Econometrics 2 Project Royal Holloway University of London

Introduction:

The Research Question: does the provision of startup support through local innovation hubs increase the number of new startup businesses?

The aim is to evaluate a government policy to open 500 local Innovation hubs, located randomly across the country. It was introduced in 2016. These hubs provide a startup support programme which comprises of the following:

- legal advice and administrative help with incorporation,
- help with writing a business plan, and subsequent evaluation of the plan
- free training for those intending to start a business regarding branding, marketing, accounting, and business strategy
- free mentoring during the first year of operation by serial startup investors

The goal of the policy is that the government wants to encourage innovation, particularly among young people. Therefore, the innovation hubs are available only to individuals who have not yet reached the age of 30 on 5th April 2015.

Data:

First dataset is a panel dataset from the Department for Business (called Register) containing a sample of 500,000 UK citizens between the ages of 20 and 40 years old, for the years 2015 and 2017. The data captures whether the individual registered or reregistered a business, some socio-economic characteristics like age, gender, education, and the distance to the nearest Innovation Hub. The dataset contains one observation per individual and year, and the following variables:

- an individual identifier (personid),

- year of survey interview (year)

- annual income (income, measured in real 2015 pounds)

- gender (male) - education (educ, measured as years of education)

- distance between the individual's home and the nearest Innovation hub (distance, in km)

- whether they registered a business in the year (business)

- if yes, the business identifier (businessid)

- if yes, the industry the business operates in (industry, measured as 1-digit industry classification)

2. Second dataset is panel dataset from the Innovation Hub (called Hub) for the year 2017. The dataset contains cross-sectional data on

- an individual identifier (personid)

- whether the individual attended the startup support programme (attend)

- age (age, Note: this measures age as recorded on April 5th, 2015, and is expressed in decimals)

Part 1: Understanding the Data

Register: The dataset has 1,000,000 observations, one per person id and year.

Hub: The dataset has 500,000 observations, one per person id.

The register dataset is a panel dataset. We have one individual and two time periods.

Merged data: The dataset has 1,000,000 observations, one per person id and year.

When the data is merged the variable "age2015" in year 2015 and 2017 are the same. Thus, we will assume that individuals aged exactly 2 years from 2015 to 2017. We will also change name of the variable to just age to avoid confusion.

Furthermore, we will assume for simplicity, that one person cannot have multiple businesses, and know that businesses need to re-register each year they operate.

Part 2: Descriptives

Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
personid	1,000,000	250,000.5	144,337.6	1	500,000
year	1,000,000	2,016.0	1.0	2,015	2,017
income	1,000,000	30,739.2	15,344.9	2,992.4	229,116.4
male	1,000,000	0.5	0.5	0	1
educ	1,000,000	9.4	3.2	6	21
distance	1,000,000	1.6	2.1	0.01	136.0
business	1,000,000	0.1	0.3	0	1
businessid	115,880	50,185,795.0	28,865,863.0	773	99,997,722
industry	115,880	4.5	2.9	0	9
attend	1,000,000	0.1	0.3	0	1
age	1,000,000	31.0	5.9	20.0	42.0

Table 1: summary statistics for socio-economic variables

There is an equal amount of men and women in the sample. Mean income is 30,739£. Mean number of education years is 9. 10% of people attend Innovation hubs.

The proportion of individuals who registered a new business during a sample period is 2.8%

Part 3: Simple DID

The Innovation Hub system was introduced in 2016. The research question is whether attending innovation hub affects the number of new startups. We want to see the effect of attending the innovation hubs, by calculating Difference in Differences(DiD):

Treatment group – people who attended the innovation hubs.

Control group - people who did not attend the innovation hubs.

	Treatment	Control
Pre	4508	46418
Post	8655	56299

Table 2: number of businesses in treatment and control group before and after introduction of the policy.

Total number of people who attended the Innovation Hubs are 52,486.

Total number of people who did not attend the Innovation Hubs is 447,514

If we divide number of businesses in each group by the total number of people:

$$\frac{8655-4508}{52486}=0.079 \quad and \frac{56299-46418}{447514}=0.022$$

If we subtract these two, we will get 0.057 or that Innovation Hubs on average increase the number of startups by 5.7%. It is important to mention that this is a rough estimate.

The research question is the effect on number of new successful startups. However, we only possess data for 2 years, making it problematic to estimate increase in **number** of successful startups. However, we may note that percentage point increase in number of successful startups is equivalent to percentage point increase of probability for an individual to have a successful startup. (If the sample population is large enough). Thus, we will use Business as a dependent variable.

We will include additional regressors to avoid Omitted Variable Bias. We will use socio-economic factors – income, sex, age (we assume people aged 2 years from 2015 to 2017).

We run 2 regressions:

Simple multiple linear regression:

$$Bussiness_i = \beta_0 + +\beta_1 Income_i +$$

$$\beta_2 Age_i + \beta_3 Male_i + u_i$$

DiD regression:

$$Bussiness_i = \beta_0 + \beta_1 Post + \beta_2 Treat + \beta_3 Post \times Treat + \beta_4 Income_i +$$

$$\beta_5 Age_i + \beta_6 Male_i + u_i$$

Post – 1 if innovation hubs are introduced, 0 otherwise.

Treat – 1 if person attends the innovation hub program, 0 otherwise.

Naive OLS and DiD Results

	Dependent variable:	
	business	
	Naive OLS	DiD
	(1)	(2)
attend	**0.089*****	
	(0.002)	
income	0.00000***	
	(0.00000)	
age	0.008***	
	(0.0001)	
post		0.022***
		(0.001)
treat		-0.019***
		(0.001)
male	0.015***	0.015***
	(0.001)	(0.001)
post:treat		**0.057*****
		(0.002)
Constant	-0.258***	0.096***
	(0.003)	(0.001)
Observations	500,000	1,000,000
R^2	0.069	0.003
Adjusted R^2	0.069	0.003
Residual Std. Error	0.324 (df = 499995)	0.320 (df = 999995)
F Statistic	9,326.131*** (df = 4; 499995)	826.316*** (df = 4; 999995)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3: Naive OLS and DiD Results

DiD regression yields = 0.054, which can be interpreted as a treatment group on average has 5.5% higher chance of a successful startups than control group.

There is a threat to internal validity – failure of randomization. In DiD analysis we rely on parallel trends assumption, here, however, the treatment might not be allocated randomly. Socioeconomic factors can influence whether an individual attends the Innovation Hubs, which might create bias.

These two models seem to have too many issues. However, we can conclude a positive relationship between successful startups and innovation hubs.

Part 4: Eligibility and take-up

Now, we want to test whether person attending innovation hub is independent of their socio-economic status. First, we calculate the take-up rate from the data:

Here we will only be looking on year 2015 where people choose to attend or not to attend Innovation Hub.

$$take\ up\ rate = \frac{eligigble\ \cap\ attended}{eligible} \times 100 = \mathbf{21}\%$$

We will regress attend on income, age2015, male, distance.

$$Attend_i = \beta_0 + \beta_1 Income_i + \beta_2 age_i + \beta_3 male_i + \beta_4 distance_i + u_i$$

We will use Probit and Logit to estimate our model. For take-up rate only 2015 data is relevant, thus, we will use 2015 subset for the estimation.

Probit:

$$\Pr\left(Attend = 1 | Income, age2015, male, distance\right) = \Phi(\beta_0 + \beta_1 Income + \beta_2 age2015 + \beta_3 male + \beta_4 distance + u_i)$$

Logit:

$$\Pr\left(Attend = 1 | Income, age2015, male, distance\right) = \Lambda(\beta_0 + \beta_1 Income + \beta_2 age2015 + \beta_3 male + \beta_4 distance + u_i)$$

Results of Binary Models

	Dependent variable:	
	attend	
	probit	*logistic*
	(1)	(2)
income	0.00001***	0.00003***
	(0.00000)	(0.00000)
male	0.165***	0.303***
	(0.006)	(0.011)
educ	0.027***	0.050***
	(0.001)	(0.002)
age	-0.152***	-0.271***
	(0.001)	(0.001)
distance	-0.776***	-1.477***
	(0.005)	(0.010)
Constant	2.910***	5.251***
	(0.021)	(0.038)
Pseudo R-squared		
0.3		
0.29		
Observations	500,000	500,000
Log Likelihood	-117,508.600	-118,553.300
Akaike Inf. Crit.	235,029.200	237,118.700
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4:Results of Binary models

Both models seem to suggest that there is a positive connection between socio-economic variables and whether an individual attends the Innovation Hub.

This suggests that treatment received in question 3 DiD model is not random. Which violates the parallel trends assumption as control and treatment groups differ socioeconomically.

The DiD coefficient still might represent some portion of causal effect of Innovation hubs on successful startups, however, it might be very biased.

Part 5: Instrumental variables

From question 4 we know that there is evidence to suggest that treatment is not allocated randomly. We can use an instrumental variable which only influences probability of opening business through attending Innovation Hubs.

The distance variable will fit this role well as it is random for every individual and only affects success of the startup through attend variable.

Distance arguably is randomly assigned. To ensure that the use of the instrument is appropriate we will check whether it satisfies the following conditions:

Instrument Relevance:

We can note that Logit and Probit tables suggest a negative relationship between Attend and Distance. Also:

$$Cor(Attend, Distance) = -0.16$$

Intuitively, the further individual lives from Innovation Hubs the less likely they are to attend.

Instrument Exogeneity:

It seems that Distance is uncorrelated with the error term as it is arguably random.

To sum up, it seems that all conditions for a valid instrument are satisfied.

We will also add covariates to our model. The same covariates we used for Naïve OLS and DID models.

Model:

$$Business_i = \beta_0 + \beta_1 Attend_i + \beta_2 Income_i + \beta_3 Male_i + \beta_4 Age_i + u_i$$

With instrument:

$$Attend_i = \pi_0 + \pi_1 Distance_i + v_i$$

We estimate the IV model; we will only use data for 2017 for regression.

IV Results

	Dependent variable:
	business
	IV Regression
attend	**0.054*****
	(0.009)
age	0.007***
	(0.0002)
income	0.00000***
	(0.00000)
male	0.016***
	(0.001)
Constant	-0.237***
	(0.006)
Observations	500,000
R^2	0.069
Adjusted R^2	0.069
Residual Std. Error	0.324 (df = 499995)
Note:	*p<0.1; **p<0.05; ***p<0.01

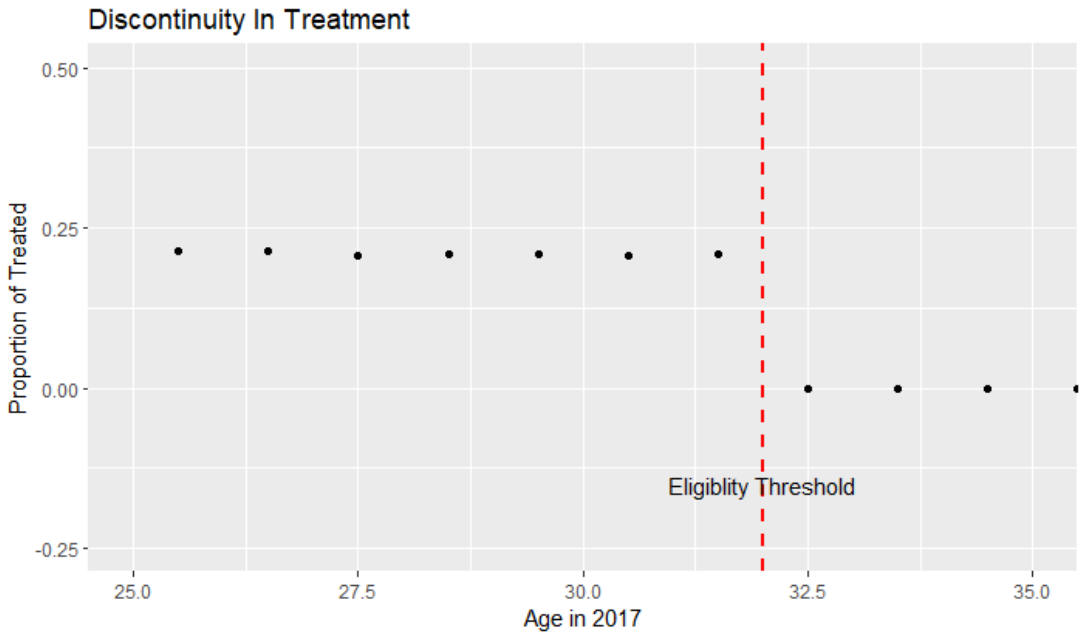
Table 5: Results of Instrumental Variable Regression

We estimate local average treatment effect(LATE) at 0.054. We can interpret this as attending Innovation Hub increases probability of opening business in 2017 by 5.4%. This is very similar result to Naïve OLS and DiD models, suggesting that all these results are significant.

Part 6: Regression discontinuity design

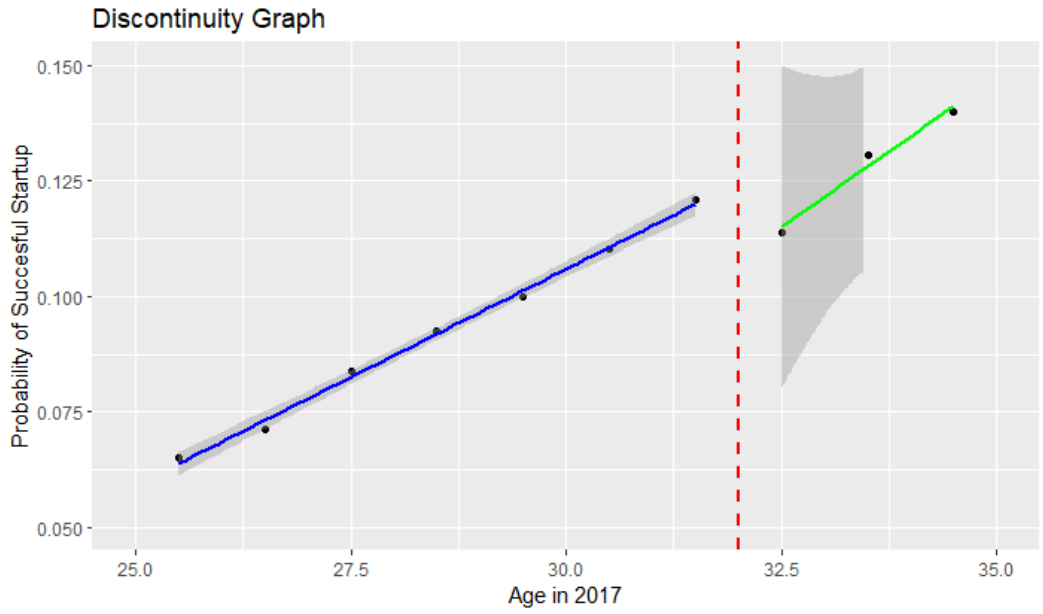
Now use an RD design to estimate the causal effect of Innovation Hubs on business formation. As the Innovation Hubs are only available to people below the age of 30, there will be discontinuity in probability of opening business for individuals around that age; given that innovation hubs have any effect on probability of opening business.

Below is the graph suggesting potential discontinuity.



Graph 1: Discontinuity as the result of eligibility condition

On y-axis we have proportion of individuals who attended Innovation Hubs per age group. Individuals who are 32 in 2017 are the threshold as they would be 30 in 2015 and would be eligible for the program. We observe the clear discontinuity in treatment.



Graph 2: Discontinuity in probability of opening a successful startup

Note the drop in probability of successful startup after the age of 32. As people who were 32 in 2017 were at the eligibility threshold in 2015 for the Innovation Hubs program. From eyeballing we clearly see the "jump" before the threshold.

We have a fuzzy regression as no one above the age of 30 attended innovation hubs in 2015 and some people below the age of 30 do not attend.

Considering functional form specification, from the graphs above it seems that linear model will fit the best.

$$business_i = \beta_0 + \beta_1 Attend_i + \beta_2(Age_i - 32) + \beta_3 Eligible_i \times (Age_i - 32) + u_i$$

And we will use eligible as an instrumental variable:

$$Attend_i = \pi_0 + \pi_1 Eligible_i + \pi_2(Age_i - 32) + \pi_3 Eligible \times (Age_i - 32) + v_i$$

RDD results

	Dependent variable:		
	business		
	(1)	(2)	(3)
attend	**0.044*****	**0.070*****	**0.091*****
	(0.006)	**(0.009)**	**(0.011)**
I(age - 32)	0.014***	0.012***	0.011***
	(0.0001)	(0.0003)	(0.001)
I(age - 32):eligible	-0.006***	-0.001**	0.003**
	(0.0002)	(0.001)	(0.001)
Constant	0.107***	0.110***	0.111***
	(0.001)	(0.001)	(0.001)
Observations	950,210	500,549	300,371
R^2	0.036	0.014	0.010
Adjusted R^2	0.036	0.014	0.010
Residual Std. Error	0.319 (df = 950206)	0.320 (df = 500545)	0.318 (df = 300367)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6: Discontinuity Regressions Results

In the table above we estimate models for window of observations of 10,5, and 3 years, respectively. It seems that the model is very sensitive to chosen size of the window of observations(around 1 percentage point decrease per 1 year extension of window of observations).

Nevertheless, we may infer that the effect is positive and between 4 and 10 %.

Part 7: Conclusion and Policy Recommendation:

Final Results

	Dependent variable:			
	business			
	OLS		*instrumental variable*	
	Naive OLS	DiD	IV Model	RDD window = 10
	(1)	(2)	(3)	(4)
attend	**0.089*****		**0.054*****	**0.044*****
	(0.002)		**(0.009)**	**(0.006)**
income	0.00000***		0.00000***	
	(0.00000)		(0.00000)	
age	0.008***		0.007***	
	(0.0001)		(0.0002)	
post		0.022***		
		(0.001)		
treat		-0.019***		
		(0.001)		
male	0.015***	0.015***	0.016***	
	(0.001)	(0.001)	(0.001)	
post:treat		**0.057*****		
		(0.002)		
I(age - 32)				0.014***
				(0.0001)
I(age - 32):eligible				-0.006***
				(0.0002)
Constant	-0.258***	0.096***	-0.237***	0.107***
	(0.003)	(0.001)	(0.006)	(0.001)
Observations	500,000	1,000,000	500,000	950,210
R^2	0.069	0.003	0.069	0.036
Adjusted R^2	0.069	0.003	0.069	0.036
Residual Std. Error	0.324 (df = 499995)	0.320 (df = 999995)	0.324 (df = 499995)	0.319 (df = 950206)
F Statistic	9,326.131*** (df = 4; 499995)	826.316*** (df€ = 4; 999995)		
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 7: Final Results

We have 4 estimates from 4 different models.

Naïve OLS suggests 8.9% increase; however, it might be that the estimate is biased because of the selection bias: people who attend innovation hubs are the ones who would start a startups even without innovation hubs.

DiD suggests 5.7% increase, however, it might be overestimating the effect due to selection bias and due to potential parallel trends assumption violation.

IV suggest 5.4% increase, however, there might be concerns about distance being weak instrument.

RDD with 10 year window of observation suggests 4.4% increase, however, from the RDD table we see that results differ significantly depending on window of observations.

Also, it seems that adding covariates in all models significantly increases the effect of Innovation Hubs. This might be the case, because covariates are significant to model and without them the independent variable is correlated with the error term.

It is also important to note that all models estimate the **probability** of an individual having a successful startup not the **number** of successful start-ups. However, that given large enough sample size probability of a successful start-up is equivalent to a number of new start-ups.

$probability\ of\ a\ startup \times number\ of people \sim number\ of\ startups$ (Given large enough N)

Moreover, it seems that selection bias is not very significant as IV estimate and DiD estimates are close to each other. Naïve OLS might have higher estimates due to not accounting for differences in Control and Treatment groups.

To summarize, all models suggest positive connection between attending Innovation Hubs and successful startup. And ¾ of our models suggest around 5% increase.

Instrumental variable model seems to be the most reliable one as it does not have issues other models possess.

IV addresses many potential threats to internal validity.

Furthermore, the instrument Distance seems to satisfy all the conditions for the good instrument. It is as-if-random, and only affects the probability of successful startup through attendance to innovation hubs.

Naïve OLS seems to overestimate the effect. DiD faces problems with its key assumption. And RDD seems to be overly dependent on the window of observations.

Even though, all 4 models provide a valuable insight into the relationship between Innovation Hubs and successful startups. Ultimately, Instrumental Variable Model would be the suggestion for the government to estimate causal effects.