MIT (14.32) **Spring 2023**
J. Angrist

PRACTICE Midterm

1 True/False/Uncertain (30 points, 5 questions * 6 each)

Indicate whether the following are true, false, or uncertain, with a brief explanation.

1. Suppose both X_i and Y_i are dummy variables. The slope from a bivariate regression of Y_i on X_i is a difference in conditional probabilities.

2. Consider the residual from a bivariate regression: $e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$, where hats denote OLS estimates. The slope coefficient from a regression of e_i on X_i depends on the value of $\hat{\beta}$.

3. Regression estimates are made less precise by highly variable regressors.

4. Consider the long regression $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, and the short regression $Y_i = \alpha^* + \beta_1^* X_{1i} + \nu_i$. When $\beta_2 \neq 0$,the short coefficient, β_1^*, surely differs from the long coefficient, β_1.

5. Studies of the RAND Health Insurance Experiment show a strong relationship between health insurance coverage and health.

2 Short Answer (30 points, 10 each)

1. A survey of 100 male and 400 female MIT grads reveals that 60 of the men and 200 of the women work in a Big Tech firm like Google or Facebook. Show how to use this information to compute the standard error for the male-female difference in Big Tech employment rates by sex. Call your estimate se; show how how to use se to construct an approximate 95% confidence interval for the difference.

2. Many MIT students are torn between an Economics (Course 14) major and a Computer Science (Course 6) major.

   (a) Let Y_{1i} represents the potential post-graduation earnings of MIT students who major in Course 14 and let Y_{0i} represent the potential post-graduation earnings of MIT students who major in Course 6. Let D_i be a dummy that indicates the Course 14 majors in a population limited to Course 14 and Course 6 majors. Assuming Y_{1i} and Y_{0i} are defined for everyone regardless of major choice, use this notation to define the average causal effect of doing a Course 14 major on Course 14 majors' earnings.

   (b) Let
   $$\beta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0].$$
   Show that β is likely to be a biased measured of the causal effect of majoring in 14.

3. Use the fact that residuals have expectation zero and are uncorrelated with regressors to show that the slope in a regression of Y_i on X_i is $\frac{C(Y_i,X_i)}{V(X_i)}$. Suggestion: Work out your answer on a separate piece of paper. Try to keep it brief. Then, writing in Gradescope, use either latex for mathematical symbols or the symbols a for intercept, b for slope, and $Ybar$ and $Xbar$ for means of Y_i and X_i.

3 Longer Questions (40 points, 20 each)

1. This question refers to analysis of data on married women from the 1980 Census. The women in this sample all have at least two children; some have 3 or more. The descriptive table at the top of the log below describes the following variables: weeks worked last year (*weeksm1*), number of children (*kidcount*), a dummy indicating women with 3 or more children (*morekids*), years of schooling (*educm*), a dummy for women with a college degree or higher (*colgrad_c*). The rest of the log reports three bivariate regression results.

```
1 . sum weeksm1 kidcount morekids educm colgrad

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     weeksm1 |    254,654    19.01833    21.86728          0         52
   kidcount |    254,654    2.507799    .7693323          2         12
   morekids |    254,654    .3805634    .4855263          0          1
      educm |    254,654    12.38582    2.450936          0         20
  colgrad_c |    254,654    .1316806    .3381439          0          1

2 .
3 . regress weeksm1 morekids

      Source |       SS           df       MS      Number of obs   =   254,654
-------------+----------------------------------   F(1, 254652)    =   3696.02
       Model |  1742078.14         1  1742078.14   Prob > F        =    0.0000
    Residual |  120027337    254,652  471.338679   R-squared       =    0.0143
-------------+----------------------------------   Adj R-squared   =    0.0143
       Total |  121769415    254,653  478.177816   Root MSE        =    21.71

------------------------------------------------------------------------------
     weeksm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -5.386996   .0886093   -60.79   0.000    -5.560667   -5.213324
       _cons |   21.06843   .0546629   385.42   0.000     20.96129    21.17557
------------------------------------------------------------------------------

4 . regress weeksm1 colgrad_c

      Source |       SS           df       MS      Number of obs   =   254,654
-------------+----------------------------------   F(1, 254652)    =    68.59
       Model |  32787.929         1  32787.929   Prob > F        =    0.0000
    Residual |  121736627    254,652  478.050938   R-squared       =    0.0003
-------------+----------------------------------   Adj R-squared   =    0.0003
       Total |  121769415    254,653  478.177816   Root MSE        =    21.864

------------------------------------------------------------------------------
     weeksm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   colgrad_c |    1.06116   .128133     8.28   0.000     .8100231    1.312298
       _cons |   18.8786   .0464967   406.02   0.000     18.78747    18.96973
------------------------------------------------------------------------------

5 . regress weeksm1 educm

      Source |       SS           df       MS      Number of obs   =   254,654
-------------+----------------------------------   F(1, 254652)    =   1102.90
       Model |  525110.698         1  525110.698   Prob > F        =    0.0000
    Residual |  121244305    254,652  476.117622   R-squared       =    0.0043
-------------+----------------------------------   Adj R-squared   =    0.0043
       Total |  121769415    254,653  478.177816   Root MSE        =    21.82

------------------------------------------------------------------------------
     weeksm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      educm |   .5858939   .0176421    33.21   0.000     .5513158     .620472
       _cons |   11.76156   .2227491    52.80   0.000     11.32498    12.19814
------------------------------------------------------------------------------
```

   (a) Whats the probability a woman with 2 or more children has more than 2?

(b) Interpret (one sentence for each) the slope coefficient magnitudes in these regression results (ignore intercepts).

(c) Are any of these regressions likely to capture a causal relationship?

(d) Regressions 3 and 4 below shows results from a multivariate regression of weeks worked on *educm* and *morekids*, followed by a bivariate regression of *morekids* on educm. Use the omitted variables bias formula to explain the change in coefficients from the regression of *weeksm1* on *educm* with and without *morekids* control.

```
2 .
3 . regress weeksm1 educm morekids
```

Source	SS	df	MS	Number of obs	=	254,654
				F(2, 254651)	=	2175.42
Model	2045544.77	2	1022772.39	Prob > F	=	0.0000
Residual	119723871	254,651	470.148834	R-squared	=	0.0168
				Adj R-squared	=	0.0168
Total	121769415	254,653	478.177816	Root MSE	=	21.683

weeksm1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educm	.449545	.0176944	25.41	0.000	.4148645	.4842255
morekids	-5.079497	.0893212	-56.87	0.000	-5.254565	-4.90443
_cons	15.38342	.230329	66.79	0.000	14.93198	15.83486

```
4 . regress morekids educm
```

Source	SS	df	MS	Number of obs	=	254,654
				F(1, 254652)	=	4763.17
Model	1102.23662	1	1102.23662	Prob > F	=	0.0000
Residual	58928.6001	254,652	.231408354	R-squared	=	0.0184
				Adj R-squared	=	0.0184
Total	60030.8368	254,653	.235735832	Root MSE	=	.48105

morekids	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educm	-.026843	.0003889	-69.02	0.000	-.0276053	-.0260807
_cons	.7130359	.0049108	145.20	0.000	.7034109	.7226608

```
5 . regress weeksm1 educm agem1 kidcount
```

Source	SS	df	MS	Number of obs	=	254,654
				F(3, 254650)	=	2737.94
Model	3804978.76	3	1268326.25	Prob > F	=	0.0000
Residual	117964437	254,650	463.241455	R-squared	=	0.0312
				Adj R-squared	=	0.0312
Total	121769415	254,653	478.177816	Root MSE	=	21.523

weeksm1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educm	.1453183	.0181865	7.99	0.000	.1096732	.1809634
agem1	.8018506	.0130619	61.39	0.000	.7762497	.8274516
kidcount	-3.81788	.0570737	-66.89	0.000	-3.929742	-3.706017
_cons	2.422062	.4274962	5.67	0.000	1.584181	3.259943

(e) Regression 5 above reports results from a longer regression on *educm* that includes controls for age of mother (*agem*1) and number of kids (*kidcount*). Briefly explain how to compute the coefficient on *educm* in this model using a two-step regression procedure in which the second step is a bivariate regression.

2. School vouchers are government-funded coupons that cover the cost of schooling. Legendary economist Milton Friedman proposed replacing traditional government-run public schools with school vouchers that families can use to attend private schools. Since Friedman's original proposal, a number of US states and some Latin American countries have experimented with vouchers. Angrist, Bettinger, and Kremer (2002) and Angrist, et al. (2006) study voucher effects on educational attainment using data from a voucher lottery in Colombia. In Colombia, voucher recipients were chosen randomly from a pool of applicants (much as with health insurance in the OregonHealth Plan we read about in class).

   (a) Consider Table 1 from the 2006 paper (reproduced below). This table reports means and differences in applicant characteristics between voucher winners and losers. Variables examined include a dummy for whether the student has a valid national ID number, age, and dummies for boys and whether the applicant has a phone.

TABLE 1—CHARACTERISTICS OF ICFES MATCHING SAMPLE BY VOUCHER STATUS

	Means		Difference by voucher status (winners vs. losers)			
	Full sample (1)	Sample w/valid age (2)	Full sample (3)	Sample w/ valid age (4)	Valid ID and age (5)	Valid ID and age and has phone (6)
Won voucher	0.588	0.585				
Valid ID	0.876	0.967	−0.010	0.001	—	—
			(0.010)	(0.006)		
Age at time of application	12.7	12.7	−0.137	−0.086	−0.085	−0.091
	(1.8)	(1.3)	(0.064)	(0.045)	(0.044)	(0.047)
Male	0.487	0.493	0.004	0.011	0.012	0.008
			(0.016)	(0.017)	(0.017)	(0.018)
Phone	0.882	0.886	0.013	0.008	0.008	—
			(0.010)	(0.011)	(0.011)	
N	4,044	3,661	4,044	3,661	3,542	3,139

Notes: Robust standard errors are reported in parentheses in columns 3–6. Regression estimates of differences by voucher status in column 4 are for the sample with valid age data embedded in the national ID number. A valid age must be between 9 and 25. Column 5 reports results for a sample limited to those with a valid ID check digit and column 6 shows results for a sample further limited to those with a phone. The sample includes applicants from the 1995 lottery cohort.

   i. What proportion of the full sample has a phone? Does phone ownership differ significantly between winners and losers?

   ii. Looking at all variables and samples described in the table, how plausible is the claim that vouchers were indeed awarded randomly?

(b) Next, consider Table 2 from the same study (reproduced below). This table reports treatment effects on the likelihood of something called an "ICFES match." In Colombia, students who graduate high school and hope to go to college take a matriculation exam called the ICFES. The 2006 study therefore uses ICFES matching (test taking) as a proxy for high school graduation.

TABLE 2—VOUCHER STATUS AND THE PROBABILITY OF ICFES MATCH

	Exact ID match		ID and city match		ID and 7-letter name match		ID, city, and 7-letter match	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. All applicants (N = 3542)								
Dependent var. mean	0.354		0.339		0.331		0.318	
Voucher winner	0.072	0.059	0.069	0.056	0.072	0.059	0.068	0.056
	(0.016)	(0.015)	(0.016)	(0.014)	(0.016)	(0.014)	(0.016)	(0.014)
Male		−0.052		−0.053		−0.043		−0.045
		(0.014)		(0.014)		(0.014)		(0.014)
Age		−0.160		−0.156		−0.153		−0.149
		(0.005)		(0.005)		(0.005)		(0.005)
B. Female applicants (N = 1789)								
Dependent var. mean	0.387		0.372		0.361		0.348	
Voucher winner	0.067	0.056	0.069	0.057	0.071	0.060	0.073	0.062
	(0.023)	(0.021)	(0.023)	(0.021)	(0.023)	(0.021)	(0.023)	(0.021)
Age		−0.168		−0.164		−0.160		−0.156
		(0.006)		(0.006)		(0.006)		(0.006)
C. Male applicants (N = 1752)								
Dependent var. mean	0.320		0.304		0.302		0.288	
Voucher winner	0.079	0.063	0.071	0.055	0.074	0.059	0.065	0.050
	(0.022)	(0.020)	(0.022)	(0.020)	(0.022)	(0.020)	(0.022)	(0.020)
Age		−0.153		−0.148		−0.146		−0.141
		(0.007)		(0.007)		(0.007)		(0.006)

Notes: Robust standard errors are shown in parentheses. The sample used to construct this table includes all Bogotá applicants with valid ID numbers and valid age data (i.e., ages 9 to 25 at application). The sample is the same as in Table 1, column 5.

i. Focusing on the first two columns, what's the overall ICFES match rate?

ii. (Continue to focus on the first two columns) Do vouchers appear to boost ICFES match rates? Are the estimated effects on match rates significantly different from zero?

iii. Are the estimated treatment effects in columns 1 and 2 robust to inclusion of controls? Briefly explain why or why not.

(c) [Extra credit] Consider Table 3 from the 2002 paper (reproduced below). Each row of this table reports regression estimates of treatment effects on different outcomes, ranging from effects on scholarship use in the first row to effects on the number of years enrolled at the bottom. Columns 2-4 show voucher treatment effects for applicants from Bogota while columns 5-6 show estimates computed using data from Bogota and an additional town. Within samples, different columns report estimates from regressions with different sets of controls.

TABLE 3—EDUCATIONAL OUTCOMES AND VOUCHER STATUS

	Bogotá 1995				Combined sample	
Dependent variable	Loser means (1)	No controls (2)	Basic controls (3)	Basic +19 barrio controls (4)	Basic controls (5)	Basic +19 barrio controls (6)
Using any scholarship in survey year	0.057 (0.232)	0.509 (0.023)	0.504 (0.023)	0.505 (0.023)	0.526 (0.019)	0.521 (0.019)
Ever used a scholarship	0.243 (0.430)	0.672 (0.021)	0.663 (0.022)	0.662 (0.022)	0.636 (0.019)	0.635 (0.019)
Started 6th grade in private	0.877 (0.328)	0.063 (0.017)	0.057 (0.017)	0.058 (0.017)	0.066 (0.016)	0.067 (0.016)
Started 7th grade in private	0.673 (0.470)	0.174 (0.025)	0.168 (0.025)	0.171 (0.024)	0.170 (0.021)	0.173 (0.021)
Currently in private school	0.539 (0.499)	0.160 (0.028)	0.153 (0.027)	0.156 (0.027)	0.152 (0.023)	0.154 (0.023)
Highest grade completed	7.5 (0.960)	0.164 (0.053)	0.130 (0.051)	0.120 (0.051)	0.085 (0.041)	0.078 (0.041)
Currently in school	0.831 (0.375)	0.019 (0.022)	0.007 (0.020)	0.007 (0.020)	−0.002 (0.016)	−0.002 (0.016)
Finished 6th grade	0.943 (0.232)	0.026 (0.012)	0.023 (0.012)	0.021 (0.011)	0.014 (0.011)	0.012 (0.010)
Finished 7th grade (excludes Bogotá 97)	0.847 (0.360)	0.040 (0.020)	0.031 (0.019)	0.029 (0.019)	0.027 (0.018)	0.025 (0.018)
Finished 8th grade (excludes Bogotá 97)	0.632 (0.483)	0.112 (0.027)	0.100 (0.027)	0.094 (0.027)	0.077 (0.024)	0.074 (0.024)
Repetitions of 6th grade	0.194 (0.454)	−0.066 (0.024)	−0.059 (0.024)	−0.059 (0.024)	−0.049 (0.019)	−0.049 (0.019)
Ever repeated after lottery	0.224 (0.417)	−0.060 (0.023)	−0.055 (0.023)	−0.051 (0.023)	−0.055 (0.019)	−0.053 (0.019)
Total repetitions since lottery	0.254 (0.508)	−0.073 (0.028)	−0.067 (0.027)	−0.064 (0.027)	−0.058 (0.022)	−0.057 (0.022)
Years in school since lottery	3.7 (0.951)	0.058 (0.052)	0.034 (0.050)	0.031 (0.050)	0.015 (0.044)	0.012 (0.043)
Sample size	562		1,147		1,577	

Notes: The table reports voucher losers' means and the estimated effect of winning a voucher. Numbers in parentheses are standard deviations in the column of means and standard errors in columns of estimated voucher effects. The samples used to estimate 7th- and 8th-grade completion effects exclude Bogotá 1997. The sample size for these outcomes is 1,304 in columns (5) and (6). The regression estimates are from models that include controls for city, year of application, phone access, age, type of survey and instrument, strata of residence, and month of interview.

i. Give a rationale for reporting results on intermediate outcomes like whether winners were more likely than losers to use a scholarship or to enroll in private school.

ii. Briefly assess program impact: does it seem like vouchers boosted winners' human capital as Milton Friedman would have hoped?