Ec142, Spring 2017

Professor Bryan Graham

Review Sheet 3

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. 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, April 27th.

[1] Consider a population of high school graduates. Let Y_1 denote the earnings an individual in this population would get if they completed at least four years of college, let Y_0 denote the earnings they would get if they did not complete college. Let $D = 1$ in an individual actually completes college and zero otherwise. Let Y denote observed earnings which, given the data structure outlined above, equals

$$Y = (1 - D) Y_0 + DY_1.$$

Assume that

$$(Y_1, Y_0) \perp D | X$$

where X is a characteristic measured at the completion of high school but prior to any college attendance. Further assume that

$$\begin{aligned} \mathbb{E}[Y_1 | X] &= \alpha_1 + \gamma_1 X \\ \mathbb{E}[Y_0 | X] &= \alpha_0 + \gamma_0 X. \end{aligned}$$

[a] Across subpopulations homogenous in $X = x$ can we use Y_1 or Y_0 to predict college attendance? Would these variables be informative about college attendance unconditional on X? Can we use an individual's observed wage Y to predict whether they went to college? [4 to 6 sentences]

[b] Let $\beta_0 = E[Y_1 - Y_0 | D = 1]$. Interpret this object. Derive a representation of it in terms of α_0, α_1, γ_0, γ_1 and the distribution of X.

[c] Show that

$$\mathbb{E}[Y | X, D] = \alpha_0 + \gamma_0 X + (\alpha_1 - \alpha_0) D + (\gamma_1 - \gamma_0) DX.$$

[d] Consider the following least squares fit of log earnings on a dummy for completion

of an undergraduate degree using a sample of 1,754 white males from the NLSY79.

$$\underset{}{\text{LogEarnings}} = \underset{(0.0220)}{10.0332} + \underset{(0.0357)}{0.4879} \quad \text{UNDERGRAD} \quad . \tag{1}$$

Now consider the least squares fit which additionally includes an individual's AFQT percentile score and its interaction with UNDERGRAD:

$$\begin{aligned} \underset{}{\text{LogEarnings}} = & \underset{(0.0573)}{9.8231} + \underset{(0.0010)}{0.0040} \quad \text{AFQT} \\ & + \underset{(0.1584)}{0.1898} \quad \text{UNDERGRAD} + \underset{(0.0020)}{0.0023} \quad \text{UNDERGRAD} \times \text{AFQT} \end{aligned} \tag{2}$$

Finally consider the least squares fit of AFQT on a constant and UNDERGRAD:

$$\underset{}{\text{AFQT}} = \underset{(0.73)}{52.27} + \underset{(1.05)}{28.35} \quad \text{UNDERGRAD} \quad . \tag{3}$$

Assume that $X = AFQT$. Using these results compute an estimate of β_0. Justify and explain your calculations. How does your estimate compare with the coefficient on UNDERGRAD in (1)? Comment on an differences and provide an explanation for them (if needed).

[2] Consider the following (mean squared error minimizing) linear predictor

$$\mathbb{E}^*[Y|X] = \alpha + \beta X.$$

Let $X^* = X + U$ with U independent of (X, Y).

[a] Show that $\mathbb{C}(X^*, Y) = \mathbb{C}(X, Y)$.

[b] Show that $\mathbb{V}(X^*) = \mathbb{V}(X) + \mathbb{V}(U)$

[c] Let $\mathbb{E}^*[Y|X^*] = a + bX^*$. Derive an expression for b in terms of β, $\mathbb{V}(X)$ and $\mathbb{V}(U)$.

[d] Comment on the implications of your analysis for regression with mismeasured covariates. How might your analysis change if U covaries with X or Y?

[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

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

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

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

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

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

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

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

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

[e] Let $Q\left(1 - Q\right) = 1/10$, $\sigma^2 = 3/10$ and $\mathbb{E}\left[Y\middle|\, D = 1\right] - \mathbb{E}\left[Y\middle|\, D = 0\right] = \mathbb{E}\left[X\middle|\, D = 1\right] - \mathbb{E}\left[X\middle|\, D = 0\right] = 3$. Provide a numerical value for $\mathbb{V}\left(X\right)$ 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] On your birthday your parents give you a copy of Stata/SE version 12. "SE" stands for "special edition" and your parents claim to give you this version because you are indeed special. In fact they have ulterior motives. It turns out that your father has been secretly collecting weekly data on lawn growth rates (in millimeters) given 10 different levels of water input. Let Y be the amount of lawn growth in a given week and $X \in \{x_1, \ldots, x_{10}\}$ the corresponding water input. Your father asserts that the conditional median of Y given $X = x$ is quadratic in x (specifically an inverted 'U'). We would like you to estimate the level of water that leads to maximal (median) growth.

[a] For each of the ten water levels you have 100 observations of lawn growth. Describe (in detail) how to calculate the conditional median of Y given each of the ten values of $X = x$.

[b] Describe (in detail) how you could construct an estimate of the sampling variability of each of your part (a) estimates.

[c] Let $Q_{Y|X}(1/2|X) = \alpha_{1/2} + \beta_{1/2}X + \gamma_{1/2}X^2$. Describe (in detail) a procedure for estimating the parameters indexing this family of conditional median functions.

[d] Your father moved to Los Angeles decades ago to pursue his dream of become an action movie star. Although this dream did not come true he did grow to love California. Nevertheless he misses his native Kentucky. As an homage to his home state he plants Kentucky Bluegrass in his backyard. Unfortunately this type of grass is not drought resistant. Can you suggest a more conservative approach to choosing a watering level?

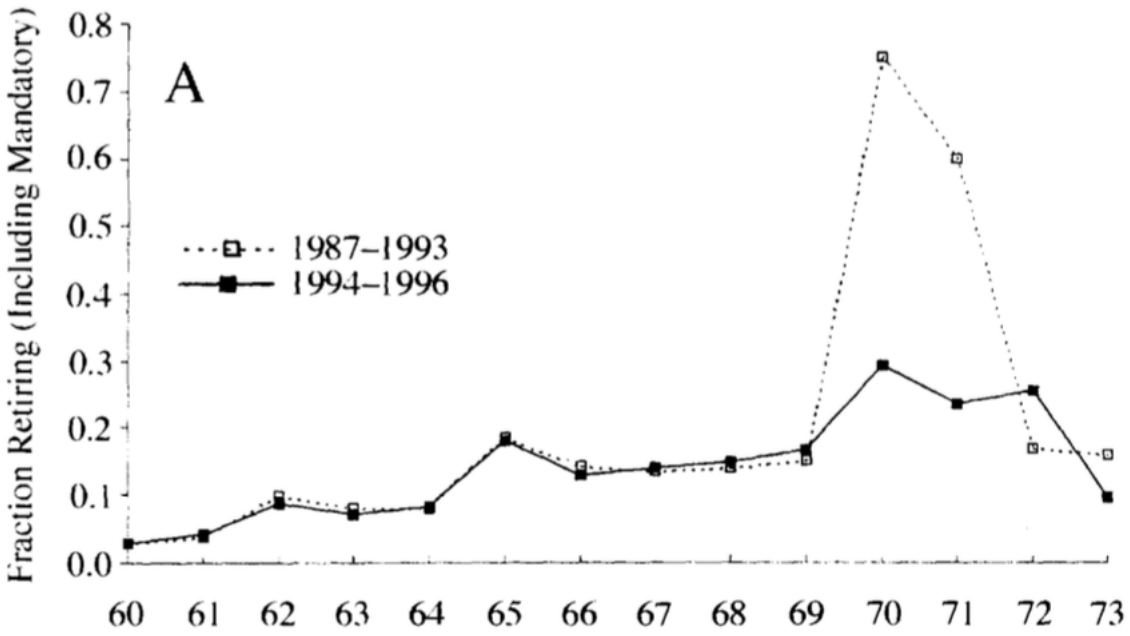
[5] Prior to 1994 colleges and universities in the United States were exempt from laws prohibiting mandatory retirement, consequently many institutions forced faculty to retire at age 70. After 1994 mandatory retirement rules were prohibited by Congress. Ashenfelter and Card (*AER*, 2002) study the effects of this exemption expiration on faculty retirement behavior in a sample of 104 colleges and universities. Let Y^* equal age of retirement, let C equal the age at which a faculty-member is lost to follow-up, $D = 1(Y^* \leq C)$ be a censoring indicator, and $Y = \min(Y^*, C)$ be the observed age at exit from the sample. Let T_y (for $y = 60, 61 \ldots, 72, 73$) equal the calendar year during which an individual was age y. So, for example, an individual who turned sixty in 1992 would have $Y_{60} = 1992$, while one who did so in 1999 would have $T_{60} = 1999$.

[a] Let $X = 0$ if $T_{70} < 1994$ and $X = 1$ if $T_{70} \geq 1994$. Assume that

$$\lambda(y; X) = \Pr(Y^* = y | Y^* \geq y, X) = \frac{\exp(\alpha_y + X\beta_0 + X \times \mathbf{1}(y \geq 70)\gamma_0)}{1 + \exp(\alpha_y + X\beta_0 + X \times \mathbf{1}(y \geq 70)\gamma_0)}. \quad (4)$$

In the context of the Ashenfelter and Card (*AER*, 2002) study interpret the hazard function $\lambda(y; X)$ when $X = 0$ and when $X = 1$. Interpret β_0 and γ_0 in terms of the hazard function.

[b] Reference the figure below when answering the following questions (justify your answers). What signs do you expect β_0 and γ_0 to take? Does the evidence appear consistent with the hypothesis that $\beta_0 = 0$ and $\gamma_0 < 0$?



[c] Assume that $D \perp Y^* | X$. Interpret this assumption. Decribe how it could be violated.

[d] Assume the first four lines of the Ashenfelter and Card (*AER*, 2002) dataset equal

	Y	D	X
1	65	0	0
2	72	1	0
3	61	1	1
4	70	0	1

What are these units' contributions to the corresponding "person-period" dataset (in the sense described in the Singer and Willett book)? Describe, in detail, how you could use this person period dataset to construct estimates of $\alpha_{60}, \alpha_{61}, \ldots, \alpha_{73}$, β_0 and γ_0 .

[e] Let $S(y; X) = \Pr(Y^* > y | X)$. Does $\Pr(Y > y | X) = S(y; X)$? If not, does $\Pr(Y > y | X) > S(y; X)$ or $\Pr(/Y > y | X) < S(y; X)$? Why? Describe a method for constructing an estimate of $S(y; X)$. Describe, in detail, how you could use this estimate to compute the effect of the end of mandatory retirement on median retirement age. Use the information in the table below to implement your procedure.

TABLE 2—AGE-SPECIFIC RETIREMENT RATES, BEFORE AND AFTER 1994

Age	Number of observations	Percentage post-1994	Average retirement rate		Change in retirement rate	
			1987–1993	1994–1996	Unadjusted	Adjusted from logit
60	7,343	31.8	3.3	3.0	−0.3	−0.2
			(0.3)	(0.4)	(0.4)	(0.5)
61	7,027	32.4	4.1	4.4	0.3	0.3
			(0.3)	(0.4)	(0.5)	(0.5)
62	6,665	32.9	10.3	8.9	−1.4	−1.4
			(0.5)	(0.6)	(0.8)	(0.8)
63	5,838	34.5	8.5	7.3	−1.3	−1.1
			(0.5)	(0.6)	(0.7)	(0.8)
64	5,222	35.4	8.4	8.5	0.1	0.1
			(0.5)	(0.7)	(0.8)	(0.8)
65	4,650	35.1	19.3	18.1	−1.2	−1.4
			(0.7)	(1.0)	(1.2)	(1.3)
66	3,653	35.1	14.7	13.0	−1.7	−1.9
			(0.7)	(0.9)	(1.2)	(1.3)
67	2,969	34.2	13.8	14.0	0.1	−0.1
			(0.8)	(1.1)	(1.3)	(1.4)
68	2,453	34.2	14.3	14.6	0.4	0.7
			(0.9)	(1.2)	(1.5)	(1.5)
69	2,004	33.7	15.4	16.7	1.3	0.6
			(1.0)	(1.4)	(1.7)	(1.7)
70	1,598	35.1	75.6	29.1	−46.5	−43.7
			(1.3)	(2.0)	(2.4)	(2.5)
71	502	58.6	60.6	23.8	−36.8	−32.2
			(3.4)	(2.5)	(4.2)	(4.0)
72	182	67.0	16.7	25.4	8.7	−3.7
			(4.9)	(4.0)	(6.3)	(7.2)

Notes: Retirement rates expressed as percent per year. Estimated standard errors are in parentheses. An individual's retirement age is measured as of September 1 following the date of retirement. The adjusted change in retirement rates is the normalized regression coefficient from a logit model for the event of retirement, fit by age and including a total of 19 covariates: gender, Ph.D., nonwhite race, region (three dummies), Carnegie classification and public/private status of institution, and six department dummies.