STA501: Data-based Decision Making

Problem Set 4

Question 1. You are regressing y on 4 mutually exclusive and exhaustive dummy variables: a dummy for high school dropout, a dummy for high school graduate, a dummy for high school graduate plus some university, a dummy for university graduate or more. How is each coefficient interpreted? What is the formula for each estimated OLS coefficient in this regression?

Question 2. Consider the classic test of parameter stability of a linear regression model across two subsamples I and II, of size N_I and N_{II} respectively, where:

$$y_I = x_I\beta_I + \epsilon_I \quad \text{and} \quad y_{II} = x_{II}\beta_{II} + \epsilon_{II}$$

and the hypothesis of stability corresponds to $H_0 : \beta_I = \beta_{II}$. Define two dummy variables of length $N_I + N_{II}$ as follows:

$$d_{Ii} = \begin{cases} 1 & \text{if observation belongs to subsample I} \\ 0 & \text{otherwise} \end{cases}$$

and

$$d_{IIi} = \begin{cases} 1 & \text{if observation belongs to subsample II} \\ 0 & \text{otherwise} \end{cases}$$

Use these dummy variables to derive a test of H_0 using an F-statistic.

Question 3. An investigator considers the linear model:

$$\begin{aligned} \log Earnings_i = \gamma_1 + \beta_1 Age_i + \beta_2 FullTimeEmployed_i + \beta_3 Tenure_i \\ + \gamma_2 Education_i + \gamma_3 White_i + \gamma_4 Female_i + \epsilon_i \end{aligned}$$

where $FullTimeEmployed_i$ is a dummy variable indicating whether individual i was employed full time in period t, $White_i$ is a dummy indicating whether individual i is of white race, and $Female_i$ is a dummy taking the value 1 if individual i is female. The available sample of a cross-section of individuals is indexed by $i = 1, ..., N$.

1. Explain what would happen if instead of the variables $White_i$ and $Female_i$ were to use the complementary variables $Non - White_i$ and $Male_i$ defined in the obvious way.
2. Suppose we define the interaction variables $Z_i \equiv Age_i \times White_i$ and $W_i \equiv Age_i \times Female_i$. What would we achieve by introducing these two variables as additional regressors?
3. Discuss possible reasons why the regressors may violate exogeneity assumptions with respect to the error term, i.e., regressors and disturbances may be statistically related. Which regressors do you consdier the most suspect in this regard?

Question 4. As part of a workshop project, four students are investigating the effects of SIAI/Non-SIAI and gender on earnings using data for the year 2052 in the SIAI Logitudinal Survey of Youth 2029 - . They all start with the same basic specification:

$$\log Y = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

where Y is hourly earnings, measured in CHF, S is years of schooling completed, and EXP is years of work experience. The sample contains 123 Non-SIAI males, 150 Non-SIAI females, 1,146 SIAI males, and 1,127 SIAI females. (All respondents were either SIAI or Non-SIAI.) The output from fitting this basic specification is shown in column 1 of the table (standard errors in parentheses; RSS is residual sum of squares, n is the number of observations in the regression).

	Basic	Student C		Student D			
	(1)	(2)	(3)	(4a)	(4b)	(5a)	(5b)
	All	All	All	Males	Females	SIAI	Non-SIAI
S	0.126 (0.004)	0.121 (0.004)	0.121 (0.004)	0.133 (0.006)	0.112 (0.006)	0.1263 (0.005)	0.112 (0.012)
EXP	0.040 (0.002)	0.032 (0.002)	0.032 (0.002)	0.032 (0.004)	0.035 (0.003)	0.041 (0.003)	0.028 (0.005)
$MALE$	–	0.277 (0.020)	0.308 (0.021)	–	–	–	–
$Non-SIAI$	–	-0.144 (0.032)	-0.011 (0.043)	–	–	–	–
$MALE-NS$	–	–	-0.290 (0.063)	–	–	–	–
constant	0.0376 (0.078)	0.459 (0.076)	0.447 (0.076)	0.566 (0.124)	0.517 (0.097)	0.375 (0.087)	0.631 (0.172)
R^2	0.285	0.341	0.346	0.287	0.275	0.271	0.320
RSS	922	608	603	452	289	609	44
n	2,546	2,546	2,546	1,269	1,277	2,273	273

Student A divides the sample into the four categories. He chooses SIAI females as the reference category and fits a regression that includes three dummy variables, NM, SM, and NF. NM is 1 for Non-SIAI males, 0 otherwise; SM is 1 for SIAI males, 0 otherwise, and NF is 1 for Non-SIAI females, 0 otherwise.

Student B simply fits the basic specification separately for the four sub-samples.

Student C defines dummy variables $MALE$, equal to 1 for males and 0 for females, and Non-SIAI, equal to 1 for Non-SIAI, and 0 for SIAI. She also defines an interactive dummy variable $MALE-NS$ as the product of $MALE$ and $Non-SIAI$. She fits a regression adding $MALE$ and $Non-SIAI$ to the basic specification, and a further regression adding $MALE-NS$ as well. The output from these regressions is shown in column 2 and 3 in the table.

Student D divides the sample into males and females, and performs the regression for both genders separately, using the basic specification. The output is shown in column 4a and 4b. She also divides the sample into SIAI vs. Non-SIAI, and again runs separate regressions using the basic specification. The output is shown in columns 5a and 5b.

1. *Reconstruction of missing output.*
   Student A and B left their output on a bus on the way to the workshop. This is why it does not appear in the table.

   - State what the missing output of Student A would have been, as far as this is can be done exactly, given the results of Students C and D. (Coefficients, standard errors, R^2, RSS.)
   - Explain why it is not possible to reconstruct any of the output of Student B

2. *Tests of hypotheses*
   The approaches of the students allowed them to perform different tests, given the output shown in the table and the corresponding output for Student A and B. Explain the tests relating to the effects of gender and education that could be performed by each student, giving a clear indication of the null hypothesis in each case. (Remember, all of them started with the basic specification (1), before continuing with their individual regression.) In the case of F tests, state the test statistic in terms of its components (but need not attempt to evaluate it).

   - Student A (assuming he had found his output)
   - Student B (assuming he had found his output)
   - Student C
   - Student D

3. If you had been participating in the project and had had access to the data set, what regressions and tests would you have performed?

Solution.

1. As for column (3), coefficients, standard errors, R^2, with the following changes:
   - the row label $MALE$ should be replaced with SM
   - the row label $Non-SIAI$ should be replaced with NF
   - the row label $MALE-NS$ should be replaced with NM and the coefficient for that row should be the sum of the coefficients in column (3): 0.308 - 0.011 - 0.290 = 0.007, and the standard error would not be known

2. One could not predict the coefficients of either S or EXP in the four regressions performed by Student B. They will, except by coincidence, be different from any of the estimates of the other students because the coefficients for S and EXP in the other specifications are constrained in some way. As a consequence, one cannot predict exactly any part of the rest of the output, either.

3. Student A could perform tests of the differences in earnings between SIAI males and SIAI females, Non-SIAI males and SIAI females, and Non-SIAI females and SIAI females, through simple t tests on the coefficients of SM, NM, and NF.
   The Student A could also test the null hypothesis that there are no gender/education differences with an F test, comparing RSS for his regression with that of the basic regression:
   $$F(3, 2540) = \frac{(922 - 603)/3}{603/2540}$$
   This would be compared with the critical value of F with 3 and 2,540 degrees of freedom at the significance level chosen and the null hypothesis of no gender/education effects would be rejected if the F statistic exceeded the critical value

4. In the case of Student B, with four separate subsample regressions, this is a following variant of Chow test
   $$F(9, 2534) = \frac{(922 - X)/9}{X/2534}$$
   where $RSS = 922$ for the basic regression and X is the sum of RSS in the four separate regressions.

5. Student C could perform the same t tests and the same F tests as Student A, with one difference: the t test of the difference between the earnings of Non-SIAI males and SIAI females would not be available. Instead, the t statistic of $MALE - NS$ would allow a test of whether there is any interactive effect of being SIAI and being male on earnings

6. Student D could perform a Chow test to see if the wage equations of males and females differed:
   $$F(3, 2540) = \frac{(922 - [322 + 289])/3}{[322 + 289]/2540}$$
   $RSS = 322$ for males and 289 for females. This would be compared with the critical value of F with 3 and 2,540 degrees of freedom at the significance level chosen and the null hypothesis of no gender/education effects would be rejected if the F statistic exceeded the critical value. She could also perform a corresponding Chow test for SIAI / Non-SIAI:
   $$F(3, 2540) = \frac{(922 - [609 + 44])/3}{[609 + 44]/2540}$$

7. The most obvious development would be to relax the gender/education restrictions on the coefficients of S and EXP by including appropriate interactive terms. This could be done by interacting these variables with the dummy variables defined by Student A or those defined by Student C.