King's College London

6SSMN961: APPLIED ECONOMETRICS

COURSEWORK 2019/20

INSTRUCTIONS TO CANDIDATES:

1. The coursework has two questions. You must answer both questions.

2. The deadline for submission of the coursework is **Friday 20th December at 10:00 am**. Work should be submitted on KEATS.

3. The file that you upload on KEATS should contain two parts:

   - Short written answers to the questions
   - The Stata output in pdf format

   You can merge two pdf files using Acrobat Professional or an online pdf merger.

4. You must complete the coursework coversheet. This is very important to ensure that your work can be identified. In addition, you should name the file with your candidate number as follows: **Candidatenumber.pdf**.

5. To avoid collusion, each student is given a unique version of the datasets. This means that you should answer the questions with the datasets that have been provided to you. If your answers or the Stata output file are based on the datasets given to another student, you will lose marks and face an allegation of collusion.

Question 1 (50 marks)

This question is based on a paper by Draca, Machin and Witt (2011) which studies the causal effect of police presence on crime using evidence from the July 2005 terror attacks in London. Police activity in central London increased in the six weeks following the terror attacks.

The dataset **crime.dta** contains data on crime rates and the level of police deployed for 28 London boroughs for the period from January 2004 to December 2005. The variable **treat** in the dataset is equal to 1 for five London boroughs that were either affected by the terror attacks or considered to be potential terrorist targets (Westminster, Camden, Islington, Kensington and Chelsea and Tower Hamlets) and 0 for other boroughs. To look at the effect during the six-week period of increased police presence, the dataset contains the variable **policy6**, which is equal to 1 in the period July 7, 2005 to August 18, 2005 and is equal to 0 in the equivalent period the year before (July 8, 2004 to August 19, 2004). To look at the effect over a longer period, the dataset contains the variable **post**, which is equal to 1 for the whole period after July 7, 2005 and 0 otherwise.

a) Complete the table below with the average levels of police deployment (**police** in the dataset) and crime rate (**crime** in the dataset) for treated and control boroughs in the six-week period after the attacks (post-period) and the equivalent six-week period the year before (use the variable **policy6** in the dataset).

	Police deployment (hours worked per 1,000 population)				Crime rate (crimes per 1,000 population)		
	Pre-period	Post-period	Difference (post-pre)		Pre-period	Post-period	Difference (post-pre)
Treatment							
Control							

What are the difference-in-differences (DD) estimates of the effect of the terror attacks on police deployment and crime? Use a simple regression to calculate the standard errors of the DD estimates. Explain what type of standard errors you are using and why. Are the estimates statistically significant? Comment on the results.

b) You are interested in the causal effect of police on crime and estimate the following model by OLS:

$$c_{bt} = \alpha^{OLS} + \beta^{OLS} POST_t + \delta^{OLS} p_{bt} + \varepsilon_{bt}$$

The dependent variable c_{bt} is the log of the crime rate and the regressor p_{bt} is the log of the level of police deployment. The variable $POST_t$ (**post** in the dataset) is equal to 1 for the whole period after July 7, 2005 and 0 otherwise.
Comment on the results. What are the potential sources of bias in this regression?

c) Estimate the following model for the effect of the terror attacks on police deployment:

$$\Delta_{52} p_{bt} = \alpha_1 + \beta_1 POST_t + \delta_1 (T_b \times POST_t \ ) + \Delta_{52} \varepsilon_{1bt}$$

The dependent variable $\Delta_{52} p_{bt}$ is the change in the log level of police deployment. The change is measured relative to the same week one year before, to account for seasonality in crime. The treatment variable is T_b (**treat** in the dataset). The variable $POST_t$ is as defined in part b).
Comment on the results.

d) Estimate the following model for the effect of the terror attacks on crime:

$$\Delta_{52} c_{bt} = \alpha_2 + \beta_2 POST_t + \delta_2 (T_b \times POST_t \ ) + \Delta_{52} \varepsilon_{2bt}$$

The dependent variable $\Delta_{52} c_{bt}$ is the change in the log of the crime rate. The other variables are as defined in part c).
Comment on the results.

e) Instead of using OLS, you decide to estimate an instrumental variables (IV) model for the causal impact of police on crime using $T_b \times POST_t$ as an instrument for the change in the log level of police deployment $\Delta_{52} p_{bt}$:

$$\Delta_{52}c_{bt} = \alpha^{IV} + \beta^{IV} POST_t + \delta^{IV}\Delta_{52}p_{bt} + \Delta_{52}\varepsilon_{bt}$$

1. Estimate the model and comment on the results. What is the relation between the coefficient δ^{IV} and the coefficients δ_1 and δ_2 in the parts c) and d)? Explain.
2. Under what assumptions is $T_b \times POST_t$ a valid instrument for the change in the log level of police deployment? Do those identification assumptions seem plausible? Explain in detail, with reference to the findings in the paper.

References:

Mirko Draca, Stephen Machin and Robert Witt (2011), "Panic on the Streets of London: Police, Crime and the July 2015 Terror Attacks", *American Economic Review*, Vol. 110, No. 5., pp. 2157-2181.

Question 2 (50 marks)

Lee (2008) uses a regression discontinuity design (RDD) to study whether the incumbent party in a district has an electoral advantage in the United States House of Representatives. He is interested in studying whether winning an election has a causal influence on the probability that the candidate will run for office again and eventually win the next election.

The dataset **election.dta** contains data on vote shares from 1946 to 1998. The main vote share variable is the Democratic vote share minus the vote share of the strongest opponent (which in most cases is a Republican nominee). The Democrat wins the election when this variable "Democratic vote share margin of victory" (**difshare** in the dataset) crosses the zero threshold and loses the election otherwise. Two measures of the success of the party in the subsequent election are used: the probability that the party's candidate will both become the party's nominee and win the election (**mmyoutcomenext** in the dataset) and the probability that the party's candidate will become the nominee in the election (**mrunagain** in the dataset).

He considers the following RDD model:

$$v_{it+1} = \alpha w_{it} + \beta v_{it} + \gamma d_{it+1} + e_{it+1}$$

The dependent variable v_{it+1} is a measure of success of the Democratic candidate in district i in election year t+1 (either **mmyoutcomenext** or **mrunagain** in the dataset). The variable v_{it} is the vote share margin of victory for the Democratic candidate in district i in election year t (**difshare** in the dataset). The vector w_{it} is a set of characteristics of the candidate in year t. d_{it+1} is an indicator variable equal to 1 if the Democrats are the incumbent party during the electoral race in year t+1. This is a deterministic function of whether the Democrats won election t:

$$d_{it+1} = \begin{cases} 1 \; if \; v_{it} > 0 \\ 0 \; if \; v_{it} \leq 0 \end{cases}$$

NOTE: for all regressions in this question, you should use robust standard errors with no clustering.

a) Suppose that you run an OLS regression of the probability that the Democratic candidate wins the election in year t+1 on the vote share margin of victory of the Democratic candidate in year t. Would this regression identify the causal effect of incumbency on the vote share? Explain.

b) Under what identification assumption does RDD lead to a valid estimate of the causal effect of incumbency on the vote share? Explain in detail.

c) Estimate the RDD model for the probability of a Democrat both running in and winning election t+1 (**mmyoutcomenext** in the dataset). You should regress this variable on a fourth-order polynomial in the democratic vote share margin of victory, separately for each side of the threshold. Replicate figure 2 (a) in the paper showing a scatter plot of the data and the fitted line of this RDD model. Does the incumbent candidate appear to have an electoral advantage? Explain.

d) Repeat the analysis in part c) for the probability that the Democrat remains the nominee for the party in election t+1 (**mrunagain** in the dataset). Replicate figure 3 (a) in the paper showing a scatter plot of the data and the fitted line of this RDD model. Does the incumbent candidate appear to have an electoral advantage? Explain.

e) Now estimate the RDD model separately for two dependent variables that have already been determined as of election t: the average number of terms the candidate has served in Congress (**mofficeexp** in the dataset) and the average number of times he has been a nominee (**melectexp** in the dataset). You should run two separate regressions of each of these variables on a fourth-order polynomial in the democratic vote share margin of victory, separately for each side of the threshold. Replicate figures 2 (b) and 3 (b) in the paper showing a scatter plot of the data and the fitted line of these RDD models. How do these results inform the validity of the identification assumption that you discussed in part b)? Explain in detail.

f) Now repeat the analysis in parts c) and d) adding the pre-determined variables **mofficeexp** and **melectexp** as controls in the regressions. Generate the two figures again using these estimates. Do the results change significantly when you control for these pre-determined variables? Explain in detail.

References:

Lee, David S. (2008), “Randomized experiments from non-random selection in U.S. House elections”, *Journal of Econometrics*, Vol. 142, pp. 675–697.