Re-analysis using Inverse Probability Weighting and Multiple Imputation of Data from the Southampton Women's Survey

Shaun Seaman and Ian White

July 2008

MRC Biostatistics Unit
Institute of Public Health
Forvie Site
Robinson Way
Cambridge CB2 0SR
United Kingdom

shaun.seaman@mrc-bsu.cam.ac.uk

Abstract

In this document I describe a re-analysis of data from the Southampton Women's Survey (SWS). The original complete-case analysis implicitly assumed that the data are missing completely at random. Inverse-probability weighting (IPW) and multiple imputation (MI) are more sophisticated methods for handling missing data, which make the weaker assumption that the data are missing at random. We sought to determine whether these methods changed the conclusions of the original analysis. We describe IPW and MI, give the results of applying them to the SWS and, where the estimates differ from the original estimates, provide intuition as to why these differences arise. We find that, assuming the missing data are missing at random, the conclusions of the original complete-case analysis remain unchanged. Thus, confidence in these original conclusions is strengthened. We hope this document may prove helpful to researchers wishing to use IPW or MI on their own data.

1 Introduction

In this document we reanalyse data described in Crozier et al. (2009). In Section 2 the patterns of missingness in the SWS will be described. Section 3 describes the analyses that we have performed. In Section 4 the IPW method will be briefly described and we shall explain how we implemented it for the SWS data. Section 5 does the same thing for MI. In Section 6 the results from IPW and MI will be compared with the original results of Crozier et al. Section 7 contains a summary of the findings.

2 Data Description

A total of 1987 women in the SWS have been pregnant and delivered by the end of 2003. These form our sample. All women were interviewed before pregnancy (the PP visit). It was planned that women also be interviewed in early pregnancy (EP visit, at approximately 11 weeks gestation) and in late pregnancy (LP visit, at approximately 34 weeks gestation). Table 1 shows the numbers of women who actually were interviewed at these times. 1553 women were interviewed at EP, 1893 at LP, and 1490 at both EP and LP. Of the 444 women not interviewed at EP, 409 were asked later (at either 19 weeks or at LP) about their smoking habits and degree of nausea at the time of EP (although not about other behaviour at EP).

Note that if the 63 women with EP data but no LP data are excluded, the missing-data pattern becomes monotone: all 1987 women have PP data, 1893 of them have LP data and, of those 1893, 1490 have EP data. This fact will be used in the IPW analysis.

Note that of the women who attended the EP visit, 96% also attended the LP visit, whereas of the women who missed the EP visit, only 93% attended the LP visit ($p = 0.01$ for the difference).

Most variables are continuous or binary, but some are ordinal or nominal categorical. The three nausea variables (at PP, EP and LP) and awexam are ordinal, but will be treated as continuous. awethnc is nominal, but the derived binary variable, nonwhite, will be used in its place, as many of the categories of awethnc are very sparse.

Table 2 shows the variables available. Table 3 shows, for each variable measured at PP, EP and LP, the number of women with missing data on that variable. In addition, of the 409 women who missed their EP visit but who were asked later about some of their behaviour at the time of EP, 408 had data on dnausea and 396 on dcrsmok. Therefore, these two variables were observed for a total of 1960 and 1947 women, respectively.

It can be seen that the main missing-data problem is missing units (women) rather than missing items, and that the greatest missingness is at EP.

	LP		
EP	Yes	No	
Yes	1490	63	1553
at 19wks	286	21	307
at LP	101	1	102
No	16	9	25
	1893	94	1987

Table 1: Numbers of women interviewed at EP and LP. 409 women were asked later (307 at 19 weeks and 102 at LP) about smoking and nausea at EP.

Variable name	Meaning
General variables	
jwage	Age of mother at conception (years)
zawbmi	Fisher-yates transformed BMI
awethnc	Ethnicity
awexam	Qualification level (1=none, ..., 6=degree)
Variables relating to PP	
agest	Gestation, i.e. date of PP minus date of conception (in weeks)
aalcunit	Units of alcohol per wk
aalcg2	Indicator of > 4 units alcohol per wk
acrsmok	Indicator of smoking at EP
acaffmg	mg of caffeine per day
acaffmgg2	Indicator of > 300mg caffeine per day
afruitveg	Portions of fruit and veg per day
afruitvegg2	Indicator of < 5 fruit and veg per day
Variables relating to EP	
dgest, etc	Analogous to agest, aalcunit, aalcg2, etc
dnausea	Degree of nausea at PP (0=none, ..., 3=severe)
Variables relating to LP	
hgest, etc	Analogous to agest, aalcunit, aalcg2, etc
hnausea	Degree of nausea at LP (0=none, ..., 3=severe)
Extra variables derived by me	
nonwhite	Indicator of being non-white (derived from awethnc)
acaffmgr3	Cubic root of acaffmg
acaffmg0	Indicator of 'zero' caffeine (< 10mg per day)
aalcunitlog	Log of aalcunit
aalcunit0	Indicator of 'zero' alcohol (< 0.1 units per day)
afruitvegr2	Square root of afruitveg, dfruitveg and hfruitveg
dcaffmgr3, etc	Analogous to acaffmgr3, acaffmg0, aalcunitlog, etc.
hcaffmgr3, etc	Analogous to acaffmgr3, acaffmg0, aalcunitlog, etc.

Table 2: List of variables

Variable	No. Missing
1987 women at PP	
agest	25
zawbmi	9
awexam	5
jwage	24
7 others*	1
1553 women at EP	
dgest	15
dcrsmok	2
3 others*	1
1893 women at LP	
hgest	11

Table 3: Numbers of women with missing data on each variable.
* One woman has missing data on seven variables at EP, but complete data on agest, zawbmi, awexam and jwage. A different woman has missing data on dgest, dcrsmok, dalcunit (and so dalcg2) and dnausea.

3 Analyses

Crozier et al. performed a complete-case analysis. That is, they analysed just the 1490 women with data at PP, EP and LP.[1] They used backward stepwise elimination to select the preferred models. So, 497 women (25%) were excluded, as well as women with a missing value for any of the variables included in the selected model.

We have repeated their analysis (in Section 6 this analysis is referred to as 'Comp') and also performed the following analyses:

- The same analysis as Crozier et al. apart from one slight change. jwage (age in years at conception) is missing for 24 women because, although awage (age at PP) is observed, agest (time between PP and conception) is missing, and jawage is calculated from awage and agest. The range of agest is only about 200 weeks, which is small compared to the range of awage. So, we imputed the 24 missing jwage values using the mean value of agest. These 24 imputed values were also retained for the complete-case and IPW analyses that follow. Results from this analysis were, unsurprisingly, very similar to those of Crozier et al. and we do not present them here.

- An analysis using all possible women, rather than just the 1490 with data at each visit. This will use 1987 women for regressions of outcomes at PP, 1553 for outcomes at EP and 1893 for outcomes at LP (minus women with missing values on any variables included in the selected model). In Section 6.2 this

[1] Strictly speaking, this is not a 'complete-case' analysis, because women who did attend all three visits but who nevertheless have missing value(s) for at least one variable needed in some of the regressions will be included in regressions that do not need those variables. A true 'complete-case' analysis would exclude these women.

analysis is referred to as 'All'. It is an 'available-cases' analysis.[2]

- For EP and LP only, an IPW analysis. This uses the same women as the preceding complete-case analysis, but weights them according to the inverse of their probability of attending the EP/LP visit. In Section 6.2 this analysis is referred to as 'IPW' or 'IPW2'.
- An MI analysis with imputation performed using chained equations. This uses all 1987 women and is referred to as 'MI' in Section 6.2.

4 IPW Method

For simplicity, the account below ignores the fact that women can have missing items for visits that they did attend. Instead, it is assumed that, if a woman attends a particular visit, she will have no missing values on variables measured at that visit. As we saw in Section 2, this is mostly true: the majority of the missingness is due to missing visits.

The validity (i.e. unbiasedness of estimated RRs and estimated SEs) of any of the original complete-case regressions relies on a missing completely at random (MCAR) assumption. This is the assumption that the probability that a woman attends both the EP and the LP visits (rather than neither or one of them) is independent of (her values on) all the variables in the regression model.[3] If the marginal mean model implicit in this quasi-likelihood (i.e. the log-linear model for risk) is correct, the MCAR assumption can be replaced by a slightly weaker assumption: that the probability of a woman attending both visits is conditionally independent of the outcome variable in the Poisson regression model given the covariates in that regression model.

Table 4 shows a comparison of the variables measured at PP (and dcrsmok and dnausea) in two sets of women: those who made both EP and LP visits ('included'), and those who did not ('excluded'). Also shown is a comparison of the variables measured at LP in two sets of women: those who made both EP and LP visits ('included'), and those who made only the LP visit ('excluded'). As can be seen from the table, there are significant differences between the included and excluded women, indicating that the MCAR assumption is invalid.

In general, MCAR is quite a strong assumption The missing at random (MAR) assumption is weaker. It assumes that, conditional on the some set of variables that are fully observed, the probability that a woman attends a visit (the EP visit, for analyses of outcomes at EP; the LP visit, for analyses of outcomes at LP) is independent of all the variables in the Poisson regression model. If the MAR assumption is true (and the logistic regression model for the probability of missingness, described below, is correctly specified), the IPW analysis will be valid. Thus, the IPW analysis relies on weaker assumptions than the original analysis of Crozier et al.[4]

[2] In fact, for the outcome dcrsmok this analysis does not use all the available cases and so is not a true 'available-cases' analysis (see Section 6).

[3] In this analysis the Poisson regression, used with robust estimation, is actually a maximum quasi-likelihood approach.

[4] It should be noted that, although usually the MCAR assumption is nested within the MAR

Variable name	Mean/median/% in Included	Excluded	p-value
jwage	30.1	29.8	0.070
zawbmi	0.000	−0.030	0.511
nonwhite	3.8%	10.3%	0.000
awexam	4	3	0.000
agest	−62.7	−71.2	0.003
aalcunit	4.69	3.89	0.010
aalcg2	53.8%	49.0%	0.067
acrsmok	26.6%	32.9%	0.008
acaffmg	246	217	0.068
acaffmgg2	38.9%	35.5%	0.180
afruitveg	5.49	5.49	0.985
afruitvegg2	46.6%	44.6%	0.434
dcrsmok	14.6%	22.9%	0.000
dnausea	1	1	0.137
hgest	34.7	34.5	0.006
halcunit	0.25	0.25	0.903
halcg2	5.1%	4.7%	0.753
hcrsmok	14.0%	22.3%	0.000
hcaffmg	146	153	0.225
hcaffmgg2	19.3%	20.8%	0.478
hfruitveg	5.70	5.94	0.087
hfruitvegg2	43.7%	40.0%	0.178
hnausea	0	0	0.773

Table 4: Differences between means or medians of variables in included and excluded women. P-values calculated from t-tests, chi-square tests or Mann-Whitney tests, as appropriate

The IPW method works as follows. Let R_{EP} and R_{LP} be indicator variables of whether a woman attends the EP and LP visits, respectively. Let Y_{PP}, Y_{EP} and Y_{LP} be the vectors of variables measured on a woman at PP, EP and LP, respectively, and let X be the subvector of Y_{EP} made up of dcrsmok and dnausea only. The variable R_{EP} is regressed (using logistic regression) on a number of explanatory variables that are fully observed. For example, older women may be more likely to attend than younger women, or vice versa, so jwage would be included as an explanatory variable. The fitted model gives a predicted probability for each woman that a woman with her characteristics (i.e. values on the explanatory variables) would attend the EP visit. If, for example, the predicted probability is 0.5 for a particular woman who did attend the visit, one can suppose that, on average, there will be another identical woman who did not attend the visit. The woman who attended the visit can then be given weight 2 in the Poisson regression of outcome at EP, in order that she 'represent' both herself and the hypothetical identical woman who did not attend the EP visit. Of course, we have not really observed two women, so this must be accounted for by using a robust estimator of the SEs.[5] In general, each woman who does attend the EP visit is given weight p^{-1}, where p is her fitted probability of attending from the logistic regression. In STATA the weights used are *probability weights*.

Weights are constructed in the same way for the LP visit.

The explanatory variables used in the logistic regressions were Y_{PP} and X, i.e. all variables measured at PP, and dcrsmok and dnausea. These variables were observed for all women, apart from some missing items on a few women (see the final paragraph of this section). Backward stepwise elimination was used, with a significance threshold of $p = 0.2$, to remove non-significant variables from the model. awexam, dnausea and hnausea were treated as categorical variables. Alcohol and caffeine consumption were included after log transformation, and binary indicator variables for zero alcohol and zero caffeine consumption were also included.

This IPW analysis for outcomes at EP is based on the MAR assumption that

$$p(R_{\text{EP}} \mid Y_{\text{PP}}, Y_{\text{EP}}, Y_{\text{LP}}) = p(R_{\text{EP}} \mid Y_{\text{PP}}, X)$$

and the IPW analysis for outcomes at LP is based on the MAR assumption that

$$p(R_{\text{LP}} \mid Y_{\text{PP}}, Y_{\text{EP}}, Y_{\text{LP}}) = p(R_{\text{LP}} \mid Y_{\text{PP}}, X). \tag{1}$$

For the EP visit only, a second IPW analysis was also performed, using a different set of weights. In this, a monotone missingness pattern was forced on the data by

assumption, this is not true in this case. This is because this MCAR assumption is about the probability that a woman attends both visits, whereas the MAR assumption is about the probabilty of attending a particular visit. Thus, the MCAR assumption could theoretically be valid without the MAR assumption being true.

[5]The SEs should also take into account the uncertainty in the inverse-probability weights, given that these are estimated rather than known, but this is often not done and, in particular, is not done by STATA's function 'poisson' with the option 'robust', which we have used here. On page 53 of Little and Rubin's 'Statistical Analysis with Missing Data' (2002), they say that the practical impact on the SEs from ignoring this source of variability is unclear.

excluding the 63 women with data at EP but not LP. So, weights are now calculated from the following equation:

$$p(R_{\mathrm{EP}} \mid Y_{\mathrm{PP}}, Y_{\mathrm{EP}}, Y_{\mathrm{LP}}) = p(R_{\mathrm{LP}} \mid Y_{\mathrm{PP}}, Y_{\mathrm{EP}}, Y_{\mathrm{LP}}) \times p(R_{\mathrm{EP}} \mid R_{\mathrm{LP}}, Y_{\mathrm{PP}}, Y_{\mathrm{EP}}, Y_{\mathrm{LP}}). \quad (2)$$

The first probability on the RHS of (2) was calculated using the same logistic regression model as in the first IPW analysis. This model contains variables measured at PP and also dcrsmok and dnausea. The second probability on the RHS of (2) was calculated from another logistic regression model, for the conditional probability that a women attends the EP visit given that she attends the LP visit. This logistic regression model contains not only variables measured at PP, and dcrsmok and dnausea, but also variables measured at LP. So, it is assumed in this second IPW analysis that equation (1) is true and that

$$p(R_{\mathrm{EP}} \mid Y_{\mathrm{PP}}, Y_{\mathrm{EP}}, Y_{\mathrm{LP}}) = p(R_{\mathrm{EP}} \mid Y_{\mathrm{PP}}, Y_{\mathrm{LP}}, X) \quad (3)$$

Although it may appear that this second IPW analysis has an advantage over the first because the probability of attending the EP visit is allowed to depend on variables measured at PP and LP, rather than just those measured at PP, it does require the additional assumption that missingness at LP depends only on variables measured at PP (and dcrsmok and dnausea). The first IPW analysis of outcomes at EP makes no assumptions about missingness at LP. On the other hand, the number of missing LP visits is considerably smaller than the number of missing EP visits, so the gain achieved by allowing missingness at EP to depend on variables measured at LP may outweigh the loss caused by requiring the additional assumption about missingness at LP. A further disadvantage of this second IPW analysis is that it requires that EP data on 63 women be excluded, in order to make the missingness pattern monotone.

For the calculation of weights only, several missing values were imputed (as being equal to the mean value of the variable) in variables measured at the PP and LP visits in women who did attend those visits. For example, the 25 missing agest values were imputed as the mean value of agest. The 23 missing dnausea and 40 missing dcrsmok values were also imputed as their corresponding means. This was done for the fitting of the logistic regression models and the calculation of weights only; these imputed values were not used in the Poisson regressions.

5 MI

MI produces complete sets of data by filling in all missing values. Each dataset thus produced consists of 1987 women with complete data at all three visits. Using iterative chained equations (the 'ICE' command in STATA), we generated 100 completed datasets. We then analysed them separately and then combined the results in the standard way (using 'Rubin's Rules'). This is done automatically in STATA by using the 'mim' functions.

MI has the potential to reduce SEs, because it uses all the observed data. It will be valid if the MAR assumption described in the next paragraph is valid.

Let $Y = (Y_{\text{PP}}, Y_{\text{EP}}, Y_{\text{LP}})$ be the vector of all variables measured on a woman. Let R denote the missingness pattern for a woman. R is a vector of the same length as Y and each element of R is an indicator variable showing whether or not the corresponding element of Y is observed for that woman. Let Y_{obs} denote the subvector of Y made up of variables whose values are observed for the woman, and let Y_{miss} denote the subvector of Y made up of variables whose values are missing for the woman. So, $Y = (Y_{\text{obs}}, Y_{\text{miss}})$. The MAR assumption is that

$$p(R \mid Y_{\text{obs}}, Y_{\text{miss}}) = p(R \mid Y_{\text{obs}}).$$

for every woman. That is, for each woman, the probability of each *possible* missingness pattern is independent of her *actual* missing data given her *actual* observed data.

To understand better the implications of this assumption for the SWS data, let us pretend for a moment that the missingness is due purely to missed visits (which is indeed almost true). So, a woman attending a particular visit, for example the LP visit, has data on all variables measured at that visit. Let us also pretend that the missingness is monotone, i.e. no woman attends the EP visit without also attending the LP visit, and that dcrsmok and dnausea are observed on all women (again, this is almost true). Finally, assume that all women attend the PP visit (which is true). The MAR assumption now becomes equivalent to equations (1) and (3), the same assumptions made in the second IPW analysis of outcomes at EP.

The MI analysis would be expected to be more efficient than the second IPW analysis, because it does not discard 63 women in order to make the missingness monotone and because it does not discard women with missing values in variables in a given Poisson regression model. It also makes use of observed dcrsmok and dnausea values in the women who miss the EP visit, variables which may, through correlations with other variables, provide information about their missing values.

Before beginning MI, we set dgest equal to missing for the women whose EP data were collected at 19 weeks or LP (for these women dgest is recorded as approximately 19 or 34), in order that dgest can represent the time that their EP visit *would* have taken place had they attended it (i.e. around 11 weeks).

The command we used in STATA was

```
ice agest aalcunitlog acrsmok acaffmgr3 acaffmg0 afruitvegr2 dgest
dalcunitlog dcrsmok dcaffmgg2 dfruitvegr2 hgest halcunitlog hcrsmok
hcaffmgr3 hcaffmg0 hfruitvegr2 jwage awhtcm zawbmi nonwhite awexam
dnausea hnausea, m(100) cmd(awexam dnausea hnausea: ologit)
conditional(acaffmgr3:acaffmg0 @0, hcaffmgr3:hcaffmg0 @0)
match(agest aalcunitlog dalcunitlog halcunitlog) genmiss(M_)
seed(20) saving(completed.dta)
```

acaffmgr3 and hcaffmgr3 are the cubic roots of acaffmg and hcaffmg. aalcunitlog, dalcunitlog and halcunitlog are the log transforms of aalcunit, dalcunit and halcunit. afruitvegr2, dfruitvegr2 and hfruitvegr2 are the square roots of afruitveg, dfruitveg and hfruitveg. These three transforms (cubic root, log and square root) are the Box-Cox transformations after excluding zero values.

The ordinal variables awexam, dnausea and hnausea were imputed by ordinal logistic regression, but were treated as continuous variables for the imputation of other variables.

The Poisson regression models involve the binary variables for alcohol, caffeine and fruit & veg consumption (e.g. aalcg2, acaffmgg2 and afruitvegg2). So, one could just impute these binary variables directly. However, in order to use as much information as possible for the imputation of other variables, we imputed the continuous variables instead (e.g. dalcunitlog, dcaffmgr3 and dfruitvegr2). The values of the binary variables were derived afterwards from these continuous variables.

The 'success' of the MI was checked by comparing, for each variable, the distribution of the imputed values with the corresponding distribution of observed values. If these are similar, this provide some reassurance that the imputation has worked OK. If they are different, an explanation for this difference needs to be found. As the continuous variables for alcohol, caffeine and fruitveg are used to calculate the binary variables, it is particularly important to check the distributions of the imputed binary variables, and not just the distributions of the imputed continuous variables. If the distributions of the imputed values of these binary variables seem to be different from the corresponding distributions of observed values, it might be safer to use prediction matching or impute the binary variables directly.

After comparing the distributions of the imputed dfruitvegg2 and hfruitvegg2 values with the observed values, it was decided that prediction matching was probably not required for dfruitvegr2 and hfruitvegr2: for both the continuous and derived binary variables the distributions of imputed and observed values were similar. An attempt was made to treat the caffeine and alcohol variables (acaffmgr3 and aalcunitlog, etc) as semi-continuous variables and use 'conditional imputation'. In this approach it is assumed that, conditional on the values of the other variables in the dataset, the distribution of cube-root caffeine (or log alcohol) consumption is a mixture of a normal distribution and a point mass at zero. The mass associated with the zero and the mean of the normal distribution depend on the values of the other variables. This method ensures that only non-negative values are generated and takes into account the large number of zeroes. This appeared to work well for hcaffmgr3. However, it was found to lead to some rather extreme imputed values for the dalcunitlog and halcunitlog and a slightly longer tail in the distribution of imputed dcaffmggr3 values than in the distribution of observed values, which in turn led to different distributions of the binary variables dalcg2, halcg2 and dcaffmgg2 between the imputed and observed values. For this reason, prediction matching was used instead for the dalcunitlog and halcunitlog. Prediction matching was also used for agest, as no satisfactory transformation to normality could be found.

When prediction matching was used for dcaffmgr3 there was still a difference between the distributions of dcaffmgg2 in the imputed and observed values. 16.3% of observed dcaffmgg2 values were equal to one, versus 21.2% of imputed values (based on five imputed datasets). This compared with 22.1% of imputed values when dcaffmgr3 was treated as semi-continuous and 'conditional imputation' used. We were a bit worried about this, as we could see no reason why the difference should arise. In the set of women whose dcaffmgg2 values were imputed acaffmgg2 was ac-

Variable	OR	p-value
dnausea=1	3.92	
dnausea=2	2.67	
dnausea=3	3.82	0.041
nonwhite	2.10	0.040
jwage	0.92	0.006
awexam=2	0.29	
awexam=3	0.41	
awexam=4	0.31	
awexam=5	0.28	
awexam=6	0.30	0.117

Table 5: Predictors of missingness at the LP visit

tually less likely to equal one than in the set of women whose dcaffmgg2 values were observed (35.5% versus 38.9%); although hcaffmgg2 was slightly more likely to equal one (20.8% versus 19.3%). Next we tried imputing the binary variable dcaffmgg2 directly. Now 19.9% of imputed values were equal to one. We were more confident that this difference between the observed proportion (16.3%) and this imputed proportion (19.9%) was genuine, rather than being due to a misspecified imputation model for dcaffmgr3, and decided to adopt this method.

6 Results

6.1 Missingness models in IPW

Table 5 shows the odds ratios for the predictors of missingness at the LP visit. It can be seen that women who experience nausea at EP, those who have no qualifications, and non-white and younger women are less likely to attend the LP visit.

Table 6 shows the odds ratios for the predictors of missingness at the EP visit in the first IPW approach. It can be seen that, unlike for the LP visit, women who experience nausea at EP are more likely to attend the EP visit. Women drinking more caffeine at PP, eating at least the recommended quantity of fruit & veg at PP, smoking at EP, having no qualifications and who are non-white are less likely to attend the EP visit.

Figure 1 shows the distribution of weights for LP in the women attending the LP visit. The coefficient of variation of these weights is 0.032. The mean weight in women who do attend the LP visit is 1.050; in the women who do not attend, it is 1.070. Pseudo R^2, a measure of variance explained is 0.041. Figure 2 shows, from the first IPW approach, the distribution of weights for EP in the women attending the EP visit. The coefficient of variation of these weights is 0.138. The mean weight in the women who do attend the EP is 1.277; in the women who do not attend, it is 1.424. Pseudo R^2 is 0.052.

As the weights — especially for the LP visit — are not particularly variable, it might

Variable	OR	p-value
dnausea=1	0.43	
dnausea=2	0.56	
dnausea=3	0.81	0.000
acaffmglog	1.31	
acaffmg0*	0.30	0.021
acaffmgg2	0.56	0.001
dcrsmok	1.39	0.032
awexam=2	0.44	
awexam=3	0.46	
awexam=4	0.35	
awexam=5	0.28	
awexam=6	0.34	0.002
afruitvegg2	0.84	0.161
nonwhite	2.79	0.000
agest	0.99	0.003

Table 6: Predictors of missingness at the EP visit.

* acaffmg0 is an indicator variable for consumption of < 10mg caffeine per day. 137 (6.9%) women drink < 10mg caffeine per day. For them acaffmglog was set equal to zero. The estimated coefficients for acaffmglog and acaffmg0 (1.31 and 0.30) imply that a women with average log caffeine consumption (the average log caffeine consumption is 5) has an *adjusted* OR of 12.9 of missing her EP visit relative to a women which drinks < 10mg caffeine ($\exp(-\log 0.30 + 5 \log 1.31) = 12.9$). The p-value (0.021) reported for acaffmg0 is for the null hypothesis that the ORs of acaffmglog and acaffmg0 both equal one.

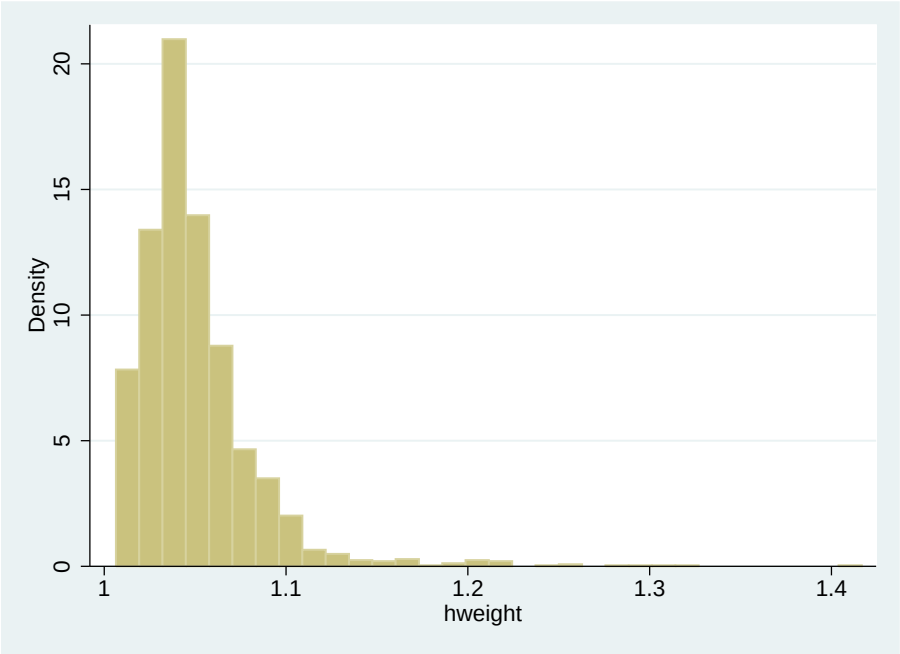

Figure 1: Distribution of inverse-probability weights for LP in women attending LP visit

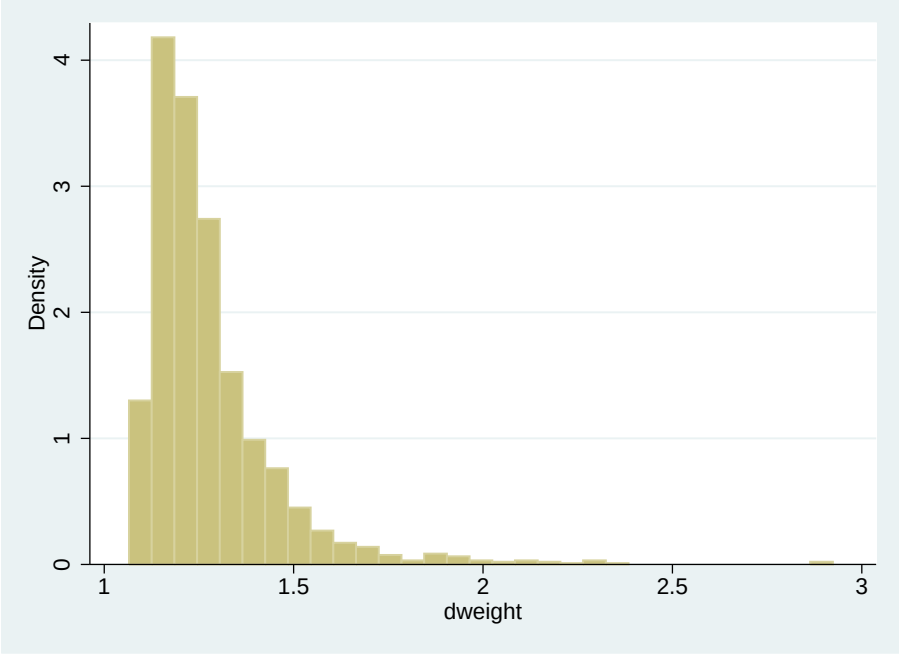

Figure 2: Distribution of inverse-probability weights for EP in women attending EP visit

be anticipated that IPW will not dramatically change the results of the Poisson regressions.

6.2 Comparisons of estimated relative risks

Table 7 shows the relative risks for outcomes measured at the PP visit, estimated by the various methods. Tables 8 and 9 show the relative risks for outcomes measured at the EP and LP visits. A dot means the variable was eliminated from the regression model by the backward stepwise elimination procedure. 'Comp' means the original analysis done by Crozier et al.The involves using only women with data at PP, EP and LP (1490 vomen) and excluding any women with missing data on a variable

Outcome		age	exam	bmi
acrsmok	Comp	0.942	0.735	.
	All	0.941	0.741	.
aalc	Comp	.	1.075	0.930
	All	.	1.073	0.956
afruitveg	Comp	0.966	0.831	.
	All	0.970	0.838	0.952
acaffmg	Comp	.	0.892	.
	All	.	0.890	.

Table 7: Relative risks for outcomes at the PP

Outcome		age	exam	bmi	naus	gest
dcrsmok	Comp	0.932	0.591	.	.	.
	All	0.936	0.591	.	.	.
	IPW	0.939	0.595	.	.	.
	IPW2	0.936	0.594	.	.	.
	MI	0.938	0.609	.	.	.
dalc	Comp	.	.	0.850	0.813	.
	All	.	.	0.819	.	.
	IPW	.	.	0.819	.	.
	IPW2	.	.	0.850	0.807	.
	MI	.	.	0.849	.	.
dfruitveg	Comp	0.972	0.828	.	1.089	.
	All	0.972	0.822	.	1.079	.
	IPW	0.973	0.822	.	1.076	.
	IPW2	0.973	0.829	.	1.085	.
	MI	0.972	0.840	.	1.084	.
dcaffmg	Comp	.	0.735	.	.	0.809
	All	.	0.722	.	0.866	0.820
	IPW	.	0.716	.	0.865	0.828
	IPW2	.	0.729	.	.	0.821
	MI	.	0.723	.	0.842	0.816

Table 8: Relative risks for outcomes at the EP

Outcome		age	exam	bmi	naus
hcrsmok	Comp	0.943	0.611	.	1.334
	All	0.947	0.618	.	1.281
	IPW	0.947	0.621	.	1.284
	MI	0.946	0.621	.	1.278
halc	Comp	1.109	.	.	.
	All	1.127	.	.	.
	IPW	1.126	.	.	.
	MI	1.116	.	.	.
hfruitveg	Comp	0.965	0.798	1.102	.
	All	0.963	0.812	1.098	.
	IPW	0.964	0.814	1.097	.
	MI	0.962	0.811	1.096	.
hcaffmg	Comp	.	0.744	.	.
	All	.	0.729	.	.
	IPW	.	0.730	.	.
	MI	.	0.731	.	.

Table 9: Relative risks for outcomes at the LP

Outcome	Comp	All	IPW	IPW2
acrsmok	1479	1981	.	.
dcrsmok	1478	1547	1547	1484
hcrsmok	1479	1888	1887	.
aalc	1477	1966	.	.
dalc	1480	1543	1543	1480
halc	1483	1893	1892	.
afruitveg	1479	1966	.	.
dfruitveg	1479	1548	1548	1485
hfruitveg	1470	1874	1873	.
acaffmg	1486	1981	.	.
dcaffmg	1479	1534	1534	1479
hcaffmg	1486	1888	1887	.

Table 10: Numbers of women included in regression models. For MI all 1987 women are included.

included in the selected regression model. 'All' means using all women except those with missing data on a variable included in the regression model. For outcomes at PP this means approximately 1987 women; for outcomes at EP, approximately 1553 women; for outcomes at LP, approximately 1893 women. 'IPW' means the IPW method. For the EP visit only, 'IPW' means IPW with the first set of weights and 'IPW2' means IPW with the second set of weights, i.e. those calculated after forcing the missingness pattern to be monotone. The first IPW analysis uses the same women as the 'All' analysis (apart from one woman excluded from analyses of outcomes at LP because of her unknown weight). The second IPW analysis excludes some women to make the missingness monotone. 'MI' means multiple imputation.

Table 10 shows the numbers of women included in each of the analyses.

With three exceptions all methods select the same predictors for each of the twelve outcome variables and yield very similar estimates of RR. The three exceptions are afruitvegg2, dalcg2 and dcaffmgg2. For afruitvegg2, using all 1966 women with the necessary data results in a different model that when just the subset of 1479 women with EP and LP visits are used: zawbmi becomes significant ($p = 0.028$). For dalcg2, the variable dnausea can cease to be significant. However, this is not too surprising, as it was only of borderline significance in the analysis of Crozier et al. ('Comp') and in the IPW2 analysis ($p = 0.039$ and $p = 0.038$, respectively). For dcaffmgg2, dnausea can become significant: $p = 0.043$ for 'All'; $p = 0.039$ for IPW; $p = 0.007$ for MI.

6.3 Comparisons of estimated standard errors

Tables 11, 12 and 13 show the ratios (as a percentage) of the SEs of the log relative risks obtained by Crozier et al. (the 'Comp' analysis) and the SEs obtained by the other methods. 'ave' means the average for that row of the table.

Outcome		age	exa	bmi	ave
acrsmok	All	124	123	.	123
aalc	All	.	114	114	114
afruitveg	All	114	116	.	115
acaffmg	All	.	116	.	116

Table 11: For PP visit: ratios (as %) of SEs of log relative risks from 'Comp' analysis and SEs obtained from other methods

Outcome		age	exa	bmi	nau	ges	ave
dcrsmok	All	103	105	.	.	.	104
	IPW	105	107	.	.	.	106
	IPW2	101	104	.	.	.	102
	MI	127	128	.	.	.	128
dalc	All	.	.	102	.	.	102
	IPW	.	.	99	.	.	99
	IPW2	.	.	95	97	.	96
	MI	.	.	100	.	.	100
dfruitveg	All	104	104	.	103	.	103
	IPW	102	104	.	102	.	103
	IPW2	98	100	.	99	.	99
	MI	104	108	.	107	.	106
dcaffmg	All	.	101	.	.	104	102
	IPW	.	101	.	.	105	103
	IPW2	.	100	.	.	102	101
	MI	.	109	.	.	111	110

Table 12: For EP visit: ratios (as %) of SEs of log relative risks from 'Comp' analysis and SEs obtained from other methods.

Outcome		age	exa	bmi	nau	ave
hcrsmok	All	123	122	.	123	123
	IPW	123	122	.	123	122
	MI	125	126	.	124	125
halc	All	109	.	.	.	109
	IPW	109	.	.	.	109
	MI	108	.	.	.	108
hfruitveg	All	111	113	112	.	112
	IPW	111	113	111	.	112
	MI	112	114	112	.	113
hcaffmg	All	.	116	.	.	116
	IPW	.	117	.	.	117
	MI	.	117	.	.	117

Table 13: For LP visit: ratios (as %) of SEs of log relative risks from 'Comp' analysis and SEs obtained from other methods.

Looking at Table 11, we see that, except for acrsmok, the SE is about 15% greater when only data on women with EP and LP visits are used, compared to when data on all women are used. This suggests a relative efficiency of about 133% ($1.15^2 = 1.33$), which is what would be expected if the data were MCAR, given the approximately 33% increase in the number of women ($1987/1490 = 1.33$). The reduction in the SE is greater for acrsmok, suggesting that the data are not MCAR. The greater gain in efficiency for acrsmok may be because the women who are excluded from the 'Comp' analysis are more informative about predictors of smoking at PP than are the women who are included: among those excluded, 33% smoke at PP; among those included, only 27% smoke; and for a binary outcome Z with $P(Z = 1) < 0.5$, cases with $Z = 1$ will be more informative than cases with $Z = 0$. For alcohol and fruit & veg consumption, the proportions drinking too much or eating too little are about the same in both the included and the excluded women. For caffeine consumption, the proportion drinking too much is actually less in the excluded than in the included women (35% versus 39%), making the excluded women potentially lightly less informative than the included women.

The pattern of efficiency of 'All' relative to 'Comp' is similar at EP (see Table 12). This time, 'All' uses about 1540 women, compared to approximately 1480 used by 'Comp', so the relative efficiency would be expected to be about 4% ($1540/1480 = 1.04$), translating as a SE 2% bigger in 'Comp' than in 'All'. For dcrsmok, the increase is 4%. Again, the excluded women may be more informative than the included ones: 22.9% of excluded women smoke at EP, versus only 14.6% of included women.

The SEs from IPW and IPW2 appears to be similar to those of the corresponding unweighted analyses ('All'). This is presumably because the weights are not very variable, as one would normally expect SEs to increase when weights are introduced.

For dalcg2 MI makes little difference compared to 'All'. For dfruitvegg2 it does reduce the SE, suggesting that some information is available in the observed variables for the imputation of missing dfruitvegg2 values.

For the analysis of dcrsmok, the SE is 28% greater for 'Comp' than for 'MI', suggesting a relative efficiency of about 164%. All of the variables in this regression model are actually observed for 1922 of the 1987 women (because dcrsmok was elicited at 19 weeks or at the LP visit for most of those who missed the EP visit). So, the relative efficiency might be expected to be about 130% ($1922/1478 = 1.30$). We suggest that the reason for the difference between these two figures, 164% and 130%, is again that the excluded women may be more informative than the included ones (22.9% of excluded women smoked at EP, versus 14.6% of included women). The same regression model was also fitted to the 1922 women with observed data, excluding the 65 with missing data. Unsurprisingly, the SEs for jwage and awexam thus obtained were almost identical (1% bigger) to those from the MI analysis.

The reduction in the SEs for dcaffmgg2 when MI is used is also greater than expected. Again, we suggest this may be because the excluded women are more informative: 19.6% of imputed dcaffmgg2 values equal one, compared to only 16.3% of observed values.

For outcomes at LP the pattern of SEs is similar to that for outcomes at PP and

EP. The relative efficiency of the 'All' analysis would be expected to be about 128% compared to the 'Comp' analysis (1890/1480 = 1.28). This corresponds to the SEs of the 'Comp' analysis being about 13% greater ($\sqrt{1.28} = 1.13$) than those of the 'All' analysis. Except for hcrsmok, this is what is observed in Table 13. The greater reduction in the SEs for hcrsmok when all women are used may again be that the excluded women are more informative than the included ones (22.3% of excluded women smoked at LP, versus 14.0% of included women). When IPW is used, the SEs change little from the corresponding unweighted analysis ('All'), which is unsurprising given the lack of variability in the weights. When MI is used, the reduction in the SEs is very small, which is consistent with the small number of missing values needing to be imputed.

7 Summary

The original (complete-case) analyses performed by Crozier et al. rely on the MCAR assumption: that the 1490 women who attend both the EP and LP visits are representative of all 1987 pregnant women in the study. An examination of variables measured at PP in all women reveals that this assumption is not valid. IPW and MI were used to assess whether replacing this MCAR assumption with a weaker MAR assumption changes the results of the analyses.

It was found that the results changed little when IPW and MI were used. There were 12 outcome variables and, with three exceptions, all methods selected the same predictors for each of these outcome variables and yielded very similar estimates of RR. The three exceptions are afruitvegg2, dalcg2 and dcaffmgg2. For afruitvegg2, the variable zawbmi can become marginally significant. For dalcg2, the variable dnausea, which was marginally significant in the original analysis, can cease to be significant. For dcaffmgg2, dnausea can become significant.

We found, as expected, that an available-cases analysis, that is, using all women with data on the variables in a regression model, increases the precision of the estimates. MI was found to reduce the SEs still further, but not by much. This reduction is because MI enables women to be included who did attend the visits but for whom some variables in the regression model were not recorded. The reduction is small because the number of such women is small. IPW is expected to increase the SEs relative to the corresponding unweighted analysis, but in this study it made little difference. This was presumably because the inverse-probability weights were not very variable.

The advantage of the complete-case analysis of Crozier et al. is that (almost) the same women are used in the regressions for all 12 outcome variables (the 'almost' is because a small number of women with missing items are excluded from some analyses but not others). Thus, the estimates from all 12 regressions are for the same sample of 1490 women. However, the complete-case analysis is potentially inefficient and could lead to biased estimates if the data are not MCAR. The advantage of using IPW or MI is that — provided the data are MAR — the bias should be removed and greater efficiency achieved. The weighting or imputation ensures that the estimates from all 12 regressions are for the same sample of 1987 women.

As the results from MI and IPW are little different from the original results, we can have more confidence in those original results.

The data may not be MAR: they may be missing not at random (MNAR). To assess the effect of deviations from MAR, fuller sensitivity analyses would be needed. These would involve making a variety of plausible assumptions about how the data might depart from MAR.

Reference

Do Women Change Their Health Behaviours in Pregnancy? Findings from the Southampton Women's Survey. Crozier SR, Robinson SM, Borland SE, Godfrey KM, Cooper C, Inskip, HM. Paediatric and Perinatal Epidemiology, 2009; 23: 446–453.