Concerning Politics, Are You Happy?

A European Social Survey on How Political Attitude and Behavior Influence Subjective Happiness

Ngan Le	Jesse Haulk	Jingyu Li
ngan3@gatech.edu	jesse.haulk@gatech.edu	alanli@gatech.edu

Abstract Subjective happiness is important for human beings. Our project focus on explore how political attitude and behavior influence subjective happiness based on the data set of European Social Survey. As the response variable is ordinal variable, we apply Ordinal Logistic Regression in our research. Modeling upon the whole data set (~42k observations), we get a full model in which almost all variables are statistically significant. Considering the size of the data set, we apply subsampling method to do variable selection and to check the significance of different explanatory variables in order to achieve more robust results. Based on the results from subsampling, we find that individual's confidence in own ability to participate in politics, general satisfaction on social environment and the degree of political behavior involvement have positive effect on happiness, while participants' general trust to political entities has negative effect on their subjective happiness.

1. Introduction

Along the history, happiness has been considered as the highest good and ultimate motivation for human behavior in philosophy. In modern social science (e.g. Psychology or Sociology), subjective happiness (also defined as subjective well-being) is also a popular concept which is widely researched. Ed Diener reviewed large amounts of previous research on measurement of subjective well-being and proposed that subjective well-being describes individual's overall emotional reactions and cognitive judgements on the quality of their life (Diener, 1984).

1.1. Political Attitude and Behavior on Subjective Happiness

As subjective happiness is an overall perception of life quality, factors related to human's life will have influence on it. Previous research revealed that biological factors (e.g. gene, health, and physical pain etc.), demographic variables (e.g. income, marriage, education and employment etc.), and psychological states and behaviors (e.g. personality, emotion, social involvement, and job satisfaction etc.) are all valid features which have positive or negative effect on subjective happiness (Diener etc., 2002).

Politics is an important issue in human society. National policies and general political environment connect with every individual's life closely. We propose that to investigate how individual's political attitude and related behavior affect their subjective happiness will be an interesting and meaningful topic. Researches on this field are scattered. There is a relative dearth of empirical research on the role of political factors on subjective happiness. The previous research we surveyed investigate the influence related to welfare policy (Pacek & Radcliff, 2008), trust towards political institution (Hudson, 2006) and voting participation (Lorenzini, 2015).

In our project, we aim at investigating a wide range of factors in individual's political attitudes and behaviors and exploring how these factors influence subjective well-being. We try to provide a more comprehensive understanding on the relationship between how people participate in politics in society and their happiness.

1.2. Framework of Model and Variable Measurement

The targeted response variable in our project is subjective happiness. It's a single-item measurement, which asks participants to indicate "how happy would you say you are?" on a 11-point Likert scale (from 0 to 10, 0 represents "Extremely unhappy" and 10 represents "Extremely happy").

As for the independent variables, first, we add several demographic variables as control variables based on the findings from previous research. Those include gender, age, level of education, household's total net income, employed or not and general health. Then the following four dimensions of features on individual's political attitude and behavior are under investigation (*see Appendix I for detailed description on variables and measurement*):

- **Possibility of Political Participation:** four independent items on how people feel they have the possibility and ability to participate in politics.
- **General Trust to Political Entities:** a summed score on the degree of trust towards different political entities (e.g. country's parliament, the legal system, the police and political parties etc.).
- **General Satisfaction on Social Environment:** a summed score on the degree of satisfaction on social environment (e.g. present state of economy in country, the national government and the way democracy works in country etc.).

- **Political Behavior Involvement**: the degree an individual conducting political related behaviors in the past 12 months.

2. Method

2.1. Data Description

2.1.1. Data Source

Data set we use is the European Social Survey Round 8 Data (ESS Round 8, 2016). ESS is an academically driven cross-national survey that has been conducted across 30 European countries since its establishment in 2001. Every two years, an-hour-long face-to-face interviews are conducted with newly selected, cross-sectional samples. In order to achieve as close as possible to a controlled experiment setup, a strict random probability sampling, a minimum target response of 70% and rigorous translation protocols are utilized. This data set contains the results of the 8th round of the ESS surveyed in 2016 and 2017, with a total of 44,387 participants over 535 variables. We select the variables relating to our research problem (*see Appendix I for detailed*).

2.1.2. Data Pre-processing

Processing Missing Value

Participants whose number of missing values exceed 5 (33 features in total) are removed in the first step. The total number of removed observations is 829, 1.87% of the sample size. Then we use mode of the variables to impute the missing value in ordinal variables "psppsgva", "actrolga", "psppipla", "cptppola" and use mean of the variables to impute the missing value in "trstprl", "trstlgl", "trstplc", "trstplt", "trstprt", "trstep", "trstun", "stflife", "stfeco", "stfgov", "stfdem", "stfedu", "stfhlth" (score from 0 to 11, considered as continuous variable in Psychology research). For household income ("hinctnta"), we add a new category called "missing" for the missing values. The number of missing values in other variables (e.g. "gndr", "agea") are relatively small, so we remove them directly (*see Appendix II for detailed*). In total, the data set we use for building the model contains 42,647 observations, 96.1% of the original sample size.

Feature Transformation

According to the design of the survey and other related researches, we sum the score of trust to different political entities ("trstprl", "trstlgl", "trstplc", "trstplt", "trstprt", "trstep", "trstun") as a new variable "General Trust to Political Entities". Also, we sum the score of satisfaction on social environment ("stflife", "stfeco", "stfgov", "stfdem", "stfedu", "stfhlth") as a new variable "General Satisfaction on Social Environment". On the other hand, participants reported whether they involve in different political behaviors in last 12 months. We count the number of behaviors each participant did and assign them into following groups: 0 = Not at all, 1/2 = Low level, 3/4/5 = Median level, 6/7/8 = High level (*see Appendix I for detailed*). Besides, we transform the age into age bins. In total, 13 independent variables are included in the data set for modeling.

2.2. Ordinal Logistic Regression

When the response variable is categorical, a logistic regression model is applied to model conditional probabilities of a certain class or event. Dependent variable can be classified into Binary and Dichotomous, corresponding to three main types of logistic regression.

- **Binary Logistic Regression**: the response variable has only two possible outcomes. For example, when modeling students getting admitted to Georgia Tech, there are 2 possible outcomes, including admitted or rejected.
- **Multinomial Logistic Regression**: the response variable has more than 2 possible outcomes without ordering. For example, when evaluating classes of vertebrates (animals with backbones), there are 5 possible outcomes, including mammals, birds, fish, reptiles and amphibians.
- **Ordinal Logistic Regression**: the response variable has more than 2 possible outcomes with ordering. In other words, there is an association between levels of response variables. For example, when studying severity of Autism Spectrum Disorder, there are 3 levels in increasing order of severity, including “Require Support” (level 1), “Require Substantial Support” (level 2) and “Requiring Very Substantial Support” (level 3) (ASD, 2019).

In this project, the response variable “subjective happiness ” is a discrete variable with multiple ordered levels. Hence, ordinal logistic regression model is selected.

Mathematical Annotation:

Data: $\{(x_{11}, x_{12}, \ldots, x_{1p}), Y_1\}, \{(x_{21}, x_{22}, \ldots, x_{2p}), Y_2\}, \ldots, \{(x_{n1}, x_{n2}, \ldots, x_{np}), Y_n\}$

where $Y_1, Y_2, \ldots, Y_n$ are dichotomous responses with j levels

Model: Probability of being at or below a specific outcome level, conditional on a collection of explanatory variables.

$$p = p(x_1, x_2, x_3, \ldots, x_p) = P_r\,(Y \leq j \mid x_1, x_2, x_3, \ldots, x_p)$$

Link p to the predicting variables through logit link function:

$$g(p) = \ ln\left(\frac{p}{1-p}\right) = \ ln\left(\frac{P_r\,(Y \leq \ j \mid x_1, x_2, x_3, \dots, x_p)}{1 - P_r\,(Y \leq \ j \mid x_1, x_2, x_3, \dots, x_p)}\right) = \ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Note that the formula above uses logit/logistic link function. Ordinal logistic regression can be constructed using different link functions. In this project, we test the 4 most common link functions, including probit, logistic, loglog and cloglog.

2.3. Subsampling Method Over Large Scale Data set

When dealing with a large dataset, conclusions drawn from such datasets can be misleading. Formally, p-values have the limiting behavior of the probability of the distribution of the parameter estimator being the same as the true parameter value (Shmueli etc., 2013).

$$\lim_{n\to\infty} p\text{–}\,value \ = \ \lim_{n\to\infty} P(\left|\hat{\beta} - \beta)\right| < \varepsilon = \{0 \; if \; \beta \neq 0, 1 \; if \; \beta \ = \ 0\}$$

To calculate a p-value, since the data is categorical, the chi-squared test is used and the statistic given by:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

The Chi-squared statistic χ^2 is a summation term. Large test statistics corresponds to small p-value, leading to the rejection of the null hypothesis and the conclusion that the variable is statistically significant. Large sample size can result in misleading interpretation of the test statistics and statistical significance of the variable of interest. With so much data (as n increases), although each individual perturbation (the difference in the observed values (O) and the expected values (E)) is small, the summation of n values of them results in a large test statistic. As a result, with large datasets, all features mostly appear significant, which leads to complex models. To alleviate this issue, the data is subsampled many times to generate a distribution of p-values. If the null hypothesis is true, all p-values are equally likely and would be approximately uniformly distributed. Otherwise, the p-value should be clustered around some small value (Bland, 2013). In this project, subsampling is done using 20% of the data, and run 100 times to see the frequency in which each variable is considered statistically significant.

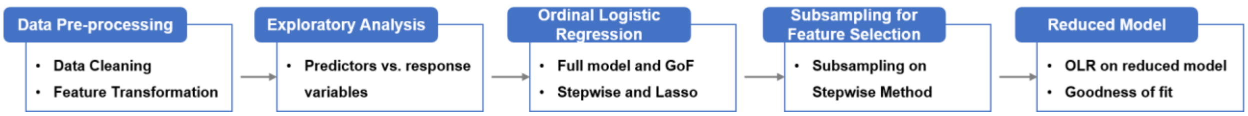
Instead of p-values, confidence intervals can be used for variable's significance testing. Since values within the confidence interval are valid coefficients, a value of 0 within the confidence interval implies that it is plausible that the corresponding variable is not statistically significant. Commonly, the method for determining confidence intervals is referred to as Wald-type Confidence Intervals given by: parameter estimate

± percentile * standard error of the parameter estimate assuming normality. However, the profile likelihood confidence interval does not have a normality assumption of the estimator and performs better if the parameter estimator is skewed (Royston, 2007).

2.4. Data Analytics Procedures

Following are the general procedure in data analytics in our project:

Figure 1. Data Analytics Procedures

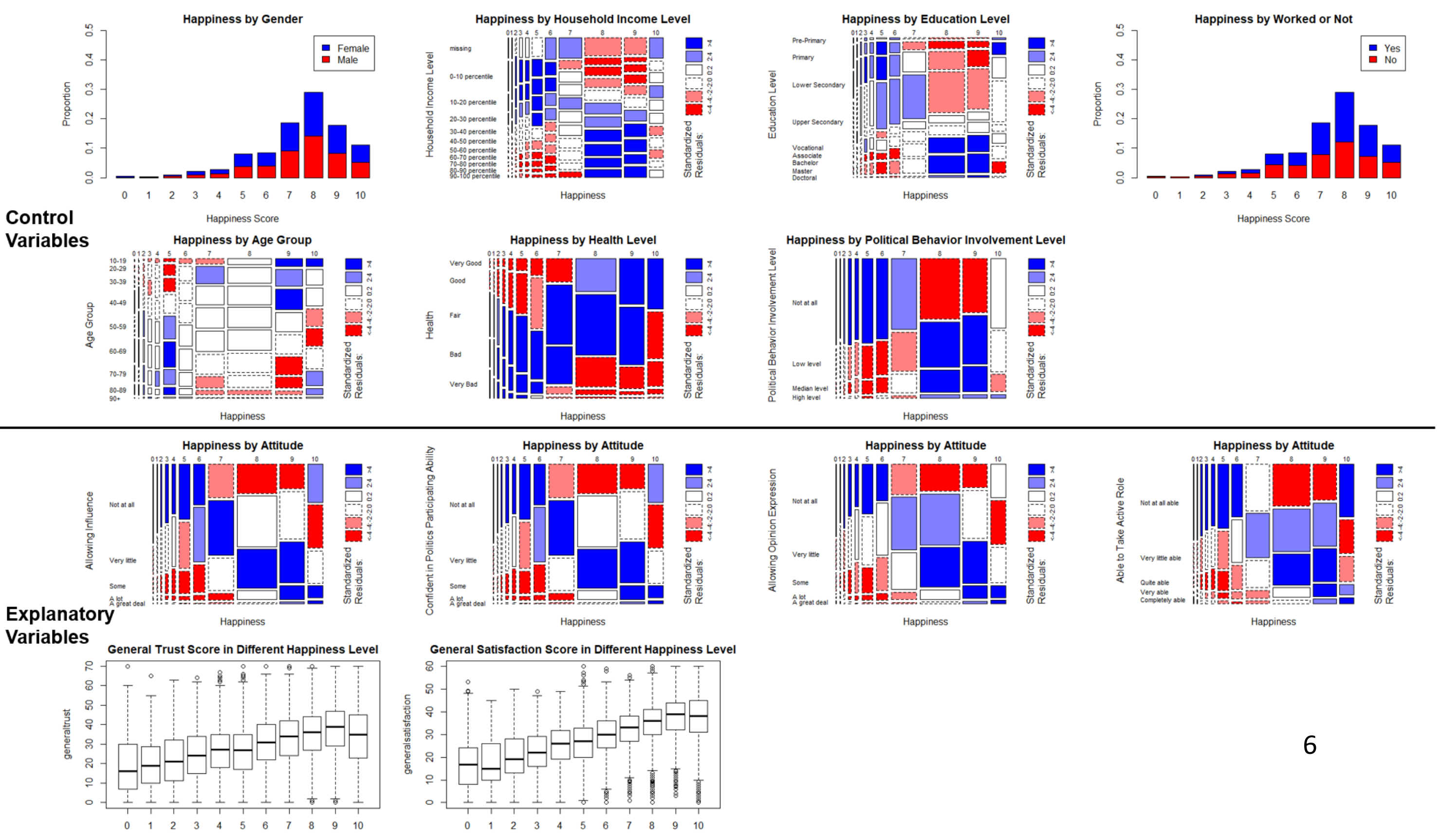


3. Results

3.1. Data Exploratory Analysis

We draw the plot for each control variable and explanatory variable against response variable: subjective happiness. Considering that response variable is an ordinal variable, the mosaic plots indicate that other categorical variables may correlated with subjective happiness. On the other hand, the box plots show that the average value of general trust and general satisfaction (numerical variables) vary among different level of happiness score, which also indicates a potential correlation.

Figure 2. Exploratory Plots for Predictors vs. Response Variable. Mosaic plots are used for categorical predictors while box plots are used for numerical predictors.

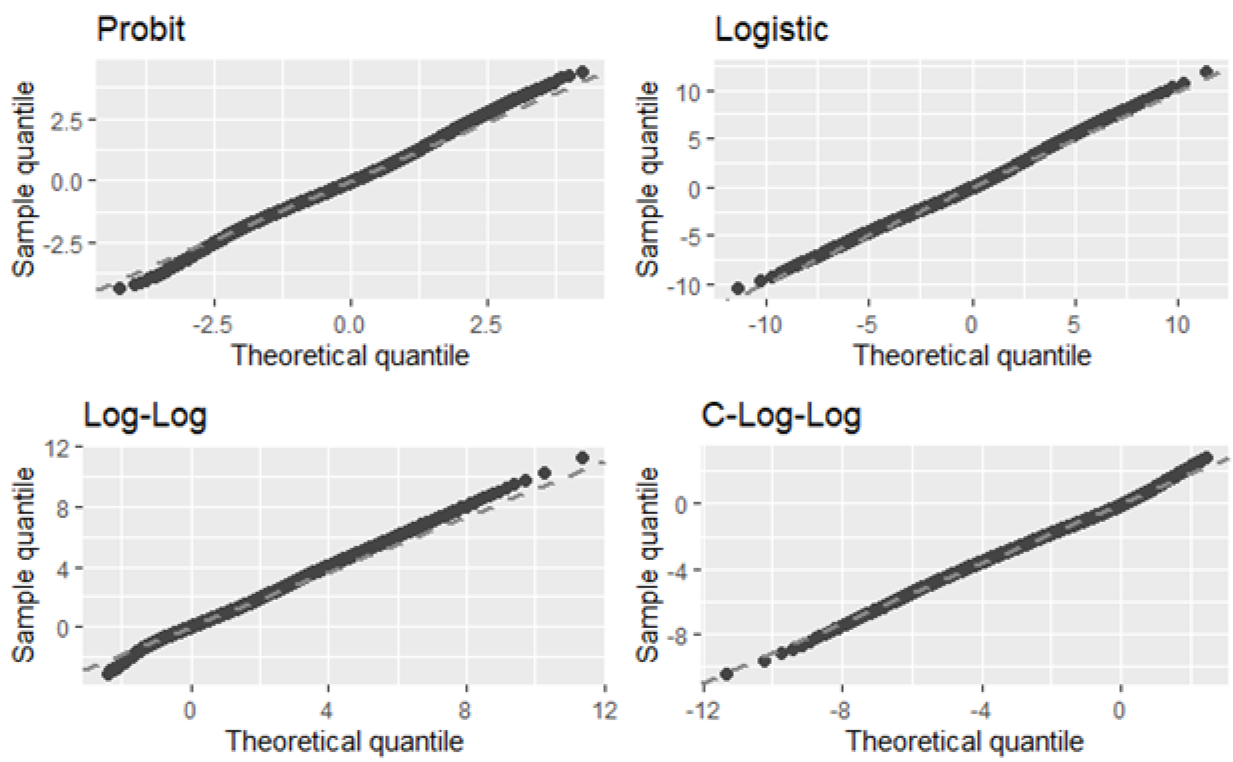


3.2. Ordinal Logistic Regression Modeling

3.2.1. Test different link functions

We fit the model by using different link functions. Evaluating different link functions in ordinal logistic regression can be performed through examining the distribution of surrogate residuals (Boehmke, 2018). From Figure 3, we can see that there is not much difference between different link functions. We choose ordinal logistic regression model with logit link function as it shows surrogate residuals follow closely normal distribution.

Figure 3. Q-Q plots of the residuals for various cumulative link models fit to simulated data with Gumbel (max) errors. Top left: A model with probit link. Top right: A model with logit link. Bottom left: A model with log-log link. Bottom right: A model with complementary log-log link.



3.2.2. Results of Full Model

We build an ordinal logistic regression model with logit link function using the whole dataset (42,647 observations). This model is referred as "Full Model". Evaluating specifications (mean, std, confidence interval and p-value) of variable coefficient βs provides statistical significance of the variable. If a variable has p-value $\leqslant$ 0.05, we conclude that the variable is statistically significant at 95% CI. Based on the results, most variables are statistically significant at 95% CI except for the following 17 variables: agegroup20-29, agegroup30-39, agegroup40-49, agegroup50-59, agegroup90+, edulvlbPrimary, edulvlbBachelor, hinctnta30-40 percentile, hinctnta40-50 percentile, pdwrkYes, psppsgvaA great deal, actrolgaQuite able, actrolgaVery able, actrolgaCompletely able, psppiplaVery little, psppiplaA lot, psppiplaA great deal.

As our project is mainly focus on the explanatory power of political attitude and behaviors on subject happiness. Instead of control variables, we are more interested in explanatory variables. The following table illustrates the coefficients for explanatory variables (*detailed results refer to Appendix IV*).

Table 1. Coefficients and Statistical Testing of Explanatory Variables in Full Model

Variable Name	Value	Std. Error	t value	p value	Significant? (95% CI)
psppsgvaVery little	**-0.091**	**0.027**	**-3.416**	**0.001**	**TRUE**
psppsgvaSome	**-0.23**	**0.031**	**-7.335**	**0**	**TRUE**
psppsgvaA lot	**-0.233**	**0.047**	**-5.005**	**0**	**TRUE**
psppsgvaA great deal	-0.093	0.09	-1.034	0.301	FALSE
actrolgaVery little able	**-0.059**	**0.025**	**-2.341**	**0.019**	**TRUE**
actrolgaQuite able	-0.058	0.031	-1.878	0.06	FALSE
actrolgaVery able	-0.044	0.045	-0.99	0.322	FALSE
actrolgaCompletely able	0.022	0.067	0.33	0.741	FALSE
psppiplaVery little	-0.039	0.026	-1.485	0.138	FALSE
psppiplaSome	**-0.104**	**0.032**	**-3.22**	**0.001**	**TRUE**
psppiplaA lot	-0.072	0.048	-1.488	0.137	FALSE
psppiplaA great deal	0.137	0.107	1.28	0.2	FALSE
cptppolaa little confident	**0.122**	**0.025**	**4.829**	**0**	**TRUE**
cptppolaQuite confident	**0.275**	**0.031**	**8.96**	**0**	**TRUE**
cptppolaVery confident	**0.4**	**0.044**	**9.177**	**0**	**TRUE**
cptppolaCompletely confident	**0.631**	**0.068**	**9.255**	**0**	**TRUE**
generaltrust	**-0.016**	**0.001**	**-17.465**	**0**	**TRUE**
generalsatisfaction	**0.097**	**0.001**	**73.503**	**0**	**TRUE**
behaviorinvolvmentLow level	**0.227**	**0.02**	**11.132**	**0**	**TRUE**
behaviorinvolvmentMedian level	**0.27**	**0.028**	**9.782**	**0**	**TRUE**
behaviorinvolvmentHigh level	**0.566**	**0.064**	**8.781**	**0**	**TRUE**
0|1	-4.012	0.127	-31.667	0	
1|2	-3.388	0.118	-28.804	0	
2|3	-2.572	0.111	-23.12	0	
3|4	-1.69	0.108	-15.63	0	
4|5	-1.042	0.107	-9.731	0	
5|6	0.001	0.107	0.007	0.995	
6|7	0.698	0.107	6.553	0	
7|8	1.783	0.107	16.687	0	
8|9	3.279	0.108	30.492	0	
9|10	4.612	0.108	42.525	0	

3.3. Goodness of Fit on Full Model

There are four main assumptions when evaluating the goodness of fit for ordinal logistic regression.

- **Assumption #1**: The dependent variable should be measured at the ordinal level. In our case, the response variable "happy" is a categorical variable following 11-point Likert scale in which 0 corresponds to "Extremely unhappy" and 10 corresponds to "Extremely happy". Hence, this assumption holds.
- **Assumption #2**: One or more of the independent variables are either continuous, categorical or ordinal. In our case, we have 13 independent variables, including 2 continuous variables, 3 categorical variables and remaining 8 ordinal variables. Refer to Appendix I for more details. As a result, this assumption also holds.
- **Assumption #3**: There is no multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated with each other. One approach is to change type of the dependent variable from ordinal discrete to numeric continuous and create a multiple linear regression model. Then, we can perform a Variance Inflation Factor (VIF) test to check if multicollinearity exists. Since none of the predictors has VIF larger than $max(10, \frac{1}{1-R^2})$ (Table 2), we can conclude that there is no multicollinearity in the dataset and assumption 3 is met.

Table 2. VIF Values for Testing of Multicollinearity Assumption

Variables	GVIF	Df	GVIF^(1/(2*Df))
gndr	1.05	1	1.03
agegroup	2.34	8	1.05
edulvlb	1.62	8	1.03
hinctnta	1.43	10	1.02
pdwrk	1.81	1	1.34
health	1.38	4	1.04
psppsgva	2.85	4	1.14
actrolga	3.40	4	1.17
psppipla	3.34	4	1.16
cptppola	3.27	4	1.15
generaltrust	1.96	1	1.40
generalsatisfaction	1.93	1	1.39
behaviroinvolvment	1.33	3	1.05

- **Assumption #4**: Response variable has proportional odds (PO assumption). This means that the effect of an independent variable on the ordinal dependent variable is uniform over all of the categories or levels of the dependent variable. An example of violation of this assumption would be modeling smoking levels based on age predictor. The dependent variable is different levels of smoking, with the lowest category being zero - no smoking and the highest category being 5 - heavy smoker. A unit increase in predicting variable such as age could have a dramatic effect on whether a person is a smoker or not, which is corresponding to the move from level 0 to a higher level in response variable. However, the unit increase in age has a smaller effect on the move from level 1 to a higher level in the response variable. This is due to the fact that there are no significant differences between different smoking levels once a person is a smoker. In our project, we use function "brant" to test proportional odds assumption. The "omnibus" p-value for our full model is approximately 0, indicating the rejection of the null hypothesis. Hence, we conclude that the proportional odds assumption is not met. However, test of the PO assumption has been described as "anti-conservative, that is it nearly always results in rejection of the proportional odds assumption, particularly when the number of explanatory variables is large (Brant, 1990), the sample size is large (Allison, 1999; Clogg and Shihadeh, 1994) or there is a continuous explanatory variable in the model (Allison, 1999)." (O'Connell, 2006, p29). In our data, we have a large sample size (~42,647 observations) which potentially leads to the violation of PO assumption. Hence, subsampling is needed to make better judgement on the PO assumption.

3.4. Variable Selection on Full Model: Stepwise and Lasso Regression

We first try stepwise and Lasso regression in variable selection. In stepwise regression, both forward and backward method omit the variable "actrolga", which indicates how able participants think they are to take an active role in a group involved with political issues. In the meanwhile, by using Lasso regression, none of coefficients of the explanatory variables are zero. On the other hand, we notice the coefficient of "pdwrk" is zero, but we cannot remove it because it is a control variable in our model (*detailed please refer to Appendix IV*). Stepwise and Lasso seems not efficient enough for variable selection due to the large sample size. We then apply subsampling method to explore more on variable selection and modeling.

3.5. Subsampling Method

By using subsampling, we compare the distribution of p-values of the hypothesis testing on each coefficient (H_0: $\beta_i = 0$) against uniform distribution. Following are the example plots, for detailed output please refer to Appendix III. We also calculate the average confidence intervals for each coefficient (in Table 3). All the results indicate that “actrolga” and “psppipla” are not significant in the model.

Figure 4. Examples of p-value distribution. Shown here ‘Some possibility of Influence’ is not uniformly distributed and more values are clustered in the lower end of the distribution, while ‘Completely able to participate’ is uniformly distributed. This means ‘Some possibility of Influence’ is statistically significant while ‘Completely able to participate’ is not.

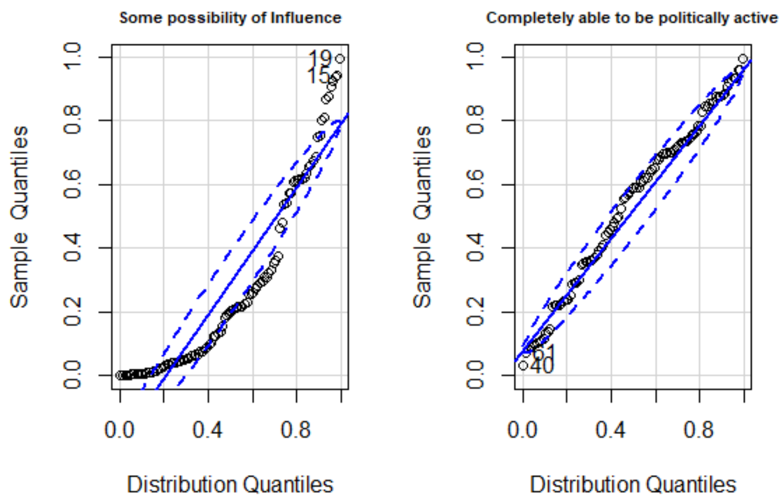


Table 3. Average Value of Coefficients’ Confidence Intervals

Variable Name	Average 2.5%	Average 97.5%
psppsgvaVery little	-0.196	0.041
psppsgvaSome	**-0.361**	**-0.083**
psppsgvaA lot	**-0.451**	**-0.040**
psppsgvaA great deal	-0.414	0.360
actrolgaVery little able	-0.197	0.026
actrolgaQuite able	-0.223	0.052
actrolgaVery able	-0.295	0.098
actrolgaCompletely able	-0.314	0.279
psppiplaVery little	-0.159	0.074
psppiplaSome	-0.232	0.052
psppiplaA lot	-0.267	0.158
psppiplaA great deal	-0.384	0.542
cptppolaa little confident	**0.008**	**0.230**
cptppolaQuite confident	**0.153**	**0.426**
cptppolaVery confident	**0.203**	**0.587**
cptppolaCompletely confident	**0.264**	**0.866**

generaltrust	**-0.021**	**-0.013**
generalsatisfaction	**0.091**	**0.103**
behaviorinvolvmentLow level	**0.194**	**0.379**
behaviorinvolvmentMedian level	**0.245**	**0.482**
behaviorinvolvmentHigh level	**0.247**	**0.678**

From another perspective, we also use sub-samples to run stepwise regression to count how many times each variable is selected. As seen in Figure 5, Political Involvement, General Satisfaction of Social Services, General Trust in Political Entities, and Ability to Participate in Politics were included in all 100 models. The Possibility to Influence Politics and Freedom of Expression were included in less than half, and Ability to be Politically Active close to 10%. This suggests that Ability to be Politically Active is not a statistically significant feature, and likely that the Freedom of Expression and Possibility to Influence Politics are also not statistically significant features. This is consistent with the results indicated by average confidence interval above.

Figure 5. Frequency of Being Selected in Stepwise Regression

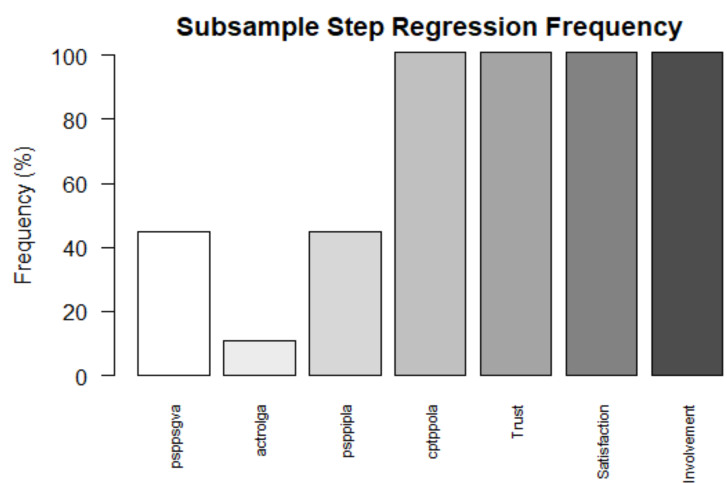
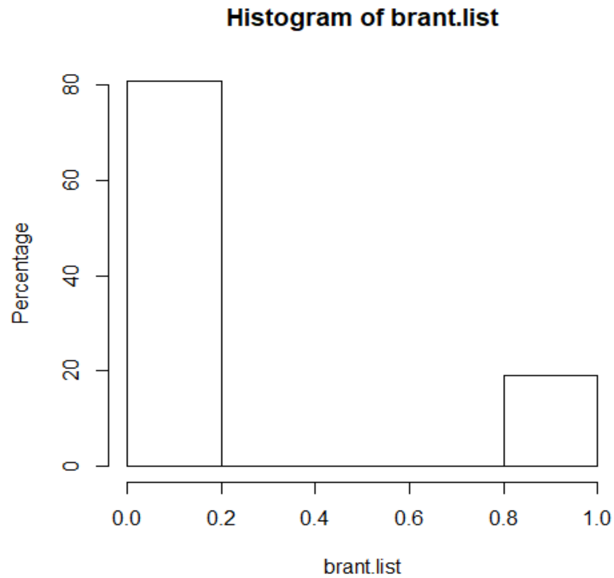


Figure 6. Distribution of Omnibus p-value in brant test in Subsampling. Average omnibus p-value from brant test is 0.20.



As for the proportional odds assumption, we observed the average omnibus p-value of 0.2 which is relatively large. This indicates the model is a good fit.

To summarize, individual's attitude and perception on Allowance in Expression ("psppsgva"), Able to Participate ("actrolga") and Influence Possibility ("psppipla") are not sufficient explanatory variables for subjective happiness. And for the other explanatory variables:

- **Ability to Participate** ("cptppolaa"): individual's confidence in own ability to participate in politics have significant influence on happiness. Compared to baseline (not at all confident), other groups have higher level of happiness. And the higher the confidence, the happier the person will be.
- **General Trust to Political Entities** ("generaltrust"): general trust has a negative coefficient, which indicates that for the individuals who trust the political entities more, they are relatively not happy.
- **General Satisfaction on Social Environment** ("generalsatisfaction"): general satisfaction has a positive coefficient, which indicates that if a person is more satisfied with the social environment, he will also be happier.

  Political Behavior Involvement ("behaviorinvolvment"): compared to baseline (not at all), the coefficients of other groups are positive, which means some degree of involvement will increase subjective happiness.

4. Conclusion and Discussion

4.1. Summary: Political Attitude and Behavior Influence Subjective Happiness

Happiness is important for every human being. Our research focus on how political attitude and behavior influence subjective happiness. With the control variables in the model, we find that individual's confidence in own ability to participate in politics, general trust to political entities, general satisfaction on social environment and political behavior involvement significantly influence subjective happiness. In the meantime, individual's attitude on the degree of political system allows people to have a say in what government does, the degree of being able to take active role in political group and the degree of political system allows people to have influence on politics are not significantly explain the variance in happiness.

The most interesting findings in our project is that general trust to political entities has a negative influence on subjective happiness. We propose that the potential explanation for this is that political entities cannot satisfy all the need and requirement from individuals. Individuals may have negative reactions if their trust is betrayed by the political entities.

4.2. Discussion

There are two tasks we can look into for future work. The first one is goodness of fit - proportional odds assumption. After subsampling, we observe two extreme values of omnibus p-value; omnibus values are either 0 or 1 with the average of approximately 0.20 (as in Figure 6). Although the average omnibus p-value of 0.2 is relatively large indicating the model is a good fit, the fact that individual omnibus p-value is always at the range end prompts more investigation. It is possible that sample size may not be the only reason causing the violation of the PO assumption in the full model.

Recall that rejection of the PO assumption implies that at least one of the explanatory variables may be having a differential effect across the outcome levels. In other words, there exists an interaction between one or more of the independent variables and the derived splits to the data (O'Connell, 2006, p.29). The next step is to determine which variable(s) may be contributing to rejection of this overall test. One reasonable approach is to examine whether the effects of the independent variables are relatively stable or not across the cumulative logits. This can be done through comparison of variable effects across the separate logistic regression models that correspond to the ordinal model being considered as in Table 4. The response variable has 11 levels, corresponding to 10 binary logistic regression models. For each binary logit model jth, the response variable has 2 levels, one includes all “happy” levels smaller or equal to j and one include all “happy” levels larger than j.

Table 4. New approach to evaluate Proportional Odds Assumption

Binary Model No.	New level 1 - Cumulative Odds (Ascending)	New Level 2 - Cumulative Odds (Descending)
1	“Happy” Category 0	“Happy” Category 1 and above
2	“Happy” Category 0 and 1	“Happy” Category 2 and above
...	...	...
10	“Happy” Category 0, 1, .., 9	“Happy” Category 10

For the Proportional Odds assumption to hold, we expect to see relatively similar coefficients for each predictor across all 10 binary models. If a predictor has statistically different values across 10 models, we can conclude that the specific predictor has differential effects across the outcome levels. We can then remove that predictor and re-evaluate the Proportional Odds assumption.

We can also try fitting a multinomial regression model on our data. The multinomial regression model is used to model categorical response variable without order, hence, there is no requirement for the proportional odds assumption to be met. We can repeat the same procedure of goodness of fit, variable selection and subsampling to examine the statistical significance of each predictor to the response variable “happy”. We can compare results from multinomial regression to ones from ordinal logistic regression and see whether the same statistically significant variables are identified.

Secondly, when doing subsampling, the data is only sampled 100 times. Ideally, this number should be much higher. In the future, subsampling the data 1000 or more times would give a better representation of the data without having the misleading p-value issue. Likewise, subsampling for confidence intervals is only done 5 times, but conveys the same information as the p-value distribution which is sampled 100 times.

Reference

Adeleke, K. A., & Adepoju, A. A. (2010). Ordinal logistic regression model: An application to pregnancy outcomes. Journal of Mathematics and Statistics, 6(3), 279-285.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. Sociological methods & research, 28(2), 186-208.

Autism Spectrum Disorder (ASD) A Clear Practical Approach for Parents. (n.d.). Retrieved December 4, 2019, from http://www.childbrain.com/pdd_print.shtml.

Bland M (2013) Do Baseline P-Values Follow a Uniform Distribution in Randomised Trials? PLoS ONE 8(10): e76010.

Boehmke, B., & Greenwell, B. (2018). Introduction to surrogate residuals in R. Retrieved December 4, 2019, from https://koalaverse.github.io/sure/articles/sure.html#references.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 1171-1178.

Brid, R. S. (2018, October 17). Logistic Regression. Retrieved December 4, 2019, from https://medium.com/greyatom/logistic-regression-89e496433063.

Bruin, J. 2006. newtest: command to compute new test. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/stata/ado/analysis/.

Clogg, C. C., & Shihadeh, E. S. (1994). Statistical models for ordinal variables (Vol. 4). Sage Publications, Inc.

Diener, E. (1984). Subjective well-being. Psychological bulletin, 95(3), 542.

Diener, E., Lucas, R. E., & Oishi, S. (2002). Subjective well-being: The science of happiness and life satisfaction. Handbook of positive psychology, 2, 63-73.

ESS Round 8: European Social Survey Round 8 Data (2016). Data file edition 2.1. NSD-Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

Hudson, J. (2006). Institutional trust and subjective well-being across the EU. Kyklos, 59(1), 43-62.

Lee, E. (2019, May 29). Ordinal Logistic Regression on World Happiness Report. Retrieved December 4, 2019, from https://medium.com/evangelinelee/ordinal-logistic-regression-on-world-happiness-report-221372709095.

Liu, X., O'Connell, A. A., & Koirala, H. (2011). Ordinal regression analysis: Predicting mathematics proficiency using the continuation ratio model. Journal of Modern Applied Statistical Methods, 10(2), 11.

Lorenzini, J. (2015). Subjective well-being and political participation: A comparison of unemployed and employed youth. Journal of Happiness Studies, 16(2), 381-404.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables (Vol. 146). Sage.

Ordinal Regression using SPSS Statistics. (n.d.). Retrieved December 4, 2019, from https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php

Pacek, A. C., & Radcliff, B. (2008). Welfare policy and subjective well-being across nations: An individual-level assessment. Social Indicators Research, 89(1), 179-191.

Royston, P. (2007). Profile Likelihood for Estimation and Confidence Intervals. The Stata Journal, 7(3), 376–387

Shmueli, Lin, and Lucas: Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research, Articles in Advance, pp. 1–12, © 2013 INFORMS

Appendix I

Table. List of Variables and Measurement

Feature	Name in Dataset	Type	Measurement
Happiness	happy	Ordinal	11-point Likert scale: 0-Extremely unhappy, 10-Extremely happy
Gender	gndr	Nominal	Male or Female
Age	agegroup	Ordinal	Bins: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+
Education Level	edulvlb	Nominal	Highest level of education: Pre-Primary, Primary, Lower Secondary, Upper Secondary, Vocational, Associate, Bachelor, Master, Doctoral
Household Income	hinctnta	Ordinal	Subjective report on percentile bins compared to society: 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100
Employed	pdwrk	Nominal	Yes or No
Health	health	Ordinal	Subjective report on 5-point Likert scale: 1-Very good, 2-Good, 3-Fair, 4-Bad, 5-Very bad
Allowance in Expression	psppsgva	Ordinal	Political system allows people to have a say in what government does 1-Not at all, 2-Very little, 3-Some, 4-A lot, 5-A great deal
Able to Participate	actrolga	Ordinal	Able to take active role in political group 1-Not at all able, 2-Very little able, 3-Quite able, 4-Very able, 5-Completely able
Influence Possibility	psppipla	Ordinal	Political system allows people to have influence on politics 1-Not at all, 2-Very little, 3-Some, 4-A lot, 5-A great deal
Ability to Participate	cptppola	Ordinal	Confident in own ability to participate in politics 1-Not at all confident, 2-a little confident, 3-Quite confident, 4-Very confident, 5-Completely confident
General Trust to Political Entities	generaltrust	Numeric	11-point Liker scale to measure the trust in country's parliament, the legal system, the police, politicians, political parties, the European Parliament, and the United Nations Sum up the scores
General Satisfaction on Social Environment	generalsatisfaction	Numeric	11-point Liker scale to measure satisfaction with life as a whole, present state of economy in country, the national government, the way democracy works in country, state of education in country nowadays, and state of health services in country nowadays Sum up the scores
Political Behavior Involvement	behaviorinvolvment	Ordinal	Participants report whether or not they did these things in last 12 months: 1) Contacted politician or government official; 2) Worked in political party or action group; 3) Worked in another organisation or association; 4) Worn or displayed campaign badge/sticker; 5) Signed petition; 6) Taken part in lawful public demonstration; 7) Boycotted certain products; 8) Posted or shared anything about politics online. We count the number of behaviors each individual did, and assign them into following groups: 0 = Not at all, 1/2 = Low level, 3/4/5 = Median level, 6/7/8 = High level

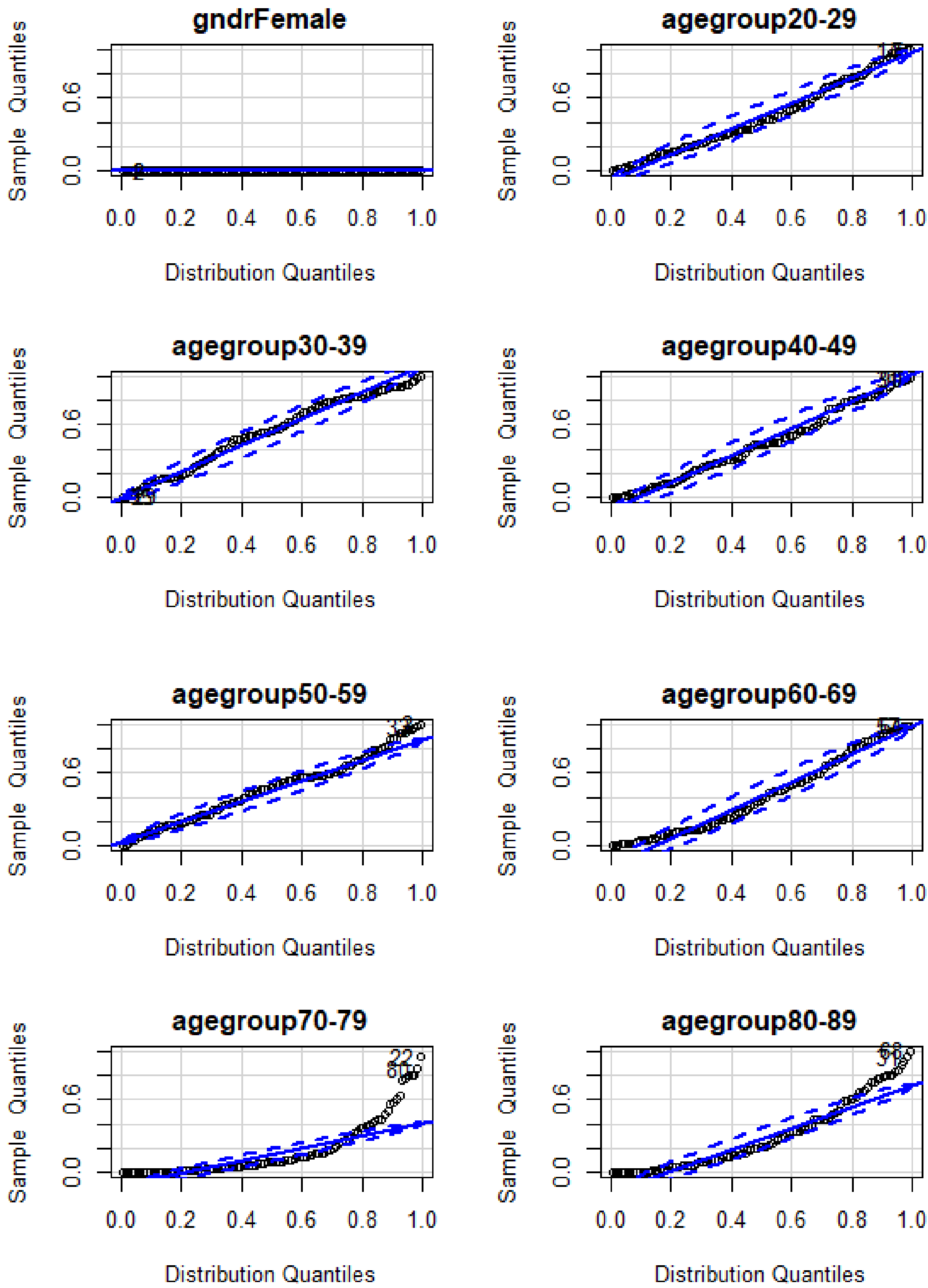
Appendix II

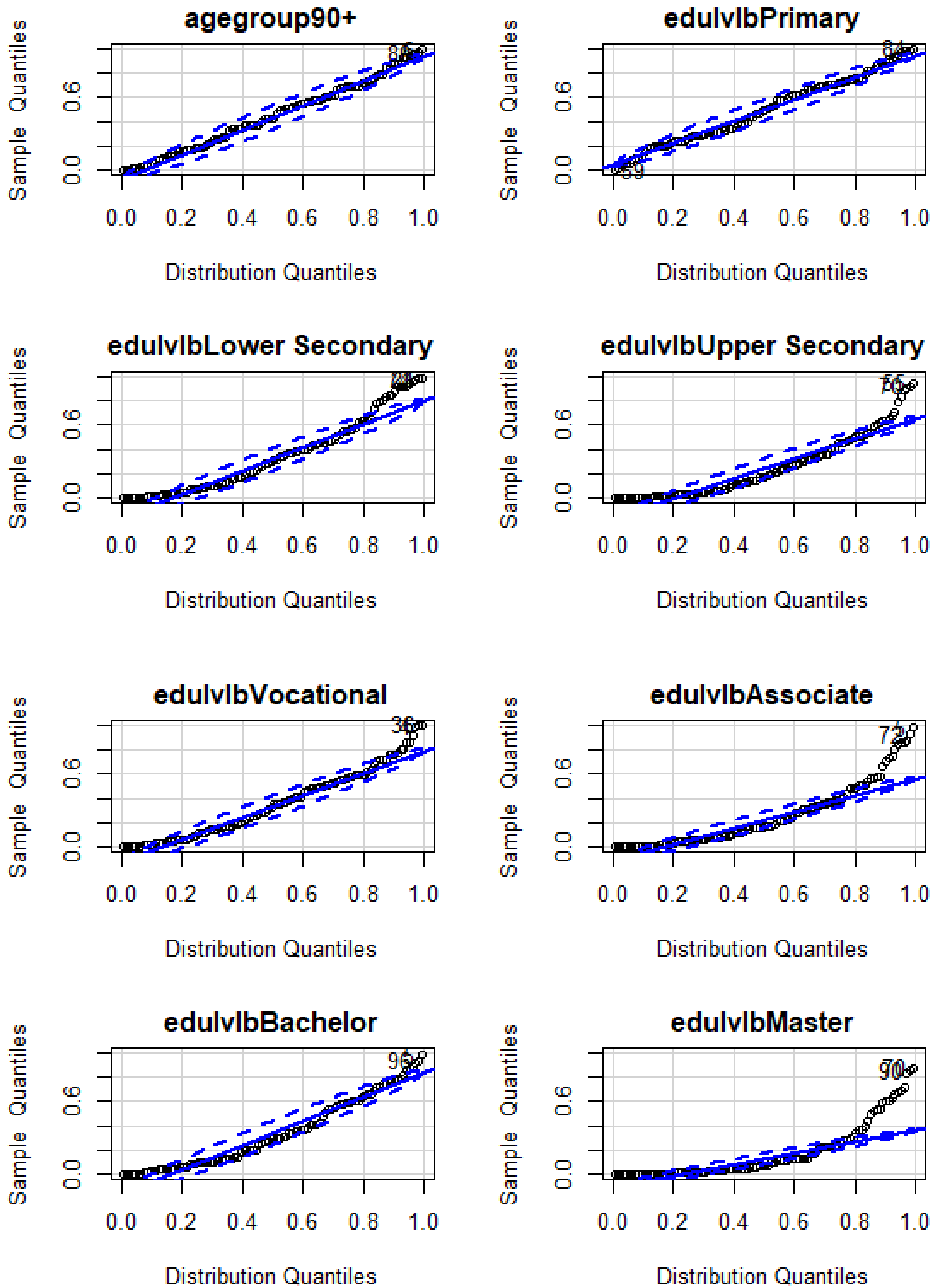
Table. Missing Value Counting and Processing Summary

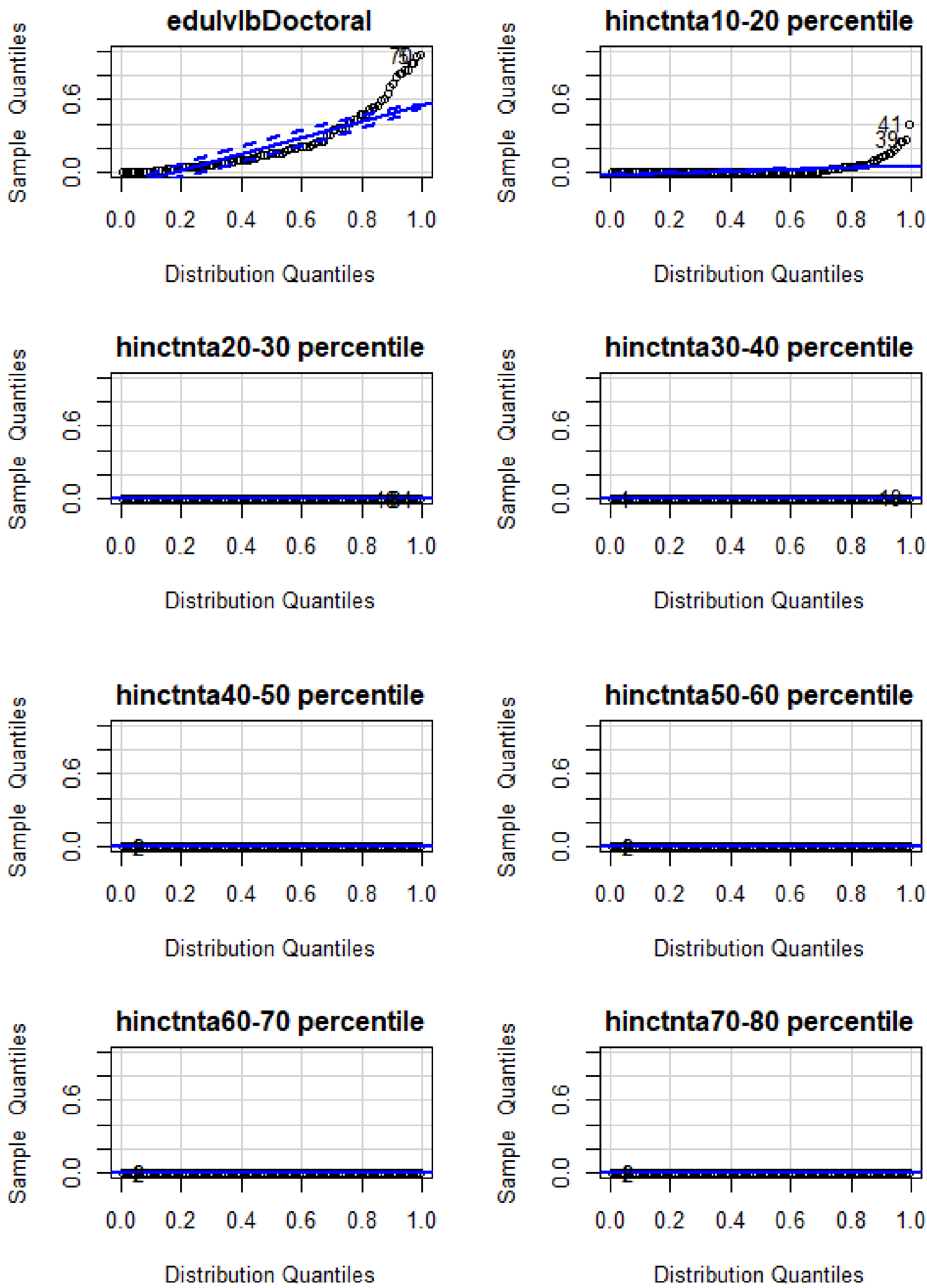
Variables Name	Number of Missing Value	% of Total Sample Size	Processing Method
happy	215	0.48%	Remove
gndr	9	0.02%	Remove
agea	155	0.35%	Remove
edulvlb	217	0.49%	Remove
hinctnta	7942	17.89%	Add a new category: “missing”
pdwrk	0	0.00%	/
health	59	0.13%	Remove
psppsgva	958	2.16%	Imputation with mode
actrolga	935	2.11%	Imputation with mode
psppipla	842	1.90%	Imputation with mode
cptppola	998	2.25%	Imputation with mode
trstprl	873	1.97%	Imputation with mean
trstlgl	849	1.91%	Imputation with mean
trstplc	320	0.72%	Imputation with mean
trstplt	646	1.46%	Imputation with mean
trstprt	885	1.99%	Imputation with mean
trstep	3463	7.80%	Imputation with mean
trstun	3459	7.79%	Imputation with mean
stflife	187	0.42%	Imputation with mean
stfeco	886	2.00%	Imputation with mean
stfgov	1161	2.62%	Imputation with mean
stfdem	1491	3.36%	Imputation with mean
stfedu	1569	3.53%	Imputation with mean
stfhlth	317	0.71%	Imputation with mean
contplt	122	0.27%	Remove
wrkprty	112	0.25%	Remove
wrkorg	129	0.29%	Remove
badge	126	0.28%	Remove
sgnptit	188	0.42%	Remove
pbldmn	135	0.30%	Remove
bctprd	214	0.48%	Remove
pstplonl	176	0.40%	Remove

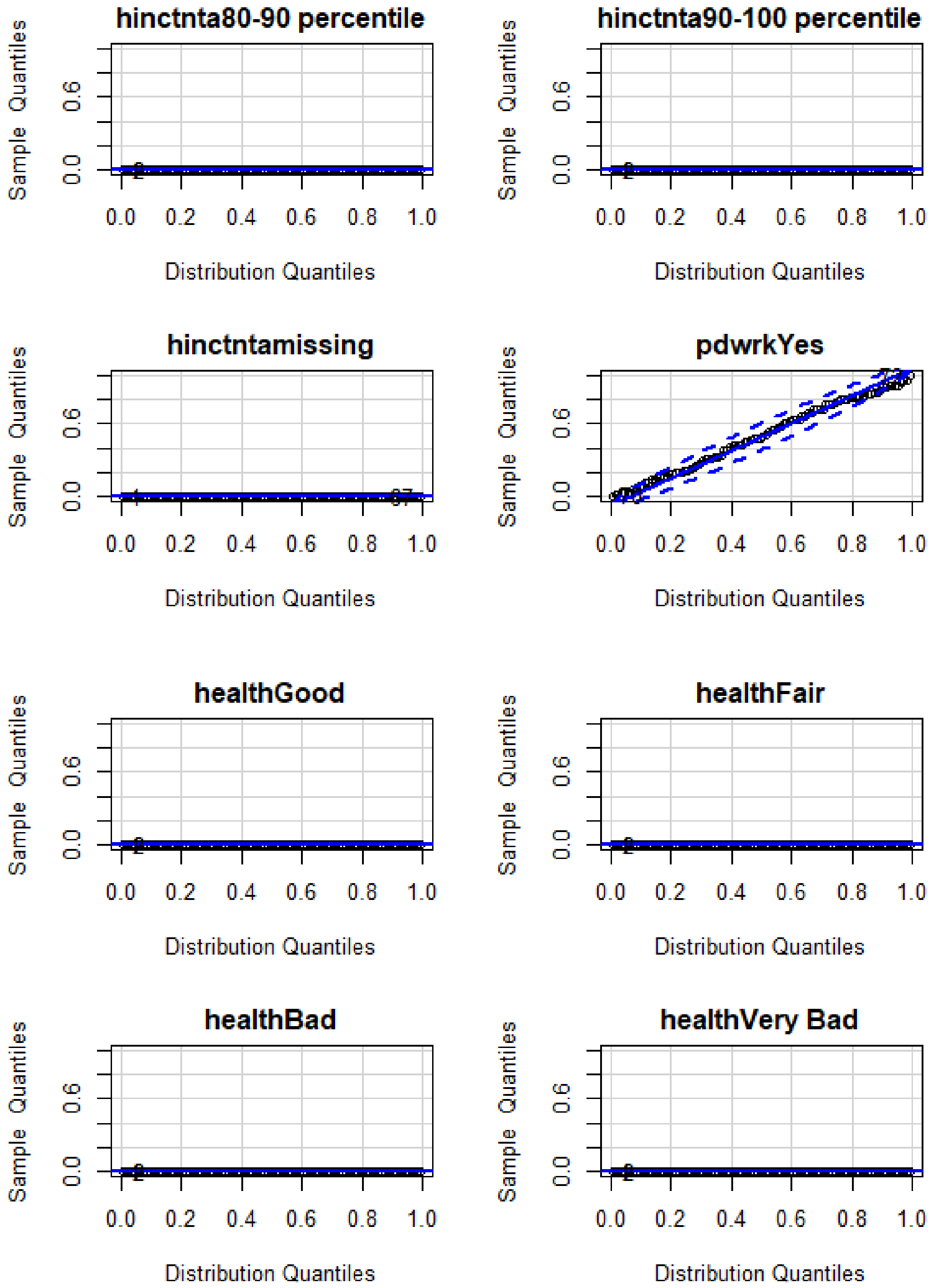
Appendix III

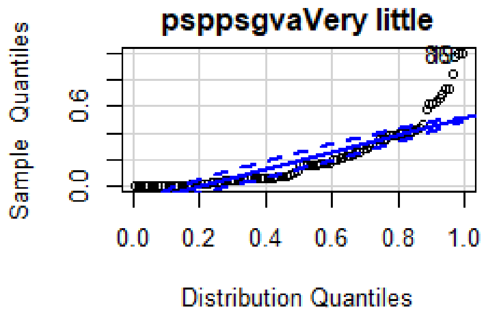
20% of the data was subsampled 100 times to generate a distribution of p-values. The p-values were plotted against the uniform distribution. The p-values should not be uniformly distributed and should be centered around some small value.

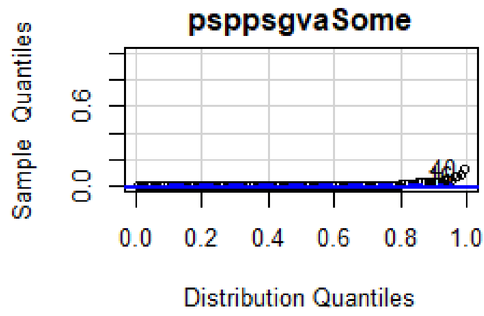


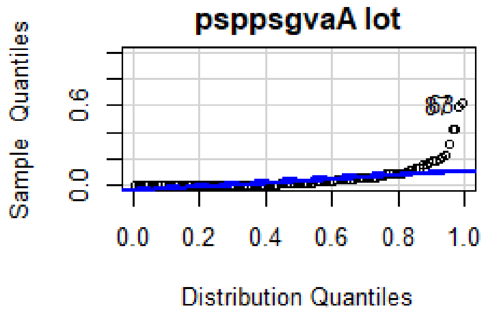

agegroup90+
edulvlbPrimary
edulvlbLower Secondary
edulvlbUpper Secondary
edulvlbVocational
edulvlbAssociate
edulvlbBachelor
edulvlbMaster
Sample Quantiles
Distribution Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0

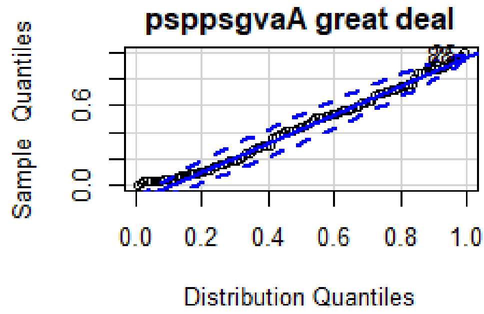

edulvlbDoctoral
hinctnta10-20 percentile
hinctnta20-30 percentile
hinctnta30-40 percentile
hinctnta40-50 percentile
hinctnta50-60 percentile
hinctnta60-70 percentile
hinctnta70-80 percentile
Sample Quantiles
Distribution Quantiles
0.0
0.2
0.4
0.6
0.8
1.0

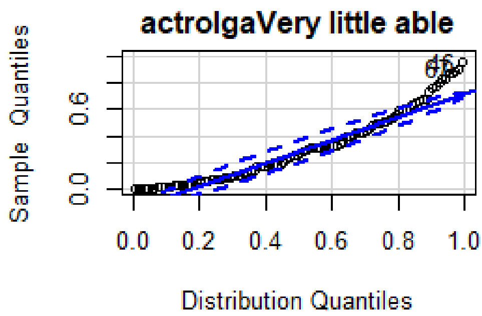

hinctnta80-90 percentile
hinctnta90-100 percentile
hinctntamissing
pdwrkYes
healthGood
healthFair
healthBad
healthVery Bad
Sample Quantiles
Distribution Quantiles
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.6

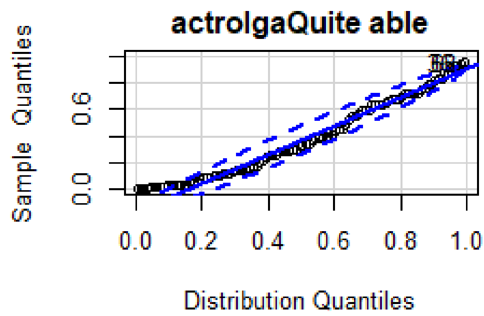

psppsgvaVery little
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles

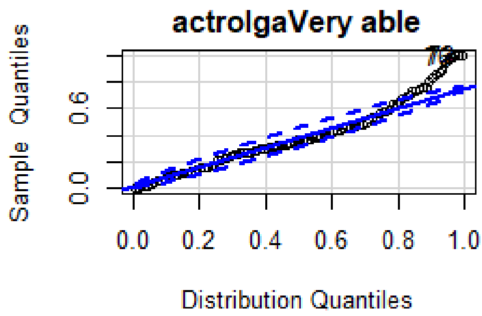

psppsgvaSome
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles

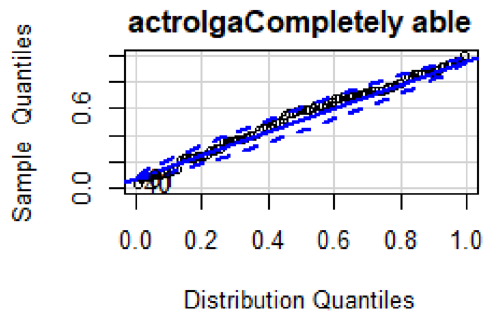

psppsgvaA lot
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles

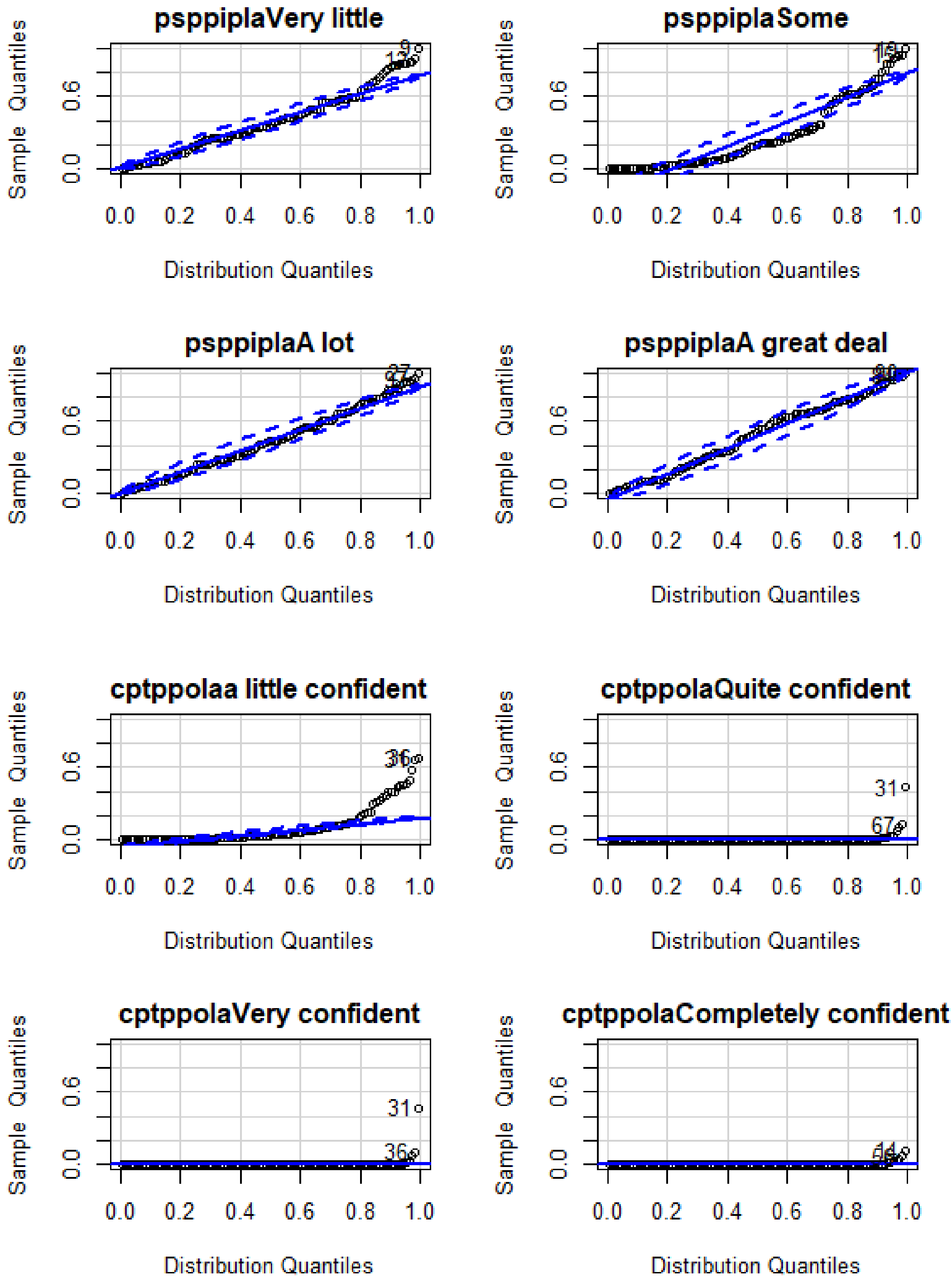

psppsgvaA great deal
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles

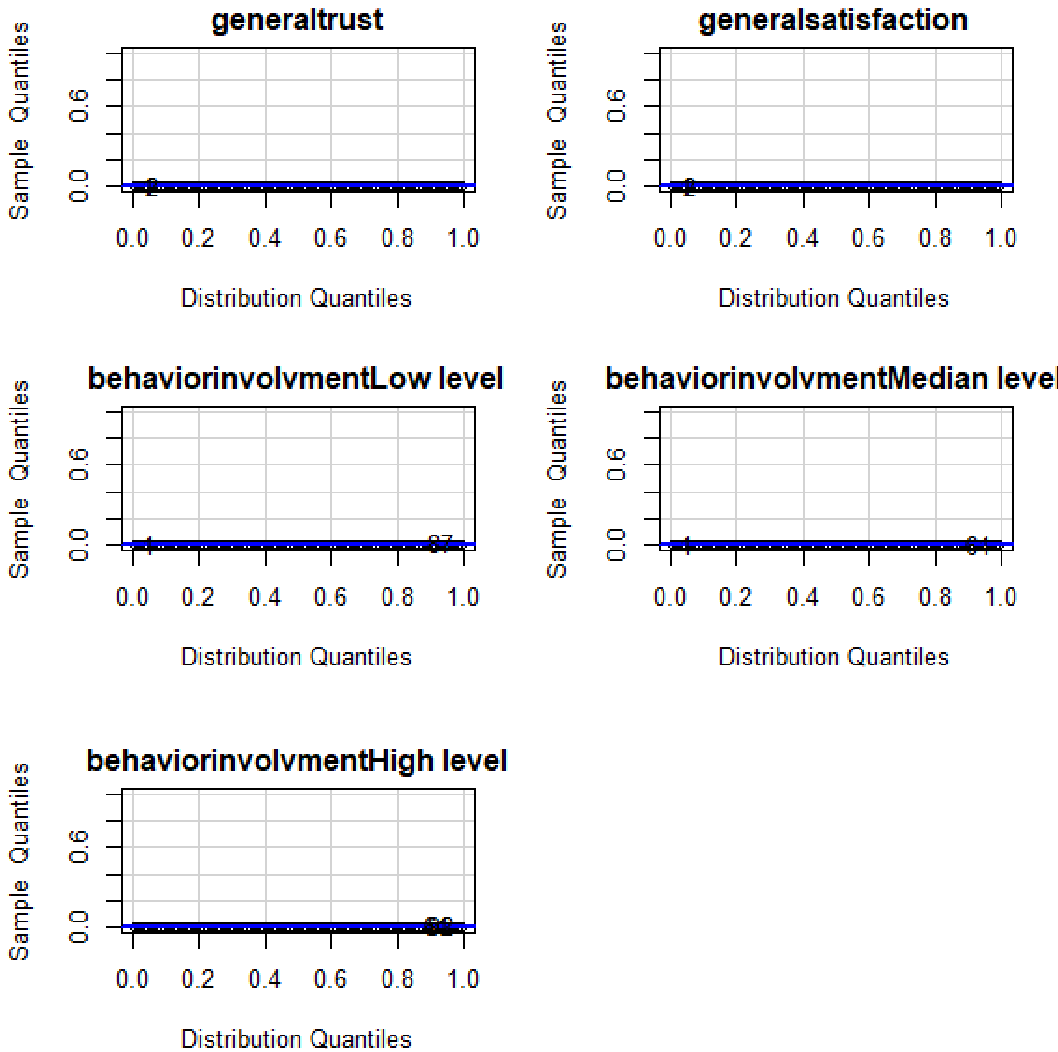

actrolgaVery little able
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles


actrolgaQuite able
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles


actrolgaVery able
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles


actrolgaCompletely able
Sample Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0
Distribution Quantiles


psppiplaVery little
psppiplaSome
psppiplaA lot
psppiplaA great deal
cptppolaa little confident
cptppolaQuite confident
cptppolaVery confident
cptppolaCompletely confident
Sample Quantiles
Distribution Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0


generaltrust
generalsatisfaction
behaviorinvolvmentLow level
behaviorinvolvmentMedian level
behaviorinvolvmentHigh level
Sample Quantiles
Distribution Quantiles
0.0
0.6
0.0 0.2 0.4 0.6 0.8 1.0