Homework 4: Making Inferences

- A. Short answer
- B. Empirical exercise
- Submission

Start Over

B. Empirical exercise

In this exercise, we will continue our gender gap analysis with the March 2009 CPS, turning attention to the connection between wages and hours.

Reading and transforming

Question 1:

Like before, we start by loading the data.

R Code Start Over Run Code Submit Answer

```r
library(readxl)
cps_mar <- read_xlsx("./data/cps09mar.xlsx")

```

That's marvelous! Correct!

Next, let's do the transformations we'll need for this week's exercise. First, pipe the data to `filter` and keep only those observations with positive earnings, then filter to ages 23-62. Then, pipe the filtered data to `mutate` to create `gender`, `wage`, `lwage`, `over40` (indicator for working over 40 hours per week), and `CollDeg` (indicator for having a college degree) variables. We are not checking your work here, but you must get this right to move on.

Continue

Average wages and hours by gender

Now, let's establish some basic empirical facts about the sample of 23-62 year-olds on which our analysis is based. Create a table of summary statistics (N, Mean, and SD) for `wage`, `hours` `over40`, and `CollDeg` by gender and answer a few questions based on the table.

R Code Start Over Run Code

```r
cps_mar_2362 <- cps_mar %>%
  filter(earnings > 0, age >= 23, age <= 62) %>%
  mutate(
    gender = ifelse(female == 1, "Female", "Male"),
    wage = earnings / (hours * week),  # Updated wage calculatio
    lwage = log(wage),
    over40 = ifelse(hours > 40, 1, 0),
    CollDeg = ifelse(education >= 16, 1, 0)
  )
# Create a nice table
datasummary(
    wage + hours + over40 + CollDeg ~ gender * (N + Mean + SD),
  data = cps_mar_2362,
  title = "Table 1. Summary statistics by gender, 23-62 year-old
)

# Calculate the percentage of women in the dataset
percentage_women <- (sum(cps_mar_2362$gender == "Female") / nrow

# Round to the nearest whole number
percentage_women_rounded <- round(percentage_women)
```

Table 1. Summary statistics by gender, 23-62 year-olds

	Female			Male		
	N	Mean	SD	N	Mean	SD
wage	20392	20.26	15.25	27279	27.21	23.53
hours	20392	42.38	6.24	27279	45.02	8.53
over40	20392	0.23	0.42	27279	0.39	0.49
CollDeg	20392	0.38	0.49	27279	0.36	0.48

```
Women comprise 43 % of our sample of 23-62 year-olds.
```

```
Men work on average 2.64 more hours per week than women. (Rep
ort to two decimals.)
```

```
The percentage wage gap is roughly 34 %.
```

Question 2:

Women comprise ____ % of our sample of 23-62 year-olds. (Round to the nearest whole number.)

43

Correct!

Question 3:

Men work on average _____ more hours per week than women. (Report to two decimals.)

2.64

Correct!

Question 4:

The percentage wage gap is roughly _____ %.

34

Correct!

Question 5:

Wages are more variable among _____ and hours are more variable among _____.

males, males

Correct!

Question 6:

Men are _____ percentage points more likely to work more than 40 hours.

16

Correct!

Question 7:

Women are _____ percentage points more likely to have a college degree.

2

Correct!

Continue

College degrees and long hours

Now, let's return to the end of the Week 4 slide deck, where the focus was on 23-30 year-olds who work more than 40 hours per week, and drill down further to college graduates.

The first task will be to filter down to this group and check average wages and hours by gender and answer a few questions.

R Code Start Over Run Code

```r
cps_mar_2330_coll40 <- cpsmar_2362 %>%
  filter(
    age>= 23 & age<= 30,
    over40  == 1,
    CollDeg == 1
  )
datasummary(
   wage + hours + over40 + CollDeg ~ gender * (N + Mean + SD),
  data=cps_mar_2330_coll40,
  title="Table 2. Summary statistics, 23-30 year-olds with colle
  )

  # Count the total number of individuals in this group
```

Table 2. Summary statistics, 23-30 year-olds with college degrees and working >40 hours

	Female			Male		
	N	Mean	SD	N	Mean	SD
wage	376	19.92	9.55	492	24.45	16.59
hours	376	50.54	7.02	492	52.79	9.90
over40	376	1.00	0.00	492	1.00	0.00
CollDeg	376	1.00	0.00	492	1.00	0.00

```
There are 868 individuals in the sample of 23-30 year-olds wi
th college degrees working more than 40 hours.
```

```
Women comprise 43 % (rounded to the nearest whole number) of
this group.
```

```
Men in this group work on average 2 hours and 15 minutes per
week more than women.
```

```
Among the group documented in Table 2, the gender wage gap is
roughly 23 %.
```

Question 8:

There are _____ individuals in the sample of 23-30 year-olds with college degrees working more than 40 hours and women comprise _____% (round to the nearest whole number) of this group.)

868, 43

Correct!

Question 9:

Men in this group work on average _____ hours and _____ minutes per week more than women.)

2, 15

Correct!

Question 10:

Among the group documented in Table 2, the gender wage gap is roughly _____ %. (Round to the nearest whole number.)

23

Correct!

Next, use `group_by` and `summarise` to estimate $E(logwage|age)$ for women and men in this group, including 95% confidence intervals for each conditional mean. Then print the data frame.

R Code Start Over Run Code

```r
cef_2330_coll40 <- cps_mar_2330_coll40 %>%
  mutate(age=age-23) %>%
  group_by(age, gender) %>%
  summarise(
    lwage_bar =   mean(lwage, na.rm = TRUE),
    lwage_se  =   sd(lwage, na.rm = TRUE)/sqrt(n()),
    upper = lwage_bar  + lwage_se  *1.96,
    lower = lwage_bar  - lwage_se  *1.96
  )
cef_2330_coll40
```

```
`summarise()` has grouped output by 'age'. You can override u
sing the `.groups`
argument.
```

age	gender	lwage_bar	lwage_se	upper	lower
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
0	Female	2.433198	0.14082014	2.709206	2.157191
0	Male	2.743054	0.07823491	2.896394	2.589713
1	Female	2.673079	0.06781115	2.805989	2.540169
1	Male	2.956772	0.10115294	3.155031	2.758512
2	Female	2.829135	0.06254741	2.951728	2.706542
2	Male	2.910640	0.06692109	3.041805	2.779474
3	Female	2.862429	0.05742124	2.974975	2.749884
3	Male	3.084154	0.07168547	3.224658	2.943651
4	Female	2.841282	0.08190148	3.001809	2.680755
4	Male	2.921660	0.07459176	3.067860	2.775460

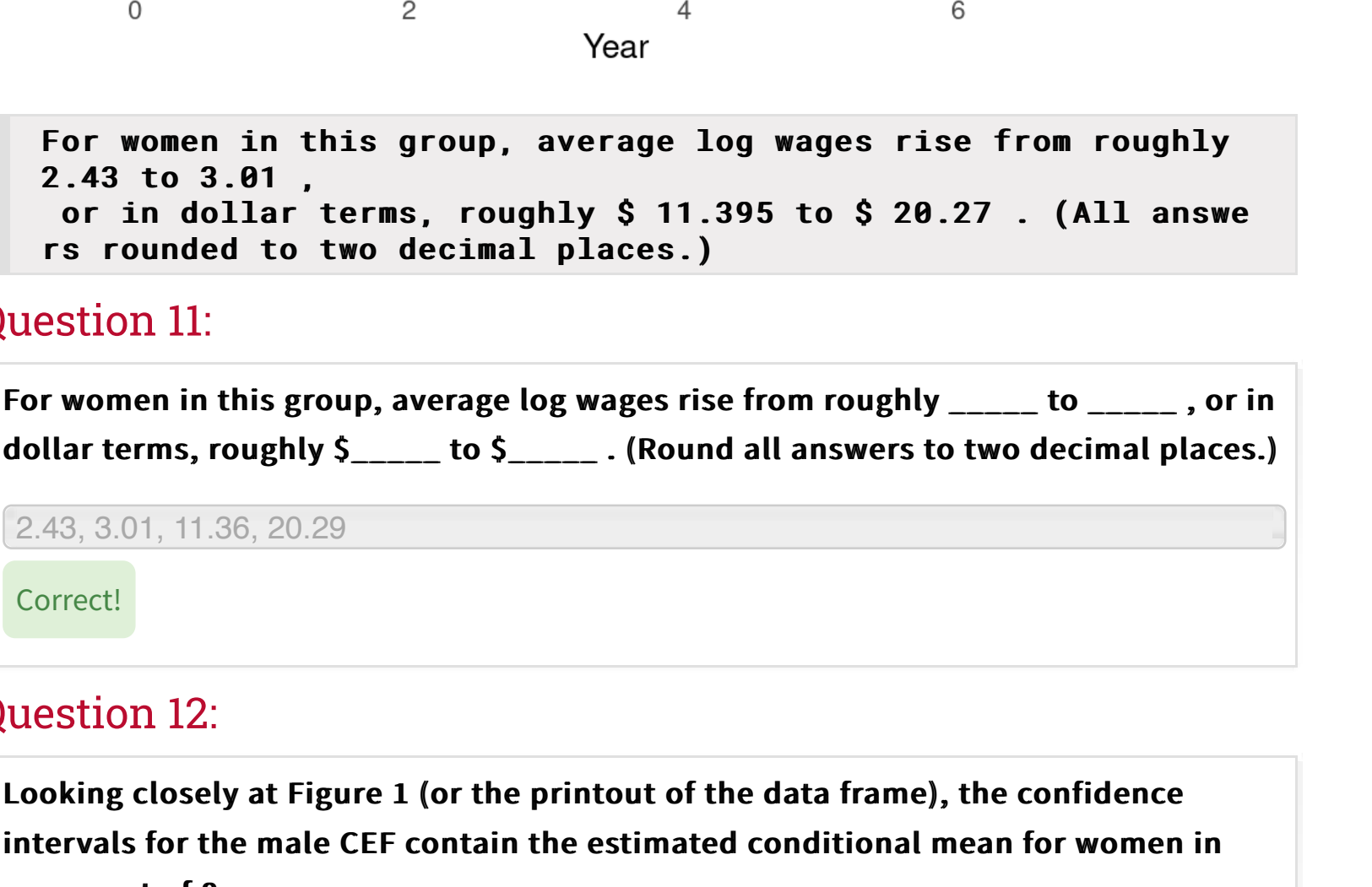
1-10 of 16 rows Previous 1 2 Next

Now, plot the CEFs with confidence intervals and answer a couple questions.

R Code Start Over Run Code

```r
ggplot(cef_2330_coll40, aes(x = age, y = lwage_bar, color = gend
  geom_pointrange(aes(ymin = lower, ymax = upper)) +
  labs(
    title="Figure 1. Log wage profiles by gender, 23-30 year-old
    x="Year",
    y="Average log wage"
    )

    # Filter the data for women
```




```
For women in this group, average log wages rise from roughly
2.43 to 3.01 ,
 or in dollar terms, roughly $ 11.395 to $ 20.27 . (All answe
rs rounded to two decimal places.)
```

Question 11:

For women in this group, average log wages rise from roughly _____ to _____ , or in dollar terms, roughly $_____ to $_____ . (Round all answers to two decimal places.)

2.43, 3.01, 11.36, 20.29

Correct!

Question 12:

Looking closely at Figure 1 (or the printout of the data frame), the confidence intervals for the male CEF contain the estimated conditional mean for women in _____ out of 8 years.

4

Correct!

Finally, let's test the null of no gender gap in log wages among this group and interpret the result.

R Code Start Over Run Code

```r
t.test(lwage ~ gender, data=cps_mar_2330_coll40)


```

```

	Welch Two Sample t-test

data:  lwage by gender
t = -3.7584, df = 865.31, p-value = 0.0001825
alternative hypothesis: true difference in means between grou
p Female and group Male is not equal to 0
95 percent confidence interval:
 -0.21391772 -0.06714274
sample estimates:
mean in group Female   mean in group Male
            2.881276             3.021806
```

Question 13:

The absolute value of the test statistic is _____. (Round to two decimal places.)

3.76

Correct!

Question 14:

The value of the test statistic suggests that the null is _____ at the 5% level.

rejected

Correct!

Continue

Previous Topic Next Topic