DS-GA 3001: Applied Statistics (Fall 2023-24)
Final, Thursday December 21st
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) In semiparametric statistics, let $s^{\theta}_{(\theta_0,\eta_0)}(y)$ and $s^{\eta}_{(\theta_0,\eta_0)}(y)$ be the score functions for the target parameter θ and nuisance parameter η, respectively. Write down the definition of the efficient score function $s^{\text{eff}}_{(\theta_0,\eta_0)}(y)$ for θ, and show that

$$\mathbb{E}_{(\theta_0,\eta_0)}[s^{\text{eff}}_{(\theta_0,\eta_0)}(y)] = 0.$$

Solution: The efficient score function is given by

$$s^{\text{eff}}_{(\theta_0,\eta_0)}(y) = s^{\theta}_{(\theta_0,\eta_0)}(y) - \alpha \cdot s^{\eta}_{(\theta_0,\eta_0)}(y),$$

where

$$\alpha = \frac{\mathbb{E}_{(\theta_0,\eta_0)}[s^{\theta}_{(\theta_0,\eta_0)}(y)s^{\eta}_{(\theta_0,\eta_0)}(y)]}{\mathbb{E}_{(\theta_0,\eta_0)}[s^{\eta}_{(\theta_0,\eta_0)}(y)^2]}.$$

Since score functions have zero mean, we have

$$\mathbb{E}_{(\theta_0,\eta_0)}[s^{\text{eff}}_{(\theta_0,\eta_0)}(y)] = \mathbb{E}_{(\theta_0,\eta_0)}[s^{\theta}_{(\theta_0,\eta_0)}(y)] - \alpha \cdot \mathbb{E}_{(\theta_0,\eta_0)}[s^{\eta}_{(\theta_0,\eta_0)}(y)] = 0.$$

(b) Explain the *unconfoundedness* in the potential outcome model of causal inference. Propose a real-life scenario where this assumption is violated.

Solution: Unconfoundedness assumption:

$$(Y(0), Y(1)) \perp\!\!\!\perp W \mid X.$$

An example where this assumption is violated could be in a study examining the effect of a training program on future earnings. If individuals who choose to participate in the training program are more motivated and have better unmeasured job-seeking skills than those who do not participate, and these characteristics also affect earnings, then the assumption of unconfoundedness is violated. This is because there are unmeasured confounders (motivation and job-seeking skills) that affect both the treatment (participation in the training program) and the outcome (future earnings), which are not accounted for in the analysis.

(c) Consider the nonparametric regression problem on $[0, 1]$, and below we list several estimators covered in class. Which of the following operations will *increase* the bias (and consequently *decrease* the variance)?

i. increase the bandwidth h in the Nadaraya–Watson estimator;
ii. increase the polynomial degree k in the local polynomial regression;
iii. increase the regularization parameter λ in cubic smoothing spline regression;
iv. increase the number of kept terms m in the Fourier projection estimator;
v. increase the threshold t in the wavelet soft-thresholding estimator.

Write Y (Yes) or N (No) for each operation, without explanations.

Solution:

i. Y. This is the most classical example we covered in class.
ii. N. When k increases, the polynomial approximation error decreases, so the bias decreases.
iii. Y. Just like ridge regression, when the regularization parameter λ increases, the solution shrinks to zero and incurs a larger bias.
iv. N. The bias-variance tradeoff covered in class is $O(m^{-2k} + m/n)$, so that the bias $O(m^{-2k})$ decreases with m.
v. Y. Increasing the threshold t makes the estimates shrink more to zero, and therefore incurs a larger bias.

(d) For a class of continuous functions $\{\phi_i(x)\}_{i\geq 1}$ on $[0,1]$, write down the definition of these functions being *orthonormal*. Show that if $\{\phi_i(x)\}_{i\geq 1}$ and $\{\psi_j(y)\}_{j\geq 1}$ are two orthonormal classes of functions on $[0,1]$, then the class of bivariate functions $\{f_{i,j}(x,y)=\phi_i(x)\psi_j(y)\}_{i,j\geq 1}$ are orthonormal on $[0,1]\times[0,1]$.

Solution: Definition of orthonormality:

$$\int_0^1 \phi_i(x)\phi_j(x)dx = \begin{cases} 1 & \text{if } i=j, \\ 0 & \text{if } i\neq j. \end{cases}$$

The orthonormality of $\{f_{i,j}(x,y)\}_{i,j\geq 1}$:

$$\begin{aligned}
&\iint_{[0,1]\times[0,1]} f_{i,j}(x,y)f_{k,\ell}(x,y)dxdy \\
&= \iint_{[0,1]\times[0,1]} \phi_i(x)\psi_j(y)\phi_k(x)\psi_\ell(y)dxdy \\
&= \left(\int_0^1 \phi_i(x)\phi_k(x)dx\right)\left(\int_0^1 \psi_j(y)\psi_\ell(y)dy\right) \\
&= \begin{cases} 1 & \text{if } i=k \text{ and } j=\ell, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

2. Estimation of causal functionals. *(30 points)*

Consider the following model for causal inference: let X be the covariate, $W \in \{0,1\}$ be the binary indicator of treatment with $\mathbb{E}[W \mid X = x] = e(x)$, and Y be the observed outcome with $\mathbb{E}[Y \mid X = x] = \mu(x)$. The target is to estimate the causal functional

$$\psi = \mathbb{E}[\mathsf{Cov}(W, Y \mid X)],$$

while treating $(e(x), \mu(x))$ as nuisance parameters.

Throughout this problem the following covariance definitions will be useful:

$$\begin{aligned}\mathsf{Cov}(W, Y \mid X) &= \mathbb{E}[WY \mid X] - \mathbb{E}[W \mid X]\mathbb{E}[Y \mid X] \\ &= \mathbb{E}[(W - \mathbb{E}[W \mid X])(Y - \mathbb{E}[Y \mid X]) \mid X].\end{aligned}$$

(a) Show that

$$f_{(\psi,e,\mu)}(X, W, Y) = WY - e(X)\mu(X) - \psi$$

is an estimating equation, i.e. $\mathbb{E}[f_{(\psi,e,\mu)}(X, W, Y)] = 0$. *(10 points)*

Solution: Use the first definition of the covariance,

$$\begin{aligned}\psi &= \mathbb{E}[\mathsf{Cov}(W, Y \mid X)] \\ &= \mathbb{E}[\mathbb{E}[WY \mid X] - \mathbb{E}[W \mid X]\mathbb{E}[Y \mid X]] \\ &= \mathbb{E}[\mathbb{E}[WY \mid X] - e(X)\mu(X)] \\ &= \mathbb{E}[WY] - \mathbb{E}[e(X)\mu(X)].\end{aligned}$$

This implies that

$$\mathbb{E}[f_{(\psi,e,\mu)}(X, W, Y)] = \mathbb{E}[WY - e(X)\mu(X)] - \mathbb{E}[WY - e(X)\mu(X)] = 0.$$

(b) Show that

$$g_{(\psi,e,\mu)}(X,W,Y) = (W - e(X))(Y - \mu(X)) - \psi$$

is also an estimating equation, i.e. $\mathbb{E}[g_{(\psi,e,\mu)}(X,W,Y)] = 0$. *(10 points)*

Solution: Use the second definition of the covariance,

$$\begin{aligned}
\psi &= \mathbb{E}[\mathsf{Cov}(W,Y \mid X)] \\
&= \mathbb{E}\left\{\mathbb{E}[(W - \mathbb{E}[W \mid X])(Y - \mathbb{E}[Y \mid X]) \mid X]\right\} \\
&= \mathbb{E}[(W - \mathbb{E}[W \mid X])(Y - \mathbb{E}[Y \mid X])] \\
&= \mathbb{E}[(W - e(X))(Y - \mu(X))].
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathbb{E}[g_{(\psi,e,\mu)}(X,W,Y)] &= \mathbb{E}[(W - e(X))(Y - \mu(X))] - \mathbb{E}[(W - e(X))(Y - \mu(X))] \\
&= 0.
\end{aligned}$$

(c) Show that $g_{(\psi,e,\mu)}$ is doubly robust, i.e. for any nuisance estimates $(\widehat{e}, \widehat{\mu})$, it holds that

$$\begin{aligned}\mathbb{E}[g_{(\psi,\widehat{e},\mu)}(X,W,Y)] = 0,\\ \mathbb{E}[g_{(\psi,e,\widehat{\mu})}(X,W,Y)] = 0.\end{aligned}$$

(10 points; hint: it might be easier to work on the difference $\mathbb{E}[g_{(\psi,\widehat{e},\mu)} - g_{(\psi,e,\mu)}]$*.)*

Solution: For the first identity, note that

$$\begin{aligned}\mathbb{E}[g_{(\psi,\widehat{e},\mu)} - g_{(\psi,e,\mu)}] &= \mathbb{E}[(e(X)-\widehat{e}(X))(Y-\mu(X))]\\ &= \mathbb{E}\left\{(e(X)-\widehat{e}(X))\cdot\mathbb{E}[(Y-\mu(X)) \mid X]\right\}\\ &= \mathbb{E}\left\{(e(X)-\widehat{e}(X))\cdot(\mathbb{E}[Y \mid X]-\mu(X))\right\}\\ &= 0,\end{aligned}$$

so the claim follows from (b). Similarly, for the second identity we have

$$\begin{aligned}\mathbb{E}[g_{(\psi,e,\mu)} - g_{(\psi,e,\widehat{\mu})}] &= \mathbb{E}[(W-e(X))(\mu(X)-\widehat{\mu}(X))]\\ &= \mathbb{E}\left\{(\mu(X)-\widehat{\mu}(X))\cdot\mathbb{E}[(W-e(X)) \mid X]\right\}\\ &= \mathbb{E}\left\{(\mu(X)-\widehat{\mu}(X))\cdot(\mathbb{E}[W \mid X]-e(X))\right\}\\ &= 0,\end{aligned}$$

so the claim also follows from (b).

3. Nonparametric functional estimation. *(30 points + 5 bonus points)*

Let f be an unknown density on $[0,1]$, and we observe i.i.d. $X_1, \cdots, X_n \sim f$. Instead of estimating f itself, suppose that our target is to estimate the quadratic functional

$$Q(f) = \int_0^1 f(x)^2 dx.$$

(a) Given an estimator $\widehat{f} \geq 0$ for f, write down the plug-in estimator $\widehat{Q}$ for $Q(f)$. Suppose that $\int_0^1 |\widehat{f}(x) - f(x)| dx \leq \varepsilon$, and $\max\{f(x), \widehat{f}(x)\} \leq L$ for all $x \in [0,1]$. Show that your above estimator $\widehat{Q}$ satisfies

$$|\widehat{Q} - Q(f)| \leq C\varepsilon,$$

for a constant $C > 0$ depending only on L. *(10 points)*

Solution: The plug-in estimator $\widehat{Q}$ is given by

$$\widehat{Q} = Q(\widehat{f}) = \int_0^1 \widehat{f}(x)^2 dx.$$

For the analysis, we have

$$\begin{aligned}
|\widehat{Q} - Q(f)| &= \left| \int_0^1 f(x)^2 dx - \int_0^1 \widehat{f}(x)^2 dx \right| \\
&\leq \int_0^1 |f(x)^2 - \widehat{f}(x)^2| dx \\
&= \int_0^1 (f(x) + \widehat{f}(x)) |f(x) - \widehat{f}(x)| dx \\
&\leq 2L \int_0^1 |f(x) - \widehat{f}(x)| dx \\
&\leq 2L\varepsilon.
\end{aligned}$$

(b) Given an estimator $\widehat{f}$, another plug-in estimator is defined as

$$\widehat{Q}_1 = \frac{2}{n}\sum_{i=1}^{n} \widehat{f}(X_i) - \int_0^1 \widehat{f}(x)^2 dx.$$

Show that

$$\mathbb{E}[\widehat{Q}_1] = Q(f) - \int_0^1 (\widehat{f}(x) - f(x))^2 dx.$$

We assume that $\widehat{f}$ is independent of $(X_1, \cdots, X_n)$ (e.g. constructed from another sample via sample splitting) and thus treated as *fixed*. *(10 points)*

Solution: By the independence of $\widehat{f}$ and $(X_1, \cdots, X_n)$, it holds that

$$\mathbb{E}[\widehat{f}(X_i)] = \int_0^1 \widehat{f}(x) f(x) dx.$$

Consequently,

$$\begin{aligned}\mathbb{E}[\widehat{Q}_1] &= 2\int_0^1 \widehat{f}(x)f(x)dx - \int_0^1 \widehat{f}(x)^2 dx \\ &= \int_0^1 \left[f(x)^2 - (f(x) - \widehat{f}(x))^2 \right] dx \\ &= Q(f) - \int_0^1 (\widehat{f}(x) - f(x))^2 dx.\end{aligned}$$

(c) Under the setting of (b), suppose that $\int_0^1 (\widehat{f}(x) - f(x))^2 dx \le \varepsilon^2$, and $0 \le \widehat{f}(x) \le L$ for all $x \in [0,1]$. By analyzing the bias and variance separately, show that

$$\mathbb{E}[(\widehat{Q}_1 - Q(f))^2] \le \varepsilon^4 + \frac{C}{n},$$

for a constant $C > 0$ depending only on L.

Solution: It has been shown in (b) that

$$|\mathrm{Bias}(\widehat{Q}_1)| = |\mathbb{E}[\widehat{Q}_1] - Q(f)| = \int_0^1 (\widehat{f}(x) - f(x))^2 dx \le \varepsilon^4.$$

As for the variance, first note that

$$\mathrm{Var}(\widehat{f}(X_i)) \le \mathbb{E}[\widehat{f}(X_i)^2] \le L^2.$$

Consequently, by the i.i.d. structure of $(X_1, \cdots, X_n)$, we have

$$\mathrm{Var}(\widehat{Q}_1) = \frac{4}{n^2} \sum_{i=1}^n \mathrm{Var}(\widehat{f}(X_i)) \le \frac{4L^2}{n}.$$

In summary,

$$\mathbb{E}[(\widehat{Q}_1 - Q(f))^2] = |\mathrm{Bias}(\widehat{Q}_1)|^2 + \mathrm{Var}(\widehat{Q}_1) \le \varepsilon^4 + \frac{4L^2}{n}.$$

(d) Now suppose the target is to estimate the entropy functional

$$h(f) = -\int_0^1 f(x) \log f(x) dx,$$

where log is the natural logarithm. Given an estimator $\widehat{f}$ independent of $(X_1, \cdots, X_n)$, propose an estimator $\widehat{h}$ of $h(f)$ in a similar spirit to (b). Justify your answer. *(5 bonus points)*

Solution: Apply the Taylor expansion of $-f \log f$ around $f \approx \widehat{f}$:

$$-f \log f = -\widehat{f} \log \widehat{f} - (1 + \log \widehat{f})(f - \widehat{f}) + O((f - \widehat{f})^2).$$

Integration over $x \in [0, 1]$ gives

$$\begin{aligned} h(f) &= h(\widehat{f}) - \int_0^1 (1 + \log \widehat{f}(x))(f(x) - \widehat{f}(x))dx + O(\|f - \widehat{f}\|_2^2) \\ &= h(\widehat{f}) - \int_0^1 (f(x) - \widehat{f}(x)) \log \widehat{f}(x) dx + O(\|f - \widehat{f}\|_2^2) \\ &= -\int_0^1 f(x) \log \widehat{f}(x) dx + O(\|f - \widehat{f}\|_2^2). \end{aligned}$$

This suggests to use the following unbiased estimator for the main term:

$$\widehat{h} = -\frac{1}{n} \sum_{i=1}^n \log \widehat{f}(X_i).$$