Problem Set 3
Econ 140, Spring 2020

Due by 2pm on Tu March 10. No late problem sets accepted, sorry!

Problem 1. Multivariate linear regression.

This problem will create a dataset by having Stata generated variables in the same way as we did in Problem Set 2. A main advantage of such an exercise is that we can control the true data generating process ("DGP"), which is not possible in practical econometric analysis.

(a) Set the sample size at 1,000 and generate an error term, u, by randomly selecting from a normal distribution with mean 0, and variance $25 = 5^2$. Draw an explanatory variable, X_{1i}, from a standard normal distribution, $N(0, 1)$, and then define a second explanatory variable, X_{2i}, to be equal to $\exp(X_{1i})$ for all i. Finally, set the dependent variable to be linearly related to the two regressors plus an additive error term: $Y_i = 2 + 4X_{1i} - 6X_{2i} + u_i$. Note that, by construction, the error term of this multivariate linear regression is homoskedastic.

(b) Regress Y on X_1 with homoskedasticity-only standard errors. Do the same analysis for Y and X_2. Compare the results with the true data generating process. Explain why differences arise between the population slopes and the estimated slopes, if there are any.

(c) Next, regress Y on both X_1 and X_2. Compare the estimation results with those in (b), especially the model with only the regressor X_1. Examine differences across the three regressions in terms of the coefficient estimates, their standard errors, the R^2, and the adjusted R^2.

(d) Generate a third regressor: $X_{3i} = 1 + X_{1i} - X_{2i} + v_i$ where v_i is drawn from a normal distribution with mean 0 and variance $0.25 = 0.5^2$. Estimate the model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + w_i$. Compare the result with (c). Do changes in OLS estimates, standard errors, the R^2, and the adjusted R^2 make sense to you? Explain why or why not. [Hint: think about the concept of "imperfect multicollinearity."]

(e) [Extra Credit] This question explores the sampling distribution of the OLSEs. Repeat the regression in part (c) 1,000 times, drawing new values for the Y, X_1, X_2, and u variables each time as specified above. Compute the means of the estimated coefficients and also their correlations. *Do not invest a lot of time on this question*, but you may draw inspiration from the example Stata .do file posted along with the problem set. In that .do file, the Central Limit Theorem is simulated by using the "looping" routine to take a random sample 1,000 times as you are asked to do here.

Problem 2. Teaching ratings.

Recall the data file **TeachingRatings.dta** which contains data on course evaluations, course characteristics, and professor characteristics for 463 courses at the University of Texas at Austin. [1] As you will also recall, one of the characteristics is an index of the professor's "beauty" as rated by a panel of six judges. The variable `course_eval` is an overall teaching evaluation score, on a scale of 1 (very unsatisfactory) to 5 (excellent). In this exercise, you will investigate how course evaluations are related to the professor's beauty.

(a) Run a regression of `course_eval` on `beauty` using robust standard errors. What is the estimated slope? Is it statistically significant?

(b) Run a regression of `course_eval` on `beauty`, including some additional variables to control for the type of course and professor characteristics. In particular, include as additional regressors `intro`, `onecredit`, `female`, `minority`, and `nnenglish`. What is the estimated effect of beauty on `course_eval`? Does the regression in (a) suffer from important omitted variable bias (OVB)? What happens with the R^2? Based on the confidence interval from the regression, can you reject the null hypothesis that the effect of beauty is the same as in part (a)? What can you say about the effect of the new variables included?

(c) Estimate the coefficient on `beauty` for the multiple regression model in (b) using the three-step process in Appendix 6.3 (the *Frisch-Waugh theorem*). Verify that the three-step process yields the same estimated coefficient for `beauty` as that obtained in (b). Comment.

(d) Professor Smith is a black male with average beauty and is a native English speaker. He teaches a three-credit upper-division course. Predict Professor Smith's course evaluation.

[1] These data were provided by Prof. Hammermesh of the University of Texas at Austin and were used in his paper with Amy Parker, "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review*, August 2005, 24(4): 369–376.

Problem 3. Education and distance to college.

The data file **CollegeDistance.dta** contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. [2] A detailed description is given in **CollegeDistance_Description.pdf**. In this exercise, you will use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student's high school to the nearest four-year college.

(a) What do you expect for the sign of the relationship and what mechanism can you think about to explain it?

(b) Run a regression of years of completed education (`yrsed`) on distance to the nearest college (`dist`), measured in tens of miles (For example, `dist` = 2 means that the distance is 20 miles). What is the estimated slope? Is it statistically significant? Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

(c) Now run a regression of `yrsed` on `dist`, but include some additional regressors to control for characteristics of the student, the student's family, and the local labor market. In particular, include as additional regressors: `bytest`, `female`, `black`, `hispanic`, `incomehi`, `ownhome`, `dadcoll`, `momcoll`, `cue80`, and `stwmfg80`. What is the estimated effect of `dist` on `yrsed`? Is it substantively different from the regression in (b)? Based on this, does the regression in (b) seem to suffer from important omitted variable bias?

(d) The value of the coefficient on `dadcoll` is positive. What does this coefficient measure? Interpret this effect.

(e) Explain why `cue80` and `stwmfg80` appear in the regression. Are the signs of their estimated coefficients what you would have believed? Explain.

(f) Bob is a black male. His high school was 20 miles from the nearest college. His base-year composite test score (`bytest`) was 58. His family income in 1980 was $26,000, and his family owned a home. His mother attended college, but his father did not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was $9.75. Predict Bob's years of completed schooling using the regression in (c).

[2] These data were used in the paper by C. Rouse (1995) "Democratization or Diversion? The Effect of Community Colleges on Educational attainment," *Journal of Business and Economic Statistics.*

Problem 4. The sheepskin effect.

The table of results on the last page of the problem set is copied from a paper by Jaeger and Page (1996) entitled "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*. The question is whether employers pay in relation to years of education or whether there is an additional premium for obtaining a degree. Such premium might be called the "sheepskin effect" (because diplomas at one time were printed on a sheet of sheepskin) or the "diploma effect." The Jaeger and Page paper estimates the magnitude of this effect. [Note: an empty cell in the table means that variable is not included in the regression].

(a) Why do you think Jaeger and Page estimate their model using only data on white men (at least in this table)?

(b) Look at column (3) of the table. In words, interpret the coefficient on the dummy variable "9" [Hint: note that "12" is the omitted category].

(c) Why do you think the effect of the 14th year of education is larger than that of the 15th?

(d) Now look at column (4). Think about a student who is currently a senior (as is the case for many of you who are taking this course). What is the average difference in her/his wage now and the one she/he could get at the end of the year following graduation?

(e) Based on the results presented in this column, would you rather choose to complete a PhD or a professional degree? Explain.

(f) Using the results from columns (3) and (4), how would you test the presence of a "diploma effect"? Carry out the test at a 5% significance level. [Hint: you may find some of the information you need in the footnote of the table.]

TABLE 2.—ESTIMATED DIPLOMA EFFECTS FOR WHITE MEN USING DIFFERENT SPECIFICATIONS

	Model			
Coefficient	(1)	(2)	(3)	(4)
Completed Years of Education (Spline)				
Years of Education (S)	0.076	0.076		
	(0.018)	(0.018)		
$S \geq 8$	−0.141	−0.112		
	(0.080)	(0.078)		
$(S \geq 8) \cdot (S - 8)$	0.002	−0.022		
	(0.027)	(0.023)		
$S \geq 12$	0.034			
	(0.053)			
$(S \geq 12) \cdot (S - 12)$	−0.006	−0.019		
	(0.022)	(0.017)		
$S \geq 16$	0.114			
	(0.035)			
$S = 17$	−0.055			
	(0.042)			
$S = 18$	−0.006			
	(0.031)			
Completed Years of Education (Dummy)				
9			−0.227	−0.109
			(0.049)	(0.061)
10			−0.164	−0.046
			(0.040)	(0.054)
11			−0.137	−0.044
			(0.043)	(0.051)
12			ref.	ref.
13			0.089	0.020
			(0.027)	(0.033)
14			0.167	0.073
			(0.022)	(0.031)
15			0.166	0.052
			(0.038)	(0.044)
16			0.406	0.178
			(0.019)	(0.045)
17			0.422	0.164
			(0.039)	(0.057)
18 or more			0.544	0.224
			(0.023)	(0.054)
Diploma Effects				
High School		0.106		0.123
		(0.037)		(0.041)
Marginal Effect Over High School				
Some College, No Degree		0.074		0.083
		(0.022)		(0.027)
Occupational Associate's		0.074		0.076
		(0.039)		(0.043)
Academic Associate's		0.188		0.191
		(0.042)		(0.046)
Bachelor's		0.273		0.245
		(0.038)		(0.045)
Marginal Effect Over Bachelor's				
Master's		0.032		0.050
		(0.030)		(0.041)
Professional		0.271		0.286
		(0.050)		(0.059)
Doctoral		0.052		0.067
		(0.058)		(0.067)
R^2	0.145	0.153	0.147	0.154
Adjusted R^2	0.144	0.151	0.145	0.151
Mean Square Error	0.372	0.369	0.372	0.369

Note: Dependent variable is log hourly wages. Estimated using ordinary least squares. Standard errors are in parentheses. Calculated from a matched sample of individuals 25 to 64 years old from the 1991 and 1992 March Current Population Survey. Model also includes Potential Experience and Potential Experience Squared as covariates. Columns (3) and (4) also include dummy variables for zero through eight completed years of education. Sample size is 8,957.