Using Sampling Matching Methods to Remove Selectivity in Survey Analysis with Categorical Data

Han Zheng (s1950142)

Supervisor: Dr. Ton de Waal (CBS)
Second Supervisor: Prof. Willem Jan Heiser (Leiden University)

MASTER THESIS

Defended on Month Day, 2019

Specialization: Data Science

STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

A problem for survey datasets is that the data may cone from a selective group of the population. This is hard to produce unbiased and accurate estimates for the entire population. One way to overcome this problem is to use sample matching.

In sample matching, one draws a sample from the population using a well-defined sampling mechanism. Next, units in the survey dataset are matched to units in the drawn sample using some background information. Usually the background information is insufficiently detaild to enable exact matching, where a unit in the survey dataset is matched to the same unit in the drawn sample. Instead one usually needs to rely on synthetic methods on matching where a unit in the survey dataset is matched to a similar unit in the drawn sample.

This study developed several methods in sample matching for categorical data. A selective panel represents the available completed but biased dataset which used to estimate the target variable distribution of the population. The result shows that the exact matching is unexpectedly performs best among all matching methods, and using a weighted sampling instead of random sampling has not contributes to increase the accuracy of matching. Although the predictive mean matching lost the competition against exact matching, with proper adjustment of transforming categorical variables into numerical values would substantial increase the accuracy of matching. All the matches are used in reducing overfitting of machine learning, and the results show that all matches are able to increase the prediction precision.

Contents

1 Introduction

In survey sampling, population characteristics can be estimated by using probability sampling due to the randomization principle which guarantees that each individual of the population has a known and non-zero selection probability. Therefore, probability sampling can generate unbiased estimates for population quantities. However, if the dataset is gathered from a special group within the population, the dataset may not be representative which may lead to biased estimates. A dataset like this is called a selective dataset. For example, an online survey result could be selective if the target population contains people who are unable to get access to the internet, and the sample from a specific region could also be selective to estimate the population of the whole country. A selective dataset can lead to a biased estimate and affect policy decisions. In order to reduce the selectivity, the sample from the population of a survey should be wide and representative. The nonresponse problem can also lead to the issue of selectivity. Nonresponse means that no information is obtained from a number of elements in the sample (Bethlehem, 2015). Usually, nonresponse can be modeled by assigning a probability of response to every element in the population based on the background variables (Bethlehem, 2015). However, since the probability of getting an observation is unknown and the background variables of nonresponding units may also be missing, it is difficult to model the nonresponse probability based on background variables. When the nonresponse probability is correlated with background variables, the available feedback of a survey could result in a very selective group of the population. Hence, it becomes impossible to produce unbiased and accurate estimates for the entire population.

The best known example of selectivity is the Survivorship Bias, which is the logical error of concentrating on the units that passed a certain selection process and overlooking those that did not, typically because of their lack of visibility. The logical error can lead to false conclusions in several different ways, which is a form of selection bias. The general idea to overcome the selectivity issue is to remove the bias during the selection of target units and ensure the selected units are as representative as possible. This selection process requires that during the collection of research units, researchers need to take conditions into account that could affect the representation of the population, like the regional effect, age composition and others. However, if there are only selective, or biased targets available, one can remove the selectivity through matching. The matching technique is widely used in clinical research (Chintan, et al., 2017). The majority of clinical studies are implemented in controlled experiments. In controlled experiments one divides research participants into a treated group and a control group; the treated group is treated with treatment and the control group is left blank or treated with a placebo. The goal of matching is, for every treated participant, to find one (or more) non-treated participant(s) with similar observable characteristics against whom the effect of the treatment can be assessed. By matching treated participants to similar non-treated participants, matching enables a comparison of outcomes among treated and non-treated units to estimate the effect of the treatment, thus reducing bias due to confounding. In the case of removing selectivity, one can match a representative dataset to a selective dataset based on similar observable characteristics, locate units in the selective dataset which are most similar to the units in the representative dataset, and use these units for obtaining population estimates.

Bethlehem (2014) has implemented a pilot study by using the matching method to solve the non-response problem in a survey. In order to simulate the response distribution, Bethlehem matched individuals in a group of people who possess high probability of response in the survey with individuals in a group of people with low probability of response, and asked matched individuals in the high response rate group to complete the questionnaire instead of the low response individuals. His simulation study showed that the rate of response substantially shifted to a higher rate through matching. Following his research, we developed the idea of using matching to remove the selectivity issue in survey analysis. If a dataset of the research is selective, one can match all units in the population to the units in the selective dataset to generate a representative dataset, and implement probability sampling from the representative dataset to estimate the target parameters. However, matching each unit in the population to the selective dataset is time-consuming; therefore, the sample matching method is more suitable to solve this problem. The aim of sample matching is to remove the selectivity in the sample, and to reduce the bias of population estimates. In sample matching, one draws a sample from the population using a well-defined mechanism. Next, units in the selective dataset are matched to the units in the sample using background information. Finally, the sampling weights obtained from the sampling mechanism are used to weight the information from the selective dataset in order to obtain estimates for population totals.

Why sample matching?

The first question that people might come up with is why use matching to estimate the target parameters. Usually, building models based on the output variable and predict the population distribution is a more reliable method, because regression analysis uses the correlation relationship between background variables and output variables. However, in both regression and machine learning models, it is assumed that the part of the data that is used to fit the model has the same distribution as the test set on output variable Y. If the distribution of the output variable of the training set and test set are different, precision of the model prediction will decrease due to the bias. If the survey data are collected through a selective dataset, the distribution of the output variable will be biased. And if the models are developed through the selective dataset, the prediction on the test set will be biased. Under this circumstance, if the model was optimized to have lower loss on the training dataset, it will result in overfitting to the selective dataset. In order to reduce the overfitting caused by the difference between the distribution of the training set and the test set, one can use matching to shift the training distribution close to the test set distribution.

There are several methods to reduce overfitting in maching learning in terms of manipulation of the distribution of dataset. One of the most commonly used methods is data augmentation. The general idea of data augmentation is to enlarge the frequency of the units which do not very commonly appear in the dataset, therefore the dataset will become a large representative dataset which reduces overfitting. Enlightened by this idea, in the scenario that the training set is selective, one can shift the distribution of the training set in a similar way to the test set by matching similar units in the training set to the test set, and train the model based on the adjusted training set.

1.1 Principle of Sample Matching

Sample matching is a two-stage process: first by applying the principle of probability sampling a random sample is drawn from the population which is called the target sample. The values of the target variables are missing partly or totally for this target sample. Step two is matching each unit in the target sample to the selective dataset with outcomes, which is called the panel. In order to find similar units in the panel, a set of auxiliary variables are required: these variables are known in both the population and the panel, and are called background variables (Bethlehem, 2015). Sample matching is to locate units in the panel which are similar to the units from the sample, generating a matched dataset which can simulate the population distribution (Rivers, 2007).

The key of sample matching is locating the most similar unit in the panel for each unit in the sample. Therefore the most important step of sample matching is the matching (Rivers, 2007). Ideally, by sample matching a selective dataset can be shifted to a representative sample of the population. We assume that the linear correlation between the target variable and background variables is the same in the population and the panel, thus theoretically by matching the sample and the panel, we are able to shift the distribution of the target variable to the distribution of the sample frame (this depends on the quality of the matching). In order to find the most similar unit in the panel, the background variables are utilized to quantify the similarity between two objects, the most commonly used measurement method is the use of similarity functions like Euclidean distance, Manhattan Distance or Cosine similarity. For a numerical dataset, measuring the similarity consists of calculating the distances between units and selecting the unit with minimum distance. For categorical variables, the usual method is to find the unit which has the maximum number of background variables with the same category. However, this could results in a huge number of candidates who possess the same values for the categorical variables. Therefore, developing methods to quantify the similarity in a categorical dataset is essential for sample matching.

1.2 Matching Methods for Categorical Variables

Rivers (2007) has summarized the general methods for matching: exact matching, proximity matching and propensity score matching. Proximity matching for numerical data is the only method that has been widely used. However, there are no research works implemented on categorical datasets. In this study, we will develop several sample matching methods in sample matching with categorical variables.

In most cases, similarity between units is measured by distance functions like the Euclidean distance, Manhattan distance, etc. However, calculating distance functions is not viable for categorical variables datasets, therefore other methods are needed. Exact matching is to match units with the same values of auxiliary variables, which might appear to be the most suitable method for categorical variables but has too many limitations. Propensity score matching is to calculate the propensity scores of all units, and match units which have similar values of propensity score. Proximity matching is also viable when proper numerical values are assigned to the units and the distance function can be calculated, which is also known as scaling.

1.3 Research Outline

The goal of this research is to develop sample matching methodology to remove the selectivity of a categorical dataset, and observe whether sample matching can remove the selectivity phenomenon in a population estimate. Also, we would like to compare whether the prediction accuracy can be improved by using sample matching to remove the selectivity of the training dataset. We assume that the selective dataset is complete on both target variable and background variables. The selected sample is representative but the values of the target variable are missing. Statistically, the traditional regression method or machine learning can be used to predict the population estimate for a target variable if the training set of the data is also representative. However, if a selective dataset is used as the training set of the model, it is believed that the prediction of the model will be biased. In this research, we will solve this problem by using the sample matching method, and develop different matching methods for categorical datasets to obtain the population estimate. We will also try to reduce the prediction bias and improve the prediction accuracy of the model by using sample matching to remove the selectivity of the training dataset.

We are interested in how the following factors affect the performance of removing selectivity: (i) the size of the selective dataset and the population, (ii) the methods of matching, (iii) the methods of drawing the sample from the population. In this research, we will design and implement an experiment based on these factors, compare the results and select the best combination for removing the selectivity through sample matching.

The thesis is organized in the following way: Chapter 2 will explain the situation of selectivity. Chapter 3 will explain the developed matching methods and design of the experiment. The matching methods include exact matching, proximity matching and propensity score matching. A comparison of sampling methods (weighted sampling) is also designed and explained in Chapter 3. In chapter 4, we use sampling matching to reduce the selectivity of the training dataset and test dataset, build prediction models based on the matched dataset and selective dataset, and compare the two population estimates to observe whether sampling matching can reduce the prediction bias due to the selectivity of the training dataset. Chapter 5 will present the comparison of the prediction model and sample matching methods, and also gives the conclusion and discussion of the study.

2 Removing the Selectivity Problem

2.1 The Selectivity Problem

The selectivity problem can be caused due to several reasons in the survey. Normally, two of the most common reasons are that the panel data is selective due to different participation of the population, and the nonresponse problem.

2.1.1 A Selective Panel

In order to study a target feature of the population, conducting survey sampling is often required. The most effective and economical method for survey sampling is sampling from online information. However, an online dataset may be selective due to the different possibilities for participating for different people, which can generate a selective dataset. Most online panels are selective which can resultf in biased estimates. For example, the result of an online web survey from social applications like Facebook or Twitter will be selective since the target population through web survey only covers people who are able to get access to the internet. The composition of the demography is assumed highly selective on younger aged people; therefore, using an online panel to estimate a target variable $\boldsymbol{Y}$ can be biased. Assuming that each unit i in the population has an unknown probability ρ_i to use the online social application, the expected sample mean of the target variable is given by

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{N} \rho_i y_i, \tag{1}$$

where y_i is the value of the target variable for unit i, N is the population size and $n = \sum_{i=1}^{N} \rho_i$ is the expected sample size.

If the selectivity problems do not exist, the expected sample mean of a numerical target variable would be

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{2}$$

It is expected that for different social applications the composition of age would be different. For example, young students are expected to be the majority users of Facebook, and the majority of the users of a babysitting website could be middle aged working people; also information from an old people‘s house is mostly provided by retired old people. The difference is caused by the different demographic ratio of the dataset, which caused the selectivity of the dataset.

2.1.2 Nonresponse Problem

Another situation which can cause selectivity is the non-response problem. In survey sampling, if the nonresponse cases are random, it will not affect the distribution which will not cause selectivity.

However, most of the nonresponse cases are related to several background variables, which can lead to biased estimates similar to the online survey problem.

Assigning a response probability r_i to each unit i of the population, the expected sample mean of target variable is therefore given by

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{N} r_i y_i, \tag{3}$$

where $n = \sum_{i=1}^{N} r_i$.

And if the nonresponse problems do not exist, the expected sample mean of a numerical target variable would be

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{4}$$

In a nonresponse case, the selectivity is caused by different rates of response for different groups, i.e, the distribution of the background variables is selective. The difference between the nonresponse problem and the fundamental problem with a selective panel is that in the nonresponse problem it is usually known who responded, but it is often unknown who participated in a selective panel.

2.1.3 Bias Caused by Selectivity

In machine learning, the object is to study and construct algorithms that can make predictions on input data. Such algorithms work by making data-driven predictions through building mathematical models from input data (Bishop, 2006). The data used to build the final model usually comes from multiple datasets (James, 2013): training set, validation set and test set. The training set is the data set that allows one to learn the model and fit parameters, the validation set is used to tune the hyperparameters of a classifier or regression model. A test dataset is a dataset that is independent of the training set but follows the same probability distribution as the training set. A good classifier requires to fit both the training set and test set well, which allows minimal overfitting to take place. Apart from the test set, the prediction of models also assumes that the dataset of the predictions follows the same probability distribution as the training set. If the distributions of the training set and test set are different, the trained model will be overfitted to the biased training dataset and result in biased prediction estimates.

In an online survey case, the available training dataset which is used to fit the models is selective while we would like obtain an unbiased estimate. In this senario, it is highly likely that the prediction on the population results in a huge bias due to the fact that the training dataset and test dataset follow different distributions. In deep learning, data augmentation is used to enlarge the training dataset in order to enable the training set to cover as many categories as possible. For example, in images one can rotate the image, change the lighting conditions to crop it differently. In this way the training dataset is enlarged and the overfitting of a classifier will be reduced (Gareth, 2013). In our case, we have a selectivity panel, which results in a selective training dataset while the test set (population) is representative. One can use sample matching to change the distribution of the training set as closely as possible to the distribution of the test set, and use the matched dataset as the training set to fit the model. In this way, we remove the difference between the training and test set distribution, the overfitting will be reduced and the precision of the prediction is expected to improve.

2.2 Modeling the Bias due to Selectivity

2.2.1 Bias caused by selective panel

Suppose the survey is conducted from a selective panel, the expected mean value of the target variable is denoted as:

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{N} \rho_i y_i, \tag{5}$$

where N and $n = \sum_{i=1}^{N} \rho_i$ represent the size of the population and expected size of the panel respectively, ρ_i represents the probability that the panel covers unit i in the population, and y_i stands for the value of the target variable. In other words,

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\rho_i}{\bar{\rho}} y_i, \tag{6}$$

where

$$\bar{\rho} = \frac{1}{N}\sum_{\kappa=1}^{N} \rho_\kappa \tag{7}$$

is the mean of the ρ_κ over all individuals κ from the population. Bethlehem, Cobben and Schouten (2011) showed that the bias of the estimate based on the panel is equal to

$$B = \frac{R_{Y\rho} S_\rho S_Y}{\bar{\rho}}, \tag{8}$$

where $R_{Y\rho}$ is the correlation between the probability of being covered by the panel and the value of the target variable (Bethlehem, Cobben and Schouten, 2011). In this case the probability of being covered by the panel is the probability of using the online social application. S_Y is the standard deviation of the target variable Y, and S_ρ is the standard deviation of the probability of using social apps. A number of conclusions can be drawn from the expression of the bias function: The bias of the estimate will be large if there is a large correlation between the target variable and the probability of using the social apps. And the bias is small if the probability of being covered by the panel is large. There will be no bias if there is no correlation between the target variable and the probability of using social apps.

2.2.2 Bias caused by nonresponse

Assume that a survey was conducted and there is nonresponse in the survey, the expected mean value of the target variable is denoted as:

$$E(\hat{y}) = \frac{1}{n}\sum_{i=1}^{N} r_i y_i, \tag{9}$$

where N and $n = \sum_{i=1}^{N} r_i$ represents the size of the population and panel respectively, r_i represents the probability of response for each unit i in the population, and y_i stands for the value of target variable. That is,

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} \frac{r_i}{\bar{r}} y_i, \tag{10}$$

where

$$\bar{r} = \frac{1}{N}\sum_{\kappa=1}^{N} r_\kappa, \tag{11}$$

is the mean of response probabilities over all individuals in the survey population. The bias of the estimator is equal to:

$$B = \frac{R_{Yr} S_r S_Y}{\bar{r}}, \tag{12}$$

where R_{Yr} is the correlation between the value of the target variable and the response probability. S_r is the standard deviation of the response probabilities and, S_Y is the standard deviation of the target variable (Bethlehem, Cobben and Schouten, 2011). From the expression of the bias we come to the conclusion that the bias is large if the correlation between the target variable and response probability is large. The bias will be small if the average response rate is large.

Summarizing, in order to reduce the selectivity of a survey from an online dataset, it is essential to use panels which are as representative as possible. But the problem of selectivity cannot be avoided since panels always represent part of the population. Another suggestion for panel selection is to use only panels where participation is hardly related to the target variable.

2.3 Removing Selectivity

The assumption is that the selectivity issue is caused by the skewness of background variables, thus the main idea of removing the selectivity of the distribution of a target variable is to eliminate the selectivity caused by the background variables. For example, when the selectivity of the survey population distribution is caused by the different participation probabilities of the background variables, say age, one can remove the selectivity by using post stratification. Consider a survey that studies a population with size N with target variable y and estimates its mean value $\hat{y}$ from an online panel, say Facebook for example. The selectivity is caused by the fact that differently aged people have different probabilities to use Facebook, say with age category $1, 2, ...C$. The distribution of age in the population is $N_1, N_2...N_C$. A simple random sample from the Facebook panel would results in a skewed age group distribution $n_1, n_2...n_C$ which affects the mean estimator. In order to remove the selectivity in the online panel, one can use post stratification. The population mean then is estimated by post stratification estimator $\hat{y_{st}} = \frac{N_1}{N}\hat{y_1} + \frac{N_2}{N}\hat{y_2} + ... + \frac{N_C}{N}\hat{y_C}$, where $\hat{y_1}, \hat{y_2}...\hat{y_C}$ are the observed mean values in the Facebook panel for age groups 1 to C.

Ideally, by using post stratification the bias caused by background variables is removed (Bethlehem, 2015). However, in real life cases, there could be more than one background variable which caused the selectivity of the target variable in a selective dataset, and it is also difficult to confirm which background variable caused the selectivity problem. When there are multiple background variables affecting the representation of the target variable, it will be difficult to conduct post stratification. Therefore, post stratification is hard to implement under this circumstance. In order to remove the selectivity by generating a representative sample from the selective dataset based on known background variables, one way is to use sample matching. In sample matching, one draws a sample from the population with a specific mechanism, and matches each unit from the sample to a unit in the panel based on background variables.

In an online web survey, sample matching is a purposive method to generate a sample when a large but not representative response dataset is available. Implementation of sample matching requires two ingredients (Vavreck and Rivers, 2008):

(i) A sampling frame: the sampling frame is required to cover the target population of the survey. The sampling frame is also required to contain a set of auxiliary variables (background variables) for each individual. The set of auxiliary variables should also cover the set of auxiliary variables of the selective dataset.

(ii) A large panel: usually the panel is selective either because of the nonresponse problem or a selective group of the population participates. The panel contains auxiliary variables which can be used as background variables for matching, also the value of the target variable is observed for every unit.

Applying the principle of probability sampling, a sample is selected from the frame. Each unit in the sample is matched to the most similar unit in the panel based on the auxiliary variables. The selected units in the panel are used to estimate population quantities (Rivers and Bailey, 2009).

2.4 Matching Methodology

2.4.1 Exact Matching

The principle of exact matching is simple: find the units in the panel whose values or categories of background variables are exactly the same as a unit from the sample (Chmura, et al, 2013). Exact matching is the most precise matching method for categorical variables but highly limited due to the curse of dimensionality. Exact matching is interesting to study because it is expected to generate the highest accuracy in sample matching estimation, especially in cases with a small amount of background variables. Also, when the panel contains only a small number of background variables, exact matching is viable to implement.

To be able to locate similar units in the panel, a set of auxiliary variables is required. Background variables in sample matching are used as auxiliary variables, and the values of these variables should be available in both population and panel. In exact matching, one draws a sample from the population. Each unit in the sample is matched to the unit in the panel that has the exact same value. Therefore, the size of the matched set is the same as the sample, the only difference is that values of the target variable come from the panel. In a situation where for each unit in the sample there are multiple candidates with the same value as the sample unit, a procedure is needed to select one unit from the group of similar candidates. One approach is to select a unit at random from the group.

There are two obvious limitations of exact matching, one is that the exact matching is only viable for categorical datasets. The second limitation is the curse of dimensionality: the combination of categories of background variables increase exponentially with the increase of the number of variables. The huge number of combined categories can lead to numerous blanks in candidates for the matching.

2.4.2 Proximity Matching

In most cases, exact matching is impossible to implement because of the dimensionality of the background variables. The matching need not be exact — matching is usually performed using a distance function that measures the similarity between a pair of respondents — if the pool of available respondents is sufficiently large and diverse, the matched sample is guaranteed to have approximately the same joint distribution of the matching variables as the target sample.

Measuring the similarity based on a distance function is the most commonly used method to find similar items in data mining. In missing data imputation the measurement of similarity between two objects in distance hot deck imputation is also based on a distance function. The most popular method to calculate the distance between categorical variables is to construct a vector and assign dummy variables to all categories, and the distance between two dummy variables will be 0 if the two units have the same value and 1 otherwise (Chmura, 2013). However, this method of measurement assumes an equal dissimilarity for each category, which reduces the precision of the distance function. One method to measure distances for categorical data is scaling the categories by assigning numerical values to all categories, and calculating the distance based on the numerical values. The advantage of scaling is that the distance function quantifies differences between categories, which results in more precise matching than obtained by applying dummy variables.

With conventional probability sampling, a simple random sample of size n is drawn from the population $(Y_1, Y_2, ...Y_n)$ and the population mean can be estimated using

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{13}$$

$$n^{1/2}(\hat{\theta} - \theta_0) \sim N(0, \sigma_0^2), \tag{14}$$

$$\sigma_0^2 = V(Y) = E(Y - \theta_0)^2. \tag{15}$$

(Chmura, 2013). In our case, it is easy to draw a sample from the population. Let $X_1, X_2, ...X_n$ denote the background variables of the units of such a random sample from the population P, and for each element of the target sample we find the closest matching element in the panel. If $X_i = x$, the index of the closest observation in the panel is denoted by

$$M(x) = m, \tag{16}$$

iff

$$|\tilde{X}_m - x| \leq |\tilde{X}_l - x|, l = 1, 2, ...N_{panel}, \tag{17}$$

and let $X_i^* = \tilde{X_m}$ denote the closest match to X_i in the panel, and Y_i^* is the target variable of the closest match (Chmura, 2013). When the distribution of $\tilde{X}$ is continuous, the closest match is often unique.

We define the matching estimator $\tilde{\theta}$ to be the mean of the matched sample

$$\tilde{\theta} = n^{-1}\sum_{i=1}^{n} y_i^*. \tag{18}$$

It is possible to observe how closely X_i matches X_i^*. Ideally, if the match is tight, the distribution of Y_i^* should be close to the distribution of Y_i. However, it is not expected that the distribution

Y_i and Y_i^* are highly similar since the conditional variance of Y given X may be large in the panel and the population (Chmura, et al, 2013).

In our study, as the dataset is composed of categorical data, calculating distances is not viable. Inspired by Siddique and Belin(2007) and Meulma and Van der Kooij (2013), we combined the predictive mean matching and optimal scaling regression to calculate the predictive distances between each units in the dataset. In the following section, this method will be represented by predictive mean matching via optimal scaling regression. In section 3.4, this method will be introduced in detail.

2.4.3 Propensity Score Matching

In clinical research, propensity score matching is a technique that attempts to simulate the random assignment of treatment and control group by matching treated subjects to untreated subjects that were similarly likely to be in the same group (Rosenbaum and Rubin, 1983). In the case of removing selectivity, one can assign treatment of selectivity to all units in the selective dataset, and regard the population as the untreated group. For each unit in the untreated group we calculate the propensity score of being treated, and match the unit to the unit in the treated group with the most similar value in propensity score.

The estimated propensity score $e(x_i)$ for subject i is the conditional probability of being assigned to a particular group given a vector of observed covariates X_i. In clinical research the units are assigned to a treatment and a control group. The relevant propensity score on given covariates is denoted as:

$$e(x_i) = P(Z_i = 1|x_i), \tag{19}$$

where $Z_i = 1$ or 0 is used to represent the treatment and control group (Caliendo and Kopeinig, 2008).

Since the propensity is a probability, it ranges in value from 0 to 1. In case of selectivity, we can use propensity scores to characterize the probability of being assigned to the panel group. For example, if we assume that younger aged people spend more time on the internet, we have a higher probability to collect the information of this age demography by means of a web application. If the panel is selective while the sample is representative, the selectivity of the panel can hopefully be removed by using propensity score matching.

3 Methodology

3.1 Variable Selection

D‘Orazio, et al (2006) suggested two main methods for choosing matching variables in terms of categorical variables: nonparametric measures of association and CART selection.

3.1.1 Categorical Association

A table of nonparametric measures of association is Table 1 (D’Orazio, 2006):

Table 1: Non parametric measurement of association

X measurement scale	Y measurement scale	Association measure
Nominal	Nominal	χ^2
		Φ
		Contingency coefficient
		Cramer’s V
		Uncertainty coefficient
		Concentration coefficient
		Λ
Ordinal	Ordinal	Γ
		Somer d
		Kendall τ_b
		Stuart τ_c
Ordinal	Interval	Pearson η
		Point biserial (when X is dichotomous)
Interval	Interval	Pearson‘s correlation coefficient
		Spearman rank correlation coefficient

Classical Pearson chi-squared statistic is used to measure the association of categorical variables, the Pearson chi-squared statistic is denoted as:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \tag{20}$$

where n_{ij} and m_{ij} represent the observed and expected cell frequencies of two categorical variables, $i = 1, 2, ...I$ and $j = 1, 2, ...J$, where I and J represents the number of categories of two categorical variables. However, when it comes to the different sample size, the Pearson χ^2 is not comparable. Therefore we are using another statistic Cramer’V, which is normalized by the sample size and the number of category of two variables:

$$V = \sqrt{\frac{\chi^2/n}{min(I-1, J-1)}}. \tag{21}$$

Cramer's V has a value between 0 and 1. Usually a value between 0 and 0.25 indicates a weak association between 2 categorical variables, and a value between 0.25 to 0.35 represents medium association. If the value is above 0.35 the correlation between variables is considered strong.

The idea of using categorical association is to calculate the value of Cramer's V for each background variable and the target variable, and use several background variables with high values of Cramer's V for matching. In our study, for operational purposes we will use background variables with Cramer's V value over 0.35, which are considered strong correlation between background variables and target variables.

3.1.2 CART Variable Selection

For the selection of matching variables in statistical matching, classification and regression trees (CART) are useful when a nonlinear relationship is believed to exist between a univariate Y and X. It is suggested to select variables that appear in the higher part of the tree which are believed to have greater explanatory power. CART tree is a method commonly used in data mining. It is used to assign inputs to a certain category of the target variable based on multiple covariates. CART algorithm is non-parametric and capable of classifying a categorical target variable without imposing a parametric structure assumption.

Constructing a decision tree is done top-down by choosing a variable that best splits the set of the items. In the CART algorithm, Gini impurity is used to measure the homogeneity of the target variable within the subset. Gini impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classified according to the distribution of class labels from the data set. The definition of Gini impurity is given as:

$$I_G(P) = \sum_{i=1}^{m} P_i(1 - P_i) = 1 - \sum_{i=1}^{m} P_i^2, \tag{22}$$

where P_i stands for the probability that event i happened conditional on an event G, and m stands for the total number of events with condition G. A smaller Gini impurity represents a better classification of the event.

A smaller value of Gini impurity represents a better classification ability of the background variables towards the target variable. As all variables in our experiment are categorical, following the guidance of D'Ozario, et al, we use the CART method to select the background variables for the matching. The matching variable selection by CART is implemented as follows,

•Build a classification tree using the panel with the output variable Y as the response and the others as explanatory variables, list the order of explanatory variables from the top to the bottom of the tree, and select the first 4 to 5 variables as important variables for the matching nad further steps.

3.2 Weighted Sampling

A commonly applied correction technique to remove bias is weighting adjustment, which assigns an adjustment weight to each survey respondent. Units in under-represented groups receive a higher weight and units in the groups that are over-represented receive a lower weight. In the computation of an estimate, not just the values of the variables are used, but also the survey weights. In sample matching, in order to use as much information as possible from the panel, the sample should probably resemble the panel as much as possible so that it is relatively easy to match a unit from the sample to a similar unit in the panel. Therefore, in the sample step we draw a weighted sample from the population rather than a simple random sample.

In our study, we would like to compare the matching results of a weighted sample with a simple random sample. Suppose N represents the total number of units in the population, and N_{panel} is the number of total units in the panel. Let N_c stands for the number of category c in the population and $N_{panel,c}$ is the number of category c in the panel. In simple random sampling, the probability of being sampled for each unit i from the population is $P_i = n/N$, where n stands for the sample size. However, in weighted sampling we first define the sampling weight for category c based on the panel:

$$w_c = \frac{N_c * N_{panel}}{n * N_{panel,c}}. \tag{23}$$

Therefore, for each unit in the population with category c, sampling weight w_c is assigned. The sampling probability for units with category c in the population is then

$$P_c = \frac{n * N_{panel,c}}{N_c * N_{panel}}. \tag{24}$$

We use the sampling probabilities to generate a weighted sample with total units n from the population. In practice, we draw a sample with each categories number equals to $n * N_{panel,i}/N_{panel}$. After generating the weighted sample, units in the sample are matched to the units in the panel.

Theoretically, one can use all background variables to generate a weighted sample. However, it is difficult in practice when there are too many combinations of categories in background variables. In order to use the maximum information from the panel, one could draw a weighted sample from the population by using only a few background variables. Weighting adjustment assigns an adjustment weight to each survey unit. During the computation of the statistics, the weights are used to estimate the population statistics.

3.3 Exact Matching on Categorical Variables

Suppose datasets D_1 and D_2 have the population size N_1 and N_2 respectively, and have the same background variables $X_1, X_2, ...X_K$ (where K stands for the number of background variables), the procedure of exact matching is that for each unit d_{i1} in dataset D_1, we find the matches in dataset D_2 such that each background variable has the same category or value, which is denoted by

$$M(d_{i1}) = d_{j2}, \tag{25}$$

iff

$$x_{i11} = x_{j21}, x_{i12} = x_{j22}, ...x_{i1K} = x_{j2K}, \tag{26}$$

$$j = 1, 2, ...N_2 \tag{27}$$

(Rivers, 2007). In most cases, the exact matching method can generate multiple candidates for each unit in dataset D_1, all matches for unit d_{i1} possess the exact same values or categories on all background variables. Multiple candidates for unit d_{i1} form the candidate set C_i which contains all units in dataset D_2 which have the same values on background variables as unit d_{i1}. In order to match all units from D_1 to D_2, for each candidate set C_i, randomly select one unit as the match for the unit d_{i1} in dataset D_1 to form the match set M. In our study, the online panel is considered as dataset D_1 and the sample is regarded as D_2. The match set M is generated by using exact matching and other matching methods introduced as following. The estimation of matching is introduced in section 3.7.1, and the results of all matching methods is compared.

3.4 Predictive Mean Matching with Optimal Scaling Regression

Predictive mean matching (PMM) is an attractive way for multiple imputation in a dataset with missing values, especially for quantitative variables that are not normally distributed (Allison, 2015). In the sample matching case, we combine the sample and the panel to generate a new dataset, with missing values on output variable Y. Due to the situation that the output variable Y as unknown in the sample but exist in the panel, the combined dataset contain missing values on output variable Y. Therefore, we can implement PMM to match the sample and the panel.

The procedure of PMM which used in our study is inspired by Allison (2015), with the following procedure:

•For cases with no missing values, estimate a linear regression model of Y on X, producing a set of coefficients β

• Predict the outcome variable Y with coefficients β for all units in the dataset, including units with missing values and without missing values as $\hat{Y}$.

• For each unit with a missing value on outcome variable Y (sample), identify a group of cases with observed Y whose predictive values $\hat{Y}$ are closest to the predicted value of the missing unit. Among those closest cases, randomly select one unit as the matching and form the matched set M.

PMM uses linear regression to construct a metric for matching cases with missing data to similar cases with observed data. However, for a categorical dataset, it is impossible to build a linear relationship between background variables X and target variable Y. In this circumstance, optimal scaling regression can be considered to build the regression model for predicting the measure metrics. In the case of sample matching, Y denotes the target variable and X denotes the vector of background variables. Because all variables are categorical, it is hard to assume the distribution and construct linear models. In optimal scaling regression, categorical variables are transformed into numerical values by creating indicator functions (Meulman and Van der Kooij, 2016). Outcome variable Y is transformed into q_y and each background variable x_j is transformed into q_j, the optimal scaling regression model is:

$$\vartheta(y) = \sum_{j=1}^{K} \beta_j \varphi_j(x_j) + e. \tag{28}$$

We use q_j to denote the numerical value of transformation function $\varphi(x_j)$ of background variable X_j, q_j is obtained by the multiplication of an indicator matrix G_j which indicates the categories and a quantification vector v_j which contains the numerical values of all categories for variable X_j (Meulman and Van der Kooij, 2016). In order to avoid confusion, in this part X_j stands for the j column of background variable X, x_j stands for a vector of X_j, and x_{ij} is the ith value in vector x_j. Elements of indicator G_j for variable X_j are defined by the following rules (Meulman and Van der Kooij, 2016): $g^j_{ik} = 1$ if $x_{ij} = k$, and $g^j_{ik} = 0$ otherwise $(k = 1, ..., C_j, i = 1, ...N)$. The indicator matrix uses a dummy variable indicating category. Eg, suppose the variable X_j has 4 categories $1, 2, 3, 4$. Given one small part of the dataset with $x_j = [1, 4, 3, 4, 2]$, we use an indicator matrix

G_j to represent x_j (Meulman and Van der Kooij, 2016):

$$x_j = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 4 \\ 2 \end{bmatrix} \Rightarrow G_j = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{29}$$

The quantification vector v_j is denoted as $v_j = [v_{j1}, v_{j2}, v_{j3}, v_{j4}]$, which assigns categorical values in x_j a numerical value v_j. Therefore, the result of the transformed variable is denoted as $q_j = G_j v_j$ for background variables and $\vartheta(y) = G_y v_y$ for the target variable. Function (30) gives an example of the quantification transformation of a vector x_j for variable X_j.

$$x_j = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 4 \\ 2 \end{bmatrix} \Rightarrow G_j v_j = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} v_{j1} \\ v_{j2} \\ v_{j3} \\ v_{j4} \end{bmatrix} = q_j = \begin{bmatrix} v_{j1} \\ v_{j4} \\ v_{j3} \\ v_{j4} \\ v_{j2} \end{bmatrix}. \tag{30}$$

In optimal scaling regression, all values of the categorical variables Y and X are all represented by numerical transformation vectors v and indicator matrix G. The regression equation is $\vartheta(y) = \sum_{j=1}^{N} \beta_j G_j v_j$. In order to quantify the goodness of transformation and regression, we define the residual sum of squares $\sum_{i=1}^{n} (\vartheta(y_i) - \sum \beta_j G_{ij} v_{ij})^2$ which is the sum over all observations as the loss function of the optimal scaling regression, which is written as

$$L(\beta, \varphi, \vartheta) = ||\vartheta(y) - \sum_{j=1}^{K} \beta_k G_j v_j||^2. \tag{31}$$

where $\vartheta(y)$ represents the transformation of target variables Y, and $G_j v_j$ represents the numerical transformation of the categorical variable X_j, K represents the dimension of the background variables. Furthermore, the transformation ignores the ordinal information of categories, we call this transformation a non-monotonic function. In summary, the transformation function $\varphi(x_j) = G_j v_j$ is a non-monotonic step function. The loss function can also be written as

$$L(\beta, \varphi, \vartheta) = ||\vartheta(y) - \sum_{j=1}^{K} \beta_j \varphi(x_j)||^2. \tag{32}$$

The loss function has to be minimized by finding optimal parameters $\beta = \beta_j$ for $j = 1, 2...K$ and the optimal numerical quantification v_{ij} for $j = 1, 2...K$. After finding the values of v_{ij}, each category of the variables will be assigned a numerical value. The proximity between units can be calculated through the numerical transformation of categorical variables.

3.4.1 Ordinal and Nominal Transformation

By using the transformation $\varphi(x_j) = G_j v_j$ we assume that the categorical variables are all nominal variables, where we merely maintain the classification information in the quantified variable $G_j v_j$.

If two units a and b belong to the same category of variable j, the transformed information is:

$$x_{a,j} = x_{b,j} \Rightarrow \varphi(x_{a,j}) = \varphi(x_{b,j}). \tag{33}$$

If a categorical background variable is ordered, the variable is a ordinal variable. The order information should be preserved during the numerical transformation:

$$x_{a,j} < x_{b,j} \Rightarrow \varphi(x_{a,j}) < \varphi(x_{b,j}). \tag{34}$$

The order information is preserved by restricting the ordinal quantifications in v_j so that

$$x_{a,j} < x_{b,j} \Rightarrow v_{a,j} < v_{b,j}. \tag{35}$$

We define a ordinal transformation function $\varphi_{ord}(x_j)$ to transform categorical variable x_j to numerical values, this function maintaines the ordinal information between categorical variables. In the following algorithms we introduce one method to maintain the ordinal information during the transformation. During the implementation of optimal scaling regression, we use the optimal scaling regression function in SPSS. SPSS contains several ordinal transformation functions $\varphi_{ord}(x_j)$.

3.4.2 Regression Weights and Transformation Parameters

The regression parameters $\beta_j (j = 1, 2, ...K)$ and quantification vectors $v_j (j = 1, 2, ...K)$ are found by minimizing the loss function $L(\beta, \varphi, \vartheta) = ||\vartheta(y) - \sum_{j=1}^{K} \beta_j G_j v_j||^2$. In order to find the value for each $\varphi(x_j) = G_j v_j$, we separate a variable and its parameter β from the linear combination of predictors, isolating the current target part $\beta_j \varphi_j$ from the remainder, denoted as $\sum_{l \neq j} \beta_l \varphi_l(x_l)$. This method is called the blocking method. The loss function can be rewritten as

$$L(\beta, \varphi, \vartheta) = ||\vartheta(y) - \sum_{l \neq j} \beta_l \varphi_l(x_l) - \beta_j \varphi_j(x_j)||^2. \tag{36}$$

We then define an auxiliary variable u_j:

$$u_j = \vartheta(y) - \sum_{l \neq j} \beta_l \varphi_l(x_l), \tag{37}$$

thus u_k is the partial residual. Next, the loss function is denoted as

$$L(\beta_j, \varphi_j, \vartheta) = ||u_j - \beta_j \varphi_j(x_j)||^2, \tag{38}$$

which is a function of β_j and φ_j only. By minimizing the function (36) the optimum value of quantification and regression parameters can be found. The standardization of the transformed variable $\varphi_j(x_j)$ allows us to compute the regression parameters β_j separately from the transformation. The current value of the regression parameter β_j is

$$\hat{\beta}_j = u_j^T \varphi_j(x_j). \tag{39}$$

Next, the loss function (36) is minimized over all $\varphi_j \in C_j(x_j)$ where $C_j(x_j)$ stands for the set of all categorical variables. Updating all βs is one iteration cycle, the loss function will be optimized by multiple iterations.

The estimation algorithm of nominal transformation is implemented as follows:

Algorithm 1 Nominal Transformation

```
1. Initialize $\beta$ and $v_j$ by randomly assigning numbers between 0 to 1, compute $G_j$ based on predictors $x_j$ for each predictor variable. Define the maximum number of iterations (maxiter) and minimum decrease in loss per iteration (crititer).
2. Determine current loss ($=SS_{res} = ||\vartheta(y) - \sum_{j=1}^{K} \beta_j G_j v_j||^2$). Set iteration $i = 1$, set $j = 1$.
3. For variable $x_j$, define auxiliary variable:
       $u_j = y - \sum_{l \neq j} \beta_l G_l v_l.$        (40)
4. Minimize
       $||u_j - \beta_j G_j v_j||^2,$        (41)
   over quantification $v_j$, giving
       $\tilde{v_j} = (\beta_j) D_j^{-1} G_j^T u_j,$        (42)
   where $D_j = G_j^T G_j$.
5. Standardize
       $v_j = N^{1/2} \tilde{v_j} (\tilde{v_j^T} D_j \tilde{v_j})^{1/2},$        (43)
   so that $v_j^T D_j v_j = N$, where $N$ is the number of total units.
6. Fix quantification $v_j$ and update regression coefficient
       $\tilde{\beta_j} = N^{-1} (G_j v_j)^T u_j.$        (44)
Repeat step 3 to 6 for all $K$ predictor variables, with $j = 2 ... M$. After one cycle of updating $\beta_j$ and $v_j$, calculate the loss $SS_{res} = ||\vartheta(y) - \sum_{j=1}^{P} \beta_j G_j v_j||^2$
7. Check whether $SS_{res_{before}} - SS_{res_{after}} \leq critier$, and $i \leq maxiter$:
if true, set $i = i + 1$, set $j = 1$ and repeat steps 3 to 7;
if false, stop the algorithm.
```

For an ordinal transformation, different from the nominal transformation, the ordinal information of all categorical variables are maintained during the transformation. The algorithm of ordinal step quantification is illustrated as below:

Algorithm 2 Nominal Transformation

1.Initialize β and v_j by randomly assigning numbers between 0 to 1, compute G_j based on predictors x_j for each predictor variable. Define the maximum number of iterations (maxiter) and minimum decrease in loss per iteration (crititer).

2. Determine current loss ($SS_{res} = ||\vartheta(y) - \sum_{j=1}^{K} \beta_j G_j v_j||^2$). Set iteration $i = 1$, set $j = 1$.

3. For variable x_j, define auxiliary variable:

$$u_j = y - \sum_{l \neq j} \beta_l G_l v_l. \tag{45}$$

4. Minimize

$$||u_j - \beta_j G_j v_j||^2, \tag{46}$$

over quantification v_j, giving

$$\tilde{v_j} = (\beta_j) D_j^{-1} G_j^T u_j, \tag{47}$$

where $D_j = G_j^T G_j$.

5. Standardize

$$v_j = N^{1/2} \tilde{v_j} (\tilde{v_j^T} D_j \tilde{v_j})^{1/2}, \tag{48}$$

so that $v_j^T D_j v_j = N$, where N is the number of total units.

6. Check the ordinal information: compute the weighted average of quantification that are in the wrong order, assign the average quantification values to the incorrectly ordered categories. E.g. if $x_{aj} > x_{bj}$ but the transformed numerical value $v_{aj} < v_{bj}$, assign the average value $\varphi(x_{aj}) = \varphi(x_{bj}) = \frac{v_{aj}+v_{bj}}{2}$.

7. Fix quantification v_j and update regression coefficient

$$\tilde{\beta_j} = N^{-1} (G_j v_j)^T u_j. \tag{49}$$

Repeat step 3 to 7 for all K predictor variables, with $j = 1...K$. After one cycle of updating β_j and v_j, calculate the loss $SS_{res} = ||\vartheta(y) - \sum_{j=1}^{P} \beta_j G_j v_j||^2$

8. Check whether $SS_{res_{before}} - SS_{res_{after}} \leq critier$, and $i \leq maxiter$:

if true, set $i = i + 1$, set $j = 1$ and repeat steps 3 to 8;

if false, stop the algorithm.

The optimal scaling regression is used on the panel to construct numerical transformation for all categorical variables, and obtain the prediction function $\vartheta(y) = \sum_{j=1}^{K} \beta_j \varphi_j(x_j) + e$. After the numerical transformation, all categorical variables in both sample and panel are assigned numerical values based on the transformation, and predictive outcome variable $\hat{Y}$ are calculated based on the prediction function. Finally, we are able to use PMM method to match the panel to the sample, with whose steps are introduced above.

3.5 Propensity Score Matching

Propensity scores matching (PSM) is an alternative method to estimate the effect of receiving a treatment when the treatment to a subject cannot be applied, by pairing the treated and untreated units with similar values on the propensity score. The matched untreated units can be used to estimate the effectiveness of the treatment. In sample matching, the treatment and control can be regarded as the units in the sample, respectively the panel.

3.5.1 Calculating Propensity Score

There are two general methods to calculate the propensity score: logistic regression and CART (D'Agostino, 1998). Logistic regression is most widely used to estimate propensity scores. Several adjusted methods to estimate propensity scores such as bagged CART, boosted CART and random forest are introduced to improve the propensity score matching.

Logistic regression is the most commonly used method for estimating propensity scores. It is a model to predict the probability that an event occurs.

$$log\frac{e^{x_i}}{1-e^{x_i}} = log\frac{P(z_i=1|x_i)}{1-P(z_i=1|x_i)} = \alpha + \beta^T x_i, \tag{50}$$

Where x_i denotes the value of background variable X. In logistic regression, the dependent variable is binary, $z_i = 1$ represents the treatment and $z_i = 0$ stands for the control. In sample matching, we define $z_i = 1$ for a unit from the panel and $z_i = 0$ for a unit from the sample.

CART represents a promising alternative to conventional logistic regression for propensity score estimation. CART does not make any assumptions towards the distribution of the explanatory variables, nor does it assume a linear relationship between the treatment and covariates. There are several approaches to improve the estimation of propensity scores based on CART including boosted CART, random forest and bagged CART.

In this study, the following methods are used to estimate the propensity scores:

- Logistic regression: standard logistic regression with a main effect for each covariate.
- CART: recursive partitioning using the *rpart* package with default parameters.
- Bagged CART: bootstrap aggregated CART is implemented using the *ipred* package. We used multiple bootstrap replicates based on empirical evidence suggesting that with more replicates, misclassification rates improve and test errors are more stable.
- Random forests: random forests are implemented using the *randomForest* package with the

default parameters.

• Boosted CART: boosted regression trees are implemented using the *twang* package. We used the parameters recommended by McCaffrey et al., with $20,000$ iterations and a shrinkage parameter of 0.0005, with an iteration stopping point that minimizes the mean of the Kolmogorov-Smirnov test statistic.

3.5.2 Adjustment for Propensity Score Matching

Propensity score matching tries to find 1 (or more) individual(s) with a similar propensity score in the treatment and control groups. There are various methods to match individuals. Once researchers obtain estimated propensity scores, proper matching techniques can be applied. The basic method is a 1:1 nearest neighborhood matching, while many of the matching methods incorporate the caliper method to improve the quality of matching (Caliendo and Kopeinig, 2008).

• Nearest Neighborhood Matching

In nearest neighborhood matching, the units from the control group are matched with T units from the treated group with the minimum difference on propensity score. In this method, the absolute difference between the estimated propensity scores for the control and treatment groups is minimized.

$$C(P_i) = min_j|P_i - P_j|, \tag{51}$$

where $C(P_i)$ represents control subject j matched to treated subject i (on the estimated propensity score), P_i is the estimated propensity score for the treated subject i and P_j is the estimated propensity score for the control subject j.

• Caliper Matching

A pre-determined range of values e is defined usually within one-quarter of the standard error (0.25s) of the estimated propensity scores, matched units that fall outside of that range are removed:

$$|P_i - P_j| < e, \tag{52}$$

where P_i and P_j represent the estimated propensity score of the control and treated subject i and j, e is the pre-determined value. The caliper matching can also be combined with nearest neighborhood matching, by selecting nearest neighborhoods with a limited difference in estimated propensity score.

After calculating the propensity scores of all units, we can match all units in the sample to the units in the panel to form the matching set M, and estimate the matching results based on the methods introduced in section 3.7.

As there are no previous studies on PSM in sample matching, this method is not expected to be more precise than the other two methods. We regarding the PSM as an explorative matching method. If the performance of PSM is not substantially performs better than other methods, we wouldn't recommend PSM in sample matching.

3.6 Reducing Bias for Random Forest

Random forest is an ensemble method for classification that constructs multiple decision trees at training time and outputs the class that is the mode of the classification of the individual trees. Random forest is a modified bagging method which selects a random subset of features as the candidates split in the learning process. Due to the random selection of both the features and training set, the algorithm itself is able to reduce overfitting and have high prediction precision. In the cases that a selective panel is available, the training model would be easily overfitted due to the selectivity of the training set, which could cause a large prediction bias on the prediction dataset. In this situation, we will use the sample matching method to remove the selectivity of the available training panel, and use the matched dataset to train the model to predict the representative population dataset. (Hastie, Tibshirani and Friedman, 2009)

Three methods of sample matching will be implemented to generate matched datasets, here represented by Ms along with and without weighted sampling, therefore we obtain 6 Ms in total. We assume that the selectivity in the matched set has been removed by sample matching, and use the newly generated matched sets as training sets to train the prediction models. Random forest is used to train the model, and we use all models to predict the distribution on the representative population dataset. The results of prediction with and without sample matching will be compared. It is expected that prediction models with sample matching will substantially increase the accuracy and reduce the bias of the estimate.

3.7 Experimental Design

3.7.1 Population Estimate

The aim of this study is to find out the best sample matching method which can remove the selectivity in the panel and reduce bias of population estimates. In this study, we will estimate the probability distribution of outcome variable "Economical Status". For each category c in the target variable Y, the estimate is defined as

$$\hat{p_c} = \frac{n_c}{n}, \tag{53}$$

where n_c standards for the number of category c in the match and n is the total number of units of the matched dataset. The estimate by sample matching is compared with the true distribution of the population, which represents by the value of the proportion of each category of target variable p_c:

$$p_c = \frac{N_c}{N}, \tag{54}$$

N_c stands for the number of units for category c in the population, and N is the population size. To measure the accuracy of each match, we use a sum of square difference (SSD) to measure the difference between the population and matches. The SSD of the population estimate is defined as

$$SSD = \frac{1}{n}\sum_{c=1}^{C}((p_c - \hat{p_c})^2), \tag{55}$$

where C is the number of categories of the target variable, p_c represents the proportion of category c in the population, and $\hat{p_c}$ is the proportion of category c in the match.

In weighted sampling, the calculation of the population estimate is different from the estimate with random sampling. As we assign a sampling weight to each unit in the population (calculation of sampling weights and sampling probability as given in function (23) and (24)), the population estimate of weighted sampling is

$$\hat{p_c} = \frac{\sum_{k\in s_c} w_k}{\sum_{k=1}^{N} w_k}, \tag{56}$$

where w_k is the sampling weight of category k, and s_i is the set of units in the matched dataset with category i and q represents all background variables category. The SSD of weighted sampling is the same as random sampling in function (55). In the following experiments, each match of the methods will estimate the frequency distribution of the population, and SSD will be used to assess the matching quality.

3.7.2 Design of the Experiment

The performance of the sample matching methods are measured under a factorial experiment design of several factors: the size of the sample and panel, the different methods of matching and sampling.

Size of the sample and panel: In order to investigate the effect of the size of sample and the panel, different size of the samples and panels are generated from the population. For the sample, 7 different sample sizes are generated from the population ($0.5\%, 1\%, 2\%, 3\%, 5\%, 10\%$). Selective panels of three different sizes (5%, 10%, 20%) are generated.

Matching methods: All three matching methods will be implemented based on the different sizes of samples and panels. In PMM, ordinal transformation and nominal transformation have been implemented and compared. For the propensity score matching, different methods of calculating and adjusting the propensity scores have been implemented, and the results of these methods are compared. The results of three matching methods are compared and the best matching method based on the results is selected.

Method of sampling: There are two ways of sampling in the experiment: simple random sampling from the population and weighted sampling based on two background variables. The methods of selecting the background variables include CART and categorical "correlation". It is expected that the result of weighted sampling will perform better than simple random sampling.

SSD assessment: For each method in section 3, the results of the matching will be assessed by means of the sum of squared differences, see (55).

The experimental procedure is implemented as follows:

- For each size of the panel (5%, 10% and 20% of the population size), calculate the Cramer's V and Gini impurity of each background variable towards the target variable, choose the two background variables with the lowest value of Gini impurity and highest value of Cramer's V as the background variables for weighted sampling. If variables are selected by Cramer's V and Gini impurity are different, the experiment is implemented on both of the two groups. Draw a series of random samples and weighted samples with size from 0.5% to 10% from the population.
- For each group of the panel and its samples, 3 methods of matching are implemented to estimate the distribution of the target variable.
- Estimate the SSD of matching, the values will be compared for different methods of matching, different methods of sampling and different sizes of samples and panels.

3.7.3 Implementation of Exact matching

The implementation of exact matching is relatively simple: For each panel, draw a series of differently sized samples and match each unit in each sample to the units from the panel.

- For each panel, draw a series of samples with simple random sampling and weighted sampling from the population.
- Start the matching with one background variable with the highest value of Carmer‘s V and lowest value of Gini impurity. For units in each sample, match each unit to the units in the panel with exactly the same values of background variables. If there are multiple candidates, randomly select

one unit from the multiple candidates as the match.

- Add one background variable with the second highest value of Cramer's V and lowest value of Gini impurity. Conduct multiple exact matches by increasing the number of background variables. The matches in the panel are used to estimate the population distribution, SSD is calculated to measure the matching.

3.7.4 Implementation of Predictive Mean Matching

The procedure of implementing the scaling PMM as follows:

- For each panel, conduct the optimal scaling regression of the outcome variable towards background variables; both for ordinal and nominal levels of scaling using the panel. Each category of the background variables will be transformed into a numerical value which can be used to calculate the distance between units.
- Assign the transformed numerical values to both the background variables in the sample and the panel.
- Predict the numerical values of the outcome variable with the optimal scaling regression model and the relevant numerical values of each unit in both the sample and the panel.
- The predictive mean matching: match each unit in the sample to one unit in the panel, based on the predicted outcome $\hat{Y}$.
- The matches in the panel are used to estimate the population distribution of the target variable. SSD is calculated for the assessment.

3.7.5 Implementing Propensity Score Matching

The procedure of the propensity score matching is implemented as follows:

- For each panel, generate a series of samples of different sizes $(0.5\%, 1\%, 2\%, 3\%, 5\%, 10\%)$, merge the panel and the sample in one dataset.
- Calculate the propensity score of each unit in the new dataset with different methods: logistic regression, CART and adjusted CART (boosted CART, random forest).
- Once the estimated propensity scores are calculated, the units in the panel are matched to the units that have the same or similar propensity scores, the matching follows a 1-to-1 match. The unmatched subjects are discarded from the analysis.
- Use the matches in the panel to estimate the population distribution, SSD is used as the assessment

3.7.6 Comparison of Sampling

Apart for the size of samples, the size of panels and matching methods, the method of sampling is another factor which could affect the result of removing selectivity. The weighted sampling is a well-defined sampling mechanism by first determining the sampling probabilities of population units, and drawing a sample from the population based on their sampling probabilities. After matching with the panel, the population estimate is weighted by using the sampling weights (i.e. the inverse of the sampling probabilities) obtained from the sampling mechanism.

The procedure of weighted sampling can be implemented as follow:

- Determine the sampling probabilities based on the selected background variables.
- Conduct probability sampling using the assigned sampling probabilities to generate a series of samples of different sizes.
- Use weighted samples for the matching, test all matching methods with different weighted samples to generate the matched datasets.
- Estimate the population distribution with the matches and the sampling weights, the estimates are weighted by the sampling weights.

3.7.7 Increase Prediction Accuracy with Sample Matching

In order to validate our assumption that when a model is developed by using a representative matched training set which is similar to the prediction dataset, this could substantially increase the prediction accuracy, we develop 3 sets of random forest models: models trained by using the original selective dataset, by using matched datasets obtained by random sampling and obtained by weighted sampling. Three sets of models are tested on the same representative prediction set and the SSD of predictions is calculated. The general procedure is implemented as follows:

- Generate differently sized selective panels from the whole dataset (5%, 10%, 20%), build random forest models based on all these selective datasets and set the group of models as RF_1.
- For each panel, generate a series of random samples of different sizes $(0.5\%, 1\%, 2\%, 3\%, 5\%, 10\%)$. For each sample, use 3 different matching methods to match the units in the panel and generate the matched datasets. Random forest is used to build prediction models based on these matched datasets, set the group of models as RF_2
- For each panel, generate a series of weighted samples of different sizes $(0.5\%, 1\%, 2\%, 3\%, 5\%, 10\%)$ by using the best variable combination (lowest value on SSD in 3.6.5). For each sample, use 3

matching methods to match the units in the panel and generate the matched datasets. Use random forest to build prediction models based on the matched datasts and set the group of models as RF_3.

• Use the three groups of models RF_1, RF_2 and RF_3 to predict the target variable for the test set, calculate the SSD of all the prediction estimates and compare the results of the SSD.

4 Results

The results of the sample matching methods introduced so far are shown in this chapter. Methods of sample matching will include exact matching, scaled predictive mean matching and propensity score matching. All matching methods will be tested combined with simple random sampling and weighted sampling. The results of random forest prediction based on a selective dataset and a matched dataset are also interpreted in this chapter.

The results of the experiments will be presented in the following order: the first part will introduce the dataset, give a description of selective panels and the scenario. In the second part, the results of variable selection will be shown. Here we use the Cramer's V value to choose the background variables for weighted sampling, and use the CART method to choose the matching variables. The third part will show the results of matching methods together with different methods of sampling. The SSD will be used to measure the results and the results will be shown by line graphs. In order to observe the variation of the methods, each experiment will be replicated 100 times and the variance of SSD will be calculated. Due to the different procedures of matching methods, the replications of the experiments are also distinguished: In exact matching and propensity score matching, the experiments are duplicated by generating different samples for the matching and calculating propensity scores. For scaled predictive mean matching, the duplication will be implemented by generating different samples and multiple coefficients β based on the variance of β.

Lastly, we will compare the random forest prediction with a selective dataset and a matched dataset. The matched dataset is generated with the best sample matching methods which produce the minimum value of SSD. Random forest models will be developed on the selective panel and matches, and will be used to predict the outcome variable. The results are also measured in SSD. The details are explained in the following sections.

4.1 Dataset, panel and variable selection

The dataset used in this study is the Dutch Population Census of 2001 provided by Statistics Netherlands. The dataset contains data on 1% of the Dutch population. The data is freely available for researchers. The data contains 190,000 individuals of the Dutch population, with 13 categorical variables. Apart for the last variable “weight”, all other variables are used in the study. The variables contain information on gender, age, position in the household, size of the household, residential area, nationality, country of birth, educational level, economical status, occupation, working field and marital status. The chosen target variable is Economical status. The details of the dataset can be found in Appendix I.

4.1.1 Population and Panel

The first step of the experiment is to generate the population dataset and panel dataset. The population dataset is generated by sampling 60,000 units from the 190,000 units from the original dataset, with the population distribution of target variable "Economical status" given by:

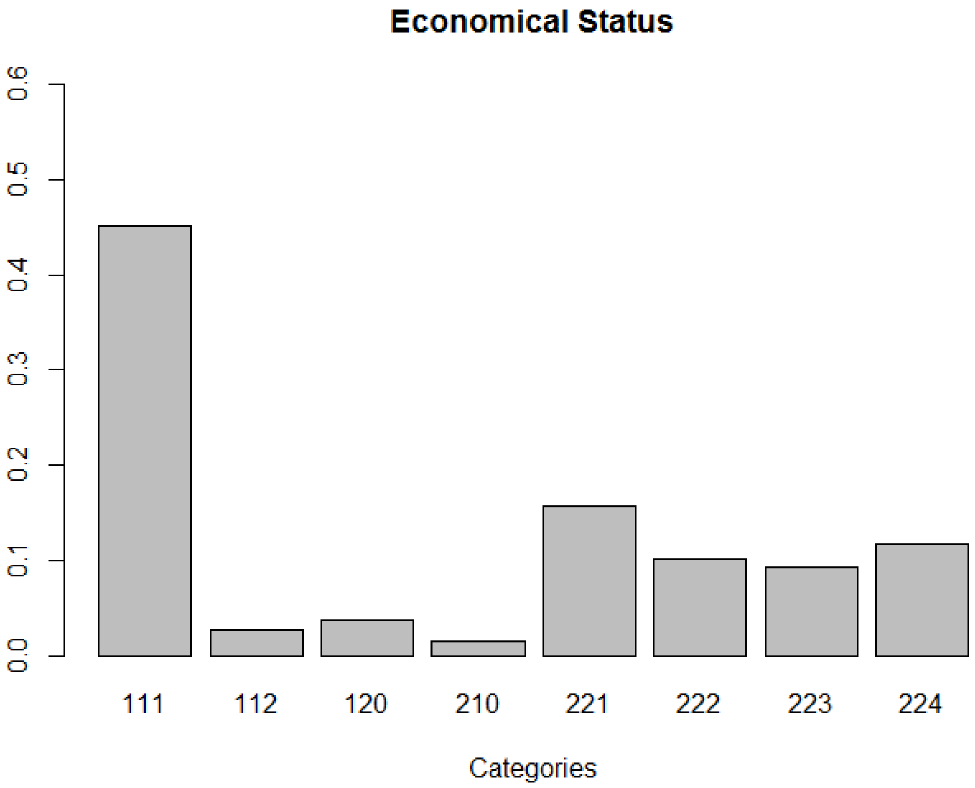


Figure 1: Population Distribution of Economic Status

Table 2: Population distribution of Economical Status

Categories	Employee	Student	Independent	Unemployed	Edu-related	Retired	Housewife	Others
Index	111	112	120	210	221	222	223	224
Frequencies	0.452	0.027	0.036	0.014	0.157	0.102	0.093	0.117

The majority of units has the category "Employee" which contains approximately 45%, the unemployed demography contains approximately 14.25%. In the panel dataset, we generate a selective panel with lower percentage of employed units and higher percentage of unemployed units.

In order to observe the panel size effect towards the matching, 3 selective panels are generated by stratified sampling from the rest of the orginal dataset with 20,000, 40,000 and 60,000 units. In the selective panel the percentage of employment declines to 36% and unemployment increases to 20%. All fractions of other categories increase or decrease differently, the details of compositions of the panels are shown in Table 3 (From Panel 1 to Panel 3).

Table 3: Panel distribution of Economival status

Employee	Student	Independent	Unemployed	Edu-related	Retired	Housewife	Others
0.369	0.0223	0.030	0.019	0.129	0.141	0.126	0.162
0.369	0.023	0.029	0.019	0.128	0.142	0.129	0.162
0.368	0.023	0.029	0.019	0.129	0.141	0.128	0.162

4.1.2 Variable Selection

The variable selection in this study consists of two parts: choosing background variables for weighted sampling and choosing variables for matching. In the simulation, the value of target variable "Economical Status" is only available in the panels but unknown for the rest of the population. The Cramer's V values and Gini values are calculated from the panel and the results are shown in Appendix II. Based on the Cramer's V value, background variables which are highly correlated are Age, Educational level, Occupation, Working field and Matrital status. However, for the CART results, the important variables are different with the change of panel size. Based on the result, we found variable "Age" and "Occupation" always appear in a high position of the tree and have a high Cramer's V value simultaneously, therefore we use these two variables to conduct weighted sampling. As for the matching variables, due to the unstable results of the CART, we use the backgroung variables which selected by Cramer‘s V value, therefore variable age, occupation, educational level and working field.

4.2 SSD and matches

The SSD of the estimates of all methods are compared under each experimental condition. In exact matching, at most 3 background variables are used for matching due to the curse of dimensionality. In PMM, SSD under ordinal and nominal scaling levels are compared. For propensity score matching, the ratio of the sample size and panel size are compared because the different sizes of panels and samples could affect the propensity scores. All methods are compared under different sizes of panels and samples.

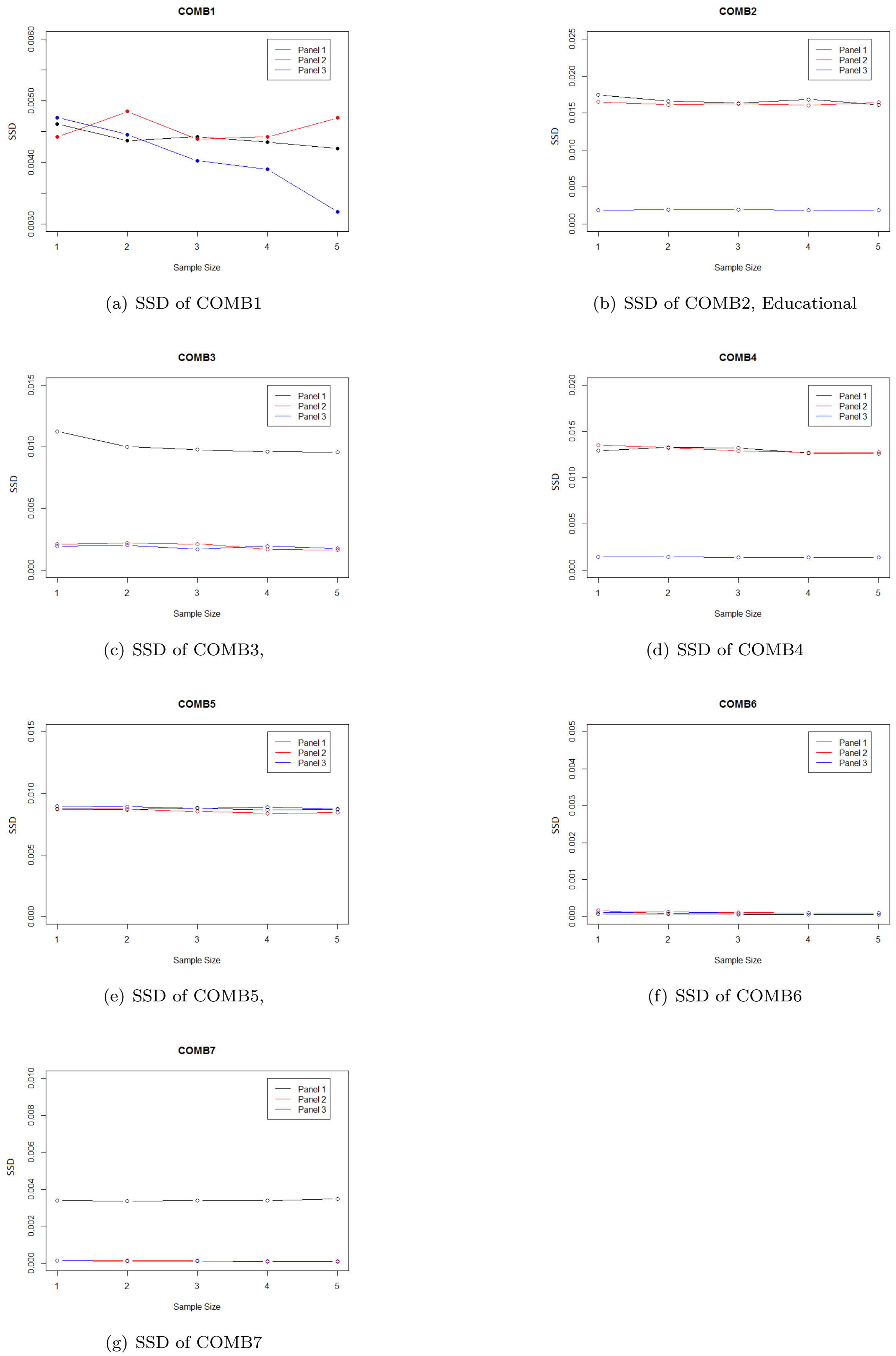
4.2.1 Results of Exact Matching

Figure 2 gives the comparison of the panel size and sample size effect for groups of background variables, Table 4 gives the index of background variables combinations. Results of exact matching are shown for different groups of background variables. The used background variables in exact matching include age, occupation, working field and education level, with 7 groups of combinations.

In general, factors that are most obvious for exact matching are background variables, panel size and methods of sampling. SSD of all matches have not shown substantial differences with the increase of sample size, but has substantial variation with the size of the panel. In all groups, the sample size shows an unstable and chaotic effect towards the matching, while the effect of panel size seems to be more obvious. In most groups the SSD of exact matching tends to be lower with a larger panel size, the results are obvious in group 2, 3, 4, and 7. In COMB2, the SSD of panel1 and panel2 are around 0.017 while it dropped to 0.001 for panel3. In group 3, 4, and 7, SSD showed a substantial decrease when the panel size goes to 40,000 (panel2), as shown in the graphs. In other groups, although the SSD does not show a substantial decrease with the increase of the panel size, one can observe that in most cases, SSD slightly decreases as we enlarge the panel size. This phenomenon shows that in order to obtain a better result for exact matching, the size of the panel should approximately be equal to the population size. While a larger panel could be helpful for a better matching, the effect might not as substantial as the factor of combination of background variables and sampling method. Furthermore, influencing the panel size is difficult or even impossible in most cases, it is suggested that focusing on the influence of background variables and sampling methods could increase the matching accuracy in exact matching.

Table 4: Background variables combinations

COMB1	Educational level, Working field
COMB2	Age, Educational level
COMB3	Age, Working field
COMB4	Age, Working field, Educational level
COMB5	Age, Occupation, Educational level
COMB6	Age, Occupation
COMB7	Age, Working field, Occupation



(a) SSD of COMB1

(b) SSD of COMB2, Educational

(c) SSD of COMB3,

(d) SSD of COMB4

(e) SSD of COMB5,

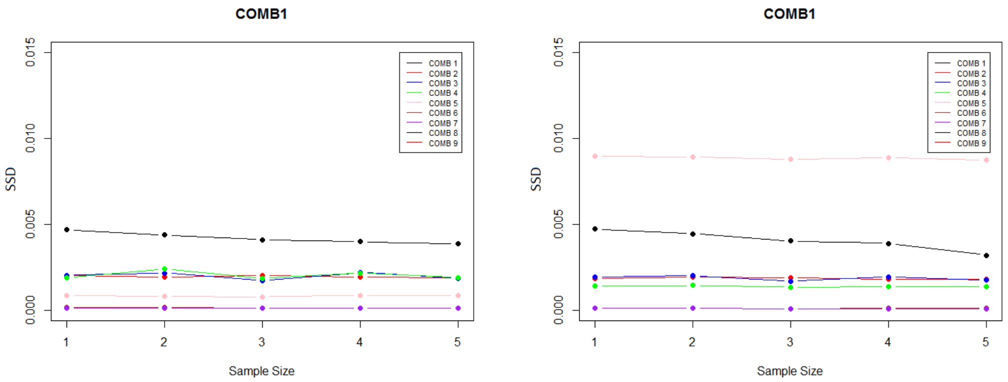
(f) SSD of COMB6

(g) SSD of COMB7

Figure 2: SSD of exact matching with random sampling

The most obvious factor that affects the matching in this experiment is the background variables.

As the results show, an enormous difference can be observed between different combinations of background variables. A large panel size will lead to a better matching result with lower SSD but a suitable panel size is different when using different background variables. As Figure 3 shows, in group 2, 3, 4, and 7, the effect of the panel size turns out to be more substantial than for other combinations. Therefore finding the best combination of background variables is essential for reducing SSD in exact matching. As a larger panel is helpful in reducing the SSD, in the graph the SSD for all combinations of background variables with the largest panel is shown. Results of the graph show that the best combinations of background variables are group 6 (age and working field) and group 7 (age with occupation). The SSD of these groups declined to around $8*10^{-5}$. The SSD of combinations 2, 3 and 4 are between 0.001 to 0.002, while combination 1 has a value around 0.005. The worst case among all groups is combination 5 whose SSD values are around 0.009, which is much higher than for other groups. When variable age is included as matching variable, the matching result outperforms all other combinations without age. The best combination of background variables in exact matching is age and working field or age and occupation. However, if all three background variables are used (age, occupation and working field), the SSD of increases, which indicates a worse combination of background variables. This could be caused by the curse of dimensionality. As variable age has 17 categories, while occupation and working field have 9 and 13 categories respectively, there are total 64 combined categories available in the dataset (Due to the curse of dimensionality, several combinations of these categories do not exist. The total number of combinations is 64 instead of $17 \times 9 \times 13$). Therefore the curse of dimensionality decreased the accuracy of exact matching since there is too much background information. As observed in the line graphs, if variable age was excluded from the matching variables, all combinations of other background variables perform worse than matches with variable age. There are not too many differences between the results with 2 combinations (educational level and occupation, educational level and working field, working field and occupation) while with all three variables included, the SSD has a substantial decrease compared to a combination with two of these variables. The reason for that could be that the total combination of categories is smaller compared with variable age, therefore the effect of curse of dimensionality does not occur.

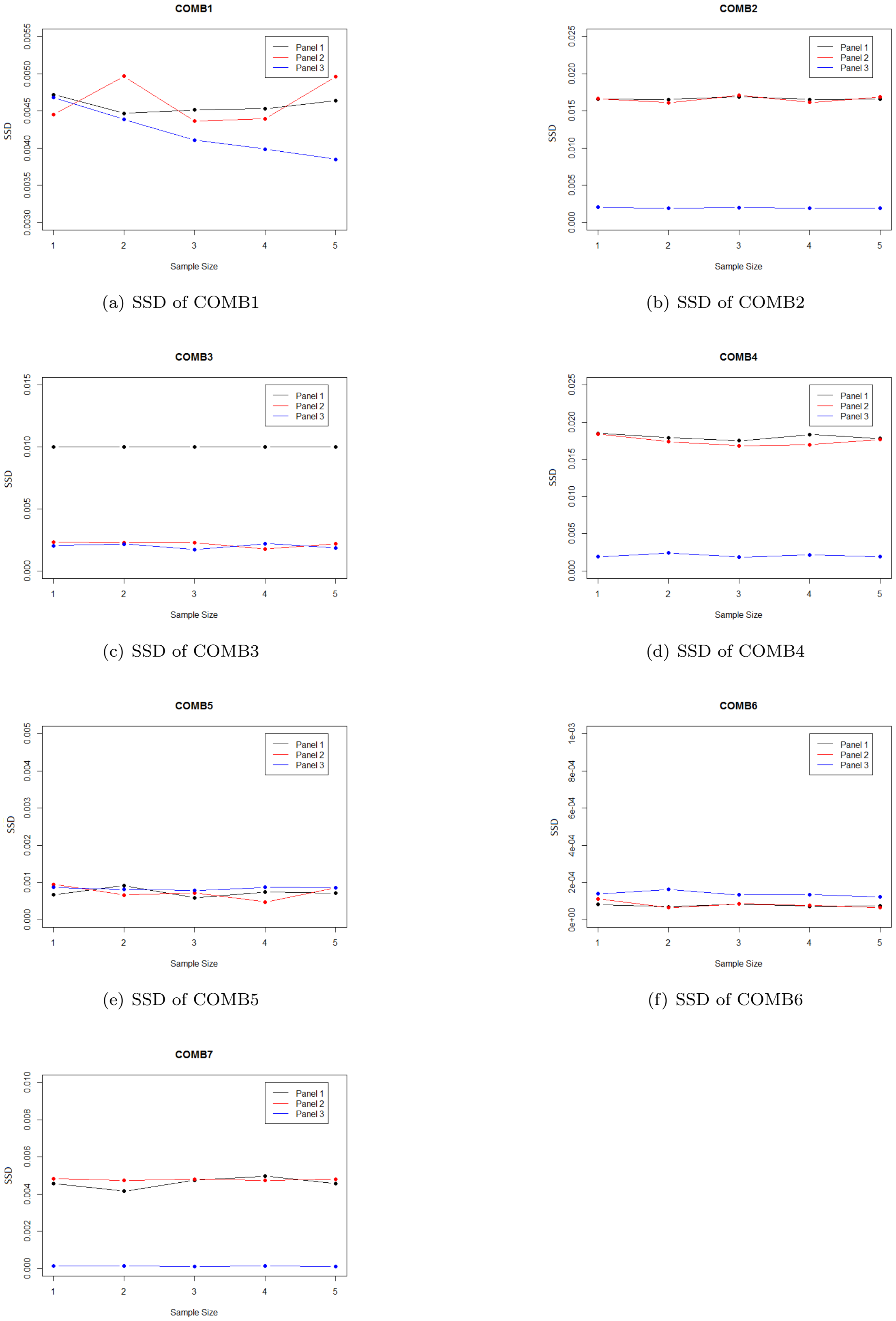


(a) SSD of all combinations for Panel2

(b) SSD of all combinations for Panel3

Figure 3: Pairwise comparison of panels for all background variables combinations

The last factor which affects the results of exact matching is the method of sampling. Figure 4 gives the results of weighted sampling with all other conditions the same as experiments in random sampling. The comparison show that rather than improving the matching and decreasing the SSD, weighted sampling shows a drawback. In the majority of groups the SSDs of weighted sampling are higher than those of simple random samples, and this phenomenon is especially clear in COMB4 and COMB6 from which we can observe that the SSD of weighted samples are higher than those of random samples. The only outlier is COMB5, which shows SSDs of random samples around 0.009 while the SSDs of weighted samples drop to around 0.001. In the second graph, the comparison of background variables of weighted samples shows a similar trend as for random samples with COMB6 (age and working field) and COMB7 (age and occupation) outperforming other combinations, and graphs 2, 3, and 4, having similar values for SSDs. COMB5 is an exception since the SSD of COMB5 is substantially improved by weighted sampling.



(a) SSD of COMB1

(b) SSD of COMB2

(c) SSD of COMB3

(d) SSD of COMB4

(e) SSD of COMB5

(f) SSD of COMB6

(g) SSD of COMB7

Figure 4: SSD of exact matching with weighted sampling

4.2.2 Results of Predictive Mean Matching

In predictive mean matching, the observed conditions are different in sample and panel size, methods of sampling and different scaling level during the optimal scaling regression. Although in general SSD of PMM is higher than that estimated through exact matching when the optimum combination of background variables is used, PMM shows a substantial improvement when we use different scaling level and panel size.

- *Sample and Panel Size*

Different from exact matching, PMM is more sensitive with respect to the size of the sample. Figure 5 shows the parallel comparison of different panel sizes in terms of SSD. As the line graph shows, in all individual panels the SSDs generally decrease with the increase of sample size: SSDs decreased from 0.065 to 0.052 in panel1, and decreased from 0.050 to 0.041 in panel2 while in panel3 the SSD decreased from 0.036 to 0.034. This seems to be the general trend in all groups of experiments: with other conditions fixed, a larger sample size results in smaller values for SSD. However, the panel size effect appears more obvious than the effect due to the sample size. In the comparison, the SSDs decreased from around 0.06 in $panel1$(20,000 units) to 0.034 in $panel3$(60,000 units). In the groups two and three, the effect of sample size and panel size appears identical to the first group (see second and third group). Therefore, a conclusion can be drawn that in PMM sample size slightly affects the matching results while the size of the panel appears to be a substantial condition that can improve the quality of matching and reduce the SSDs. A large panel would be more effective in predictive mean matching.

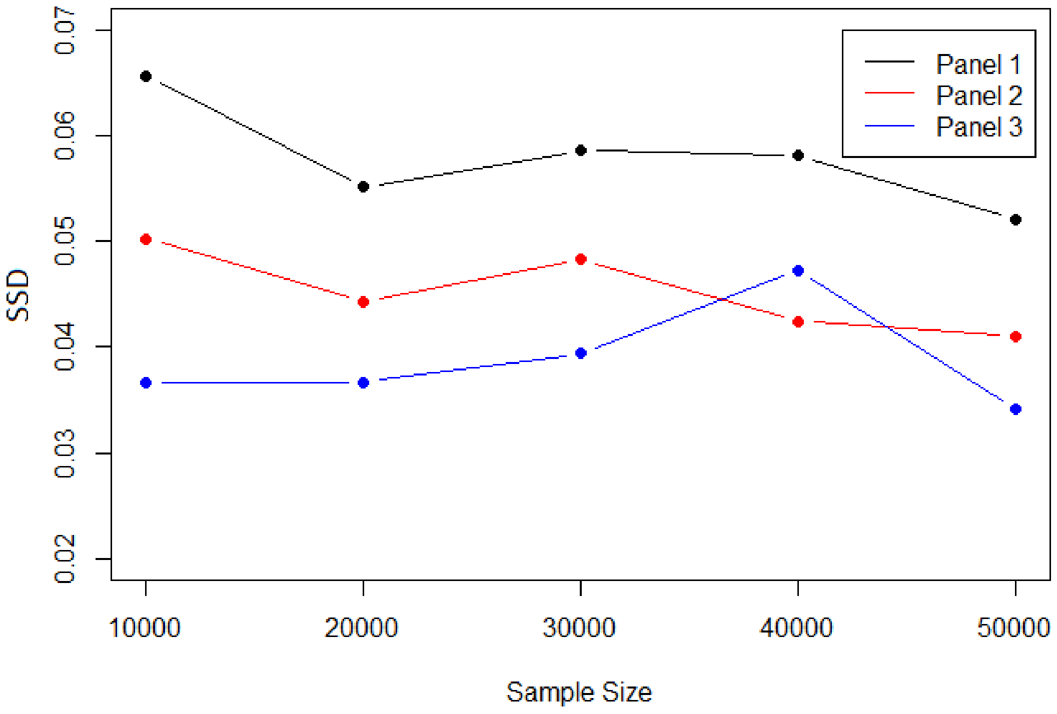
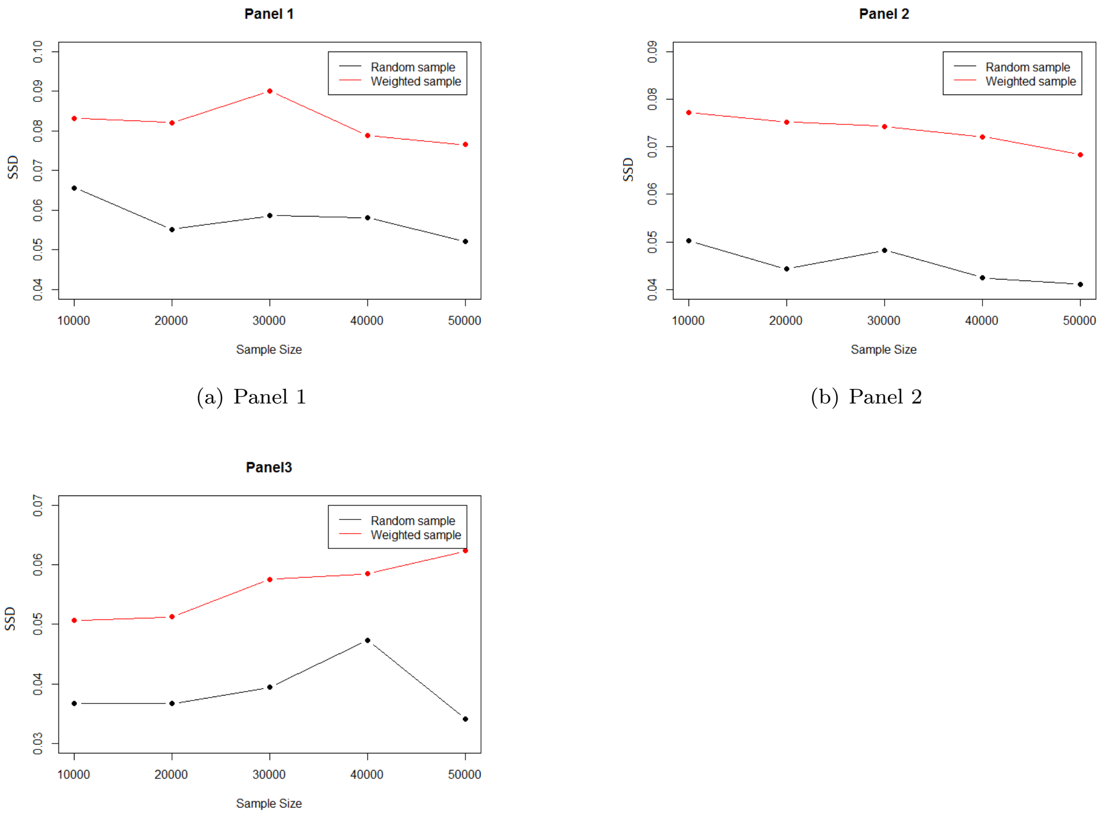


Figure 5: SSD of PMM for random sampling

•*Method of sampling*

This experimental condition differs from above in the method of sampling. In this section we compared PMM results by using random sampling with results obtained by weighted sampling. Theoretically, weighted sampling is able to generate a sample that closer resembles the panel distribution, therefore it is expected to result in higher accuracy. However, as in the graph (Figure 6) shows, the SSDs of weighted samples have substantial drawbacks compared to the random samples. Among all comparisons, the SSDs of weighted samples are substantially higher than those of random samples. In contrast to the original hypothesis, weighted sampling is not helpful in increasing the matching accuracy.



(a) Panel 1

(b) Panel 2

(c) Panel 3

Figure 6: Sampling methods comparison of PMM for different panels

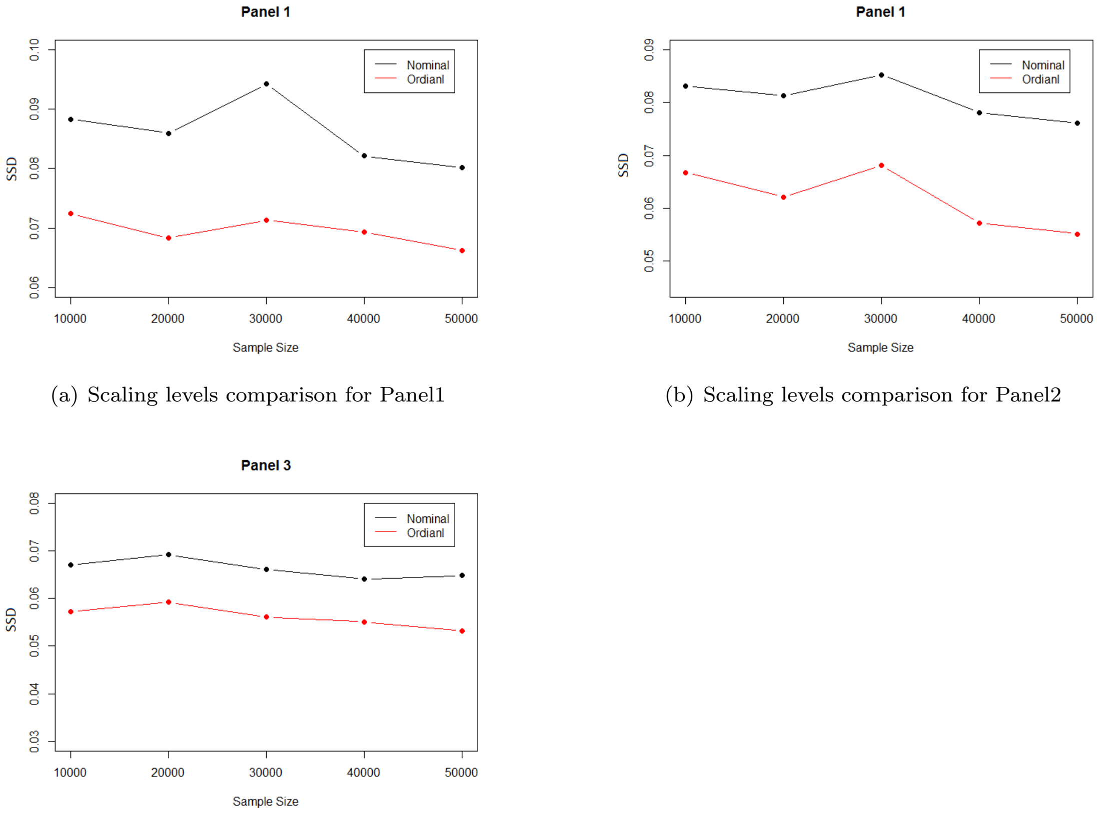
- *Scaling Levels*

The following graph (Figure 7) shows the PMM results for different scaling levels. In ordinal scaling regression, background variables "Educational Level" and "Age" are set as ordinal variables. By setting these 2 background variables as ordinal ones, it is expected that the ordinal information inside the variables contribute to the regression precision. On the other hand, more restrictions have been added which could affect the regression.

Figure 7 is the comparison of ordinal scaling regression and nominal scaling regression. We can see that the differences between these two scaling levels enlarge as the sample size increases. For sample sizes 10,000 and 20,000, even though the results are very close, we can still observe that the ordinal level performs slightly better than the nominal level. This difference starts to become obvious as the sample size reaches 40,000, and becomes stable and clear with 60,000 sample units. The results indicate that the ordinal information will be helpful for PMM. In the lower graph of Figure 7, a comparison of different scaling levels based on a smaller panel is shown. In this graph the general trend remains the same as the graph in Figure 7, with ordinal scaling outperforming the nominal scaling level and the differences becoming more obvious when the sample size increases. In all three panels, the trend of the scaling levels' effect towards the matching accuracy is almost identical. The usage of ordinal information increases the performence of matching.

Appendix III gives the transformation plots of optimal scaling regression. As the transformation plots show, some categories in variables "Occupation" and "Working field" are transformed into numeric values which are quite similar, which could reduce the accuracy of regression and matching. The reason is that during the transformation a step function was used to transform the categories. One way to solve the phenomenon that several categories are transformed into similar values is to add restrictions to the transformation by using a spline function for the scaling. By using a spline function to scale the categories, restrictions are added during the transformation. Therefore different categories can be transformed into numeric values which are not too close. However, adding restrictions could reduce the transformation accuracy which leads to a decrease of the accuracy of matching. We use a spline function to adjust the numerical scaling. The adjusted transformation plots are shown together with all transformation results in Appendix II. As the plots show, categories in variable Occupation and Working Field have all been transformed into individual numeric values and the values are not too close.

Figure 8 presents the results for adjusted scaling matching. By using spline function restrictions, the spline ordinal level always performs more accurately than nominal spline levels. Figure 8 shows that with the same level of scaling, scaling with spline restrictions performs better than without these restrictions. With all scaling level comparisons in the graph, the spline ordinal scaling has the lowest value of SSD in predictive mean matching.

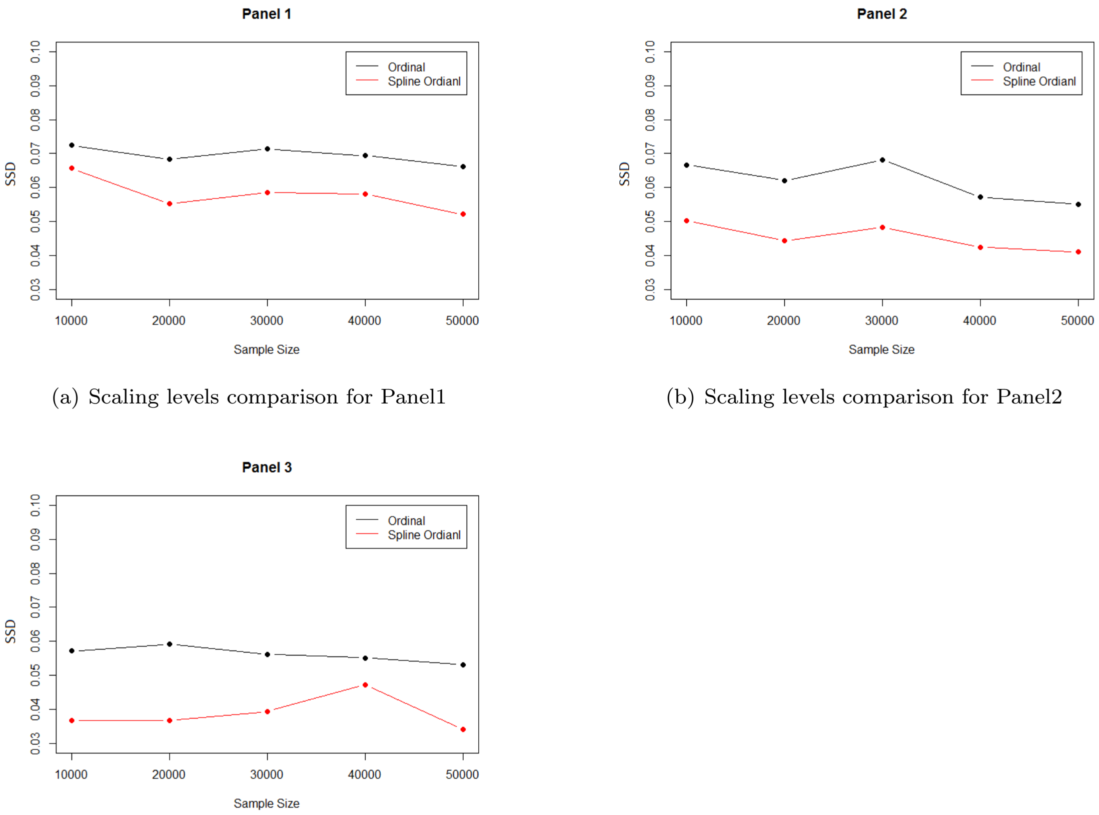


(a) Scaling levels comparison for Panel1

(b) Scaling levels comparison for Panel2

(c) Scaling levels comparison for Panel3

Figure 7: SSD of nominal and oridinal scaling levels for PMM



(a) Scaling levels comparison for Panel1

(b) Scaling levels comparison for Panel2

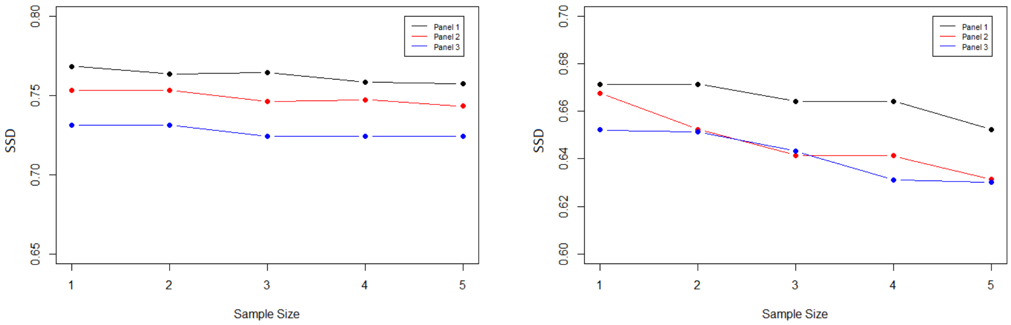
(c) Scaling levels comparison for Panel3

Figure 8: SSD of ordinal and spline ordinal scaling levels for PMM

4.2.3 Results of Propensity Score Matching

With the propensity score matching method the units are matched based on the most similar value of propensity scores. All the matches with PSM returned a much higher value of SSD compared to exact matching and PMM. In PSM, we used two sets of combined background variables to calculate the propensity scores. Apart from the variables chosen by CART (educational level, occupation, working field, age), variable marital status is also included for a comparison considering that its Cramer's V value is higher than 0.35 which indicates a medium correlation towards the target variable.

Comparing the results of PSM with other matching methods, PSM performs the worst with an average SSD over 0.6. The results of PSM also show that the sample size and panel size seem to have no effect towards the matching. Furthermore, the additional variable marital status shows no effect towards the matching results, so this variable can be excluded.



(a) PSM results with random sampling

(b) PSM results for weighted sampling

Figure 9: SSD of propensity score matching

4.3 Results of Random Forest Prediction

The SSDs of PMM were substantially higher than that of exact matching in all experiments. In the study of predictions, we expect that models built based on a matched dataset will result in lower prediction bias than the selective panels due to the reason that the distribution of matches are closer to the distribution of the test dataset. In this section, prediction SSD of a same test set from panel models and matched models will be compared. Considering that the datasets are purely categorical, random forest is used to train the model and the population is used for testing. All matches are designed to shift the distribution of the training dataset closer to the distribution of the test set (population).

We use random samples for our test. In exact matching, the aim is to test whether a model built based on a match with lower SSD for the estimated proportion would be more accurate. Three groups experiments are implementd: Group I consists of background variables with “Age" and "Occupation" which is the best combination of background variables among all matching methods. Group II consists of variables "Age", "Occupation" and "Working Field", with slightly higher SSD than Group I. Group III uses the match with background variables "Occupation", "Working Field" and "Educational Level". The selected three groups have different SSDs for the exact matching estimate. The aim of this experiment is to test whether a matching dataset with lower SSD would improve the accuracy of the prediction models.

The accuracy of the matching dataset is compared with predictions directly based on the selective panels, and SSDs are used to measure the accuracy of predictions. The results under each experiment condition are shown in Figure 8.

As the plot shows, accuracy of predictions under the matched datasets has been improved compared to the predictions based on the selective panels, but the differences are not too obvious. Under most panel size conditions, a matched set of Group II in exact matching performs better than any other matches and results in the lowest value of SSD in prediction. Group I which is the most accurate matching estimate of the population performs slightly better than Group III, thought sometimes SSDs are even higher than Group III. The PMM dataset shows no obvious difference with the panel predictions, but for some panel size conditions, the SSDs of PMM show a decrease compared with direct predictions based on panels.

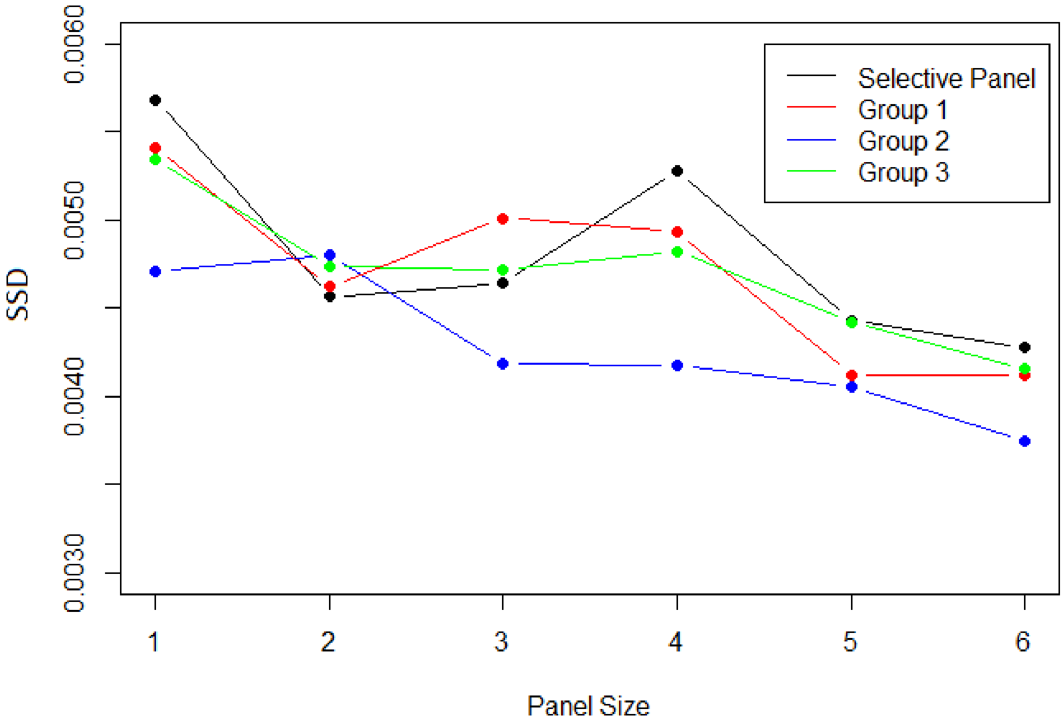


Figure 10: MSE of random forest predictions

Result of the comparison show that predicting models based on a matched dataset could increase the accuracy of models, however our original hypothesis that a model built on a matched dataset which is closer to the population will have more accurate prediction is incorrect. In the above result, Group III of exact matching has the lowest SSD while the performance with respect to removing selectivity is worse than Group II.

5 Discussion

In this thesis, we investigated possible methods of sample matching to remove selectivity in a survey dataset. Methods of sample matching include two parts: the way of drawing samples from the population and the way of matching. The main topic is methods of matching which include exact matching, predictive mean matching and propensity score matching. All matching methods are developed specifically for a categorical dataset. The results of sample matching, which are called the matches, are used to estimate the population distribution of target variable "Economic Status". It is expected that sample matching is able to remove the selectivity issue in a dataset and the matches can improve the accuracy for further analysis (e.g. regression predictions). The performance of all methods are compared with each other by the means of SSD, and models are developed based on relevant matches to predict the distribution of target variables. The performances of the methods are tested under several different conditions. In the method of exact matching, experiments examine different combinations of background variables. Predictive mean matching simulations differ with respect to scaling levels. For propensity score matching, the main differences are the size of the selective panel and representative samples. All methods are tested under difference sizes of panel and sample. The object is to observe how different methods and conditions affect the matching result, and whether sample matching is able to reduce the prediction bias in random forest classification. The target variable was the population proportion of "Economic Status" from the Dutch Population Census 2001 dataset. Assuming that the distribution of the target variable is available but selective, and the rest of the dataset was used as the population distribution whose values of the target variable are unknown.

The result was that among all the methods, exact matching with random sampling is able to generate the most accurate estimate of the target variable distribution of the population. The involved background variable is the combination of "age" and "occupation". In general, exact matching outperforms all other methods of matching, and exact matching with random sampling can improve the accuracy of the matching. Estimates of PMM are not as accurate as exact matching in all cases, but one interesting phenomenon is that by adjusting the conditions in PMM, the accuracy can be substantially improved. The experiment results show that propensity score matching is not suitable since this results in biased estimates.

In this study, we have compared the scaling methods of stepwise function and monotonic spline function. The study shows that adding a constraint into the scaling (using spline transformation functions) can substantially improve the accuracy of the PMM estimate. Furthermore, maintaining the ordinal information between categorical values also positively affects the results of the PMM estimate. This interesting phenomenon indicates that a further study on the proper methods of scaling could be conducted to improve the matching accuracy. To further explore PMM matching, perhaps more matching variables characteristics and their influence on the scaling can be investigated. This requires a deep study on matching variable selection, or the methods could be implemented in another dataset whose background variables X have a stronger explanatory ability. The scaling methods for PMM might outperform the exact matching.

As the random forest results show, removing selectivity by sample matching is able to reduce the bias of random forest prediction compared to the models based on a selective panel. However, the results of experiments contradict the original hypothesis that a more accurate matching dataset (with smaller SSD value) would also be more accurate with respect to reducing prediction bias. The most accurate prediction is not based on the most accurate match, but an intermediate match which has a relatively small SSD. Further study on reducing prediction bias can focus on which levels of accuracy of a match can optimally fit the population predictions.

In summary, the sample matching method is able to remove the estimation bias due to the selectivity of the data set. For purely categorical datasets we have not discovered matching methods that perform better than exact matching. The key of exact matching lies in the selection of matching variables. If the selected background variables fail to cover sufficient information towards the target variable, the matches would result in high bias. However, if too many matching variables are chosen, this will also harm the matching results due to the curse of dimensionality. The newly developed scaling predictive mean matching has not shown a better performance than exact matching, however the methods of scaling remains an interesting topic for further study. In the final experiment, it is shown that removing selectivity in the panel by sample matching is able to reduce the overfitting issue, while the best fitted matching requires a further study.

6 Bibliography

Allison, P. (2015). Imputation by Predictive Mean Matching: Promise & Peril. https://statisticalhorizons.com/predictive-mean-matching.

Bethlehem, J.G. (2015). Solving the Nonresponse Problem with Sample Matching? *Social Science Computer Review,* 34(1).

Bethlehem, J.G., Cobben, F. & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys.* John Wiley & Sons, Hoboken, NJ.

Bishop, C.M. (2006). *Pattern Recognition & Machine Learning.* New York: Springer. p.vii. ISBN 0-387-31073-8.

Caliendo, M. & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys.* Vol. 22, No. 1, pp.31-72.

Chintan, P., James, C., Julian, D., Achille, F., Aditya, K., Aaron, K., Li, M., Edith, S. & Kavitha, S.(2007). Matching Patient Records to Clinical Trials Using Ontologies. *The Semantic Web,* pp 816-829.

Chmura, L., Rivers, D., Bailey, D., Piercea, C. & Bell, C. (2013), Modeling a Probability Sample? An Evaluation of Sample Matching for an Internet Measurement Panel. *The Nielsen Company & YouGov.*

D'Agostino, R, B. (1998). Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomized Control Group. *Statistics in Medicine,* 17, 2265—2281, (1998).

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). *Statistical Matching: Theory and Practice.* p.167-171. ISBN 978-0470023532.

Hastie, T., Tibshirani, R & Friedman, J. 2009. *The Elements of Statistical Learning Second Edition.* p.587-592. ISBN 978-0387848570.

James, G. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer. p.176. ISBN 978-1461471370.

Meulman, J., J. & Van der Kooij, A., J. (2016), ROS Regression: Integrating Regularization with Optimal Scaling Regression.*arXiv:1611.05433* [*stat.ML*]

Rivers, D. (2007), Understanding People Sample Matching. *T 650.462.8000*

Rivers, D. (2007), Sampling for Web Surveys. *2007 Joint Statistical Meetings,* Salt Lake City, UT, August 1, 2007.

Rivers, D. & Bailey, D. (2009), Inference from Matched Samples in the 2008 U.S. National Elections. *American Association of Public Opinion Research,* JSM 2009.

Rosenbaum, P.R. & Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika.* 70 (1): 41–55.

Siddique, J. & Belin, T,R. (2007). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine,*2008; 27:83-102.

Vavreck, L. & Rivers, D. (2008), The 2006 Cooperative Congressional Election Study. *Journal of Elections 18,* pp.355-366.

7 Appendix

7.1 Appendix I: Dutch Census 2001 Data

Variable	Category	Description
Economical Status	111	Employee
	112	Student with job
	120	Independent worker
	210	Unemployed
	221	Education-related
	222	Retired
	223	Houseman/Housewoman
	224	Other inactive
	998	unknown
Gender	1	Man
	2	Woman
	8	Unknown
Age	1	0-4 years
	2	5-9 years
	3	10-14 years
	4	15-19 years
	5	20-24 years
	6	25-29 years
	7	30-34 years
	8	35-39 years
	9	40-41 years
	10	45-49 years
	11	50-54 years
	12	55-59 years
	13	60-64 years
	14	65-69years
	15	70-74 years
	16	75-79 years
	17	80 years older
	98	unknown
Position in the household	1110	Child
	1121	Married without children
	1122	Married with children
	1131	Living together without children
	1132	Living together with children
	1140	Alone living old person
	1210	Living alone
	1220	Different household
	9998	Unknown
Size of household	111	1 person
	112	2 people
	113	3 people
	114	4 people
	125	5 people
	126	6 people or more
	998	Unknown
Residential last year	1	Same COROP area
	2	Other COROP area, or outside the Netherlands
	9	Not applicable (person less than 1 year old)
	998	Unknown

Nationality	1	The Netherlands
	2	From other countries in Europe
	3	Others
	998	Unknown
Country of Birth	1	The Netherlands
	2	From other countries in Europe
	3	Others
	998	Unknown
Educational level	0	Pre-primary
	1	Primary
	2	Lower secondary
	3	Upper secondary
	4	Post secondary
	5	Tertiary
	6	Without any education
	98	Unknown
Occupation	1	ISCO 1; legislators, senior officials and managers
	2	ISCO 2; professionals
	3	ISCO 3; technicians and assistant professionals
	4	ISCO 4; clerks
	5	ISCO 5; service, shop, market sales workers
	6	Other
	7	ISCO 7; craft and relative workers
	8	ISCO 8; plant and machine operators and assistants
	9	ISCO 9; elementary occupations
	998	Unknown
	999	Not working
Working field	111	NACE A+B. Agriculture, hunting, forestry and fishing
	122	NACE C+D+E; mining, manufacturing and electricity
	124	NACE F; construction
	131	NACE G; wholesale, retail trade, repair
	132	NACE H; hotels and restaurants
	133	NACE I; transport, storage, communication
	134	NACE J; financial intermediation
	135	NACE K; real estate, renting and business activities
	136	NACE L; public administration, defence
	137	NACE M; education
	138	NACE N; health, social work
	139	NACE O; other community, social personal service activities
	200	Not working
	998	Unknown
Marital status	1	Unmarried
	2	Married
	3	Widow
	4	Divorced
	8	Unknown

7.2 Appendix II: Cramers'V Value and Gini Impurity

Table 5: Cramer's V value for Panel 1

Target Variable	Background Variables	Cramers'V
	Gender	0.3489
	Age	0.5579
	Position in the household	0.3340
	Size of Household	0.2319
	Residential last year	0.2123
	Nationality	0.0920
	Country of Birth	0.0952
	Educational level	0.3498
	Occupation	0.4033
	Working field	0.4091
	Marital status	0.4464

Table 6: Cramer's V value for Panel 2

Target Variable	Background Variables	Cramers'V
	Gender	0.3481
	Age	0.5597
	Position in the household	0.3348
	Size of Household	0.2306
	Residential last year	0.2034
	Nationality	0.0836
	Country of Birth	0.0893
	Educational level	0.3574
	Occupation	0.4057
	Working field	0.4076
	Marital status	0.4429

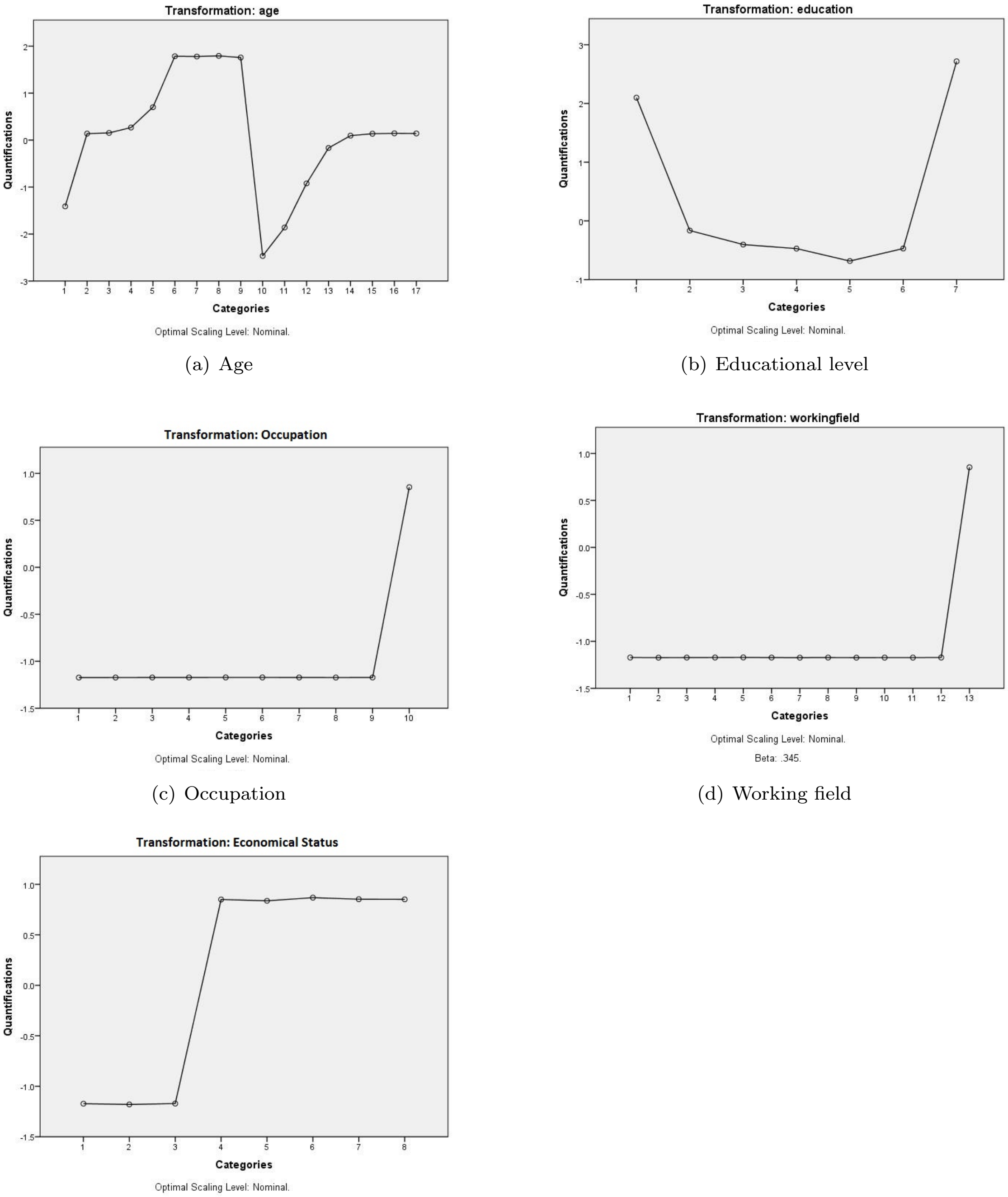
Table 7: Cramer's V value for Panel 3

Target Variable	Background Variables	Cramers'V
	Gender	0.3504
	Age	0.5579
	Position in the household	0.3327
	Size of Household	0.2298
	Residential last year	0.1908
	Nationality	0.0811
	Country of Birth	0.0855
	Educational level	0.3534
	Occupation	0.4047
	Working field	0.4104
	Marital status	0.4336

Table 8: Variable Importance for CART

Panels	Variable importance
Panel 1	Age, Occupation, Position in the household, Educational level, Working field, Gender.
Panel 2	Occupation, Working field, Age, Educational level, Marital status, Position in the household.
Panel 3	Age, Occupation, Educational level, Marital status, Position in the household, Gender.

7.3 Appendix III: Transformation Plots



(a) Age

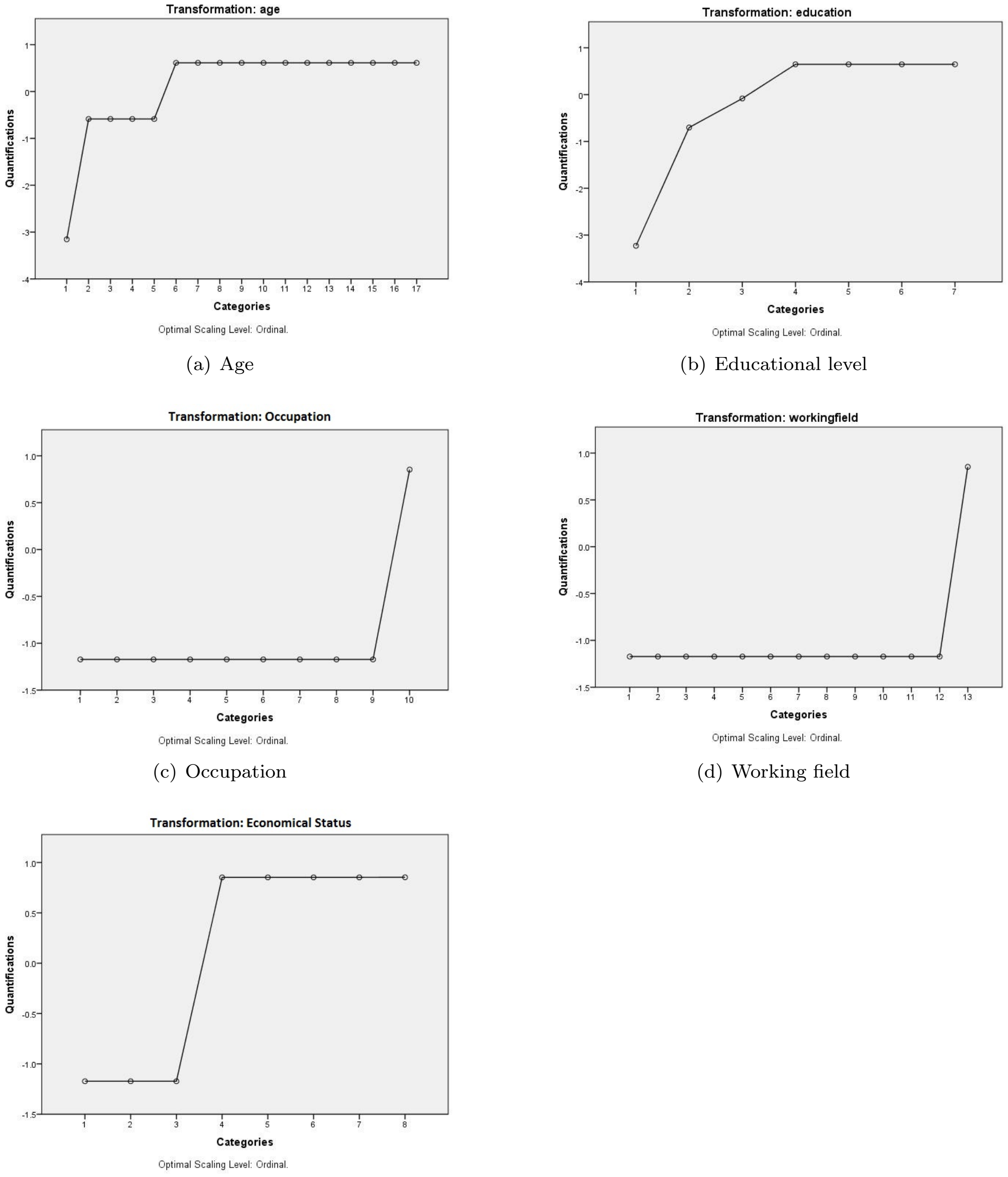
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 11: Panel 1: Nominal Transformation



(a) Age

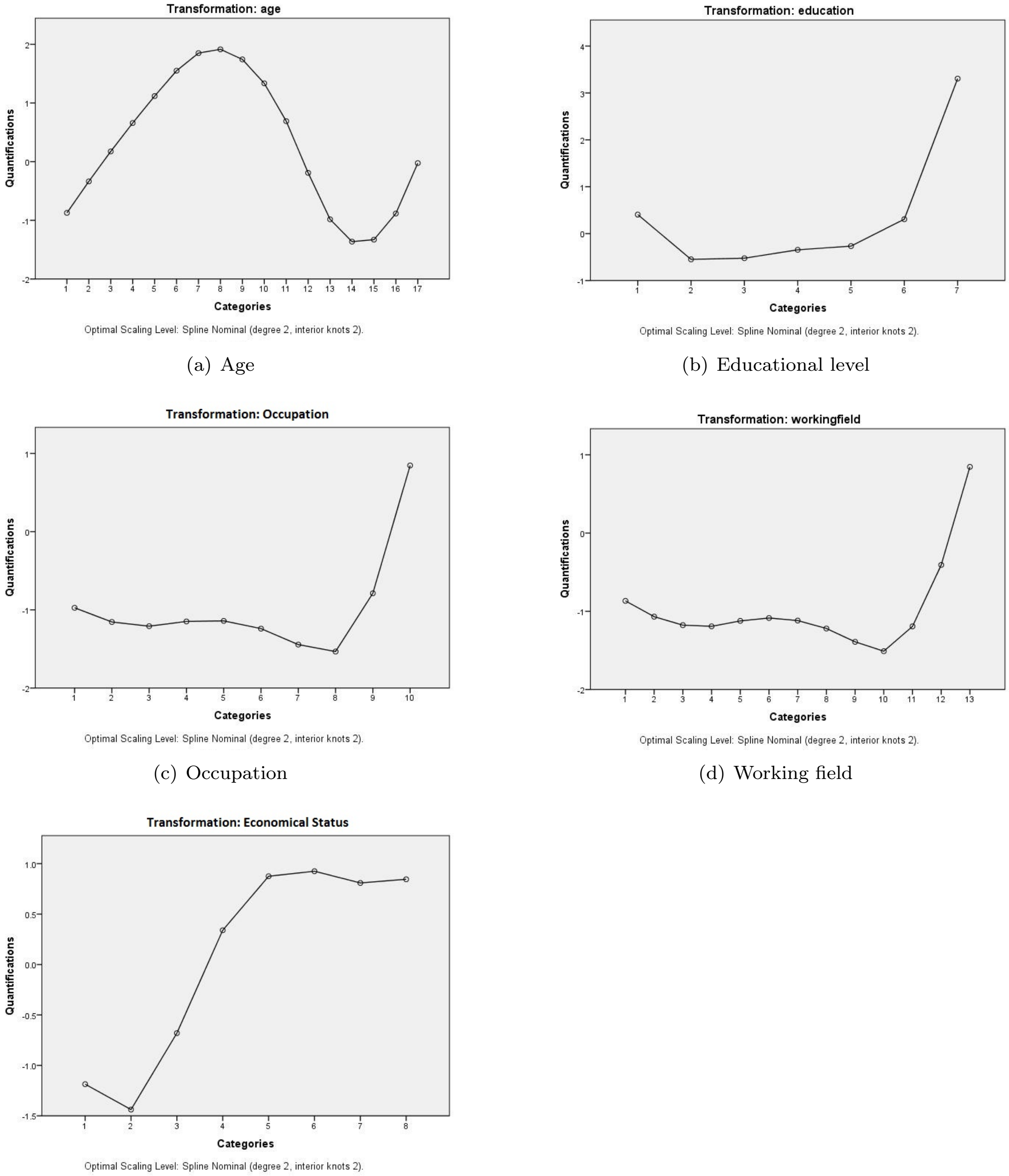
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 12: Panel 1: Ordinal Transformation



(a) Age

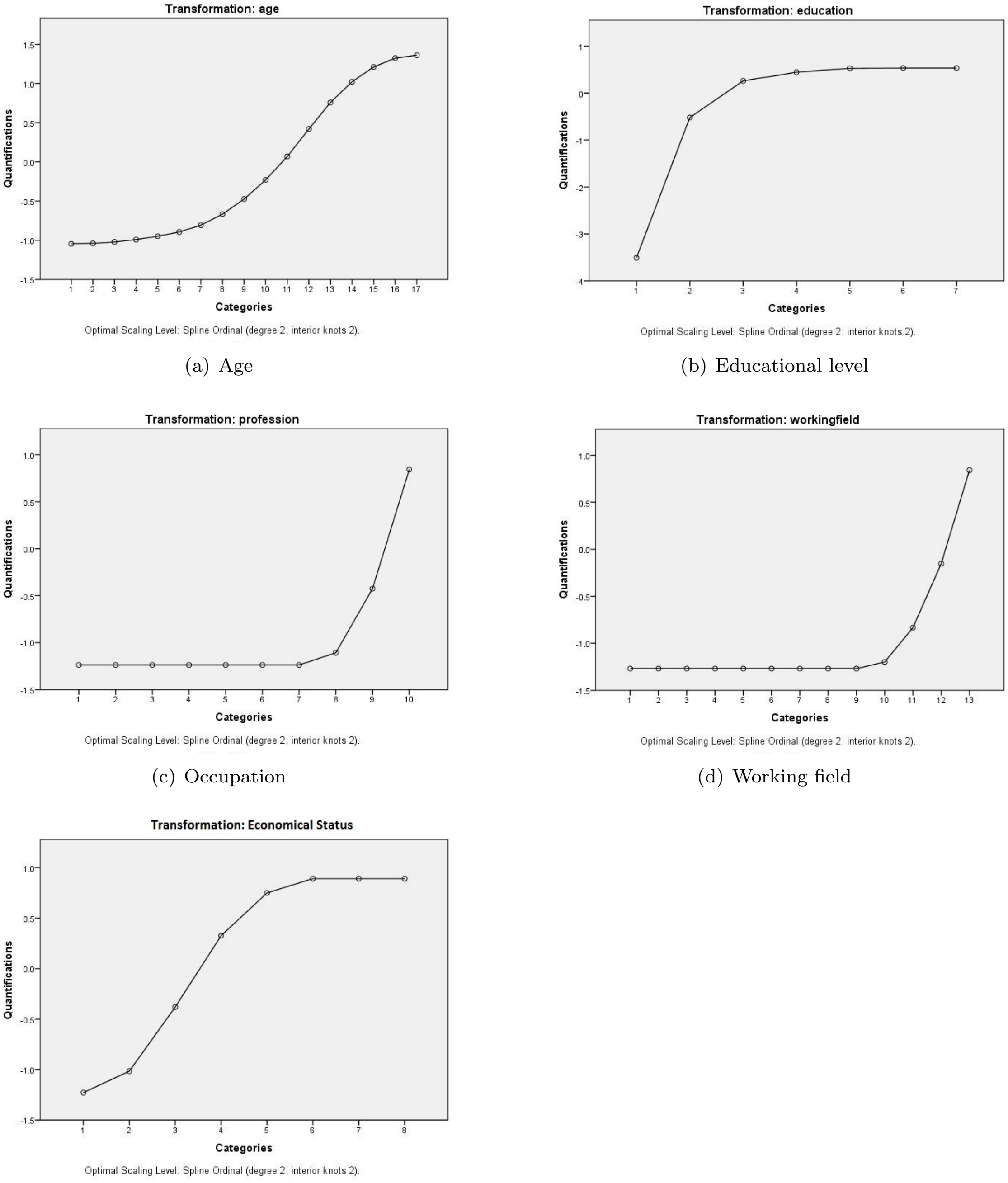
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 13: Panel 1: Spline Nominal Transformation



(a) Age

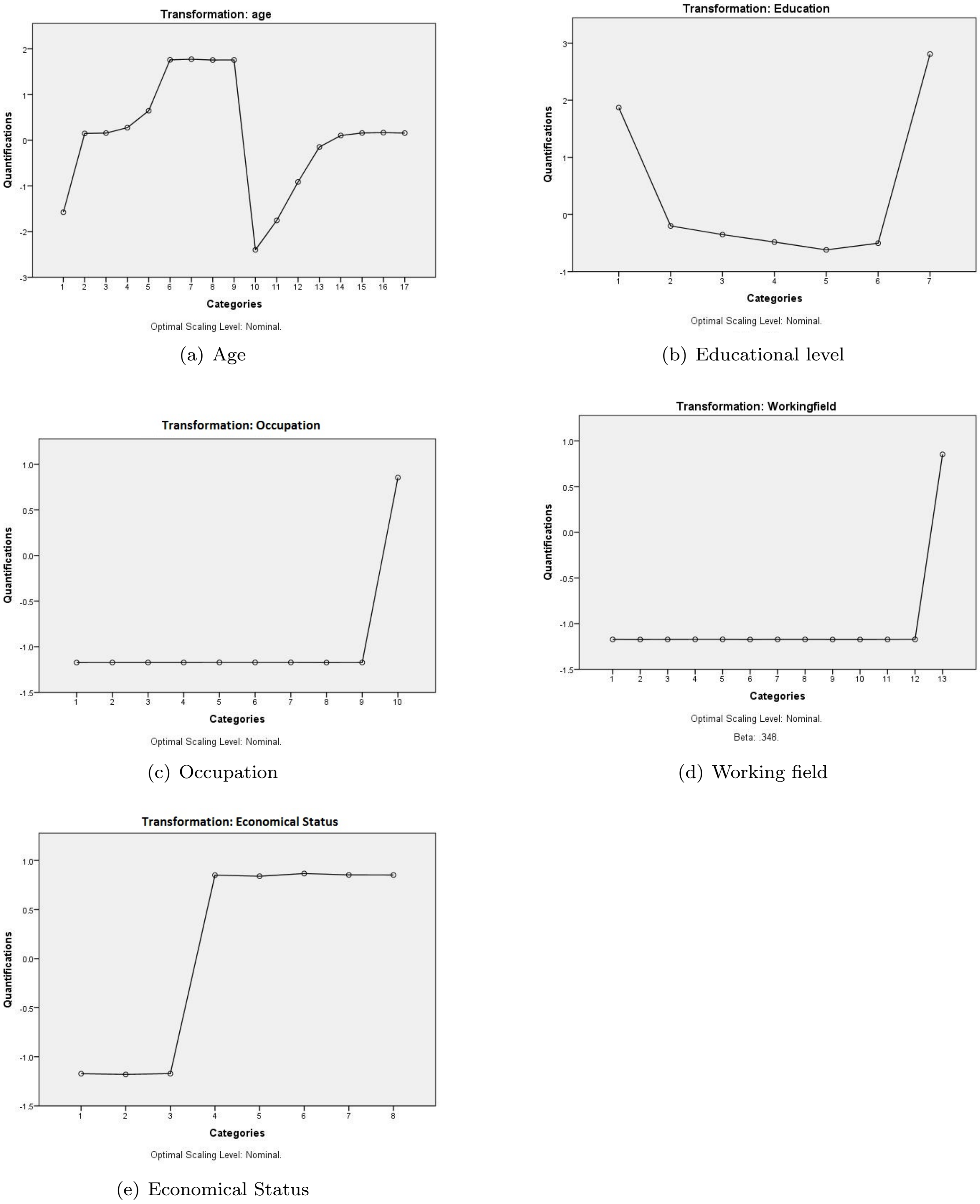
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 14: Panel 1: Spline Ordinal Transformation



(a) Age

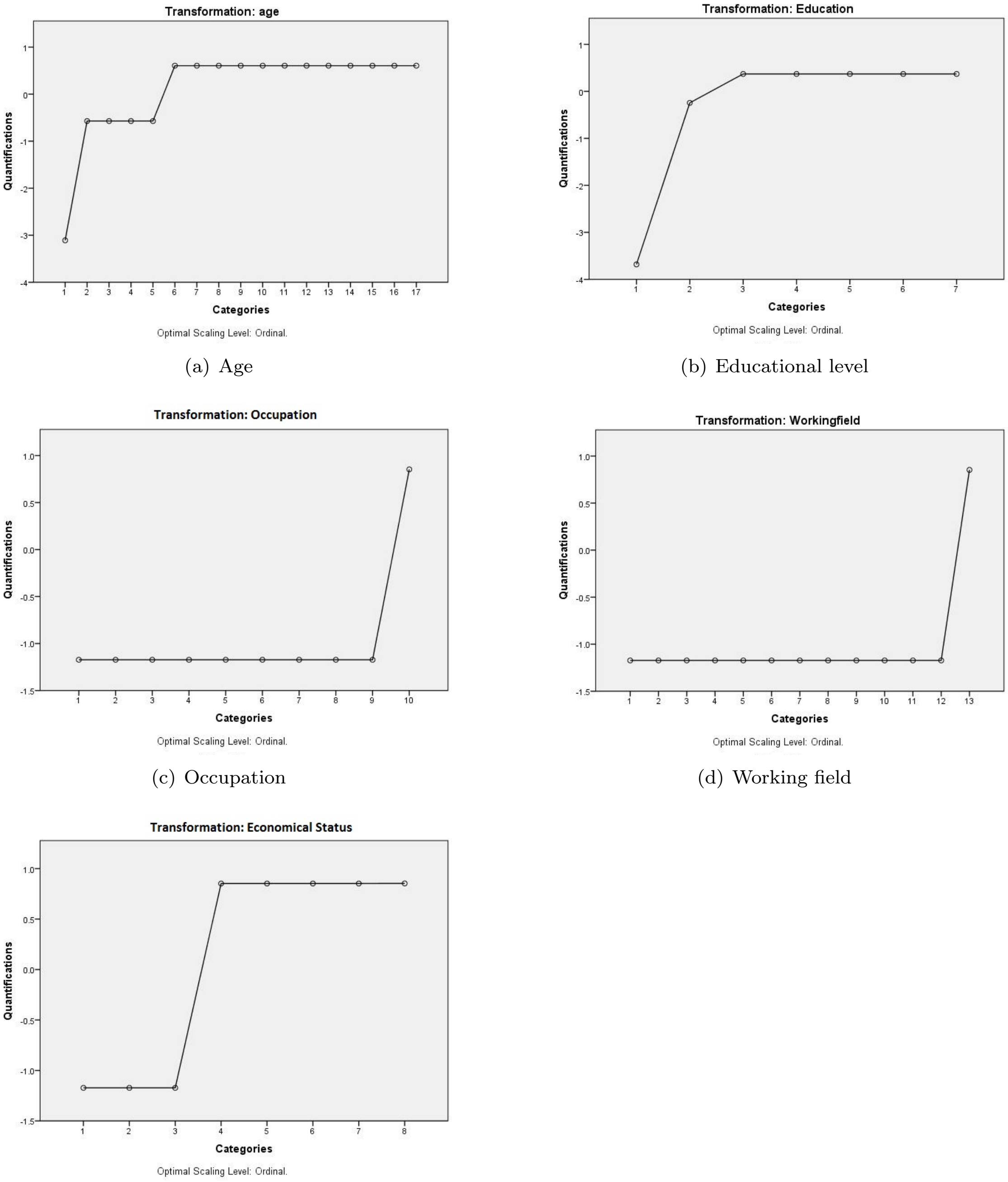
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 15: Panel 2: Nominal Transformation



(a) Age

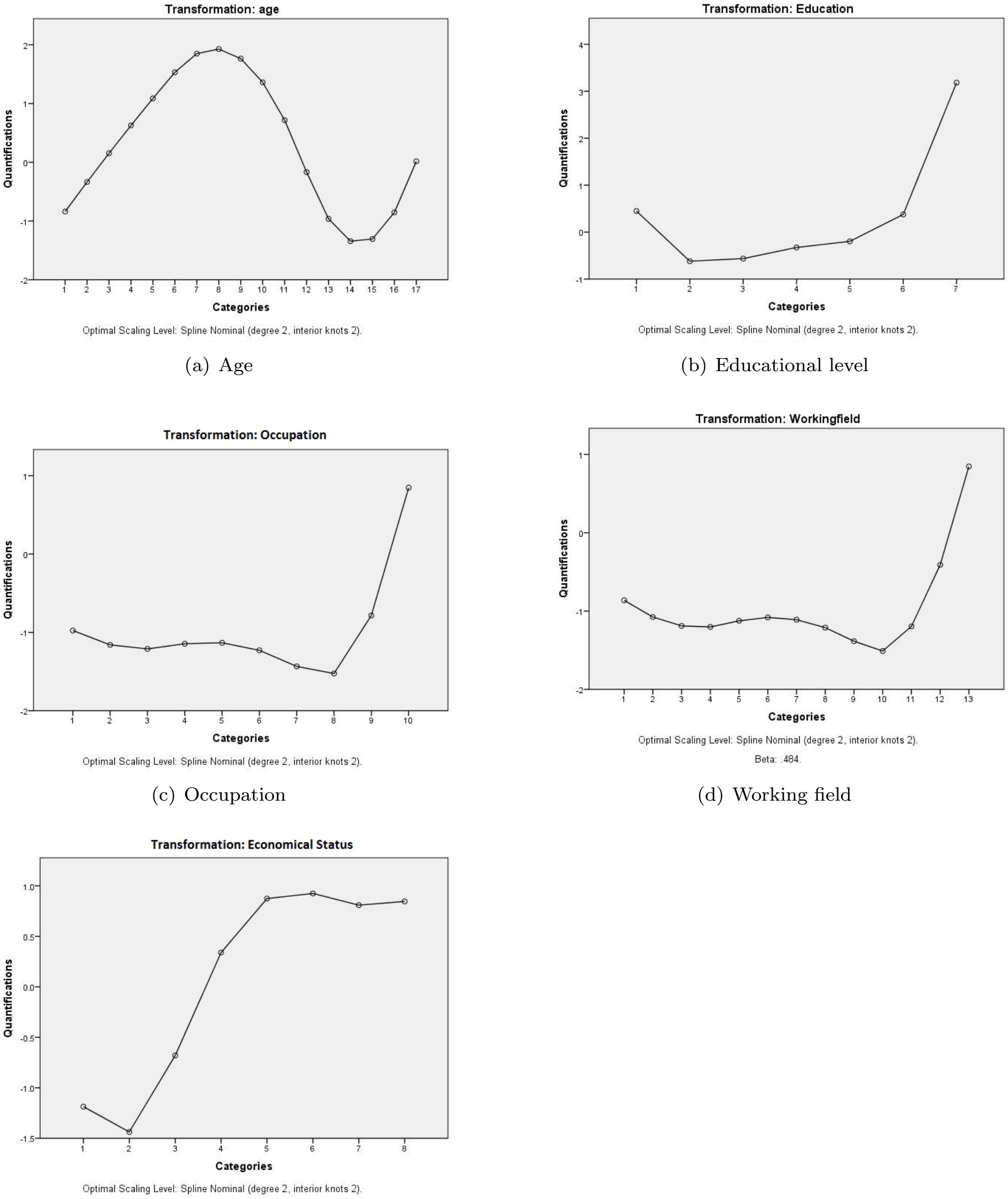
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 16: Panel 2: Ordinal Transformation



(a) Age

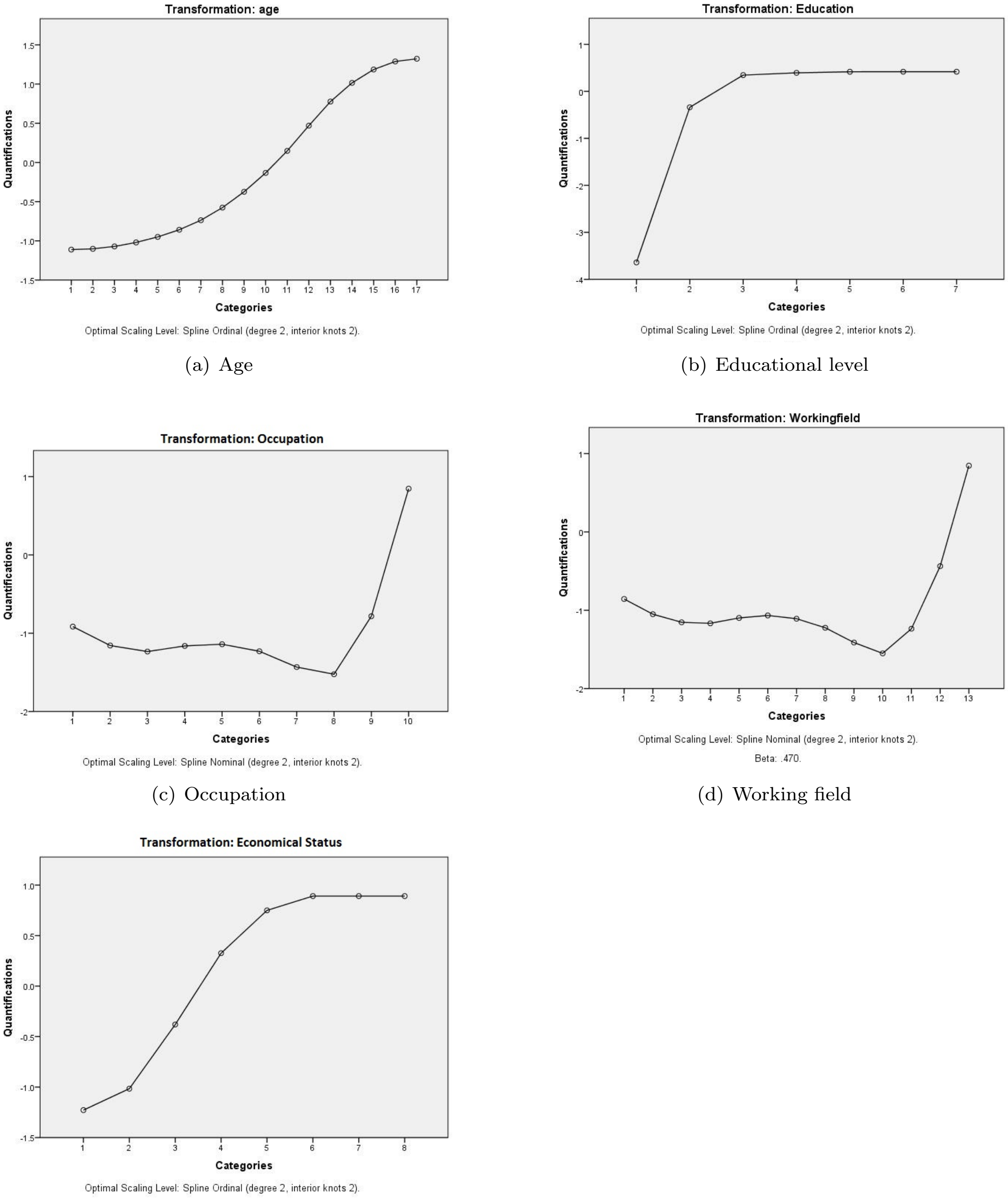
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 17: Panel 2: Spline Nominal Transformation



(a) Age

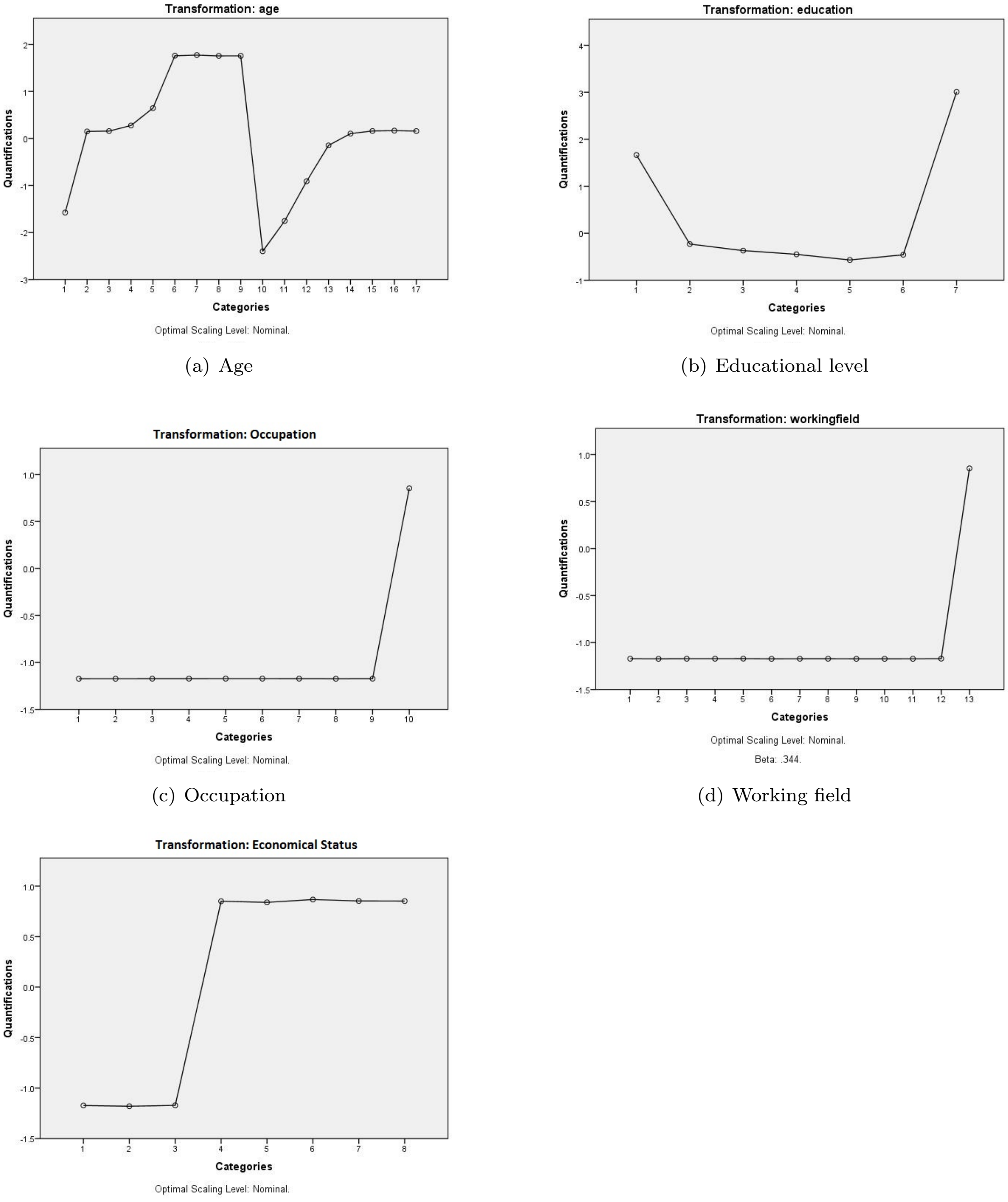
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 18: Panel 2: Spline Ordinal Transformation



(a) Age

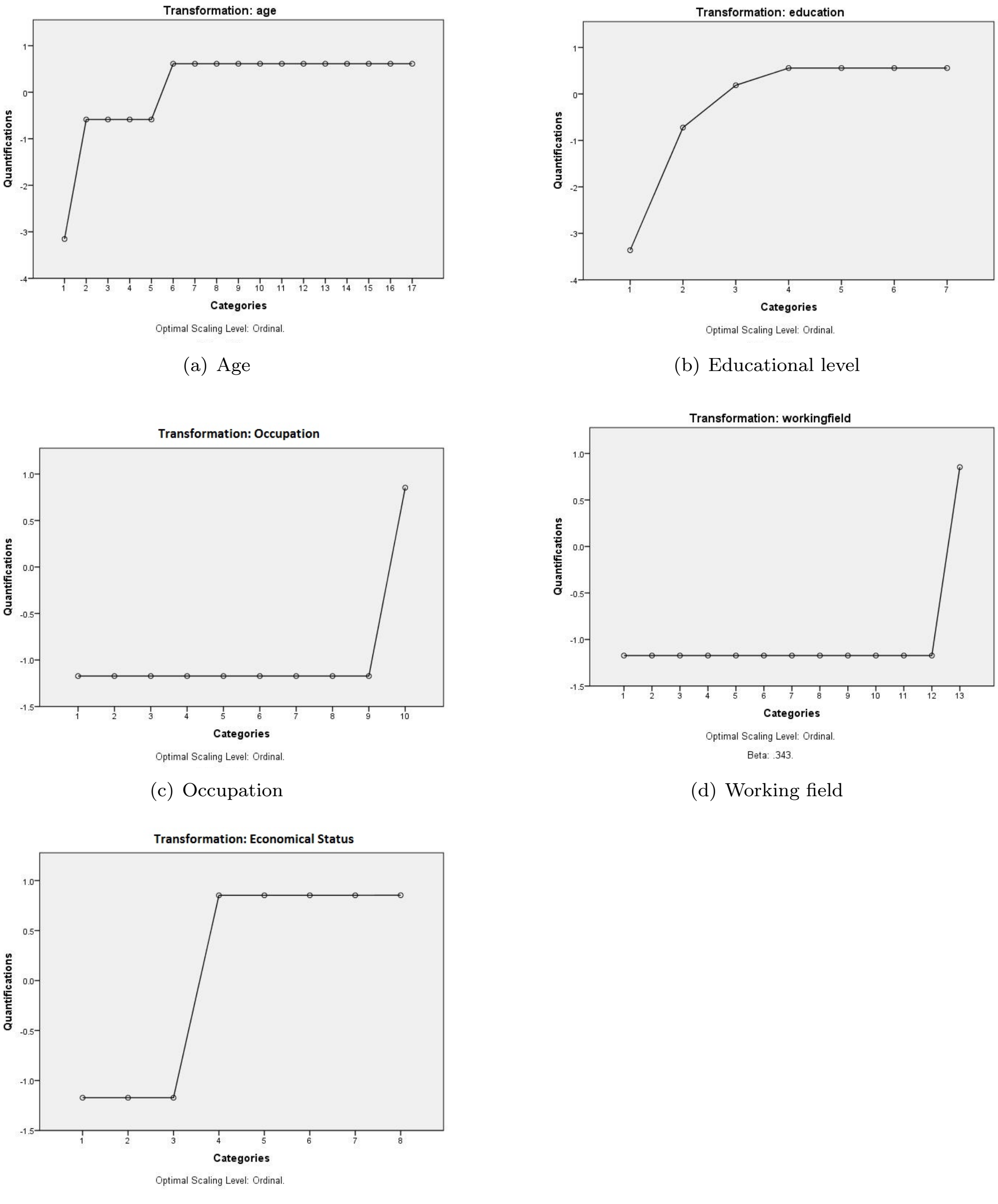
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 19: Panel 3: Nominal Transformation



(a) Age

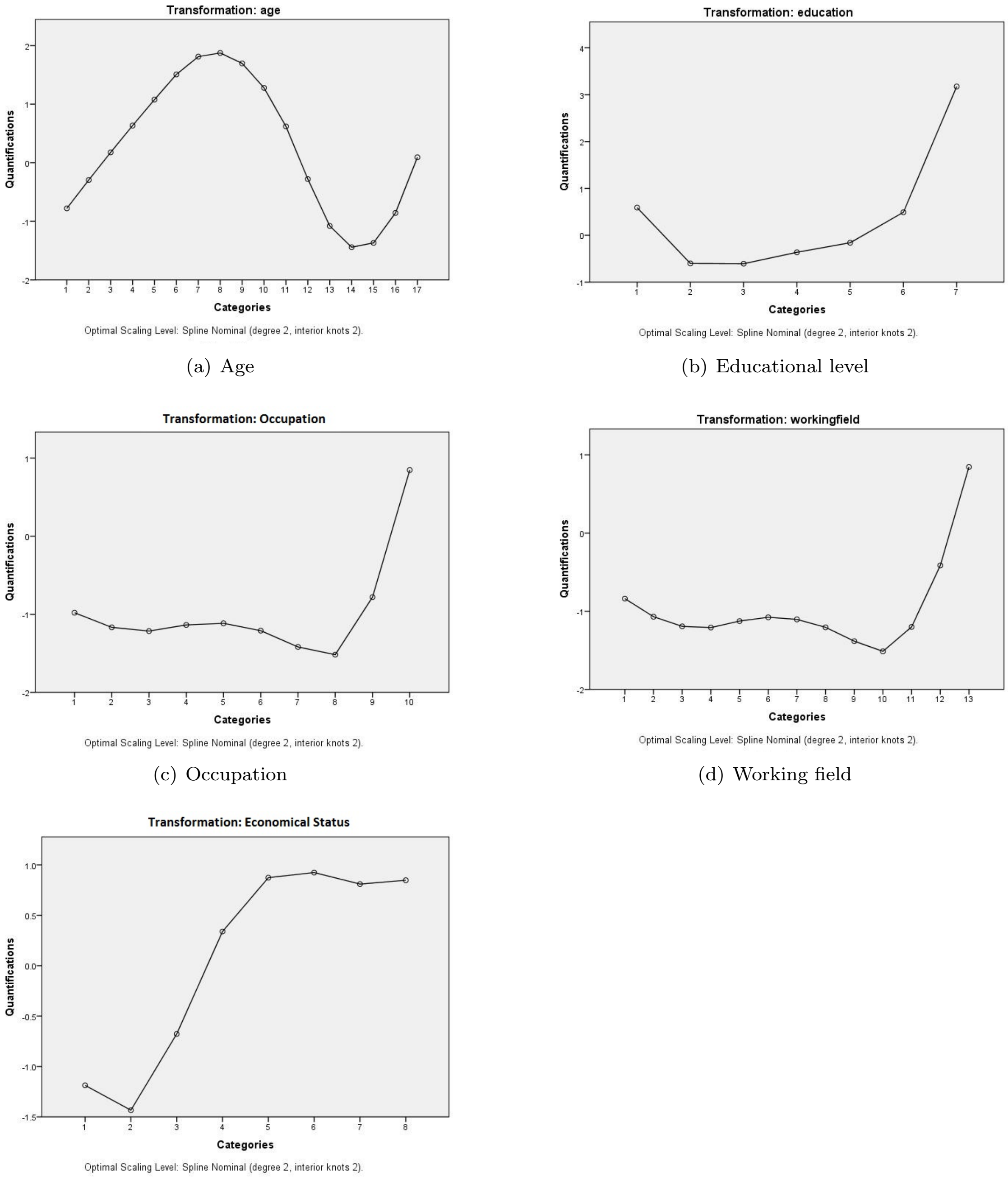
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 20: Panel 3: Ordinal Transformation



(a) Age

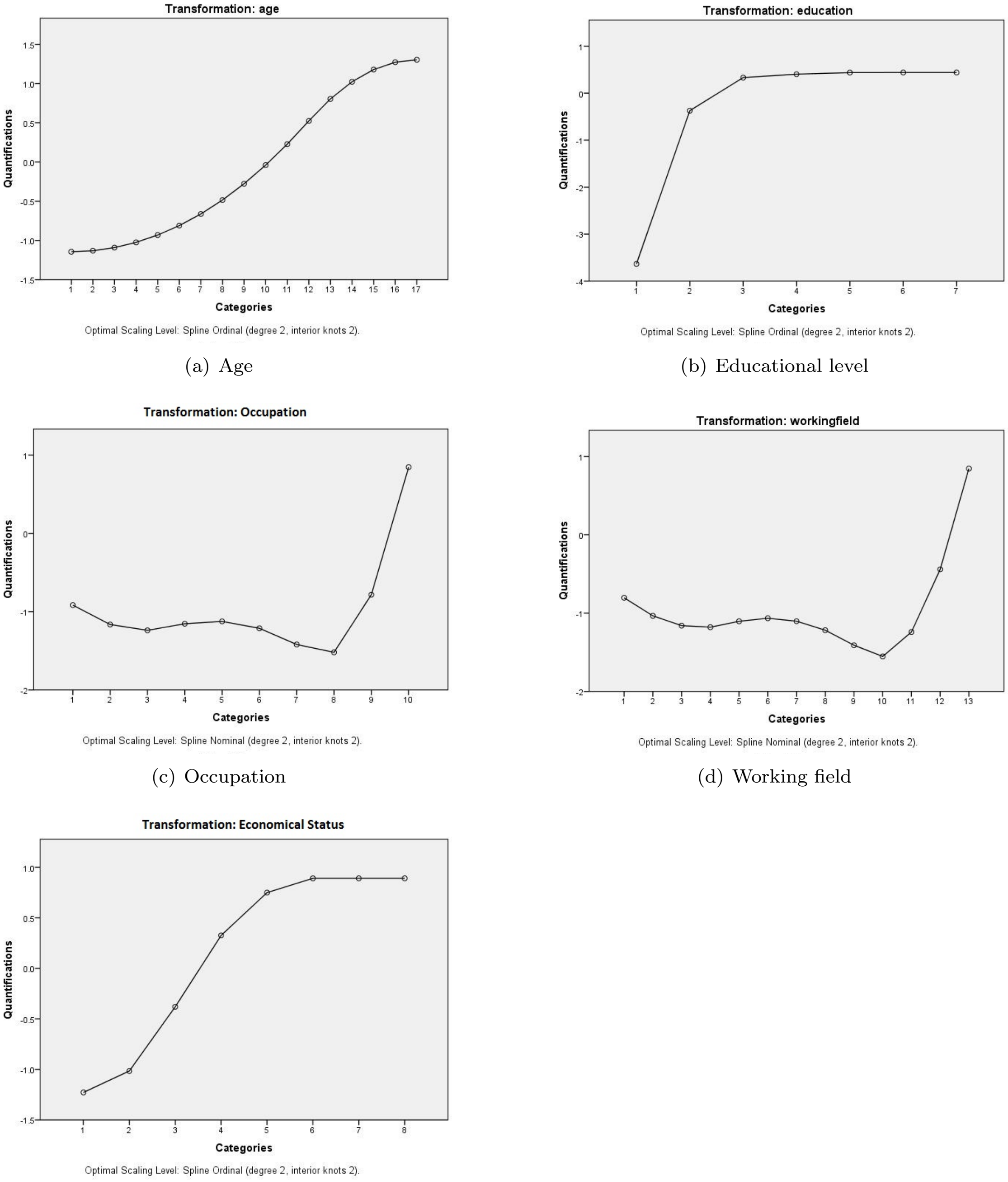
(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 21: Panel 3: Spline Nominal Transformation



(a) Age

(b) Educational level

(c) Occupation

(d) Working field

(e) Economical Status

Figure 22: Panel 3: Spline Ordinal Transformation