ECMA 31380 - Causal Machine Learning - Homework 2

Fernando Rocha Urbano

Autumn 2024

1 Propensity Score Weighting & ATT Estimation

This is a continuation from homework 2.

Assume that the random variables $(Y_1, Y_0, T, X')' \in \mathbb{R} \times \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ obey $\{Y_1, Y_0\} \perp\!\!\!\perp T \mid X$. The researcher observes $(Y, T, X')'$, where $Y = Y_1 T + Y_0(1-T)$. Define the propensity score $p(x) = \mathbb{P}[T = 1 \mid X = x]$ and assume it is bounded inside $(0,1)$. Define $\mu_t = \mathbb{E}[Y(t) \mid T = 1]$ and $\mu(x) = \mathbb{E}[Y(t) \mid X = x]$. The average treatment effect on the treated (ATT) is $\tau = \mu_1 - \mu_0$.

Assume that the propensity score is correctly specified as a logistic regression: for a d-vector θ_0, it holds that $p(x) = (1 + \exp\{-\theta_0' x\})^{-1}$.

1.a Estimating θ_0 Using Maximum Likelihood

(a) Consider estimating θ_0 using maximum likelihood, denote the estimator $\hat{\theta}_{\text{MLE}}$. Write down the objective function that is solved by the estimator and the equations that characterize the solution.

The maximum likelihood estimator is:

$$\ell(\theta) = \prod p(X_i)^{y_i} \times (1 - p(X_i))^{(1-y_i)}, \quad \text{for } y_i \in \{0,1\}$$

The maximum log-likelihood estimator $\hat{\theta}_{\text{MLE}}$ is obtained by maximizing the log-likelihood function:

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \left[T_i \log p(X_i) + (1-T_i)\log(1-p(X_i))\right] \\ &= \sum_{i=1}^{n} T_i \log p(X_i) + \sum_{i=1}^{n}(1-T_i)\log(1-p(X_i))\end{aligned}$$

where $p(X_i) = \frac{1}{1+\exp\{-\theta' X_i\}}$.

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} T_i \log\left(\frac{1}{1+\exp\{-\theta' X_i\}}\right) + \sum_{i=1}^{n}(1-T_i)\log\left(1-\frac{1}{1+\exp\{-\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n} T_i \log\left(\frac{1}{1+\exp\{-\theta' X_i\}}\right) + \sum_{i=1}^{n}(1-T_i)\log\left(1-\frac{1}{1+\exp\{-\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n} T_i \log\left(\frac{1}{1+\exp\{-\theta' X_i\}}\right) + \sum_{i=1}^{n}\log\left(\frac{\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right) - \sum_{i=1}^{n} T_i \log\left(\frac{\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n} T_i \left[\log\left(\frac{1}{1+\exp\{-\theta' X_i\}}\right) - \sum_{i=1}^{n}\log\left(\frac{\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right)\right] + \sum_{i=1}^{n}\log\left(\frac{\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n} T_i \left[\log\left(\frac{1-\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right)\right] + \sum_{i=1}^{n}\log\left(\frac{\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n} T_i \log\left(\exp\{\theta' X_i\}\right) + \sum_{i=1}^{n}\log\left(\frac{\exp\{-\theta' X_i\}}{1+\exp\{-\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n} T_i\theta' X_i + \sum_{i=1}^{n}\log\left(\frac{1}{1+\exp\{\theta' X_i\}}\right) \\ &= \sum_{i=1}^{n}\left[T_i\theta' X_i + \log\left(\frac{1}{1+\exp\{\theta' X_i\}}\right)\right] \\ &= \sum_{i=1}^{n}\left[T_i\theta' X_i - \log\left(1+\exp\{\theta' X_i\}\right)\right]\end{aligned}$$

The first-order conditions that characterize the solution is:

$$\nabla_\theta \ell(\theta) = \sum_{i=1}^{n}\left[T_i - p(X_i)\right] X_i = 0.$$

Which translates that for every parameter $\theta_i \in \theta$:

$$\frac{\partial \ell(\theta)}{\partial \theta_i} = \sum_{i=1}^{n}\left[T_i - p(X_i)\right] X_i = 0.$$

The result is derived from:

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= \nabla_\theta \left(\sum_{i=1}^{n} \left[T_i \theta' X_i - \log\left(1 + \exp\left\{\theta' X_i\right\}\right) \right] \right) \\
&= \sum_{i=1}^{n} T_i X_i - \sum_{i=1}^{n} \left(\frac{1}{1 + \exp\left\{\theta' X_i\right\}} \right) \exp\left\{\theta' X_i\right\} X_i \\
&= \sum_{i=1}^{n} T_i X_i - \sum_{i=1}^{n} \left(\frac{1}{1 + \exp\left\{-\theta' X_i\right\}} \right) X_i \\
&= \sum_{i=1}^{n} \left[T_i X_i - \left(\frac{1}{1 + \exp\left\{-\theta' X_i\right\}} \right) X_i \right] \\
&= \sum_{i=1}^{n} \left[T_i X_i - p(X_i) X_i \right] \\
&= \sum_{i=1}^{n} \left[T_i - p(X_i) \right] X_i
\end{aligned}$$

1.b Influence Function for $\hat{\theta}_{\textbf{MLE}}$

(b) Derive the influence function for $\hat{\theta}_{\text{MLE}}$.

To derive the influence function for $\hat{\theta}_{\text{MLE}}$, we start with the score function (gradient of the log-likelihood with respect to θ) for a single observation (T, X):

$$s(T, X; \theta_0) = [T - p(X; \theta_0)]X,$$

where $p(X; \theta_0) = \frac{1}{1+\exp\left\{-\theta_0' X\right\}}$.

M-Estimators are estimators defined as solutions for optimization problems, often involving minimization of sum of loss functions. The $\hat{\theta}_{\text{MLE}}$ is an M-estimator.

The influence function for an M-estimator is defined as:

$$\text{IF}(z; \hat{\theta}_{\text{MLE}}, F) = J^{-1} s(z; \theta_0),$$

where J is the expected information matrix given by:

$$J = E\left[\frac{\partial s(T, X; \theta_0)}{\partial \theta'} \right] = E\left[p(X; \theta_0)[1 - p(X; \theta_0)] X X' \right].$$

Therefore, the influence function for $\hat{\theta}_{\text{MLE}}$ is:

$$\text{IF}(T, X; \hat{\theta}_{\text{MLE}}, F) = J^{-1}[T - p(X; \theta_0)]X.$$

The IF provides a linear approximation of how the estimator θ responds to small changes in data distribution. We take the derivative with respect to θ because $\hat{\theta}$ is viewed as a functional estimator, meaning that it maps from the space of the probability distribution F to the parameter space. Calculating IF answers how much $\hat{\theta}(F)$ changes as a distribution of F is perturbed.

1.c Estimating θ_0 Using Nonlinear Least Squares

(c) Consider estimating θ_0 using nonlinear least squares, denote the estimator $\hat{\theta}_{\text{NLS}}$. Write down the objective function that is solved by the estimator and the equations that characterize the solution.

The nonlinear least squares estimator $\hat{\theta}_{\text{NLS}}$ minimizes the sum of squared differences between the observed treatment indicator and the predicted propensity score. The objective function is:

$$\hat{\theta}_{\text{NLS}} = \arg\min_{\theta} \sum_{i=1}^{n} \left[T_i - p(X_i; \theta)\right]^2,$$

where the propensity score $p(X_i; \theta)$ is given by:

$$p(X_i; \theta) = \frac{1}{1 + \exp\{-\theta' X_i\}}.$$

The equations that characterize the solution are obtained by taking the gradient of the objective function with respect to θ and setting it to zero:

$$\nabla_\theta \sum_{i=1}^{n} \left[T_i - p(X_i; \theta)\right]^2 = -2 \sum_{i=1}^{n} \left[T_i - p(X_i; \theta)\right] p(X_i; \theta)[1 - p(X_i; \theta)]X_i = 0.$$

1.d Influence Function for $\hat{\theta}_{\text{NLS}}$

(d) Derive the influence function for $\hat{\theta}_{\text{NLS}}$. Compare it to the one for $\hat{\theta}_{\text{MLE}}$.

To derive the influence function for $\hat{\theta}_{\text{NLS}}$, we begin by expressing the estimator as an M-estimator. The nonlinear least squares estimator minimizes the objective function:

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[T_i - p(X_i; \theta)\right]^2,$$

where $p(X_i;\theta) = \frac{1}{1+\exp\{-\theta' X_i\}}$.

The first-order condition (gradient) of this objective function with respect to θ is:

$$\Psi_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = -\frac{2}{n}\sum_{i=1}^{n} [T_i - p(X_i;\theta)]\, p(X_i;\theta)[1-p(X_i;\theta)]X_i = 0.$$

At the population level, the expectation of the gradient function is:

$$\Psi(\theta) = E\left[-2\,[T - p(X;\theta)]\, p(X;\theta)[1-p(X;\theta)]X\right] = 0.$$

The influence function for an M-estimator is given by:

$$\text{IF}(Z;\hat{\theta}_{\text{NLS}}, F) = -A^{-1}\psi(Z;\theta_0),$$

where:

- $Z = (T, X)$ is an observation from the population,
- $\psi(Z;\theta) = -2\,[T - p(X;\theta)]\, p(X;\theta)[1-p(X;\theta)]X$ is the influence function's numerator,
- $A = E\left[\frac{\partial \psi(Z;\theta_0)}{\partial \theta'}\right]$ is the expected derivative matrix evaluated at the true parameter θ_0.

First, compute the derivative matrix A:

$$\begin{aligned} A &= E\left[\frac{\partial \psi(Z;\theta_0)}{\partial \theta'}\right] \\ &= E\left[-2\left\{[T - p(X;\theta_0)] \cdot \frac{\partial}{\partial \theta'}\left(p(X;\theta_0)[1-p(X;\theta_0)]X\right) - p(X;\theta_0)[1-p(X;\theta_0)]XX'\right\}\right] \end{aligned}$$

Since $E[T \mid X] = p(X;\theta_0)$, the term involving $[T - p(X;\theta_0)]$ vanishes in expectation. Therefore, A simplifies to:

$$A = 2E\left[p(X;\theta_0)[1-p(X;\theta_0)]\left(p(X;\theta_0)[1-p(X;\theta_0)]XX'\right)\right].$$

Simplifying further:

$$A = 2E\left[p(X;\theta_0)^2[1-p(X;\theta_0)]^2 XX'\right].$$

Now, the influence function becomes:

$$\text{IF}(Z;\hat{\theta}_{\text{NLS}}, F) = -A^{-1}\psi(Z;\theta_0) = 2A^{-1}[T - p(X;\theta_0)]p(X;\theta_0)[1-p(X;\theta_0)]X.$$

Comparing this to the influence function for the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$:

$$\text{IF}(Z;\hat{\theta}_{\text{MLE}},F)=J^{-1}[T-p(X;\theta_0)]X,$$

where $J=E\left[p(X;\theta_0)[1-p(X;\theta_0)]XX'\right]$.

The key differences between the two influence functions are:

- For $\hat{\theta}_{\text{NLS}}$, the influence function includes an additional factor of $2p(X;\theta_0)[1-p(X;\theta_0)]$ in both the numerator and the inverse of A. In contrast, $\hat{\theta}_{\text{MLE}}$ involves the Fisher information matrix J without these extra terms.
- The NLS influence function gives more weight to observations where $p(X;\theta_0)[1-p(X;\theta_0)]$ is large, emphasizing data points with propensity scores near 0.5. The MLE influence function weights observations uniformly in terms of $[T-p(X;\theta_0)]X$.
- The MLE is asymptotically efficient under correct model specification, whereas the NLS estimator may be less efficient due to the additional weighting.

The NLS estimator's influence function includes extra weighting factors derived from the logistic function's properties. This leads to differences in the estimators' asymptotic variances and efficiency.

Now we turn to ATT estimation and inference. Combining the moment conditions (see homework 2), the ATT obeys

$$\tau=\mu_1-\mu_0=\mathbb{E}[Y(1)\mid T=1]-\mathbb{E}[Y(0)\mid T=1]=\mathbb{E}\left[\frac{TY}{\mathbb{E}[T]}\right]-\frac{1}{\mathbb{E}[T]}\mathbb{E}\left[\frac{(1-T)p(X)Y}{(1-p(X))}\right].$$

For an estimator $\hat{p}(x)$ of the propensity score, we will estimate the ATT using the sample analogue of the above moment condition. Let $\hat{p}=\sum_{i=1}^{n}t_i/n$ and define the estimator

$$\hat{\tau}=\hat{\mu}_1-\hat{\mu}_0=\frac{1}{n}\sum_{i=1}^{n}\frac{t_iy_i}{\hat{p}}-\frac{1}{1-\hat{p}}\frac{1}{n}\sum_{i=1}^{n}\frac{(1-t_i)\hat{p}(x_i)y_i}{(1-\hat{p}(x_i))}.$$

1.e Influence Function for Estimator Using Maximum Likelihood

(e) Derive the influence function of your estimator assuming that you use maximum likelihood to estimate the propensity score.

To derive the influence function of the estimator $\hat{\tau}$ when the propensity score is estimated using maximum likelihood, we need to consider both the variability from the sample data and the estimation error from $\hat{\theta}_{\text{MLE}}$. The estimator $\hat{\tau}$ is given by:

$$\hat{\tau}=\hat{\mu}_1-\hat{\mu}_0,$$

where:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{p}},$$

$$\hat{\mu}_0 = \frac{1}{1-\hat{p}} \cdot \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)\hat{p}(X_i)Y_i}{1-\hat{p}(X_i)}.$$

Here, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} T_i$ is the sample proportion of treated units, and $\hat{p}(X_i) = p(X_i; \hat{\theta}_{\text{MLE}})$ is the estimated propensity score using maximum likelihood.

The influence function $\text{IF}(Z; \hat{\tau}, F)$ for $\hat{\tau}$ can be expressed as:

$$\text{IF}(Z_i; \hat{\tau}, F) = \phi_{\hat{\tau}}(Z_i) = \phi_{\mu_1}(Z_i) - \phi_{\mu_0}(Z_i),$$

where $\phi_{\mu_1}(Z_i)$ and $\phi_{\mu_0}(Z_i)$ are the influence functions for $\hat{\mu}_1$ and $\hat{\mu}_0$, respectively.

- *Influence Function for $\hat{\mu}_1$:*

  Since $\hat{\mu}_1$ is the sample average of Y_i among treated units, its influence function is:

  $$\phi_{\mu_1}(Z_i) = \frac{T_i}{p}[Y_i - \mu_1],$$

  where $p = \mathbb{E}[T]$.

- *Influence Function for $\hat{\mu}_0$:*

  The estimator $\hat{\mu}_0$ depends on the estimated propensity score $\hat{p}(X_i)$. Its influence function involves two components:

  – The influence from the sample data.

  – The influence from the estimation of $\hat{\theta}_{\text{MLE}}$.

  We can write $\phi_{\mu_0}(Z_i)$ as:

  $$\phi_{\mu_0}(Z_i) = \phi^{(1)}_{\mu_0}(Z_i) + \phi^{(2)}_{\mu_0}(Z_i),$$

  where:

  – $\phi^{(1)}_{\mu_0}(Z_i)$ accounts for the variability in Y_i and T_i.

  – $\phi^{(2)}_{\mu_0}(Z_i)$ accounts for the estimation error in $\hat{\theta}_{\text{MLE}}$.

- *First Component $\phi^{(1)}_{\mu_0}(Z_i)$:*

  $$\phi^{(1)}_{\mu_0}(Z_i) = \frac{(1-T_i)p(X_i)}{(1-p)(1-p(X_i))}[Y_i - \mu_0].$$

- *Second Component $\phi_{\mu_0}^{(2)}(Z_i)$:*

  We need to compute the derivative of μ_0 with respect to θ:

$$\frac{\partial \mu_0}{\partial \theta'} = \frac{1}{1-p} \cdot \mathbb{E}\left[(1-T) \cdot \frac{\partial}{\partial \theta'}\left(\frac{p(X)}{1-p(X)}\right) Y\right].$$

  Since:

$$\frac{\partial}{\partial \theta'}\left(\frac{p(X)}{1-p(X)}\right) = \frac{p(X)}{1-p(X)} X,$$

  we have:

$$\frac{\partial \mu_0}{\partial \theta'} = \frac{1}{1-p} \cdot \mathbb{E}\left[(1-T) \cdot \frac{p(X)}{1-p(X)} XY\right].$$

  The influence function of $\hat{\theta}_{\text{MLE}}$ is:

$$\text{IF}(Z_i; \hat{\theta}_{\text{MLE}}, F) = J^{-1}[T_i - p(X_i)]X_i,$$

  where J is the expected Fisher information matrix:

$$J = \mathbb{E}\left[p(X)[1-p(X)]XX'\right].$$

  Therefore, the second component of $\phi_{\mu_0}(Z_i)$ is:

$$\phi_{\mu_0}^{(2)}(Z_i) = \left(\frac{\partial \mu_0}{\partial \theta'}\right) \text{IF}(Z_i; \hat{\theta}_{\text{MLE}}, F) = \left(\frac{\partial \mu_0}{\partial \theta'}\right) J^{-1}[T_i - p(X_i)]X_i.$$

Combining the components, the influence function for $\hat{\mu}_0$ is:

$$\phi_{\mu_0}(Z_i) = \frac{(1-T_i)p(X_i)}{(1-p)(1-p(X_i))}[Y_i - \mu_0] + \left(\frac{\partial \mu_0}{\partial \theta'}\right) J^{-1}[T_i - p(X_i)]X_i.$$

The final influence function for $\hat{\tau}$ is then:

$$\phi_{\hat{\tau}}(Z_i) = \frac{T_i}{p}[Y_i - \mu_1] - \frac{(1-T_i)p(X_i)}{(1-p)(1-p(X_i))}[Y_i - \mu_0] - \left(\frac{\partial \mu_0}{\partial \theta'}\right) J^{-1}[T_i - p(X_i)]X_i.$$

- The first term represents the variability in $\hat{\mu}_1$ due to sampling.
- The second term captures the variability in $\hat{\mu}_0$ from the sample data.
- The third term adjusts for the estimation error in $\hat{\theta}_{\text{MLE}}$ when estimating $\hat{\mu}_0$.

In conclusion, the influence function of the estimator $\hat{\tau}$ when using maximum likelihood to estimate the propensity score is given by:

$$\mathrm{IF}(Z_i;\hat{\tau},F) = \frac{T_i}{p}[Y_i - \mu_1] - \frac{(1-T_i)p(X_i)}{(1-p)(1-p(X_i))}[Y_i - \mu_0] - \left(\frac{\partial \mu_0}{\partial \theta'}\right) J^{-1}[T_i - p(X_i)]X_i.$$

This expression accounts for both the sampling variability and the additional uncertainty introduced by estimating the propensity score via maximum likelihood.

1.f Influence Function for Estimator Using Nonlinear Least Squares

(f) Derive the influence function of your estimator assuming that you use nonlinear least squares to estimate the propensity score.

To derive the influence function of the estimator $\hat{\tau}$ when the propensity score is estimated using nonlinear least squares (NLS), we start with the estimator:

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0,$$

where:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{p}},$$

$$\hat{\mu}_0 = \frac{1}{1-\hat{p}} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)\hat{p}(X_i)Y_i}{1-\hat{p}(X_i)}.$$

Here, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} T_i$ is the sample proportion of treated units, and $\hat{p}(X_i) = p(X_i;\hat{\theta}_{\mathrm{NLS}})$ is the estimated propensity score using NLS.

The influence function $\mathrm{IF}(Z_i;\hat{\tau},F)$ for $\hat{\tau}$ can be expressed as:

$$\mathrm{IF}(Z_i;\hat{\tau},F) = \phi_{\mu_1}(Z_i) - \phi_{\mu_0}(Z_i),$$

where $\phi_{\mu_1}(Z_i)$ and $\phi_{\mu_0}(Z_i)$ are the influence functions for $\hat{\mu}_1$ and $\hat{\mu}_0$, respectively.

- *Influence Function for $\hat{\mu}_1$:*

  Since $\hat{\mu}_1$ is the sample average of Y_i among treated units, its influence function is:

  $$\phi_{\mu_1}(Z_i) = \frac{T_i}{p}[Y_i - \mu_1],$$

  where $p = \mathbb{E}[T]$.

- *Influence Function for $\hat{\mu}_0$:*

  The estimator $\hat{\mu}_0$ depends on the estimated propensity score $\hat{p}(X_i)$. Its influence function involves two components:

  - The influence from the sample data.
  - The influence from the estimation of $\hat{\theta}_{\text{NLS}}$.

  We can write $\phi_{\mu_0}(Z_i)$ as:

  $$\phi_{\mu_0}(Z_i) = \phi_{\mu_0}^{(1)}(Z_i) + \phi_{\mu_0}^{(2)}(Z_i),$$

  where:

  - $\phi_{\mu_0}^{(1)}(Z_i)$ accounts for the variability in Y_i and T_i:

    $$\phi_{\mu_0}^{(1)}(Z_i) = \frac{(1-T_i)p(X_i)}{(1-p)(1-p(X_i))}[Y_i - \mu_0].$$

  - $\phi_{\mu_0}^{(2)}(Z_i)$ accounts for the estimation error in $\hat{\theta}_{\text{NLS}}$.

- *Derivative of μ_0 with Respect to θ:*

  We compute the derivative:

  $$\frac{\partial \mu_0}{\partial \theta'} = \frac{1}{1-p} \cdot \mathbb{E}\left[(1-T) \cdot \frac{\partial}{\partial \theta'}\left(\frac{p(X)}{1-p(X)}\right) Y\right].$$

  Since:

  $$\frac{\partial}{\partial \theta'}\left(\frac{p(X)}{1-p(X)}\right) = \frac{p(X)}{[1-p(X)]}X,$$

  it follows that:

  $$\frac{\partial \mu_0}{\partial \theta'} = \frac{1}{1-p} \cdot \mathbb{E}\left[(1-T) \cdot \frac{p(X)}{[1-p(X)]}XY\right].$$

- *Influence Function of $\hat{\theta}_{NLS}$:*

  The influence function for the NLS estimator $\hat{\theta}_{\text{NLS}}$ is:

  $$\text{IF}(Z_i; \hat{\theta}_{\text{NLS}}, F) = A^{-1}\psi(Z_i; \theta_0),$$

  where:

  $$\psi(Z_i; \theta_0) = -2[T_i - p(X_i)]p(X_i)[1-p(X_i)]X_i,$$

  and

  $$A = 2\mathbb{E}\left[p(X)^2[1-p(X)]^2 XX'\right].$$

- *Second Component $\phi^{(2)}_{\mu_0}(Z_i)$:*

  Using the influence function of $\hat{\theta}_{\text{NLS}}$, we have:

$$\phi^{(2)}_{\mu_0}(Z_i) = \left(\frac{\partial \mu_0}{\partial \theta'}\right) \text{IF}(Z_i; \hat{\theta}_{\text{NLS}}, F) = \left(\frac{\partial \mu_0}{\partial \theta'}\right) A^{-1}\psi(Z_i; \theta_0).$$

  Substituting $\psi(Z_i; \theta_0)$:

$$\phi^{(2)}_{\mu_0}(Z_i) = -2\left(\frac{\partial \mu_0}{\partial \theta'}\right) A^{-1}[T_i - p(X_i)]p(X_i)[1 - p(X_i)]X_i.$$

Combining the Components, the influence function for $\hat{\mu}_0$ is:

$$\phi_{\mu_0}(Z_i) = \frac{(1 - T_i)p(X_i)}{(1 - p)(1 - p(X_i))}[Y_i - \mu_0] - 2\left(\frac{\partial \mu_0}{\partial \theta'}\right) A^{-1}[T_i - p(X_i)]p(X_i)[1 - p(X_i)]X_i.$$

Final Influence Function for $\hat{\tau}$:

Subtracting $\phi_{\mu_0}(Z_i)$ from $\phi_{\mu_1}(Z_i)$, we obtain:

$$\text{IF}(Z_i; \hat{\tau}, F) = \phi_{\mu_1}(Z_i) - \phi_{\mu_0}(Z_i).$$

Substituting the expressions:

$$\text{IF}(Z_i; \hat{\tau}, F) = \frac{T_i}{p}[Y_i - \mu_1] - \frac{(1 - T_i)p(X_i)}{(1 - p)(1 - p(X_i))}[Y_i - \mu_0] + 2\left(\frac{\partial \mu_0}{\partial \theta'}\right) A^{-1}[T_i - p(X_i)]p(X_i)[1 - p(X_i)]X_i.$$

Interpretation:

- The first term represents the variability in $\hat{\mu}_1$ due to sampling:

$$\frac{T_i}{p}[Y_i - \mu_1].$$

- The second term captures the variability in $\hat{\mu}_0$ from the sample data:

$$-\frac{(1 - T_i)p(X_i)}{(1 - p)(1 - p(X_i))}[Y_i - \mu_0].$$

- The third term adjusts for the estimation error in $\hat{\theta}_{\text{NLS}}$:

$$+2\left(\frac{\partial \mu_0}{\partial \theta'}\right) A^{-1}[T_i - p(X_i)]p(X_i)[1 - p(X_i)]X_i.$$

Conclusion:

The influence function of the estimator $\hat{\tau}$ when using nonlinear least squares to estimate the propensity score is given by:

$$\mathrm{IF}(Z_i;\hat{\tau},F)=\frac{T_i}{p}[Y_i-\mu_1]-\frac{(1-T_i)p(X_i)}{(1-p)(1-p(X_i))}[Y_i-\mu_0]+2\left(\frac{\partial\mu_0}{\partial\theta'}\right)A^{-1}[T_i-p(X_i)]p(X_i)[1-p(X_i)]X_i.$$

This expression accounts for both the sampling variability and the additional uncertainty introduced by estimating the propensity score via nonlinear least squares.

Comparison with Maximum Likelihood Estimation:

Compared to the influence function when using maximum likelihood (MLE) estimation, the key differences are:

- The scaling factor in the third term is different due to the influence function of $\hat{\theta}_{\text{NLS}}$ involving an extra factor of $2p(X_i)[1-p(X_i)]$.
- The matrix A in NLS is:

  $$A=2\mathbb{E}\left[p(X)^2[1-p(X)]^2XX'\right],$$

  whereas in MLE, the Fisher information matrix is:

  $$J=\mathbb{E}\left[p(X)[1-p(X)]XX'\right].$$

- The presence of $2p(X_i)[1-p(X_i)]$ in the NLS influence function emphasizes observations with propensity scores near 0.5 more than in the MLE case.

In summary, the influence function for $\hat{\tau}$ when using NLS differs from that using MLE due to the different weighting and scaling factors inherent in the NLS estimation method.

1.g Simulation Study

(g) Conduct a simulation study where you use both first step estimation methods. Your study should verify the derivations above as well as compare the two estimators. Which performs better? Explore different sample sizes, dimensions of X, noise levels, etc., i.e., vary different aspects of the simulation design.

We conducted a simulation study to compare the performance of the average treatment effect on the treated (ATT) estimators using maximum likelihood estimation (MLE) and nonlinear least squares (NLS) for estimating the propensity score. The study varied sample sizes, dimensions of X, and noise levels. The results are summarized in the table provided.

From the simulation results, we observe the following:

Table 1: Simulation Results Comparing MLE and NLS Estimators for ATT

Dim	SS	NL	Bias MLE	Bias NLS	Var MLE	Var NLS	MSE MLE	MSE NLS
2	500	1	-0.00932	-0.01588	0.01961	0.02185	0.01950	0.02188
2	500	2	-0.02962	-0.02358	0.06793	0.06445	0.06813	0.06437
2	500	5	0.00070	0.00400	0.20787	0.19905	0.20579	0.19708
2	1000	1	-0.00034	0.00143	0.01175	0.01284	0.01163	0.01272
2	1000	2	-0.00518	-0.00591	0.02166	0.02158	0.02147	0.02140
2	1000	5	-0.01149	-0.00750	0.14905	0.14579	0.14769	0.14439
2	5000	1	0.00526	0.00336	0.00152	0.00180	0.00154	0.00180
2	5000	2	-0.00588	-0.00439	0.00504	0.00499	0.00503	0.00495
2	5000	5	-0.00009	0.00278	0.03048	0.03032	0.03017	0.03002
5	500	1	0.00160	-0.03872	0.08885	0.14498	0.08797	0.14503
5	500	2	-0.06337	-0.11231	0.26856	0.41486	0.26989	0.42332
5	500	5	-0.00751	NaN	0.37166	NA	0.36800	NaN
5	1000	1	-0.00464	-0.03373	0.04965	0.07184	0.04918	0.07226
5	1000	2	0.06409	0.04796	0.05941	0.06751	0.06292	0.06913
5	1000	5	-0.02145	-0.05645	0.24894	0.31807	0.24691	0.31808
5	5000	1	0.00311	-0.00395	0.01185	0.01460	0.01174	0.01447
5	5000	2	-0.03152	-0.03556	0.01799	0.02229	0.01881	0.02333
5	5000	5	0.01152	0.01018	0.04560	0.04740	0.04528	0.04703
10	500	1	0.06400	NaN	0.35931	NA	0.35981	NaN
10	500	2	0.01739	NaN	0.44957	NA	0.44538	NaN
10	500	5	0.05529	NaN	1.15303	NA	1.14456	NaN
10	1000	1	0.06698	NaN	0.42974	NA	0.42993	NaN
10	1000	2	-0.08432	NaN	0.49769	NA	0.49982	NaN
10	1000	5	0.14828	NaN	0.39085	NA	0.40893	NaN
10	5000	1	-0.00669	NaN	0.07871	NA	0.07796	NaN
10	5000	2	0.02409	NaN	0.05547	NA	0.05550	NaN
10	5000	5	-0.03186	-183907.8	0.15921	3.3e+12	0.15863	3.3e+12

- *Bias*:
  - For lower dimensions ($d = 2$), both MLE and NLS estimators exhibit small biases across different sample sizes and noise levels.
  - As the dimension increases to $d = 5$, the bias of the NLS estimator increases significantly, especially at smaller sample sizes and higher noise levels.
  - In the highest dimension ($d = 10$), the NLS estimator often fails to produce valid results (indicated by `NaN` values), suggesting convergence issues in the NLS estimation method. The MLE estimator, however, maintains reasonable bias levels.
- *Variance*:
  - The variance of both estimators decreases with increasing sample size, as expected.
  - The MLE estimator consistently shows lower variance compared to the NLS estimator across most settings.
  - In higher dimensions and noise levels, the variance of the NLS estimator becomes substantially larger, indicating less reliable estimates.
- *Mean Squared Error (MSE)*:
  - The MSE of the MLE estimator is generally lower than that of the NLS estimator, indicating better overall performance.
  - In cases where the NLS estimator fails (evidenced by `NaN` or extremely large values), the MSE is significantly higher, reinforcing the instability of the NLS method in those settings.
- *Estimator Performance*:
  - The MLE-based ATT estimator performs better than the NLS-based estimator in terms of bias, variance, and MSE, especially as the dimension of X increases.
  - The NLS estimator encounters convergence issues in higher dimensions and with higher noise levels, leading to unreliable estimates.
  - The MLE estimator remains robust across different simulation settings, providing consistent and accurate estimates of the ATT.
- *Effect of Sample Size*:
  - Increasing the sample size generally improves the performance of both estimators by reducing variance and MSE.
  - The benefit of larger sample sizes is more pronounced for the MLE estimator, which continues to provide accurate estimates even in challenging settings.
- *Effect of Noise Level*:
  - Higher noise levels increase the variance and MSE of both estimators.
  - The NLS estimator is more adversely affected by higher noise levels compared to the MLE estimator.

Based on the simulation study, the ATT estimator using maximum likelihood estimation for the propensity score outperforms the estimator using nonlinear least squares. The MLE method demonstrates better accuracy (lower bias), precision (lower variance), and overall reliability (lower MSE) across various dimensions, sample sizes, and noise levels. The NLS estimator struggles in higher-dimensional settings and with higher noise, often failing to converge or producing invalid results.

These findings align with the theoretical derivations of the influence functions. The MLE estimator is asymptotically efficient under correct model specification, as it directly maximizes the likelihood function. In contrast, the NLS estimator introduces additional weighting factors that can lead to inefficiencies and convergence issues, especially in complex settings.

```r
# Function to simulate data
simulate_data <- function(n, d, theta_0, beta0, tau_true, noise_level) {
  # Generate X ~ N(0, I_d)
  X <- matrix(rnorm(n * d), nrow = n, ncol = d)

  # Compute propensity scores
  p <- 1 / (1 + exp(-X %*% theta_0))

  # Generate treatment assignment T ~ Bernoulli(p)
  T <- rbinom(n, 1, p)

  # Generate potential outcomes
  Y0 <- X %*% beta0 + rnorm(n, mean = 0, sd = noise_level)
  Y1 <- Y0 + tau_true

  # Observed outcome
  Y <- T * Y1 + (1 - T) * Y0

  data <- data.frame(Y = Y, T = T, X)
  colnames(data)[-(1:2)] <- paste0("X", 1:d)
  return(data)
}

# Function to estimate propensity score via MLE (logistic regression)
estimate_propensity_mle <- function(data, d) {
  formula <- as.formula(paste("T ~", paste(paste0("X", 1:d), collapse = " + ")))
  model <- glm(formula, data = data, family = binomial(link = "logit"))
  data$propensity_mle <- predict(model, type = "response")
  return(data)
}

# Function to estimate propensity score via NLS
estimate_propensity_nls <- function(data, d) {
  # Define the logistic function
  logistic_function <- function(theta, X) {
    1 / (1 + exp(-X %*% theta))
  }

  # Objective function for NLS
  nls_objective <- function(theta, T, X) {
    p_hat <- logistic_function(theta, X)
    sum((T - p_hat)^2)
  }

  # Initial guess for theta
  theta_init <- rep(0, d)

  # Optimize theta using nonlinear least squares
  nls_result <- optim(
    theta_init,
    nls_objective,
    T = data$T,
    X = as.matrix(data[, paste0("X", 1:d)]),
    method = "BFGS"
  )

  # Estimated propensity scores
  data$propensity_nls <- logistic_function(nls_result$par, as.matrix(data[, paste0("X", 1:d)]))

  return(data)
}

# Function to compute ATT estimator
compute_att <- function(data, method = c("mle", "nls")) {
  method <- match.arg(method)
  if (method == "mle") {
    p_hat <- mean(data$T)
    data$propensity <- data$propensity_mle
  } else if (method == "nls") {
    p_hat <- mean(data$T)
    data$propensity <- data$propensity_nls
  }

  # Compute ATT estimator
  mu1_hat <- mean(data$T * data$Y) / p_hat
  mu0_hat <- (1 / (1 - p_hat)) * mean(((1 - data$T) * data$propensity * data$Y) / (1 - data$propensity))
  att_hat <- mu1_hat - mu0_hat
  return(att_hat)
}

# Simulation parameters
n_sim <- 100
sample_sizes <- c(500, 1000, 5000)
dimensions <- c(2, 5, 10)
noise_levels <- c(1, 2, 5)

# True parameter values
tau_true <- 2    # True ATT

# Store results
results <- data.frame()

for (d in dimensions) {
  theta_0 <- rep(0.5, d)  # True theta_0
  beta0 <- rep(1, d)      # Coefficients for Y0

  for (n in sample_sizes) {
    for (noise_level in noise_levels) {
      att_mle_estimates <- numeric(n_sim)
      att_nls_estimates <- numeric(n_sim)

      for (sim in 1:n_sim) {
        # Simulate data
        data <- simulate_data(n, d, theta_0, beta0, tau_true, noise_level)

        # Estimate propensity scores
        data <- estimate_propensity_mle(data, d)
        data <- estimate_propensity_nls(data, d)

        # Compute ATT estimators
        att_mle <- compute_att(data, method = "mle")
        att_nls <- compute_att(data, method = "nls")

        # Store estimates
        att_mle_estimates[sim] <- att_mle
        att_nls_estimates[sim] <- att_nls
      }

      # Compute biases and variances
      bias_mle <- mean(att_mle_estimates - tau_true)
      bias_nls <- mean(att_nls_estimates - tau_true)
      var_mle <- var(att_mle_estimates)
      var_nls <- var(att_nls_estimates)
      mse_mle <- mean((att_mle_estimates - tau_true)^2)
      mse_nls <- mean((att_nls_estimates - tau_true)^2)

      # Store results
      results <- rbind(results, data.frame(
        Dimension = d,
        SampleSize = n,
        NoiseLevel = noise_level,
        Bias_MLE = bias_mle,
        Bias_NLS = bias_nls,
        Variance_MLE = var_mle,
        Variance_NLS = var_nls,
        MSE_MLE = mse_mle,
        MSE_NLS = mse_nls
      ))

      # Print progress
      cat("Completed: Dimension =", d, "Sample Size =", n, "Noise Level =", noise_level, "\n")
    }
  }
}

# Display results
print(results)
```

Listing 1: Simple Regression on Price Experiment

Recommendations:

- For practical applications, especially when dealing with higher-dimensional covariates or noisy data, the MLE method for estimating the propensity score is preferred.
- The NLS method may be acceptable in low-dimensional, low-noise settings but should be used with caution due to potential convergence problems.
- Further investigation into regularization techniques or alternative estimation methods may be warranted to improve the performance of the NLS estimator in challenging scenarios.

2 Nonparametric Density Estimation

Density estimation isn't as useful as nonparametric regression, in general and for causal inference in particular, but all the conceptual lessons learned here carry over to regression.

We have an i.i.d. sample $\{x_1, \ldots, x_n\}$ from a scalar random variable $X \in \mathbb{R}$, where X has the cdf $F(x)$ and the (Lebesgue) density $f(x)$. Assume X has compact, connected support and that $f(x)$ is bounded and bounded away from zero. Our goal in this problem is to learn $F(x)$ and $f(x)$ at a single point x.

2.a Empirical Distribution Function

(a) Consider the empirical distribution function

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\}.$$

Motive this estimator as the sample analogue of the population cdf. Prove that $\hat{F}(x)$ is unbiased and compute its variance. Establish that the estimator is consistent.

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\}$$

serves as the sample analogue of the population cumulative distribution function $F(x) = P(X \leq x)$, because it represents the proportion of observed data points less than or equal to x.

To show that $\hat{F}(x)$ is unbiased, we compute its expected value:

$$\begin{aligned} E[\hat{F}(x)] &= E\left[\frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[1\{x_i \leq x\}] \\ &= E[1\{X \leq x\}] \\ &= P(X \leq x) \\ &= F(x). \end{aligned}$$

Thus, $\hat{F}(x)$ is an unbiased estimator of $F(x)$.

Next, we compute the variance of $\hat{F}(x)$:

$$\begin{aligned}
\mathrm{Var}(\hat{F}(x)) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(1\{x_i \leq x\}) \quad \text{(since the indicators are independent)} \\
&= \frac{1}{n^2} \cdot n \cdot \mathrm{Var}(1\{X \leq x\}) \\
&= \frac{1}{n}\left[F(x)(1 - F(x))\right].
\end{aligned}$$

Therefore, the variance of $\hat{F}(x)$ decreases at a rate of $1/n$.

To establish consistency, observe that as $n \to \infty$:

$$\mathrm{Var}(\hat{F}(x)) \to 0.$$

Since $\hat{F}(x)$ is unbiased, it converges in mean square to $F(x)$. By the Weak Law of Large Numbers, $\hat{F}(x)$ also converges in probability to $F(x)$. Therefore, $\hat{F}(x)$ is a consistent estimator of $F(x)$.

2.b Asymptotic Normality

(b) Prove, including providing sufficient conditions, that $\sqrt{n}(\hat{F}(x) - F(x)) \to_d \mathcal{N}(0, \Omega)$. Characterize the variance Ω and provide a consistent estimator.

To prove that

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, \Omega),$$

we consider the indicator variables

$$Y_i = 1\{x_i \leq x\}, \quad i = 1, 2, \ldots, n.$$

Each Y_i is an independent and identically distributed (i.i.d.) Bernoulli random variable with success probability $p = F(x)$:

$$E[Y_i] = F(x), \quad \mathrm{Var}(Y_i) = F(x)(1 - F(x)).$$

By the Central Limit Theorem (CLT), if the following sufficient conditions are met:

- The Y_i are i.i.d. random variables.
- The variance $\mathrm{Var}(Y_i)$ is finite.

then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - E[Y_i]\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(Y_i)).$$

Substituting back, we have

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} \mathcal{N}\left(0, F(x)(1 - F(x))\right).$$

Thus, the asymptotic variance is

$$\Omega = F(x)(1 - F(x)).$$

A consistent estimator for Ω is obtained by replacing $F(x)$ with $\hat{F}(x)$:

$$\hat{\Omega} = \hat{F}(x)\left(1 - \hat{F}(x)\right).$$

Since $\hat{F}(x)$ is a consistent estimator of $F(x)$, $\hat{\Omega}$ is a consistent estimator of Ω.

2.c Normal Distribution Assumption

(c) Suppose that you know that $X \sim \mathcal{N}(\mu, \sigma^2)$. Use the sample mean and variance to provide an estimator of the cdf, call it $\tilde{F}(x)$. Prove that this estimator is consistent and asymptotically Normal.

Given that $X \sim \mathcal{N}(\mu, \sigma^2)$, we can estimate the cumulative distribution function at point x using the sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$:

$$\tilde{F}(x) = \Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}.$$

Should it be $1/(n-1)$ in the standard deviation instead of n?

Since $\hat{\mu}$ and $\hat{\sigma}$ are consistent estimators of μ and σ respectively, we have

$$\hat{\mu} \xrightarrow{p} \mu, \quad \hat{\sigma} \xrightarrow{p} \sigma \quad \text{as } n \to \infty.$$

The function $\Phi\left(\frac{x-\mu}{\sigma}\right)$ is continuous in both μ and σ. By the Continuous Mapping Theorem,

$$\tilde{F}(x) = \Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right) \xrightarrow{p} \Phi\left(\frac{x - \mu}{\sigma}\right) = F(x).$$

Therefore, $\tilde{F}(x)$ is a consistent estimator of $F(x)$.

To establish the asymptotic normality, we use the Delta Method. Let $\theta = (\mu, \sigma)$ and $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$. Define the function

$$h(\theta) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

The first-order Taylor expansion of $\tilde{F}(x)$ around θ is

$$\sqrt{n}\left(\tilde{F}(x) - F(x)\right) \approx \nabla h(\theta)^T \sqrt{n}(\hat{\theta} - \theta),$$

where the gradient $\nabla h(\theta)$ is given by

$$\nabla h(\theta) = \begin{pmatrix} \frac{\partial h}{\partial \mu} \\ \frac{\partial h}{\partial \sigma} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \\ -\frac{x-\mu}{\sigma^2}\phi\left(\frac{x-\mu}{\sigma}\right) \end{pmatrix},$$

and $\phi(\cdot)$ is the standard normal probability density function.

Under the assumption of normality, the sample mean and sample variance are asymptotically independent and satisfy

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right), \quad \sqrt{n}(\hat{\sigma} - \sigma) \xrightarrow{d} \mathcal{N}\left(0, \tfrac{\sigma^2}{2}\right).$$

Thus, the asymptotic distribution of $\tilde{F}(x)$ is

$$\sqrt{n}\left(\tilde{F}(x) - F(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \Omega\right),$$

where the asymptotic variance Ω is

$$\Omega = \left(\frac{\partial h}{\partial \mu}\right)^2 \sigma^2 + \left(\frac{\partial h}{\partial \sigma}\right)^2 \frac{\sigma^2}{2}.$$

Substituting the derivatives, we have

$$\begin{aligned} \Omega &= \left(\frac{1}{\sigma}\phi(z)\right)^2 \sigma^2 + \left(\frac{x-\mu}{\sigma^2}\phi(z)\right)^2 \frac{\sigma^2}{2} \\ &= \phi(z)^2 + \frac{(x-\mu)^2}{2\sigma^2}\phi(z)^2, \end{aligned}$$

where $z = \frac{x-\mu}{\sigma}$.

Simplifying, since $(x-\mu)^2/\sigma^2 = z^2$, we get

$$\Omega = \phi(z)^2\left(1 + \frac{z^2}{2}\right).$$

Therefore, $\tilde{F}(x)$ is asymptotically normal with mean $F(x)$ and variance Ω/n:

$$\tilde{F}(x) \approx \mathcal{N}\left(F(x), \frac{\Omega}{n}\right).$$

The estimator $\tilde{F}(x) = \Phi\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)$ is both consistent and asymptotically normal, converging to the true cumulative distribution function $F(x)$ as $n \to \infty$, with an asymptotic variance that can be consistently estimated by replacing μ and σ with $\hat{\mu}$ and $\hat{\sigma}$ in Ω:

$$\hat{\Omega} = \phi\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\left(1 + \frac{\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2}{2}\right).$$

2.d Simulation Study

(d) Conduct a simulation study to examine the empirical performance of both $\hat{F}(x)$ and $\tilde{F}(x)$. Evaluate the consistency and the variance (i.e., the CLT) for both estimators. If the true distribution is Normal, which is more efficient? What happens when the distribution is not Normal? Try several different distributions as well as different parameters for those distributions. Choose three representative values x at which to study $F(x)$. Study what happens as n changes.

Now we turn to estimating the density $f(x)$. The density is the derivative of the cdf, and therefore is given by

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}.$$

2.e Plug-In Estimator

(e) Use (1) and (2), for a fixed h, to give a plug-in estimator for $f(x)$ denoted $\hat{f}(x)$.

We aim to construct a plug-in estimator $\hat{f}(x)$ for the density $f(x)$ using the definition of the density as the derivative of the cumulative distribution function:

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}.$$

For a fixed small $h > 0$, we approximate $f(x)$ by:

$$f(x) \approx \frac{F(x+h) - F(x)}{h}.$$

Using the empirical distribution function $\hat{F}(x)$, the plug-in estimator $\hat{f}(x)$ becomes:

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x)}{h}.$$

Substituting the expression for $\hat{F}(x)$, we get:

$$\begin{aligned}\hat{f}(x) &= \frac{1}{h}\left(\frac{1}{n}\sum_{i=1}^{n} 1\{x_i \le x+h\} - \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \le x\}\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} \left(1\{x_i \le x+h\} - 1\{x_i \le x\}\right).\end{aligned}$$

Simplifying, note that $1\{x_i \le x+h\} - 1\{x_i \le x\}$ equals 1 if $x < x_i \le x+h$ and 0 otherwise. Therefore, the estimator counts the number of observations falling in the interval $(x, x+h]$:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} 1\{x < x_i \leq x+h\} = \frac{N_h(x)}{nh},$$

where $N_h(x)$ is the number of observations in $(x, x+h]$.

Thus, the plug-in estimator for $f(x)$ is:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} 1\{x < x_i \leq x+h\}.$$

2.f Bias of the Estimator

(f) For fixed h, compute the bias of $\hat{f}(x)$. Prove that the bias vanishes as $h \to 0$.

To compute the bias of $\hat{f}(x)$ for fixed h, we start by finding its expected value:

$$\begin{aligned} E[\hat{f}(x)] &= E\left[\frac{1}{nh}\sum_{i=1}^{n} 1\{x < x_i \leq x+h\}\right] \\ &= \frac{1}{h}E\left[1\{x < X \leq x+h\}\right] \\ &= \frac{1}{h}\left[F(x+h) - F(x)\right]. \end{aligned}$$

Using a Taylor series expansion of $F(x+h)$ around x:

$$F(x+h) = F(x) + f(x)h + \frac{1}{2}f'(x)h^2 + o(h^2).$$

Subtracting $F(x)$ and dividing by h:

$$\frac{F(x+h) - F(x)}{h} = f(x) + \frac{1}{2}f'(x)h + o(h).$$

Therefore, the expected value of $\hat{f}(x)$ is:

$$E[\hat{f}(x)] = f(x) + \frac{1}{2}f'(x)h + o(h).$$

The bias of $\hat{f}(x)$ is:

$$\text{Bias}[\hat{f}(x)] = E[\hat{f}(x)] - f(x) = \frac{1}{2}f'(x)h + o(h).$$

As $h \to 0$, the bias approaches zero:

$$\lim_{h\to 0} \text{Bias}[\hat{f}(x)] = \lim_{h\to 0} \left(\frac{1}{2}f'(x)h + o(h)\right) = 0.$$

Thus, the bias of $\hat{f}(x)$ vanishes as $h \to 0$.

2.g Bias Order

(g) Assume that $f(x)$ is twice continuously differentiable. Prove that the bias of $\hat{f}(x)$ is $O(h)$ and characterize the constant. That is, show that

$$\mathbb{E}[\hat{f}(x) - f(x)] = Kh + o(h)$$

and give the precise form of K.

We assume that $f(x)$ is twice continuously differentiable. Our goal is to show that the bias of $\hat{f}(x)$ is $O(h)$ and to find the constant K such that:

$$\mathbb{E}[\hat{f}(x) - f(x)] = Kh + o(h).$$

Starting from the expression for the expected value of $\hat{f}(x)$:

$$\mathbb{E}[\hat{f}(x)] = \frac{1}{h}\left(F(x+h) - F(x)\right).$$

Using the Taylor expansion of $F(x+h)$ around x:

$$F(x+h) = F(x) + f(x)h + \frac{1}{2}f'(x)h^2 + \frac{1}{6}f''(x)h^3 + o(h^3).$$

Subtracting $F(x)$ and dividing by h:

$$\frac{F(x+h) - F(x)}{h} = f(x) + \frac{1}{2}f'(x)h + \frac{1}{6}f''(x)h^2 + o(h^2).$$

Therefore, the expected value of $\hat{f}(x)$ is:

$$\mathbb{E}[\hat{f}(x)] = f(x) + \frac{1}{2}f'(x)h + o(h).$$

The bias of $\hat{f}(x)$ is:

$$\mathbb{E}[\hat{f}(x) - f(x)] = \frac{1}{2}f'(x)h + o(h).$$

Thus, the bias is $O(h)$, and the constant K is given by:

$$K = \frac{1}{2}f'(x).$$

2.h Variance of the Estimator

(h) For fixed h, compute the variance denoted $\Sigma = \mathbb{V}[\hat{f}(x)]$. Provide a consistent estimator.

We compute the variance of $\hat{f}(x)$ for fixed h. Recall that $\hat{f}(x)$ is given by:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} 1\{x < x_i \leq x + h\}.$$

Define the indicator variables:

$$Y_i = 1\{x < x_i \leq x + h\}, \quad i = 1, 2, \ldots, n.$$

Each Y_i is an independent Bernoulli random variable with success probability:

$$p = P(x < X \leq x + h) = F(x + h) - F(x).$$

The variance of $\hat{f}(x)$ is:

$$\begin{aligned}
\mathbb{V}[\hat{f}(x)] &= \mathbb{V}\left(\frac{1}{nh}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{(nh)^2}\sum_{i=1}^{n}\mathbb{V}[Y_i] \\
&= \frac{1}{(nh)^2}\cdot n \cdot p(1-p) \\
&= \frac{p(1-p)}{nh^2}.
\end{aligned}$$

To express $\Sigma = \mathbb{V}[\hat{f}(x)]$ in terms of $f(x)$, we approximate p for small h:

$$\begin{aligned}
p &= F(x + h) - F(x) \\
&= \int_x^{x+h} f(t)\,dt \\
&= f(x)h + \frac{1}{2}f'(x)h^2 + o(h^2).
\end{aligned}$$

Therefore, for small h, we have $p \approx f(x)h$. Then, $p(1-p) \approx f(x)h(1 - f(x)h) \approx f(x)h$, since h is small.

Substituting back into the variance:

$$\mathbb{V}[\hat{f}(x)] \approx \frac{f(x)h}{nh^2} = \frac{f(x)}{nh}.$$

Thus, the variance is:

$$\Sigma = \mathbb{V}[\hat{f}(x)] = \frac{f(x)}{nh} + o\left(\frac{1}{nh}\right).$$

To provide a consistent estimator of Σ, we estimate p using the sample proportion:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i = nh\hat{f}(x) \cdot \frac{1}{n} = h\hat{f}(x).$$

Then, the estimated variance is:

$$\begin{aligned}\hat{\Sigma} &= \frac{\hat{p}(1-\hat{p})}{nh^2} \\ &= \frac{h\hat{f}(x)\left(1 - h\hat{f}(x)\right)}{nh^2} \\ &= \frac{\hat{f}(x)\left(1 - h\hat{f}(x)\right)}{nh}.\end{aligned}$$

Since h is small, $h\hat{f}(x)$ is negligible, and we can approximate:

$$\hat{\Sigma} \approx \frac{\hat{f}(x)}{nh}.$$

Therefore, a consistent estimator of the variance Σ is:

$$\hat{\Sigma} = \frac{\hat{f}(x)}{nh}.$$

2.i Mean Square Error

(i) Compute the mean square error of your estimator and find the value of h that minimizes it. Characterize precisely what happens to this optimal h as $n \to \infty$. How would you choose h in an application for the goal of estimation?

The mean square error (MSE) of the estimator $\hat{f}(x)$ is given by the sum of the squared bias and the variance:

$$\text{MSE}(h) = \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2 + \mathbb{V}[\hat{f}(x)].$$

From previous results, the bias is approximately:

$$\text{Bias} = \mathbb{E}[\hat{f}(x)] - f(x) = \frac{1}{2}f'(x)h + o(h).$$

The variance is approximately:

$$\mathbb{V}[\hat{f}(x)] = \frac{f(x)}{nh} + o\left(\frac{1}{nh}\right).$$

Ignoring higher-order terms, the MSE becomes:

$$\text{MSE}(h) = \left(\frac{1}{2}f'(x)h\right)^2 + \frac{f(x)}{nh} = \frac{1}{4}[f'(x)]^2h^2 + \frac{f(x)}{nh}.$$

To find the value of h that minimizes the MSE, take the derivative of MSE(h) with respect to h and set it equal to zero:

$$\frac{d}{dh}\text{MSE}(h) = \frac{1}{2}[f'(x)]^2h - \frac{f(x)}{nh^2} = 0.$$

Solving for h:

$$\frac{1}{2}[f'(x)]^2h = \frac{f(x)}{nh^2},$$

$$\frac{1}{2}[f'(x)]^2nh^3 = f(x),$$

$$h^3 = \frac{2f(x)}{[f'(x)]^2n}.$$

Therefore, the optimal bandwidth h that minimizes the MSE is:

$$h_{\text{opt}} = \left(\frac{2f(x)}{[f'(x)]^2n}\right)^{1/3}.$$

As $n \to \infty$, the optimal h behaves like:

$$h_{\text{opt}} \propto n^{-1/3}.$$

This means that the optimal bandwidth decreases at the rate of $n^{-1/3}$ as the sample size increases.

In an application aiming for estimation, we should choose h proportional to $n^{-1/3}$ to balance the bias and variance, minimizing the MSE. Specifically:

$$h = Cn^{-1/3},$$

where C is a constant that may depend on estimates of $f(x)$ and $f'(x)$. Since $f(x)$ and $f'(x)$ are typically unknown, we can use pilot estimates or assume reasonable values based on prior knowledge to select h.

2.j Asymptotic Normality for Fixed h

(j) For fixed h, prove that

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\Sigma^{1/2}} \to_d \mathcal{N}(0, 1).$$

We aim to prove that for fixed h:

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\Sigma^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\Sigma = \mathbb{V}[\hat{f}(x)]$.

Recall that:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i,$$

with $Y_i = 1\{x < x_i \leq x + h\}$. The Y_i are independent and identically distributed (i.i.d.) Bernoulli random variables with success probability:

$$p = P(x < X \leq x + h) = F(x + h) - F(x).$$

The mean and variance of Y_i are:

$$\mathbb{E}[Y_i] = p, \quad \mathbb{V}[Y_i] = p(1 - p).$$

The expected value and variance of $\hat{f}(x)$ are:

$$\mathbb{E}[\hat{f}(x)] = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{p}{h},$$

$$\mathbb{V}[\hat{f}(x)] = \frac{1}{(nh)^2} \sum_{i=1}^{n} \mathbb{V}[Y_i] = \frac{p(1 - p)}{nh^2} = \Sigma.$$

Define the standardized version of $\hat{f}(x)$:

$$Z_n = \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\Sigma^{1/2}} = \frac{\frac{1}{nh}\sum_{i=1}^n Y_i - \frac{p}{h}}{\left(\frac{p(1-p)}{nh^2}\right)^{1/2}} = \frac{\sum_{i=1}^n (Y_i - p)}{\sqrt{np(1-p)}}.$$

Since the Y_i are i.i.d. with finite variance, by the Central Limit Theorem:

$$\frac{\sum_{i=1}^n (Y_i - p)}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Therefore,

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\Sigma^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

This completes the proof.

2.k Sufficient Conditions for Asymptotic Normality

(k) Provide sufficient conditions so that

$$\frac{\hat{f}(x) - f(x)}{\Sigma^{1/2}} \to_d \mathcal{N}(0,1).$$

Characterize precisely the requirements that h must obey as $n \to \infty$.

We are to provide sufficient conditions such that:

$$\frac{\hat{f}(x) - f(x)}{\Sigma^{1/2}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\Sigma = \mathbb{V}[\hat{f}(x)]$.

From earlier results:

The bias of $\hat{f}(x)$ is approximately:

$$\mathbb{E}[\hat{f}(x)] - f(x) = \frac{1}{2}f'(x)h + o(h).$$

The variance of $\hat{f}(x)$ is approximately:

$$\Sigma = \mathbb{V}[\hat{f}(x)] = \frac{f(x)}{nh} + o\left(\frac{1}{nh}\right).$$

The standard deviation is:

$$\Sigma^{1/2} = \sqrt{\frac{f(x)}{nh}} + o\left(\sqrt{\frac{1}{nh}}\right).$$

To ensure that the standardized estimator converges in distribution to a standard normal, the bias must be negligible compared to the standard deviation. Specifically, we require:

$$\frac{\mathbb{E}[\hat{f}(x)] - f(x)}{\Sigma^{1/2}} \to 0 \quad \text{as } n \to \infty.$$

Computing the standardized bias:

$$\begin{aligned}\frac{\mathbb{E}[\hat{f}(x)] - f(x)}{\Sigma^{1/2}} &\approx \frac{\frac{1}{2}f'(x)h}{\sqrt{\frac{f(x)}{nh}}} \\ &= \frac{1}{2}f'(x)h \cdot \sqrt{\frac{nh}{f(x)}} \\ &= \frac{1}{2}\frac{f'(x)}{\sqrt{f(x)}}\sqrt{nh^3}.\end{aligned}$$

Therefore, to have the standardized bias tend to zero, we need:

$$\sqrt{nh^3} \to 0 \quad \text{as } n \to \infty.$$

This implies:

$$nh^3 \to 0 \quad \text{as } n \to \infty.$$

At the same time, to ensure that the variance Σ shrinks to zero (i.e., the estimator becomes more precise), we require:

$$nh \to \infty \quad \text{as } n \to \infty.$$

In summary, the sufficient conditions are:

- $h \to 0$ as $n \to \infty$,
- $nh \to \infty$ as $n \to \infty$,
- $nh^3 \to 0$ as $n \to \infty$.

Characterizing the Requirements on h:

Let us consider h of the form:

$$h = n^{-\beta},$$

for some $\beta > 0$.

We analyze the conditions:

1. $h \to 0$:

$$h = n^{-\beta} \to 0 \quad \text{if } \beta > 0.$$

2. $nh = n \cdot n^{-\beta} = n^{1-\beta} \to \infty$:

$$nh \to \infty \quad \text{if } 1 - \beta > 0 \quad \text{or} \quad \beta < 1.$$

3. $nh^3 = n \cdot n^{-3\beta} = n^{1-3\beta} \to 0$:

$$nh^3 \to 0 \quad \text{if } 1 - 3\beta < 0 \quad \text{or} \quad \beta > \frac{1}{3}.$$

Combining these conditions, we require:

$$\frac{1}{3} < \beta < 1.$$

Therefore, choosing h such that:

$$h = n^{-\beta}, \quad \text{with} \quad \beta \in \left(\frac{1}{3}, 1\right),$$

satisfies all the sufficient conditions.

Thus, For the asymptotic normality:

$$\frac{\hat{f}(x) - f(x)}{\Sigma^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1),$$

to hold, it is sufficient that:

- The bandwidth h decreases to zero at a rate $h = n^{-\beta}$ with $\beta \in \left(\frac{1}{3}, 1\right)$.
- This ensures $h \to 0$, $nh \to \infty$, and $nh^3 \to 0$ as $n \to \infty$.

2.1 Comparison of Requirements for h

(l) Compare the requirements on h in part (k) to what you found in part (i). Discuss what you find. How would you choose h in an application for the goal of inference?

In part (i), we found that the bandwidth h that minimizes the mean square error (MSE) of the estimator $\hat{f}(x)$ is:

$$h_{\text{opt}} = \left(\frac{2f(x)}{[f'(x)]^2 n} \right)^{1/3} \propto n^{-1/3}.$$

This implies that to minimize the MSE, we should choose h proportional to $n^{-1/3}$.

In part (k), we determined sufficient conditions for the asymptotic normality of the standardized estimator:

$$\frac{\hat{f}(x) - f(x)}{\Sigma^{1/2}} \xrightarrow{d} \mathcal{N}(0,1),$$

which require that:

- $h \to 0$ as $n \to \infty$,
- $nh \to \infty$ as $n \to \infty$,
- $nh^3 \to 0$ as $n \to \infty$.

These conditions are satisfied when $h = n^{-\beta}$ with β in the interval $\left(\frac{1}{3}, 1\right)$.

Comparing these results, we observe that:

- The optimal h for minimizing MSE is $h_{\text{opt}} \propto n^{-1/3}$, which corresponds to $\beta = \frac{1}{3}$.
- The asymptotic normality requires $\beta > \frac{1}{3}$.

This indicates a trade-off between bias and variance:

- Choosing h proportional to $n^{-1/3}$ minimizes the MSE but does not satisfy the condition $nh^3 \to 0$, since $nh^3 = n \cdot (n^{-1/3})^3 = 1$, which does not converge to zero.
- To achieve asymptotic normality for inference purposes, we need h to decrease slightly faster than $n^{-1/3}$, i.e., $h \propto n^{-\beta}$ with $\beta > \frac{1}{3}$.

In practice, when the goal is estimation (minimizing MSE), we might choose $h \propto n^{-1/3}$. However, for inference (e.g., constructing confidence intervals), we need the standardized estimator to be asymptotically normal. Therefore, we should choose h such that:

$$h = n^{-\beta}, \quad \text{with} \quad \beta \in \left(\frac{1}{3}, 1 \right).$$

By selecting β slightly greater than $\frac{1}{3}$, we ensure that:

- The bias becomes negligible compared to the standard deviation.
- The conditions $nh \to \infty$ and $nh^3 \to 0$ are satisfied.

Conclusion: For the goal of inference, we would choose h proportional to $n^{-\beta}$ with β slightly greater than $\frac{1}{3}$. This choice balances the need for the estimator to be asymptotically normal (which facilitates valid statistical inference) while controlling the bias and variance.

Therefore, in an application focused on inference, we would select h such that:

$$h = n^{-\beta}, \quad \text{where} \quad \beta = \frac{1}{3} + \varepsilon, \quad \varepsilon > 0.$$

This ensures that the standardized estimator converges in distribution to a normal distribution, enabling us to construct confidence intervals and perform hypothesis tests reliably.

2.m Simulation Study on Empirical Performance

(m) Conduct a simulation study to examine the empirical performance of $\hat{f}(x)$. Evaluate the bias and variance of your estimator and the quality of the Normal approximation. Compute the empirical coverage and length of 95% confidence intervals. Study what happens as you vary n, h, the true distribution, and the evaluation point x.

3 Application

The file `Banerji-Berry-Shotland_2017_AEJ.csv` contains data from a recent paper.

The outcome is a (normalized) child's test, in `caser_total_norm`. `treatment` has four different values, indicating different trainings for mothers. The first is the baseline/control. There are six X variables (dummies) and three W variables (continuous). We want to explore the impact of each treatment relative to the baseline (`treatment`=1).

LASSO & Discrete Heterogeneity

3.a Run a Single Regression

(a) Run a single linear regression that provides estimates and inference for $\mu_t = \mathbb{E}[Y(t)]$, $t = 1, 2, 3, 4$. Add covariates to the regression to see if efficiency is improved. First add the covariates directly and then do it demeaned and interacted. Try adding interactions among the X and W.

3.b LASSO to Select Controls

(b) Use the LASSO to select controls in one of the models you ran above. Leave the treatment coefficients unpenalized. Is precision improved?

3.c Inference for the Heterogeneous

(c) Choose one of the X variables out of the six. Run a single linear regression that provides estimates and inference for the heterogeneous effects $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$ for $x = \{0, 1\}$ (i.e., eight total numbers).

3.d LASSO to All Variables

(d) Add all the other X variables, and the W variables, and interactions and polynomials, and apply the lasso to select controls while still giving inference on the eight $\mu_t(x)$. Is precision improved?

3.e Sample A and Sample B

(e) Split the data randomly in two pieces, call them sample A and sample B. In sample A, use the lasso to identify the most impacted subgroups based on X and interactions in X (go up to only two- or three-way interactions). Use sample B to validate the size of these impacts and do hypothesis testing. Discuss the role played by sample splitting in this case.

Binsreg & Continuous Heterogeneity

3.f Other Controls

For $j = 1, 2, 3$, define $\omega_t(w_j) = \mathbb{E}[Y(t) \mid W_j = w_j]$.

(f) Use `binsreg` to plot all possible $\omega_t(w_j)$ (probably not in one picture). What did you specify for the other controls and why?

3.g Substantive

(g) Pick one W_j and use confidence bands to assess a substantive question about $\omega_t(w_j)$, $t = 1, 2, 3, 4$. For example, is it monotonic? Are there decreasing returns? Etc.

Deep Nets and Forests

3.h Conditions for Identification

(h) Consider the model

$$Y_i = \sum_{t=1}^{4} \mu_t(x, w)\mathbf{1}\{T_i = t\} + \epsilon_i.$$

Provide conditions under which the functions $\mu_t(x, w) = \mathbb{E}[Y(t) \mid X = x, W = w]$ are identified.

3.i Random Forest Full Flexibility

(i) Apply random forests to learn $\mu_t(x, w)$ full flexibly. For each one, create a partial dependence plot for each continuous w_j. How do these compare to what you found in (f)? For each $\mu_t(x, w)$, create and discuss the variable importance plot. Do these make sense to you for this application?

3.j Neural Networks Full Flexibility

(j) Use neural networks to learn $\mu_t(x, w)$ full flexibly. Try several different architectures for your deep nets. Select a single one as the best and justify your choice.

3.k Inference with Influence Function

(k) Conduct inference on the treatment effect of treatment t compared to baseline, $\mathbb{E}[\mu_t(X, W) - \mu_0(X, W)]$, using the influence function based estimation from class and preliminary estimates from both (i) and (j).