Exam ECON3150/4150: Introductory Econometrics – Spring 2022

1. (80%) Suppose you have the following data from the American "Current Population Survey" from 1992, with average hourly earnings (`ahe`), a dummy variable (`bachelor`) that equals one if a person holds at least a bachelor degree and is zero for those with only a high-school degree, and finally age (`age`):

```
##                mean     SD      min     max    N
## year      1992.0000 0.0000 1992.000 1992.00 7612
## ahe         11.6168 5.6195    1.243   46.63 7612
## bachelor     0.3891 0.4876    0.000    1.00 7612
## age         29.7105 2.8063   25.000   34.00 7612
```

You estimate the following OLS regression:

```
reg = feols(ahe ~ bachelor + age + I(age^2), df, vcov="hetero")
reg
```

```
## OLS estimation, Dep. Var.: ahe
## Observations: 7,612
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -17.33751   6.990530  -2.480 0.0131545 *
## bachelor      4.34001   0.128357  33.812 < 2.2e-16 ***
## age           1.50138   0.479201   3.133 0.0017363 **
## I(age^2)     -0.01947   0.008155  -2.388 0.0169773 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 5.13253   Adj. R2: 0.16536
```

```
vcov(reg)
```

```
##             (Intercept)    bachelor       age    I(age^2)
## (Intercept)    48.86751 -0.04748831 -3.346487  0.05678874
## bachelor       -0.04749  0.01647545  0.002553 -0.00003688
## age            -3.34649  0.00255322  0.229634 -0.00390420
## I(age^2)        0.05679 -0.00003688 -0.003904  0.00006650
```

a. Interpret the estimated coefficient on `bachelor`.

   ANSWER HINT: at a given age bachelors earn 4.34 USD more per hour than those with only a high-school degree

b. Construct and interpret the 68 percent confidence interval for the estimate in 1.a.

   ANSWER HINT: using the critical value this is ca $4.34 \pm 1 \times SE \approx (4.21, 4.47)$

c. Can we give the estimate in 1.a a causal interpretation? Motivate your answer.

   ANSWER HINT: this estimate probably not causal because of omitted variable bias. people with a bachelor education are for example probably more able, motivated, connected, from a dense labor market etc. than people with only a high-school diploma.

d. What is the interpretation of the Intercept?

   ANSWER: the hourly wage of a 0-year-old (without a bachelor)

e. How much does a 25-year-old with a bachelor degree earn on average per hour?

ANSWER: -17.34 + 4.34 + 1.50 * 25 + -0.0195 * 25^2 = 12.31

f. Compute the average marginal effect of `age`.

ANSWER: $\beta_{age} + 2 * \beta_{age^2}\bar{age}$ which equals 1.5 - 0.0195 * 2 * 29.7105 = 0.3413

g. Compute the standard error of the estimate in 1.f.

ANSWER HINT: Use key concept 2.3 Stock and Watson:

$$var(\beta_{age} + 2 * \beta_{age^2}\bar{age}) = var((\beta_{age}) + (2\bar{age})^2 var(\beta_{age^2}) + 2(2\bar{age})cov(\beta_{age}, \beta_{age^2})$$
$$= 0.229634 + (2 * 29.7105)^2 * 0.00006650 + 2 * (2 * 29.7105) * -0.00390420 = 0.0004529$$

and the standard error is the square root: $\sqrt{0.0004529} \approx 0.02128$

h. Suppose you want to test the joint significance of the age profile at the 5% level. Explain how you would go about testing this and what exact critical value you would use.

ANSWER HINT: Perform an F-test of the null-hypothesis $H_0 : \beta_{age} = \beta_{age^2} = 0$ agains the alternative that at least one of these coefficients is non-zero. Under the null there are 2 restrictions, and there are many observations. Which means that we need the critical value from the $F_{2,\infty}$ distribution which is 3.00.

A friend suggests to estimate the following regression instead:

```
feols(log(ahe) ~ bachelor + age, df, vcov="hetero")
```

```
## OLS estimation, Dep. Var.: log(ahe)
## Observations: 7,612
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1.38303   0.055445   24.94 < 2.2e-16 ***
## bachelor     0.37464   0.010504   35.67 < 2.2e-16 ***
## age          0.02722   0.001856   14.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.453515   Adj. R2: 0.154567
```

i. Your friend claims that this regression is better. What do you reply?

ANSWER HINT: that they are both probably wrong in the sense of being non-causal. in terms of fit we cannot say anything about this question with the information provided because we cannot compare R-squares when the dependent variables are different.

j. Interpret the estimated coefficient on `age`. Is this similar to your results above?

ANSWER HINT: increasing age by 1 year is associated with a 2.7% higher hourly wage. Above we found that the average partial effect of age equalled 0.3413 which is 0.3413/11.6168 = 0.029 or about 2.9% of the average hourly wage and therefore very similar.

k. In a next step your friend wants to investigate the hypothesis that people with a bachelor degree have *steeper* age profiles than those with only a high-school degree. Explain how to do this.

ANSWER HINT: Estimate a regression with an interaction:

$$\log(ahe) = \beta_0 + \beta_1 age + \beta_2 bachelor + \beta_3 age \cdot bachelor + u$$

and test of $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 < 0$.

2. (20%) In the Netherlands applicants to medical school used to be admitted solely based on a random lottery. Below you find data on such lotteries with information on admissions (`admitted`), medical school degrees (`medschool`), and later income when applicants were about 35 years old (`income`):

```
xtabs( ~ medschool + admitted, df)
```

```
##           admitted
## medschool    0    1
##         0 1162  154
##         1  965 2409
```

```
aggregate(income ~ medschool + admitted, df, mean)
```

```
##   medschool admitted income
## 1         0        0  21.59
## 2         1        0  20.80
## 3         0        1  24.05
## 4         1        1  23.26
```

Use these data to

a. estimate the causal effect of medical school on income,

   ANSWER HINT: we need to compute the first-stage & reduced-form estimate and use these to compute the IV (Wald) estimate.

   First-stage = 2409 /(2409+154) - 965 / (965 + 1162) = 0.4862

   Reduced form = (23.26 * 2409 + 24.05 * 154) / (2409+154) - (20.8 * 965 + 21.59 * 1162)/ (965 + 1162) = 2.076

   IV = 2.076 / 0.4862 = 4.27

b. state your assumptions and

   ANSWER HINT: To give the estimate a causal interpretation admission need to satisfy the exclusion restriction in the sense that it is exogenous (does not correlate with potential outcomes), and does not have an independent effect on income (apart from its effect on medical school completion). admission must also have an impact of medical school completion (relevance)

c. discuss their validity in the current setting, providing support from the data where possible, and

   ANSWER HINT: the randomization of admission through the lottery mechanically satisfies the exogeneity of the instrument. It could violate exclusion if losing the lottery leads to disappointment which in turn affects earnings.

   To verify relevance we need a statistically significant first-stage.

   With a binary variable where $p = \Pr(Y = 1)$ the variance equals $p(1 - p)$.

   We also know that the variance of an estimator equals the estimated variance divided by N

   Let $p_1 = 2409/(2409 + 154) =$ and $N_1 = 2563$, $p_0 = 965/(965 + 1162)$ and $N_0 = 2127$ then the first stage equals

   $$p_1 - p_0$$

   the variance of this estimate equals

   $$p_1(1 - p_1)/N_1 + p_0(1 - p_0)/N_0$$

   where there is no covariance because these groups are independent (lottery is random).

   This gives a SE of the first stage that is about 0.01177.

d. interpret your findings.

We see that admissing increases the likelihood of completing medical school by about 49 percentage points, and causally increases income by 2.076.

Completing medical school has a causal effect of 4.27 on income which is substantial compared to for example the average income of application who where not admitted (about 21.2)