Diff in Diff

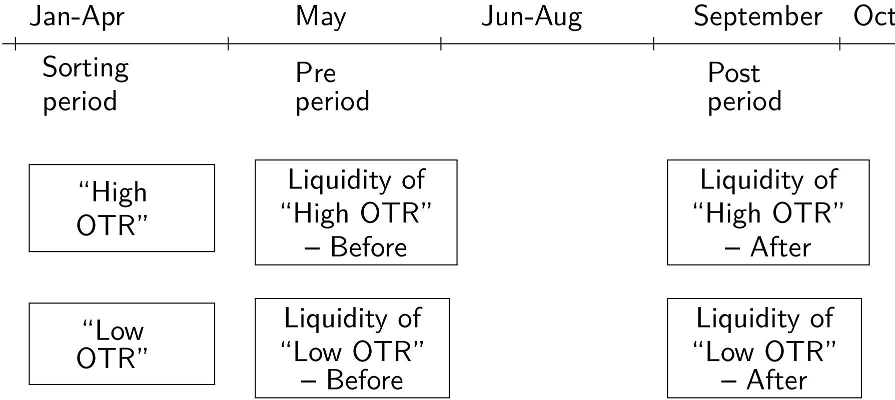
Idea of a diff in diff:
Group into

- One group affected by an intervention (treated)
- One group not affected by the intervention (nontreated)

This paper:
Intervention: Introduction of a “congestion tax” – traders on the Oslo Stock Exchange pays fee if the number of “orders” (messages into the exchanges limit order book: orders, cancellations, updating of orders) over the number of “trades” (OTR - Order to Trade Ratio) exceeds 70.
So, want to investigate if this “tax” affects negatively the working of the stock exchange.
Look at trade quality around intervention. To capture the exact effect of the intervention, need to (hopefully) control for general movements in liquidity around intervention.

Fact: Not all stocks at the OSE are heavily traded.
Most (the smaller stocks) only traded at the OSE,
but some (the larger stocks) also traded elsewhere (Nasdaq OMX (Stockholm), BATS, Chi-X, Turquoise).
Idea: The smaller stocks have low OTR's anyway, will not be affected by the tax. Use as nontreated.
The stocks with high OTRs (implemented as OTR above 50) proxied as the treated sample.

Illustrating the difference in difference analysis



Jan-Apr

May

Jun-Aug

September

Oct

Sorting period

Pre period

Post period

"High OTR"

"Low OTR"

Liquidity of "High OTR" – Before

Liquidity of "Low OTR" – Before

Liquidity of "High OTR" – After

Liquidity of "Low OTR" – After

The OTR ratio regulation was introduced at the end of May '12, to be implemented first time for September '12.

We use the first part of the year (January-April) to choose a set of stocks not likely to be affected by the new regulation.

To proxy for that we measure the OTR for each stock for each day in the period January–April, and choose as the set of stocks not likely to be affected by the regulation the stocks with their maximal OTR in the period lower than fifty.

We term this group the "Low OTR" stocks.

This is then compared to stocks with an observed OTR higher than fifty in the same time period (Jan-Apr), which we term the "High OTR" stocks.

Estimation of the diff in diff is based on regressions of the type

$$y = \beta_0 + \beta_1 d_{treated} + \beta_2 d_{time} + \delta d_{treated} \times d_{time} + \alpha \mathbf{X} + \varepsilon \quad (1)$$

where
y is the variable of interest (i.e. liquidity),
$d_{treated}$ is a dummy variable for whether an element belongs to the treatment or the control group (high vs low OTR),
d_{time} a time dummy for the second period.

$$y = \beta_0 + \beta_1 d_{treated} + \beta_2 d_{time} + \delta d_{treated} \times d_{time} + \alpha \mathbf{X} + \varepsilon$$

The coefficient of interest, δ, multiplies the interaction term, which is the same as a dummy variable equal to one for the observations in the treatment group in the second period.

The coefficient δ measures the direct effect of the intervention.

In the regression we allow for additional covariates **X**.

In the reported regression we control for size differences between the high and low OTR groups by including log firm size as an additional explanatory variable.

Finally, we adjust for the panel data nature of the data by including fixed date and stock effects, and adjusting the standard errors in the panel for clustering.

The estimation is done using the `R` library `plm`. Calculation of standard errors is described in Croissant and Millo (2008).

Results
Estimates of Difference in Difference investigation of "Low OTR" vs "High OTR" stocks

	Dependent variable:					
	OTR	Quoted(Rel) Spread	Effective Spread	Realized Spread	Roll	RV
β_2 d(Post Period)	-13.489	40.229^{***}	11.038^{***}	40.229^{***}	17.393^{***}	9.203^{***}
	(27.870)	(0.102)	(0.140)	(0.102)	(0.143)	(0.160)
β_1 d(high OTR)	-3.574	6.265^{***}	1.020^{***}	6.265^{***}	0.508^{***}	-0.054
	(7.829)	(0.152)	(0.044)	(0.152)	(0.033)	(0.052)
δ Interaction	-11.331	0.014	0.055	0.014	0.008	0.016
	(8.438)	(0.210)	(0.047)	(0.210)	(0.061)	(0.063)
ln(Firm Size)	1.588	-2.059^{***}	-0.533^{***}	-2.059^{***}	-0.790^{***}	-0.402^{***}
	(1.568)	(0.001)	(0.007)	(0.001)	(0.006)	(0.009)
Observations	5,710	7,061	4,977	7,061	2,703	4,746
Adjusted R^2	0.554	0.745	0.717	0.745	0.487	0.390
Note:						*p$<$0.1; **

Estimates of the regression $y = \beta_0 + \beta_2 d_{time} + \beta_1 d_{treated} + \delta d_{treated} \times d_{time} + \alpha \mathbf{X} + \varepsilon$, where y is the various liquidity measures. $d_{treated}$ (d(high OTR)) is a dummy variable for treatment, where treatment is proxied by the maximal OTR in Jan-Apr '12 being above 50. d_{time} (d(Post Period) is equal to one if the observations is in the second period (September '12) and zero otherwise. **X** are additional covariates. The analysis is performed for the the Order to Trade Ratio quoted (relative) spread, the effective spread, the realized spread, the Roll measure, the Realized Volatility, , and the Depth. The spread measures, the Roll measure and the Realized Volatility in percent. Depth

Doing the analysis in R

The necessary libraries

```
> library(plm)
> library(stargazer)
```

Reading the data. Replace the in and out directories with ones you use, but always try to keep results separate.

```
> indir <- "../../results/2016_12_dump_for_diff_in_diff/"
> outdir <- indir
>
> filename  <- paste0(indir,"liquidity_and_other_variables.
> data      <- read.table(filename,header=TRUE,sep=";")
> data$Date      <- as.Date(as.character(data$Date),format=
> data$EffSpread <- 100.0*data$EffSpread # use numbers in p
> head(data)
       ID       Date maxOTR   FirmSize EffSpread RelSpread
1 1249085 2012-05-02 267.61 3711568806    0.4253  0.014543
2 1249085 2012-05-03 267.61 3711568806    0.5504  0.013497
3 1249085 2012-05-04 267.61 3711568806    0.5705  0.012979
4 1249085 2012-05-07 267.61 3711568806    1.0331  0.029051
5 1249085 2012-05-08 267.61 3711568806    0.4895  0.011465
6 1249085 2012-05-09 267.61 3711568806    0.3592  0.006616
        OTR       RV     Roll  Depth
1 320.92857 0.002929       NA 151176
2  78.04167 0.006812       NA 163500
```

Creating the dummy variables

```
> dSecondPeriod <- as.numeric(data$Date>=as.Date("2012-09-0
> dTreated      <- as.numeric(data$maxOTR>50)
> dInteraction  <- dSecondPeriod * dTreated
```

Need to remove missing obs from firm size before taking logs

```
> merged    <- na.omit(data.frame(data$ID,
+                                 data$Date,
+                                 data$EffSpread,
+                                 dSecondPeriod,
+                                 dTreated,
+                                 dInteraction,
+                                 data$FirmSize))
> names(merged) <- c("ID","Date","EffSpread","dSecondPerio
+                             "dTreated","dInteraction","F
> lnFirmSize    <- log(merged$FirmSize)
```

Running an OLS regression

```
> regrEffSpreadOLS <- lm(merged$EffSpread ~ merged$dSecond
+                            + merged$dInteraction + lnFirm
```

With results

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           6.19145    0.09125  67.855   <2e-16 ***
merged$dSecondPeriod -0.18943    0.02112  -8.968   <2e-16 ***
merged$dTreated      -0.14914    0.02176  -6.855    8e-12 ***
merged$dInteraction   0.07458    0.03081   2.420   0.0155 *  
lnFirmSize           -0.25022    0.00425 -58.871   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5424 on 4972 degrees of freedom
Multiple R-squared:  0.4307,Adjusted R-squared:  0.4302
F-statistic: 940.4 on 4 and 4972 DF,  p-value: < 2.2e-16
```

Doing similar analysis using the panel utilities

```
> PanelData <- data.frame(merged$ID,
+                         merged$Date,
+                         merged$EffSpread,
+                         merged$dSecondPeriod,
+                         merged$dTreated,
+                         merged$dInteraction,
+                         lnFirmSize)
> names(PanelData) <- c("ID","Date","EffSpread","dSecondPeriod",
+                                    "dInteraction","lnFirmSize"
> regrEffSpreadPanel1 <- plm(EffSpread ~ 0
+                               + dSecondPeriod + dTreated + dIn
+                               + lnFirmSize,
+                            data = PanelData,
+                            model = "pooling")
```

With results

```
Unbalanced Panel: n=181, T=1-40, N=4977

Coefficients :
                 Estimate  Std. Error t-value  Pr(>|t|)
dSecondPeriod -0.03324317  0.02913674 -1.1409 0.2539513
dTreated      -0.11291150  0.03018287 -3.7409 0.0001854 ***
dInteraction  -0.06164543  0.04267004 -1.4447 0.1486050
lnFirmSize     0.03432806  0.00096138 35.7069 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-Squared:      0.077742
Adj. R-Squared: 0.077186
F-statistic: -109.416 on 4 and 4973 DF, p-value: 1
```

With fixed effects (using factors for each data and ID).

```
> regrEffSpreadPanel2 <- plm(EffSpread ~ 0
+                              + dSecondPeriod + dTreated +
+                              + lnFirmSize
+                              + factor(Date) + factor(ID),
+                              data = PanelData,
+                              model = "pooling")
```

with results

```
> summary(regrEffSpreadPanel2)

Unbalanced Panel: n=181, T=1-40, N=4977

Coefficients :
                         Estimate Std. Error t-value  Pr(>|t|)
dSecondPeriod          11.0384713  1.4097414  7.8301 5.966e-15 *
dTreated                1.0203672  0.3868604  2.6376 0.0083777 *
dInteraction            0.0551399  0.0222091  2.4828 0.0130710 *
lnFirmSize             -0.5334836  0.0802801 -6.6453 3.366e-11 *
factor(Date)2012-05-02 11.1313884  1.4098697  7.8953 3.568e-15 *
factor(Date)2012-05-03 11.0758747  1.4098831  7.8559 4.872e-15 *
....
factor(ID)6037          0.2623546  0.0849443  3.0885 0.0020230 *
factor(ID)6059          2.4338387  0.6312505  3.8556 0.0001170 *
.....
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-Squared:      0.72999
Adj. R-Squared: 0.7175
F-statistic: 58.1815 on 221 and 4756 DF, p-value: < 2.22e-16
```

Pretty printing. Note the replacing of the OLS variance covariance matrix with a panel robust one

```
>                                         # pretty printing for
> labels <- c("$\\beta_2$ d(Post Period)","$\\beta_1$ d(high OTR
                      "$\\delta$ Interaction","ln(Firm Size)")
> filename  <- paste0(outdir,"example_panel_regr_eff_spread.tex"
> stargazer(regrEffSpreadPanel2,
+           keep=c(1,2,3,4), # do not display fixed effects
+           out=filename,
+           float=FALSE,
+           covariate.labels=labels,
+           se = list(sqrt(diag(vcovHC(regrEffSpreadPanel2)))),
+           omit.stat=c("f","rsq","ser"))
```

	Dependent variable:
	EffSpread
β_2 d(Post Period)	11.038^{***}
	(0.140)
β_1 d(high OTR)	1.020^{***}
	(0.044)
δ Interaction	0.055
	(0.047)
ln(Firm Size)	-0.533^{***}
	(0.007)
Observations	4,977
Adjusted R^2	0.717
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

Yves Croissant and Giovanni Millo. Panel data econometrics in R: the plm package, 2008. R vignette, available at CRAN.