Empirical Bayesian Estimation of Treatment Effects

Christopher P. Adams*
Federal Trade Commission
Email: cadams@ftc.gov

October 9, 2018

Abstract

This paper considers estimation of the treatment effect where the researcher has data from a large number of related experiments. Within each experiment it is assumed that the treatment allocation is unconfounded. This situation can be thought of as an idealized panel data set where each individual cross sectional unit is such an experiment. The paper considers two cases. In the first, there are a large number of treated units. The paper shows that standard analog estimate of higher moments of the treatment effect is biased and not consistent as the number of experiments gets large. The empirical Bayesian estimator of the treatment distribution is unbiased and consistent. In the second, the number of treated units is small. The standard analog estimator is not consistent. Conditioning on the observed data, the mean of the posterior distribution is the expected value of the treatment effect. However, the posterior distribution is not known. The paper shows that under certain assumptions, the empirical Bayesian estimator is an unbiased and consistent estimate of the unknown posterior distribution.

*I'm grateful for continued discussions about this problem with Emek Basker, Dan Greenfield, Dan Hosken, Nick Kreisle, Yana Petrova, Joris Pinske, Jeremy Sandford, Chris Taylor and Nathan Wilson. As well as participants at the International Association for Applied Econometrics 2018 conference in Montreal QC. Thanks also to Chuck Manski for the pointer to Robbins (1956). The returns to school data comes from Bill Greene's data archive. Note that this paper does not necessarily represent the views of the Commission or any individual Commissioners. All remaining errors are my own.

The empirical Bayesian estimators are illustrated using simulations and two conical data sets.

1 Introduction

Standard econometric analysis provides information on the average treatment effect or possibly the local average treatment effect (Imbens, 2010). However, such measures may understate or overstate the value of a particular policy. The distribution of the treatment effect may provide a more accurate representation of the likely effect of the policy. Alternatively, there may be only one treated unit and so the value of interest is the individual treatment effect on the treated. In either case, estimation may be improved by the use of a large number of related experiments where the variation in the treatment effect is unconfounded within each experiment. This paper considers the case of a panel data set with a large number of cross-sectional units and time-series variation in the treatment. The paper considers two cases. In the first, there are a large number of treated units. The paper shows that the standard analog estimate of the treatment effect distribution is both biased and not consistent as the number of cross-sectional units gets large. The paper proposes an alternative analog estimator called an empirical Bayesian estimator. The paper presents conditions for which this estimator is unbiased and consistent. In the second case, there is only one treated unit. In this case the standard analog estimator is not consistent as the number of cross sectional units gets large. If an a priori distribution exists, then the expected treatment effect conditional on the observed data is equal to the mean of the posterior distribution. The paper shows that while the posterior distribution is not known, under certain conditions there is an empirical Bayesian estimator that is an unbiased and consistent estimate of the posterior distribution. The proposed estimator is illustrated using simulated data, analysis of returns to schooling and the impact of German reunification.

Robbins (1956) considers the case where there is a large number of related but independent experiments. In addition Robbins (1956) assumes that there exists an a priori distribution. Given this Bayesian assumption, Robbins (1956) points out that the observed density of parameters from the set of experiments can be written as a mixture distribution. The set of likelihood functions mapping from the true parameters to the observed parameter estimates is mixed by an a priori density over the true parameters. Robbins (1956) suggests that if there is a unique solution to this mixture equation, then the a priori distribution can be estimated. Solving

this convolution problem to estimate the prior is called "g-modelling" (Efron and Narasimhan, 2016). This is an analog estimator of a Bayesian object. Unfortunately, in general there is no unique solution to the mixture equation (Adams, 2016). However, there is a unique solution if the likelihood function is known and it satisfies a rank condition (Efron, 2014). Efron and Narasimhan (2016) point out that if the outcome measure is finite, then the likelihood function is known and is a multinomial function. This paper will consider the case where the support of the outcome and treatment can be reasonably approximated with a finite set.

The standard analog estimate of the distribution of the treatment effect is straightforward to estimate given the data and the assumption that the treatment is unconfounded at the individual level. However, the estimates of the higher moments are biased and inconsistent. This is due to sampling error at the level of each individual experiment (Louis, 1991). Following Robbins (1956), this observed distribution of the treatment effect can be written as a mixture distribution where the weighting density is the true distribution of the treatment effect. Unfortunately, we don't know the function that maps from the true treatment effect to the estimated treatment effect and thus we cannot estimate the a priori distribution over the treatment effect. Fortunately, there is a related mixture equation which can be solved. From each individual sample, we can estimate the joint distribution over outcomes and treatment levels assuming finite support. From the set of individuals we can estimate the distribution over these sample estimated distributions. This distribution can be written as a mixture distribution, where the likelihood function is a multinomial function. Given a rank condition on the likelihood function we can uniquely determine and estimate the a priori distribution over the joint distributions of outcomes and treatments. Given this prior over the joint distributions we can calculate an a priori distribution over treatment effects.

The statistician, Thomas Louis, a student of Robbins, argues for the use of the empirical Bayesian approach for estimating heterogenous treatment effects. Louis (1991) considers a randomized controlled trial with multiple centers and argues that the treatment effect varies across these centers. He points out that the observed treatment effect at each center will not be equal to the true treatment effect by center, due to sampling variation. Louis (1991) proposes an EM algorithm to estimate a non-parametric maximum likelihood function. Louis (1991) also argues

that researchers using these methods must account for the fact that the prior and posterior are estimated, when making inference. Like here, Louis (1991) assumes that the treatment is unconfounded by center. This paper uses results from Efron (2014) and Efron and Narasimhan (2016) and an algorithm based on one developed for the estimating finite mixture models (Benaglia et al., 2009). The paper presents an analog method for determining confidence intervals more similar to a bootstrap than the parametric approach of Louis (1991).

In economics, there has been some work using empirical Bayesian techniques to estimate individual growth rates using panel data (Liu et al., 2016).[1] In panel data with so-called "short" panels, there is an incidental parameters problem. The fixed effect parameters are allowed to be different for each individual cross-sectional unit. However, each of these parameter estimated with only a small sample and are not consistent. There are two standard approaches to the problem. The first is to ignore it. This is more or less what occurs in standard fixed effects approaches. Often these parameters are not of direct interest, and it can be shown that parameters of interest can be consistently estimated even when the incidental parameters cannot (Cameron and Trivedi, 2005). The second is to assume that the parameters of interest can be drawn from a single distribution. This approach is often called random effects. Under certain assumptions it can be shown that the parameters of this distribution are consistently estimated (Cameron and Trivedi, 2005).

The empirical Bayesian approach fits somewhere between the two. Like the fixed effects approach, the empirical Bayesian approach will estimate each individual unit separately. Like the random effects approach, the empirical Bayesian approach assumes that the parameters are drawn from a single distribution. The empirical Bayesian approach is more general than the standard random effects model in that allows the errors to also vary by individual cross sectional unit. The estimator is also non-parametric although a finite support assumption is required for the estimator used here (Efron and Narasimhan, 2016).

Koop and Tobias (2004) estimate a random effects model using a standard Bayesian estimator. The authors propose that the panel data structure can be used

[1] Instead of the Tweedie formula used in Liu et al. (2016), this paper uses the finite mixture idea suggested at the end of Robbins (1956) and developed in Efron (2014) and Efron and Narasimhan (2016).

to estimate heterogeneity in the treatment effect. As stated below, the distribution of the treatment effect is potentially identified if the treatment level is unconfounded at the individual unit level. That is, the timing of an individual's increase in educational attainment is "as if" random. Unfortunately, it is unclear from the presentation exactly what their estimate is. Despite the lack of clarity, the estimated distribution under the random effects assumption is much tighter than for the standard analog estimator and much tighter the empirical Bayesian estimator presented here.

In addition to providing a method for estimating the distribution of the treatment effect, the paper provides a method for estimating the individual treatment effect. The standard analog estimator is not consistent but is unbiased, at least if you do not condition on the observed data. Conditional on the observed data the expected value of the individual treatment effect is the mean of the posterior distribution. Under some parametric assumptions, the mean of the posterior is a weighted average of the analog estimator and the true mean of the a priori distribution (Louis, 1991). There is a sense in which this approach "shrinks" the estimator. Under certain parametric assumption this estimator has lower expected loss (risk) than the analog estimator (Louis, 1991). Unfortunately, the true posterior is unknown. This paper shows that the empirical Bayesian estimator is an unbiased and consistent estimator of the posterior distribution. In the empirical example there is only one treated unit. For this case the empirical Bayesian approach requires an assumption that each sample conditional on the individual cross sectional unit and the treatment level is drawn from a distribution whose parameters are independently drawn from the same a priori distribution. An immediate implication of the assumption is that the a priori expected value of the treatment effect is zero.

As Louis (1991) points out, it is not correct to use the standard Bayesian approach of using the posterior to determined the confidence intervals. The reason is that the posterior is estimated and thus there is sampling uncertainty associated with it. This paper uses the estimated prior and estimated posterior as analogs of the true prior and true posterior (Manski, 1988). As discussed further below, the standard errors are calculated as draws from these estimated functions.

This adjusted fixed effect method stands in contrast to the synthetic control and related approaches to the problem of estimating an individual treatment effect (Abadie et al., 2010; Doudchenko and Imbens, 2016). The fixed effect model assumes

that time series variation can be accounted for with a simple average of the control units. Abadie et al. (2010) and others suggest using the pre-period to construct optimal weights. The empirical Bayesian estimator developed here could be used in conjunction with these other methods (Adams, 2018). This approach is similar to the interactive fixed effects approach used in macroeconomics panel data models (Bai, 2009).

The paper proceeds as follows. Section two presents the theoretical characteristics of the standard estimator and the proposed empirical Bayesian estimator. Section three presents the estimator and simulation results. Section 4 presents analysis of the distribution of returns to schooling. Section 5 presents analysis of the impact of German reunification on per capita growth rates. Section 6 concludes.

2 Theory

2.1 Model

Consider the following data generating process.

$$Y_{it} = F_i(X_{it}) + G(W_{it}) + \epsilon_{it} \tag{1}$$

where $Y_{it} \in \Re$ is the observed outcome of interest for individual $i \in \{1, ..., N\}$ in time $t \in \{1, ..., T\}$, $X_{it} \in \Re^J$ is the level of the treatment of interest, W_{it} are other observable characteristics of interest and ϵ_{it} is some unobserved component that determines the observed outcome. The treatment effect is $F_i : \Re^J \to \Re$ which may be vector valued and may vary across individuals. The other observed characteristics are assumed to affect all individuals the same and affect outcomes independently of the treatment. Note that for ease of exposition this term will often be dropped in the discussion below.

To simplify things, assume $F_i(\mathbf{X}_i) \approx \mathbf{X}_i\beta_i$, where $\mathbf{X}_i$ is a $T \times J$ matrix. Given this approximation, the "treatment effect" is measured by the vector β_i.

Assumption 1. *For each $i \in \mathcal{N}$, $X_{it} \perp \epsilon_{it}$ for all $t \in \{1, ..., T\}$.*

Assumption 1 is the main assumption motivating the paper and the use of panel data for analysis of treatment effects. The assumption states that at the level of

the individual, the allocation of the treatment is unconfounded. The assumption does not rule out the possibility that individuals that receive the treatment differ systematically from those that don't receive the treatment. In general, the set of individuals for which the treatment effect is identified, is some subset of the total number of individuals in the sample.

Definition 1. *For each* $i' \in \mathcal{N}' \subset \mathcal{N}$, $\mathbf{X}_{i'}$ *is full rank.*

Definition 1 is the formal statement that there exists a subset of the sample, although not necessarily a random subset, for which the treatment effect is identified. Note that just because the treatment effect is identified for a particular individual, it may not be well estimated.

Assumption 2. ϵ_{it} *is drawn iid conditional on* X_{it}.

Assumption 2 states that there error term is drawn independently. This assumption simplifies the theoretical analysis of both the standard analog estimators and the proposed empirical Bayesian estimators.

2.2 Treatment Distribution

In the first case we are interested in estimating a particular moment of the treatment effect distribution.

$$\mu_n = \int_\beta \beta^n g_\beta(\beta) \tag{2}$$

where g_β is the distribution of the treatment effect in the population. The finite sum version is as follows.

$$m_n = \frac{1}{N} \sum_{i=1}^{N} \beta^n \tag{3}$$

The standard analog estimator is then given by the following finite sum.

$$\hat{m}_n = \frac{1}{N'} \sum_{i'=1}^{N'} \hat{\beta}_{i'}^n \tag{4}$$

Note that $\hat{\beta}_{i'}$ is only observed for the subset of individuals where the treatment effect is identified ($i' \in \mathcal{N}'$).

To see the issue with standard analog estimator, note that

$$\beta_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'Y_i + (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i \tag{5}$$

Note that this can be re-written as follows.

$$\hat{\beta}_i = \beta_i - (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i \tag{6}$$

That is, the observed treatment effect for individual i may differ substantially from the true treatment effect for that individual. The extent of the divergence is dependent on the number observations, in this case, the number of time periods. The following theorem summarizes how this divergence affects the standard analog estimator of the distribution of the treatment effect.

Theorem 1. *Given Assumption 1, 2,* $E(\epsilon_{it}) = 0$, $E(|\epsilon_{it}|) < 0$, $T < \infty$ *and* $\sum_{i=1}^{\infty} \frac{E(\epsilon_i^2)}{i} < \infty$

1. *if* $n = 1$, $E(\hat{m}_n - m_n) = 0$,
2. *if* $n = 1$, $(\hat{m}_n - m_n) \xrightarrow{a.s.} 0$, *and*
3. *if* $n > 1$ *and* n *is even,* $(\hat{m}_n - m_n) > 0$

Proof. Step 0. From above

$$\hat{m}_n = \frac{1}{N}\sum_{i=1}^{N}(\beta_i - (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i)^n \tag{7}$$

Step 1. Let $n = 1$

$$\begin{aligned} E(\hat{m}_n) &= E\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\beta_i\right) - E(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i) \\ &= m_n - \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_{it}E(\epsilon_{it}) \\ &= m_n \text{ by assumption that } E(\epsilon_{it}) = 0 \end{aligned} \tag{8}$$

where ω_{it} is the weight associated with the tth column of $(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$

Step 2. Let $n = 1$.

Step 2.0. Let $\upsilon_i = \sum_{t=1}^{T}\omega_{it}\epsilon_{it}$. $E(\upsilon_i) = \sum_{t=1}^{T}\omega_{it}E(\epsilon_{it}) = 0$ and $E(|\upsilon_i|) \leq \sum_{t=1}^{T}|\omega_{it}|E(|\epsilon_{it}|) = T\bar{\omega}E(|\epsilon_{it}|) < \infty$ as $E(|\epsilon_{it}|) < \infty$ and $T < \infty$ and $\bar{\omega} = \max_{it}|\omega_{it}| < \infty$.

Step 2.1. $\hat{m}_1 - m_1 = -\frac{1}{N}\sum_{i=1}^{N} \upsilon_i$. So by Theorem A.8 and (2.0), we have the result (Cameron and Trivedi, 2005).

Step 3. Let $n > 1$ and n is even.

$$\begin{aligned} \hat{m}_n &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\beta_i + (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i)^n \\ &\geq \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\frac{2^n}{2}(\beta_i^n + (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i)^n) \text{ by Jensen's inequality} \\ &> m_n \end{aligned} \tag{9}$$

□

The theorem states that while the standard analog estimator can be used to provide an unbiased and consistent estimator of the mean of the distribution, the estimate of the second moment is both biased and not consistent.

The empirical Bayesian approach considers Equation (2) and proposes to estimate g_β directly. To be clear, such an approach requires that there exists such an a priori distribution.

Assumption 3. *$\beta_i \overset{iid}{\sim} g_\beta$ and $\{Y_i, X_i\} \overset{iid}{\sim} g_P$ for all $i \in \{1, ..., N\}$.*

Assumption 3 states that there exists an a priori distribution for both the underlying joint distribution of the outcomes and treatment levels (g_P), as well as the derived distribution of the treatment effect (g_β). Assumption 3 can be thought of as a "Bayesian" assumption in that it is consistent with the Likelihood Principle (Berger, 1985; Louis, 1991). Given Assumption 3, Robbins (1956) notes the following relationship holds.

$$f_\beta(\hat{\beta}) = \int_\beta h_\beta(\hat{\beta}|\beta) g_\beta(\beta) \tag{10}$$

where $f_\beta(\hat{\beta})$ is the observed distribution of the estimated treatment effects and $h_\beta(\hat{\beta}|\beta)$ is a likelihood function. Note that from (3) of Theorem 1, $f_\beta(\hat{\beta}) \neq g_\beta(\beta)$. The notation $\hat{\beta}$ means the sample estimate of the parameter β along with the number of data points used in the estimation.

Efron (2014) notes that if h_β is known and a rank-condition holds, then there exists a g_β that uniquely solves the mixture model. Unfortunately, it is not clear what h_β is. Therefore, the paper considers a related problem. However, before stating that problem consider the following assumption.

Assumption 4. *Let the following hold for the data generating process.*

1. *The set $\mathcal{Y} \times \mathcal{X}$ has $K < \infty$ elements.*
2. *Let $P_i = \{p_{i1}, p_{i2}, ..., p_{iK}\} \in \mathcal{P}$ and p_{ik} be the associated probability of the kth set occurring given P_i is the true distribution for individual i.*
3. *Let $\hat{P}_i = \{\hat{p}_{i1}, \hat{p}_{i2}, ..., \hat{p}_{iK}\} \in \mathcal{P}$ and $\hat{p}_{ik}$ be the observed frequency of the kth set occurring for individual i.*
4. *Let $|\mathcal{P}| = L < \infty$*

Assumption 4 significantly simplifies the exposition. The finiteness of the support of the joint distribution over the outcome and the treatment level is required for identification. However, this could be thought of as a reasonable approximation of the true support. The assumption follows the assumptions in Efron and Narasimhan (2016), who states that are made for computational convenience.

Again, given Assumption 3, Robbins (1956) states that we can write out the observed probabilities as a mixture model.

$$f_P(\hat{P}) = \int_P h_P(\hat{P}|P) g_P(P) \tag{11}$$

Note that for expositional simplicity it is assumed that $T_i = T$. Given this, the notation $\hat{P}$ provides information both about the frequencies in the data and the number of observations (T). Given Assumption 2, h_P is known, it is a multinomial function (Efron and Narasimhan, 2016). So if the rank condition on h_P holds, g_P is uniquely determined by the observed f_P. Note that the number of different frequencies that can be observed is dependent on the size of the sample (**?**).

As stated above the joint distribution of the observed outcome and the treatment level is directly related to the distribution of the treatment effect. The latter can be constructed from the former by stacking the appropriate arguments for each $x \in \mathcal{X}$ and drawing Y from the P_x, which is the marginal distribution of Y conditional on x. So $g_\beta(\beta') = \int_P 1|\beta(P) = \beta' | g_P(P)$.

The following summarizes the main theoretical result of the paper. The empirical Bayesian estimator is an unbiased and consistent estimator of any moment of the treatment distribution.

Theorem 2. *Given Assumption 1, 2, 3, 4 hold and the rank condition on h_P holds, and $E(\epsilon_{it}) = 0$ and $E(|\epsilon_{it}|) < \infty$,*

1. $E(\hat{\mu}_n - \mu_n) = 0$
2. $(\hat{\mu}_n - \mu_n) \xrightarrow{a.s.} 0$

Proof. Step 0.

$$\beta(P) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y_{P|X} \tag{12}$$

where $\mathbf{X}$ is a matrix with all possible treatment allocations and $Y_{P|X}$ is a vector of outcomes drawn from distribution P conditional on the treatment in the corresponding row of $\mathbf{X}$.

Given this relationship between the treatment effect β and the joint distribution P, we can rewrite Equation (2).

$$\begin{aligned} \mu_n &= \textstyle\int_\beta \beta^n g_\beta(\beta) d\beta \\ &= \textstyle\int_\beta \beta^n \int_{P\in\mathcal{P}} 1|\beta(P) = \beta| g_P(P) dP g_\beta(\beta) d\beta \\ &= \textstyle\int_{P\in\mathcal{P}} \int_\beta \beta^n 1|\beta(P) = \beta| g_\beta(\beta) d\beta g_P(P) dP \\ &= \textstyle\int_{P\in\mathcal{P}} \beta(P)^n g_P(P) dP \end{aligned} \tag{13}$$

Given Assumption 4

$$\mu_n = \beta' g_P \tag{14}$$

where β is $L \times 1$ vector and g_P is an $L \times 1$ vector.

The analog estimator is then

$$\hat{\mu}_n = \beta' \hat{g}_P \tag{15}$$

where $\hat{g}_P = \mathbf{H}_P^{-1} \hat{f}_P$. This is the solution to the analog of Equation (11) and $\mathbf{H}_P$ is the $L \times L$ matrix representing the multinomial likelihood function.

Step 1. Let f_{lP} denote the lth row of f_P referring to the probability of observing $\hat{P}_l \in \mathcal{P}$

$$\begin{aligned} |E(\hat{\mu}_n) - \mu_n| &= |E(\beta' \mathbf{H}_P^{-1} \hat{f}_P - \beta' \mathbf{H}_P^{-1} f_P)| \\ &= |\textstyle\sum_{l=1}^L \omega_l (E(\hat{f}_{lP}) - f_{lP})| \\ &\leq \textstyle\sum_{l=1}^L |\omega_l| |E(\frac{1}{N} \sum_{i=1}^N h_P(\hat{P}_{il}|P_i)) - f_{lP}| \\ &= \textstyle\sum_{l=1}^L |\omega_l| |\frac{1}{N} \sum_{i=1}^N \int_{P_i\in\mathcal{P}} h_P(\hat{P}_{il}|P_i) g_P(P_i) dP_i - f_{lP}| \\ &= \textstyle\sum_{l=1}^L |\omega_l| |\frac{1}{N} \sum_{i=1}^N f_{lP} - f_{lP}| \\ &= 0 \end{aligned} \tag{16}$$

Step 2. Let $\epsilon' = \max_l |\omega_l|\epsilon$ and $\delta' = L\delta$

Step 2.1 As $E(P) < \infty$ and $E(|P|) < \infty$, by Theorem A.8 (Cameron and Trivedi, 2005)

$$\lim_{N\to\infty} \Pr(|\hat{f}_{lP} - f_{lP}| > \epsilon) < \delta \tag{17}$$

for all $l \in \{1, ..., L\}$.

Step 2.2

$$\begin{aligned} \Pr(|\hat{\mu}_n - \mu_n| > \epsilon') &= \Pr(|\beta'\mathbf{H}(\hat{f}_P - f_P)| > \epsilon') \\ &\leq \textstyle\sum_{l=1}^{N} \Pr(|\omega_l||\hat{f}_{lP} - f_{lP}| > \epsilon') \\ &= \textstyle\sum_{l=1}^{N} \Pr(|\hat{f}_{lP} - f_{lP}| > \frac{\max_l |\omega_l|\epsilon}{|\omega_l|}) \\ &\leq \textstyle\sum_{l=1}^{N} \Pr(|\hat{f}_{lP} - f_{lP}| > \epsilon) \\ &< L\delta \\ &= \delta' \end{aligned} \tag{18}$$

□

Theorem 2 states that the empirical Bayesian estimator is both unbiased and a consistent estimate of the distribution of the treatment effect.

2.3 Individual Treatment Effect

In addition to estimating the distribution of the treatment effect, we may be interested in estimating the treatment effect for a particular individual i. This case is illustrated below with analysis of the impact of German reunification on per-capita GDP growth. There is only one treated unit, Germany. Therefore we are interested in the individual treatment effect on the treated.

The standard analog estimator is the observed $\hat{\beta}_i$ for individual i. The following theorem formally states that this estimator is not consistent and whether it is unbiased depends on whether the observed data is conditioned upon.

Theorem 3. *Given Assumptions 1, 2 and 3,* $E(\epsilon_{it}) = 0$ *and* $T < \infty$*, then*

1. $E(\hat{\beta}_i) = \beta_i$
2. $|\hat{\beta}_i - \beta_i| > 0$

3. $E(\hat{\beta}_i|\{Y_{i1}, X_{i1}\}, ..., \{Y_{iT}, X_{iT}\}) = \int_\beta \beta_i \gamma(\beta_i|\hat{\beta}_i) d\beta_i$

Proof. Step 1. From Equation (6)

$$\begin{aligned} E(\hat{\beta}_i) &= \beta_i - \textstyle\sum_{t=1}^{T} \omega_t E(\epsilon_{it}) \\ &= \beta_i \end{aligned} \tag{19}$$

where ω_t denotes the weights associated with the matrix operation in Equation (6).

Step 2. If $T < \infty$ then

$$\begin{aligned} |\hat{\beta}_i - \beta_i| &= |(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\epsilon_i| \\ &> 0 \end{aligned} \tag{20}$$

Step 3. By the Law of Total Expectations

$$E(\hat{\beta}_i|\{Y_{i1}, X_{i1}\}, ..., \{Y_{iT}, X_{iT}\}) = E(\beta_i|\{Y_{i1}, X_{i1}\}, ..., \{Y_{iT}, X_{iT}\}) \tag{21}$$

The rest follows from Bayes' rule and the maintained assumptions. Note that in the function γ_β the notation $\hat{\beta}_i$ represents the observed sample estimate and other sample properties such as the sample size. □

As stated above, this estimator is unbiased but not consistent. Actually, whether it is unbiased depends on whether the expectation is conditional on the observed data. If the expectation is unconditional and it is assumed that $E(\epsilon_{it}) = 0$, the standard analog estimator is unbiased. However, if the expectation is conditional on the observed data for that individual and we make the Bayesian assumption above (Assumption 3), then (unobtainable) true posterior is unbiased. Note that in general the mean of the posterior is not equal to the analog estimator. Under some parametric assumptions it is a weighted sum of the analog estimator and the mean of the prior distribution (Louis, 1991). The remainder of the section shows that the estimated posterior is an unbiased and consistent estimate of this true posterior.

The nth moment of the true posterior is given by the following expectation.

$$\mu_{ni} = \int_{\beta_i} \beta_i^n \gamma_\beta(\beta_i|\hat{\beta}_i) \tag{22}$$

where γ_β denotes the posterior distribution of the treatment effect for individual i conditional on observing the data which is denoted $\hat{\beta}_i$.

$$\gamma_\beta(\beta|\hat{\beta}) = \frac{h_\beta(\hat{\beta}|\beta)g_\beta(\beta)}{f_\beta(\hat{\beta})} \tag{23}$$

As above the empirical Bayesian estimator is an analog estimator.

$$\hat{\mu}_{ni} = \int_{\beta_i} \beta_i^n \hat{\gamma}_\beta(\beta_i|\hat{\beta}_i) \tag{24}$$

Corollary 1. *Given the assumptions of Theorem 2 hold and* $Cov\left(\hat{g}_\beta(\beta_i), \frac{1}{\hat{f}_\beta(\hat{\beta}_i)}\right) = 0$, *then*

1. $|E(\hat{\mu}_{ni} - \mu_{ni})| = 0$
2. $|\hat{\mu}_{ni} - \mu_{ni}| \overset{a.s.}{\to} 0$

Proof. Step 1. Note that in this analysis $\hat{\beta}_i$ is fixed and represents the data available regarding individual i. The expected difference between the two is then.

$$\begin{aligned} |E(\hat{\mu}_{ni} - \mu_{ni})| &= |E(\hat{\mu}_{ni} - \mu_{ni})| \\ &= \left|\int_{\beta_i} \beta_i^n \left(\gamma_\beta(\beta_i|\hat{\beta}_i) - E(\hat{\gamma}_\beta(\beta_i|\hat{\beta}_i))\right)\right| \\ &\leq \int_{\beta_i} |\beta_i^n||h_\beta(\hat{\beta}_i|\beta_i)| \left|\frac{g(\beta_i)}{f_\beta(\hat{\beta}_i)} - E\left(\frac{\hat{g}_\beta(\beta_i)}{\hat{f}_\beta(\hat{\beta}_i)}\right)\right| \\ &= 0 \text{ if } Cov\left(\hat{g}_\beta(\beta_i), \frac{1}{\hat{f}_\beta(\hat{\beta}_i)}\right) = 0 \text{ and Theorem 2} \end{aligned} \tag{25}$$

Step 2. Similar to Step (1) noting that $\left|\frac{\hat{g}_\beta(\beta_i)}{\hat{f}_\beta(\hat{\beta}_i)} - \frac{g_\beta(\beta_i)}{f_\beta(\hat{\beta}_i)}\right| \overset{\text{a.s.}}{\to} 0$ by Theorem 4.9 (Greene, 2000) □

The empirical Bayesian estimator is an unbiased estimator of the posterior distribution. Note however, that $\mu_{1i} \neq \beta_i$. The expected value of the posterior distribution of the parameter is not necessarily equal to the true parameter value.

Assumption 5. $\beta_i \overset{iid}{\sim} g_\beta$ *and* $\{Y_{ix}\} \overset{iid}{\sim} g_P$ *for all* $i \in \{1, ..., N\}$ *and* $x \in \Re^J$.

In the illustrative example below there is only one treated unit. In this case we need to make Assumption 5 instead of Assumption 3. The assumption states that for each individual and each treatment level, the distribution of outcomes is drawn from the same a priori distribution. An implication of the assumption is that the a priori expected treatment effect is zero. That is, given this assumption the empirical Bayesian estimator will tend to "shrink" the treatment effect toward zero.[2]

2.4 Sample Variation

As stated above these are analog estimators not Bayesian estimators (Louis, 1991; Carlin and Louis, 2000). As is standard practice, we want to indicate to the reader or policy maker the extent to which the estimate varies due to sampling variation (Louis, 1991).

For example, we may want to estimate the second moment of the estimate. The standard classical approach states that if we knew the true distribution of the treatment effect we could take repeated draws of sets of the appropriate sample size and construct a distribution of sample estimates.

$$\begin{aligned} E(\hat{\mu}_n^2) &= E\left(\left(\int_\beta \beta^n \hat{g}_\beta(\beta)\right)^2\right) \\ &\approx \tfrac{1}{M}\sum_{m=1}^{M}\left(\int_\beta \beta^n \hat{g}_{\beta m}(\beta)\right)^2 \text{ for large } M \\ &= \tfrac{1}{M}\sum_{m=1}^{M}\left(\tfrac{1}{N}\sum_{i=1}^{N}\beta_{im}^n\right)^2 \text{ where } \beta_{im} \overset{\text{iid}}{\sim} g_\beta \end{aligned} \tag{26}$$

Unfortunately we don't know g_β. However, we can follow the standard practice in classical econometrics and substitute g_β with $\hat{g}_\beta$ (Goldberger, 1991).

Similarly, for the individual treatment case $\beta_{im} \overset{\text{iid}}{\sim} \gamma_\beta(|\hat{\beta}_i)$, where its analog, $\hat{\gamma}_\beta(|\hat{\beta}_i)$, is used.

[2]See discussion above that given certain parametric assumptions the empirical Bayesian estimator is a weighted sum of the analog estimator and the mean of the prior (Louis, 1991).

3 Estimation

3.1 Estimator

The paper uses an estimator based on the mixture model estimator of Benaglia et al. (2009). The algorithm assumes a finite number of "types." Here, a type is a true distribution generating the observed sample data. The algorithm solves for the a priori distribution iteratively. It guesses a prior and then based on the initial guess and the data, it determines the posterior distribution for each experiment. It then aggregates of the estimated posterior distributions to determine the next proposed prior. It repeats until it converges. Note that I'm unaware of any result stating that the algorithm will infact converge. That said, given the uniqueness condition stated above, if the algorithm converges it converges to the true estimated a priori distribution.

The estimating algorithm is as follows.

1. For each experiment $i \in \{1, ..., N\}$, determine the sample distribution $\hat{P}_{iK} = \{\hat{p}_{i1}, ..., \hat{p}_{iK}\}$ and the number of observations T_i. Note that for an unbalanced panel, the number of observations will vary.

2. Determine the set of true distributions generating the data, $P \in \mathcal{P}$. Note that this set is finite. The algorithm used chooses $\mathcal{P}$ around the observed set of sample frequencies in the data.

3. Calculate $h_P(\hat{P}_i|P_i)$ for each $\hat{P}_i \in \mathcal{P}$ and P_i for all $i \in \{1, ..., N\}$. This is a K-nomial function given the assumption of finite support on the set of outcomes and treatment levels.

4. Choose an initial distribution over $\mathcal{P}$, $g_{0P}(P)$.

5. Given this prior, determine the posterior.

$$\gamma_{1P}(P|\hat{P}_i) = \frac{h_P(\hat{P}_i|P)g_{0P}(P)}{\sum_{Q\in\mathcal{P}} h_P(\hat{P}_i|Q)g_{0P}(Q)} \tag{27}$$

6. Aggregate previous posterior to get the new prior.

$$g_{(s+1)P}(P) = \frac{1}{N}\sum_{i=1}^{N} \gamma_{sP}(P|\hat{P}_i) \tag{28}$$

where s denotes the iteration.

7. Repeat steps (6) and (7) until $|g_{(s+1)P}(P) - g_{sP}(P)| < \epsilon$ for all P.

The are two approximations made for computational reasons. Both could lead to biased estimates. In regards to the choice of K, the finite support of the observed outcomes and treatment levels, it is relatively straightforward to compare the estimates of the distribution $f_\beta(\beta)$ given the approximation to the observed distribution. The choice of the finite of L, the finite support of $\mathcal{P}$, it is more difficult to test. The algorithm chooses L as random deviations from the observed set of sample distributions. The researcher is encouraged to try different values for L and different amounts of deviation from the observed set.

3.2 Simulation

The simulations provide a sense of the finite sample properties of the empirical Bayesian estimator of the treatment distribution. The simulations consider a case with two levels of treatment that are randomly allocated to individuals and across time. The simulations use the algorithm presented above to estimate the a prior distribution over the parameters describing the joint distribution over outcomes and the treatment level. Note that the true data generating process has a continuous outcome.

The support of the joint distribution is $\mathcal{Y} \times \mathcal{X} = \Re \times [0, 1]$. The treatment effect, $\beta_i \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = [-1, 2]$ and $\sigma_0^2 = \sigma_1^2 = 0.1$ and $\rho = 0$. Also $T = 5$ and $\epsilon_{it} \sim \mathcal{N}(0, 1)$.

In the algorithm used the support of $\mathcal{Y} \times \mathcal{X}$ assumes $K = 20$. The cell borders are determined by the deciles of the observed outcomes $Y_{it} \in \mathcal{Y}$ and the values are set to the mean of the cells. The support of $\mathcal{P}$ assumes $L = 20,000$. The set is chosen at random based on deviations around the observed sample distributions. Table 1 presents simulation results for the estimator.

4 Returns to Schooling

Koop and Tobias (2004) is interested in estimating the distribution of the returns to schooling. In the returns to school literature, the focus is on estimating the "beta,"

	True	$N = 500$	$N = 1000$	$N = 2000$	$N = 5000$
Fixed Effects	2	2.000	1.998	2.003	2.000
Standard mean	2	1.996	2.000	2.004	1.999
EB mean	2	1.988	1.996	2.003	2.012
Standard std. dev.	0.31	1.037	1.031	1.038	1.030
EB std. dev.	0.31	1.023	1.016	1.016	0.991

Table 1: Estimates of the first two moments. If applicable, standard errors are in parenthesis below the estimate. Note that the EB standard errors are calculated as described above. Results from Model 1 (Koop and Tobias, 2004). Two EB estimates are presented, the first using all the observations and the second using only observations in which β_i is identified. Note that for Model 1 and OLS, the treatment effect is assumed to be constant across all individuals .

the linear average of the increase in log wages associated with one year increase in years of schooling. Estimates tend to vary from 0.07 to 0.14 (Card, 2001). Of note is that the local average treatment effect estimates tend to be higher which may be due to variation in the treatment effect (Koop and Tobias, 2004). It may be that "compliers" have higher returns to schooling than other unobserved groups. Of particular interest is whether the returns are positive for everyone.

Consider the following data generating process.

$$Y_{it} = \alpha_i + \beta_i X_{it} + \gamma W_{it} + \epsilon_{it} \tag{29}$$

$$\alpha_i, \beta_i \overset{\text{iid}}{\sim} g_\beta(\alpha_i, \beta_i) \tag{30}$$

where Y_{it} is log hourly wages for individual i in year t, X_{it} is years of schooling for individual i as of time t, W_{it} are other time varying and time invariant variables such as a year indicator and experience measures. It is assumed that the intercept and slope variables, α_i and β_i respectively, can vary across individuals and are distributed iid from g_β. This model is less general than the Koop and Tobias (2004) version in that other variables only enter additively into the log wage equation. Koop and Tobias (2004) allow endogeniety in that there are unobserved characteristics (υ_{it}) that determine the years of schooling and are correlated with unobserved characteristics determining wages (ϵ_{it}). The authors argue that under standard parametric

assumptions, the "shape" of the distribution of returns is correct even if the mean is biased. Here, it is assumed that variation in X_{it} across time is exogenous conditional on individual i.

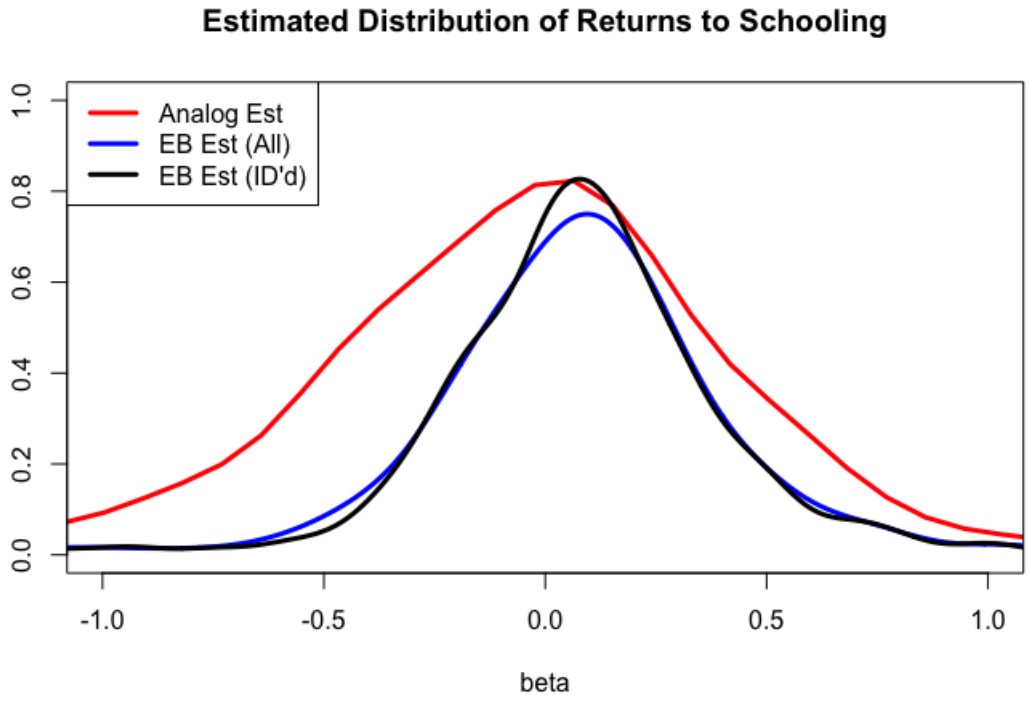

Figure 1: Estimated density of the treatment effect. The figure presents the standard analog estimator, the empirical Bayesian estimator on all individuals and the empirical Bayesian estimator on the identified individuals.

As described above, the paper cannot estimate the distribution of the treatment effect (g_β) directly. Rather the paper estimates the joint distribution of the outcome variable and the treatment level. To determine the outcome variable a first stage regression is run with log wages in the current time period on yearly time dummies, the current measure of potential experience (age - education level - 6) and potential experience squared. The residuals from this regression are used as the outcome measure. The joint distribution is binned into $K = 100$ subsets. The outcome

measure is binned into 10 subsets using quantiles of the measure, with the value taken as the mean of the subset. The treatment levels are binned as 9, 10, 11, 12, 13, 14, 15, 16, 17, more than 17. The values are taken as the mean of the subsets. As the treatment effect is not identified for every individual, the analysis is conducted separately on all individuals and only the individuals for which the treatment effect is identified. In the computation of estimator it is assumed that there are $L = 50,000$ distributions for the case when all the data is used and $L = 200,000$ when the identified subset is used. Each distribution is drawn at random around the set of observed sample estimated joint distributions.

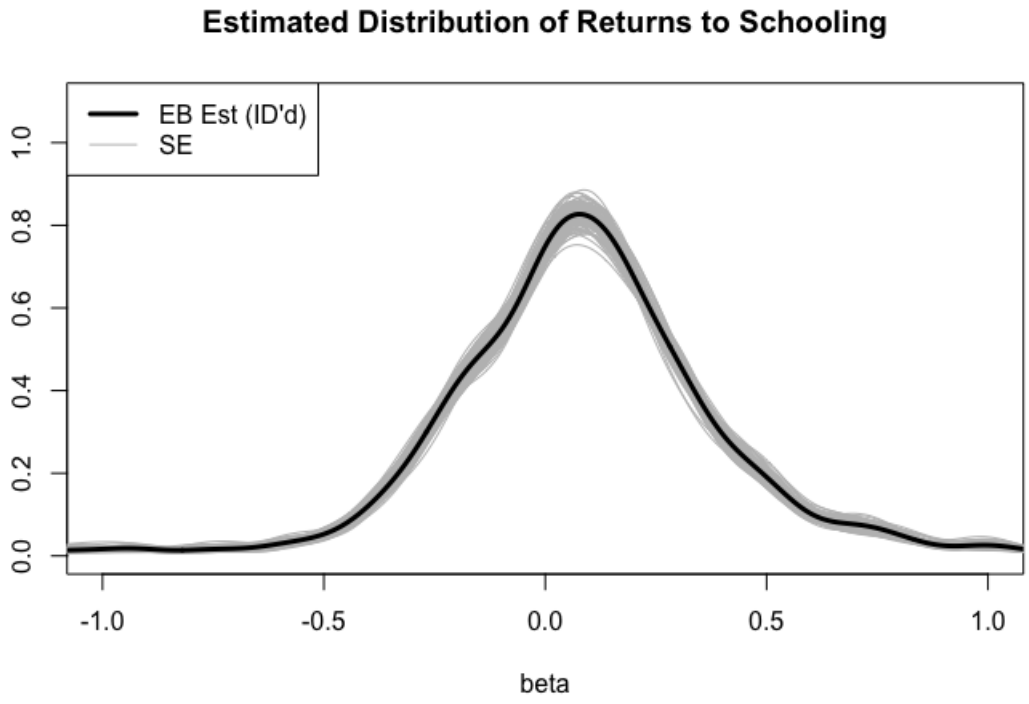

Figure 2: Estimated density of beta from the empirical Bayesian estimator on the identified individuals. The standard error is calculated using the method described above.

Koop and Tobias (2004) present a horse-race between various models of treatment

effect heterogeneity using data from NLSY. This paper uses the exact same data, however it does not use the local labor market information available to the authors. Table 2 presents results presented in Koop and Tobias (2004) from Model 1.[3] The first two moments of the distribution of β_i are presented. Note that in Model 1, the variance in the treatment effect is zero by assumption. The table also presents results based on the models presented above. For comparison purposes, the OLS estimate of β is presented. It is the same as for Model 1. This despite slight differences in the specification including the omission of local labor market information and the use of time-dummies. The standard analog estimate are calculated with two steps. The first regresses log wages on experience, experience squared and time-dummies. The second step runs separate OLS regressions of the residual from the first step and years of education, for each individual in the data ($N = 2178$). Note that these regression are only run if there is variance in education levels across time for the same individual. Lastly, the table presents results from the empirical Bayesian analog estimates. In the first case it presents results using all of the individuals. In the second case, the results are based on individuals for which the education level varies.

Table 2 shows that the basic estimates of the average treatment effect are the same despite slight differences in the specifications. Figure 1 presents the estimated densities of returns to schooling. The results are consistent with the theoretical and simulation results that suggest the analog estimate of the second moment is biased upwards. The empirical Bayesian estimates show that variance in the treatment is about half of the standard analog estimate. While the empirical Bayesian estimates of the variance of returns to schooling is much smaller than the analog estimate, it is still large. The estimate suggests that for a significant proportion of the population, the returns to schooling are negative. Figure 2 presents the estimate of the distribution with information on the variation due to sampling error on the estimation of the prior.

[3]The authors state that more general models provide estimates similar to Model 2. However, the paper's presentation of the estimated variance is unclear. On page 838 it states that the standard deviation of β is 0.006, while on the next page and the remainder of the paper it conducts the analysis as if the variance of β is 0.006.

	$E\beta$	$E\beta^2$
K-T Model 1	0.105	0.011
OLS	0.105	0.011
	(0.003)	-
Standard Analog	0.075	0.110
EB (All data)	0.077	0.104
	(0.012)	(0.007)
EB (ID data)	0.081	0.099
	(0.006)	(0.005)

Table 2: Estimates of the first two moments. If applicable, standard errors are in parenthesis below the estimate. Note that the EB standard errors are calculated as described above. Results from Model 1 (Koop and Tobias, 2004). Two EB estimates are presented, the first using all the observations and the second using only observations in which β_i is identified. Note that for Model 1 and OLS, the treatment effect is assumed to be constant across all individuals .

5 Reunification of Germany

On November 9, 1989, a spokesman mistakenly announced that citizens of the German Democratic Republic would be allowed unrestricted travel to West Germany. This led to a series of events that ended with the reunification of Germany. We are interested in whether this major change in the political institutions and population had an impact on economic growth.

Consider the following data generating process.

$$\begin{aligned} Y_{it} &= \alpha_i + \beta_i X_{it} + \gamma W_{it} + \epsilon_{it} \\ X_{it} &= \begin{cases} 1 \text{ if Germany and post 1990} \\ 0 \text{ otherwise} \end{cases} \end{aligned} \tag{31}$$

where Y_{it} is the measure of the growth rate used. It is the difference in the log of annual GDP per capita in constant 2010 US dollars, from the World Bank's Data Bank between years for country i in the time period t. The treated country is Germany. Here, X_{it} is equal to 1 for Germany post 1990 and 0 otherwise. Note that β_i is only

potentially identified for Germany. The variable W_{it} is a set of dummy variables for time periods.

There are 217 countries in this data set, although not all countries have data for all periods. The final data set includes growth rate estimates for 197 countries. Abadie et al. (2010) are careful to estimate the impact on West Germany, rather than Germany. Here we use the definition of Germany used by the World Bank. In order to account for serial correlation in the data, the data is aggregated up into 5 year periods from 1961 to 2016.[4]

The empirical Bayesian estimator relies on the estimation of the prior distribution of the treatment effect. Unlike the previous example, there is only one treated unit. Therefore, to proceed it is necessary to assume that the distribution of outcomes for each country and each treatment level is drawn from the same a priori distribution (Assumption 5). In practice, there are 198 countries in data with Germany split into a pre-Reunification and a post-Reunification country.

Given this data, the paper runs a first stage regression of observed growth rates on time-dummies and then uses the residuals to estimate the prior distribution. Prior to estimation, the observed residuals are binned into $K = 20$ subsets.[5] In the estimation, the set of distributions is assumed to be $L = 10,000$, which are drawn randomly. The algorithm described above outputs the posterior distribution for both pre-Reunification Germany and post-Reunification Germany. The posterior of the treatment effect is calculated as the difference in these two distributions.[6]

The estimated treatment effect from the difference in difference model is $\hat{\beta}_i = -0.014$ with a standard error of 0.030. Using the Abadie et al. (2010) data, $\hat{\beta}_i = -0.010$ with a standard error of 0.012. Although neither estimate is statistically different from zero at standard levels, there is still concern that the point estimate is not close to the true magnitude.

Figure 3 presents the posterior density estimator of the treatment effect. The solid-black line is the posterior density using the World Bank definition of Germany pre and post 1990. The grey lines represent the standard error of the estimate and are

[4]Data that is aggregated up to shorter time periods shows signs of serial correlation.

[5]The subsets are equidistant from the smallest to the largest value of the residual. The value of each bin is determined by the mid-point.

[6]This calculation assumes that the two distribution are independent, which is based on Assumption 5.

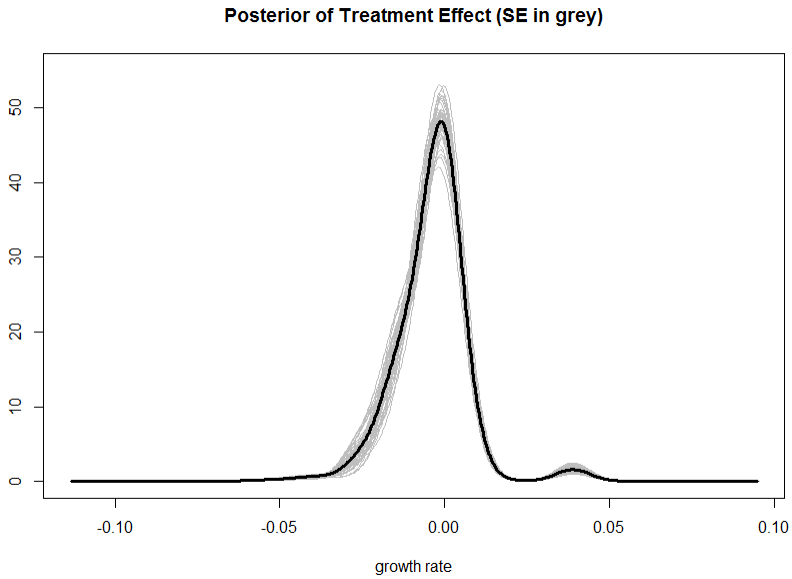

Figure 3: Posterior density of the treatment effect. The solid black line represents the the posterior of the treatment effect and the grey lines represent the standard error and are draws from the estimated posterior distribution .

calculate as sample draws from the estimated posterior distribution. The standard error are calculated following the method presented above. The results show that the weight of the distribution is below zero suggesting that German reunification reduced per-capita GDP growth. That said, a large portion of the posterior is above zero suggesting that this data and method does not provide convincing evidence that reunification in fact reduced German per-capita GDP growth.

5.1 Serial Correlation

Bertrand et al. (2004) raises the concern that there is often a substantial amount of serial correlation in the data. This is of particular concern here because the proposed estimator relies heavily on the assumption that the observations are drawn independently conditional on the country's type.

Figure 4 shows the correlation across adjacent time periods within the same country.

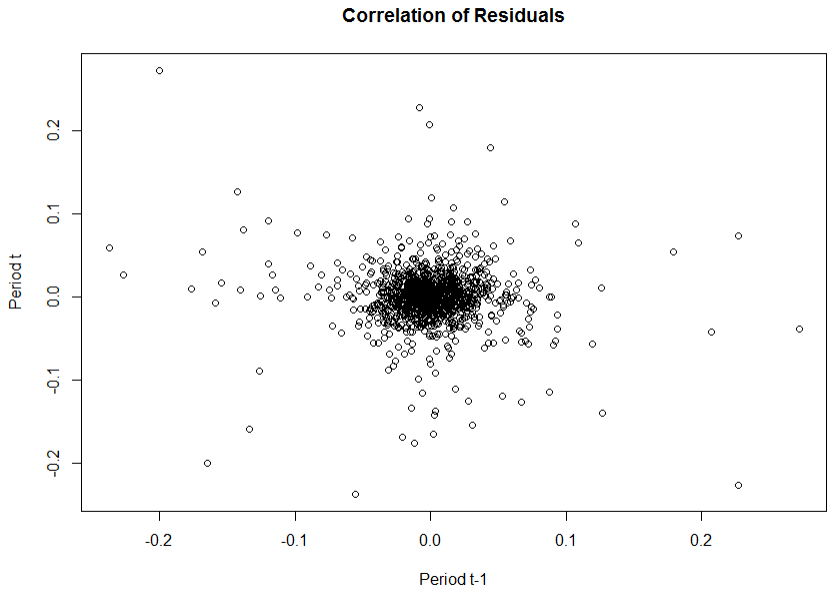

Figure 4: Plot of country level residuals across time periods.

The figure and regression results (not shown) suggest that the serial correlation is not present in the data.

6 Conclusion

This paper presents an alternative approach to estimating heterogenous treatment effects with panel data. The paper considers a problem where the researcher has access to a large number of experiments and the treatment level is allocated "exogenously" within each experiment. The paper considers two case. In the first the researcher has access to large number of treated units. This data set allows the researcher to estimate the distribution of the treatment effect. This is what Efron and Narasimhan (2016) calls the "f-modeling". In general this is not equal to distribution of interest. Using results from Robbins (1956), Efron (2014) and Efron and Narasimhan (2016) this paper follows the "g-modeling" approach, which provides a classical estimate of the true distribution of the treatment effect. In the second case, the researcher has

data with only one treated unit.

The paper shows that the standard analog estimator of the treatment effect distribution is biased and not consistent. It also shows that the standard analog estimate of the individual treatment effect is not consistent and biased when conditioning on the observed data. The paper shows that the empirical Bayesian estimator of the treatment effect distribution and the individual treatment effect is both unbiased and consistent. The theoretical results are illustrated with simulations and analysis of returns to schooling and the impact of the reunification of Germany on economic growth.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Christopher P. Adams. Finite mixture models with one exclusion restriction. *The Econometrics Journal*, 19:150–165, 2016.

Christopher P. Adams. Measuring treatment effects with big n panels. Federal Trade Commission, July 2018.

Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77: 1229–1279, 2009.

Tatiana Benaglia, Didier Chauveau, and David R Hunter. An em-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition, 1985.

M. Bertrand, E. Duflo, and S. Mullainathan. How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119, 2004.

A. Colin Cameron and Pravin K. Trivedi. *Microconometrics: Methods and Applications*. Cambridge University Press, 2005.

David Card. Estimating the returns to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, September 2001.

Bradley P. Carlin and Thomas A. Louis. Empirical bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289, December 2000.

Nikolay Doudchenko and Guido W. Imbens. Balancing, Regression, Difference-in-Difference and Synthetic Control Methods: A synthesis. NBER Working Paper 22791, October 2016.

Bradley Efron. Two modeling strategies for empirical bayes estimation. *Statistical Science*, 29:285–301, 2014.

Bradley Efron and Balasubramanian Narasimhan. A G-Modeling Program for Deconvolution and Empirical Bayes Estimation. Standford, 2016.

Arthur Goldberger. *A Course in Econometrics*. Harvard University Press, 1991.

William Greene. *Econometric Analysis*. Prentice Hall, fourth edition, 2000.

Guido W. Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, 48(2):399–423, June 2010.

Gary Koop and Justin L. Tobias. Learning about heterogeneity in returns to schooling. *Journal of Applied Econometrics*, 19(7):827–849, Nov - Dec 2004.

Laura Liu, Hyungsik Roger Moon, and Frank Schorfheide. Forecasting with dynamic panel data models. University of Pennsylvania, December 2016.

Thomas A. Louis. Using empirical bayesian methods in biopharmaceutical research. *Statistics in Medicine*, 10:811–829, 1991.

Charlies Manski. *Analog Estimation Methods in Econometrics*. Chapman and Hall, 1988.

Herbert Robbins. The empirical bayes approach to statistical decision problems. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 157–163, 1956. University of California Press.