Effect of Job Training Partnership Act

Eric Cheung

2024-03-05

```
library(readxl)
library(car)
```

```
## Loading required package: carData
```

```
library(ivreg)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::recode() masks car::recode()
## ✖ purrr::some()   masks car::some()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

1)

```
data <- read_xlsx("jtpa.xlsx")
```

```
data |>
  summarise(count = n(),
            treatment_group_mean = mean(offered_training),
            received_training_mean = mean(received_training))
```

```
## # A tibble: 1 × 3
##   count treatment_group_mean received_training_mean
##   <int>                <dbl>                  <dbl>
## 1  9872                0.671                  0.448
```

The two groups are not equal meaning that compliance with the treatment is imperfect. As people some people that were offered training didn't receive the training or some people that weren't offered training received training anyways.

2)

```
data <- data |>
  mutate(logincome = log(income))

glimpse(data |>
  group_by(received_training) |>
  summarise(mean_logincome = mean(logincome),
            mean_hsorged = mean(hsorged),
            mean_black = mean(black),
            mean_hispanic = mean(hispanic),
            mean_married = mean(married),
            mean_wkless13 = mean(wkless13),
            mean_afdc = mean(afdc),
            mean_age2225 = mean(age2225),
            mean_age2629 = mean(age2629),
            mean_age3035 = mean(age3035),
            mean_age3644 = mean(age3644),
            mean_age4554 = mean(age4554)))
```

```
## Rows: 2
## Columns: 13
## $ received_training <dbl> 0, 1
## $ mean_logincome    <dbl> 8.990993, 9.259150
## $ mean_hsorged      <dbl> 0.7024944, 0.7378261
## $ mean_black        <dbl> 0.2634478, 0.2549153
## $ mean_hispanic     <dbl> 0.0982192, 0.1125424
## $ mean_married      <dbl> 0.2709508, 0.2948685
## $ mean_wkless13     <dbl> 0.4373933, 0.4377163
## $ mean_afdc         <dbl> 0.1615568, 0.1873446
## $ mean_age2225      <dbl> 0.2390307, 0.2474576
## $ mean_age2629      <dbl> 0.2098403, 0.2063277
## $ mean_age3035      <dbl> 0.2362768, 0.2490395
## $ mean_age3644      <dbl> 0.1969892, 0.1859887
## $ mean_age4554      <dbl> 0.08298146, 0.07186441
```

The groups look quite similar except in some variables where they differ by a few percentage points. In baseline characteristics such as hsorged, hispanic, afdc, and age4554 there is a slight difference between the groups but we cannot be sure without testing.

3)

```
data |>
  group_by(received_training) |>
  summarise(mean_logincome = mean(logincome),
            sd = sd(logincome),
            count = n(),
            se = sqrt((1.423213^2)/4425 + (1.596272^2)/5447),
            t_test = (9.259150 - 8.990993)/se)
```

```
## # A tibble: 2 × 6
##   received_training mean_logincome    sd count     se t_test
##               <dbl>          <dbl> <dbl> <int>  <dbl>  <dbl>
## 1                 0           8.99  1.60  5447 0.0304   8.81
## 2                 1           9.26  1.42  4425 0.0304   8.81
```

reject the null hypothesis at 95% level of confidence

The average logincome of individuals who received training does not equal the average logincome of individuals who did not receive training. The difference in the means do not measure a causal effect effect of training on logincome as there is imperfect compliance with the treatment meaning that there could be selection problems which makes individuals in the received training different from those in the control group.

4)

```
# mean of male and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_male = mean(male),
            sd = sd(male),
            count = n(),
            se = sqrt((0.4984927^2)/6620 + (0.4991303^2)/3252),
            t_test = (0.4607251 - 0.4692497)/0.01068389)
```

```
## # A tibble: 2 × 6
##   offered_training mean_male    sd count     se t_test
##              <dbl>     <dbl> <dbl> <int>  <dbl>  <dbl>
## 1                0     0.469 0.499  3252 0.0107 -0.798
## 2                1     0.461 0.498  6620 0.0107 -0.798
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of hsorged and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_hsorged = mean(hsorged),
            sd = sd(hsorged),
            count = n(),
            se = sqrt((0.4332721^2)/6620 + (0.4398952^2)/3252),
            t_test = (0.7230403 - 0.7087456)/0.009373444)
```

```
## # A tibble: 2 × 6
##   offered_training mean_hsorged    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0        0.709 0.440  3252 0.00937   1.53
## 2                1        0.723 0.433  6620 0.00937   1.53
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of black and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_black = mean(black),
            sd = sd(black),
            count = n(),
            se = sqrt((0.4391457^2)/6620 + (0.4370870^2)/3252),
            t_test = (0.2608761 - 0.2570726)/0.009374337)
```

```
## # A tibble: 2 × 6
##   offered_training mean_black    sd count      se t_test
##              <dbl>      <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0      0.257 0.437  3252 0.00937  0.406
## 2                1      0.261 0.439  6620 0.00937  0.406
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of hispanic and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_hispanic = mean(hispanic),
            sd = sd(hispanic),
            count = n(),
            se = sqrt((0.3051895^2)/6620 + (0.3079983^2)/3252),
            t_test = (0.1039275 - 0.1060886)/0.00657573)
```

```
## # A tibble: 2 × 6
##   offered_training mean_hispanic    sd count      se t_test
##              <dbl>         <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0         0.106 0.308  3252 0.00658 -0.329
## 2                1         0.104 0.305  6620 0.00658 -0.329
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of married and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_married = mean(married),
            sd = sd(married),
            count = n(),
            se = sqrt((0.4394176^2)/6620 + (0.4293692^2)/3252),
            t_test = (0.2882695 - 0.2682405)/0.009265958)
```

```
## # A tibble: 2 × 6
##   offered_training mean_married    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0        0.268 0.429  3252 0.00927   2.16
## 2                1        0.288 0.439  6620 0.00927   2.16
```

reject the null hypothesis at 95% level of confidence

```
# mean of wkless13 and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_wkless13 = mean(wkless13),
            sd = sd(wkless13),
            count = n(),
            se = sqrt((0.4699491^2)/6620 + (0.4701505^2)/3252),
            t_test = (0.4415291 - 0.4294137)/0.01006639)
```

```
## # A tibble: 2 × 6
##   offered_training mean_wkless13    sd count     se t_test
##              <dbl>         <dbl> <dbl> <int>  <dbl>  <dbl>
## 1                0         0.429 0.470  3252 0.0101   1.20
## 2                1         0.442 0.470  6620 0.0101   1.20
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of afdc and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_afdc = mean(afdc),
            sd = sd(afdc),
            count = n(),
            se = sqrt((0.3781112^2)/6620 + (0.3789434^2)/3252),
            t_test = (0.1728097 - 0.1737392)/0.008108837)
```

```
## # A tibble: 2 × 6
##   offered_training mean_afdc    sd count      se t_test
##              <dbl>     <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0     0.174 0.379  3252 0.00811 -0.115
## 2                1     0.173 0.378  6620 0.00811 -0.115
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of age2225  and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_age2225 = mean(age2225),
            sd = sd(age2225),
            count = n(),
            se = sqrt((0.4278671^2)/6620 + (0.4307462^2)/3252),
            t_test = (0.2412387 - 0.2460025)/0.009203746)
```

```
## # A tibble: 2 × 6
##   offered_training mean_age2225    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0        0.246 0.431  3252 0.00920 -0.518
## 2                1        0.241 0.428  6620 0.00920 -0.518
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of age2629 and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_age2629 = mean(age2629),
            sd = sd(age2629),
            count = n(),
            se = sqrt((0.4106121^2)/6620 + (0.3964651^2)/3252),
            t_test = (0.2146526 - 0.1952645)/0.008590888)
```

```
## # A tibble: 2 × 6
##   offered_training mean_age2629    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0        0.195 0.396  3252 0.00859   2.26
## 2                1        0.215 0.411  6620 0.00859   2.26
```

reject the null hypothesis at 95% level of confidence

```
# mean of age3035 and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_age3035 = mean(age3035),
            sd = sd(age3035),
            count = n(),
            se = sqrt((0.4257462^2)/6620 + (0.4334338^2)/3252),
            t_test = (0.2377644 - 0.2506150)/0.009227657)
```

```
## # A tibble: 2 × 6
##   offered_training mean_age3035    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0        0.251 0.433  3252 0.00923  -1.39
## 2                1        0.238 0.426  6620 0.00923  -1.39
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of age3644 and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_age3644 = mean(age3644),
            sd = sd(age3644),
            count = n(),
            se = sqrt((0.3943699^2)/6620 + (0.3931176^2)/3252),
            t_test = (0.1925982 - 0.1909594)/0.008427074)
```

```
## # A tibble: 2 × 6
##   offered_training mean_age3644    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0        0.191 0.393  3252 0.00843  0.194
## 2                1        0.193 0.394  6620 0.00843  0.194
```

do not reject the null hypothesis at 95% level of confidence

```
# mean of age4554 and test for if treatment means equals control means
data |>
  group_by(offered_training) |>
  summarise(mean_age4554 = mean(age4554),
            sd = sd(age4554),
            count = n(),
            se = sqrt((0.2664341^2)/6620 + (0.2717345^2)/3252),
            t_test = (0.07688822 - 0.08025830)/0.005781786)
```

```
## # A tibble: 2 × 6
##   offered_training mean_age4554    sd count      se t_test
##              <dbl>        <dbl> <dbl> <int>   <dbl>  <dbl>
## 1                0       0.0803 0.272  3252 0.00578 -0.583
## 2                1       0.0769 0.266  6620 0.00578 -0.583
```

do not reject the null hypothesis at 95% level of confidence

The means of the married and age2629 variables were not the same across the treatment and control group. Which could indicate that the assignment to treatment was not actually random.

5)

```
data |>
  group_by(male, offered_training) |>
  summarise(mean_logincome = mean(logincome),
            sd = sd(logincome),
            count = n()) |>
  filter(male == 0) |>
  mutate(se = sqrt((sd^2)/count + (sd[1]^2)/count[1]),
         t_test = (8.988229 - 8.869295)/se)
```

```
## `summarise()` has grouped output by 'male'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2 × 7
## # Groups:   male [1]
##    male offered_training mean_logincome    sd count     se t_test
##   <dbl>            <dbl>          <dbl> <dbl> <int>  <dbl>  <dbl>
## 1     0                0           8.87  1.53  1726 0.0522   2.28
## 2     0                1           8.99  1.51  3570 0.0448   2.66
```

reject the null hypothesis at 95% level of confidence

```
data |>
  group_by(male, offered_training) |>
  summarise(mean_logincome = mean(logincome),
            sd = sd(logincome),
            count = n()) |>
  filter(male == 1) |>
  mutate(se = sqrt((sd^2)/count + (sd[1]^2)/count[1]),
         t_test = (9.308423 - 9.278250)/se)
```

```
## `summarise()` has grouped output by 'male'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2 × 7
## # Groups:   male [1]
##    male offered_training mean_logincome    sd count     se t_test
##   <dbl>            <dbl>          <dbl> <dbl> <int>  <dbl>  <dbl>
## 1     1                0           9.28  1.46  1526 0.0530  0.569
## 2     1                1           9.31  1.54  3050 0.0467  0.646
```

do not reject the null hypothesis at 95% level of confidence

I do not think that Y0 - YN measures the causal effect of training on logincome because of the imperfect compliance of randomization which could cause selection problems. However, Y0 - YN is the intention to treat which measures the causal effect of being assigned to the treatment group on logincome. But because there is imperfect compliance with the randomization process this does not measure the causal effect of training on logincome.

6)

```
data |>
  group_by(male, offered_training) |>
  summarise(mean_received_training = mean(received_training),
            sd = sd(received_training),
            count = n()) |>
  filter(male == 0) |>
  mutate(se = sqrt((sd^2)/count + (sd[1]^2)/count[1]),
         t_test = (0.67507003 - 0.01738123)/se)
```

```
## `summarise()` has grouped output by 'male'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2 × 7
## # Groups:   male [1]
##    male offered_training mean_received_training    sd count      se t_test
##   <dbl>            <dbl>                  <dbl> <dbl> <int>   <dbl>  <dbl>
## 1     0                0                 0.0174 0.131  1726 0.00445  148.
## 2     0                1                 0.675  0.468  3570 0.00845   77.9
```

```
data |>
  group_by(male, offered_training) |>
  summarise(mean_received_training = mean(received_training),
            sd = sd(received_training),
            count = n()) |>
  filter(male == 1) |>
  mutate(se = sqrt((sd^2)/count + (sd[1]^2)/count[1]),
         t_test = (0.64491803 - 0.01179554)/se)
```

```
## `summarise()` has grouped output by 'male'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2 × 7
## # Groups:   male [1]
##    male offered_training mean_received_training    sd count      se t_test
##   <dbl>            <dbl>                  <dbl> <dbl> <int>   <dbl>  <dbl>
## 1     1                0                 0.0118 0.108  1526 0.00391  162.
## 2     1                1                 0.645  0.479  3050 0.00910   69.6
```

R0 - RN measures the difference for people in that got offered training and received training and people that did not get offered training but still received training. R0 does not equal RN. This is important because we are testing the effect of offering training on income so it is important that only people that got offered training received training while those who were not offered training did not get it. And in order for the IV estimate to have causal effect it must meet the first stage assumption where the instrument does have an effect on the treatment variable. R0 - RN is also called the first stage.

7)

```
female_iv_estimate = (8.988229 - 8.869295)/(0.67507003 - 0.01738123)
female_iv_estimate
```

```
## [1] 0.1808363
```

```
male_iv_estimate = (9.308423 - 9.278250)/(0.64491803 - 0.01179554)
male_iv_estimate
```

```
## [1] 0.04765744
```

IV estimates requires the first stage, independence, exclusion, and monotonicity assumptions. The first stage assumption is met as the instrument has a causal effect on treatment. The independence assumption could be met as the instrument is not related to ommited variables but it might not be met as there is evidence to suggest that offered training was not fully randomly assigned. The exclusion assumption is met as being offered training does not affect income but one needs to actually receive training to affect income. The monotonicity assumption is met as everyone that was offered training had a higher change at receiving training. This IV estimate is likely not the same as ATT because IV estimate is a local average treatment effect. It is the causal effect of receiving training for those who received training after being offered training and those who do not receive training after not being offered training.

8)

```
female <- data |>
  filter(male == 0)

summary(lm(logincome ~ received_training + hsorged + black + hispanic + married
           + wkless13 + afdc + age2225 + age2629 + age3035 + age3644 + age4554,
           data = female))
```

```
##
## Call:
## lm(formula = logincome ~ received_training + hsorged + black +
##     hispanic + married + wkless13 + afdc + age2225 + age2629 +
##     age3035 + age3644 + age4554, data = female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1940 -0.6367  0.3905  1.0206  2.9521
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.79337    0.10987  80.034  < 2e-16 ***
## received_training  0.23347    0.04073   5.732 1.05e-08 ***
## hsorged            0.24636    0.04768   5.167 2.47e-07 ***
## black             -0.02166    0.04959  -0.437  0.66224
## hispanic          -0.09000    0.06675  -1.348  0.17761
## married           -0.03457    0.05351  -0.646  0.51821
## wkless13          -0.55440    0.04533 -12.229  < 2e-16 ***
## afdc              -0.34393    0.04994  -6.887 6.35e-12 ***
## age2225            0.24826    0.10862   2.286  0.02231 *
## age2629            0.25432    0.10909   2.331  0.01977 *
## age3035            0.28645    0.10773   2.659  0.00786 **
## age3644            0.29997    0.10944   2.741  0.00614 **
## age4554            0.23873    0.12130   1.968  0.04911 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.472 on 5283 degrees of freedom
## Multiple R-squared:  0.06469,    Adjusted R-squared:  0.06257
## F-statistic: 30.45 on 12 and 5283 DF,  p-value: < 2.2e-16
```

```
male <- data |>
  filter(male == 1)

summary(lm(logincome ~ received_training + hsorged + black + hispanic + married
           + wkless13 + afdc + age2225 + age2629 + age3035 + age3644 + age4554,
           data = male))
```

```
##
## Call:
## lm(formula = logincome ~ received_training + hsorged + black +
##     hispanic + married + wkless13 + afdc + age2225 + age2629 +
##     age3035 + age3644 + age4554, data = male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0554 -0.6529  0.3702  0.9947  2.7841
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.89672    0.13644  65.208  < 2e-16 ***
## received_training  0.30666    0.04364   7.027 2.42e-12 ***
## hsorged            0.25600    0.04921   5.203 2.05e-07 ***
## black             -0.16823    0.05205  -3.232  0.00124 **
## hispanic           0.03256    0.07464   0.436  0.66267
## married            0.42768    0.04867   8.788  < 2e-16 ***
## wkless13          -0.49455    0.04790 -10.324  < 2e-16 ***
## afdc              -0.15484    0.10501  -1.474  0.14042
## age2225            0.22771    0.13376   1.702  0.08875 .
## age2629            0.26677    0.13527   1.972  0.04866 *
## age3035            0.20358    0.13366   1.523  0.12779
## age3644            0.04372    0.13495   0.324  0.74595
## age4554            0.02751    0.14819   0.186  0.85275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 4563 degrees of freedom
## Multiple R-squared:  0.07231,    Adjusted R-squared:  0.06987
## F-statistic: 29.64 on 12 and 4563 DF,  p-value: < 2.2e-16
```

These OLS estimates do not measure the causal effect of training on logincome because there is imperfect compliance with the randomization of treatment and received_training could be related to the error term which means it could be exogenous.

9)

```
# first stage for females
fsls_females <- lm(received_training ~ offered_training + hsorged + black + hispanic +
             married + wkless13 + afdc + age2225 + age2629 + age3035 + age3644 +
             age4554, data = female)
summary(fsls_females)
```

```
##
## Call:
## lm(formula = received_training ~ offered_training + hsorged +
##     black + hispanic + married + wkless13 + afdc + age2225 +
##     age2629 + age3035 + age3644 + age4554, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79177 -0.07604  0.03360  0.31753  1.03705
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.012469   0.029634   0.421  0.67394
## offered_training  0.658856   0.011439  57.596  < 2e-16 ***
## hsorged           0.040819   0.012612   3.237  0.00122 **
## black            -0.040601   0.013126  -3.093  0.00199 **
## hispanic          0.037956   0.017671   2.148  0.03176 *
## married           0.022748   0.014163   1.606  0.10830
## wkless13         -0.024301   0.012002  -2.025  0.04294 *
## afdc              0.067552   0.013206   5.115 3.24e-07 ***
## age2225          -0.029675   0.028755  -1.032  0.30212
## age2629          -0.043616   0.028881  -1.510  0.13105
## age3035          -0.001583   0.028522  -0.056  0.95573
## age3644          -0.049734   0.028970  -1.717  0.08608 .
## age4554          -0.068817   0.032106  -2.143  0.03213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3897 on 5283 degrees of freedom
## Multiple R-squared:  0.3904, Adjusted R-squared:  0.389
## F-statistic: 281.9 on 12 and 5283 DF,  p-value: < 2.2e-16
```

```
linearHypothesis(fsls_females, "offered_training = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## offered_training = 0
##
## Model 1: restricted model
## Model 2: received_training ~ offered_training + hsorged + black + hispanic +
##     married + wkless13 + afdc + age2225 + age2629 + age3035 +
##     age3644 + age4554
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   5284 1305.89
## 2   5283  802.19  1     503.7 3317.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# first stage for males
fsls_males <- lm(received_training ~ offered_training + hsorged + black + hispanic +
             married + wkless13 + afdc + age2225 + age2629 + age3035 + age3644 +
             age4554, data = male)
summary(fsls_males)
```

```
##
## Call:
## lm(formula = received_training ~ offered_training + hsorged +
##     black + hispanic + married + wkless13 + afdc + age2225 +
##     age2629 + age3035 + age3644 + age4554, data = male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75777 -0.06301  0.01567  0.34596  1.05343
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.022199   0.037400   0.594   0.5528
## offered_training  0.632765   0.012399  51.034   <2e-16 ***
## hsorged           0.033095   0.013311   2.486   0.0129 *
## black             0.019472   0.014086   1.382   0.1669
## hispanic          0.040477   0.020194   2.004   0.0451 *
## married           0.029233   0.013165   2.221   0.0264 *
## wkless13         -0.008318   0.012968  -0.641   0.5213
## afdc              0.005221   0.028423   0.184   0.8543
## age2225          -0.026740   0.036204  -0.739   0.4602
## age2629          -0.075632   0.036613  -2.066   0.0389 *
## age3035          -0.045175   0.036174  -1.249   0.2118
## age3644          -0.062646   0.036523  -1.715   0.0864 .
## age4554          -0.058513   0.040103  -1.459   0.1446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3951 on 4563 degrees of freedom
## Multiple R-squared:  0.3664, Adjusted R-squared:  0.3647
## F-statistic: 219.9 on 12 and 4563 DF,  p-value: < 2.2e-16
```

```
linearHypothesis(fsls_males, "offered_training = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## offered_training = 0
##
## Model 1: restricted model
## Model 2: received_training ~ offered_training + hsorged + black + hispanic +
##     married + wkless13 + afdc + age2225 + age2629 + age3035 +
##     age3644 + age4554
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   4564 1118.63
## 2   4563  712.16  1    406.48 2604.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The received training variable is not a weak instrument as the F statistic calculated for both genders are well above the 100 needed for a strong instrument.

10)

```
d_hat_females <- fitted.values(fsls_females)
tsls_female <- lm(logincome ~ d_hat_females + hsorged + black + hispanic + married
           + wkless13 + afdc + age2225 + age2629 + age3035 + age3644 + age4554,
           data = female)
summary(tsls_female)
```

```
##
## Call:
## lm(formula = logincome ~ d_hat_females + hsorged + black + hispanic +
##     married + wkless13 + afdc + age2225 + age2629 + age3035 +
##     age3644 + age4554, data = female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1194 -0.6367  0.3970  1.0215  2.7956
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.81920    0.11241  78.458  < 2e-16 ***
## d_hat_females  0.17422    0.06574   2.650  0.00807 **
## hsorged        0.24999    0.04790   5.219 1.87e-07 ***
## black         -0.02352    0.04974  -0.473  0.63632
## hispanic      -0.08845    0.06692  -1.322  0.18633
## married       -0.03216    0.05368  -0.599  0.54910
## wkless13      -0.55538    0.04545 -12.219  < 2e-16 ***
## afdc          -0.34036    0.05015  -6.786 1.28e-11 ***
## age2225        0.24627    0.10890   2.261  0.02377 *
## age2629        0.25253    0.10936   2.309  0.02097 *
## age3035        0.28556    0.10799   2.644  0.00821 **
## age3644        0.29703    0.10973   2.707  0.00682 **
## age4554        0.23474    0.12164   1.930  0.05369 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.475 on 5283 degrees of freedom
## Multiple R-squared:  0.06012,    Adjusted R-squared:  0.05799
## F-statistic: 28.16 on 12 and 5283 DF,  p-value: < 2.2e-16
```

```
d_hat_males <- fitted.values(fsls_males)
tsls_male <- lm(logincome ~ d_hat_males + hsorged + black + hispanic + married
           + wkless13 + afdc + age2225 + age2629 + age3035 + age3644 + age4554,
           data = male)
summary(tsls_male)
```

```
##
## Call:
## lm(formula = logincome ~ d_hat_males + hsorged + black + hispanic +
##     married + wkless13 + afdc + age2225 + age2629 + age3035 +
##     age3644 + age4554, data = male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2004 -0.6492  0.3853  1.0003  2.7952
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.00877    0.13927  64.688  < 2e-16 ***
## d_hat_males  0.03644    0.07279   0.501  0.61665
## hsorged      0.26566    0.04951   5.366 8.47e-08 ***
## black       -0.16304    0.05234  -3.115  0.00185 **
## hispanic     0.04596    0.07509   0.612  0.54058
## married      0.44060    0.04901   8.991  < 2e-16 ***
## wkless13    -0.49345    0.04816 -10.246  < 2e-16 ***
## afdc        -0.15164    0.10558  -1.436  0.15100
## age2225      0.22478    0.13448   1.671  0.09470 .
## age2629      0.25375    0.13603   1.865  0.06219 .
## age3035      0.19451    0.13439   1.447  0.14787
## age3644      0.03068    0.13571   0.226  0.82115
## age4554      0.01050    0.14903   0.070  0.94382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.467 on 4563 degrees of freedom
## Multiple R-squared:  0.06232,    Adjusted R-squared:  0.05986
## F-statistic: 25.27 on 12 and 4563 DF,  p-value: < 2.2e-16
```

The estimated coefficient on the first stage regression measure the effect of the instrument on the treatment. It is different from the received_training variable because this new coefficient is a function of offered_training and the baseline characteristics which are all exogenous which would make the new coefficient exogenous as well.

11)

For 2SLS to have a causal effect my instrument must be correlated with the treatment and it must be exogenous meaning it must be uncorrelated with the error term. We can test these assumptions via the causal chain that offered_training -> received_training -> logincome but offered_training has no effect on logincome which is only possible if those 2 above assumptions are met. I believe that offered_training can only affect logincome through received_training therefore the assumptions are met.

12)

```
# "proper" ivreg for females
ivreg_female <- ivreg(logincome ~ received_training + hsorged + black + hispanic
                      + married  + wkless13 + afdc + age2225 + age2629 + age3035
                      + age3644 + age4554 | offered_training + hsorged + black
                      + hispanic + married + wkless13 + afdc + age2225 + age2629
                      + age3035 + age3644 + age4554, data = female)
summary(ivreg_female)
```

```
##
## Call:
## ivreg(formula = logincome ~ received_training + hsorged + black +
##     hispanic + married + wkless13 + afdc + age2225 + age2629 +
##     age3035 + age3644 + age4554 | offered_training + hsorged +
##     black + hispanic + married + wkless13 + afdc + age2225 +
##     age2629 + age3035 + age3644 + age4554, data = female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1651 -0.6435  0.3947  1.0170  2.9220
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.81920    0.11216  78.633  < 2e-16 ***
## received_training  0.17422    0.06559   2.656  0.00793 **
## hsorged            0.24999    0.04779   5.230 1.76e-07 ***
## black             -0.02352    0.04963  -0.474  0.63556
## hispanic          -0.08845    0.06677  -1.325  0.18534
## married           -0.03216    0.05356  -0.600  0.54821
## wkless13          -0.55538    0.04535 -12.246  < 2e-16 ***
## afdc              -0.34036    0.05004  -6.801 1.15e-11 ***
## age2225            0.24627    0.10865   2.267  0.02346 *
## age2629            0.25253    0.10912   2.314  0.02069 *
## age3035            0.28556    0.10775   2.650  0.00807 **
## age3644            0.29703    0.10949   2.713  0.00669 **
## age4554            0.23474    0.12137   1.934  0.05316 .
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 5283  3317.262  <2e-16 ***
## Wu-Hausman          1 5282     1.329   0.249
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.472 on 5283 degrees of freedom
## Multiple R-Squared: 0.06432, Adjusted R-squared: 0.06219
## Wald test: 28.29 on 12 and 5283 DF,  p-value: < 2.2e-16
```

```
# "proper" ivreg for males
ivreg_male <- ivreg(logincome ~ received_training + hsorged + black + hispanic
                    + married + wkless13 + afdc + age2225 + age2629 + age3035
                    + age3644 + age4554 | offered_training + hsorged + black
                    + hispanic + married + wkless13 + afdc + age2225 + age2629
                    + age3035 + age3644 + age4554, data = male)
summary(ivreg_male)
```

```
##
## Call:
## ivreg(formula = logincome ~ received_training + hsorged + black +
##     hispanic + married + wkless13 + afdc + age2225 + age2629 +
##     age3035 + age3644 + age4554 | offered_training + hsorged +
##     black + hispanic + married + wkless13 + afdc + age2225 +
##     age2629 + age3035 + age3644 + age4554, data = male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1770 -0.6508  0.3841  1.0005  2.7824
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.00877    0.13910  64.763  < 2e-16 ***
## received_training  0.03644    0.07270   0.501  0.61623
## hsorged            0.26566    0.04946   5.372 8.18e-08 ***
## black             -0.16304    0.05228  -3.119  0.00183 **
## hispanic           0.04596    0.07501   0.613  0.54011
## married            0.44060    0.04895   9.001  < 2e-16 ***
## wkless13          -0.49345    0.04811 -10.258  < 2e-16 ***
## afdc              -0.15164    0.10546  -1.438  0.15052
## age2225            0.22478    0.13432   1.673  0.09432 .
## age2629            0.25375    0.13587   1.868  0.06189 .
## age3035            0.19451    0.13423   1.449  0.14740
## age3644            0.03068    0.13555   0.226  0.82094
## age4554            0.01050    0.14886   0.071  0.94376
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    1 4563   2604.42  < 2e-16 ***
## Wu-Hausman          1 4562     21.98 2.83e-06 ***
## Sargan              0   NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 4563 degrees of freedom
## Multiple R-Squared: 0.06452, Adjusted R-squared: 0.06206
## Wald test: 25.33 on 12 and 4563 DF,  p-value: < 2.2e-16
```

My estimates are the same as the ones i obtained from question 10 because they are the same regression but the question 10 regression was done manually while the regression in this question was done with the ivreg function.

13)

Training has a significant effect on logincome for females as the t-statistic is higher than the 1.96 critical value for 95% level of confidence. While training has no significant effect on logincome for males as the t-statistic is lower than the 1.96 critical value for 95% level of confidence.