well-being questions are measured on a 0 to 10 scale, where 0 is “not at all” and 10 is “completely”. For the three positively framed questions (questions 1 to 3 above), a score of 4 or less is deemed to be “poor”, and for the anxiety question (question 4 above), a score of 6 or more is defined as “poor” (as it indicates higher anxiety). In this research, individuals defined as having poorest well-being are those who reported life satisfaction, worthwhile and happiness scores of 4 or less, in addition to an anxiety score of 6 or more.

Of the 280,003 respondents who answered all four personal well-being questions, 3,135 reported poorest personal well-being – approximately 1% of the sample. Similarly, with survey weighting taken into account, this represents about 1% of the UK population. A binary variable was derived to flag respondents with or without poorest personal well-being, allowing for the characteristics of those with poorest personal well-being to be compared with those who reported higher personal well-being.

Missing data and bias

As noted, cases with missing data for variables included in the logistics regression and LCA model were excluded from analysis. Missing data can produce biased estimates and invalid conclusions, particularly if data are not “missing at random” or, in other words, if there is some (unknown) patterning to that “missingness” (Graham, 2009).

People with certain characteristics, for example, may be less likely to answer the personal well-being questions accurately. The three variables with the largest proportion of missing data were: education (17.0%), sexual orientation (9.9%) and disability status (7.0%).

3 . Logistic regression

Logistic regression analysis allows for the relationship between an explanatory variable and the outcome variable to be examined, while at the same time accounting for other explanatory variables that influence the outcome. It is used when looking at categorical outcomes. While it is possible to conduct multinomial logistic regression with multiple categorical outcomes, logistic regression with binary outcomes was chosen to increase ease of understanding (with the predicted outcomes either “poorest personal well-being” or “higher personal well-being”) and for consistency with the latent class analysis (LCA) which can only be applied to categorical data.

Procedure

This analysis was carried out using R. The package used for the logistic regression was mlogit. After removing those cases where there were missing data in the predictor variables, 192,567 cases were included. Variables were then added one-by-one to build the logistic regression model.

Goodness of fit

Goodness of fit describes how well a model fits the data from which it is generated. After the addition of each variable to the model, goodness of fit and change in the coefficients were assessed. The variables tested included sex, age, marital status, self-reported health, self-reported disability, socio-economic activity, education, housing tenure, ethnicity, sexual identity and religion.

Causality

Regression analysis can identify relationships between factors; however, it cannot tell us about causality. While, for some factors, causality is fairly clear based on prior knowledge (for example, poorest personal well-being does not cause someone to become widowed, however, becoming widowed can cause poorest personal well-being), for others the relationship between cause and effect is more blurred (for example, having very bad or bad health can cause poorest personal well-being, but also poorest personal well-being can negatively impact on health). Therefore, where prior knowledge does not make the direction of causality clear, it is important to note that causality can operate in either direction (or both).

Weighting

Weights were included in the logistic regression to compensate for unequal selection probabilities and differential non-response. Our regression models take the weights into account. For more information about how the Annual Population Survey (APS) datasets are weighted to reflect the size and composition of the general population, please see Personal well-being in the UK Quality and Methodology Information .

Interpretation of the results

Odds are the probability of an event occurring divided by the probability of the event not occurring. The odds ratio, which is the ratio between two sets of odds, is the usual output from logistic regression. The odds ratio for each variable in the model is obtained by exponentiating the estimate. For this analysis, the odds ratio represents the odds of reporting poorest personal well-being for given predictor variables relative to the reference category while holding all other variables constant. This reveals how personal characteristics and circumstances relate to odds of reporting poorest personal well-being.

Table 1: The odds of people reporting the poorest personal well-being for different characteristics

Factor	Reference	Category	Log-odds	Std error (log-odds)	Odds
Self-reported health	Good	Bad	2.61	0.01	13.63
		Fair	1.37	0.01	3.93
Economic activity (ILO definition)	Student	Employee	0.25	0.02	1.29
		Family work	1.42	0.03	4.12
		ILO unemployed	1.3	0.02	3.69
		Inactive	0.6	0.02	1.83
		Inactive (LT sick or disabled)	1.08	0.02	2.94
		Retired	0.02	0.02	1.02
		Self-employed	0.27	0.02	1.31
Age	70+	16 to 29	0.43	0.04	1.54
		30 to 39	0.89	0.04	2.43
		40 to 49	1.09	0.04	2.98
		50 to 59	1.04	0.04	2.82
		60 to 69	0.48	0.04	1.61
Marital status	Married or civil partnership	Separated	0.78	0.01	2.18
		Single	0.71	0	2.04
		Divorced	0.73	0.01	2.08
		Widowed	0.79	0.01	2.21
Self-reported disability	No disability	Disability	0.63	0.01	1.87
Socio-economic activity	Managerial	Intermediate or lower superv	0.11	0.01	1.11
		Semi or routine occupation	0.28	0.01	1.33
		Small employer or own account	0.2	0.01	1.22
		Never worked or unemployed	0.16	0.01	1.18
Sexual identity	Heterosexual	Non-heterosexual	0.24	0.01	1.27
Ethnicity	Not White British	White British	0.22	0.01	1.25
Education	A-level	Basic or none	0.13	0.01	1.14
		Degree or professional	0.14	0.01	1.15
		GCSE	0.08	0.01	1.08
		Other qualification	0.16	0.01	1.17
Housing tenure	Mortgage	Owned	0.08	0.01	1.08
		Rent	0.15	0.01	1.16
Religion	Religious	Not religious	0.14	0	1.15

Sex	Female	Male	0.11	0	1.12
	Constant		-8.31	0.04	0

Notes:

1. All odds significant except 'retired' (p<0.01).

4 . Latent class analysis

Latent class analysis (LCA) is a technique used to identify sub-groups within a population. It classifies individuals into mutually exclusive groups or “types” based on patterns of characteristics represented as categorical variables. LCA was used to group individuals with similar characteristics including:

- age
- self-reported health
- self-reported disability (as defined by the Equality Act 2010)
- housing tenure
- economic activity
- socio-economic activity

Table 2 presents the social characteristics of each class. In Class 8, for example, 82.4% were found to self-report a disability whereas 5.8% in Class 7 were found to self-report a disability.

Table 2: Characteristics by class

		Class							
Factor	**Level**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age group	16 to 59	18.78	0.61	99.94	86.5	96.01	82.03	97.76	0.6
	60 and over	81.22	99.39	0.06	13.5	3.99	17.97	2.24	99.4
General health	Fair, bad or very bad	10.2	3.95	6.57	98.4	34.41	15.5	5.22	93.59
	Good or very good	89.8	96.05	93.43	1.6	65.59	84.5	94.78	6.41
Disability	No disability	86.96	82.35	92.53	22.63	67.09	85.01	94.18	17.59
	Disabled	13.04	17.65	7.47	77.37	32.91	14.99	5.82	82.41
Tenure	Own home	92	86.82	8.58	19.4	12.74	28.52	13.75	61.39
	Mortgage	6.71	4.44	16.85	31.47	9.98	43.91	53.4	4.88
	Renting	1.3	8.74	74.57	49.13	77.28	27.57	32.85	33.73
Economic activity	Employee	68.76	0	19.56	89.31	0	0	97.92	0
	Inactive	16.87	1.22	2.72	8.12	23.62	1.39	0.65	1.34
	Retired	13.16	98.74	0	0.39	0.03	0.09	0	98.64
	Self-employed	0	0.01	2.14	0	0	97.92	0	0
	Student	0.08	0	65.88	0.13	1.57	0.02	0.04	0
	Unemployed	1.13	0.03	9.7	2.05	74.77	0.58	1.39	0.02
Socio-economic classification	Employers or self-employed	0.06	5.99	0.12	0.07	3.36	98.71	0.01	2.71
	Full-time student	0.03	0.03	94.99	2.13	1.28	0.1	2.26	0
	Higher-lower professional or intermediate	66.02	24.14	0.68	43.72	2.05	1.17	66.24	6.86
	Never worked or long-term unemployed	0	1.4	3.01	0.03	47.83	0	0	2.39
	Routine	33.86	12.94	1.05	54	33.53	0	31.43	8.1
	Other	0.03	55.5	0.14	0.06	11.95	0.02	0.07	79.94
Marital status	Married or civil partnership	67.73	62	5.41	39.94	21.61	55.96	46.68	47.09
	Separated	2.19	1.41	0.85	5.1	4.37	3.57	3.18	1.97
	Single	9.22	5.87	92.8	38.16	63.46	27.71	41.49	6.28
	Divorced	12.69	9.44	0.82	13.97	8.92	10.66	7.72	12.47
	Widowed	8.18	21.28	0.12	2.83	1.64	2.1	0.93	32.19
Proportion of LCA sample		7.45	14.46	2.82	6.96	3.22	10.07	44.61	10.42

Notes:

1. Class numbering is arbitrary.

2. Estimates take the APS weights into account.

3. Marital status was not included in the LCA model, as it was not a key variable to help identify the classes, but was added as a descriptive variable for further examination of the groups.

Logistic regression was used to calculate the odds of reporting poorest personal well-being for members of each group (Table 3).

Table 3: Estimated odds of reporting poorest personal well-being by class

	Log odds		Odds
Class	Estimate	Std. Error	Fractional
1	-6.22900	0.22383	1/508
2	-6.62675	0.14443	1/756
3	-6.08313	0.25849	1/439
4	-3.67916	0.05524	1/41
5	-3.43928	0.06684	1/32
6	-5.22588	0.09078	1/187
7	-5.84813	0.05697	1/348
8	-4.24246	0.05604	1/71

The model shows that individuals are at significantly different risk of reporting poorest personal well-being, depending on which latent class they belong to. As 1% of the UK population have poorest personal well-being, before any characteristics are taken into account an individual selected at random has a 1 in 100 chance of reporting poorest personal well-being. With individual characteristics taken into account, those at greatest risk of having the poorest personal well-being are in Class 4 (1 in 41 chance), Class 5 (1 in 32 chance) and Class 8 (1 in 71 chance). In the article, Understanding well-being inequalities: Who has the poorest personal well-being , only these classes have been reported as they represent the main focus of the analysis; Class 4 are "Employed renters with self-reported health problems or disability", Class 5 are "Unemployed or inactive renters with self-reported health problems/disability" and Class 8 are "Retired homeowners with self-reported health problems or disability".

Optimal number of classes

LCA analysis reduces complexity by splitting a dataset up into meaningful sub-groups based on the specified characteristics. The process involves running the algorithm on the same data with different numbers of classes specified. The analyst first specifies one group, then two groups, then three and so on. To better ensure data quality, this process was first applied to random sub-samples of the dataset with 60,000, 80,000 and 100,000 cases to ensure consistency. The final specification was then applied to the full sample.

With each run a goodness of fit statistic, the Bayes Information Criterion (BIC), is produced. All other things equal, a lower BIC value suggests a better fitting model (Lin and Dayton 1997).

Although a model with a lower BIC value may suggest a better fitting model, separation into greater numbers of classes has disadvantages. Doing so can increase complexity, making interpretation and communication of findings more difficult, while splitting the dataset into more groups can mean fewer respondents fall into each class thereby potentially reducing statistical power of the model.

The BIC value fell continuously as the number of classes specified increased. However, past eight groups the composition of characteristics associated with poorest personal well-being changed little. As such, eight classes was selected as the most useful model for this release.