

DS-GA 3001: Applied Statistics (Fall 2023-24)
Practice Final

Instructions:

- You have **110 minutes**, 4:00PM - 5:50PM
- The exam has 3 problems, totaling 100 points (+5 bonus points).
- Please answer each problem in the space below it.
- You are allowed to carry the textbook, your own notes and other course related material with you. Electronic devices are not allowed.
- Please read the problems carefully.
- Unless otherwise specified, you are required to provide explanations of how you arrived at your answers.
- You can use previous parts of a problem even if you did not solve them.
- The problems may not be arranged in an increasing order of difficulty. If you get stuck, it might be wise to try other problems first.
- Good luck and enjoy!

Full name: _____

N number: _____

1. Short questions. (40 points)

Provide a short answer to each of the questions. Each question is worth 10 points.

- (a) Consider the potential outcome model with observations (X, W, Y) and potential outcomes $(Y(1), Y(0))$, where $\mathbb{E}[W \mid X = x] = e(x)$ and $\mathbb{E}[Y(1) \mid X = x] = \mu_1(x)$. The following chain of equations holds:

$$\begin{aligned}\mathbb{E}[YW] &= \mathbb{E}\{\mathbb{E}[YW \mid X]\} \\ &\stackrel{(1)}{=} \mathbb{E}\{\mathbb{E}[Y(1)W \mid X]\} \\ &\stackrel{(2)}{=} \mathbb{E}\{\mathbb{E}[Y(1) \mid X]\mathbb{E}[W \mid X]\} \\ &= \mathbb{E}[\mu_1(X)e(X)].\end{aligned}$$

Justify the steps (1) and (2), by providing the assumptions used (SUTVA, unconfoundedness, etc.) and/or the mathematical reasoning behind them.

- (b) In linear regression with endogeneity, one has the regression model $Y = \beta X + \varepsilon$, while with $\mathbb{E}[X\varepsilon] \neq 0$. A common way to estimate β in this scenario is to find an *instrumental variable* Z such that $\mathbb{E}[Z\varepsilon] = 0$ and $\mathbb{E}[ZX] \neq 0$.

Show that for such a Z , the function $f_\beta(X, Y, Z) = Z(Y - \beta X)$ is an estimating function. Explain why we need $\mathbb{E}[ZX] \neq 0$ when using $f_\beta(X, Y, Z)$ to estimate β .

- (c) Consider the nonparametric regression problem with a uniform grid $x_i = i/n$. An estimator \hat{f} is a mapping from the observations (y_1, \dots, y_n) to a function, and it is called *linear* if for any $(y_1, \dots, y_n), (y'_1, \dots, y'_n)$ and $\alpha, \beta \in \mathbb{R}$,

$$\hat{f}(\alpha y_1 + \beta y'_1, \dots, \alpha y_n + \beta y'_n) = \alpha \hat{f}(y_1, \dots, y_n) + \beta \hat{f}(y'_1, \dots, y'_n).$$

In other words, the estimator \hat{f} is a linear function of (y_1, \dots, y_n) .

Below we list several estimators covered in class. Which of the following are *linear* estimators?

- i. the Nadaraya–Watson estimator (with fixed K, h);
- ii. the local polynomial regression (with fixed k, K, h);
- iii. the cubic smoothing spline regression (with fixed λ);
- iv. the Fourier projection estimator (with fixed m);
- v. the wavelet soft-thresholding estimator (with fixed threshold t).

Write L (Linear) or N (Nonlinear) for each estimator, without explanations.

(d) Consider the Haar wavelet discussed in class, with father and mother wavelets

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad \psi(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1/2, \\ -1 & \text{if } 1/2 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Write down the expressions of $\phi_{1,0}(x)$ and $\psi_{2,1}(x)$. Verify that they are orthonormal on $[0, 1]$:

$$\int_0^1 \phi_{1,0}(x)^2 dx = \int_0^1 \psi_{2,1}(x)^2 dx = 1, \quad \int_0^1 \phi_{1,0}(x) \psi_{2,1}(x) dx = 0.$$

2. Causal inference with discrete covariates. (30 points + 5 bonus points)

Consider the following setting of a potential outcome model: let $X \in \{1, 2, \dots, K\}$ be a discrete covariate with $\mathbb{P}(X = k) = p_k$, $W \in \{0, 1\}$ be a binary indicator of treatment with $\mathbb{E}[W \mid X = k] = e_k$, and Y be the observed outcome. Here the potential outcomes are assumed to be binary, i.e. $Y \in \{0, 1\}$, with

$$\begin{aligned}\mathbb{P}(Y = 1 \mid X = k, W = 1) &= \mu_{1,k}, \\ \mathbb{P}(Y = 1 \mid X = k, W = 0) &= \mu_{0,k}.\end{aligned}$$

The learner is given a dataset $\{(X_i, W_i, Y_i)\}_{i=1}^n$.

(a) Based on the dataset, a natural estimator for p_k is the empirical distribution

$$\hat{p}_k = \frac{\#\{1 \leq i \leq n : X_i = k\}}{n}.$$

Using the definition of $(e_k, \mu_{1,k}, \mu_{0,k})$ and the plug-in approach, justify the following estimators for them:

$$\begin{aligned}\hat{e}_k &= \frac{\#\{1 \leq i \leq n : X_i = k, W_i = 1\}}{\#\{1 \leq i \leq n : X_i = k\}}, \\ \hat{\mu}_{1,k} &= \frac{\#\{1 \leq i \leq n : X_i = k, W_i = 1, Y_i = 1\}}{\#\{1 \leq i \leq n : X_i = k, W_i = 1\}}, \\ \hat{\mu}_{0,k} &= \frac{\#\{1 \leq i \leq n : X_i = k, W_i = 0, Y_i = 1\}}{\#\{1 \leq i \leq n : X_i = k, W_i = 0\}}.\end{aligned}$$

We assume that the denominators are always non-zero. (10 points)

(b) Suppose that the target is to estimate the average treatment effect

$$\tau = \mathbb{E}[\mu_{1,X} - \mu_{0,X}] = \sum_{k=1}^K p_k(\mu_{1,k} - \mu_{0,k}).$$

A natural estimator for τ is based on outcome regression:

$$\hat{\tau}_R = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{1,X_i} - \hat{\mu}_{0,X_i}),$$

where $(\hat{e}_k, \hat{\mu}_{1,k}, \hat{\mu}_{0,k})$ are defined in (a). Show that

$$\hat{\tau}_R = \sum_{k=1}^K \hat{p}_k (\hat{\mu}_{1,k} - \hat{\mu}_{0,k}).$$

(10 points: hint: $\hat{\mu}_{1,X_i} = \sum_{k=1}^K \mathbb{1}(X_i = k) \hat{\mu}_{1,k}$.)

(c) Another estimator for τ is the IPW estimator:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i W_i}{\hat{e}_{X_i}} - \frac{Y_i (1 - W_i)}{1 - \hat{e}_{X_i}} \right).$$

Show that this estimator is identical to the regression estimator in (b), i.e. $\hat{\tau}_{\text{R}} = \hat{\tau}_{\text{IPW}}$. (10 points)

(d) The double robust estimator for τ is given by

$$\hat{\tau}_{\text{DR}} = \frac{1}{n} \sum_{i=1}^n \left(\frac{W_i(Y_i - \hat{\mu}_{1,X_i})}{\hat{e}_{X_i}} + \hat{\mu}_{1,X_i} - \frac{(1 - W_i)(Y_i - \hat{\mu}_{0,X_i})}{1 - \hat{e}_{X_i}} - \hat{\mu}_{0,X_i} \right).$$

Show that this estimator is also identical to the previous estimators, i.e. $\hat{\tau}_{\text{DR}} = \hat{\tau}_{\text{R}}$.
(5 bonus points)

3. Optimal kernel and bandwidth. (30 points)

Consider the nonparametric regression problem (X, Y) with $X \sim \text{Unif}[0, 1]$, $\mathbb{E}[Y \mid X = x] = f(x)$, and $\text{Var}(Y \mid X = x) \equiv 1$. If f is twice continuously differentiable, in class we showed that solving the local linear regression

$$(\hat{\theta}_0, \hat{\theta}_1) = \arg \min_{(\theta_0, \theta_1)} \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2 \cdot \frac{1}{h} K\left(\frac{x_0 - x_i}{h}\right)$$

and estimating $f(x_0)$ by $\hat{f}(x_0) = \hat{\theta}_0 + \hat{\theta}_1 x_0$ achieves the MSE $O(h^4 + 1/(nh))$. A more accurate characterization of the MSE was obtained in Fan (1993): for large n ,

$$\begin{aligned} |\text{Bias}(\hat{f}(x_0))| &\approx \frac{|f''(x_0)|h^2}{2} \cdot \int_{-\infty}^{\infty} t^2 K(t) dt, \\ \text{Var}(\hat{f}(x_0)) &\approx \frac{1}{nh} \cdot \int_{-\infty}^{\infty} K(t)^2 dt. \end{aligned}$$

(a) Using these approximations, show that for fixed kernel K , choosing the bandwidth

$$h^* = \left(\frac{\int_{-\infty}^{\infty} K(t)^2 dt}{n f''(x_0)^2 (\int_{-\infty}^{\infty} t^2 K(t) dt)^2} \right)^{1/5}$$

minimizes the MSE of $\hat{f}(x_0)$, and the smallest MSE is

$$\frac{5 f''(x_0)^{2/5}}{4 n^{4/5}} \left(\int_{-\infty}^{\infty} t^2 K(t) dt \right)^{2/5} \left(\int_{-\infty}^{\infty} K(t)^2 dt \right)^{4/5}.$$

(10 points; hint: use first-order condition to find the minimum of $h \mapsto a^2 h^4 + b/h$.)

- (b) The smallest MSE in (a) also provides guidelines for how to choose the kernel K . Consider the Epanechnikov kernel

$$K^*(t) = \begin{cases} a(1 - t^2) & \text{if } |t| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here a is a normalization factor such that $\int_{-\infty}^{\infty} K^*(t) dt = 1$. Find a , and compute the values of

$$\int_{-\infty}^{\infty} t^2 K^*(t) dt \quad \text{and} \quad \int_{-\infty}^{\infty} K^*(t)^2 dt.$$

(10 points)

- (c) It turns out that the Epanechnikov kernel gives the smallest MSE, and we prove a weaker claim here. Let K be another kernel supported on $[-1, 1]$ (i.e. $K(t) = 0$ if $|t| > 1$), with

$$\int_{-1}^1 K(t)dt = 1, \quad \int_{-1}^1 t^2 K(t)dt = \frac{1}{5}.$$

Show that

$$\int_{-1}^1 K(t)^2 dt \geq \int_{-1}^1 K^*(t)^2 dt.$$

(10 points; hint: check that $\int_{-1}^1 (K(t) - K^*(t))K^*(t)dt = 0$.)