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

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:

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

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

Solution: Step (1) follows from $YW = Y(1)W$: if $W = 0$ both sides are zero, if $W = 1$ we have $Y = Y(1)$ by SUTVA.

Step (2) follows from unconfoundedness, i.e. $Y(1) \perp\!\!\!\perp W \mid X$, so that $\mathbb{E}[Y(1)W \mid X] = \mathbb{E}[Y(1) \mid X]\mathbb{E}[W \mid X]$.

(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.

Solution: Estimating function:

$$\mathbb{E}[f_\beta(X, Y, Z)] = \mathbb{E}[Z(Y - \beta X)] = \mathbb{E}[Z\varepsilon] = 0.$$

The idea of estimating β based on this function is to use

$$\beta = \frac{\mathbb{E}[ZY]}{\mathbb{E}[ZX]},$$

so we need $\mathbb{E}[ZX] \neq 0$ to ensure that the denominator is not zero.

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

$$\widehat{f}(\alpha y_1 + \beta y_1', \cdots, \alpha y_n + \beta y_n') = \alpha \widehat{f}(y_1, \cdots, y_n) + \beta \widehat{f}(y_1', \cdots, y_n').$$

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

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.

Solution:

i. L. The Nadaraya–Watson estimator takes the form $\widehat{f}(x_0) = \sum_{i=1}^{n} w(x_i, x_0) y_i$ for some weights independent of $(y_1, \cdots, y_n)$.
ii. L. It is equivalently a weighted least squares problem, and linear in $(y_1, \cdots, y_n)$.
iii. L. It is equivalently a ridge regression problem, and linear in $(y_1, \cdots, y_n)$.
iv. L. Both the Fourier transform and projection operation are linear in $(y_1, \cdots, y_n)$.
v. N. Although the wavelet transform is linear in $(y_1, \cdots, y_n)$, the thresholding operation applied to $(y_1, \cdots, y_n)$ is nonlinear.

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

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

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

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

Solution: Expressions:

$$\phi_{1,0}(x) = 2^{1/2}\phi(2x) = \begin{cases} \sqrt{2} & \text{if } 0 \le x \le 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

$$\psi_{2,1}(x) = 2^{2/2}\psi(4x - 1) = \begin{cases} 2 & \text{if } 1/4 \le x \le 3/8, \\ -2 & \text{if } 3/8 < x \le 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Orthonormality:

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

$$\int_0^1 \psi_{2,1}(x)^2 dx = \int_{1/4}^{1/2} 2^2 dx = 1,$$

$$\int_0^1 \phi_{1,0}(x)\psi_{2,1}(x) dx = \int_{1/4}^{3/8} 2\sqrt{2}dx + \int_{3/8}^{1/2} (-2\sqrt{2})dx = 0.$$

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

Consider the following setting of a potential outcome model: let $X \in \{1, 2, \cdots, K\}$ 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

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

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

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

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

$$\widehat{e}_k = \frac{\#\{1 \le i \le n : X_i = k, W_i = 1\}}{\#\{1 \le i \le n : X_i = k\}},$$
$$\widehat{\mu}_{1,k} = \frac{\#\{1 \le i \le n : X_i = k, W_i = 1, Y_i = 1\}}{\#\{1 \le i \le n : X_i = k, W_i = 1\}},$$
$$\widehat{\mu}_{0,k} = \frac{\#\{1 \le i \le n : X_i = k, W_i = 0, Y_i = 1\}}{\#\{1 \le i \le n : X_i = k, W_i = 0\}}.$$

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

Solution: For the propensity score e_k, we have

$$e_k = \mathbb{P}(W = 1 \mid X = k) = \frac{\mathbb{P}(W = 1, X = k)}{\mathbb{P}(X = k)}.$$

Note that natural estimators for $\mathbb{P}(W = 1, X = k)$ and $\mathbb{P}(X = k)$ are

$$\frac{\#\{1 \le i \le n : X_i = k, W_i = 1\}}{n} \quad \text{and} \quad \frac{\#\{1 \le i \le n : X_i = k\}}{n},$$

respectively, the plug-in approach then gives the target estimator $\widehat{e}_k$. The reasonings for the remaining estimators are entirely similar.

(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:

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

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

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

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

Solution: It holds that

$$\begin{aligned}
\widehat{\tau}_{\mathrm{R}} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}(X_i = k)(\widehat{\mu}_{1,k} - \widehat{\mu}_{0,k}) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \mathbb{1}(X_i = k)(\widehat{\mu}_{1,k} - \widehat{\mu}_{0,k}) \\
&= \frac{1}{n}\sum_{k=1}^{K} \#\{i : X_i = k\} \cdot (\widehat{\mu}_{1,k} - \widehat{\mu}_{0,k}) \\
&= \sum_{k=1}^{K} \widehat{p}_k(\widehat{\mu}_{1,k} - \widehat{\mu}_{0,k}).
\end{aligned}$$

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

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

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

Solution: It holds that

$$\begin{aligned}
\widehat{\tau}_{\mathrm{IPW}} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(X_i = k)\left(\frac{Y_i W_i}{\widehat{e}_k} - \frac{Y_i(1-W_i)}{1-\widehat{e}_k}\right) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{1}(X_i = k)\left(\frac{Y_i W_i}{\widehat{e}_k} - \frac{Y_i(1-W_i)}{1-\widehat{e}_k}\right) \\
&= \frac{1}{n}\sum_{k=1}^{K}\left(\frac{\#\{i : X_i = k, W_i = 1, Y_i = 1\}}{\widehat{e}_k} - \frac{\#\{i : X_i = k, W_i = 0, Y_i = 1\}}{1-\widehat{e}_k}\right) \\
&= \sum_{k=1}^{K}\widehat{p}_k\left(\frac{\#\{i : X_i = k, W_i = 1, Y_i = 1\}}{\#\{i : X_i = k, W_i = 1\}} - \frac{\#\{i : X_i = k, W_i = 0, Y_i = 1\}}{\#\{i : X_i = k\} - \#\{i : X_i = k, W_i = 1\}}\right) \\
&= \sum_{k=1}^{K}\widehat{p}_k\left(\frac{\#\{i : X_i = k, W_i = 1, Y_i = 1\}}{\#\{i : X_i = k, W_i = 1\}} - \frac{\#\{i : X_i = k, W_i = 0, Y_i = 1\}}{\#\{i : X_i = k, W_i = 0\}}\right) \\
&= \sum_{k=1}^{K}\widehat{p}_k(\widehat{\mu}_{1,k} - \widehat{\mu}_{0,k}).
\end{aligned}$$

By (b), we have $\widehat{\tau}_{\mathrm{R}} = \widehat{\tau}_{\mathrm{IPW}}$.

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

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

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

Solution: Since

$$\widehat{\tau}_{\mathrm{DR}} = \widehat{\tau}_{\mathrm{R}} + \widehat{\tau}_{\mathrm{IPW}} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i\widehat{\mu}_{1,X_i}}{\widehat{e}_{X_i}} - \frac{(1-W_i)\widehat{\mu}_{0,X_i}}{1-\widehat{e}_{X_i}}\right),$$

it suffices to prove that the last term is equal to $\widehat{\tau}_{\mathrm{R}}$. Indeed,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i\widehat{\mu}_{1,X_i}}{\widehat{e}_{X_i}} - \frac{(1-W_i)\widehat{\mu}_{0,X_i}}{1-\widehat{e}_{X_i}}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(X_i = k)\left(\frac{W_i\widehat{\mu}_{1,k}}{\widehat{e}_k} - \frac{(1-W_i)\widehat{\mu}_{0,k}}{1-\widehat{e}_k}\right)\\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{1}(X_i = k)\left(\frac{W_i\widehat{\mu}_{1,k}}{\widehat{e}_k} - \frac{(1-W_i)\widehat{\mu}_{0,k}}{1-\widehat{e}_k}\right)\\
&= \frac{1}{n}\sum_{k=1}^{K}\left(\#\{i : X_i = k, W_i = 1\}\frac{\widehat{\mu}_{1,k}}{\widehat{e}_k} - \#\{i : X_i = k, W_i = 0\}\frac{\widehat{\mu}_{0,k}}{1-\widehat{e}_k}\right)\\
&= \frac{1}{n}\sum_{k=1}^{K}\left(\#\{i : X_i = k\}\widehat{\mu}_{1,k} - \#\{i : X_i = k\}\widehat{\mu}_{0,k}\right)\\
&= \sum_{k=1}^{K}\widehat{p}_k(\widehat{\mu}_{1,k} - \widehat{\mu}_{0,k}),
\end{aligned}$$

so this equals $\widehat{\tau}_{\mathrm{R}}$ as desired.

3. Optimal kernel and bandwidth. *(30 points)*

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

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

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

$$|\mathsf{Bias}(\widehat{f}(x_0))| \approx \frac{|f''(x_0)| h^2}{2} \cdot \int_{-\infty}^{\infty} t^2 K(t) dt,$$
$$\mathsf{Var}(\widehat{f}(x_0)) \approx \frac{1}{nh} \cdot \int_{-\infty}^{\infty} K(t)^2 dt.$$

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

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

minimizes the MSE of $\widehat{f}(x_0)$, 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* $h \mapsto a^2 h^4 + b/h$.)

Solution: For $h \mapsto a^2 h^4 + b/h$, the first-order condition gives

$$4a^2 h^3 - \frac{b}{h^2} = 0 \Longrightarrow h = \left(\frac{b}{4a^2}\right)^{1/5} \Longrightarrow \text{Opt} = \frac{5}{4} (2a)^{2/5} b^{4/5}.$$

Since MSE $= \mathsf{Bias}^2 + \mathsf{Var}$, plugging

$$a = \frac{|f''(x_0)|}{2} \int_{-\infty}^{\infty} t^2 K(t) dt, \qquad b = \frac{1}{n} \int_{-\infty}^{\infty} K(t)^2 dt$$

into the above result gives the claimed answer.

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

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

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

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

(10 points)

Solution: The value of a:

$$1 = \int_{-\infty}^{\infty} K^\star(t)dt = \int_{-1}^{1} a(1-t^2)dt = a\left(t - \frac{t^3}{3}\right)\Big|_{t=-1}^{t=1} = \frac{4a}{3} \Longrightarrow a = \frac{3}{4}.$$

The other integrals:

$$\int_{-\infty}^{\infty} t^2 K^\star(t)dt = \int_{-1}^{1} \frac{3}{4}t^2(1-t^2)dt = \frac{3}{4}\left(\frac{t^3}{3} - \frac{t^5}{5}\right)\Big|_{t=-1}^{t=1} = \frac{1}{5},$$
$$\int_{-\infty}^{\infty} K^\star(t)^2 dt = \int_{-1}^{1} \frac{9}{16}(1-t^2)^2 dt = \frac{9}{16}\left(t - \frac{2t^3}{3} + \frac{t^5}{5}\right)\Big|_{t=-1}^{t=1} = \frac{3}{5}.$$

(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

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

Show that

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

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

Solution: First note that

$$\begin{aligned}
\int_{-1}^{1}(K(t)-K^\star(t))K^\star(t)dt &= \frac{3}{4}\int_{-1}^{1}\left[K(t)-\frac{3}{4}(1-t^2)\right](1-t^2)dt \\
&= \frac{3}{4}\left[\int_{-1}^{1}K(t)dt - \int_{-1}^{1}t^2K(t)dt - \int_{-1}^{1}\frac{3(1-t^2)^2}{4}dt\right] \\
&= \frac{3}{4}\left[1-\frac{1}{5}-\frac{3}{4}\cdot\frac{16}{15}\right] = 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int_{-1}^{1} K(t)^2 dt &= \int_{-1}^{1}[K^\star(t)+(K(t)-K^\star(t))]^2 dt \\
&= \int_{-1}^{1}[K^\star(t)^2 + 2(K(t)-K^\star(t))K^\star(t) + (K(t)-K^\star(t))^2]dt \\
&= \int_{-1}^{1} K^\star(t)^2 dt + \underbrace{2\int_{-1}^{1}(K(t)-K^\star(t))K^\star(t)dt}_{=0} + \underbrace{\int_{-1}^{1}(K(t)-K^\star(t))^2 dt}_{\geq 0} \\
&\geq \int_{-1}^{1} K^\star(t)^2 dt.
\end{aligned}$$