MRU Mastering Econometrics **M. Joshway**
Spring 2020

Problem Set 2

1. Suppose the CEF of Y_i given X_i is linear:

$$E[Y_i|X_i] = a + bX_i. \tag{1}$$

We know from the regression-CEF theorem that $a = \alpha$ and $b = \beta$, where Greek letters denote the regression intercept and slope. Typically, we estimate β with the OLS estimator, $\hat{\beta}_{OLS}$. But there are many ways to fit a line. Here's one: (a) split the data in half by dividing the sample into observations with values above and below median X_i. Compute above-median and below-median average Y_i and X_i; call these $\bar{y}_1, \bar{x}_1$ for means above and $\bar{y}_0, \bar{x}_0$ for means below. Define an alternative slope estimator

$$\hat{\beta}_w = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

   (a) Show that $\hat{\beta}_w$ is an unbiased estimator of the regression slope (hint: use the law of iterated expectations)

   (b) Treating the X_i as fixed in repeated random samples, derive a formula for the sampling variance of $\hat{\beta}_w$.

   (c) Suppose the X_i are fixed in repeated random samples and the residuals in (1) are homoskedastic. (i) Compare the sampling variance of $\hat{\beta}_{OLS}$ and $\hat{\beta}_w$ (Hint: no math required).

   (d) Challenging: using the same assumptions as in (c), confirm your claim by deriving a formula for the sampling variance of $\hat{\beta}_w$ and comparing this formula to the sampling variance of $\hat{\beta}_{OLS}$ (hint: $(\bar{x}_1 - \bar{x}_0)^2$ is proportional to the variance of fitted values from a regression of X_i on a dummy that indicates values of X_i above the median).

2. Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of y_i on $x_{i1}, \ldots, x_{ik}$ for $i = 1, 2, \ldots, n$, where $\hat{\beta}_0$ is the intercept. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$ for $i = 1, 2, \ldots, n$ are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. (Hint: Use the fact that the $\tilde{\beta}_j$ satisfy OLS first order conditions for the rescaled dependent and independent variables.)

3. Regression in practice

   (a) This question uses replication data from Angrist, Lang, and Oreopoulos (2009) posted in the Angrist Data Archive (https://economics.mit.edu/faculty/angrist/data1/data/angrist,1)

      i. Using Stata's `ttest` command for t-tests, compare fall grades (*grade_20059_fall*) in the control group to each of the 3 treatment groups (*ssp*, *sfp*, and *sfsp*). For each test, report the difference in means, the standard error of the difference in means, and the t-statistic. Are any of these differences statistically significant?

      ii. Use Stata's regression command (`regress`) to compute the same pairwise comparisons and the associated test statistics

   (b) Estimate a multivariate regression of fall grades on all three treatment dummies (*ssp*, *sfp*, and *sfsp*).

      i. How do these coefficients compare to those obtained in part (ii) above?

      ii. Compare your results here to the estimates reported in the first column (top panel) of Table 5 in ALO (2009). Why do your estimates differ?

   (c) Code a dummy variable for students belonging to either the SFP group or the SFSP group, and regress fall grades on the SSP dummy and this dummy. Interpret the magnitude of the coefficient on this new variable.

(d) Add controls for gender, high school GPA, mother education, and father education to the regression you run in part (d). Does including these controls matter for estimated treatment effects from the STAR experiment? Explain.

4. Practice makes perfect

(a) Return to the NHIS data used in PS1. As before, start by selecting the sample used to produce MM Table 1.1. Use regression to compare average health for husbands with and without health insurance. Construct the relevant confidence intervals and comment on the precision of this estimate.

(b) Differences in health between those with and without insurance may be due to differences between the insured and uninsured population that exist in the absence of insurance. Regression-adjust your comparison by insured status by sequentially adding controls for age, years of education, and income. Explain and interpret changes in the insurance coefficient as you add controls.

(c) Return to the RAND HIE dataset "rand_initial_sample_2.dta" used in PS1. Define a dummy variable called *anyins*, which is equal to 1 for individuals with Plan Types 1-3 ("any insurance") and equal to 0 for individuals with Plan Type 4 (those with only "catastrophic" insurance). Regress the general health index *ghindx* (a health index similar to that in the NHIS, scaled differently) on *anyins*. Comment on the impact of including these controls on the *anyins* coefficient. How and why do the consequences of additional controls differ from what you saw in question 3b?

- More awesome apps

1. This question asks about regression results that explore the relationship between childbirth and labor force participation for women aged 25-44 in the 2020 ACS. The variables *working, has_kids,* and *college* are dummies for, respectively, being employed, having kids living at home, and having a BA.

```
.
. * Summary Statistics:
. sum working has_kids college
```

Variable	Obs	Mean	Std. Dev.	Min	Max
working	381,199	.7801883	.4141195	0	1
has_kids	381,199	.6038447	.4890981	0	1
college	381,199	.4340017	.4956257	0	1

```
.
. * Regression 1
. reg working has_kids
```

Source	SS	df	MS	Number of obs	=	381,199
				F(1, 381197)	=	4048.14
Model	686.942347	1	686.942347	Prob > F	=	0.0000
Residual	64686.5957	381,197	.169693349	R-squared	=	0.0105
				Adj R-squared	=	0.0105
Total	65373.5381	381,198	.171494966	Root MSE	=	.41194

working	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
has_kids	-.0867938	.0013641	-63.62	0.000	-.0894675	-.0841202
_cons	.8325983	.00106	785.44	0.000	.8305206	.834676

```
.
. * Regression 2
. reg has_kids college
```

Source	SS	df	MS	Number of obs	=	381,199
				F(1, 381197)	=	5103.03
Model	1204.60805	1	1204.60805	Prob > F	=	0.0000
Residual	89984.3971	381,197	.236057464	R-squared	=	0.0132
				Adj R-squared	=	0.0132
Total	91189.0052	381,198	.239216904	Root MSE	=	.48586

has_kids	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
college	-.1134211	.0015877	-71.44	0.000	-.116533	-.1103091
_cons	.6530696	.001046	624.36	0.000	.6510195	.6551197

   (a) Interpret the coefficient on *has_kids* in Regression 1. Is this likely to be a causal relationship? Briefly explain why or why not.

   (b) What does Regression 2 show? Why is this important for our understanding of Regression 1?

   (c) In the covariance matrix below, *resid* is the residual obtained from a regression of *has_kids* on *college*, while *fitted* are the corresponding fitted values. Use this covariance matrix to compute the coefficient on *has_kids* in a model for fertility effects on employment that controls for *college.* Check your work against Regression 3, below.

```
. * Covariance:
. corr working has_kids college resid fitted, cov
(obs=381,199)
```

	working	has_kids	college	resid	fitted
working	.171495				
has_kids	-.020763	.239217			
college	.036191	-.027861	.245645		
resid	-.016658	.236057	-2.2e-09	.236057	
fitted	-.004105	.00316	-.027861	2.4e-10	.00316

```
.
.
. * Regression 3
. reg working has_kids college
```

Source	SS	df	MS		
Model	2480.63452	2	1240.31726	Number of obs	= 381,199
Residual	62892.9036	381,196	.164988362	F(2, 381196)	= 7517.60
				Prob > F	= 0.0000
				R-squared	= 0.0379
				Adj R-squared	= 0.0379
Total	65373.5381	381,198	.171494966	Root MSE	= .40619

working	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
has_kids	-.0705667	.0013541	-52.11	0.000	-.0732206	-.0679128
college	.1393261	.0013362	104.27	0.000	.1367071	.1419451
_cons	.7623319	.0012437	612.97	0.000	.7598943	.7647694

2. In many poor countries, women marry and begin childbearing at young ages. Economists are interested in whether this behavior reflects a lack of employment opportunities for women. Jensen (2012) uses a randomized trial to estimate effects of employment opportunities on women in rural India. This experiment offered job search assistance to women in randomly selected villages.[1]

   Estimated treatment effects on on women's marital and fertility status appear in columns (1) and (2) of the table below. These are from regressions with no independent variables other than a treatment dummy.

TABLE V

EFFECT OF THE INTERVENTION ON MARRIAGE AND FERTILITY, AGES 18–24 IN ROUND 2

	(1) Married	(2) Had child	(3) Desired fertility
Panel A: Women			
Treatment	−0.051**	−0.057**	−0.35***
	(0.024)	(0.026)	(0.078)
R^2	0.003	0.003	0.018
Observations	1,278	1,278	1,226
Control group mean	0.71	0.43	3.0

   (a) What fraction of women in the control group were married at follow up? What fraction were married in the treatment group? Is the difference by treatment status statistically significant?

[1] Jensen, Robert. (2012). "Do Labor Market Opportunities Affect Young Women's Work and Family Decisions? Experimental Evidence from India." *The Quarterly Journal of Economics*, 127(2): 753–92.

(b) The table below reports treatment effects from regressions that control for a number of variables measured at baseline, including log per capita expenditure (denoted *log(expend pc)*), parents' education, and family size. The effects of treatment on martial and fertility status conditional on these covariates are reported in columns (4) and (5) below. Why are these similar to the estimates with no controls in columns (1) and (2) of the previous table? (Note: "gave birth" in this table is the same as "had child" in the previous table).

APPENDIX B

EFFECTS OF THE INTERVENTION ON WOMEN: ADDING BASELINE CONTROLS AND USING CHANGES

	Working age (18–24)						School age (as indicated)		
	(1) Works BPO	(2) Works for pay	(3) Training	(4) Married	(5) Gave birth	(6) Desired fertility	(7) In school (6–17)	(8) BMI for age (5–15)	(9) Height for age (5–15)
Panel A: Baseline Controls									
Treatment	0.048***	0.023**	0.028***	−0.055**	−0.061**	−0.37***	0.050***	0.20***	0.058
	(0.009)	(0.011)	(0.008)	(0.024)	(0.026)	(0.076)	(0.015)	(0.070)	(0.064)
log(expend pc)	0.023***	0.018	0.023***	−0.069***	−0.064***	0.083	0.038***	0.11*	0.17***
	(0.007)	(0.019)	(0.008)	(0.019)	(0.021)	(0.062)	(0.015)	(0.058)	(0.057)
Head's educ	−0.003	0.004	0.002*	−0.006	−0.011**	−0.024**	0.004*	0.018	0.008
	(0.002)	(0.004)	(0.001)	(0.004)	(0.005)	(0.012)	(0.002)	(0.012)	(0.010)
Spouse's educ	0.004	0.005	−0.002*	−0.003	−0.004	−0.011	0.006	−0.034*	0.035**
	(0.003)	(0.006)	(0.001)	(0.005)	(0.005)	(0.015)	(0.004)	(0.017)	(0.014)
Family size	0.002	0.004	0.005*	−0.006	−0.007	−0.011	0.002	−0.011	0.025**
	(0.001)	(0.005)	(0.002)	(0.005)	(0.005)	(0.015)	(0.003)	(0.013)	(0.012)
R^2	0.036	0.023	0.031	0.11	0.078	0.029	0.070	0.017	0.06
Observations	1,278	1,278	1,278	1,278	1,278	1,226	2,264	2,031	2,031

3. The MM site contains a CSV data set (PS4.csv) with observations on log weekly wages, log hourly wages, age, sex (1=male), race (1=White, 2=Black, 3=Native American, 4= Asian or Pacific Islander, 5=Other), and years of schooling for men and women aged 25-50 in the March 1992 CPS.

   Labor economists often replace age with potential experience, which is potential time working adjusting for schooling. Wages generally increase as we get more experience working, though not necessarily at a constant rate. Construct a measure of potential work experience by defining:

$$experience = age - education - 6$$

   (a) Why does this constructed experience variable not measure actual labor market experience? Check the distribution of potential experience. Set implausible values to missing, or to a plausible value that seems consistent with the underlying data.

   (b) Compute the multivariate regression of log hourly wages on a dummy for women, a full set of race dummies, a quadratic function of potential work experience, and years of schooling.

   (c) Plot the estimated experience profile, and use calculus to compute the level of experience at which earnings peak (according to this model). How old is a college graduate who reaches this level of experience?

   (d) Re-estimate the model allowing the relationship between potential experience and wages to differ by sex. Construct an F-test for the null hypothesis that the relationship between potential experience and wages is the same for men and women. How does the female main effect (coefficient on F_i) change when potential experience coefficients are free to differ by sex? Can you explain this?

   (e) Allow schooling coefficients to differ by both race and sex, with a common experience profile for all. Use an F-test to evaluate the hypothesis that schooling coefficients are the same across racial groups, allowing for differences by sex.

4. Ashenfelter and Rouse (1998) use data collected from identical twins to estimate returns to schooling. The idea is to compare education and income within pairs of twins, thereby controlling for their shared family background and similar (or even identical) genetic heritage.

(a) Download data file ar98.dta the MM site. These file has individual-level wages and schooling for 340 twin pairs (680 total observations). Regress log wages (*lwage*) on years of schooling (*educ*), age (*age*), age-squared (*age2*), and dummies for female (*female*) and white (*white*). Interpret the estimated schooling and age coefficients.

(b) Consider the regression model

$$\ln Y_{if} = \alpha' X_{if} + \beta S_{if} + \gamma A_f + \epsilon_{if},$$

where f stands for family and subscript $i = 1, 2$ indexes twins in family f. The vector X_{if} includes the covariates from part (a), while A_f is an *unobserved* ability variable assumed to be fixed within families.

i. What's the rationale for this model?

ii. Show (mathematically) that a regression of the within-family difference in log wages on the corresponding difference in schooling eliminates family-specific ability bias. What's the key assumption behind this trick?

(c) Run the regression suggested by your analysis in part (b). What do these results suggest about the direction of ability bias in the undifferenced model?

(d) What should the constant be in the differenced model? How does this work out in practice?

5. This problem asks you to replicate and extend the Krueger (1993) study of the effects of computer use on wages using CPS extracts posted on MM in k93.dta. Many variables are constructed already, but you should note that the K93 paper employs sample restrictions (e.g. age and work status for all tables, plus maximum/minimum wages allowed in sample for Table 2 onward) that you'll need to use as well (see Appendix A of K93 for details, though note also that the \$999 vs. \$1,923 top-coding issue is taken care of).

Report your replication alongside the original results in a format similar to the original with added columns for the replication. You won't get exactly the same results. Still, your coefficient estimates should be close to the original and of similar precision. For example, a coefficient of 0.090 with a standard error of (0.025) is close to a coefficient of 0.095 with a standard error of (0.029).

(a) Reproduce Table I, with the exception of the occupation means (the definition of these variables is not clear). The K93 part-time variable differs from ours, but everything else should closely match.

(b) Reproduce Table II. You should be able to match this closely except for the "other race" coefficient and the intercept.

(c) Reproduce Table III. You should be able to match this reasonably closely.

(d) Re-estimate the regressions in Table II, Column 6 without region dummies. Do the computer-use effects change much? Use Table I to explain why. Use the R^2 version of the F-test to test the joint significance of region effects (check this with Stata's `test` command).

(e) Is serial correlation an issue in Tables II and III? Why or why not? What about heteroskedasticity? For the estimates in Columns 1 and 4 of Table II, report old-fashioned (regular) and heteroskedasticity-consistent (robust) standard errors. Is this a big deal?

(f) Estimate versions of the Table II models in columns 2 and 5 allowing both the computer-use and schooling coefficients to vary by sex. Test these interactions one at a time and jointly. Briefly discuss your findings. Finally, allow the schooling coefficient to differ both by computer use and sex. Use Stata `lincom` to estimate the returns to schooling for male and female computer users (hint: you'll need a triple interaction term).

(g) Estimate the *change* in the returns to schooling and computer use by pooling the models from columns 2 and 5 and adding interactions with year (Restrict effects of variables other than schooling and computer use to be additive). Test the significance of these changes jointly. Compare your estimated changes to those implied by the results in Table II.

6. The Tennessee Student/Teacher Achievement Ratio experiment (aka Project STAR) is a legendary randomized trial investigating the relationship between class size and student achievement. Kindergarten students and their teachers were randomly assigned to one of three class size groups at the beginning of the 1985-1986 school year. This question uses data from Krueger (1999), a study evaluating STAR.

   (a) Download data file k99.dta from the MM site. This contains information about STAR participants and includes data on their test scores (*pscore*), class size (*cs*), and an identifier for each class (*classid*). Regress test scores on class size. Interpret the results.

   (b) Re-estimate the relationship between test scores and class size using heteroskedasticity-robust standard errors. Compare these robust standard errors with the old-fashioned standard errors from part (a).

   (c) Re-estimate the relationship between test scores and class size, clustering your standard errors at the class-level. Why should you cluster? What happens when you do?

   (d) Re-estimate the relationship between test scores and class size by collapsing the data from student-level to class-level, and regressing average scores on class size, weighting by the number of observations in each class. How do these standard errors compare to those estimated in parts (b) and (c)?