ECON3150/4150: Introductory Econometrics – Exam Spring 2024

Be brief and to the point. Always motivate your answers. All sub-questions have equal weight.

The following study

Altonji, J.G., Elder, T.E, Taber, C.R. (2005). An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling. Journal of Human Resources 40(4), 791-821.

investigated the effect of Catholic schooling on school outcomes. Below you find definitions of some variables in their dataset as well as descriptive statistics:

```
# id:       person identifier
# read12:   reading standardized score
# math12:   mathematics standardized score
# female:   =1 if female
# asian:    =1 if Asian
# hispan:   =1 if Hispanic
# black:    =1 if black
# motheduc: mother's years of education (8,11,11.5,12,14,16,17,18)
# fatheduc: father's years of education (8,11,11.5,12,14,16,17,18)
# lfaminc:  log of annual family income (USD)
# hsgrad:   =1 if graduated from high school by 1994
# cathhs:   =1 if attended Catholic high school
# parcath:  =1 if a parent reports being Catholic
```

```
##                  mean           SD        min        max    N
## id       4648285.08191 2721625.8314 124902.000 7979086.00 5970
## read12        51.54848       9.3953     29.150      68.09 5970
## math12        51.85018       9.4783     30.140      71.37 5970
## female         0.52764       0.4993      0.000       1.00 5970
## asian          0.06281       0.2426      0.000       1.00 5970
## hispan         0.11139       0.3146      0.000       1.00 5970
## black          0.07755       0.2675      0.000       1.00 5970
## motheduc      13.24154       2.0022      8.000      18.00 5970
## fatheduc      13.54397       2.2630      8.000      18.00 5970
## lfaminc       10.31288       0.8019      6.215      12.35 5970
## hsgrad         0.93032       0.2546      0.000       1.00 5970
## cathhs         0.07203       0.2586      0.000       1.00 5970
## parcath        0.36265       0.4808      0.000       1.00 5970
```

```
# Results to Question 1.
```

```
##                          reg1           reg2           reg3
## Dependent Var.:        hsgrad         hsgrad         hsgrad
##
## cathhs          0.060 (0.007)  0.039 (0.006)  0.029 (0.007)
## fatheduc = 11                 -0.051 (0.027) -0.052 (0.027)
## fatheduc = 11.5                0.023 (0.036)  0.005 (0.036)
## fatheduc = 12                  0.058 (0.024)  0.042 (0.023)
## fatheduc = 14                  0.070 (0.023)  0.043 (0.023)
## fatheduc = 16                  0.110 (0.023)  0.069 (0.023)
## fatheduc = 17                  0.103 (0.024)  0.055 (0.024)
## fatheduc = 18                  0.111 (0.024)  0.049 (0.024)
## black                         -0.025 (0.014) -0.003 (0.014)
## asian                          0.027 (0.009)  0.031 (0.009)
## hispan                        -0.018 (0.013) -0.008 (0.013)
## lfaminc                                      0.045 (0.006)
## Constant        0.926 (0.004)  0.870 (0.023)  0.429 (0.064)
## _______________ _____________  _____________  _____________
## S.E. type       Heteros.-rob. Heterosk.-rob. Heterosk.-rob.
## SSR                    385.57         371.59         365.68
## Observations            5,970          5,970          5,970
```

```
# Results to Question 1j.
cor(predict(reg3), catholic$hsgrad)^2
cor(residuals(reg3), catholic$cathhs)
```

1. (60%) We are interested in estimating the effect of attending a Catholic high school on high school graduation. To answer this question we estimated the models in the R-output above:

   a. Interpret the coefficient on `cathhs` in the 1st results column (`reg1`).

      ANSWER HINT: people who attended a Catholic high school are 6 percentage points more likely to graduate from high school than people who didn't.

   b. Interpret the coefficient on `fatheduc = 11` in the 2nd column (`reg2`). Is this estimate significant at the 1% level?

      ANSWER HINT: people with a father who has 11 years of schooling are 5.1 percentage points less likely to graduate from high school **than people whose father has 8 years of schooling** while **adjusting for catholic school attendance and ethnicity**.

   c. Interpret the coefficient on `cathhs` in the 2nd column. Explain why the size dropped compared to the first column and use the omitted-variable-bias formula to motivate your answer.

      ANSWER HINT: people who attended a Catholic high school are 3.9 percentage points more likely to graduate from high school than people who didn't, **adjusting for their ethnicity and fathers education**. The intuition of the OVB formula tells us that the difference between the coef in 1a. and 1.c involves the association between ethnicity + fathers education and high-school completion, times a term whose sign depends on the correlation between these variables and cathss. Since there is a positive bias this implies that on average people with high(er) educated fathers and/or from ethnicities that are more likely to complete are also more likely to have attended catholic school.

   d. What is the interpretation of the intercept in the 2nd column (`reg2`)? And in the 3rd column (`reg3`)?

      ANSWER HINT: column 2: E[hsgrad | cathss=0, fatheduc=8, white=1]. column 3: E[hsgrad | cathss=0, fatheduc=8, white=1, lfaminc=0]. Note that lfaminc=0 implies a family income of 1 dollar.

   e. The third column (`reg3`) adds the logarithm of family income to the specification. What is the predicted value of high-school graduation for an asian student who attended a catholic high-school, whose father has 16 years of schooling, and with a family income of 100,000 USD? What do you conclude?

      ANSWER HINT: $0.029 + 0.069 + 0.031 + 0.045 * \log(100000) + 0.429 = 1.076082$. Since this is a predicted probability (which cannot logically exceed 1) the model is misspecified and a probit or logit would have been more appropriate (at least for prediction purposes).

   f. Compute the R-squared of the regression in the 3rd column (`reg3`).

ANSWER HINT: R2 = 1 - RSS/TSS = 1 - 365.68 / ((5970-1)*(0.2546)^2) = 0.05489.

g. Interpret the coefficient on `lfaminc` in the 3rd column (`reg3`). How would the coefficients in that regression change if family income was not measured in US dollars but in 1000s of US dollars?

ANSWER HINT: A 10% increase in family income is associated with a 0.1* 0.045 = 0.0045 percentage point increase in the high school graduation rate, keeping fathers income and ethnic background fixed. Changing the unit of measurement only affects the intercept (and thus not the coefficient on log family income) since log(a*y) = log(a) + log(y).

h. Perform a joint two-sided F-test (assuming homoskedasticity) of the null-hypothesis that the coefficient on the race variables and father's education in the 2nd column (`reg2`) are zero. Do you reject the null hypothesis at the 1% level?

ANSWER HINT: F = ((RSS_restr - RSS_unrestr) / #restr) / (RSS_unrestr / #dof). F = ((385.57 - 371.59)/10)/(371.59/(5970 - 12)) = 22.42. The critical value at the 1% level is 2.3209 and we thus reject the null-hypthesis since the F statistic is larger.

i. Explain under what assumption the coefficient on `cathhs` in the 3rd column (`reg3`) estimates the causal effect of Catholic high-school attendance on high-school graduation. Do you believe this assumption?

ANSWER HINT: We need to assume that after ajusting for fathers education, ethnic background and (log) family income there are no other determinants of high-school graduation left that correlate with catholic school attendance. In other words, conditional on those variables cathhs needs to be as good as randomly assigned. This is not clear at all. For example, more able kids may be send to catholic schools for all we know.

j. What two numbers (rounded to 3 decimals) does the R-code above in the output (marked Question 1j.) produce?

ANSWER HINT: the first line computes the square of the correlation between y and yhat in regression 3 which correspond to the R-squared which we saw above equalled 0.05. The second line is the correlation between the residual and one of the regressors in reg3 which is mechanically zero because that (zero correlation between residuals and regressors) is what ols **uses** to estimate the coefs.

k. Using the dataset above and starting from the specification in the first column (`reg1`) you want to jointly estimate how i) high-school completion differs for kids with a mother who has at least a high-school education (ie at least 12 years of schooling), and ii) how this gap is different in Catholic high-schools. What is the regression specification that you would estimate?

ANSWER HINT: Let x equal 1 if the mother has at least high-school and is zero

otherwise. We then want to estimate

$$hsgrad = \beta_0 + \beta_1 cathhs + \beta_2 x + \beta_3 x \cdot cathss + e$$

. (Note that testing for zero in i) corresponds to $H_0 : \beta_2 = \beta_3 = 0$ and ii) $H_0 : \beta_3 = 0$).

l. Write the R-code that implements the estimation proposed in 1k.

ANSWER HINT: `feols(hsgrad~cathhs*I(motheduc ⩾ 12), catholic)`

```r
# Results to Question 2.
t(aggregate( . ~ parcath, catholic, mean))
```

```
##                  [,1]          [,2]
## parcath       0.00000       1.00000
## id       4707410.34218 4544372.09561
## read12       51.45293      51.71641
## math12       51.55579      52.36758
## female        0.53850       0.50855
## asian         0.06807       0.05358
## hispan        0.03679       0.24249
## black         0.11222       0.01663
## motheduc     13.29566      13.14642
## fatheduc     13.53679      13.55658
## lfaminc      10.28121      10.36855
## hsgrad        0.92116       0.94642
## cathhs        0.01340       0.17506
```

2. (40%) Some studies have used religious affiliation as exogenous source of variation to estimate the causal effect of Catholic high-school attendance on outcomes. Taking inspiration from this approach we plan to use whether a parent reports being Catholic as a so-called instrumental variable. The output below reports sample averages separately for children without a catholic parent (`parcath=0`) and with a catholic parent (`parcath=1`).

   a. The output above shows that kids with a catholic parent are 0.94642 - 0.92116 = 0.02526 or about 2.5 percentage points more likely to graduate from high school than those without a catholic parent. Explain what we need to assume to for this to be a causal effect.

      ANSWER HINT: We need to assume that catholicism is as good as randomly assigned and that there are no other unobserved and omitted factors (genes, environment, resources, etc) that cause differences in high-school graduation and that are not caused by whether the parent is catholic or not.

   b. Given the output above, do you think the assumption in 2a. is valid?

      ANSWER HINT: We observe differences in terms of education, ethnicity and income between catholic and non-catholic parents. This suggests that catholicism is not as good as randomly assigned and that there there might be OVB in a naive estimate in 1a. This is in particular true for ethnicity which is immutable, While catholicism could affect education and/or income.

   c. Maintain the assumption in 2a. What is the (estimated) causal effect of having a Catholic parent on attending a catholic high-school?

      ANSWER HINT: 0.17506 - 0.01340 = 0.1617

   d. Use the Wald estimator to estimate the effect of attending a Catholic high-school on high-school graduation.

      ANSWER HINT: IV = RF/FS = 0.02526 / 0.1617 = 0.1562

   e. What do you need to assume for the estimator in 2d. to be consistent? Discuss in the context of this application.

      ANSWER HINT: We need to assume exogeneity and exclusion. The first assumption is already discussed in 2b. Exclusion means that the catholicism of the parent only affects high-school graduation through the choice of a catholic high school and not in other ways. This may be a strong assumption if catholic parents are more generally concerned with education and adjust their behavior accordingly. Finally, we also need to check whether we have a significant first stage.

   f. Explain how you can improve on your estimate in 2d. using the data above.

      ANSWER HINT: We at least want to adjust for ethnicity (and possibly for education and/or income to adjust for potential OVB) in the estimation. This would mean that we need to estimate the causal effect using 2SLS controlling for these variable(s).

Critical Values for the $F_{m,\infty}$ Distribution

```
## Rows denote degrees of freedom (m), and columns significance level (%)

##          10%     5%     1%

##
##  1 :  2.7055 3.8415 6.6349
##  2 :  2.3026 2.9957 4.6052
##  3 :  2.0838 2.6049 3.7816
##  4 :  1.9449 2.3719 3.3192
##  5 :  1.8473 2.2141 3.0173
##  6 :  1.7741 2.0986 2.8020
##  7 :  1.7167 2.0096 2.6393
##  8 :  1.6702 1.9384 2.5113
##  9 :  1.6315 1.8799 2.4073
##  10 :  1.5987 1.8307 2.3209
##  11 :  1.5705 1.7886 2.2477
##  12 :  1.5458 1.7522 2.1847
##  13 :  1.5240 1.7202 2.1299
##  14 :  1.5046 1.6918 2.0815
##  15 :  1.4871 1.6664 2.0385
##  16 :  1.4714 1.6435 2.0000
##  17 :  1.4570 1.6228 1.9652
##  18 :  1.4439 1.6038 1.9336
##  19 :  1.4318 1.5865 1.9048
##  20 :  1.4206 1.5705 1.8783
##  21 :  1.4102 1.5557 1.8539
##  22 :  1.4006 1.5420 1.8313
##  23 :  1.3916 1.5292 1.8104
##  24 :  1.3832 1.5173 1.7908
##  25 :  1.3753 1.5061 1.7726
##  26 :  1.3678 1.4956 1.7554
##  27 :  1.3608 1.4857 1.7394
##  28 :  1.3541 1.4763 1.7242
##  29 :  1.3478 1.4675 1.7099
##  30 :  1.3419 1.4591 1.6964
```