Advanced Statistics and Programming

Group Assignment 2

Group 7

Eline van Groningen
Valery Maasdamme
Paola Priante
Yuhu Wang

October 2020

1 Difference-in-Difference Analysis: Female Labor Force Participation

In order to conduct a difference-in-difference (DiD) analysis, a dummy variable *dPeriod* is created to indicate the time before and after the policy. The EITC was implemented in 1993. Therefore, the years 1991 and 1992 take a value of 0, and 1993 and onward take a value of 1.

Furthermore, the EITC policy applies to women with children. A dummy variable *cChildren* is created to indicate the different groups that are compared: women with and without children. Women with no children take a value of 0 (non-treated group), whereas women with at least one child child take a value of 1 (treated group).

1.1 Difference-in-difference plots

As can be seen in the first and second plot of figure 1, there is an increase in the annual earnings and annual family income of women with children after the EITC was implemented. The annual earnings and annual family income of women without any children seems more or less constant, with a relatively small increase in 1995. The third plot of figure 1 shows an increase of women with children that are employed after the implementation of the EITC. On the other hand, there seems to a decrease of the number of women without children that are employed.

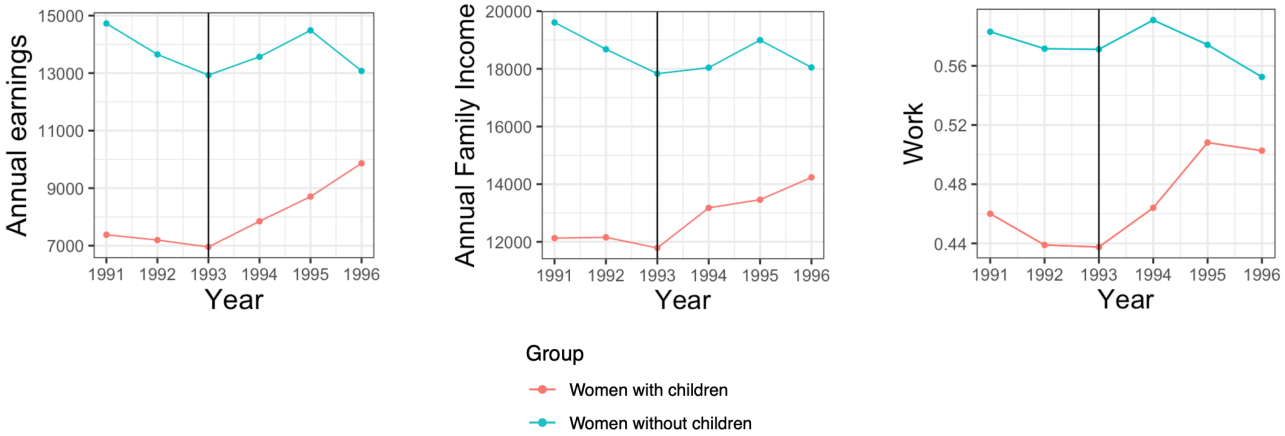

Figure 1: Plot of earn, finc and work

Based on the plots above, the EITC seems to have a positive effect on annual earnings (*earn*), annual family income (*finc*) and rate of employment (*work*).

1.2 Summary statistics

The summary statistics are shown in table 1. The data set contains 13,746 observations in the data set. Approximately 56,9% of the women have children. Moreover, the average number of children women in the data set have is 1.193, with a standard deviation of 1.382. Approximately 60.1% of the women are nonwhite. The average annual earnings and annual family income are and 10,432.48 and 15,255.32 US Dollars, with a standard deviation of 18,200.760 and 19,444.250 US Dollars respectively. The unearned income is on average 4.823 US Dollars with a standard deviation of 7.123 US Dollars. The women in the data set are on average 35 years old, with a standard deviation of 10.157 years. The years of education is on average 8.806 years with a standard deviation of 2.636 years. Approximately half of the women in the data set are employed, as the average employment rate equals 0.513.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
urate	13,746	6.762	1.462	2.600	5.700	7.700	11.400
children	13,746	1.193	1.382	0	0	2	9
nonwhite	13,746	0.601	0.490	0	0	1	1
finc	13,746	15,255.320	19,444.250	0.000	5,123.418	18,659.180	575,616.800
earn	13,746	10,432.480	18,200.760	0.000	0.000	14,321.220	537,880.600
age	13,746	35.210	10.157	20	26	44	54
ed	13,746	8.806	2.636	0	7	11	11
work	13,746	0.513	0.500	0	0	1	1
unearn	13,746	4.823	7.123	0	0	6.9	134
dPeriod	13,746	0.632	0.482	0	0	1	1
cChildren	13,746	0.569	0.495	0	0	1	1

1.3 Difference-in-difference effects

As can be seen in the table 2, the average annual earnings of employed women without children decreased with 695.997 US Dollars after the implementation of the EITC. However, the average annual earnings of women with children increased with 986.813 US Dollars after the EITC was implemented. The DiD effect of the EITC on annual earnings of employed women is $986.813 - (-695.997) = 1,682.81$.

Similarly, as can be seen in table 3 the average annual family income of employed women without children decreased with 940.239 US Dollars after the implementation of the EITC, whereas the average family income of women with children increased with 970.796 US Dollars. The DiD effect of the EITC on annual family income of employed women is $970.796 - (-940.239) = 1,911,035$.

The average rate of employment of women without any children decreased with 0.005 after the EITC was implemented, as shown in table 4. On the other hand, the average rate of employment of women with children increased with 0.026 after the implementation of the EITC. The DiD effect of the EITC on the number of women that are employed is $0.026 - (-0.005) = 0.031$.

Table 2: Average Annual Earnings

	dPeriod	Women without children (0)	Women with children (1)
Before	0	14,203.900	7,290.383
After	1	13,507.900	8,277.196
Difference Difference		−695.997	986.813

Table 3: Average Indicator Annual Family Income

	dPeriod	Women without children (0)	Women with children (1)
Before	0	19,159.190	12,140.900
After	1	18,218.950	13,111.690
Difference		−940.239	970.796

Table 4: Average Indicator Work Status

	dPeriod	Women without children (0)	Women with children (1)
Before	0	0.577	0.450
After	1	0.573	0.476
Difference		−0.005	0.026

1.4 Difference-in-difference regression

The estimated difference-in-difference models are shown in equation 1a-c.

$$Earn = \beta_0 + \beta_1 cChildren + \beta_2 dPeriod + \beta_3 cChildrendPeriod + \varepsilon_i \quad (1a)$$

$$Finc = \beta_0 + \beta_1 cChildren + \beta_2 dPeriod + \beta_3 cChildrendPeriod + \varepsilon_i \quad (1b)$$

$$Work = \beta_0 + \beta_1 cChildren + \beta_2 dPeriod + \beta_3 cChildrendPeriod + \varepsilon_i \quad (1c)$$

Table 5: DiD regressions for *earn*, *finc* and *work*

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
Constant	14,203.900***	19,159.190***	0.577***
	(387.548)	(414.751)	(0.011)
cChildren	−6,913.517***	−7,018.295***	−0.128***
	(510.988)	(546.857)	(0.014)
dPeriod	−695.997	−940.239*	−0.005
	(485.413)	(519.486)	(0.013)
cChildren:dPeriod	1,682.810***	1,911.035***	0.031*
	(642.099)	(687.171)	(0.018)
Observations	13,746	13,746	13,746
R^2	0.026	0.022	0.012
Adjusted R^2	0.026	0.022	0.012
Residual Std. Error (df = 13742)	17,965.670	19,226.750	0.497
F Statistic (df = 3; 13742)	121.691***	105.245***	54.906***
Note:			*p<0.1; **p<0.05; ***p<0.01

The results of the regression analyses of the DiD effect are shown in table 5. As can be derived from the table, the DiD effect of the EITC on annual earnings is 1,682.810. This indicates an increase in annual earnings of employed women with at least one child after the implementation of the EITC. The coefficient is significant (p <0.01).

Similarly, the effect of the EITC on annual family income is 1,911.035. The coefficient is significant (p <0.01). This indicates an increase in annual family income of employed women with at least one child after the EITC was implemented.

The DiD effect of the EITC on the rate of employment of women is 0.031. The coefficient is significant (p <0.1). This indicates an increase of employed women with at least one child after the implementation of the EITC. That is, an increase of approximately 3%.

1.4.1 Control variables

Control variables are added to the regressions to control for other variables that may have an influence on the DiD effect of the EITC. The resuls are shown in table 6.

A possible explanation for the obtained regression results is that state specific characteristics are not taken into account. Therefore, the state unemployment rate (*urate*) is added as a control variable, considering that women living in a state with a high unemployment rate are likely to be unemployed. Furthermore, the unearned income (*unearn*) is added as a control variable. Women with a high income gained through other resources than work, e.g. inheritance or investments, may be less inclined to work.

The years of education (*ed*) are added as a control variable. Years of education may impact work status, as women whom have enjoyed longer education may have an increased chance of employment. Additionally, the ethnicity indicator (*nonwhite*) is added as a control variable, as members of a minority group might be disadvantaged when going into a job interview.

It is expected that women with more children work less than women with a few children, as they may have a harder time balancing work and family life. Consequently, the number of children women have (*children*) is added as a control variable.

Table 6: DiD regression for earn, finc and work with control variables

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
Constant	$13,537.960^{***}$	$13,537.960^{***}$	0.693^{***}
	(1,080.611)	(1,080.611)	(0.029)
cChildren	$-4,057.741^{***}$	$-4,057.741^{***}$	-0.023
	(621.362)	(621.362)	(0.016)
dPeriod	-548.805	-548.805	-0.023^{*}
	(498.933)	(498.933)	(0.013)
urate	123.855	123.855	-0.015^{***}
	(114.415)	(114.415)	(0.003)
unearn	-25.552	974.448^{***}	-0.017^{***}
	(21.705)	(21.705)	(0.001)
ed	45.604	45.604	0.012^{***}
	(58.946)	(58.946)	(0.002)
nonwhite	$-1,008.532^{***}$	$-1,008.532^{***}$	-0.052^{***}
	(328.317)	(328.317)	(0.009)
children	$-1,314.326^{***}$	$-1,314.326^{***}$	-0.049^{***}
	(169.342)	(169.342)	(0.004)
cChildren:dPeriod	$1,699.596^{***}$	$1,699.596^{***}$	0.036^{**}
	(640.477)	(640.477)	(0.017)
Observations	13,746	13,746	13,746
R^2	0.032	0.151	0.097
Adjusted R^2	0.031	0.151	0.096
Residual Std. Error (df = 13737)	17,916.640	17,916.640	0.475
F Statistic (df = 8; 13737)	55.923^{***}	306.471^{***}	183.767^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

In the regression on annual earnings, the control variables *nonwhite* and *children* are significant (p <0.01). When controlling for ethnicity and the number of children, the DiD-effect of the EITC on annual earnings increased from 1,682.810 to 1,699.596. This effect remains significant (p <0.01).

In the regression on annual family income, the control variables *unearn*, *nonwhite* and *children* are significant (p <0.01). When controlling for the unearned income, ethnicity and the number of children, the DiD-effect of the EITC on annual family income decreased from 1,911.035 to 1,699.596. The effect remains significant (p <0.01).

In the regression on the work status of women, all the control variables are significant (p <0.01). The DiD-effect of the EITC increased from 0.031 to 0.036. The statistical significance of this effect increased as well (p <0.05). This indicates that, when the unemployment rate, unearned income, years of education, ethnicity

and the number of children are controlled for, the EITC led to a larger growth in employment of women with children (approximately 4%).

1.4.2 Robust standard errors

The Breusch-Pagan test for heteroskedasticity was used to assess whether robust standard errors are necessary. For the *earn*, *finc* and *work* models, the Breusch-Pagan test is equal to 142.79 ($p < 2.2e-16$), 142.79 ($p < 2.2e-16$) and 1254.6 ($p < 2.2e-16$) respectively. The statistical significance of these results indicate a correlation between the predicted Y values and the error terms. Heteroskedasticity is thus detected. This can be controlled for with White heteroskedasticity consistent standard errors or clustered standard errors (clustered on *state*).

2 Instrumental Variable Analysis: Effect of Compulsory Schooling on Wages

2.1 Bias of the estimated education effect

Two examples that could bias the effect.

Example One: the education level of parents could bias the estimated education effect, as parents with high education level will be more likely to urge their children to have and finish an education. On the other hand, children with parents who have higher education level will be more likely to get higher salary when they go to work, as they may have had a better education environment from their family. Therefore, the education level of parents could bias the education effect.

Example Two: the living area could be a factor that biases the education effect on wages as well. People from a rich area are likely to have more opportunities to get more years of education. Consequently, these people are likely to get higher wages when they go to work, as they may have a better network in this area, which causes a biased estimation of the education effect on wages.

2.2 Summary statistics

The summary statistics of the variables are shown in table 7. The data set consists of 329,509 observations. The age is on average 44.645 years with a standard deviation of 2.940 years. On average the years of education is 12.770 with a standard deviation of 3.281 years. The minimum years of education was 0 year and the maximum 20 years. The log weekly earnings is on average 5.900 with a standard deviation of 0.679. The average year born is approximately 1934 with a standard deviation of 2.905 years. Moreover, 289,111 of the residents live in an urban area, which is approximately 81.37% and 61,398 of the residents live elsewhere. Furthermore, 284,221 are married and 45,288 are not, which is 86.3% and 13.7% respectively.

Frequency table for *qob*, quarter of birth, is shown in table 8. Regarding the quarter of birth, 81,671 are born in the first quarter, 80,138 are in the second quarter, 86,856 are in the third quarter and 80,844 are in the fourth quarter, which is approximately 24,79%, 24.32%, 26.36% and 26.36%, respectively.

Table 7: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	329,509	44.645	2.940	40	42	47	50
educ	329,509	12.770	3.281	0	12	15	20
lnwage	329,509	5.900	0.679	−2.342	5.637	6.257	10.532
married	329,509	0.863	0.344	0	1	1	1
SMSA	329,509	0.186	0.389	0	0	0	1
yob	329,509	1,934.603	2.905	1,930	1,932	1,937	1,939

Table 8: Frequency table *qob*

quarter	Freq
1	81,671
2	80,138
3	86,856
4	80,844

2.3 Testing relevant criterion of instrumental variable Quarter of Birth (qob)

In order to test if the quarter of birth (*qob*) would be a good instrument variable for years of education (*edu*), a box plot is plotted. The box plot is shown in figure 2. However, the box plot hardly shows signs of a strong correlation between the instrumental variable *qob* and the causal variable *educ*. A weak instrument test is performed. The test has statistically significant results (p <0.01). We can reject the null-hypothesis, which indicates there should be a strong correlation between *qob* and *educ*. Therefore, we can claim that variable *qob* meets the relevant criterion of instrumental variable.

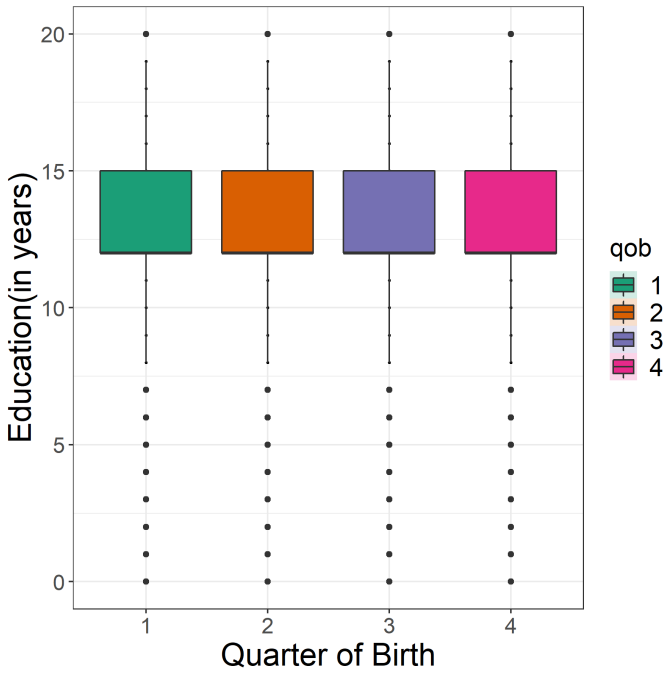

Figure 2: Box plot educ and qob.

2.4 Instrumental variable regression analysis

IV regression analysis of the effect of education (in years) on log wages using quarter of birth as instrument variable: The quarter of birth partially influences the years of education. We use the variation that *qob* has on education to estimate the effect of education on weekly earnings.

The results of the IV regression are shown in table 9. Model 1 shows that an extra year of education for people in the United States born between 1930-1939, leads to a 0.103% increase in weekly earnings. With a $p < 0.05$, the estimated coefficient is statistically significant. Moreover, model 2 shows the results of the model when adding marital status and an indicator of living situation as control variables. This model presents a decrease on the estimated coefficient from 0.103% to 0.1% with slightly lower standard errors still significant at 5% significance level.

Moreover, a regression with robust standard errors has been performed. The results are shown in table 10, model 3. However, the use of robust standard errors does no have a critical effect on the statistical inference of the estimated coefficient of $Education$ nor the control variables $Married$ and $SMSA$.

Table 9: IV Regressions and Robust Standard Errors

	Dependent variable:		
	lnwage		
	(1)	(2)	(3)
Constant	4.590***	4.425***	4.425***
	(0.249)	(0.248)	(0.248)
educ	0.103***	0.100***	0.100***
	(0.020)	(0.019)	(0.020)
married1		0.255***	0.255***
		(0.006)	(0.007)
SMSA1		−0.148***	−0.148***
		(0.022)	(0.022)
Observations	329,509	329,509	329,509
R^2	0.094	0.120	0.120
Adjusted R^2	0.094	0.120	0.120
Residual Std. Error	0.646 (df = 329507)	0.637 (df = 329505)	0.637 (df = 329505)
Note:			*p<0.1; **p<0.05; ***p<0.01

2.5 Comparison with OLS regressions

Following the previous analysis, the same regressions are conducted with OLS as with IV. The results are shown in table 10, model 1 and 2.

In order to determine for which of the models, OLS or IV, there is a relative preference, a Wu-Hausman test should be performed. A significant Wu-Hausman test indicates that the IV model is preferred. For both model 1 and 2 the Wu-Hausman's test is equal to 2.722 ($p = 0.099$) and 2.988 ($p = 0.0839$), respectively. The results are marginally significant (at 10%), which suggest that there is a mild preference for using the IV estimator.

A model is over-identified if the model has more instrumental variables than endogenous variables. In order to test if the over-identified restriction violates the independence assumption of the instruments a Sagan-Hansen test should be preformed. It is important to note that the Sagan-Hansen test can only be formed when the model is indeed over-identified. The Sagan-Hansen test tests whether by adding one variable the model violates the assumption that the instruments are uncorrelated with the error term. A significant Sagan-Hansen test indicates that the independence violation is violated. However, both of the IV models do not have more instrumental variables than endogenous variables. Thus, the Sagan-Hansen test cannot be performed. Therefore, an over identified model is defined with both *qob* and *age* as instrumental variables. The results of this model are shown in table 11, model 3. The Sargan-Hansen's test is equal to 22.86 ($p = 4.31e - 05$), which is significant (at 1%). This result indicates that the validity (exogeneity) of the instruments is violated.

Table 10: OLS Regression and IV Regression (over-identification)

	Dependent variable:		
	lnwage		
	OLS		*instrumental variable*
	(1)	(2)	(3)
Constant	4.995***	4.847***	5.511***
	(0.004)	(0.005)	(0.076)
educ	0.071***	0.067***	0.015**
	(0.0003)	(0.0003)	(0.006)
married1		0.265***	0.280***
		(0.003)	(0.004)
SMSA1		−0.185***	−0.243***
		(0.003)	(0.007)
Observations	329,509	329,509	329,509
R^2	0.117	0.145	0.083
Adjusted R^2	0.117	0.145	0.083
Residual Std. Error	0.638 (df = 329507)	0.628 (df = 329505)	0.650 (df = 329505)
F Statistic	43,782.560*** (df = 1; 329507)	18,646.640*** (df = 3; 329505)	
Note:			*p<0.1; **p<0.05; ***p<0.01

2.6 Possible concerns instrumental variable identification strategy

There are several possible causes of concern that could render the instrumental variable identification strategy invalid. Firstly, in the US children are able to start with school when their 5 years old, however, it is not mandatory until they are of age 6. Therefore, children do not necessarily start in the same year even if they turn 6 in the same year. Thus, some children could have more years of education because they started school earlier. Moreover, it is a possibility that children drop out of school even before they reach the legal age due to certain circumstances which will affect the years of education regardless of what is mandatory. Next to that, the legal age for dropouts in America varies per state. Thus, students are able to drop out of school at ages varying from 16 to 18 depending on where they live, resulting in a difference of two years of mandatory education. Furthermore, it is also possible that a person who dropped out of school decides to restart their education in a later period in time. Therefore, dropping out does not imply that the years of education cannot increase once one drops out of school. Concluding, considering the possible causes of concern the measurement of school years for dropout students could be inaccurate.

3 Panel data modeling: time to export

3.1 Population model

The dependent variable of the model is $ExportTime$. The independent variables are $ExportCost$, $ExportGoodsServices$, $GDPPerCap$ and $MerchanidiseGDP$. The definitions of the variables are shown in table 11. The population regression model is shown in equation 2.

$$TimeExport_{it} = \alpha_{it} + \beta_{1,it} CostExport_{it} + \beta_{2,it} ExportGoodsServices_{it} + \beta_{3,it} GDPPerCap_{it} + \beta_{4,it} MerchandiseGDP_{it} + u_{it} \quad (2)$$

Table 11: Variables

Statistic	Variable
ExportCost	Cost to export, documentary compliance (US$).
ExportTime	Time to export, border compliance (hours).
ExportGoodsServices	Exports of goods and services (% of GDP).
GDPPerCap	GDP per capita (current US$).
MerchandiseGDP	Merchandise trade (% of GDP).

3.2 Summary statistics

The original data set consists of 217 unique countries. However, in order to perform the regression analyses a balanced data set needs to be constructed. A balanced data set contains only the countries for which all records have been completely observed during the period 2014-2019. All countries that do not meet this criterion have been deleted form the data set. As a result 75 countries have been deleted and the balanced data set contains observations for 142 unique countries.

The summary statistics of the dependent and independent variables are shown in table 12. The data set contains data on 142 unique countries for a six year period, between 2014-2019, resulting in 852 observations. The cost of export are on average 110.207 US$ with a standard deviation of 169.287 US$. Time of export is on average 54.546 hours with a standard deviation of 59.884 hours. Export of goods and services is on average 44.851% of the GDP with a standard deviation of 33.057%. The GDP per capita has a mean of 14,895.070 US$ and a standard deviation of 20,263.220 US$. Merchandise trade in on average 68.093% of the GDP and had a standard deviation of 42.798%.

Table 12: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
CostExport	852	110.207	169.287	0.000	35.000	138.900	1,800.000
TimeExport	852	54.546	59.884	0	9	78	515
ExportGoodsServices	852	44.851	33.057	5.710	24.916	52.599	221.197
GDPPerCap	852	14,895.070	20,263.220	261.247	1,972.168	17,723.010	118,823.600
MerchandiseGDP	852	68.093	42.798	17.065	42.067	81.031	385.953

3.3 Plot of Time of Export and Cost of Export

Figure 3 shows the total variation between dependent variable $TimeExport$ and explanatory variable $CostExport$ for a subset of all countries included in the analyses. The graph shows that there is a linear relationship between these variables. Moreover, it can be observed that India has the highest variation in $TimeExport$ and $CostExport$. Similarily, Brazil, Thailand and China show some variation in $TimeExport$ across the years studied. The between variation is shown by the color country labels which present a high variation between country averages regarding both variables. $TimeExport$ varies from 0 to 110 hours while $CostExport$ varies from 0 to 275 US$ depending on the country.

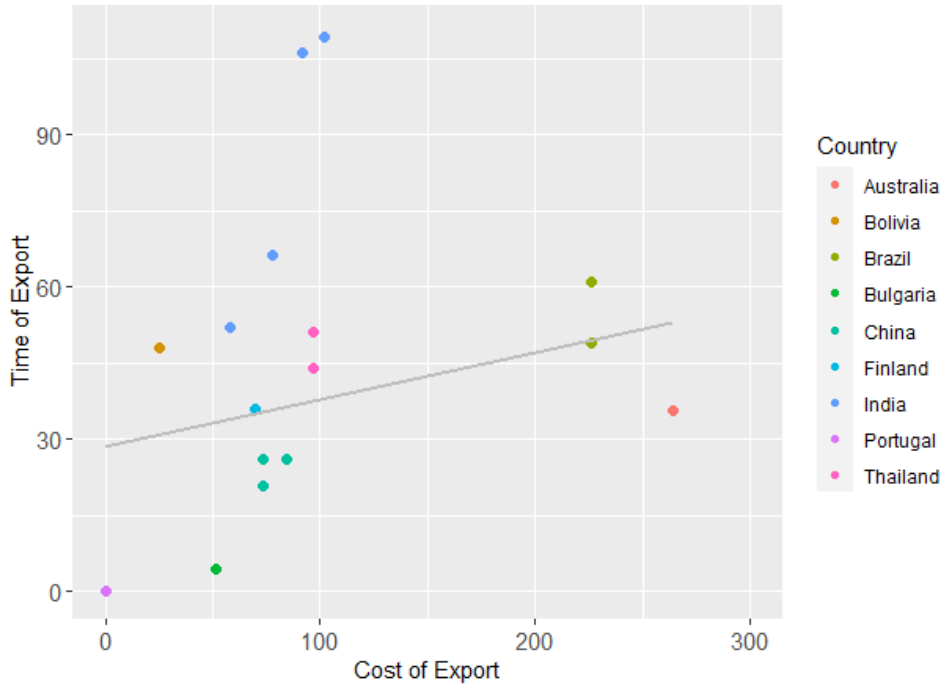

Figure 3: Relationship diagram TimeExport and CostExport

3.4 Pooled, Between, Fixed effect and Random effect regressions

Table 13

	Dependent variable:			
	TimeExport			
	OLS		*panel linear*	
	(1)	(2)	(3)	(4)
Constant	71.059***	71.472***		57.212***
	(3.883)	(9.519)		(6.752)
GDPPerCap	−0.001***	−0.001***	0.00001	−0.0003**
	(0.0001)	(0.0003)	(0.0002)	(0.0001)
CostExport	0.093***	0.093***	0.125***	0.113***
	(0.011)	(0.026)	(0.047)	(0.023)
ExportGoodsServices	0.309***	0.330	−0.150	−0.110
	(0.087)	(0.215)	(0.109)	(0.096)
MerchandiseGDP	−0.369***	−0.383**	−0.017	−0.073
	(0.060)	(0.149)	(0.068)	(0.062)
Observations	852	142	852	852
R^2	0.255	0.264	0.014	0.050
Adjusted R^2	0.251	0.243	-0.189	0.045
Residual Std. Error	51.817 (df = 847)	51.497 (df = 137)		
F Statistic	72.402*** (df = 4; 847)	12.312*** (df = 4; 137)	2.487** (df = 4; 706)	44.379***
Note:				*p<0.1; **p<0.05; ***p<0.01

Results for Pooled (1), Between estimator (2), Fixed (3) and Random (4) effect models are shown in table 13. The pooled model shows that there is a positive relationship between dependent variable $TimeExport$ and explanatory variables $CostExport$ and $ExportGoodsServices$. A 1% increase in this variables leads to a 0.093% and a 0.309% increase in Time of Export respectively. Variables $GDPPerCap$ and $MerchandiseGDP$ show a negative relationship as a 1% increase in this variables leads to a decrease of -0.001% and -0.369% in Time of Export respectively. With $p < 0.01$ coefficients of this model are all statistically significant. When comparing to the between estimator model, it is know that the pooled model is a weighted regression between within and between variations, the later having a much higher weight. Consequently, there is no critical variation between

the coefficients of the models. However, it is observed that in the between model there is an increase in standard errors, which decreases the significance level of all variables, specially that of $ExportGoodsServices$ which is no longer significant.

Moreover, when looking at the Fixed Effect model, a more drastic difference in results can be observed. This model takes into account the impact of country level characteristics by controlling for unobserved time-invariant variables which are potentially correlated with the explanatory variables. The only variable with an statistically significant coefficient is $CostExport$ which compared to the pooled and between model, has a higher coefficient. This is attributed to $CostExport$ being more strongly correlated to $TimeExport$ for some specific countries overtime. The same situation can be observed for the rest of the variables coefficients which are significant where accounting for all countries, but not when controlling for country specific characteristics. In addition, this model does not include a constant as the constant differs per country.

Last, Random Effect model shows different results when compared to the pooled model. Although they both use weighted regression between within and between variation, pooled model uses ordinary least squares and Random Effect uses generalized least squares, which takes into account the panel data information $Country$ and $Year$ which OLS method disregards. This makes it a better estimator as it accounts for different sources of variation. This model shows that only $GDPPerCap$ and $CostExport$ have statistically significant coefficients. Compared to the pooled results, there is a lower decrease in $TimeExport$ given a 1% change in $GDPPerCap$ and a higher increase in $TimeExport$ given a 1% change in $CostExport$. When compared to Fixed Effect model, this model should be used if the time-invariant unobserved effects are uncorrelated to the dependent and independet variables. A test to conclude which is the best model estimator will be performed in the following section.

3.5 Tests to evaluate the most suitable model

In order to decide which of the estimated models is the preferred specification for the analysis of time to export, the different models are compared. First, in order to compare the fixed effect model versus the pooled regression model a F-test is performed. F_{obs} is equal to 125.75 ($p < 0.001$), therefore, the null hypothesis is rejected. The result of the F-test implies that the pooled model is rejected in favor for the fixed country effects model.

Second, Hausman's test compares the random model and the fixed model. The result of Hausman's test is equal to 14.87 ($p = 0.004975$). This result is significant (at 1%), which indicates correlation between disturbance and explanatory variables, therefore, only the fixed effect model is consistent.

Concluding, after reviewing the results of both the F-test and the Hausman test the fixed effect model is the most suitable for the analysis of time to export.

Appendix Difference-in-Difference Analysis: Female Labor Participation

Cleaning data

```
# Load libraries
library(tidyverse)
library(stargazer)
library(dagitty)
library(gridExtra)
library(tinytex)
library(ggplot2)
library(tidyr)
library(dplyr)
library(plyr)
library(reshape2)
library(sandwich)

# Set directories
dir <- "/Users/valeriemaasdamme/Desktop/BAM_ASP_A2"
dirProg <- paste0(dir, "/programs/")
dirData <- paste0(dir, "/data/")
dirResults <- paste0(dir, "results/")

# Load csv file
dfDiD <- read.csv(file=paste0(dirData, "DiD_dataset.csv"))

# Make dummy variable for period and children
dfDiD$dPeriod = ifelse(dfDiD$year >= 1993, 1, 0)
dfDiD$cChildren = ifelse(dfDiD$children >= 1, 1, 0)
```

1 Visual evidence

```
## Variable == earn
# Compute group average for both women with and without children
earn.agg = aggregate(dfDiD$earn, list(dfDiD$year, dfDiD$cChildren == 1),
                     FUN = mean, na.rm = TRUE)

# Rename column names
names(earn.agg) = c("Year","Children","Earn")

# New variable with group name
earn.agg$Group[1:6] = "Women without children"
earn.agg$Group[7:12] = "Women with children"
```

```
# Make and save plot
Earn.plot <- qplot(Year, Earn, data=earn.agg, geom=c("point","line"),
                   colour = Group, xlab="Year", ylab="Annual earnings") +
  geom_vline(xintercept = 1993) +
  theme_bw(base_size = 15) +
  theme(axis.title = element_text(size = 19))
Earn.plot
```

```
ggsave(file="Earn.pdf", width=7, height=4)

# Variable == finc
# Compute group average for both women with and without children
finc.agg = aggregate(dfDiD$finc, list(dfDiD$year, dfDiD$cChildren == 1),
                     FUN = mean, na.rm = TRUE)
names(finc.agg) = c("Year","Children","Finc")

# New variable with group name
finc.agg$Group[1:6] = "Women without children"
finc.agg$Group[7:12] = "Women with children"

# Make and save plot
Finc.plot <- qplot(Year, Finc, data=finc.agg, geom=c("point", "line"),
                   colour = Group, xlab="Year", ylab="Annual Family Income") +
  geom_vline(xintercept = 1993) +
  theme_bw(base_size = 15) +
  theme(axis.title = element_text(size = 19))
Finc.plot
```

```
ggsave(file="Finc.pdf", width=7, height=4)

# Variable == work
# Compute group average for both women with and without children
work.agg = aggregate(dfDiD$work, list(dfDiD$year, dfDiD$cChildren == 1),
                     FUN = mean, na.rm = TRUE)
names(work.agg) = c("Year","Children","Work")

# New variable with group name
work.agg$Group[1:6] = "Women without children"
work.agg$Group[7:12] = "Women with children"

# Make and save plot
Work.plot <- qplot(Year, Work, data=work.agg, geom=c("point", "line"),
                   colour = Group, xlab="Year", ylab="Work") +
  geom_line() +
  geom_vline(xintercept = 1993) +
  theme_bw(base_size = 15) +
  theme(axis.title = element_text(size = 19))
Work.plot
```

```
ggsave(file="Work.pdf", width=7, height=4)
```

2 Summary statistics of the data

```
# Convert categorical variables to vector
dfDiD$year <- as.factor(dfDiD$year)
dfDiD$state <- as.factor(dfDiD$state)
# dfDiD$nonwhite <- as.factor(dfDiD$nonwhite)
# dfDiD$dWork <- as.factor(dfDiD$work)
# dfDiD$cChildren = as.factor(dfDiD$cChildren)
# dfDiD$dPeriod <- as.factor(dfDiD$dPeriod)

# Check for converting categorical variables
str(dfDiD)

# Tabulate summary statistics
stargazer(dfDiD, title="Summary Statistics")

# Generate frequency tables categorical variables
year_freq <- as.data.frame(table(dfDiD$year))
state_freq <- as.data.frame(table(dfDiD$state))
nonwhite_freq <- as.data.frame(table(dfDiD$nonwhite))
work_freq <- as.data.frame(table(dfDiD$work))
dPeriod_freq <- as.data.frame(table(dfDiD$dPeriod))
cChildren_freq <- as.data.frame(table(dfDiD$cChildren))

# Tabulate frequency tables with stargazer
stargazer(year_freq, summary=FALSE, title="year")
stargazer(state_freq, summary=FALSE, title="state")
stargazer(nonwhite_freq, summary=FALSE, title="nonwhite")
stargazer(work_freq, summary=FALSE, title="work")
stargazer(dPeriod_freq, summary=FALSE, title="period")
stargazer(cChildren_freq, summary=FALSE, title="cChildren")
```

3 Difference in difference effect

```
avgEarn <- ddply(dfDiD, .(dPeriod, cChildren), summarise,
                 avgEarn = mean(earn, na.rm=TRUE))

avgFinc <- ddply(dfDiD, .(dPeriod, cChildren), summarise,
                 avgFinc = mean(finc, na.rm=TRUE))

avgWork <- ddply(dfDiD, .(dPeriod, cChildren), summarise,
                 avgWork = mean(work, na.rm=TRUE))

# Remodel the avg table from long to wide, add  additional row for the
# difference in averages and rename the rows
avgtable.Earn <- dcast (avgEarn, dPeriod ~ cChildren, value.var = "avgEarn")
avgtable.Earn <- rbind(avgtable.Earn, avgtable.Earn[2,] - avgtable.Earn[1,])
rownames(avgtable.Earn) <- c("Before", "After", "Difference")
colnames(avgtable.Earn) <- c("dPeriod", "Women without children (0)",
                             "Women with children (1)")
avgtable.Earn[3, "dPeriod"] <- NA
```

```r
avgtable.Finc <- dcast (avgFinc, dPeriod ~ cChildren, value.var = "avgFinc")
avgtable.Finc <- rbind(avgtable.Finc, avgtable.Finc[2,] - avgtable.Finc[1,])
rownames(avgtable.Finc) <- c("Before", "After", "Difference")
colnames(avgtable.Finc) <- c("dPeriod", "Women without children (0)",
                             "Women with children (1)")
avgtable.Finc[3, "dPeriod"] <- NA

avgtable.Work <- dcast (avgWork, dPeriod ~ cChildren, value.var = "avgWork")
avgtable.Work <- rbind(avgtable.Work, avgtable.Work[2,] - avgtable.Work[1,])
rownames(avgtable.Work) <- c("Before", "After", "Difference")
colnames(avgtable.Work) <- c("dPeriod", "Women without children (0)",
                             "Women with children (1)")
avgtable.Work[3, "dPeriod"] <- NA

# Tabulate DiD tables with stargazer
stargazer(avgtable.Earn, summary=FALSE, align = TRUE,
          title = "Average Annual Earnings")
stargazer(avgtable.Finc, summary=FALSE, align = TRUE,
          title = "Average Indicator Annual Family Income")
stargazer(avgtable.Work, summary=FALSE, align = TRUE,
          title = "Average Indicator Work Status")
```

4 Regression analysis

```r
mdlEarn <- earn ~ cChildren + dPeriod + cChildren:dPeriod
rsltOLSEarn <- lm(mdlEarn, data=dfDiD)

mdlFinc <- finc ~ cChildren + dPeriod + cChildren:dPeriod
rsltOLSFinc <- lm(mdlFinc, data=dfDiD)

mdlWork <- work ~ cChildren + dPeriod + cChildren:dPeriod
rsltOLSWork <- lm(mdlWork, data=dfDiD)

stargazer(rsltOLSEarn, rsltOLSFinc, rsltOLSWork,
          intercept.bottom = FALSE, align = TRUE, no.space=TRUE)
```

4.1 Regression with control variables

```r
# DiD regression earn with control variables
mdl.control.earn <- earn ~ cChildren + dPeriod + cChildren:dPeriod + urate +
  unearn + ed + nonwhite + children
rsltOLS.control.earn <- lm(mdl.control.earn, data=dfDiD)

# DiD regression earn with control variables
mdl.control.finc <- finc ~ cChildren + dPeriod + cChildren:dPeriod + urate +
  unearn + ed + nonwhite + children
rsltOLS.control.finc <- lm(mdl.control.finc, data=dfDiD)

# DiD regression earn with control variables
```

```
mdl.control.work <- work ~ cChildren + dPeriod + cChildren:dPeriod + urate +
  unearn + ed + nonwhite + children
rsltOLS.control.work <- lm(mdl.control.work, data=dfDiD)

stargazer(rsltOLS.control.earn, rsltOLS.control.finc, rsltOLS.control.work,
           intercept.bottom = FALSE, align = TRUE, no.space=TRUE, type="text")

stargazer(rsltOLS.control.earn, rsltOLS.control.finc, rsltOLS.control.work,
           intercept.bottom = FALSE, align = TRUE, no.space=TRUE,
          title="DiD regression for earn, finc and work with control variables")
```

4.2 Regression with robust standard errors

```
#Plots
ggplot(data = data.frame(fit = fitted(rsltOLS.control.earn),
  rsid = residuals(rsltOLS.control.earn)),
  aes(fit, rsid)) +
  geom_point() +
  stat_smooth(se = F) +
  theme_bw() +
  labs(x = "Results OLS Fitted") +
  labs(y = "Residuals")
```

```
ggplot(data = data.frame(fit = fitted(rsltOLS.control.finc),
  rsid = residuals(rsltOLS.control.finc)),
  aes(fit, rsid)) +
  geom_point() +
  stat_smooth(se = F) +
  theme_bw() +
  labs(x = "Results OLS Fitted") +
  labs(y = "Residuals")
```

```
ggplot(data = data.frame(fit = fitted(rsltOLS.control.work),
  rsid = residuals(rsltOLS.control.work)),
  aes(fit, rsid)) +
  geom_point() +
  stat_smooth(se = F) +
  theme_bw() +
  labs(x = "Results OLS Fitted") +
  labs(y = "Residuals")
```

```
# BP test for the three models
lmtest::bptest(rsltOLS.control.earn)
lmtest::bptest(rsltOLS.control.finc)
lmtest::bptest(rsltOLS.control.work)
```

section2_IVA

Eline van Groningen, Paola Priante, Valery Maasdamme, Yuhu Wang

9/25/2020

Instrumental Variable Analysis: Effect of Compulsory Schooling on Wages

Downloading the libraries

```
# Load libraries
library(tidyverse)
library(stargazer)
library(dagitty)
library(gridExtra)
library(tinytex)
library(stargazer)
library(AER)
library(ivpack)

# Set working director
setwd("C:/Users/Administrator/Desktop/NewStart/Courses/AdvancedStatisti
csandProgramming/assignment2/github/BAM_ASP_A2/data")

# Load csv and generate subset containing only variables for interest
da.IV <- read.csv("IV_dataset.csv", header = TRUE)
da.IV <- subset(da.IV, select = c("age", "educ", "lnwage", "married", "
qob",
                                  "SMSA", "yob"))

## Subset the data set so that we could focus on the variables above ac
cording to the order
da.IV <- read.csv("IV_dataset.csv", header = TRUE)
da.IV <- subset(da.IV,select = c("age","educ","lnwage","married","qob",
"SMSA","yob"))
## Subset the dataset so that we could focus on the variables above acc
ording to the order

stargazer(da.IV,type = "text")
summary(as.factor(da.IV$married))

# Convert to factor variables
da.IV$married <- as.factor(da.IV$married)
da.IV$qob <- as.factor(da.IV$qob)
da.IV$SMSA <- as.factor(da.IV$SMSA)
```

```r
da.IV$yob <- as.factor(da.IV$yob)

# To change those variables which should be factor variables into facto
r variables
g1.1 <- ggplot(data = da.IV, aes(qob, educ)) +
  geom_point(size = 0.5) +
  geom_smooth(method = "lm", color = "blue", alpha = 0.2)  +
  theme_bw() +
  labs(caption = "Figure 2.1") +
  geom_boxplot() +
  theme(plot.caption =  element_text(hjust = 0.5, size = 12, face = "bo
ld")) +
  labs(x = "Quarter of Birth", y = "Education(in years)")
g1.1



rsltIV <- ivreg(lnwage ~ educ|qob,data = da.IV)
summary(rsltIV, diagnostics = TRUE)

library(ivreg)
rslt2SLS.A <- ivreg(lnwage ~ educ | qob, data=da.IV)
summary(rslt2SLS.A)
stargazer(rslt2SLS.A, type= "text")

rslt2SLS.B <- ivreg(lnwage ~ educ + married + SMSA | married + SMSA + q
ob,
                    data=da.IV)
summary(rslt2SLS.A)
stargazer(rslt2SLS.A, rslt2SLS.B)

#Robust standard errors
modelIV <- ivreg(lnwage ~ educ + married + SMSA | married + SMSA + qob
,
                 data=da.IV)
summary(modelIV)

#Standard  errors (superfluous  in the  case of  seBasic)
seBasic  <- sqrt(diag(vcov(modelIV)))
seWhite  <- sqrt(diag(vcovHC(modelIV , type="HC0")))
library(vcov)
# Make  table  with  stargazer
stargazer(modelIV , modelIV ,align=TRUE , no.space=TRUE ,intercept.bott
om = FALSE ,se = list(seBasic , seWhite), type= "text")

da.IV_sub <- subset(da.IV,select = c("age", "educ", "lnwage", "married"
, "qob",
                                     "SMSA", "yob"))

# Convert to factor variables
da.IV_sub$married <- as.factor(da.IV_sub$married)
```

```r
da.IV_sub$qob <- as.factor(da.IV_sub$qob)
da.IV_sub$SMSA <- as.factor(da.IV_sub$SMSA)
da.IV_sub$yob <- as.factor(da.IV_sub$yob)

# Define OLS models
rsltOLS.A <- lm(lnwage ~ educ, data=da.IV_sub)
rsltOLS.B <- lm(lnwage ~ educ + married + SMSA, data=da.IV_sub)

# Define IV model
rsltSLS.A <- ivreg(lnwage ~ educ | qob, data=da.IV_sub)
rsltSLS.B <- ivreg(lnwage ~ educ + married + SMSA | married + SMSA + qo
b,
                   data=da.IV_sub)
rsltSLS.C <- ivreg(lnwage ~ educ + married + SMSA | married + SMSA + ag
e + qob,
                   data=da.IV_sub)

# Generate table containing both models
stargazer(rsltOLS.A, rsltOLS.B, rsltSLS.A, rsltSLS.B, rsltSLS.C, type="
text")

# Test for violation over-identification
summary(rsltSLS.A, diagnostics = TRUE)
summary(rsltSLS.B, diagnostics = TRUE)
summary(rsltSLS.C, diagnostics = TRUE)
```

section3_PDM

Eline van Groningen, Paola Priante, Valery Maasdamme, Yuhu Wang

9/25/2020

```
# Load libraries
library(tidyverse)
library(stargazer)
library(wbstats)
library(ggplot2)
library(plyr)
library(plm)

# Load world bank data
dfExport <- wb_data(indicator=c("IC.EXP.TMBC",       # Time to export
                                "NY.GDP.PCAP.CD",    # GDP per capita
                                "TG.VAL.TOTL.GD.ZS", # Merchandise trad
e % GDP
                                "NE.EXP.GNFS.ZS",    # Exports of goods
 and services (% of GDP)
                                "IC.EXP.CSDC.CD"),   # Cost to export
                                country = "countries_only",
                                start_date = 2014,
                                end_date = 2019)

# Rename column names
colnames(dfExport)[colnames(dfExport) == "date"]              <- "Year"
colnames(dfExport)[colnames(dfExport) == "country"]           <- "Count
ry"
colnames(dfExport)[colnames(dfExport) == "date"]              <- "Year"
colnames(dfExport)[colnames(dfExport) == "IC.EXP.TMBC"]       <- "TimeE
xport"
colnames(dfExport)[colnames(dfExport) == "NY.GDP.PCAP.CD"]    <- "GDPPe
rCap"
colnames(dfExport)[colnames(dfExport) == "TG.VAL.TOTL.GD.ZS"] <- "Merch
andiseGDP"
colnames(dfExport)[colnames(dfExport) == "NE.EXP.GNFS.ZS"]    <- "Expor
tGoodsServices"
colnames(dfExport)[colnames(dfExport) == "IC.EXP.CSDC.CD"]    <- "CostE
xport"

# Subset complete observations, and implement an admittedly arbitrary
# observation period
dfExport.sub <- dfExport[complete.cases(dfExport),]

# Generate list with all countries with complete observations
```

```
complete <- dfExport.sub %>%
  dplyr::count(Country) %>%
  filter(n == 6)
completeCountry <- as.vector(complete$Country)

# Generate data frame only containing countries with complete observati
ons
dfExport.sub.cmplt <- dfExport.sub %>%
  filter(Country %in% completeCountry)

# Convert to data frame
dfExport.sub.cmplt <- as.data.frame(dfExport.sub.cmplt)

# Generate table with summary statistics
stargazer(dfExport.sub.cmplt)

# Plot Cost Export

subCountries <- c("Australia", "Bolivia", "Brazil", "Portugal", "Thaila
nd",
                  "Zimbabwe", "Bangladesh", "Bulgaria", "China", "Denma
rk",
                  "France", "Finland", "India")

dfExport.sub.cmplt <-
  dfExport.sub.cmplt[dfExport.sub.cmplt$Country %in% subCountries,]

ggplot(dfExport.sub.cmplt, aes(x=CostExport, y=TimeExport))+
  #add the annual outcomes coloured by Country
  geom_point(aes(color=Country), size=1)+
  #add regression lines for the countries
  geom_smooth(method="lm", se=FALSE, colour="dark grey")+
  #label the axis
  xlim(0, 300) + ylim(0, 70)+
  xlab("Cost of Export")+
  ylab("Time of Export")+
  theme(axis.title= element_text(size=rel(1)),
        axis.text= element_text(size=rel(1)))+
  guides(colour = guide_legend(override.aes = list(size=1)))
```

Preparing data for regression

```
# Determine country averages of the included variables, as well as the
number of
# non missing observations during the selected observation period
dfExport.sub.cmplt.avg <-
  ddply(dfExport.sub.cmplt, .(Country), summarise,
        avg.TimeExport   = mean(TimeExport, na.rm=TRUE),
```

```
        avg.GDPPerCap    = mean(GDPPerCap, na.rm=TRUE),
        avg.CostExport   = mean(CostExport, na.rm=TRUE),
        avg.ExportGoodsServices     = mean(ExportGoodsServices, na.rm=T
RUE),
        avg.MerchandiseGDP   = mean(MerchandiseGDP, na.rm=TRUE),
        numValid         = length(Country))

# Merge averages in dfWorld.avg with dfWorld.sub (this can be done with

# 'mutate', but then the concise data frame with country average will n
ot be
# made available
dfExport.sub.cmplt <- merge(dfExport.sub.cmplt, dfExport.sub.cmplt.avg,

                            by="Country")

attach(dfExport.sub.cmplt)
dfExport.sub.cmplt$diff.TimeExport   <- TimeExport   - avg.TimeExport
dfExport.sub.cmplt$diff.GDPPerCap    <- GDPPerCap    - avg.GDPPerCap
dfExport.sub.cmplt$diff.CostExport   <- CostExport   - avg.CostExport
dfExport.sub.cmplt$diff.ExportGoodsServices <- ExportGoodsServices  -
  avg.ExportGoodsServices
dfExport.sub.cmplt$diff.MerchandiseGDP   <- MerchandiseGDP   -
  avg.MerchandiseGDP
detach(dfExport.sub.cmplt)
```

Pooled Regression

```
#Formulate the model (very ad hoc)
mdlA <- TimeExport ~ GDPPerCap + CostExport + ExportGoodsServices +
  MerchandiseGDP

#Make between and within group data frames

#For convenience two datasets are made that contain the model
#variables for the within group differences and the between
#group difference

# find the variable of interest
mdlvars <- all.vars(mdlA)
mdlvars.avg <- paste0("avg.", mdlvars)
mdlvars.diff <- paste0("diff.", mdlvars)

# Select variables from the data frames
dfExport.between <- dfExport.sub.cmplt.avg[mdlvars.avg]
dfExport.within <-dfExport.sub.cmplt[mdlvars.diff]

# Rename column names in order to make use  of the same model specifica
tion
```

```
# mdlA, and to conveniently merge the regression objects in stargazer

colnames(dfExport.within) <-
  gsub("diff\\.", "", colnames(dfExport.within))
colnames(dfExport.between) <-
  gsub("avg\\.", "", colnames(dfExport.between))

## Estimation of the pooled model
rsltPool <- lm(mdlA, data= dfExport.sub.cmplt)
summary(rsltPool)
stargazer::stargazer(rsltPool, align=TRUE, no.space=TRUE,
                     intercept.bottom=FALSE, type="text")
```

Between regression

```
rsltwithin <- lm(mdlA, data= dfExport.within)
summary(rsltwithin)
rsltBetween <- lm (mdlA, data= dfExport.between)
summary(rsltBetween)

stargazer::stargazer(rsltPool, rsltBetween, aling=TRUE, no.space=TRUE,
                     intercept.bottom= FALSE, type= "text")
```

Fixed Effect Regression

```
rsltFE.Country <- plm(mdlA, data= dfExport.sub.cmplt,
                      index= c("Country", "Year"), model="within")
#Tabulate the results
summary(rsltFE.Country)
stargazer::stargazer(rsltPool, rsltFE.Country, align=TRUE, no.space=TRU
E,
                     intercept.bottom=FALSE, type="text")
#Explore the estimated intercepts
summary(fixef(rsltFE.Country, type="dmean"))
```

Random Effect Regression

```
#Estimate random effect model ('random')
rsltRE.Country <- plm(mdlA, data=dfExport.sub.cmplt,
                      index=c("Country", "Year"), model= "random")

#Tabulate the results
summary(rsltRE.Country)
stargazer::stargazer(rsltPool, rsltFE.Country, rsltRE.Country,
                     align=TRUE, no.space=TRUE, intercept.bottom=FALSE,

                     type="text")

# Evaluate the fixed effects model versus the pooled regression model
# Last minute of tutorial #4 Panel Data
# An insignificant tests tells that all models are consistent
```

```r
# A significant tests rejects the hypothesis in favor of the fix effect
s model
pFtest(rsltFE.Country, rsltPool)

# How do we now when to use fixed and when to use random?
# Hausman test: compare random and fixed effects models
# Under H0, no correlation between disturbance and explanatory variable
s,
# both RE and FE are consistent (though FE is not efficient), under H1,

# correlation between disturbance, only FE consistent
# Last two minutes of tutorial #5 Panel Data
phtest(rsltFE.Country, rsltRE.Country)
```