Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due September 9, 2025
61 total points

Problem Set 2

Instructions: Answer the following questions in a Stata do-file, and submit your resulting log file via email to `sean.corcoran@vanderbilt.edu`, preferably as a .pdf or .txt file. Use your last name and problem set number as the filename (e.g., *Kelce PS2.pdf*). The resulting log should include the questions below (commented), your commands, output, and written responses. Edit this file as appropriate, with any requested interpretations of your output. Graphical output can be submitted separately, preferably as a PDF file. Working together is encouraged, but all submitted work should be that of the individual student.

This problem set will use the National Education Longitudinal Study (NELS-88) data and matching methods to estimate the academic benefits, if any, to attending a Catholic high school. The variable definitions in this dataset should be self-explanatory, but if you have any questions, just ask.

You can read the data into Stata directly using this syntax:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/catholic.dta, clear
```

1. Provide some basic descriptive information about students in this dataset. How many observations are there? What proportion attended a Catholic high school? What proportion graduated high school on time? What proportion entered post-secondary education after high school? What are the overall means and standard deviations for 12th grade math and reading scores, respectively? (**5 points**)

2. Create a few additional variables for the analysis (**5 points**):

   - The family income variable *faminc8* is an ordinal categorical variable with 12 categories. Create a "continuous" version of the family income variable *faminc8* by assigning a dollar amount equal to the midpoint of each interval. For example, \$4,000 for \$3,000-\$4,999.

   - Create a "collapsed" version of the family income variable *faminc8* in which 1=≤\$19,999, 2=\$20,000 to 34,999, and 3=\$35,000 to 74,900, corresponding to *Lo, Med,* and *Hi* income. This will allow you to replicate Tables 12.1 and 12.2 in Murnane & Willett.

   - Create a categorical version of the 8th grade math achievement variable (*math8*) with four categories corresponding to *Lo, MLo, MHi,* and *Hi* achievement. The cut points for these four categories should be 38, 44, and 51. Hint: I like to use the `egen` *`varname`*`=cut(`*`varname2`*`)` command for creating ordered categorical

variables and quantiles. This will allow you to replicate Table 12.2 in Murnane & Willett.

- Create dummy variables for each parent's *highest* level of education (<HS grad, HS grad, some college, college+). Also create dummy variables that indicate the maximum of the two parents' highest education.

3. Use this dataset to replicate the statistics found in Table 12.1 in Murnane & Willett (in the lecture notes and reproduced below). Specifically, report (**8 points**):

   - Mean (continuous) income by income strata, separately for public and Catholic school students. Also conduct t-tests for significant differences within each strata. Does income appear balanced within each strata? Note: M&W used the ordinal income variable here; you should use the continuous one you created.

   - Mean 12th grade math scores by income strata, separately for public and Catholic school students. Also conduct t-tests for significant differences within each strata.

   - The ATE and ATT estimates by calculating differences within each strata and weighting appropriately. Compare this to the simple difference in means.

4. Now replicate the statistics reported in Table 12.2 in Murnane & Willett (in the lecture notes and reproduced below), where the strata are income (3 categories) and 8th grade math achievement (4 categories). Specifically, report (**8 points**):

   - Mean 12th grade math scores by income and baseline achievement strata. Also conduct t-tests for significant differences within each strata.

   - The ATE and ATT estimates by calculating differences within each strata and weighting appropriately.

5. Use `teffects` to exact match on the 3-category family income variable used in #3 and calculate the ATE and ATT. How do these compare to your estimates in #3? What are the minimum and maximum number of exact matches? (**5 points**)

6. After exact matching in #5 use `tebalance summarize` to check for balance on your continuous family income measure (in dollars), and 8th grade math and reading scores. Note you *can* conduct balance checks on variables that were not part of your original exact matching algorithm. Explain how to read the results here. How do the Catholic and public schools students in the matched sample compare on their distributions of these variables? Note: do this after requesting the ATT, not ATE, as the results will differ. (**5 points**)

7. Do the same as #5 and #6 but exact match on both the 3-category family income variable and 4-category baseline math achievement variables used in #4. How do these compare to your answer in #4? What are the minimum and maximum number of exact matches? How do the Catholic and public school students in the matched sample compare now? (**5 points**)

8. Estimate the ATT of attending a Catholic school on two later outcomes: high school graduation and enrollment in post-secondary education. Use nearest neighbor matching (with Mahalanobis distance) on the following covariates: 8th grade math achievement, 8th grade reading achievement, family income (continuous), and the highest educational attainment of either parent. For now, just use the ordinal version of parent's educational attainment. Interpret the point estimates. What is the minimum and maximum number of nearest neighbors used? (**5 points**)

9. After nearest neighbor matching in #8 use `tebalance summarize` and `tebalance box` to check for balance on your matching variables. Use `tebalance density` to compare distributions of the two test score variables. How do the distributions compare? (**5 points**)

10. Repeat #8 but force an exact match on parent's educational attainment. Try `tebalance summarize` again. How did the exact match affect the balance, if at all? (**5 points**)

11. Repeat #8 but force an exact match on parent's educational attainment *and* increase the number of nearest neighbors to 5. Include the Abadie & Imbens bias correction for the continuous covariates. Try `tebalance summarize` again. How did the exact match affect the balance, if at all? What happened to the standard error of your ATE? (**5 points**)

12. What is the assumption necessary to interpret the matching estimator in #11 as causal? Do you believe it holds in this case? Why or why not? (**5 points**)

NOT REQUIRED (but recommended): Work through the do file on Github called *Mahalanobis distance example* and its accompanying handout by the same name. This walks you through an example using generated data that contrasts the Mahalanobis distance measure with Euclidean distance.

Table 12.1 Descriptive statistics on annual family income, by stratum, overall and by type of high school attended, and average twelfth-grade mathematics achievement by income stratum and by high-school type (n = 5,671)

Stratum		Average Base-Year Annual Family Income *(1988 dollars, 15-point ordinal scale)*			Cell Frequencies		Average Mathematics Achievement (12th grade)		
Label	Income Range	Sample Variance	Sample Mean		Public	Catholic (% of stratum total)	Public	Catholic	Diff.
			Public	Catholic					
Hi_Inc	$35,000 to $74,999	0.24	11.38	11.42	1,969	344 (*14.87%*)	53.60	55.72	2.12***,†
Med_ Inc	$20,000 to $34,999	0.22	9.65	9.73	1,745	177 (*9.21%*)	50.34	53.86	3.52***,†
Lo_Inc	≤$19,999	3.06	6.33	6.77	1,365	71 (*4.94%*)	46.77	50.54	3.76***,†
							Weighted Average ATE		***3.01***
							Weighted Average ATT		***2.74***

~p <0.10; *p <0.05; **p <0.01; ***p <0.001
†One-sided test.

Table 12.2 Sample frequencies and average twelfth-grade mathematics achievement, by high-school type, within 12 strata defined by the crossing of stratified versions of base-year annual family income and mathematics achievement (n = 5,671)

Stratum		Cell Frequencies		Average Mathematics Achievement (12th Grade)		
Base-Year Family Income	Base-Year Mathematics Achievement	Public	Catholic	Public	Catholic	Diff.
Hi_Inc	*Hi_Ach*	1,159	227	58.93	59.66	0.72
	MHi_Ach	432	73	49.18	50.71	1.53*,†
	MLo_Ach	321	38	42.75	44.23	1.48
	Lo_Ach	57	6	39.79	40.40	0.62
Med_Inc	*Hi_Ach*	790	93	57.42	59.42	2.00**,†
	MHi_Ach	469	49	47.95	50.14	2.19**,†
	MLo_Ach	390	33	41.92	44.56	2.64*,†
	Lo_Ach	96	2	37.94	39.77	1.83
Lo_Inc	*Hi_Ach*	405	36	56.12	56.59	0.47
	MHi_Ach	385	13	47.12	48.65	1.53
	MLo_Ach	433	21	40.99	41.70	0.71
	Lo_Ach	142	1	36.81	42.57	5.76
				Weighted Average ATE		***1.50***
				Weighted Average ATT		***1.31***

~p <0.10; *p <0.05; **p <0.01; ***p <0.001
†One-sided test.