MY457: Problem Set 3 - Difference in Differences

Pedro Torres-Lopez, Michael Ganslmeier, Daniel de Kadt

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 5pm on Sat/17/May. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a study of the effect of a treatment $D_i \in 0, 1$ on Y_i for all $i \in 1, ..., N$. In this case, treatment occurs across two dimensions: *i)* treatment group $G_i \in 0, 1$, and *ii)* time $t \in 0, 1$.

1.1. In this setting we can denote the following potential outcomes:

- $Y_{it}(0)$: potential outcome for unit i in period t when untreated
- $Y_{it}(1)$: potential outcome for unit i in period t when treated

Write out the realisations of these potential outcomes as observed data. Which are observed, when, and for which groups?

In this case the realisation of each potential outcome would be written as Y_{it}. Now, we know that $Y_{it} = Y_{it}(0)$ if in period t individual is not in the treated group. On the other hand, $Y_{it} = Y_{it}(1)$ if individual i is in the treated group in the post period, and $Y_{it}(0)$ if in the pre-period.

Note that in this setting, before getting treatment, all individuals will have $Y_{it}(0)$. It is not until treatment has been given that we observe $Y__{it}(1) for the treated group.

1.2 What is the main assumption in a canonical two-period difference-in-differences setting? Explain how violations of this assumption can impact the validity of the estimated treatment effect.

The main assumption in the canonical Diff-in-Diff setting is the parallel trends assumption. We assume that if no treatment happened the outcomes between treated and control would trend in parallel.

When this assumption is violated, the estimate of our treatment effect is going to be biased. We would incorrectly assume a counterfactual trend (one that does not accurately capture what would have happened without treatment), and therefore get a point estimate that is not correct.

1.3 Given repeated cross-sectional data, we can estimate a canonical two-period difference-in-differences design with the following regression specification:

$$Y_i = \hat{\alpha} + \hat{\gamma} G_i + \hat{\delta} T_i + \hat{\tau}(G_i \times T_i) + \hat{\varepsilon}_i$$

Explain the parameter (estimand) that each coefficient in the specification estimates.

$\hat{\alpha}$ is the mean for the control group in the pre-treatment period. $\hat{\alpha} + \hat{\gamma}$ the same for the treated group. $\hat{\delta}$ is the change in the mean for the control group in the post-treatment period compared to the pre-treatment period. $\hat{\tau}$ is the change in the the mean for the treatment group in the post-treatment period compared to the pre-treatment period, hence the treatment effect.

2 Simulations

In this question we will use simulated data to test some of our intuitions about difference-in-differences. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Explain the code below and relate it to a difference-in-differences data generating process. What kind of data (panel or repeated-cross sectional) is this?

```
set.seed(123)

n_units <- 1000

tau  <-  25000

G = rbinom(n_units, 1, 0.5)
for (i in 1:2) {
  data <- tibble(
    ID = 1:n_units,
    G  = G,
    T = ifelse(i == 2, 1, 0)
  )

  if (i == 1) {
    sim_data <- data
  } else {
    sim_data <- rbind(sim_data, data)
  }
}

Y0 <- rnorm(n_units, 50000, 2500)

data <- sim_data %>% mutate(
  Y0 = c(Y0, Y0*(1+1/10)),
  Y0 = ifelse(G == 1, Y0 + 10000, Y0),
  Y1 = Y0 + tau,
  Y = ifelse(G == 1 & T == 1, Y1, Y0)
)
```

We are creating a data frame with a 100 units. We randomly assign units into treatment and control groups (G) and specify two periods, one pre-treatment and one post-treatment (T).

We create a continuous potential outcome under control which is normally distributed. In the post-period, we increase this by 10%. We define our treatment effect to be 25,000 and add it

to the potential outcome under treatment.

Lastly, we define our realized outcome based on treatment group and treatment period (G and T).

2.2 Without using a regression, estimate the canonical two-period difference-in-differences using only `Y`, `G`, and `t`. What do you find?

Period	Control	Treatment	Diff
Pre	50025.7147832091	60034.0672387196	10008.3524555105
Post	55028.28626153	90037.4739625916	35009.1877010616
Diff	5002.57147832091	30003.406723872	25000.8352455511

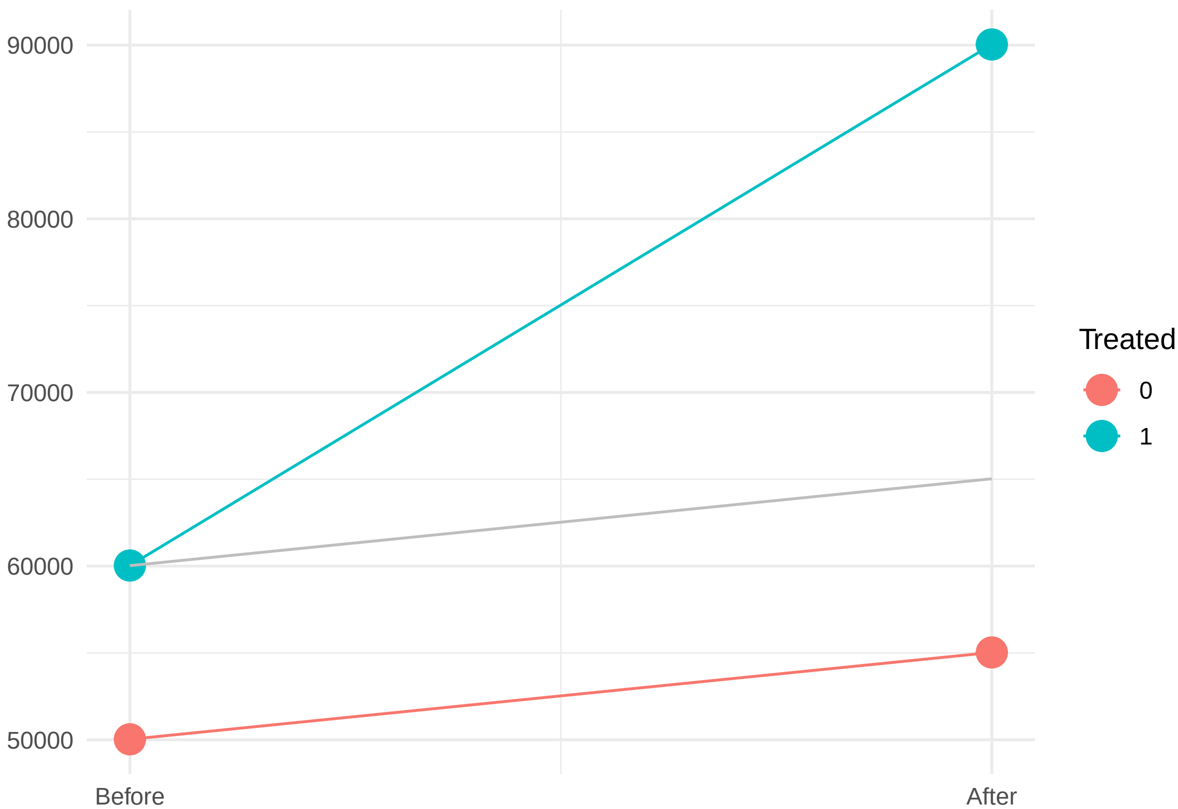
When estimating the canonical Diff-in-Diff, we find an estimated treatment effect of 25,000.935.

2.3 Now estimate the difference-in-differences design using linear regression. Do you find any differences to your previous estimation? Why or why not?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50025.715	116.9478	427.76116	0
G	10008.352	166.5592	60.08888	0
T	5002.571	165.3891	30.24728	0
G:T	25000.835	235.5502	106.13803	0

We find the same estimated treatment effect.

2.4 Using the potential outcomes in our simulated data, create a plot visualizing the difference-in-differences estimator.



2.5 Now consider a new data generating process, given by the simulation code below. Explain how this code is different to the code in question 2.1.

```r
set.seed(123)

n_obs <- 1000

n_periods <- 20

tau_values <- c(1000, 3000, 3000, 2000, 5000, 3000, 9000, 6000, 7000, 10000,
                9000, 8000, 6000, 3000, 7000, 2000, 5000, 2000, 1000)

tau <- setNames(tau_values, paste0("tau_", 1:19))

G = rbinom(n_obs, 1, 0.5)

for (i in 1:20) {
  treated_units <- ifelse(i > 5, sample(1:n_obs, size = floor(1/40*n_obs)), NA)
  if (i == 1) {
    treated <- treated_units
  } else {
    treated <- c(treated, treated_units)
  }

  data <- tibble(
      ID = 1:n_obs,
      G  = G,
      P  = i,
      T  = ifelse(ID %in% treated, 1, 0)
    )

  if (i == 1) {
    sim_data <- data
  } else {
    sim_data <- rbind(sim_data, data)
  }
}

Y0 <- rnorm(n_obs, 50000, 2500)

sim_data <- sim_data %>%
  mutate(
    Y0 = (1 + P/10) * Y0 + if_else(G == 1, 10000, 0),
    Y1 = case_when(
      P %in% 1:19 ~ Y0 + tau[paste0("tau_", P)],
      TRUE ~ Y0
    ),
    Y = if_else(G == 1 & T == 1, Y1, Y0),
    D = T * G
  )

data <- sim_data
```

Now we allow for staggered treatment adoption. We allow for different individuals to start receiving treatment at different periods. With this, we also change the magnitude of treatment depending on when an individual is treated – that is, we build in heterogeneity in treatment effects over time.

2.6 Using the new simulated data, estimate the difference-in-differences design using a two-way fixed effects linear regression. You can do this in multiple ways: using `lm` and `factor()`, using `lm` on de-meaned data, using `plm` with `model = "within"` and `effect = "twoways"`, or using `fixest`.

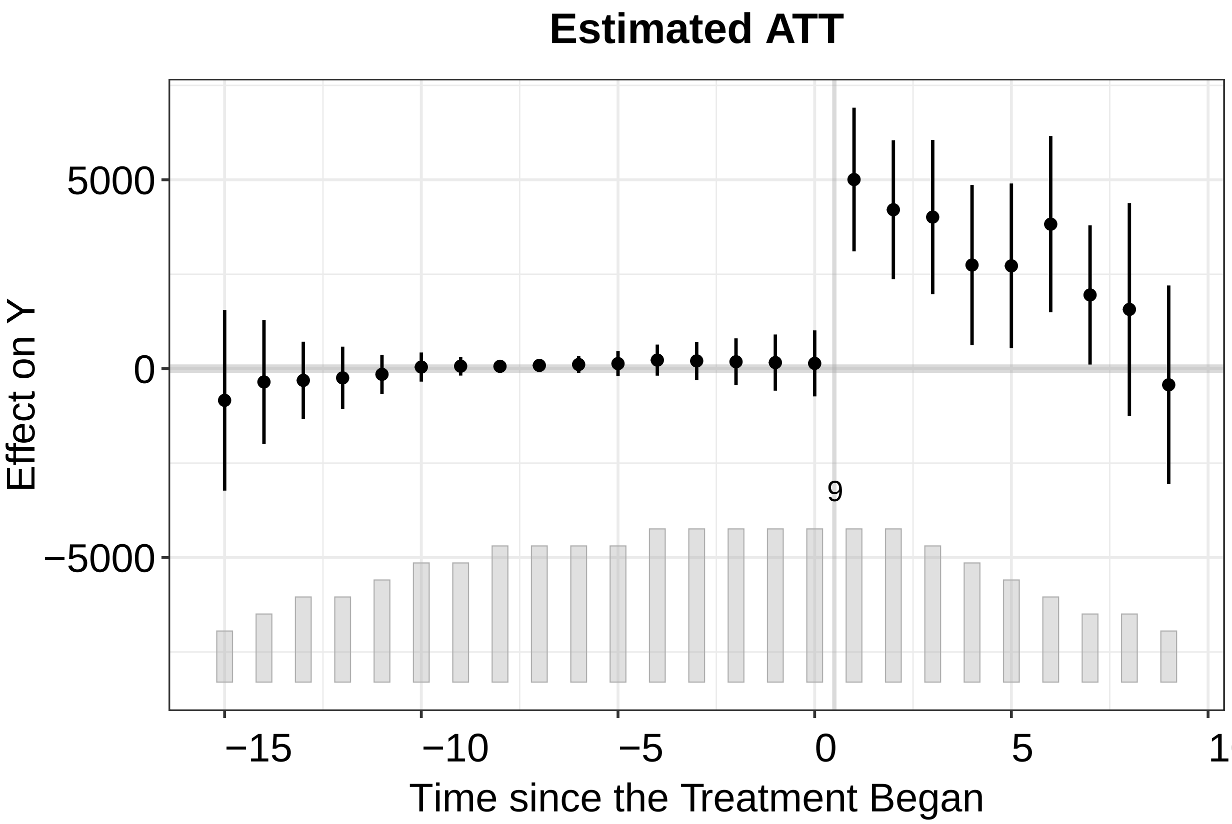
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59394.61461	336.71674	176.3934131	0.0000000
D	2731.50390	256.74008	10.6391798	0.0000000
factor(P)2	4994.71515	66.71237	74.8694028	0.0000000
factor(P)3	9989.43030	66.71237	149.7388057	0.0000000
factor(P)4	14984.14545	66.71237	224.6082085	0.0000000
factor(P)5	19978.86060	66.71237	299.4776114	0.0000000
factor(P)6	24973.84425	66.71286	374.3482668	0.0000000
factor(P)7	29974.55940	66.71286	449.3070529	0.0000000
factor(P)8	34966.27455	66.71286	524.1309324	0.0000000
factor(P)9	39961.98970	66.71286	599.0147705	0.0000000
factor(P)10	44966.97334	66.71434	674.0225660	0.0000000
factor(P)11	49959.68849	66.71434	748.8597726	0.0000000
factor(P)12	54957.67214	66.71681	823.7454521	0.0000000
factor(P)13	59949.65579	66.72027	898.5223542	0.0000000
factor(P)14	64932.37094	66.72027	973.2030323	0.0000000
factor(P)15	69947.35458	66.72472	1048.2975009	0.0000000
factor(P)16	74916.33823	66.73015	1122.6759959	0.0000000
factor(P)17	79931.32188	66.73657	1197.7139492	0.0000000
factor(P)18	84904.30552	66.74398	1272.0893751	0.0000000
factor(P)19	89889.28917	66.75237	1346.6081617	0.0000000
factor(P)20	94875.00432	66.75237	1421.2978692	0.0000000
factor(ID)2	5215.43287	471.72768	11.0560247	0.0000000
factor(ID)3	-8013.44791	471.72768	-16.9874449	0.0000000
factor(ID)4	1877.12836	471.72768	3.9792627	0.0000694
factor(ID)5	5615.67968	471.72768	11.9044948	0.0000000
factor(ID)6	-757.35521	471.72768	-1.6054924	0.1084021
factor(ID)7	13987.95647	471.72768	29.6526092	0.0000000
factor(ID)8	13979.08280	471.72768	29.6337982	0.0000000
factor(ID)9	12923.45786	471.72768	27.3960136	0.0000000
factor(ID)10	-9423.58307	471.72768	-19.9767440	0.0000000
factor(ID)11	8796.20796	471.72768	18.6467922	0.0000000
factor(ID)12	-6852.53873	471.72768	-14.5264716	0.0000000
factor(ID)13	3486.93589	471.72768	7.3918408	0.0000000
factor(ID)14	6694.32979	471.72768	14.1910897	0.0000000
factor(ID)15	3315.74940	471.72768	7.0289482	0.0000000
factor(ID)16	-934.39071	471.72768	-1.9807841	0.0476299
factor(ID)17	6653.56157	471.72768	14.1046665	0.0000000
factor(ID)18	-5180.86270	471.72768	-10.9827406	0.0000000
factor(ID)19	-7029.63779	471.72768	-14.9018981	0.0000000
factor(ID)20	9975.36089	471.72768	21.1464397	0.0000000
factor(ID)21	4108.49049	471.72768	8.7094540	0.0000000
factor(ID)22	11227.88343	471.72768	23.8016211	0.0000000
factor(ID)23	5862.48666	471.72768	12.4276928	0.0000000
factor(ID)24	7121.18887	471.72768	15.0959742	0.0000000
factor(ID)25	5631.61449	471.72768	11.9382745	0.0000000
factor(ID)26	9172.61337	471.72768	19.4447216	0.0000000
factor(ID)27	4911.21286	471.72768	10.4111187	0.0000000
factor(ID)28	2515.79629	471.72768	5.3331539	0.0000001
factor(ID)29	-1115.65673	471.72768	-2.3650440	0.0180379
factor(ID)30	-10459.90498	471.72768	-22.1736088	0.0000000
factor(ID)31	9931.90387	471.72768	21.0543166	0.0000000
factor(ID)32	8714.00229	471.72768	18.4725271	0.0000000
factor(ID)33	-398.96934	471.72768	-0.8457620	0.3976962
factor(ID)34	9073.48859	471.72768	19.2345902	0.0000000
factor(ID)35	-8848.07694	471.72768	-18.7567475	0.0000000
factor(ID)36	1277.23620	471.72768	2.7075710	0.0067838
factor(ID)37	61.12372	471.72768	0.1295742	0.8969047
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

	Estimate	Std. Error	t-value	Pr(>|t|)
D	2731.504	256.7401	10.63918	0

	x
D	2731.504

2.7 Using the new data and either the *fect* package or the *did* package, estimate dynamic period-specific ATTs and provide an event study plot. What do you find?

```
## Call:
## fect.formula(formula = Y ~ D, data = sim_data, index = c("ID",
##     "P"), force = "two-way", method = "fe", se = TRUE, nboots = 200)
##
## ATT:
##                             ATT  S.E. CI.lower CI.upper    p.value
## Tr obs equally weighted    2661 624.6     1437     3885 0.00002037
## Tr units equally weighted 2603 596.0     1435     3771 0.00001253
```



The estimated treatment effect is a bit lower than what we estimated in section 2.6, a bit closer to the truth. This has to do with the fact that in 2.6 the ATT is an average effect over the rolling units, since different units adopt treatment at different times, the TWFE estimator might be somewhat biased.

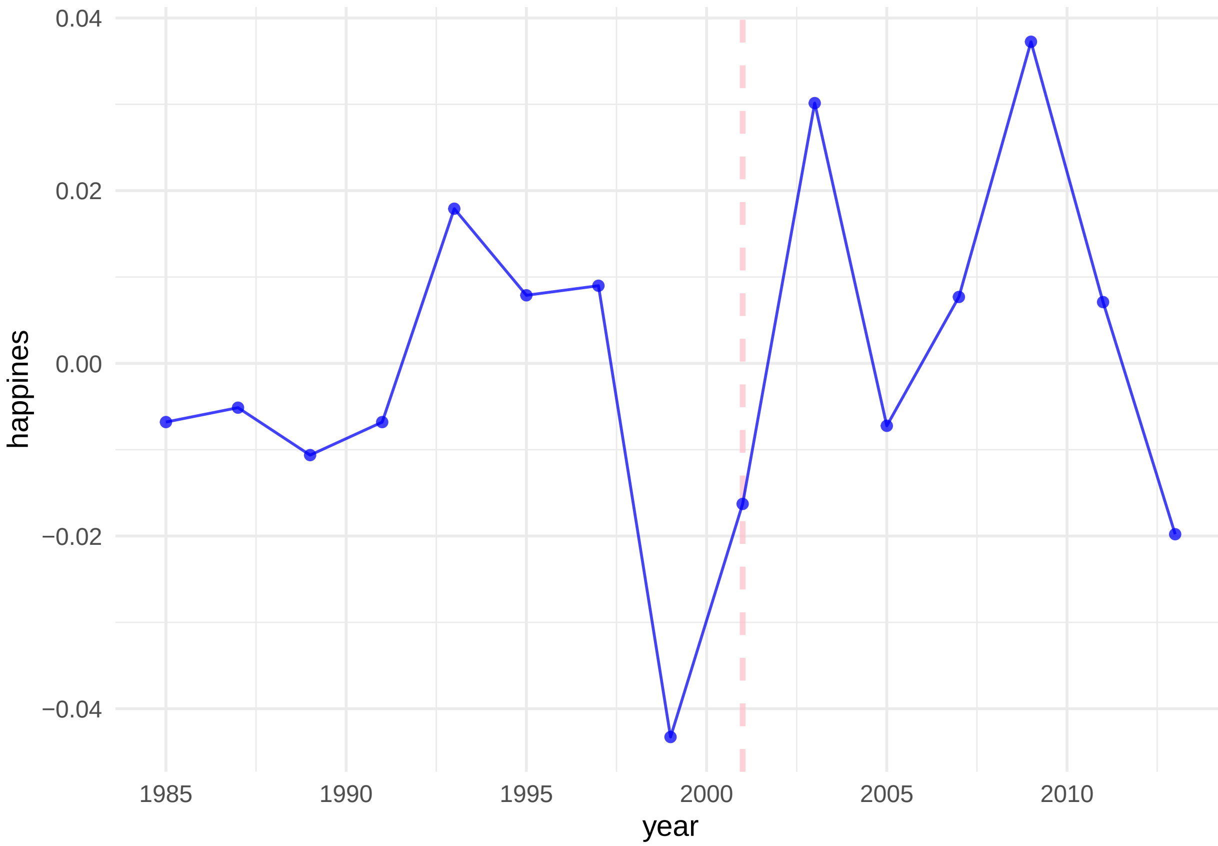
3 Replication

In this section, we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment.*

In recent decades, there has been an increasing push towards higher transparency in income, wealth, and earnings. Transparency facilitates comparisons between individuals. In 2001, Norwegian tax records became accessible online allowing individuals to have access to these easily, assuming they had access to internet.

The author uses this setting to analyze the effect of salary transparency on the subjective well-being of individuals across the income distribution.

3.1 Read into `R` the replication data set (`Norway-MSD.dta`) and visualise the trend in Norwegian happiness (`po_happy`) over the years. Include a vertical line to indicate when treatment came into effect.



3.2 Explain, simply and in your own words, the causal inference problem faced by the authors (i.e., what confounding are they concerned about?). Then explain, simply and in your own words, the author's research design and how it mitigates the problems identified.

The authors want to analyse whether making the tax records of individuals easily accessible had an effect on the well-being of individuals. They specifically want to focus on the gap between income ranks in happiness measures.

To do this, they rely on a quasi-experiment, where in 2001, the Norwegian tax records became easily accessible online. They employ a Diff-in-Diff design (among others) to show that people with higher incomes report on being happier compared to people with lower incomes after this tax records became easily accessible.

3.3 In what way is the author's design a difference-in-differences, and how does it differ from the cases we have typically seen in the lecture? Do you have any potential concerns about the plausibility of the underlying assumptions? You might benefit from reading section II of the paper closely.

It is a Diff-in-Diff design since we have individuals with different incomes G and a pre and post period T. However, in this case, treatment is not binary and is rather continuous, reflecting higher income when the value is higher.

3.4 Estimate the baseline specification as given in equation (1) in the paper. In addition to the difference-in-differences components, the regression should include a dummy variable for each year, and should control for marital status, education, household size, household workers, female, age and age squared. Hint: remember to include categorical variables as `factors()` where appropriate.

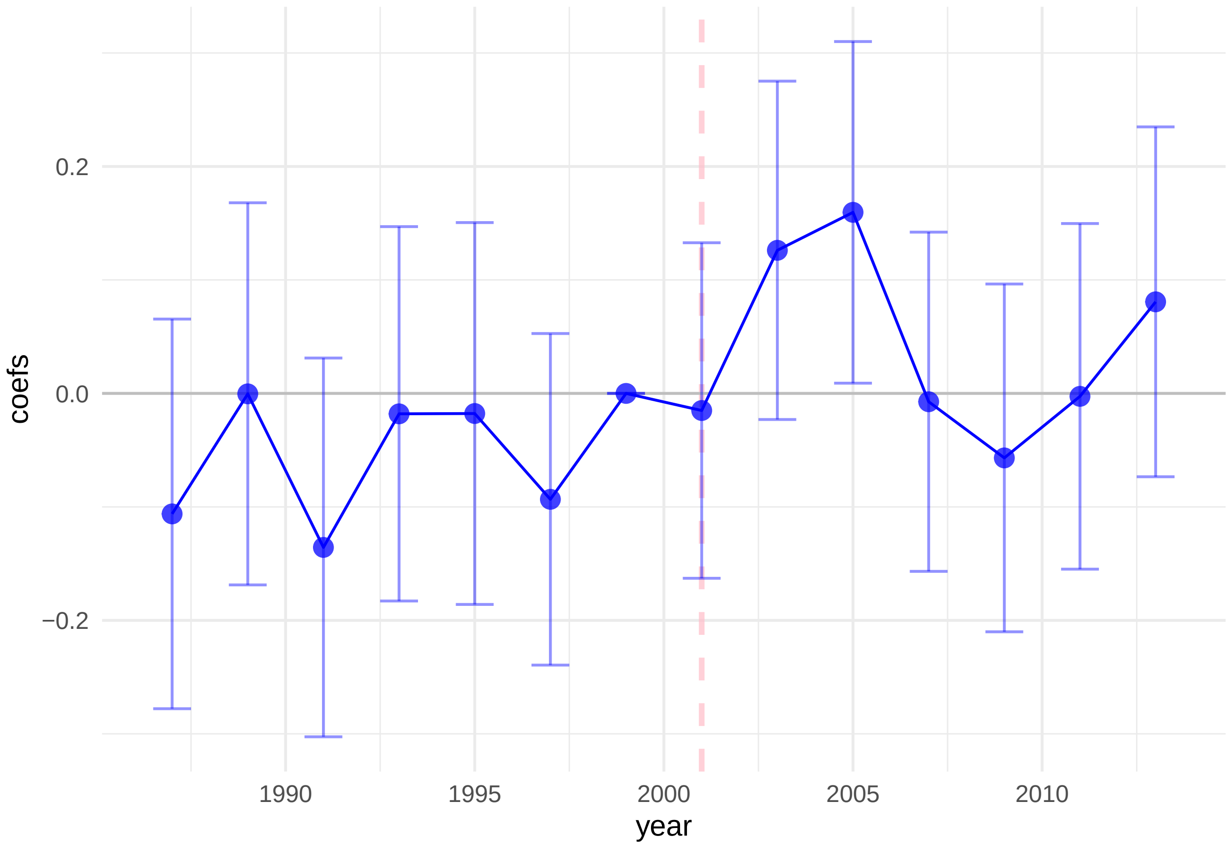
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0287713	0.0328310	-0.8763473	0.3808456
imp_hh_rank	0.3108112	0.0270175	11.5040890	0.0000000
post_2001	-0.0701098	0.0351527	-1.9944347	0.0461102
factor(year)1987	-0.0098852	0.0292977	-0.3374044	0.7358136
factor(year)1989	-0.0086877	0.0289631	-0.2999588	0.7642098
factor(year)1991	-0.0518002	0.0328077	-1.5789039	0.1143646
factor(year)1993	-0.0105482	0.0328346	-0.3212515	0.7480212
factor(year)1995	-0.0135428	0.0331743	-0.4082319	0.6831053
factor(year)1997	-0.0003249	0.0311054	-0.0104440	0.9916671
factor(year)1999	-0.0455749	0.0313510	-1.4536975	0.1460367
factor(year)2001	0.0219451	0.0233033	0.9417139	0.3463438
factor(year)2003	0.0657877	0.0234240	2.8085580	0.0049784
factor(year)2005	0.0422628	0.0235964	1.7910693	0.0732884
factor(year)2007	0.0729919	0.0234565	3.1117934	0.0018606
factor(year)2009	0.0647433	0.0239121	2.7075569	0.0067804
factor(year)2011	0.0411850	0.0225570	1.8258156	0.0678842
factor(marital_status)2	-0.1185882	0.0151073	-7.8497265	0.0000000
factor(marital_status)3	-0.5255272	0.0197736	-26.5772565	0.0000000
factor(marital_status)4	-0.4446700	0.0201541	-22.0634511	0.0000000
factor(marital_status)5	-0.3596434	0.0292096	-12.3125004	0.0000000
factor(education)2	-0.0218209	0.0198973	-1.0966781	0.2727876
factor(education)3	0.0025993	0.0187681	0.1384954	0.8898495
factor(education)4	0.0294296	0.0193499	1.5209166	0.1282873
factor(education)5	0.0800133	0.0324779	2.4636253	0.0137574
factor(hh_size)2	-0.0604507	0.0184597	-3.2747410	0.0010583
factor(hh_size)3	-0.1070978	0.0205572	-5.2097419	0.0000002
factor(hh_size)4	-0.1352898	0.0223149	-6.0627627	0.0000000
factor(hh_size)5	-0.0761222	0.0261616	-2.9096912	0.0036195
factor(hh_size)6	-0.0609511	0.0438796	-1.3890545	0.1648226
factor(hh_size)7	-0.1295010	0.0898495	-1.4413097	0.1495036
factor(hh_workers)1	0.0424444	0.0174297	2.4351743	0.0148882
factor(hh_workers)2	0.0878256	0.0188040	4.6705731	0.0000030
factor(hh_workers)3	0.0403690	0.0275326	1.4662227	0.1425941
factor(hh_workers)4	0.0906973	0.0483622	1.8753769	0.0607469
factor(hh_workers)5	-0.0379070	0.1071292	-0.3538437	0.7234575
female	0.0984352	0.0089958	10.9423318	0.0000000
poly(age, 2)1	-30.3955611	1.5224937	-19.9643262	0.0000000
poly(age, 2)2	27.6063336	1.2121694	22.7743191	0.0000000
imp_hh_rank:post_2001	0.0897253	0.0314065	2.8569058	0.0042797

3.5 Estimate the same specification, but separately on two different subgroups in the data. First estimate the effect for those who have high access to internet, then for those who do not. Do you find any differences? What do you conclude from this exercise?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2733015	0.1255861	2.1762090	0.0295492
imp_hh_rank	0.2800640	0.0405323	6.9096544	0.0000000
post_2001	-0.1503790	0.1058500	-1.4206804	0.1554225
factor(year)1987	-0.0001510	0.0410329	-0.0036795	0.9970642
factor(year)1989	-0.0119116	0.0406172	-0.2932648	0.7693222
factor(year)1991	-0.1019043	0.1037739	-0.9819836	0.3261177
factor(year)1993	-0.0209050	0.1039130	-0.2011784	0.8405608
factor(year)1995	-0.0140621	0.1038587	-0.1353961	0.8922998
factor(year)1997	-0.0004058	0.1029775	-0.0039411	0.9968555
factor(year)1999	-0.0670409	0.1030058	-0.6508461	0.5151520
factor(year)2001	-0.0080487	0.0323433	-0.2488531	0.8034765
factor(year)2003	0.0827094	0.0325453	2.5413588	0.0110484
factor(year)2005	0.0235391	0.0327333	0.7191176	0.4720754
factor(year)2007	0.0638032	0.0324607	1.9655525	0.0493617
factor(year)2009	0.0168963	0.0330631	0.5110323	0.6093331
factor(year)2011	0.0436867	0.0317555	1.3757198	0.1689211
factor(marital_status)2	-0.1250942	0.0198685	-6.2960940	0.0000000
factor(marital_status)3	-0.5890247	0.0291787	-20.1868129	0.0000000
factor(marital_status)4	-0.4488229	0.0393736	-11.3990689	0.0000000
factor(marital_status)5	-0.3663070	0.1279757	-2.8623176	0.0042091
factor(education)2	-0.2540096	0.1173551	-2.1644519	0.0304394
factor(education)3	-0.1837640	0.1164916	-1.5774872	0.1146965
factor(education)4	-0.1638767	0.1169829	-1.4008602	0.1612687
factor(education)5	-0.1554284	0.1290225	-1.2046607	0.2283461
factor(hh_size)2	-0.1035725	0.0362430	-2.8577223	0.0042706
factor(hh_size)3	-0.1375580	0.0362543	-3.7942560	0.0001484
factor(hh_size)4	-0.1823835	0.0371067	-4.9151062	0.0000009
factor(hh_size)5	-0.1144005	0.0394115	-2.9027179	0.0037027
factor(hh_size)6	-0.0936666	0.0533378	-1.7561028	0.0790835
factor(hh_size)7	-0.6758218	0.2069484	-3.2656538	0.0010936
factor(hh_workers)1	0.0668731	0.0969255	0.6899429	0.4902367
factor(hh_workers)2	0.1261134	0.0963368	1.3090892	0.1905165
factor(hh_workers)3	0.0875970	0.0990779	0.8841224	0.3766389
factor(hh_workers)4	0.1573078	0.1071165	1.4685667	0.1419633
factor(hh_workers)5	-0.0263394	0.1430801	-0.1840882	0.8539458
female	0.0653311	0.0131242	4.9779158	0.0000006
poly(age, 2)1	-32.6517319	1.4630655	-22.3173414	0.0000000
poly(age, 2)2	10.1267923	1.2758511	7.9372837	0.0000000
imp_hh_rank:post_2001	0.2166007	0.0516744	4.1916411	0.0000278

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1423225	0.0436693	-3.2590942	0.0011192
imp_hh_rank	0.3368182	0.0415337	8.1095111	0.0000000
post_2001	-0.0583974	0.0470543	-1.2410659	0.2145934
factor(year)1987	-0.0255565	0.0418790	-0.6102464	0.5417043
factor(year)1989	-0.0148224	0.0414182	-0.3578708	0.7204431
factor(year)1991	-0.0442600	0.0440642	-1.0044456	0.3151739
factor(year)1993	-0.0399125	0.0442129	-0.9027326	0.3666768
factor(year)1995	-0.0510114	0.0446378	-1.1427848	0.2531392
factor(year)1997	-0.0378289	0.0416246	-0.9088117	0.3634586
factor(year)1999	-0.0634111	0.0419976	-1.5098741	0.1310886
factor(year)2001	0.0310937	0.0346957	0.8961841	0.3701634
factor(year)2003	0.0273706	0.0348889	0.7845071	0.4327503
factor(year)2005	0.0430337	0.0349734	1.2304713	0.2185326
factor(year)2007	0.0683365	0.0346075	1.9746141	0.0483233
factor(year)2009	0.0979001	0.0353640	2.7683529	0.0056383
factor(year)2011	0.0369758	0.0321067	1.1516525	0.2494753
factor(marital_status)2	-0.1166376	0.0248215	-4.6990644	0.0000026
factor(marital_status)3	-0.4514937	0.0308809	-14.6204816	0.0000000
factor(marital_status)4	-0.4132895	0.0271761	-15.2078407	0.0000000
factor(marital_status)5	-0.3303903	0.0343595	-9.6156988	0.0000000
factor(education)2	-0.0051198	0.0215520	-0.2375570	0.8122267
factor(education)3	-0.0044230	0.0211470	-0.2091536	0.8343301
factor(education)4	0.0388998	0.0240481	1.6175790	0.1057664
factor(education)5	0.1073623	0.0405543	2.6473720	0.0081173
factor(hh_size)2	-0.0105905	0.0246032	-0.4304519	0.6668708
factor(hh_size)3	-0.0806508	0.0302410	-2.6669393	0.0076596
factor(hh_size)4	-0.0667344	0.0359465	-1.8564928	0.0633955
factor(hh_size)5	-0.0575172	0.0574786	-1.0006724	0.3169952
factor(hh_size)6	-0.3074327	0.2043159	-1.5046927	0.1324162
factor(hh_size)7	0.0579644	0.1019486	0.5685643	0.5696571
factor(hh_workers)1	0.0528043	0.0186232	2.8353977	0.0045807
factor(hh_workers)2	0.0924111	0.0231395	3.9936581	0.0000653
factor(hh_workers)3	0.0463531	0.0600011	0.7725381	0.4398033
factor(hh_workers)4	-0.1566388	0.1792052	-0.8740752	0.3820860
factor(hh_workers)5	1.8865660	0.9704498	1.9440119	0.0519056
female	0.1262843	0.0134912	9.3604633	0.0000000
poly(age, 2)1	-12.8716786	1.5853241	-8.1192727	0.0000000
poly(age, 2)2	19.0490686	1.1322162	16.8245861	0.0000000
imp_hh_rank:post_2001	-0.0073654	0.0516588	-0.1425783	0.8866244

3.6 Test for parallel pre-trends using the event study design. What do you find?



None of the coefficients in the pre-treatement period are statistically significant as shown by the confidence intervals. This suggests that the parallel trends assumption holds.

3.7 (Extra credit): What do you think of the research design used in this paper? Do you have any suggestions for how it could have been improved, or extra falsification tests the author could have tried?

The authors find a clever way to estimate the Diff-in-Diff. However, it is not clear exactly who is in the treatment group and who is in the control group. Another problem they face is their measure of income. They do not observe direct income but rather impute the rank based on a binned question. Binned questions tend to be biased since they normally have a reference number to which individuals tend to agglomerate. Another porblem is that the definition of these bins has changed over the years as they mention i page 1036.

A possible solution would be to define clear treatment and control groups, for example, people above the 5th decile are treated and the rest are control. Another solution to the second problem could be to directly impute income from other surveys.

4 Appendix

```r
# you can include your libraries here:
library(tidyverse)
library(knitr)
library(haven)
library(fect)
library(plm)
library(fixest)
```

```r
# and any other options in R:
options(scipen=999)

# 2 -----
## 2.2
(mean(data$Y[data$G == 1 & data$T == 1]) - mean(data$Y[data$G == 1 & data$T == 0])) - (mean(data$Y[data$

## 2.3
lm(Y~G*T, data = data) %>% summary()

## 2.4
data %>% group_by(G, T) %>%
    summarise(Y = mean(Y)) %>%
    ggplot() + aes(x = T, y = Y, color = factor(G)) +
    geom_point(size = 5) +
    geom_line() +
    geom_line(data = . %>% filter(G == 0), aes(y = Y + 10000), color = "grey") +
    labs(color = "Treated") +
    scale_x_continuous(breaks = c(0 , 1), labels = c("Before", "After")) +
    theme_minimal() +
    xlab("") +
    ylab("")

## 2.6
### OLS (Dummy)
lm(Y~D + factor(P) + factor(ID), data = data) %>% summary()

### Fixed Effects (De-Meaned)
plm(Y ~ D, data = sim_data,
           index = c("ID",  "P"),
           model = "within", effect = "twoways") %>% summary()

feols(Y ~ D | ID + P, data = data) %>% summary()

## 2.7
out.fect <- fect(Y~D, data = sim_data, index = c("ID", "P"),
  method = "fe", force = "two-way", se = TRUE, nboots = 200)

print(out.fect)

plot(out.fect)

# 3 ----
df <- read_dta("./Norway-MSD.dta")

## 3.1
df %>% group_by(year) %>%
  summarise(
    happines = mean(po_happy, na.rm = T)
  ) %>%
  ggplot() + aes(x = year, y = happines) +
  geom_vline(xintercept = 2001, color = "pink", linetype='dashed', alpha = 0.75, size = 1) +
  geom_point(color = "blue", alpha = 0.75) +
```

```
  geom_line(color = "blue", alpha = 0.75) +
  theme_minimal()

## 3.4
lm(po_happy~imp_hh_rank*post_2001+factor(year)+factor(marital_status)+
            factor(education)+factor(hh_size)+factor(hh_workers)+
            female+poly(age,2), data = df) %>% summary()

## 3.5
lm(po_happy~imp_hh_rank*post_2001+factor(year)+factor(marital_status)+
            factor(education)+factor(hh_size)+factor(hh_workers)+
            female+poly(age,2), data = filter(df, higher_internet == 1)) %>% summary()

lm(po_happy~imp_hh_rank*post_2001+factor(year)+factor(marital_status)+
            factor(education)+factor(hh_size)+factor(hh_workers)+
            female+poly(age,2), data = filter(df, higher_internet == 0)) %>% summary()


## 3.6
df <- df %>% mutate(
  year = factor(year, levels = c(1999, 1987, 1989, 1991, 1993, 1995, 1997, 2001, 2003, 2005, 2007, 2009
)
reg1 <- lm(po_happy~imp_hh_rank*factor(year)+factor(marital_status)+
            factor(education)+factor(hh_size)+factor(hh_workers)+
            female+poly(age,2), data = df)
coefs <- coef(reg1)[38:50]
upper <- confint.lm(reg1, vcov. = vcovHC(reg1, type = "HC0"))[38:50, 2]
lower <-  confint.lm(reg1, vcov. = vcovHC(reg1, type = "HC0"))[38:50, 1]
estudy <- data.frame(
  cbind(
    coefs, lower, upper
  )
) %>% mutate(
    year = c(1987, 1989, 1991, 1993, 1995, 1997, 2001, 2003, 2005, 2007, 2009, 2011, 2013)
)

new_row <- data.frame(
  coefs = 0,
  lower = 0,
  upper = 0,
  year = 1999
)

estudy <- rbind(estudy, new_row) %>% data.frame()

estudy %>% ggplot() + aes(x = year, y = coefs) +
  geom_vline(xintercept = 2001, color = "pink", linetype='dashed', alpha = 0.75, size = 1) +
  geom_hline(yintercept = 0, color = "grey") +
  geom_point(color = "blue", size = 3, alpha = 0.75) +
  geom_line(color = "blue") +
  geom_errorbar(aes(ymin = lower, ymax = upper), color = "blue", width = 0) +
  theme_minimal()
```