Second Midterm Review Sheet (Part 2 of 2)

Ec240a – Second Half, Fall 2021

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] 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

$$Y_i = (1 - D_i) Y_{0i} + D_i Y_{1i}. \tag{1}$$

That us 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

$$(Y_{0i}, Y_{1i}) \perp D_i | X_i = x_l, \; l = 1, \ldots, L. \tag{2}$$

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

$$0 < \kappa \leq e(x_l) \leq 1 - \kappa < 1, \; l = 1, \ldots, L, \tag{3}$$

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

[a] Use (2) and (2) to show that

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

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

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

and hence that

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

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

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

[c] Use (1), (2) and (3), your results from parts (a) and (b) to show that

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

Comment on the role played by (3).

[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 allow 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

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

[2] 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 $f\left(y_{it}|\,\theta_i\right)$ is Gaussian, the municipality mean is also Gaussian:

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

[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

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

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

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

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

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

[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:

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

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

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

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 $\text{SURE}\left(\bar{Y},\lambda\right)$. 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.

[3] Let Y denote log-earnings and X years of completed schooling for a cohort of workers. Assume a random sample of size N is available from this population. Let $D_x=1$ if $X=x$ and zero otherwise. Assume that $X\in\{0,\ldots,16\}$ with positive probability attached to each support point.

[a] Let

$$\mathbb{E}^*\left[Y|\,D_1,\ldots,D_L\right]=\alpha_0+\sum_{l=1}^{16}\gamma_{0l}D_l.$$

What is the relationship between this linear predictor and $\mathbb{E}\left[Y|\,X=x\right]$?

[b] Assume that $\Pr\left(X=6\right)=0$. Is the linear predictor defined in part [a] still well-defined? Why or why not?

[c] You hypothesize that $\mathbb{E}\left[Y|\,X=x\right]$ is linear in x. Consider the linear predictor in part [a] and let $\beta=\left(\alpha,\gamma_1,\ldots,\gamma_{16}\right)'$. Show how your hypothesis may be equivalently expressed as set of linear restrictions of the form $C\beta_0=c$. Provide explicit expressions for C and c. Describe how you would construct a test statistics for your hypothesis. What is the asymptotic sampling distribution of your statistics under the null? Assume that your have a consistent estimate $\hat{\Lambda}$ of the asymptotic variance-covariance matrix of $\sqrt{N}\left(\hat{\beta}-\beta\right)$, with $\hat{\beta}$ the least squares estimate.

[d] After attending the labor lunch you now believe that $\mathbb{E}\left[Y|\,X=x\right]$ is linear in x but with discrete jumps at $X=12$ and $X=16$. Describe, in detail, how you would evaluate this new hypothesis?

[4] Consider the population of married men. Let Y denote log earnings for a generic random draw from this population, X his years of completed schooling and W the schooling of his spouse. Assume that the conditional mean of own log earnings given own and spouse's schooling is

$$\mathbb{E}\left[Y|\,X,W\right]=\alpha_0+\beta_0X+\gamma_0W,$$

while the best linear predictor of spouse's schooling given own schooling is

$$\mathbb{E}^*\left[W|\,X\right]=\delta_0+\zeta_0X.$$

You may assume that the joint distribution of (W,X,Y) is such that these objects are well-defined.

[a] Show that $\zeta_0=\rho_{WX}\frac{\sigma_W}{\sigma_X}$, with ρ_{WX} the correlation of W with X, and σ_W and σ_X respectively the standard deviation of W and X. Further show that $\delta_0=\mu_W-\rho_{WX}\frac{\sigma_W}{\sigma_X}\mu_X$ with μ_W and μ_X denoting the population means of W and X.

[b] Using your answers in [a] above, as well as the form of $\mathbb{E}\left[Y|\,W,X\right]$, provide an expression for $\mathbb{E}^*\left[Y|\,X\right]$.

[c] Consider another population of married men where $F_{Y|W,X}\left(y|\,W=w,X=x\right)$, $F_W\left(w\right)$ and $F_X\left(x\right)$ coincide with those for the population described above, but where $F_{W,X}\left(w,x\right)$ differs. Assume that in this alternative population $\rho_{WX}=0$. Solve for $\mathbb{E}\left[Y|\,X,W\right]$, $\mathbb{E}^*\left[W|\,X\right]$ and $\mathbb{E}^*\left[Y|\,X\right]$. Use the notation established in parts [a] and [b] to formulate your answer.

[d] Assume that $F_W\left(w\right)$ and $F_X\left(x\right)$ are identical and that marriage is homogamous in terms of education so that $W=X$ for all couples (i.e., individuals choose partners with identical levels of education). Show that in this world $\rho_{WX}=1$. Solve for $\mathbb{E}\left[Y|\,X,W\right]$, $\mathbb{E}^*\left[W|\,X\right]$ and $\mathbb{E}^*\left[Y|\,X\right]$. Use the notation established in parts [a] and [b] to formulate your answer.

[e] Compare the form of $\mathbb{E}^*\left[Y|\,X\right]$ in the original population with that in the two alternative populations of parts [c] and [d]. In which population does log earnings rise most steeply with years of schooling? Provide some intuition for your answer (5 sentences).

[f] Assume that schooling is binary valued, taking on the values 0,1. Let R_W be a 2×1 vector equal to $(1,0)'$ if $W=0$ and $(0,1)'$ if $W=1$. Let S_X be the analogous 2×1 vector defined using X. Let

$T_{WX} = (R_W \otimes S_X)$ and

$$\mathbb{E}^* \left[Y | T_{WX} \right] = T'_{WX} \pi,$$

where a constant is not included and $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})'$. Show that

$$\pi_{jk} = \mathbb{E} \left[Y | W = j, X = k \right].$$

[g] Consider the null hypothesis that $\mathbb{E} \left[Y | W, X \right] = \alpha_0 + \beta_0 X + \gamma_0 W$. Maintaining this null find an explicit expression for each component of π in terms of α_0, β_0 and γ_0. Express this null in the form $C\pi = c$ for some matrix of constants C and vector of constants c.

[h] Let $W = 1$ if a wife has completed primary school and zero otherwise, let $X = 1$ if a husband has completed primary school and zero otherwise. A least squares fit, loosely based on data from Brazil, of log husband's earnings on T_{WX} as defined in [f] using a random sample of size $N = 50,000$ yields point estimate of

$$\hat{\pi} = \begin{pmatrix} 5.50 \\ 6.00 \\ 5.00 \\ 7.00 \end{pmatrix}$$

with an estimated asymptotic variance-covariance matrix of

$$\hat{\Lambda} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Can you reject the null hypothesis (at the $\alpha = 0.05$ level) formulated in part [g] on the basis of this sample? For your reference the 0.95 quantiles of χ^2 random variables with parameters 1, 2 and 3 are, respectively, 3.84, 5.99 and 7.81.