Uncertainty Intervals and Sensitivity Analysis for Missing Data

Minna Genbäck

Department of Statistics
Umeå School of Business and Economics
Umeå University 2016

Doctoral Thesis
Department of Statistics
Umeå School of Business and Economics
Umeå University
SE-901 87 Umeå

Statistical Studies No. 50
ISBN: 978-91-7601-555-1
ISSN: 1100-8989
Cover photo: Mimosa sensitiva by Eric Hunt
Electronic version available at http://umu.diva-portal.org/

Printed by: Print & Media
Umeå, Sweden 2016

Contents

List of Papers

The thesis is based on the following papers:

I. Genbäck, M., E. Stanghellini and X. de Luna (2015). Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome. Statistical papers 56(3), 829-847.

II. Genbäck, M., N. Ng, E. Stanghellini and X. de Luna (2016). Predictors of decline in self-reported health: addressing non-ignorable dropout in longitudinal studies of ageing. *Manuscript.*

III. Genbäck, M. and X. de Luna (2016). Bounds and sensitivity analysis when estimating average treatment effects with imputation and double robust estimators. *Manuscript.*

IV. Genbäck, M. (2016). Uncertainty intervals for mixed effects models with non-ignorable missingness. *Manuscript.*

Abstract

In this thesis we develop methods for dealing with missing data in a univariate response variable when estimating regression parameters. Missing outcome data is a problem in a number of applications, one of which is follow-up studies. In follow-up studies data is collected at two (or more) occasions, and it is common that only some of the initial participants return at the second occasion. This is the case in Paper II, where we investigate predictors of decline in self reported health in older populations in Sweden, the Netherlands and Italy. In that study, around 50% of the study participants drop out. It is common that researchers rely on the assumption that the missingness is independent of the outcome given some observed covariates. This assumption is called *data missing at random* (MAR) or *ignorable missingness mechanism*. However, MAR cannot be tested from the data, and if it does not hold, the estimators based on this assumption are biased. In the study of Paper II, we suspect that some of the individuals drop out due to bad health. If this is the case the data is not MAR. One alternative to MAR, which we pursue, is to incorporate the uncertainty due to missing data into interval estimates instead of point estimates and uncertainty intervals instead of confidence intervals. An uncertainty interval is the analog of a confidence interval but wider due to a relaxation of assumptions on the missing data. These intervals can be used to visualize the consequences deviations from MAR have on the conclusions of the study. That is, they can be used to perform a sensitivity analysis of MAR.

The thesis covers different types of linear regression. In Paper I and III we have a continuous outcome, in Paper II a binary outcome, and in Paper IV we allow for mixed effects with a continuous outcome. In Paper III we estimate the effect of a treatment, which can be seen as an example of missing outcome data.

KEYWORDS: missing data; missing not at random; non-ignorable; set identification; uncertainty intervals; sensitivity analysis; self reported health; average causal effect; average causal effect on the treated; mixed-effects models

Sammanfattning

I den här avhandlingen utvecklar vi metoder för att hantera inkomplett data i en responsvariabel vid skattning av en regressionsmodell. Det finns många orsaker till inkomplett data, varav ett exempel är uppföljningsstudier där egenskaper mäts hos individer vid flera tillfällen. Där är det vanligt att inte alla individer som deltagit vid första tillfället kommer tillbaka till det senare tillfället. I Papper II vill vi hitta riskfaktorer för försämring av självskattad hälsa hos ett slumpmässigt urval av individer äldre än 50 år i Sverige, Nederländerna och Italien. Den här studien har ett omfattande bortfall, ca 50%, och vi tror att en del av de som inte kommer tillbaka gör det p.g.a. dålig hälsa. Om bortfallet skett helt slumpmässigt, finns många existerande metoder för att analysera datamaterialet på ett korrekt sätt. Den här avhandlingen handlar om när bortfallet är relaterat till responsvariabeln. De flesta metoder för att hantera den typen av data bygger på att man gör andra starka antaganden, vilket leder till en god precision i skattningarna om antagandena är uppfyllda. Problemet med dessa metoder är att det utifrån datamaterialet är omöjligt att avgöra om antagandena är uppfyllda eller inte. Vi föreslår metoder för att ta hänsyn till bortfallet utan att göra dessa starka antaganden, genom att intervallskatta istället för att punktskatta. Detta synliggör osäkerheten från bortfallet samt vilka konsekvenser det får för slutsatserna av studien.

Avhandlingen innehåller metoder för att hantera inkomplett data i olika typer av linjära regressionsmodeller, med kontinuerligt utfall (Papper I och III), binärt utfall (Papper II) och blandade effekter (Papper IV). I Papper III utvecklas metoder för skattning av effekten av en behandling, vilket också är ett exempel på inkomplett data i en responsvariabel.

Preface

Being a PhD-student has been a pleasant journey for me, for which I have quite a few people to thank. First and foremost, I would like to thank Xavier de Luna, my main supervisor, for keeping me focused on what is most important and not letting me spend too much time on what is not. Also, thank you for giving me interesting problems to solve, for letting me find answers on my own, and for constantly finding the weak spots in our work (in the best possible sense). Without your constant support and knowledge this thesis could not have been written. I would also like to thank Elena Stanghellini, my co-supervisor, for being so generous with your time and knowledge. I have enjoyed and learned so much from working with you. I would especially like to thank you for exploring the theory for the first paper together with me, which was important both for this thesis and for my development as a researcher. Also, thank you Nawi for bringing a much needed perspective into Paper II.

Next, I would like to thank the entire department of statistics. I honestly cannot imagine a better place to work. I love how open, generous and supportive we are in research and in teaching. I would especially like to thank Anders for all your helpful comments, and Ingrid, your support has meant a great deal to me. I would also like to thank all my fellow PhD-students from the department, especially Philip and Anita, for making this so much more fun, for all the great advice, and for being so thoughtful.

Finally, thank you to all my family and friends. Thank you Anton for proofreading a part of this thesis. Thank you Moa and Astrid for keeping my mind off work and on more important things. Last but not least, thank you Markus for being my partner in life, with all that it entails... I love you!

Umeå, October 2016
Minna Genbäck

1 Introduction

This thesis focuses on estimating regression parameters (within different contexts) when the outcome has missing data. Missing outcome data is a problem in a number of applications, for instance in follow-up studies. Hence, it has been researched extensively in both statistics and econometrics. In follow-up studies, information about the participants is gathered at two or more occasions, and usually some of the participants will not show up to the follow-up session(s), rendering an incomplete dataset. Another practical example of missing outcome data is in evaluation of the effect of a treatment, which is covered in detail in Section 4.

It is common to assume that the missingness mechanism is independent of the outcome, given some covariates, called missing at random or ignorable missingness, in order to analyze and draw conclusions from data with missing outcomes. This thesis contains methods for investigating how sensitive these conclusions are to the assumption of ignorable missingness. In all papers, we derive uncertainty intervals with a sensitivity parameter (see Section 3) to investigate departures from ignorability. In Paper I we derive uncertainty intervals in a linear regression context with a continuous outcome, and in Paper III these results are further developed to a causal inference context. In Paper IV we propose uncertainty intervals within linear mixed effects models (LME). Finally, in Paper II uncertainty intervals are derived in an applied study with a binary outcome.

This thesis is structured as follows. It begins with a brief overview of missing data mechanisms and some common approaches for analysis of incomplete data in Section 2, followed by an introduction to uncertainty intervals and partial identification in Section 3. An introduction to the causal framework, including how it can be viewed as missing data, is given in Section 4, followed by a summary of the enclosed Papers I-IV.

2 Missing outcome data

A large part of the missing data literature focuses on missing outcome data (i.e. if we have covariates, there are no missing data in them).

Definitions for the missing outcome data mechanisms were originally introduced by Rubin (1976) (with a multivariate outcome, not univariate as we have here) and can also be found in Little and Rubin (2002) and Molenberghs et al. (2015).

2.1 Missing data mechanisms

Let z_i be an indicator variable that is 1 if the outcome y_i is observed and 0 otherwise, and let $\mathbf{x}_i$ be a vector of covariates observed for each individual i. Then the data is said to be *missing completely at random* (MCAR) if

$$\Pr(z_i|y_i, \mathbf{x}_i) = \Pr(z_i).$$

Hence, if the data is MCAR the individuals without missing data (complete cases) can be viewed as a random sample of the target population. That is, using only the complete cases will give valid estimates and inference, while the only loss due to missing data is a smaller sample size.

A weaker assumption is that the data is *missing at random* (MAR),

$$\Pr(z_i|y_i, \mathbf{x}_i) = \Pr(z_i|\mathbf{x}_i).$$

Here the missing data mechanism does not depend on the outcome conditionally on the observed covariates $\mathbf{x}_i$. If interest lies in the regression of y_i on $\mathbf{x}_i$, the estimated parameters in a complete case analysis are unbiased under MAR, since we adjust for $\mathbf{x}_i$. However, complete case analyses are generally not valid under MAR. For instance, the sample mean and variance will be biased estimates of the corresponding moments in the target population.

Note that the definition of MAR above is not the most common one. However, in applications where the outcome is univariate, the definition above is usually what is meant by MAR.

2.2 Inference for an ignorable missingness mechanism

Under MCAR and/or MAR the missingness mechanism is said to be *ignorable* and many standard statistical techniques are valid (Molenberghs et al., 2015).

As mentioned earlier, estimating regression parameters based on the complete cases will yield unbiased estimates under MAR. Likelihood

based techniques are also valid as long as the distribution of y_i is correctly specified. A large strain of the literature regarding missing data is imputation techniques, where values are imputed instead of the missing data using parametric models (Rubin, 1996). Finally, weighting methods, such as inverse probability weighting and doubly robust estimators (Tsiatis, 2006), have become popular.

2.3 Inference for a non-ignorable missingness mechanism

If MCAR and MAR are not fulfilled, the missing data mechanism is called *non-ignorable*, or alternatively the data is called *missing not at random* (MNAR). Estimates made with the techniques in Section 2.2 are then biased and other methods of inference need to be considered. The non-ignorable statistical methods are usually divided into those based on selection models and those based on pattern mixture models.

Pattern mixture models are predominantly used in longitudinal studies, and involve modeling the outcome for different patterns of missingness. Pattern mixture models require additional assumptions in order to achieve identification: different specific assumptions about the distribution of the missing outcomes given the observed ones are commonly used (Molenberghs et al., 2015). The methods of this thesis are not within the pattern mixture framework and therefore they are not further discussed here. For more information about pattern mixture models see Daniels and Hogan (2008); Little (2009); or Molenberghs et al. (2015).

Selection models are built by estimating a model for selection (the missingness mechanism) in order to identify the distribution of y_i given $\mathbf{x}_i$ and are common in the econometric literature. Selection models with a univariate outcome were developed by Heckman (1979), and can be fitted by maximum likelihood or a two-stage regression. Selection models usually require additional assumptions, such as exclusion restrictions (instrumental variables), in order to identify the parameters of interest.

Most of the literature recommends a sensitivity analysis of the untestable assumptions that are a necessity in order to identify parameters in presence of missing data. In this thesis we model the missingness mechanism in the same way as selection models, although we avoid making additional identifying assumptions and get partial identification (explained in Section 3) instead of point identification.

3 Uncertainty intervals and partial identification

When a dataset is incomplete, the observed data could come from many different complete data sets, and the different complete datasets would lead to different estimates and inference. Instead of making untestable assumptions to achieve point identification, it is possible to make weaker or no assumptions to achieve partial identification (Manski, 2003). Formal definitions of partial identification (identification intervals or ignorance intervals) and uncertainty intervals are found in Vansteelandt et al. (2006). Below we give a less formal but hopefully more accessible account.

Suppose that the parameter of interest is $\theta \in \Theta$, and that $\theta = \theta_0$ for the target population. Then we say that θ is *identified* if

$$\hat{\theta}_\infty = \theta_0,$$

where $\hat{\theta}_\infty$ is an estimator of θ with a sample size that tends to infinity. Similarly we say that θ is *partially identified* if we can estimate an interval which converges to an interval that contains θ_0 and is a subset of the parameter space Θ, i.e,

$$\hat{\theta}^L_\infty = \theta^L,\ \hat{\theta}^U_\infty = \theta^U, \text{ and } \theta_0 \in [\theta^L,\ \theta^U] \subset \Theta.$$

The interval $[\theta^L,\ \theta^U]$ is called an *identification interval*, or an *ignorance interval*. This is an interval containing θ_0 which is consistent with the observed data law.

A $(1-\alpha)100\%$ *uncertainty interval* (UI) is an interval that covers the identification interval with a probability of $(1-\alpha)$.

The difference between a confidence interval (CI) and an UI is that a CI covers a point θ_0, and the UI covers an interval $[\theta^L,\ \theta^U]$, with a certain probability. Hence, the uncertainty interval will cover θ_0 with a probability of at least $(1-\alpha)$. Uncertainty intervals can be constructed with non-parametric or parametric models, with or without the use of sensitivity parameters (discussed in Section 3.1). To help understand the different terms let us look at a simple example.

Example 1. We ask a Yes or No question to a sample of 1000 and want to estimate p, the proportion of Yes answers in the population. Suppose

that we get 450 individuals answering Yes, 300 answering No and 250 that did not answer.

If we do a complete case analysis, we would estimate the parameter p by $\hat{p} = \frac{450}{750} = 0.60$, which is only unbiased if the data is missing completely at random (MCAR). However, the only information we have for certain is that between 450 and 700 out of the 1000 would have answered Yes, i.e. $0.45 \leq \hat{p} \leq 0.70$. Hence, without further assumptions (for instance MCAR) p cannot be point identified. However, p is partially identified and an estimate of the identification interval containing p is $[0.45, 0.70]$.

Since we only observed a sample, we might want to make a 95% confidence interval for p, which is $(0.56, 0.64)$ assuming MCAR. The analog of a 95% confidence interval for the identification interval is called a 95% uncertainty interval. We can derive the UI by adding lower (and upper) confidence bounds to the lower (and upper) bound of the estimated identification interval. The 95% uncertainty interval for p is (0.42, 0.73). Note that the UI is wider than the CI since it does not require assumptions about the missing data.

3.1 Sensitivity parameters

A sensitivity parameter is a parameter that can be used to narrow down the identification interval, by letting it describe some extra available information. Continuing with Example 1, let us use the proportion of Yes answers among the non-respondents ($p_{z=0}$) as a sensitivity parameter. We know for certain that $p_{z=0} \in [0, 1]$. If we do not want to impose restrictions on the sensitivity parameter, an estimate of the identification interval is $[0.45, 0.70]$, as above. If, however, we assume MCAR, this is equivalent to assuming that $p_{z=0} = p_{z=1}$, and we get a point estimate of 0.60 as above. Now suppose that we have subject matter knowledge that indicates that the proportion of Yes answers is greater among the non-responders compared to the responders ($p_{z=0} \geq p_{z=1}$). We want to use this knowledge and assume that $p_{z=0} \in [p_{z=1}, 1]$. The resulting estimate of the identification interval is then $[0.60, 0.70]$ and the resulting 95% UI is $[0.56, 0.73]$. By using sensitivity parameters, we have a tool that can use subject matter knowledge, while allowing for weaker assumptions than MCAR.

3.2 Uncertainty intervals for regression parameters

The theme of this thesis, estimation of regression parameters when the data is missing in a non-random way, is a considerably more complicated problem than the one in Example 1. Usually in a regression there are many variables (covariates) and a model is built to estimate their association with the outcome. The theoretic framework of partial identification can, however, also be used here since the outcome is partially missing. In this thesis we use parametric regression models and a sensitivity parameter to build uncertainty intervals.

The work by Horowitz and Manski (2006) could be used in the same setting, if the outcome and covariates are binary (or discretized) and bounded. They derive a similar type of set estimates as in Example 1 for regression parameters. Their approach allows for missing values both in covariates and outcome. They propose an algorithm that gives almost the same results as if you impute all possible combinations of values instead of the missing data, and run the regression for all these completed datasets. By doing this, you get estimated identification intervals for each regression parameter. This approach is appealing since it requires no assumptions on the missing data. However, it is computationally intensive and the resulting set estimates are often quite wide. Horowitz and Manski (2006) do not use sensitivity parameters and hence can not use subject matter knowledge to get narrower intervals.

3.3 Uncertainty intervals for sensitivity analysis

Uncertainty intervals using sensitivity parameters can be used for sensitivity analyses of ignorability (MAR) in two slightly different ways.

One way is to derive UI:s based on reasonable assumptions on the sensitivity parameter and checking if the conclusions differ from the MAR analysis. In practice, this means checking for all significant variables if the UI contains 0. If it does, the variable is sensitive to non-ignorability (MNAR).

Another way is to find out for which values of the sensitivity parameter the conclusions from the MAR analysis hold for each significant variable. In practice, this means finding out for which values of the sensitivity parameter the UI does not contain 0. There will generally be different ranges of the sensitivity parameter where the conclusions

hold for each variable. Similarly to the first sensitivity analysis, there will then be a judgment of which ranges of the sensitivity parameter are reasonable, in order to decide which variables are sensitive to non-ignorability.

4 Causal inference

The problem of correctly evaluating a treatment in observational studies can be viewed as a missing data problem, which will be explained later in this section. Treatment is defined broadly: it can be a medical treatment, being enrolled into a labor market program, getting divorced, etc. Also in this section, we have a running example to help explain terminology.

Example 2. An unemployment office has the possibility to assign their employment seekers into a labor market program. In this example, treatment is defined as being enrolled into the program. Ideally we want to know, for each employment seeker, if the labor market program was beneficial for them and by how much. Note that beneficial is an unspecific outcome variable and difficult to measure, so instead we use something easily measurable, income (the total work income in the first two years after completing the program).

4.1 Potential outcome framework

To formally define average causal effects it is useful to use the potential outcome framework, introduced by Neyman (1923) in a random experiment context and established by Rubin (1974) as a framework for causal inference in observational studies.

Let z be a variable indicating if an individual is enrolled in the program or not, taking value 1 if an individual is enrolled and 0 otherwise. Now the potential outcomes are: y^1 income if enrolled in the program ($z = 1$), and y^0 income if not enrolled ($z = 0$). The individual causal effect of the program on income is then defined as:

$$y_i^1 - y_i^0 \qquad i = 1, ..., n.$$

The two potential outcomes are defined for each individual in the study although only one is observed (y^1 is observed when $z = 1$, and y^0 is observed when $z = 0$). Therefore, calculating the individual causal effect from observed data is impossible (Holland, 1986). We can, however, estimate the average causal effect for the employment seekers, also called average treatment effect:

$$\tau = \mathrm{E}(y^1 - y^0),$$

or the average causal effect of the ones that are enrolled in the program, called average causal effect on the treated (or average treatment effect on the treated):

$$\tau^1 = \mathrm{E}(y^1 - y^0 | z = 1).$$

4.2 Identification of causal effects

Since we are missing half of the outcomes when estimating τ and some of the outcomes when estimating τ^1, additional assumptions are needed for identification of τ and τ^1. For this purpose, most of the literature relies on three assumptions.

The first assumption is that the potential outcome(s) are MAR, i.e.

$$\Pr(z_i | y_i^0, \mathbf{x}_i) = \Pr(z_i | \mathbf{x}_i) \text{ for } \tau^1,$$

$$\Pr(z_i | y_i^0, y_i^1, \mathbf{x}_i) = \Pr(z_i | \mathbf{x}_i) \text{ for } \tau,$$

where $\mathbf{x}_i$ are confounders observed for each individual i. This assumption is usually called *no unobserved confounders*, *unconfoundedness*, or *ignorability*, within the causal inference literature and it is not fulfilled if we do not observe all confounders. A confounder is a variable that affects both treatment assignment and outcome. For instance, if the employment office only assigns individuals they see as unmotivated into the program, and if motivated employment seekers have a different income from unmotivated ones (with or without the program), then motivation is a confounder.

The second assumption is *overlap*, which means that the probability of being enrolled into the program cannot be 0 or 1 for any individual. Formally this is written as:

$$\Pr(z = 0 \mid x) > 0, \forall x \in \mathcal{X} \text{ for } \tau^1,$$

$$0 < \Pr(z = 1 \mid x) < 1, \forall x \in \mathcal{X} \text{ for } \tau,$$

where $\mathcal{X}$ is the support of x.

Finally, the third assumption is the *stable unit treatment value assumption* (Rubin, 1980), which states that there is only one version of treatment and that an individuals outcome is not affected by which other individuals are treated. This assumption is not fulfilled if, in our example, the program is divided into groups with different content in some way, or if an individuals income is affected by whether or not their friends were enrolled into the program. In Paper III we derive uncertainty intervals in order to relax one of these assumptions: the MAR assumption.

5 Summary of papers

5.1 Paper I

In the first paper we propose a sensitivity analysis of the assumption that the missing continuous outcome data is MAR for the estimation of regression parameters $\boldsymbol{\beta}$ (ordinary least squares, OLS). The sensitivity analysis consists of first estimating $\boldsymbol{\beta}$ with a complete case analysis and then estimating the bias of this estimate (due to non-ignorability). The bias is expressed as a function of a sensitivity parameter ρ, which is the correlation between the missingness mechanism and the outcome given the covariates. If we assume that ρ is in an interval, the estimate of $\boldsymbol{\beta}$ and its bias can be used to derive an UI for $\boldsymbol{\beta}$.

The results of the paper are illustrated with a study on predictors of body mass index (BMI) change in middle age men allowing us to identify possible predictors of BMI change even when assuming little about the missingness mechanism. A simulation study comparing our proposed method with Heckman's two-step procedure (Heckman, 1979) is also performed.

5.2 Paper II

The second paper is an applied paper using data from the Survey of Health, Ageing and Retirement in Europe (SHARE). This is a panel

survey conducted on individuals aged 50 or older in several European countries. The objective of this study is to identify predictors of declining self-reported health in an older population (50 years or older) in Sweden, the Netherlands, and Italy. Decline in self-reported health in this study is defined as a binary variable with the levels either declining from good health to bad or staying in good health. For this, we use the baseline wave conducted in 2004, and the latest follow-up conducted in 2013 and fit six logistic regression models, one for each country and sex. Since only 2893 out of the 5653 participants from 2004 participated in 2013, we study whether the results are sensitive to the high dropout rate. We suspect that some might drop out due to bad health, which would mean that the missing data mechanism is non-ignorable. Hence, we develop a method for sensitivity analysis when the outcome is binary. The sensitivity analysis is based on uncertainty intervals that are derived by maximizing the likelihood for different values of the sensitivity parameter ρ, and is closely related to the approach of Copas and Li (1997), and Stingo et al. (2011).

We found that age and a greater number of chronic diseases were positively associated with a decline in self-reported health in the three countries. Maximum grip strength was associated with decline in self-reported health in Sweden and Italy, and higher body mass index and self-reported limitations in normal activities due to health problems was associated with decline in self-reported health in Sweden.

5.3 Paper III

In the third paper we propose a sensitivity analysis of the unconfoundedness assumption when estimating average causal effects using an imputation or doubly robust estimator. The sensitivity analysis is based on uncertainty intervals, allowing for unobserved confounders, for the causal effect of interest. The UI:s are derived, like in Paper I, from the bias of the estimators using a complete case analysis, expressed as a function of a sensitivity parameter.

In this paper we also contrast the size of potential bias due to violation of the unconfoundedness assumption and to the misspecification of the models used to explain outcome with the observed covariates. While the latter bias can in principle be made arbitrarily small with increasing sample size (by increasing the flexibility of the models used), the bias

due to unobserved confounding does not disappear with increasing sample size. Through numerical experiments we illustrate the relative size of the biases due to unobserved confounders and model misspecification, as well as the empirical coverage of the uncertainty intervals.

5.4 Paper IV

The fourth paper adapts the results from Paper I to linear mixed effects models (LME). LME models are useful for datasets that consist of many small groups, where the members of the group share some characteristic that is unmeasured. The number of parameters required to control for this grouping using ordinary least squares is large while a LME only requires a few parameters. On the other hand, LME requires a distributional assumption about the group effect. In LME the difference between the groups is not important in itself. It is regarded as a nuisance that needs to be accounted for in order to estimate the other effects correctly. Paper IV contains theoretical results and a simulation study.

6 Further research

This thesis proposes sensitivity analysis methods for missing not at random outcome data, in different scenarios. All simulations and almost all data analyses were performed with `R` statistical software, a natural continuation would be to write an `R-package` containing these sensitivity analyses. Although `R-code` is already produced for this thesis and can be provided on request, this code is not user friendly enough for an inexperienced user of `R` and the computing time can probably be optimized further, hence the need for an `R-package`.

Other possibilities for further research within the area is to allow for several levels of missingness or more than two levels of treatment.

Finally, a natural extension of Paper IV would be to consider generalized linear mixed models, allowing e.g. for a binary outcome.

References

Copas, J. and Li, H. (1997). Inference for non-random samples, *Journal of the Royal Statistical Society: Series B* **59**(1): 55–95.

Daniels, M. and Hogan, J. (2008). *Missing Data In Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman and Hall/CRC.

Heckman, J. (1979). Sample selection bias as a specification error, *Econometrica* **47**(1): 153–161.

Holland, P. W. (1986). Statistics and causal inference, *Journal of the American statistical Association* **81**(396): 945–960.

Horowitz, J. and Manski, C. (2006). Identification and estimation of statistical functionals using incomplete data, *Journal of Econometrics* **132**(2): 445–459.

Little, R. (2009). Selection and pattern-mixture models, *in* G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs (eds), *Longitudinal Data Analysis*, Chapman and Hall/CRC.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*, 2 edn, Wiley and Sons.

Manski, C. (2003). *Partial Identification of Probability Distributions*, Springer.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. and Verbeke, G. (2015). *Handbook of missing data methodology*, CRC Press.

Neyman, J. (1923). On the application of probability theory to agricultural experiments, essay on principles., Roczniki nauk Rolczych *X*, 1-51. In Polish, English translation by D.M. Dabrowska and T. P. Speed in *Statstical Science*, *5*, 465-472, 1990.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies., *Journal of educational Psychology* **66**(5): 688.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**(3): 581–592.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment, *Journal of the American Statistical Association* **75**(371): 591–593.

Rubin, D. B. (1996). Multiple imputation after 18+ years, *Journal of the American statistical Association* **91**(434): 473–489.

Stingo, F. C., Stanghellini, E. and Capobianco, R. (2011). On the estimation of a binary response model in a selected population, *Journal of statistical planning and inference* **141**(10): 3293–3303.

Tsiatis, A. (2006). *Semiparametric theory and missing data*, Springer Science & Business Media.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, *Statistica Sinica* pp. 953–979