ECMA 31380 - Causal Machine Learning - Homework 3

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 $\hat{\theta}_{\mathrm{MLE}}$

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

To derive the influence function for $\hat{\theta}_{\mathrm{MLE}}$, 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 $\hat{\theta}_{\mathrm{MLE}}$ is an M-estimator.

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

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

where J is the expected information matrix given by:

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

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:

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

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

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

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,
- $\psi(Z;\theta) = -2\left[T - p(X;\theta)\right] p(X;\theta)[1-p(X;\theta)]X$ 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:

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

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

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

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:

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

  where:

  - $\phi_{\mu_0}^{(1)}(Z_i)$ accounts for the variability in Y_i and T_i.
  - $\phi_{\mu_0}^{(2)}(Z_i)$ accounts for the estimation error in $\hat{\theta}_{\text{MLE}}$.

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

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

- *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:

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

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 $\hat{p}(X_i) = p(X_i; \hat{\theta}_{\text{NLS}})$ is the estimated propensity score using NLS.

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

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

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:

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

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:

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

  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
```

```r
  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 $F(x) = \mathbb{P}(X \leq x)$, 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:

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

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

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

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

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.

Consider the indicator variables

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

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

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

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

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

In this case, we can use n in the denominator of $\hat{\sigma}$ or $n-1$. the denominator with n is the result derived from the MLE function. For finite samples, $\hat{\sigma}_{\text{MLE}}$ is biased. Nonetheless, assyntotically, it has the properties as the estimator using $n-1$ and it is unbiased:

$$\hat{\sigma}^2_{\text{MLE}} \xrightarrow{p} \sigma^2 \quad \text{and} \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2 \quad \text{as } n \to \infty.$$

We continue with $\hat{\sigma}_{\text{MLE}}$ to simplify further calculations.

Again, 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.

The Delta Method states that if:

- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, meaning that the estimator $\hat{\theta}$ is asymptotically Normal with mean θ and covariance matrix Σ.
- $h(\theta)$ is a function that is continuously differentiable at θ.

Then:

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} \mathcal{N}(0, \nabla h(\theta)^\top \Sigma \nabla h(\theta))$$

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 partial derivatives of $h(\theta)$ are

$$\frac{\partial h}{\partial \mu} = \frac{\partial\left(\frac{x-\mu}{\sigma}\right)}{\partial \mu}\phi\left(\frac{x-\mu}{\sigma}\right) = -\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\frac{\partial h}{\partial \sigma} = \frac{\partial\left(\frac{x-\mu}{\sigma}\right)}{\partial \sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = -\frac{x-\mu}{\sigma^2}\phi\left(\frac{x-\mu}{\sigma}\right)$$

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

Therefore, the $\nabla h(\theta)$ is defined as:

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

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.

In this simulation, we use 9 different distributions:

- Normal(1, 1)
- Exponential(2)

- Gamma(3, 2)
- Gamma(1, 1)
- Log-Normal(1, 1)
- Log-Normal(1, 2)
- T-Student(10)
- T-Student(70)
- T-Student(150)

We run simulations for n between 10 and 400, with intervals of 5. For each n, we run 500 simulations. In total, we run 355 thousand simulations.

For $x = 0.5$, $x = 1.5$ and $x = 2.5$ we provide estimates for:

- Confidence Interval: a raw estimate of the CI, using $\hat{\text{CDF}} \pm 2 \times \text{Var}(\hat{\text{CDF}})$ for that particular sample size and distribution. The estimate is not statistically precise, but provides intuition. We define:
$$\hat{\text{CDF}} = \frac{1}{n}\sum_{s}^{S} \hat{\text{CDF}}_s$$
where S is the number of simulations for each n.
- Bias: how much the estimated CDF differs from the actual CDF.
$$\text{Bias} = \hat{\text{CDF}} - \text{CDF}$$
- Variance:
$$\text{Variance} = \frac{1}{S-1}\sum_{s}^{S}\left(\hat{\text{CDF}}_s - \hat{\text{CDF}}\right)$$
- MSE (Mean Squared Error):
$$\text{MSE} = \text{Variance} + \text{Bias}^2$$

We present the following takeaways:

- When the data generating process has normal distribution $\tilde{F}(x)$ outperforms $\hat{F}(x)$.
- The variance of $\tilde{F}(x)$ is almost always smaller. Nonetheless, the $\tilde{F}(x)$ estimator is biased for most distributions, leading to considerably higher MSE.
- MSE of $\tilde{F}(x)$ only outperforms the MSE of $\hat{F}(x)$ for distributions similar to normal. We see that in T-Student distribution with high *df* (e.g. 70 and 150), distributions extremely similar to a normal.
- In all distributions, with the exception of T-Student with high *df* and Normal, increase in n leads in most cases to equal or bigger bias when using $\tilde{F}(x)$.

In conclusion, $\tilde{F}(x)$ is only preferable in situations where we have a strong case to believe the data generating process follows a normal or t-student with high df.

REMOVE COMMENT

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:

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

Simplifying, note that $1\{x_i \leq x + h\} - 1\{x_i \leq x\}$ equals 1 if $x < x_i \leq x + h$ 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 $h \to 0$.

In a more informal way, it is interesting to view the bias as:

$$\begin{aligned}
\mathbb{E}[\hat{f}(x) - f(x)] &= \mathbb{E}\left[\frac{\hat{F}(x+h) - \hat{F}(x)}{h}\right] - \mathbb{E}\left[\frac{F(x+h) - F(x)}{h}\right] \\
&= \mathbb{E}\left[\frac{\hat{F}(x+h)}{h}\right] - \mathbb{E}\left[\frac{F(x+h)}{h}\right] \quad \text{(from the results of 2.e)} \\
&= \frac{1}{h}\left(\mathbb{E}\left[\hat{F}(x+h)\right] - \mathbb{E}\left[F(x+h)\right]\right) \\
&= \frac{1}{h}\left(\mathbb{E}\left[\frac{1}{n}\sum_i^n 1\{x_i < x+h\}\right] - F(x+h)\right) \\
&= \frac{1}{h}\left(\frac{1}{n}\left[\sum_i^n \mathbb{E}[1\{x_i \leq x\} + 1\{x < x_i \leq x+h\}]\right] - \mathbb{P}(X \leq x+h)\right) \\
&= \frac{1}{h}\left(\frac{1}{n}\left[\sum_i^n (\mathbb{E}[1\{x_i \leq x\}] + \mathbb{E}[1\{x < x_i \leq x+h\}])\right] - \mathbb{P}(X \leq x+h)\right) \\
&= \frac{1}{h}\left(\mathbb{P}(X < x) + \frac{n_h}{n} - \mathbb{P}(X \leq x+h)\right) \\
&= \frac{1}{h}\left(\frac{n_h}{n} - \mathbb{P}(x < X \leq x+h)\right)
\end{aligned}$$

Thus,

$$\lim_{h \to 0} \frac{1}{h}\left(\frac{n_h}{n} - \mathbb{P}(x < X \leq x+h)\right) = 0$$

CHECK: Is it worth to keep this last part?

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.

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:

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

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 $\text{MSE}(h)$ 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:

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

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

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

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

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

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:

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

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.

This choice balances the need for the estimator to be asymptotically normal (which facilitates valid statistical inference) while controlling the bias and variance.

The optimal choice for h is:

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

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.

For this question, we use the following calculations:

$$p_{x,h} = \frac{n_h}{n}, \quad \text{where } n_h \text{ is the n. of obs in } h \text{and } n \text{is the n. of obs in the sample simulation}$$

$$\hat{f}(x)_h = \frac{p_{x,h}}{h}$$

$$\text{Var}[\hat{f}(x)_h] = \frac{p_{x,h}(1-p_{x,h})}{nh^2} = \hat{\sigma}^2_{x,h}$$

$$\text{CI} = \left[\hat{f}(x)_h - t_{n-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_{x,h}, \quad \hat{f}(x)_h + t_{n-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_{x,h}\right]$$

We run simulations using:

- Sample sizes n of: 10, 25, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000
- Interval h of:

  $$h_i = \frac{1}{n_i^{1/3}} + \varepsilon \quad \text{for } n_i \text{ in sample sizes available and } \varepsilon = 0.0001$$

  The intervals tested coincide with the optimal values found for the previous solutions.
- The following distributions:
  - T-Student(30)
  - T-Student(50)
  - T-Student(100)
  - Normal(0, 1)
  - Normal(1, 1)
  - Gamma(1, 1)
  - Gamma(2, 3)
  - Log-Normal(1, 1)
- 1000 simulations for every combination of n, h and distribution, using the results for $x = 0.5$ and $x = 1$, totaling 1.792 million simulations.

We check the performance of the results based on the efficiency of the CI.

Given that we always use CI of 95%, we hope that in 95% of the simulations the true $f(x)$ falls within the ranges determined.

Our results agree with the theoretical conclusions: for smaller samples, bigger h yields more precise results. As n increases, the results using bigger h has a considerable decrease in performance.

The best results, as expected, are found for big n and, on average, $h = \frac{1}{n^{1/3}}$, as specified to provide good inference.

REMOVE COMMENT

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.

Regression without Covariates

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_data %>% select(caser_total_norm,
    t2, t3, t4))

Residuals:
    Min      1Q  Median      3Q     Max
-1.3966 -0.8517 -0.2407  0.7092  2.2221

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.18733    0.01643  11.403  < 2e-16 ***
t2           0.05233    0.02325   2.250 0.024434 *
t3           0.07789    0.02353   3.310 0.000936 ***
t4           0.10038    0.02331   4.307 1.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9998 on 14570 degrees of freedom
Multiple R-squared:  0.001408, Adjusted R-squared:  0.001202
F-statistic: 6.847 on 3 and 14570 DF,  p-value: 0.0001319
```

Listing 2: Regression without Covariates

The estimators for the treatment are:

- $\hat{\mu}_1 = \mathbb{E}[Y(1)] = 0.18733$
- $\hat{\mu}_2 = \mathbb{E}[Y(2)] = 0.18733 + 0.05233 = 0.23966$
- $\hat{\mu}_3 = \mathbb{E}[Y(3)] = 0.18733 + 0.07789 = 0.26522$
- $\hat{\mu}_4 = \mathbb{E}[Y(4)] = 0.18733 + 0.10038 = 0.28771$

All treatment are statistically significanly from t_1 (control).

Regression with Covariates

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1702 -0.2850 -0.0686  0.2346  3.1656

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.047014   0.024572   1.913  0.05573 .
age                   0.010872   0.002372   4.584 4.60e-06 ***
state                 0.007818   0.008483   0.922  0.35678
bl_caser_total_norm   0.852219   0.004639 183.720  < 2e-16 ***
boy                   0.052580   0.007636   6.886 5.96e-12 ***
number_of_kids       -0.008028   0.002508  -3.201  0.00137 **
mother_educ           0.120762   0.011264  10.721  < 2e-16 ***
factor_educ           0.073858   0.008155   9.057  < 2e-16 ***
mother_age30         -0.017244   0.007926  -2.176  0.02961 *
farmingIncome         0.034616   0.008106   4.270 1.96e-05 ***
t2                    0.014434   0.010533   1.370  0.17057
t3                    0.025175   0.010667   2.360  0.01828 *
t4                    0.055961   0.010569   5.295 1.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4525 on 14561 degrees of freedom
Multiple R-squared:  0.7955,␣Adjusted R-squared:  0.7954
F-statistic:  4721 on 12 and 14561 DF,  p-value: < 2.2e-16
```

Listing 3: Regression with Covariates

When adding the covariates without demeaning them, we cannot recover directly the influence of the treatment. This happens because the mean of the covariates is not zero and, thus, influences the parameters of t_2, t_3, t_4.

Regression with Demeaned Covariates

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_demeaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1702 -0.2850 -0.0686  0.2346  3.1656

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.220804   0.007444  29.660  < 2e-16 ***
age                  0.010872   0.002372   4.584 4.60e-06 ***
state                0.007818   0.008483   0.922  0.35678
bl_caser_total_norm  0.852219   0.004639 183.720  < 2e-16 ***
boy                  0.052580   0.007636   6.886 5.96e-12 ***
number_of_kids      -0.008028   0.002508  -3.201  0.00137 **
mother_educ          0.120762   0.011264  10.721  < 2e-16 ***
factor_educ          0.073858   0.008155   9.057  < 2e-16 ***
mother_age30        -0.017244   0.007926  -2.176  0.02961 *
farmingIncome        0.034616   0.008106   4.270 1.96e-05 ***
t2                   0.014434   0.010533   1.370  0.17057
t3                   0.025175   0.010667   2.360  0.01828 *
t4                   0.055961   0.010569   5.295 1.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4525 on 14561 degrees of freedom
Multiple R-squared:  0.7955, Adjusted R-squared:  0.7954
F-statistic:  4721 on 12 and 14561 DF,  p-value: < 2.2e-16
```

Listing 4: Regression with Demeaned Covariates

The estimators for the treatment are:

- $\hat{\mu}_1 = \mathbb{E}[Y(1)] = 0.220804$
- $\hat{\mu}_2 = \mathbb{E}[Y(2)] = 0.220804 + 0.014434 = 0.23524$
- $\hat{\mu}_3 = \mathbb{E}[Y(3)] = 0.220804 + 0.025175 = 0.24598$
- $\hat{\mu}_4 = \mathbb{E}[Y(4)] = 0.220804 + 0.055961 = 0.27677$

Regression with Demeaned Covariates and Interaction with Treatment

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_interaction_with_treatment_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3.15439 -0.28318 -0.06778  0.23171  3.14488

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.2206781  0.0074665  29.556  < 2e-16 ***
age                       0.0078702  0.0047159   1.669 0.095165 .
state                    -0.0157254  0.0167048  -0.941 0.346531
bl_caser_total_norm       0.8560157  0.0093911  91.151  < 2e-16 ***
boy                       0.0587387  0.0150555   3.901  9.6e-05 ***
number_of_kids           -0.0053473  0.0053995  -0.990 0.322031
mother_educ               0.1036680  0.0228584   4.535  5.8e-06 ***
factor_educ               0.0614496  0.0163795   3.752 0.000176 ***
mother_age30             -0.0239711  0.0159705  -1.501 0.133387
farmingIncome             0.0360261  0.0157747   2.284 0.022398 *
t2                        0.0145792  0.0105516   1.382 0.167082
t3                        0.0251895  0.0106849   2.357 0.018412 *
t4                        0.0566208  0.0105900   5.347  9.1e-08 ***
`age:t2`                 -0.0012798  0.0066639  -0.192 0.847706
`state:t2`                0.0166727  0.0236875   0.704 0.481531
`bl_caser_total_norm:t2`  0.0088660  0.0130646   0.679 0.497385
`boy:t2`                  0.0120852  0.0214209   0.564 0.572642
`number_of_kids:t2`      -0.0051515  0.0074093  -0.695 0.486899
`mother_educ:t2`          0.0435934  0.0324624   1.343 0.179328
`factor_educ:t2`         -0.0105803  0.0229996  -0.460 0.645507
`mother_age30:t2`         0.0385381  0.0223856   1.722 0.085171 .
`farmingIncome:t2`       -0.0076130  0.0225113  -0.338 0.735229
`age:t3`                 -0.0009993  0.0067380  -0.148 0.882104
`state:t3`                0.0328596  0.0243263   1.351 0.176785
`bl_caser_total_norm:t3` -0.0027395  0.0132984  -0.206 0.836794
`boy:t3`                 -0.0425143  0.0216588  -1.963 0.049676 *
`number_of_kids:t3`      -0.0001652  0.0071542  -0.023 0.981581
`mother_educ:t3`          0.0227090  0.0316270   0.718 0.472753
`factor_educ:t3`          0.0444950  0.0234417   1.898 0.057701 .
`mother_age30:t3`         0.0277102  0.0226109   1.226 0.220397
`farmingIncome:t3`        0.0079026  0.0230456   0.343 0.731669
`age:t4`                  0.0153258  0.0066877   2.292 0.021940 *
`state:t4`                0.0435920  0.0236954   1.840 0.065836 .
`bl_caser_total_norm:t4` -0.0242199  0.0131799  -1.838 0.066137 .
`boy:t4`                  0.0025854  0.0214199   0.121 0.903930
`number_of_kids:t4`      -0.0049219  0.0073465  -0.670 0.502890
`mother_educ:t4`          0.0010198  0.0320488   0.032 0.974616
`factor_educ:t4`          0.0191965  0.0229212   0.837 0.402327
`mother_age30:t4`        -0.0400953  0.0224720  -1.784 0.074407 .
`farmingIncome:t4`       -0.0006717  0.0227106  -0.030 0.976406
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4522 on 14534 degrees of freedom
Multiple R-squared:  0.7962, Adjusted R-squared:  0.7957
F-statistic:  1456 on 39 and 14534 DF,  p-value: < 2.2e-16
```

Listing 5: Regression with Demeaned Covariates and Interaction with Treatment

Regression with Demeaned Covariates and Interaction with Discrete Variables

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_interaction_with_discrete_features_
    data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1965 -0.2816 -0.0662  0.2292  3.1190

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          0.2334465  0.0076312  30.591  < 2e-16 ***
age                                  0.0120631  0.0025196   4.788 1.70e-06 ***
state                                0.0063704  0.0085238   0.747 0.454855
bl_caser_total_norm                  0.8569115  0.0047928 178.792  < 2e-16 ***
boy                                  0.0525410  0.0076121   6.902 5.33e-12 ***
number_of_kids                      -0.0093497  0.0025912  -3.608 0.000309 ***
mother_educ                          0.1145596  0.0119436   9.592  < 2e-16 ***
factor_educ                          0.0716927  0.0081653   8.780  < 2e-16 ***
mother_age30                        -0.0174407  0.0079114  -2.204 0.027505 *
farmingIncome                        0.0303828  0.0081049   3.749 0.000178 ***
t2                                   0.0126451  0.0104912   1.205 0.228106
t3                                   0.0254664  0.0106236   2.397 0.016536 *
t4                                   0.0555816  0.0105252   5.281 1.30e-07 ***
`boy:age`                            0.0021971  0.0045344   0.485 0.628002
`factor_educ:age`                   -0.0010452  0.0048498  -0.216 0.829377
`farmingIncome:age`                  0.0016801  0.0048539   0.346 0.729247
`mother_age30:age`                  -0.0174581  0.0047981  -3.639 0.000275 ***
`mother_educ:age`                   -0.0170993  0.0075371  -2.269 0.023302 *
`state:age`                         -0.0057160  0.0054432  -1.050 0.293685
`boy:bl_caser_total_norm`           -0.0270815  0.0089213  -3.036 0.002405 **
`factor_educ:bl_caser_total_norm`   -0.0363072  0.0099022  -3.667 0.000247 ***
`farmingIncome:bl_caser_total_norm` -0.0121865  0.0093312  -1.306 0.191575
`mother_age30:bl_caser_total_norm`   0.0091714  0.0090524   1.013 0.311010
`mother_educ:bl_caser_total_norm`    0.0022111  0.0125935   0.176 0.860632
`state:bl_caser_total_norm`          0.0652783  0.0098139   6.652 3.00e-11 ***
`boy:number_of_kids`                 0.0011213  0.0048828   0.230 0.818374
`factor_educ:number_of_kids`        -0.0067374  0.0051856  -1.299 0.193878
`farmingIncome:number_of_kids`      -0.0092593  0.0054203  -1.708 0.087610 .
`mother_age30:number_of_kids`       -0.0008231  0.0050972  -0.161 0.871722
`mother_educ:number_of_kids`        -0.0017150  0.0077143  -0.222 0.824073
`state:number_of_kids`              -0.0216544  0.0055545  -3.899 9.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4503 on 14543 degrees of freedom
Multiple R-squared:  0.7978, Adjusted R-squared:  0.7974
F-statistic:  1913 on 30 and 14543 DF,  p-value: < 2.2e-16
```

Listing 6: Regression with Demeaned Covariates and Interaction with Discrete Variables

The estimators for the treatment are:

- $\hat{\mu}_1 = \mathbb{E}[Y(1)] = 0.2334465$
- $\hat{\mu}_2 = \mathbb{E}[Y(2)] = 0.2334465 + 0.0126451 = 0.2460916$

- $\hat{\mu}_3 = \mathbb{E}[Y(3)] = 0.2334465 + 0.0254664 = 0.2589129$
- $\hat{\mu}_4 = \mathbb{E}[Y(4)] = 0.2334465 + 0.0555816 = 0.2890281$

Comparison

Model	$t_1 + t_2$	p-value t_2	Std Error t_2
W/out Cov	0.18733 + 0.05233 = 0.239659	0.024434	0.02325
W/ Dem Cov	0.220804 + 0.014434 = 0.235238	0.17057	0.010533
W/ Dem Cov Disc Interac	0.2334465 + 0.0126451 = 0.2460916	0.228106	0.0104912

Table 2: ATE for t_2

t_2 significance level decreases as we add covariates, but the standard error decreases with the addition of more covariates.

Model	$t_1 + t_2$	p-value t_2	Std Error t_2
W/out Cov	0.18733 + 0.05233 = 0.239659	0.024434	0.02325
W/ Dem Cov	0.220804 + 0.014434 = 0.235238	0.17057	0.010533
W/ Dem Cov Disc Interac	0.2334465 + 0.0126451 = 0.2460916	0.228106	0.0104912

Table 3: ATE for t_2

t_3 significance level decreases as we add covariates. t_3 is still significant after adding covariates, showing relevant statistical difference between the control level and the treatment 3 level. The standard error decreases with the addition of more covariates.

Model	$t_1 + t_4$	p-value t_4	Std Error t_4
W/out Cov	0.18733 + 0.10038 = 0.28771	1.66 e -05	0.02331
W/ Dem Cov	0.220804 + 0.055961 = 0.276765	1.21 e -07	0.010569
W/ Dem Cov Disc Interac	0.2334465 + 0.0555816 = 0.2890281	1.30 e -07	0.0105252

Table 4: ATE for t_4

t_4 significance level increases as we add covariates. t_4 is highly significant even at 1% level regardless of the covariates added, showing relevant statistical difference between the control level and the treatment 4 level. The standard error decreases with the addition of more covariates.

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?

LASSO Model without Interactions

```
[1] "Best lambda: 0.00172896895443247"
13 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          0.220644695
age                  0.018719744
state                .
bl_caser_total_norm  0.862970413
boy                  0.024063319
number_of_kids      -0.010475785
mother_educ          0.041056669
factor_educ          0.034546832
mother_age30        -0.006418082
farmingIncome        0.013878908
t2                   0.014654484
t3                   0.025316624
t4                   0.056238208
```

Listing 7: Regression with Demeaned Covariates and Interaction with Discrete Variables

Model without Interactions controlling for LASSO

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_interaction_with_discrete_features_
    data %>%
    select(all_features %>% colnames()) %>% .[, non_zero_features] %>%
    cbind(banerji_interaction_with_discrete_features_data %>%
        select(caser_total_norm)))

Residuals:
    Min      1Q  Median      3Q     Max
-3.1670 -0.2845 -0.0686  0.2359  3.1580

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.220710   0.007444  29.650  < 2e-16 ***
age                  0.011235   0.002339   4.803 1.57e-06 ***
bl_caser_total_norm  0.851664   0.004599 185.169  < 2e-16 ***
boy                  0.052720   0.007634   6.906 5.20e-12 ***
number_of_kids      -0.007833   0.002499  -3.134  0.00173 **
mother_educ          0.121976   0.011186  10.904  < 2e-16 ***
factor_educ          0.072551   0.008031   9.034  < 2e-16 ***
mother_age30        -0.016881   0.007917  -2.132  0.03299 *
farmingIncome        0.032076   0.007623   4.208 2.59e-05 ***
t2                   0.014645   0.010530   1.391  0.16431
t3                   0.025218   0.010667   2.364  0.01808 *
t4                   0.056082   0.010568   5.307 1.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4525 on 14562 degrees of freedom
Multiple R-squared:  0.7955,␣Adjusted R-squared:  0.7954
F-statistic:  5150 on 11 and 14562 DF,  p-value: < 2.2e-16
```

Listing 8: Model without Interactions controlling for LASSO

LASSO Model with Interactions

```
[1] "Best lambda: 0.00189044654131057"
31 x 1 sparse Matrix of class "dgCMatrix"
                                              s1
(Intercept)                         0.2211187665
age                                 0.0190715846
state                               .
bl_caser_total_norm                 0.8684083028
boy                                 0.0241497877
number_of_kids                     -0.0126412496
mother_educ                         0.0396514364
factor_educ                         0.0332970434
mother_age30                       -0.0066400340
farmingIncome                       0.0121875354
boy:age                             .
factor_educ:age                     .
farmingIncome:age                   .
mother_age30:age                   -0.0117502693
mother_educ:age                    -0.0087160941
state:age                          -0.0017946149
boy:bl_caser_total_norm            -0.0103103377
factor_educ:bl_caser_total_norm    -0.0167799656
farmingIncome:bl_caser_total_norm -0.0033334836
mother_age30:bl_caser_total_norm    0.0002676178
mother_educ:bl_caser_total_norm     .
state:bl_caser_total_norm           0.0302192207
boy:number_of_kids                  .
factor_educ:number_of_kids         -0.0024424862
farmingIncome:number_of_kids       -0.0037319979
mother_age30:number_of_kids         .
mother_educ:number_of_kids          .
state:number_of_kids               -0.0128826236
t2                                  0.0129975670
t3                                  0.0256619453
t4                                  0.0556894608
```

Listing 9: LASSO Model with Interactions

Model with Interactions controlling for LASSO

```
Call:
lm(formula = caser_total_norm ~ ., data = selected_data %>% select(all_features %>%
    colnames()) %>% .[, non_zero_features] %>% cbind(selected_data %>%
    select(caser_total_norm)))

Residuals:
    Min      1Q  Median      3Q     Max
-3.1951 -0.2812 -0.0653  0.2289  3.1138

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           0.233349   0.007563  30.856  < 2e-16 ***
age                                   0.012288   0.002465   4.985 6.28e-07 ***
bl_caser_total_norm                   0.856703   0.004724 181.333  < 2e-16 ***
boy                                   0.052549   0.007604   6.911 5.02e-12 ***
number_of_kids                       -0.009203   0.002570  -3.581 0.000343 ***
mother_educ                           0.116805   0.011232  10.399  < 2e-16 ***
factor_educ                           0.070433   0.008030   8.771  < 2e-16 ***
mother_age30                         -0.017145   0.007881  -2.176 0.029599 *
farmingIncome                         0.028104   0.007600   3.698 0.000218 ***
`mother_age30:age`                   -0.017289   0.004740  -3.647 0.000266 ***
`mother_educ:age`                    -0.016632   0.005926  -2.807 0.005013 **
`state:age`                          -0.006574   0.004949  -1.328 0.184072
`boy:bl_caser_total_norm`            -0.024780   0.007481  -3.312 0.000927 ***
`factor_educ:bl_caser_total_norm`    -0.037068   0.008080  -4.588 4.52e-06 ***
`farmingIncome:bl_caser_total_norm` -0.010240   0.007865  -1.302 0.192914
`mother_age30:bl_caser_total_norm`    0.008967   0.008914   1.006 0.314423
`state:bl_caser_total_norm`           0.066593   0.009486   7.020 2.32e-12 ***
`factor_educ:number_of_kids`         -0.006854   0.005053  -1.357 0.174946
`farmingIncome:number_of_kids`       -0.009296   0.005347  -1.738 0.082151 .
`state:number_of_kids`               -0.021912   0.005491  -3.991 6.61e-05 ***
t2                                    0.012783   0.010482   1.220 0.222674
t3                                    0.025435   0.010616   2.396 0.016592 *
t4                                    0.055499   0.010519   5.276 1.34e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4501 on 14551 degrees of freedom
Multiple R-squared:  0.7978, Adjusted R-squared:  0.7975
F-statistic:  2610 on 22 and 14551 DF,  p-value: < 2.2e-16
```

Listing 10: Model with Interactions controlling for LASSO

Comparison

Model	$t_1 + t_2$	p-value t_2	Std Error t_2
W/out Cov	0.18733 + 0.05233 = 0.239659	0.024434	0.02325
W/ Dem Cov	0.220804 + 0.014434 = 0.235238	0.17057	0.010533
W/ Dem Cov Disc Interac	0.2334465 + 0.0126451 = 0.2460916	0.228106	0.0104912
LASSO W/ Dem Cov	0.220710 + 0.014645 = 0.235355	0.16431	0.010530
LASSO W/ Dem Cov Disc Interac	0.233349 + 0.012783 = 0.246132	0.222674	0.010482

Table 5: ATE for t_2

Model	$t_1 + t_3$	p-value t_3	Std Error t_3
W/out Cov	0.18733 + 0.07789 = 0.26522	1.66 e -05	0.02353
W/ Dem Cov	0.220804 + 0.025175 = 0.245979	0.01828	0.010667
W/ Dem Cov Disc Interac	0.2334465 + 0.0254664 = 0.2589129	0.016536	0.0106236
LASSO W/ Dem Cov	0.220710 + 0.025218 = 0.245928	0.01808	0.010667
LASSO W/ Dem Cov Disc Interac	0.233349 + 0.025435 = 0.258784	0.016592	0.010616

Table 6: ATE for t_3

Model	$t_1 + t_4$	p-value t_4	Std Error t_4
W/out Cov	0.18733 + 0.10038 = 0.28771	1.66 e -05	0.02331
W/ Dem Cov	0.220804 + 0.055961 = 0.276765	1.21 e -07	0.010569
W/ Dem Cov Disc Interac	0.2334465 + 0.0555816 = 0.2890281	1.30 e -07	0.0105252
LASSO W/ Dem Cov	0.220710 + 0.056082 = 0.276792	1.13 e -07	0.010568
LASSO W/ Dem Cov Disc Interac	0.233349 + 0.055499 = 0.288848	1.34 e -07	0.010519

Table 7: ATE for t_4

Considering standard error as proxy for precision, for t_2, t_3 and t_4, there is improvement in precision of the estimators by using controls selected by LASSO.

Specifically, we see that there is improvement using only demeaned covariates and when using demeaned covariates with intereaction.

The best model for all three ATE regarding standard error estimate is the model with demeaned covariates and interaction with covariates being selected by LASSO.

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).

Model with Interactions controlling for LASSO

```
Call:
lm(formula = caser_total_norm ~ ., data = banerji_data %>% select(t2,
    t3, t4, selected_discrete_variable, caser_total_norm) %>%
    purrr::set_names("t2", "t3", "t4", "selected_discrete_variable",
        "caser_total_norm") %>% mutate(across(c("t2", "t3", "t4"),
    ~. * selected_discrete_variable, .names = "{.col}:{selected_discrete_variable}"))
     %>%
    rename_with(~gsub("selected_discrete_variable", selected_discrete_variable,
        .)))

Residuals:
    Min      1Q  Median      3Q     Max
-1.8028 -0.8262 -0.2202  0.6946  2.2778

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.13164    0.01743   7.551 4.57e-14 ***
t2                 0.04140    0.02467   1.678   0.0934 .
t3                 0.04969    0.02519   1.972   0.0486 *
t4                 0.10044    0.02477   4.055 5.03e-05 ***
mother_educ        0.41335    0.04750   8.703  < 2e-16 ***
`t2:mother_educ`   0.08350    0.06729   1.241   0.2147
`t3:mother_educ`   0.09925    0.06541   1.518   0.1292
`t4:mother_educ`  -0.01516    0.06687  -0.227   0.8207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9869 on 14566 degrees of freedom
Multiple R-squared:  0.02718, Adjusted R-squared:  0.02671
F-statistic: 58.14 on 7 and 14566 DF,  p-value: < 2.2e-16
```

Listing 11: Model with Interactions controlling for LASSO

The estimators for the treatment are:

- $\mu_1(0) = \mathbb{E}[Y(1) \mid X = 0]$
$$\hat{\mu}_1(0) = \beta_0 = 0.13164$$
- $\mu_1(1) = \mathbb{E}[Y(1) \mid X = 1]$
$$\hat{\mu}_1(1) = \beta_0 + \beta_{\text{ME}} = 0.13164 + 0.41335 = 0.54499$$
- $\mu_2(0) = \mathbb{E}[Y(2) \mid X = 0]$
$$\hat{\mu}_2(0) = \beta_0 + \beta_{t_2} = 0.13164 + 0.04140 = 0.17304$$
- $\mu_2(1) = \mathbb{E}[Y(2) \mid X = 1]$
$$\hat{\mu}_2(1) = \beta_0 + \beta_{t_2} + \beta_{\text{ME}} + \beta_{\text{ME},t_2} = 0.13164 + 0.41335 + 0.04140 + 0.08350 = 0.66989$$
- $\mu_3(0) = \mathbb{E}[Y(3) \mid X = 0]$
$$\hat{\mu}_3(0) = \beta_0 + \beta_{t_3} = 0.13164 + 0.04969 = 0.18133$$
- $\mu_3(1) = \mathbb{E}[Y(3) \mid X = 1]$
$$\hat{\mu}_3(1) = \beta_0 + \beta_{t_3} + \beta_{\text{ME}} + \beta_{\text{ME},t_3} = 0.13164 + 0.41335 + 0.04969 + 0.09925 = 0.693930$$
- $\mu_4(0) = \mathbb{E}[Y(4) \mid X = 0]$
$$\hat{\mu}_4(0) = \beta_0 + \beta_{t_4} = 0.13164 + 0.10044 = 0.23208$$
- $\mu_4(1) = \mathbb{E}[Y(4) \mid X = 1]$
$$\hat{\mu}_4(1) = \beta_0 + \beta_{t_4} + \beta_{\text{ME}} + \beta_{\text{ME},t_4} = 0.13164 + 0.41335 + 0.10044 - 0.01516 = 0.63027$$

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?

```
[1] "Best lambda: 0.000502677636886109"
40 x 1 sparse Matrix of class "dgCMatrix"
                                           s1
(Intercept)                      2.082162e-01
age                              4.462536e-02
state                            1.562346e-02
bl_caser_total_norm              8.864615e-01
boy                              1.030960e-02
number_of_kids                   .
factor_educ                      3.959780e-02
mother_age30                     .
farmingIncome                    1.028717e-02
boy:age                          1.333859e-02
factor_educ:age                  .
farmingIncome:age                3.259882e-03
mother_age30:age                -1.148611e-02
state:age                        1.901602e-02
boy:bl_caser_total_norm         -8.192345e-03
factor_educ:bl_caser_total_norm -1.730713e-02
farmingIncome:bl_caser_total_norm -4.040476e-03
mother_age30:bl_caser_total_norm 5.834013e-03
state:bl_caser_total_norm        1.008217e-01
boy:number_of_kids               2.900538e-03
factor_educ:number_of_kids      -5.950095e-03
farmingIncome:number_of_kids    -3.256761e-04
mother_age30:number_of_kids      .
state:number_of_kids            -3.115408e-02
age:bl_caser_total_norm         -6.646937e-02
age:number_of_kids               .
bl_caser_total_norm:number_of_kids 2.036422e-02
age^2                           -1.983193e-04
bl_caser_total_norm^2           -1.408387e-01
number_of_kids^2                 4.607524e-05
age^3                           -5.922764e-02
bl_caser_total_norm^3            6.522308e-02
number_of_kids^3                 1.463515e-02
t2                               6.249520e-03
t3                               1.847526e-02
t4                               5.433661e-02
mother_educ                      9.763798e-02
t2:mother_educ                   3.958140e-02
t3:mother_educ                   2.775945e-02
t4:mother_educ                   8.679767e-03
```

Listing 12: LASSO for CATE

```
Call:
lm(formula = caser_total_norm ~ ., data = as_tibble(all_features) %>%
    .[, non_zero_features] %>% cbind(selected_data %>% select(caser_total_norm)))

Residuals:
    Min      1Q  Median      3Q     Max
-3.0586 -0.2659 -0.0567  0.2295  3.2402

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           0.208270   0.007833  26.587  < 2e-16 ***
age                                   0.504941   0.072963   6.920 4.69e-12 ***
state                                 0.025152   0.020941   1.201  0.22973
bl_caser_total_norm                   0.873575   0.028988  30.136  < 2e-16 ***
boy                                   0.009539   0.017786   0.536  0.59175
factor_educ                           0.053964   0.010594   5.094 3.56e-07 ***
farmingIncome                         0.024912   0.018485   1.348  0.17779
`boy:age`                             0.016003   0.016868   0.949  0.34280
`farmingIncome:age`                   0.007598   0.017426   0.436  0.66284
`mother_age30:age`                   -0.012303   0.004341  -2.834  0.00460 **
`state:age`                           0.041751   0.028233   1.479  0.13922
`boy:bl_caser_total_norm`            -0.013141   0.006775  -1.939  0.05246 .
`factor_educ:bl_caser_total_norm`    -0.022588   0.006391  -3.534  0.00041 ***
`farmingIncome:bl_caser_total_norm`  -0.008421   0.006796  -1.239  0.21534
`mother_age30:bl_caser_total_norm`    0.007703   0.005355   1.438  0.15034
`state:bl_caser_total_norm`           0.097842   0.015695   6.234 4.68e-10 ***
`boy:number_of_kids`                  0.002086   0.010578   0.197  0.84366
`factor_educ:number_of_kids`         -0.020477   0.011029  -1.857  0.06339 .
`farmingIncome:number_of_kids`       -0.021116   0.012049  -1.753  0.07971 .
`state:number_of_kids`               -0.068763   0.017570  -3.914 9.14e-05 ***
`age:bl_caser_total_norm`            -0.045611   0.024053  -1.896  0.05795 .
`bl_caser_total_norm:number_of_kids`  0.020335   0.010434   1.949  0.05131 .
`age^2`                              -0.884192   0.135085  -6.545 6.13e-11 ***
`bl_caser_total_norm^2`              -0.143036   0.012749 -11.219  < 2e-16 ***
`number_of_kids^2`                    0.080684   0.033884   2.381  0.01727 *
`age^3`                               0.365861   0.065426   5.592 2.29e-08 ***
`bl_caser_total_norm^3`               0.072101   0.014852   4.854 1.22e-06 ***
`number_of_kids^3`                   -0.027417   0.021997  -1.246  0.21265
t2                                    0.006672   0.011059   0.603  0.54633
t3                                    0.018026   0.011307   1.594  0.11091
t4                                    0.054896   0.011111   4.941 7.87e-07 ***
mother_educ                           0.097025   0.021555   4.501 6.81e-06 ***
`t2:mother_educ`                      0.038410   0.030138   1.274  0.20252
`t3:mother_educ`                      0.027552   0.029308   0.940  0.34719
`t4:mother_educ`                      0.007047   0.029940   0.235  0.81393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4416 on 14539 degrees of freedom
Multiple R-squared:  0.8056, Adjusted R-squared:  0.8051
F-statistic:  1772 on 34 and 14539 DF,  p-value: < 2.2e-16
```

Listing 13: CATE with Controls selected by LASSO

Parameter	Std. Error Model Without Cov.	Std. Error Model with LASSO Cov.
t_2	0.0934	0.54633
t_3	0.0486	0.11091
t_4	5.03 e -05	7.87 e -07
t_2 : `mother_educ`	0.2147	0.20252
t_3 : `mother_educ`	0.1292	0.34719
t_4 : `mother_educ`	0.8207	0.81393

Table 8: ATE for t_2

Parameter	Std. Error Model Without Cov.	Std. Error Model with LASSO Cov.
t_2	0.02467	0.011059
t_3	0.02519	0.011307
t_4	0.02477	0.011111
t_2 : `mother_educ`	0.06729	0.030138
t_3 : `mother_educ`	0.06541	0.029308
t_4 : `mother_educ`	0.06687	0.029940

Table 9: ATE for t_2

We see improvement in precision after adding covariates for all parameters (using Standard Error) as proxy for precision.

For t_4 and t_4 : `mother_educ` there is improvement even in p-value.

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.

```
100 x 1 sparse Matrix of class "dgCMatrix"
                                                         s1
(Intercept)                                    0.2144797531
age                                            0.0281165901
bl_caser_total_norm                            0.8414957100
number_of_kids                                 0.0031073010
factor_educ                                    0.0121030944
farmingIncome                                 -0.0002860080
age:bl_caser_total_norm                       -0.0431832017
age:boy                                        0.0185544106
state:bl_caser_total_norm                      0.1908771300
state:factor_educ                              0.0124303091
state:farmingIncome                            0.0007226657
bl_caser_total_norm:mother_age30              -0.0001074328
boy:number_of_kids                             0.0120670221
number_of_kids:mother_age30                    0.0047433751
age:state:bl_caser_total_norm                 -0.1160607119
age:state:number_of_kids                      -0.0364662876
age:state:factor_educ                          0.0262208697
age:state:farmingIncome                        0.0230873583
age:bl_caser_total_norm:boy                   -0.0279366353
age:bl_caser_total_norm:factor_educ           -0.0657800701
age:bl_caser_total_norm:mother_age30          -0.0216465918
age:bl_caser_total_norm:farmingIncome         -0.0101131068
age:boy:number_of_kids                         0.0000133068
age:mother_age30:farmingIncome                -0.0288624155
state:bl_caser_total_norm:boy                  0.0146324285
state:bl_caser_total_norm:number_of_kids       0.0195089515
state:bl_caser_total_norm:factor_educ          0.0213374333
state:bl_caser_total_norm:mother_age30        -0.0042999453
state:boy:farmingIncome                        0.0145076285
state:factor_educ:mother_age30                -0.0035105213
state:factor_educ:farmingIncome                0.0012287151
state:mother_age30:farmingIncome               0.0243074119
bl_caser_total_norm:boy:factor_educ            0.0066880420
bl_caser_total_norm:number_of_kids:factor_educ 0.0173389292
bl_caser_total_norm:factor_educ:farmingIncome -0.0038277620
boy:factor_educ:farmingIncome                 -0.0128410264
boy:mother_age30:farmingIncome                -0.0085520226
number_of_kids:mother_age30:farmingIncome     -0.0036848997
t2                                            -0.0040694660
t3                                             0.0139670529
t4                                             0.0498399624
mother_educ                                    0.0821242381
t2:mother_educ                                 0.0572264862
t3:mother_educ                                 0.0223986859
t4:mother_educ                                 0.0172743524
```

Listing 14: CATE with Controls selected by LASSO - Only Showing Non-Zero Variables

We keep 47% of the features (48 out of 99), and the features related to treatment.

```
Call:
lm(formula = caser_total_norm ~ ., data = sample_b %>% .[, non_zero_features] %>%
    cbind(sample_b %>% select(caser_total_norm)))

Residuals:
    Min      1Q  Median      3Q     Max
-3.0048 -0.2661 -0.0586  0.2327  2.6275

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        0.2033015  0.0111148  18.291  < 2e-16 ***
age                                                0.0353443  0.0137153   2.577 0.009986 **
bl_caser_total_norm                                0.9518833  0.0794556  11.980  < 2e-16 ***
number_of_kids                                     0.0228890  0.0115025   1.990 0.046638 *
factor_educ                                       -0.0139027  0.0167671  -0.829 0.407039
farmingIncome                                      0.0088017  0.0197687   0.445 0.656165
`age:bl_caser_total_norm`                         -0.0913389  0.0941372  -0.970 0.331942
`age:boy`                                          0.0572287  0.0165222   3.464 0.000536 ***
`state:bl_caser_total_norm`                        0.1933415  0.0829593   2.331 0.019804 *
`state:factor_educ`                                0.0556608  0.0275777   2.018 0.043594 *
`state:farmingIncome`                              0.0070487  0.0357285   0.197 0.843609
`bl_caser_total_norm:mother_age30`                -0.0981152  0.0437718  -2.242 0.025023 *
`boy:number_of_kids`                              -0.0272818  0.0225362  -1.211 0.226096
`number_of_kids:mother_age30`                     -0.0074939  0.0102350  -0.732 0.464079
`age:state:bl_caser_total_norm`                   -0.1703170  0.0798217  -2.134 0.032899 *
`age:state:number_of_kids`                        -0.0397825  0.0150788  -2.638 0.008350 **
`age:state:factor_educ`                            0.0096676  0.0233021   0.415 0.678241
`age:state:farmingIncome`                         -0.0123397  0.0260546  -0.474 0.635792
`age:bl_caser_total_norm:boy`                     -0.0519926  0.0258961  -2.008 0.044708 *
`age:bl_caser_total_norm:factor_educ`             -0.0536542  0.0294783  -1.820 0.068781 .
`age:bl_caser_total_norm:mother_age30`             0.0844550  0.0339642   2.487 0.012920 *
`age:bl_caser_total_norm:farmingIncome`           -0.0227074  0.0164834  -1.378 0.168373
`age:boy:number_of_kids`                           0.0019349  0.0264597   0.073 0.941708
`age:mother_age30:farmingIncome`                  -0.0198906  0.0214441  -0.928 0.353669
`state:bl_caser_total_norm:boy`                    0.0239092  0.0229671   1.041 0.297901
`state:bl_caser_total_norm:number_of_kids`         0.0324951  0.0279518   1.163 0.245054
`state:bl_caser_total_norm:factor_educ`            0.0261915  0.0260080   1.007 0.313941
`state:bl_caser_total_norm:mother_age30`           0.0104609  0.0263450   0.397 0.691326
`state:boy:farmingIncome`                          0.0182192  0.0120889   1.507 0.131829
`state:factor_educ:mother_age30`                  -0.0189089  0.0089674  -2.109 0.035010 *
`state:factor_educ:farmingIncome`                  0.0001149  0.0110288   0.010 0.991689
`state:mother_age30:farmingIncome`                 0.0192365  0.0193029   0.997 0.319011
`bl_caser_total_norm:boy:factor_educ`             -0.0155427  0.0122739  -1.266 0.205440
`bl_caser_total_norm:boy:mother_age30`            -0.0022791  0.0112973  -0.202 0.840129
`bl_caser_total_norm:number_of_kids:factor_educ`  -0.0104403  0.0196029  -0.533 0.594334
`bl_caser_total_norm:number_of_kids:mother_age30`  0.0107413  0.0217029   0.495 0.620669
`bl_caser_total_norm:factor_educ:farmingIncome`   -0.0025192  0.0125900  -0.200 0.841410
`bl_caser_total_norm:mother_age30:farmingIncome`   0.0253755  0.0118507   2.141 0.032286 *
`boy:factor_educ:farmingIncome`                   -0.0045536  0.0101141  -0.450 0.652564
`boy:mother_age30:farmingIncome`                   0.0009734  0.0090861   0.107 0.914688
`number_of_kids:mother_age30:farmingIncome`       -0.0097655  0.0158369  -0.617 0.537497
t2                                                 0.0141120  0.0156937   0.899 0.368569
t3                                                 0.0282517  0.0160594   1.759 0.078587 .
t4                                                 0.0583243  0.0158014   3.691 0.000225 ***
mother_educ                                        0.1043280  0.0300798   3.468 0.000527 ***
`t2:mother_educ`                                   0.0139198  0.0424748   0.328 0.743133
`t3:mother_educ`                                   0.0276729  0.0417285   0.663 0.507246
`t4:mother_educ`                                  -0.0036344  0.0418646  -0.087 0.930823
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4432 on 7239 degrees of freedom
Multiple R-squared:  0.802, Adjusted R-squared:  0.8008
F-statistic:   624 on 47 and 7239 DF,  p-value: < 2.2e-16
```

Listing 15: CATE with Controls selected by LASSO

```
Call:
lm(formula = caser_total_norm ~ ., data = sample_a %>% .[, non_zero_features] %>%
    cbind(sample_a %>% select(caser_total_norm)))

Residuals:
     Min       1Q   Median       3Q      Max
-2.93306 -0.27227 -0.06359  0.22800  2.98463

Coefficients:
                                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        0.214795   0.011217  19.150  < 2e-16 ***
age                                                0.048690   0.014336   3.396 0.000686 ***
bl_caser_total_norm                                0.931734   0.078250  11.907  < 2e-16 ***
number_of_kids                                     0.024298   0.011893   2.043 0.041075 *
factor_educ                                        0.007698   0.017057   0.451 0.651776
farmingIncome                                     -0.019565   0.019813  -0.987 0.323450
`age:bl_caser_total_norm`                         -0.144175   0.093398  -1.544 0.122714
`age:boy`                                          0.021883   0.017005   1.287 0.198179
`state:bl_caser_total_norm`                        0.211926   0.081203   2.610 0.009077 **
`state:factor_educ`                                0.018981   0.028304   0.671 0.502479
`state:farmingIncome`                              0.005286   0.035813   0.148 0.882666
`bl_caser_total_norm:mother_age30`                -0.106590   0.044475  -2.397 0.016571 *
`boy:number_of_kids`                               0.008049   0.022604   0.356 0.721771
`number_of_kids:mother_age30`                      0.009692   0.010404   0.932 0.351603
`age:state:bl_caser_total_norm`                   -0.115286   0.078902  -1.461 0.144027
`age:state:number_of_kids`                        -0.074388   0.015713  -4.734 2.24e-06 ***
`age:state:factor_educ`                            0.030671   0.024028   1.277 0.201817
`age:state:farmingIncome`                          0.034374   0.026194   1.312 0.189467
`age:bl_caser_total_norm:boy`                     -0.039625   0.025441  -1.558 0.119380
`age:bl_caser_total_norm:factor_educ`             -0.066516   0.030183  -2.204 0.027572 *
`age:bl_caser_total_norm:mother_age30`             0.052568   0.034138   1.540 0.123633
`age:bl_caser_total_norm:farmingIncome`           -0.012081   0.016307  -0.741 0.458795
`age:boy:number_of_kids`                           0.004044   0.026598   0.152 0.879154
`age:mother_age30:farmingIncome`                  -0.057466   0.021356  -2.691 0.007144 **
`state:bl_caser_total_norm:boy`                    0.020592   0.022147   0.930 0.352516
`state:bl_caser_total_norm:number_of_kids`         0.004860   0.027344   0.178 0.858936
`state:bl_caser_total_norm:factor_educ`            0.019396   0.025230   0.769 0.442066
`state:bl_caser_total_norm:mother_age30`          -0.005029   0.025956  -0.194 0.846382
`state:boy:farmingIncome`                          0.018288   0.011921   1.534 0.125058
`state:factor_educ:mother_age30`                  -0.014249   0.008992  -1.585 0.113079
`state:factor_educ:farmingIncome`                  0.004510   0.011033   0.409 0.682756
`state:mother_age30:farmingIncome`                 0.057053   0.019041   2.996 0.002743 **
`bl_caser_total_norm:boy:factor_educ`              0.010406   0.011979   0.869 0.385064
`bl_caser_total_norm:boy:mother_age30`            -0.005557   0.011207  -0.496 0.619988
`bl_caser_total_norm:number_of_kids:factor_educ`   0.021801   0.019380   1.125 0.260659
`bl_caser_total_norm:number_of_kids:mother_age30`  0.065252   0.021045   3.101 0.001938 **
`bl_caser_total_norm:factor_educ:farmingIncome`   -0.013234   0.012416  -1.066 0.286514
`bl_caser_total_norm:mother_age30:farmingIncome`   0.023015   0.011616   1.981 0.047602 *
`boy:factor_educ:farmingIncome`                   -0.014610   0.010094  -1.447 0.147822
`boy:mother_age30:farmingIncome`                  -0.013009   0.008949  -1.454 0.146049
`number_of_kids:mother_age30:farmingIncome`       -0.012496   0.015883  -0.787 0.431447
t2                                                -0.004855   0.015838  -0.307 0.759182
t3                                                 0.015108   0.016170   0.934 0.350146
t4                                                 0.051222   0.015882   3.225 0.001265 **
mother_educ                                        0.079796   0.031758   2.513 0.012006 *
`t2:mother_educ`                                   0.055800   0.043576   1.281 0.200407
`t3:mother_educ`                                   0.017537   0.042092   0.417 0.676958
`t4:mother_educ`                                   0.015046   0.043660   0.345 0.730384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4458 on 7239 degrees of freedom
Multiple R-squared:  0.8057, Adjusted R-squared:  0.8044
F-statistic: 638.5 on 47 and 7239 DF,  p-value: < 2.2e-16
```

Listing 16: CATE with Controls selected by LASSO

Parameter	Std Error LASSO Sample A	Std Error LASSO Sample B
t_2	0.015838	0.0156937
t_3	0.016170	0.0160594
t_4	0.015882	0.0158014
t_2 : `mother_educ`	0.043576	0.0424748
t_3 : `mother_educ`	0.042092	0.0417285
t_4 : `mother_educ`	0.043660	0.0418646

Table 10: ATE for t_2

Parameter	Std Error LASSO Sample A	Std Error LASSO Sample B
Intercept	0.2033015	0.21479
t_2	0.0141120	-0.004855
t_3	0.0282517	0.015108
t_4	0.0583243	0.051222
β_{ME}	0.1043280	0.079796
t_2 : `mother_educ`	0.0139198	0.055800
t_3 : `mother_educ`	0.0276729	0.017537
t_4 : `mother_educ`	-0.0036344	0.015046

Table 11: ATE for t_2

Precision is similar between Sample A and Sample B (standard errors). Nonetheless, the estimate parameters are relatively different in the two samples. In t_4 : `mother_educ` and t_2, the parameters even have different signs.

Binsreg & Continuous Heterogeneity

3.f Binsreg

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?

We use `binsreg`:

- Without controls (left).

- With the controls identified from question (3.b) using LASSO. In it, we saw great improvement in the estimator precision (right).

For each continous variable, we provide the `binsreg` for each of the treatment.

- BL Caser Total Norm has very similar $\omega_t(w_j) = \mathbb{E}[Y(t) \mid W_j = w_j]$ for all t.
- number of kids with and without control has more heterogeneous impact on the target when number of kids is larger for all t.
- Number of kids has more homogeneous impact on the target for t_3 and t_4 when focusing on the lower variables of the feature number of kids than t_1 and t_2.
- The impact of age seems to follow a quadratic form, meaning that fitting a model with $\texttt{age}^2$ and `age` should be beneficial. The quadratic effect is smaller when controlling for selected covariates, but still visible.

REMOVE COMMENT

3.g Confidence Interval in Binsreg

(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.

Again, We use `binsreg` with CI:

- Without controls (left).
- With the controls identified from question (3.b) using LASSO. In it, we saw great improvement in the estimator precision (right).

- We again find that age has a non-linear function with respect to the target and controlling and not controlling for covariates.
- The effect for all the continous covariates is more heterogeneous is the upper quantile of their distribution. It is particularly visible for age and number of kids controlling or not for covariates.
- BL Caser Total Norm is almost always monotonic increasing with the exception of a point in the upper quantile in which we see an heterogeneous uncertain effect.
- Controlling for covariates makes the effect of number of kids more homogeneous vis-a-vis not controlling. The opposite is true for age.

REMOVE COMMENT

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.

To identify the functions $\mu_t(x,w) = \mathbb{E}[Y(t) \mid X = x, W = w]$, the following conditions must be satisfied:

1. Conditional Independence (Unconfoundedness): The potential outcomes $Y_i(t)$ are independent of the treatment assignment T_i given the covariates $X_i = x$ and $W_i = w$. Formally,

$$Y_i(t) \perp T_i \mid X_i = x, W_i = w, \quad \text{for all } t \in \{1,2,3,4\}.$$

   This assumption ensures that, conditional on X_i and W_i, the treatment assignment is as good as random and there are no unobserved confounders affecting both the treatment and the outcome.

2. Positivity (Overlap): For all values of x and w in the support of X_i and W_i, there is a positive probability of receiving each treatment level. That is,

$$0 < P(T_i = t \mid X_i = x, W_i = w) < 1, \quad \text{for all } t \in \{1,2,3,4\}.$$

   This condition guarantees that there is sufficient variation in treatment assignments across all covariate patterns to estimate $\mu_t(x,w)$.

3. Consistency: The observed outcome corresponds to the potential outcome under the received treatment. Formally,

$$Y_i = Y_i(t) \quad \text{if } T_i = t.$$

   This assumption implies that there are no interference or spillover effects between units and that the treatment is well-defined. This is also referred to as the Stable Unit Treatment Value Assumption (SUTVA).

4. Correct Model Specification: The functional form of $\mu_t(x,w)$ correctly captures the relationship between the covariates and the potential outcomes. Additionally, the error term ϵ_i satisfies the following condition:

$$\mathbb{E}[\epsilon_i \mid T_i, X_i, W_i] = 0.$$

   This ensures that there are no systematic errors in the model and that all relevant covariates are included.

5. Measurability and Integrability: The functions $\mu_t(x,w)$ must be measurable and integrable with respect to the joint distribution of X_i and W_i. This ensures that the expectations are well-defined and finite.

Under these conditions, the functions $\mu_t(x, w)$ are identified because the conditional expectation of the observed outcomes equals the conditional expectation of the potential outcomes:

$$\mu_t(x, w) = \mathbb{E}[Y_i \mid T_i = t, X_i = x, W_i = w].$$

This equality allows us to estimate $\mu_t(x, w)$ directly from the observed data, facilitating the assessment of treatment effects across different covariate profiles.

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).
