Utility of synthetic microdata generated using tree-based methods

Beata Nowok

Administrative Data Research Centre - Scotland, University of Edinburgh, School of GeoSciences, Drummond Street, Edinburgh EH8 9XP, United Kingdom, e-mails: beata.nowok@ed.ac.uk

Abstract. Synthetic data methods aim to allow for the release of high-quality microdata without compromising confidentiality. The risks of disclosures are minimized by replacing some or all of the data values with simulations from statistical models estimated from the original confidential data. The usefulness of the disseminated synthetic data depends, however, on the correct specification of these models which can be a difficult and complex task. In this paper we assess the performance of non-parametric tree-based methods (classification and regression trees, bagging and random forests) for generating synthetic microdata. To evaluate utility of synthetic data we use both analysis-specific and general measures. The former focus on comparing regression estimates from the original and synthetic data, whereas the latter are based on propensity scores and evaluate similarities between overall structures of the confidential data and their synthetic version. All synthetic data are produced and investigated using the *synthpop* package for **R**.

1 Introduction

Wide access to microdata is crucial to the advancement of research and evidence-based policy but it is often constrained by confidentiality concerns. The protection of sensitive information is usually achieved by restricting access or modifying data before they are released to users. Traditional means of disclosure control such as recoding, data swapping or noise adding may lead, however, to significant information loss. A promising alternative in the form of synthetic data was proposed by Rubin (1993) but applications are limited due to the complexity of generating synthetic data. This may change with recent advances in the field which can simplify the synthesising process considerably. First, some machine learning techniques have been successfully adopted for generating synthetic data, which was initiated by Reiter (2005) who suggested using the classification and regression trees (CART) for this purpose. Second, software for data synthesis (a *synthpop* package for **R**) which implements such methods is under development.

In this paper we evaluate utility of synthetic data produced using tree-based methods available in the current version of the software (*synthpop 1.2-0*). They include CART, bagging and random forest and they are described in Section 3 after

brief presentation of the implemented synthesising method in Section 2. Section 4 presents utility measures which are used for empirical evaluation of synthetic data sets in Section 5. Section 6 contains concluding remarks.

2 Synthesising algorithm

Variables are synthesised one-by-one by fitting a sequence of regression models to original data and drawing synthetic values from the corresponding predictive distributions. The fitted models are conditioned on the original variables that are earlier in the synthesis sequence, so the number of covariates increases for subsequent variables with the last being conditioned on all other variables. We use a series of conditional distributions to approximate the joint probability distribution of the data because specifying the latter is usually unrealistic in real data applications. Similar conditional specification approaches are used in most implementations of synthetic data generation. They provide a simple way to define models for each variable separately and to take into account complex interdependencies between variables such as logical constraints or missing data patterns.

3 Tree-based methods for data synthesis

Machine learning techniques such as tree-based methods provide a promising alternative to parametric methods for data synthesis. In principle they can automatically detect patterns in data and use them for prediction, which can significantly simplify the process of developing synthesising models. Classification and regression trees (CART) were suggested for generation of synthetic data by Reiter (2005) and random forest by Caiola and Reiter (2010). Drechsler and Reiter (2011) evaluated performance of CART, bagging, random forest and support vector machines and concluded that synthesis based on regression trees can produce data that give reliable estimates. Below we present tree-based methods that are implemented in the *synthpop* package and will be assessed in this paper. They include two CART approaches with different splitting rules, bagging and random forest.

3.1 CART

CART (Breiman et al. 1984) are an algorithmic modelling approach that can be applied to any type of data. The basic idea is to recursively split a data set into groups with increasingly homogeneous outcome. The splits, which can be represented by a tree structure, are specified as yes-no questions referring to the predictor space. The values in each final group (leaf) approximate the conditional distribution of the predicted variable for units with predictors meeting the criteria that define that group. The synthetic values are generated by sampling from an appropriate group.

The first CART implementation performs splits based on testing the global null hypothesis of independence between any of the explanatory variables and the out-

come variable. If this hypothesis cannot be rejected, the splitting process stops. Otherwise an independent variable with strongest association with the response variable is used to define splitting criterion. Then these steps are repeated recursively for each node. The alternative CART algorithm use Gini index and deviance to find best splits for classification and regression respectively.

3.2 Bagging and random forest

Bagging and random forest use CART as building blocks to construct prediction model with improved accuracy. In bagging, called also bootstrap aggregation, a number of bootstrap samples is drawn from the original data set and for each sample a decision tree is constructed. Then the results are combined by averaging for regression or simple voting for classification. For synthesis we do not combine the predictions from single trees. Instead, we sample synthetic values from donors from all trees. For bootstrapping we use sampling with replacement and sample sizes are the same as the size of the original data.

As bagging, random forest uses multiple trees constructed on bootstrapped samples but an algorithm for building these trees is different. At each potential split a new random sample of predictors is chosen as split candidates from the full set of p predictors. For classification number of variables randomly sampled is equal to the square root of the total number of predictors ($\sqrt{p}$) and for regression to one third of them ($p/3$).

4 Assessment of synthetic data utility

We use both analysis-specific and general utility measures to assess and compare quality of synthetic data produced by different modelling approaches. Researchers usually fit regression models to data and therefore analysis-specific measures focus on comparing regression estimates from the original and synthetic data. We apply a measure based on overlap of confidence intervals as suggested by Karr et al. (2006) and a measure based on standardized differences between point estimates. General measures evaluate similarities between overall structures of the confidential data and their synthetic version. They rely on discrimination methods that are used to check the possibility to distinguish the two datasets. We use an indicator based on propensity scores that was evaluated as the most promising one by Woo et al. (2009). A similar measure was also suggested by Snoke et al. (in progress). We model propensity scores using CART model.

5 Empirical evaluation of tree-based methods

5.1 Data set and its synthetic versions

For empirical evaluation, we use a subset of individual-level data from the Scottish Health Survey (SHeS) 2008 and 2010, which is a repeated cross-sectional study. The

subset includes 4661 individuals aged 16 years and older who were part of version A of the core sample and 46 variables of different types. Some of the variables are derived ones and are not present in the original study (see Appendix A for a complete list of variables).

To generate synthetic data we use function *syn()* from the **R** package *synthpop*. All parameters are set to default if not stated otherwise. We replace all values for all variables in our data set. Missing data are not imputed prior to synthesis and their pattern is synthesised instead. The synthesising order, which is the same for all syntheses, was chosen randomly except the first three variables: *sex*, *age* and *age2* (see Appendix A for a complete sequence). We create 10 synthetic data sets for each synthesising method. We use all tree-based methods described in Section 3 ($CART_1$, $CART_2$, BAG - bagging, RF - random forest) and also default parametric models ($PARA$) and simple random sampling with replacement ($SAMP$) for comparison purposes. $CART_1$ uses *ctree()* function from *party* package and $CART_2$ uses function *rpart()* from *rpart* package. The first splitting criterion described in Section 3.1 is implemented in the former function and the second (Gini index and deviance) in the latter. For BAG and RF *randomForest()* function from *randomForest* package is used and 100 trees are constructed. When constructing a single classification tree the Gini impurity is used as the splitting criterion.

5.2 Inference models

In order to evaluate analysis-specific utility of synthetic data generated using various approaches we fit three regression models. They differ in terms of model specification (linear, logistic and Poisson regression) and the scope of explanatory variables, which provides a thorough test for analytical validity of synthetic data. Details of general model specification are presented below but for description of all 27 variables used in the analysis see Appendix A.

1. Model 1: Linear regression of mental well-being
   Ywemwbs $\sim$ *Sex + age + age2 + sptinPAg + hrstot10 + AllNature + Gym + Other + Longill08 + simd15_09 + maritalg*

2. Model 2: Logistic regression of overall life satisfaction
   Ylsg $\sim$ *Sex + ag16g10 + ParentInf + hrsptg10 + eqvinc + maritalg + XOwnRnt08 + econac08 + drkcat3 + cigst3 + porftvg3 + AllNature*

3. Model 3: Poisson regression of average time of sport activities per week
   hrsspt10 $\sim$ *Sex + ag16g10 + hrstot10 + URINDSC + XCare + Longill08 + eqvinc + maritalg + Active + ParentInf + econac08*

5.3 General and analysis-specific utility

The aggregate measures of general and analysis-specific utility of synthetic data produced using different approaches are presented in Table 1. Analysis-specific

measures are calculated for each model separately and they include mean standardized absolute difference in coefficient estimates and mean overlap in the 95% confidence intervals obtained using the observed and synthetic data. Boxplots for those measures are displayed in Figure 1 and Figure 2. Coefficient estimates and 95% confidence intervals for linear, logistic and Poisson model for the original and selected synthetic data ($PARA$, $CART_2$, $CART_1$ and BAG) can be consulted in Appendix B in Figure 3, Figure 4 and Figure 5 respectively.

	$CART_1$	$CART_2$	BAG	RF	$PARA$	$SAMP$
General utility						
Propensity score measure	0.14	**0.11**	0.12	0.13	0.15	0.21
Analysis-specific utility						
Model 1: linear						
Mean $\hat{\beta}$ stand. difference	8.40	2.76	4.93	6.51	**1.77**	11.63
Mean 95% CI overlap	0.42	0.77	0.60	0.52	**0.86**	0.31
Model 2: logistic						
Mean $\hat{\beta}$ stand. difference	6.95	3.72	3.73	5.30	**1.72**	10.98
Mean 95% CI overlap	0.50	0.71	0.70	0.60	**0.86**	0.33
Model 3: Poisson						
Mean $\hat{\beta}$ stand. difference	7.32	4.29	4.04	4.01	**3.89**	9.48
Mean 95% CI overlap	0.47	0.66	0.62	0.64	**0.69**	0.39

NOTE: BAG - bagging, RF - random forest, $PARA$ - parametric, $SAMP$ - random sampling with replacement; bold indicates best model

Table 1: General and analysis-specific utility measures for various synthesising approaches and inference models.

According to propensity score measure, which evaluates general utility and provides relative values for synthetic data producer, $CART_2$ generates synthetic data of best quality and as expected randomly sampled synthetic data are the least useful. Parametric approach is outperformed by all tree-based methods in reproducing overall structure of the original data. Nonetheless, it performs best at the analysis-specific level for all estimated models. The parametric approach is followed by $CART_2$, which performs best at the general level. $CART_1$ is much less effective in preserving relationships in the data than $CART_2$ which indicates the importance of splitting criteria for growing trees. Random forest and bagging show improvement over $CART_1$ but not $CART_2$. Note that relatively good performance of parametric approach in terms of analysis-specific utility measures can be related to the types of models that were fitted. Linear and logistic model specification were also used for modelling the outcome variables and the synthesis stage. For Poisson model, which was not used for synthesis, synthetic data generated using parametric models give

results comparable with those produced using non-parametric techniques ($CART_2$, BAG and RF).

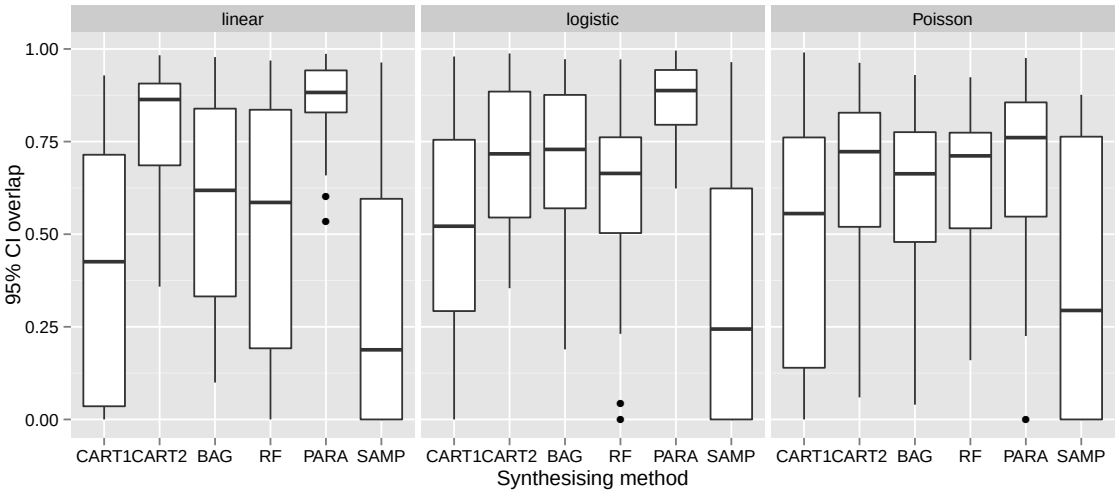

Figure 1: Relative overlap in the 95% confidence interval for $\hat{\beta}$ coefficients from a linear, logistic and Poisson regression fitted to original and synthetic data.

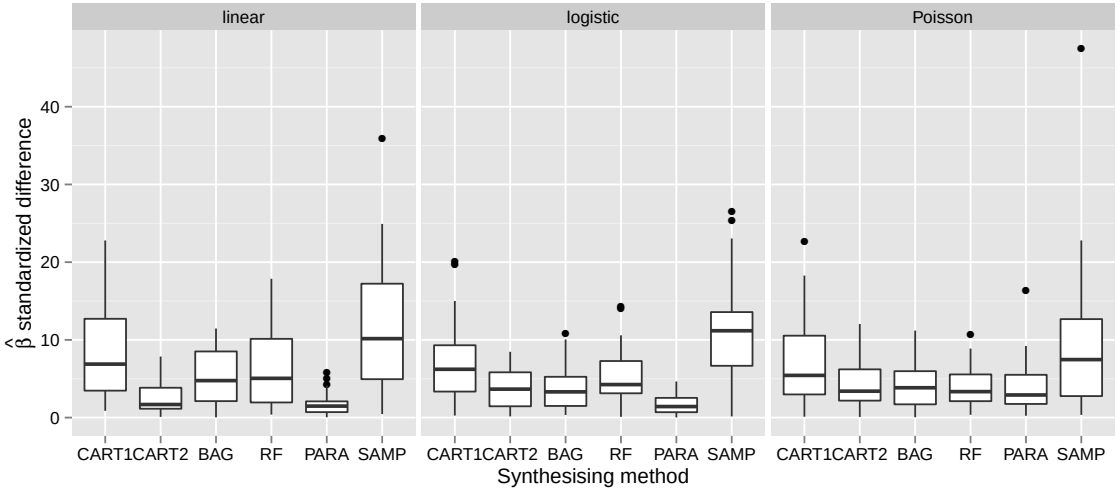

Figure 2: Standardised absolute difference in $\hat{\beta}$ coefficients from a linear, logistic and Poisson regression fitted to original and synthetic data.

6 Concluding remarks

Synthetic data techniques have been developed to allow for the release of high-quality microdata without compromising confidentiality and as illustrated by the empirical example in this paper it is possible to produce useful completely synthetic data using automated methods. However, more research need to be done to validate these and other machine learning methods, especially in more complex settings with regard to both data structure and inference models.

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Caiola, G. and Reiter, J.P. (2010) Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, **3**, 27–42.

Drechsler, J. and Reiter J.P. (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, **55**, 3232–3243.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. and Sanil, A. P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, **60**, 224–232.

Nowok, B., Raab, G.M. and Dibben, C. (2015) *synthpop: Bespoke creation of synthetic data in **R***. Package vignette, `http://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf`.

R Core Team (2015) ***R**: A language and environment for statistical computing*, **R** Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org`.

Reiter, J.P. (2005) Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, **21**, 441–462.

Rubin, D.B. (1993) Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, **9(2)**, 461–468.

Snoke, J., Raab, G.M., Nowok, B., Dibben, C. and Slavkovic, A.B. (in progress) Data utility for synthetic data: Improved general and specific measures.

Woo, M., Reiter, J.P., Oganian, A. and Karr, A.F. (2009) Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, **1(1)**, 111–124.

A Appendix - A list of variables and visit sequence

Visit sequence: Sex, age, age2, Parent, Longill08, maritalg, econac08, ForestInNatureg, Forest, porftvg3, XOwnRnt08, URINDSC, AllNature, nssec8, Other, SYear, YLifeSat, Gym, ag16g10, SIMD5_SG, drkcat3, vera0810wt, Active, simd15_09, hrsspt10, XCare, drating, hrstot10, Nature, HHSize, eqvinc, ParentInf, sptinPAg, Ylsg, Yghq12scr, alclim, cigdyal, hrtotg10, eqv5, cigst3, YGHQg2, Ywemwbs, sptinPA, hedqul08, hrsptg10, YGenHelf

B Appendix - Coefficient estimates and 95% confidence intervals for models fitted to observed and synthetic data

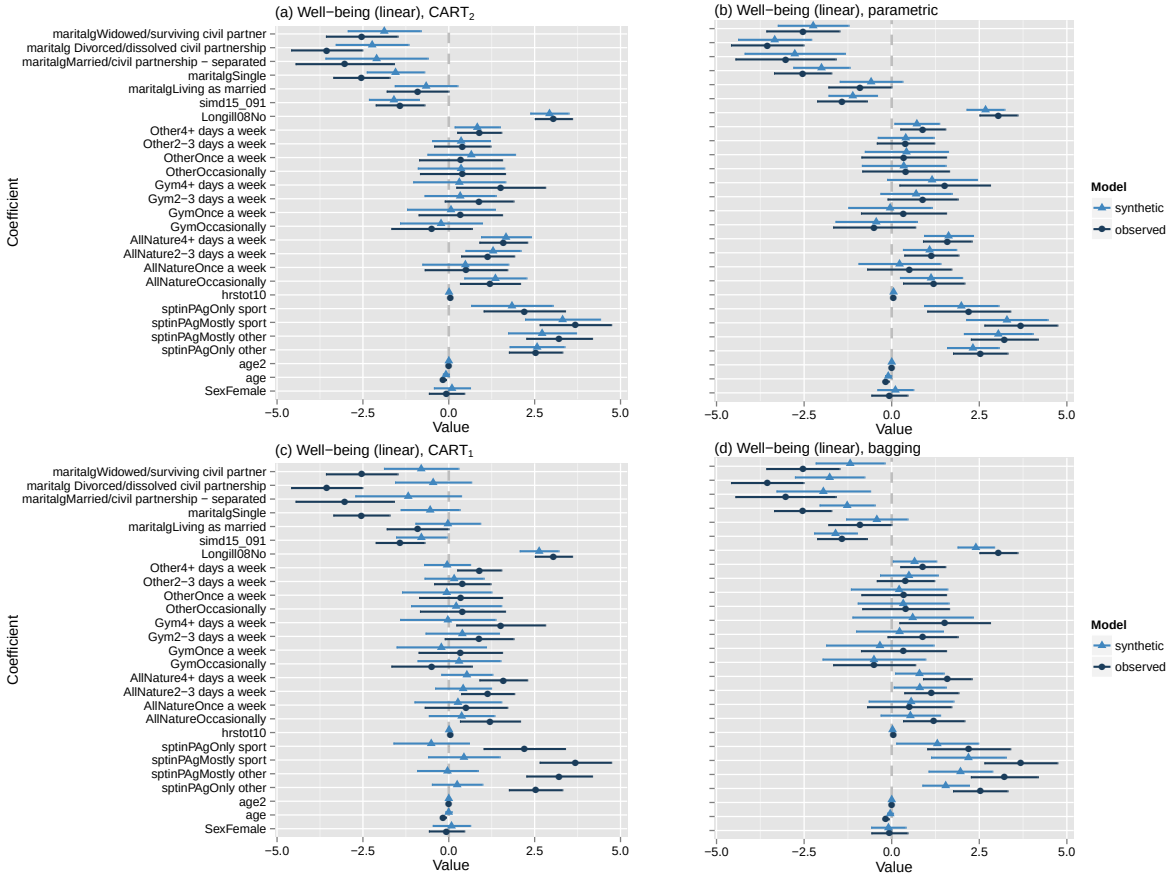

Figure 3: Estimates and 95% confidence intervals for $\hat{\beta}$ from a linear regression of well-being for observed and synthetic data (Model 1).

No.	Variable name	Description
1	Sex	Sex
2	age	Age
3	*age2*	Age squared
4	ag16g10	Age in ten year bands
5	maritalg	Marital status
6	HHSize	Household Size
7	*Parent*	Whether parent/guardian of any child 0-15 in household
8	*ParentInf*	Whether parent/guardian of any infant (0-2) in household
9	Longill08	Whether has long-standing illness
10	hedqul08	Highest educational qualification
11	*Ywemwbs*	WEMWBS score
12	*YLifeSat*	Satisfaction with life as a whole
13	*Ylsg*	Satisfaction with life as a whole - grouped (0-6,7-10)
14	*Yghq12scr*	GHQ Score - 12 point scale
15	*YGHQg2*	GHQ Score - grouped (0,1-3,4+)
16	*YGenHelf*	Self-assessed general health
17	econac08	Economic status of respondent
18	eqvinc	Equivalised Income
19	eqv5	Equivalised income quintiles
20	nssec8	NS-SEC 8 category classification (individual)
21	simd15_09	SIMD flag lowest 15%
22	SIMD5_SG	SIMD 2009 quintiles
23	*XOwnRnt08*	Accommodation occupancy status
24	URINDSC	Urban/Rural Indicator (Scotland)
25	cigst3	Cigarette smoking status
26	cigdyal	Number of cigarettes smoke a day
27	drkcat3	Weekly drinking category (non/moderate/hazardous or harmful)
28	drating	Total units of alcohol/week
29	alclim	Whether exceeding government recommendations on alcohol consumption
30	porftvg3	Grouped portions of fruit (inc.fruit juice) & veg (5/less than 5/none)
31	*XCare*	Hours spent per week providing help or unpaid care (grouped)
32	Active	Level of physical activity at work
33	hrsspt10	Average hours doing sport per week
34	hrsptg10	Average hours doing sports per week (grouped)
35	hrstot10	Average hours doing all physical activities per week
36	hrtotg10	Average hours doing all physical activities per week (grouped)
37	*sptinPA*	Share of sport in time spent on physical activity
38	*sptinPAg*	Share of sport in time spent on physical activity (grouped)
39	*Forest*	Frequency of physical activities using woodland etc.
40	*Nature*	Frequency of physical activities using open space/park., country paths, beach, river etc.
41	*AllNature*	Frequency of physical activities using nature (*Forest* and *Nature* combined)
42	*ForestInNatureg*	Share of *Forest* in *AllNature* (grouped)
43	*Gym*	Frequency of physical activities using sports field, swimming pool, gym/sport centre, etc.
44	*Other*	Frequency of physical activities using pavement/streets, home/garden and other places
45	SYear	Survey year
46	vera0810wt	Version A weight

NOTE: Italic indicates recoded or new derived variables

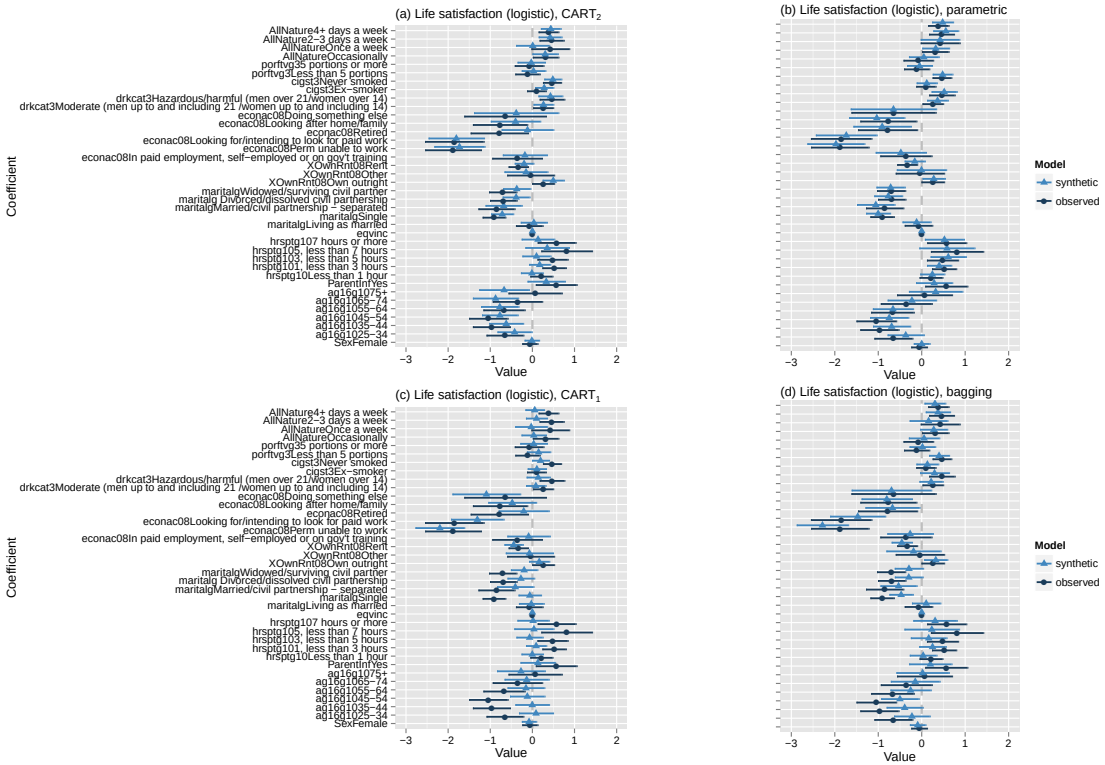

Figure 4: Estimates and 95% confidence intervals for $\hat{\beta}$ from a logistic regression of overall life satisfaction for observed and synthetic data (Model 2).

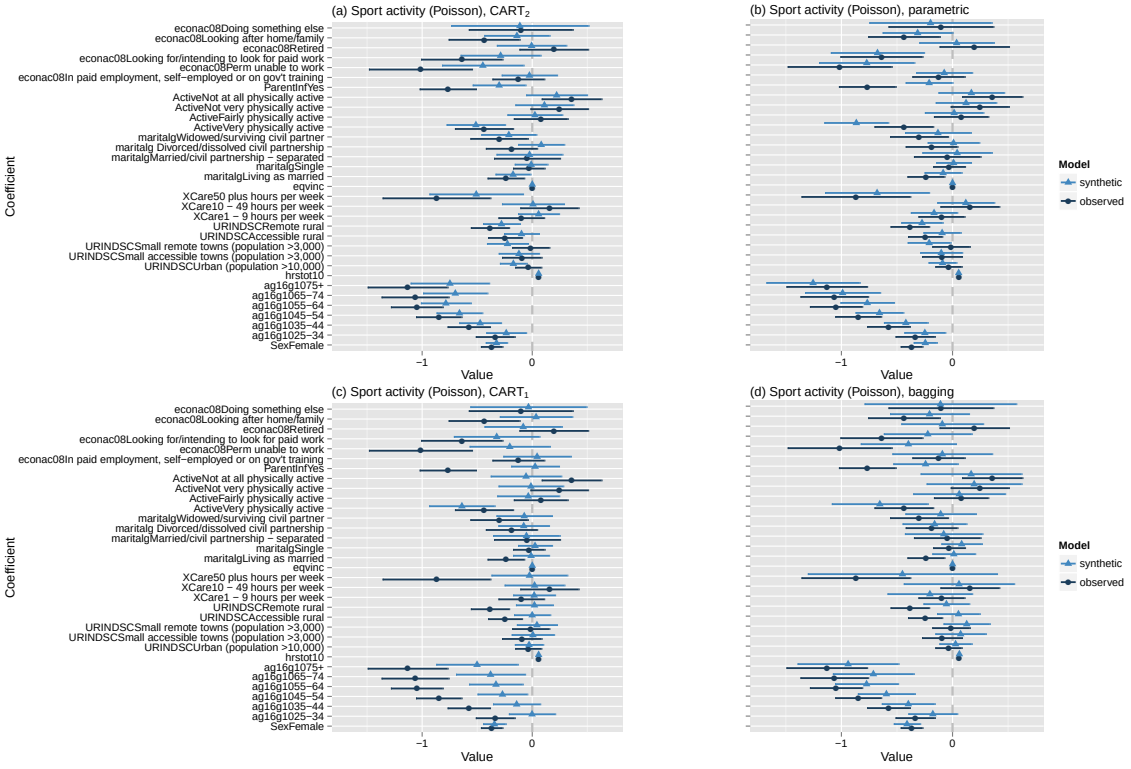

Figure 5: Estimates and 95% confidence intervals for $\hat{\beta}$ from a Poisson regression of average time of sport activities for observed and synthetic data (Model 3).