

Confidentialising microdata using multiple imputation

Patrick Graham

Bayesian Research and University
of Otago, Christchurch

Acknowledgements

- Project team
SNZ: Olena Rodnyanskiy, Richard Penny,
Mike Camden
Ex-SNZ: Lisa Henley
- Funded by OS Research

What is the issue?

- Official statistics agencies seek to release data to external analysts (users) but need to protect respondent confidentiality.
- We assume analysts wish to conduct analysis involving some statistical sophistication.
- Want released data to support a range of analyses. Unit-record data (microdata) is likely to be the most flexible option.

Methods for creating confidentialised microdata

- Masking methods (traditional)
 - e.g. top-coding, rounding, noise, data-swapping.
- Sufficiency based perturbation (Muralidhar & Sarathy, 2006)
 - modify data values subject to preserving sufficient statistics.
- Multiply imputed synthetic data (Rubin, 1993, Reiter, 2002)

A Potential Issue with Masking methods

- In theory, masking is not always ignorable - should be accounted for in analysis.
- E.G. Adding noise and data swapping produce data with measurement error and mis-classification. There is an extensive literature on methods for dealing with such data, but methods not routinely included in standard software.
- Top-coding and bottom-coding are like censoring so could should use censored and/or coarsened data analysis methods.
- Currently , entirely, up to the user whether to account for masking in the data.

What can happen if masking ignored in analysis?

Simulation study: Coverage probabilities (%) for
95% confidence intervals–logistic interaction
model (selected parameters)

		Randomly rounded	
	Real data	Naïve	E-M
No quals	94.8	93.8	94.9
Post sec	95.3	94.2	95.5
Maori (Ma)	95.9	93.4	96.4
Pacific (PI)	94.7	84.6	95.6
PI x sex	95.1	87.7	95.0
Ma x no qual	95.6	92.2	95.8
Ma x psec	95.5	90.3	95.5
PI x no qual	94.9	87.8	95.8

MI synthetic data - background

- Comes out of Bayesian theory for finite population inference.
- Theory guides both creation of synthetic data by agencies and analysis of synthetic data by users.
- In practice:
 - i. OS Agency generates predictions (imputations) for a new sample.
 - ii. Multiple imputations to account for the fact the new sample is generated, not observed.
 - iii. User runs standard analyses on each imputation and combines results.
- “Users have their own science to worry about” (Rubin, 1993)

Generating imputed datasets

Need an imputation model – a model for the data.

Standard to consider a parametric model, F_θ , say.

Then given observed data Y^{obs} , the imputation process is

- Fit Model to Data
- Repeat for $m = 1 \cdots M$:

Draw θ_m from $p(\theta | Y^{\text{obs}})$

Generate Y_m^{new} from F_{θ_m}

i.e draw Y_m^{new} from the posterior predictive distribution of Y , under the model F_θ .

Imputation Model

- Choice of the imputation model is a crucial step in creating synthetic data.
- Conventional models (e.g. linear or logistic regression, other glms) constrain predicted values to follow the specified functional form.
- Hierarchical models allow for some model uncertainty and lead to a compromise between data and conventional models.

Comparison of conventional and hierarchical Poisson log-linear model (LLM)

GLM	HB GLM
$Y_i \lambda_i \sim \text{Poisson}(\lambda_i)$	$Y_i \lambda_i \sim \text{Poisson}(\lambda_i)$
$\ln(\lambda_i) = X_i \beta$	$\lambda_i \sim \text{Gamma}(\xi, \xi / \mu_i)$
	$\ln(\mu_i) = X_i \beta$
	$\beta, \xi \sim \pi$
$\hat{\lambda}_i = \exp(X_i \hat{\beta})$	$\hat{\lambda}_i = w_i \hat{\mu}_i + (1 - w_i) Y_i$ $w_i = \hat{\xi} / (\hat{\xi} + \hat{\mu}_i)$

Value of hierarchical imputation models.

For categorical data simulation studies demonstrate superior frequentist performance for synthetic data created under HB Poisson LLM than under non-hierarchical Poisson LLM

-Less bias, better confidence interval coverage, more robustness to prior model specification

Graham, Young & Penny (2008),

<http://www.statisphere.govt.nz/official-statistics-research/series/vol-3.htm>

Graham Young & Penny (2009, JOS, In press)

Building imputation models for data comprising a mix of categorical and numerical variables

Build on HB LLM for categorical data by modelling numerical variables conditionally on categorical variables.

$$D = (Y^{\text{num}}, X^{\text{cat}})$$

$$p(D | \theta) = p(X^{\text{cat}} | \theta^{\text{cat}}) p(Y^{\text{num}} | X^{\text{cat}}, \theta^{\text{num}})$$

e.g. could consider a HB multivariate normal regression model for $p(Y^{\text{num}} | X^{\text{cat}}, \theta^{\text{num}})$

Hierarchical multivariate normal model

$$[Y_{ij}^{num} \mid X_i^{cat}, \mu_i, V_i] \sim MVN(\mu_i, V_i), j = 1 \cdots n_i, i = 1 \cdots K$$

$$\mu_i \sim MVN(X_i^{cat} \beta, \Sigma)$$

$$p(\beta) \propto 1$$

$$\Sigma \sim MVUS(\tilde{V}), \text{ where } \tilde{V} \text{ is some summary of the } V_i$$

Note

$$E(\mu_i \mid \beta, \Sigma) = W_i(X_i^{cat} \beta) + (I - W_i)\bar{Y}_i^{num}$$

$$\text{where } W_i = \bar{V}_i(\bar{V}_i + \Sigma)^{-1} \text{ and } \bar{V}_i = V_i / n_i$$

Building imputation models – fully parametric

$[X^{\text{cat}} | \theta^{\text{cat}}]$ - HB Poisson LLM for cell counts

$[Y^{\text{num}} | X^{\text{cat}}, \theta^{\text{num}}]$ – HB MVN regression model

OK for cell counts because cell counts are sufficient statistics for all categorical data analyses.

But means are not sufficient for all analyses of numerical data, hence good estimation of means does not ensure good performance for all possible analyses

Building imputation models cont'd

- HB MVN provides estimates of means which are a compromise between data and model but imposes an assumption of normality on the data, which is too restrictive.
- Assumes linear model for relationships between numerical variables.
- Need a more flexible imputation model for numerical variables.

More General Imputation Model: Non-parametric Bayes

Data Y drawn from some distribution, F

Instead of introducing parametric assumptions, F_θ and estimating θ , treat F as an unknown and assign a prior over the space of distribution functions

Then, in principle, synthetic data is generated by:

For $m = 1 \cdots M$

draw F_m from $F | Y^{obs}$

draw Y_m^{new} from F_m

Combining Non-parametric and Hierarchical Bayes

In fact, the prior for F is usually centered on some Standard parametric distribution, e.g. normal or MVN.

The parameters of this centering distribution are typically unknown and hence must be assigned priors which are updated by the data.

We can use a hierarchically structured prior for the parameters of the centering distribution, thereby embedding the HB set-up within a broader class of distributions.

Full imputation

$[X^{cat} | \theta^{cat}]$ – HB Poisson log-linear model

for (i in 1 to K) {

$$Y_{ij}^{num} | x_i \sim F_i, j = 1 \cdots n_i$$

$$F_i \sim DP(\alpha, G(\theta_i))$$

$$G(\theta_i) \leftrightarrow MVN(\mu_i, V_i)$$

$$\mu_i | x_i, \beta, \Sigma \sim MVN(x_i \beta, \Sigma)$$

}

$$p(\beta, \Sigma)$$

$DP(\alpha, G(\theta))$ denotes a Dirichlet process prior, centred on G , with precision parameter, α .

As $\alpha \rightarrow \infty$, the model for Y^{num} reduces to the HB model.

As $\alpha \rightarrow 0$, the model for Y^{num} reduces to the Bayesian bootstrap.

Dirichlet Process Prior

$$F \sim DP(\alpha, G_\theta)$$

Indicates that for any partition of the sample space, $B = \{B_1, \dots, B_k\}$, the probabilities assigned by F to the elements of B have a Dirichlet distribution with parameters

$$\{\alpha G_\theta(B_1), \alpha G_\theta(B_2), \dots, \alpha G_\theta(B_k)\}$$

and G_θ is, usually, a standard parametric distribution.

DP implicitly defines a prior for F , by defining a prior for the set of probabilities assigned by F to any partition of the sample space.

Imputation under the full model

Conditional on all parameters, the DP prior implies the posterior predictive distribution of Y^{num} can be simulated using a generalised Polya urn.

Consequently an approximation to the posterior predictive distribution under the full model can be simulated by:

1. Draw all parameters from their posterior distribution,
 2. Draw synthetic cell counts, c_i from their predictive distribution
2. For each cell, i , generate c_i new data values from a generalised Polya urn, centered on the model implied by the parameter values simulated in (1).

Generalised Polya Urn

Suppose: observed cell size n_i ; m^{th} synthetic cell size c_i .

Initialise data urn to the n_i observed Y_i^{num} values

For j in $1 \cdots c_i$

$b_j = n_i + j - 1$ (current size of data urn)

$$p_j = \frac{\alpha}{\alpha + b_j}$$

With probability p_j , sample from model (e.g. HB MVN);

With probability $(1 - p_j)$, sample from data urn.

Add sampled values to the data urn.

Add sampled values to the imputed dataset.

Fitting the non-parametric hierarchical model for the numerical data.

Gibbs Sampler

1. $p(\boldsymbol{\mu} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}, \mathbf{V}, \boldsymbol{\alpha}, \mathbf{Y}^{num}) \propto p(\mathbf{Y}^{num} \mid \boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}) p(\boldsymbol{\mu} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X})$
2. $p(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}, \mathbf{V}, \boldsymbol{\alpha}, \mathbf{Y}^{num}) \propto p(\boldsymbol{\mu} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}) p(\boldsymbol{\beta})$
3. $p(\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \mathbf{X}, \mathbf{V}, \boldsymbol{\alpha}, \mathbf{Y}^{num}) \propto p(\boldsymbol{\mu} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}) p(\boldsymbol{\Sigma})$

Step 1 Makes use of the Polya Urn representation of the data distribution under a DP prior - the likelihood for the cell-specific means can be derived from the Polya sampling model.

This likelihood replaces the multivariate normal likelihood which arises in the fully parametric case

Steps 2 and 3 are just as for the fully parametric case

Example (1)

- Using subset of 2003 Income Survey CURF as the observed data.
- Restricted to ages 25-64, positive income, positive average weekly hours worked.
- Other variables: age, sex, ethnic group, education.
- Treated as SRS.
- Results below are for income.

Performance metrics –synthetic versus real data

- Absolute differences (means, medians, quartiles, model parameter estimates).
- Relative credible interval length.
- Data utility index (Karr, 2006, Am Stat)
(measure of overlap of 95% credible intervals, defined so that 0.95 is perfect).

Estimation of mean weekly income and income quartiles.

36 groups – not all mutually exclusive,
various sample sizes

Full sample,

Males, females

Ethnic groups,

Educational achievement groups

Sex by ethnicity

Sex by educational achievement

NPHB = non-parametric hierarchical Bayes model

Imputation models specified on original scale

Comparison of MI methods: imputation modelling on original scale

Estimating mean income for 36 groups

parametric HB

NPHB, alpha=19

NPHB, alpha=9

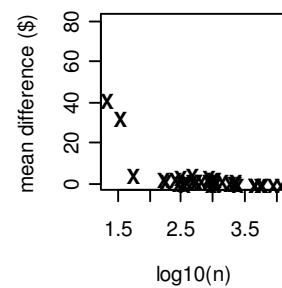
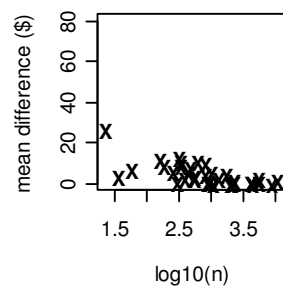
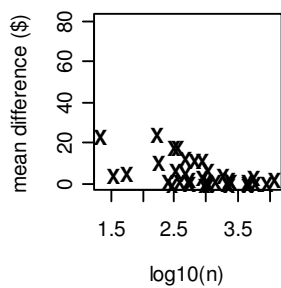
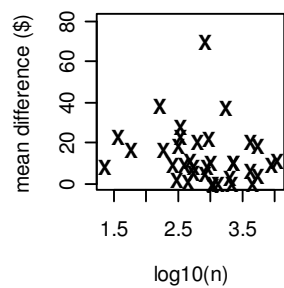
Bayes bootstrap

absolute mean diff

absolute mean diff

absolute mean diff

absolute mean diff

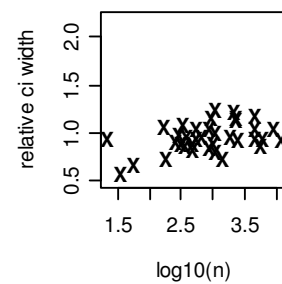
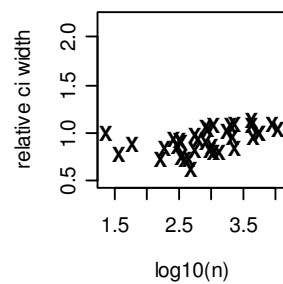
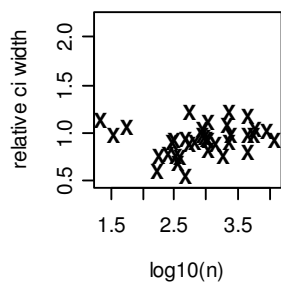
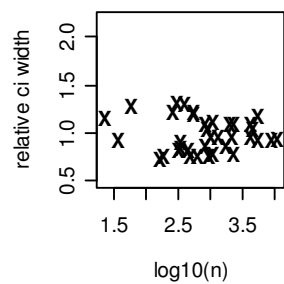


relative ci width

relative ci width

relative ci width

relative ci width

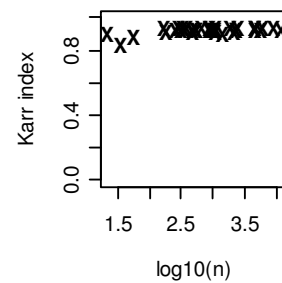
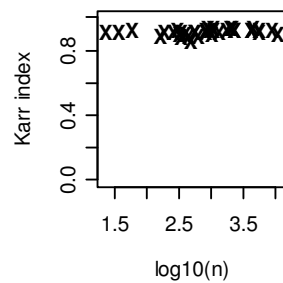
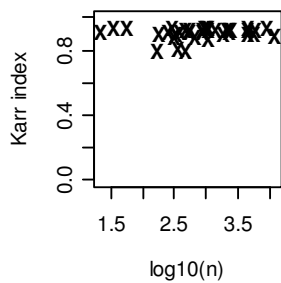
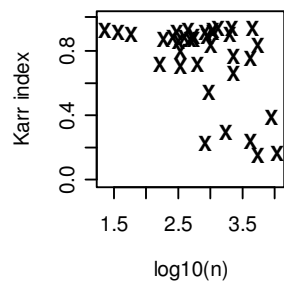


data utility

data utility

data utility

data utility



Comparison of MI and SBP methods: imputation modelling on original scale

Estimating income quartiles for 36 groups

parametric HB

NPHB, alpha=19

NPHB, alpha=9

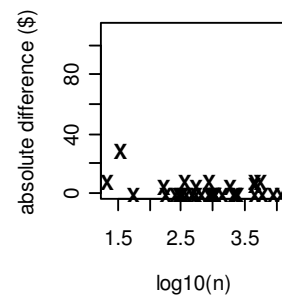
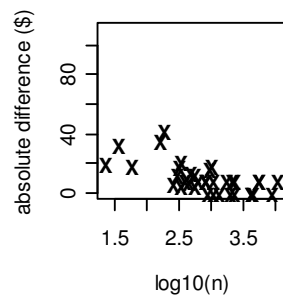
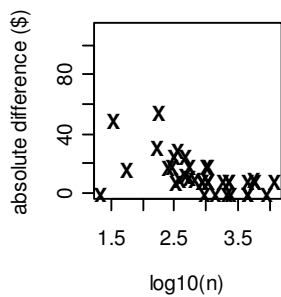
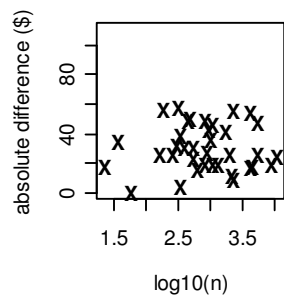
Bayes bootstrap

difference in medians

difference in medians

difference in medians

difference in medians

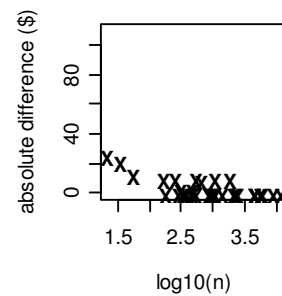
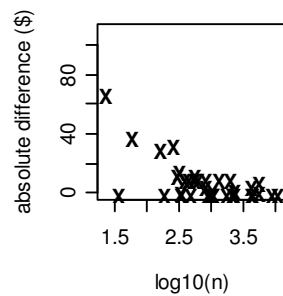
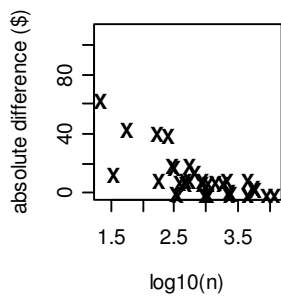
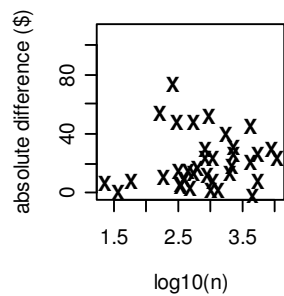


difference in LQ

difference in LQ

difference in LQ

difference in LQ

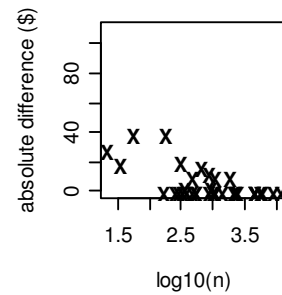
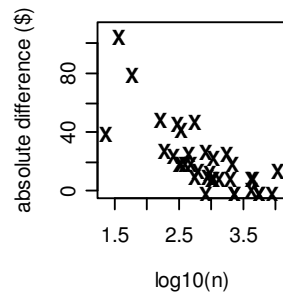
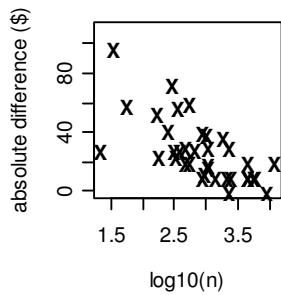
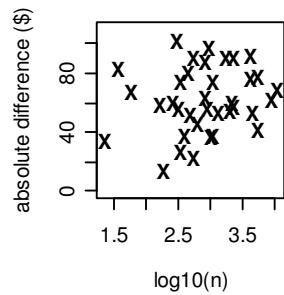


difference in UQ

difference in UQ

difference in UQ

difference in UQ



Imputation models specified on log scale

Comparison of MI methods: imputation modelling on log scale

Estimating mean income for 36 groups

parametric HB

NPHB, alpha=19

NPHB, alpha=9

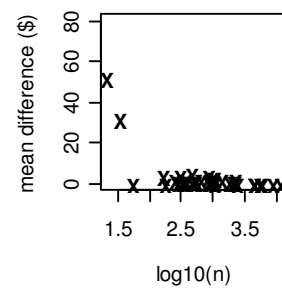
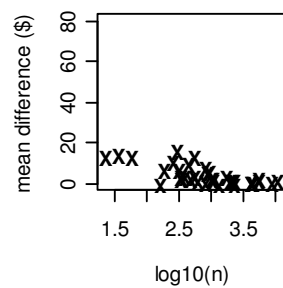
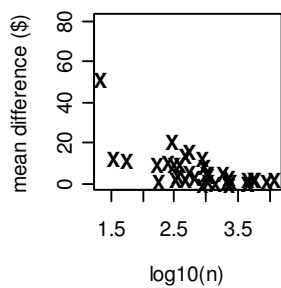
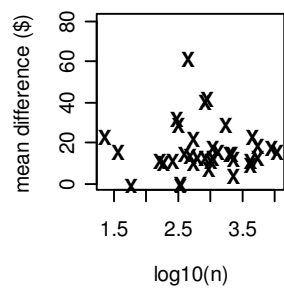
Bayes bootstrap

absolute mean diff

absolute mean diff

absolute mean diff

absolute mean diff

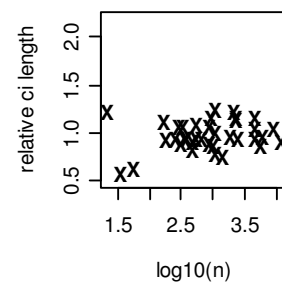
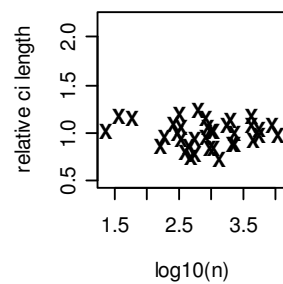
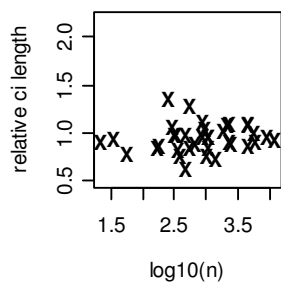
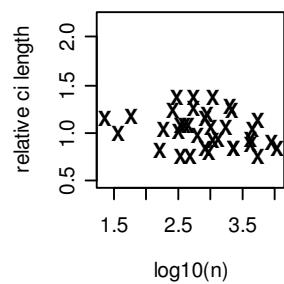


relative ci length

relative ci length

relative ci length

relative ci length

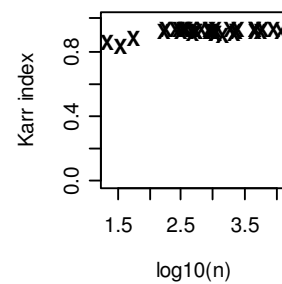
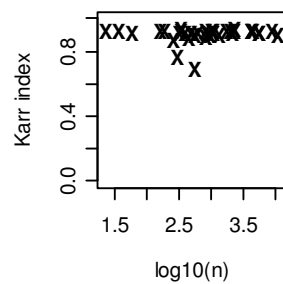
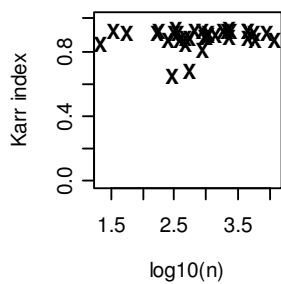
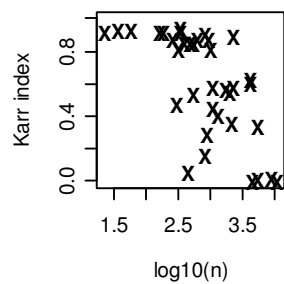


data utility

data utility

data utility

data utility



Comparison of MI methods: imputation modelling on log scale

Estimating income quartiles for 36 groups

parametric HB

NPHB, alpha=19

NPHB, alpha=9

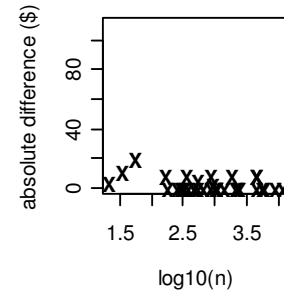
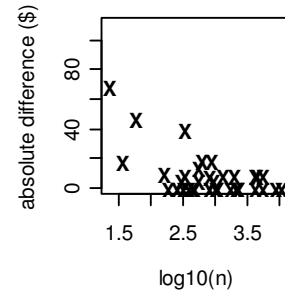
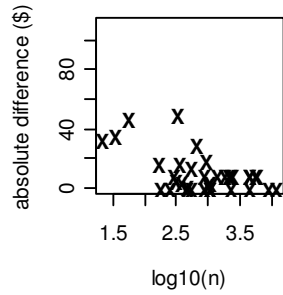
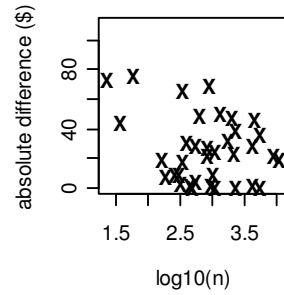
Bayes bootstrap

difference in medians

difference in medians

difference in medians

difference in medians

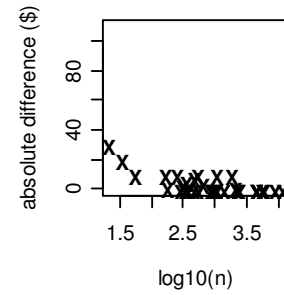
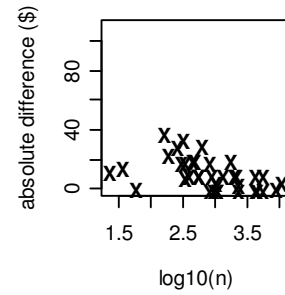
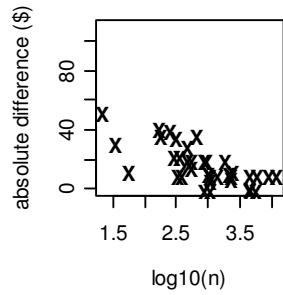
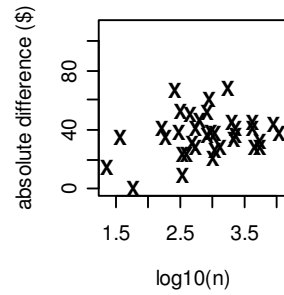


difference in LQ

difference in LQ

difference in LQ

difference in LQ

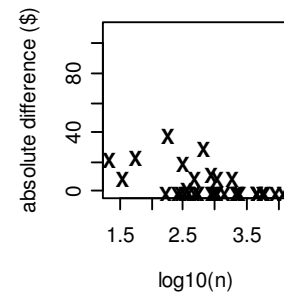
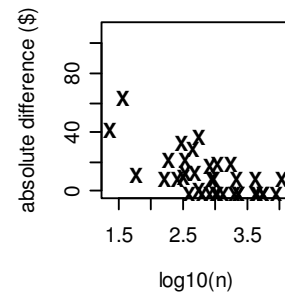
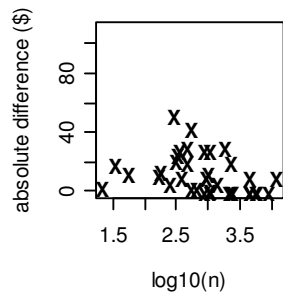
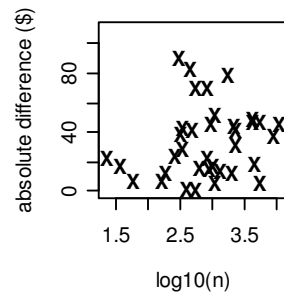


difference in UQ

difference in UQ

difference in UQ

difference in UQ



Estimating log-linear regressions for income under imputation models specified on the log-scale

1. main effects model simpler than the imputation model

Regressors: log-hours worked, age group, sex, ethnicity, educational qualifications

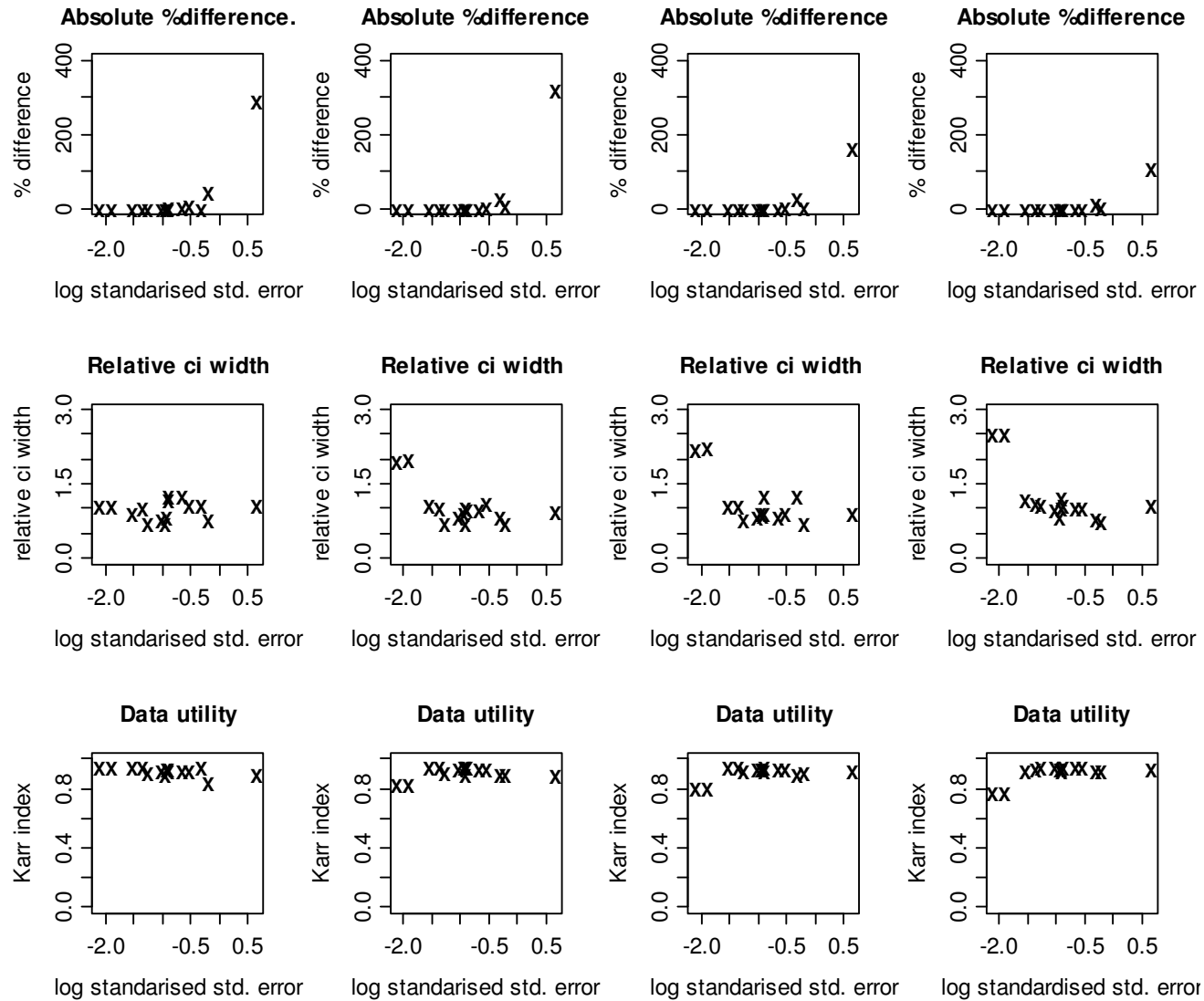
**Comparison of MI methods: Imputation modelling on log scale.
 Estimating income main effects log-linear regression, working hours log transformed**

Parametric HB

NPHB ($\alpha = 19$)

NPHB ($\alpha = 9$)

Bayes bootstrap



2. More complex log-linear regression model, including log working hours and some interaction terms not represented in the imputation model (specified on the log scale).

Comparison of MI methods: Imputation modelling on log scale.

Estimating income log-linear regression with interactions, working hours log transformed.

Parametric HB

NPHB ($\alpha = 19$)

NPHB ($\alpha = 9$)

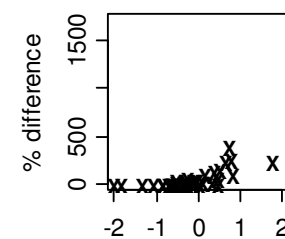
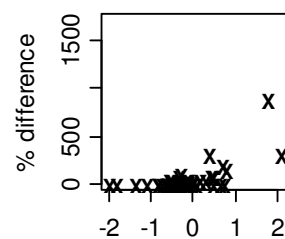
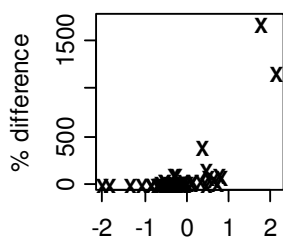
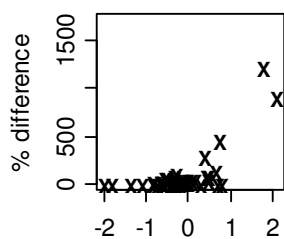
Bayes bootstrap

Absolute %difference.

Absolute %difference

Absolute %difference

Absolute %difference



log standardised std. error

log standardised std. error

log standardised std. error

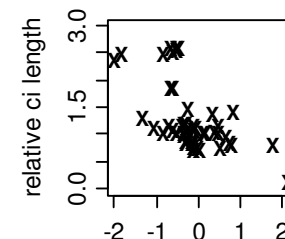
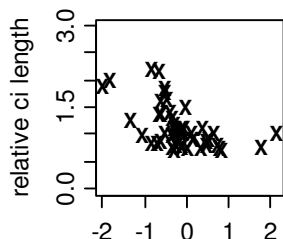
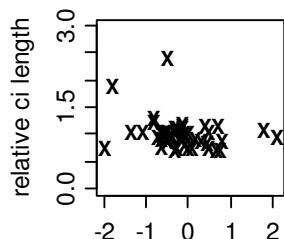
log standardised std. error

Relative ci length

Relative ci length

Relative ci length

Relative ci length



log standardised std. error

log standardised std. error

log standardised std. error

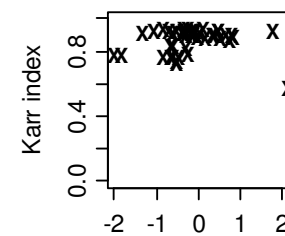
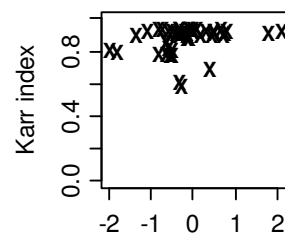
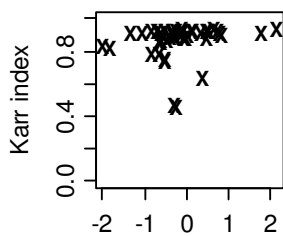
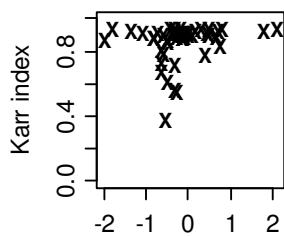
log standardised std. error

Data utility

Data utility

Data utility

Data utility



log standardised std. error

log standardised std. error

log standardised std. error

log standardised std. error

Estimating complex log-linear regression with working hours included on original scale and including some interaction terms not included in the imputation models (specified on log-scale)

Comparison of MI methods: Imputation modelling on log scale.

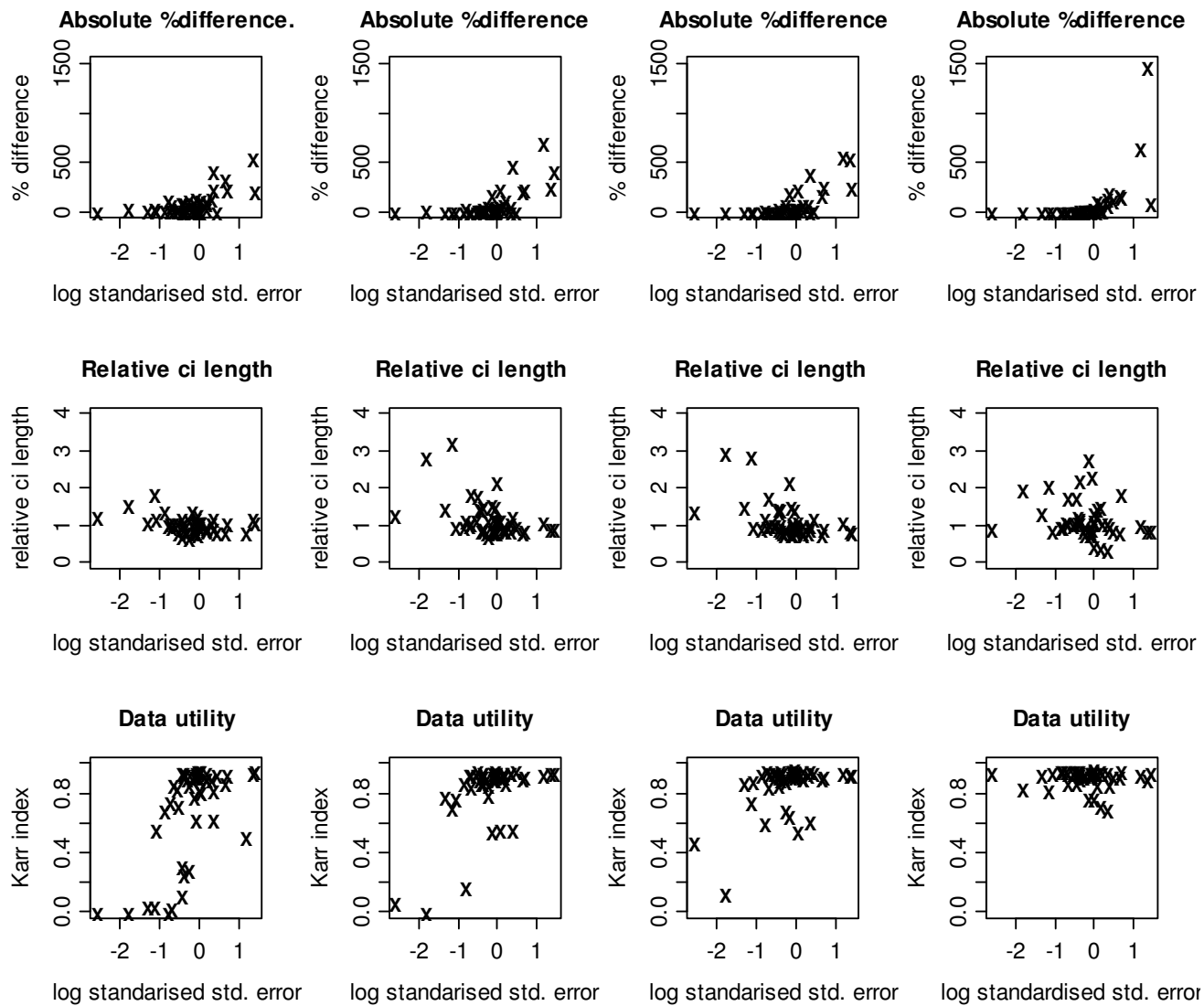
Estimating income log-linear regression with interactions, working hours untransformed.

Parametric HB

NPHB ($\alpha = 19$)

NPHB ($\alpha = 9$)

Bayes bootstrap



There is a problem with low data utilities for three points corresponding to the intercept, working hours and sex parameters.

The problem seems due to using working hours as a regressor, when the imputation model used log-working hours

When the working hours variable is removed

Comparison of MI methods: Imputation modelling on log scale.

Estimating income log-linear regression with interactions, working hours omitted

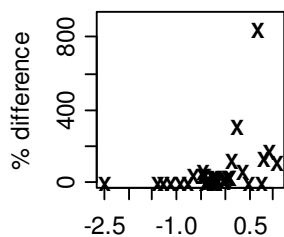
Parametric HB

NPHB ($\alpha = 19$)

NPHB ($\alpha = 9$)

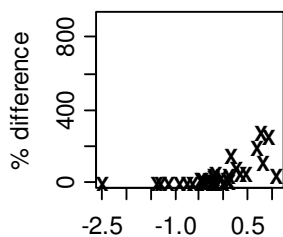
Bayes bootstrap

Absolute %difference.



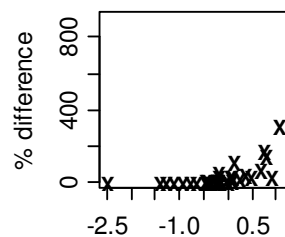
log standardised std. error

Absolute %difference



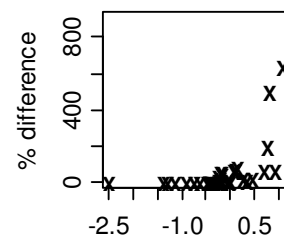
log standardised std. error

Absolute %difference



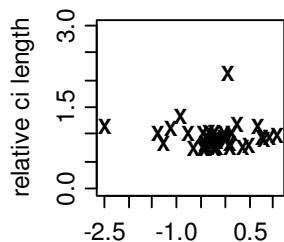
log standardised std. error

Absolute %difference



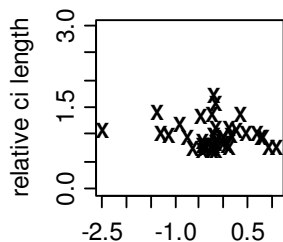
log standardised std. error

Relative ci length



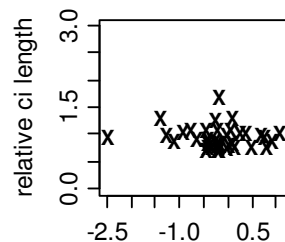
log standardised std. error

Relative ci length



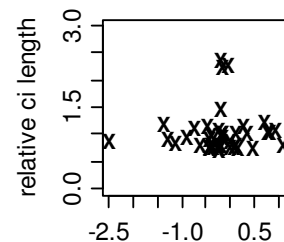
log standardised std. error

Relative ci length



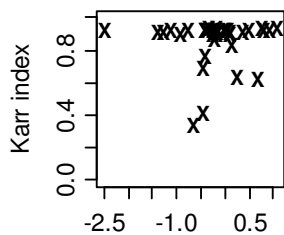
log standardised std. error

Relative ci length



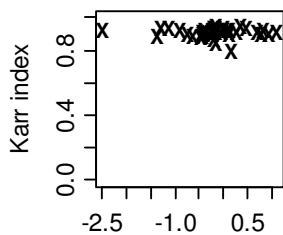
log standardised std. error

Data utility



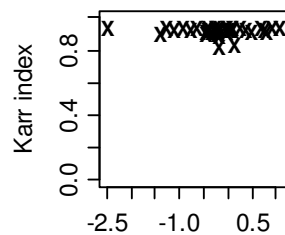
log standardised std. error

Data utility



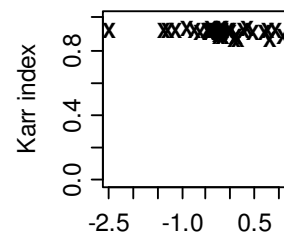
log standardised std. error

Data utility



log standardised std. error

Data utility



log standardised std. error

Disappointing that NPHB did not do a better job of protecting against mis-modelling of working hours (would need lower alpha, or different centering model).

However, NPHB improves on fully parametric HB in protecting against other types of imputation model – analysis model discrepancy (e.g. additional interaction terms)

3. Estimation log linear regressions for income under imputation models specified on the original scale.

Worst case here is complex analysis model including log transformed working hours and interactions omitted from the imputation model.

Comparison of MI methods: imputation modelling on original scale
Estimating income log-linear regression with interactions, work hours log transformed

parametric HB

NPHB, alpha=19

NPHB, alpha=9

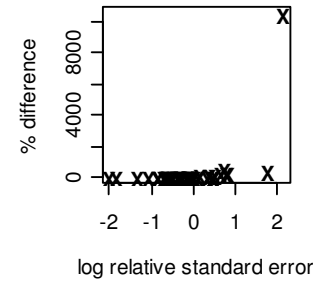
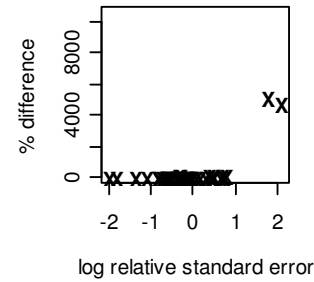
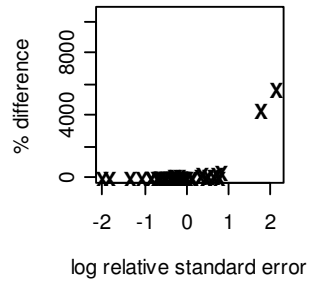
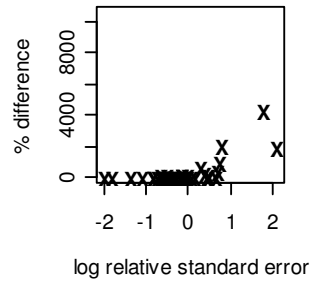
Bayes bootstrap

Absolute pct. difference

Absolute pct. difference

Absolute pct. difference

Absolute pct. difference

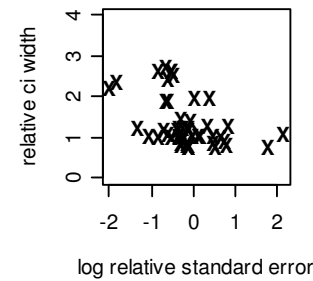
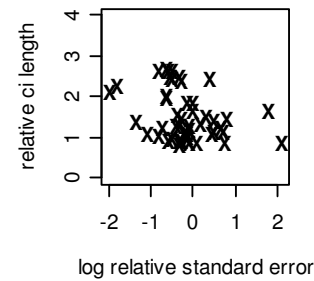
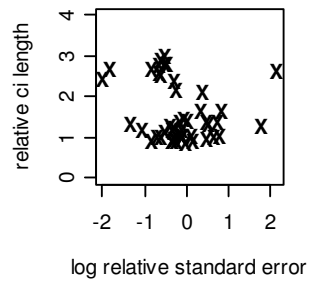
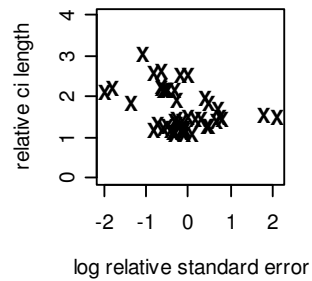


Relative ci length

relative ci length

relative ci length

relative ci width

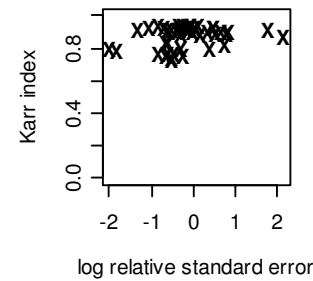
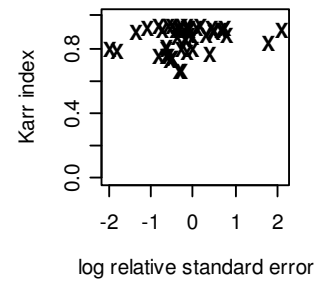
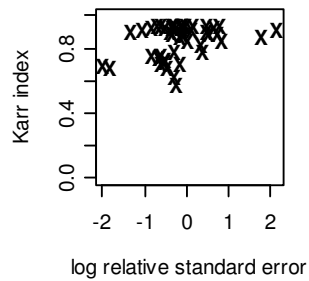
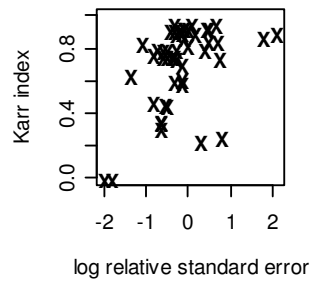


Data utility

Data utility

Data utility

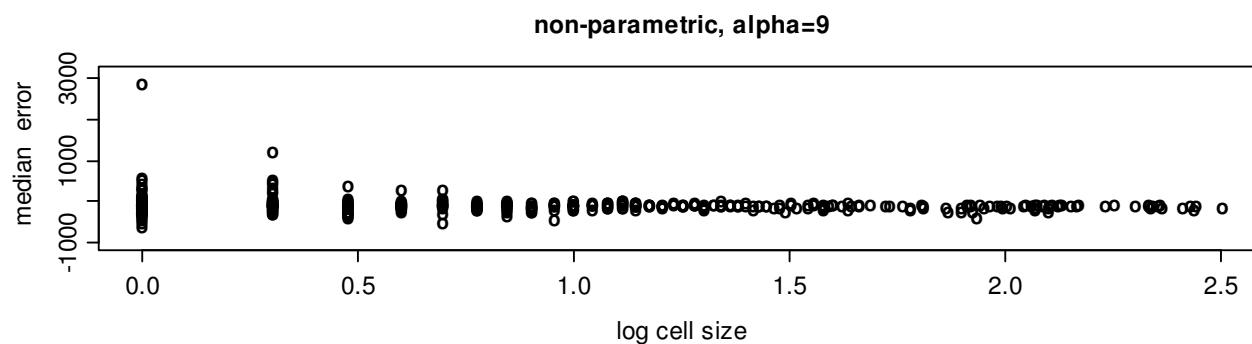
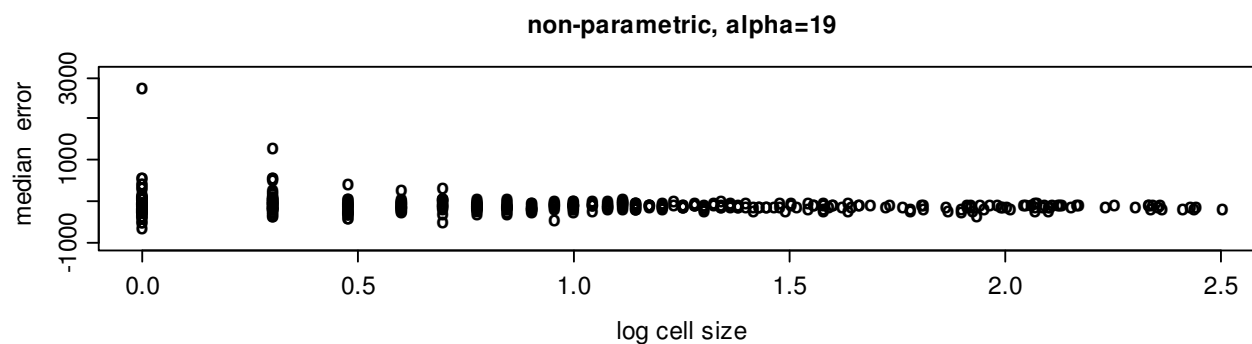
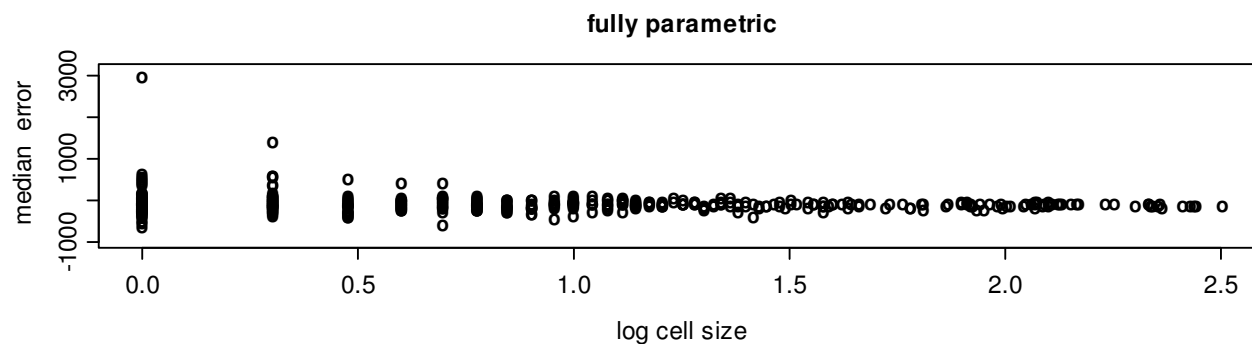
Data utility



Disclosure Risks

Comparison of median prediction errors for 405 cells, by cell size

Imputation models fitted on original scale



Distribution of median prediction errors among 54 uniques

Imputation model	min	25%	50%	75%	max
Parametric	-582.10	-172.20	-43.26	132.23	3034.00
NPHB ($\alpha=19$)	-569.00	-141.70	-4.91	131.80	2385.00
NPHB ($\alpha=9$)	-541.30	-160.80	0.01	145.20	2897.00
Bayes bootstrap	0.00	0.00	0.00	0.00	0.00

Percent of predictions within specified bounds of target for 54 uniques

Imputation model	% within \$10	% within \$20	% within \$50
Parametric	0.0	5.6	20.4
NPHB ($\alpha=19$)	7.4	7.4	20.4
NPHB ($\alpha=9$)	13.0	18.5	20.4
Bayes bootstrap	100.0	100.0	100.0

Example (2)

- Evaluate CURF, multiply imputed synthetic data and sufficiency based perturbation against real IS 2003.
- Same restricted subset as for Example (1):
- Restricted to ages 25-64, positive income, positive average weekly hours worked.
- Other variables: age, sex, ethnic group, education.
- Treated as SRS.
- Results below are for income.

Estimation of mean weekly income and income quartiles.

36 groups – not all mutually exclusive,
various sample sizes

Full sample,

Males, females

Ethnic groups,

Educational achievement groups

Sex by ethnicity

Sex by educational achievement

|

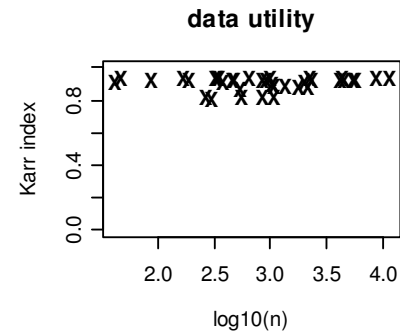
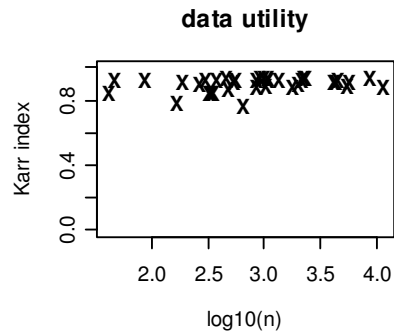
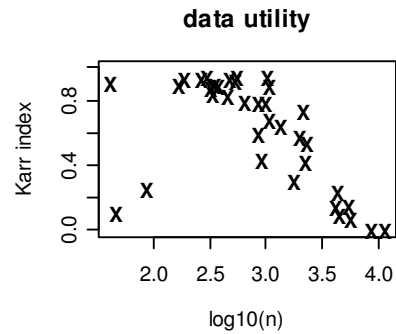
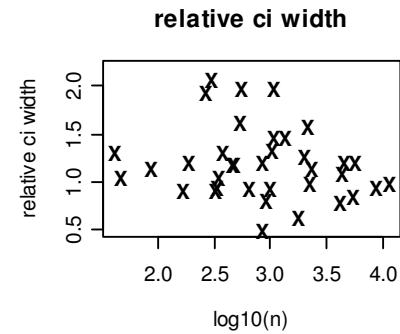
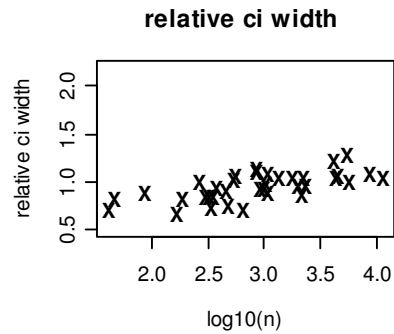
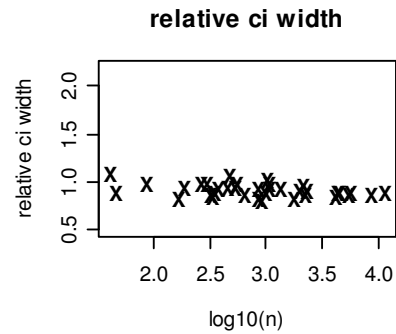
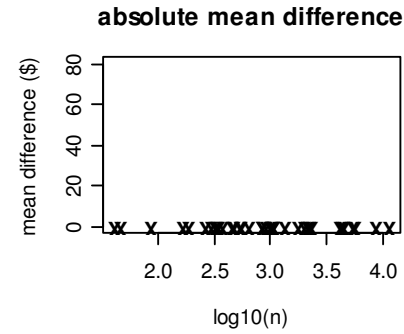
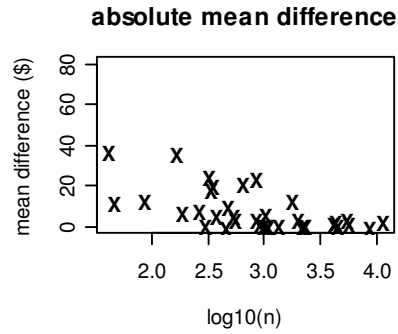
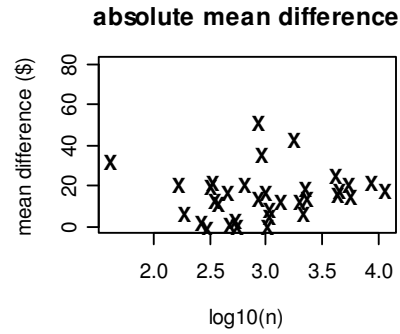
Imputation model specified on original scale.

**Comparison of CURF, MI (alpha=19) and SBP: Estimating mean income for 36 groups.
Imputation/perturbation modelling on original scale.**

CURF

MI: non-parametric Bayes

perturbation

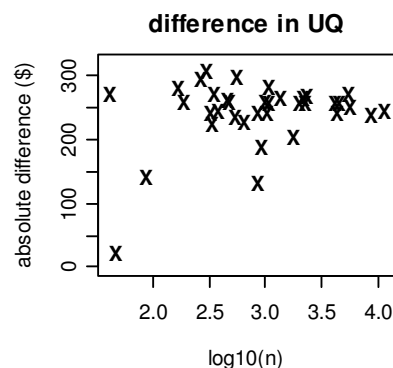
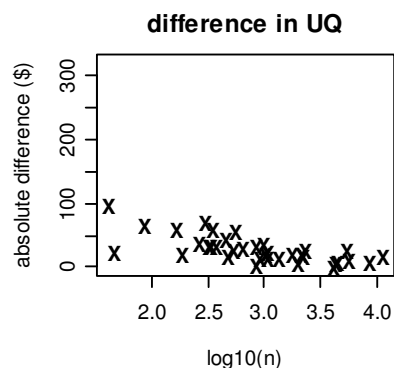
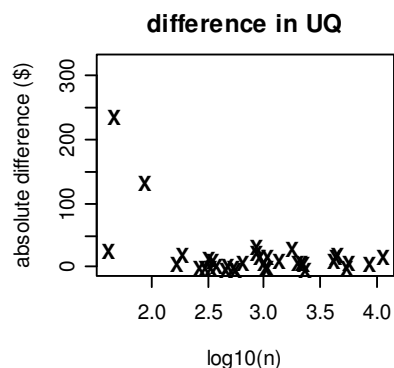
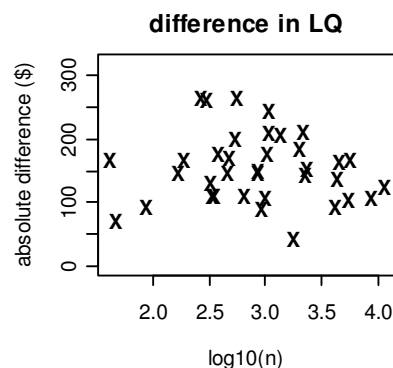
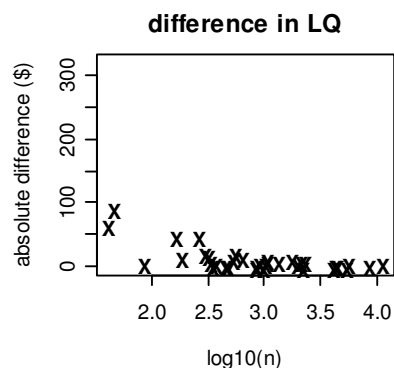
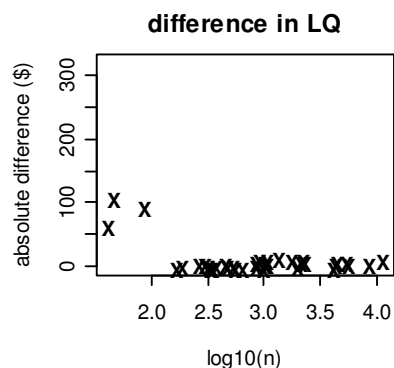
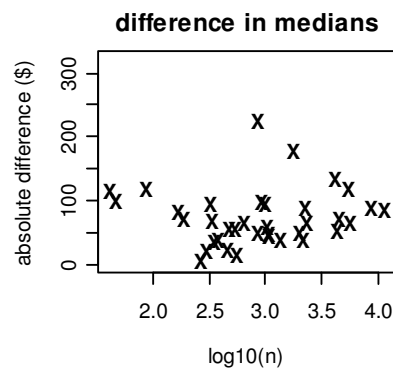
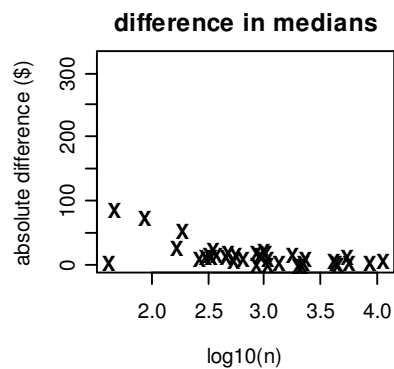
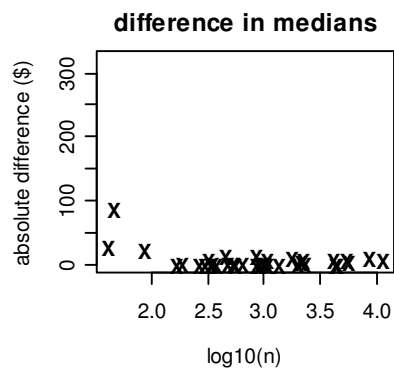


**Comparison of CURF, MI (alpha=19) and SBP: Estimating income quartiles for 36 groups.
Imputation/perturbation modelling on original scale.**

CURF

MI: non-parametric Bayes

perturbation



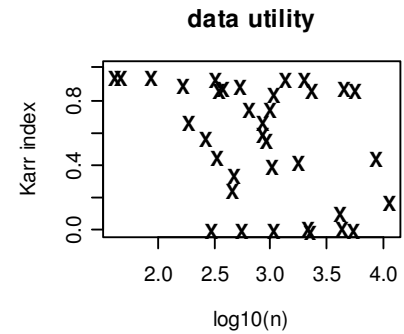
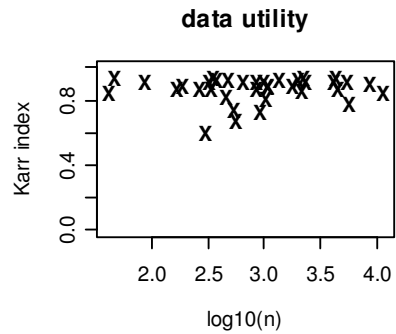
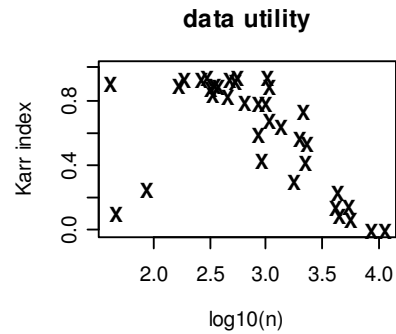
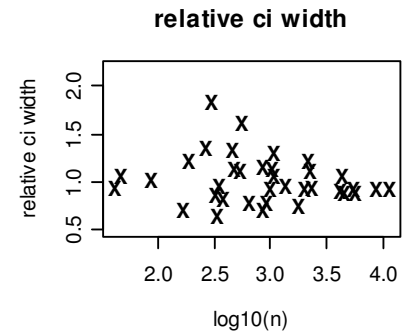
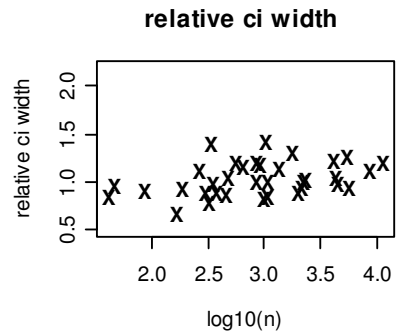
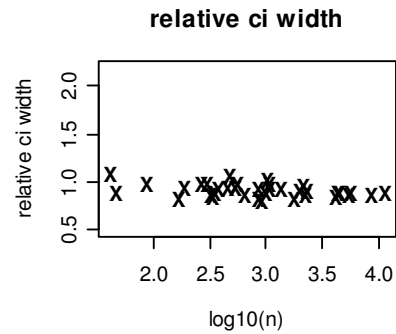
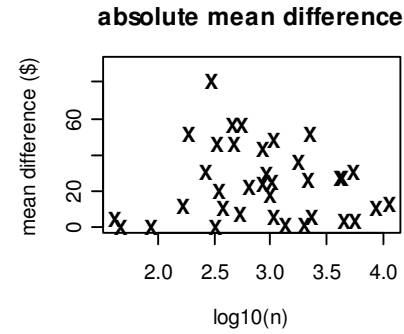
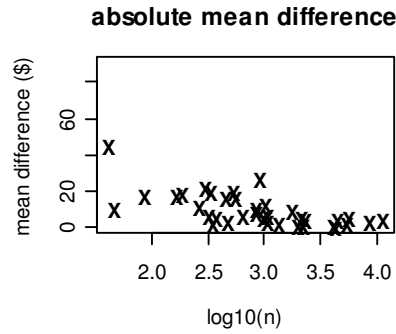
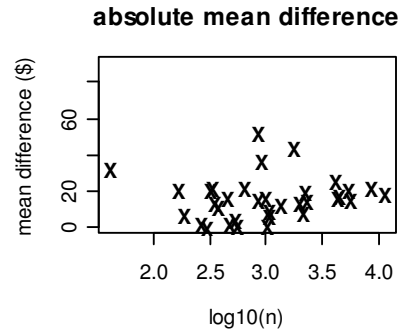
Imputation model specified on log scale.

**Comparison of CURF, MI (alpha=19) and SBP: Estimating mean income for 36 groups.
Imputation/perturbation modelling on log scale.**

CURF

MI: non-parametric Bayes

perturbation

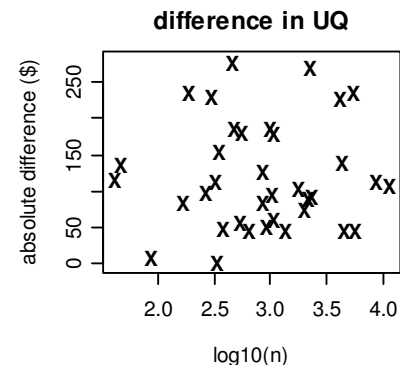
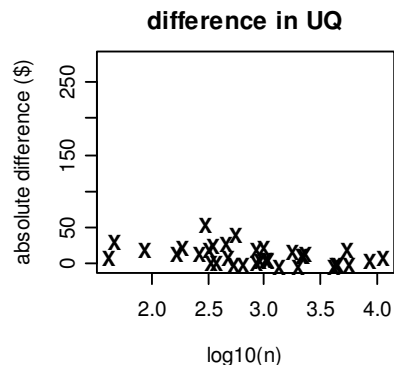
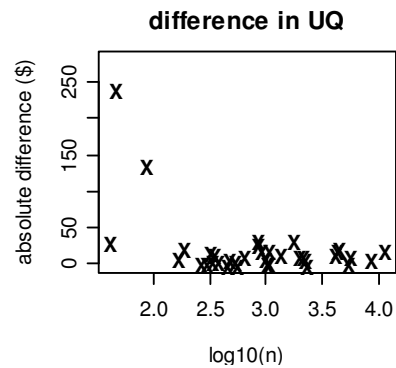
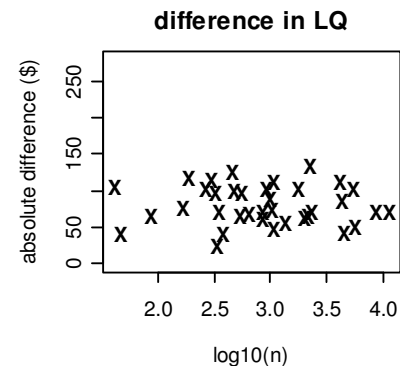
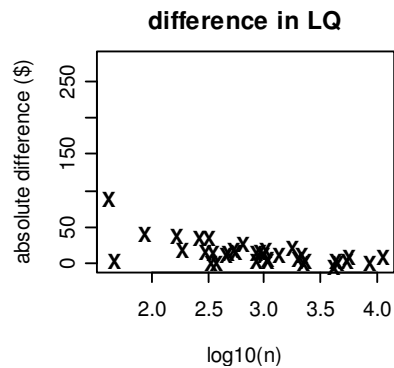
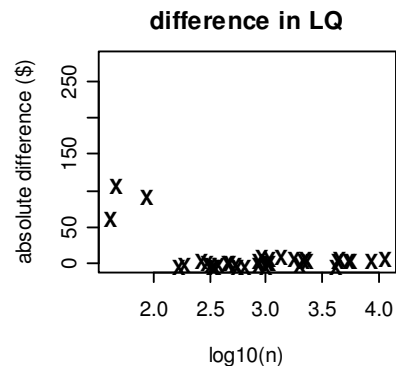
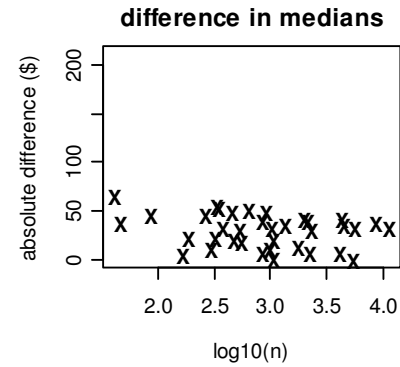
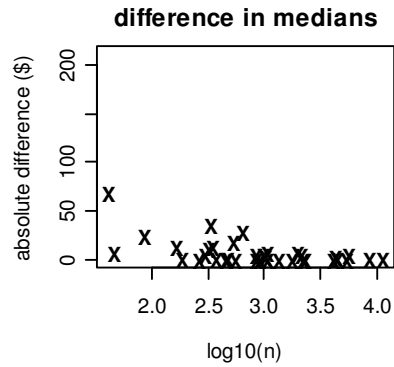
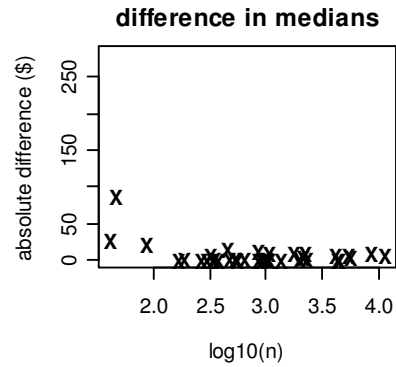


**Comparison of CURF, MI (alpha=19) and SBP: Estimating income quartiles for 36 groups.
Imputation/perturbation modelling on log scale.**

CURF

MI: non-parametric Bayes

perturbation



Estimating log-linear regressions for income under imputation / perturbation models specified on log-scale

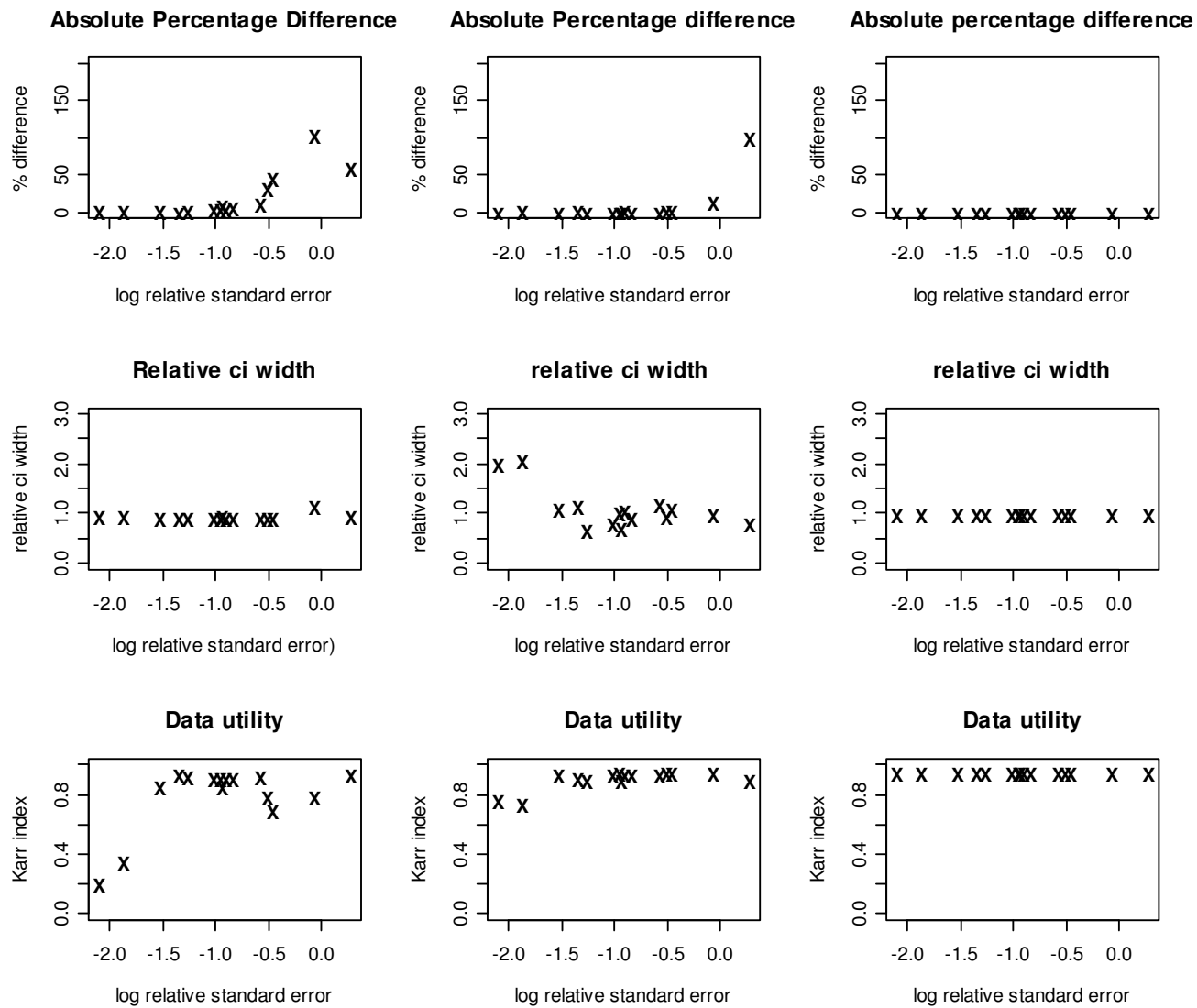
1. Main effects model including log working hours

**Comparison of MI methods: imputation modelling on log scale.
 Estimating income main effects regression, working hours on log scale**

CURF

non-parametric Bayes, alpha=19

perturbation



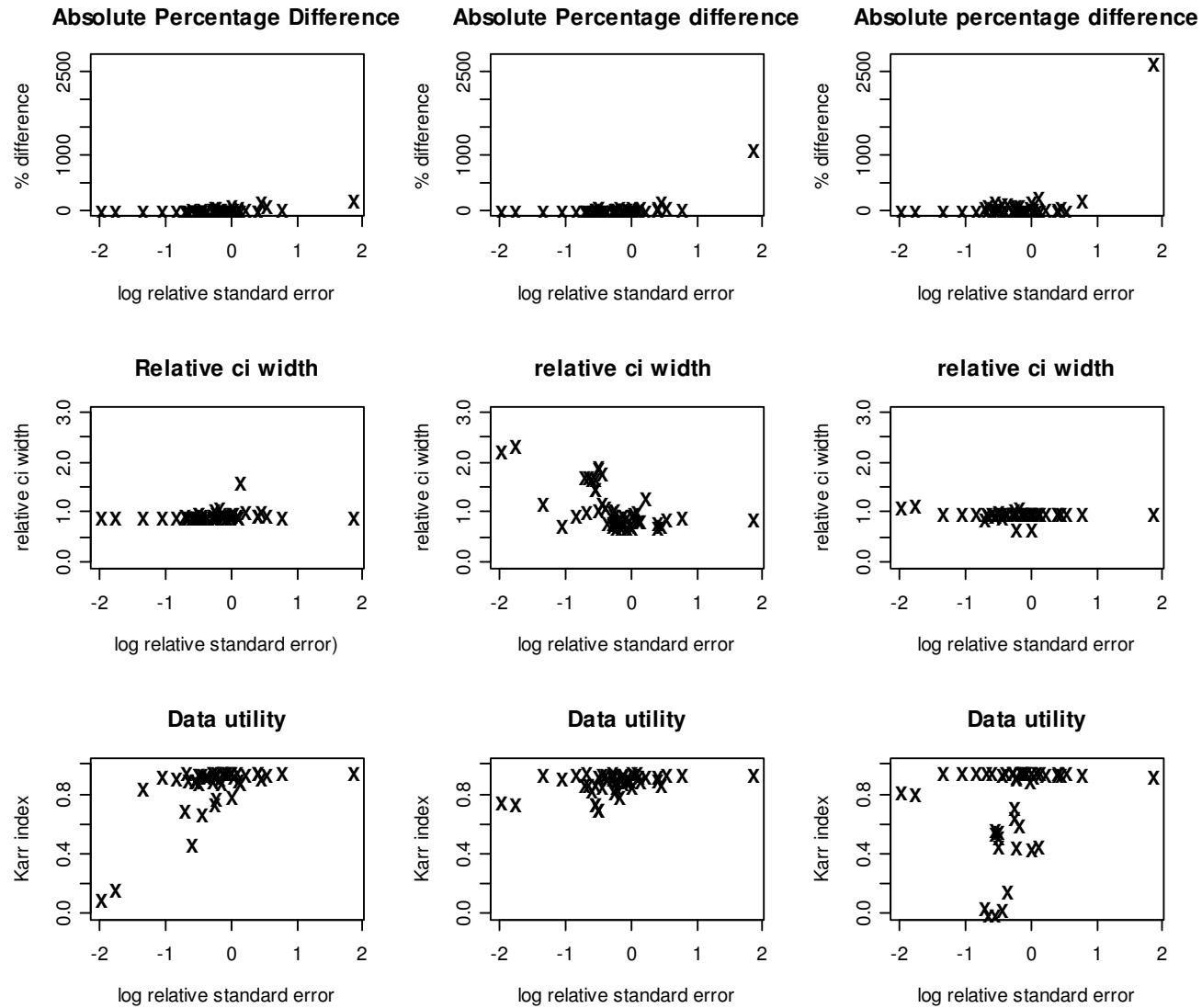
2. More complex log-linear regression model, including log working hours and some interaction terms not represented in the imputation model.

**Comparison of MI methods: imputation modelling on log scale.
 Estimating income regression with interactions, working hours log transformed**

CURF

non-parametric Bayes, alpha=19

perturbation



3. Estimating complex log-linear regression with working hours included on original scale and including some interaction terms not included in the imputation models (specified on log-scale).

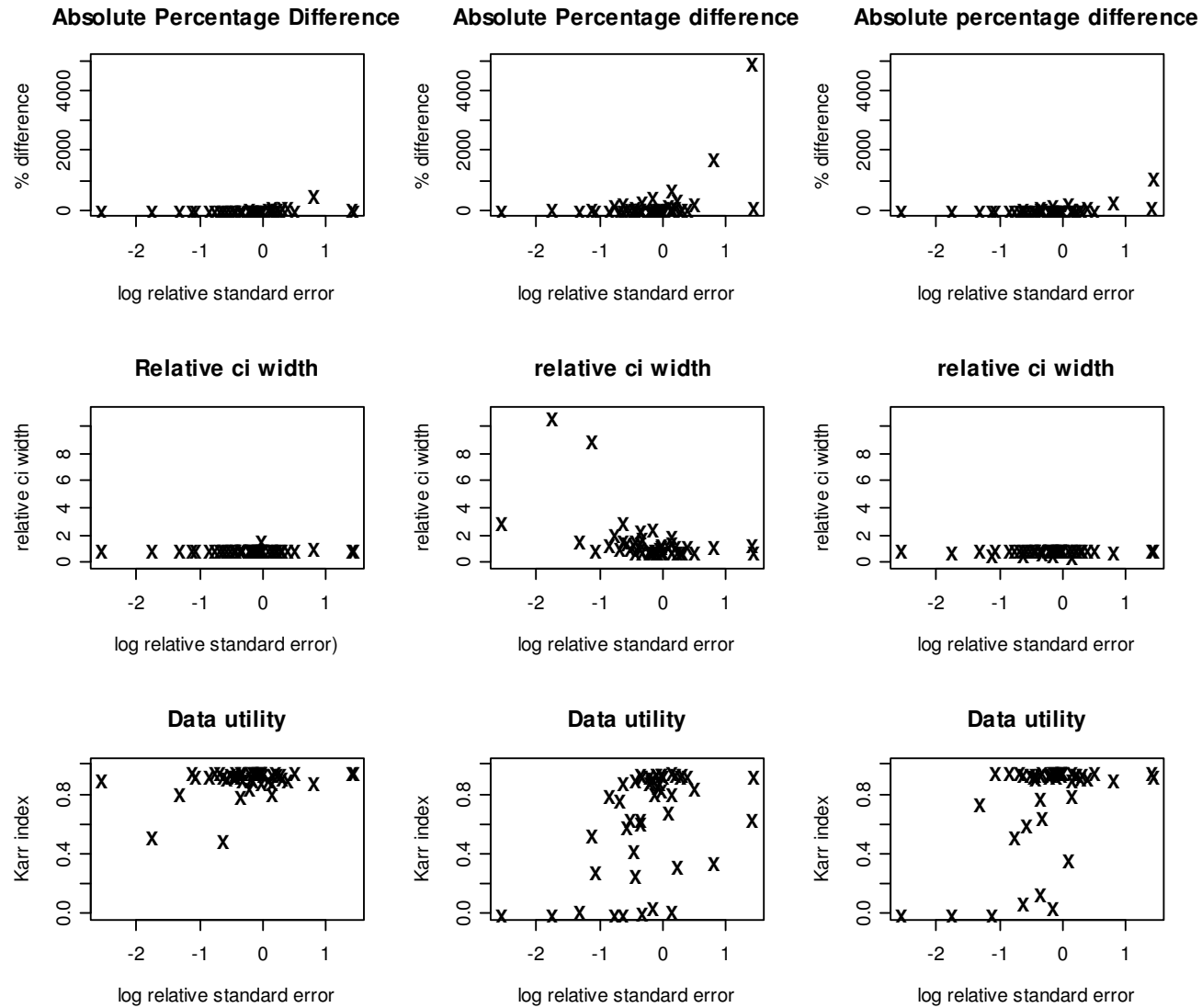
Worst case analysis model for log-scale imputation model.

**Comparison of MI methods: imputation modelling on log scale.
 Estimating income regression with interactions, working hours untransformed**

CURF

non-parametric Bayes, alpha=19

perturbation



Estimating log-linear regression including log-working hours and interaction terms not represented in the imputation / perturbation model which is specified on the original scale.

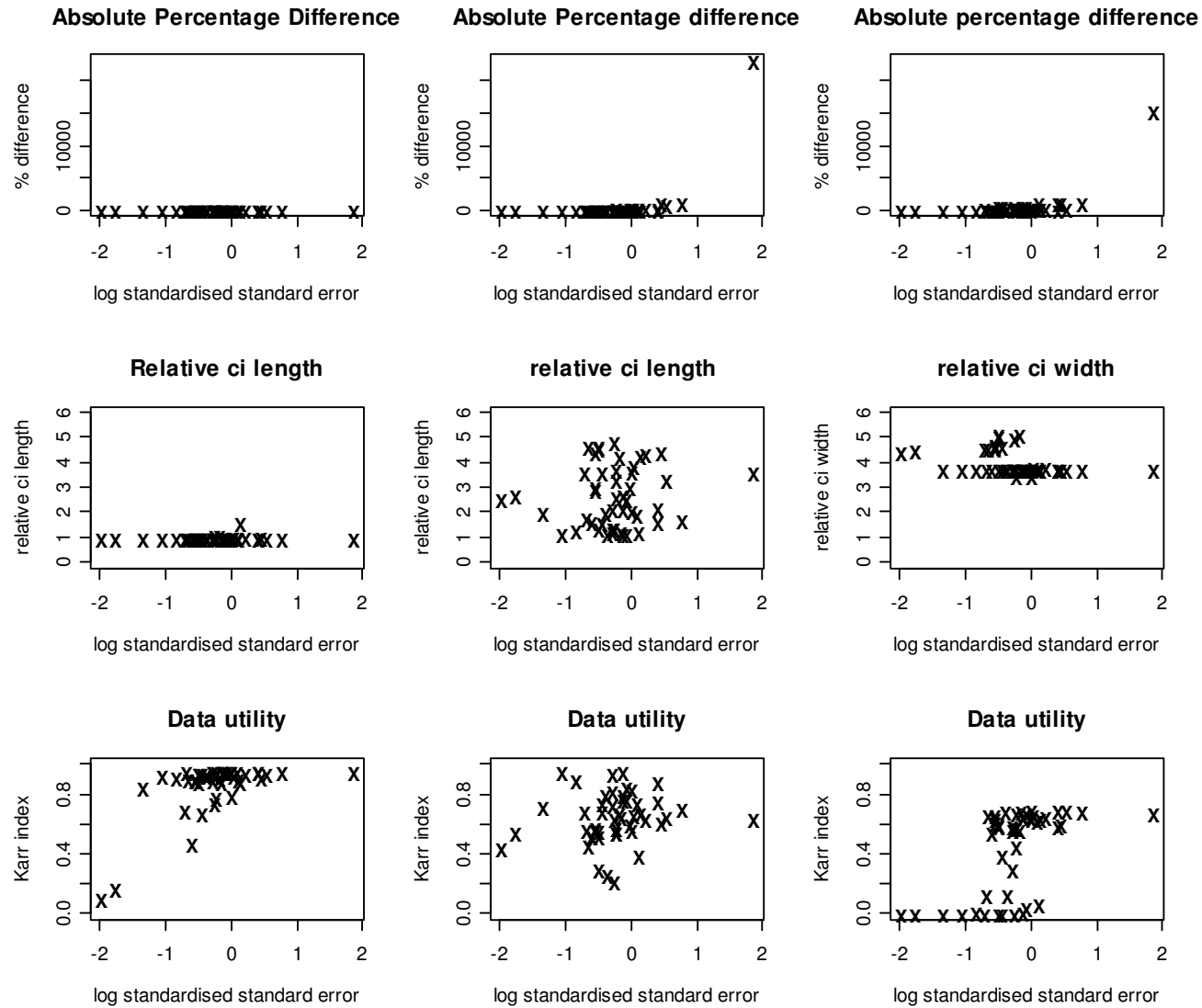
Worst case analysis model for imputation / perturbation model specified on the original scale

**Comparison of MI methods: imputation modelling on original scale.
 Estimating income regression with interactions, working hours log transformed**

CURF

non-parametric Bayes, alpha=19

perturbation



So, how do the methods compare.

CURF

Strengths

- Accepted
- In general performs well in preserving inferences

Weaknesses

- Occasional poor performance in estimation
- Hard to automate
- Disclosure risks ?

Multiply imputed synthetic data, using non-parametric hierarchical Bayesian models

Strengths

- Extendable
- Robust to some types of discrepancy between imputation and analysis models

Weaknesses

- Implementation needs sophisticated knowledge of Bayesian modelling.
- Current implementation (HB MVN centering model) is sensitive to certain types of discrepancy between imputation analysis models.

Sufficiency based perturbation.

Strengths

- Easy implementation
- Good protection of individual values.

Weaknesses

- Extreme sensitivity to discrepancies between perturbation and analysis models.

Thanks

Summary

- Extending HB imputation models to mixed categorical and numerical data seems to need more than HB MVN regression model.
- Non-parametric Bayesian approach looks promising.
- Non-parametric Bayesian approach leads to Polya sampling at final stage of imputation, so “synthetic” data is actually a mix of real and synthetic data – may have some appeal, but presumably increases disclosure risk

Next steps

- Non-parametric Bayesian model at second stage of the HB model rather than the first?
- Adapt to complex survey data.
- Scale up to higher dimensional problems