BT2101 GA2 Group 67 Submission

Lo Zhi Hao

2022-10-16

1 Econometric Analysis using R II

Please use the injury dataset from the Wooldridge package in R to answer this question. Please use data only from Kentucky. Please carefully read the document of data description. Meyer, Viscusi, and Durbin (1995) (hereafter, MVD) studied the effect of injury compensation on the duration of “injury leave”.

On July 15, 1980, Kentucky raised the cap on weekly earnings that were covered by workers’ compensation. Such an increase in the cap has no effect on the benefit of low-income workers, but it makes it less costly for a high-income worker to stay on “injury leave” instead of working. In other words, they still earn a good salary in terms of injury compensation if they choose to stay out of work.

For the sake of this group assignment, consider the control group as low-income workers, and the treatment group as high-income workers. Using random samples both before and after the policy change, MVD was able to test whether more generous workers’ compensation causes people to stay out of work longer, ceteris paribus.

```
## Setting up the environment for further studies

## install.packages("wooldridge")
## install.packages("dplyr")
## install.packages("MASS")
## install.packages("ggplot2")

library(wooldridge)
library(dplyr)
library(knitr)
library(corrplot)
library(ggplot2)

## Downloading the dataset
data('injury')
?injury

## filter to keep only kentucky data, and rename columns to make it more clearer

injury.filtered <- injury %>%
  filter(ky == 1) %>%
  rename(duration = durat, log_duration = ldurat,
         after_1980 = afchnge) %>%
  mutate(highearn = as.factor(highearn), after_1980 = as.factor(after_1980))
```

```
## inspecting the dataset

str(injury.filtered)
```

```
## 'data.frame':    5626 obs. of  30 variables:
##  $ duration    : num  1 1 84 4 1 1 7 2 175 60 ...
##  $ after_1980  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ highearn    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ male        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ married     : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ hosp        : int  1 0 1 1 0 0 0 1 1 1 ...
##  $ indust      : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ injtype     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age         : int  26 31 37 31 23 34 35 45 41 33 ...
##  $ prewage     : num  405 644 398 528 529 ...
##  $ totmed      : num  1188 361 8964 1100 373 ...
##  $ injdes      : int  1010 1404 1032 1940 1940 1425 1110 1207 1425 1010 ...
##  $ benefit     : num  247 247 247 247 212 ...
##  $ ky          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ mi          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ log_duration: num  0 0 4.43 1.39 0 ...
##  $ afhigh      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ lprewage    : num  6 6.47 5.99 6.27 6.27 ...
##  $ lage        : num  3.26 3.43 3.61 3.43 3.14 ...
##  $ ltotmed     : num  7.08 5.89 9.1 7 5.92 ...
##  $ head        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ neck        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ upextr      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trunk       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lowback     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lowextr     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ occdis      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ manuf       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ construc    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ highlpre    : num  6 6.47 5.99 6.27 6.27 ...
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

```
head(injury.filtered)
```

```
##   duration after_1980 highearn male married hosp indust injtype age  prewage
## 1        1          1        1    1       0    1      3       1  26 404.9500
## 2        1          1        1    1       1    0      3       1  31 643.8250
## 3       84          1        1    1       1    1      3       1  37 398.1250
## 4        4          1        1    1       1    1      3       1  31 527.8000
## 5        1          1        1    1       1    0      3       1  23 528.9375
## 6        1          1        1    1       1    0      3       1  34 614.2500
##      totmed injdes  benefit ky mi log_duration afhigh lprewage     lage
## 1 1187.5732   1010 246.8375  1  0     0.000000      1 6.003764 3.258096
## 2  361.0786   1404 246.8375  1  0     0.000000      1 6.467427 3.433987
## 3 8963.6572   1032 246.8375  1  0     4.430817      1 5.986766 3.610918
## 4 1099.6483   1940 246.8375  1  0     1.386294      1 6.268717 3.433987
## 5  372.8019   1940 211.5750  1  0     0.000000      1 6.270870 3.135494
## 6  211.0199   1425 176.3125  1  0     0.000000      1 6.420402 3.526361
##    ltotmed head neck upextr trunk lowback lowextr occdis manuf construc
## 1 7.079667    1    0      0     0       0       0      0     0        0
## 2 5.889095    1    0      0     0       0       0      0     0        0
## 3 9.100934    1    0      0     0       0       0      0     0        0
## 4 7.002746    1    0      0     0       0       0      0     0        0
## 5 5.921047    1    0      0     0       0       0      0     0        0
## 6 5.351953    1    0      0     0       0       0      0     0        0
##   highlpre
## 1 6.003764
## 2 6.467427
## 3 5.986766
## 4 6.268717
## 5 6.270870
## 6 6.420402
```

(a) Estimate the impact of the policy based on a difference-in-differences (DiD) regression without including any other control variables.

Guidelines:

• Use durat as the dependent variable.

• Your specification should include three terms: the control/treatment group dummy, before/after intervention time dummy, and a term capturing the interaction between control/treatment and before/after the intervention.

Please interpret all the coefficient estimates in your regression table.

```
## creating the linear model

linear.model <- lm(duration ~ highearn + after_1980 + highearn * after_1980, data = injury.filtered)
summary(linear.model)
```

```
##
## Call:
## lm(formula = duration ~ highearn + after_1980 + highearn * after_1980,
##     data = injury.filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.644  -6.787  -4.272  -0.272 175.728
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.2716     0.5229  11.994  < 2e-16 ***
## highearn1               4.9050     0.8071   6.077  1.3e-09 ***
## after_19801             0.7658     0.7607   1.007    0.314
## highearn1:after_19801   0.9513     1.1654   0.816    0.414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.59 on 5622 degrees of freedom
## Multiple R-squared:  0.01577,    Adjusted R-squared:  0.01524
## F-statistic: 30.02 on 3 and 5622 DF,  p-value: < 2.2e-16
```

The relationship is as follows:

$$\texttt{duration} = 6.2716 + 4.9050 \times \texttt{highearn} + 0.7658 \times \texttt{after\_1980} + 0.9513 \times \texttt{highearn:after\_1980}$$

Multiple R squared is 0.01577, and Adjusted R squared is 0.01524.

The coefficient for afchnge is 0.7658, which means that duration of benefits is 0.7658 weeks more on average for people after the policy change compared to before the policy change. The coefficient for highearn is 4.91, which means that the duration of benefits is 4.91 weeks more on average for worker with high income compared to low-income worker. The coefficient of the interaction variable between afchnge and highearn is 0.9513. The average treatment effect is 0.9513 weeks higher for high-income workers after the policy change.

Thus, for different groups of people, the duration is as follows:

Low earners before policy change in 1980 (in weeks) = 6.2716

Low earners after policy change in 1980 (in weeks) = 6.2716 + 0.7658 = 7.0374

High earners before policy change in 1980 (in weeks) = 6.2716 + 4.9050 = 11.1766

High earners after policy change in 1980 (in weeks) = 6.2716 + 4.9050 + 0.7658 + 0.9513 = 12.8937

(b) Change the dependent variable into ldurat, and repeat a similar DiD regression as the question (a).Please interpret all the coefficient estimates in this regression table.

```
## creating the second linear model

linear.model2 <- lm(log_duration ~ highearn + after_1980 + highearn * after_1980, data = injury.filtered)
summary(linear.model2)
```

```
##
## Call:
## lm(formula = log_duration ~ highearn + after_1980 + highearn *
##     after_1980, data = injury.filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9666 -0.8872  0.0042  0.8126  4.0784
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.125615   0.030737  36.621  < 2e-16 ***
## highearn1             0.256479   0.047446   5.406 6.72e-08 ***
## after_19801           0.007657   0.044717   0.171  0.86404
## highearn1:after_19801 0.190601   0.068509   2.782  0.00542 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.269 on 5622 degrees of freedom
## Multiple R-squared:  0.02066,    Adjusted R-squared:  0.02014
## F-statistic: 39.54 on 3 and 5622 DF,  p-value: < 2.2e-16
```

The relationship is as follows:

log_duration = 1.125615 + 0.256479 x highearn + 0.007657 x after_1980 + 0.190601 x highearn :after_1980

Multiple R squared is 0.02066, and Adjusted R squared is 0.02014.

The intercept of 1.1256 showed that the mean duration of benefits of low earners(control group) before the change in benefits is e^1.1256 unit. The coefficient for afchnge is 0.007657, which means that after the policy change the duration of benefits is 0.7657% higher compared to before the policy change for both high-income workers and low-income workers. The coefficient for highearn is 0.2565, which means that a high-income worker is associated with a 25.65% increase on average in duration of benefits compared to low income workers before the policy change in 1980. The coefficient for the interaction variable is 0.1906, which suggests the average treatment effect is 19.1% more in duration for high earners after the policy change in 1980.

(c) Using ldurat as the dependent variable, and the independent variables already used in the previous question, now add more control variables: male, married, and the full set of industry and injury type dummy variables. How does the coefficient of interaction term change when these other factors are controlled? Is the estimate still statistically significant? Please explain the changes, if any.

```
## creating the new linear model

linear.model3 <- lm(log_duration ~ highearn + after_1980 + afhigh + male + married + as.factor(indust) + as.facto
r(injtype) , data = injury.filtered)
summary(linear.model3)
```

```
##
## Call:
## lm(formula = log_duration ~ highearn + after_1980 + afhigh +
##     male + married + as.factor(indust) + as.factor(injtype),
##     data = injury.filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3436 -0.8541  0.0989  0.7856  4.4372
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.57135    0.10266   5.565 2.74e-08 ***
## highearn1            0.17576    0.05175   3.397 0.000687 ***
## after_19801          0.01063    0.04492   0.237 0.812974
## afhigh               0.23088    0.06952   3.321 0.000904 ***
## male                -0.09794    0.04455  -2.198 0.027959 *
## married              0.12210    0.03912   3.121 0.001812 **
## as.factor(indust)2   0.27087    0.05867   4.617 3.98e-06 ***
## as.factor(indust)3   0.16067    0.04090   3.928 8.67e-05 ***
## as.factor(injtype)2  0.78381    0.15617   5.019 5.36e-07 ***
## as.factor(injtype)3  0.33536    0.09234   3.632 0.000284 ***
## as.factor(injtype)4  0.64035    0.10087   6.348 2.36e-10 ***
## as.factor(injtype)5  0.50530    0.09281   5.445 5.42e-08 ***
## as.factor(injtype)6  0.39361    0.09356   4.207 2.63e-05 ***
## as.factor(injtype)7  0.78661    0.20703   3.800 0.000147 ***
## as.factor(injtype)8  0.51390    0.12928   3.975 7.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.251 on 5334 degrees of freedom
##   (277 observations deleted due to missingness)
## Multiple R-squared:  0.0412, Adjusted R-squared:  0.03868
## F-statistic: 16.37 on 14 and 5334 DF,  p-value: < 2.2e-16
```

The coefficient of interaction changed by 0.23088 - 0.190601 = 0.040279, it also means that after the change, the duration of benefits of high earners increased by 4.11% compared to the previous model. The estimate of this model is still statistically significant compared to the previous model, as the p-value of the interaction term goes from 0.00542 to 0.000904. A reason of the change might be industry type being a confounding variable. Different industry type have different wages thus result in the distinction of high earner and low earner. Working in different industry will also have different level of injury risks, which will affect the duration of “injury leave”. Other than that, the new model also controls for gender, marriage, industry type and injury type, the new coefficients are actually the expected result of the reference group of these dummy variables.

(d) Your colleague argues that we cannot draw a causal inference due to the small magnitude of the R-squared and adjust R-squared in question (c). How will you respond to this argument? Explain.

The low R-squared can only showed that the explanatory power of the model is weak, and cannot determine the causal inference as the value of R-squared only shows correlation, not causation. Furthermore, low R-squared value might also means that the data set has high variability or low prediction accuracy, and these do not necessarily indicates no causality. This does not suggest that the coefficients are not meaningful, and thus should not be excluded when we are drawing a causal inference.

For example, for a model with small R-squared like the one above, if all the independent variables are statistically significant, it may imply that there is a small but statistically significant effect from each independent variables on the dependent variable. As such, these variables should not be ruled out and it is still possible to draw a causal inference out of this despite its small R-squared values.

(e) What is the most critical assumption of the difference-in-differences model? Even if you cannot provide conclusive proof, can you use the data to offer some qualitative support/opposition to the validity of this assumption in this dataset? Using your own words, discuss what plots and/or statistics would help you support/oppose this assumption. Construct/compute these plots/statistics and make

a concluding statement describing your support/opposition to the validity of this critical assumption in this dataset

A major assumption for Difference-In-Difference assumption is the Parallel Path assumption, which states that in the absence of treatment or intervention, the unobserved difference between treatment and control group will stay the same over time.

Even though there are no ways to statistically inspect the assumption, we can judge via visual inspections. A way of doing so is by producing before / after control / treatment graphs and tables to showcase the difference in values after the treatment on the treatment group.

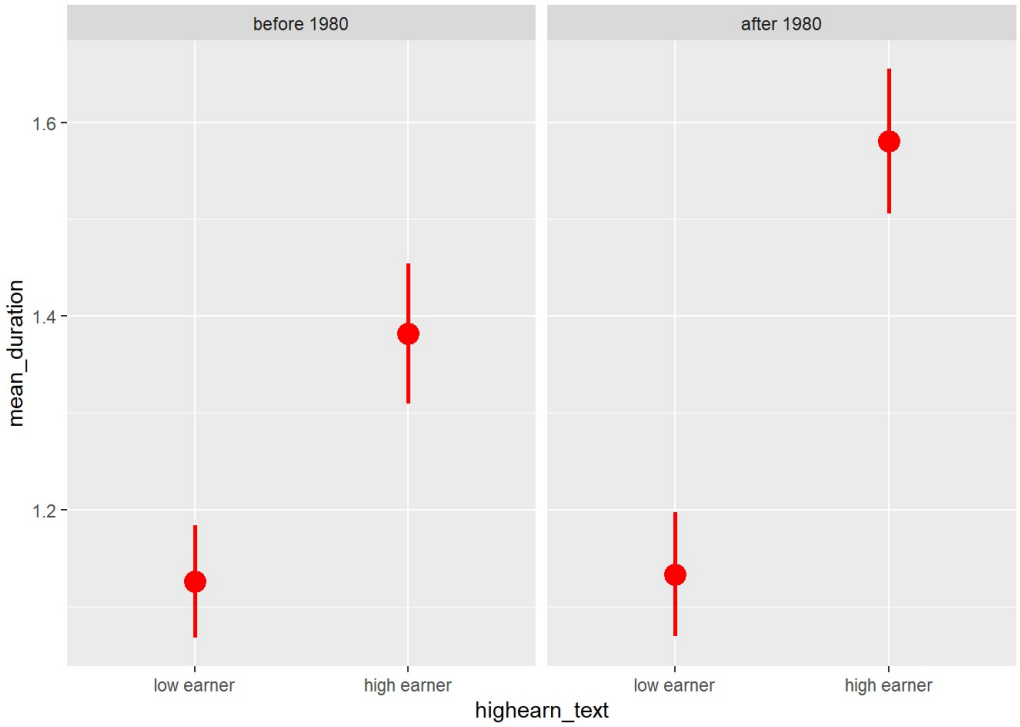
```
## inspecting on the difference in mean and 95% confidence intervals in control and treatment groups before and a
fter treatment

# manipulating data to make it look better in the plot

injury.toplot <- injury.filtered %>%
  mutate(highearn_text = factor(highearn, labels = c("low earner", "high earner"))) %>%
  mutate(after_1980_text = factor(after_1980, labels = c("before 1980", "after 1980"))) %>%
  group_by(highearn_text, after_1980_text) %>%
  summarize(mean_duration = mean(log_duration),
            se_duration = sd(log_duration) / sqrt(n()),
            upper = mean_duration + (1.96 * se_duration),
            lower = mean_duration + (-1.96 * se_duration))
```

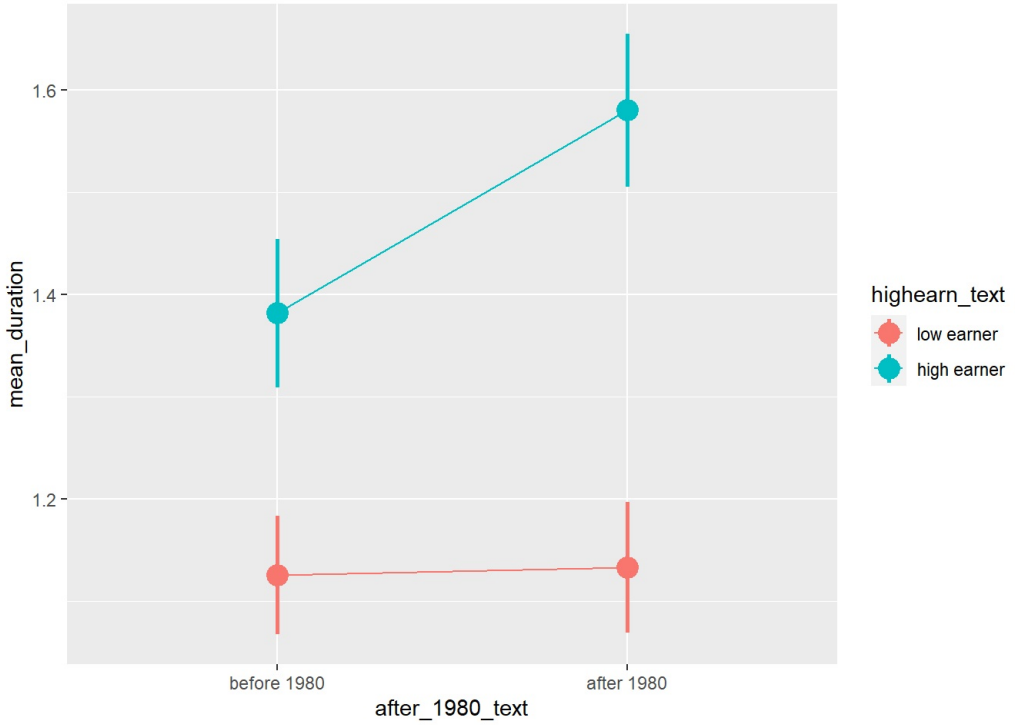
```
## `summarise()` has grouped output by 'highearn_text'. You can override using the
## `.groups` argument.
```

```
ggplot(injury.toplot, aes(x = highearn_text, y = mean_duration)) +
  geom_pointrange(aes(ymin = lower, ymax = upper),
                  color = "red", size = 1) +
  facet_wrap(vars(after_1980_text))
```




```
## next, we can compute the before / after control / treatment graphs to visualise the difference

ggplot(injury.toplot, aes(x = after_1980_text, y = mean_duration, color = highearn_text)) +
  geom_pointrange(aes(ymin = lower, ymax = upper), size = 1) +
  # The group = highearn here makes it so the lines go across categories
  geom_line(aes(group = highearn_text))
```




```
# calculating difference between control and treatment, and between before and after 1980

injury.diffs <- injury.filtered %>%
  group_by(after_1980, highearn) %>%
  summarize(mean_log_duration = mean(log_duration),
            # Calculate average with regular duration too, just for fun
            mean_duration = mean(duration))
```

```
## `summarise()` has grouped output by 'after_1980'. You can override using the
## `.groups` argument.
```

```
kable(injury.diffs)
```

after_1980	highearn	mean_log_duration	mean_duration
0	0	1.125615	6.271554
0	1	1.382094	11.176602
1	0	1.133273	7.037328
1	1	1.580353	12.893626

```
# pulling out values from the table

before_treatment <- injury.diffs %>%
  filter(after_1980 == 0, highearn == 1) %>%
  pull(mean_log_duration)

before_control <- injury.diffs %>%
  filter(after_1980 == 0, highearn == 0) %>%
  pull(mean_log_duration)

after_treatment <- injury.diffs %>%
  filter(after_1980 == 1, highearn == 1) %>%
  pull(mean_log_duration)

after_control <- injury.diffs %>%
  filter(after_1980 == 1, highearn == 0) %>%
  pull(mean_log_duration)

## calculating beta1

beta1 <- after_treatment - before_treatment
beta1
```

```
## [1] 0.1982585
```

```
## calculating beta2

beta2 <- after_control - before_control
beta2
```

```
## [1] 0.007657313
```

```
## calculating beta3

beta3 <- beta1 - beta2
beta3
```

```
## [1] 0.1906012
```

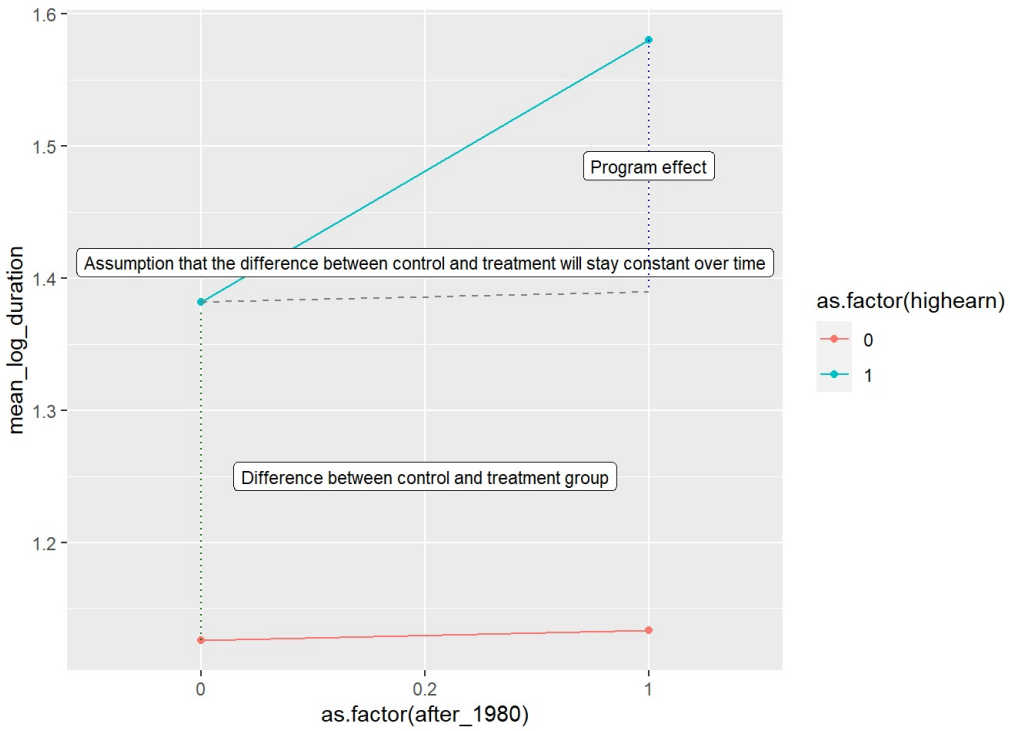
```
data.frame <- data.frame(beta1, beta2, beta3)
names(data.frame) <- c("diff_treatment_before_after", "diff_control_before_after", "diff_diff")
data.frame
```

```
##   diff_treatment_before_after diff_control_before_after diff_diff
## 1                   0.1982585                0.007657313 0.1906012
```

```
## Visualising this on the plot

ggplot(injury.diffs, aes(x = as.factor(after_1980), y = mean_log_duration, color = as.factor(highearn))) +
  geom_point() +
  geom_line(aes(group = as.factor(highearn))) +
  annotate(geom = "segment", x = "0", xend = "1",
           y = before_treatment, yend = after_treatment - beta3,
           linetype = "dashed", color = "grey50") +
  annotate(geom = "label", x = "0.2", y = before_treatment + 0.03,
           label = "Assumption that the difference between control and treatment will stay constant over time", s
ize = 3) +
  annotate(geom = "segment", x = "1", xend = "1",
           y = after_treatment, yend = after_treatment - beta3,
           linetype = "dotted", color = "blue") +
  annotate(geom = "segment", x = "0", xend = "0",
           y = before_control, yend = before_treatment - beta2,
           linetype = "dotted", color = "darkgreen", label = "beta2") +
  annotate(geom = "label", x = "0.2", y = 1.25,
           label = "Difference between control and treatment group", size = 3) +
  annotate(geom = "label", x = "1", y = after_treatment - (beta3 / 2),
           label = "Program effect", size = 3)
```

```
## Warning: Ignoring unknown parameters: label
```



From the plot we created, we can see that both control and treatment group see an increase in duration of benefits. From this, we can see that it is possible that both control and treatment groups have the same outcome trend and it is possible that in the absence of treatment the difference between treatment and control groups remain constant over time.

However, from the plots earlier, we also cannot determined conclusively whether the key assumption is met or not. This is due to the lack of baseline data to support the fact that treatment and control groups will follow the same slight upwards trend as seen in our plot. We need more data along the timeline before 1980 (when change in benefits happened) to determine if the parallel trend exist before the policy change. We cannot rule out the possibility that there are factors affecting the duration of benefits over time which are not included in the research. If more supplementary data is provided (for example data of control and treatment group duration of benefit before 1980), we can plot out another difference-in-difference graph as above to visualize the difference in mean by the two groups before and after the policy change. From this, we can come to a more conclusive statement on the validity of the assumption.

Another possible reason behind why we are unable to conclude whether this assumption is valid is due to the fact that the control group and treatment group are fundamentally different and may not be comparable. For example, the higher earners may be workers who works in offices compared to low earners who may work as more physical workers. Hence, from a logical perspective, the injuries they encounter and the duration of health benefits may be inherently different between the two. For example, office workers may encounter more chronic diseases such as diabetes and back pains, while lower wages workers may encounter more physical injuries such as fractures and sprains. As such, without historical data to support the statement, we are unable to identify the trends between the two groups, and as such are unable to conclusively determine whether that in the absence of treatment, the two groups have unobserved difference that are same over the time.