Second Midterm Review Sheet (Part 1 of 2)

Ec240a – Second Half, Fall 2024

In preparing for the exam review your lecture notes, assigned course readings, problem sets and prepare answers to the questions on the review sheet(s). You may bring to the exam a single 8.5 × 11 inch sheet of paper with notes on it. No calculation aides are allowed in the exam (e.g., calculators, phones, laptops etc.). You may work in pencil or pen, however I advise the use of pencil. Please be sure to bring sufficient blue books with you to the exam if possible. Scratch paper will be provided.

[1] You observe a simple random sample of size N from the population

$$Y_0 \sim N\left(\mu, \sigma^2\right)$$

as well as a second, independent, simple random sample, also of size N, from the population

$$Y_1 \sim N\left(\mu, 4\sigma^2\right).$$

The value of σ^2 is known. Consider the family of estimates of μ

$$\hat{\mu}\left(c_0, c_1\right) = c_0 \bar{Y}_0 + c_1 \bar{Y}_1,$$

where $\bar{Y}_0 = \frac{1}{N}\sum_{i=1}^{N} Y_{0i}$ and $\bar{Y}_1 = \frac{1}{N}\sum_{i=1}^{N} Y_{1i}$.

[a] Show that mean squared error equals

$$\mathbb{E}\left[\left(\hat{\mu}\left(c_0, c_1\right) - \mu\right)^2\right] = \frac{c_0^2\sigma^2}{N} + \frac{c_1^2 4\sigma^2}{N} + \left(1 - c_0 - c_1\right)^2 \mu^2. \quad (1)$$

[b] Derive the oracle estimator (within the family) which minimizes (1).

[c] Show that

$$\hat{R}\left(c_0, c_1\right) = \frac{c_0^2\sigma^2}{N} + \frac{c_1^2 4\sigma^2}{N} + \left(1 - c_0 - c_1\right)^2 \frac{1}{2}\left\{\bar{Y}_0^2 + \bar{Y}_1^2 - \frac{5\sigma^2}{N}\right\} \quad (2)$$

is an unbiased estimate of of (2). Can you propose another unbiased risk estimate? Why would you prefer one unbiased risk estimate over another?

[d] Describe in *words* how one might use (2) to construct an implementable estimator of μ.

[2] Let (D_i, X_i, Y_i) for $i = 1, \ldots, N$ be a simple random sample drawn from the target population of interest. Here Y_i denotes an outcome of interest, $D_i \in \{0, 1\}$ an indicator for assignment to active treatment ($D_i = 1$) or control ($D_i = 0$) and $X_i \in \{x_1, \ldots, x_L\}$ a discretely-valued pre-treatment attribute taking on L values. Let Y_{1i} denote unit i's potential outcome under treatment and Y_{0i} their potential outcome under control. The observed outcome equals

$$Y_i = \left(1 - D_i\right) Y_{0i} + D_i Y_{1i}. \quad (3)$$

That is we observe Y_{1i} if unit i is treated and Y_{0i} otherwise. The effect of treatment on unit i is

$$Y_{1i} - Y_{0i}.$$

Our target estimands are the conditional average treatment effect (CATE)

$$\alpha_l = \mathbb{E}\left[Y_{1i} - Y_{0i} \middle| X = x_l \right], \ l = 1, \ldots, L.$$

We assume that the treatment is randomly assigned within subpopulations homogenous in X such that

$$(Y_{0i}, Y_{1i}) \perp D_i | X_i = x_l, \ l = 1, \ldots, L. \tag{4}$$

We also assume that within each such some population some units are treated, while others are not:

$$0 < \kappa \leq e(x_l) \leq 1 - \kappa < 1, \ l = 1, \ldots, L, \tag{5}$$

with $e(x_l) = \Pr(D_i = 1 | X_i = x_l)$ the "propensity score".

[a] Use (4) and (4) to show that

$$\mathbb{E}\left[Y_i | X = x_l, D_i = 1 \right] = \mathbb{E}\left[Y_{1i} | X = x_l \right]$$

for $l = 1, \ldots, L$ and, similarly, that

$$\mathbb{E}\left[Y_i | X = x_l, D_i = 0 \right] = \mathbb{E}\left[Y_{0i} | X = x_l \right]$$

and hence that

$$\alpha_l = \mathbb{E}\left[Y_i | X = x_l, D_i = 1 \right] - \mathbb{E}\left[Y_i | X = x_l, D_i = 0 \right].$$

[b] Let $p = \Pr(D_i = 1)$ equal the marginal probability of treatment. Use Bayes' Rule to show that

$$f(x_l) = \frac{pf(x_l | D_i = 1)}{e(x_l)} = \frac{(1-p) f(x_l | D_i = 0)}{1 - e(x_l)}.$$

[c] Use (3), (4) and (5), your results from parts (a) and (b) to show that

$$\alpha_l = \mathbb{E}\left[\left. \frac{D_i Y_i}{e(x_l)} \right| X = x_l \right] - \mathbb{E}\left[\left. \frac{(1 - D_i) Y_i}{e(x_l)} \right| X = x_l \right].$$

Comment on the role played by (5).

[d] Assume that

$$\begin{aligned} Y_{1i} | X_i = x_l &\sim \mathcal{N}\left(\beta_{1l}, \sigma_{1l}^2\right) \\ Y_{0i} | X_i = x_l &\sim \mathcal{N}\left(\beta_{0l}, \sigma_{0l}^2\right) \end{aligned}$$

for $l = 1, \ldots, L$. Assume that in each $X_i = x_l$ in your sample there are some treated and some control units. You may also assume that σ_{0l}^2 and σ_{1l}^2 are know for $l = 1, \ldots, L$. For the remainder of this problem we will consider properties under repeated samples with covariate and treatment configurations identical to those in the sample in hand (effectively allowing us to proceed 'as if' $\{D_i, X_i\}_{i=1}^N$ were non-stochastic. Consider the conditional average treatment effect estimate

$$\hat{\alpha}_l = \sum_{i=1}^{N} \left\{ \frac{D_i \mathbf{1}\left(X_i = x_l\right) Y_i}{D_i \mathbf{1}\left(X_i = x_l\right)} - \frac{\left(1 - D_i\right) \mathbf{1}\left(X_i = x_l\right) Y_i}{\left(1 - D_i\right) \mathbf{1}\left(X_i = x_l\right)} \right\}$$

and show that, letting $N_{1l} = \sum_{i=1}^{N} D_i \mathbf{1}\left(X_i = x_l\right)$ and $N_{0i} = \left(1 - D_i\right) \mathbf{1}\left(X_i = x_l\right)$,

$$\hat{\alpha}_l \sim \mathcal{N}\left(\alpha_l, \frac{\sigma_{1l}^2}{N_{1l}} + \frac{\sigma_{0l}^2}{N_{0l}}\right),$$

with the L different estimates being conditionally independent of each other.

[e] Let $\lambda_l \in [0, 1]$ for $l = 1, \ldots, N$ and assume we wish to construct good estimates for each of α_l for $l = 1, \ldots, N$ under squared error loss:

$$L\left(\hat{\alpha}, \alpha\right) = \sum_{l=1}^{L} \left(\hat{\alpha}_l - \alpha_l\right)^2.$$

Consider the shrinkage estimate

$$\hat{\alpha}_{l,\lambda} = \lambda_l \hat{\alpha}_l$$

for $l = 1, \ldots, L$. Show that the (infeasible) risk-minimizing choice of λ_l equals

$$\lambda_l^* = \frac{\alpha_l^2}{\alpha_l^2 + \frac{\sigma_{1l}^2}{N_{1l}} + \frac{\sigma_{0l}^2}{N_{0l}}}.$$

[f] Motivate the feasible estimate which uses

$$\hat{\lambda}_l^* = 1 - \frac{\frac{\sigma_{1l}^2}{N_{1l}} + \frac{\sigma_{0l}^2}{N_{0l}}}{\hat{\alpha}_l^2}$$

for $l = 1, \ldots, L$.

[g] What insights from your analysis above might apply when considering estimation of the average treatment effect

$$\alpha = \mathbb{E}\left[Y_{1i} - Y_{0i}\right].$$

[3] You have been hired by UNICEF to estimate the prevalence of childhood stunting (low height-for-age) across municipalities in a country where childhood malnutrition is commonplace. Let Y_{it} be the height-for-age Z score of individual $t = 1, \ldots, T$ in municipality $i = 1, \ldots, N$. In each municipality you draw T children at random and compute the average height-for-age Z score

$$\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}.$$

You assume that $Y_{it} | \theta_i \sim \mathcal{N}\left(\theta_i, \sigma^2\right)$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. In this model the expected height-for-age Z score, θ_i, varies across municipalities. Your goal is to estimate the municipality (population) means $\theta_1, \theta_2, \ldots, \theta_N$. Municipalities with low θ_i estimates will be slated to receive new anti-hunger and nutrition programs. Initially you may assume that σ^2 is known (in a healthy population of children $\sigma^2 \approx 1$ since height-for-age Z scores are calibrated to have unit variance in such a setting).

[a] Explain why, if $f\left(y_{it}\middle|\,\theta_i\right)$ is Gaussian, the municipality mean is also Gaussian:

$$\bar{Y}_i\big|\,\theta_i \sim \mathcal{N}\left(\theta_i, \frac{\sigma^2}{T}\right).$$

[b] Let $\|\mathbf{m}\| = \left[\sum_{i=1}^N m_i^2\right]^{1/2}$ denote the Euclidean norm of a vector. Let $\theta = (\theta_1, \ldots, \theta_N)'$. Show that

$$\mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right] = \sum_{i=1}^N \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right],$$

with $\hat{\theta}$ some estimate – based upon the sample data $\mathbf{Y} = (Y_{11}, \ldots, Y_{1T}, \ldots, Y_{N1}, \ldots, Y_{NT})'$ – of θ. Explain why this measures *expected* estimation accuracy or *risk*? What is being averaged in the expectation?

[c] Further show that

$$\mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right] = \sum_{i=1}^N \mathbb{V}\left(\hat{\theta}_i\right) + \sum_{i=1}^N \left(\mathbb{E}\left[\hat{\theta}_i\right] - \theta_i\right)^2.$$

Interpret this expression.

[d] Consider the following family of estimators for θ_i (for $i = 1, \ldots, N$):

$$\hat{\theta}_i = (1 - \lambda)\,\bar{Y}_i + \lambda\mu,$$

with μ the country-wide mean of Y_{it} (i.e., the expected height-for-age Z score of a randomly sampled child from the full country-wide population). You may assume that μ is known (perhaps from prior research). Assume that $0 \leq \lambda \leq 1$. Interpret this estimator? Why might the estimator with $\lambda = 0$ be sensible? How might you justify the estimator when $\lambda > 0$.

[e] Show, for the family of estimates introduced in part [d], that

$$\mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right] = (1 - \lambda)^2 \frac{N}{T}\sigma^2 + \lambda^2 \sum_{i=1}^N (\theta_i - \mu)^2.$$

You hear, in the hallways of Evans, that “small λ means small bias” and “big λ means low variance”. Explain?

[f] Show that the risk-minimizing choice of λ, say λ^*, is

$$\lambda^* = \frac{N\sigma^2}{N\sigma^2 + \sum_{i=1}^N T\,(\theta_i - \mu)^2}.$$

Is an estimator based upon λ^* feasible? Why or why not? What is the optimal choice of λ^* as $T \to \infty$? Provide some intuition for your answer. What happens to the optimal choice of λ^* as σ^2 becomes large? Provide some intuition for your answer.

[g] Show that

$$\sum_{i=1}^N \mathbb{E}\left[\left(\bar{Y}_i - \hat{\theta}_i\right)^2\right] = \sum_{i=1}^N \mathbb{E}\left[\left(\bar{Y}_i - \theta_i\right)^2\right] + \sum_{i=1}^N \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right] - \frac{2\sigma^2}{T}\mathrm{df}\left(\hat{\theta}\right)$$

with the degree-of-freedom of $\hat{\theta}$ (or *model complexity*) equal to

$$\mathrm{df}\left(\hat{\theta}\right) = \sum_{i=1}^{N} \frac{T}{\sigma^2} \mathbb{C}\left(\bar{Y}_i, \hat{\theta}_i\right).$$

We call the term to the left of the first equality above *apparent error*.

[h] Show that

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \theta_i\right)^2\right] = \frac{N}{T}\sigma^2$$

and also, for the family of estimates indexed by λ introduced in part [d] above, that

$$\mathrm{df}\left(\hat{\theta}\right) = N\left(1-\lambda\right).$$

[i] Calculate apparent error and model complexity for $\hat{\theta}$ when $\lambda = 0$. Explain?

[j] Calculate apparent error and model complexity for $\hat{\theta}$ when $\lambda = 1$. Explain?

[k] You are roaming around Evans Hall looking for Professor Graham's office. You can't find his office because of the confusing floor plan. However, after a few hours of wandering around aimlessly, you bump into someone who introduces himself as Chuck Stein. He rearranges the expression you derived in part [g] above to get

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right] = -\sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \theta_i\right)^2\right] + \sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \hat{\theta}_i\right)^2\right] + \frac{2\sigma^2}{T}\mathrm{df}\left(\hat{\theta}\right).$$

He then notices your results from part [h] further imply that

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right] = -\frac{N}{T}\sigma^2 + \sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \hat{\theta}_i\right)^2\right] + 2N\frac{\sigma^2}{T}\left(1-\lambda\right).$$

Finally he says therefore an unbiased estimate of risk is:

$$\mathrm{SURE}\left(\bar{Y}, \lambda\right) = -\frac{N}{T}\sigma^2 + \sum_{i=1}^{N} \lambda^2 \left(\bar{Y}_i - \mu\right)^2 + 2\frac{N}{T}\sigma^2\left(1-\lambda\right).$$

Provide an explanation for Chuck Stein's claim. Next show that

$$\mathbb{E}\left[\left(\bar{Y}_i - \mu\right)^2\right] = \frac{\sigma^2}{T} + \left(\theta_i - \mu\right)^2$$

and hence that

$$\mathbb{E}\left[\mathrm{SURE}\left(\bar{Y}, \lambda\right)\right] = \sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right]$$

as implied by Chuck's unbiasedness claim. HINT: Don't forget the work you've done in part [e] above (and you may need to factor a quadratic equation in λ). Why is this result awesome?

[l] Let $\hat{\lambda}^*$ be the value of λ which minimizes SURE $\left(\bar{Y}, \lambda\right)$. Show that

$$\hat{\lambda}^* = \frac{\frac{N}{T}\sigma^2}{\sum_{i=1}^{N}\left(\bar{Y}_i - \mu\right)^2}.$$

Relate this feasible estimator to the infeasible oracle estimator based upon λ^* defined in part [f] above. HINT: use the expression for $\mathbb{E}\left[\left(\bar{Y}_i - \mu\right)^2\right]$ derived in part [k] to argue that when N is large enough $\hat{\lambda}^* \approx \lambda^*$. Discuss.

[m] You decide to use the estimator based upon $\hat{\lambda}^*$. For each municipality you calculate

$$\hat{\theta}_i = \left(1 - \hat{\lambda}^*\right)\bar{Y}_i + \hat{\lambda}^*\mu,$$

and then report when $\hat{\theta}_i < -1$. You tell the Minister of Health that those municipalities where $\hat{\theta}_i < -1$ should be targeted for supplemental child nutrition programs to combat stunting. A few month's later the mayor of village $i = 19$ comes to the capital as says: "You really screwed up. In my village *every single one* of the T sampled children had a height-for-age Z score, Y_{19t}, less than negative -1. My villages' mean was $\bar{Y}_{19} = -5/4$, but because your goofy econometrician decided to shrink all the village means toward the country-wide mean of $\mu = 0$ (with $\hat{\lambda}^* = 1/5$), they reported $\hat{\theta}_{19} = -1$ to you. As a result I have a bunch of hungry kids not getting the help they need. Your estimator is biased!" Write a response to this Mayor's concern.

[4] This problem is adapted from (that late) Gary Chamberlain's undergraduate class at Harvard. Consider a sub-population of borrowers. To make this problem more realistic you may imagine these borrows are homogenous in a vector of observable attributes. Let $Y = 1$ if a borrower repays their loan and $Y = 0$ if they default. We have

$$\Pr\left(Y = 1 \middle| \theta\right) = \theta, \quad \Pr\left(Y = 0 \middle| \theta\right) = 1 - \theta.$$

You work at Proxima Centauri Bank (PCB). PCB is the a largest bank on a generation starship with tens of thousands of passengers traveling to a planetary system 100 light-years from earth. If the borrower pays back the loan the gain to PCB is g, whereas if they default the loss to the bank is l. Expected profits therefore equal

$$g\Pr\left(Y = 1 \middle| \theta\right) - l\Pr\left(Y = 1 \middle| \theta\right) = g\theta - l\left(1 - \theta\right).$$

[a] Assume that θ is known. What is the minimal repayment rate (i.e., value of θ) such that is profitable to lend to this group of borrowers?

[b] Let $\mathbf{Y} = \left(Y_1, \ldots, Y_N\right)'$ be a vector of past repayment outcomes for a random sample of borrowers. For a given θ what is the ex ante probability of the event $\mathbf{Y} = \mathbf{y}$ (i.e., find an expression for the likelihood $\Pr\left(\mathbf{Y} = \mathbf{y} \middle| \theta\right) = f\left(\mathbf{y} \middle| \theta\right)$)?

[c] You have been asked to consider whether it is profitable to continue to lend to this subpopulation (you may base your decision on the dataset introduced in part [b]). You may take one of two actions

$$\mathcal{A} = \left\{a_1, a_2\right\}.$$

Action a_1 corresponds to continuing to approve loans for this subpopulation. Whereas action a_2 corresponds to no longer lending to this group (which yields a payoff of zero under all states of nature). Write down the loss function associated these two actions (i.e., an expression for $L(\theta, a_1)$ and $L(\theta, a_2)$). Assume that loss equals the negative of expected profits calculated 'as if' θ were known.

[d] You attended many applied microeconomics seminars as a graduate student. Based on this experience you decide it is best to "let the data speak" (although your econometrics instructor has indicated to you that data does not, in fact, speak). Specifically you decide that you will construct a decision rule which maps the data into actions: $d{:}\mathbb{Y} \rightarrow \mathcal{A}$. The data will speak and you, as the ultimate decision-maker, will decide. Here $\mathbb{Y} = \{0,1\}^N$ is the set of possible repayment patterns in your sample. Define *risk* and explain why it equals

$$\begin{aligned} R(\theta, d) &= \mathbb{E}\left[L(\theta, d(\mathbf{Y}))\middle|\,\theta\right] \\ &= \sum_{\mathbf{y}\in\{0,1\}^N} L(\theta, d(\mathbf{y}))\, f(\mathbf{y}|\,\theta)\,. \end{aligned}$$

[e] Prior to boarding the starship you worked in a bank in Idaho. From this experience you formed a prior about θ with density $\pi(\theta)$. Using this prior show that average risk equals

$$\begin{aligned} r(\theta, d) &= \int R(\theta, d)\, \pi(\theta)\, \mathrm{d}\theta \\ &= \sum_{\mathbf{y}\in\{0,1\}^N} \left[\int L(\theta, d(\mathbf{y}))\, f(\mathbf{y}|\,\theta)\, \pi(\theta)\, \mathrm{d}\theta\right], \end{aligned}$$

and argue that you can solve for the average-risk-minimizing decision rule "sample-wise":

$$d_0(\mathbf{y}) = \arg\min_{a\in\mathcal{A}} \int L(\theta, d(\mathbf{y}))\, \pi(\theta|\,\mathbf{y})\, \mathrm{d}\theta,$$

where

$$\pi(\theta|\,\mathbf{y}) = \frac{f(\mathbf{y}|\,\theta)\,\pi(\theta)}{\int f(\mathbf{y}|\,\theta)\,\pi(\theta)\,\mathrm{d}t}.$$

[e] Compute the posterior expected loss from making the loan. From your answer deduce the (average risk) optimal decision rule (i.e., the "Bayes' rule").

[f] Show that the likelihood can be written as

$$f(\mathbf{y}|\,\theta) = \theta^{s_N}(1-\theta)^{N-s_N}$$

with $S_N = \sum_{i=1}^N Y_i$. Assume further a prior on θ of

$$\theta \sim \mathrm{Beta}(\alpha_1, \alpha_2)$$

and hence show that

$$\theta|\,\mathbf{z} \sim \mathrm{Beta}(S+\alpha_1, N-S+\alpha_2)\,.$$

What is the average risk optimal decision rule under this prior? Why does this decision rule only depend on the data through $S_N = s_N$?

[5] Let Y be a scalar random variable, X a K vector of covariates (which includes a constant), and W a

vector of additional covariates (which excludes a constant). Consider the long (linear) regression

$$\mathbb{E}^*\left[Y|\,W,X\right] = X'\beta_0 + W'\gamma_0. \tag{6}$$

Next define the short and auxiliary regressions

$$\mathbb{E}^*\left[Y|\,X\right] = X'b_0 \tag{7}$$

$$\mathbb{E}^*\left[W|\,X\right] = \Pi_0 X. \tag{8}$$

[a] Let $V = W - \mathbb{E}^*\left[W|\,X\right]$ be the projection error associated with the auxiliary regression. Show that

$$\begin{aligned}\mathbb{E}^*\left[Y|\,V,X\right] &= \mathbb{E}^*\left[Y|\,X\right] + \mathbb{E}^*\left[Y|\,1,V\right] - \mathbb{E}\left[Y\right]\\ &= \mathbb{E}^*\left[Y|\,X\right] + \mathbb{E}^*\left[Y|\,V\right]\end{aligned}$$

where $\mathbb{E}^*\left[Y|\,1,V\right]$ denotes the linear regression of Y onto a constant and V, while $\mathbb{E}^*\left[Y|\,V\right]$ denotes the corresponding regression without a constant (Hint: Observe that $\mathbb{C}\left(X,V\right) = 0$).

[b] Next show that $\mathbb{E}^*\left[Y|\,V,X\right] = \mathbb{E}^*\left[Y|\,W,X\right]$ and hence that the coefficient on V in $\mathbb{E}^*\left[Y|\,V,X\right]$ coincides with that on W in $\mathbb{E}^*\left[Y|\,W,X\right]$.

[c] Let $U = Y - \mathbb{E}^*\left[Y|\,X\right]$ be the projection error associated with the short regression. Derive the coefficient on V in the linear regression of U onto V (excluding a constant).

[d] Discuss the possible practical value of the results shown in [b] and [c] above.