Estimation of Average Treatment Effects

Honors Thesis

Peter Zhang

Abstract

The estimation of average treatment effects is an important issue in economic evaluations of the impact of policy intervention on job employment and the effect of education and training on income. This honors thesis is concerned with studying different approaches to the estimation of an average treatment effect. Motivated by the works of Khandker, Koolwal, and Samad (2010) and Kreif, Grieve, Radice, and Sekhon (2013), we consider complete-case method, inverse probability weighting and propensity score matching methods using the propensity score, the regression method using two treatment-specific regression models, augmented inverse probability weighting and generalized method of moments methods using both the propensity score and treatment-specific regression models. These methods yield six different estimates of average treatment effects. Our empirical study is an application of these six estimates to the assessment of the National Supported Work job-training program; our analysis shows a positive impact of this job-training program with higher annual earnings for those participating in the program. We conduct a Monte Carlo simulation study that investigates these six estimates' performance. Overall, we find that the regression and generalized method of moments approaches produce least biased and most efficient estimates of an average treatment effect; yet the resulting estimates are most robust against functional form misspecification of the propensity score and treatment-specific regression models.

1. Introduction

Causality, i.e., the relationship between cause and effect, is a topic of great significance in economics and related fields such as medicine, epidemiology, and health science. For example, the impact of policy intervention on job employment, the effect of education and training on income, and the effectiveness of a drug (or whether it is effective at all) on a disease are all questions of causality. One important and commonly used measure of causality is the average treatment effect (ATE) for a binary policy or treatment on a scalar outcome, which is the mean outcome difference between

the treatment and control groups. Economic evaluations of policy or health care interventions in many observational studies often require identification and estimation of ATEs, which is challenging because randomized experiments cannot always be implemented.

The potential outcome framework, described in Khandker, Koolwal, and Samad (2010), provides a useful way of identifying and evaluating ATEs in observational studies. Under this conceptual framework, the problem of estimating an ATE becomes a missing data problem due to the absence of data on counterfactual outcomes (namely, outcomes from treated units had they not been exposed to the treatment). The presence of missing data in causal inference often leads to potential selection bias and thus complicates the relationship between cause and effect. To account for potential selection bias in treatment (program) participation that might affect evaluation of ATEs, policy makers need to isolate the effect of the policy intervention on outcomes from other observed and unobserved confounding factors affecting the outcomes. This implies that the evaluation of ATEs in observational studies often requires adjustment for differences in baseline covariates because treatment and control groups can be unbalanced on measured or unmeasured covariates.

As discussed in Khandker, Koolwal, and Samad (2010), there are a number of different approaches in the literature to handling missing counterfactuals in impact evaluation theory. Rosenbaum and Rubin (1983) introduce the notion of propensity score (PS) defined as the probability of assignment to treatment conditional on covariates. The PS gives rise to two different methods of estimating ATEs, the inverse probability weighting (IPW) method and the propensity score matching (PSM); both of which control the differences between treatment and control groups by adjusting baseline covariates and thus reduce the selection bias. The IPW method weighs each complete outcome by the inverse probability of itself being observed, whereas the PSM method constructs a matching counterfactual or control group that is as similar to the treatment group as possible in terms of observed covariates based on PS. Instead of modeling the probability of assignment on covariates, the regression method estimates the ATE by modeling the outcome on covariates, as discussed in Kreif, Grieve, Radice, and Sekhon (2013). Robins, Rotnitzky, and Zhao (1994) propose a hybrid of the PS and regression methods to estimate ATEs. Their method, called the augmented inverse probability weighting (AIPW) method, is doubly robust (DR) in the sense that it renders consistent estimates of ATEs if either one of the PS or regression models is correctly specified, but not necessarily both. Compared to the PS and regression methods, the AIPW method is less sensitive to the

misspecification of the PS or regression model. For other approaches to estimating the ATE in the literature, see, for example, Khandker, Koolwal, and Samad (2010).

The generalized method of moments (GMM) method, due to Hansen (1982), is an effective and widely used moments-based approach to parameter estimation in microeconometrics studies. The focus of this thesis is to explore an alternative approach to estimation of ATEs from a GMM perspective by exploiting the PS and regression models. The goal of this thesis is two-fold. The first goal is to apply the GMM-based estimate of ATEs to the program evaluation in econometrics by analyzing a real dataset from a randomized evaluation of a labor market training program, the National Supported Work (NSW) job-training program, originally analyzed by Lalonde (1986) and subsequently by Dehejia and Wahba (1999). It is also of interest to compare the GMM-based estimate of ATEs with other existing ATE estimates described in Khandker, Koolwal, and Samad (2010). Thus, our second goal is to perform a simulation study to examine the merits of the GMM-based and other related methods.

This thesis is ordered as follows. Section 2 describes various methodologies of estimating ATEs. The results of the analysis of the NSW job-training data are reported in Section 3. Section 4 presents simulations results comparing various ATE estimates. Finally, the thesis is concluded with discussion in Section 5.

2. Methodology

For the ith participant with $i = 1, \ldots, n$, let T_i represent the treatment assignment taking value 1 if participant i is treated and value 0 if participant i is not treated. Furthermore, let $Y_i(1)$ denote the potential outcome under treatment and $Y_i(0)$ denote the potential outcome when there is no treatment; then the ATE is given by ATE $= E[Y_i(1) - Y_i(0)]$. The potential outcomes $Y_1(1)$ and $Y_i(0)$ cannot be observed simultaneously for each participant i; instead, we can only observe $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$, the observed outcome from participant i. In addition to observe T_i and Y_i, we can also observe X_i, a set of other baseline covariates or characteristics of individual i, such as his or her household and local environment. Thus, the observed data is composed of (T_i, Y_i, X_i) for $i = 1, \ldots, n$. Without adjustment for baseline covariates, a complete-case estimator of ATE is

simply the mean difference in observed outcomes by treatment status:

$$\widehat{\text{ATE}}_{\text{CC}} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1-T_i) Y_i}{\sum_{i=1}^{n} (1-T_i)}.$$

It is widely known that $\widehat{\text{ATE}}_{\text{CC}}$ is biased in observational studies and requires adjustment for differences in baseline covariates.

One important assumption in causal inference is the unconfoundedness or conditional independence assumption (Rosenbaum and Rubin, 1983), which states that conditional on a set of observable covariates X_i that are not affected by treatment, the potential outcomes $Y_i(1)$ and $Y_i(0)$ are independent of treatment assignment T_i; symbolically, the unconfoundedness assumption can be expressed as

$$(Y_i(1), Y_i(0)) \perp T_i | X_i.$$

Under the unconfoundedness assumption, the propensity score, which is the probability of treatment assignment conditional on covariates, is given by $p(X_i) = P(T_i = 1|X_i) = P(T_i = 1|Y_i(1), Y_i(0), X_i)$. As in Kreif, Grieve, Radice, and Sekhon (2013), a commonly adopted parametric model for PS is the logistic regression model given by

$$p(X_i, \eta) = \frac{e^{\eta^T X_i}}{1 + e^{\eta^T X_i}},$$

where η is a vector parameter and can be estimated by the maximum likelihood estimate $\hat{\eta}$ based on the observed data (T_i, X_i), $i = 1, \ldots, n$.

As pointed by Kreif, Grieve, Radice, and Sekhon (2013), the IPW estimate of ATE is obtained below by reweighting the observed outcomes for treatment and control samples using the inverse of the estimated probability of the observed treatment:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} T_i w_i Y_i - \frac{1}{n} \sum_{i=1}^{n} (1-T_i) w_i Y_i,$$

where $w_i = 1/p(X_i, \hat{\eta})$. The validity of this IPW estimate relies on the correct specification of the PS model $p(X_i, \eta)$.

Kreif, Grieve, Radice, and Sekhon (2013) also consider employing two generalized linear models $g_1(X_i, \beta_1)$ and $g_0(X_i, \beta_0)$ to estimate ATE as given below:

$$\widehat{\text{ATE}}_{\text{REG}} = \frac{1}{n} \sum_{i=1}^{n} g_1(X_i, \hat{\beta}_1) - \frac{1}{n} \sum_{i=1}^{n} g_0(X_i, \hat{\beta}_0),$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are the maximum likelihood estimates or the least squares estimates of the parameter vectors β_1 and β_0. The validity of this regression estimate depends on the correct specification of the two treatment-specific regression models $g_1(X_i, \beta_1)$ and $g_0(X_i, \beta_0)$.

Combining the models for the PS and for the potential outcomes, Kreif, Grieve, Radice, and Sekhon (2013) consider the following AIPW estimate of ATE:

$$\widehat{\text{ATE}}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} T_i w_i [Y_i - g_1(X_i, \hat{\beta}_1)] - \frac{1}{n}\sum_{i=1}^{n}(1-T_i) w_i [Y_i - g_0(X_i, \hat{\beta}_0)]$$
$$+ \frac{1}{n}\sum_{i=1}^{n} g_1(X_i, \hat{\beta}_1) - \frac{1}{n}\sum_{i=1}^{n} g_0(X_i, \hat{\beta}_0).$$

The validity of this AIPW estimate relies on the correct specification of either the PS model or the outcome regression models (but not necessarily both); the AIPW estimate is thus doubly robust. Moreover, as pointed out by Kreif, Grieve, Radice, and Sekhon (2013), the AIPW estimate is semiparametrically efficient when models for the PS and for the potential outcomes are all correctly specified.

In this thesis, we apply the GMM method to estimate ATE based on the aforementioned PS and regression models. According to the GMM theory, the GMM estimate $\widehat{\text{ATE}}_{\text{GMM}}$ of ATE minimizes the quadratic function $U(\text{ATE})^T S^{-1} U(\text{ATE})$, where S is the estimated variance of $U(\text{ATE})$ and

$$U(\text{ATE}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{T_i w_i Y_i - (1-T_i) w_i Y_i - \text{ATE},\ (T_i - w_i^{-1})\Big(w_i g_1(X_i, \hat{\beta}_1),\ \frac{g_0(X_i, \hat{\beta}_0)}{1-w_i^{-1}},\ \frac{w_i \frac{\partial p(X_i, \hat{\eta})}{\partial \eta^T}}{1-w_i^{-1}}\Big)\Big\}.$$

Since the GMM estimate is constructed based on the PS and regression models, it is interesting to know if its validity is dependent on either one component or both components of these models; in other words, it is interesting to know if the GMM estimate is doubly robust. It is also of interest to compare the GMM estimate with the IPW, regression, AIPW, matching estimates of ATE in terms of their bias and root mean square error (RMSE) through a Monte Carlo simulation study.

3. Re-analysis of National Supported Work training program data

In this section, we revisit the National Supported Work job-training data collected by Lalonde (1986) for evaluating labor training programs for the substantially disadvantaged in the labor market. One interesting question is to determine if the National Supported Work job-training program is successful in increasing post-intervention earnings among those who participate in the program. Our objective is to apply the GMM method to the evaluation of the job-training program's impact and effect.

The specific data set we analyze here is the one discussed by Dehejia and Wahba (1999), which is a subset of the Lalonde (1986) data. The population is composed of men who were substantially disadvantaged (those who had very poor labor market histories) in the labor market and were eligible for the NSW training program. A sample of 445 men was collected from this population, of whom 185 were assigned to the job training program and 260 were assigned to the control group. Each observation of a man in this sample consists of data on baseline covariates (background characteristics), including age in years (`age`), years of education or schooling (`educ`), indicator variable for blacks (`black`), indicator variable for Hispanics (`hisp`), indicator variable for martial status (`married`), indicator variable for high school diploma (`nodegr`), pre-training real earnings in 1974 and 1975 (`re74` and `re75`), indicator variable for earnings in 1974 being zero (`u74`), indicator variable for earnings in 1975 being zero (`u75`), and indicator variable for treatment status (`treat`). The outcome variable of interest is the post-intervention labor market experience represented by the post-program real earnings in 1978 (`re78`). All earning variables `re74`, `re75`, and `re78` are in dollars.

Figure 1 displays the histograms of earnings for treatment (`treat` $= 1$) and control (`treat` $= 0$) groups, whereas Figure 2 exhibits the boxplots of earnings in 1978 for treatment and control groups. Both histograms and boxplots show that the distributions of the outcome (earnings in 1978) in the treatment and control groups are skewed to the right (high earnings). Furthermore, the two histograms indicate the high proportion of men with earnings around \$40,000 or less for the treatment group and the high proportion of men with earnings around \$25,000 or less for the control group, thus demonstrating an increasing of NSW post-training earnings in 1978. Indeed, the average annual earning in 1978 is \$6349.15 in the treatment group and \$4554.80 in the control group, yielding the simple difference $\$6349.15 - \$4554.80 = \$1794.34$ in sample means. This is simply the complete-case

estimator $\widehat{\text{ATE}}_{\text{CC}}$ of the average effect of the training program on annual earnings in 1978 without adjustment for differences in the aforementioned baseline covariates.

To calculate other estimates of ATE presented in the last section that adjust for all baseline covariates, we employ stepwise regression procedures to select the propensity score and treatment-specific regression models. The fitted propensity score is found to be

$$\log \frac{p(X, \hat{\eta})}{1 - p(X, \hat{\eta})} = -0.951897 + 0.005678 \times \texttt{educ}^2.$$

On the other hand, the two fitted treatment-specific regression models are given by

$$\begin{aligned}
g_1(X, \hat{\beta}_1) &= 8583.73 + 810.65 \times \texttt{age} - 875.90 \times \texttt{educ} - 27212.19 \times \texttt{hisp} - 9389.81 \times \texttt{married} \\
&\quad - 3439.67 \times \texttt{nodegr} - 0.6822 \times \texttt{re74} + 2.0413 \times \texttt{re75} - 87452.57 \times \texttt{u74} \\
&\quad - 71956.40 \times \texttt{u75} - 11.7292 \times \texttt{age}^2 - 0.0001 \times \texttt{re75}^2 + 1266.11 \times \texttt{age} \times \texttt{hisp} \\
&\quad - 0.1170 \times \texttt{age} \times \texttt{re75} + 6805.45 \times \texttt{educ} \times \texttt{u74} - 5737.69 \times \texttt{educ} \times \texttt{u75} \\
&\quad + 2.5092 \times \texttt{married} \times \texttt{re75} + 11556.68 \times \texttt{married} \times \texttt{u75} + 27992.80 \times \texttt{nodegr} \times \texttt{u74} \\
&\quad - 27513.17 \times \texttt{nodegr} \times \texttt{u75} + 0.0001 \times \texttt{re74} \times \texttt{re75} + 1.8222 \times \texttt{re74} \times \texttt{u75}, \\
g_0(X, \hat{\beta}_0) &= 7694.11 + 72.17 \times \texttt{age} - 3506.01 \times \texttt{black} + 4452.92 \times \texttt{hisp} - 13025.27 \times \texttt{married} \\
&\quad - 0.5059 \times \texttt{re74} + 0.9641 \times \texttt{re75} - 2277.99 \times \texttt{u74} - 290.08 \times \texttt{age} \times \texttt{hisp} \\
&\quad + 12973.58 \times \texttt{black} \times \texttt{married} + 0.5806 \times \texttt{black} \times \texttt{re74} - 1.2528 \times \texttt{black} \times \texttt{re75} \\
&\quad + 13952.29 \times \texttt{hisp} \times \texttt{married}.
\end{aligned}$$

Table 1 presents six estimates of the average treatment effect of the NSW job-training program on annual earnings in 1978, including the matching estimate $\widehat{\text{ATE}}_{\text{M}}$ described in Kreif, Grieve, Radice, and Sekhon (2013). It is seen from Table 1 that the matching and complete-case methods produce appreciably larger estimates of ATE than the other four methods, all of them are positive. Therefore, we conclude that the NSW job-training program has a positive intervention on annual earnings in 1978

Since the NSW job-training program analysis displays positive impact on labor training and post-intervention earnings, this can help policymakers create more novel programs that will help the substantially disadvantaged so that they can return to or enter the labor force. This can help reduce

the cost of the government welfare program and lower the unemployment rate. This analysis can also encourage people who are currently unemployed to undergo job training to possibly pursue another career in a different field. This can not only significantly reduce the unemployment rate, but can also increase household family earnings to improve quality of life.

4. Monte Carlo simulation study

In this section, we conduct a Monte Carlo simulation study to examine the relative performance of complete-case, IPW, regression, AIPW, matching, and GMM methods for estimating ATE in terms of bias, standard deviation (SD), and root mean square error (RMSE); the matching method is described in Kreif, Grieve, Radice, and Sekhon (2013). Bias, SD, and RMSE provide useful information about the accuracy and the precision of the estimated ATE across the various methods. Throughout our simulation study, bias is calculated as the difference between the mean of the estimated ATEs and the true ATE, SD is the sample standard deviation of the simulated estimated ATEs, and RMSE is the square root of the mean square differences between the estimated and true ATEs. We consider eight scenarios and simulate 1000 datasets for each scenario, each with 2000 subjects, using the data generating process described below.

For each subject, let (Z_1, Z_2) and (Z_3, Z_4) be generated from independent bivariate normal distributions with $E(Z_1) = E(Z_3) = 2$, $E(Z_2) = E(Z_4) = 4$, $\text{Var}(Z_1) = \text{Var}(Z_2) = \text{Var}(Z_3) = \text{Var}(Z_4) = 1$, and $\text{Cov}(Z_1, Z_2) = \text{Cov}(Z_3, Z_4) = 0.2$; in other words, (Z_1, Z_2) and (Z_3, Z_4) are simulated from the following bivariate normal distributions:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\Big\{ \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \Big\}, \qquad \begin{pmatrix} Z_3 \\ Z_4 \end{pmatrix} \sim N\Big\{ \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \Big\}.$$

The treatment assignment indicator T is generated from two different Bernoulli distributions with parameters or propensity scores $P(T = 1)$ determined by two different logistic regression models with nonlinear terms in the logit, as described by logit models (I) and (II) below:

$$\text{logit}[P(T = 1)] = \frac{\log[P(T = 1)]}{1 - \log[P(T = 1)]} = 0.4 - Z_1 + 0.5Z_2 + 0.025Z_2^2 - 0.25Z_3 - 0.1Z_4, \qquad \text{(I)}$$

$$\text{logit}[P(T = 1)] = \frac{\log[P(T = 1)]}{1 - \log[P(T = 1)]} = 1.5 - 2Z_1 + Z_2 + 0.05Z_2^2 - 0.5Z_3 - 0.2Z_4. \qquad \text{(II)}$$

As pointed by Kreif, Grieve, Radice, and Sekhon (2013), propensity score (I) produces stable IPW weights, whereas propensity score (II) gives rise to unstable IPW weights in that a large number of true propensity scores are close to 0 or 1.

The potential outcomes $Y(1)$ and $Y(0)$ are, respectively, drawn from two normal distributions with common variance 0.2, namely, $Y(1) \sim N(\mu_1, 0.2)$ and $Y(0) \sim N(\mu_0, 0.2)$, where

$$\mu_1 = 9.4 + 0.1Z_1 - 0.05Z_2 + 0.05Z_3 - 0.05Z_4,$$

$$\mu_0 = 9 + 0.1Z_1 - 0.05Z_2 + 0.05Z_3 - 0.05Z_4.$$

These two treatment-specific regression models imply that the treatment effect is additive on average and that the true ATE is equal to 0.4.

As in Kreif, Grieve, Radice, and Sekhon (2013), we adopt a common functional form of misspecification for both the propensity score and the treatment-specific regression models. Instead of the true covariates or confounders Z_1, Z_2, Z_3, Z_4, their nonlinear functions X_1, X_2, X_3, X_4 are observed and are used for modeling the propensity score and outcome regression models. These observed covariates are defined as

$$\begin{aligned}
X_1 &= \exp\Big(\frac{Z_1}{3}\Big), \\
X_2 &= Z_2(1 + Z_1) + 10, \\
X_3 &= \Big(\frac{Z_3}{25} + 0.6\Big)^2, \\
X_4 &= \Big(\frac{Z_2}{10} + Z_4 + 20\Big)^2.
\end{aligned}$$

We consider eight simulation scenarios, resulting from different combinations of stable and unstable propensity scores (I) and (II), correct and incorrect specifications of propensity scores, and correct and incorrect specifications of treatment-specific outcome regression models. Specifically, for each of the logit models (I) and (II), we consider four scenarios: (a) the propensity score and regression models are all correctly specified; (b) the propensity score is correctly specified but the two regression models are misspecified; (c) the two regression models are correctly specified but the propensity score is misspecified; (d) the propensity score and regression models are all misspecified. Reported in Tables 2 to 9 are simulation results comparing the GMM estimate with the complete-case, IPW, regression, AIPW, and matching estimates of ATE under these eight scenarios, in terms

of their biases, standard deviations (SDs), and root mean square errors (RMSEs). Also included in Tables 2 to 9 are the relative bias and the relative root mean square errors, which are, respectively, calculated as the bias of the indicated estimate divided by the bias in absolute value for the GMM estimate and the RMSE of the indicated estimate divided by the RMSE of the GMM estimate. The simulation results in Tables 2 to 9 are based on 1000 replications with sample size 2000 for the eight different scenarios and are summarized in the following four subsections. This section is closed with a subsection providing an overall summary of the simulation results.

4.1. *Correct specification of stable propensity score (I)*

Table 2 shows simulation results when both propensity score (I) and two treatment-specific regression models are correctly specified. In this case, all estimates of ATE are valid except for the complete-case estimate because it does not adjust for differences in baseline covariates in observational studies and is thus invalid for the estimation of ATE. Indeed, the bias of the complete-case estimate is significantly larger than those of the other estimates and is, in particular, at least 53 times larger than the GMM estimate in absolute value. The regression, AIPW, and GMM estimates provide similar and the smallest standard deviations and root mean square errors than other estimates, whereas the IPW estimate has large standard deviation and root mean square error with RMSE about 7 times as large as the GMM estimate. In terms of having smaller SD and RMSE, the performance of the matching estimate is better than the IPW estimate, but is not as good as the regression, AIPW, and GMM estimates. Thus, the regression, AIPW, and GMM estimates perform the best under correctly specified stable propensity score (I) and correctly specified treatment-specific regression models.

By contrast, Table 3 displays simulation results for correct specification of propensity score (I) and incorrect specifications of two regression models, so that all estimates of ATE are valid except for the complete-case and regression estimates. In fact, the regression estimate has a larger bias than the IPW, AIPW, matching, and GMM estimates and has a bias about 6 times as large as the GMM estimate. However, the regression estimate has the smallest standard deviation, rendering a root mean square error as large as those of the AIPW and GMM estimates and yet smaller than the IPW and matching estimates. Although the IPW estimate is a valid estimate of ATE when propensity score (I) is correct, its RMSE is significantly higher than those of the other estimates and is noticeably about 7 times higher than those of the regression, AIPW, and GMM estimates.

Thus, the regression, AIPW, and GMM estimates perform the best under correctly specified stable propensity score (I) and misspecified treatment-specific regression models, though the IPW estimate should perform better than the regression estimate according to the standard econometrics theory. This implies that the regression estimate is quite robust against the possible misspecification of treatment-specific regression models.

4.2. *Incorrect specification of stable propensity score (I)*

Table 4 shows simulation results when two treatment-specific regression models are correctly specified, but propensity score (I) is misspecified. In this case, the regression and AIPW estimates are known to be valid estimates of ATE. The biases of the complete-case and IPW estimates are much larger than those of the other estimates; in particular, the bias of the IPW estimate is about 469 times higher than the GMM estimate in absolute value. Furthermore, the bias of the GMM estimate is smaller than those of the AIPW and matching estimates, suggesting that the GMM estimate is valid and appears to be doubly robust. In terms of estimation precision, the regression estimate has the smallest SD and RMSE, whereas the GMM estimate follows next. Moreover, the IPW estimate has the highest SD and RMSE, whose RMSE is about 188 times higher than the GMM estimate, implying that the IPW estimate is quite sensitive to the misspecification of propensity score (I). On the other hand, the matching estimate works quite well in that it has slightly larger SD and RMSE than the regression and GMM estimates. Thus, the regression, GMM, and matching estimates perform the best under misspecified stable propensity score (I) and correctly specified treatment-specific regression models, suggesting that the GMM and matching estimates are quite robust against the possible misspecification of propensity score (I).

Table 5 presents simulation results for neither correct specification of propensity score (I) nor correct specifications of two regression models, so that all estimates of ATE are invalid. The biases of the regression, matching, and GMM estimates are significantly smaller than those of the other estimates; the IPW estimate has the highest bias, which is about 212 times as high as that of the GMM estimate. Moreover, the regression, matching, and GMM estimates have much smaller standard deviations and root mean square errors than other estimates; the regression estimate has the lowest SD and RMSE, followed by the GMM and matching estimates. By contrast, the IPW estimate has the highest SD and RMSE, significantly higher than other estimates with its RMSE about 181

times higher than the GMM estimate. Thus, the regression, GMM, and matching estimates perform the best when neither propensity score (I) nor treatment-specific regression models are correctly specified. These results indicate that the regression, GMM, and matching estimates are more robust against model misspecifications than the IPW and AIPW estimates, whereas the IPW estimate relies heavily on the correct specification of propensity score (I) and is hence least robust against its misspecification.

4.3. *Correct specification of unstable propensity score (II)*

For unstable IPW weights, Table 6 presents simulation results for correct specifications of both propensity score (II) and treatment-specific regression models. In this case, the complete-case estimate is the only invalid estimate of ATE; it has a very large bias that is at least 49 times as large as the GMM estimate in absolute value. The IPW estimate also has a quite large bias, which is about 7 times larger than the GMM estimate in absolute value. All other estimates have very small biases, though the regression estimate has the smallest one. In terms of estimation precision, the regression estimate has the lowest SD and RMSE, followed by the GMM, AIPW, and matching estimates. The consequence of using unstable IPW weights reflects that the IPW estimate has the highest SD and RMSE; its RMSE is about 19 times higher than that of the GMM estimate. This is in sharp comparison of the IPW estimate using the stable IPW weights in Table 2, whose RMSE is about 7 times higher than that of the GMM estimate. These results indicate that the IPW estimate is sensitive to unstable IPW weights, even though propensity score (II) is correctly specified.

For unstable IPW weights, Table 7 shows simulation results for correct propensity score (II) and incorrect treatment-specific regression models. In this case, the complete-case and regression estimates are invalid estimates of ATE; their biases are significantly higher than those of the AIPW, GMM, and matching estimates, but the bias of the regression estimate is lower than that of the IPW estimate. On the other hand, the regression estimate has the lowest SD and RMSE, followed by the GMM, matching, and AIPW estimates; there is a 35% reduction in RMSE compared to the GMM estimate. All these estimates have much smaller SDs and RMSEs than those of the IPW estimate. Although the IPW estimate is a valid estimate of ATE using the unstable IPW weights under correct propensity score (II), its bias and RMSE are about 7 times and 19 times as high as those of the GMM estimate, respectively. These results affirm that the IPW estimate is sensitive to unstable

IPW weights under correct propensity score (II), whether or not treatment-specific regression models are correctly specified.

4.4. *Incorrect specification of stable propensity score (II)*

For unstable IPW weights, Table 8 displays simulation results for incorrect propensity score (II) and correct treatment-specific regression models. In this case, the regression and AIPW estimates are known to be valid estimates of ATE. The biases of the regression and GMM are appreciably smaller than that of the matching estimate, which is in turn much smaller than the AIPW, complete-case, and IPW estimates, suggesting that the GMM estimate is valid and appears to be doubly robust. Moreover, the bias of the IPW estimate is about 2222 times as high as the GMM estimate. As for the estimation precision, the regression, matching, and GMM estimates have much smaller SDs and RMSEs than the AIPW and IPW estimates; the RMSE of the IPW estimate is about 670 times larger than the GMM estimate. By contrast, the regression estimate has the smallest RMSE; there is a 58% reduction in RMSE compared to the GMM estimate. These results demonstrate that the performance of the IPW estimate is substantially worse than other estimates in terms of having much higher bias, SD, and RMSE, when incorrect propensity score (II) is employed to generate unstable IPW weights and when two treatment-specific regression models are correctly specified.

Finally, Table 9 shows simulation results for both incorrect propensity score (II) and incorrect treatment-specific regression models with unstable IPW weights. As in Table 5, all estimates of ATE are invalid in this case. The matching estimate has the lowest bias, followed by the GMM and regression estimates; the biases of these three estimates are considerably lower than those of the other estimates. With unstable IPW weights, the IPW estimate has the highest bias, which is about 622 times as high as that of the GMM estimate and is thus substantially higher than those of the other estimates. In terms of estimation precision, the regression estimate has the lowest SD and RMSE, followed by the matching and GMM estimates; all three estimates have significantly smaller SDs and RMSEs than other estimates. Using unstable IPW weights, the IPW estimate has the highest SD and RMSE, substantially higher than other estimates; its RMSE is about 660 times higher than the GMM estimate. These results again confirm that whether or not treatment-specific regression models are correctly specified, the IPW estimate can produce substantially large bias, SD, and RMSE and can thus perform wildly by employing unstable IPW weights generated by incorrect propensity score (II).

4.5. *Summary*

Combining simulation results in Tables 2–9 reveals that the regression, GMM, and matching estimates have lower biases, SDs, and RMSEs than other estimates in most cases. Overall, the regression estimate has the lowest SD and RMSE across the eight scenarios, even when two treatment-specific regression models are misspecified. The GMM estimate also performs quite well and appears to be doubly robust, having smaller bias, SD, and RMSE than the other four estimates in most cases. The matching and AIPW estimates always reduce bias, SD, and RMSE compared to the IPW estimate. The IPW estimate is a valid estimate of ATE when the propensity score is correctly specified; its performance is, however, very unstable. It often produces very large biases, SDs, and RMSEs across most of the eight scenarios, especially when the propensity score is misspecified or generates unstable IPW weights. As pointed out at the beginning of this section, the complete-case estimate is always an invalid estimate of ATE because it fails to adjust for differences in baseline covariates in observational studies.

In connection with the analysis of the NSW job-training program data in the last section, the indicated high accuracy and precision of the regression and GMM estimates lead to a more refined assessment of the amount of the positive impact of the NSW job-training program. Indeed, based on the regression and GMM estimates, we conclude with high confidence that the annual earnings on average is increased from \$1646.41 to \$1662.99 for those participating in this job-training program; this is in comparison with the increase from \$1633.41 to \$1890.84 based on the six estimates in Table 1. The range of improvement, $\$1662.99 - \$1646.41 = \$16.58$ using the regression and GMM estimates, is much more confined than that of the six estimates, $\$1890.84 - \$1633.41 = \$257.43$.

5. Discussion

This thesis has considered different methods for estimating ATEs in observational studies and has studied six estimates of ATE based on the propensity score and two treatment-specific regression models. The complete-case estimate is simply the sample mean difference in observed treatment and control outcomes and is thus not a valid estimate of ATE because it does not adjust for differences in baseline covariates. The IPW and matching estimates are constructed using the propensity score, whereas the regression estimate is constructed using two treatment-specific regression models. By

contrast, the AIPW and GMM estimates are constructed using both the propensity score and two treatment-specific regression models. We have applied the six estimates to the analysis of the NSW job-training data and have found that the NSW job-training program has increased the annual earnings in 1978 on average from \$1633.41 to \$1890.84 for those participating in this job-training program. We have also conducted a Monte Carlo simulation study to compare these six estimates in terms of their biases, SDs, and RMSEs under eight simulation scenarios. The simulation results in Tables 2–9 indicate that the regression estimate has the smallest SD and RMSE across all scenarios and thus performs the best among all the estimates, followed by the GMM and matching estimates. The AIPW estimate is relatively sensitive to functional form misspecification of the propensity score. One noticeable finding in the simulation results is that the GMM estimate appears to be doubly robust. Another is that the IPW estimate has reported very high biases, SDs, and RMSEs, compared to other estimates; suggesting that the IPW estimate is very sensitive to misspecification of the propensity score and is particularly sensitive to unstable IPW weights caused by estimated propensity score values close to 0 or 1. Overall, we have observed that the regression and GMM estimates are superior to other estimates and are less sensitive to misspecification of both the propensity score and the treatment-specific regression models. This observation via simulation has further improved our assessment for the positive intervention of the NSW job-training program, leading to a more accurate conclusion that the annual earnings on average is increased from \$1646.41 to \$1662.99 for those participating in this job-training program.

For future research, it is of interest to employ new developments in machine learning methods for the estimation of the propensity score and the treatment-specific regression models. The machine learning-based approach can effectively reduce bias due to functional form misspecification of these models. Therefore, the machine learning-based causal inference procedures for estimating average treatment effects warrant further and careful consideration.

Acknowledgements

I am very grateful to Professor Linda Tesar and the students in Economics 495 for giving a very careful reading and for providing many constructive and helpful comments and suggestions on three earlier versions of this honors thesis, which led to a substantially improved presentation.

REFERENCES

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American Statistical Association*, **94**, 1053-1062.

Hansen, L. P. (1982). Large sample properties of generalized method of moment estimators, *Econometrica*, **50**, 1029-1954.

Khandker, S. R., Koolwal, G. B., and Samad, H. A. (2010). *Handbook on Impact Evaluation : Quantitative Methods and Practices*, World Bank, Washington, D.C.

Kreif, N., Grieve, R., Radice, R., and Sekhon, J. S. (2013). Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation, *Health Services and Outcomes Research Methodology*, **13**, 174-202.

Lalonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review*, **76**, 604-620.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (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 Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41-55.

Table 1 Six estimates of ATE

Method	Estimate
Complete-case	1794.34
Regression	1662.99
IPW	1633.41
AIPW	1661.67
Matching	1890.84
GMM	1646.41

Table 2. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (I) and both regression models are all correctly specified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.09300	0.02021	0.09518	−53.44828	3.95266
Regression	−0.00066	0.02231	0.02236	−0.37931	0.92857
IPW	−0.00673	0.18381	0.18393	−3.86782	7.63829
AIPW	−0.00123	0.02390	0.02387	−0.70690	0.99128
Matching	−0.00152	0.02820	0.02828	−0.87356	1.17442
GMM	−0.00174	0.02403	0.02408	−1.00000	1.00000

Table 3. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (I) is correctly specified, but both regression models are misspecified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.09300	0.02021	0.09518	−59.61538	3.95266
Regression	−0.00922	0.02222	0.02408	−5.91026	1.00000
IPW	−0.00673	0.18381	0.18393	−4.31410	7.63829
AIPW	−0.00095	0.02414	0.02408	−0.60897	1.00000
Matching	−0.00152	0.02820	0.02828	−0.97436	1.17442
GMM	−0.00156	0.02408	0.02408	−1.00000	1.00000

Table 4. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (I) is misspecified, but both regression models are correctly specified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.09300	0.02021	0.09518	−25.47945	3.76206
Regression	−0.00066	0.02231	0.02236	−0.18082	0.88379
IPW	1.71265	4.44221	4.76093	469.21918	188.17905
AIPW	−0.00755	0.34345	0.34353	−2.06849	13.57826
Matching	−0.00538	0.02740	0.02793	−1.47397	1.10395
GMM	−0.00365	0.02499	0.02530	−1.00000	1.00000

Table 5. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (I) and both regression models are all misspecified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.09300	0.02021	0.09518	−11.55280	3.62314
Regression	−0.00922	0.02222	0.02408	−1.14534	0.91663
IPW	1.71265	4.44221	4.76093	212.75155	181.23068
AIPW	−0.05431	0.42023	0.42372	−6.74658	16.12943
Matching	−0.00538	0.02740	0.02793	−0.66832	1.06319
GMM	−0.00805	0.02500	0.02627	−1.00000	1.00000

Table 6. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (II) and both regression models are all correctly specified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.13166	0.02063	0.13327	−49.68302	2.92901
Regression	−0.00047	0.02734	0.02739	−0.17736	0.60198
IPW	0.01868	0.89723	0.89743	7.04906	19.72374
AIPW	−0.00210	0.05386	0.05394	−0.79245	1.18549
Matching	−0.00317	0.05115	0.05128	−1.19623	1.12703
GMM	−0.00265	0.04538	0.04550	−1.00000	1.00000

Table 7. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (II) is correctly specified, but both regression models are misspecified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.13166	0.02063	0.13327	−49.49624	2.93611
Regression	−0.01323	0.02654	0.02966	−4.97368	0.65345
IPW	0.01868	0.89723	0.89743	7.02256	19.77154
AIPW	−0.00205	0.05707	0.05710	−0.77068	1.25799
Matching	−0.00317	0.05115	0.05128	−1.19173	1.12976
GMM	−0.00266	0.04535	0.04539	−1.00000	1.00000

Table 8. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (II) is misspecified, but both regression models are correctly specified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.13166	0.02063	0.13327	−27.95329	2.03466
Regression	−0.00047	0.02734	0.02739	−0.09979	0.41817
IPW	10.46651	42.66240	43.92753	2222.18896	670.64931
AIPW	−0.10095	2.71203	2.71391	−21.43312	41.43374
Matching	−0.01207	0.04613	0.04764	−2.56263	0.72733
GMM	−0.00471	0.06534	0.06550	−1.00000	1.00000

Table 9. Monte Carlo simulation results: bias, SD, RMSE, relative bias, and relative RMSE. Propensity score (II) and both regression models are all misspecified.

Estimate	Bias	SD	RMSE	Relative bias	Relative RMSE
Complete-case	−0.13166	0.02063	0.13327	−10.85408	2.00225
Regression	−0.01323	0.02654	0.02966	−1.09068	0.44561
IPW	10.46651	42.66240	43.92753	862.86150	659.96890
AIPW	−0.37686	3.51311	3.53327	−31.06843	53.08398
Matching	−0.01207	0.04613	0.04764	−0.99505	0.71575
GMM	−0.01213	0.06543	0.06656	−1.00000	1.00000

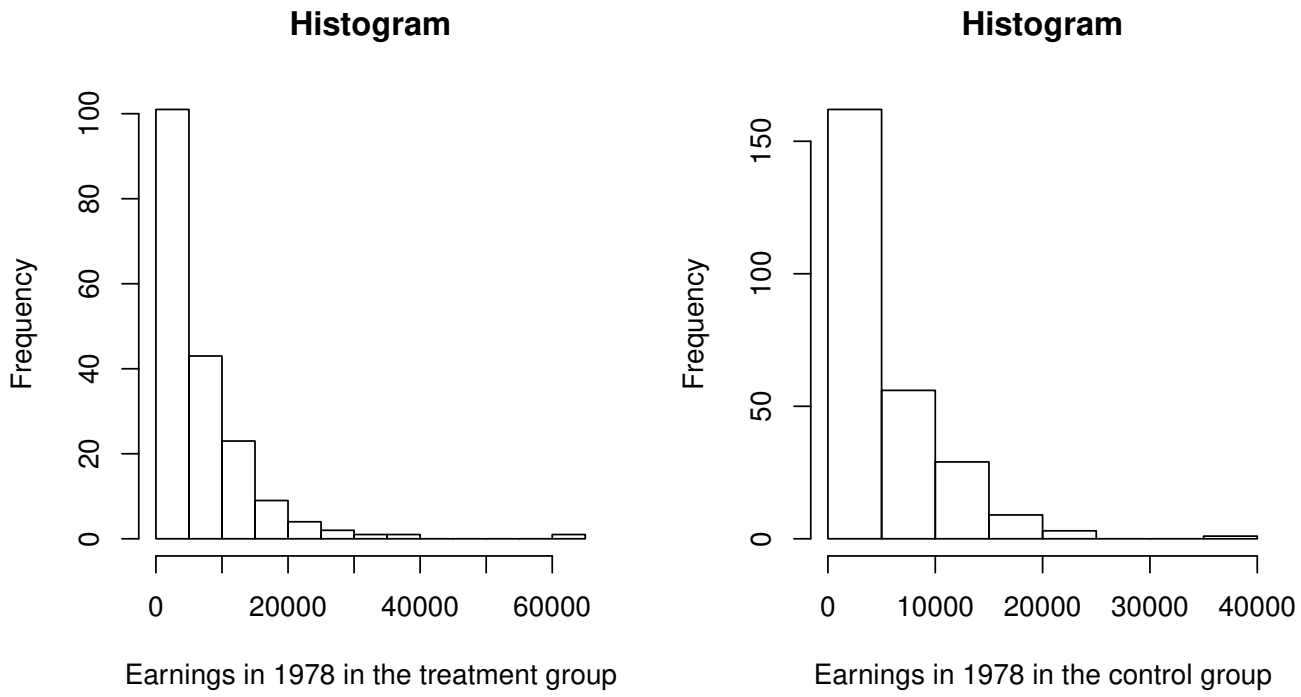

Figure 1. Histograms of earnings in 1978 for treatment (left panel) and control (right panel) groups.

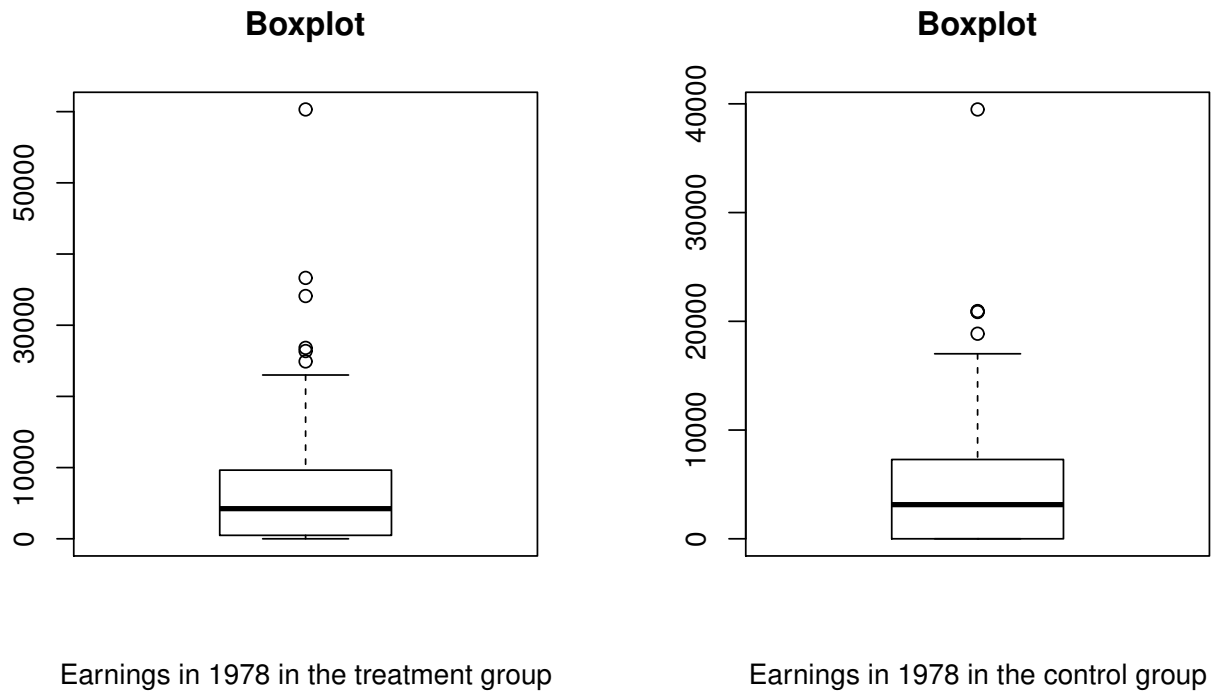

Figure 2. Boxplots of earnings in 1978 for treatment (left panel) and control (right panel) groups.