Machine Learning for Econometrics: exercise solutions

Christophe Gaillac
University of Geneva
CREST

Jérémy L'Hour
Capital Fund Management
CREST

Version: May 17, 2025

This document regroups elements of solution for the exercises in Chapter 15 of “Machine Learning for Econometrics".

15.1 Regression as a Weighting Estimator

1. A first argument is that $DY = DY_1$ almost surely. A second argument is:
$$\mathbb{E}\left[DY_0\right] = \mathbb{E}\left[DX'\beta_0 + D\varepsilon\right] = \mathbb{E}\left[DX\right]'\beta_0.$$

2. The theoretical value is
$$\beta_0 = \mathbb{E}\left[(1-D)XX'\right]^{-1}\mathbb{E}\left[(1-D)XY\right].$$
Its empirical counterpart is given by:
$$\widehat{\beta} = \left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iY_i\right].$$
$\widehat{\beta}$ is also obtained from a regression of Y on X using the sample of untreated units.

3. Starting from the first question, we can show that:
$$\begin{aligned}\pi\,\mathbb{E}\left[Y_1 - Y_0 \mid D=1\right] &= \mathbb{E}\left[D(Y_1-Y_0)\right]\\ &= \mathbb{E}[DY] - \mathbb{E}[DX]'\beta_0\\ &= \mathbb{E}[DY] - \mathbb{E}[DX]'\mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[(1-D)XY]\\ &= \mathbb{E}[DY] - \mathbb{E}\left[(1-D)\,\mathbb{E}[DX]'\,\mathbb{E}[(1-D)XX']^{-1}XY\right]\\ &= \mathbb{E}\left[DY - (1-D)W_0Y\right]\end{aligned}$$
where we define
$$W_0 := \mathbb{E}[DX]'\,\mathbb{E}[(1-D)XX']^{-1}X.$$

4. Since W_0 is scalar-valued, we can write:

$$\begin{aligned}\mathbb{E}\left[(1-D)XW_0\right] &= \mathbb{E}\left[(1-D)XW_0'\right]\\ &= \mathbb{E}\left[(1-D)XX'\mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[DX]\right]\\ &= \mathbb{E}[(1-D)XX']\,\mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[DX]\\ &= \mathbb{E}[DX].\end{aligned}$$

The reweighted control group characteristics have the same first moment as those of the treated group.

5. The weights are given by

$$\omega_j = n_1^{-1}\left[\sum_{i=1}^{n} D_i X_i\right]'\left[\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]^{-1} X_j.$$

We use the same trick as before:

$$\begin{aligned}\sum_{j:D_j=0} X_j\,\omega_j &= \sum_{j=1}^{n}(1-D_j)X_j\,\omega_j'\\ &= n_1^{-1}\sum_{j=1}^{n}(1-D_j)X_jX_j'\left[\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]^{-1}\left[\sum_{i=1}^{n}D_iX_i\right]\\ &= n_1^{-1}\sum_{i=1}^{n}D_iX_i\\ &= n_1^{-1}\sum_{i:D_i=1}X_i.\end{aligned}$$

Suppose the first element of X is a constant term, then

$$\sum_{j:D_j=0}\omega_j = n_1^{-1}\sum_{i:D_i=1}1 = 1.$$

Consequently, the sum of the weights equals one.

6. In the synthetic control estimator, the weights sum to one and are positive, which is not necessarily the case here: although the weights sum to one, they can be negative. Moreover, by construction this estimator reproduces the characteristics of the treated group, whereas this may not be the case with synthetic control. Finally, although it is not directly visible here, the synthetic control produces a sparse solution, which is not true here.

15.2 Orthogonal Score for the Treatment Effect on the Treated

1. (a) First, note that $\mathbb{E}[DY^{obs}] = \mathbb{E}[DY_1]$. Next:

$$\begin{aligned}\mathbb{E}[\exp(X'\beta_0)(1-D)Y^{obs}] &= \mathbb{E}\left[\frac{p(X)}{1-p(X)}(1-D)Y_0\right] \\ &= \mathbb{E}\left[\frac{p(X)}{1-p(X)}\mathbb{E}[(1-D)Y_0 \mid X]\right] \\ &= \mathbb{E}\left[\frac{p(X)}{1-p(X)}\mathbb{E}[(1-D) \mid X]\mathbb{E}[Y_0 \mid X]\right] \\ &= \mathbb{E}\left[p(X)\mathbb{E}[Y_0 \mid X]\right] \\ &= \mathbb{E}\left[DY_0\right].\end{aligned}$$

Thus, $\mathbb{E}[DY^{obs}]-\mathbb{E}[\exp(X'\beta_0)(1-D)Y^{obs}] = \mathbb{E}[D(Y_1-Y_0)]$. Finally, note that $\mathbb{E}[D(Y_1 - Y_0)] = \tau_0 P[D = 1] = \mathbb{E}[D\tau_0]$.

(b) The propensity score $p(X)$ can be estimated by logistic regression, which yields an estimator for β_0. Then, $\hat{\tau}$ can be estimated by:

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(D_i - (1-D_i)\exp(X_i'\hat{\beta})\right)Y_i^{obs}}{\frac{1}{n}\sum_{i=1}^{n}D_i}.$$

2. (a) One can show that:

$$\mathbb{E}[\partial_\beta m(W, \tau_0, \beta_0)] = -\mathbb{E}[XDY_0] \neq 0.$$

(b) The most efficient method is maximum likelihood estimation, given by the equation (LOGIT) in section (2.3.2) of the book, where Y_i is replaced by D_i. The estimator $\hat{\tau}$ of τ_0 found in question 1.a, using the maximum likelihood estimator $\hat{\beta}$, will be asymptotically normal. This is straightforward to show.

(c) In high-dimensional settings, it will be necessary to use machine learning methods whereby the estimator of β_0 will not be asymptotically normal (for example, a Lasso logistic regression as mentioned in the prompt). This will introduce bias in $\hat{\tau}$ since generally:

$$\mathbb{E}[\partial_\beta m(W, \tau_0, \beta_0)] = -\mathbb{E}[XDY_0] \neq 0.$$

This is shown in Chapter 4.

3. (a) The Conditional Independence Assumption (CIA) implies $Y_0 - X'\gamma_0 \perp D \mid X$, hence

$$\mathbb{E}[DX(Y_0 - X'\gamma_0)] = \mathbb{E}[X\mathbb{E}[\varepsilon \mid X]\mathbb{E}[D \mid X]] = 0.$$

(b) Drawing inspiration from Chapter 5, we propose:

$$\psi(W,\tau,\beta,\gamma) = (D_i - \exp(X_i'\beta)(1-D_i))\left(Y_i^{obs} - X_i'\gamma\right) - D_i\tau.$$

Moreover,

$$\mathbb{E}[\partial_\beta \psi(W,\tau_0,\beta_0,\gamma_0)] = -\mathbb{E}\left[X\exp(X'\beta_0)(1-D)(Y^{obs} - X'\gamma_0)\right] = 0,$$

and

$$\mathbb{E}[\partial_\gamma \psi(W,\tau_0,\beta_0,\gamma_0)] = -\mathbb{E}\left[(D-(1-D)\exp(X'\beta_0))\,X\right] = 0.$$

4. One can rely on the estimators proposed in section 5.3 and use Theorem 5.1.

15.3 Voting Model

1. A regression using honest random forests (since D is continuous), with random-split, is particularly well-suited to the shape of the basis functions in $\mathcal{F}_{p,q}$ (hypercubes). This estimator, however, can only be used in low-dimensional cases, because the convergence rate $n^{-1/(1+p\alpha_2\delta)}$ does not allow one to have $p+q \gg \log(n)$.

2. In this case, one can use a Lasso with transformed regressors:

$$\begin{aligned}\widetilde{X}_{t,i} &= 1\{X_t \in C_{a_i,\epsilon}\}\\ \widetilde{Z}_{t,i} &= 1\{Z_t \in C_{b_i,\epsilon}\}\end{aligned}$$

Thus, the linear model

$$D_t = \gamma_0^\top \widetilde{X}_t + \delta_0^\top \widetilde{Z}_t + u_t,$$

with $\widetilde{X}_t \in \mathbb{R}^p$, $\widetilde{Z}_t \in \mathbb{R}^q$, is particularly well-adapted. It can be estimated via:

$$(\widehat{\gamma}_0, \widehat{\delta}_0) \in \underset{(\gamma_0,\delta_0)\in\mathbb{R}^{p+q}}{\arg\min} \quad \frac{1}{n}\sum_{t=1}^{n}\left(D_t - \widetilde{X}_t'\gamma_0 - \widetilde{Z}_t'\delta_0\right)^2 + \frac{\lambda}{n}\left\|\widehat{Y}(\gamma_0,\delta_0)'\right\|_1,$$

where

$$\left\|\widehat{Y}(\gamma_0,\delta_0)'\right\|_1 = \sum_{j=1}^{p}\left|\widehat{Y}_j\gamma_{0,j}\right| + \sum_{j=1}^{q}\left|\widehat{Y}_{j+p}\delta_{0,j}\right|.$$

The matrices $\widehat{Y} \in \mathcal{M}_{p+q,p+q}(\mathbb{R})$ are penalty parameters.

3. As usual, using a logistic regression:

$$\tilde{S}_t := \ln\left(\frac{S_t}{1-S_t}\right) = g(X_t'\beta_0) + \tau_0 D_t + \xi_{L,t}.$$

4. We add the following linear equation:

$$\widetilde{Z}_t = \Pi \widetilde{X}_t + \zeta_t, \quad \zeta_t \perp X_t, \quad \Pi \in \mathcal{M}_{p+q,p+q}(\mathbb{R}).$$

Thus, we have:

$$\begin{aligned} D_t &= \widetilde{X}_t' \gamma_0 + \widetilde{X}_t' \Pi' \delta_0 + u_t + \zeta_t' \delta_0 \\ &= \widetilde{X}_t'(\gamma_0 + \Pi' \delta_0) + \rho_t^d, \quad \text{and hence} \\ D_t &= \widetilde{X}_t' \nu_0 + \rho_t^d, \quad \rho_t^d \perp X_t \end{aligned} \tag{15.1}$$

We denote the nuisance parameter by $\eta = (\beta_0, \nu_0, \delta_0, \gamma_0)$. Define

$$\begin{aligned} m(W_t, \eta, \tau_0) = &\left(\tilde{S}_t - \widetilde{X}_t'(\tau_0 \nu_0) - g(\widetilde{X}_t' \beta_0) - \tau_0 (D_t - \widetilde{X}_t' \nu_0) \right) \\ &\times \left(\widetilde{X}_t' \gamma_0 + \widetilde{Z}_t' \delta_0 - \widetilde{X}_t' \nu_0 \right). \end{aligned}$$

Then the two equations are satisfied, in particular:

$$\begin{aligned} \mathbb{E}\left[\partial_{\beta_0} m(W_t, \eta, \tau_0)\right] &= -\mathbb{E}\left[\zeta_t' \delta_0 \, \widetilde{X}_t g(\widetilde{X}_t' \beta_0) \right] = 0 \quad (\text{since } \zeta_t \perp X_t), \\ \mathbb{E}\left[\partial_{\nu_0} m(W_t, \eta, \tau_0)\right] &= -\mathbb{E}\left[\widetilde{X}_t \xi_t \right] = 0. \end{aligned}$$

5. In the corresponding chapter, the asymptotic normality of $\widehat{\tau}$ is proven for the affine-quadratic model, which requires that:
   (a) either g is an affine function. In this case, we return to the case seen in the corresponding chapter.
   (b) or g is a quadratic function, which is allowed by the theorem but requires using an estimator for β_0 from a nonlinear index model in the first stage.

15.4 Heterogeneity of the Gender Wage Gap

1. $\mathbb{E}[\ln W_i | X_i, F_i = 1] - \mathbb{E}[\ln W_i | X_i, F_i = 0]$ is the average wage gap between men and women for the population with characteristics X_i.

2. (a) X_i can contain many variables: hours worked, experience, experience squared, age, type of education, years of education, geographic location, nationality, marital status, number of young children, total number of children, industry, psychological traits such as conscientiousness and openness, etc.
   (b) In this case, a simple OLS estimator works thanks to the exogeneity assumption.
   (c) No, this is not the case. Given the sparse structure, one must use the double selection procedure seen in class, employing a Lasso in the first two steps — a brief description of the procedure is necessary here.

(d) $\mathbb{E}[\ln W_i|X_i, F_i = f] = \theta f + X_i'\beta$. This means the wage gap is constant over the entire support of X_i, which is probably unreasonable.

3. (a) This is true. Indeed, in this case, $\mathbb{E}[\ln W_i|X_i, F_i = 1] - \mathbb{E}[\ln W_i|X_i, F_i = 0] = \theta(Z_i) = \sum_{k=1}^{K} \theta_k Z_{i,k}$. Thus, the wage gap varies by θ_k percentage points when $Z_{i,k}$ changes by one unit. Given that the overall wage gap is negative, a positive value of θ_k means that (for example) the wage gap is lower than baseline in the population for which $Z_k = 1$.

   (b) Using the notation $\theta = (\theta_k)_{k=1,\dots,K}$, we have:

   $$\ln W_i = \alpha + F_i Z_i'\theta + X_i'\beta + \varepsilon_i,$$

   We thus have a linear model with $p + K$ covariates $(F_i Z_i', X_i')'$ and the $p + K$ normal equations are:

   $$\mathbb{E}\left[(\ln W_i - \alpha - F_i Z_i'\theta - X_i'\beta + \varepsilon_i)(F_i Z_i', X_i')'\right] = 0.$$

   If we focus only on θ, then we keep only

   $$\mathbb{E}\left[(\ln W_i - \alpha - F_i Z_i'\theta - X_i'\beta + \varepsilon_i) F_i Z_i\right] = 0,$$

   considering β as a nuisance parameter.

   (c) An orthogonal (or "immunized") moment function ψ for $(\theta_1, \dots, \theta_K)$ takes the form

   $$\mathbb{E}\left[(\ln W_i - \alpha - F_i Z_i'\theta - X_i'\beta + \varepsilon_i)(F_i Z_i - X_i'\gamma)\right] = 0,$$

   with derivatives with respect to β equal to zero. This corresponds to using the "double-selection" procedure but with K parameters of interest. This will require $K + 2$ steps:

   i. The first K steps consist of regressing each element of Z_i on X_i for the subsample of women using a Lasso,
   ii. The $(K + 1)^{\text{th}}$ step is a Lasso regression of $\ln W_i$ on X_i,
   iii. The final step is a regression of $\ln W_i$ on $F_i Z_i$ and the union of all elements of X_i previously selected.

4. Example: Compared to the baseline group, having a child aged 18 or older increases the wage gap by 5 percentage points (the wage gap is more negative for them). Thus, women with a child aged 18 or younger earn 5 percentage points less compared to men than other women do.

5. There is a multiple testing problem. (These tables already correct for multiple testing, but you could not have known this).

6. Using that $\mathbb{E}\left[\hat{\mu}(x; X_1, \ldots, X_n)\right] = \mathbb{E}\left[T(x; X_1, \ldots, X_n)\right]$. Therefore, we have:

$$\begin{aligned}
&\left|\mathbb{E}\left[\hat{\mu}(x; X_1, \ldots, X_n)\right] - \mu(x)\right| \\
&= \left|\mathbb{E}\left[T(x; X_1, \ldots, X_n)\right] - \mu(x)\right| \\
&= \left|\sum_{i \in \{i_1, \ldots, i_s\}} \mathbb{E}\left[\frac{1\{X_i \in L(x)\}}{s|L(x)|} \ln W_i\right] - \mu(x)\right| \\
&= \left|\mathbb{E}\left[\ln W_i \mid X_i \in L(x)\right] \frac{1}{s} \sum_{i \in \{i_1, \ldots, i_s\}} \mathbb{E}\left[\frac{1\{X_i \in L(x)\}}{|L(x)|} \mid X_i \in L(x)\right] - \mu(x)\right| \\
&= \left|\mathbb{E}\left[\ln W_i \mid X_i \in L(x)\right] - \mathbb{E}\left[\ln W_i \mid X_i = x\right]\right| \leq C\, \mathrm{Diam}(L(x)).
\end{aligned}$$

- We choose these two methods based on the performance indicators Λ and $\overline{\Lambda}$, which measure the degree of heterogeneity captured by the procedure. The table clearly shows that Random Forest and Elastic Net are the best. Table 2 shows that the average of $\mathbb{E}[\ln W_i|X_i, F_i = 1] - \mathbb{E}[\ln W_i|X_i, F_i = 0]$ (i.e., β_1) is negative, which means that women earn on average less than men, but the slope of the BLP is significantly positive and close to 1; therefore, there is heterogeneity and its profile is quite well described by the proxies from the Elastic Net and Random Forest.

- According to Figure 15.1 and Table 3, we see that there is a group of women for whom there is no wage gap. These are women with fewer children and less experience than average. Here, the parameter of interest depends on the accuracy of the ML proxy, as well as on the splits. We can thus only learn about the characteristics of $\mathbb{E}[\ln W_i|X_i = \cdot, F_i = 1] - \mathbb{E}[\ln W_i|X_i = \cdot, F_i = 0]$ (heterogeneity, subgroups that benefit the least and the most, and their characteristics) and not about this quantity itself or as previously done. Since this Generic ML procedure depends on the sample splitting, the p-value must be adjusted to take into account this additional randomness.

- (Bonus)

15.5 Drought and incentives to save water

1. (a) $\beta_1 = \mathbb{E}[\tau(X)] = \mathbb{E}[Y_1 - Y_0]$, the average treatment effect.

   (b) β_2 is the best linear predictive coefficient (see the course for the formula); the test H_0: "$\beta_2 = 0$" provides a heterogeneity test. When this hypothesis is rejected, we know that there is both heterogeneity and that the ML proxy helps to partially capture it. When this hypothesis is NOT rejected, the conclusion is unclear: this may be either because there is no heterogeneity or because the proxy predictor is weak (uncorrelated with the CATE).

2. (a) It can be expressed as a function of the squared correlation between the ML proxy and the true CATE multiplied by the variance of the CATE. By maximizing this quantity, we ensure selecting the algorithm most correlated with the true CATE.

   (b) Gradient Boosting Machine, because its corresponding Λ is the largest.

3. (a) Yes, the test of H_0: "$\beta_1 = 0$" is rejected at any confidence level. On average, households that received an incentive to save water consume about 952 gallons (3604 liters) less water than untreated households, all else equal.

   (b) The hypothesis that $\beta_2 = 0$ is rejected for algorithm 1. For algorithm 2, it is not rejected (for a test at the 5% confidence level, for example), which means either that there is no heterogeneity or that algorithm 2 is too weak. Since algorithm 2 is dominated by algorithm 1 in terms of Λ, more weight should be given to algorithm 1.

   Note that, as explained in the course, for a test at level α, the p-value displayed here must be less than $\alpha/2$ to be rejected, in order to account for the random splitting of the data.

4. (a) See the regression for the GATES:

$$w(X)(D - p(X))Y = \sum_{k=1}^{5} \gamma_k G_k + \varepsilon,$$

   where the most affected treatment effect is γ_1, and the treatment effect for the least affected is estimated by γ_5. Indeed, a quick calculation shows that:

$$\gamma_k = \mathbb{E}[Y_1 - Y_0 \mid G_k].$$

   This regression can be estimated by OLS. It is simply an average of $w(X)(D - p(X))Y$ within each group.

   (b) This question cannot receive a rigorous answer because we test by group, so even if the average treatment effect is significantly different from zero in the least affected group, the treatment effect (as measured by the CATE) may be zero for some individuals in this group. Moreover, groups are based on the values of the proxy predictors that are not perfectly correlated with the true CATE. Thus, even in the most affected group (as defined by the values of the proxy predictor), the true CATE may be zero for some individuals.

   However, we have accepted the conclusion that, with algorithm 1, even the least affected individuals have a nonzero treatment effect, which implies that the treatment effect is significant for most people.

   (c) Not really, the test of the difference is not rejected at any commonly accepted level (5%, 10%).

5. (a) See the regression for the CLAN in Chernozhukov et al. (2017):

$$X = \sum_{k=1}^{5} \theta_k G_k + \varepsilon,$$

where the average characteristic of the most affected is θ_1 and the average characteristic of the least affected is estimated by θ_5. It can be estimated by OLS.

(b) First, note that these coefficients estimate $P[VOTE = 1 \mid G_1 = 1]$ and $P[VOTE = 1 \mid G_5 = 1]$, but since $P[G_1] = P[G_5] = 0.2$, Bayes' theorem implies that:

$$\frac{P[VOTE = 1 \mid G_1 = 1]}{P[VOTE = 1 \mid G_5 = 1]} = \frac{P[G_1 = 1 \mid VOTE = 1]}{P[G_5 = 1 \mid VOTE = 1]},$$

so we can interpret that households where voting is more frequent are more likely to be among the most affected by the prosocial campaign compared to being among the least affected. The difference is significant.

– The same remark applies, and it follows that Democrats are more likely to be among the most affected (the difference is significant) compared to the least affected. This difference is not significant for Republican households.

Note that in all three cases, the test compares the most affected and the least affected groups, and not each group to the general population. It could thus be that Democrats are relatively more likely to be among the most affected than the least affected, but that they are underrepresented in both groups relative to the general population. [This is not the case in this application, but one cannot deduce it solely from the tables].

– For $j = 0, 1$:

$$\widehat{\alpha}_j \in \arg\max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{D_i = j\}(Y_i - X_i'\alpha)^2 + \lambda \|\alpha\|_1,$$

The CATE can then be estimated by:

$$\widehat{\tau}(X) = X'(\widehat{\alpha}_1 - \widehat{\alpha}_0).$$

The problem is that this estimator consists of two estimators computed separately and may not be optimal for estimating the CATE.

– The proposed solution for the Conditional Random Forest (CRF) is to split according to a joint criterion (not one for $D = 1$ and another for $D = 0$) specially designed to target the treatment effect, and not the conditioning of the treatment on D.

- Here, $\gamma = \beta - \mathbb{E}[D]\delta$, $\tau(X) = X'\delta$, $\widehat{\tau}(X) = X'\widehat{\delta}$, and we use

$$\mathbb{E}\left[(Y - X'\beta - (D - \mathbb{E}[D])X'\delta)\begin{pmatrix} X \\ X(D - \mathbb{E}[D]) \end{pmatrix}\right] = 0.$$

- β is a nuisance parameter. The estimator is orthogonal (immunized) because, from the equation related to δ_j,

$$\begin{aligned}\partial_\beta \mathbb{E}\left[(Y - X'\beta - (D - \mathbb{E}[D])X'\delta)X_j(D - \mathbb{E}[D])\right] &= -\mathbb{E}[XX_j(D - \mathbb{E}[D])] \\ &= -\mathbb{E}[XX_j]\mathbb{E}[(D - \mathbb{E}[D])] \\ &= 0.\end{aligned}$$

- Here, the two coefficients are estimated simultaneously, which is likely to yield a better estimation of the CATE.

- This estimator is relevant if there is sparsity in both β_0 and δ_0, i.e., only a few components of X are relevant for predicting the baseline outcome and the heterogeneity of the treatment effect. Moreover, both the CATE and the regression function for the baseline outcome are linear. The CRF is more suitable in a context where the regression function for the baseline outcome is piecewise constant.

- We replace $\mathbb{E}[D]$ by $Z'\gamma$ and add a penalty as well as a potential sparsity assumption to handle the potential high dimensionality of γ, which is a nuisance parameter.

$$\begin{aligned}(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta}) = \arg\min_{\beta,\gamma,\delta} \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i'\beta - (D_i - Z_i'\gamma)X_i'\delta)^2 \\ + \lambda_\beta\|\beta\|_1 + \lambda_\gamma\|\gamma\|_1 + \lambda_\delta\|\delta\|_1.\end{aligned}$$

  It is orthogonal (in the sense of the course definition) because, from the equation related to δ_j,

$$\begin{aligned}\partial_\beta \mathbb{E}\left[(Y - X'\beta - (D - Z'\gamma)X'\delta)X_j(D - Z'\gamma)\right] &= -\mathbb{E}[XX_j(D - Z'\gamma)] \\ &= -\mathbb{E}[XX_j]\mathbb{E}[(D - Z'\gamma)] \\ &= 0,\end{aligned}$$

  and

$$\partial_\gamma \mathbb{E}\left[(Y - X'\beta - (D - Z'\gamma)X'\delta)X_j(D - Z'\gamma)\right]$$

  is the sum of two terms: the first

$$\mathbb{E}[Z(X'\delta)X_j(D - Z'\gamma)] = -\mathbb{E}[Z(X'\delta)X_j]\mathbb{E}[\zeta] = 0,$$

  and the second

$$-\mathbb{E}\left[(Y - X'\beta - (D - Z'\gamma)X'\delta)X_jZ\right] = -\mathbb{E}[\epsilon X_jZ] = 0.$$

15.6 Synthetic Control and Regularization

1. This minimization program finds the parameters that best reproduce the behavior of the time series of unit 1 before treatment by a linear combination of the untreated units, hoping that $\hat{\mu} + \sum_{i=2}^{N+1} \hat{\omega}_i Y_{i,T+1}^{obs}$ will be a good counterfactual for $Y_{1,T+1}(0)$.

2. (a) This is a least squares program.
   We use the notation $\mathbf{Y}_t = \big(1, Y_{2,t}^{obs}, \ldots, Y_{N+1,t}^{obs}\big)$. Then the solution $(\hat{\mu}, \hat{\boldsymbol{\omega}})$ is given by:

   $$\left[\frac{1}{T}\sum_{t=1}^{\prime} \mathbf{Y}_t \mathbf{Y}_t'\right]^{-1} \left[\frac{1}{T}\sum_{t=1}^{\prime} \mathbf{Y}_t Y_{1,t}^{obs}\right],$$

   provided that $\frac{1}{T}\sum_{t=1}^{\prime} \mathbf{Y}_t \mathbf{Y}_t'$ is invertible. This condition cannot be verified, for example, if $N > T$, so we must be in a setting with long panel data.

   (b) $\hat{\omega}_i$ is the weight given to untreated unit i in reproducing the treated unit. It corresponds to the partial correlation of the time series of that particular unit with that of the treated unit. It can be negative if they are negatively correlated.

   (c) Several: it cannot always be calculated, it allows for extrapolation.

3. (a) $\hat{\omega}_1 = 1/N$.

   (b) $\hat{\mu} = (1/T)\sum_{t=1}^{\prime} Y_{1,t}^{obs} - \sum_{i=2}^{N+1} Y_{i,t}^{obs}/N$.

   (c) Consequently:

   $$\hat{\theta} = \left[Y_{1,T+1}^{obs} - \frac{1}{T}\sum_{t=1}^{\prime} Y_{1,t}^{obs}\right] - \frac{1}{N}\left[\sum_{i=2}^{N+1} Y_{i,T+1}^{obs} - \frac{1}{T}\sum_{t=1}^{\prime}\sum_{i=2}^{N+1} Y_{i,t}^{obs}\right],$$

   which is the difference-in-differences estimator.

4. For this question, we add to (OBJ) the three constraints: $\omega_i \geq 0$ for $i = 2, \ldots, N+1$, $\sum_{i=2}^{N+1} \omega_i = 1$, and $\mu = 0$.

   (a) This is the synthetic control estimator.

   (b) No, it is generally not unique.

5. (a) This helps to regularize the problem.

   (b) $\alpha = 0$ corresponds to the Ridge estimator:

   $$\left[\frac{1}{T}\sum_{t=1}^{T} \mathbf{Y}_t \mathbf{Y}_t' + \lambda I_{N+1}\right]^{-1} \left[\frac{1}{T}\sum_{t=1}^{T} \mathbf{Y}_t Y_{1,t}^{obs}\right],$$

   but the first element of I_{N+1} is zero.
   When $\alpha = 1$, it is the Lasso solution. Some weights will be large and others exactly zero.

(c) Some form of cross-validation using either the temporal dimension or the cross-sectional dimension among the untreated units.