Ec141, Spring 2019

Professor Bryan Graham

Review Sheet 1

This review sheet is designed to assist you in your exam preparations. I suggest preparing written answers to each question. You may find it useful to study with your classmates (indeed I encourage you to do so and also to be generous with one another as you prepare). In the exam you may bring in a single 8.5 x 11 sheet of notes. No calculators or other aides will be permitted. Please bring blue books to the exam. The midterm exam will occur in class on Thursday, March 22nd.

[1] Let W, X be a pair of regressors with the property that $\mathbb{C}(W, X) = 0$. Show that, for outcome, Y,

$$\mathbb{E}^{*}[Y|W, X] = \mathbb{E}^{*}[Y|W] + \mathbb{E}^{*}[Y|X] - \mathbb{E}[Y].$$

You may assume that all objects in the above expression are well-defined (i.e., all necessary moments exist and so on).

[a] First show that

$$\mathbb{E}^{*}[\mathbb{E}^{*}[Y|W]|X] = \mathbb{E}^{*}[\mathbb{E}^{*}[Y|X]|W] = \mathbb{E}[Y]$$

[b] Second verify the result using the Projection Theorem.

[c] Finally show that

$$\mathbb{E}^{*}[Y|W, X] = \mathbb{E}[Y] + \frac{\mathbb{C}(Y, W)}{\mathbb{V}(W)}(W - \mathbb{E}[W]) + \frac{\mathbb{C}(Y, X)}{\mathbb{V}(X)}(X - \mathbb{E}[X]).$$

[2] You've been hired by the Government of Honduras to assess the efficacy of treatment for decompression sickness among lobster divers in La Moskitia. In this region of Honduras lobsters are harvested by divers who, on occasion, get decompression sickness which may result in partial paralysis or worse. You are provided the following table of information about 300 diving accident victims.

		$Y = 0$ (No Limp)	$Y = 1$ (Limp)
$X = 0$ (Untreated)	$W = 0$ (Depth $<$ 75')	90	10
	$W = 1$ (Depth $\geq$ 75')	10	40
$X = 1$ (Treated)	$W = 0$ (Depth $<$ 75')	30	20
	$W = 1$ (Depth $\geq$ 75')	50	50

[a] What is the probability of a victim walking with a limp conditional on treatment ($X = 1$) and non-treatment ($X = 0$)?

[b] What is the probability of a victim receiving treatment conditional on having dived "deep" ($W = 1$) vs. "shallow" ($W = 0$)?

[c] A government official worries that treatment is harming the divers and thinks it would be better to do nothing. Present a counter-argument to this official.

[d] Let $Y(0)$ and $Y(1)$ denote a divers potential outcome given non-treatment and treatment respectively. Discuss the conditional independence assumption assumption

$$(Y(0), Y(1)) \perp X| W = 0, 1.$$

Make a positive and negative argument for this assumption.

[e] Using the assumption in part [d] construct the IPW estimate of the average treatment effect (ATE) on the outcome. Report your result to the government official. Your report should include an explanation for why and how your are adjusting for accident depth. Is treatment effective?

[f] Say instead you were given the table:

		$Y = 0$ (No Limp)	$Y = 1$ (Limp)
$X = 0$ (Untreated)	$W = 0$ (Depth $<$ 75')	90	10
	$W = 1$ (Depth $\geq$ 75')	0	0
$X = 1$ (Treated)	$W = 0$ (Depth $<$ 75')	30	20
	$W = 1$ (Depth $\geq$ 75')	75	75

Can you compute the ATE is this case? Why or why not?

[3] You are given a random sample from South Africa in the late 1980s. Each record in this sample includes, Y, an individual's log income at age 40, X the log permanent income of their parents, and D a binary indicator equaling 1 if the respondent is White and zero if they are Black. Let the best linear predictor of own log income at age forty given parents' log permanent income and own race be

$$\mathbb{E}^* \left[Y | X, D \right] = \alpha_0 + \beta_0 X + \gamma_0 D.$$

[a] Let $Q = \Pr(D = 1)$, assume that $\mathbb{V}(X | D = 1) = \mathbb{V}(X | D = 0) = \sigma^2$ and recall the analysis of variance formula $\mathbb{V}(X) = \mathbb{V}(\mathbb{E}[X | D]) + \mathbb{E}[\mathbb{V}(X | D)]$. Show that

$$\mathbb{V}(X) = Q(1 - Q) \left\{ \mathbb{E}[X | D = 1] - \mathbb{E}[X | D = 0] \right\}^2 + \sigma^2.$$

[b] Let $\mathbb{E}^* [D | X] = \kappa + \lambda X$. Show that

$$\lambda = \frac{Q(1 - Q) \left\{ \mathbb{E}[X | D = 1] - \mathbb{E}[X | D = 0] \right\}}{Q(1 - Q) \left\{ \mathbb{E}[X | D = 1] - \mathbb{E}[X | D = 0] \right\}^2 + \sigma^2}.$$

[c] Assume that $\beta_0 = 0$. Show that in this case $\gamma_0 = \mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0]$.

[d] Let $\mathbb{E}^* [Y | X] = a + bX$. Maintaining the assumption that $\beta_0 = 0$ show that

$$b = \frac{Q(1 - Q) \left\{ \mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0] \right\} \left\{ \mathbb{E}[X | D = 1] - \mathbb{E}[X | D = 0] \right\}}{Q(1 - Q) \left\{ \mathbb{E}[X | D = 1] - \mathbb{E}[X | D = 0] \right\}^2 + \sigma^2}.$$

[e] Let $Q(1 - Q) = 1/10$, $\sigma^2 = 3/10$ and $\mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0] = \mathbb{E}[X | D = 1] - \mathbb{E}[X | D = 0] = 3$. Provide a numerical value for $\mathbb{V}(X)$ and b.

[f] On the basis of β_0 a member of the National Party argues that South Africa is a highly mobile society. One the basis of b a member of the African National Congress argues that it is a highly immobile one. Comment on the relative merits of these two assertions.

[4] Let $C_t = 1$ if an individual (child) went to college and zero otherwise. Let $C_{t-1} = 1$ if the corresponding individual's parent went to college and zero otherwise. The following table gives the joint distribution of father and sons' college attendance:

	$C_t = 0$	$C_t = 1$
$C_{t-1} = 0$	0.60	0.20
$C_{t-1} = 1$	0.10	0.10

For example 20% percent of the population consists of pairs with a father who did not attend college, but a son who did.

[a] Among children of college graduates, what fraction go on to complete college themselves? Among children of non-graduates, what fraction go on to complete college themselves?

[b] Let $\mathbb{E}^*\left[C_t|\, C_{t-1}\right] = a + bC_{t-1}$; calculate a and b.

[c] The following table gives child's adult earnings, Y_t, for each of the four subpopulations introduced above

	$C_t = 0$	$C_t = 1$	
$C_{t-1} = 0$	\$8,000	\$60,000	.
$C_{t-1} = 1$	\$14,000	\$30,000	

What is the average earnings level of college graduates in this economy? What is the average earnings of non-college graduates? What is the overall average earnings level? Express your answers symbolically using the notation of (conditional) expectations and also provide a numerical answer.

[d] Let $\pi_{c_{t-1}} = \Pr\left(C_{t-1} = c_{t-1}|\, C_t = 1\right)$. Consider the estimand

$$\beta = \sum_{c_{t-1}=0,1} \left\{\mathbb{E}\left[Y|\, C_t = 1, C_{t-1} = c_{t-1}\right] - \mathbb{E}\left[Y|\, C_t = 0, C_{t-1} = c_{t-1}\right]\right\} \pi_{c_{t-1}}.$$

In what sense does β adjust for "covariate differences" between college and non-college graduates? Evaluate β and compare your numerical answer with the raw college - non-college earnings gap you calculated in part (c). Why are these two numbers different?

[e] Gavin Newsom is considering a community college expansion policy. You have been tasked to asked to predict the earnings gain associated with completing a college degree. Gavin estimates that after the community college expansion the distribution of college attendance in California will look like

	$C_t = 0$	$C_t = 1$	
$C_{t-1} = 0$	0.40	0.40	.
$C_{t-1} = 1$	0.05	0.15	

Calculate average earnings in this new economy (you may assume that the mapping from background and education into earnings introduced in part (c) remains the same)? Assume a state tax rate of 10 percent. What is the long run predicted increase in annual tax revenue from the community college expansion? Treat this revenue as a perpetuity and assume a discount rate of 0.05. What is the present value of the increase in tax revenue that is expected to be generated by the community college expansion?

[5] Let $X = (1, R')'$. The mean squared error minimizing linear predictor of Y given X is

$$\mathbb{E}^*\left[Y|\, X\right] = X'\eta_0,\ \eta_0 = \mathbb{E}\left[XX'\right]^{-1} \times \mathbb{E}\left[XY\right].$$

You may assume that all objects in the above expression are well-defined (i.e., all necessary moments exist and so on).

[a] Show that

$$\mathbb{E}\left[XX'\right] = \left[\begin{array}{cc} 1 & \mathbb{E}\left[R'\right] \\ \mathbb{E}\left[R\right] & \mathbb{E}\left[RR'\right] \end{array}\right].$$

[b] Use the partitioned inverse formula to show that

$$\mathbb{E}\left[XX'\right]^{-1} = \left[\begin{array}{cc} 1 - \mathbb{E}\left[R'\right]' \mathbb{V}\left(R\right)^{-1} \mathbb{E}\left[R'\right] & -\mathbb{E}\left[R'\right]' \mathbb{V}\left(R\right)^{-1} \\ -\mathbb{V}\left(R\right)^{-1} \mathbb{E}\left[R'\right] & \mathbb{V}\left(R\right)^{-1} \end{array}\right].$$

[c] Next show that

$$\eta_0 = \left(\begin{array}{c} \alpha_0 \\ \beta_0 \end{array}\right) = \left(\begin{array}{c} \mathbb{E}\left[Y\right] - \mathbb{E}\left[R\right]' \beta_0 \\ \mathbb{V}\left(R\right)^{-1} \mathbb{C}\left(R, Y\right) \end{array}\right).$$

[d] Finally show that

$$\left\|Y - X'\eta_0\right\|^2 = \mathbb{V}\left(Y\right) - \mathbb{C}\left(Y, R'\right) \mathbb{V}\left(R\right)^{-1} \mathbb{C}\left(R, Y\right).$$

[e] Let $\rho^2 = 1 - \frac{\left\|Y - X'\eta_0\right\|^2}{\mathbb{V}(Y)}$ be the *coefficient of determination*. Interpret this object.

[6] Let $\mathcal{H}$ be a (pre-) Hilbert space.

[a] Prove that for all $(h_1, h_2) \in \mathcal{H} \times \mathcal{H}$,

$$\left|\langle h_1, h_2 \rangle\right| \leq \left\|h_1\right\| \left\|h_2\right\|$$

with equality if, and only if, $h_1 = \alpha h_2$ for some real scalar α or $h_2 = 0$.

[b] Prove that for all$(h_1, h_2) \in \mathcal{H} \times \mathcal{H}$,

$$\left\|h_1 + h_2\right\| \leq \left\|h_1\right\| + \left\|h_2\right\|.$$

[c] Let y be a fixed vector within $\mathcal{H}$. Show that for each $\epsilon > 0$ that there exists a $\delta > 0$ such that

$$\left|\langle h_1, y \rangle - \langle h_2, y \rangle\right| \leq \epsilon$$

for all $h_1, h_2 \in \mathcal{H}$ where $\left\|h_1 - h_2\right\| \leq \delta$ (HINT: use the inequality shown in [a] above).

[7] The Undergraduate Dean has been collecting data on the high school GPA (X) of incoming students for a long, long time. She has also kept track of 1st semester GPA (Y) for incoming students over the same period of time. She would like to be able to predict 1st semester GPA for incoming students using their high school GPA. She reports to you the following means, variances and a covariance for X and Y:

$$\mu_X = 3, \mu_Y = 2$$

and

$$\sigma_X^2 = 1/2, \sigma_Y^2 = 1/4, \sigma_{XY} = 1/3.$$

Because she has collected such a large sample you are free to treat these numbers as if they were population quantities.

[a] Calculate the α and β associated with the (mean square error minimizing) linear predictor of Y given X, $\mathbb{E}^*[Y|X] = \alpha + \beta X$?

[b] What is the coefficient of determination associated with $\mathbb{E}^*[Y|X]$?

[8] Consider the following statistical model for the earnings of Berkeley students

$$Y = \alpha + \beta G + \gamma A + U, \; \mathbb{E}[U|G, A] = 0,$$

where G equals one if the student graduated and zero if they dropped out and A equals one if at least one of the student's parents graduated from college and zero otherwise.

[a] You read in the Oakland Tribune newspaper that Berkeley graduates earn an average of \$55,000 per year nationwide, while the earnings of dropouts average only \$35,000. Express this population earnings difference between Berkeley graduates and dropouts in terms of the statistical model given above.

[b] Under what conditions is it true that $\beta = 20,000$? Do you think these conditions are likely to be true in practice? Briefly explain your answer.

[c] The same article reports that among Berkeley graduates 80 percent come from families where at least one parent completed college, while among all former students (i.e., graduates and dropouts) only 50 percent come from such families. It also states that the overall (i.e., unconditional) graduation rate at Berkeley is 50 percent. What fraction of dropouts come from families where at least one parent completed college?

[d] Assume $\gamma = 10,000$. Using your answers in parts (a) and (c) solve for β. What is the expected earnings gain associated with graduating from Berkeley holding parent's education (i.e., A) constant? Briefly comment on why your answer differs from the earnings gap between graduates and dropouts reported by the Tribune.

[e] You are considering dropping out of Cal to spend more time on Telegraph Avenue. What is the (approximate) expected earnings loss associated with this decision?

[f] You move to Oakland upon graduation, your neighbor to the left tells you that he dropped out of Berkeley during the Free Speech Movement, your neighbor to the right tells you that he graduated from Berkeley about the same time. What is your expectation of the annual earnings of your two neighbors?

[9] The World Health Organization has contracted you to design a randomized experiment evaluating the efficacy of zinc supplements on diarrhea prevalence (measured as the number of episodes in the one hundred days prior to surveying). Let $Y(1)$ be the potential number of episodes of diarrhea if taking zinc supplements and $Y(0)$ the control potential outcome. A baseline survey of your target population yields a diarrhea prevalence of 10 days per one hundred days with a standard deviation of 5 days. Let N be your target sample size and assume that half of respondents will be randomly assigned to treatment. Assume that the variance of $Y(1)$ and $Y(0)$ are equal to each other. Also assume that no respondents in your baseline survey were taking zinc supplements.

[a] Derive an expression for the ex ante probability (β) that you reject the null of no effect in favor of a *one-sided* alternative of a positive effect. Let α denote the size of your test and θ the ATE. Carefully explain your reasoning and notation.

[b] Assume that $\theta = 4$. How large would N need to be to ensure an ex ante rejection probability of 95 percent (for a test with size $\alpha = 0.05$).

[c] You ultimately design an experiment with power of $\beta = 0.90$ and size $\alpha = 0.05$. In the end you find no effect of zinc supplements on the prevalence of diarrhea (i.e., you fail to reject the null of no effect). Prior to the experiment you believed that the probability that zinc supplements reduced the prevalence of diarrhea was 0.9. What is your belief after your null finding?