Exercise 1 (40 points)

In their paper "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration" published in the Journal of Political Economy in 2014, Abramitzky, Boustan, and Eriksson study the assimilation of European immigrants in the United States labor market during 1850-1913. This is a period that witnessed one of the largest migration episodes in modern history: almost thirty million, mostly European, immigrants moved to the US. The common view among scholar has been that European immigrants held substantially lower-paid occupations than natives upon first arrival, but that they converged with the native born after spending some time in the US. The paper provides new light to this issue.

To this end, the authors construct two datasets. They start by obtaining three random samples of migrants and natives from the United States census of the population, one for the year 1900, one for the year 1910 and finally one for the year 1920. For the second dataset, they first obtain a random sample of individuals (migrants and natives) in 1900 and then they search these individuals by their name, surname and age in the 1910 Census and, again, in the 1920 Census. In doing this exercise, their matching rate is about 25%. In other words, of the original 1900 random sample, the authors are able to follow over time 25% of the individuals. To sum up, the author have two sets of data: a series of cross-sectional random samples, and a panel dataset.

Next, they estimate different versions of the following model:

$$\begin{aligned} Earnings_{it} = \beta_0 + \beta_1 Y0\_5_{it} + \beta_2 Y6\_10_{it} + \beta_3 Y11\_20_{it} + \beta_4 Y21\_30_{it} + \beta_5 Y30_{it} + \\ + \beta_6 age_{it} + \beta_7 age_{it}^2 + \beta_8 age_{it}^3 + \beta_8 age_{it}^4 + \beta_9 after1890 + \theta_t + \alpha_j + a_i + u_{it} \end{aligned} \tag{1}$$

In equation (1), i indicates individuals and t time (1900, 1910, 1920). $Earnings_{it}$ measures the earnings of individuals in 2010 dollars; $Y0\_5$ is an indicator that equals one if the migrant has spent in the US zero to 5 years and zero otherwise, $Y6\_11$ is an indicator that equals one if the migrant has spent in the US 6 to 10 years and zero otherwise, and so on, until $Y30$, which is an indicator that equals one if the migrant has spent in the US more than 30 years. The omitted category here is being a US born. age_{it} indicates the age of the individual, which appears as a fourth-degree polynomial; $after1890$ is an indicator that equals one if the migrant arrived after 1890. Finally, θ_t indicates Census year fixed effects, α_j indicates country of birth fixed effects and a_i indicates individual fixed effects.

The results from their analysis are reported in Table 1 on page 4. In column (1) and (2) the model is estimated with a standard OLS estimator. In column (3) the model is estimated using a fixed-effect estimator applied to the panel sample.

(a) (5 points) Interpret the coefficients for the years 0 to 20 reported in column (1). What can you conclude about the earnings profile of the immigrants?

Solution. Other things being equal, immigrants just arrived (0-5 years in the US) are expected to earn \$1255.73 below natives of similar age, similarly those in the US for 6-10 years are expected to earn \$734.51 less than natives of similar age and finally those in the US for 11-20 years are expected to earn \$352.94 less than similar natives. Immigrants appear to completely make up this gap over time.

(b) (10 points) Focus only on the results reported in column (1) and (2). Why does the immigrant-native earnings gap in the first five years since arrival ($\hat{\beta}_1$) shrink?

Solution. Column (2) introduces an indicator that controls for arrival after 1890. The coefficient of that variable indicates that these later cohorts are expected to earn $739.18 less than earlier arrivals, ceteris paribus. This indicator suggests that, even within the SAME sending country (all regressors control for country of birth), the initial gap in the pooled cross section is due to the lower occupational skills of immigrants who arrived after 1890.

The estimated gap shrinks because in column (1) we were omitting this variable, which is both relevant and correlated with years since migration. In other words, in column (1) we had:

$$Earnings_{it} = a_0 + a_1 Y0\_5_{it} + \boldsymbol{x}'\gamma + \tilde{u}_{it},$$

where $\boldsymbol{x}'$ is the vector containing all the regressors in equation (1) except the $after1890$ indicator and $\tilde{u}_{it} = \beta_9 after1890 + u_{it}$. The expected value of OLS estimator for β_1 in the previous equation would then be:

$E(\hat{a}_1) = \beta_1 + \delta\beta_9$

where δ is the coefficient from a regression of $after1890$ on $Y0\_5$. Looking at the direction of the bias we can conclude that $\delta\beta_9 < 0$.

(c) (10 points) A young referee suggests that part of the reasons that explain different assimilation profiles could be permanent differences in earnings between immigrant women and other groups, i.e. discrimination against immigrant women. The referee therefore proposes to include in the model of column (3) an indicator for being a female migrant. Provide a brief comment to this proposal.

Solution. Such an indicator is time-invariant. Its inclusion in column (3) is unnecessary. In fact, the basic model in column (3) is:

$$Earnings_{it} = \beta_0 + \beta_1 FemMig_i + \boldsymbol{x}'\gamma + a_i + u_{it},$$

where a_i represents individual heterogeneity, $FemMig_i$ the proposed indicator and $\boldsymbol{x}$ all the characteristics listed in (1). The fixed effect estimator, also called within estimator, applied to the regression subtracts the time-average from each variable:

$$\tilde{y} = Earnings_{it} - \sum_{t=1900}^{1920} Earnings_{it}/3$$

$$FemMig_i - 3 * FemMig_i/3 = 0$$

$$\tilde{x} = x_{it} - \sum_{t=1900}^{1920} x_{it}/3$$

$$a_i - 3 * a_i/3 = 0$$

and estimates: $\tilde{y} = \gamma_0 + \gamma_1\tilde{x} + \tilde{u}$. As seen, the indicator would be dropped from the analysis.

(d) (15 points) The abstract of the paper reads along these lines: "Prior cross-sectional work finds that immigrants initially held lower-paid occupations than natives but converged over time. In newly-assembled panel data, we show that, in fact, the average immigrant did not face a

substantial earnings penalty upon first arrival and experienced earnings growth at the same rate as natives. Cross-sectional patterns are driven by biases from declining arrival cohort quality and departures of negatively-selected return migrants." Discuss whether the conclusion that immigrant did not face substantial earnings penalties upon arrival could be questioned. Next, discuss whether cross-sectional patterns could be explained by other reasons besides cohort quality. Remember to base your statements on the concepts learned in class.

Solution. As shown in point (b), the omission of $after1890$ caused serious omitted variable bias in the estimates of cross-sectional earnings profiles. I agree hence with the statement that cross-sectional patterns are partly driven by biases from declining cohort quality.

The results in column (3) also strikingly show that immigrants upon arrival earn more than natives. This sample include those migrants who have not left the US in the 1900-1920 periods. Hence the authors conclude that return migration is explaining most of the gap. I agree also with this statement.

However, there are other reasons why the results might differ. In column (3) the authors are now estimating with a fixed effect estimator the following model:

$$Earnings_{it} = \beta_0 + \beta_1 Y0\_5_{it} + \boldsymbol{x}'\gamma + a_i + u_{it}.$$

The fixed effect estimator, compared to the simple OLS, is capturing all time-invariant characteristics (a_i) that were unobserved in the repeated cross-section. Hence, the results are also explained by controlling for time-invariant individual-specific heterogeneity, such as ability or motivation. In addition, to deliver unbiased estimates of β_1 we need a set of assumptions. First, linearity (which we assume throughout the course), second random sampling, third enough variability in the Xs (not an issue here), fourth a zero conditional mean assumption. The panel estimates in this case are not based on a random sample. In fact, as introduced on page 1, finding individuals over time resulted in matching rate around 25%. It is easy to imagine that the migrants that were not matched were a non-random sample of the initial set of individuals. In addition to these observations, one could further argue that we are disregarding important other characteristics that are time-varying and related with time since migration, for example getting married or having children.

To conclude, while return migration and cohort effects indeed explain part of the results, other reasons might be equally important:

- time-invariant unobserved heterogeneity
- non-randomness of the sample
- OVB

The last two reasons making it difficult to judge the final conclusions.

Table 1: Age-earnings profile for natives and foreign-born, 1900-1920

	OLS	OLS	FE
	Earnings	Earnings	Earnings
	(1)	(2)	(3)
0-5 Years in US	-1,255.73	-384.48	139.68
	(143.44)	(187.30	(57.96)
6-10 Years in US	-734.51	-2.89	313.81
	(147.44)	(172.05)	(113.61)
11-20 Years in US	-352.93	173.83	175.55
	(131.27)	(132.02)	(200.49)
21-30 Years in US	-294.87	128.44	-79.49
	(142.10)	(138.93)	(150.33)
30+ Years in US	22.41	155.77	78.07
	(184.65)	(178.49)	(186.55)
Arrive after 1890	-	-739.18	-
	-	(106.99)	-
Age controls	Yes	Yes	Yes
θ_t	Yes	Yes	Yes
α_j	Yes	Yes	-
N	205,458	205,458	65,804

Source: selected results from Table 4, Abramitzky et al. (2014), p. 484.
Earnings are measured in 2010 dollars.

Exercise 2 (40 points)

Romer (1993) proposes theoretical models of inflation that imply that more open countries should have lower inflation rates. His empirical analysis explains average annual inflation rates (since 1973) in terms of the average share of imports in gross domestic (or national) product since 1973–which is his measure of openness. In addition to estimating the key equation by OLS, he uses instrumental variables. While Romer does not specify both equations in a simultaneous system, he has in mind a two-equation system:

$$inf = \beta_{10} + \alpha_1 open + \beta_{11} lpcinc + \beta_{12} oil + u_1 \quad (1)$$

$$open = \beta_{20} + \alpha_2 inf + \beta_{21} lpcinc + \beta_{22} land + u_2, \quad (2)$$

where inf measures the annual inflation rate, *open* measures imports as % of GDP, *lpcinc* is the log of 1980 per capita income measured in U.S. dollars, *lland* is the log of land area of the country measured in square miles and *oil* is a dummy variable that takes value of one if the country is an oil producer.

The do-file of the analysis is reported on page 8 and the relative log-file starts on page 9.

(a) (5 points) Consider first the analysis reported on line 17-21 of the do-file. Using a 5% significance

level, test for heteroskedasticity.

Solution. The question asks to perform a Breush-Pagan test. The test is based on the squared residuals regressed on all characteristics:

$u^2 = \delta_0 + \delta_1 oil + \delta_2 lpcinc + \delta_3 lland + v$

$H_0 : \delta_1 = \delta_2 = \delta_3 = 0$

H_1 : not H_0

One can perform an F test or an LM test. As the statistic is 6.1332 and the critical value is 7.81 we fail to reject H_0 and conclude that the model is homoskedastic.

(b) (10 points) Discuss under which conditions the estimator reported on line 27 and 28 of the do-file can identify the parameters α_1 and α_2. If needed, base your answer on appropriate tests.

Solution. This requires two t-tests. For inflation to be identified, we need *lland* to be a valid and relevant instrument for open. Assuming validity, we have the following hypothesis

$H_0 : \beta_{22} = 0$

$H_1 : \beta_{22} \neq 0$

The t-statistic is:

$$t - stat = \frac{\hat{\beta}_{22}}{se(\hat{\beta}_{22})} = -9.25.$$

Since this is well above the 3.2 Stock and Yogo suggested critical value, we reject the null hypothesis and conclude that the equation is identified.

Similarly, for identification of *open* we need *oil* to be a good predictor of inf. Hence:

$H_0 : \beta_{12} = 0$

$H_1 : \beta_{12} \neq 0$

The t-statistic is:

$$t - stat = \frac{\hat{\beta}_{12}}{se(\hat{\beta}_{12})} = -0.69.$$

Since this is well below the 3.2 Stock and Yogo suggested critical value, we fail to reject the null hypothesis and conclude that the equation is not identified. Note that for identification we need the two instruments to be valid. In other words one also needs to discuss if $Cov(IV, u) = 0$ for the two IVs and the relative error terms.

(c) (5 points) A commentator suggests to use *land* in levels, rather than in logs (*lland*), as an instrument for *open*. Provide at least two strategies based on which to decide which regressor to include.

Solution. This question is about model specification and relates to the functional form of regressors. Since the dependent variable is the same (*open*) and the question is whether to use *lland* or *land* as a regressor one could:

- Compare the R-squared from two different first-stage regressions:

$$open = \pi_0 + \pi_1 lland + \pi_2 lpcinc + \pi_3 oil + u$$

$$open = \gamma_0 + \gamma_1 land + \gamma_2 lpcinc + \gamma_3 oil + \epsilon$$

  and choose the one with the highest R-squared.
- One could use *land* and use a RESET test for general model misspecification.
- One could also pick the functional form that delivers the strongest instrument (the one with the largest t-statistic in the first stage).

(d) (10 points) How would you test whether the OLS and IV estimates on the equation for *open* are statistically different?

Solution. Use the Hausman test from Chapter 15. In particular, let $\hat{v}_2$ be the OLS residuals from the reduced form regression of open on *lpcinc* and *lland*. Then, use an OLS regression of inf on open, lpcinc, oil, and $\hat{v}_2$ and compute the t statistic for significance of the coefficient on $\hat{v}_2$. If $\hat{v}_2$ is significant, the IV and OLS estimates are statistically different and the IV is needed.

(e) (10 points) Explain whether you think the estimator used on line 27 of the do-file for *inf* is consistent and efficient.

Solution. We have studied that an IV, if weak, could have a larger small sample bias than the OLS. In part b we showed however that for this equation the instrument passes the rule of thumb and, even if the sample is small, the bias in the IV should also be small. In terms of efficiency, in point (a) we showed that the reduced form model did not exhibit heteroskedasticity. It is possible that the model is homoskedastic and the use of "robust standard errors" then might be unnecessary. If this was the case, the proposed IV would be less efficient than an IV calculated under the homoskedasticity assumption.

Exercise 3 (20 points)

In their article "Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack." published in the American Economic Review (2004), Di Tella and Schargrodsky wirte: "An important challenge in the crime literature is to isolate causal effects of police on crime. Following a terrorist attack on the main Jewish center in Buenos Aires, Argentina, in July 1994, all Jewish institutions received police protection. Thus, this hideous event induced a geographical allocation of police forces that can be presumed exogenous in a crime regression. Using data on the location of car thefts and police forces before and after the attack, we find a large deterrent effect of observable police on crime".

(a) (10 points) Let $CarTheft_{it}$ indicate the number of care theft in location i at time t and $NewPolice_{it}$ the number of policemen allocated to location i at time t. Explain which model you think the authors are using to pin down the causal effect of police on crime.

Solution. This is a typical DiD set-up. The model takes the form:

$$CarTheft_{it} = \beta_0 + \beta_1 AfterJuly_t + \beta_2 NewPolice_i + \beta_3 (AfterJuly_t \times NewPolice_i) + u_{it} \quad (3)$$

This is clear from the statement "Using data on the location of car thefts and police forces before and after the attack", which rules out a simple regression and also an IV estimation.

(b) (10 points) Interpret the coefficient(s) you have mentioned at point (a).

Solution. here we have, other things being equal:

β_0 = expected car thefts before the terrorist attack in control areas;

β_1 = additional increase/reduction in car thefts in control areas after July 1994

β_2 = increase/reduction in car thefts in the affected areas before July 1994, compared with control areas

β_3 = expected change in car thefts over time in treated areas compared to control areas.

```
1   ********************************************
2
3   * Exam Spring 2019: Question 2 Do-File
4
5   ********************************************
6
7   clear all
8   cd "/Users/co/Documents/Teaching/Courses Taught/Econometrics/NTNU/Exam"
9
10  log using exam_s19, replace text
11
12
13  use OPENNESS.DTA
14
15
16  ***
17  reg inf  oil lpcinc lland
18  predict res, res
19  gen ressq= res*res
20
21  reg ressq oil lpcinc lland
22
23
24  ***
25  reg inf oil lpcinc lland
26  reg open  oil lpcinc lland
27  ivregress 2sls inf lpcinc oil (open =  lland), robust
28  ivregress 2sls open lpcinc land (inf =  oil), robust
29
30
31
32  log close
33
34
35
36
```

```
      name:  <unnamed>
       log:  C:\Users\costanzb\Documents\Teaching\Courses Taught\Econometrics\NTNU\Exam\exam_s19.
  log type:  text
 opened on:   6 May 2019, 14:57:10

1 .
2 .
3 . use OPENNESS.DTA

4 .
5 .
6 . ***
7 . reg inf  oil lpcinc lland

      Source |       SS           df       MS      Number of obs   =       114
-------------+----------------------------------   F(3, 110)       =      2.04
       Model |  3436.23107         3  1145.41036   Prob > F        =    0.1119
    Residual |  61637.1906       110  560.338097   R-squared       =    0.0528
-------------+----------------------------------   Adj R-squared   =    0.0270
       Total |  65073.4217       113  575.870989   Root MSE        =    23.671

------------------------------------------------------------------------------
         inf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         oil |  -6.690162   9.709693    -0.69   0.492    -25.9325     12.55217
      lpcinc |   .6279128   2.084876     0.30   0.764   -3.503823     4.759648
       lland |   2.549812   1.083075     2.35   0.020    .4034102     4.696213
       _cons |  -15.50321   21.49386    -0.72   0.472     -58.099     27.09257
------------------------------------------------------------------------------

8 . predict res, res

9 . gen ressq= res*res

10 .
11 . reg ressq oil lpcinc lland

      Source |       SS           df       MS      Number of obs   =       114
-------------+----------------------------------   F(3, 110)       =      0.71
       Model |  23172800.5         3  7724266.82   Prob > F        =    0.5493
    Residual |  1.2002e+09       110  10910513.4   R-squared       =    0.0189
-------------+----------------------------------   Adj R-squared   =   -0.0078
       Total |  1.2233e+09       113  10825922.7   Root MSE        =    3303.1

------------------------------------------------------------------------------
       ressq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         oil |  -619.0514   1354.887    -0.46   0.649    -3304.119     2066.016
      lpcinc |   115.9505   290.9229     0.40   0.691    -460.5903     692.4913
       lland |   208.7068   151.1319     1.38   0.170    -90.80121     508.2148
       _cons |  -2631.407   2999.245    -0.88   0.382    -8575.206     3312.392
------------------------------------------------------------------------------
```

```
12 .
13 .
14 . ***
15 . reg inf oil lpcinc lland
```

Source	SS	df	MS	Number of obs	=	114
				F(3, 110)	=	2.04
Model	3436.23107	3	1145.41036	Prob > F	=	0.1119
Residual	61637.1906	110	560.338097	R-squared	=	0.0528
				Adj R-squared	=	0.0270
Total	65073.4217	113	575.870989	Root MSE	=	23.671

inf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
oil	-6.690162	9.709693	-0.69	0.492	-25.9325	12.55217
lpcinc	.6279128	2.084876	0.30	0.764	-3.503823	4.759648
lland	2.549812	1.083075	2.35	0.020	.4034102	4.696213
_cons	-15.50321	21.49386	-0.72	0.472	-58.099	27.09257

```
16 . reg open  oil lpcinc lland
```

Source	SS	df	MS	Number of obs	=	114
				F(3, 110)	=	29.84
Model	28607.1395	3	9535.71318	Prob > F	=	0.0000
Residual	35150.8507	110	319.553188	R-squared	=	0.4487
				Adj R-squared	=	0.4336
Total	63757.9902	113	564.230002	Root MSE	=	17.876

open	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
oil	.3989381	7.332499	0.05	0.957	-14.13235	14.93023
lpcinc	.5204514	1.574443	0.33	0.742	-2.599724	3.640627
lland	-7.566865	.8179095	-9.25	0.000	-9.18777	-5.945961
_cons	117.2567	16.23158	7.22	0.000	85.08954	149.4239

```
17 . ivregress 2sls inf lpcinc oil (open =  lland), robust
```

Instrumental variables (2SLS) regression

Number of obs	=		114
Wald chi2(3)	=		6.98
Prob > chi2	=		0.0725
R-squared	=		0.0349
Root MSE	=		23.471

inf	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
open	-.3369707	.1500285	-2.25	0.025	-.6310211	-.0429202
lpcinc	.8032896	1.527911	0.53	0.599	-2.19136	3.797939
oil	-6.555731	3.708423	-1.77	0.077	-13.82411	.7126437
_cons	24.00886	11.28069	2.13	0.033	1.899121	46.11861

Instrumented: open
Instruments: lpcinc oil lland

```
18 . ivregress 2sls open lpcinc land (inf =  oil), robust

Instrumental variables (2SLS) regression          Number of obs   =        114
                                                  Wald chi2(3)    =      17.30
                                                  Prob > chi2     =     0.0006
                                                  R-squared       =          .
                                                  Root MSE        =     23.952

------------------------------------------------------------------------------
             |               Robust
        open |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inf |   .2660628   1.500991     0.18   0.859    -2.675825    3.207951
      lpcinc |   4.160504   2.557307     1.63   0.104    -.8517257    9.172735
        land |  -.0000138   9.36e-06    -1.47   0.141    -.0000321    4.57e-06
       _cons |    4.73574   39.00215     0.12   0.903    -71.70707    81.17855
------------------------------------------------------------------------------
Instrumented:  inf
Instruments:   lpcinc land oil

19 .
20 .
21 .
22 . log close
      name:  <unnamed>
       log:  C:\Users\costanzb\Documents\Teaching\Courses Taught\Econometrics\NTNU\Exam\exam_s19.
  log type:  text
 closed on:   6 May 2019, 14:57:10
-------------------------------------------------------------------------------
```