Econometrics 2, Assignment 2

Henry Haustein

Section A

(a) The research question is: *What are the long-term labour market consequences (in terms of earnings) for Vietnam war veterans?* This question is relevant because the previous research was maybe biased by selection bias that's why the author wants to use an IV approach. Previous research also had different outcomes when it comes to the question whether army veterans get an adequate compensation for the risks they take. The research could lead to new policies if negative consequences are found.

(b) The problem is the selection bias. Who volunteers for army? Maybe men with relatively few civilian opportunities but often certain types men (think of venturesome, assertive, could fast adapt to new situations etc.). In some branches men with this abilities could climb up the career leader faster which results in more income.

(c) A simple indicator would measure the effect of being eligible on income. But that's not the thing the author is interested in because being eligible doesn't mean that you're a veteran since there is avoidance behaviour (going to college, moving out of the country, etc.). During the Vietnam war draft lottery it is estimated that college enrolment rates are 6 to 7 percent higher [2]. Because the veteran status is not endogenous the author uses a IV approach. The outcome equation is

$$\text{Earnings}_i = \alpha_1 + \lambda \cdot \mathbf{1}(\text{veteran}_i) + \varepsilon_i$$

First stage equation:

$$\text{veteran}_i = \alpha_2 + \phi \cdot \mathbf{1}(\text{drawn by lottery}_i) + \xi_i$$

Reduced form equation

$$\text{Earnings} = \alpha_3 + \rho \cdot \mathbf{1}(\text{drawn by lottery}_i) + \psi_i$$

We could estimate λ by $\hat{\lambda} = \frac{\hat{\rho}}{\hat{\phi}}$. That is done in equation (2) in the paper.

(d) The IV assumptions are:

- Instrument has an effect on treatment ($\phi \neq 0$): In Table 2 we can see that the $\hat{p}^e - \hat{p}^n$ column is significant except 1953 for whites in the SIPP dataset. For non-white people the value of $\hat{p}^e - \hat{p}^n$ is never significant 1953 in the DMDC/CWHS dataset. Because of that the author is conducting his research only on the 1950-1952 whites people.
- Random assignment: The instrument has to be random but there are no balance checks in the paper to support this assumption. On the other hand the assignment from dates to numbers was done by lottery and birthday dates are mostly random the draft lottery should be overall random.

- Exclusion restriction: The lottery could only affect earnings through the veteran status because the lottery results are only used for that. The are no other benefits from being eligible. But on the other hand the lottery might have an effect on future career plans because it changes education plans. But there is more research needed.
- No defiers: This point is not discussed in the paper although this behaviour would be highly irrational: If you want to go to army then there is always the chance and if you don't want you can mostly avoid it. There is no point in volunteering for army if you are not eligible but avoiding army if you are eligible.

(e) As mentioned in (c) the estimator for the treatment effect is $\hat{\lambda} = \frac{\hat{\rho}}{\hat{\phi}}$. And since *drawn by lottery* and *veteran* are binomial the estimator becomes

$$\hat{\lambda} = \frac{\bar{Y}^e - \bar{Y}^n}{\hat{p}^e - \hat{p}^n}$$

which is called a Wald estimator, after Abraham Wald. The paper doesn't talk about compliers, it mentions always- and never-takers.

(f) There are two main critiques I have on this paper:

- Method: I would have liked it if the author has done some balance checks to see whether the random-assignment-assumption is true. Especially since there is doubt on the fairness of the lottery (see [3], p. 101).
- Results: There are no consequences drawn from this paper. Since we have a 15% gap would it be enough just to raise the pay of soldiers by 15%? Or, when we can't overcome the gap, just participate in less conflicts and build the army up on volunteers? It would be interesting to see how the 15% gap evolved over time. Interestingly Angrist and Chen have checked for gap in year 2000 data and found that the gap is gone (see [1]). Maybe in the original paper the gap could not be estimated after 20 years but after 5, 10, 15 and 20 years or even for every year if there is such data. This would make it easier to make policies to overcome the gap.

Section B

Load the data

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library(haven)
library(tidyverse)
library(psych)
library(mfx)
library(margins)

data = read_dta("Smoking.dta")
```

(a) Since the *smoker* variable is binary we can use `mean`:

```
mean(data$smoker)
```

results in 0.2423.

(b) After using a filter on `smoker` and running `describe` we get

	mean	standard deviation	min	max
smoker	1.00	0.00	1	1
	0.00	0.00	0	0
smkban	0.53	0.50	0	1
	0.63	0.48	0	1
age	37.96	11.61	18	78
	38.93	12.26	18	88
hsdrop	0.14	0.35	0	1
	0.08	0.26	0	1
hsgrad	0.42	0.49	0	1
	0.30	0.46	0	1
colsome	0.28	0.45	0	1
	0.28	0.45	0	1
colgrad	0.11	0.31	0	1
	0.22	0.42	0	1
black	0.08	0.27	0	1
	0.08	0.27	0	1
hispanic	0.10	0.30	0	1
	0.12	0.32	0	1
female	0.54	0.50	0	1
	0.57	0.49	0	1

First line is for smokers, second line for non smokers

We see that non smokers are on average older and there are 10 people which are older than the oldest smoker. That fits to my intuition that non smokers life longer. We also see that non smokers have a higher education (less high school drop, higher college graduation). Both smokers and non smokers have about the same proportion of Black and Hispanic people and gender seems to be equally distributed too.

(c) Just do 2 t-tests

```
1 t.test(smoker$hsdrop, nonSmoker$hsdrop, var.equal = TRUE)
2 t.test(smoker$female, nonSmoker$female, var.equal = TRUE)
```

The first t-test is giving a p-value of $< 2.2 \cdot 10^{-16}$ and the other has a p-value of 0.001654 so in both cases we reject the null hypothesis which means that there are differences in *hsdrop* and *female*.

(d) The biggest problem I see here is that smoking is not randomly assigned so a standard linear regression or DiD won't work here. You might try an IV approach but then you need a good instrument.

(e) I have a bit of a problem with this question. For me "affected" means that someone has to smoke and has a smoking ban, so in the group of a affected people we have 100% smokers. Not affected by a smoking ban are people that either don't smoke or don't have a smoking ban. We can test for differences:

```
1 affected = data %>% filter(smoker == 1 & smkban == 1)
```

```
notAffected = data %>% filter(smoker == 0 | smkban == 0)
t.test(affected$smoker, notAffected$smoker, var.equal = TRUE)
```

giving us a p-value of $< 2.2 \cdot 10^{-16}$ (means: 1.0000000 vs. 0.1297806). But after reading (f) I came back to this question and thought that it might give us more insights if we split the data at *smkban*:

```
smokeBan = data %>% filter(smkban == 1)
noSmokeBan = data %>% filter(smkban == 0)
t.test(smokeBan$smoker, noSmokeBan$smoker, var.equal = TRUE)
```

which also gives a p-value of $< 2.2 \cdot 10^{-16}$ (means: 0.2120367 vs. 0.2895951). So we can say that a smoking ban seems to have an effect on smoking.

(f) Fitting the LPM:

```
lpm = lm(smoker ~ ., data = data)
summary(lpm)
```

gives us a effect of -0.0453435. So if there is a smoking ban you have a 4.53% less chance to be a smoker. If we compare this with the differences in means from (e) which is -0.0775584 we can see differences in the estimated effects. This could be due to pre-emptive behaviour: If a company has a smoking ban then it might attract more non smokers because the smokers will go to companies without a smoking ban. One other reason is that in (e) we had far less variables that could explain the difference in smoking status. So the effect on *smkban* is higher.

(g) Running

```
describe(lpm$fitted.values)
```

gives

	mean	**standard deviation**	**min**	**max**
fitted values	0.24	0.1	-0.08	0.49

We see some negative probabilities in our fitted values! That is not good since negative probabilities don't make any sense. We can check and see that in total 34 people have a negative probability. On the other hand we got the same mean as in (a).

(h) Fit a probit model:

```
lpmProbit = glm(smoker ~ ., data = data, family = binomial(
    link = "probit"))
summary(lpmProbit)
```

we get a coefficient of *smkban* of -0.151762 but since only the sign of this coefficient is interpretable when it comes to marginal effects we know that a smoking ban reduces the probability of being a smoker.

(i) The marginal effects are

```
probitmfx(smoker ~ ., data = data, atmean = TRUE)
summary(margins(lpmProbit, type = "response"))
```

the first function calculates the MEA (marginal effect at the average, or sometimes called MEM: marginal effect at the mean) which is -0.04652485. The other function gives the average marginal

effect (AME) which is here -0.0449. These two numbers are not the same because MEA and AME are not the same: MEA is the marginal effect for an average person from the dataset, the AME is the average over all the treatment effects from all the persons in the dataset.

(j) We can see that the effects we got from the LPM and the probit model are all around 4.5% to 4.6%. I always prefer easier approaches I would be fine with the LPM although it has the problem with negative probabilities in the fitted values. But these probabilities are barely below 0 so in some way acceptable. We can also take into account the computing time, on my machine the `probitmfx` and `margins` functions take significantly longer to compute than the `lm` function.

References

[1] Joshua Angrist and Stacey Chen. "Long-term economic consequences of Vietnam-era conscription: Schooling, experience and earnings". In: *SSRN Electronic Journal* (2008). DOI: 10.2139/ssrn.1214917.

[2] Gaddis Smith, Lawrence M. Baskir, and William A. Strauss. "Chance and circumstance: The draft, the war, and the Vietnam Generation". In: *Foreign Affairs* 57.1 (1978), p. 221. DOI: 10.2307/20040081.

[3] H. C. Tijms. *Understanding probability*. Cambridge University Press, 2012.