IV PART 3

Doron Zamir

5/9/2021

Getting things ready

Load Packages

```
if (!require("pacman")) install.packages("pacman")

pacman::p_load(
  tidyverse,
  vip,
  here,
  readxl,
  DataExplorer,
  GGally,
  np,
  ivtools
)
```

Load Data

Using data from Card (1993) that can be found here (https://davidcard.berkeley.edu/data_sets.html)

```
schooling_raw <- read.table("Data/nls.dat") %>%
  as_data_frame()
```

Change Variables Names

Tidy up the data

Selecting variables to work with, remove missing data, and create a new dummy for collage proximity (regardless if it’s a 4 year or 2 year collage) for later analysis

```
schooling <- schooling_raw %>%
  select(
    ed76,      # Education - The treatment (t_i)
    nearc4,    # 4 year collage proximity - The IV (u_i)
    lwage78,   # log wage in 78 - The output (y_i)
    black,     # Dummy for black - control (w_i)
    age76      # Age at 76       - control (w_i)
    ) %>%
  filter(lwage78 != ".") %>%                      # remove missing values
  mutate_at(vars(lwage78),funs(as.numeric)) %>%    #make lwage numeric
  mutate_at(vars(black,nearc4),funs(as.factor)) #set dummys as factor
```

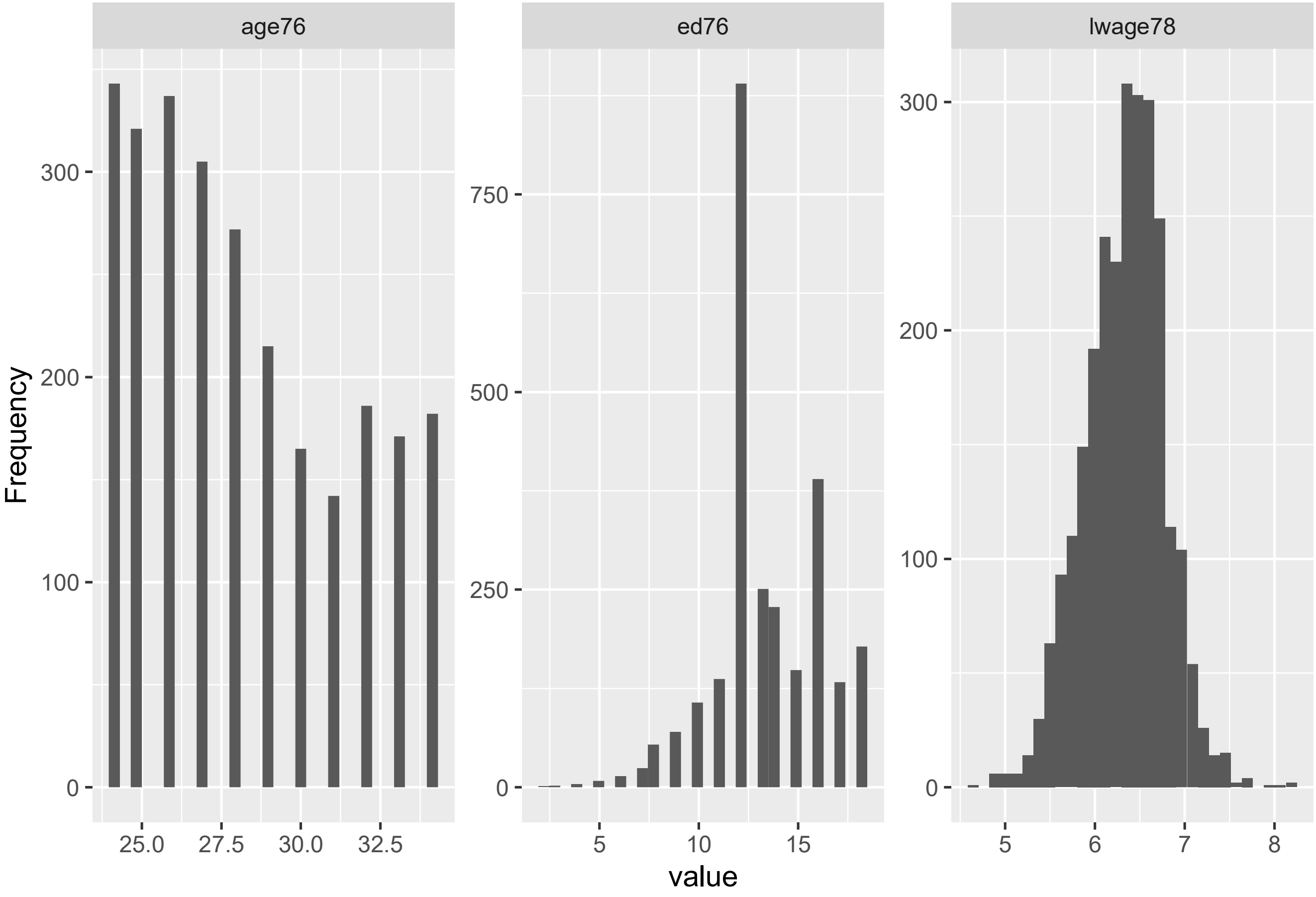
Exploratory Data Analysis

using `DataExplorar` and `GGally` packages

Histograms

looking at histograms of `lwage78, ed76`

```
plot_histogram(schooling)
```



It seems that there are a lot of observations with education less than 10, which might add noise to our analysis.

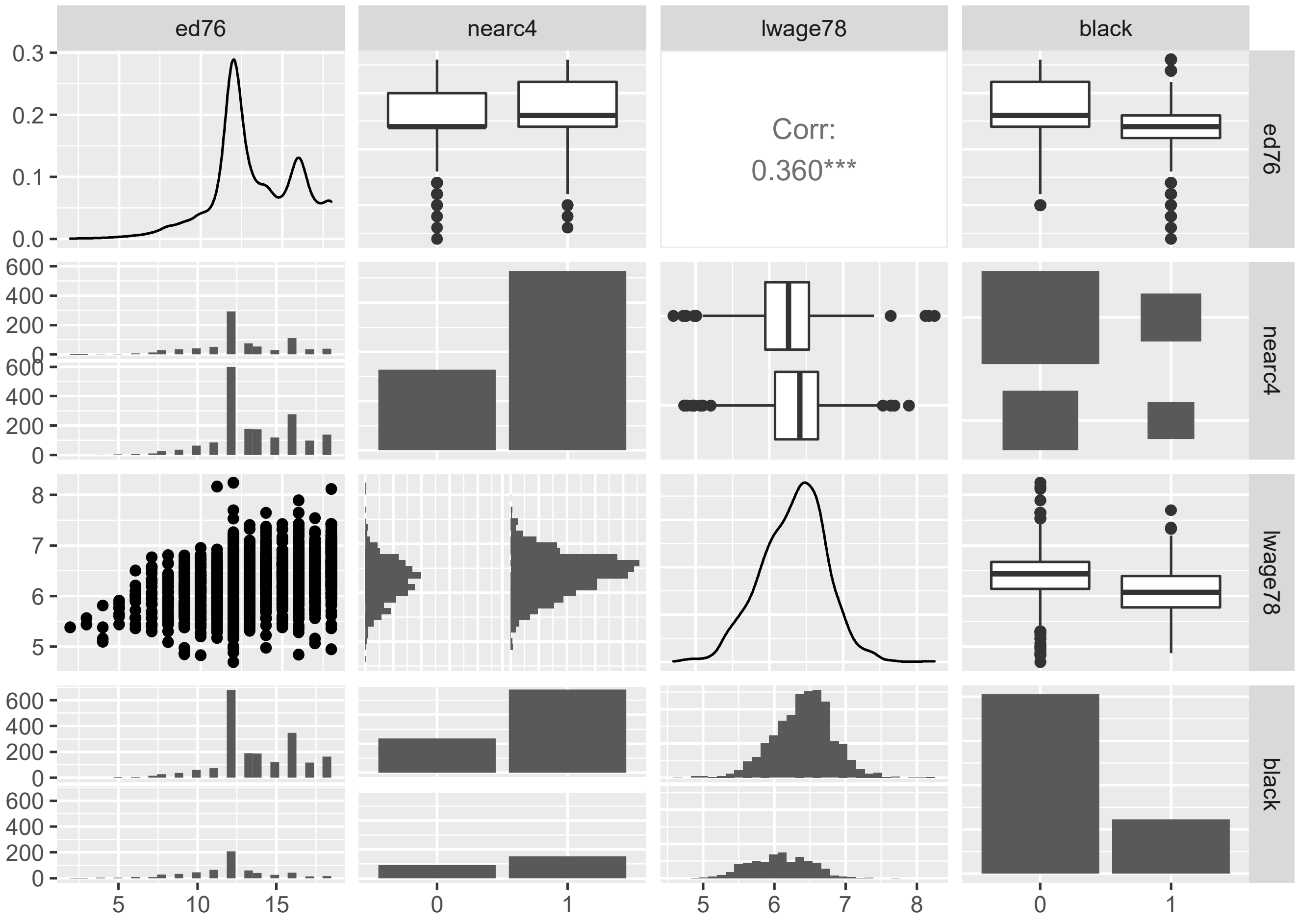
we might need to consider that later.

Correlations:

I used `ggparis()` from `GGally` package to plot every pair of variables.

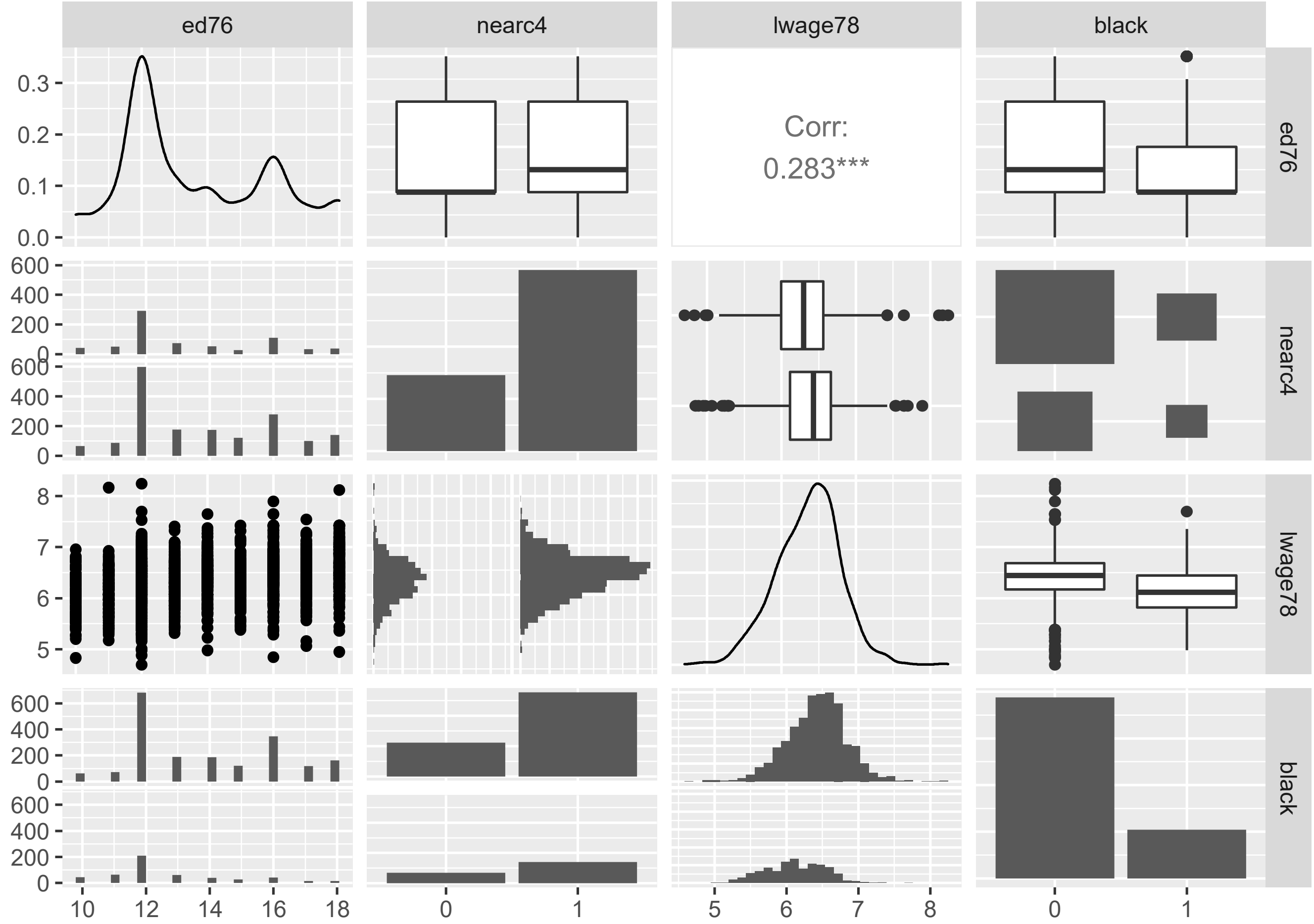
This is for the entire sample:

```
schooling %>%
  select(-c(age76)) %>%
  ggpairs()
```



Since we seen that `ed76` has a long tail from the left, let's see how's the correlation differ when trimming it

```
schooling %>%
  filter(ed76 >=10 ) %>%
  select(-c(age76)) %>%
  ggpairs()
```



Looking at the differences between the two plots, it seems that when trimming for education > 10, we get a stronger correlation between the instrument and the treatment. Then, trimming the data will lead to better results:

```
schooling <- schooling %>%
  filter(ed76 >= 10)
```

Non Parametric Estimation

General - boundaries for the output

Set K_0, K_1 s.t $\forall i, y_i(t) \in [K_0, K_1]$

```
k_0 <- min(schooling$lwage78)
k_1 <- max(schooling$lwage78)
```

This will be useful for later calculations.

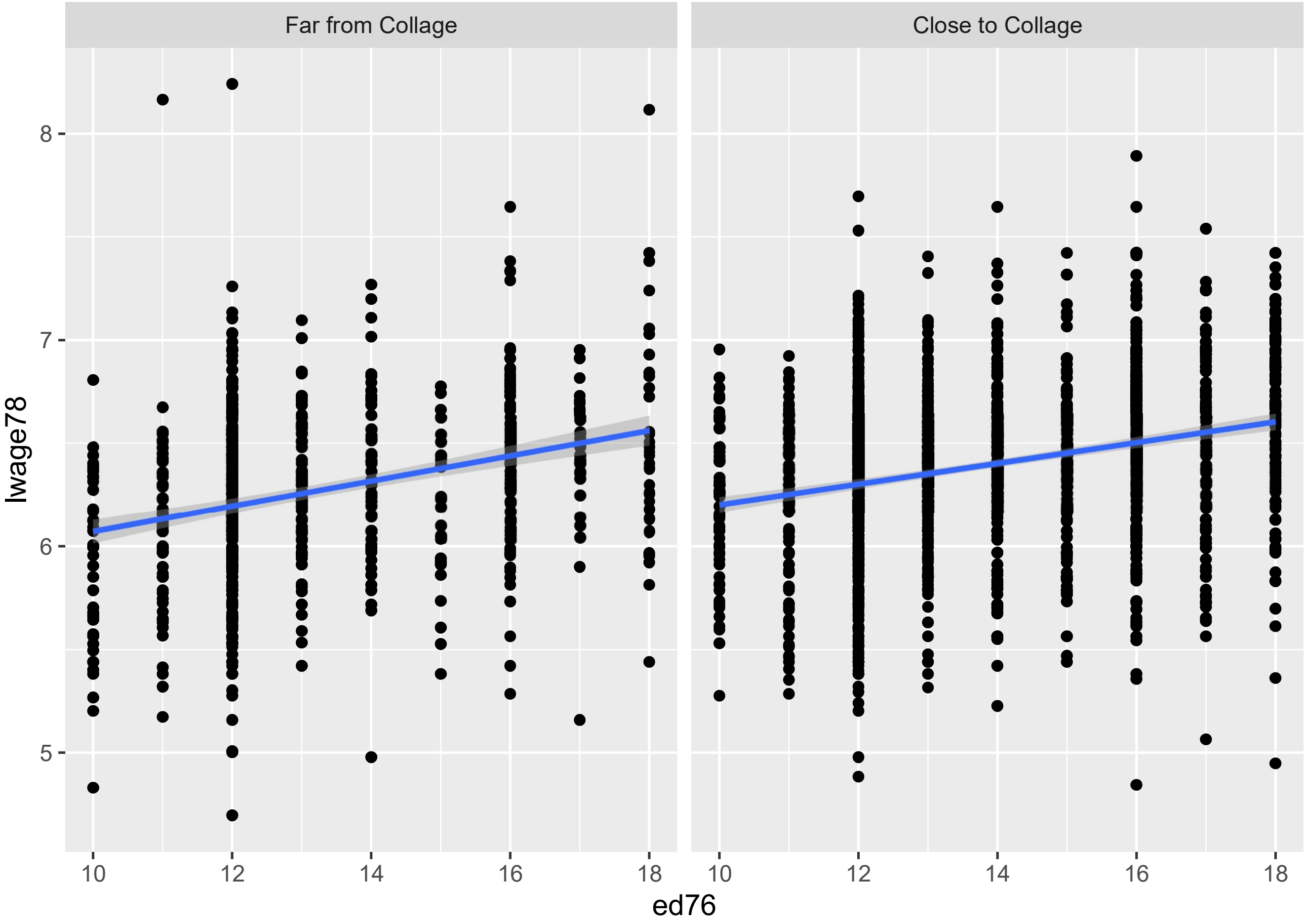
IV assumption

First, we assume that the connection between education and wage is the same, regardless of collage proximity. More specially, we want to check if

$$E(lwage78|ed76, nearc4 = 1) = E(lwage78|ed76, nearc4 = 0)$$

let’s check if this stands in the sample:

```
collage.labs <- c("Far from Collage", "Close to Collage")
names(collage.labs) <- c(0,1)

schooling %>%
  ggplot(aes(ed76,lwage78)) +
    geom_point() +
    geom_smooth(formula = y ~ x) +
    facet_wrap("nearc4", labeller =labeller(nearc4 = collage.labs))
```



seems like the assumption holds: the trend line for (`lwage78 ~ ed76`) looks the same in both plots, i.e. return for schooling doesn’t seem to differ between those who grew up close to 4 years collage and those who didn’t.

Calculating bounds

I calculated bounds related to the IV assumption. That is, for every treatment level $t \in T$, I calculate bounds $UB(t), LB(t)$ such that:

$$LB(t) \equiv \max_{u \in V} E[y|z = t, v = u] \cdot P(z = t|v = u) + K_0 \cdot P(z \neq t|v = u)$$
$$UB(t) \equiv \min_{u \in V} E[y|z = t, v = u] \cdot P(z = t|v = u) + K_1 \cdot P(z \neq t|v = u)$$

note that in this section, the treatment is the years of education, i.e., $t \in T = \{10, 11, 12...\}$

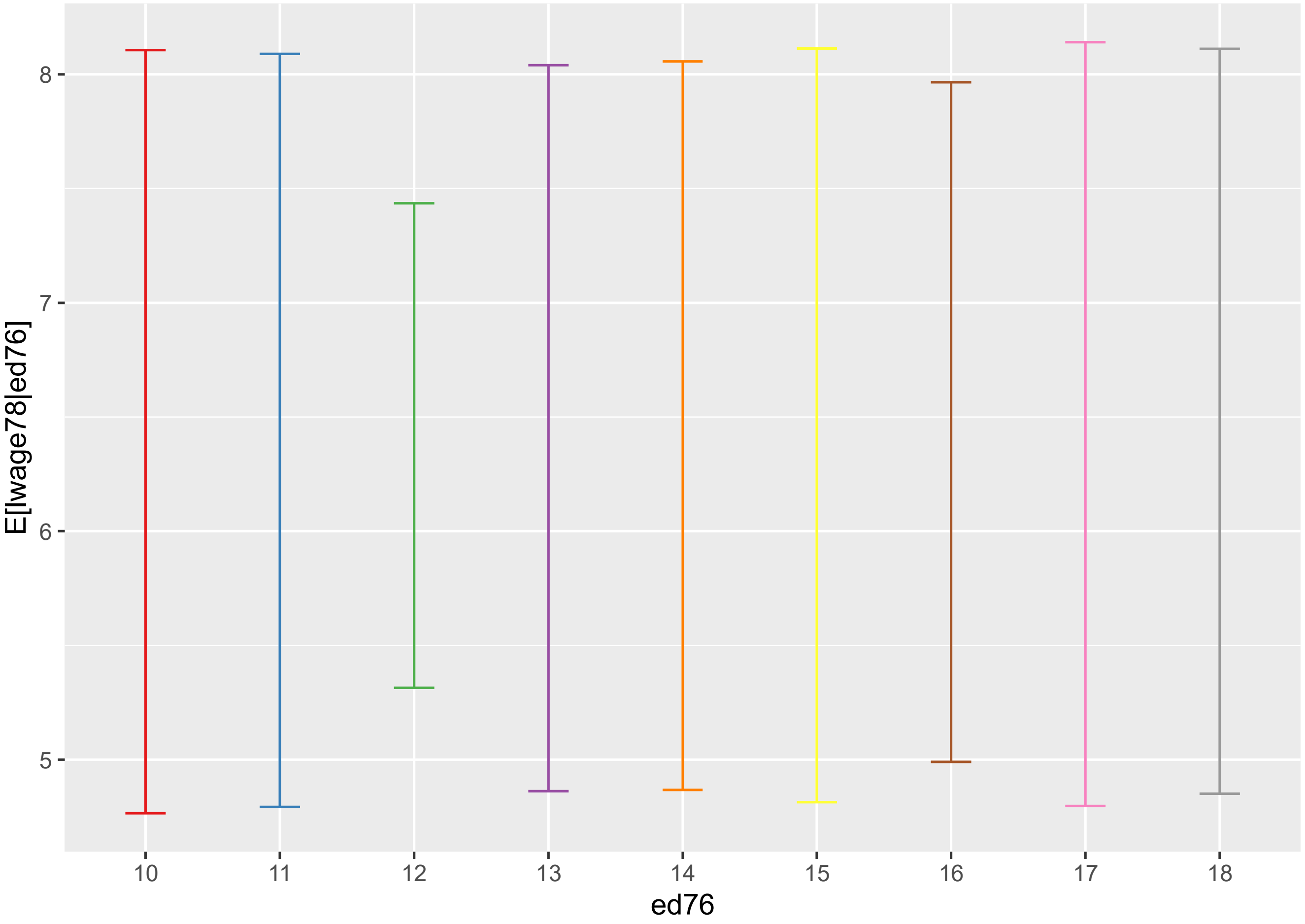
```
iv_bounds <- schooling %>%
  select(-c(black, age76)) %>%     # Remove controls
  group_by(ed76, nearc4) %>%       # Grouping by treatment and variable - (t_i,u_i)
  summarise(
    exp = mean(lwage78),           # calculate expected value of Y for each group
    count = n()) %>%               # counting observation in each group
  pivot_wider(                     # splitting count and exp vars by u_i
    values_from = c(exp,count),
    names_from = nearc4,
    values_fill = 0)%>%            # replacing missing data(no obs) with 0
    ungroup %>%
  mutate(                          # calculating probabilities
    prob_0 = count_0 / sum(count_0),
    prob_1 = count_1 / sum(count_1)
    ) %>%
    mutate(                                          #calculating bounds
      l_bound_0 = prob_0 * exp_0 + k_0 *(1-prob_0), #lower bound for u_i = 0
      l_bound_1 = prob_1 * exp_1 + k_0 *(1-prob_1), #lower bound for u_i = 1
      u_bound_0 = prob_0 * exp_0 + k_1 *(1-prob_0), #upper bound for u_i = 0
      u_bound_1 = prob_1 * exp_1 + k_1 *(1-prob_1)  #upper bound for u_i = 1
      ) %>%
    group_by(ed76) %>%                              #grouping by ed76 level
    mutate(l_bound = max(l_bound_0,l_bound_1),      #computing bounds
           u_bound = min(u_bound_0,u_bound_1)) %>%
    select(- contains(c("_0","_1"))) # removing uneeded variables
```

this are the results for upper and lower bounds:

```
print(iv_bounds)
```

```
## # A tibble: 9 x 3
## # Groups:   ed76 [9]
##    ed76 l_bound u_bound
##   <int>   <dbl>   <dbl>
## 1    10    4.77    8.11
## 2    11    4.79    8.09
## 3    12    5.32    7.44
## 4    13    4.86    8.04
## 5    14    4.87    8.06
## 6    15    4.81    8.11
## 7    16    4.99    7.97
## 8    17    4.80    8.14
## 9    18    4.85    8.11
```

```
ggplot(iv_bounds,aes(x = as.factor(ed76), color = as.factor(ed76))) +
  geom_errorbar(aes(ymin = l_bound,ymax = u_bound, width = 0.3)) +
  labs( x= "ed76", y ="E[lwage78|ed76]" ) +
  theme(legend.position = "none")+
   scale_colour_brewer(palette = "Set1")
```



we get quit big margins for every school of education. remembering that the two picks of `ed76` ’s distribution were in 12 and 16 years, that what led into smaller margins for these levels of treatment.

Calculating treatment effect (TE)

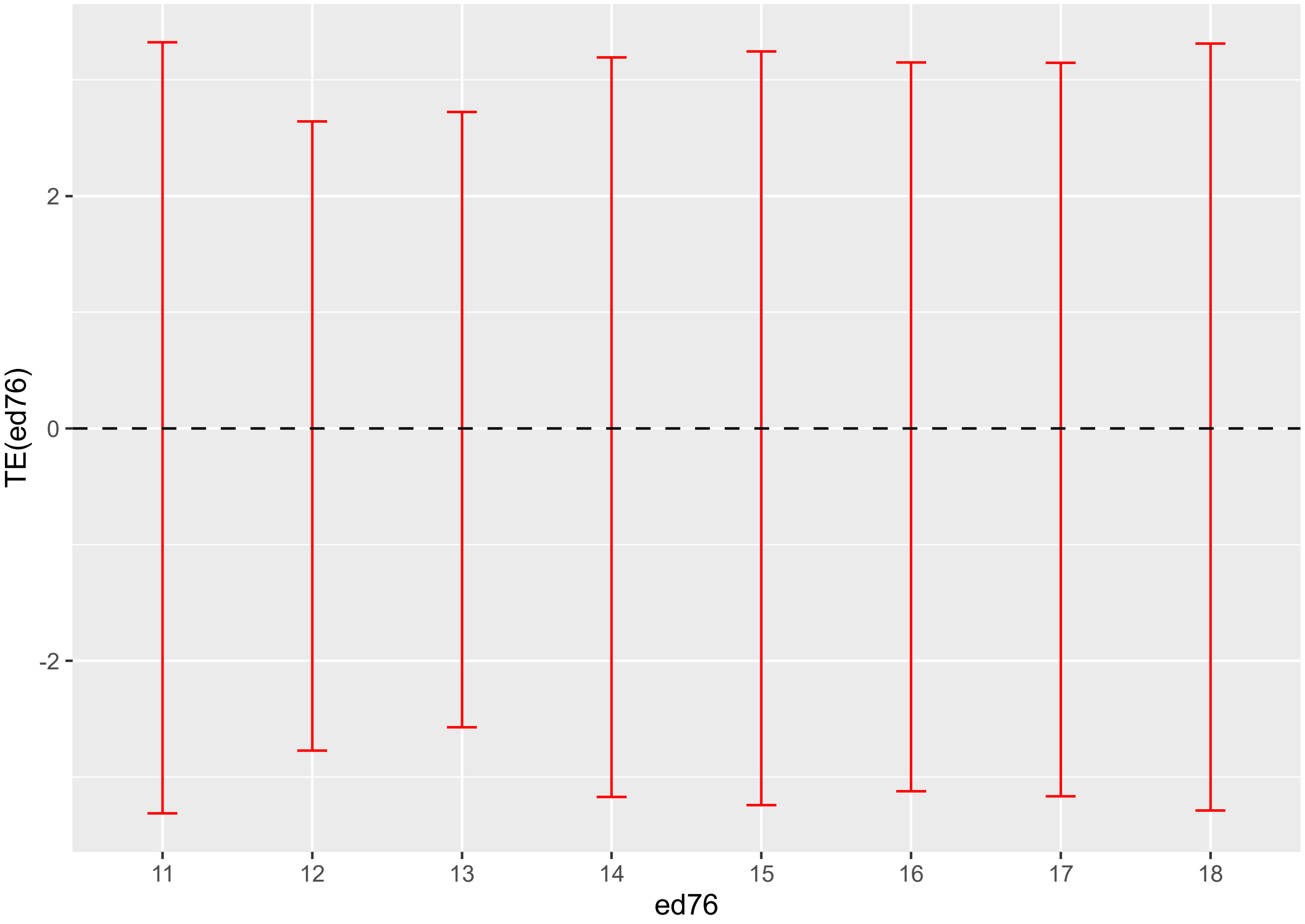
And for every $t \in T$, meaning for every value of `ed76` , set the treatment effect to be the difference between having treatment t, or $t-1$.

$$TE(t) = E[y(t) - y(t-1)]$$
$$\forall t > inf(T):$$
$$LB(t) - UB(t-1) \leq E[y(t) - y(t-1)] \leq UB(t) - LB(t-1)$$

```
iv_te <- iv_bounds[-1,]
colnames(iv_te) <- c("ed76","min_te","max_te" )  # calculating the diff acording to t
he formula above
for (i in 1:length(iv_te$ed76)) {
  iv_te$min_te[i] = iv_bounds$l_bound[i+1] - iv_bounds$u_bound[i]
  iv_te$max_te[i] = iv_bounds$u_bound[i+1] - iv_bounds$l_bound[i]
}

iv_te
```

```
## # A tibble: 8 x 3
## # Groups:   ed76 [8]
##    ed76 min_te max_te
##   <int>  <dbl>  <dbl>
## 1    11  -3.31   3.32
## 2    12  -2.77   2.64
## 3    13  -2.57   2.72
## 4    14  -3.17   3.19
## 5    15  -3.24   3.25
## 6    16  -3.12   3.15
## 7    17  -3.17   3.15
## 8    18  -3.29   3.31
```

```
iv_te %>% ggplot(aes(
  x = as.factor(ed76)
)) + geom_errorbar(aes(
  ymin = min_te,
  ymax = max_te
), width= 0.2, color = "red") +
  labs(x = "ed76", y = "TE(ed76)") +
  geom_hline(yintercept=0, linetype="dashed")
```



Discussion

The results does not seem to be useful: the margins for TE are very wide, and also contains negative effects, which are the opposite of what you would expect to get.

In the next part, I'll try to better define the treatment, and use MIV assumption.

MIV assumption

Redefining the problem

In the previous section, I tried to non parameticaly estimate the treatment effect of every year of extra education on wage, using proximity to 4 year collage as an IV.

In this section, the output is still `lwage78` , but the treatment is now set to be attendance to collage. more specifically. let $t \in \{0, 1\}$

$$t = \begin{cases} 0 & \text{ed76} < 12 \\ 1 & \text{ed76} > 12 \end{cases}$$

So, we want to examine how going to collage affects wages, and we want to do that using porximity to collage as an IV. We will not limit proximity tocollage only to a 4 yrs collage, but rather to both 4 and 2 yrs.

Again, selecting data

```
schooling_miv <- schooling_raw %>%
  filter(lwage78 != ".",ed76 >= 10) %>%  # remove missing values, only obs with 10 yrs or more of education
  mutate(
    nearc = as.numeric(nearc4 | nearc2), #new dummy for collage proximity
    attendc = as.numeric(ed76 > 12)      #new dummy for collage attendence
  ) %>%
  select(
    attendc,  # attendenc to collage - The treatment (t_i)
    nearc,    #  collage proximity - The IV (u_i)
    lwage78,  # log wage in 78 - The output (y_i)
    black     # Dummy for black - control (w_i)
      ) %>%
  mutate_at(vars(lwage78),funs(as.numeric)) %>%    #make lwage numeric
  mutate_at(vars(black,nearc,attendc),funs(as.factor)) #set dummys as factor
```

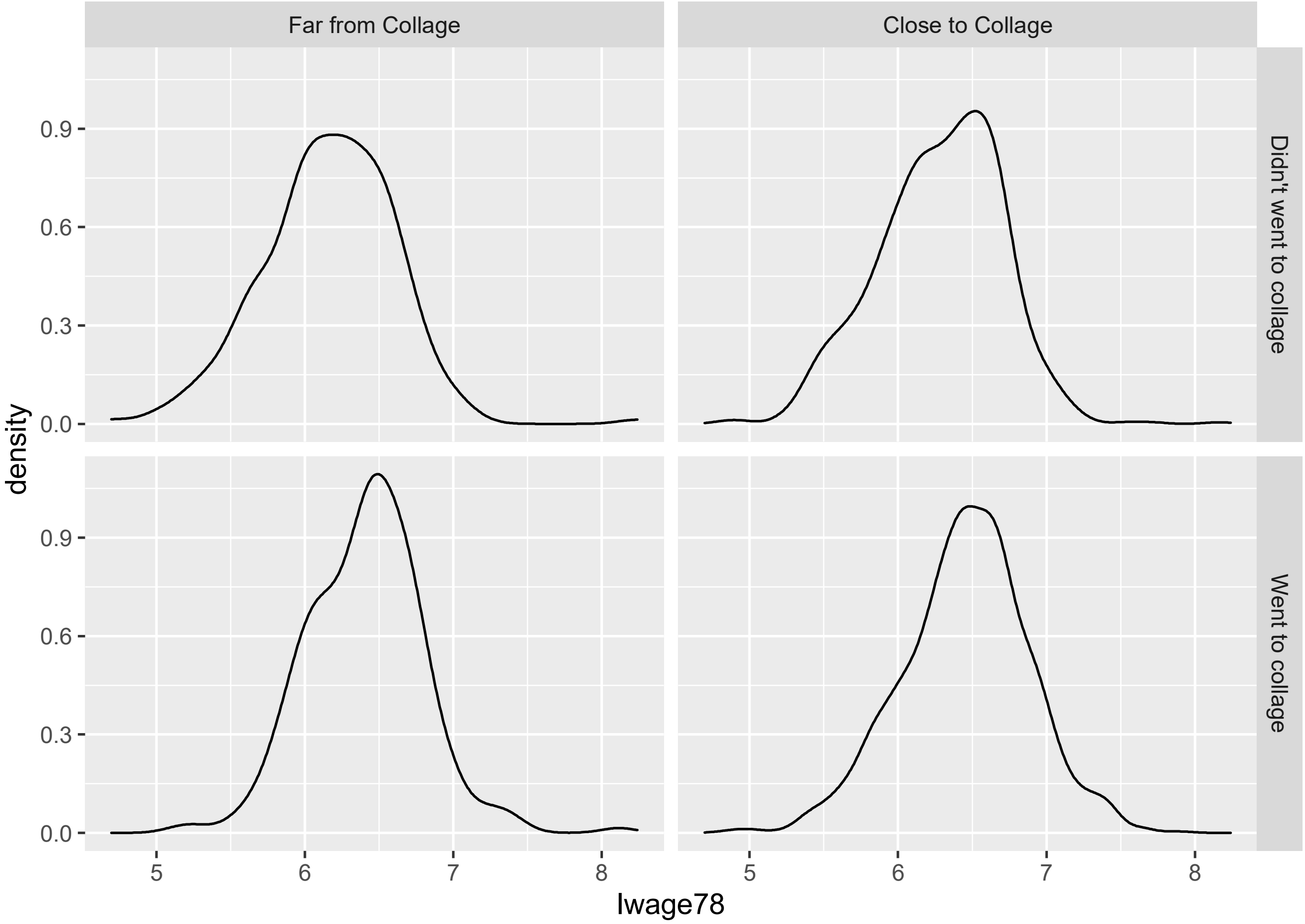
The MIV assumption

the MIV assumption, in our case, suggest that proximity to collage is positively correlated with wages when controlling on education. Specially:

$$\forall t \in T, \\ E[lwage78|z = t, nearc = 1] \geq E[lwage78|z = t, nearc = 0]$$

There are many reasons to assume that, specially the environment and education might be better when closer to collage, due to academic staff leaving nearby. But, it’s interesting to see if this is true in our sample:

```
attend.labs <- c("Didn't went to collage", "Went to collage")
names(attend.labs) <- c("0","1")

schooling_miv %>%
  ggplot(aes(lwage78)) +
  geom_density()+
  facet_grid(attendc~nearc,labeller = labeller(nearc = collage.labs, attendc =attend.labs))
```



The difference is notable, but not very significant.

MIV Boundaries

I calculated bounds related to the MIV assumption. That is, for every treatment level $t \in T$, I calculate bounds $UB(t), LB(t)$ such that:

$$LB(t) \equiv \sum_{u \in V} [P(v = u) \cdot \max_{u_1 \leq u} \underline{b}(t, u_1)]$$
$$UB(t) \equiv \sum_{u \in V} [P(v = u) \cdot \min_{u \leq u_2} \bar{b}(t, u_2)]$$

where

$$\underline{b}(t, u) \equiv E[y|z = t, v = u] \cdot P(z = t|v = u) + K_0 \cdot P(z \neq t|v = u)$$
$$\bar{b}(t, u) \equiv E[y|z = t, v = u] \cdot P(z = t|v = u) + K_1 \cdot P(z \neq t|v = u)$$

note that since $u \in \{0, 1\}$, and our assumption that

$$\forall t \in T,$$
$$E[lwage78|z = t, u = 1] \geq E[lwage78|z = t, u = 0]$$

we get that:

$$LB(t) \equiv \sum_{u \in V}[P(v=u) \cdot \max_{u_1 \leq u} \underline{b}(t,u_1)] = P(u=0) \cdot \max_{u_1 \leq 0} \underline{b}(t,u_1) + P(u=1) \cdot \max_{u_1 \leq 1} \underline{b}(t,u_1)$$

$$= P(u=0) \cdot \underline{b}(t,0) + P(u=1) \cdot \max_{u_1 \leq 1} \underline{b}(t,u_1)$$

$$UB(t) \equiv \sum_{u \in V}[P(v=u) \cdot \min_{u \leq u_2} \bar{b}(t,u_2)] = P(u=0) \cdot \min_{0 \leq u_2} \bar{b}(t,u_2) + P(u=1) \cdot \min_{1 \leq u_2} \bar{b}(t,u_2)$$

$$P(u=0) \cdot \min_{0 \leq u_2} \bar{b}(t,u_2) + P(u=1) \cdot \bar{b}(t,1)$$

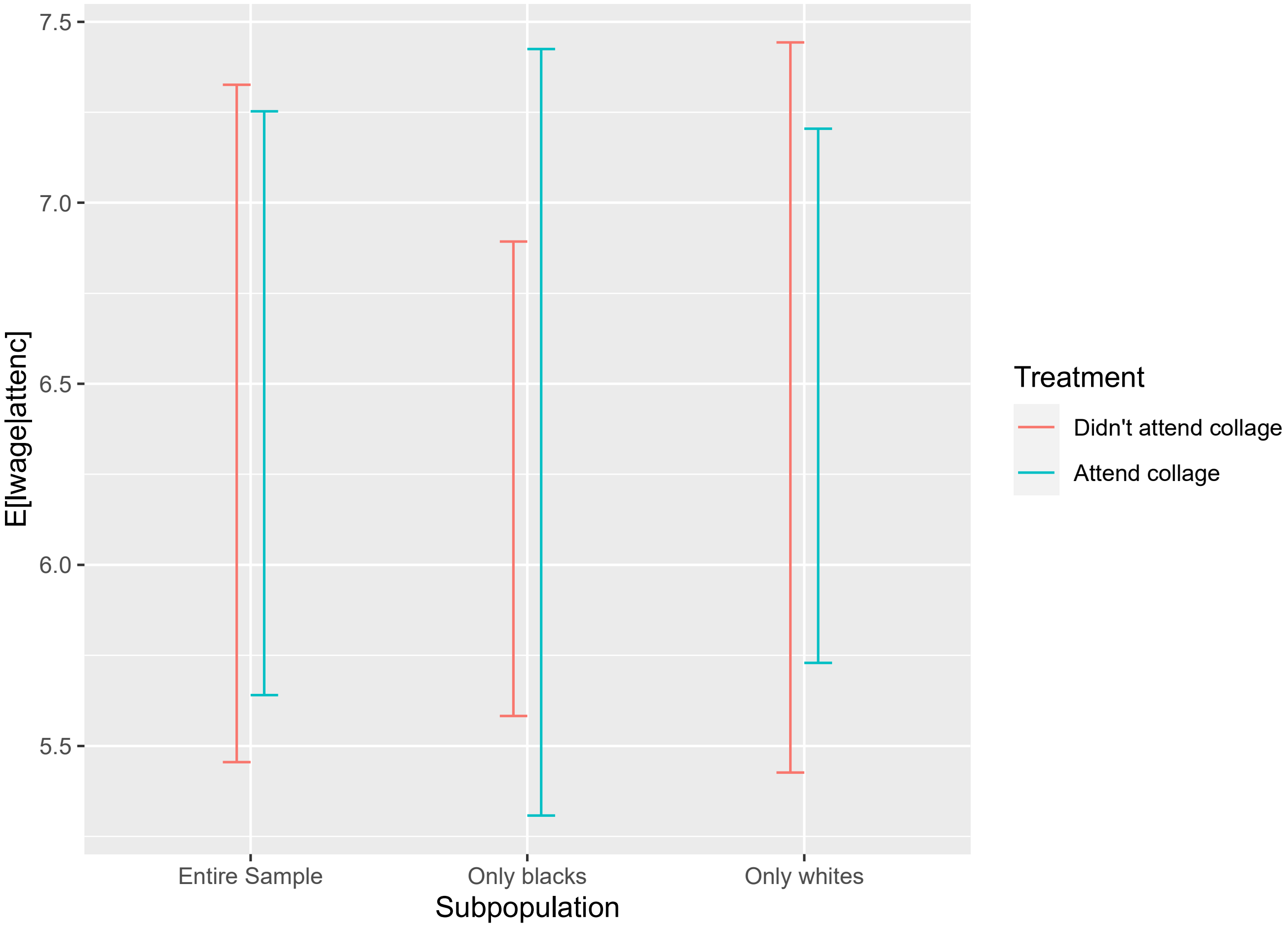
```
probs <- schooling_miv %>%
  select(nearc) %>% group_by(nearc) %>%
  summarise(count = n()) %>% mutate(
    prob = count/ sum(count)
  )
pr_0 <- as.numeric(probs[1,"prob"])
pr_1 <- 1- pr_0

MIV_bounds_all_pop <-
  schooling_miv %>%
  select(-c(black)) %>%     # Remove controls
  group_by(attendc, nearc) %>%     # Grouping by treatment and variable - (t_i,u_i)
  summarise(
    exp = mean(lwage78),           # calculate expected value of Y for each group
    count = n()) %>%               # counting observation in each group
  pivot_wider(                     # splitting count and exp vars by u_i
    values_from = c(exp,count),
    names_from = nearc,
    values_fill = 0)%>%            # replacing missing data(no obs) with 0
    ungroup %>%
  mutate(                          # calculating probabilities
    prob_0 = count_0 / sum(count_0),
    prob_1 = count_1 / sum(count_1)
    ) %>%
    mutate(                                          #calculating bounds
      l_bound_0 = prob_0 * exp_0 + k_0 *(1-prob_0), #lower bound for u_i = 0
      l_bound_1 = prob_1 * exp_1 + k_0 *(1-prob_1), #lower bound for u_i = 1
      u_bound_0 = prob_0 * exp_0 + k_1 *(1-prob_0), #upper bound for u_i = 0
      u_bound_1 = prob_1 * exp_1 + k_1 *(1-prob_1)  #upper bound for u_i = 1
      ) %>%
     group_by(attendc) %>%             #grouping by attendence level
     mutate(                           #computing bounds
       lb = pr_0 *l_bound_0 +pr_1 * max(l_bound_0,l_bound_1),
       ub = pr_0 *min(u_bound_0,u_bound_1) + pr_1 * u_bound_1
           ) %>%
     select(- contains(c("_0","_1"))) %>%  # removing uneeded variables
    pivot_wider(
      values_from = lb:ub,
      names_from = attendc
    ) %>% mutate(pop = "Entire Sample")
```

For comparison, I will do the same calculation for sub populations: black and white people.

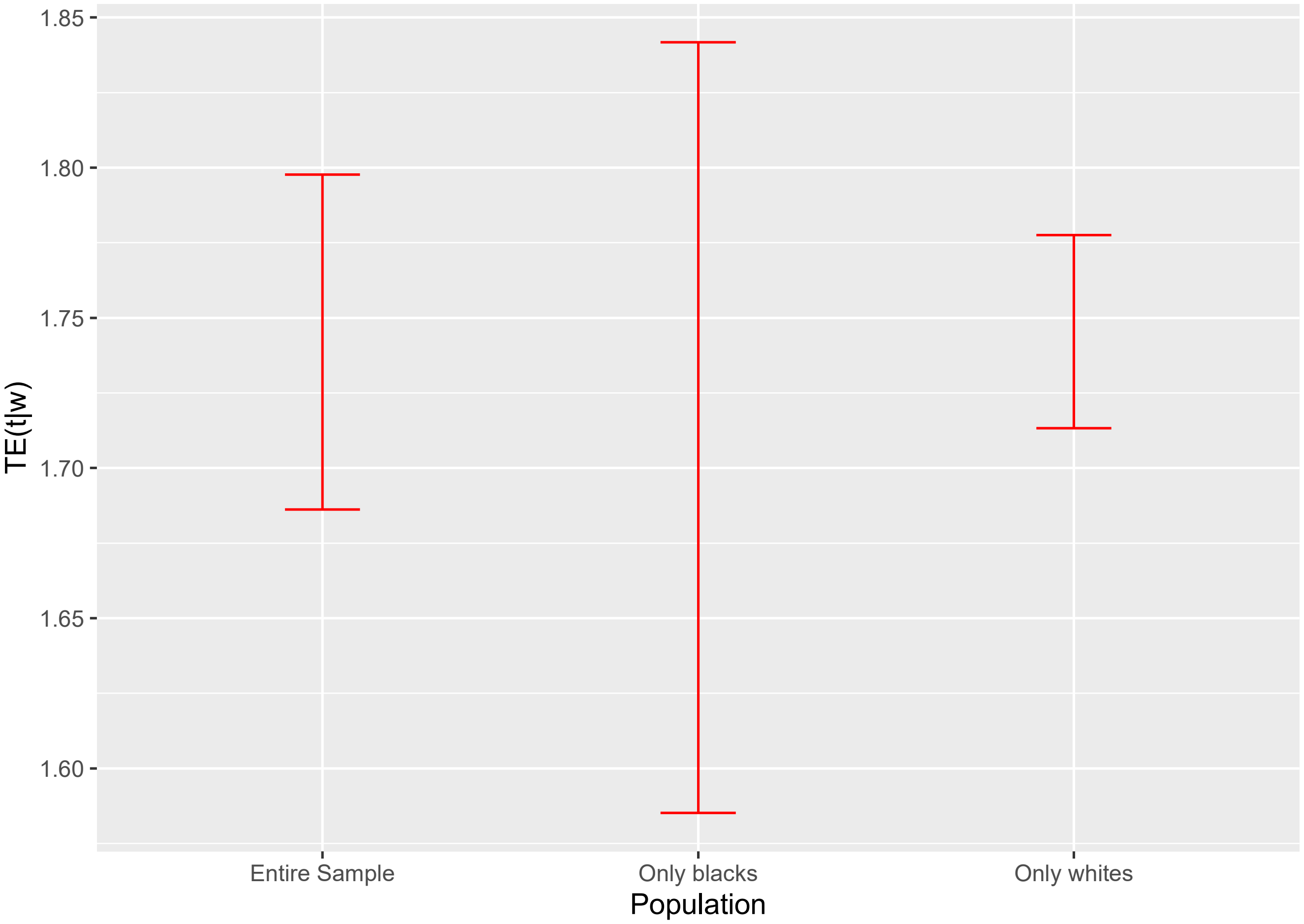
let $w \in W$ be a vector of covarients, in this example $W = \{black, white\}$, the calculation is the same but this time the probabilities and expectations are conditional on w.

$$LB(t,w) \equiv \sum_{u \in V} [P(v = u|w) \cdot \max_{u_1 \leq u} \underline{b}(t, u_1, w)]$$
$$UB(t,w) \equiv \sum_{u \in V} [P(v = u|w) \cdot \min_{u \leq u_2} \bar{b}(t, u_2, w)]$$

```
miv_bounds <- MIV_bounds_all_pop %>%  # this df is for storing the bounds
  bind_rows(MIV_bounds_only_blacks,MIV_bounds_only_whites)

miv_bounds_plot <- miv_bounds %>%   # this df is for storing the bounds, in a way good for plotting
  pivot_longer(cols = !"pop",
               names_to = c(".value","treatment"),
               names_pattern = "(.*)_(.)") %>%
  mutate(treatment = factor(treatment,levels=c(0,1), labels=c("Didn't attend collage", "Attend collage")))

miv_bounds_plot %>% ggplot(aes(
  x = pop)) +
  geom_errorbar(width = 0.2, aes(
    ymin = lb,
    ymax = ub,
    colour = treatment),
    position = "dodge") +
  labs(x = "Subpopulation", y = "E[lwage|attenc]", color = "Treatment")
```



```
miv_te  <-  miv_bounds %>% group_by(pop) %>%  #calculating treatmnet effects
  summarise(min_te = ub_0 - lb_1,
            max_te = ub_1 - lb_0)

miv_te %>% ggplot(aes(x = pop)) +
  geom_errorbar(width = 0.2, aes(
    ymin = min_te,
    ymax = max_te),
    color = "red") +
  labs( x= "Population", y = "TE(t|w)")
```



Discussion

We obtained nonparaetric bounds on the ATE for two sub populations.

We can see that the return for going to collage is strictly positive for all subgroups, and that the segment in which the TE lies is bigger for blacks then for whites. This corresponds to the assumption that the process determining wages is sensitive to race (throw discrimination, profession selection. etc.) as well as to education.