Chapter 4

Regularization and Confounding in Linear Regression for Treatment Effect Estimation

Some analysis and text in this chapter follows Hahn et al. [2018a]. We will reintroduce the main ideas described in that paper and discuss ongoing research.

4.1 Introduction

While the previous chapters focused on regularization and its purpose within a formal model selection setting, this chapter considers its use as solely a shrinkage/bias inducing technology. Specifically, we are interested in regularization's role in treatment effect estimation with observational (and potentially clustered) data.

A treatment effect – the amount a response variable would change if a treatment variable were changed by one unit – is appropriately estimated only when all other *confounding* variables are taken into account. Confounding variables are given such a name because they explain a portion of the correlation between the treatment and response variables; effectively masking (confounding) the true relationship the data analyst wishes to measure. The

models considered in this chapter specify a linear relationship between the response Y_i, and the treatment and covariates Z_i and X_i respectively:

$$Y_i = \alpha Z_i + \mathrm{X}_i^T \beta + \nu_i. \tag{4.1}$$

For notation, let letters denote vectors, boldfaced letters denote matrices, and italicized letters denote scalars. Let β be a p-length vector of coefficient parameters, and α be the scalar treatment effect parameter. The errors ν are normally distributed with zero mean and unknown variance. In observational studies, there may be many covariates, i.e.: p may be large. For example, corporate finance studies often involve observations that are firms and covariates that are firm characteristics taken from financial statements. The relationship of interest may be the effect of a firm's cash flow (Z_i) on its debt-to-equity ratio (Y_i), and there are a plethora of firm characteristics one can include as part of X_i. Although including all covariates may mitigate bias in the estimate for α, this is at the expense of increased variance of the estimator for α. Practically, interval lengths of the treatment effect for this naive approach will be large, and discovering statistical significance will be difficult.

One solution to the "many covariate" problem is to hand-select a subset of variables from X_i to control for in Model (4.1) and toss out the remaining covariates. Leamer [1983] describes how this procedure is an unsatisfyingly ad-hoc reaction to a practical data analysis issue. After hand-selecting covariates, how does the analyst truly know if all information from X_i is taken into account? This chapter provides a solution to this problem using statistical

regularization. Specifically, we propose using information from marginal likelihoods to narrow down our potentially large list of covariates. The aim is to replace an ad-hoc selection approach with one that is informed by the data, and our desire is to appeal to a broad base of researchers estimating linear treatment effects from observational data.

Exploring the use of regularization in treatment effect estimation and providing a procedure was the main contribution of Hahn et al. [2018a]. We extend these ideas to an empirical-Bayes setting and where model errors may be dependent across clusters of data. The forthcoming sections are speculative, but compare this new approach to Hahn et al. [2018a] and discuss areas of future work.

4.1.1 Previous literature

Treatment effect estimation is an important topic with a deep literature base. This work focuses on one slice: The use of Bayesian regularized regression models for effect estimation. Li and Tobias [2014] and Heckman et al. [2014] provide review articles of Bayesian approaches to this problem. Careful attention will be given to the impact of regularization on the estimation of treatment effects and new ways for characterizing the estimates' standard errors.

Hahn et al. [2018a] contributed to the small but growing literature on Bayesian approaches to treatment effect estimation via linear regression with many potential controls. They proposed a conceptual and computational

refinement of ideas first explored in Wang et al. [2012], where Bayesian adjustment for confounding is addressed via hierarchical priors. Their method can be seen as an alternative to Wang et al. [2012], with certain conceptual and computational advantages, namely ease of prior specification and posterior sampling. Other related papers include Wang et al. [2015], Lefebvre et al. [2014] and Talbot et al. [2015]; see also Jacobi et al. [2016]. Zigler and Dominici [2014] and An [2010] focus on Bayesian propensity score models (for use with binary treatment variables). Wilson and Reich [2014] takes a decision theoretic approach to variable selection of controls. Again, each of these previous approaches cast the problem as one of selecting appropriate controls; posterior treatment effect estimates are obtained via model averaging. Here, we argue that if the goal is estimation of a certain regression parameter (corresponding to the treatment effect, provided the model is correctly specified), then questions about which specific variables are necessary controls is a means to an end rather than an end in itself. Other recent papers looking at regularized regression for treatment effect estimation include Ertefaie et al. [2015] and Ghosh et al. [2015], but even here the focus is on variable selection via the use of 1-norm penalties on the regression coefficients.

There are several books dealing with the broader topic of causal inference, including Imbens and Rubin [2015], Morgan and Winship [2014], and Angrist and Pischke [2008]. Similar to Wang et al. [2012] where there is a focus on the joint modeling of the treatment and response variables as a function of covariates, the following papers have approached the problem similarly:

Rosenbaum and Rubin [1983], Robins et al. [1992], and McCandless et al. [2009].

Equally important and vast is the literature dealing with clustered inference. We defer this literature review to a final section on the application of our approach to the clustered data setting.

4.2 Regularization-induced confounding

What happens when regularization is naively used in treatment effect estimation? In this section, we illustrate this important phenomena, referred to as "regularization-induced confounding" (RIC). Hahn et al. [2018a] and Hahn et al. [2017] provide intuition for this issue within Bayesian linear regression and heterogenous treatment effect estimation using random forests, respectively. We recapitulate their exposition here since RIC is a central issue of this chapter. It is expected that regularization will introduce bias in coefficient estimates from a regression. What is not obvious is that bias will still exist in an *unregularized* treatment effect estimate if the treatment and covariates are correlated. For illustration, suppose regularization is introduced via a ridge penalty over parameters. A similar theoretical demonstration of RIC is presented in Hahn et al. [2017].

Returning to Model (4.1):

$$Y_i = \alpha Z_i + \mathrm{X}_i^T \beta + \nu_i,$$

the overall goal is to properly estimate the treatment effect α. We assume

that the error term is mean zero Gaussian and a ridge estimator is placed on the regression coefficients. Define observed data as $\tilde{\mathbf{X}} = (\mathrm{Z} \ \mathbf{X})$ and consider a ridge matrix $\mathbf{M}$. Following the seminal work of Hoerl and Kennard [1970], the ridge estimator for coefficients $\theta = \begin{pmatrix} \alpha & \beta^T \end{pmatrix}^T$ is $\hat{\theta}_{\text{ridge}} = (\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathrm{Y} = (\mathbb{I}_{p+1} - (\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{M})\hat{\theta}$ where $\hat{\theta} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathrm{Y}$ is the maximum likelihood estimator for θ. Taking expectation of the ridge estimator yields $\mathbb{E}[\hat{\theta}_{\text{ridge}}] = (\mathbb{I}_{p+1} - (\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{M})\theta$, so we have the bias as the second term within the parentheses:

$$\text{bias}(\hat{\theta}_{\text{ridge}}) = -(\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{M}\theta \tag{4.2}$$

Consider a diagonal ridge matrix that leaves the treatment effect unregularized $\mathbf{M} = \begin{bmatrix} 0 & 0 \\ 0 & \lambda\mathbb{I}_p \end{bmatrix}$. Using this ridge matrix and the block inversion formula for $(\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$, the bias for the treatment effect may be expressed as:

$$\text{bias}(\hat{\alpha}_{\text{ridge}}) = -(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_p - \mathbf{X}^T\hat{\mathbf{X}}_{\mathrm{Z}}\right)^{-1}\lambda\beta \tag{4.3}$$

where $(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathbf{X}$ is a p-length vector of coefficients from p univariate regressions of each X_j on Z and $\hat{\mathbf{X}}_{\mathrm{Z}} = \mathrm{Z}(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathbf{X}$ are the predicted values from these regressions. For all $\lambda > 0$, we see that Equation (4.3) will be nonzero; especially in the case when the X_j's are correlated with the treatment Z (confounding exists). Also pointed out by Hahn et al. [2018a], the treatment effect bias is not a function of the true treatment α, but instead the unknown (and likely nonzero) coefficient vector β. Of course, the OLS estimate of α is obtained when $\lambda = 0$, in which case the estimator is unbiased. In sum, Equation

(4.3) analytically highlights the issue of bias in the treatment effect estimate should a practitioner choose to regularize a linear treatment effect model.

4.2.1 Mitigating regularization-induced confounding

How can we avoid RIC in treatment effect estimation from observational data? We will describe two approaches: (*i*) Controlling for the propensity of Z, and (*ii*) Replacing the treatment with a proxy for random treatment variation *excluding* X. Both of these approaches require the addition of an equation to Model (4.1) that accounts for the relationship between Z and X:

$$\begin{aligned} &\text{Selection equation:} \quad Z_i = \mathrm{X}_i^T\gamma + \epsilon_i, && \epsilon_i \sim N(0, \sigma_\epsilon^2) \\ &\text{Response equation:} \quad Y_i = \alpha Z_i + \mathrm{X}_i^T\beta + \nu_i, && \nu_i \sim N(0, \sigma_\nu^2), \end{aligned} \tag{4.4}$$

thereby learning about confounding through the parameter γ. The first is called the selection equation since it determines which X_i's should be "selected" for controls, and the second is the original response equation. First, we briefly show how including an estimate of the propensity function from the selection equation: $\hat{\mathrm{Z}} \approx \mathbf{X}\hat{\gamma}$ can mitigate bias from RIC. Suppose we augment our covariates with predicted values for the treatment $\hat{\mathbf{X}}_{\text{new}} = \left(\mathrm{Z}\ \hat{\mathrm{Z}}\ \mathbf{X}\right)$. Effectively, we are including information about the predictable variation in the treatment described by the original controls $\mathbf{X}$. Using the same calculations to arrive at Equation (4.3) now unpenalizing the coefficients associated with *both* Z and $\hat{\mathrm{Z}}$, the bias of the treatment effect can be written as:

$$\text{bias}(\hat{\alpha}_{\text{ridge}}) = -\{(\tilde{\mathrm{Z}}^T\tilde{\mathrm{Z}})^{-1}\tilde{\mathrm{Z}}^T\mathbf{X}\}_1 \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_p - \mathbf{X}^T\hat{\mathbf{X}}_{\mathrm{Z}}\right)^{-1} \lambda\beta \tag{4.5}$$

where $\tilde{\mathrm{Z}} = \left(\mathrm{Z}\ \hat{\mathrm{Z}}\right)$ and $\{\cdot\}_1$ corresponds to the top row of the matrix $\{\cdot\}$. $\{(\tilde{\mathrm{Z}}^T\tilde{\mathrm{Z}})^{-1}\tilde{\mathrm{Z}}^T\mathbf{X}\}_1$ are the coefficients on Z in the two variable regressions of each X_i on $\left(\mathrm{Z}\ \hat{\mathrm{Z}}\right)$. Since the propensity estimate $\hat{\mathrm{Z}}$ accounts for variation in Z due to the controls, the coefficient on Z in these univariate regressions is approximately zero which renders the bias of the treatment effect close to zero. This feature will be illustrated in simulations to follow.

Hahn et al. [2018a] discuss a reparameterization of Model (4.6) that allows for regularization via Bayesian shrinkage priors in both equations while mitigating RIC – an alternative to controlling for the propensity of Z. The following parameter transformation

$$\begin{pmatrix} \alpha \\ \beta + \alpha\gamma \\ \gamma \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta_d \\ \beta_c \end{pmatrix}$$

results in a new formulation of Model (4.6):

$$\begin{aligned} &\text{Selection equation:} \quad Z_i = \mathrm{X}_i^T\beta_c + \epsilon_i, && \epsilon_i \sim N(0, \sigma_\epsilon^2) \\ &\text{Response equation:} \quad Y_i = \alpha(Z_i - \mathrm{X}_i^T\beta_c) + \mathrm{X}_i^T\beta_d + \nu_i, && \nu_i \sim N(0, \sigma_\nu^2). \end{aligned} \tag{4.6}$$

Conveniently, β_c and β_d nicely separate the roles covariates play in treatment effect estimation. A covariate X_{ij} that is distinctly predictive of the response will have $\beta_{dj} \neq 0$ and $\beta_{cj} = 0$. As common in medicine, this covariate may also be called prognostic. Alternatively, the covariate may be a confounder, in which case $\beta_{cj} \neq 0$, $\beta_{dj} \neq 0$. This formulation provides an intuitive interpretation of the treatment effect. The selection equation provides the variation of the treatment excluding X ($\epsilon_i = Z_i - \mathrm{X}_i^T\beta_c$) that is then used to infer α. In

other words, the residual ϵ_i may be thought of as a "randomized experiment" that we then use to infer the treatment effect.

Examining the bias of the treatment effect under a ridge matrix that leaves the treatment effect unregularized ($\mathbf{M} = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \mathbb{I}_p \end{bmatrix}$) yields:

$$\text{bias}(\hat{\alpha}_{\text{ridge}}) = -(\mathrm{R}^T\mathrm{R})^{-1}\mathrm{R}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_p - \mathbf{X}^T\hat{\mathbf{X}}_{\mathrm{Z}}\right)^{-1}\lambda\beta_d \tag{4.7}$$

where $\mathrm{R} = \mathrm{Z} - \mathbf{X}\beta_c$. Further, $(\mathrm{R}^T\mathrm{R})^{-1}\mathrm{R}^T\mathbf{X}$ will be close to the zero vector since $R_i = Z_i - \mathrm{X}_i^T\beta_c$ is independent of X_i. In this case, the treatment likelihood given by the selection equation is crucial in providing information on β_c and thus R_i.

4.3 Regularization using the marginal likelihood

What remains to be discussed is how an analyst should choose the level of regularization. In the ridge regression case, this would amount to choosing two λ's for ridge priors on the coefficients in the treatment and response models shown in Model (4.6). Hahn et al. [2018a] approach this in a Bayesian regression context by regularizing using a variant of the horseshoe prior on the regression coefficients from Carvalho et al. [2010b]:

$$\begin{aligned} \pi(\beta_j) &\propto \frac{1}{v}\log\left(1 + \frac{4}{(\beta_j/v)^2}\right), \\ \pi(v) &\sim \mathrm{C}^+(0,1), \end{aligned} \tag{4.8}$$

where v is a global scale parameter common across all elements $j = 1, \ldots p$, and $\mathrm{C}^+(0,1)$ denotes a folded standard Cauchy distribution. Such priors have