Problem Set 1

Scott Cohn

Last compiled on 05 March, 2021

Set up

The data cleaning is done in python. See `ipums_clean.py` in my GitHub repository here (link available in PDF).

To run the cleaning script, run the following chunk in R:

```
library(reticulate)

py_run_file("ipums_clean.py")
```

Once these data are cleaned, we import the cleaned version:

```
ipums_fl <- read.csv("data/ipums_FL.csv")
```

Part A

> *Produce detailed summary statistics that describe the sample distributions of weekly hours of work, annual hours of work, and hourly wages — separately for males and females. Examine your data to make sure that the variables in your data set take on meaningful values for all of the observations in your analysis samples.*

Note that for this sample, we are limited to using `WKSWORK2` which records weekly hours in intervals.

```
ipums_fl %>%
  filter(sex == "male") %>%
  select(wkswork2, annl_hrs_wrkd, hourly_wage) %>%
  modelsummary::datasummary_skim(
    title = "Summary statistics -- Male",
    notes = "Variables include weeks worked, annual hours worked, and hourly wage",
    histogram = F,
    output = "kableExtra"
    ) %>%
  kable_paper(full_width = F)
```

Table 1: Summary statistics – Male

	Mean	SD	P0	P25	P50	P75	P100
wkswork2	4.7	2.3	0.0	4.0	6.0	6.0	6.0
annl_hrs_wrkd	2074.1	682.1	7.0	2040.0	2040.0	2295.0	5049.0
hourly_wage	369.0	3298.9	-24.3	21.8	38.7	66.2	178571.4

Variables include weeks worked, annual hours worked, and hourly wage

Table 2: Summary statistics – Female

	Mean	SD	P0	P25	P50	P75	P100
wkswork2	4.3	2.5	0.0	2.0	6.0	6.0	6.0
annl_hrs_wrkd	1835.3	674.3	7.0	1632.0	2040.0	2040.0	5049.0
hourly_wage	164.6	1660.2	-20.6	23.1	41.7	73.5	119047.6

Variables include weeks worked, annual hours worked, and hourly wage

```
ipums_fl %>%
  filter(sex == "female") %>%
  select(wkswork2, annl_hrs_wrkd, hourly_wage) %>%
  modelsummary::datasummary_skim(
    title = "Summary statistics -- Female",
    notes = "Variables include weeks worked, annual hours worked, and hourly wage",
    histogram = F,
    output = "kableExtra"
    ) %>%
  kable_paper(full_width = F)
```

We can also use graphs.

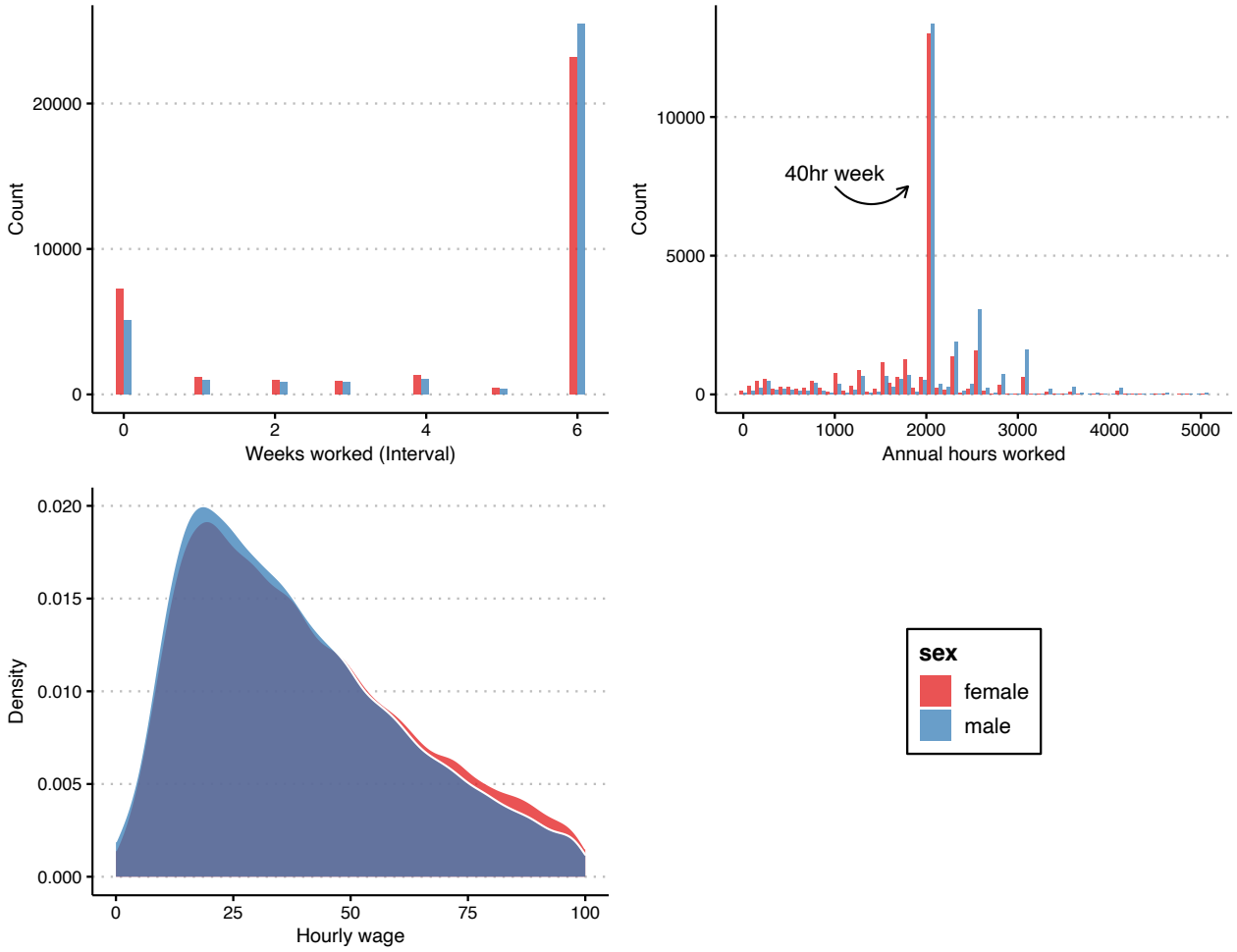
```
p1 <-
  ipums_fl %>%
  select(sex, wkswork2, annl_hrs_wrkd, hourly_wage) %>%
  ggplot() +
  geom_histogram(aes(wkswork2, fill = sex),
                 position = "dodge", alpha = 0.75) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "Weeks worked (Interval)", y = "Count") +
  theme_clean() +
  theme(plot.background = element_rect(color = "white"))

df <- data.frame(x1 = 1000, x2 = 1800, y1 = 7500, y2 = 7500)

p2 <-
  ipums_fl %>%
  select(sex, wkswork2, annl_hrs_wrkd, hourly_wage) %>%
  ggplot() +
  geom_histogram(aes(annl_hrs_wrkd, fill = sex), bins = 60,
                 position = "dodge", alpha = 0.75) +
  annotate(geom = "text", x = 1000, y = 8000, label = "40hr week") +
  geom_curve(
    aes(x = x1, y = y1, xend = x2, yend = y2),
    data = df,
    arrow = arrow(length = unit(0.03, "npc"))
    ) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "Annual hours worked", y = "Count") +
  theme_clean() +
  theme(plot.background = element_rect(color = "white"),
        legend.position = "none")
```

```
p3 <-
  ipums_fl %>%
  select(sex, wkswork2, annl_hrs_wrkd, hourly_wage) %>%
  ggplot() +
  geom_density(aes(hourly_wage, fill = sex), color = "white", alpha = 0.75) +
  xlim(0, 100) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "Hourly wage", y = "Density") +
  theme_clean() +
  theme(plot.background = element_rect(color = "white"),
        legend.position = "none")

p1 + p2 + p3 + guide_area() +
  plot_layout(guides = 'collect')
```



Part B

Estimate labor supply models — one for males and one for females — using only data on individuals who worked positive hours in the previous year in paid, wage employment (i.e., not self-employed). Your measure of labor supply (the dependent variable) should be the natural log of annual hours worked. Your explanatory variables should be the natural log of the hourly wage, the natural log of nonlabor income, and a vector of any additional explanatory variables that you believe might control for systematic differences across individuals in "tastes for market work" at a given wage

and nonlabor income.

```
mod1 <- lm_robust(
  formula = log_annl_hrs_wrkd ~ log_hourly_wage + log_nonlabor_inc + female +
    nchild + age + hispan_d + black_d + asian_d,
  data = ipums_fl,
  subset = annl_hrs_wrkd > 0,
  se_type = "stata"
)

# control for ed
mod2 <- lm_robust(
  formula = log_annl_hrs_wrkd ~ log_hourly_wage + log_nonlabor_inc + female +
    nchild + age + hispan_d + black_d + asian_d + educ2_d + educ3_d + educ4_d + educ5_d,
  data = ipums_fl,
  subset = annl_hrs_wrkd > 0,
  se_type = "stata"
)

# control for occ
mod3 <- lm_robust(
  formula = log_annl_hrs_wrkd ~ log_hourly_wage + log_nonlabor_inc + female +
    nchild + age + hispan_d + black_d + asian_d +
    educ2_d + educ3_d + educ4_d + educ5_d +
    occ_comp_eng + occ_edu_leg_art + occ_health_tech + occ_serv + occ_sales +
    occ_office + occ_farm_fish + occ_constr + occ_maintn + occ_prod + occ_transport,
  data = ipums_fl,
  subset = annl_hrs_wrkd > 0,
  se_type = "stata"
)

models = list(
  "Model 1" = mod1,
  "Model 2" = mod2,
  "Model 3" = mod3
)

modelsummary(models, stars = T,
             title = "Part B Models",
             notes = "Dependent variable is logged annual hours worked") %>%
  row_spec(9:10, bold = TRUE, background = 'pink')
```

Table 3: Part B Models

	Model 1	Model 2	Model 3
(Intercept)	7.032***	6.827***	7.080***
	(0.036)	(0.038)	(0.039)
age	0.006***	0.006***	0.005***
	(0.000)	(0.000)	(0.000)
asian_d	-0.027**	-0.057***	-0.057***
	(0.014)	(0.013)	(0.013)
black_d	-0.100***	-0.060***	-0.038***
	(0.008)	(0.008)	(0.008)
female	**-0.193*** **	**-0.219*** **	**-0.216*** **
	(0.005)	**(0.005)**	**(0.005)**
hispan_d	-0.091***	-0.068***	-0.055***
	(0.006)	(0.006)	(0.006)
log_hourly_wage	-0.501***	-0.523***	-0.530***
	(0.007)	(0.007)	(0.007)
log_nonlabor_inc	0.219***	0.228***	0.232***
	(0.005)	(0.005)	(0.005)
nchild	0.004*	0.001	-0.001
	(0.002)	(0.002)	(0.002)
educ2_d		0.072***	0.039***
		(0.014)	(0.014)
educ3_d		0.167***	0.101***
		(0.014)	(0.014)
educ4_d		0.277***	0.171***
		(0.014)	(0.014)
educ5_d		0.403***	0.282***
		(0.014)	(0.015)
occ_comp_eng			-0.045***
			(0.010)
occ_constr			-0.261***
			(0.014)
occ_edu_leg_art			-0.242***
			(0.009)
occ_farm_fish			-0.272***
			(0.049)
occ_health_tech			-0.040***
			(0.010)
occ_maintn			-0.193***
			(0.013)
occ_office			-0.189***
			(0.009)
occ_prod			-0.196***
			(0.015)
occ_sales			-0.193***
			(0.009)
occ_serv			-0.319***
			(0.008)
occ_transport			-0.263***
			(0.012)
Num.Obs.	43481	43481	43481
R2	0.399	0.430	0.455
R2 Adj.	0.398	0.430	0.455
se_type	HC1	HC1	HC1

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable is logged annual hours worked

Part C

1. The coefficient for `log_hourly_wage` gives an estimate of the uncompensated (i.e., with nonlabor income held constant) wage elasticity of supply. We can interpret that coefficient as the percentage change in labor supply for a 1% change in the (gross) wage. Yes, the sign makes sense: An increase in wage would lead to a decrease in supply holding all else fixed.
2. The coefficient for `log_nonlabor_inc` gives an estimate of the (nonlabor) income elasticity of supply. Yes the sign makes sense, or at least it would be ambiguous. It would lead to an income effect, but the size of the substitution effect is unclear.
3.