Estimating correlations between socio-demographic and socio-economic factors and self-reported health while discounting objective health metrics: An analysis of the Continuous National Health and Nutrition Examination Survey (2001-2010)

Introduction

While self-reported health (SRH) metrics have been accepted as reliable approximations for objective health status (Miilunpalo et al., 1997, p.526; Gilmore et al., 2002, p. 2177), some studies examine their correlation with socio-demographic and socio-economic variables in combination with objective health metrics (Szwarcward et al., 2005; Gallagher et al., 2016). Since socio-demographic and socio-economic factors are proven correlates of objective health outcomes, and vice versa, this is problematic. (Link and Phelan, 1996, pp. 3-4). Therefore, more research is needed to determine socio-demographic and socio-economic correlates of SRH, while discounting objective health metrics.

Research Question

In this paper, we wish to determine socio-demographic and socio-economic correlates of SRH, in the absence of objective metrics. Particularly, which socio-demographic and socio-economic correlates of SRH should policymakers and researchers consider when making public health decisions?

In determination, objective health metrics will be omitted, as they are correlated with socio-demographic and socio-economic characteristics. Subsequently, further studies should reassess correlations found here, or uncover SRH's objective health correlates, while discounting socio-demographic and socio-economic factors. Ultimately, SRH's socio-demographic, socio-economic, and objective health correlates should be tested for multicollinearity. From these tests, a definitive list of the most predictive, independent SRH metrics can be generated. This list would allow researchers and policy-makers to implement studies and policies with the greatest impacts on SRH.

First, a literature review will address the feasibility of focusing solely on socio-demographic and socio-economic variables and will examine which factors have been proven SRH correlates in previous studies. Next, data will be presented and summarized. The subsequent methodology section will explain why multinomial logistic regression (mlogit) is used as opposed to a multinomial probit regression (mprobit). A Brant test will show that data is better specified for unordered mlogit than ordinal mlogit, and hypotheses will be stated. Finally, results will be presented, discussion of implications and extensions will ensue, and, in conclusion, the research question will be answered.

Literature Review

Generally, studies on correlates of SRH fall into two categories: those that examine objective health metrics, socio-demographic, and socio-economic variables in conjunction and those that use either objective health metrics or a combination of socio-demographic and socio-economic variables. Those fixated on socio-demographic and socio-economic factors usually consider age, gender, socioeconomic status (SES), ethnicity, and educational attainment (Franks, et al., 2003, p. 2508; Lubetkin and Jia, 2017, pp. 3-5). Often, studies considering objective measures as well focus on the same factors but include diagnostic metrics, such as body mass index or glucose levels (Szwarcwald et al., 2005, pp. S56-S60; Gallagher et al., 2016, p. 3-4). These approaches are problematic because socio-demographic socio-economic factors and objective health measures are inherently linked. Despite warnings of this in 1996 (Link and Phelan), Gallagher et al. are the only authors cited here that mention independent variable correlation; they indicate multicollinearity was not examined but is likely present. They posit that further health metric studies should investigate independent variables' collinearity.

While not explicitly reporting correlation matrices, researchers have investigated correlations among socio-demographic and socio-economic variables. For instance, those with a lower social economic status (SES) are more likely to be in worse health (Stringhini et al., 2017, pp. 1233-1235). Additionally, those with a lower SES are more likely to be smokers and more likely to be in worse health (Reijneveld, 1998, p. 35). Likewise, black individuals are more likely to have a lower SES, and thus are more likely than non-black individuals to be in worse health (Do et al., 2012, p. 1389). Based on these findings, correlations between socio-demographic and socio-economic factors are apparent. Thus, including objective health metrics would increase collinearity among predictors, which is likely already present. For this reason, only socio-demographic and socio-economic factors will be considered when determining SRH correlates.

In addition to ethnicity and SES, practically all studies use age, gender, and educational attainment when determining SRH variations. It is well-documented that women report being in worse health than men but have a lower mortality rate (Cherepanov et al., 2010, p. 1119). Research has also determined that those with less education are more likely to report poor health (Baker et al., 1997, p. 1030), as are those who are lacking citizenship (Gallagher et al., 2016, p. 8). Counterintuitively, older people disproportionately report better health, relative to younger counterparts (Idler, 1993, pp. S298-S299). Based on these findings, the socio-demographic and socio-economic factors used as predictors in this paper are age, gender, SES, ethnicity, educational attainment, and citizenship.

In addition to influencing variable choices, the references made throughout the paper also impacted the model of choice. While a range of methods were used, almost all used regression. Multivariate nonparametric, ordinary least squares and logistic regressions were used by some researchers, whereas most used a form of multinomial logistic or stepwise logistic regression. For this reason, multinomial logistic regression will be used in this paper and will be described in depth in the methodology section.

Data

Data from the Continuous National Health and Nutrition Examination Survey (Continuous NHANES) will be used to determine socio-demographic and socio-economic correlates of SRH. In the Continuous NHANES, a country-wide survey, all respondents contribute to the questionnaire portion of the survey, whereas some are selected for a physical examination as well.

The survey contains demographic, dietary, physical examination, laboratory, and questionnaire data from 1999-2016, but, due to collection inconsistencies, only demographic and questionnaire data from 2001-2010 will be considered. Particularly, demographic data supply age, gender, ethnicity, annual household income, education, and citizenship information, while the questionnaire portion contains the SRH measure[1] (National Center for Health Statistics, 2018a; 2018b). Annual household income is considered a proxy for SES.

The initial dataset covers 48,499 individuals. For the considered variables, those who refused to answer or failed to respond were removed. Additionally, the educational variable is only applicable for those twenty years or older, so individuals younger than twenty are omitted. Ages are censored at eighty-years for some periods and at eighty-five for others, so records for individuals eighty and older were also omitted. Finally, salaries were censored at $75,000 or $100,000 depending on the year, so those that made $75,000 or above were combined into one group.

Table 1: Socio- traits of the sample (*N*=20,453 twenty- to seventy-nine-year-old persons)

	N	Percent		N	Percent
Annual Household Income			Education		
$0 - $4,999	394	1.9	Less than 9th grade	2,377	11.6
$5,000 - $9,999	966	4.7	9-12th grade, without diploma	3,228	15.8
$10,000 - $14,999	1,596	7.8	HS diploma or GED equivalent	4,892	23.9
$15,000 - $19,999	1,522	7.4	Some college or Associate degree	5,722	28.0
$20,000 - $24,999	1,727	8.4	College graduate and above	4,234	20.7
$25,000 - $34,999	2,751	13.5	Age bands, in years		
$35,000 - $44,999	2,152	10.5	20-29	3,909	19.1
$45,000 - $54,999	1,900	9.3	30-39	3,690	18.0
$55,000 - $64,999	1,423	7.0	40-49	3,772	18.4
$65,000 - $74,999	1,120	5.5	50-59	3,135	15.3
$75,000 +	4,902	24.0	60-69	3,425	16.7
Citizenship			70-79	2,522	12.3
Citizen	17,768	86.8	Age statistics		
Non-citizen	2,685	13.1	Mean	47.2	
Ethnicity			Median	46	
Mexican American	3,983	19.5	Standard deviation	16.76	
Other Hispanic	1,318	6.4	Gender		
Non-Hispanic White	10,174	49.7	Male	10,013	49.0
Non-Hispanic Black	4,134	20.2	Female	10,440	51.0
Other/Multi-Racial	844	4.1			

Table 2: Self-reported health outcomes (*N*=20,453 twenty- to seventy-nine-year-old persons)

	N	Percent
Self-Reported Health		
Excellent	2,227	10.9
Very Good	5,824	28.5
Good	7,806	38.2
Fair	3,817	18.7
Poor	779	3.8

Besides age, which is continuous, all categorical variables were converted to binaries, with the least representative case (Multi-Racial, Male, etc.) being excluded. After data preparation, 20,453 individuals remain. The variables that will be modelled are: self-reported health (SRH), citizenship (citizen), ethnicity (MexAmerican, Hispanic, White, Black), education (9-12th, diploma, some college, college grad), age (age), gender (female), and annual income

[1] NHANES IDs: age-ridageyr, gender-riagendr, ethnicity-ridreth1, annual household income-indhhin2, education-dmdeduc2, citizenship-dmdcitzn, SRH-hsd010

($5k-$9,999; $10k-$14,999; $15k-$19,999; $20k-$24,999; $25k-$34,999; $35k-$44,999; $45k-$54,999; $55k-$64,999; $65k-$74,999; $75k+).

Methodology

Since SRH has five categories, the specified model should predict the likelihood of a respondent to report a category, based on their traits. The logistic regression satisfies this requirement using maximum likelihood estimation to determine independent variables' correlation with outcomes but is only feasible for two outcome categories (Cox, 1958, pp. 216-217).

An extension of the simple logistic model is the mlogit, and, for this case, the conditional mlogit. This model uses maximum likelihood estimation to determine the probability of occupying an outcome category, given a record's traits (Sperandei, 2013, pp. 13-14); however, it allows for more than two outcomes. Furthermore, a conditional mlogit is preferable to the unconditional form, because the goal is to estimate predictor coefficients individually for each SRH outcome, not cumulatively (Kuo et al., 1980, p. 2). While the conditional mlogit is suitable, mprobit should be considered.

The main difference between multinomial probit and multinomial logistic regressions is that mlogit assumes independence of irrelevant alternatives (IIA). For instance, if another category, say "Very ill", was added to SRH, it should not affect respondents' choices in other categories. While this additional choice *would* likely affect the choices of some in the "Poor" category, research has found that mlogit results are only marginally affected when IIA is violated and the sample is small (Dow and Endersby, 2004, pp. 119-120). Furthermore, IIA tests (Hausman-McFadden, etc.) have been found insufficient (Cheng and Long, 2007, p 598). Considering the size of the sample (*N*=20,453), mlogit is the preferred choice.

While most models assume normality, homoscedasticity, and linearity, mlogit does not. However, it does require absence of multicollinearity (Starkweather and Moske, 2011, p. 1). Figure A illustrates that multicollinearity is present among some independent variables, but since it is below (above) ±0.50, all variables will be included. Future analyses should consider the impact of omitting marginally correlated variables.

Figure A: Correlation matrix of variables with correlations greater (less) than ±0.30

	SRH	**MexAmerican**	**Hispanic**	**White**	**Black**	**Citizen**	**diploma**	**some college**	**college grad**	**$75k+**
SRH	1	0.17	0.03	-0.2	0.06	-0.12	0.03	-0.08	-0.25	-0.23
MexAmerican	0.17	1	-0.13	***-0.49***	-0.25	***-0.44***	-0.06	-0.1	-0.16	-0.12
Hispanic	0.03	-0.13	1	-0.26	-0.13	-0.14	-0.02	-0.01	-0.04	-0.04
White	-0.2	***-0.49***	-0.26	1	***-0.5***	***0.33***	0.06	0.05	0.17	0.16
Black	0.06	-0.25	-0.13	***-0.5***	1	0.14	0.01	0.04	-0.07	-0.06
Citizen	-0.12	***-0.44***	-0.14	***0.33***	0.14	1	0.08	0.12	0.09	0.12
diploma	0.03	-0.06	-0.02	0.06	0.01	0.08	1	***-0.35***	-0.29	-0.1
some college	-0.08	-0.1	-0.01	0.05	0.04	0.12	***-0.35***	1	***-0.32***	0.04
college grad	-0.25	-0.16	-0.04	0.17	-0.07	0.09	-0.29	***-0.32***	1	***0.33***
$75k+	-0.23	-0.12	-0.04	0.16	-0.06	0.12	-0.1	0.04	***0.33***	1

The final decision is whether mlogit should be ordered. Ordinal regressions estimate the cumulative likelihood of membership for one response category relative to all others but require proportional log-odds. Unordered models estimate likelihoods relative to a reference case but do not require proportional log-odds (Manor et al., 2000, pp. 150-151). For ordered regressions, if the Brant test indicates that the probability of any variable is below 0.05, then the proportional log-odds or parallel regression assumption is violated (Brant, 1990, pp. 1172-1174). Like other SRH studies (Xu and Jensen, 2006, p. 511), we find this assumption is violated by nine variables (see Figure B). Therefore, an unordered mlogit is preferable.

Figure B: Brant test results indicating violation of the parallel regression assumption

	X2	**df**	**probability**
female	2.93	3	0.40
age	15.60	3	***0.00***
MexAmerican	12.59	3	***0.01***
Hispanic	14.58	3	***0.00***
White	4.72	3	0.19
Black	14.78	3	***0.00***
citizen	77.91	3	***0.00***
9-12th	9.18	3	***0.03***
diploma	31.12	3	***0.00***
some college	36.81	3	***0.00***
college grad	33.64	3	***0.00***
$5k-$9,999	2.99	3	0.39
$10k-$14,999	0.54	3	0.91
$15k-$19,999	0.34	3	0.95
$20k-$24,999	0.84	3	0.84
$25k-$34,999	2.76	3	0.43
$35k-$44,999	2.91	3	0.41
$45k-$54,999	1.67	3	0.64
$55k-$64,999	6.92	3	0.07
$65k-$74,999	3.74	3	0.29
$75k+	4.04	3	0.26

For a conditional unordered mlogit, Y is an outcome category, taking values $1, \dots, N$, where N is the number of categories—in this case, 5. The response variables $y_i = (y_{i1}, \dots, y_{iN})$ are representative of the i-th subgroup and a multinomial distribution of $Mn(n_i, p_{i1}, \dots, p_{iN})$, where n_i is the value of an independent predictor and p_{iN} is the probability of outcome N for the i-th subgroup. Independent predictors' actual values for subgroup i are represented by $x_i = (x_{i1}, \dots, x_{jk})'$, where x_{ik} is the subgroup's k-th predictor. Regression coefficients are $\beta_j = (\beta_{0j}, \dots, \beta_{kj})'$, relative to the reference (j^*-th) case (Žežula, 2010, pp. 67-71). Subsequently, the general multinomial regression model is

$$\log\left(\frac{p_{ij}}{p_{ij^*}}\right) = x_i{}'\beta_j$$

where

$$j \neq j^*$$

Cox's original logistic regression estimated the probability of being in one of two categories with a binary logistic regression (1958), but in mlogit, $N - 1$ binary logistic regressions are calculated, where the unmodeled category is the reference case. With software advancements, the $N - 1$ regressions are now run simultaneously, thus reducing unexplained error (Grace-Martin, 2018).

Based on these findings, a conditional mlogit will be conducted. For each predictor and category, the null hypothesis will hold if the predictor's impact is indistinguishable from zero (i.e. there is no correlation with SRH). The alternative hypothesis is satisfied if a predictor's explanatory power in determining category membership is significantly different from zero. Predictors' p-values indicate the probability of seeing the results, or something more extreme. Therefore, for a 95% confidence, if a predictor's p-value is greater than 0.05 the null hypothesis cannot be rejected. However, the function used to conduct the model in R (`nnet`'s `multinom`) does not produce p-values. Instead, two-tailed Wald tests will be used to test the statistical significance of predictors from zero, by using `multinom`'s estimated coefficients and standard errors (Gouriéroux et al., 1982, pp. 63-66). These p-values will then be used to test null hypotheses.

Results

Table 3.1: Multinomial logistic regression results (part one)

SRH	Intercept	female	age	MexAmerican	Hispanic	White	Black	citizen	9-12th	diploma	some college	college grad
Very Good												
Coef.	0.400	0.128	0.005	-0.308	-0.553	-0.250	-0.408	0.520	0.278	0.446	0.413	0.025
Std. Err.	0.304	0.050	0.002	0.149	0.165	0.134	0.144	0.100	0.142	0.134	0.132	0.133
p-value	0.187	0.011	0.001	0.039	0.001	0.058	0.004	0.000	0.050	0.001	0.002	0.848
		**	***	**	***	*	***	***	**	***	***	
Good												
Coef.	1.739	0.163	0.014	-0.130	-0.444	-0.661	-0.325	0.006	0.167	0.144	-0.112	-0.861
Std. Err.	0.283	0.049	0.002	0.143	0.157	0.131	0.139	0.090	0.125	0.119	0.117	0.120
p-value	0.000	0.001	0.000	0.365	0.005	0.000	0.019	0.943	0.184	0.227	0.341	0.000
	***	***	***		***	***	**					***
Fair												
Coef.	1.259	0.259	0.026	0.289	-0.202	-0.694	0.004	-0.057	-0.283	-0.634	-1.047	-2.084
Std. Err.	0.304	0.057	0.002	0.167	0.183	0.157	0.164	0.099	0.128	0.123	0.122	0.133
p-value	0.000	0.000	0.000	0.083	0.269	0.000	0.979	0.562	0.026	0.000	0.000	0.000
	***	***	***	*		***			**	***	***	***
Poor												
Coef.	-0.728	0.305	0.035	-0.114	-0.658	-0.727	-0.412	0.705	-0.546	-0.891	-1.309	-2.773
Std. Err.	0.410	0.087	0.003	0.251	0.285	0.237	0.248	0.160	0.161	0.158	0.162	0.224
p-value	0.076	0.000	0.000	0.650	0.021	0.002	0.096	0.000	0.001	0.000	0.000	0.000
	*	***	***		**	***	*	***	***	***	***	***

Significance Levels: (p < 0.10 *) (p < 0.05 **) (p < 0.01 ***)

Table 3.2: Multinomial logistic regression results (part two)

SRH	$5K- $9,999	$10k- $14,999	$15k- $19,999	$20k- $24,999	$25k- $34,999	$35k- $44,999	$45k- $54,999	$55k- $64,999	$65k- $74,999	$75k+
Very Good										
Coef.	-0.243	-0.185	-0.232	-0.248	-0.106	-0.118	-0.034	-0.046	-0.243	-0.266
Std. Err.	0.284	0.269	0.264	0.260	0.253	0.254	0.255	0.258	0.259	0.246
p-value	0.393	0.492	0.381	0.340	0.674	0.642	0.894	0.858	0.349	0.279
Good										
Coef.	-0.340	-0.143	-0.331	-0.370	-0.252	-0.444	-0.461	-0.505	-0.677	-0.883
Std. Err.	0.266	0.251	0.247	0.244	0.237	0.238	0.240	0.243	0.245	0.231
p-value	0.202	0.569	0.181	0.129	0.287	0.063	0.054	0.038	0.006	0.000
						*	*	**	***	***
Fair										
Coef.	-0.078	-0.208	-0.533	-0.593	-0.656	-0.918	-0.890	-1.166	-1.335	-1.602
Std. Err.	0.274	0.260	0.257	0.254	0.247	0.250	0.252	0.259	0.264	0.244
p-value	0.776	0.425	0.038	0.019	0.008	0.000	0.000	0.000	0.000	0.000
			**	**	***	***	***	***	***	***
Poor										
Coef.	-0.029	-0.287	-0.605	-0.888	-1.054	-1.477	-1.350	-1.977	-2.003	-2.181
Std. Err.	0.330	0.317	0.317	0.317	0.310	0.322	0.326	0.370	0.381	0.319
p-value	0.931	0.365	0.056	0.005	0.001	0.000	0.000	0.000	0.000	0.000
			*	***	***	***	***	***	***	***

Significance Levels: (p < 0.10 *) (p < 0.05 **) (p < 0.01 ***)

Discussion

For variables denoted by “**” or “***”, the null hypothesis for that SRH outcome can be rejected, and we say that they have a statistically significant correlation with SRH, for that SRH outcome. Since the “Excellent” SRH category is the reference case, coefficients are interpreted relative to “Excellent”. For instance, we would say that a female should expect her multinomial log-odds for “Very good” relative to “Excellent” to increase by 0.128, over a male with identical, modelled traits.

It is apparent that individuals with higher household incomes ($55k+) have reduced chances of reporting “Good”, “Fair”, or “Poor” SRH, relative to “Excellent”. While this finding supports established SES and SRH relations (Stringhini et al., 2017, pp. 1233-1235), further analysis should determine how binary income-level estimates compare to those of one continuous variable.

Likewise, coefficients for age increase as SRH worsens, indicating that older individuals have greater log-odds probabilities of reporting “Poor” health than any other category, relative to “Excellent”. This finding directly contradicts previous studies (Idler, 1993, pp. S298-S299), implying that further research should seek to clarify the age-SRH correlation.

Ultimately, results are favorable: almost every predictor has a case in which one of its null hypotheses is rejected. This indicates future studies should work to either corroborate or refute the proposed correlations. Furthermore, other data may require a different model: the Brant test may remain unviolated, indicating ordinal regression is viable, or researchers may feel differently about the IIA assumption of mlogit and may wish to reassess using mprobit. We feel these extensions would be valuable, as they would produce much needed literature on correlations between socio-demographic or socio-economic factors and SRH.

Conclusion

The results indicate that all predictors are correlated with SRH for at least one outcome, besides two annual household income level ($5K-$9,999 and $10k-$14,999). However, because other income levels are significant, we conclude that gender, age, ethnicity, citizenship, education, and SES are correlates of SRH. We hope these results and those from future studies can eventually be used in conjunction to determine independent correlates of SRH, so that public health policies and studies can be conducted more efficiently and result in greater impacts.

Word Count: 1,989

References

Baker, D.W. et al. (1997). “The Relationship of Patient Reading Ability to Self-Reported Health and Use of Health Services”. *American Journal of Public Health*, 87(6), pp. 1027-1030.

Brant, R. (1990). “Assessing Proportionality in the Proportional Odds model for Ordinal Logistic Regression”. *Biometrics*, 46(4), pp. 1171-1178.

Cheng, S. and Long, J.S. (2007). “Testing for IIA in the Multinomial Logit Model”. *Sociological Methods & Research*, 35(4), pp. 583-600.

Cherepanov, D. et al. (2010). “Gender differences in health-related quality-of-life are partly explained by sociodemographic and socioeconomic variation between adult men and women in the US: evidence from four US nationally representative data sets”. *Quality of Life Research*, 19(8), pp. 1115-1114.

Cox, D.R. (1958). “The Regression Analysis of Binary Systems”. *Journal of the Royal Society, Series B (Methodological)*, 20(2), pp. 215-242.

Do, D.P. et al. (2012). “Does SES explain more of the black/white health gap than we thought? Revisiting out approach toward understanding racial disparities in health”. *Social Science & Medicine*, 74(1), pp. 1385-1393.

Dow, J.K. and Endersby, J.W. (2004). “Multinomial probit and multinomial logit: a comparison of choice models for voting research”. *Electoral Studies*, 23(1), pp. 107-122.

Franks, P. et al. (2003). “Sociodemographics, self-rated health, and mortality in the US”. *Social Science & Medicine*, 56(1), pp. 2505-2514.

Gallagher, J.E. et al. (2016). “Factors associated with self-reported health: implications for screening level community-based health and environmental studies”. *BMC Public Health*, 16(640), pp. 1-15.

Gilmore, A.B. et al. (2002). “Determinants of and inequalities in self-perceived health in Ukraine”. *Social Science & Medicine*, 55(1), pp. 2177-2188.

Grace-Martin, K. (2018). “Logistic Regression Models for Multinomial and Ordinal Variables”. *The Analysis Factor*. Available at: https://www.theanalysisfactor.com/logistic-regression-models-for-multinomial-and-ordinal-variables/

Gouriéroux, C. et al. (1982). “Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters”. *Econometrica*, 50(1), pp. 63-80.

Idler, E.L. (1993). “Age Difference in Self-Assessments of Health: Age Changes, Cohort Differences, or Survivorship?”. *Journal of Gerontology*, 48(6), pp. S289-S300.

Kuo, C-L. et al. (2018). “Unconditional or Conditional Logistic Regression Model for Age Matched Case-Control Data?”. *Front Public Health*, 6(57), pp. 1-11.

Link, B.G. and Phelan, J.C. (1996). “Editorial: Understanding Sociodemographic Differences in Health – The Role in Fundamental Social Causes”. *American Journal of Public Health*, 86(4), pp. 471-472.

Lubetkin, E.I. and Jia, H. (2017). “Burden of disease associated with lower levels of income among US adults aged 65 and older”. *BMJ Open*, 7(e013720), pp. 1-6.

Manor, O. et al. (2000). “Dichotomous or categorical response? Analysing self-rated health and lifetime social class”. *International Journal of Epidemiology*, 29(1), pp. 149-157.

Miilunpalo, S. et al. (1997). “Self-Rated Health Status as a Health Measure: The Predictive Value of Self-Reported Health Status on the Use of Physician Services and on Mortality in the Working-Age Population”. *Journal of Clinical Epidemiology*, 50(5), pp. 517-528.

National Center for Health Statistics. (2018a). “NHANES Demographics Data: Demographic Variables and Sample Weights: Sets B-F (2001-2010)”. *Centers for Disease Control and Prevention*. Available at: https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics

National Center for Health Statistics. (2018b). “NHANES Questionnaire Data: Current Health Status: Sets B-F (2001-2010)”. *Centers for Disease Control and Prevention*. Available at: https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire

Reijneveld, S.A. (1998). “The impact of individual and area characteristics on urban socioeconomic difference in health and smoking”. *International Journal of Epidemiology*, 27(1), pp. 33-40.

Sperandei, S. (2013). “Understanding logistic regression analysis”. *Biochemia Medica*, 24(1), pp. 12-18

Starkweather, J. and Moske, A.K. (2011). “Multinomial Logistic Regression”. *University of North Texas*, lecture slideshow, pp. 1-6. Available at: https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf

Stringhini, S. et al. (2017). “Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women”. *The Lancet*, 389(10075), pp. 1229-1237.

Szwarcwald, C.L. et al. (2005). “Socio-demographic determinants of self-rated health in Brazil”. *Cadernos de Saúde Pública*, 21(1), pp. S54-S64.

Xu, X. and Jensen, G.A. (2006). “Health Effects of Managed Care Among the Near-Elderly”. *Journal of Aging and Health*, 18(4), pp. 507-533.

Žežula, I. (2010). “Logistic, multinomial, and ordinal regression”. *P.J. Šafárik University, Košice*, lecture slideshow. Available at: http://www.karlin.mff.cuni.cz/~antoch/robust10/PREDNASKY/CTVRTEK_DOPOLEDNE/zezula.pdf