MY457/MY557: Causal Inference for Experimental and Observational Studies

Class 3: Difference in Differences

```
# read in required packages
# install.packages("plm") # Uncomment and run (once) if not already installed
# ... similarly for the other packages if needed

library(dplyr)
library(ggplot2)
library(knitr)
library(markdown)
library(plm)
```

1. In-class exercise: Examples of different types of analysis

In this part of the exercise we show examples of how basic difference-in-differences estimators and more general fixed effects estimators that were discussed in the lecture in week 5 can be implemented in R. This is done using a single simulated dataset, for demonstration purposes. The comments on each of the methods are fairly limited, so some of the R steps may seem a little mysterious. Please refer to the help files of the functions, and ask me during the computer class.

```
# Data (actually the same sim_data that will be generated again later in the class exercise)
sim_data <- readRDS("simdata1.rds")
print(sim_data,n=30) # Showing what the data look like
```

```
## # A tibble: 5,000 x 17
##       id     t     g    t1    t2    t3    t4    t5    t6    t7    t8    t9   t10
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     1     1     0     1     0     0     0     0     0     0     0     0     0
##  2     1     2     0     0     1     0     0     0     0     0     0     0     0
##  3     1     3     0     0     0     1     0     0     0     0     0     0     0
##  4     1     4     0     0     0     0     1     0     0     0     0     0     0
##  5     1     5     0     0     0     0     0     1     0     0     0     0     0
##  6     1     6     0     0     0     0     0     0     1     0     0     0     0
##  7     1     7     0     0     0     0     0     0     0     1     0     0     0
##  8     1     8     0     0     0     0     0     0     0     0     1     0     0
##  9     1     9     0     0     0     0     0     0     0     0     0     1     0
## 10     1    10     0     0     0     0     0     0     0     0     0     0     1
## 11     2     1     0     1     0     0     0     0     0     0     0     0     0
## 12     2     2     0     0     1     0     0     0     0     0     0     0     0
## 13     2     3     0     0     0     1     0     0     0     0     0     0     0
## 14     2     4     0     0     0     0     1     0     0     0     0     0     0
## 15     2     5     0     0     0     0     0     1     0     0     0     0     0
```

```
## 16     2     6     0     0     0     0     0     0     1     0     0     0     0
## 17     2     7     0     0     0     0     0     0     0     1     0     0     0
## 18     2     8     0     0     0     0     0     0     0     0     1     0     0
## 19     2     9     0     0     0     0     0     0     0     0     0     1     0
## 20     2    10     0     0     0     0     0     0     0     0     0     0     1
## 21     3     1     1     1     0     0     0     0     0     0     0     0     0
## 22     3     2     1     0     1     0     0     0     0     0     0     0     0
## 23     3     3     1     0     0     1     0     0     0     0     0     0     0
## 24     3     4     1     0     0     0     1     0     0     0     0     0     0
## 25     3     5     1     0     0     0     0     1     0     0     0     0     0
## 26     3     6     1     0     0     0     0     0     1     0     0     0     0
## 27     3     7     1     0     0     0     0     0     0     1     0     0     0
## 28     3     8     1     0     0     0     0     0     0     0     1     0     0
## 29     3     9     1     0     0     0     0     0     0     0     0     1     0
## 30     3    10     1     0     0     0     0     0     0     0     0     0     1
## # i 4,970 more rows
## # i 4 more variables: y0 <dbl>, y1 <dbl>, d1 <dbl>, y <dbl>
```

Let us first consider the simple difference-in-differences setting where we have observations in two periods, before and after the intervention for some units.

Consider first the estimation formulated as difference-in-differences, first estimated explicitly using differences of means and then, equivalently, implemented using linear regression modelling a few different ways.

Note: These estimators do not formally require actual panel (longitudinal) data for the same units, but can be calculated even when we have separate samples of units from the same population in the two periods. But then we need in effect to further assume that the composition of units in that population has not changed in any relevant way between the periods.

```
##############################################################################################################
# First illustration: Suppose we had only observed data just before and after the intervention (periods
sdata2 <- sim_data[sim_data$t==7 | sim_data$t==8,]
print(sdata2,n=20) # Here observations3, 5, 9 and 10 received the treatment between periods 7 and 8
```

```
## # A tibble: 1,000 x 17
##       id     t     g    t1    t2    t3    t4    t5    t6    t7    t8    t9   t10
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     1     7     0     0     0     0     0     0     0     1     0     0     0
##  2     1     8     0     0     0     0     0     0     0     0     1     0     0
##  3     2     7     0     0     0     0     0     0     0     1     0     0     0
##  4     2     8     0     0     0     0     0     0     0     0     1     0     0
##  5     3     7     1     0     0     0     0     0     0     1     0     0     0
##  6     3     8     1     0     0     0     0     0     0     0     1     0     0
##  7     4     7     0     0     0     0     0     0     0     1     0     0     0
##  8     4     8     0     0     0     0     0     0     0     0     1     0     0
##  9     5     7     1     0     0     0     0     0     0     1     0     0     0
## 10     5     8     1     0     0     0     0     0     0     0     1     0     0
## 11     6     7     0     0     0     0     0     0     0     1     0     0     0
## 12     6     8     0     0     0     0     0     0     0     0     1     0     0
## 13     7     7     1     0     0     0     0     0     0     1     0     0     0
## 14     7     8     1     0     0     0     0     0     0     0     1     0     0
## 15     8     7     0     0     0     0     0     0     0     1     0     0     0
## 16     8     8     0     0     0     0     0     0     0     0     1     0     0
## 17     9     7     1     0     0     0     0     0     0     1     0     0     0
```

```
## 18     9     8     1     0     0     0     0     0     0     0     1     0     0
## 19    10     7     1     0     0     0     0     0     0     1     0     0     0
## 20    10     8     1     0     0     0     0     0     0     0     1     0     0
## # i 980 more rows
## # i 4 more variables: y0 <dbl>, y1 <dbl>, d1 <dbl>, y <dbl>
```

```
# Difference-in-differences estimator of the treatment effect

## Calculated explicitly as difference-in-differences (of means)

(mean(sdata2[sdata2$t==8 & sdata2$g==1,]$y)-mean(sdata2[sdata2$t==8 & sdata2$g==0,]$y))-
  (mean(sdata2[sdata2$t==7 & sdata2$g==1,]$y)-mean(sdata2[sdata2$t==7 & sdata2$g==0,]$y))
```

```
## [1] 8.144188
```

```
## More conveniently, calculated using different regerssion formulations:
###
sdata2 <- sdata2 %>% mutate(y_diff = y - dplyr::lag(y))
tail(sdata2)
```

```
## # A tibble: 6 x 18
##      id     t     g    t1    t2    t3    t4    t5    t6    t7    t8    t9   t10
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   498     7     0     0     0     0     0     0     0     1     0     0     0
## 2   498     8     0     0     0     0     0     0     0     0     1     0     0
## 3   499     7     1     0     0     0     0     0     0     1     0     0     0
## 4   499     8     1     0     0     0     0     0     0     0     1     0     0
## 5   500     7     0     0     0     0     0     0     0     1     0     0     0
## 6   500     8     0     0     0     0     0     0     0     0     1     0     0
## # i 5 more variables: y0 <dbl>, y1 <dbl>, d1 <dbl>, y <dbl>, y_diff <dbl>
```

```
summary(lm(y_diff~g,data=sdata2,subset=(t==8)))
```

```
##
## Call:
## lm(formula = y_diff ~ g, data = sdata2, subset = (t == 8))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1976 -0.9698  0.0060  1.0174  6.7440
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8993     0.1004   8.961   <2e-16 ***
## g             8.1442     0.1446  56.339   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.615 on 498 degrees of freedom
## Multiple R-squared:  0.8644, Adjusted R-squared:  0.8641
## F-statistic:  3174 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
###
summary(lm(y~g*t8,data=sdata2))
```

```
##
## Call:
## lm(formula = y ~ g * t8, data = sdata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6281 -0.7099 -0.0179  0.6612  4.3084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.08405    0.06994  86.995   <2e-16 ***
## g            4.85885    0.10073  48.235   <2e-16 ***
## t8           0.89931    0.09890   9.093   <2e-16 ***
## g:t8         8.14419    0.14246  57.169   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 996 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9594
## F-statistic:  7861 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
###
summary(lm(y~g+t8+d1,data=sdata2))
```

```
##
## Call:
## lm(formula = y ~ g + t8 + d1, data = sdata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6281 -0.7099 -0.0179  0.6612  4.3084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.08405    0.06994  86.995   <2e-16 ***
## g            4.85885    0.10073  48.235   <2e-16 ***
## t8           0.89931    0.09890   9.093   <2e-16 ***
## d1           8.14419    0.14246  57.169   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 996 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9594
## F-statistic:  7861 on 3 and 996 DF,  p-value: < 2.2e-16
```

Consider then the same estimation using a fixed-effects model with fixed effects for the individual units (and times). This does require panel data for (at least some) units.

```r
## Estimated using a fixed-effects regression model with fixed effects for the 500 individual units (an

### Explicitly as a linear model with dummy variables for the units
lm.fe.model <- lm(y~factor(id)+factor(t)+d1,data=sdata2)
lm.fe.model
```

```
##
## Call:
## lm(formula = y ~ factor(id) + factor(t) + d1, data = sdata2)
##
## Coefficients:
##   (Intercept)    factor(id)2    factor(id)3    factor(id)4    factor(id)5
##     6.1410343      0.6865779      6.3861151     -0.9190970      3.8995142
##   factor(id)6    factor(id)7    factor(id)8    factor(id)9   factor(id)10
##    -0.3616517      3.1944743      0.1416432      3.1970009      5.5841963
##  factor(id)11   factor(id)12   factor(id)13   factor(id)14   factor(id)15
##     0.2770122     -0.8963509      4.8422383      4.8188563      5.3125957
##  factor(id)16   factor(id)17   factor(id)18   factor(id)19   factor(id)20
##     4.6142962      0.4568294      4.7991584      0.9117876      5.8366171
##  factor(id)21   factor(id)22   factor(id)23   factor(id)24   factor(id)25
##    -0.6491809      0.6756303     -0.3056531     -1.3428265      5.2378175
##  factor(id)26   factor(id)27   factor(id)28   factor(id)29   factor(id)30
##     5.5934162      5.1165142      4.3383696     -0.5945491     -0.4750476
##  factor(id)31   factor(id)32   factor(id)33   factor(id)34   factor(id)35
##     4.8231690      4.8824513     -1.3297401     -0.4365491      4.6676371
##  factor(id)36   factor(id)37   factor(id)38   factor(id)39   factor(id)40
##     6.5062612     -1.3713513     -0.9496785      4.6812485     -0.3155455
##  factor(id)41   factor(id)42   factor(id)43   factor(id)44   factor(id)45
##    -0.7628557      4.1042903      3.2218208      5.7506425     -1.5639333
##  factor(id)46   factor(id)47   factor(id)48   factor(id)49   factor(id)50
##     6.1146259      0.7035256     -0.3159405      3.6514774      0.1790707
##  factor(id)51   factor(id)52   factor(id)53   factor(id)54   factor(id)55
##     0.3180039     -1.4963408      0.6726569      3.4283254      4.5293201
##  factor(id)56   factor(id)57   factor(id)58   factor(id)59   factor(id)60
##     5.9610489      5.7193639      5.4592963      0.4424120      4.2017468
##  factor(id)61   factor(id)62   factor(id)63   factor(id)64   factor(id)65
##     0.4938129      5.2653718      0.3808656      6.1823557      4.0106465
##  factor(id)66   factor(id)67   factor(id)68   factor(id)69   factor(id)70
##    -0.2855825     -0.1242681      1.2923448      6.3898079      5.4741015
##  factor(id)71   factor(id)72   factor(id)73   factor(id)74   factor(id)75
##     6.7798002      6.6453582      0.0001435      0.7775179      5.5451684
##  factor(id)76   factor(id)77   factor(id)78   factor(id)79   factor(id)80
##     5.7382296      3.6205797     -0.2472551      1.2063917      5.1895796
##  factor(id)81   factor(id)82   factor(id)83   factor(id)84   factor(id)85
##     5.4237982      6.6360347     -0.5513404     -0.7197424     -1.9415713
##  factor(id)86   factor(id)87   factor(id)88   factor(id)89   factor(id)90
##     0.1472769      3.5439306     -0.3511322      0.3649367     -0.0728301
##  factor(id)91   factor(id)92   factor(id)93   factor(id)94   factor(id)95
##    -0.1783271      0.4796860      4.5472779      0.4883812      4.2861848
##  factor(id)96   factor(id)97   factor(id)98   factor(id)99  factor(id)100
##     4.2298211      4.1369902      5.7050990      3.8294755      0.8748057
## factor(id)101  factor(id)102  factor(id)103  factor(id)104  factor(id)105
##     3.6137082      3.7833245      5.0904845      4.3807130     -0.1117879
```

```
## factor(id)106 factor(id)107 factor(id)108 factor(id)109 factor(id)110
##    -0.4646770     0.2332367    -0.5548987     4.0250301     0.7620231
## factor(id)111 factor(id)112 factor(id)113 factor(id)114 factor(id)115
##     4.9350033    -0.2507942     4.9338310     4.0597079     1.2762840
## factor(id)116 factor(id)117 factor(id)118 factor(id)119 factor(id)120
##    -0.3203931     5.7827740    -0.7914162     0.5103166    -0.0126200
## factor(id)121 factor(id)122 factor(id)123 factor(id)124 factor(id)125
##     5.2301398    -0.3171279     3.6568086     0.4754003     0.5270181
## factor(id)126 factor(id)127 factor(id)128 factor(id)129 factor(id)130
##     4.9757278    -0.6235884     5.2910462     4.8242474     4.3368139
## factor(id)131 factor(id)132 factor(id)133 factor(id)134 factor(id)135
##     3.1848106     0.5537936     4.2428279     5.4002518    -0.8196692
## factor(id)136 factor(id)137 factor(id)138 factor(id)139 factor(id)140
##    -0.5264871    -0.6195031     0.6934165     5.2291911    -0.0611892
## factor(id)141 factor(id)142 factor(id)143 factor(id)144 factor(id)145
##     0.7681749     4.9343967     5.3407316     5.6764278     5.0774270
## factor(id)146 factor(id)147 factor(id)148 factor(id)149 factor(id)150
##     4.1752105    -0.6055690    -1.3630817     0.7607637     5.7344323
## factor(id)151 factor(id)152 factor(id)153 factor(id)154 factor(id)155
##     4.4736972     4.8122547     4.8187070    -0.2539495     4.7398413
## factor(id)156 factor(id)157 factor(id)158 factor(id)159 factor(id)160
##     4.7881165     4.1752271     3.2375760     1.2352501     4.7920281
## factor(id)161 factor(id)162 factor(id)163 factor(id)164 factor(id)165
##     4.4289050     4.7415753     0.3982108     0.2791662     0.0881559
## factor(id)166 factor(id)167 factor(id)168 factor(id)169 factor(id)170
##    -0.3125047     4.3682290    -0.0914652     1.0525771     0.4662471
## factor(id)171 factor(id)172 factor(id)173 factor(id)174 factor(id)175
##    -0.7613431     0.5226099     0.2471822    -0.1450449    -0.5123341
## factor(id)176 factor(id)177 factor(id)178 factor(id)179 factor(id)180
##     5.2698899     4.6790500     4.8065967     0.7064342     0.5364565
## factor(id)181 factor(id)182 factor(id)183 factor(id)184 factor(id)185
##     0.0607706     5.5047349     4.7565496     6.5797542     4.5036784
## factor(id)186 factor(id)187 factor(id)188 factor(id)189 factor(id)190
##     3.5646500    -0.4936805     0.9003966    -0.3969665     4.9898909
## factor(id)191 factor(id)192 factor(id)193 factor(id)194 factor(id)195
##     5.7356009     4.2192577     0.2780132    -0.1259163     3.7226478
## factor(id)196 factor(id)197 factor(id)198 factor(id)199 factor(id)200
##     4.7071642     0.8788964     3.9801302     0.2530425     0.2543012
## factor(id)201 factor(id)202 factor(id)203 factor(id)204 factor(id)205
##     5.8706765     4.8287233     1.0303149     3.8428171     3.2513360
## factor(id)206 factor(id)207 factor(id)208 factor(id)209 factor(id)210
##     5.0064428    -0.2311760     4.7735742    -1.1133667     1.3023484
## factor(id)211 factor(id)212 factor(id)213 factor(id)214 factor(id)215
##     0.1502596     4.5495982     5.8045210    -0.8039800     4.5083976
## factor(id)216 factor(id)217 factor(id)218 factor(id)219 factor(id)220
##    -1.0611713    -0.2373064    -1.3248366    -0.4041306    -0.6701515
## factor(id)221 factor(id)222 factor(id)223 factor(id)224 factor(id)225
##    -0.9169103     0.4300695     0.1743148     0.4117602     7.2229553
## factor(id)226 factor(id)227 factor(id)228 factor(id)229 factor(id)230
##     0.6347518    -0.1589705     4.5588238     0.2100972     0.1860781
## factor(id)231 factor(id)232 factor(id)233 factor(id)234 factor(id)235
##     5.3709547     4.0114169     0.6622245     4.6632808    -0.2344624
## factor(id)236 factor(id)237 factor(id)238 factor(id)239 factor(id)240
##    -0.9120705     0.4150402    -0.0403125     0.7471047     4.0328004
```

```
## factor(id)241  factor(id)242  factor(id)243  factor(id)244  factor(id)245
##     0.1631267      3.9549158     -0.0960649     -0.7098693      4.4354944
## factor(id)246  factor(id)247  factor(id)248  factor(id)249  factor(id)250
##     4.2546621     -1.3519544     -0.2819352     -0.1149956      4.8913117
## factor(id)251  factor(id)252  factor(id)253  factor(id)254  factor(id)255
##     0.0089714      4.5469296     -0.2494297     -0.7876267     -0.2715924
## factor(id)256  factor(id)257  factor(id)258  factor(id)259  factor(id)260
##    -1.2009721      3.8499883     -0.0713356      0.9239970      4.1207641
## factor(id)261  factor(id)262  factor(id)263  factor(id)264  factor(id)265
##    -0.7219637      5.4974882      4.0493066     -0.4564286      0.1314508
## factor(id)266  factor(id)267  factor(id)268  factor(id)269  factor(id)270
##     5.6746545      4.8000033      1.7372715      5.2366938      0.2455515
## factor(id)271  factor(id)272  factor(id)273  factor(id)274  factor(id)275
##     0.0817639      3.9206669     -0.2618827      5.3681171     -0.9208282
## factor(id)276  factor(id)277  factor(id)278  factor(id)279  factor(id)280
##     4.1321626      4.4874864      4.4931962      4.7714303      4.7062144
## factor(id)281  factor(id)282  factor(id)283  factor(id)284  factor(id)285
##     3.8499904      1.3605045     -0.5841131      5.9559130      5.8966022
## factor(id)286  factor(id)287  factor(id)288  factor(id)289  factor(id)290
##     3.9559370      0.3495762      0.1905093     -0.4430695      5.7723746
## factor(id)291  factor(id)292  factor(id)293  factor(id)294  factor(id)295
##     5.7920882     -0.1111647      4.0014097     -0.1518526      6.3673478
## factor(id)296  factor(id)297  factor(id)298  factor(id)299  factor(id)300
##     0.0818729      0.3456592     -0.6346447      4.6168905      4.4563871
## factor(id)301  factor(id)302  factor(id)303  factor(id)304  factor(id)305
##     3.6308537      0.5750827      1.0356252     -0.4980329      3.8888483
## factor(id)306  factor(id)307  factor(id)308  factor(id)309  factor(id)310
##     6.2095610     -1.0909795     -0.7893003      0.1830482     -0.4325096
## factor(id)311  factor(id)312  factor(id)313  factor(id)314  factor(id)315
##     0.3948000      3.8133552     -0.3503441      5.6641806      4.0280288
## factor(id)316  factor(id)317  factor(id)318  factor(id)319  factor(id)320
##    -0.2278863      2.6470168      3.9793188     -0.7388619      4.4023703
## factor(id)321  factor(id)322  factor(id)323  factor(id)324  factor(id)325
##     0.6040638     -1.0811789      0.3610724     -0.6864090      5.6883747
## factor(id)326  factor(id)327  factor(id)328  factor(id)329  factor(id)330
##    -1.3145305      0.2516629      0.3222241      5.3309189      5.4603218
## factor(id)331  factor(id)332  factor(id)333  factor(id)334  factor(id)335
##     4.6653660      4.2545411      6.3563484      5.3820656      5.2423583
## factor(id)336  factor(id)337  factor(id)338  factor(id)339  factor(id)340
##     5.4412259      0.2528495      5.1204487     -0.5894344      5.7914754
## factor(id)341  factor(id)342  factor(id)343  factor(id)344  factor(id)345
##     5.4166374      0.4811847      4.8542195     -0.8030292      4.0952847
## factor(id)346  factor(id)347  factor(id)348  factor(id)349  factor(id)350
##     4.9533065      0.2533923      3.5144055     -1.6772375     -0.2035220
## factor(id)351  factor(id)352  factor(id)353  factor(id)354  factor(id)355
##     4.5101968      4.9770940      0.0455455      0.0335868      0.5995412
## factor(id)356  factor(id)357  factor(id)358  factor(id)359  factor(id)360
##    -0.1646965     -0.0057681      0.8878975      0.2362875      0.5034131
## factor(id)361  factor(id)362  factor(id)363  factor(id)364  factor(id)365
##    -0.2546354      4.8252310     -0.3636726      5.3348366      7.7271135
## factor(id)366  factor(id)367  factor(id)368  factor(id)369  factor(id)370
##     3.9107804      5.1162296      5.6036526      0.4989609     -0.2339864
## factor(id)371  factor(id)372  factor(id)373  factor(id)374  factor(id)375
##     3.5272120      3.1271735      4.5055219      0.6763260     -0.5604916
```

```
## factor(id)376  factor(id)377  factor(id)378  factor(id)379  factor(id)380
##    -0.4311821     -0.4563750      0.0307734      3.9246311     -0.9619203
## factor(id)381  factor(id)382  factor(id)383  factor(id)384  factor(id)385
##    -0.1752324     -0.8222454      4.8015299      5.2539224      5.7921687
## factor(id)386  factor(id)387  factor(id)388  factor(id)389  factor(id)390
##     4.2690218      0.0177621      0.4812496      4.4838140     -0.1841508
## factor(id)391  factor(id)392  factor(id)393  factor(id)394  factor(id)395
##     0.0404777      5.7498008     -0.8333256      1.2195586      1.4490793
## factor(id)396  factor(id)397  factor(id)398  factor(id)399  factor(id)400
##     0.0509663     -0.2265219     -0.1333167      1.0323648     -0.3794988
## factor(id)401  factor(id)402  factor(id)403  factor(id)404  factor(id)405
##     0.4954012      5.3092999     -0.5156582      0.2660335      4.5554803
## factor(id)406  factor(id)407  factor(id)408  factor(id)409  factor(id)410
##     2.9517395      4.1519406      5.0147432      5.2237917      6.4680996
## factor(id)411  factor(id)412  factor(id)413  factor(id)414  factor(id)415
##    -0.5816006     -0.2182264     -0.6866406      4.1083465      0.5193638
## factor(id)416  factor(id)417  factor(id)418  factor(id)419  factor(id)420
##     0.2133016      0.0674298      1.4028342      4.8279785     -0.4370642
## factor(id)421  factor(id)422  factor(id)423  factor(id)424  factor(id)425
##    -0.0083686      4.5967306      6.5607779      0.7060088      3.6019397
## factor(id)426  factor(id)427  factor(id)428  factor(id)429  factor(id)430
##     4.6128699      5.3037536      5.4441715      5.6836334      6.3963696
## factor(id)431  factor(id)432  factor(id)433  factor(id)434  factor(id)435
##     5.2457803      5.5871584      4.2279521     -0.4955982      0.9930488
## factor(id)436  factor(id)437  factor(id)438  factor(id)439  factor(id)440
##    -1.0714544      5.4353584      0.4118821     -0.4090712     -0.8909916
## factor(id)441  factor(id)442  factor(id)443  factor(id)444  factor(id)445
##     4.6638291      5.8375560     -0.6156941     -1.4713459      0.2066465
## factor(id)446  factor(id)447  factor(id)448  factor(id)449  factor(id)450
##     0.0694383      5.3213159      0.4021762      5.3154390      5.5812724
## factor(id)451  factor(id)452  factor(id)453  factor(id)454  factor(id)455
##     0.6328950      3.9273016     -0.5444056      5.1328128      3.9783218
## factor(id)456  factor(id)457  factor(id)458  factor(id)459  factor(id)460
##     3.5122653      3.7943274     -0.6578158      0.0283587     -0.1843274
## factor(id)461  factor(id)462  factor(id)463  factor(id)464  factor(id)465
##     4.5489921      0.3220907      5.5327761      4.8060407      4.2672178
## factor(id)466  factor(id)467  factor(id)468  factor(id)469  factor(id)470
##     4.7090549      4.5611588     -0.0330577     -0.0562265      0.3227091
## factor(id)471  factor(id)472  factor(id)473  factor(id)474  factor(id)475
##    -1.5440286     -0.3674916     -1.6008639      4.3350043      0.3803445
## factor(id)476  factor(id)477  factor(id)478  factor(id)479  factor(id)480
##     4.6217202      4.5978662      5.2545025      5.6523308      1.1330366
## factor(id)481  factor(id)482  factor(id)483  factor(id)484  factor(id)485
##     0.5358103      3.6565051      5.3389467      4.0477134      0.5400447
## factor(id)486  factor(id)487  factor(id)488  factor(id)489  factor(id)490
##     4.0746949      0.0002477     -0.3123528      1.0641016      4.7584822
## factor(id)491  factor(id)492  factor(id)493  factor(id)494  factor(id)495
##     5.7812968     -0.7231068      1.7726119     -0.0479134      3.5988722
## factor(id)496  factor(id)497  factor(id)498  factor(id)499  factor(id)500
##     6.2420475      2.6122757      0.1760690      4.7195430     -0.0612990
##    factor(t)8             d1
##     0.8993100      8.1441879
```

```
summary(lm.fe.model)$coefficients[499:502,] # Estimated unit fixed effects for two units, time effect a
```

```
##                   Estimate Std. Error     t value      Pr(>|t|)
## factor(id)499  4.71954300  1.1443624  4.12416807  4.359643e-05
## factor(id)500 -0.06129902  1.1420776 -0.05367325  9.572170e-01
## factor(t)8     0.89930998  0.1003600  8.96083917  6.474318e-18
## d1             8.14418789  0.1445564 56.33918690 3.395446e-218
```

```
### Using a dedicated function for fixed effects estimation (from the plm package).
### This deals with the unit fixed effects without actually
### having to estimate them (but can show them afterwards)
fe.model <- plm(y~factor(t)+d1,data=sdata2,index=c("id","t"),model="within",effect="individual")
summary(fe.model) # This displays only the estimated time and treatment effects
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = y ~ factor(t) + d1, data = sdata2, effect = "individual",
##     model = "within", index = c("id", "t"))
##
## Balanced Panel: n = 500, T = 2, N = 1000
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -3.3720e+00 -4.9835e-01  4.4409e-16  4.9835e-01  3.3720e+00
##
## Coefficients:
##            Estimate Std. Error t-value  Pr(>|t|)
## factor(t)8  0.89931    0.10036  8.9608 < 2.2e-16 ***
## d1          8.14419    0.14456 56.3392 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    10609
## Residual Sum of Squares: 649.56
## R-Squared:      0.93877
## Adj. R-Squared: 0.87718
## F-statistic: 3817.95 on 2 and 498 DF, p-value: < 2.22e-16
```

```
fixef(fe.model,effect="individual",type="dfirst")[498:499] # These are the unit fixed effects, with tha
```

```
##         499         500
##  4.71954300 -0.06129902
```

```
fixef(fe.model,effect="individual",type="dmean")[499:500] # These are the unit fixed effects, with thei
```

```
##       499       500
##  2.434559 -2.346283
```

The ideas and methods of fixed-effects estimation can also be used with more general structures of panel data (see the lecture for more on this). To illustrate this, let us use the dataset with all ten periods included:

```r
#############################################################################################################
# Second illustration: Using data on all 10 periods.
# Note: In this dataset the intervention still happens (if it does) only once, and always between perio
# However, the fixed effects model could also be fitted to datasets with other patterns of observation.

## Fixed effects model, with separate fixed effect for each time
fe10.model <- plm(y~factor(t)+d1,data=sim_data,index=c("id","t"),model="within",effect="individual")
summary(fe10.model) # This displays only the estimated time and treatment effects
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = y ~ factor(t) + d1, data = sim_data, effect = "individual",
##     model = "within", index = c("id", "t"))
##
## Balanced Panel: n = 500, T = 10, N = 5000
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.455803 -0.675807 -0.014423  0.653519  3.690947
##
## Coefficients:
##              Estimate Std. Error t-value  Pr(>|t|)
## factor(t)2   0.997565   0.066590  14.981 < 2.2e-16 ***
## factor(t)3   3.016404   0.066590  45.298 < 2.2e-16 ***
## factor(t)4  -1.933278   0.066590 -29.032 < 2.2e-16 ***
## factor(t)5  -2.924890   0.066590 -43.924 < 2.2e-16 ***
## factor(t)6   4.126240   0.066590  61.965 < 2.2e-16 ***
## factor(t)7   1.086470   0.066590  16.316 < 2.2e-16 ***
## factor(t)8   2.023986   0.073598  27.501 < 2.2e-16 ***
## factor(t)9  -2.960981   0.073598 -40.232 < 2.2e-16 ***
## factor(t)10  3.006585   0.073598  40.852 < 2.2e-16 ***
## d1           8.064923   0.065027 124.023 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    67608
## Residual Sum of Squares: 4977.4
## R-Squared:      0.92638
## Adj. R-Squared: 0.91803
## F-statistic: 5649.76 on 10 and 4490 DF, p-value: < 2.22e-16
```

```r
## Same, but estimated so that we can see the estimated fixed effects for each time
fe10B.model <- plm(y~d1,data=sim_data,index=c("id","t"),model="within",effect="twoways") # The "twoways
summary(fe10B.model)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = y ~ d1, data = sim_data, effect = "twoways", model = "within",
##     index = c("id", "t"))
##
## Balanced Panel: n = 500, T = 10, N = 5000
```

```
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.455803 -0.675807 -0.014423  0.653519  3.690947
##
## Coefficients:
##    Estimate Std. Error t-value  Pr(>|t|)
## d1 8.064923   0.065027  124.02 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    22029
## Residual Sum of Squares: 4977.4
## R-Squared:      0.77405
## Adj. R-Squared: 0.74844
## F-statistic: 15381.8 on 1 and 4490 DF, p-value: < 2.22e-16
```

```
fixef(fe10B.model, effect="time", type="dfirst") # Estimated time effects as differences to first perio
```

```
##        2        3        4        5        6        7        8        9
##  0.99756  3.01640 -1.93328 -2.92489  4.12624  1.08647  2.02399 -2.96098
##       10
##  3.00658
```

```
fixef(fe10B.model, effect="time", type="level")  # Estimated time effects for each period
```

```
##      1      2      3      4      5      6      7      8      9     10
## 4.9564 5.9540 7.9728 3.0231 2.0315 9.0826 6.0429 6.9804 1.9954 7.9630
```

```
## Same model,but fitted using lm
lm.fe10A.model <- lm(y~factor(id)+factor(t)+d1,data=sim_data)
summary(lm.fe10A.model)$coefficients[-(2:500),]
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  4.9564052 0.33606511  14.74835  4.143491e-48
## factor(t)2   0.9975646 0.06659007  14.98068  1.518660e-49
## factor(t)3   3.0164040 0.06659007  45.29811  0.000000e+00
## factor(t)4  -1.9332778 0.06659007 -29.03253 5.473581e-170
## factor(t)5  -2.9248896 0.06659007 -43.92381  0.000000e+00
## factor(t)6   4.1262403 0.06659007  61.96480  0.000000e+00
## factor(t)7   1.0864701 0.06659007  16.31580  3.493108e-58
## factor(t)8   2.0239858 0.07359779  27.50064 5.264811e-154
## factor(t)9  -2.9609811 0.07359779 -40.23193 1.646388e-302
## factor(t)10  3.0065846 0.07359779  40.85156 1.680397e-310
## d1           8.0649228 0.06502740 124.02345  0.000000e+00
```

```
## Instead of separate time effects for each period, we can also fit more parsimonious functions of tim
## Here linear and quadratic time effects.
## Note: This is just for illustration. We skip an examination of whether these smooth forms of time de

lm.fe10lin.model <- lm(y~factor(id)+t+d1,data=sim_data)
summary(lm.fe10lin.model)$coefficients[-(2:500),]
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 5.1274914  0.8732296  5.871871 4.619220e-09
## t           0.0859498  0.0162435  5.291336 1.271991e-07
## d1          7.7009639  0.1466470 52.513597 0.000000e+00
```

```
lm.fe10quad.model <- lm(y~factor(id)+t+I(t^2)+d1,data=sim_data)
summary(lm.fe10quad.model)$coefficients[-(2:500),]
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  5.77833503 0.876244526  6.594432 4.761537e-11
## t           -0.27529597 0.061358521 -4.486679 7.414779e-06
## I(t^2)       0.03470151 0.005685563  6.103442 1.125441e-09
## d1           7.36726634 0.155957130 47.239048 0.000000e+00
```

```
##############################################################################################################
```

2. Further in-class exercise and demonstration

Note: This part of the exercise is from 2021. I expect that we will not have time to go through it (or all of it) during the class.But it provides very useful additional information and demonstration, so I have left it here for your self-study.

Here we will use a set of simulated panel data with an exogenous intervention to illustrate the various ways that difference-in-difference estimates can be obtained. This is the same simulated dataset which was used in part 2 above. The initial exploration here give more information about those data, and the estimation then repeats some of the same methods results we demonstrated above.

Learning objectives:

- Understand the data structures and variable codings required to obtain a difference-in-difference estimate
- Understand why different estimators lead to similar or even identical estimates of the difference-in-differences results
- Understand how to interpret the difference-in-differences estimate
- Understand how to use lags and leads to explore the parallel trends assumption

First, we will load in some required packages:

Causal identification under selection on *un*observables

Recall that in the potential outcomes framework, we are almost always using a strategy aimed at characterizing the form that the counterfactuals take in order to identify a causal relationship. This will almost always involve making an argument about the data generating proces that led to the observed distribution of the treatment indicator. Randomization is one such argument that gives us causal identification via independence between treatment and potential outcomes. Selection-on-observables is another such argument that gives us causal identification via conditional independence between treatment and potential outcomes.

The motivation for selection on *un*observables arises when we do not have randomization, and we do not have a good argument for conditional independence because there is an unobserved factor that is associated with selection into treatment and control. Because such a factor is unobserved, at first glance it may seem as if there is nothing that we can do. But we may be able to exploit some feature of the design setting to render the effects of unobserved factors impotent.

Difference in differences is a particular strategy that becomes available when some minimum data requirements are satisfied, and we believe in the veracity of an untestable assumption. Specifically, when we observe trends in an outcome over time, and an exogenous intervention that affects some units but not others, it is reasonable to think that unobserved factors may contribute to both baseline differences between units in the treatment and control groups, as well as natural change over time. But such a situation allows us to exploit a simple trick to eliminate the effects of the unobserved factor and isolate the effect of the intervention on the outcome, if we believe that in the absence of the intervention, the treated units would have behaved exactly like the control units (the so-called *parallel trends assumption*). If we have outcome measurements in at least one pre-intervention and one post-intervention time period and the intervention assigns some units to treatment and some to control, and if we believe the parallel trends assumption holds, then the difference-in-difference estimate gives us the average treatment effect on the treated (ATT).

Different ways of estimating the ATT from a difference in difference procedure

Now we will simulate some data to illustrate how various ways of calculating the difference-in-differences estimate lead to the same results. We focus on the simplest difference-in-differences setting in which all units receive the intervention at the same time. However, in this example we observe cases over ten time periods, setting up the ability to explore the longer time trend for the purpose of assessing the quality of the parallel trends assumption.

Begin by setting the seed so that any random process in your code is reproducible.

```
# set seed
set.seed(71036)
```

Then we generate a sample of size $N = 500$, each observed over 10 time periods, with potential outcomes Y_0 and Y_1 constructed as a linear function of unit-specific random variation, experimental group-specific fixed effects, and time-specific trends. Furthermore, let us assume that a random exogenous treatment occurs between time period 7 and time period 8.

Specifically, we imagine that Y_0 is generated by:

$Y_{0it} = \beta_0 + \gamma * g + \lambda_t * s_t + \epsilon_{it}$,

where $i = 1, \ldots, 500$ indexes units, $t = 1, \ldots, 10$ indexes time periods, g is a dummy indicator for whether a unit is treated by the intervention, $s_2, \ldots, s_{10}$ are dummy indicators for time periods $2, \ldots, 10$, and ϵ_{it} is random error term (in this case, distributed normally with a mean of 0 and standard deviation of 10).

Y_{1it} is then generated as $Y_{1it} = Y_{0it} + \bar{Y}_0 + \varepsilon_{it}$, where ε_{it} is a random error term distributed normally with a mean of 0 and standard deviation of 1.

```
# sample size
N <- 500

# randomly assigned treatment indicator
g <- sample(0:1, N, replace = TRUE)

# id placeholder
id <- 1:N

# time periods
t <- c(rep(1, N), rep(2, N), rep(3, N), rep(4, N), rep(5, N), rep(6, N),
       rep(7, N), rep(8, N), rep(9, N), rep(10, N))

# put into dataframe
```

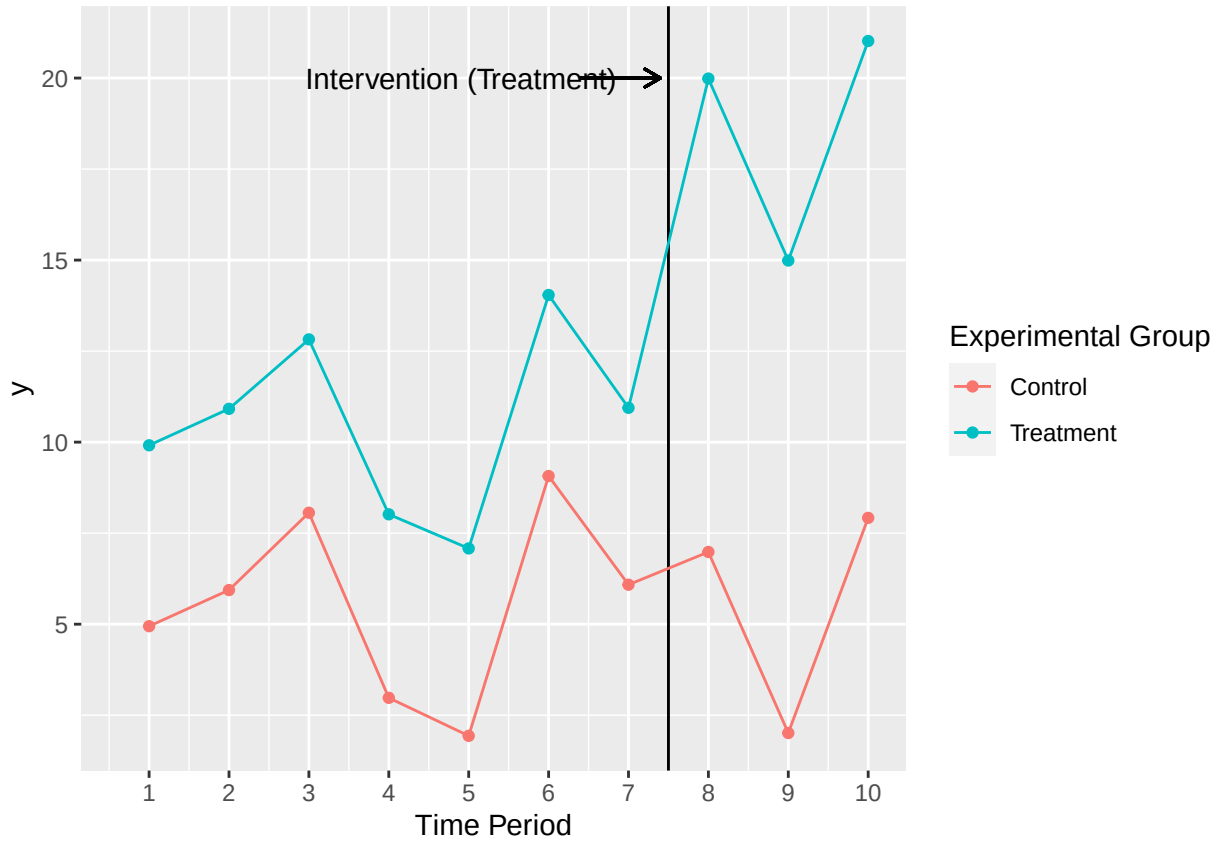
```
sim_data <- cbind(id = rep(id, 10), t, g = rep(g, 10)) %>% as_tibble()

# add time period dummy variables
sim_data <- sim_data %>%
  mutate(t1 = if_else(t == 1, 1, 0), t2 = if_else(t == 2, 1, 0),
         t3 = if_else(t == 3, 1, 0), t4 = if_else(t == 4, 1, 0),
         t5 = if_else(t == 5, 1, 0), t6 = if_else(t == 6, 1, 0),
         t7 = if_else(t == 7, 1, 0), t8 = if_else(t == 8, 1, 0),
         t9 = if_else(t == 9, 1, 0), t10 = if_else(t == 10, 1, 0))

# parameters for equation to generate simulated outcomes
b0 <- 5
gamma <- 5
lambda_t_minus5 <- 1
lambda_t_minus4 <- 3
lambda_t_minus3 <- -2
lambda_t_minus2 <- -3
lambda_t_minus1 <- 4
lambda_t_0 <- 1
lambda_t_1 <- 2
lambda_t_2 <- -3
lambda_t_3 <- 3

# generate y0
y0_panel <- b0 + gamma * sim_data$g + lambda_t_minus5 * sim_data$t2 +
  lambda_t_minus4 * sim_data$t3 + lambda_t_minus3 * sim_data$t4 +
  lambda_t_minus2 * sim_data$t5 + lambda_t_minus1 * sim_data$t6 +
  lambda_t_0 * sim_data$t7 + lambda_t_1 * sim_data$t8 +
  lambda_t_2 * sim_data$t9 + lambda_t_3 * sim_data$t10 + rnorm(N * 10)

# generate y1
y1_panel <- y0_panel + mean(y0_panel) + rnorm(N)

# add y0, y1, and treatment indicator to dataframe
sim_data <- sim_data %>%
  mutate(y0 = y0_panel, y1 = y1_panel, d1 = if_else(g == 1 & t >= 8, 1, 0))

# generate y as y0 when d1 = 0 and y1 when d1 = 1
sim_data <- sim_data %>% mutate(y = if_else(d1 == 1, y1, y0))

# sort data into standard panel format
sim_data <- sim_data %>% arrange(id, t)
```

Now we can plot the trend lines for the the means of Y by experimental group.

```
# plot data
avg_lineplot_sim_data <- sim_data %>% group_by(t, g, d1) %>%
  summarize(y = mean(y)) %>%
  mutate(group = if_else(g == 0, "Control", "Treatment"))
```

```
## `summarise()` has grouped output by 't', 'g'. You can override using the
## `.groups` argument.
```

```
p1 <- ggplot(avg_lineplot_sim_data, aes(x = t, y = y, group = group)) +
  geom_vline(xintercept = 7.5, color = 'black') +
  geom_segment(x = 6.4, y = 20, xend = 7.4, yend = 20,
               arrow = arrow(length = unit(.1, "inches"))) +
  annotate("text", x = .62, y = 20, label = "Intervention (Treatment)", hjust = -.6) +
  geom_line(aes(color = group)) + geom_point(aes(color = group)) +
  scale_x_continuous(breaks = 1:10) +
  labs(x = "Time Period", color = "Experimental Group")
p1
```



We see that, as should be the case, the two experimental groups start from different baselines, then experience over-time trends that largely mirror one another, then diverge when the intervention comes into play. In other words, the simulated data sets up an ideal situation for analyzing treatment effects using difference-in-differences. For the illustrations that follow, recall that the difference-in-differences procedure, in combination with the parallel trends assumption, produces the ATT. Therefore, our benchmark is the true ATT:

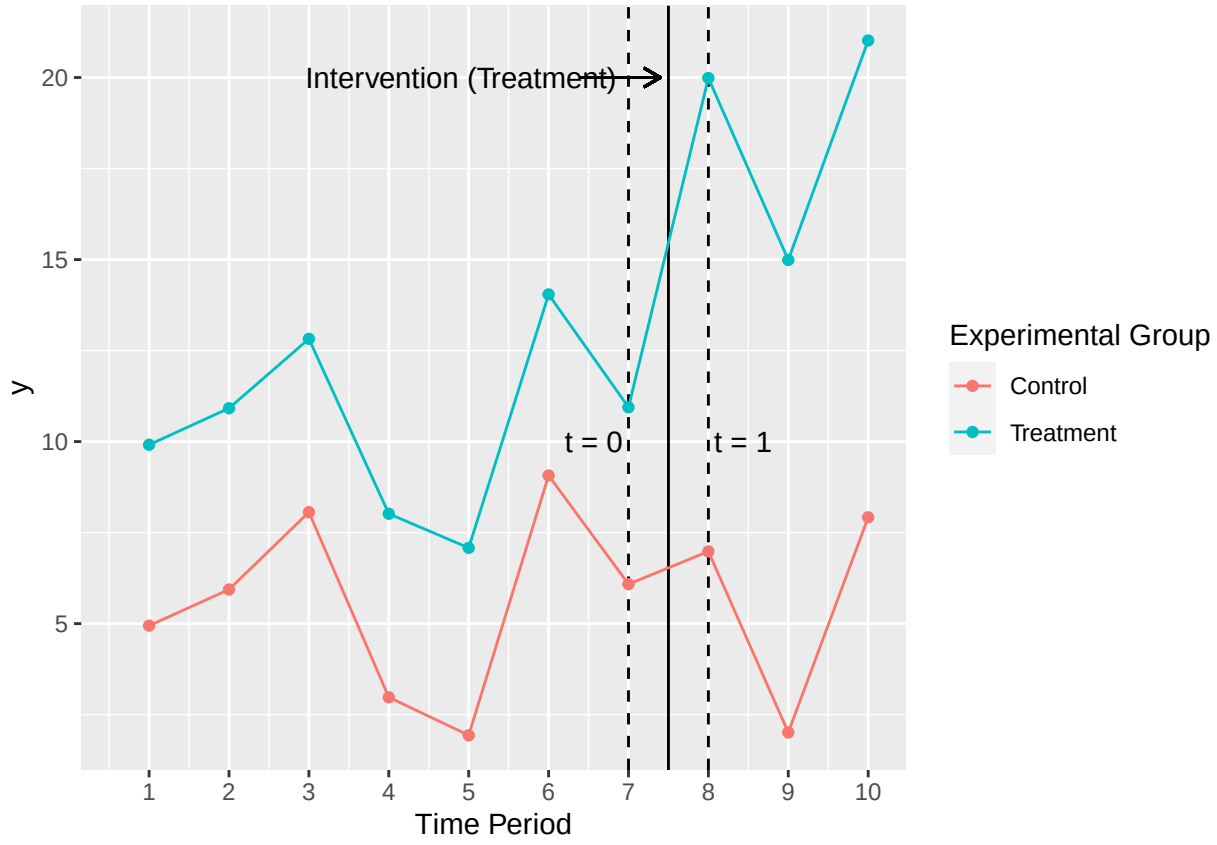
```
true_att <- mean(sim_data[sim_data$t == 8 & g == 1, ]$y1) -
            mean(sim_data[sim_data$t == 8 & g == 1, ]$y0)
true_att
```

```
## [1] 8.091875
```

Now we consider the various ways that the difference-in-differences estimate can be calculated, given the data structure. Since the intervention takes effect for all units in the treatment group at the same time, we can first ignore the fact that we have multiple pre-intervention and post-intervention observations. In

other words, just focusing on the minimum data requirements, we can simply examine the data at the final pre-intervention time period as compared to the first post-intervention time period (usually denoted as $t = 0$ and $t = 1$, respectively).

```
p2 <- ggplot(avg_lineplot_sim_data, aes(x = t, y = y, group = group)) +
  geom_vline(xintercept = 7.5, color = 'black') +
  geom_vline(xintercept = 7, color = 'black', lty = 2) +
  geom_vline(xintercept = 8, color = 'black', lty = 2) +
  geom_segment(x = 6.4, y = 20, xend = 7.4, yend = 20,
               arrow = arrow(length = unit(.1, "inches"))) +
  annotate("text", x = .62, y = 20, label = "Intervention (Treatment)",
           hjust = -.6) +
  annotate("text", x = 7, y = 10, label = "t = 0", hjust = 1.1) +
  annotate("text", x = 8, y = 10, label = "t = 1", hjust = -.1) +
  geom_line(aes(color = group)) + geom_point(aes(color = group)) +
  scale_x_continuous(breaks = 1:10) +
  labs(x = "Time Period", color = "Experimental Group")
p2
```



We can begin by simply performing the difference-in-differences calculation by hand. For this, we have two options:

1. Take the average difference between treatment and control in the post-intervention period, the average difference between treatment and control in the pre-intervention period, and the difference between these two differences; or

2. Take the average difference between treated units in the post-intervention period and treated units in the pre-intervention period, the average difference between control units in the post-intervention period and control units in the pre-intervention period, and the difference between these two differences.

If you are not already convinced, at this point convince yourself that these calculations are indeed identical.

```
# treatment-control differences post, treatment-control differences pre, and the
#   difference between them
treat_control_pre_diff <-
  mean(sim_data[sim_data$t == 7 & sim_data$g == 1, ]$y) -
  mean(sim_data[sim_data$t == 7 & sim_data$g == 0, ]$y)
treat_control_post_diff <-
  mean(sim_data[sim_data$t == 8 & sim_data$g == 1, ]$y) -
  mean(sim_data[sim_data$t == 8 & sim_data$g == 0, ]$y)
did_est01 <- treat_control_post_diff - treat_control_pre_diff

# treatment group pre-post differences, control group pre-post differences, and
#   the difference between them
treat_post_pre_diff <-
  mean(sim_data[sim_data$t == 8 & sim_data$g == 1, ]$y) -
  mean(sim_data[sim_data$t == 7 & sim_data$g == 1, ]$y)
control_post_pre_diff <-
  mean(sim_data[sim_data$t == 8 & sim_data$g == 0, ]$y) -
  mean(sim_data[sim_data$t == 7 & sim_data$g == 0, ]$y)
did_est02 <- treat_post_pre_diff - control_post_pre_diff

# difference-in-difference hand calculations
did_est01
```

```
## [1] 8.144188
```

```
did_est02
```

```
## [1] 8.144188
```

In addition to being identical to one another, these calculations are identical to three different versions of a regression estimate of the difference-in-differences ATT. Specifically, we should get identical estimates as the hand calculations from each of the following:

1. $Y = \mu + \gamma * g + \delta * t + \alpha * g * t + \varepsilon$,
2. $Y = \mu + \gamma * g + \delta * t + \alpha * D + \varepsilon$, and
3. $\Delta Y = \mu + \alpha * \Delta D$,

where g is defined as above, $t = 0$ in the pre-intervention period and $t = 1$ in the post-intervention period, $D = 1$ for treated units in the post-intervention period and 0 otherwise, and the Δ notation indicates having taken the first difference, such that $\Delta Y = Y_{t=1} - Y_{t=0}$ and $\Delta D = 1$ for treated cases and 0 for control cases. In all cases, the coefficient estimate for α will represent the ATT.

```
## regressions based on panel format
# 1.
did_reg_panel01 <-
```

```
  lm(y ~ g*t8, data = sim_data[sim_data$t %in% c(7, 8), ])
# 2.
did_reg_panel02 <-
  lm(y ~ g + t8 + d1, data = sim_data[sim_data$t %in% c(7, 8), ])

## regression based on first differences
# 3.
sim_data <- sim_data %>% mutate(y_diff = y - dplyr::lag(y, 1))
did_reg_firstdiff <- lm(y_diff ~ d1, data = sim_data[sim_data$t == 8, ])

summary(did_reg_panel01)
```

```
##
## Call:
## lm(formula = y ~ g * t8, data = sim_data[sim_data$t %in% c(7,
##     8), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6281 -0.7099 -0.0179  0.6612  4.3084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.08405    0.06994  86.995   <2e-16 ***
## g            4.85885    0.10073  48.235   <2e-16 ***
## t8           0.89931    0.09890   9.093   <2e-16 ***
## g:t8         8.14419    0.14246  57.169   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 996 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9594
## F-statistic:  7861 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
summary(did_reg_panel02)
```

```
##
## Call:
## lm(formula = y ~ g + t8 + d1, data = sim_data[sim_data$t %in%
##     c(7, 8), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6281 -0.7099 -0.0179  0.6612  4.3084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.08405    0.06994  86.995   <2e-16 ***
## g            4.85885    0.10073  48.235   <2e-16 ***
## t8           0.89931    0.09890   9.093   <2e-16 ***
## d1           8.14419    0.14246  57.169   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 996 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9594
## F-statistic:  7861 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
summary(did_reg_firstdiff)
```

```
##
## Call:
## lm(formula = y_diff ~ d1, data = sim_data[sim_data$t == 8, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1976 -0.9698  0.0060  1.0174  6.7440
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8993     0.1004   8.961   <2e-16 ***
## d1            8.1442     0.1446  56.339   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.615 on 498 degrees of freedom
## Multiple R-squared:  0.8644, Adjusted R-squared:  0.8641
## F-statistic:  3174 on 1 and 498 DF,  p-value: < 2.2e-16
```

The next natural question is how we can exploit the additional information from the multiple pre-intervention periods. Keeping in mind that there is no direct test of the parallel trends assumption, observations at pre-intervention time points can provide some additional evidence in favor of parallel trends between treated and control observations. This essentially involves performing a series of placebo tests (comparisons between future treated and untreated cases) at each of the two-period time trends before the intervention comes into effect. In this case, it would take the form of placebo tests at time periods 1 and 2, 2 and 3, 3 and 4, 4 and 5, 5 and 6, 6 and 7, and then a calculation of the treatment effect between time periods 7 and 8, followed by post-intervention tests of the persistence of the treatment effect.

Specifically, we would estimate a model like the following:

$$Y = \mu + \gamma * g + \lambda_t * s_t + \\ \alpha_{-6} * g * s_2 + \alpha_{-5} * g * s_3 + \alpha_{-4} * g * s_4 + \alpha_{-3} * g * s_5 + \alpha_{-2} * g * s_6 + \alpha_{-1} * g * s_7 + \alpha_0 * g * s_8 + \alpha_1 * g * s_9 + \alpha_2 * g * s_{10}$$

In this formulation, the coefficient estimates for $\alpha_{-6}, \ldots, \alpha_{-1}$ are placebo tests for all of the pre-intervention periods, while α_0 is the ATT, and α_1 and α_2 are estimates of the post-intervention persistence effects. If this is to serve as evidence in favor of the parallel trends assumption, we should find a series of null effects for all of the placebo tests. In general, there is no expectation about the coefficient estimates capturing persistence effects (though in these simulated data, Y_1 was drawn for treated cases in all post-intervention periods, indicating persistence).

```
# inclusion of lags and leads to test parallel trends
did_reg_laglead <-
  lm(y ~ g*t2 + g*t3 + g*t4 + g*t5 + g*t6 + g*t7 + g*t8 + g*t9 + g*t10,
     data = sim_data)
summary(did_reg_laglead)
```

```
##
## Call:
## lm(formula = y ~ g * t2 + g * t3 + g * t4 + g * t5 + g * t6 +
##     g * t7 + g * t8 + g * t9 + g * t10, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6281 -0.7024 -0.0021  0.6812  4.3084
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.944372   0.066160  74.734   <2e-16 ***
## g            4.969246   0.095295  52.146   <2e-16 ***
## t2           0.993497   0.093564  10.618   <2e-16 ***
## t3           3.116614   0.093564  33.310   <2e-16 ***
## t4          -1.967457   0.093564 -21.028   <2e-16 ***
## t5          -3.010801   0.093564 -32.179   <2e-16 ***
## t6           4.124553   0.093564  44.083   <2e-16 ***
## t7           1.139683   0.093564  12.181   <2e-16 ***
## t8           2.038993   0.093564  21.792   <2e-16 ***
## t9          -2.933677   0.093564 -31.355   <2e-16 ***
## t10          2.976092   0.093564  31.808   <2e-16 ***
## g:t2         0.008439   0.134768   0.063    0.950
## g:t3        -0.207905   0.134768  -1.543    0.123
## g:t4         0.070912   0.134768   0.526    0.599
## g:t5         0.178239   0.134768   1.323    0.186
## g:t6         0.003501   0.134768   0.026    0.979
## g:t7        -0.110400   0.134768  -0.819    0.413
## g:t8         8.033788   0.134768  59.612   <2e-16 ***
## g:t9         8.008275   0.134768  59.423   <2e-16 ***
## g:t10        8.128186   0.134768  60.313   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.065 on 4980 degrees of freedom
## Multiple R-squared:  0.9586, Adjusted R-squared:  0.9584
## F-statistic:  6065 on 19 and 4980 DF,  p-value: < 2.2e-16
```

As expected the placebo tests all produce null results, indicating no differences in the trends in the pre-intervention periods, a result that would be helpful for convincing ourselves that the parallel trends assumption is reasonable.

We see that the treatment effect calculation, α_0, for the interaction between g and s_8 is close, but not exactly the same as the previous calculations. This difference is due to the fact that the regression is now estimating the treatment effect after controlling for differences in time trends across the ten time periods.

The purpose here has basically been to demonstrate that there are many ways to estimate a difference-in-difference treatment effect depending on the data structure and the preferences of the researcher. This is important for illustrative purposes. But most important is to always keep in mind that, though the mathematics and minimum data requirements necessary to obtain a difference-in-difference estimate are quite simple, our ability to interpret that estimate as an ATT rests solely on the quality of the parallel trends assumption. If parallel trends does not hold, we have no way of characterizing counterfactuals, and hence no way of eliminating potential competing explanations. Examining pre-intervention placebo tests is the best indirect test we have at our disposal when we have multiple pre-intervention periods, it is not a direct test of parallel trends. And, indeed, there is no direct test of parallel trends.

Homework Assignment: The Impact of Minimum Wage on Teenage Employment

Classical economic theory states that raises in minimum wages hurt employment, especially teenage employment since teenage wages are often set at the minimum wage. Such is the main argument of those who oppose raising minimum wages. In this exercise, we are going to put this economic theory to test by exploiting a natural experiment in New Jersey and Pennsylvania. In 1992, New Jersey's minimum wage increased from \$4.25 to \$5.05 while the minimum wage in Pennsylvania remained at \$4.25. In their seminal study, Card and Krueger (1994) used data on employment at fast-food establishments in New Jersey and Pennsylvania before and after the increase in the minimum wage to measure the impact of the increase in minimum wage on teenage employment.

The units is here a fast-food restaurant, and the populations of them that we consider are these groups of such restaurants in Pennsylvania and New Jersey. The causal effect of interest is the effect of increase of minimum wage on employment among the New Jersey restaurants.

The following variables are included in the `CardKrueger.csv` dataset from the Card and Krueger (1994) paper. You can find the paper itself on the MY457 Moodle page, under week 5.

Variable Name	Variable Description
`emptot`	Full-time Equivalent (FTE) Employment Before Minimum Wage Increase in New Jersey: count of number of full-time workers plus 0.5 times the count of the number of part-time workers
`emptot2`	Full-time Equivalent (FTE) Employment After Minimum Wage Increase in New Jersey: count of number of full-time workers plus 0.5 times the count of the number of part-time workers
`nj`	1 if NJ; 0 if PA (Treatment Indicator)
`pa`	1 if PA; 0 if NJ (Control Indicator)
`southj`	1 if in southern NJ (Subset of Treated Cases)
`centralj`	1 if in central NJ (Subset of Treated Cases)
`pa1`	1 if in PA, northeast suburbs of Philadelphia (Subset of Control Cases)
`pa2`	1 if in PA, Easton, etc. (Subset of Control Cases)
`wage_st`	Starting Wage (\$/hr) Before Minimum Wage Increase in New Jersey
`wage_st2`	Starting Wage (\$/hr) After Minimum Wage Increase in New Jersey
`hrsopen`	Hours Open Weekday Before Minimum Wage Increase in New Jersey
`hrsopen2`	Hours Open Weekday After Minimum Wage Increase in New Jersey
`bk`	1 if Burger King; 0 Otherwise
`kfc`	1 if KFC; 0 Otherwise
`roys`	1 if Roy Rogers; 0 Otherwise
`wendys`	1 if Wendys; 0 Otherwise
`pmeal`	Price of Full Meal Before Minimum Wage Increase in New Jersey
`pmeal2`	Price of Full Meal After Minimum Wage Increase in New Jersey
`closed`	Closed Permanently After Minimum Wage Increase in New Jersey
`co_owned`	1 if company owned; 0 Otherwise

Preliminaries

These are panel data, but the data are formatted in a "wide" form where the two observations for the same restaurant are in different columns (variables) on the same row of the data.To employ the kinds of models we introduced above it is necessary to first convert the data into a "long" format which has one row per period per restaurant. The code below shows how this can be done.

```
ckdata <- read.csv('CardKrueger.csv', stringsAsFactors = FALSE)
tail(ckdata)
```

```
##     emptot emptot2 nj pa southj centralj pa1 pa2 wage_st wage_st2 hrsopen
## 405   6.50   16.00  1  0      0        0   0   0    4.75     5.05    11.0
## 406   9.00   23.75  1  0      0        0   0   0    4.95     5.25    12.0
## 407   9.75   17.50  1  0      0        0   0   0    4.75     5.25    11.0
## 408  24.50   20.50  1  0      0        0   0   0    4.25     5.05    19.0
## 409  14.00   20.50  1  0      0        0   0   0    4.75     5.05    12.5
## 410  19.50   25.00  1  0      0        0   0   0    4.62     5.14    12.5
##     hrsopen2 bk kfc roys wendys pmeal pmeal2 closed co_owned
## 405     11.0  0   1    0      0  4.25   4.17      0        0
## 406     11.0  0   1    0      0  4.25   4.21      0        1
## 407     14.0  0   1    0      0  4.25   4.31      0        1
## 408     18.0  0   0    1      0  3.26   3.21      0        1
## 409     12.5  0   0    0      1  4.07   4.27      0        0
## 410     12.5  0   0    0      1  3.10   4.05      0        0
```

```
ckdata2 <- reshape(ckdata,direction="long",
      varying=list(c("emptot","emptot2"),c("wage_st","wage_st2"),
                   c("hrsopen","hrsopen2"),c("pmeal","pmeal2")),
        v.names=c("emptot","wage_st","hrsopen","pmeal"),
      idvar="restaurant",ids=as.numeric(rownames(ckdata)))
ckdata2 <- ckdata2[order(ckdata2$restaurant,ckdata2$time),]
ckdata2$timepost <- as.numeric(ckdata2$time==2)
ckdata2$treated <- ckdata2$nj*ckdata2$timepost
tail(ckdata2)
```

```
##       nj pa southj centralj pa1 pa2 bk kfc roys wendys closed co_owned time
## 408.1  1  0      0        0   0   0  0   0    1      0      0        1    1
## 408.2  1  0      0        0   0   0  0   0    1      0      0        1    2
## 409.1  1  0      0        0   0   0  0   0    0      1      0        0    1
## 409.2  1  0      0        0   0   0  0   0    0      1      0        0    2
## 410.1  1  0      0        0   0   0  0   0    0      1      0        0    1
## 410.2  1  0      0        0   0   0  0   0    0      1      0        0    2
##       emptot wage_st hrsopen pmeal restaurant timepost treated
## 408.1   24.5    4.25    19.0  3.26        408        0       0
## 408.2   20.5    5.05    18.0  3.21        408        1       1
## 409.1   14.0    4.75    12.5  4.07        409        0       0
## 409.2   20.5    5.05    12.5  4.27        409        1       1
## 410.1   19.5    4.62    12.5  3.10        410        0       0
## 410.2   25.0    5.14    12.5  4.05        410        1       1
```

Q1. The homework this time is short and simple. Calculate the difference-in-differences estimate of the effect of the increase in minimum wage on FTE employment in these restaurants in New Jersey. Calculate this estimator both using a linear-regression formulation which does not use a fixed-effects formulation of the analysis, and using a fixed-effects model.

Check that your results match that reported in Table 3 (row 4) of Card and Krueger (1994). What is the substantive conclusion from these results?

Note 1: The estimated standard errors of your estimated effects will be different from the ones in the paper. For the linear-model estimator this is because it assumes homoscedasticity (constant residual variance) within

each cell defined by a combination of state and period, while the estimates in the paper (which are derived using basic formulas for differences of sample means) allow these variances to be different from each other. The standard errors from the fixed effects model (or a linear model for first differences) are here substantially smaller than the standard errors from the linear regression models (apart from the first-differences model). This is because by controlling for the individual fixed effects this model reflects the fact that we actually have panel data of individual restaurants (i.e. the estimation effectively uses within-restaurant differences over time rather than differences of group means). Here the values of employment vary much more between than within restaurants, so controlling for the between-restaurant variation with the fixed effects substantially increases the precision of the estimates.

Note 2: The answer document for the homework also includes some further analysis of these data (from a previous round of teh course in 2021) to give you further examples of such analyses. It is not part of the homework.