Sofia Guo PSET 5 Econ 142

Sofia Guo

3/1/2019

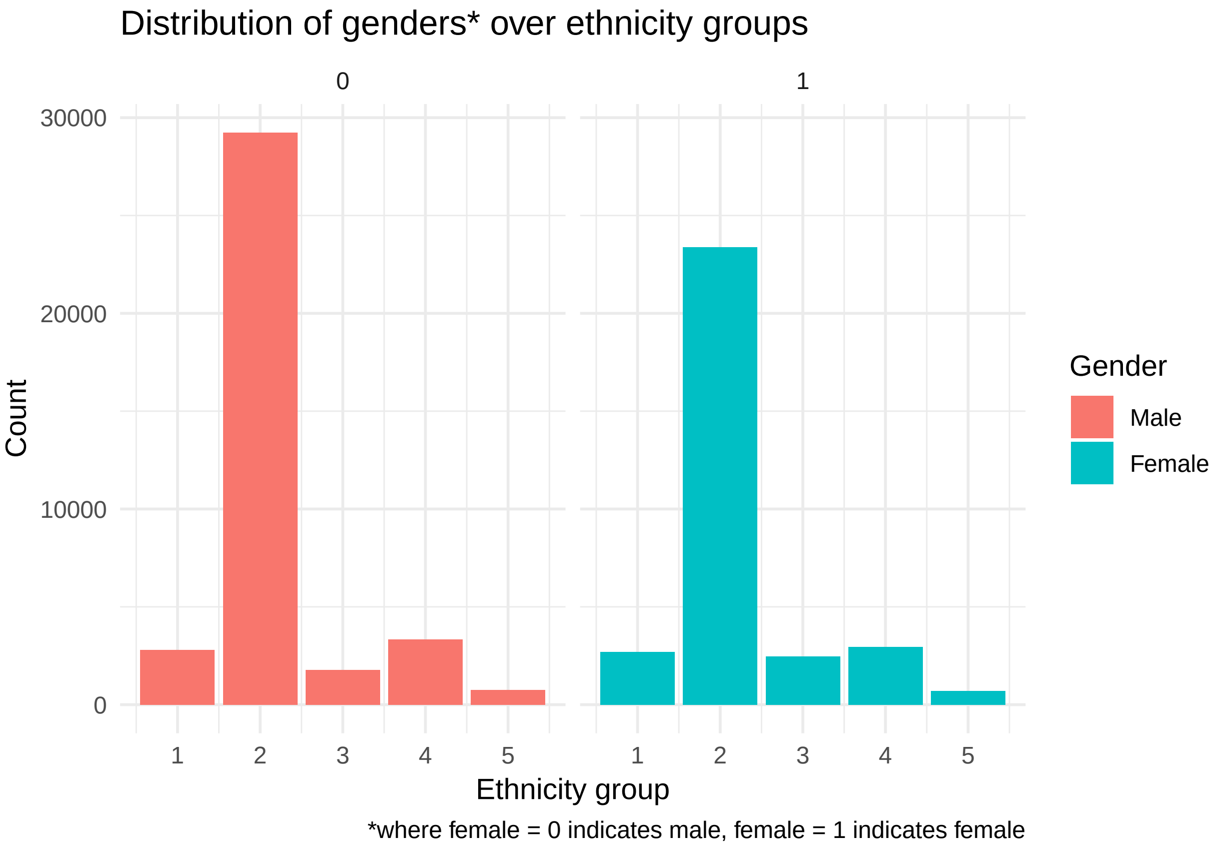
```
#load libraries
library(dplyr)
library(tidyr)
library(magrittr)
library(sandwich)
library(ggplot2)
library(stargazer)
library(kableExtra)
library(janitor)
library(modelr)
```

1. Constructing the 5-group ethnicity variable and distribution

We construct the ethnicity variable categories as the following:

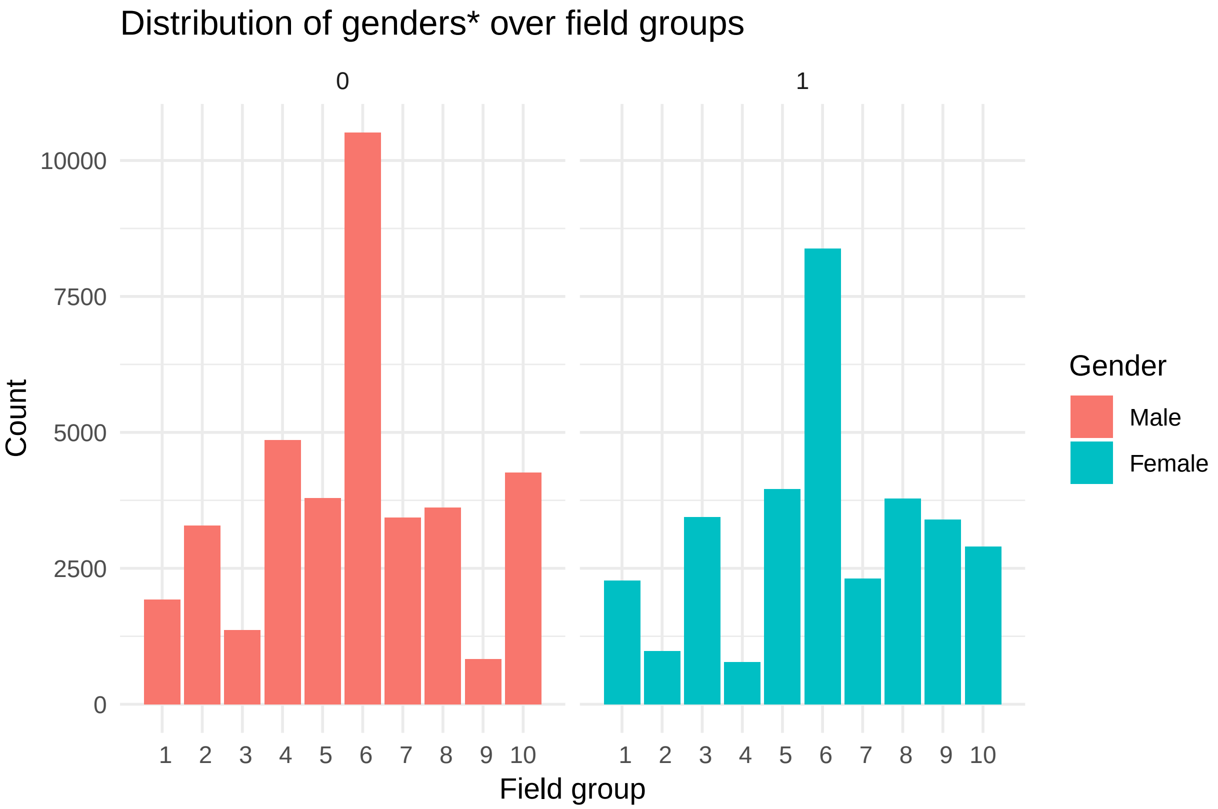
1. HA = hispanic (any race)
2. WNH = white non-hispanics
3. BNH = black non-hispanics
4. ANH = asian non-hispanics
5. ONH = other non-hispanics

```
#read in data
deg <- read.csv('/Users/sofia/Box/Cal (sofiaguo@berkeley.edu)/2018-19/Spring 2019/Econ 142/PSETS/PSET 5/

#construct ethnicity variable
deg_eth <- deg %>%
  mutate(HA = 1*as.numeric(hispanic == 1),
         WNH = 2*as.numeric(hispanic == 0 & race == 1),
         BNH = 3*as.numeric(hispanic == 0 & race == 2),
         ANH = 4*as.numeric(hispanic == 0 & race == 3),
         ONH = 5*as.numeric(hispanic == 0 & race == 4),
         eth = HA + WNH + BNH + ANH + ONH,
         male = as.numeric(female ==0))

#show distributions of males and females in these 5 categories
ggplot(deg_eth, aes(eth)) +
  geom_histogram(aes(fill = factor(female)), stat = "count", show.legend = T, binwidth = 1) +
  labs(x = 'Ethnicity group', y = 'Count',
       title = 'Distribution of genders* over ethnicity groups',
       caption = '*where female = 0 indicates male, female = 1 indicates female')+
 scale_fill_discrete(name="Gender",
                        breaks=c("0", "1"),
                        labels=c("Male", "Female")) +
  facet_wrap(~female) +
  theme(text = element_text(size=10, family="LM Roman 10")) +
  theme_minimal()
```



2(a). Fraction of M/F in each category of "mfield" and mean log-wage of both groups

```
#Display distributions first
ggplot(deg_eth, aes(mfield)) +
  geom_histogram(aes(fill = factor(female)), stat = "count", show.legend = T, binwidth = 1) +
  labs(x = 'Field group', y = 'Count',
       title = 'Distribution of genders* over field groups',
       caption = '*where female = 0 indicates male, female = 1 indicates female')+
    scale_fill_discrete(name="Gender",
                        breaks=c("0", "1"),
                        labels=c("Male", "Female")) +
  facet_wrap(~female) +
  scale_x_discrete(limits = 1:10) +
  theme_minimal()
```



*where female = 0 indicates male, female = 1 indicates female

```
#obtain fractions of males and females in each mfield category
wg_frac <- deg_eth %>%
  group_by(mfield, female) %>%
  summarise (n = n(),
             avgwage = mean(logwage)) %>%
  mutate(genfraction = n / sum(n)) %>%
  spread(female, avgwage) %>%
  select( -n)

wg_frac_male <- wg_frac %>%
  filter(`0` != "NA")

wg_frac_fem <- wg_frac %>%
  filter(`1` != "NA")

wg_frac_all <- merge(wg_frac_male, wg_frac_fem, by = "mfield") %>%
  remove_empty("cols")
```

Table 1: Fraction of genders and average logwage by field group

	Male		Female	
Field Group	Fraction	Mean log(wage)	Fraction	Mean log(wage)
1	0.46	7.11	0.54	7.02
2	0.77	7.34	0.23	7.17
3	0.28	6.89	0.72	6.77
4	0.86	7.40	0.14	7.32
5	0.49	7.20	0.51	6.97
6	0.56	7.23	0.44	7.07
7	0.60	7.09	0.40	6.99
8	0.49	7.05	0.51	6.92
9	0.20	7.32	0.80	7.19
10	0.59	7.09	0.41	6.88

```
#display table
kable(wg_frac_all,"latex", caption = "Fraction of genders and average logwage by field group",  digits =
  kable_styling(position = "center")  %>% add_header_above(c(" " = 1, "Male" = 2, "Female" = 2))
```

2(b). Weighted average mean logwage by gender

```
wtmlavg <- mean(wg_frac_all$genfraction.x * wg_frac_all$`0.x`)
wtmlavg
```

```
## [1] 3.814006
```

```
wtflavg <- mean(wg_frac_all$genfraction.y * wg_frac_all$`1.y`)
wtflavg
```

```
## [1] 3.291108
```

```
ml_fl_gap <- wtmlavg - wtflavg
ml_fl_gap
```

```
## [1] 0.5228976
```

The male-female log wage gap is approximately 0.5229.

```
#compute the counterfactual mean for females if they had the male distribution

counterf_fem <- sum(wg_frac_all$genfraction.x * wg_frac_all$`1.y`)/10
counterf_fem
```

```
## [1] 3.739719
```

```
ml_fl_gap_counterf <- wtmlavg - counterf_fem
ml_fl_gap_counterf
```

```
## [1] 0.07428688
```

```
ml_fl_gap_counterf/ml_fl_gap
```

```
## [1] 0.1420677
```

Approximately 0.0742/0.5229 = 0.142 or 14.2% of the female-male log wage gap is “explained” by mfield.

2(c) Counterfactual mean for males

```
#compute the counterfactual mean for males if they had the female distribution

counterf_male <- sum(wg_frac_all$genfraction.y * wg_frac_all$`0.x`)/10
counterf_male
```

```
## [1] 3.357493
```

```
ml_fl_gap_counterf <- counterf_male - wtflavg
ml_fl_gap_counterf
```

```
## [1] 0.06638542
```

```
ml_fl_gap_counterf/ml_fl_gap
```

```
## [1] 0.1269568
```

Approximately 0.0664/0.5229 = 0.1270 or 12.7% of the female-male log wage gap is “explained” by mfield. This answer is 1.5% less than the answer in part (b) because of the variation between mean logwage unexplained by the variation in distribution across mfield (the gender wage premium).

3(e) Fitting a logistic model

Fit the model:

$$p_i = p(m_i = 1|male) = \theta_0 + \theta_1 age_i + \theta_2 age_i^2 + \gamma_i D_{i_{ethnicity}} + \lambda_i D_{i_{mfield}} + \theta_3 age_i * D_{i_{mfield}} + \mu_i \tag{1}$$

```
#fit logistic model for the probability of being male as a function of age, race and mfield

log_mod <- (glm(factor(male) ~ poly(AGEP, 2) + factor(eth) + factor(mfield) + AGEP*factor(mfield), data

stargazer(log_mod, type = "latex", title = "Logistic model for probability of being a male",
          header = F,
          multicolumn = F,
          column.sep.width = '0.1pt',
          single.row = T,
          omit = c("eth"))
```

```
#store the predicted values
p_i <- log_mod$fitted.values

#make the weight
w_i <- p_i/(1-p_i)

#add weights to data frame
deg_eth_pred <- deg_eth %>%
  mutate(weights = w_i)
```

Table 2: Logistic model for probability of being a male

	Dependent variable:
	factor(male)
poly(AGEP, 2)1	75.885^{***} (8.352)
poly(AGEP, 2)2	-9.051^{***} (2.139)
factor(mfield)2	3.812^{***} (0.373)
factor(mfield)3	0.421 (0.341)
factor(mfield)4	2.664^{***} (0.377)
factor(mfield)5	0.928^{***} (0.294)
factor(mfield)6	1.837^{***} (0.262)
factor(mfield)7	0.625^{*} (0.320)
factor(mfield)8	0.793^{***} (0.296)
factor(mfield)9	-1.861^{***} (0.400)
factor(mfield)10	1.559^{***} (0.300)
AGEP	
factor(mfield)2:AGEP	-0.072^{***} (0.011)
factor(mfield)3:AGEP	-0.037^{***} (0.010)
factor(mfield)4:AGEP	-0.019^{*} (0.011)
factor(mfield)5:AGEP	-0.025^{***} (0.009)
factor(mfield)6:AGEP	-0.044^{***} (0.008)
factor(mfield)7:AGEP	-0.003 (0.010)
factor(mfield)8:AGEP	-0.021^{**} (0.009)
factor(mfield)9:AGEP	0.017 (0.012)
factor(mfield)10:AGEP	-0.031^{***} (0.009)
Constant	-0.306^{***} (0.042)
Observations	70,079
Log Likelihood	$-44{,}088.990$
Akaike Inf. Crit.	88,227.980
Note:	*p<0.1; **p<0.05; ***p<0.01

3(a) Check the weights:

As given in lecture, we know that $\bar{y}_{counterf} = \frac{\sum_i w_i y_i}{\sum_i w_i}$. Thus we can calculate the counterfactual mean for females in each category of mfield:

```
#calculate the weighted avg fraction of females in each category of mfield
deg_eth_pred_wt <- deg_eth_pred %>%
  mutate(ycounterf = weights*female)

weight_fem_frac <- sum(deg_eth_pred_wt$ycounterf)/sum(deg_eth_pred_wt$weights)
weight_fem_frac
```

```
## [1] 0.3192937
```

```
#compare to the unweighted male and female fractions
unweight_fem_frac <- mean(wg_frac_all$genfraction.y)
unweight_fem_frac
```

```
## [1] 0.4700586
```

```
unweight_male_frac <- mean(wg_frac_all$genfraction.x)
unweight_male_frac
```

```
## [1] 0.5299414
```

3(b) Regress log weekly wage on female dummy

```
gap <- lm(logwage ~ female, data = deg_eth)

gap$coefficients #is not equal to the gap in 2(b)?
```

```
## (Intercept)      female
##   7.1959397  -0.1916514
```

3(c) Re-run the regression using weighted least squares with weights

```
gap_wt <- lm(logwage ~ female, data = deg_eth, weights = w_i)

gap_wt$coefficients
```

```
## (Intercept)      female
##   7.3025706  -0.2353854
```

3(d) Run unweighted regression

```
unweighted_reg <- lm(logwage ~ female + poly(AGEP, 2) + factor(eth) + factor(mfield), data = deg_eth)

unweighted_reg$coefficients
```

```
##      (Intercept)           female   poly(AGEP, 2)1   poly(AGEP, 2)2
##      7.041727099     -0.125704882     34.313834189     -7.116619756
##     factor(eth)2     factor(eth)3     factor(eth)4     factor(eth)5
##      0.118666868     -0.039014060      0.146352338      0.051591774
##   factor(mfield)2  factor(mfield)3  factor(mfield)4  factor(mfield)5
##      0.182386453     -0.246393402      0.246522483      0.007819645
##   factor(mfield)6  factor(mfield)7  factor(mfield)8  factor(mfield)9
##      0.071115629     -0.048448467     -0.086697067      0.164419256
## factor(mfield)10
##     -0.076658215
```

Part 2: Twins Dataset

1. Simple model

We estimate the model:

$$logwage = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp^2 + \gamma_1 D_{married_i} + \gamma_2 D_{female_i} + \mu_i \tag{2}$$

```
#load data
twins_raw <- read.csv("/Users/sofia/Box/Cal (sofiaguo@berkeley.edu)/2018-19/Spring 2019/Econ 142/PSETS/

#estimate simple model
simple_mod <- lm(lw ~ educ + exp + I(exp^2) + married + female, data = twins_raw)
```

2. Fit model separately for men and women

```
#get female twins
twins_fem <- twins_raw %>%
  filter(female ==1)

#get male twins
twins_male <- twins_raw %>%
  filter(female ==0)

#run the model separately
simple_mod_fem <- lm(lw ~ educ + exp + I(exp^2) + married + female, data = twins_fem)

simple_mod_male <- lm(lw ~ educ + exp + I(exp^2) + married + female, data = twins_male)
```

```
#display results
stargazer(simple_mod,
          simple_mod_fem,
          simple_mod_male,
          type = "latex", title = "Simple Model Regression Results by gender",
          header = F)
```

The separate regressions look different because they drop the female variable and the marriage variable drops its statistical significance for women but becomes even more significantly positive for men. In addition, education has a stronger positive effect on women than men, while experience does not change/differ between the two genders.

Table 3: Simple Model Regression Results by gender

	Dependent variable:		
		lw	
	(1)	(2)	(3)
educ	0.124^{***}	0.137^{***}	0.103^{***}
	(0.008)	(0.010)	(0.013)
exp	0.052^{***}	0.053^{***}	0.059^{***}
	(0.004)	(0.006)	(0.007)
I(expˆ2)	-0.001^{***}	-0.001^{***}	-0.001^{***}
	(0.0001)	(0.0001)	(0.0002)
married	0.069^{**}	-0.040	0.199^{***}
	(0.034)	(0.041)	(0.056)
female	-0.334^{***}		
	(0.031)		
Constant	0.346^{***}	-0.085	0.520^{**}
	(0.132)	(0.160)	(0.205)
Observations	1,074	574	500
R^2	0.315	0.303	0.255
Adjusted R^2	0.311	0.298	0.249
Residual Std. Error	0.506 (df = 1068)	0.443 (df = 569)	0.556 (df = 495)
F Statistic	98.021^{***} (df = 5; 1068)	61.915^{***} (df = 4; 569)	42.424^{***} (df = 4; 495)
Note:			*p<0.1; **p<0.05; ***p<0.01

3(a). Construct the mean family marriage rate for each person in the dataset

```
#add a mean marriage rate column with values 0, 0.5, or 1
twins_married <- twins_raw %>%
  mutate(meanmarriage = (married+omarried)/2)

#verify regressing marriage of twin on avg fract of married siblings -> coeff = 1

twins_married_reg <- lm(omarried ~ meanmarriage, data = twins_married)

twins_married_reg$coefficients #verified
```

```
##   (Intercept)  meanmarriage
## -4.498898e-15  1.000000e+00
```

3(b). Add mean fraction of siblings married to gender specific models

```
#separate genders
twins_married_fem <- twins_married %>%
  filter(female ==1)

twins_married_male <- twins_married %>%
  filter(female ==0)

#run the model separately
simple_mod_fem_mar <- lm(lw ~ educ + exp + I(exp^2) + married + meanmarriage + female, data = twins_marr

simple_mod_male_mar <- lm(lw ~ educ + exp + I(exp^2) + married + meanmarriage + female, data = twins_mar
```

```
#display results
stargazer(simple_mod_fem_mar,
          simple_mod_male_mar,
          type = "latex", title = "Simple Model Regression Results with average marriage",
          header = F)
```

Table 4: Simple Model Regression Results with average marriage

	Dependent variable:	
	lw	
	(1)	(2)
educ	0.137^{***}	0.103^{***}
	(0.010)	(0.013)
exp	0.053^{***}	0.057^{***}
	(0.006)	(0.007)
I(expˆ2)	-0.001^{***}	-0.001^{***}
	(0.0001)	(0.0001)
married	-0.022	0.030
	(0.071)	(0.098)
meanmarriage	-0.027	0.250^{**}
	(0.088)	(0.119)
female		
Constant	-0.083	0.510^{**}
	(0.160)	(0.204)
Observations	574	500
R^2	0.303	0.262
Adjusted R^2	0.297	0.254
Residual Std. Error	0.443 (df = 568)	0.554 (df = 494)
F Statistic	49.472^{***} (df = 5; 568)	35.054^{***} (df = 5; 494)

Note: *p<0.1; **p<0.05; ***p<0.01

The addition of *meanmarriage* makes the coefficient on *marriage* non-statistically significant for both genders. However, the coefficient on *meanmarriage* becomes larger and statistically significant for men only. An interpretation of this pattern is that male twins see a much larger boost to their weekly logwages from their twin having a spouse than women; this makes sense because taking the mean of a marriage indicator variable would probably boost the unmarried twin's wages simply because their twin has a higher dual income with their spouse, or because twins with married counterpart twins benefit from social implications such as higher wage expectations or abilities than two single twins. Thus on average, these types of married twins would have higher earnings than twins where neither is married.