ECON3150/4150: Introductory Econometrics – Postponed Exam Spring 2022

1. (80%) Suppose you have data from from the 1977–1978 Australian Health Survey. Descriptive statistics of your data stored in data frame `df` are as follows:

```
##           mean     SD  min   max    N
## visits  0.3007 0.7932 0.00  9.00 5111
## gender  0.5189 0.4997 0.00  1.00 5111
## age     0.4076 0.2050 0.19  0.72 5111
## income  0.5922 0.3645 0.01  1.50 5111
## private 0.4424 0.4967 0.00  1.00 5111
## health  1.2105 2.1181 0.00 12.00 5111
```

   where

   1. **visits** Number of doctor visits in the past 2 weeks.
   2. **gender** indicates gender (1=female, 0=male)
   3. **age** Age in years divided by 100.
   4. **income** Monthly income in tens of thousands of dollars.
   5. **private** Does the individual have private health insurance? (1=yes, 0=no)
   6. **health** General health questionnaire score (a higher score implies worse health)

   You perform the following analysis:

```
df$age = df$age - 0.4076
reg = feols(visits ~ private + age + log(income), df, vcov="hetero")
reg
```

```
## OLS estimation, Dep. Var.: visits
## Observations: 5,111
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  0.25085    0.02019  12.427  < 2.2e-16 ***
## private      0.02384    0.02265   1.053 2.9259e-01
## age          0.44587    0.05700   7.822 6.2774e-15 ***
## log(income) -0.05262    0.01849  -2.847 4.4342e-03 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## RMSE: 0.785756   Adj. R2: 0.017835
```

```
vcov(reg)
```

```
##              (Intercept)     private       age log(income)
## (Intercept)   0.0004075 -0.00028544 0.0002609  0.00025841
## private      -0.0002854  0.00051285 0.0001088 -0.00009224
## age           0.0002609  0.00010880 0.0032493  0.00015137
## log(income)   0.0002584 -0.00009224 0.0001514  0.00034172
```

a. Interpret the estimated coefficient on `log(income)`.

   ANSWER: a 1% increase in income is associated with a 0.01*0.05 decrease in the nr of doctors visits, keeping age and private insurance fixed.

b. Construct (briefly explain your steps) and interpret the 86 percent confidence interval for the estimate in 1.a.

   ANSWER: qnorm(.07) = -1.475791, so the 86% CI is approx (b - 1.48*se, b + 1.48*se) or (-0.080, 0.025) which covers the true value of b in 86% of all possible random samples.

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

   ANSWER: probably not because of omitted variable bias: we do not adjust for things that correlate with (log) income keep age and private fixed and that themselves determine the outcome. an example is health which affects the number of doctor visits and correlates with income (people with more income are on average more healthy)

d. What is the interpretation of the Intercept?

   ANSWER: the outcome setting all variables equal to zero. note that the variable age is recentered at 0.41 and therefore zero at approx the sample average, log income is zero when income is 1 which corresponds to a monthly income of 10,000 dollars. so the intercept is the average nr of doctor visits for someone who has an average age, has a monthly income of 10K and does not have private insurance.

e. Predict the outcome for a 60 year old female with private health insurance who earns 5,000.

   ANSWER: 0.25085 + 0.02384 * 1 +0.44587 * (60 - 0.4076) -0.05262 * log(0.5) = 26.9. Note that this is the prediction for both males and females since the regression does not adjust for gender.

f. Compute the marginal effect of income:

$$\partial E[visits|gender, age, income]/\partial income$$

   when income=2,000.

   ANSWER: partial effect is betahat / income = -0.05262 / 0.2 = -0.2631

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

   var(betahat$c$) = c^2 var(betahat). so se is c * se(betahat). which is 0.01849 / 0.2 = 0.09245

A friend is concerned that the estimate of having private health insurance on the number of doctor visits suffers from omitted variable bias. She suggests to use instrumental variable estimation instead with gender as the instrumental variable, arguing that women are more risk averse than men and therefore more likely to take out private health insurance.

Your friend provides the following OLS regressions:

```
feols(health ~ gender + age + log(income), df, vcov="hetero")
```

```
## OLS estimation, Dep. Var.: health
## Observations: 5,111
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   0.9396    0.04840  19.414     < 2.2e-16 ***
## gender        0.1735    0.06253   2.775 0.00554663216 **
## age          -0.1683    0.15471  -1.088 0.27671289807
## log(income)  -0.2421    0.04660  -5.194 0.00000021368 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 2.10766   Adj. R2: 0.009073
```

```
feols(visits ~ gender + age + log(income), df, vcov="hetero")
```

```
## OLS estimation, Dep. Var.: visits
## Observations: 5,111
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error t value          Pr(>|t|)
## (Intercept)  0.23604    0.01737  13.588         < 2.2e-16 ***
## gender       0.06554    0.02453   2.672 0.007555494394212 **
## age          0.40841    0.05996   6.811 0.000000000010775 ***
## log(income) -0.04105    0.01870  -2.195 0.028203192221163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.78523   Adj. R2: 0.019149
```

```
feols(private ~ gender + age + log(income), df, vcov="hetero")
```

```
## OLS estimation, Dep. Var.: private
## Observations: 5,111
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.51631    0.01171  44.102 < 2.2e-16 ***
```

```
## gender       0.12592    0.01400   8.994 < 2.2e-16 ***
## age         -0.04873    0.03459  -1.409   0.15898
## log(income)  0.18636    0.01023  18.222 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.476583   Adj. R2: 0.078707
```

i. Compute and interpret the IV estimate of the effect of having private health insurance on the number of doctor visits in the past 2 weeks.

ANSWER: reduced form divided by first-stage: 0.06554/0.12592 = 0.5205. private health insurance causally increases the number of doctor visits last week by 0.5.

j. What are the assumptions that are necessary for the IV estimator in 1.i to be internally valid and explain **why** they need to hold.

ANSWER: instrument relevance: gender affects private insurance. if the instrument does not affect the treatment (private insurance) then there is no variation in treatment status, and without variation in treatment we cannot compute outcomes with and without treatment. exogeneity/exclusion: gender affects the outcome only through insurance. if this doesn't hold then we cannot be sure that the instrument induced difference in outcomes (the reduced form effect) is only explained by variation in treatment status.

k. Do you believe that the assumptions in 1.j hold in this specific application? Explain your answer.

ANSWER: relevance: true, t-stat of 9 (F = 9*9=81). exog/excl: not true. from the first regression we see that women have better health. they might also are more likely to visit a doctor keeping health fixed.

2. (20%) Briefly outline and motivate the analysis you would perform in the following cases. Make sure to explain the assumption(s) that need to hold for internal validity in the context of the application.

   a. From 2007 workers in the municipality of Mandal could self-certify their sickness, meaning that they did not need a doctor's certificate to be absent from work. You want to estimate the causal effect of the self-certification policy on sickness absence. You have a dataset for the years 2000 to 2015. In the data you have information on workers' annual sickness absence, their municipality of residence, education, age and gender.

ANSWER: difference in differences comparing sickness absence in Mandal before and after 2007 to sickness absence in other municipalities before and after 2007. the key assumption here is that Mandal would have experienced the same change in sickness absence as the other municipalities if they would not have introduced self-certification. this is the common trend assumption. we can relax this assumption by controlling for education, age and gender (if we believe this changes differentially over time across municipalities). the most basic regression would be

$$absence_{it} = \beta_1 * post2007 + \beta_2 * mandal_{it} + \beta_3 * mandal_{it} * post2007 + e_{it}$$

where $mandal_{it}$ is a binary variable that indicates whether person i lives in Mandal in year t. and $post2007$ is a dummy that equals one for the treatment years (post 2007) and is zero otherwise.

   b. A number of secondary schools in the Netherlands run a program for gifted students. Students are graded on a 0 to 10 point scale, and those who score at least an 8 participate in the program. You are interested in estimating the causal impact of participating in the program on student's final exam score in mathematics. You have a dataset with students' average grade used to determine their eligibility for the program, their gender, parents' education, and final exam score in mathematics.

ANSWER: here treatment is assignment based on a variable crossing a threshold. this is known as the regression discontinuity design. the basic idea is that very close to the threshold people are comparable, but above the threshold people are treated while below not. this comparison thus gives a local estimate of the causal effect of the treatment. the most basic regression implementing this idea is

$$exam_i = \delta * treatment_i + \beta * grade_i + e_i$$

the comparability is key and we therefore need to be able to rule out that students or schools can sort themselves (by manipulating grades) around the threshold. we can check this by comparing peoples gender or parental education around the threshold.