Racial Differences in Earnings in the United States

Tom Vogl

This coding example estimates racial differences in earnings using data from the National Longitudinal Survey of Youth '79. Let's load the packages we need as well as the data.

```
rm(list=ls())
library(tidyverse)
library(fixest)

# load nlsy79.Rdata
load(url("https://github.com/tomvogl/econ121/raw/main/data/nlsy79.rds"))
```

To get started, Let's look at the structure of the dataset.

```
glimpse(nlsy79)
```

```
## Rows: 12,686
## Columns: 19
## $ caseid     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ perweight  <dbl> 563563, 763795, 536272, 565820, 764753, 636938, 674417, 148~
## $ age79      <dbl> 20, 20, 17, 16, 19, 18, 14, 20, 15, 18, 19, 19, 20, 15, 15,~
## $ region79   <fct> NORTHEAST, NORTHEAST, NORTHEAST, NORTHEAST, NORTHEAST, NORT~
## $ foreign    <dbl> 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ urban14    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ mag14      <dbl> 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1,~
## $ news14     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ lib14      <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ educ_mom   <dbl> 8, 5, 10, 11, 12, 12, 12, 9, 12, 12, 12, 15, 12, 12, 12, 12~
## $ educ_dad   <dbl> 8, 8, 12, 12, 12, 12, 12, 6, 10, 12, 12, 12, 16, 12, 12, 12~
## $ numsibs    <dbl> 1, 8, 3, 3, 1, 1, 1, 7, 4, 3, 1, 3, 2, 2, 1, 3, 2, 2, 3, 2,~
## $ afqt81     <dbl> NA, 12, 51, 62, 90, 99, 33, 43, 55, 27, 71, 94, 78, 88, 83,~
## $ laborinc18 <dbl> NA, 25000, 80000, 0, NA, 117000, NA, 51313, NA, NA, NA, NA,~
## $ hours18    <dbl> NA, 1820, 2244, 2765, NA, 2080, NA, 2600, NA, NA, NA, NA, N~
## $ educ       <dbl> 12, 12, 12, 14, 18, 16, 12, 14, 14, 9, 16, 16, 16, 19, 16, ~
## $ black      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hisp       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ male       <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0,~
```

What are the mean and SD of labor income?

```
mean(nlsy79$laborinc18,na.rm=TRUE)
```

```
## [1] 44887.57
```

```
sd(nlsy79$laborinc18,na.rm=TRUE)
```

```
## [1] 65078.64
```

How about percentiles?

```r
summary(nlsy79$laborinc18)
```

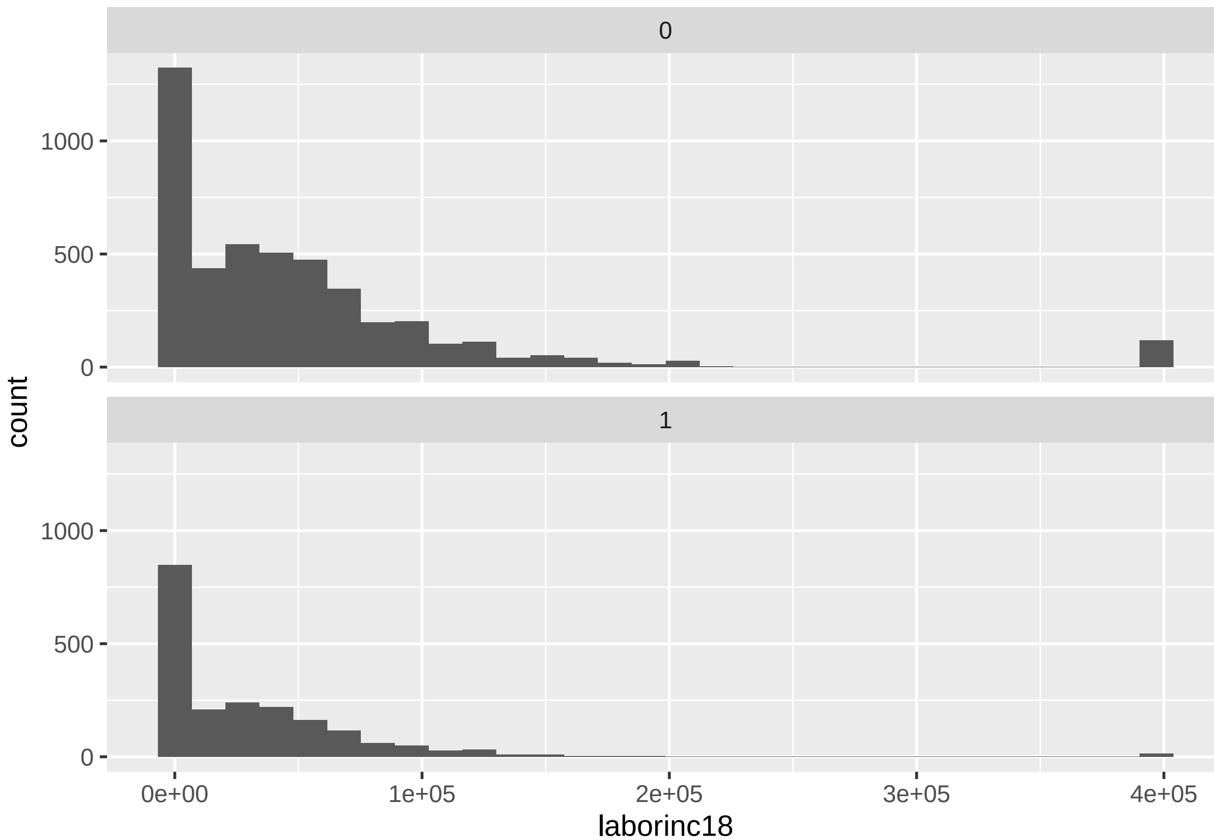
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       0   30000   44888   60000  396970    6115
```

We can see more detail when we plot histograms by race.

```r
nlsy79 %>%
  ggplot(aes(x = laborinc18)) +
    geom_histogram() +
    facet_wrap(~black, ncol=1) # separate graphs by race, stacked into one column
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 6115 rows containing non-finite values (`stat_bin()`).
```



We will estimate differences in mean income between blacks and non-blacks. Let's look at means by race.

```r
nlsy79 %>%
  drop_na(laborinc18) %>% # removes NA values so we don't need to use na.rm below
  group_by(black) %>%
  summarize(mean=mean(laborinc18),
            sd=sd(laborinc18),
            n=n())
```

```
## # A tibble: 2 x 4
##   black   mean     sd     n
##   <dbl>  <dbl>  <dbl> <int>
## 1     0 50798. 70856.  4558
## 2     1 31505. 46907.  2013
```

These results give us all the information we need to test for differences by race. The difference is:

```r
50798-31505
```

```
## [1] 19293
```

And the t-statistic is

```r
(50798-31505)/sqrt(70856^2/4558 + 46907^2/2013)
```

```
## [1] 13.02358
```

which is well above 1.96, so statistically significant by the usual standards.

An alternative way to run this test is the t-test with unequal variances:

```r
t.test(laborinc18 ~ black, data = nlsy79)
```

```
##
##  Welch Two Sample t-test
##
## data:  laborinc18 by black
## t = 13.023, df = 5599.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  16388.20 22196.42
## sample estimates:
## mean in group 0 mean in group 1
##        50797.69        31505.38
```

Equivalently, we can run a regression with heteroskedasticity-robust SEs, using feols() from fixest package

```r
feols(laborinc18 ~ black, data = nlsy79, vcov = 'hetero')
```

```
## NOTE: 6,115 observations removed because of NA values (LHS: 6,115).
```

```
## OLS estimation, Dep. Var.: laborinc18
## Observations: 6,571
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  50797.7    1049.56  48.3989 < 2.2e-16 ***
## black       -19292.3    1481.35 -13.0234 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 64,463.1   Adj. R2: 0.018528
```

Note that lm() is the base-R way to estimate a regression, but it doesn't directly allow for robust standard errors, and you need to use summary() to even see classical standard errors. feols() from fixest is more convenient.

```r
model1 <- lm(laborinc18 ~ black, data = nlsy79)
model1
```

```
##
## Call:
## lm(formula = laborinc18 ~ black, data = nlsy79)
##
## Coefficients:
## (Intercept)        black
##       50798       -19292
```

```r
summary(model1)
```

```
##
## Call:
## lm(formula = laborinc18 ~ black, data = nlsy79)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50798 -32798 -15798  15495 365465
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    50798        955   53.19   <2e-16 ***
## black         -19292       1725  -11.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64470 on 6569 degrees of freedom
##   (6115 observations deleted due to missingness)
## Multiple R-squared:  0.01868,    Adjusted R-squared:  0.01853
## F-statistic:   125 on 1 and 6569 DF,  p-value: < 2.2e-16
```

It is actually uncommon to test for average differences in the level (rather than log) of earnings, including zeros from the non-employed. It would be much more typical to restrict to employed individuals. So let's restrict to people restrict to people who worked for pay for at least 1000 hours: equivalent to a part-time job of 20 hours per week for 50 weeks.

```r
summary(nlsy79$hours18)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       0    2040    1513    2172    8736    5866
```

```r
nlsy79_workers <-
  nlsy79 %>%
  filter(hours18>=1000 & laborinc18>0)

summary(nlsy79_workers$hours18)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2080    2080    2245    2515    8736
```

Means by race in the workers sample:

```r
nlsy79_workers %>%
  drop_na(laborinc18) %>%
  group_by(black) %>%
  summarize(mean=mean(laborinc18),
            sd=sd(laborinc18),
            n=n())
```

```
## # A tibble: 2 x 4
##   black   mean     sd     n
##   <dbl>  <dbl>  <dbl> <int>
## 1     0 73786. 76195.  3015
## 2     1 55632. 52303.  1078
```

Still a $19k difference.

Now let's look at log earnings.

```r
nlsy79_workers <-
  nlsy79_workers %>%
    mutate(loginc18 = log(laborinc18))

nlsy79_workers %>%
  drop_na(loginc18) %>%
  group_by(black) %>%
  summarize(mean=format(mean(loginc18, na.rm = TRUE)),   # the format() function is just to report more
            sd=sd(loginc18, na.rm = TRUE),
            n=n())
```

```
## # A tibble: 2 x 4
##   black mean        sd     n
##   <dbl> <chr>    <dbl> <int>
## 1     0 10.85106 0.867  3015
## 2     1 10.61642 0.849  1078
```

The difference is:

```r
10.851-10.616
```

```
## [1] 0.235
```

This difference in logs can by roughly interpreted as a 23.5% gap in earnings, although this interpretation relies on calculus [dln(y)/dx]. Since we are doing a comparison by a discrete variable, we can think of 23.5% as an approximation .

The t-statistic is now:

```r
(10.851-10.616)/sqrt(.867^2/3015 + .849^2/1078)
```

```
## [1] 7.756312
```

Again well above 1.96, so statistically significant by the usual standards.

As an alterantive way to do the same thing, we can run a t-test with unequal variances:

```r
t.test(loginc18 ~ black, data = nlsy79_workers)
```

```
##
##  Welch Two Sample t-test
##
## data:  loginc18 by black
## t = 7.7464, df = 1935.7, p-value = 1.514e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.1752323 0.2940397
## sample estimates:
## mean in group 0 mean in group 1
##        10.85106        10.61642
```

Or run a regression with heteroskedasticity-robust standard errors:

```r
feols(loginc18 ~ black, data = nlsy79_workers, vcov = 'hetero')
```

```
## OLS estimation, Dep. Var.: loginc18
## Observations: 4,093
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error   t value   Pr(>|t|)
```

```
## (Intercept) 10.851056   0.015799 686.82440  < 2.2e-16 ***
## black       -0.234636   0.030285  -7.74749 1.1741e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.862284   Adj. R2: 0.013921
```

Same results. That is to say, a regression on a "dummy variable" for black leads to the same results as a difference of means Note that the t-statistic is very slightly different from what we computed "by hand." That's likely due to rounding errors.