Econometrics 2, Assignment 1

Henry Haustein

Section A: Proof of Variance

(a) We start by showing $\sum(W_i - \bar{W})^2 = N\bar{W}(1-\bar{W}) = \left(\frac{1}{N_c} + \frac{1}{N_t}\right)^{-1}$:

$$\begin{aligned}\sum_{i=1}^N (W_i - \bar{W})^2 &= \sum_{i=1}^N W_i^2 - 2\bar{W}\underbrace{\sum_{i=1}^N W_i}_{N\bar{W}} + N\bar{W}^2 \\ &= \sum_{i=1}^N W_i^2 - N\bar{W}^2 \\ &= N\bar{W} - N\bar{W}^2 \qquad (1) \\ &= N\bar{W}(1-\bar{W}) \\ &= N\frac{N_t}{N}\cdot\frac{N_c}{N} \\ &= \frac{N_t N_c}{N} \qquad (2) \\ &= \frac{N_t N_c}{N_t + N_c} \\ &= \left(\frac{1}{N_c} + \frac{1}{N_t}\right)^{-1}\end{aligned}$$

where we use in (1) the fact that for $W \in \{0,1\}$ $W = W^2$.

We can now show that $V^{const} = V^{homo}$:

$$\begin{aligned}
V^{const} &= s^2\left(\frac{1}{N_c} + \frac{1}{N_t}\right) \\
&= \frac{1}{N-2}\left((N_c - 1)\frac{1}{N_c - 1}\sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 + (N_t - 1)\frac{1}{N_t - 1}\sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2\right)\left(\frac{1}{N_c} + \frac{1}{N_t}\right) \\
&= \frac{1}{N-2}\left(\sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 + \sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2\right)\left(\frac{1}{N_c} + \frac{1}{N_t}\right) \\
&= \frac{1}{N-2}\left(\sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 + \sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2\right)\left(\frac{1}{N_c} + \frac{1}{N_t}\right) \cdot \frac{\left(\frac{1}{N_c} + \frac{1}{N_t}\right)^{-1}}{\left(\frac{1}{N_c} + \frac{1}{N_t}\right)^{-1}} \\
&= \frac{\frac{1}{N-2}\left(\sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 + \sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2\right)}{\left(\frac{1}{N_c} + \frac{1}{N_t}\right)^{-1}} \\
&= \frac{\frac{1}{N-2}\left(\sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 + \sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2\right)}{\sum_{i=1}^N (W_i - \bar{W})^2} \\
&= \frac{\hat{\sigma_\varepsilon^2}}{\sum_{i=1}^N (W_i - \bar{W})^2} \\
&= V^{homo}
\end{aligned}$$

(b) We show $V^{hetero} = V^{neyman}$:

$$\begin{aligned}
V^{hetero} &= \frac{\sum_{i=1}^N \hat{\varepsilon}_i^2 \cdot (W_i - \bar{W})^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2} \\
&= \frac{\sum_{i=1}^N (Y_i - \alpha^{OLS} - \beta^{OLS} W_i)^2 \cdot (W_i - \bar{W})^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2} \\
&= \frac{\sum_{i=1}^N (Y_i - \bar{Y} - (\bar{Y}_t - \bar{Y}_c)(W_i - \bar{W}))^2 \cdot (W_i - \bar{W})^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2}
\end{aligned}$$

We are now using the fact that $\bar{Y} = \bar{W}\bar{Y}_t + (1 - \bar{W})\bar{Y}_c$

$$\begin{aligned}
&= \frac{\sum_{W_i:i=0}(Y_i - \bar{W}\bar{Y}_t - (1 - \bar{W})\bar{Y}_c + \bar{Y}_t\bar{W} - \bar{Y}_c\bar{W})^2 \cdot (-\bar{W})^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2} \\
&+ \frac{\sum_{W_i:i=1}(Y_i - \bar{W}\bar{Y}_t - (1 - \bar{W})\bar{Y}_c - \bar{Y}_t + \bar{Y}_t\bar{W} + \bar{Y}_c - \bar{Y}_c\bar{W})^2 \cdot (1 - \bar{W})^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2} \\
&= \frac{\sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 \cdot (-\bar{W})^2 + \sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2 \cdot (1 - \bar{W})^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2} \\
&= \frac{\left(\frac{N_t}{N}\right)^2 \sum_{W_i:i=0}(Y_i - \bar{Y}_c)^2 + \left(\frac{N_c}{N}\right)^2 \sum_{W_i:i=1}(Y_i - \bar{Y}_t)^2}{\left(\sum_{i=1}^N (W_i - \bar{W})^2\right)^2}
\end{aligned}$$

using (2)

$$
\begin{aligned}
&= \frac{\frac{N_t^2}{N^2}\sum_{W_i:i=0}(Y_i-\bar{Y}_c)^2 + \frac{N_c^2}{N^2}\sum_{W_i:i=1}(Y_i-\bar{Y}_t)^2}{\left(\frac{N_tN_c}{N}\right)^2} \\
&= \frac{1}{N_c^2}\sum_{W_i:i=0}(Y_i-\bar{Y}_c)^2 + \frac{1}{N_t^2}\sum_{W_i:i=1}(Y_i-\bar{Y}_t)^2 \\
&= \frac{\tilde{s}_c^2}{N_c} + \frac{\tilde{s}_t^2}{N_t} \\
&= V^{neyman}
\end{aligned}
$$

(c) Homogeneous treatment effects mean that the treatment effect is the same for every individual. Then the variance of the outcome Y is only coming from the error term ε_i (and this variance is the same for every one). This means we have homoscedasticity.

Section B: Power Calculation

We can calculate the power of a t-test pretty fast with R's `power.t.test` function (we have to set $n = 400$ because the n we pass to the `power.t.test` function is the n for each group: sample size of 800 and 2 groups $\Rightarrow n = 400$):

```r
power.t.test(n = 400, delta = 0.207)
```

returns a power of 0.8324603. But there's a problem: We don't split our sample size in 2 *equal* groups. One group has 480 dyads (60% of 800), the other one has 320 dyads (40% of 800). We could make our life easy and do a linear approximation of the power function between $n = 320$ and $n = 480$:

```r
power.t.test(n = c(320,480), delta = 0.207)
```

gives a power of `c(0.7435918, 0.8931858)` so on average a power of: $\frac{0.7435918+0.8931858}{2} = 0.8183888$.

How are these numbers calculated? As the power of a test is the probability of rejecting H_0 when H_1 is true, we draw our sample from a t distribution with $2(n-1)$ degrees of freedom and a non-centrality-parameter of $\sqrt{\frac{n}{2}} \cdot \frac{\delta}{\sigma}$. We are interested that our realisation will fall in the rejection region of H_0 which is outside of $[t_{df,\alpha/2}, t_{df,1-\alpha/2}]$[1]. The critical values are

```r
qt(0.025, 798)
qt(0.975, 798)
```

which are -1.962941 and 1.962941. Then the probability of being outside of $[t_{df,\alpha/2}, t_{df,1-\alpha/2}]$ (when H_1 is true) is

$$
\begin{aligned}
\mathbb{P}(X \notin [t_{df,\alpha/2}, t_{df,1-\alpha/2}]) &= 1 - \mathbb{P}(X \in [t_{df,\alpha/2}, t_{df,1-\alpha/2}]) \\
&= 1 - [\mathbb{P}(X \leq, t_{df,1-\alpha/2}) - \mathbb{P}(X \leq t_{df,\alpha/2})]
\end{aligned}
$$

where $X \sim \mathcal{T}(df, ncp)$ and $ncp = \sqrt{\frac{n}{2}} \cdot \frac{\delta}{\sigma} = \sqrt{\frac{400}{2}} \cdot \frac{0.207\sigma}{\sigma} = 2.927422$

```r
ncp = sqrt(200)*0.207
1 - (pt(1.962941, df = 798, ncp = ncp) - pt(-1.962941, df = 798,
    ncp = ncp))
```

[1] As I learned stats in Germany I'm following the German notation which might be different that the notation taught here. With $t_{df,\alpha/2}$ I mean the $\alpha/2$ quantile of the t-distribution with df degrees of freedom.

returns 0.8324609.

But we can get more accurate: Following [1] the non-centrality-parameter for groups with an unequal size is

$$\begin{aligned} ncp &= \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\ &= \frac{0.207\sigma}{\sigma\sqrt{\frac{1}{320}+\frac{1}{480}}} \\ &= 2.868276 \end{aligned}$$

This leads then to

```r
ncp = 0.207/sqrt(1/320 + 1/480)
1 - (pt(1.962941, df = 798, ncp = ncp) - pt(-1.962941, df = 798, ncp = ncp))
```

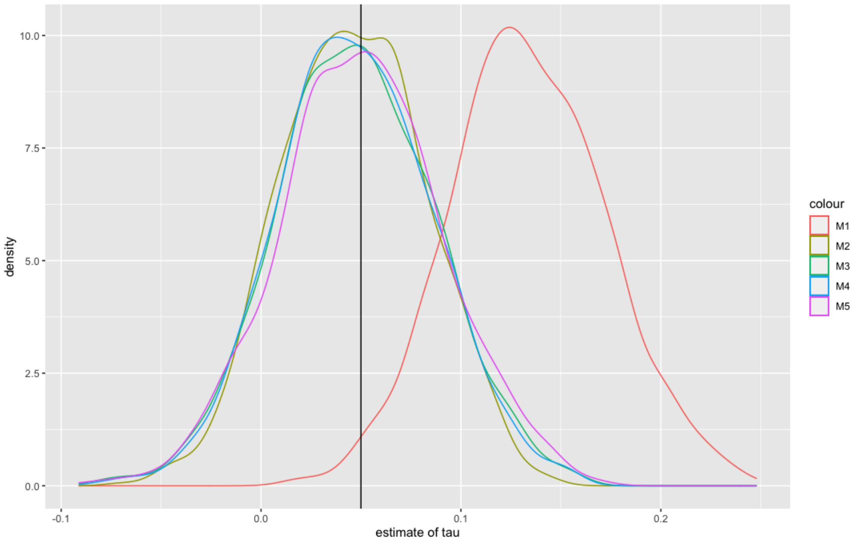
which is 0.8172306.

Section C: Simulation Exercise

The 5 model specifications are:

```r
# Model 1
m1 = lm(Y ~ W)

# Model 2
m2 = lm(Y ~ W + age + female + age*female) # this is the same as m2 = lm(Y ~ W + age*female)

# Model 3
agematrix = matrix(0, nrow = observations, ncol = 65-20+1)
for (i in 1:observations) {
  currentAge = age[i]
  agematrix[i, currentAge - 20 + 1] = 1
}
m3 = lm(Y ~ W + agematrix + 0) # could have achieved the same with declaring age as a factor

# Model 4
logitModel = glm(W ~ age, family = binomial(logit))
propScore = predict.glm(logitModel, newdata = as.list(age), type = "response")
lambda = 1/(propScore^W * (1 - propScore)^(1-W))
m4 = lm(Y ~ W, weights = lambda)

# Model 5
logitModel2 = glm(W ~ agematrix + 0, family = binomial(logit))
propScore2 = predict.glm(logitModel2, newdata = as.list(age), type = "response")
lambda2 = 1/(propScore2^W * (1 - propScore2)^(1-W))
m5 = lm(Y ~ W, weights = lambda2)
```

You can see the whole R code in the attached file. The kernel-density plot is



where we can see:

- Model 1 has a omitted variable bias since we didn't include `female` and `age`. We can check from what variable the OVB comes with `cor.test()`:

  ```
  cor.test(W, female)
  cor.test(W, age)
  ```

  The first test results in a p-value of 0.2356; the second test has a p-value of $2.314 \cdot 10^{-8}$. This means in the first test we can't reject the Null hypothesis but we reject it in the second test. So there is a significant correlation between `W` and `age` but not between `W` and `female`. The OVB comes from `age`.

- Model 2 is the same specification then the data generation process so we except no OVB and see that there is no OVB.

- Model 3 doesn't show any OVB either but this is not a surprise after we've seen in the analysis of model 1 that `female` has no significant correlation with `W`. So removing `female` from the model and using a saturated model with `age` can capture all the variance from `female` (which comes from `age` by construction) and we see the same KDE for $\hat{\tau}$ as model 2.

- Model 4 has the same specification than model 1 so we would expect a OVB on `age`. But in the plot we don't see a OVB. According to [2] page 620 the method of using inverse probability weights adjusts for selection bias.

- The same is valid for model 5.

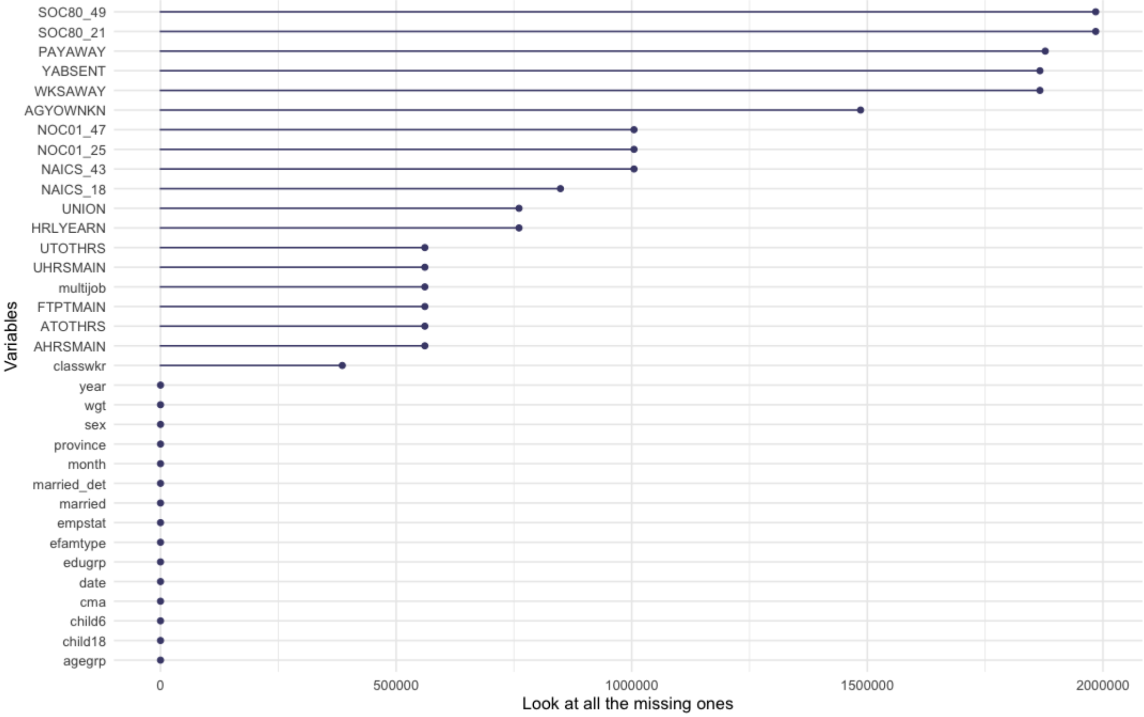
Section D: Differences-in-differences

Before going to the tasks let's have a look on the missing data:

```
library(haven)
library(naniar)
library(ggplot2)

data = read_dta("lfs_2010_2019_ages1564_20per.dta")
```

```
gg_miss_var(data) + labs(y = "Look at all the missing ones")
```

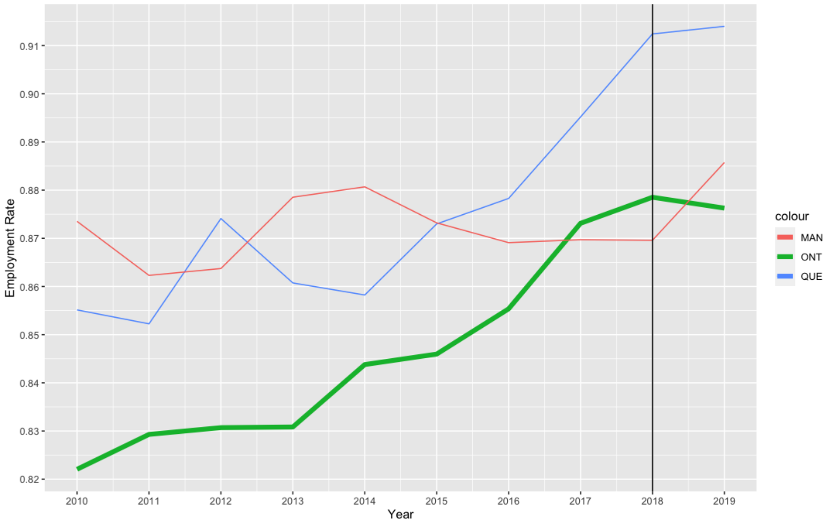
We get the following plot:



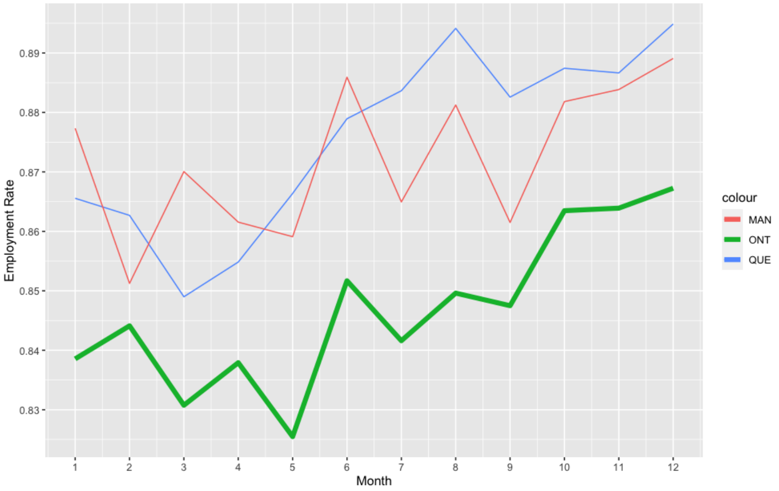
I don't want to deal with missing data and luckily I don't need to when fitting the models in (c).

(a) From the lecture notes *Like controlled experiments, natural experiments create the conditions of a treated and control group. BUT without random assignment, these groups are not directly comparably.* The policy change divides the population of Canada into a treatment and control group but there might be spillover effect since people can move from one state to another state.

(b) The plot for the yearly employment rate is



We can see a parallel trend for QUE and ONT, but not for MAN and ONT. Let's have a look on the monthly employment rate:



There is some seasonality in the employment rate, during spring is less employment but it increases till the end of the year (maybe because employers need helping hands to handle the Christmas business?). We can see parallel trends between MAN and ONT after May and some weak parallel trends between QUE and ONT the whole year.

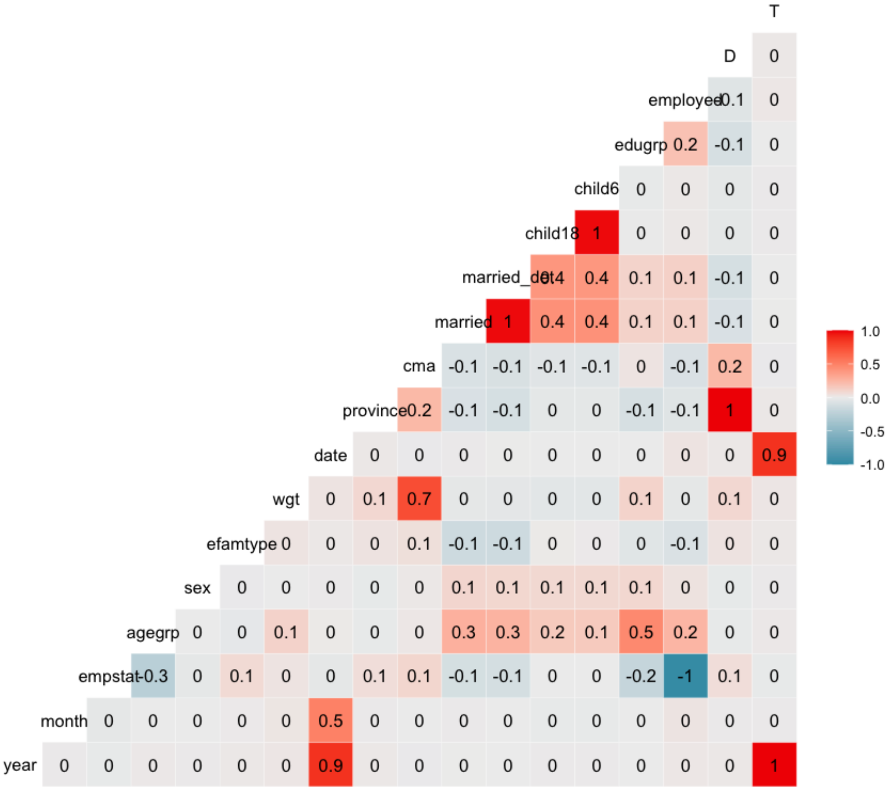
(c) Since I'll use the yearly data in the following tasks I use QUE as control group. Looking at some basic population statistics for states in Canada on Wikipedia QUE has more inhabitants than MAN and a similar population density as ONT. This means the employment rate is more stable during the years. After filtering the data and storing it in a data frame `df` we can specify the models (as usual look at the attached file for the whole R code):

```r
# Model 1
m1 = lm(employed ~ D + T + D*T, data = df)

# Model 2
monthmatrix = matrix(0, nrow = nrow(df), ncol = 12)
for (i in 1:nrow(df)) {
  currentMonth = as.integer(df$month[i])
  monthmatrix[i, currentMonth] = 1
}
monthmatrix = monthmatrix[,-1] # base month = January
m2 = lm(employed ~ D + T + D*T + monthmatrix, data = df)
```

Before we come to the results let's think about good covariates. Good covariates correlate with the outcome variable but don't correlate with each other (no collinearity).

```r
library(GGally)

# remove non-full columns
fullColumns = c()
for (colname in colnames(df)) {
  values = df[colname] %>% na.omit() %>% nrow()
  if (values == df[colname] %>% nrow()) {
  fullColumns = c(fullColumns, colname)
}
}
df = df[fullColumns]
ggcorr(df, label = TRUE)
```

Gives the following picture



If we look at the column with `employed` we can see that the following variables have a non-zero correlation with `employed`:

- `edugrp`: including in model 3
- `married`: including
- `married_det`: not including because high correlation with `married`
- `cma`: not including because `cma` is the nearby Census Metropolitan Areas but it correlates with `province` which determines (= correlates with) `D`
- `province`: not including
- `efamtype`: including
- `agegrp`: including
- `empstat`: perfect correlation, not including, would lead to collinearity

This leads to the specification for model 3

```
m3 = lm(employed ~ D + T + D*T + monthmatrix + edugrp +
    married + efamtype + agegrp, data = df)
```

The results are

	Dependent variable:		
	employed		
	(1)	(2)	(3)
D	-0.039^{***}	-0.039^{***}	-0.021^{***}
	(0.008)	(0.008)	(0.008)
T	0.047^{***}	0.046^{***}	0.044^{***}
	(0.010)	(0.010)	(0.009)
monthmatrix1		-0.031^{**}	-0.029^{**}
		(0.014)	(0.014)
monthmatrix2		0.010	0.009
		(0.014)	(0.014)
monthmatrix3		-0.014	-0.016
		(0.014)	(0.014)
monthmatrix4		0.042^{***}	0.041^{***}
		(0.014)	(0.014)
monthmatrix5		0.104^{***}	0.106^{***}
		(0.014)	(0.014)
monthmatrix6		0.142^{***}	0.143^{***}
		(0.014)	(0.014)
monthmatrix7		0.101^{***}	0.101^{***}
		(0.014)	(0.014)

monthmatrix8		0.015	0.015
		(0.014)	(0.014)
monthmatrix9		0.030^{**}	0.028^{**}
		(0.014)	(0.014)
monthmatrix10		0.022	0.019
		(0.014)	(0.014)
monthmatrix11		0.018	0.023^{*}
		(0.014)	(0.014)
edugrp			0.073^{***}
			(0.005)
married			0.073^{***}
			(0.010)
efamtype			-0.006^{***}
			(0.001)
agegrp			0.192^{***}
			(0.007)
D:T	-0.046^{***}	-0.046^{***}	-0.045^{***}
	(0.012)	(0.012)	(0.012)
Constant	0.586^{***}	0.551^{***}	0.177^{***}
	(0.007)	(0.012)	(0.015)
Observations	28,647	28,647	28,647
R^2	0.004	0.014	0.089
Adjusted R^2	0.004	0.014	0.089
Residual Std. Error	0.494	0.492	0.473
F Statistic	41.232^{***}	30.059^{***}	156.306^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01	

We see that the treatment effect (coefficient for `D:T`) is significant, negative and nearly the same in all three models. The added covariates in model 3 increase the R^2 (although it is still extremely low) and we can see the seasonality effects in summer (`monthmatrix4` is for May etc.).

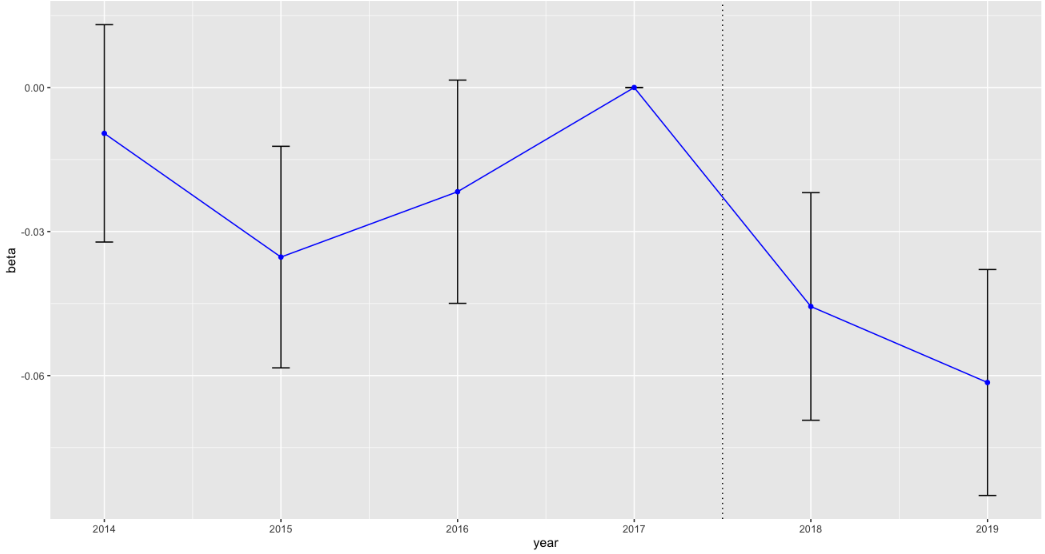
(d) Specification for model 4:

```r
timematrix = matrix(0, nrow = nrow(df), ncol = 6)
colnames(timematrix) = c("2014", "2015", "2016", "2017", "2018", "2019")
for (i in 1:nrow(df)) {
  currentYear = as.integer(df$year[i])
  timematrix[i, currentYear - 2014 + 1] = 1
}
timematrix = timematrix[,-4] # base year = 2017
monthmatrix = matrix(0, nrow = nrow(df), ncol = 12)
for (i in 1:nrow(df)) {
  currentMonth = as.integer(df$month[i])
  monthmatrix[i, currentMonth] = 1
}
monthmatrix = monthmatrix[,-1] # base month = January
m4 = lm(employed ~ D + timematrix + monthmatrix + timematrix*D, data = df)
```

We can collect the estimates and standard errors with `coef(m4)` and `coef(summary(m4))`. The the 95%-CI is (assuming a normal distribution because we have lots of observations ⇒ high degrees of freedom for t-distribution)

$$CI = \hat{\beta} + z_{0.975} \cdot SE(\hat{\beta})$$

Plotting these gives us



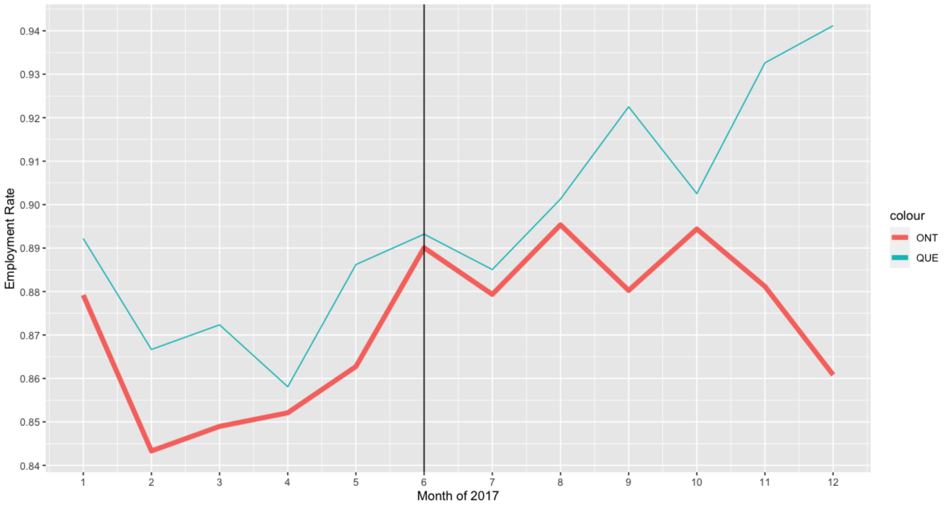
For our parallel trends assumption we want $\hat{\beta}_{2014} = \hat{\beta}_{2015} = \hat{\beta}_{2016} = 0$ which we can get for 2014 and 2016 but not for 2015. We can rule out pre-emptive behaviour because 2015 no one could have known of this policy (maybe some members of government). At least 2018 and 2019 we have clear negative treatment effects.

(e) We need to make the additional assumption that we have parallel trends on monthly trends.

(f) The best approach would be to do whole section again but this time averaging employment rate not on a yearly basis but on a monthly basis giving us coefficients $\beta_{2014,Jan}$, $\beta_{2014,Feb}$, ..., $\beta_{2019,Dez}$ and then testing for $\beta_{2017,Jun} = ... = \beta_{2017,Dec} = 0$. I'll go another approach by just plotting the employment rate of the year 2017. If there is pre-emptive behaviour we should see it:

```r
employmentRateMonthly2017 = data.frame(cbind(1:12, rep(0, 12), rep(0, 12)))
colnames(employmentRateMonthly2017) = c("Month", "QUE", "ONT")
for (prov in 5:6) {
  empRateForMonth = c()
  for (m in 1:12) {
    employed = data %>% filter(province == prov) %>% filter(month == m) %>% filter(year == 2017) %>% filter(between(agegrp, 1, 2)) %>% filter(between(empstat, 1, 2)) %>% nrow()
    population = data %>% filter(province == prov) %>% filter(month == m) %>% filter(year == 2017) %>% filter(between(agegrp, 1, 2)) %>% filter(between(empstat, 1, 3)) %>% nrow()
    empRateForMonth = c(empRateForMonth, employed/population)
  }
  if (prov == 5) employmentRateMonthly2017$QUE = empRateForMonth
  if (prov == 6) employmentRateMonthly2017$ONT = empRateForMonth
}
ggplot(employmentRateMonthly2017, aes(x = Month)) +
  geom_line(aes(y = QUE, color = "QUE")) +
  geom_line(aes(y = ONT, color = "ONT"), size = 2) +
  ylab("Employment Rate") +
  xlab("Month of 2017") +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  scale_y_continuous(breaks = seq(0.1, 1, by = 0.01)) +
  geom_vline(xintercept = 6)
```

gives us



The trends look pretty much the same but they differ when it comes to the end of the year. Instead of going up as seen in (b) for Christmas business the employment rate in ONT goes down. Maybe this is has nothing to do with the upcoming policy and Christmas business isn't booming as usual but it can also be pre-emptive behaviour.

(g) There might be bigger changes in the labour market in ONT than just the minimum wage raise. If there is than older people (older people have most likely worked for many years and therefore earn more than minimum wage) would be affected to. We can test this by looking only at ONT and using as control group people older than 24. We could go one step further and not use age as a separator but `HRLYEARN` and investigate what's happening to the employment rate of low-earners vs. high-earners. Problem here is that 25% of our dataset has no information of `HRLYEARN` (see plot of missing values at the start of this task).

(h) Literature suggests that $\beta_{2018} = \beta_{2019} = 0$ which is not the case for my estimators, the 95%-CI doesn't include 0. Since I only look at data from ONT maybe other Ontario-specific effects have influenced my result and we should use more data from other parts of the world too.
Interestingly after playing around with the data I found an interesting R command: `step()` which does a backward model selection. I start with every covariate included (and excluding covariates with perfect collinearity) and step by step covariates get removed to make the AIC smaller:

```
step(lm(employed ~ . -year - date - province + D:T, data = df)
    )
```

leads to the following model:

```
lm(formula = employed ~ month + empstat + agegrp + sex +
    efamtype + cma + married + married_det + child6 + D + T,
    data = df)
```

which doesn't include the treatment effect `D:T`. With other words: The employment status can't be enough explained by minimum wage raise or $\beta = 0$.

Section E: Synthetic Control

References

[1] Peter Dalgaard. “Power and the computation of sample size”. In: *Introductory Statistics with R*. New York, NY: Springer New York, 2008, pp. 155–162. ISBN: 978-0-387-79054-1. DOI: 10.1007/978-0-387-79054-1_9. URL: https://doi.org/10.1007/978-0-387-79054-1_9.

[2] Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. “A Structural Approach to Selection Bias”. In: *Epidemiology* 15.5 (2004), pp. 615–625. ISSN: 10443983. URL: http://www.jstor.org/stable/20485961 (visited on 11/12/2022).