Deconvolution Estimator of Treatment Effect Distribution

Ximing Wu* and Jeffrey M. Perloff**

Abstract

This study proposes a deconvolution method to estimate the entire distribution of treatment effect of a program. This method utilizes high-order moment information implied by the standard average treatment effect estimator and approximates the underlying treatment distribution using the method of maximum entropy density. Monte Carlo simulations and experiments with real data demonstrate the flexibility and effectiveness of this estimator as an alternative deconvolution method in general and a useful program evaluation tool in specific. The proposed estimator is applied to data from the U.S. Job Training Partnership Act (JTPA) experimental training program to estimate the distribution of its impacts on individual earnings.

*: Department of Economics, University of Guelph; xiwu@uoguelph.ca

**: Department of Agricultural and Resource Economics, University of California, Berkeley; perloff@are.berkeley.edu

Introduction

According to the counterfactual approach of program evaluation, each individual has two potential outcomes: with and without the program/treatment. For an individual *i*, denote the outcome with treatment as $Y_{i,1}$ and that without treatment as $Y_{i,0}$, and the difference $\Delta_i = Y_{i,1} - Y_{i,0}$ as individual treatment effect. Because an individual cannot be in both states, we do not observe both $Y_{i,0}$ and $Y_{i,1}$ at the same time. Therefore, this problem is essentially a missing data problem and Δ_i is never directly observable.

Wooldridge (2001, Chapter 18) reviews the program evaluation literature with a focus on the Average Treatment Effect (ATE). Let the distribution of outcome with and without treatment be Y_1 and Y_0 respectively, and the distribution of treatment effect be Δ. The majority of the program evaluation literature focuses on the mean of Δ. Under the assumption of Stable Unit Treatment Value Assumption (SUTVA) which rules out cases where the treatment of one unit affects another's outcome, the ATE is estimated as

$$ATE_1 = E(Y_1 - Y_0) = E(\Delta).$$

Let *D* be the treatment status indicator, with *D=1* indicating treatment and *D=0* otherwise. When the treatment status and outcome depend on covariates *X*, it is often assumed that conditional on *X*, (Y_0, Y_1) and *D* are independent. Under this Ignorability of Treatment (given observed covariates *X*) assumption, we can estimate the conditional ATE as

$$ATE_2 = E(Y_1 - Y_0 \mid X) = E(\Delta \mid X).$$

Another quantity of interest is the average treatment effect on the treated, which is defined as

$$ATE_3 = E\left(Y_1 - Y_0 \mid X, D = 1\right) = E\left(\Delta \mid X, D = 1\right).$$

The ATE_3 is the mean effect for those who actually participated in the program. Under the Ignorability of Treatment assumption or the weaker conditional mean independence assumption, $ATE_2 = ATE_3$.

Although the ATE is of fundamental importance, it only reflects one aspect of the treatment effect. As pointed out by Heckman et al. (1997), there exists interesting questions that cannot be answered by the ATE, such as the proportion of people taking the program who benefit from it or the selected quantiles of the impact distribution. Without knowing the distribution of treatment effect Δ, the researchers are not able to answer these important policy questions.

However, the estimation of the impact distribution is difficult because, for an individual *i*, we cannot observe $Y_{i,0}$ and $Y_{i,1}$ at the same time. Furthermore, even if we have complete knowledge of Y_0 and Y_1 as in the case of random experiments, these two marginal distributions do not give us Δ directly as we do not know the joint distribution (Y_0, Y_1). For example, suppose $Y_0 = Y_1 = [1,2]$, then the treatment effect is [0, 0]. If instead $Y_1 = [2,1]$, the treatment effect is [-1, 1]. This discrepancy arises from the fact that the outcome vector is numbered: permutation of the entries of Y_0 and/or Y_1 will generally change Δ. Therefore even if the marginal distributions for both periods are identical, the treatment effect may be non-zero, depending on if the rank of the entries of outcome vector in the second period remains the same as that of the first period.

Generally, without specific assumption, such as the restrictive rank invariance condition on individual outcome between two periods, we are not able to recover Δ.[1]

In this study, we propose a method to estimate Δ directly by exploiting information suggested by the ATE model.[2] The source of identification of our proposed method is the higher-order moments of the impact distribution implied by the standard method. We use a maximum entropy density approach to estimate the underlying distribution based on moment information.

We make two contributions. First, we propose an alternative deconvolution method, which outperforms the orthogonal series method in Carroll and Hall (2004), especially for the non-trivial case when the unknown target distribution is not normal. Second, we apply this new method to the program evaluation problem to estimate the entire distribution of treatment effect. The distribution function has a simple functional form yet is flexible enough to accommodate various shapes of the distribution.

The next section describes our maximum entropy deconvolution method based on moments. The third section provides results on both Monte Carlo simulations and numerical experiments using real data. The fourth section applies the proposed method to an experimental training program. Conclusions are presented in the final section.

---

[1] Instead of estimating the entire distribution, Heckman et al. (1997) discusses how to obtain bounds of var(Δ) under certain statistical and behavioral assumptions.

[2] Concerning the popular Difference-in-Difference estimator (DID), Athey and Imbens (2002) notes that "…, one could state the assumption directly in terms of the estimator, which involves only the four conditional means rather than other moments of the distribution, thus allowing for unrestricted heteroskedasticity. However, such an assumption might be harder to justify, since, for example, it treats differences between groups in moments other than the mean as uninformative about the underlying structural model.''

The Model

Under the assumptions of SUTVA and Ignorability of Treatment, let $Y_{i,j} = f_j(X_i) + e_{i,j}$, where $f_j(X_i)$ estimates $Y_{i,j}$ consistently for j=0,1. Suppose $e_{i,0}$ and $e_{i,1}$ follow the same distribution e_i, we then have

$$\begin{aligned} Y_i &= Y_{i,0} + D_i\left(Y_{i,1} - Y_{i,0}\right) \\ &= f_0(X_i) + D_i\left[f_1(X_i) - f_0(X_i)\right] + e_{i,0} + D_i\left(e_{i,1} - e_{i,0}\right) \\ &= f_0(X_i) + D_i E(\Delta_i \mid X_i) + e_i, \end{aligned}$$

where e_i is an i.i.d. error term with mean zero and $e_i \perp \Delta_i \mid X_i$.

We can use the standard *ATE* estimator to estimate $E(\Delta \mid X)$. To obtain information beyond the first moment of Δ, one possible way to proceed is to model Δ_i explicitly as a flexible function of X_i. However, this estimator will suffer from omitting variable bias if Δ_i depends on some unobserved covariates Z_i. On the other hand, if Δ_i is a random variable independent of X_i, the covariates in X_i have zero explanatory power.

Instead we take another approach. Let $r_i = D_i\Delta_i + e_i$. Because r_i is a sum of two independent random variables, we can use a deconvolution method to estimate the distribution of Δ_i if we either know or have data on the distributions of e_i and r_i. This approach is employed by Heckman and Smith (1997) and Heckman et al. (1997) in the program evaluation problem and by Horowitz and Markatou (1996) in the error component model.

The conventional numerical deconvolution method uses the ratio of nonparametrically-estimated characteristic functions of r and e to derive that of the target distribution Δ, and then invokes the inversion theorem to recover Δ. One practical drawback of this method is that it can be sensitive to the choice of bandwidth and

sometimes leads to negative estimated densities. Moreover, Carroll and Hall (2004) note that consistently estimating of the target density is practically impossible because of rather slow convergence rate. They propose two methods, involving kernel and orthogonal series respectively, that are based on low-order approximation of the target density rather than its consistent estimates. Each of these methods requires only the existence and knowledge/estimates of low-order moments of e and r.

We propose an alternative but conceptually similar error component estimator, in which we model the individual effect Δ_i as a random coefficient for the treatment indicator D_i. Existing methods often assume that the random coefficient is distributed according to a known parametric family (typically the normal distribution). We relax this assumption and show that one can use higher-order moment information to obtain a flexible estimate of the target density.

Because $\hat{f}_0(X_i)$ estimates $Y_0(X_i)$ consistently, we can obtain a consistent estimate of r_i as

$$\hat{r}_i = Y_i - \hat{f}_0(X_i) = D_i\hat{\Delta}_i + \hat{e}_i .$$

Denote μ_k and ν_k as the k^{th} moment of Δ_i and e_i respectively, we have

$$E\left((1-D_i)r_i^k\right) = E\left(e_i^k \mid D_i = 0\right) = \nu_k . \tag{1}$$

Therefore, we can estimate the moments of the error distribution from the error terms of the control group.

Since Δ and e are independent, using Binomial expansion, we obtain

$$
\begin{aligned}
E\left(D_i r_i^k\right) &= E\left(\left(\Delta_i + e_i\right)^k \mid D_i = 1\right) \\
&= E\left(\sum_{j=0}^{k} \frac{k!}{(k-j)!\,j!} \Delta_i^{\,k-j} e_i^{\,j} \mid D_i = 1\right) \\
&= \sum_{j=0}^{k} \frac{k!}{(k-j)!\,j!} \mu_{k-j} \nu_j .
\end{aligned} \tag{2}
$$

Assuming both Δ_i and e_i have finite moments up to order k, we are able to estimate μ_k from $\left(\hat{r}_i, D_i\right)$ using Equation (1) and (2). For example, let $\hat{\nu}_k = \frac{\sum_i^n \left(1 - D_i\right) \hat{r}_i^{\,k}}{\sum_i^n \left(1 - D_i\right)}$ and $\hat{\varpi}_k = \frac{\sum_i^n D_i \hat{r}_i^{\,k}}{\sum_i^n D_i}$, then we can estimate the first four moments of Δ as

$$
\begin{aligned}
\hat{\mu}_1 &= \hat{\omega}_1, \\
\hat{\mu}_2 &= \hat{\omega}_2 - \hat{\nu}_2, \\
\hat{\mu}_3 &= \hat{\omega}_3 - 3\hat{\mu}_1 \hat{\nu}_2 - \hat{\nu}_3, \\
\hat{\mu}_4 &= \hat{\omega}_4 - 6\hat{\mu}_2 \hat{\nu}_2 - 4\hat{\mu}_1 \hat{\nu}_3 - \hat{\nu}_4 .
\end{aligned}
$$

The terms involving $\hat{\nu}_1$ disappear because $\nu_1 = E\left(e_i\right) = 0$.

Certain restrictions on the estimated moments can be used as specification tests on the independence assumption $e_i \perp \Delta_i \mid X_i$, as noted by Heckman et al. (1997). For example, because

$$
\omega_2 = E\left[\left(\Delta_i + e_i\right)^2\right] = \mu_2 + \nu_2 + 2E\left[\Delta_i e_i\right],
$$

if Δ_i and e_i are negatively correlated and $\mu_2 + 2E\left[\Delta_i e_i\right] < 0$, using the relation under independence assumption $\mu_2 = \omega_2 - \nu_2$ will lead to a negative μ_2. Therefore, a negative $\hat{\mu}_2$ clearly indicates the violation of the independence assumption. Similarly, $\hat{\mu}_4$ is required to be positive.

Given estimated moments of Δ, we can use the maximum entropy (maxent) density approach to estimate its distribution. The principle of maximum entropy is a general method to assign values to probability distributions on the basis of given information, in our case, moments. This method produces the least informative and most conservative distribution in the sense that minimal assumption is made regarding the unknown distribuiton. The resulting maxent densities have simple yet flexible functional forms. The appendix provides a brief description of this approach. Interested readers can find details in Zellner and Highfield (1988) and Wu (2003).

Theoretically, we can approximate a distribution arbitrarily well by increasing the number of moments (Cobb et al., 1983). However, high-order moments can be sensitive to outliers, especially when sample size is small. In this study, we use only the first four moments. The maxent densities based on the first four moments are rather flexible, allowing both uni-modal and multi-modal distributions and straightforward extension to accommodate higher-order moments.[3] The versatile Pearson family distributions can be completely characterized by their first four moments. Biddel et al. (2003) uses the Pearson family to approximate the treatment distribution. However, the specification considered in their study only allows uni-modal distributions, which restrict the applicability of their method. By using a maxent approach, our approximation of the target density has a more flexible functional.

[3] Cobb et al. (1983) discusses the relationship between number of modes and the moment conditions.

Numerical Properties

To investigate the numerical properties of the proposed estimator, we provide both Monte Carlo simulations and experiments using pseudo program evaluation data generated from a real survey sample.

Monte Carlo Simulations

Following Carroll and Hall (2004), we set $E(\Delta) = E(e) = 0$, $\sigma_{\Delta}^{2} = 4/3$ and $\sigma_{e}^{2} = 1/3$ for all the distributions. As possible distributions for Δ, we consider the normal, skew-normal with index 5 and density function $2\phi(x)\Phi(5x)$,[4] and a normal mixture with equal probability of *N(1.2, 0.5)* and *N(-1.2, 0.5)*. The distributions of *e* include the Normal and Uniform in [-1,1]. All distributions are rescaled to have the prescribed variance. Sample sizes are *n=250* and *n=500*. Each experiment is repeated 500 times.

For comparison, we also report the results of Carroll and Hall's orthogonal series estimator, which is shown to outperform traditional methods. This method approximates the target distribution using a series expansion. In all experiments, we use a Gram-Charlier expansion based on the first four Hermite polynomials, which are constructed from the estimated moments obtained as discussed above.

[4] Denote the shape parameter as α, the skew-normal distribution with density $2\phi(x)\Phi(\alpha x)$ has mean $\mu = \frac{\alpha\sqrt{2\pi}}{\sqrt{(1+\alpha^{2})}}$ and variance $\sigma^{2} = 1 - \mu^{2}$.

Denote f_0 and $\hat{f}$ as the theoretical and estimated target density respectively, our measure of performance is the integrated squared errors $\int\left(f_0-\hat{f}\right)^2 dx$. The results are reported in column (a) of Table 1.[5]

For *n=250*, the maxent estimator outperforms the orthogonal series except when the target density is normal. The better performance of the orthogonal series estimator for the normal density is to be expected as the Gram-Charlier expansion works best when the baseline distribution is normal. However for program evaluation problem, we have no reason to assume the unknown distribution is normal, in which case we could simply use the more efficient maximum likelihood estimator. As the target density deviates from normal, the performance of the orthogonal series estimator deteriorates rapidly. When the target density is skewed, the integrated squared errors of the maxent estimator are about 85% of those of orthogonal series estimator. For the bi-modal mixed-normal density, this ratio falls to below 25%. The performance of both deconvolution estimators is slightly better when the noise distribution is uniform rather than normal, reflecting the difficulty in filtering out a Gaussian noise.

One advantage of estimating the entire distribution rather than just the ATE is that we can calculate any feature of the estimated distribution. In column (b) and (c) of Table 1, we compare the MSE of the median and inter-quartile range of the two estimators. For normal density, the orthogonal series estimator performs better for the reasons already discussed. For skew-normal distribution, the MSE of the maxent estimator is about 75% of that of the orthogonal series estimator for both the median and inter-quartile range.

[5] The integrated square errors are obtained numerically using a simple quadrature method.

When the target distribution is the bi-modal mix-normal distribution, the orthogonal series estimator estimates the median more precisely. However, this result is largely because the Gram-Charlier series is an asymptotic expansion of the normal, which happens to share the same median with the mix-normal distribution used in our experiment. As a consequence, the estimated distribution by the Gram-Charlier series maintains the normal-like bell shape. In terms of the inter-quartile range, the MSE of the maxent estimator is about 8% of that of the orthogonal series estimator, which is consistent with the results on integrated mean square errors. This example illustrates the importance of flexibility in the functional form, which to some degree prescribes the shape of the distributions.

For $n=500$, the results are similar both quantitatively and qualitatively. One notable improvement is that now two estimators perform nearly equally well when the target density is Normal.

Our experiments show that the proposed estimator is able to approximate the target density well, especially for the non-trivial cases when the target density is not normal.

Experiments with CPS Wage Data

In this section, we apply our proposed method to examine a pseudo treatment effect constructed from real data. The data used in this experiment come from the Current Population Survey Outgoing Rotation Group (ORG) file of April, 2004. Our sample is restricted to prime-aged, full-time workers with hourly wages between 5 and 100 dollars. We use the logarithm of wages in the experiments.

We construct our pseudo-data based on the standard assumptions of the popular difference-in-difference estimator. For each observation, we assigned two independent Bernoulli random variables with equal probability of success and failure, one for an indicator called "Time" (T) and the other for "Group" (G). Therefore, we randomly divide the observations into four mutually exclusive groups: a control group in the first period ($T=0, G=0$), a control group in the second period ($T=1, G=0$), a treatment group in the first period ($T=0, G=1$), and a treatment group in the second group ($T=1, G=1$). We have 6,110 observations, and each group has roughly 1,500 observations. The treatment indicator is $D = TG$.

To introduce time effect and group effect to our data, we add 0.1 to the wage of each observation with $T=1$ and 0.1 to that of those with $G=1$. We then regress the wage variable on a vector of social-economic controls, including: age, age square, education, education square, sex, union status, MSA status, plus the time and group dummies. The baseline wage used in this experiment is then constructed as the OLS fitted value plus a normally distributed error term with mean zero and standard deviation 0.1. This constructed wage w_0 has a mean of 2.59 and a standard deviation of 0.27.

We run three experiments on the estimation of the treatment impacts. First, we randomly generate a treatment effect (Δ) distributed as $N(0.1, 0.1)$ and add Δ_i to $w_{0,i}$ for the treated sample (those with $T=1$ and $G=1$). Figure 1 shows the histogram of the randomly generated treatment effect used in this experiment and the estimated treatment distribution using the proposed method. One can see that the estimate is very close to the data.

The second experiment involves a non-normal random treatment effect. The data are generated according to the log-normal distribution, whose exponential has mean -2 and standard deviation 0.5. The generated random effect sample for the treatment group has a mean 0.15 and standard deviation 0.08. The results are reported in Figure 2. The approximation is surprisingly good despite the fact the maxent density subject to the first four moments does not nest the log-normal distribution.[6]

In the third experiment, we generate a non-random heterogeneous treatment effect for the treated sample according to the following hypothetical formula

$$\Delta_i = 0.1 + 0.1 \times educ_i - 0.01 \times educ_i^2 ,$$

where *educ* is the education level. The generated treatment effect has mean 0.17 and standard deviation 0.17. Figure 3 plots the estimated treatment effect distribution, which is very close to the data used in the experiment. The two modes correspond to the clusters of high school and college graduates. The result suggests that the proposed method is also able to approximate an impact distribution that is a function of explanatory variables.

Our method relies on the higher-order moment information contained in the residuals of standard DID model and does not model the heterogeneous treatment effect as a function of potential explanatory variables. Therefore, we are able to approximate the distribution of treatment impacts regardless if the treatment effect is a random variable or a function of some other variables, and more importantly, for the latter, regardless if it is a function of observables or unobservables.

[6] The log-normal distribution is a maxent density characterized by generalized moments of $log(x)$ and $log(x)^2$.

Empirical Application

In this section, we use the proposed estimator to estimate the impact distribution on earnings by the U.S. Job Training Partnership Act (JTPA). We use data from the National JTPA Study (NJS), a recent experimental evaluation of a large scale U.S. training program. Heckman et al. (1997) gives a detailed account of this data set. In the JTPA evaluation, accepted applicants were randomly assigned to treatment or control group, with the control group being prohibited from receiving JTPA services for 18 months. The treatment includes classroom training, on-the-job training and job search assistance to the disadvantage. Following Heckman et al. (1997), we focus on the earnings of adult women. The number of observations is 4,317.

The outcome variable used in this study is the earnings 18 months after the treatment. We consider three specifications. First, we calculate the unconditional moments of the treatment effect. Second, we estimate those moments conditional on some social-economic controls, including: age, education, race, marital status, number of children, total family income, past work and welfare history, and experiment site dummies. In the third experiment, we add interaction between the treatment status and age and education.[7]

The estimated moments of the treatment distribution is reported in Table 2.[8] The conditional variance is slightly smaller than the unconditional one. The inclusion of the interaction terms with the treatment status barely changes the variance. Statistical tests do not reject the hypothesis that three variances are identical. These results suggest that

[7] The regression results are available from the authors upon request.
[8] Heckman et al. (1997) discusses using the estimated moments as a specification test on the assumption of independence between Δ and Y_0. Since all the even moments are positive, they can not reject the independence assumption.

most of the variation in the treatment effect cannot be explained by the observed covariates used in our model. Therefore in this case, modeling the treatment effect distribution as a function of observables will have little explanatory power, a situation that calls for a random estimator. The proposed estimator is suitable for this task as it imposes few restrictions on the shape of the distribution and is shown to be able to approximate the underlying distribution when it is a random variable not dependent on covariates observable to the researchers.

Since all three specifications produce similar results, we focus our discussion on conditional moment estimates without interactions. Figure 4 plots the estimated treatment distribution.[9] Also plotted is a normal distribution with identical mean and variance. The treatment distribution is shown to be more concentrated than the normal but skewed to the right. The Jarque and Bera test of normality rejects the hypothesis of normality at 1% level. Hence, the conventional error component estimator assuming a normal treatment effect distribution will fail to capture some important feature of the underlying impact distribution.

We can use this estimated treatment distribution to answer some interesting policy questions. For example, the median impact is $327, less than the average treatment effect $844. The distribution also predicts that among the treated, 51% of the population benefit from this program.[10] Therefore, about half of the benefit from the treatment, and

[9] The shape of the distribution is close to what is reported in Heckman, et al (1997), which uses the empirical characteristic function approach for deconvolution on the same data.

[10] Heckman, et al. (1997) reports that 56% of the population benefits from this program, using unconditional estimates.

consistent with the right-extended tail, the median impact is smaller than the average impact by $517.

Conclusion

The commonly used estimators for program evaluation focus on the average treatment effect. However, these estimators may fail to capture important features of the distribution of heterogenous treatment effect, knowledge of which is critical in answering important policy questions.

In this study, we propose a deconvolution method to approximate the entire distribution of treatment effect. The method uses high-order moment information implied by the standard average treatment effect estimator and employs the method of maximum entropy density to estimate a flexible distribution. Monte Carlo and numerical examples demonstrate the effectiveness of the proposed estimator as an alternative deconvolution method in general and its superior performance when applied to program evaluation problem.

We apply the proposed method to the JTPA experimental training program to estimate the distribution of the treatment effects on individual earnings. Our results suggest that little variation in the individual treatment can be explained by observables, highlighting the importance of modeling the treatment effect distribution as a flexible random process. Consistent with previous studies using the same data, the impact distribution is shown to right-skewed, with the average treatment effect larger than the median treatment effect. Slightly more than 50% of the treated population is projected to benefit from this training program.

Appendix

The principle of maximum entropy, introduced by Jaynes in 1957, states that one should choose the probability distribution, consistent with given constraints, that maximizes Shannon's entropy. According to Jaynes (1957), the maximum entropy distribution is "uniquely determined as the one which is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters."

The maxent density $p(x)$ can be obtained by maximizing Shannon's information entropy

$$W = -\int p(x) \log p(x)\, dx,$$

subject to K known moment conditions for the entire range of the distribution

$$\int p(x)dx = 1,$$

$$\int g_i(x)\, p(x)dx = \mu_i,$$

where $i = 1, 2, \ldots, K$ indexes the characterizing moments, μ_i, and their functional forms, $g_i(x)$. Here $g_i(x)$ is continuous and at least twice differentiable.

We can solve this optimization problem using Lagrange's method, which leads to a unique global maximum entropy. The solution takes the form

$$p(x,\, \lambda) = \exp\left(-\lambda_0 - \sum_{i=1}^{K} \lambda_i g_i(x) \right),$$

where λ_i is the Lagrange multiplier for the i^{th} moment constraint and $\lambda = [\lambda_0, \lambda_1, \ldots, \lambda_K]$.

Zellner and Highfield (1988), Ornermite and White (1999), and Wu (2003) discuss the estimation of maxent density subject to moment constraints. Generally the

maxent density estimation method has no analytical solution. To solve for the Lagrange multipliers, we use Newton's method to iteratively update

$$\lambda^{(1)} = \lambda^{(0)} + \mathbf{G}^{-1}\mathbf{b},$$

where **G** is the (K+1) by (K+1) Hessian matrix of the form

$$G_{ij} = \int g_i(x) g_j(x) p(x,\lambda) dx, \qquad 0 \le i,\ j \le K,$$

and

$$b_i = \mu_i - G_{i0}\left(\lambda^{(0)}\right), \quad 0 \le i \le K.$$

This maximum entropy method is equivalent to a maximum likelihood approach where the likelihood function is defined over the exponential distribution and therefore consistent and efficient (see Golan, et al., 1996 for a discussion of the duality between these two approaches). In mathematical statistics, most of the known distributions may be described as maximum entropy (maxent) densities subject to moment constraints. These characterizing moments are sufficient statistics for exponential families; the entire distribution can be summarized by the characterizing moments.

References:

Athey, Susan, and Guido Imbens. "Identification and Inference in Nonlinear Difference-in-Difference Models." Working Paper. 2002.

Biddle, Jeffrey, Les Boden and Robert Reville. "A Method for Estimating the Full Distribuiton of a Treatment Effect, with an Application to the Impact of Workplace Injury on Subsequent Earnings." Working Paper, 2003.

Carroll, Raymond J., and Peter Hall. "Low-Order Approximations in Deconvolution and Regression with Errors in Variables." *Journal of Royal Statistical Society. B* 66 (2003): 31-46.

Cobb, L., P. Koppstein, and N. H. Chen. "Estimation and moment recursion relations for multimodal distributions of the exponential family." *Journal of the American Statistical Association* 78 (1983): 124-130.

Golan, A., G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. New York: John Wiley and Sons, 1996.

Heckman, James J., and Jeffrey Smith. "Evaluating the Welfare State." In *Frisch Centenary*, edited by S. Strom. Cambridge: Cambridge University Press, 1997.

Heckman, James J., Jeffrey Smith, and Nancy Clements. "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64, no. 4 (1997): 487-535.

Horowitz, Joel L., and M. Markatou. "Semiparametric Estimation of Regression Models for Panle Data." *Review of Economic Studies* 63 (1996): 145-68.

Jaynes, E. T. "Information Theory and Statistical Mechanics." *Physics Review* 106 (1957): 620-30.

Ormoneit, Dirk, and Halbert White. "An Efficient Algorithm to Compute Maximum Entropy Densities." *Econometric Reviews* 18, no. 2 (1999): 141-67.

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, 2001.

Wu, Ximing. "Calculation of Maximum Entropy Densities with Application to Income Distribution." *Journal of Econometrics* 115 (2003): 347-54.

Zellner, Arnold, and Richard A. Highfield. "Calculation of Maximum Entropy Distribution and Approximation of Marginal Posterior Distributions." *Journal of Econometrics* 37 (1988): 195-209.

Table 1. Simulation results

n	Δ	E	Maxent			Orthog		
			(a)	(b)	(c)	(a)	(b)	(c)
250	Normal	Normal	0.002	0.008	0.011	0.001	0.005	0.006
250	Normal	Uniform	0.002	0.007	0.011	0.001	0.005	0.006
250	Skew-Normal	Normal	0.042	0.030	0.643	0.051	0.040	0.855
250	Skew-Normal	Uniform	0.042	0.030	0.644	0.050	0.040	0.837
250	Mix-Normal	Normal	0.029	0.026	0.058	0.117	0.002	0.688
250	Mix-Normal	Uniform	0.023	0.017	0.058	0.118	0.002	0.704
500	Normal	Normal	0.001	0.004	0.005	0.001	0.003	0.003
500	Normal	Uniform	0.001	0.004	0.006	0.001	0.003	0.003
500	Skew-Normal	Normal	0.041	0.025	0.601	0.051	0.034	0.861
500	Skew-Normal	Uniform	0.041	0.023	0.615	0.051	0.036	0.859
500	Mix-Normal	Normal	0.023	0.014	0.048	0.117	0.001	0.695
500	Mix-Normal	Uniform	0.019	0.011	0.048	0.118	0.001	0.698

(a): Integrated mean square errors
(b): Mean square errors of median
(c): Mean square errors of inter-quartile range

Table 2. Estimated moments of the treatment distribution (unit: $1,000)

	1st moment	2nd moment	3rd moment	4th moment	variance
Unconditional	0.77	8.75	185.86	6280.52	2.86
Conditional	0.84	8.61	173.20	7609.69	2.81
Conditional+Interaction	0.75	8.51	176.39	7712.65	2.82

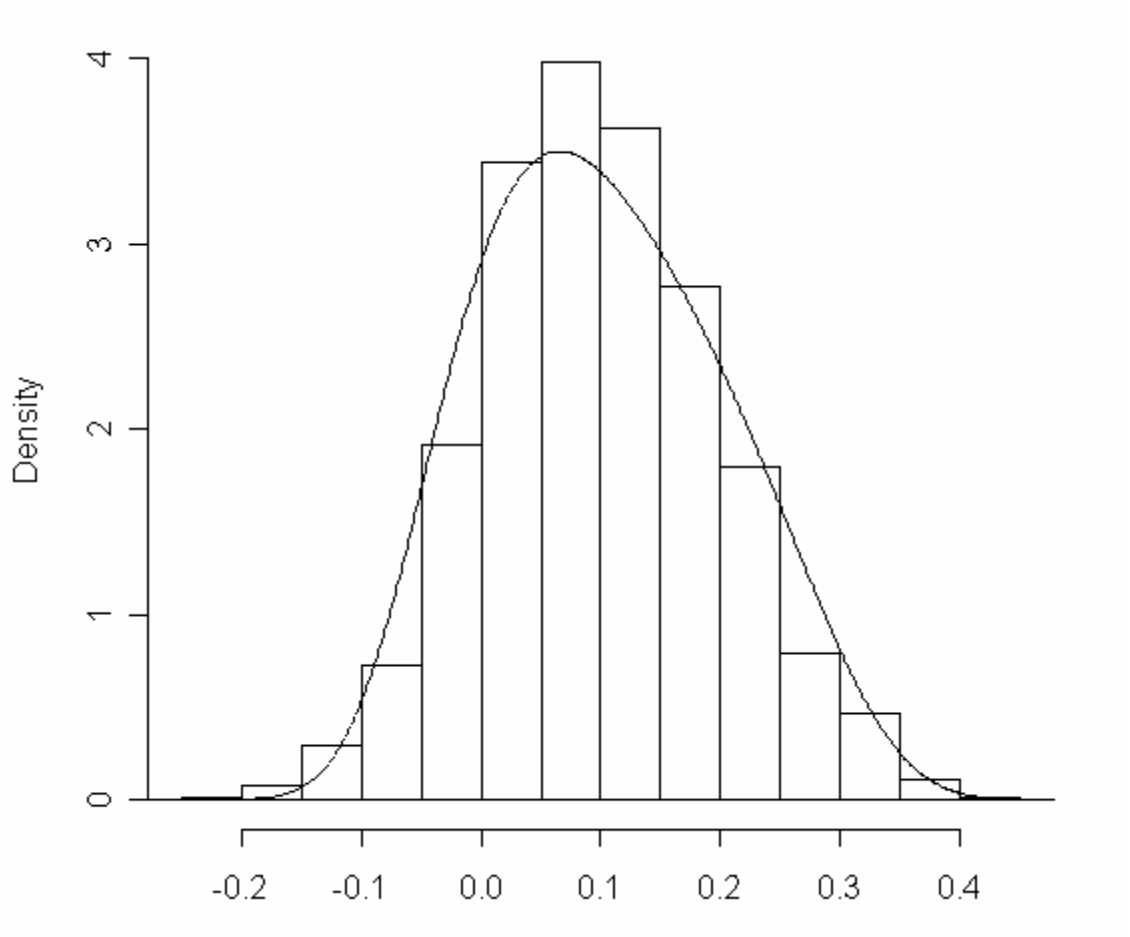

Figure 1: Estimation of normal treatment effect distribution

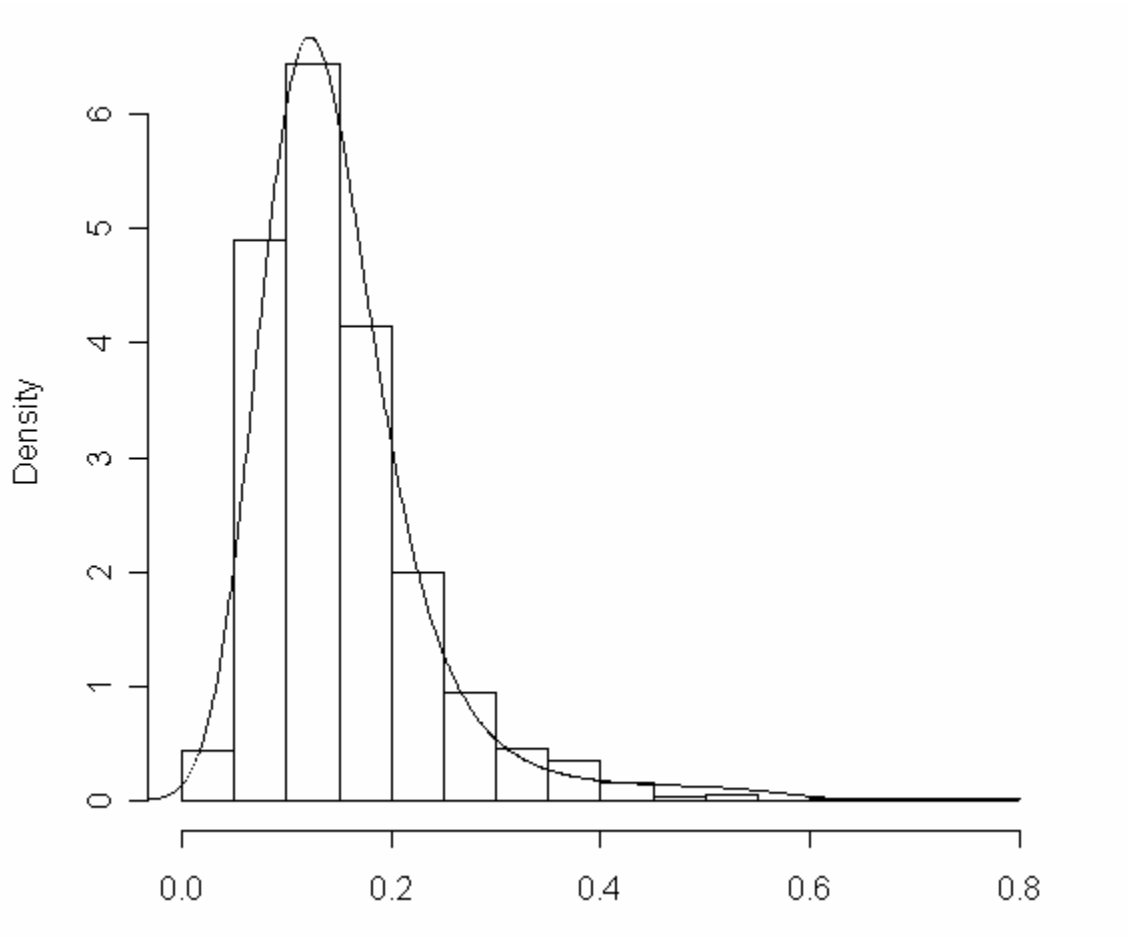

Figure 2. Estimation of log-normal treatment effect distribution

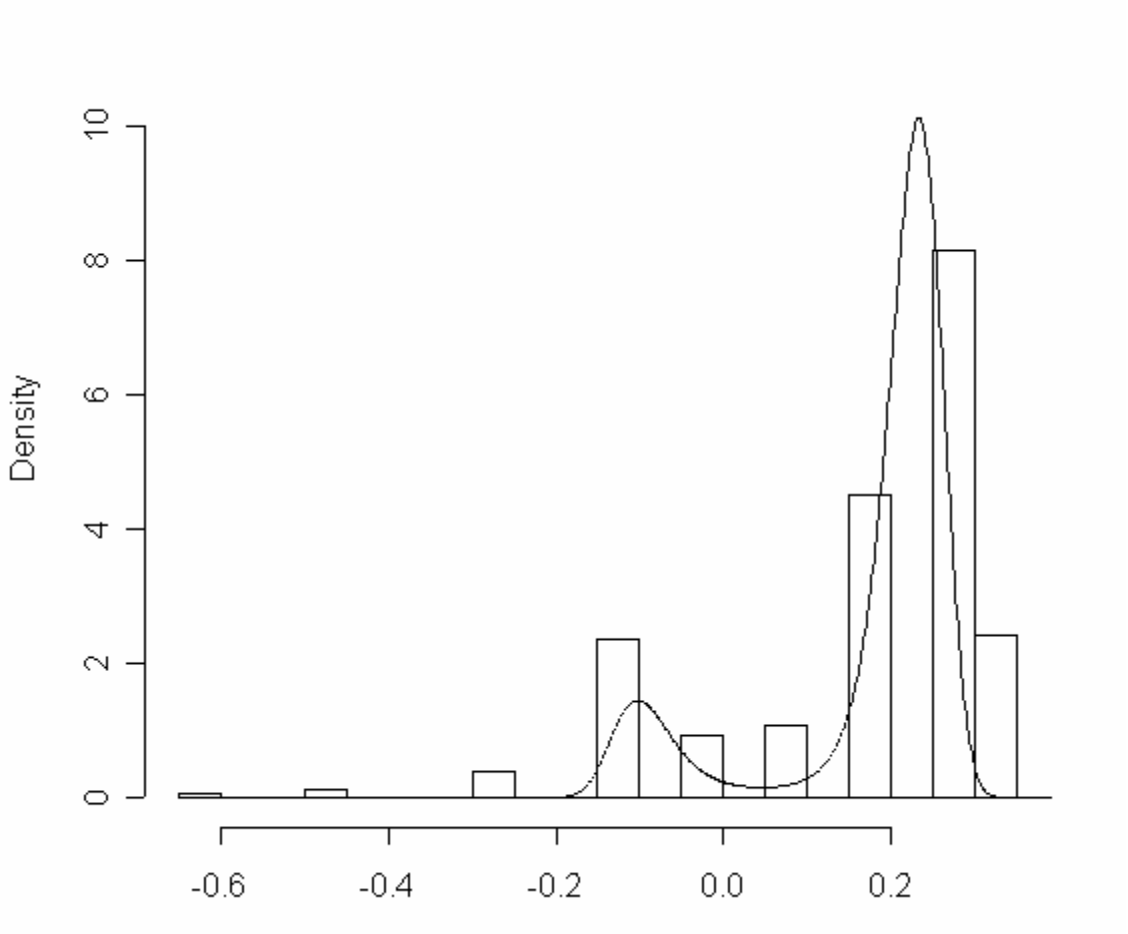

Figure 3. Estimation of heterogenous treatment effect distribution

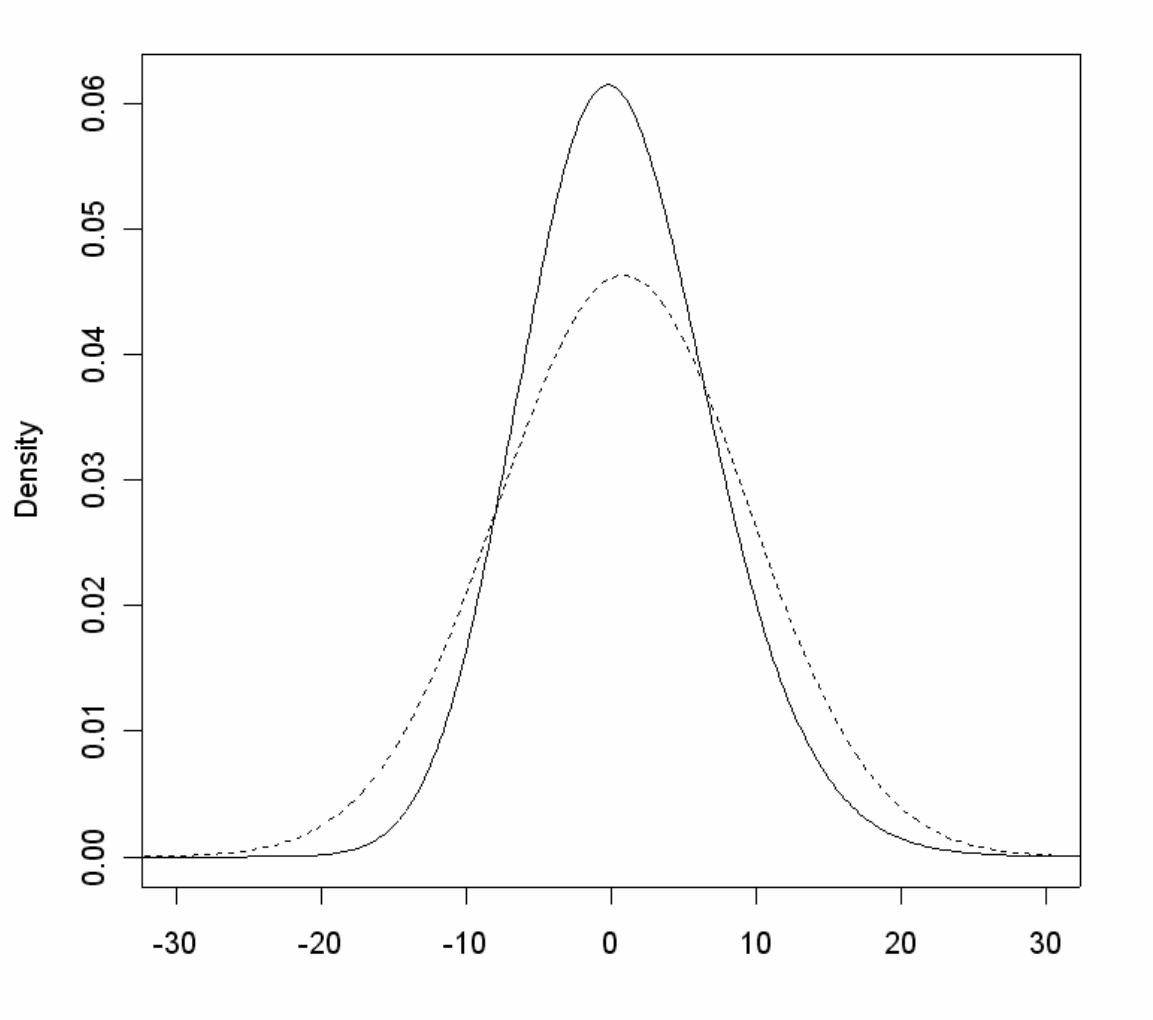

Figure 4: Estimated treatment effect distribution (solid) and normal distribution with identical mean and variance (dotted); unit: $1,000