Causal Inference in RCTs: Bayesian Analysis vs Regression

(Disclaimer: though the project was written by all for of the authors, the code is a joint effort of only Flavio Argentieri and myself)

Flavio Argentieri Alessandro Cauzzi Tommaso Crosta

Jakob Wall

August 2021

1 Introduction

Recent advances in computing power have given researchers the ability to use Bayesian statistics as a tool when tackling econometric and statistical empirical problems. One of such issues regards the question of causality, a main concern particularly among the social sciences. In this project, we explore the possibility of using Bayesian statistical analysis to make causal inference in the Potential Outcomes framework as conceptualised by Rubin (1974). For this purpose, we reassess data from Duflo, Dupas, and Kremer (2011), who analyse the impact of tracking, i.e. separating students into classes based on their ability, on school outcomes in the context of a randomized evaluation program in Kenya.

In section 2 we briefly introduce the potential outcomes approach both in general and in a Bayesian context. In section 3 we discuss the experimental setting and results presented by Duflo et al. (2011). Then, in section 4, we turn to the theoretical models we employ for our analysis and we elaborate on their Stan implementation, relative performance, and diagnostic. Finally, in section 5, we compare our results with those in the original paper and we explore treatment effect heterogeneity.

2 The problem of causal inference

2.1 The potential outcomes framework

Let us start by summarizing the most common tool used for assessing causality: the potential outcome framework (Rubin, 1974). Furthermore, let us assume there are only two levels of treatments, $D_i = 1$ if treated or $D_i = 0$ if not treated (control). Let the outcome of interest for individual i, Y_i, be equal to Y_{1i} if the individual is treated and Y_{0i} if she / he is not. The individual causal treatment effect of the treatment is then defined as $Y_{1i} - Y_{0i}$, while the observed outcome can be written as $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$. Many assumption have already been made. First, we are considering a treatment that can assume only two values (either 0 or 1). Second, we are assuming that each individual, if treated, is exposed to a treatment that is stable, i.e. always the same. Third, we are assuming that the relevant scenarios for each individual are only two: either she / he is treated, or she / he is not. Whether other individuals are treated or not, thus, does not influence the potential outcome for a given individual, which excludes the possibility of externalities. While the first assumption can be relaxed, the other two are crucial and amount to the Stable Unit Treatment Value Assumption (SUTVA). Finally, notice that we ignore time in the analysis for ease of presentation and that we focus on treatment effects defined as differences, a common choice in the literature.

In principle, we would like to estimate the individual effect for each individual exposed (or potentially exposed) to the treatment. This, however, is not possible in a frequentist approach, as only one potential outcome can be observed for a given individual. This is known as the fundamental problem of causal inference and is usually dealt with by defining aggregate causal estimands and by performing inference on those estimands. In what follows, indeed, we are going to focus on the Average Treatment Effect (ATE), defined as $ATE = \sum_{i=1}^{N} Y_{1i} - Y_{0i}$. Though this goes beyond the scope of our work, we highlight that Bayesian analyses can also provide estimates for individual treatment effects.

To reliably estimate aggregate causal estimands (however defined), the basic assumptions on which all available techniques rely (both in experimental and in observational settings) is that for each treated individual we have a non-treated group (even composed by only one other individual) that is *comparable* to her / him / it. Comparable, in such context, mean that we can consider the treatment as randomly assigned among treated and controls (maybe in subsample defined by the values assumed by covariates). We are going to focus on the most common case in Development Economics, i.e. that of a treatment randomly assigned across the villages in a sample (thus making the SUTVA more credible). In such case, we can safely make the strongest sufficient identifying assumption: $Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i$, which enables to estimate $\mathbb{E}[Y_{i1} - Y_{i0}]$ with $\mathbb{E}[\widehat{Y_i|D_i = 1}] - \mathbb{E}[\widehat{Y_i|D_i = 0}]$, where tha latter can be simply estimated using sample analogs. Indeed, under such assumption, one can write:

$$\begin{aligned}
&E[Y_{i1} - Y_{i0}] = \\
=&E[Y_{i1}] - E[Y_{i0}] = \\
=&E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 0] = \\
=&E[Y_i|D_i = 1] - E[Y_i|D_i = 0]
\end{aligned}$$

2.2 The Bayesian approach to causal inference

We next present a model based Bayesian approach to analyze data within the potential outcome framework. Potential outcomes, and consequently any causal estimand of interest, are modeled as random variables and the purpose of the analysis becomes performing inference on such estimands.

Let the discussion of such approach start with the presentation of two naive approaches to the imputation of missing outcomes and their performance in the computation of ATE (average treatment effect), a very common estimand of interest. During the whole discussion, the experimental setting will be randomization.

For a treated Unit 1, we have $Y_1^{mis} = Y_1(0)$, while for control Unit 2, we have $Y_2^{mis} = Y_2(1)$. One could set the missing potential outcome for Unit 1 equal to the average outcome for the control units, and similarly set it equal to the average outcome for the treated units for control Unit 2, i.e. set $Y_1^{mis} = \frac{\sum_{i \in c} Y_i}{\sum_{i \in c} i}$ and $Y_2^{mis} = \frac{\sum_{i \in t} Y_i}{\sum_{i \in t} i}$.

This result is unsatisfying, as it does not provide any indication about the precision of the estimates: missing potential outcomes were imputed as if there was no uncertainty in their value, hence it is only possible to compute point estimates for ATE. Moreover, randomization almost never allows to know the exact value of missing potential outcomes, but rather their distribution.

Another, less naive approach would be to use the distribution of observed potential outcomes of treated *(control)* units - notice that this might significantly differ from the true distribution of $Y_{i \in t}$ *($Y_{i \in c}$)*, with a finite sample - to randomly draw an observed potential outcome from that distribution, and impute it as the value of the missing potential outcome for some control *(treated)* Unit. For example, a randomly drawn $Y_{i \in c}$ could be assigned to treated Unit 1, while another randomly drawn $Y_{i \in t}$ could be assigned to control Unit 2. Draws would be performed with replacement until all Units have been assigned their missing potential outcome. Then, draws could be repeated until all permutations are exhausted, computing as many values for ATE as the number of possible permutations. From this point, it is possible to calculate the average of all values of ATE and also estimate the distribution of ATE and its variability. As the number of permutations grows exponentially with the number of units, it could be reasonable to only use a randomly selected subset of possible imputations.

Though this second, less naive approach is more reasonable, it is still very limited in describing the uncertainty a social scientist faces: missing data are still imputed as if the exact distribution of potential outcomes was known. Unfortunately, this is not the case in general: with a limited number of observations, relatively little is known about the distribution of possible potential outcomes.

The next step thus is to account for this uncertainty in estimation of the distribution of potential outcomes, and this requires a model for potential outcomes, both observed and missing. Here it comes the Bayesian model approach.

The goal is to derive the conditional distribution of missing outcomes, given observables:

$$f(\boldsymbol{Y}^{mis} \mid \boldsymbol{Y}^{obs}, \boldsymbol{W}) \tag{1}$$

From this model, one will be able to infer - usually using computational methods, rather than analytically - the distribution of any estimand of interest, $\tau = \tau(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{W})$ rewriting it as a function of missing and observed outcomes and assignments (treatment indicators), $\tau = \tau(\boldsymbol{Y}^{mis}, \boldsymbol{Y}^{obs}, \boldsymbol{W})$.

As it is difficult to directly specify a model for 1, two inputs must be defined first in order to derive the distribution:

1. the joint distribution of potential outcomes given the unknown parameters,

$$f(\boldsymbol{Y}(1), \boldsymbol{Y}(0) \mid \theta) = \prod_{i=1}^{N} f(Y_i(0), Y_i(1) \mid \theta) \tag{2}$$

   The specification of the distribution of potential outcomes given the unknown parameters, which can also be referred to as *the model of the Science*, requires knowledge in the research subject of investigation.

2. the prior distribution of the unknown parameters θ,

$$p(\theta) \tag{3}$$

   The choice of the prior distribution is not critical, given that the specification is not too dogmatic and data are sufficiently informative.

In practice, the specification of these two inputs is usually not critical in randomized experiments, while it might be in observational studies.

Recall that, in randomized experiments, the assignment mechanism - which gives the probability of each vector of assignment W given the Science - is known, so that there is no need to specify it as a third input, $f(\boldsymbol{W} \mid \boldsymbol{Y}(1), \boldsymbol{Y}(0))$. This would be the case, instead, in observational studies.

Starting from the two inputs, in four steps one can get to the distribution of the estimand:

1. derivation of $f(\boldsymbol{Y}^{mis} \mid \boldsymbol{Y}^{obs}, \boldsymbol{W}, \theta)$
2. derivation of $p(\theta \mid \boldsymbol{Y}^{obs}, \boldsymbol{W})$
3. derivation of $f(\boldsymbol{Y}^{mis} \mid \boldsymbol{Y}^{obs}, \boldsymbol{W}) = \int_\theta f(\boldsymbol{Y}^{mis} \mid \boldsymbol{Y}^{obs}, \boldsymbol{W}, \theta) \cdot p(\theta \mid \boldsymbol{Y}^{obs}, \boldsymbol{W})\ d\theta$
4. derivation of $f(\tau \mid \boldsymbol{Y}^{obs}, \boldsymbol{W})$

Computationally, random draws are performed from the posterior derived in Step 2 and are substituted into the conditional distribution of $\boldsymbol{Y}^{mis}$ of Step 1.

The process does not substantially change in presence of covariates. One now will have to specify the joint distribution of potential outcomes conditional on covariates, i.e. $f(\boldsymbol{Y}(1), \boldsymbol{Y}(0) \mid \boldsymbol{X}, \theta)$. At the individual level, distributions are modelled by De Finetti's theorem as i.i.d. conditional on θ. Then, it is possible to factor them into $f(Y_i(0), Y_i(1), X \mid \theta_{Y|X}, \theta_X) = f(Y_i(0), Y_i(1) \mid X, \theta_{Y|X}) \cdot f(X \mid \theta_X)$. It must then be cautiously considered whether the parameters entering the marginal distribution of the covariates are indeed distinct from those entering the conditional distribution of the potential outcomes given the covariates, in which case it is possible to write $p(\theta_{Y|X}, \theta_X) = p(\theta_{Y|X}) \cdot p(\theta_X)$. This would significantly simplify the analysis, as the model to be specified would reduce to $f(Y_i(0), Y_i(1) \mid X_i, \theta)$.

If it was also the case of a random sample from an infinite super-population, estimands would be function only of the parameters: $\tau_{sp} = \tau(\theta)$.

3 The impact of tracking on school outcomes in Kenya

Duflo et al. (2011) estimate the effect of tracking, i.e. separating students in different groups according to their abilities, on school outcomes in Kenya. To the extent that students benefit from having academically strong peers, tracking may benefit good students while hurting low-achieving ones, exacerbating human-capital inequalities. On the other hand, teachers may respond to tracking by tweaking the general level of instruction to match the average ability of the students in the different groups. Moreover, the level of effort of the teacher, which also affects school outcomes, may respond to tracking depending on various institutional factors.

Let us now describe the experiment that allowed the authors to estimate the effects of tracking on school outcomes. In May 2005, under the Extra-Teacher Program (ETP), ICS Africa provided 140 schools with funds to hire an extra first-grade teacher on a contractual basis. This program was designed to add an additional section in first grade. The authors focus their analysis on the 121 schools that only had one section before the program took place. In 61 randomly selected schools, first-grade students were randomly assigned to one of two sections. In the remaining 60 schools, pupils were assigned to the two sections based on whether their score on the first-term exam was above or below the class median. The program lasted for 18 months and in the second year, all students not repeating the grade remained in the same section and with the same teacher. At the end of the program trained enumerators administered a standardized test to random samples of students in each school.

The overall impact of tracking on end-line test scores is estimated via the following regression:

$$y_{it} = \alpha T_j + X_{ij}\beta + \varepsilon_{ij}$$

where T_j is the tracking indicator and X_{ij} contains various school and child controls, such as baseline score, whether the child was at the bottom half of the distribution, gender, age and type of teacher. Potential differential effects depending of the assigned section are estimated with the following specification:

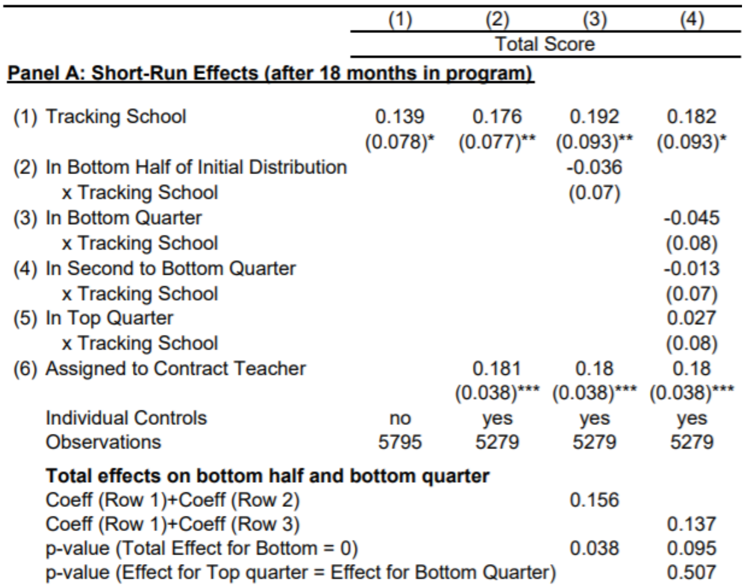
$$y_{it} = \alpha T_j + \gamma T_j \cdot B_{ij} + X_{ij}\beta + \varepsilon_{ij}$$

where B_{ij} is an indicator for whether a child is in the bottom half of the distribution of baseline scores.

In total, the authors use data from about 7,022 students enrolled in the first grade in March 2005. Due to attrition, mostly stemming from students not taking the standardized test at the end of the program, the sample drops down to 5796 for the main baseline regression, of which 5279 students also have complete data on individual control variables.

Tracking by initial achievement significantly increased test scores throughout the distribution, going against the hypothesis of tracking only benefiting good students. Figure 1 shows the main results of the paper, corresponding to a subsection of table 2 of Duflo et al. (2011).

Figure 1: Overall Effect of Tracking



	(1)	(2)	(3)	(4)
	Total Score			
Panel A: Short-Run Effects (after 18 months in program)				
(1) Tracking School	0.139	0.176	0.192	0.182
	(0.078)*	(0.077)**	(0.093)**	(0.093)*
(2) In Bottom Half of Initial Distribution			-0.036	
x Tracking School			(0.07)	
(3) In Bottom Quarter				-0.045
x Tracking School				(0.08)
(4) In Second to Bottom Quarter				-0.013
x Tracking School				(0.07)
(5) In Top Quarter				0.027
x Tracking School				(0.08)
(6) Assigned to Contract Teacher		0.181	0.18	0.18
		(0.038)***	(0.038)***	(0.038)***
Individual Controls	no	yes	yes	yes
Observations	5795	5279	5279	5279
Total effects on bottom half and bottom quarter				
Coeff (Row 1)+Coeff (Row 2)			0.156	
Coeff (Row 1)+Coeff (Row 3)				0.137
p-value (Total Effect for Bottom = 0)			0.038	0.095
p-value (Effect for Top quarter = Effect for Bottom Quarter)				0.507

In the baseline regression of standardized total scores on the tracking dummy with clustered standard errors, Duflo et al. find that students in tracking schools scored 0.138 (0.078) standard deviations more than students in non-tracking schools. This effect remains statistically significant and even increases in magnitude when considering other regression designs including individual control variables. (columns 2 to 4). Furthermore, when they interact the tracking dummy with a dummy for being in the lower half of the initial distribution, the relevant coefficient is not statistically different from 0, signalling that weaker students did not benefit less from the program compared to initially stronger students.

4 The Hierarchical Model

4.1 Standard Model

Let us consider N individuals and J schools. Let $\underset{N\times K}{y}$ be the vector of individual outcomes, $\underset{N\times K}{X}$ be the matrix of individual level covariates and $\underset{K\times 1}{\beta_j}$ be the vector of school-specific coefficients. Therefore, the slope of every covariate in the linear regression will vary for different schools. Moreover, let $\underset{N\times 1}{w}$ be the vector of treatment assignment and $\underset{J\times 1}{\alpha}$ be the vector of school-specific intercepts. Furthermore, we suppose that potential outcomes are independent, jointly normal and with heterogeneous variances, i.e.:

$$\begin{pmatrix} Y_{n,j}(0) \\ Y_{n,j}(1) \end{pmatrix} \mid \beta_j, \delta, \left(\sigma_s^2\right)_{s\in\{t,c\}}, \alpha_j \sim N\left(\begin{pmatrix} \alpha_{j(n)} + x_n'\beta_j \\ \alpha_{j(n)} + x_n'\beta_j + \delta \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{pmatrix}\right)$$

In this scenario, δ represents the stochastic causal effect, as opposed to the fixed parameter in the frequentist tradition. The notation $j(n)$ refers to the school of student n.

We also assume that the school-level coefficients $(\beta_j)_{j=1}^J$ depend linearly in expectation on some school level variables, contained in $\underset{J\times L}{U}$, via a matrix of coefficients, $\underset{L\times K}{\Gamma}$. Let $(u\cdot\gamma)_j$ be the j-th row of the matrix $\underset{J\times K}{U\cdot\Gamma}$. Lastly, we assume that each vector β_j has the same variance-covariance matrix , $\underset{K\times K}{\Sigma}$, which can be represented as follows:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\ \sigma_1^2\ \sigma_2^2 & ... & \rho\ \sigma_1^2\ \sigma_k^2 \\ \rho\ \sigma_1^2\ \sigma_2^2 & \sigma_2^2 & ... & ... \\ ... & ... & ... & ... \\ \rho\ \sigma_1^2\ \sigma_k^2 & ... & ... & \sigma_k^2 \end{pmatrix}$$

However, for the sake of simplicity, we assume that, for $j \neq k$, β_j and β_k are uncorrelated and thus independent due to normality.

Similarly, we suppose that the school-level intercepts, $(\alpha_j)_{j=1}^J$, depend on the same school-level covariates, $\underset{J\times L}{U}$, via a vector of coefficients, $\underset{L\times 1}{\eta}$. Let $(u\cdot\eta)_j$ be the j-th row of the vector $\underset{J\times 1}{U\cdot\eta}$. Each intercept has the same variance, labelled as σ_α.

Thus, the Hierarchical model in its entirety and the priors [1] assigned to each parameters are summarized as follows:

$$\begin{cases} y_{n,j} \mid \beta_j, \sigma_t, \sigma_c, \rho \sim N(\alpha_{j(n)} + x_n\beta_j + w_n\delta,\ w_n\sigma_t + (1-w_n)\sigma_c) \\ \alpha_j \mid \eta, \sigma_\alpha \sim N((u\cdot\eta)_j,\ \sigma_\alpha),\ j = 1,\cdots,J \\ \beta_j \mid \Gamma, (\sigma_i)_{i=1}^K, \rho \sim N((u\cdot\gamma)_j,\ \Sigma),\ j=1,\cdots,J \\ \sigma_i \sim Half-Cauchy(0,1),\ i = 1,\cdots,K,c,t,\alpha \\ \gamma_{h,l} \sim N(0,1),\ h=1,\cdots,L\ ;\ l=1,\cdots,K \\ \eta_v \sim N(0,1),\ v=1,\cdots,L \\ \rho \sim U(-1,1) \\ \delta \sim N(0,2) \end{cases}$$

4.1.1 Implementation

The computational complexity of obtaining the posterior densities of the parameters of interest using MCMC algorithms requires a few adjustments. For example, instead of generating random variables from Half-Cauchy distributions, it is more efficient to generate intermediate random variables from uniforms from 0 to $\pi/2$ and take the tangent [2]. Similarly, instead of generating dependent random vectors from a normal distribution, such as β_j, it is more convenient to generate independent random variables from a standard normal and multiply them by the Cholesky factor of the variance covariance matrix of interest. Moreover, it is better to use uncentered parameters when the data is deemed weakly informative (Betancourt, 2020). This procedure is explained in detail in the next section.

[1] We follow Stan's convention and assign priors to the standard deviations of normal random variables, not to their variances. Furthermore, when we write $X \sim N(0,v)$, v is the standard deviation of X.

[2] We exploit the fact that if $X \sim U(0,\pi/2)$ and $Y = c\cdot tan(X)$, then Y is distributed as a half-Cauchy with scale parameter c.

The standard model is implemented in Stan as follows:

```
// Compute beta using draws from a standard normal and Cholesky factorization
beta = gamma * u_transp + cholesky_decompose(Sigma) * z;

// Mean of school-specific intercepts
for (j in 1:J) {
    c[j] = eta' * u_transp[ ,j];
  }

// Compute alpha using draws from a standard normal
alpha = c + sigma_alpha * z_2;
```

Moreover, the likelihood function of the outcome is specified in Stan as follows:

```
// Define for each individual the mean and variance of the outcome
for (n in 1:N) {
    m[n] = alpha[g[n]] + x[n, ] * beta[, g[n]] + effect * w[n];
  }

uno = rep_vector(1, N);
d = sigma_t*w + sigma_c*(uno-w);

// Specify the model
y ~ normal(m, d);
```

4.2 Cholesky Reparametrization

4.2.1 Theory

The Choleksy decomposition is a particular case of a LU decomposition, i.e. the factorization of a matrix $\underset{N\times N}{A}$ into the product of a lower triangular matrix $\underset{N\times N}{L}$ and an upper triangular matrix $\underset{N\times N}{U}$. The Cholesky decomposition requires A to be a symmetric positive definite (SPD) matrix and factorizes it into the product of a lower triangular matrix L and its transpose, written as $A = LL^T$. It should be noted that every real-valued SPD matrix has a unique Cholesky factorization. One can derive L starting from a SPD A via the following formulas, which can be derived by induction on the dimension of A:

$$\begin{cases} L_{i,j} = \sqrt{A_{i,j} - \sum_{k=1}^{j-1} L_{i,k}^2} & i = j \\ L_{i,j} = \frac{1}{L_{j,j}}\left(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k}L_{j,k}\right) & i \neq j \end{cases} \qquad i = j, \cdots, N \qquad j = 1, \cdots, N$$

One of the most common uses of the Cholesky factorization is to ease the computation of a system of linear equations of the form $Ax = b$ when A is SPD. In particular, one may rewrite the system as $LL^Tx = b$ and divide it into two easier problems: i) solving for y in $Ly = b$, where $y \coloneqq L^Tx$ and ii) solving for x in $L^Tx = y$. These two problems are computationally simpler than the original one, as y and x can be solved for via forward and backward substitution, two very fast algorithms.

Cholesky factorization is also used to improve the efficiency and numerical stability of Monte Carlo simulations with correlated random variables. In general, this factorization can be used when sampling normally distributed correlated random variables is computationally burdensome. In fact, rather than generating a random vector $x \sim N(\mu, \Sigma)$, one can write $\Sigma = diag(\eta)\Psi diag(\eta)$, find L_Ψ such that $L_\Psi L_\Psi^T = \Psi$, generate $z \sim N(\mathbf{0}, I)$ and compute $x = \mu + diag(\eta)L_\Psi z$. This works as the following holds:

$$\mu + diag(\eta)L_\Psi z \mid \mu, \eta, L_\Psi \sim N(\mu,\ \Sigma = diag(\eta)L_\Psi L_\Psi^T diag(\eta))$$

4.2.2 Implementation

In our application, we factorize the previously described model directly in terms of Cholesky factors of Σ. More precisely, let $\Sigma = diag(\tau)\Omega diag(\tau)$, where $diag(\tau)$ is a diagonal matrix with $diag(\tau)_{k,k} = \tau_k$ and Ω is a correlation matrix. As Ω is SPD, one can find its Cholesky decomposition, i.e. a matrix L_Ω such that $L_\Omega L_\Omega^T = \Omega$, and assign priors to τ and L_Ω rather than to ρ and $(\sigma_j)_{j=1,\cdots,J}$. We introduce an additional variable, τ_{unif}, such that $\tau := tan(\tau_{unif})$ and we assume that $\tau_{unif} \sim U(0, \pi/2)$. On the other hand, we assign a LKJ Cholesky prior with shape parameter equal to 2 to L_Ω, this ensures that $\Omega = L_\Omega L_\Omega^T$'s prior is a LKJ distribution with shape parameter equal to 2. A shape parameter greater than 1 favor less correlation while a shape parameter lower than one favors more correlation.

Then, rather than sampling β_j from a normal with mean $(u{\cdot}\gamma)_j$ and covariance matrix Σ, we sample z from a multivariate standard normal and compute $\beta_j = (u{\cdot}\gamma)_j + diag(\tau)L_\Omega z$. This transformation greatly improves the performance of the code while leaving the dynamics of the model intact. This reparametrization is implemented in Stan as follows:

```
// Compute beta using draws from a standard normal and Cholesky factorization
beta = gamma * u_transp + diag_pre_multiply(tau, L_Omega) * z;

...

Model{

tau_unif ~ uniform(0,pi()/2);
L_Omega ~ lkj_corr_cholesky(2);

}
```

4.3 QR Reparametrization

4.3.1 Theory

The QR decomposition consists in factorizing a matrix $\underset{N\times K}{A}$ into the product of a orthonormal matrix $\underset{N\times K}{Q}$ and a upper triangular matrix $\underset{K\times K}{R}$. One can derive Q by applying the Gram-Schmidt process to A. This process starts from a set of linearly independent vectors, possibly the columns of A, and derives a set of orthonormal vectors that spans the same subspace. Suppose A contains linearly independent column vectors, labelled $a_1, \cdots, a_K$. Then, one can derive the k-th column of Q, q_k, via the following iterative procedure: for every $k = 1, \cdots, K$, compute $v_k = (a_k - \sum_{j=1}^{k-1} proj_{a_j} a_k)$ and let $q_k = v_k / \|v_k\|$. Having done that, one can find R by noticing that $A = QR \iff Q^T A = Q^T QR \iff R = Q^T A$ as $Q^T Q = I$ by construction.

4.3.2 Implementation

We employ QR decomposition in our model by factorizing $\underset{N\times K}{X}$ into the product of Q^* and R^*, where $Q^* := Q \cdot \sqrt{N-1}$ and $R^* := R/\sqrt{N-1}$. Then, we write $\eta = X\beta = Q^* R^* \beta$ and we define $\vartheta := R^* \beta$, which implies that $\eta = Q^* \vartheta$. Then, instead of generating $z \sim N(\mathbf{0}, I)$ and computing $\beta_j = (u \cdot \gamma)_j + diag(\tau) L_\Omega z$, as in the previous reparametrization, we compute $\theta_j = R^*(u \cdot \gamma)_j + L_{\Sigma_\theta} z$, where L_{Σ_θ} is the lower triangular matrix obtained using Cholesky decomposition on the covariance matrix of ϑ, Σ_ϑ. Recall that the following holds:

$$\beta \mid \gamma, \tau, L_\Omega \sim N(u \cdot \gamma,\ \Sigma = diag(\tau) L_\Omega L_\Omega^T diag(\tau)) \implies$$
$$\vartheta \mid \gamma, \tau, L_\Omega \sim N(R^*(u \cdot \gamma),\ \Sigma_\vartheta := R^* \Sigma {R^*}^T)$$

Moreover, despite the additional simplification, it is sufficient to maintain the priors on γ, τ and L_Ω selected previously. In addition to this, as Q^* contains orthogonal columns, one may expect the components of ϑ_j to be significantly less correlated, which makes the assumption of uncorrelated coefficients across schools more realistic. This reparametrization is implemented in Stan by defining L_{Σ_ϑ} and computing ϑ as follows:

```
// Derive the covariance matrix of theta and factorize it using Cholesky
L_Sigma = cholesky_decompose(
    quad_form(
        multiply_lower_tri_self_transpose(diag_pre_multiply(tau, L_Omega)),
        R_ast’
        ));

// Compute theta using draws from a standard normal
theta = R_ast * (gamma * u_transp) + L_Sigma * z;

// Mean of school-specific intercepts
for (j in 1:J) {
    c[j] = eta’ * u_transp[ ,j];
  }

// Compute alpha using draws from a standard normal
alpha = c + sigma_alpha * z_2;
```

Moreover, the likelihood function of the outcome is specified in Stan as follows:

```
// Define for each individual the mean and variance of the outcome
for (n in 1:N) {
    m[n] = alpha[g[n]] + Q_ast[n, ] * theta[, g[n]] + effect * w[n];
  }

uno = rep_vector(1, N);
d = sigma_t * w + sigma_c * (uno-w);
```

```
// Specify the model
y ~ normal(m, d);
```

Model reparametrized in this way tends to perform better for the following reasons: i) as Q is orthonormal, it is easier for a Markov chain to move in ϑ-space than β-space, ii) as the columns of Q^* have the same scale, a HMC algorithm tends to make fewer relatively large steps and iii) as the covariance matrix for the columns of Q^* is I, ϑ usually has reasonable scale, which further helps a HMC algorithm.

4.4 Comparison & Diagnostics

We now sample four chains for each of the three models discussed above (Vectorized, Cholesky, & Cholesky and QR) and turn to some model diagnostics in order to evaluate and compare the different approaches empirically. Even though the vectorized model uses a different prior for the variance covariance matrix of the individual coefficients, the resulting estimates are virtually the same as in the other two models. Thus, it seems fair to argue that such difference cannot account for the differences in performance that we are going to explore.

We begin by looking at the trace plots of the chains in order to assess their convergence visually. If all chains from a given model behave similarly across the trajectory this is a good indication of convergence, while big differences point to problems in this process.

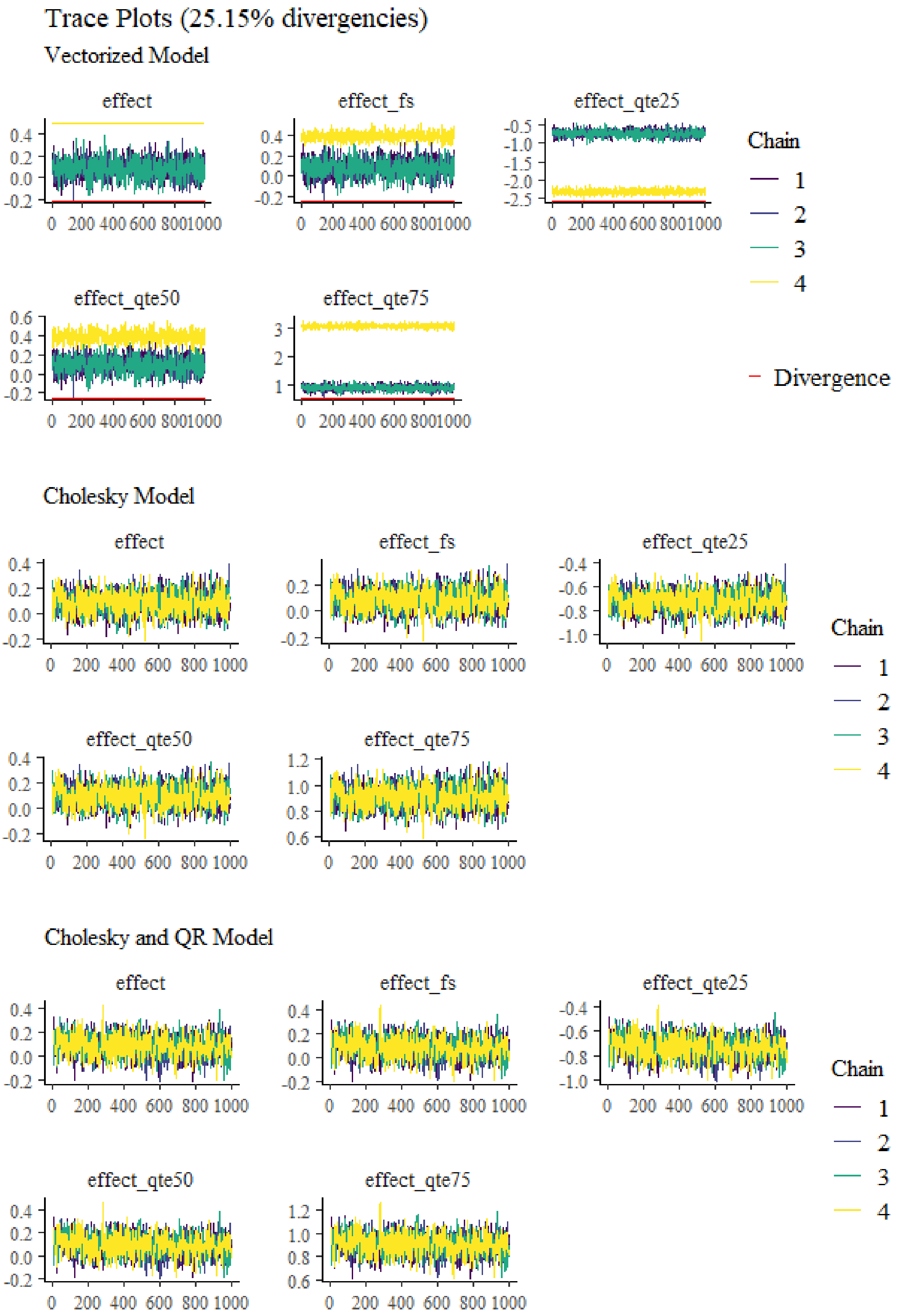


Figure 2: Trace Plots

As we can see from figure 2, our vectorized model exhibits problematic divergence behaviour as chain number 4 differs substantially from the other three realisations over the entire trajectory (25.15% convergencies). For our other two approaches on the other hand, the four chains have visually indistinguishable trace plots, indicating convergence. This holds true both for the main treatment effect as well as for the finite sample treatment effect and quantile treatment effects.

The autocorrelation plots in figure 3 show similar results, with chain number 4 of the vectorized model showing a very high degree of autocorrelation (close to 1) during the entire sample for the main effect. This is indicative of bad mixing, leading to a small effective sample size and in turn large variance of our estimator.

Both the Cholesky model as well as the Cholesky and QR model start to show low degrees of autocorrelation very quickly into their trajectories, with the Cholesky and QR model taking slightly longer to approach values close to 0.

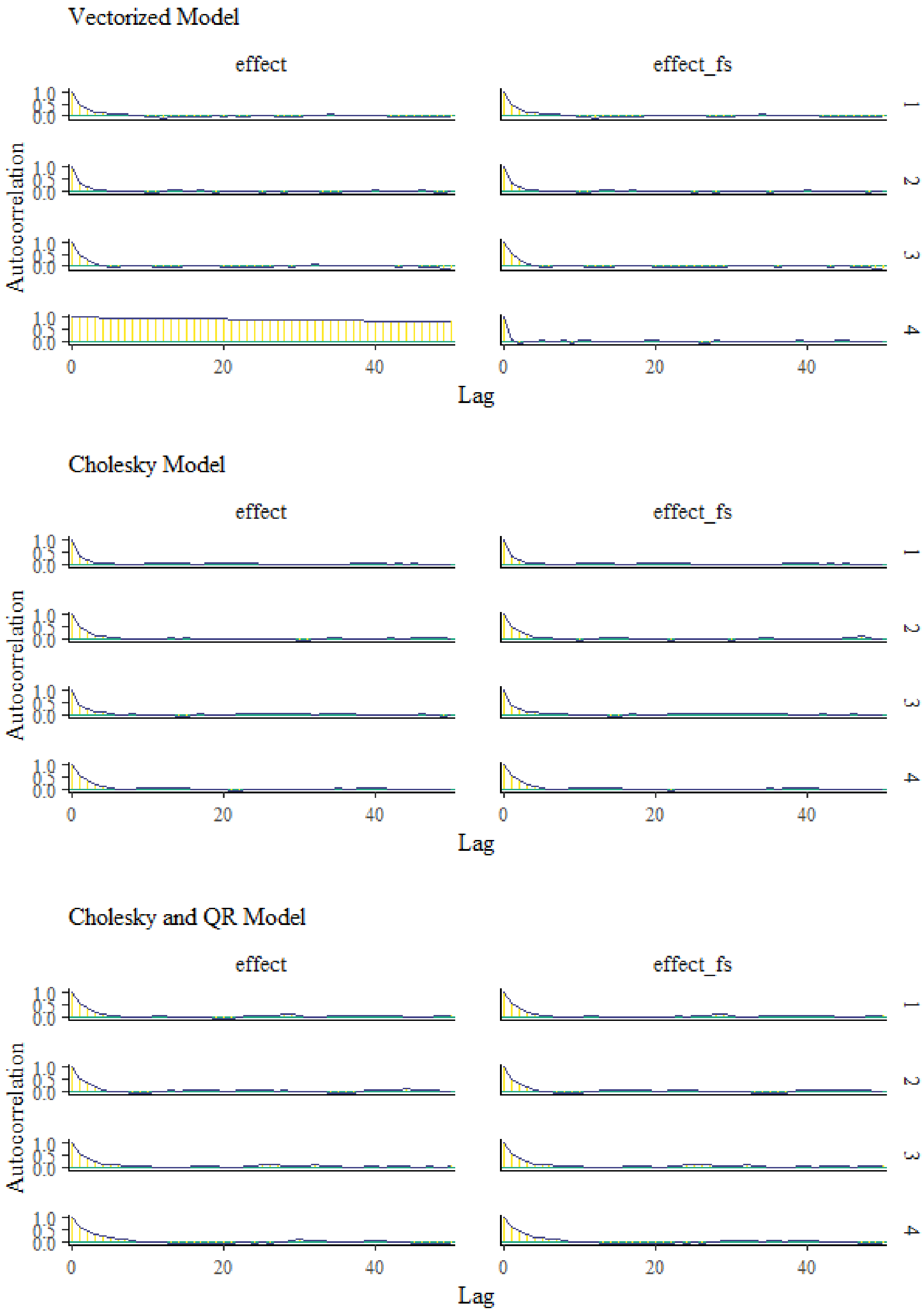


Figure 3: Autocorrelation Plots

The problem of a small effective sample size for the vectorized model is illustrated directly in figure 4, showing the proportion of effective sample size to total sample size by chain. The vectorized model exhibits an extremely low effective sample size of close to 0% of the total sample size. The Cholesky as well as the Cholesky and QR model perform much better in this regard as well, with values of more than 25%. In addition to this, in the Cholesky model, the ratio of mean effective sample size to mean number of gradient evaluations is equal to 0.0147 for the population effect and 0.0141 for the finite population effect, while the same two numbers are equal to 0.0197 and 0.0189 in the Cholesky & QR model. Thus, the effectiveness of each gradient evaluation is higher with the QR reparametrization.

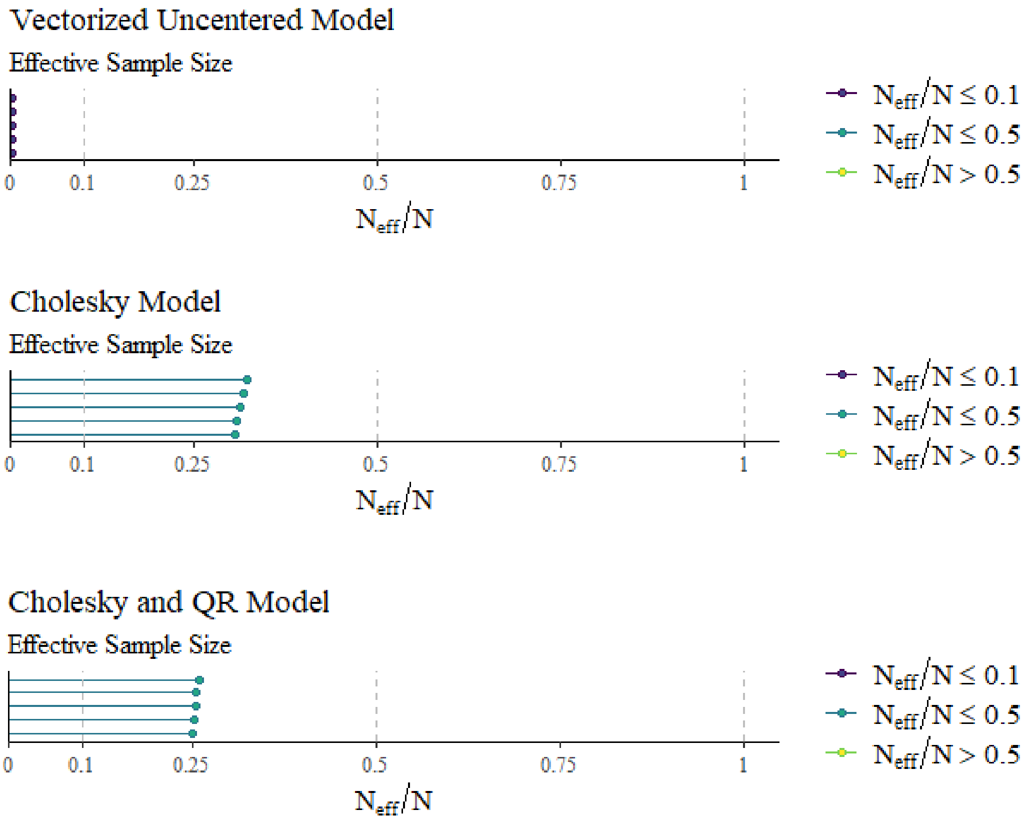


Figure 4: Plots of the Effective Sample Sizes

Next we turn to energy plots as shown in figure 5. Following Betancourt (2017), we can use these plots to assess the efficiency of our Markov Chains. In this framework, Markov Chains are conceptualized as exploring, through resampling, energy distributions. The performance of a chain is determined by how quickly the random walk can diffuse across the so called marginal energy distribution (π_E in figure 5). When the transitional energy distribution ($\pi_\Delta E$) is narrow compared to the marginal energy distribution, exploration happens very slowly, indicating low efficiency. In our cases, we see transitional energy distributions that correspond relatively well to the marginal energy distributions across all chains.

Furthermore, QR reparametrizations often require a lower number of gradient evaluations, which tends to decrease run time. In this application, the total number of gradient evaluations performed in the Cholesky model is 346336 across four chains while it equals 257824 across the four chains of the Cholesky & QR model. This decrease of approximately 25% significantly reduces run time and is an additional advantage of this reparametrization.

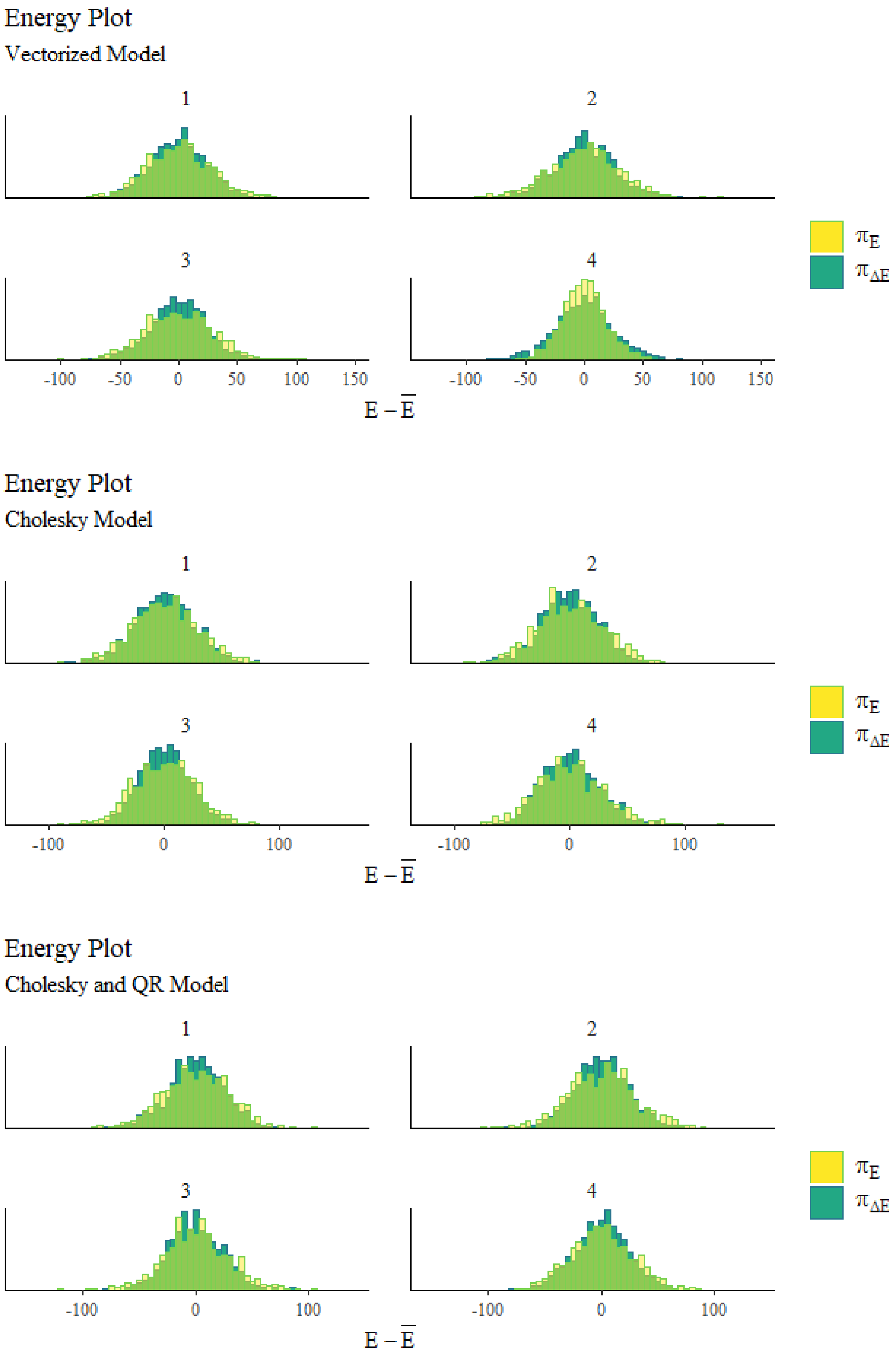


Figure 5: Energy Plots

5 Results

5.1 Comparison with regression

As mentioned in section 3, (Duflo et al., 2011) find a treatment effect of 0.138 (0.078) sd of the totalscore. Moreover, such effect is found to increase when adding covariates and interaction terms in successive columns of Figure 1. On the other hand, our Bayesian model suggests that the mean of the population treatment effect is 0.87 (0.080) and the mean of the finite sample treatment effect is 0.88 (0.079). These effects are not statistically significant using symmetric 90% credible intervals. As this approach to causal inference allows us to derive the distribution of treatment effects, we are able to notice that the densities of both treatment effects are slightly left skewed, as the means are lower than the medians. The means and credible intervals of the population treatment effect, finite sample treatment effect, three quantiles of the finite sample treatment effect, $\sigma_c|x$ and $\sigma_t|x$ are displayed in Figure 6.

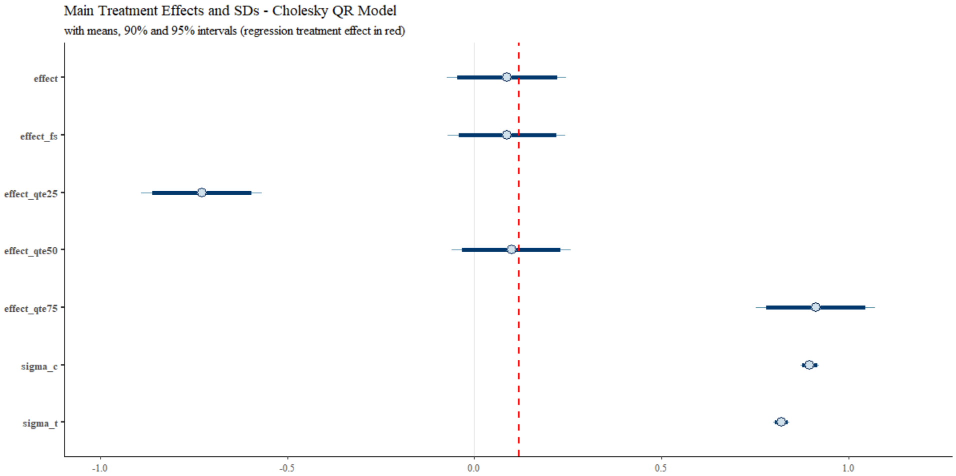


Figure 6: Summary of Mean and Quantile Treatment Effects and SDs

The differences between our estimates and (Duflo et al., 2011)'s ones are due to two separate reasons: first, unlike regression, our model is able to estimate the shape of the distribution of the treatment effect; second, (Duflo et al., 2011)'s non-baseline estimates use a different sample than the original one.

As to the first point, the distribution we estimate for the treatment effect is left-skewed. Since regression confidence intervals must be centered on the mean, the shape of the distribution we estimate makes the point estimate comparison between the two models inappropriate. Indeed, when we compare 90% credibility intervals with regression 90% confidence intervals, we can see how our estimates is $(-0.014, 0.189)$, while (Duflo et al., 2011)'s baseline estimate (which has to be centered on the mean by construction) is $(0.0102, 0.269)$. Moreover, including the variables in our model as controls in (Duflo et al., 2011)'s baseline specification, the 90% confidence interval becomes: $(0.009, 0.228)$, centered in 0.119. When considering intervals, thus, differences in the two modes of estimation are much less stark and likely due to the constraint of centered intervals for regressions.

As to the second point, in columns (2)-(4) of Figure 1, (Duflo et al., 2011) use baseline students' percentile as a covariate. Such variable, though, is observed only for a subsample amounting to 5279 observations. Thus, w.r.t. the sample for which final scores are available, specifications including percentile baseline mark as a covariate drop 526 observations. All such observations belong to the control group, which could undermine randomization assumptions if the attriting students (i.e. the dropped ones) are systematically different from the non-attriting ones. Indeed, this would result in the control group not being a good counterfactual for the treatment group anymore.

To control for differential attrition, thus, we performed a test that randomly picks 526 attriters out of the original sample and regresses the covariates included in the model on the artificial attrition dummy, capturing differences in means in terms of such variables between simulated attriters and simulated non-attriters. Then a statistic is chosen (we have used the maximum out of the t-stats on the coefficient of the attrition dummy in all such regressions) and recorded and the the artificial randomization is repeated. Finally, the observed statistics (i.e. the one resulting from the actual randomization) is compared with the distribution of the statistics resulting from the simulated randomizations, and a p-value is calculated as the proportion of permutations where the simulated statistics is greater in absolute value than the observed one. Such procedure allows to test the null hypothesis that being an attiter has an effect whatsoever on the set of covariates in our model, also accounting for multiple hypotheses testing. The test we performed rejects the null, thus suggesting that it is not safe to drop the observations for which the variable percentile is missing.

We argue that (Duflo et al., 2011)'s non-baseline results are driven by dropping of such observations. To support such claim, we imputed missing values by regressing percentile on all other covariates and then predicting their values for missing observations. Then, we ran a regression saturated in quartiles of baseline marks, quartiles of age and gender, i.e. a regression including 32 cells dummies, one for each possible combination of the above variables, and a dummy for the treatment. In such saturated model, the coefficient on the treatment dummy is a weighted average of cell-wise mean differences between treatment and control groups, with a weight for cell x proportional to $P(X = x)V(W|X = x)$. The regression was run on both the complete sample and the subsample used by (Duflo et al., 2011). What can be observed is that dropping observations slightly moves to the right the distribution and mean of cell mean differences, thus inflating the coefficient on the treatment dummy through cell-wise mean differences, as we can observe in Figure 7. More importantly, though, the dropping induces a positive correlation between mean differences and weights, meaning that in (Duflo et al., 2011)'s subsample the cells with higher mean differences are weighted more, while the opposite happens in the complete sample. In such subsample, thus, the coefficient on the treatment dummy is also inflated through higher weights on cells with higher differences.

Figure 8 shows that the regression tends to give more weight to cells with lower mean differences in the complete sample, while the opposite occurs in the dropped one. Furthermore, figure 9 shows a scatter plot of the cell-wise variations in mean differences in outcomes and regression weights, in which the values for the dropped sample are subtracted from the values for the complete sample, i.e. $w_j = w_{j,comp} - w_{j,drop}$ and $\Delta_j = (\bar{Y}_{j,t,comp} - \bar{Y}_{j,c,comp}) - (\bar{Y}_{j,t,drop} - \bar{Y}_{j,c,drop})$. As one can notice, cells with lower outcome

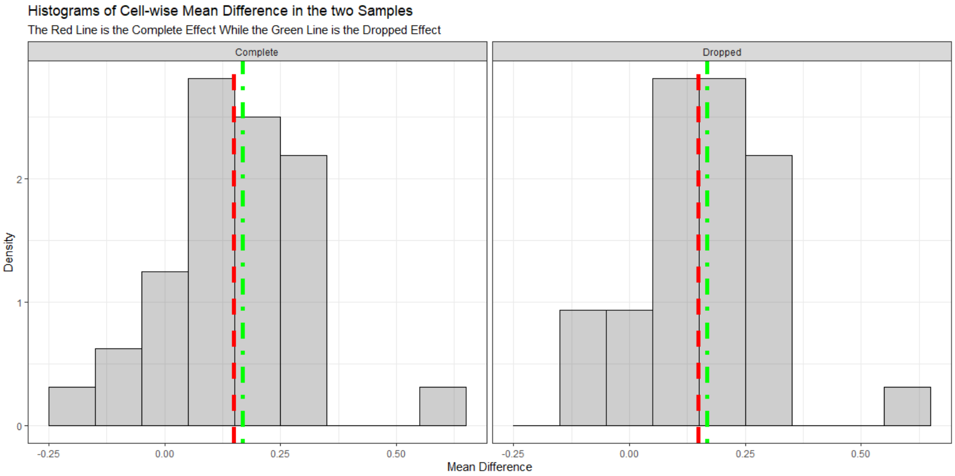


Figure 7: Histogram of Mean Differences by Cells for Both Samples

differences are weighted more in the complete sample, which explains why dropping students from the sample produces a higher treatment effect estimate.

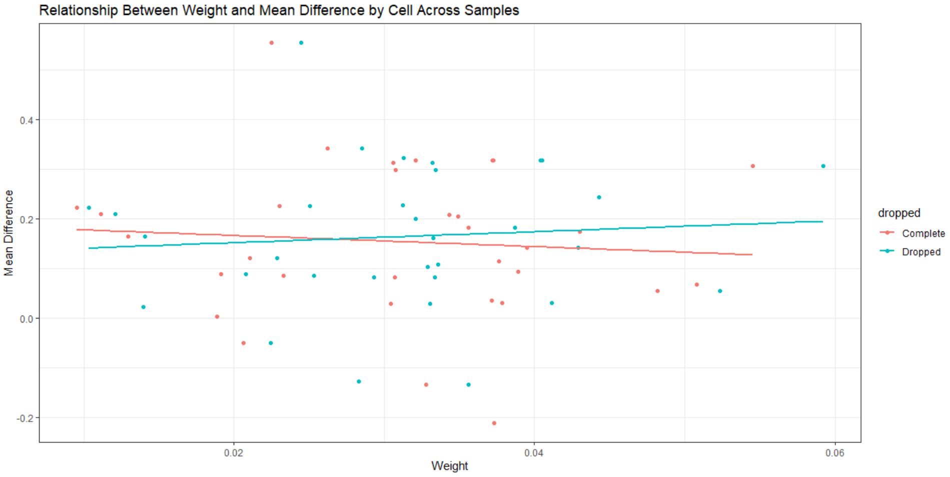


Figure 8: Scatterplot of Mean Differences and Weights by Cells Across the Two Samples

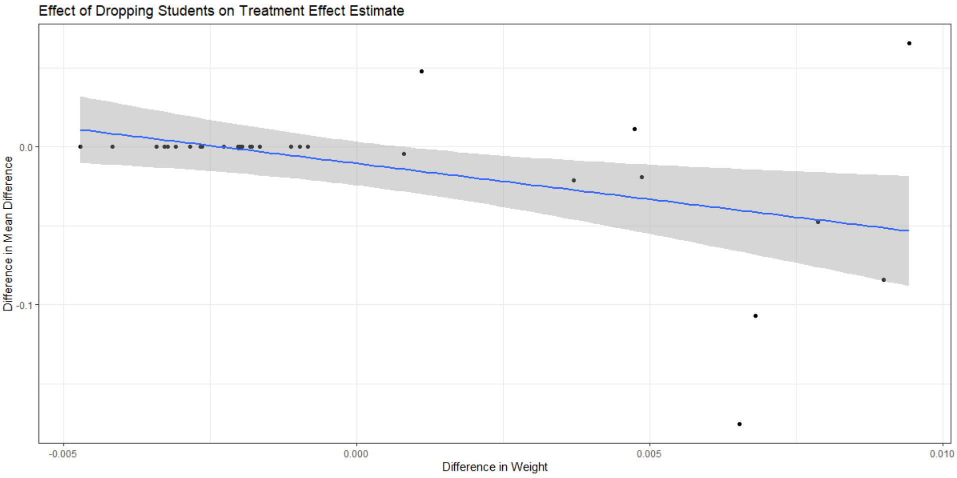


Figure 9: Scatterplot of Variations in Mean Differences and Weights Between the Completed and Dropped Samples

5.2 Treatment Effect Heterogeneity

As our approach is able to identify individual treatment effects, one can aggregate them according to some criterion, such as a partition of the covariate space, to study the heterogeneity of the treatment effect. We do this by dividing the students by (partly imputed) quartiles of the pre-treatment standardized mark to replicate the regression with the interaction between treatment and quartiles of the standardized mark performed in the paper. In addition to this aggregation, we explore the heterogeneity of the treatment effect according to several other partitions, both one- and multi-dimensional. The credible intervals of the treatment effects by quartile of standardized mark and by other partitions are displayed in figure 10.

As one can see, the only statistically significant treatment effect is the one among the students in the first and second quartiles of the standardized mark, i.e. those assigned to the lower classroom. This is in contrast with the results of (Duflo et al., 2011), which state that the treatment effect is not statistically different between the students assigned to the lower and upper classroom. Furthermore, treatment effects aggregated according to other partitions based on other covariates, such as age, gender, additional programs, etc.,

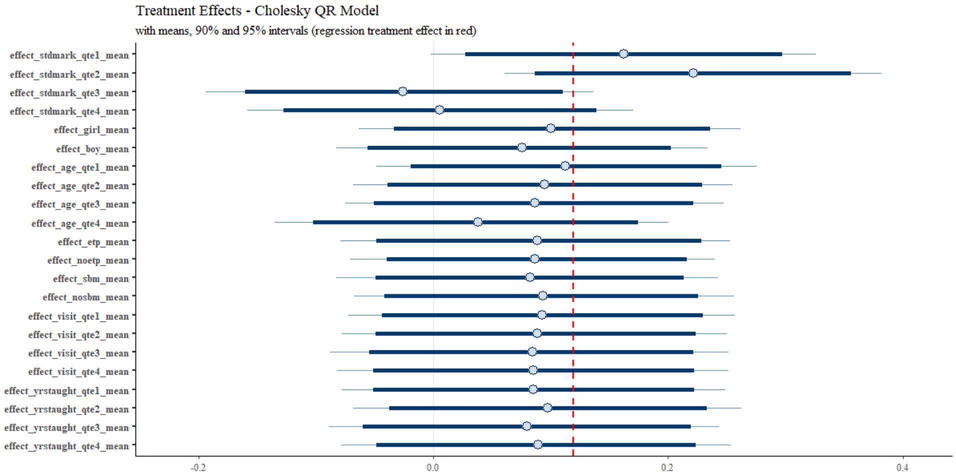


Figure 10: Summary of Mean Treatment Effects by Quartiles of Std. Mark and Other Partitions

all contain zero in their credible intervals. In addition to these results, credible intervals of treatment effects by zones all include zero.

To further explore treatment effect heterogeneity, we aggregate students by means of a three-dimensional partition of the covariate space using gender, standardized pre-treatment mark quartiles and age quartiles, thus obtaining 32 cells. Figures 11 and 12 contain 90% credible intervals of mean and median treatment effects by cell for each gender.

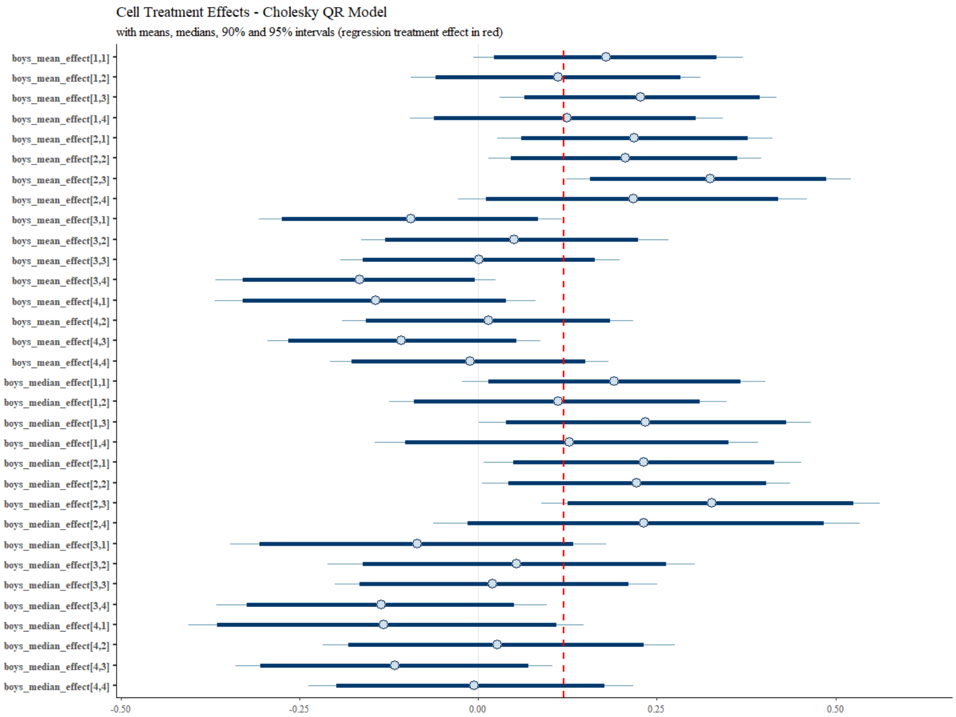


Figure 11: Summary of Mean and Median Treatment Effects by Quartiles of Std. Mark for Boys

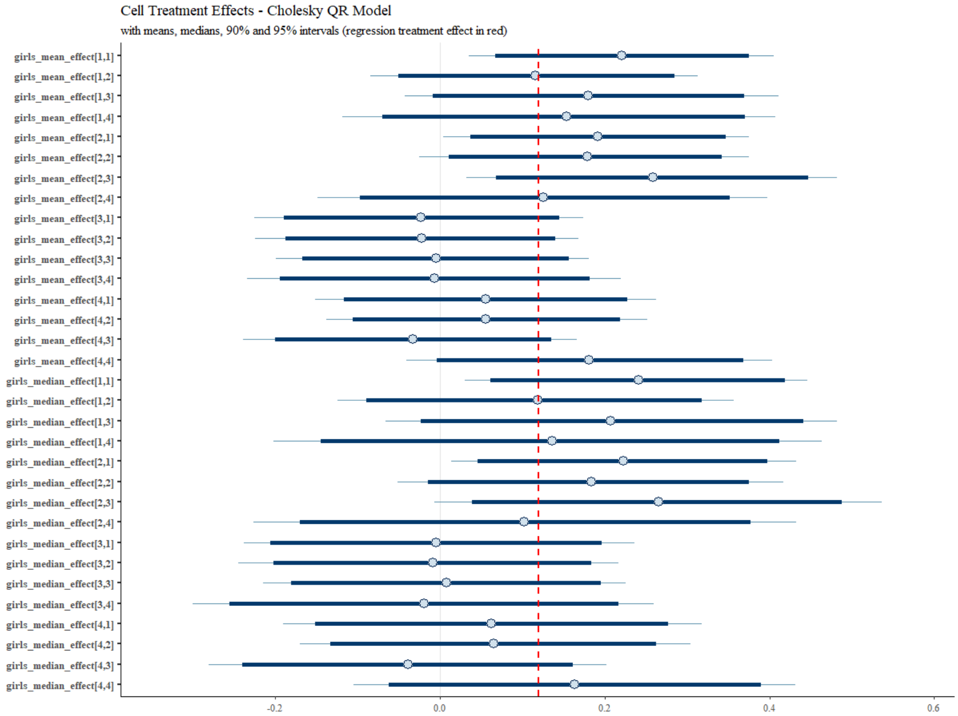


Figure 12: Summary of Mean and Median Treatment Effects by Quartiles of Std. Mark for Girls

The first number of a cell refers to the standardized mark quartile while the second number refers to the age quartile. As one can see, the only significant mean and median treatment effects belong to cells with standardized marks below the median, no matter the gender. Nonetheless, there is less evidence of significant treatment effects in the lower classroom among girls, as a greater proportion of mean and median effects are insignificant. At the same time, the absence of patterns in the credible intervals suggest no clear effect of age on the effect of tracking on school outcomes, although treatment effects in third age quartile exhibit slightly higher means across genders.

6 Conclusion

In our work, we have presented a Bayesian approach to the analysis of randomized control trials (RCTs). We have considered several parametrizations for a baseline hierarchical model and analyzed their relative performance. This allowed us to identify the best parametrization among the ones presented. Then, we used the resulting model to analyze data from an RCT studied in Duflo, Dupas, and Kremer (2011). While regression techniques grant a straightforward simple method to reliably detect average treatment effects in such contexts, we have shown that adding controls and interactions to baseline specifications may bias results in finite samples, particularly so when information on the control variables is available only for a subsample. Thus, we argue that, in order to detect treatment effect heterogeneity, a Bayesian model is to be preferred. We also argue that assigning a prior to the estimandum (i.e. the treatment effect) allows to more flexibly estimate a distribution for the quantity of interest, thus improving on regression.

As we have explored, there are many assumptions (many more than in regression) to be made, on which the results of a Bayesian analysis rely. But, as long as large samples are concerned, the influence of most assumptions fade, leading to a coincidence between Bayesian and regression point estimates. Such point estimates are reliable in RCTs, as long as the random assignment of treatment is credible. Given the different intrinsic value of Bayesian approaches vis a vis regression, with the former more precise and thorough but relying on more assumptions and the latter almost assumptions-free but reliable only as a point estimate for the average treatment effect, we argue that RCTs should present both analyses separately.

References

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv: Methodology*. Retrieved from https://arxiv.org/pdf/1701.02434.pdf

Betancourt, M. (2020). *Hierarchical modelling.* Retrieved from https://betanalpha.github.io/assets/case_studies/hierarchical_modeling.html#5_Multivariate_Normal_Hierarchical_Models

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053-1062. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473858 doi: 10.1080/01621459.1999.10473858

Dehejia, R. H., & Wahba, S. (2002, 02). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, *84*(1), 151-161. Retrieved from https://doi.org/10.1162/003465302317331982 doi: 10.1162/003465302317331982

Duflo, E., Dupas, P., & Kremer, M. (2011, August). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, *101*(5), 1739-74. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.101.5.1739 doi: 10.1257/aer.101.5.1739

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* USA: Cambridge University Press.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, *76*(4), 604–620. Retrieved from http://www.jstor.org/stable/1806062

Lee, J.-H., Feller, A., & Rabe-Hesketh, S. (2018). *Model-based Inference for Causal Effects in Completely Randomized Experments.* Retrieved from https://mc-stan.org/users/documentation/case-studies/model-based_causal_inference_for_RCT.html

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. Retrieved from https://doi.org/10.1037/h0037350

Stan Development Team. (2019). *Stan User's Guide.* Retrieved from https://mc-stan.org/docs/2_27/stan-users-guide/index.html