Estimating the Density in a Subgroup With Imputed Subgroup Indicators

Nielsen, Søren Feodor

Document Version
Final published version

Published in:
Research in Statistics

DOI:
10.1080/27684520.2023.2279720

Publication date:
2023

License
CC BY

Citation for published version (APA):
Nielsen, S. F. (2023). Estimating the Density in a Subgroup With Imputed Subgroup Indicators. *Research in Statistics*, *1*(1). https://doi.org/10.1080/27684520.2023.2279720

Research in Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/urst20

Estimating the density in a subgroup with imputed subgroup indicators

Søren Feodor Nielsen

To cite this article: Søren Feodor Nielsen (2023) Estimating the density in a subgroup with imputed subgroup indicators, Research in Statistics, 1:1, 2279720, DOI: 10.1080/27684520.2023.2279720

To link to this article: https://doi.org/10.1080/27684520.2023.2279720

RESEARCH IN STATISTICS
2023, VOL. 1, NO. 1, 2279720
https://doi.org/10.1080/27684520.2023.2279720

Estimating the density in a subgroup with imputed subgroup indicators

Søren Feodor Nielsen

Center for Statistics, Copenhagen Business School, Frederiksberg, Denmark

ABSTRACT

In this paper, I consider the problem of how to estimate the density in a subgroup when some of the subgroup indicators are missing at random. Four different imputation estimators are compared to each other and to an inverse probability weighted estimator suggested previously. An optimal estimator is derived. I also provide expressions for the asymptotic variance of the imputation estimators including terms of order $1/(nh)$ and $1/n$.

ARTICLE HISTORY
Received 31 July 2023
Accepted 30 October 2023

KEYWORDS
Kernel density estimation; imputation; Rao-Blackwellization; optimal estimation; variance estimation

1. Introduction

Missing data is a problem in many applications of statistics. Observations are incompletely observed either by accident or intentionally, and failure to take the incompleteness into account in the statistical analysis may bias the results. Strategies for analyzing incomplete data range from ad hoc methods to principled methods. The focus in this paper is on imputation, i.e., filling in suitable values for the missing observations. This leads to a complete data set, which allows us to use the estimator we would have used had we had complete data. The disadvantage of imputation is that the estimation of standard errors is complicated by the fact that the imputed data does not give the same information as the original data. Hence using the formula for the standard error of the complete data estimator will typically underestimate the uncertainty.

The problem considered in this paper is how to estimate the density of a subgroup of the sample, when it is not known for some of the observations, if they belong to the relevant subgroup or not. If the reason for the missing group membership is related to the variable for which we wish to estimate the density, then a complete case estimator—a density estimator based on observations known to be in the group of interest—will be biased. A typical example of such data comes from medical statistics, where it may too costly or too difficult to ascertain the disease status of all patients. As a consequence disease status will be more likely to be known for patients that are more likely to be ill. If we are interested in the distribution of a measurement in the diseased sub-population, those known to be ill will then be a biased sample. A proxy measurement may be available for all patients allowing us to impute the missing disease status. In business economics, data with partially missing group information occurs for instance in auditing and rejection inference. Not all records are audited, and in many cases the records selected for audit are not chosen completely at random but based on dollar amount ("sampling proportional to size") or perceived likelihood of being flawed ("non-statistical" sampling). If we are interested in e.g., the distribution of the dollar amount among the flawed records, the inspected records may be a biased sample. When applying for a loan, the application is granted or denied based on the applicant's credit status and other information. Rejection inference aims at improving credit scoring by incorporating information from rejected applicants who would have repaid the loan had it been granted. But whether an applicant would have repaid their loan is obviously missing for rejected applicants, and the non-rejected applicants will typically be a biased sample.

The problem of estimating a subgroup density with missing group membership information has previously been considered by Tang, He, and Gunzler (2012), who use inverse probability weighted kernel density estimation to get a consistent asymptotically normal estimator of the unknown density. Imputation—replacing missing group membership information with suitable predictions or simulations—is an obvious alternative approach. In this paper, I will consider a number of imputation estimators and compare them to the estimator suggested by Tang, He, and Gunzler (2012).

The remainder of this paper is structured as follows. In Section 2, I define notation and specify the regularity assumptions that will be used to derive the main asymptotic results, which are given in Section 3. The following section gives results from a small simulation experiment, and the estimators are then applied to data in Section 5. In Section 6, I present higher-order expressions for the asymptotic variance of our estimators, before I conclude in Section 7. Proofs are deferred to the Appendix.

2. Set-up

Let $Y_1, \ldots, Y_n$ be iid real random variables with an unknown density f. Let $D_1, \ldots, D_n$ iid 0–1 variables, such that $D_i = 1$ if

CONTACT Søren Feodor Nielsen sfn.mes@cbs.dk Center for Statistics, Copenhagen Business School, Frederiksberg, Denmark.

and only if the ith observation belongs to the group of interest. Hence, I am interested in estimating the conditional density, f_1, of Y_i given $D_i = 1$.

Let

$$p(y) = P\{D_i = 1|Y_i = y\}$$

and note that for all $y \in \mathbb{R}$

$$p(y)f(y) = f_1(y)p \quad \text{with} \quad p = P\{D_i = 1\}.$$

I assume that $p(y) > 0$ whenever $f(y) > 0$, so that the marginal distribution of Y_i and the conditional distribution of Y_i given $D_i = 1$ share support.

I will assume that the D_i's are missing at random (MAR), i.e., that the probability that D_i is missing does not depend on D_i but only on observed data. As Tang, He, and Gunzler (2012) do, I will allow this probability to depend not only on Y_i but also on additional variables X_i. If R_i denotes the "response indicator" such that $R_i = 1$ if and only if D_i is observed, the assumption of MAR is that

$$\begin{aligned}\pi(x,y) &= P\{R_i = 1|X_i = x, Y_i = y_i\} \\ &= P\{R_i = 1|X_i = x, Y_i = y, D_i = d\}, \quad d = 0, 1.\end{aligned} \tag{1}$$

I let π_i denote $\pi(X_i, Y_i)$.

As mentioned in Section 1 a complete case estimator of the unknown conditional density will in general be biased if the probability that D_i is missing depends on Y_i. Estimating the density using inverse probability weights requires that we can estimate the π_i's; the assumption of MAR makes this feasible. Estimating the density using imputation requires on the other hand that we can estimate the distribution of D_i given Y_i and X_i. A consequence of MAR is that

$$\begin{aligned}p(x,y) &= P\{D_i = 1|X_i = x, Y_i = y\} \\ &= P\{D_i = 1|X_i = x, Y_i = y, R_i = 1\}.\end{aligned} \tag{2}$$

Thus, the conditional probability of belonging to the group of interest can be estimated from the complete cases, i.e., observations with $R_i = 1$. I put $p_i = p(X_i, Y_i)$. I will assume throughout this paper that the vectors (Y_i, D_i, R_i, X_i), $i = 1, \ldots, n$, are iid.

As a model for the conditional probability $p(x, y)$ of $D_i = 1$ given $X_i = x$ and $Y_i = y$ I will assume a generalized linear model with linear predictor $Z_i^\top \beta$, where β is an unknown parameter and Z_i is a vector of explanatory variables constructed from Y_i and X_i. Assuming that the components of Z_i have second moments, it follows that

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i + o_P(1)$$

where

$$S_i = R_i \frac{D_i - p_i}{p_i(1 - p_i)} p_i' \mathcal{I}(\beta)^{-1} Z_i \tag{3}$$

with $\mathcal{I}(\beta)$ equal to the expected information matrix (given $R_i = 1$) and p_i' is the derivative of the inverse link function taken at the ith linear predictor, i.e., $\frac{d}{d\beta} p_i = p_i' \cdot Z_i$. The factor R_i in (3) is due to the fact that the estimation is based on complete cases only. I will assume that the inverse link function of this generalized linear model has a bounded derivative. Commonly used link functions, such as logit and probit, satisfy this assumption. I let $\hat{p}_i$ equal the fitted probabilities at (X_i, Y_i). I will refer to the assumptions on the distribution of the data made so far as assumption A.

I choose a kernel function, K, such that:

$$\begin{aligned}\int K(u)du &= 1 & \int uK(u)du &= 0 \\ \int u^2 K(u)du &= 1 & \int K^2 \equiv \int K(u)^2 du &< \infty\end{aligned} \tag{4}$$

These assumptions, which I will refer to as assumption B, are satisfied for instance if the kernel K is a bounded symmetric probability density with variance equal to 1. Finally, I need some smoothness of various functions of y: $f(y), f_1(y), p(y), E[p_i^2|Y_i = y]$, $E[\pi_i p_i|Y_i = y]$, $E[\pi_i p_i^2|Y_i = y]$ are all assumed to be twice continuously differentiable with a second derivative which is locally Lipschitz so that for each $y \in \mathbb{R}$ there is a constant (possibly depending on y) such that for $f_1(y)$

$$|f_1''(y) - f_1''(y + \delta)| \leq \text{const} \cdot \delta$$

for δ sufficiently small (possibly depending on y) and similarly for the other functions. I do not need smoothness of all of these functions for each result. Consequently I will let assumption C1 be the assumption about $f(y)$, C2 the assumption on $f_1(y)$, and so on. For completeness I have listed the assumptions in the Appendix.

3. Asymptotic results

With complete data, the kernel density estimator for f_1 is given by

$$\hat{f}_1(y) = \frac{\sum_{i=1}^n D_i K\left(\frac{y - Y_i}{h}\right)/h}{\sum_{i=1}^n D_i}. \tag{5}$$

Provided that $n \to \infty$ and $h \to 0$ such that $nh \to \infty$

$$\hat{f}_1(y) \sim \text{approx}\, N\left(f_1(y) + \frac{h^2}{2} f_1''(y), \frac{f_1(y)}{pnh} \int K^2\right).$$

With missing data, a complete case estimator, i.e., replacing D_i by $D_i R_i$ in (5), will be biased with an asymptotic mean equal to

$$\frac{1}{E[\pi_1 p_1]}\left(E[\pi_1 p_1|Y_1 = y]f(y) + \frac{h^2}{2}\frac{d^2}{dy^2}\left(E[\pi_1 p_1|Y_1 = y]f(y)\right)\right) + o(h^2),$$

the density of Y_i given that $D_i = 1$ and $R_i = 1$ plus $O(h^2)$-terms.

One way of eliminating this bias is to weight each observed D_i with the inverse probability of observing it, i.e., replace D_i in (5) by $D_i R_i/\hat{\pi}_i$, where $\hat{\pi}_i$ is an estimator of π_i. Tang, He, and Gunzler (2012) show that the asymptotic distribution of the resulting estimator is

$$N\left(f_1(y) + \frac{h^2}{2} f_1''(y), E[1/\pi_1|Y_1 = y]\frac{f_1(y)}{pnh} \int K^2\right)$$

as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$ for y in a bounded set using a logistic regression for estimating π. It

seems that they implicitly assume that $p(x, y)$ only depends on y. They also consider the case where π_i is known and show that the weighted estimator using the known π_i's has the same asymptotic distribution as when the π_i's are estimated.

In the following subsections, I will consider the asymptotic distribution of the density estimator (5), when missing D_i's are replaced by imputations.

3.1. Single imputation

The simplest imputation estimator is obtained by replacing missing D_i's by simulated 0–1 variables where the success probabilities are given by $\hat{p}_i = \hat{p}(X_i, Y_i)$, the fitted value from the generalized linear model. Hence, the estimator is given by

$$\hat{f}_{1,SI}(y) = \frac{\sum_{i=1}^n \tilde{D}_i K\left(\frac{y-Y_i}{h}\right)/h}{\sum_{i=1}^n \tilde{D}_i} \tag{6}$$

with

$$\tilde{D}_i = R_i D_i + (1 - R_i) 1_{\{U_i \le \hat{p}_i\}}$$

where $(U_i)_{i=1,\dots,n}$ is a sequence of independent uniformly distributed random variables, independent of the observations. For this estimator, I obtain the following result:

Theorem 1. Under assumptions A, B, C1, C2, C4–C6, then for any y in the interior of the support of Y_i

$$\hat{f}_{1,SI}(y) \sim \text{approx}\, N\left(f_1(y) + \frac{h^2}{2} f_1''(y), \frac{f_1(y)}{pnh}\int K^2\right)$$

as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$.

I only give results for y's in the interior of the support. At boundary points the bias of the kernel density estimator may be considerably larger. I conjecture that using a boundary kernel in this missing data set-up will give results similar to those found in complete data cases but proving this is beyond the scope of this paper. Alternatively, boundary effects can be mitigated using transformation as I do in Section 5.

The proof of Theorem 1 may be found in the Appendix. The asymptotic variance may be consistently estimated by $\hat{f}_{1,SI}(y)/(h\sum_{i=1}^n \tilde{D}_i) \cdot \int K^2$.

It is somewhat surprising that not only does the asymptotic distribution of the imputation estimator $\hat{f}_{1,SI}(y)$ not depend on the missing data mechanism in any way, it is in fact the same asymptotic distribution as that of the complete data estimator (5). In particular, it is more efficient than the inverse probability weighted estimator considered by Tang, He, and Gunzler (2012). This is due to the fact that the rate of convergence of $\hat{\beta}$ is $\sqrt{n}$ whereas the rate of convergence of the kernel density estimator is only $\sqrt{nh}$. Therefore the estimation uncertainty regarding β may be ignored asymptotically, and the imputations may be treated as if they were the actual (missing) observations. As I will discuss later, the estimation of β may well influence the precision of the imputation estimator (6) in finite samples.

3.2. Multiple imputation

Multiple imputation (Rubin 1987) is a popular alternative to the single imputation considered in the previous subsection. Rather than just generate one set of imputations ($\tilde{D}_i$ for i such that $R_i = 0$), $B > 1$ sets of imputations are generated, B imputation estimators similar to (6) are calculated and then averaged to obtain the estimator

$$\hat{f}_{1,MI}(y) = \frac{1}{B}\sum_{j=1}^B \hat{f}_{1,j}(y)$$

where

$$\hat{f}_{1,j}(y) = \frac{\sum_{i=1}^n \tilde{D}_{i,j} K\left(\frac{y-Y_i}{h}\right)/h}{\sum_{i=1}^n \tilde{D}_{i,j}}$$

with

$$\tilde{D}_{i,j} = R_i D_i + (1 - R_i) 1_{\{U_{i,j} \le \hat{p}_{i,j}\}}$$

where $(U_{i,j})_{i=1,\dots,n,j=1,\dots,B}$ are independent uniformly distributed random variables independent of the data; $\hat{p}_{i,j}$ will be discussed below. Apart from decreasing the simulation noise, multiple imputation is attractive because—under suitable regularity assumptions—the variance of the estimator may be estimated by a combination of the average of the "complete data variance estimators" and the empirical variance of the B imputation estimators:

$$\frac{1}{B}\sum_{j=1}^B \frac{\hat{f}_{1,j}(y)}{h\sum_{i=1}^n \tilde{D}_{i,j}}\int K^2 + \left(1 + \frac{1}{B}\right)\frac{1}{B-1}\sum_{j=1}^B \left(\hat{f}_{1,j}(y) - \hat{f}_{1,MI}(y)\right)^2. \tag{7}$$

Rubin (1987) suggests that a requirement for the suitability of (7) as an estimator of the asymptotic variance of the multiple imputation estimator $\hat{f}_{1,MI}(y)$ is that the imputations are generated at least "approximately" from a Bayesian model. Thus, $\hat{p}_{i,j}$ should be the predicted probability of $D_i = 1$ given (X_i, Y_i) using $\tilde{\beta}_j$ as the value of the unknown parameter, where $\tilde{\beta}_j, j = 1, \dots, B$ are independent draws from the posterior distribution of β given the observed data. This is my assumption D.

In the present case I obtain the following result:

Theorem 2. Under assumptions A, B, C1, C2, C4–C6, and D, then for any y in the interior of the support of Y_i

$$\hat{f}_{1,MI}(y) \sim \text{approx}\, N\left(f_1(y) + \frac{h^2}{2} f_1''(y), \left(v_1(y) + \frac{1}{B}v_2(y)\right)\frac{f(y)}{p^2 nh}\int K^2\right)$$

with

$$v_1(y) = E[\pi_1 p_1 + (1 - \pi_1)p_1^2 | Y_1 = y]$$
$$v_2(y) = E[(1 - \pi_1)p_1(1 - p_1) | Y_1 = y]$$

as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$. Moreover, the variance estimator (7) estimates

$$\frac{f_1(y)}{pnh}\int K^2 + \left(1 + \frac{1}{B}\right)v_2(y)\frac{f(y)}{pnh}\int K^2.$$

The asymptotic distribution of $\hat{f}_{1,MI}(y)$ does not depend on whether the $\tilde{\beta}_j$'s are from a Bayesian posterior distribution or are all equal to the MLE $\hat{\beta}$. As in the previous section, this is due to the faster rate of convergence. In fact, the Bernstein-von Mises theorem says that given the observed data

$$\tilde{\beta}_j \sim \text{approx}\, N(\hat{\beta}, \mathcal{I}(\hat{\beta})^{-1}/n). \tag{8}$$

This means that the difference between $\tilde{\beta}_j$'s and $\hat{\beta}$ is negligible compared to the $\sqrt{nh}$-rate of convergence of the kernel density estimator.

Note that $\nu_1(y) + \nu_2(y) = p(y)$ so that when $B = 1$ the asymptotic variance of $\hat{f}_{1,MI}(y)$ equals the asymptotic variance of $\hat{f}_{1,SI}(y)$ as it should. Furthermore, the asymptotic variance of $\hat{f}_{1,MI}(y)$ is strictly decreasing as B increases. Note also that the variance estimator (7) is biased. This is due to the fact that the imputations asymptotically may be treated as if they were actual observations and not simulations. An simple asymptotically unbiased estimator may in principle be obtained by

$$\frac{1}{B}\sum_{j=1}^{B} \frac{\hat{f}_{1,j}(y)}{h\sum_{i=1}^{n}\tilde{D}_{i,j}}\int K^2 + \left(\frac{1}{B}-1\right)\frac{1}{B-1}\sum_{j=1}^{B}\left(\hat{f}_{1,j}(y)-\hat{f}_{1,MI}(y)\right)^2.$$

In practice, however, this estimator will be negative with positive probability as the first term is asymptotically a constant divided by nh, and the second term is asymptotically χ^2-distributed with $B-1$ degrees of freedom, with a scale parameter inversely proportional to nh. A consistent estimator may be obtained for instance by kernel regression of $R_i\hat{p}_i + (1-R_i)\hat{p}_i^2$ and $(1-R_i)\hat{p}_i(1-\hat{p}_i)$ on Y_i to obtain estimates of $\nu_1(y)$ and $\nu_2(y)$, respectively, combined with a kernel density estimator of the marginal density $f(y)$.

3.3. Fixed imputation

The multiple imputation estimator is asymptotically equivalent to a "single imputation" estimator based on the average of the imputations, i.e., (6) with $\tilde{D}_i = \frac{1}{B}\sum_{j=1}^{B}\tilde{D}_{i,j}$. Generating the imputations using $\hat{\beta}$ instead of $\hat{\beta}_j$ and letting B tend to infinity, I obtain the estimator

$$\hat{f}_{1,FI}(y) = \frac{\sum_{i=1}^{n}\left(R_iD_i + (1-R_i)\hat{p}_i\right)K\left(\frac{y-Y_i}{h}\right)/h}{\sum_{i=1}^{n}\left(R_iD_i+(1-R_i)\hat{p}_i\right)}.$$

This is often called a fixed imputation estimator. In many cases, "random imputations" (as used in the two previous sections) are preferable to "fixed imputations", as "fixed imputations" may lead to biased estimators if a nonlinear transformation is applied to the imputations. However, "fixed imputations" do not lead to biased estimation of the density here:

Theorem 3. Under assumptions A, B, C1, C2, C4–C6, then for any y in the interior of the support of Y_i

$$\hat{f}_{1,FI}(y) \sim \text{approx}\, N\left(f_1(y) + \frac{h^2}{2}f_1''(y), E[\pi_1p_1 + (1-\pi_1)p_1^2|Y_1=y]\frac{f(y)}{p^2nh}\int K^2\right)$$

as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$.

As in the previous subsection, the asymptotic variance may be estimated from a kernel regression and a marginal density estimate.

3.4. Optimal estimation

So far I have shown that if the distribution of D_i given (X_i, Y_i) can be modeled parametrically then imputed data is asymptotically as good as observed data for estimating the conditional density of Y_i given $D_i = 1$. Improvements in the form of lower asymptotic variances are obtained by multiple imputation and fixed imputation schemes. As fixed imputations are better than random imputations and random imputations are as good as real data, it would be natural to suppose that imputing all the D_i's (using fixed imputations) regardless of whether they are missing or not will lead to an even better estimator:

$$\hat{f}_{1,RB}(y) = \frac{\sum_{i=1}^{n}\hat{p}_iK\left(\frac{y-Y_i}{h}\right)/h}{\sum_{i=1}^{n}\hat{p}_i}.$$

Indeed, I obtain:

Theorem 4. Under assumptions A, B, C1–C4, then for any y in the interior of the support of Y_i

$$\hat{f}_{1,RB}(y) \sim \text{approx}\, N\left(f_1(y) + \frac{h^2}{2}f_1''(y), E[p_1^2|Y_1=y]\frac{f(y)}{p^2nh}\int K^2\right)$$

as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$.

As such the result—that the estimator is improved if the group indicators are replaced by estimates of the conditional probabilities—is not surprising. The faster rate of convergence ensures that the estimated p_i's may be treated as known, and then the improvement is just an example of a "Rao-Blackwellization". This phenomenon occurs in other missing data examples, too. For instance, Müller (2009) shows that what she calls "full imputation" (replacing observed data with imputations) is efficient in her model, a semi-parametric regression with outcomes missing at random, when the imputations are chosen optimally. She does not consider smoothing, so her results are not directly applicable to the problem considered in this paper.

The question is if there are further improvements that may be easily obtained? I restrict attention to estimators of the form

$$\hat{f}_{1,d}(y) = \frac{\sum_{i=1}^{n} d_iK\left(\frac{y-Y_i}{h}\right)/h}{\sum_{i=1}^{n} d_i} \tag{9}$$

with $d_i = d(Y_i, X_i, R_i, D_i, U_i)$, where U_i is a (simulated) uniformly distributed random variable (so that random imputation estimators are included), such that

$$d(Y_i, X_i, 0, 1, U_i) = d(Y_i, X_i, 0, 0, U_i)$$
$$E[d_i|Y_i = y] = p(y).$$

The first of these restrictions ensures that the estimator only depends on observed data, i.e., does not depend on D_i when $R_i =$

0. The second ensures that the estimator (9) is asymptotically unbiased. The imputation estimators belong to this class if $p(x, y)$ is known as does the estimator considered by Tang, He, and Gunzler (2012) when π is known. The asymptotic variance of an estimator of this form is $E[d_1^2|Y_1 = y]\frac{f(y)}{p^2 nh}\int K^2$ and as

$$E[d_1^2|Y_1 = y] = \text{Var}\,[d_1|Y_1 = y] + p(y)^2 \geq p(y)^2$$

with equality if and only if $d_i = p(Y_i)$ almost surely, the optimal estimator (within the specified class) is given by

$$\hat{f}_1(y) = \frac{\sum_{i=1}^n p(Y_i)K\left(\frac{y-Y_i}{h}\right)/h}{\sum_{i=1}^n p(Y_i)}.$$

In practice, $p(y)$ is typically unknown, making the optimal estimator infeasible. If, however, we are able to correctly specify a parametric model satisfying the regularity conditions listed in Section 2 for this conditional probability, then Theorem 4 shows that the estimator has the same asymptotic distribution as the infeasible optimal estimator:

$$\hat{f}_1(y) \sim \text{approx}\, N\left(f_1(y) + \frac{h^2}{2}f_1''(y), p(y)^2\frac{f_1(y)}{pnh}\int K^2\right).$$

However, estimating $p(y)$ may be complicated in our set-up. If the missingness depends on X_i as well as on Y_i, then a regression of D_i on Y_i using complete cases only will give a biased estimator of $p(y)$. One option could be to use inverse probability weighting

$$\frac{1}{n}\sum_{i=1}^n \frac{R_i}{\hat{\pi}_i}\tilde{S}_i(\gamma) = 0$$

where $\tilde{S}_i(\gamma)$ is the score function for a binomial regression of D_i on Y_i. With suitable regularity assumptions (see Robins, Rotnitzky, and Zhao 1995), the fitted $\hat{p}(Y_i)$ will be $\sqrt{n}$-consistent, and we will obtain the asymptotic result outlined above.

3.5. Bandwidth selection

It is well known that the performance of a kernel estimator is highly dependent on the chosen bandwidth. For their estimator, Tang, He, and Gunzler (2012) suggest choosing the bandwidth that minimizes the mean integrated squared error derived from the asymptotic distribution, and obviously this would be a possibility for the estimators considered in this paper as well. More computationally demanding methods such as cross validation or Ruppert's (1997) EBBS-method could also be considered. These methods make the choice of bandwidth data-driven, and it should be considered to what degree this influences the asymptotic results given above. Doing this is however outside the scope of this paper. Note however that the bandwidths used in the simulations in the next section are random, as they depend on the sample standard deviation. The effect of this seems to be minor compared to the effects of the kernel estimator itself and the effect of estimating the parameters of the regression of D_i on (Y_i, X_i).

4. Simulations

I will now present some simulation results aimed at comparing the four estimators—the single imputation estimator, the multiple imputation estimator, the fixed imputation estimator, and the "Rao-Blackwellized" estimator—discussed in the previous section to each other as well as to the complete data estimator and the weighted estimator considered by Tang, He, and Gunzler (2012). Note that I present results for the complete data estimator, not for the complete case estimator, as I am interested in the comparison of different (asymptotically) unbiased estimators, not in how they improve on a biased estimator, such as the complete case estimator. The multiple imputation estimator is based on 5 sets of imputations. I have on purpose chosen a small number of imputations to make differences between the multiple imputation estimator and the single and fixed imputation estimators stand out. The simulation results are based on 2000 replications.

In the simulations, the distribution of Y given $D = d$ is a normal distribution with mean equal to $d - 1$ and standard deviation equal to 1. Hence, the density I am trying to estimate is a standard normal density. I simulate D such that $P\{D = d\} = 1/2$ for $d = 0, 1$. Thus, the two groups are of approximately equal size, and the distribution of D given $Y = y$ is given by a logistic regression. The sample size equals 400.

As "response mechanism" I use $\pi(y) = 1/(1 + \exp(-2 - 2y))$. Hence the probability of D being unobserved decreases with y, and the distribution of Y given that D is observed to be equal to 1 is a right-skewed distribution. With this missing data mechanism, approximately 22.5% of the observations with $D = 1$ have missing data; overall approximately 15% of the observations have missing group indicators.

I use a standard normal kernel function and Silverman's rule of thumb based on the complete data in the $D = 1$-group to choose the bandwidth. This will probably give the advantage to the complete data estimator. I estimate the density in 21 equally spaced points between -2 and 2. Over this grid, $\pi(y)$ decreases from 0.88 to 0.12. The results of the simulations are presented in Figures 1 and 2. In Table 1, I present the results for the estimates of the density at $y = -1, 0, 1$.

Figure 1 shows that all estimators are behaving as expected: they have similar biases, and the largest bias is found at $y = 0$, where the underlying normal density has the largest curvature. The weighted estimator has considerably larger standard deviation for small values of y where $\pi(y)$ is small and more data is missing. The Rao-Blackwellized estimator has noticeably smaller standard deviations than the other estimators.

Figure 2 provides a closer look at the bias and the standard deviations of the six estimators. The weighted estimator has smaller bias when y is small than the other estimators, but otherwise the estimators seem to have the same bias as the asymptotic results suggest. Indeed, the bias of the weighted estimator is too small compared to what the asymptotic results predict when $\pi(y)$ is small. When it comes to the standard deviation, the imputation estimators clearly outperforms the weighted estimator, in particular when $\pi(y)$ is small. For small y, the four imputation estimators have essentially the same standard deviation as the complete data estimator, but as y increases the standard deviation of the Rao-Blackwellized estimator becomes

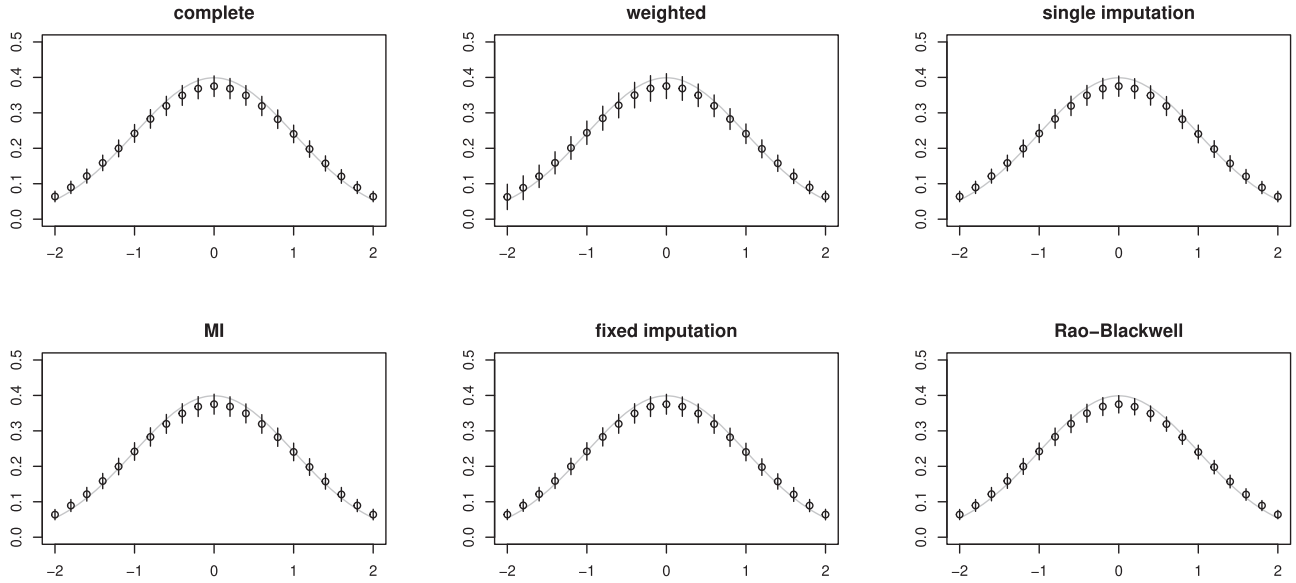

Figure 1. Simulation results. The points are the averages of 2000 simulated estimators, the vertical lines represent ± 1 times the standard deviations of the estimators. The grey curve is the density we are trying to estimate.

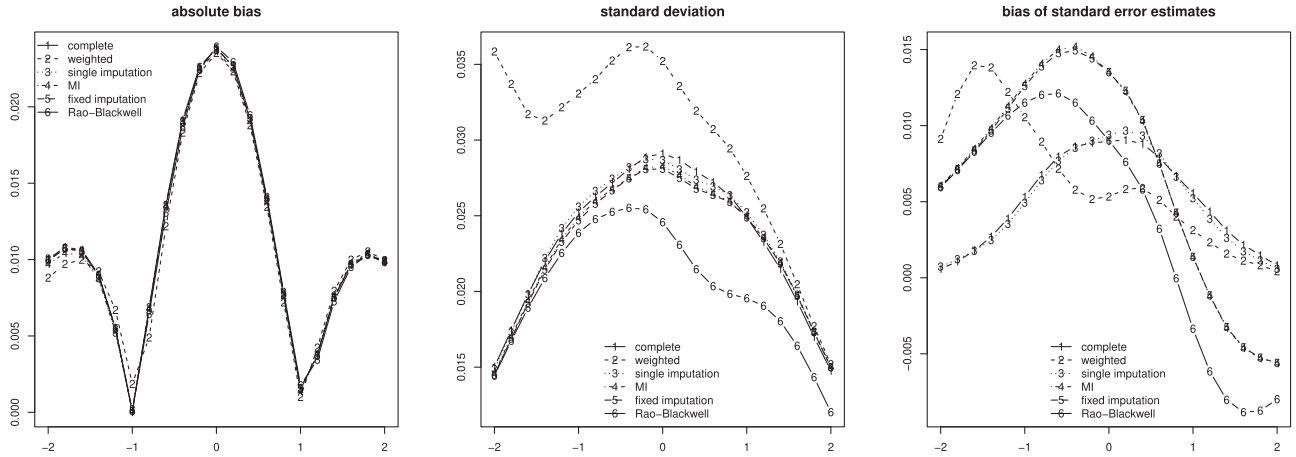

Figure 2. Simulation results. The first panel shows the absolute bias of the estimators, the second the standard deviations of the estimators. The final panel shows the bias of the estimators of the standard deviations of the estimators.

considerably smaller than the standard deviations of the others. This is due to the fact that for small y almost all the observations belong to the $D = 1$-group and consequently $\hat{p}(y) \approx 1$ when y is small. As y increases the Rao-Blackwellized estimator begins to take advantage of the observations in the second group, and this lowers the standard deviation of the estimator. There is very little difference between the four other estimators, but where there is a discernible difference the ordering of the imputation estimators is as expected: the fixed imputation estimator has smaller standard deviation than the multiple imputation estimator which is slightly better than the single imputation estimator. Contrary to what the asymptotic results suggest, the standard deviations of the single imputation estimator differ from the standard deviations of the complete case estimator. Hence, it seems that in finite samples there may well be an effect of estimating the probability of D given $Y = y$; I will investigate this in Section 6. It seems that the complete data estimator has larger standard deviation than the single imputation estimator for y close to 0. It is difficult to see from the graph, but when y is further away from 0, the complete data estimator has smaller standard deviations than the single imputation estimator.

Table 1. Simulation results for $y = -1, 0, 1$.

y	Complete	Weighted	Single	MI	Fixed	Rao-Blackwell
	0.2420	0.2438	0.2418	0.2419	0.2420	0.2421
-1	0.0252	0.0331	0.0256	0.0249	0.0247	0.0239
	0.0305	0.0436	0.0305	0.0377	0.0373	0.0354
	0.3751	0.3754	0.3752	0.3753	0.3751	0.3750
0	0.0291	0.0352	0.0287	0.0283	0.0281	0.0246
	0.0381	0.0405	0.0381	0.0418	0.0416	0.0336
	0.2405	0.2410	0.2404	0.2406	0.2404	0.2402
1	0.0249	0.0276	0.0252	0.0249	0.0249	0.0195
	0.0305	0.0308	0.0304	0.0263	0.0263	0.0162

For each value of y, the first line is the mean of the estimates, the second the standard deviation of the estimates, and the third the average of the asymptotic standard deviations.

The right-most panel of Figure 2 shows the bias of the estimators of the standard deviations of the six different estimators. As

the bias is quite large compared to the true standard deviation, the variance estimators derived may be quite misleading in finite samples. I will investigate this aspect further in Section 6.

5. Application to rejection inference data

As an illustration of the difference between the weighted estimator and the imputation estimator as well as the effect of incorporating the missing data on the results, I look at an example from rejection inference.

I use data from the online peer-to-peer lending market Lending Club previously used by Li et al. (2020). The data set contains information on 233318 loan applications of which 27157 were approved. This means that for 88.4% of the observations the information regarding whether the loan would have been repaid, had the loan been granted, is missing. The data set contains information on debt-to-income ratios, a Fair Isaac Corporation score ("FICO"), employment length and a categorized bad risk rate based on geography for each loan application. I will focus on the distribution of the FICO score for those who would repay their loan. The FICO score is one of the variables used to decide whether at loan is granted, so its distribution for those who repay, had the application been granted, will differ from its distribution among those who has been granted a loan and repaid it.

A little more than half the applications have a FICO score below 660 even though a FICO score above 660 is a requirement for a loan. Without any accepted applications with FICO scores below 660 the weighted kernel density estimator is not able to estimate the density of the FICO score below 660 in any reasonable way. The imputation estimator can, but the estimate will be based on extrapolations from the imputation model, so I remove loan applications with FICO scores below 660 and effectively estimate the density of the FICO score conditional on this being at least 660 in the group of applicants who would repay their loan. I end up with 114679 loan applications, of which 76.4% were turned down.

As the density of the truncated FICO scores is supported on the interval $[660; \infty[$ I estimate the density of the logarithm of the scores minus 660 to mitigate the boundary bias and then transform my estimate to get an estimate of the density that I am interested in.

For the imputations I use a generalized linear model with a complementary log-log link using all the explanatory variables. The debt to income ratio is log-transformed and both this variable and the FICO score are included as second order polynomials. As a model for the missing data mechanism I again use a generalized linear model with a complementary log-log link and all the explanatory variables. In this model the FICO score, the log-transformed debt to income ration and the loan amount are included as second order polynomials whereas I use a linear spline with a knot at 2 for the employment length. Both models appear to fit the observed data.

An important assumption for the weighted estimator is that the probability of response is bounded away from 0. A generalized linear model with a standard link function will not have this property unless the covariates are bounded, but in many real data examples the estimated probability of response will be sufficiently far away from 0 for this to cause any problems. In this example, however, the missing data model gives estimated probability of response very close to zero, which makes the weighted density estimator highly unstable. It spikes at the lowest FICO scores, and it is essentially zero everywhere else. As an ad hoc fix, I replace all estimated probabilities of response smaller than a cutoff by the cutoff value. The results of this are shown in Figure 3 next to the imputation estimator.

Compared to the complete case estimator (the grey curve) the imputation estimator puts more distribution mass on smaller FICO scores. This is what we should expect: low FICO scores lead to rejection, so the unapproved applications have smaller FICO scores. We only show the single imputation estimator here as all the imputation estimators are visually indistinguishable. The three following graphs show weighted kernel estimators with smaller probabilities of response replaced by a cutoff value. A cutoff value of 0.001 seems too little, and a cutoff value of 0.1 is probably too much: here the weighted density estimator is too close to the presumably biased complete case estimator. With a cutoff of 0.01 I get a density with more of the mass at very small and large FICO scores compared to the complete case estimator.

Obviously, the results depend on the chosen bandwidth. I have used the bandwidth calculated using Silverman's rule of thumb based on the complete cases for all estimators. A larger bandwidth may improve the weighted estimator with a cutoff value of 0.01. But this example shows how unstable the weighted estimator is when the probability of response varies as much as it does here. The weighted estimator also has a much larger standard error. Again this is particularly bad in this example due to the low probability of response. The imputation estimator works well even though most of the data is missing. Compared

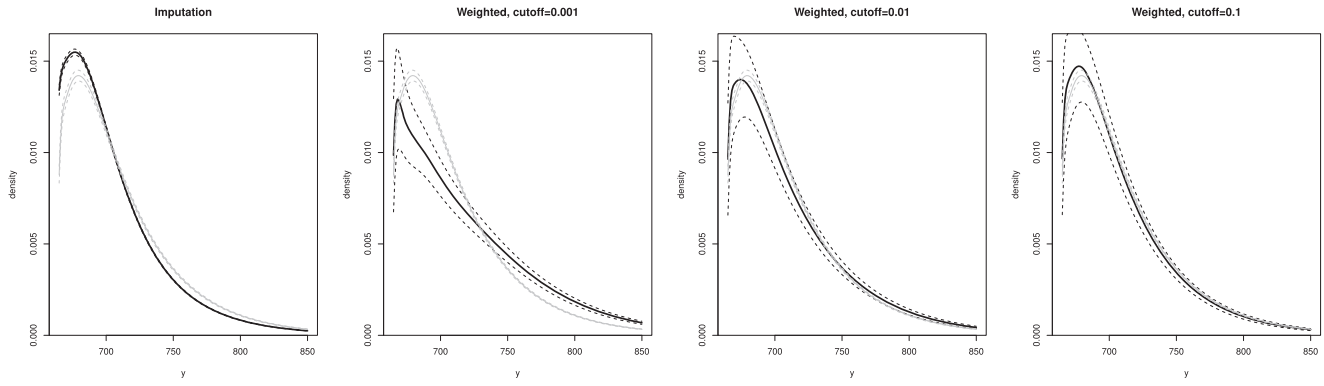

Figure 3. Estimated FICO score density. Grey curves are complete case estimates, dashed curves are asymptotic pointwise confidence bands.

to the complete case estimator, the imputation estimator moves some of the probability mass to lower values of the FICO score. It seems that low FICO scores possibly could play a smaller role in the decision of whether to grant a loan or not.

6. Higher order expressions for the asymptotic variance

Due to the faster rate of convergence, the estimation uncertainty in the estimate of β may be ignored asymptotically in the density estimator. However, as the simulations indicate, it is not ignorable in small samples where $1/n$ may not be negligible compared to $1/(nh)$. In this section I will derive an expression for the asymptotic variance which includes terms of order up to $1/n$.

Including terms of order $1/n$, the variance of the complete data estimator (5) can be written as

$$\operatorname{Var}\left[\hat{f}_1(y)\right] = \frac{f_1(y)}{pnh}\int K^2 - \frac{f_1'(y)}{pn}\int uK^2(u)du - \frac{f_1(y)^2}{pn} + o(1/n). \tag{10}$$

With a symmetric kernel K the second term of this expression is 0, and the asymptotic variance of the estimator may be written as

$$\frac{f_1(y)}{pnh}\int K^2 \cdot \left(1 - \frac{hf_1(y)}{\int K^2}\right). \tag{11}$$

If h is larger than $\int K^2/f_1(y)$, this expression is negative, and in this case the $o(1/n)$-term in (10) is not negligible. The expression for the asymptotic variance given in (11) is clearly smaller than the usual one. For instance, with f_1 equal to the standard normal density, a standard normal kernel, and $h = 0.367$ (corresponding to using Silverman's rule of thumb with $np = 200$), (11) is less than half the usual expression when $y = 0$. An unbiased estimator of the variance of a kernel density estimator (without missing data and with a nonrandom bandwidth) is easy to obtain: The kernel estimator is an average of iid terms, so the empirical variance based on these terms divided by n is an unbiased estimate.

For our imputation estimators, I obtain the following results:

Theorem 5. Under the same assumptions as in Section 3 and the additional assumptions listed in A.3, the asymptotic variances may be written as follows:

$$\begin{aligned}
\operatorname{Var}[\hat{f}_{1,SI}(y)] &= \frac{f_1(y)}{pnh}\int K^2 - \frac{f_1(y)^2}{pn} \\
&\quad + \frac{1}{p^2n}A_1(y)\mathcal{I}(\beta)^{-1}A_1(y)^\top \\
&\quad + \frac{2}{p^2n\sqrt{h}}A_1(y)\mathcal{I}(\beta)^{-1}A_2(y) + o(1/n) \\
\operatorname{Var}[\hat{f}_{1,MI}(y)] &= \frac{1}{p^2nh}(v_1(y) + v_2(y)/B)f(y)\int K^2 \\
&\quad - \frac{1}{p^2n}2(v_1(y) + v_2(y)/B)f(y)f_1(y) \\
&\quad + \frac{1}{p^2n}E[v_1(Y_i) + v_2(Y_i)/B]f_1(y)^2 \\
&\quad + (1 + \frac{1}{B})\frac{1}{p^2n}A_1(y)\mathcal{I}(\beta)^{-1}A_1(y)^\top \\
&\quad + \frac{1}{B}\frac{2}{p^2n\sqrt{h}}A_1(y)\mathcal{I}(\beta)^{-1}A_2(y) + o(1/n) \\
\operatorname{Var}[\hat{f}_{1,FI}(y)] &= \frac{1}{p^2nh}v_1(y)f(y)\int K^2 \\
&\quad - \frac{1}{p^2n}2v_1(y)f(y)f_1(y) + \frac{1}{p^2n}E[v_1(Y_i)]f_1(y)^2 \\
&\quad + \frac{1}{p^2n}A_1(y)\mathcal{I}(\beta)^{-1}A_1(y)^\top + o(1/n) \\
\operatorname{Var}[\hat{f}_{1,RB}(y)] &= \frac{1}{p^2nh}E[p_1^2|Y_1 = y]f(y)\int K^2 \\
&\quad - \frac{2}{p^2n}E[p_1^2|Y_1 = y]f(y)f_1(y) \\
&\quad + \frac{1}{p^2n}E[p_1^2]f_1(y)^2 \\
&\quad + \frac{1}{p^2n}A(y)\mathcal{I}(\beta)^{-1}A(y)^\top + o(1/n)
\end{aligned}$$

where

$$\begin{aligned}
A_1(y) &= E[(1-\pi_1)p_1'Z_1^\top|Y_1 = y]f(y) - E[(1-\pi_1)p_1'Z_1^\top]f_1(y) \\
A_2(y) &= E[\pi_1 p_1'Z_1|Y_1 = y]f(y) - E[\pi_1 p_1'Z_1]f_1(y) \\
A(y) &= E[p_1'Z_1^\top|Y_1 = y]f(y) - E[p_1'Z_1^\top]f_1(y) \\
v_1(y) &= E[\pi_1 p_1 + (1-\pi_1)p_1^2|Y_1 = y] \\
v_2(y) &= E[(1-\pi_1)p_1(1-p_1)|Y_1 = y]
\end{aligned}$$

as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$.

An expression for the variance of the weighted estimator similar to those in the theorem above may be derived from Tang, He, and Gunzler (2012)'s Theorem 3. Figure 4 provides a comparison of these formulae and the simulation results from Section 4. It is clear that the higher order expressions are better able to capture the true variability of the estimators.

As with the formula (10), the expressions for the variances in Theorem 5 may give negative values for h sufficiently large. In the (somewhat uninteresting) case, where π is a constant, it is easy to see that the sum of the last two terms of the variance of the single imputation estimator is positive (unless h is large). Thus, if the missing data mechanism is missing completely at random, the (infeasible) complete data estimator is more efficient than the single imputation estimator. In other words, there is a price to be paid when estimating β. However, the expressions for the variances given in the theorem above confirm the findings from the simulations in the previous section: In the set-up from the simulations, where π depends on y, the variance of the single imputation estimator (ignoring $o(1/n)$-terms) is smaller than the variance of the complete data estimator for y close to 0.

The expressions for the asymptotic variances in Theorem 5 are difficult to compare to each other algebraically. It is however clear from the proof that the variance of the "fixed imputation" estimator is smaller than the variance of the single and the multiple imputation estimators. The single and the multiple imputation estimators are not easy to compare. Though more imputations (larger B) clearly decreases the asymptotic variance, the variance of the multiple imputation estimator is actually

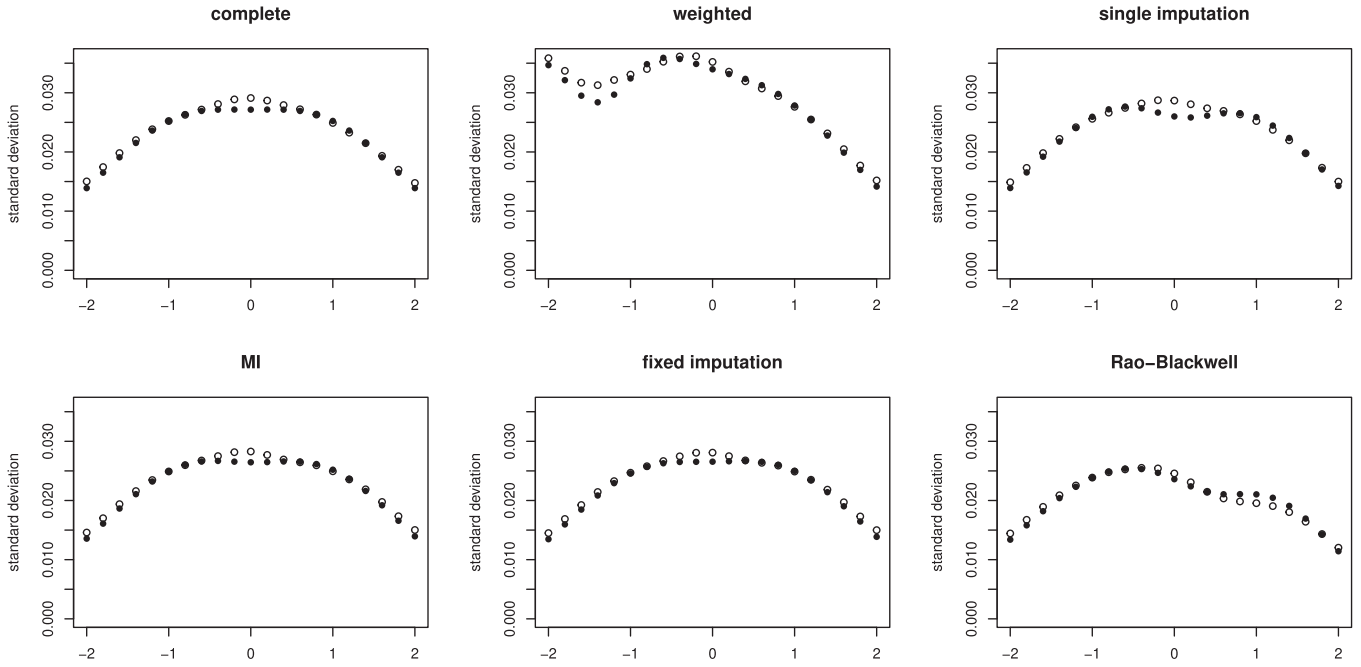

Figure 4. Standard deviations of density estimators. The open circles are the standard deviations from the simulations in Section 4, the filled circles are asymptotic standard deviations obtained from Theorem 5.

larger than the variance of the single imputation estimator, when $B = 1$. This is due to the extra randomness inherent in simulating from a Bayesian predictive distribution. As B increases, the asymptotic variance of the multiple imputation estimator approaches the variance of the fixed imputation estimator and in particular becomes smaller than the variance of the single imputation estimator. Comparing the variance of the "Rao-Blackwellized" estimator to the variance of the fixed imputation estimator, note that $E[p_1^2|Y_1 = y] \leq v_1(y)$. This implies that

$$\frac{1}{p^2nh}E[p_1^2|Y_1 = y]f(y)\int K^2 - \frac{2}{p^2n}E[p_1^2|Y_1 = y]f(y) + \frac{1}{p^2n}E[p_1^2]f_1(y)^2$$
$$\leq \frac{1}{p^2nh}v_1(y)f(f)\int K^2 - \frac{1}{p^2n}2v_1(y)f(f) + \frac{1}{p^2n}E[v_1(Y_1)]f_1(y)^2$$

for h sufficiently small ($h \leq \int K^2/(2f_1(y))$). However,

$$\frac{1}{p^2n}A(y)\mathcal{I}(\beta)^{-1}A(y)^\top \geq \frac{1}{p^2n}A_1(y)\mathcal{I}(\beta)^{-1}A_1(y)^\top$$

in the case where π_i is constant, and this will presumably be the case also when π_i is not constant, at least for some values of y. Hence, comparing these two estimators algebraically does not seem possible. Eventually, of course, as n increases and h decreases, the variance of $\hat{f}_{1,RB}(y)$ will be the smaller of the two, but I am not able to rule out that the fixed imputation may have smaller variance in small samples.

Even though it is possible to estimate the different terms in these expressions consistently using kernel regressions, so that these formulae may be used to get improved estimation of the variance of the imputation estimators, it seems too complicated for routine use. Moreover, even if such a plug-in estimator would be consistent, the finite sample bias in the kernel estimators will presumably bias the variance estimators to some degree. Bootstrapping may be a useful alternative way of obtaining accurate standard errors.

7. Conclusion

In this paper, I have considered four different imputation estimators for estimating the density in a subgroup when group membership is not known for every observation. The single random imputation estimator is attractive, because it is easily implemented and has the same asymptotic distribution as the complete data estimator. The "Rao-Blackwellized" estimator, where every subgroup indicator is replaced by its conditional mean, is more efficient but may require a little more work to implement and estimating its variance will be somewhat harder, especially if the estimated probability of belonging to the subgroup of interest depend on X_i as well as Y_i. The other two estimators—the multiple imputation estimator and the fixed imputation estimator—are more complicated to work with than the single imputation estimator and less efficient than the "Rao-Blackwellized" estimator. Thus for practical applications, either the single imputation or the Rao-Blackwellized estimator should be used. As shown, the asymptotic variance ignoring terms of smaller order than $1/(nh)$ may be quite misleading in finite samples, and I have provided more accurate formulae for the asymptotic variance. These formulae will require more work to implement in practice, where a simple bootstrap may be an easier option.

All the imputation estimators are more efficient than the weighted estimator considered by Tang, He, and Gunzler (2012). This is due to the weighting with the inverse of potentially very small probabilities, which inflates the variance. Robins, Rotnitzky, and Zhao (1995) show how the efficiency of inverse probability weighed estimators for semi-parametric regression

models may be improved, and this is probably possible also for the kernel density estimator considered here. It seems unlikely that one would be able to obtain better results using weighting than by using imputation in the situation considered in this paper, though. In general, a reason for preferring weighting to imputation is that it is easier to fit a binary regression model for the response mechanism than it is to fit an imputation model, but here the imputation model is also a binary regression, and as imputation leads to more efficient estimators there is little reason to prefer weighting.

The simplicity of the results are due to the fact that the faster rate of convergence of the parameters from the imputation model allows us to treat the imputation model as known. Even though a parametric binary regression model is a flexible tool, it would be of interest to see how the asymptotic results would change if we used a non-parametric or a semi-parametric binary regression model for the imputation model. I hope to address this in a future paper.

Disclosure statement

The author reports there are no competing interests to declare.

ORCID

Søren Feodor Nielsen http://orcid.org/0000-0002-9399-4918

References

Li Z, Hu X, Li K, Zhou F, Shen F. 2020. Inferring the outcomes of rejected loans: an application of semisupervised clustering. J R Stat Soc Ser A Stat Soc. 183:631–654.

Müller UU. 2009. Estimating linear functionals in nonlinear regression with responses missing at random. Ann Stat. 37:2245–2277.

Robins JM, Rotnitzky A, Zhao LP. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 90:106–121.

Rubin D. 1987. Multiple imputation for nonresponse in surveys. New York: Wiley.

Ruppert D. 1997. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. J Am Stat Assoc. 92:1049–1062.

Tang W, He H, Gunzler D. 2012. Kernel smoothing density estimation when group membership is subject to missing. J Stat Plan Inference. 142:685–694.

Appendix A: Proofs

A.1. Assumptions for the main results

Assumption A

The random variables (Y_i, D_i, R_i, X_i), $i = 1, \ldots, n$ are iid and the Y_i's are missing at random.

The conditional distribution of D_i given $X_i = x$ and $Y_i = y$ is from a generalized linear model with an inverse link function that has a bounded derivative.

Assumption B

The kernel function satisfies (4).

Assumptions C1–C6

C1: $f(y)$ is twice continuously differentiable with a second derivative that is locally Lipschitz.

C2: $f_1(y)$ is twice continuously differentiable with a second derivative that is locally Lipschitz.

C3: $p(y)$ is twice continuously differentiable with a second derivative that is locally Lipschitz.

C4: $E[p_i^2|Y_i = y]$ is twice continuously differentiable with a second derivative that is locally Lipschitz.

C5: $E[\pi_i p_i|Y_i = y]$ is twice continuously differentiable with a second derivative that is locally Lipschitz.

C6: $E[\pi_i p_i^2|Y_i = y]$ is twice continuously differentiable with a second derivative that is locally Lipschitz.

A.2. Asymptotic distribution

I first consider the centered single imputation estimator

$$\hat{f}_{1,SI}(y) - \left(f_1(y) + \frac{h^2}{2} f_1''(y)\right) = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{D}_i \left(K\left(\frac{y-Y_i}{h}\right)/h - f_1(y) - \frac{h^2}{2} f_1''(y)\right)}{\frac{1}{n}\sum_{i=1}^n \tilde{D}_i}.$$

The denominator converges in probability to p. To see this first note that

$$\frac{1}{n}\sum_{i=1}^n R_i D_i + (1 - R_i) 1_{\{U_i \le p_i\}} \to p \quad \text{in probability.}$$

Next, note that

$$E\left|\frac{1}{n}\sum_{i=1}^n \tilde{D}_i - \frac{1}{n}\sum_{i=1}^n R_i D_i + (1 - R_i) 1_{\{U_i \le p_i\}}\right| \le E\left[\frac{1}{n}\sum_{i=1}^n \left|1_{\{U_i \le \hat{p}_i\}} - 1_{\{U_i \le p_i\}}\right|\right] = E\left[\frac{2}{n}\sum_{i=1}^n |\hat{p}_i - p_i|\right].$$

This mean converges to 0 as the average is bounded and

$$\frac{1}{n}\sum_{i=1}^n |\hat{p}_i - p_i| \le \text{const}\, \frac{1}{n}\sum_{i=1}^n |Z_i| \cdot \left|\hat{\beta} - \beta\right|$$

which converges to 0 in probability. Similar arguments hold for the denominators of the other estimators. Thus the denominators may be treated as known.

To derive the asymptotic distribution I decompose the numerator as follows:

$$\frac{1}{n}\sum_{i=1}^n \tilde{D}_i \left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2} f_1''(y)\right)$$

$$= \frac{1}{n}\sum_{i=1}^n (R_i D_i + (1 - R_i)\hat{p}_i) \left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2} f_1''(y)\right) \quad (12)$$

$$+ \frac{1}{n}\sum_{i=1}^n (1 - R_i)(1_{\{U_i \le \hat{p}_i\}} - \hat{p}_i) \left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2} f_1''(y)\right) \quad (13)$$

The second term (13) has mean 0 and variance

$$\frac{1}{n^2}\sum_{i=1}^n (1 - R_i) \left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2} f_1''(y)\right)^2 \hat{p}_i(1 - \hat{p}_i)$$

conditional on the observed data. Moreover, as

$$\left|\hat{p}_i(1-\hat{p}_i) - p_i(1-p_i)\right| \leq \text{const}\ |Z_i| \cdot \left|\hat{\beta}-\beta\right| = |Z_i|\ O_P(1/\sqrt{n})$$

it follows that unconditionally

$$\frac{h}{n}\sum_{i=1}^{n}(1-R_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)^2 \hat{p}_i(1-\hat{p}_i) \xrightarrow{P} E[(1-\pi_i)p_i(1-p_i)|Y_i=y]f(y)\int K^2.$$

It follows by the central limit theorem for triangular arrays that given the observed data

$$\sqrt{nh}\,\frac{1}{n}\sum_{i=1}^{n}(1-R_i)(1_{\{U_i\leq\hat{p}_i\}}-\hat{p}_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right) \xrightarrow{\mathcal{D}} N\left(0, E[(1-\pi_1)p_1(1-p_1)|Y_1=y]f(y)\int K^2\right) \tag{14}$$

in probability. As the limiting distribution does not depend on the observed data, the convergence is also valid unconditionally, and (13) is asymptotically independent of (12).

Now turn to the term (12), which is decomposed as follows:

$$\begin{aligned}
&\sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(R_iD_i+(1-R_i)\hat{p}_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\\
&\quad=\sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(R_iD_i+(1-R_i)p_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\\
&\qquad+\sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(1-R_i)(\hat{p}_i-p_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right).
\end{aligned}$$

Here the second term is $o_P(1)$ as it numerically bounded by a constant times

$$\begin{aligned}
&\sqrt{h}\frac{1}{n}\sum_{i=1}^{n}|Z_i|\cdot\left|K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right|\cdot|\sqrt{n}(\hat{\beta}-\beta)|\\
&\quad= O_P(\sqrt{h}).
\end{aligned}$$

Ignoring the remainder term, it follows that (12) has mean 0 and a variance that is

$$E[\pi_1p_1+(1-\pi_1)p_1^2|Y_1=y]f(y)\int K^2 + o(1).$$

Thus, I obtain

$$\sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(R_iD_i+(1-R_i)\hat{p}_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right) \xrightarrow{\mathcal{D}} N\left(0, E[\pi_1p_1+(1-\pi_1)p_1^2|Y_1=y]f(y)\int K^2\right).$$

Combining this with (14) I obtain Theorem 1.

The other imputation estimators are handled similarly. The "fixed imputation estimator" in Section 3.3 corresponds to ignoring (13), whereas the estimators in Section 3.4 are obtained by letting $R_i = 0$ in (12) and omitting (13). The multiple imputation estimator is slightly more complicated. The U_i should be replaced by $U_{i,j}$'s, $\hat{p}_i$ should be replaced by $\hat{p}_{i,j}$ in (13) (but not in (12)) and a new term

$$\frac{1}{n}\sum_{i=1}^{n}(1-R_i)(\hat{p}_{i,j}-\hat{p}_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right) \tag{15}$$

should be included between with (12) and (13). The term (13) may be handled as in the single imputation case by conditioning on $\tilde{\beta}_j$ as well as the observed data; the term (12) is unchanged. The new term (15) is handled conditionally on the observed data:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(1-R_i)(\tilde{p}_{i,j}-\hat{p}_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}(1-R_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\tilde{p}'_{i,j}Z_i^\top(\tilde{\beta}_j-\hat{\beta})
\end{aligned}$$

where $\hat{p}'_{i,j}$ is the derivative of the inverse link function at a point between $Z_i^\top\tilde{\beta}_j$ and $Z_i^\top\hat{\beta}$. As $\tilde{\beta}_j-\hat{\beta} = O_P(1/\sqrt{n})$ the term (15) is $o_P(1/\sqrt{nh})$ and may be ignored asymptotically. Thus, the multiple imputation estimator is asymptotically equivalent to the sum of (12) and an average of B independent copies of (13) divided by p.

A.3. Variances

For Theorem 5 I need a few additional assumptions: I will assume that the derivative of the inverse link function in the model for $p(x,y)$ is Lipschitz continuous and that $E[p_i''Z_i|Y_i=y]$ and $E[\pi_ip_i''Z_i|Y_i=y]$ are twice continuously differentiable (as functions of y) with a locally Lipschitz continuous second derivative.

Focusing first on the single imputation estimator, the variance of (13) may be written as

$$\frac{1}{nh}\nu_2(y)f(y)\int K^2 - \frac{2}{n}\nu_2(y)f(y)f_1(y) + \frac{1}{n}E[\nu_2(Y_1)]f_1(y)^2 + o(1/n)$$

with $\nu_2(y)$ given in Theorem 5.

A little more work is required to obtain a higher order expression for the variance of (12), as the effect of $\hat{p}_i - p_i$ which is $O_P(1/\sqrt{n})$ has to be incorporated. Note, however, that (13) and (12) are uncorrelated. Taylor expanding p_i as a function of the linear predictor, we get

$$\hat{p}_i = p_i + p_i'Z_i^\top(\hat{\beta}-\beta) + (\hat{p}_i'-p_i')Z_i^\top(\hat{\beta}-\beta)$$

where $\hat{p}_i'$ is the derivative of the inverse link function at a point between $Z_i^\top\hat{\beta}$ and $Z_i^\top\beta$. Moreover, as

$$|\hat{p}_i'-p_i'| \leq \text{const}\,|Z_i^\top\hat{\beta}-Z_i^\top\beta| \leq \text{const}\,|Z_i|\cdot|\hat{\beta}-\beta|$$

I get

$$\hat{p}_i - p_i = p_i'Z_i^\top(\hat{\beta}-\beta) + o_p(1/\sqrt{n}) = p_i'Z_i^\top\frac{1}{n}\sum_{j=1}^{n}S_j + o_p(1/\sqrt{n}).$$

Hence, I obtain

$$\begin{aligned}
&\sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(R_iD_i+(1-R_i)\hat{p}_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\\
&\quad=\sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(R_iD_i+(1-R_i)p_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\\
&\qquad+\sqrt{h}A_1(y)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S_i + o_P(1)
\end{aligned}$$

with $A_1(y)$ given in Theorem 5, as

$$\frac{1}{n}\sum_{i=1}^{n}(1-R_i)\left(K(\tfrac{y-Y_i}{h})/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\hat{p}_i'Z_i^\top \xrightarrow{P} A_1(y).$$

It now follows that the asymptotic variance of (12) may be written as

$$\frac{1}{h}v_1(y)f(f)\int K^2 - 2v_1(y)f(y)f_1(y) + E[v_1(Y_i)]f_1(y)^2 + A_1(y)\mathcal{I}(\beta)^{-1}A_1(y)^\top + 2A_1(y)\mathcal{I}(\beta)^{-1}A_2(y)/\sqrt{h} + o(1/n)$$

with $v_1(y)$ and $A_2(y)$ given in Theorem 5 as

$$\begin{aligned}
&E\left[(R_iD_i + (1-R_i)p_i)\left(K\left(\tfrac{y-Y_i}{h}\right)/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)S_i\right]\\
=&E\left[D_iS_i\left(K\left(\tfrac{y-Y_i}{h}\right)/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\right]\\
=&E\left[E[D_iS_i|Y_i]\left(K\left(\tfrac{y-Y_i}{h}\right)/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\right]\\
=&\mathcal{I}(\beta)^{-1}A_2(y)
\end{aligned}$$

since

$$\begin{aligned}
E[D_iS_i|Y_i] &= E\left[R_iD_i\frac{(1-p_i)}{p_i(1-p_i)}p_i'\mathcal{I}(\beta)^{-1}Z_i|Y_i\right]\\
&= \mathcal{I}(\beta)^{-1}E[\pi_ip_i'Z_i|Y_i].
\end{aligned}$$

Noting that $v_2(y) + v_1(y) = p(y)$ gives the desired result.

As in the proofs of the asymptotic results, the expressions for the asymptotic variances of $\hat{f}_{1,FI}(y)$ and $\hat{f}_{1,RB}(y)$ follows easily by ignoring the expression coming from (13) and, in the case of $\hat{f}_{1,RB}(y)$, by formally putting $R_i = 0$ and noting that the "covariance term" vanishes as

$$E[p_iS_i|Y_i] = E\left[R_i\frac{D_i - p_i}{1-p_i}p_i'\mathcal{I}(\beta)^{-1}Z_i|Y_i\right] = 0.$$

Turning to the multiple imputation estimator, the term corresponding to (13) contribute

$$\frac{1}{B}\left(\frac{1}{nh}v_2(y)f(f)\int K^2 - \frac{2}{n}v_2(y)f(y)f_1(y) + \frac{1}{n}E[v_2(Y_i)]f_1(y)^2\right) + o(1/n)$$

to the variance of the average, whereas the contribution from the term (12) is unchanged. The term (15) may be written as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(1-R_i)\left(K\left(\tfrac{y-Y_i}{h}\right)/h - f_1(y) - \tfrac{h^2}{2}f_1''(y)\right)\hat{p}_{i,j}'Z_i^\top(\tilde{\beta}_{i,j} - \hat{\beta})\\
&\qquad = A_1(y)(\tilde{\beta}_{i,j} - \hat{\beta}) + o_P(1/\sqrt{n}).
\end{aligned}$$

As $E[\hat{p}_{i,j} - \hat{p}_i|\text{data}] = O_P(1)$, the correlation between this term and (12) is negligible. Thus the contribution of this term to the variance is $A_1(y)\mathcal{I}(\beta)A_1(y)^\top$.