IFAU
Institute for Evaluation of Labour
Market and Education Policy

Some aspects of propensity score-based estimators for causal inference

Ronnie Pingel

DISSERTATION SERIES 2014:5

Presented at the Department of Statistics, Uppsala University

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala. IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

IFAU also provides funding for research projects within its areas of interest. The deadline for applications is October 1 each year. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The institute has a scientific council, consisting of a chairman, the Director-General and five other members. Among other things, the scientific council proposes a decision for the allocation of research grants. A reference group including representatives for employer organizations and trade unions, as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P O Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

This doctoral dissertation was defended for the degree of Doctor in Philosophy at the Department of Statistics, Uppsala University, September 19, 2014. Paper 1 has previously been published as Working paper 2013:5.

ISSN 1651-4149

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Pingel, R. and I. Waernbaum (2014). Effects of correlated covariates on the asymptotic efficiency of matching and inverse probability weighting estimators for causal inference. *Statistics: A Journal of Theoretical and Applied Statistics*. Epub ahead of print.

II Pingel, R. and I. Waernbaum (2014). Correlation and efficiency of propensity score-based estimators for average causal effects. *Manuscript*.

III Pingel, R. (2014). Estimating the variance of a propensity score matching estimator: A new look at right heart catheterisation data. *Submitted*.

IV Pingel, R. (2014). Some approximations of the logistic distribution with application to the covariance matrix of logistic regression. *Statistics & Probability Letters 85*, 63–68.

Reprints were made with permission from the publishers.

Contents

1. Introduction

Although most researchers would agree that the question "What is the effect of [...] on [...]?" is generally more interesting than "Are [...] and [...] associated?", traditionally, most statisticians have been reluctant to get into discussions about causality. Still, suppose we really would like to know the causal effect of an education policy on academic achievement or perhaps the causal effect of a vaccine on some disease. How should we proceed? A prerequisite is to have a clear understanding and a formal statistical formulation of what we mean by causation.

In this thesis causality is defined in terms of potential outcomes, as introduced by Neyman (1923) and extended by Rubin (1974). A potential outcome can be thought of as what would have happened to an individual if he or she received a different treatment than the one actually given. In reality we can only observe the outcome for that individual under the treatment actually received. Thus, causality in this context is reduced to thinking of how to observe outcomes that would have been if the individuals received a different treatment.

Several important consequences may be derived from the idea of potential outcomes (Holland, 1986). First, instead of becoming entangled in an attempt to define what the causes of a given effect are, studies of causation should start with asking what are the effects of causes. Second, a cause always refers to something relative to another cause. Third, each individual must be potentially exposable to any one of the treatments, or at least we must be able to imagine possible exposure. Importantly, these aspects are also the core ideas in the design of randomised experiments. Observing the link between potential outcomes and randomised experiments, suggests that potential outcomes can be used in observational studies to resemble an experimental setting as closely as possible. By doing so, researchers may ask causal questions even in settings where experimentation is not feasible or ethical. Certain issues when estimating causal effects in observational studies are the main focus of this thesis.

The thesis is organised as follows. It starts with a brief introduction to the framework used in this thesis, as well as a short description of the propensity score and propensity score-based estimators for average causal effects. This section is followed by a summary of the four papers included in the thesis. Paper I and II investigate how propensity score-based estimators for average causal effects are affected by having correlated covariates. Paper III provides some guidance regarding the implementation of estimators for the variance of a propensity score matching estimator for the average causal effect, and

applies the findings when estimating the effect of right heart catheterisation (RHC) on the survival of intensive care unit patients. The fourth paper studies the covariance matrix of estimators of parameters in a logistic model when having normally distributed random regressors. The fourth paper is related to the other papers in that logistic regression is commonly used to estimate the propensity score.

1.1 A framework for causal inference

In the following section we formally introduce the Neyman-Rubin framework for causal inference (Neyman, 1923; Rubin, 1974). The aim throughout the thesis is to evaluate the effect of a binary treatment, W, on some outcome, Y. By convention, those individuals exposed to the treatment, $W = 1$, are denoted the treatment group or treated, while those exposed to the control treatment, $W = 0$, are denoted the control group or controls. A binary treatment implies two potential outcomes: Y_1 had the individual been exposed to the treatment and Y_0 had the individual been exposed to the control. In a random sample of N individuals drawn from a large population we define the individual causal effects of the treatment as the differences of the potential outcomes

$$Y_{1i} - Y_{0i}, \quad i = 1, \ldots, N.$$

The fundamental problem of causal inference is that for each individual we only observe the realised outcome $Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}$, making identification and estimation of individual causal effects impossible (Holland, 1986). Still, aggregated causal effects may be estimated and the focus of this thesis is on estimation of the population average causal effect,

$$\tau = E(Y_1 - Y_0).$$

The average causal effect is probably the most popular estimand (Imbens and Wooldridge, 2009), but other causal effects might be of interest depending on the research question. Some examples are the average causal effect of the treated $E(Y_1 - Y_0|W = 1)$, the relative causal effect $E(Y_1)/E(Y_0)$ and the median causal effect $\text{Med}(Y_1 - Y_0)$. A clear definition of what is a causal effect, without making any parametric assumptions, and with focus on the estimands, is one of the key benefits of the Neyman-Rubin framework (Imbens and Wooldridge, 2009).

Although the average causal effect is defined without considering how treatments are assigned to individuals (Imbens and Wooldridge, 2009), in order to identify it assumptions regarding treatment assignment must be made. A perfectly conducted completely randomised experiment allows for identification of the average causal effect by conditioning

$$\tau = E(Y_1 - Y_0) = E(Y_1|W = 1) - E(Y_0|W = 0) = E(Y|W = 1) - E(Y|W = 0).$$

Random assignment of the treatment ensures that $(Y_1, Y_0) \perp\!\!\!\perp W$, where $\perp\!\!\!\perp$ denotes independence, from which the second equality follows. In this setting the difference in sample means yields an unbiased estimator of the average causal effect. Though often implicitly assumed, we also require that the treatments received by one individual do not affect the outcomes for other individuals (Cox, 1958), and that there is only one version of the treatment (Neyman et al., 1935). Rubin (1980) denotes the combination of these two assumptions the *stable unit treatment value assumption*, an assumption maintained throughout this thesis.

For ethical and practical reasons, it is often not possible to perform a randomised experiment. Instead, the researcher must rely on observational data and thus estimate the average causal effect in a setting where the independence assumption is not likely to hold. Even so, the average causal effect can be identified in an observational study (or in a randomised experiment with imperfections) under certain assumptions. Let X be a vector of pre-treatment variables, referred to as covariates. The assumptions of *unconfoundedness*

$$(Y_1, Y_0) \perp\!\!\!\perp W \mid X,$$

and *overlap*

$$0 < \Pr(W = 1|X) < 1, \text{ for all } x,$$

are together referred to as the assumption of strong ignorability (Rosenbaum and Rubin, 1983), which allows for identification of the average causal (treatment) effect:

$$\begin{aligned}\tau &= E(Y_1 - Y_0) = E[E(Y_1 - Y_0|X)] = E[E(Y_1|W = 1, X) - E(Y_0|W = 0, X)] \\ &= E[E(Y|W = 1, X) - E(Y|W = 0, X)].\end{aligned}$$

We see that the average causal effect can be estimated with the observed data by comparing treated and controls conditional on X and then taking the marginal expectation. Without unconfoundedness, we have in general no way of estimating the average causal effect.

A straightforward and perhaps the oldest way of adjusting for X is to divide the data into subclasses based on the covariate values, i.e. to perform standardisation. Although intuitive, this method has two practical caveats. First, to eliminate the bias several subclasses may be necessary. Second, the number of subclasses grows dramatically as the number of covariates in X increases. In both cases, the consequence is subclasses with sparse data with the result of unreliable estimates.

An alternative estimation strategy is to use matching. In matching each individual is matched with one or several individuals having similar values on X but are in the opposite treatment group. By comparing the outcomes in matched treated and controls we are able to estimate the average causal effect. An issue is that matching on the covariates is also subject to inherent problems

of sparse data. If individuals with identical values of X cannot be found, which is likely the case when X contains a large number of covariates, and certain if we have at least one continuous covariate, matching will not perfectly control for X. This results in biased estimation due to incomplete matching. Stuart (2010) provides a comprehensive overview of matching methods. Closely related to matching is nonparametric kernel regression, but again, as the number of continuous covariates increases, the rates of convergence of kernel methods deteriorate.

One solution to overcome the multidimensionality problem when estimating the conditional means is to assume some parametric model. For instance, if the conditional means are functions linear in the parameters, we can simply use the ordinary least squares estimator. This is also the most common method used today, but some researchers would argue that such parametric assumptions are bold.

1.2 Estimators using the propensity score

Instead of comparing treated and controls with the same values on all covariates, Rosenbaum and Rubin (1983) show that it suffices to condition on the conditional probability of assignment to a particular treatment given a vector of observed covariates, $p(X) \equiv \Pr(W = 1 \mid X)$. The scalar function $p(X)$ is called the propensity score, which enables us to reformulate the previously stated unconfoundedness assumption to $(Y_1, Y_0) \perp\!\!\!\perp W \mid p(X)$. Consequently,

$$\tau = E(Y_1 - Y_0) = E(E[Y|W = 1, p(X)] - E[Y|W = 0, p(X)]).$$

The estimation of the average causal effect may still involve estimating conditional means, but recalling that $p(X)$ is a scalar there is no dimensionality problem. An option is of course to use some parametric regression, but since we only condition on a scalar, non-parametric methods lend themselves particularly well to this setting. Hence, we will not discuss parametric estimation of the conditional means further.

The analogue of subclassification on the covariates is subclassification of the sample using $p(X)$. Although not considering subclassification in this thesis, it is instructive to include it here as an example of the most basic propensity score-based estimator. This estimator creates subclasses with individuals that are homogenous in the propensity score, and if the propensity score is constant within each subclass, then the covariates are independent of the treatment indicator. In each subclass the data could be interpreted as coming from a completely randomised experiment, removing all bias due to differences in the covariates. In practice the propensity score is not constant within each subclass and bias remains. Nonetheless, Rosenbaum and Rubin (1984) show that five subclasses can often remove over 90% of the bias due to each covariate.

Instead of subclassification, we may consider matching on the propensity score. Although several propensity score matching estimators exist (e.g., with or without replacement, with or without caliper), they all impute missing the potential outcome using the outcomes of individuals of the opposite treatment group. Paper I-III include a nearest-neighbour estimator with replacement. In general there is a trade-off between bias and variance when selecting the number of matches. A single match results in the least bias, but perhaps at the cost of losing some precision. A reason for the popularity of the matching estimators is that they are intuitive to use for practitioners with a smoothing parameter (i.e. the number of matches) that is easy to interpret. In fact, the number of matches is similar to bandwidth selection in kernel regression, and kernel regression can be seen as a special case of matching.

The propensity score is one of several balancing scores (Rosenbaum and Rubin, 1983), i.e. functions of the observed covariates such that the conditional distribution of X is the same for the treated and controls. However, the propensity score is also a probability. This leads to a class of estimators based on propensity score weighting that is similar to sample weighting proposed by Horvitz and Thompson (1952). It can be shown that the average causal effect can be identified with

$$\tau = E(Y_1 - Y_0) = E\left(\frac{WY}{p(X)}\right) - E\left(\frac{(1-W)Y}{1-p(X)}\right).$$

In this setting the propensity score is used to generate the $1/p(X) - 1$ missing potential outcomes with similar characteristics, which creates a pseudo-population in which unconfoundedness holds. Thus, the inverse probability weighting corrects for disproportionality of the observed responses with respect to the potential outcomes in the population. Inverse probability weighting estimators follow from the identification formula above. In this thesis, Paper I-II study a normalised version suggested by Hirano et al. (2003).

A special case of inverse probability weighting is an augmented inverse probability weighting estimator also referred to as the doubly robust (DR) estimator (see e.g., Robins et al., 1994; Lunceford and Davidian, 2004; Cao et al., 2009). This estimator combines inverse probability weighting with two regression models where for each treatment the outcome is regressed on the covariates. As long as either the propensity score model or the outcome regression models are correctly specified, the average causal effect will be consistently estimated. The DR estimator is examined in Paper II.

Estimation of the propensity score has thus far not been mentioned. Indeed, since the propensity score is rarely known to researchers, estimation is important in practice. Still, if consistently estimated, we can estimate the average causal effect using the same estimators as described above. The propensity score may be estimated using non-parametric regression, resulting in a completely non-parametric estimator of the average causal effect. More common, however, is to use parametric regression, such as a logit or probit model, and

the estimator for the average causal effect is then considered semi-parametric. Depending on the propensity score-based estimator, the parametric specification can be more or less crucial for estimation of the average causal effect (e.g., Millimet and Tchernis, 2009 and Waernbaum, 2012), but in general propensity score-based estimators show greater robustness compared with entirely parametric estimators for the average causal effect. In this thesis we confine our focus to propensity scores estimated with logistic regression. In Paper IV some properties of logistic regression as such are studied.

One important feature of propensity score-based estimators is that they allow for a clear separation between design and analysis. Often neglected in observational studies, this enables the researcher to focus entirely on modelling the estimated propensity score so that the conditional distributions of X are similar in both treatment groups. When the researchers are satisfied with the covariate balance, they may turn their attention to the outcome. This an appealing trait in that it mimics experimental design and does not require access to the outcome data (Rubin, 2001).

2. Summary of papers

2.1 Paper I

Large databases provide researchers with wonderful opportunities when attempting to estimate causal effects of treatments. Large databases also implies that it is common for researchers to have access to data containing several covariates that are correlated. For instance, many covariates could be measuring health or socio-economic status of an individual. The covariates then describe the same characteristic of the individual, and the researcher could choose to include one or more of the correlated variables in the analysis. Another reason for correlation between variables is when having an index constructed from other variables as a separate variable (e.g., the sequential organ failure assessment score). A common practice among researchers is to include both the index itself and some of the variables included in the index in the analysis, without considering that the variables and the index are correlated.

Paper I, entitled *Effects of correlated covariates on the asymptotic efficiency of matching and inverse probability weighting estimators for causal inference*, is the first study to investigate how correlation between covariates influences estimators using the propensity score. Although Stuart (2010) conjectures that for propensity score matching estimators collinearity in a propensity score model is of no standard concern, this supposition is done without any formal support.

We study two commonly used propensity score-based estimators: the normalised inverse probability weighting estimator (IPW) suggested by Hirano et al. (2003),

$$\hat{\tau}_{\text{IPW}} = \left(\sum_{i=1}^{N} \frac{W_i}{p(X_i)} \right)^{-1} \sum_{i=1}^{N} \frac{W_i Y_i}{p(X_i)} - \left(\sum_{i=1}^{N} \frac{1-W_i}{1-p(X_i)} \right)^{-1} \sum_{i=1}^{N} \frac{(1-W_i)Y_i}{1-p(X_i)},$$

with asymptotic variance σ^2_{IPW} and the simple nearest-neighbour propensity score matching estimator (Abadie and Imbens, 2006) which is defined as follows. For individuals i and i' from opposite treatment groups we denote the distance $d_{ii'}$,

$$d_{ii'} = \mid p(X_i) - p(X_{i'}) \mid,$$

where for each i, we denote by J_i a set $J_i = \{1, 2, \ldots, i', \ldots, M\}$ of indices of the M individuals with the smallest order statistics $d_{i(i')}$, $i' \leq M$. The matching

estimator, matching treated and controls to a fixed number of matches may then be written

$$\hat{\tau}_{\text{M}} = \frac{1}{N}\sum_{i=1}^{N} W_i(Y_i - \hat{Y}_{0i}) + (1 - W_i)(\hat{Y}_{1i} - Y_i).$$

where $\hat{Y}_{0i} = \sum_{i' \in J_i} Y_{i'}/M$ and $\hat{Y}_{1i} = \sum_{i' \in J_i} Y_{i'}/M$ are the observed response means for the M individuals with the smallest absolute difference in the propensity score. We denote the asymptotic variance of the matching estimator σ^2_{M}.

To study the influence of the correlation on the asymptotic variance of the estimators we assume a data-generating process with potential outcomes and a logit that are linear in the parameters, constant causal effect, and normally distributed covariates.

First, if all correlated covariates are positively related to the treatment and the outcome (or if they are all negatively related), then an increase in the correlation between the covariates leads to an increase in σ^2_{IPW}, whereas σ^2_{M} may increase or decrease. Therefore, a main contribution of this paper is that propensity score-based estimators may respond differently to a change in the correlation. Second, a study of the asymptotic relative efficiency of the estimators reveals that the matching estimator is less affected by an increase in the correlation compared with the IPW estimator. In fact, in some cases the IPW estimator is very sensitive to an increase in the correlation, which occurs if the correlation increases an already strong treatment assignment. In such a case increasing the correlation may result in a propensity score distribution with values close to 0 or 1, inflating the variance of the estimator. Third, numerical results demonstrate that an increase in the correlation between covariates is for almost all cases more beneficial for the matching estimator relative to the IPW estimator. Fourth, we see that the strength of the confounding towards the outcome and the treatment also plays an important role. This can be seen as similar to the findings in earlier investigations (Brookhart et al., 2006) that concluded that including a covariate that is strongly related to the treatment assignment but only weakly related to the outcome may increase the variance of a propensity score-based estimator.

However, even though we show how correlation is present in components of the asymptotic variances, the effect of the correlation depends on how the covariates are related to the potential outcomes and the treatment. The effect can therefore be difficult to predict in a given dataset. The findings in Paper I are mainly theoretical rather than practical. Thus, to provide guidance to how data analysts should treat correlated covariates in practise further studies are needed. Paper I provides motivation for further scrutiny of the impact of correlation on estimators of causal effects to assist empirical scientists to use strategies that improve efficiency.

2.2 Paper II

The second paper, *Correlation and efficiency of propensity score-based estimators for average causal effects*, is an extension of Paper I but containing several important additions. We investigate the behaviour under the more realistic setting of using the estimated propensity score, $\hat{p}(X)$. We also include a third estimator, the DR estimator, in the analysis. Furthermore, we study the effect of correlation on the behaviour of the estimators under less restrictive assumptions, considering several data-generating processes (including the omission of a confounder and the inclusion of an irrelevant covariate).

The IPW and matching estimators in Paper I remain the same, except that we replace $p(X)$ with $\hat{p}(X)$. The DR estimator is defined as follows (Lunceford and Davidian, 2004). Let $m_w(X,\beta_w) = E(Y|W=w,X)$ be the regression of Y on X in group w and let $\hat{\beta}_w$ be an estimator for the regression parameter β_w using only subjects within group w. The DR estimator is

$$\begin{aligned}\hat{\tau}_{\text{DR}} =& \frac{1}{N}\sum_{i=1}^{N}\frac{W_iY_i-(W_i-\hat{p}(X_i))\,\hat{m}_1\left(X_i,\hat{\beta}_1\right)}{\hat{p}(X_i)}\\ &-\frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i+(W_i-\hat{p}(X_i))\,\hat{m}_0\left(X_i,\hat{\beta}_0\right)}{1-\hat{p}(X_i)}.\end{aligned}$$

The asymptotic variances of the estimators using the estimated propensity score are

$$\sigma^2_{\text{IPW},\hat{p}} = \sigma^2_{\text{IPW}} - a'I^{-1}a, \quad \sigma^2_{\text{DR}} = \sigma^2_{\text{IPW}} - b, \quad \sigma^2_{\text{M},\hat{p}} = \sigma^2_{\text{M}} - c'I^{-1}c,$$

where a and c are vectors, I is the Fisher information matrix of the logistic regression, and b is a positive scalar. For the IPW and matching estimator, the second term of the variance is a correction due to the estimation of the propensity score.

Using the same data-generating process as in Paper I, but with the estimated propensity score instead, the findings are comparable to those in Paper I. If the correlated covariates are all positively related to the treatment and the outcome (or if they are all negatively related), then an increase in the correlation between the covariates leads to an increase in $\sigma^2_{\text{IPW},\hat{p}}$, an increase in $\sigma^2_{\text{DR},\hat{p}}$, and an increase or a decrease in $\sigma^2_{\text{M},\hat{p}}$. We also observe that, in comparison with the matching estimator, the variance of the IPW estimator is more extremely affected by a change in correlation. The DR estimator, however, is similar to the matching estimator regarding the sensitivity to a change in correlation, albeit in different directions.

When extending the analysis we find that in a setting with non-constant treatment effect, we have cases in which an increase in the correlation may in fact decrease the asymptotic variances of the IPW and DR estimator. Also, when studying how the mean squared error (MSE) is affected by correlation

when omitting a confounder we conclude for the DR and matching estimator that all confounders should always be included in the propensity score model, regardless of the correlation. As for the IPW estimator, we observe that in some cases, when the treatment assignment is strong, it is beneficial in terms of the MSE to omit a confounder from the propensity score model. However, in most cases all covariates should be included for the IPW estimator as well.

2.3 Paper III

When analysing data to estimate the average causal effect researchers are faced with several decisions, for example, a model must be chosen, an estimator must be selected and those covariates to be included in the model have to be determined. In the third paper, *Estimating the variance of a propensity score matching estimator: A new look at right heart catheterisation data*, we address one particular aspect of the decision process for researchers deciding to use a propensity score matching estimator with replacement, namely how to efficiently estimate the variance.

Recall from the summary of Paper II that the asymptotic variance of the matching estimator can be written

$$V = \sigma^2 - c' I^{-1} c$$

when matching is based on the estimated propensity score. For ease of exposition, view σ^2 as a weighted average of conditional variances $V(Y|W, p(X))$ and the vector c as a weighted average of conditional covariances between the outcome and covariates, $\text{cov}(X, Y|W, p(X))$. The weights in σ^2 and c are different, but both in both cases the weights are constructed from the propensity score. Because the Fisher information matrix, I, and the weights are already given from the estimation of the average causal effect, what is left for the researcher to decide on when estimating the variance is how to estimate $V(Y|W, p(X))$ and $\text{cov}(X, Y|W, p(X))$.

Abadie and Imbens (2012) suggest that the conditional variances and covariances are estimated using matching with replacement. We denote the estimator

$$\hat{V}_{LL'L''} = \hat{\sigma}^2_L - \hat{c}'_{L'L''} \hat{I}^{-1} \hat{c}_{L'L''},$$

where L is the number of matches used to estimate $V(Y|W, p(X))$. To be more specific, L is the number matches selected (including the unit being matched on) in the same treatment group as the unit itself. Depending on the treatment group, $\text{cov}(X, Y|W, p(X))$ is estimated with either L' matches selected from the same treatment group (including the unit being matched on), or L'' number of matches selected from the opposite treatment group. The guidance given by Abadie and Imbens (2012), and suggested without further motivation, is that typically $L = 2$, $L' = 2$, and $L'' = 2$ which results in $\hat{V}_{222}$. Abadie

and Imbens (2008) studied the choice of L when estimating conditional variances in paired experiments, but in a restricted setting and without considering propensity score matching.

Therefore, a main contribution of this paper is the investigation of how to select the number of matches when estimating the variance. Furthermore, we study the alternative residual-based estimator,

$$\hat{V}_r = \hat{\sigma}_r^2 - \hat{c}_r' I^{-1} \hat{c}_r,$$

which uses non-parametric local linear estimation for the conditional variances and covariances.

The performance of the estimators with respect to bias and MSE is evaluated via a simulation study under 240 scenarios (five propensity score designs, 16 outcome models and three sample sizes).

When estimating σ^2, the simulations show that in terms of MSE, L should be larger than 2, and preferably in the neighbourhood of 5 to 15 matches. Regarding bias, the finite sample performance of the estimators is poor for the sample size $N = 250$, and considerable bias remains for the sample size $N = 1000$. Focusing on the largest sample size, $N = 5000$, we suggest that the best choice in terms of bias and MSE is $L = 9$. The residual-based estimator performs in general well, but displays a large bias in settings with extreme propensity scores. However, this is not necessarily a negative trait since it forces the practitioners to pay special attention to the overlap assumption.

The performance of the estimators for $c' I^{-1} c$ is for small sample sizes poor regarding bias, particularly when $c' I^{-1} c$ is small. Even for the largest sample size substantial bias remain for most estimators. The estimator with the lowest MSE and least bias is the residual-based estimator, but in larger samples an option is to use the estimator proposed by Abadie and Imbens (2012) letting $L' = L'' = 9$.

Two other estimators were included in the study, one assuming a constant variance of the error term of the outcome and one other local linear estimator. Both of these estimators were unreliable in several settings, however.

To illustrate the importance of variance estimation we evaluate the effect of RHC during the first 24 hours in an intensive care unit on 30-day survival. The analysis is based on the SUPPORT study which collected data from 5735 patients who received intensive care at five US teaching hospitals between 1989 and 1994. The estimated average causal effect is -0.0343 or -0.0388 depending on the propensity score specification, but we observe a large variation in the estimates of σ^2 and $c' I^{-1} c$. Therefore, we show either a significant negative effect on survival or no effect depending on how the variances are estimated. Still, when selecting an estimator for the variance based on the simulation results, i.e. putting $L = L' = L'' = 9$ or choosing $\hat{V}_r$, we find that RHC has no effect on survival. This result is in accord with Tan (2006) and a meta-analysis of randomised studies by Shah et al. (2005), but is contrary to the conclusions in Connors et al. (1996), Hirano and Imbens (2001), Li

et al. (2008), and Crump et al. (2009), who all find small to moderate, though significant, negative effects of RHC on patient survival.

2.4 Paper IV

The main contribution of the fourth and final paper of the thesis, *Some approximations of the logistic distribution with application to the covariance matrix of logistic regression*, is that we provide an analytic expression of the covariance matrix of a logistic regression with normally distributed random regressors. More specifically, we show that the asymptotic covariance matrix of the maximum likelihood estimators $\hat{\mu}$ and $\hat{\gamma}$ of the logistic model $\text{logit}[p(Z)] = \mu + X'\gamma = Z$, where X is a random vector from a multivariate normal distribution with zero mean and covariance matrix Σ, can be formulated

$$\Lambda = \begin{pmatrix} \Lambda^{11} & \Lambda^{12} \\ \Lambda^{21} & \Lambda^{22} \end{pmatrix}, \text{ where}$$

$$\Lambda^{11} = \frac{E[f(Z)Z^2] - 2E[f(Z)Z]\mu + E[f(Z)]\mu^2}{E[f(Z)]E[f(Z)Z^2] - (E[f(Z)Z])^2},$$

$$\Lambda^{21} = -\gamma \frac{E[f(Z)Z] - E[f(Z)]\mu}{E[f(Z)]E[f(Z)Z^2] - (E[f(Z)Z])^2}, \ (\Lambda^{12} \text{ is the transpose of } \Lambda^{21})$$

$$\Lambda^{22} = \frac{\Sigma^{-1}}{E[f(Z)]} - \frac{\gamma\gamma'}{E[f(Z)]\sigma^2} + \frac{\gamma\gamma' E[f(Z)]}{E[f(Z)]E[f(Z)Z^2] - (E[f(Z)Z])^2},$$

where $f(\cdot)$ is the standard logistic density function. Although beyond the scope of the present thesis, this result may be of potential use for sample size calculations (such as Hsieh et al., 1998), or when investigating the properties of logistic regression and comparing it with other methods (such as Efron, 1975 and Courvoisier et al., 2011).

In general, however, the expectations in Λ cannot be solved without resorting to numerical methods. To increase the applicability of Λ, we substitute $f(\cdot)$ with a two-component normal mixture density that closely resembles $f(\cdot)$. We are then able to provide an approximate closed form expression for the asymptotic covariance matrix Λ. Numerical results show that the approximation of Λ generally works well when considering the relative error between the true and approximative covariance matrix. Further, a simulation study demonstrates that in most cases the coverage rate when using the approximative covariance matrix to calculate standard errors is close to the nominal level of 0.95.

The choice of a two-component normal mixture density function is motivated by a numerical comparison of some approximations of the standard logistic distribution, and its density. It is well-known that the normal distribution approximates the logistic distribution fairly well (Haley, 1952), but it

is also known that the heavier-tailed t-distribution improves the approximation (Mudholkar and George, 1978) and that a normal mixture distribution can be constructed to approximate any distribution arbitrary well (Sorenson and Alspach, 1971). We extend previous results by considering approximations of both the distribution and density function and by regarding both the maximum absolute deviation between the functions and the square root of the sum of all squared deviations. The conclusion is that the two-component mixture model is comparable to using the t-distribution as an approximate distribution in terms of the approximation errors.

By using a more sophisticated optimisation method than the one used in this study, the unpublished work by Monahan and Stefanski (1989) reaches a somewhat smaller maximum absolute deviation error between the logistic distribution and the normal mixture distribution (i.e. 0.0005 vs. 0.0007). However, the authors do not consider approximation of the density, which is needed for the expectations in Λ, nor do they consider minimisation with respect to the square root of the sum of all squared deviations.

Acknowledgements

I often find myself thinking in terms of "what if" when reflecting on past experiences in life (which must come to no surprise for those who read this thesis). Certainly, I cannot help wondering what would've happened if some things during my Ph.D. studies would've been different. What if Johan Lyhagen hadn't been my supervisor; always available to answer every question, open-minded, giving me freedom to pursue my ideas yet providing all the necessary support when I needed it. And what if Ingeborg Waernbaum hadn't been my co-supervisor; with contagious enthusiasm, explaining to me the A to Z of writing scientific articles, seeing potential where I saw none. You have both helped me to reach the highs and overcome the lows of my Ph.D. studies and I am sure that this thesis would not have been possible without your valuable guidance. For that I am very grateful.

It is a privilege to have been a part of the Department of Statistics and I have felt nothing but friendship and encouragement from everyone (despite my previous studies in Lund). I would like thank my co-supervisor Rolf Larsson. Your perfect sense for details has significantly improved this thesis. Lisbeth Hansson, thank you for involving me in the teaching and that you patiently listen to every silly idea that I have regarding pedagogy. Lars Forsberg, Katrin Kraus and Inger Persson, thank you for lifting the spirit at the department with your enthusiasm. Bo Wallentin, I am grateful for your sincere and thoughtful advice. Adam Taube, your intellectual honesty is an inspiration. A thousand thanks to Eva Enefjord. I think that I am now finally able to fill out the form for travel expenses. Thommy Perlinger, you are a role model for teaching, but you are even more a great snooker companion, a good neighbour and a friend.

Of course, big thanks go to my fellow Ph.D. students, past and present. What if you hadn't been here? Without doubt, there would have been less moments with laughter, joy and beer. Jim, our endless discussions have contributed more to my Ph.D. studies than you might imagine. Martin and Myrsini, sharing office in the beginning was the perfect start. Xingwu, thank you for sharing your knowledge. To you (and everyone else at the department): never forget that as statisticians, we are on a mission.

I am also truly grateful to my friends. Besides providing numerous bed and breakfasts, some of you have (from time to time) showed an interest in my research. And to those of you who couldn't care less about the contents of this thesis, rest assured that you have still contributed by just being who you are, for which I am thankful. By the way, Tobias, I guess this thesis means that we didn't end up in Tibet doing research on the mysteries of the qi-force.

Finally, my deepest gratitude goes to my entire family. You provide the haven where I am able to recharge my batteries. Mum and dad, this thesis would never had come to a conclusion without your infinite generosity and support. My sister and brothers, thank you for your understanding, sense of humour and caring.

Becki, there are no words to express how grateful I am to you. You didn't hesitate to join me on this journey to the cold north. You have put up with me being physically and mentally absent. And you continue to bring happiness, beauty and adventure in my life. What if I hadn't met you?

Ronnie Pingel
Rönås, August 2014

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*, 235–267.

Abadie, A. and G. W. Imbens (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, 175–187.

Abadie, A. and G. W. Imbens (2012). Matching on the estimated propensity score. Harvard University and National Bureau of Economic Research. http://www.hks.harvard.edu/fs/aabadie/pscore.pdf. Retrieved 2014-08-01.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer (2006). Variable selection for propensity score models. *American Journal of Epidemiology 163*, 1149–1156.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*, 732–734.

Connors, A. F., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *Journal of the American Medical Association 276*, 889–897.

Courvoisier, D. S., C. Combescure, T. Agoritsas, A. Gayet-Ageron, and T. V. Perneger (2011). Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology 64*, 993–1000.

Cox, D. R. (1958). *Planning of experiments*. New York: Wiley.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*, 187–199.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association 70*, 892–898.

Haley, D. (1952). Estimation of the dosage mortality relationship when the dose is subject to error. *Applied Mathematics and Statistics Laboratory, Stanford University, Technical report* (15).

Hirano, K. and G. W. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology 2*, 259–278.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–1189.

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

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association 47*, 663–685.

Hsieh, F. Y., D. Bloch, and M. Larsen (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine 17*, 1623–1634.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*, 5–86.

Li, Q., J. S. Racine, and J. M. Wooldridge (2008). Estimating average treatment effects with continuous and discrete covariates: The case of Swan-Ganz catheterization. *The American Economic Review: Papers & Proceedings 98*, 357–362.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine 23*, 2937–2960.

Millimet, D. L. and R. Tchernis (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business and Economic Statistics 27*, 397–415.

Monahan, J. F. and L. A. Stefanski (1989). Normal scale mixture approximations to the logistic distribution with applications. *Institute of Statistics Mimeograph Series 1968*.

Mudholkar, G. and O. George (1978). A remark on the shape of the logistic distribution. *Biometrika 65*, 667–668.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. English translation by D. Dabrowska and T. Speed in *Statistical Science 5*, 465–472, 1990.

Neyman, J., K. Iwaszkiewicz, and S. Kolodziejczyk (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society 2*, 107–180.

Robins, J. M., A. Rotnitsky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*, 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1980). Discussion of randomization analysis of experimental data: the Fisher randomization test by D. Basu. *Journal of the American Statistical Association 75*, 591–593.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology 2*, 169–188.

Shah, M. R., V. Hasselblad, L. W. Stevenson, C. Binanay, C. M. O'Connor, G. Sopko, and R. M. Califf (2005). Impact of the pulmonary artery catheter in critically ill patients: Meta-analysis of randomized clinical trials. *Journal of the*

American Medical Association 294, 1664–1670.
Sorenson, H. W. and D. L. Alspach (1971). Recursive bayesian estimation using gaussian sums. *Automatica 7*, 465–479.
Stuart, E. A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science 25*, 1–21.
Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*, 1619–1637.
Waernbaum, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine*. forthcoming.