Econ 400 Problem Set 2

Travis Cao

February 9, 2021

(Travis's suggested guide on solving Q3. All Stata commands are written in red color)

(a) Before performing any analysis, let's import the dataset to Stata (make sure you first change your working directory to where the dataset is stored):

```
use "CPS96_15.dta", clear
```

(i) Use the following command to compute sample mean for AHE in 1996 and 2015:

```
tabstat ahe, by(year) statistics(mean)
```

(ii) Use the following command to compute sample standard deviation for AHE in 1996 and 2015:

```
tabstat ahe, by(year) statistics(sd)
```

(iii) Use the following commands to compute 95% confidence interval (CI) for mean of AHE in 1996 and 2015:

```
ci means ahe if year == 1996, level(95)
ci means ahe if year == 2015, level(95)
```

(iv) This is the tricky bit of the problem set. The question asks you to compute the 95% confidence interval for

$$\overline{AHE}_{2015} - \overline{AHE}_{1996}$$

The difficult bit is to calculate the standard error at this mean statistic. Let's start off by calculating the variance at this mean statistic:

$$\begin{aligned}
&Var(\overline{AHE}_{2015} - \overline{AHE}_{1996}) \\
&= Var(\overline{AHE}_{2015}) + Var(\overline{AHE}_{1996}) - 2Cov(\overline{AHE}_{2015}, \overline{AHE}_{1996}) \\
&= Var(\overline{AHE}_{2015}) + Var(\overline{AHE}_{1996}) \\
&\qquad\qquad \text{(The two statistics should be uncorrelated, so Cov term} = 0) \\
&= Var\left(\frac{1}{n_{2015}}\sum_{i=1}^{n_{2015}} AHE_{2015,i}\right) + Var\left(\frac{1}{n_{1996}}\sum_{i=1}^{n_{1996}} AHE_{1996,i}\right) \\
&= \frac{1}{n_{2015}^2} Var\left(\sum_{i=1}^{n_{2015}} AHE_{2015,i}\right) + \frac{1}{n_{1996}^2} Var\left(\sum_{i=1}^{n_{1996}} AHE_{1996,i}\right) \\
&= \frac{1}{n_{2015}^2}\sum_{i=1}^{n_{2015}} Var(AHE_{2015,i}) + \frac{1}{n_{1996}^2}\sum_{i=1}^{n_{1996}} Var(AHE_{1996,i}) \\
&\qquad\qquad \text{(Each individual observation should be uncorrelated)}
\end{aligned}$$

$$= \frac{1}{n_{2015}^2} n_{2015} Var(AHE_{2015,i}) + \frac{1}{n_{1996}^2} n_{1996} Var(AHE_{1996,i})$$

(Random sampling implies i.i.d. dataset)

$$= \frac{1}{n_{2015}} Var(AHE_{2015,i}) + \frac{1}{n_{1996}} Var(AHE_{1996,i})$$

This means that the standard error at this mean statistic is thus

$$se(\overline{AHE}_{2015} - \overline{AHE}_{1996}) = \sqrt{Var(\overline{AHE}_{2015} - \overline{AHE}_{1996})}$$
$$= \sqrt{\frac{1}{n_{2015}} Var(AHE_{2015,i}) + \frac{1}{n_{1996}} Var(AHE_{1996,i})}$$

where n_{2015} is the number of observations in year 2015, n_{1996} is the number of observations in year 1996.

Now,

- Result from (i) gives us $\overline{AHE}_{2015} - \overline{AHE}_{1996}$
- Result from (ii) gives us $sd(AHE_{2015,i})$ and $sd(AHE_{1996,i})$, which are just square root of $Var(AHE_{2015,i})$ and $Var(AHE_{1996,i})$
- Result from (iii) tells us number of observations in each year group (so we have n_{2015} and n_{1996})

So we can calculate $se(\overline{AHE}_{2015} - \overline{AHE}_{1996})$. Once this is obtained, the 95% confidence interval is constructed to be

$$[\overline{AHE}_{2015} - \overline{AHE}_{1996} - 1.96 \times se(\overline{AHE}_{2015} - \overline{AHE}_{1996}),$$
$$\overline{AHE}_{2015} - \overline{AHE}_{1996} + 1.96 \times se(\overline{AHE}_{2015} - \overline{AHE}_{1996})]$$

(b) We can adjust 1996 data for inflation so that it can be expressed in terms of 2015 dollars. Here,

$$\frac{\$ \text{ worth in } 2015}{\$ \text{ worth in } 1996} = \frac{CPI_{2015}}{CPI_{1996}}$$
$$\$ \text{ worth in } 2015 = \frac{CPI_{2015}}{CPI_{1996}} \times \$ \text{ worth in } 1996$$

So to adjust 1996 data for inflation, we use the following command in Stata:

```
replace ahe = ahe * (237.0 / 156.9) if year == 1996
```

Then we can repeat the same analysis described in (a).

(c) We'd like to use real (i.e. inflation adjusted) data for comparison, so result generated in (b) will be suitable.

(d) (i)
```
ci means ahe if year == 2015 & bachelor == 0, level(95)
```
(ii)
```
ci means ahe if year == 2015 & bachelor == 1, level(95)
```
(iii) Similar to what we did in (a), now we are interested in constructing CI for

$$\overline{AHE}_{2015,college} - \overline{AHE}_{2015,highschool}$$

- $\overline{AHE}_{2015,college} - \overline{AHE}_{2015,highschool}$ can be calculated from (i) and (ii)

- In the `ci` table created by (i) and (ii), the column `Std.Err.` was already reporting the standard error at mean. That is, the `Std.Err.` column for, say part (i) of this question, was reporting

$$Var(\overline{AHE}_{2015,highschool}) = \frac{Var(AHE_{2015,highschool})}{n_{2015,highschool}}$$

Thus,

$$\begin{aligned} &se(\overline{AHE}_{2015,college} - \overline{AHE}_{2015,highschool}) \\ &= \sqrt{Var(\overline{AHE}_{2015,college}) - Var(\overline{AHE}_{2015,highschool})} \end{aligned}$$

With $\overline{AHE}_{2015,college} - \overline{AHE}_{2015,highschool}$ and $se(\overline{AHE}_{2015,college} - \overline{AHE}_{2015,highschool})$, confidence interval can be constructed in a similar fashion as question (a)(iv).

(e) Almost identical to what you did in part (d), except that `year == 1996` here.

(f) (i) Construct confidence interval for the following to answer this question (the way to construct confidence interval is very similar to (a)(iv) and (d)(iii)):

$$\overline{AHE}_{2015,highschool} - \overline{AHE}_{1996,highschool}$$

One way to answer the question using this confidence interval: If 0 is included in this confidence interval, or if the confidence interval contains signifcant portion of negative numbers, then $\overline{AHE}_{2015,highschool} - \overline{AHE}_{1996,highschool}$ is likely not going to be positive, which means real wages of high school graduates did not increase from 1996 to 2015.

(ii) Construct confidence interval for the following to answer this question:

$$\overline{AHE}_{2015,college} - \overline{AHE}_{1996,college}$$

(iii) Construct confidence interval for the following to answer this question:

$$\overline{GAP}_{2015} - \overline{GAP}_{1996}$$

Here,

- $\overline{GAP}_{2015} = \overline{AHE}_{2015,college} - \overline{AHE}_{2015,highschool}$ from (d)(iii)
- $\overline{GAP}_{1996} = \overline{AHE}_{1996,college} - \overline{AHE}_{1996,highschool}$ from (e)(iii)
-

$$\begin{aligned} &se(\overline{GAP}_{2015} - \overline{GAP}_{1996}) \\ &= \sqrt{Var(\overline{GAP}_{2015} - \overline{GAP}_{1996})} \\ &= \sqrt{Var(\overline{GAP}_{2015}) + Var(\overline{GAP}_{1996})} \end{aligned}$$

(Again, 1996 and 2015 data are uncorrelated)

where $Var(\overline{GAP}_{2015})$ and $Var(\overline{GAP}_{1996})$ have been calculated in (d)(iii) and (e)(iii)

(g) To prepare the similar table for high school graduates, the following Stata commands will be helpful in terms of calculating certain statistics.

```
mean ahe if year == 1996 & female == 0 & bachelor == 0
mean ahe if year == 1996 & female == 1 & bachelor == 0
mean ahe if year == 2015 & female == 0 & bachelor == 0
mean ahe if year == 2015 & female == 1 & bachelor == 0
tabstat ahe if bachelor == 0 & female == 0, by(year) statistics(sd)
tabstat ahe if bachelor == 0 & female == 1, by(year) statistics(sd)
```

The technique for recreating the difference in Men vs. Women section is very similar to what we have been doing throughout this problem set.