21204276

Kadessova Ayazhan

**ECON3096 - Causal Inference Problem Set 1 (Lectures 1-3)**

**Due Date: 10 October before class (15:30 pm)**

**Consider the following three causal questions:**

* 􏰡 Many firms, particularly in southern European countries, are small, and owned and run by families. Are family owned firms growing more slowly than firms with a dispersed ownership?
* 􏰡 What is the effect of taking a job in a finance position compared to a human resources position on the salary earned by an economics graduate?
* 􏰡 What is the effect of mortgage interest rates on the number of new housing starts?  
  For each of these questions answer the following:
  + (a) What is the outcome variable and what is the treatment?
  + (b) Define the counterfactual outcomes Y01and Y1i.
  + (c) What plausible causal channel(s) runs directly from the treatment to the outcome?
  + (d) What are possible sources of selection bias in the raw comparison of outcomes by treatment status? Which way would you expect the bias

1. **Family-owned/Dispersed firms**
2. Outcome variable: growth of firms, including expansion of revenue or their output. Treatment: firm ownership type: family owned/dispersed.
3. Counterfactual outcome (Y0i) - firm growth rate when i is a dispersed firm. Counterfactual outcome (Y1i) - firm growth rate when i is a family-owned firm. 0 - Dispersed, 1 - family-owned firm.
4. There are many possible channels. On a positive note, family-owned firms may benefit from stakeholder efficiency. Stronger ties with managers, employees, and customers in dispersed firms, can result in higher trust, loyalty, and motivation, all of which are likely to lead to increased growth rates. Alternatively, family-owned firms will likely suffer from inefficient and outdated management practices. The firms may want to secure the family succession, whereas dispersed firms are likely to focus on market-determined modern management choices for a better outcome, beating family-owned firms.
5. A possible source of selection bias:

* Survivorship bias

In developing countries, family-owned firms that are legally registered and pay tax are the ones who can afford to do so, meaning they have higher growth levels than the whole group. If the research only included legally registered firms, there was a “screening contest,” which pushes out unsuccessful variables due to their lack of visibility. As a result, such a comparison will overestimate the growth of family-owned firms.

* Low/Slow growth

Family-owned firms in the research may be slow or have low growth because family-owned firms with high productivity or high growth potential attract outside investments. After such investment, these successful family-owned firms become dispersed and are never comprehended in our research, which will underestimate the growth of family-owned firms.

**2. Finance/Human Resources**

1. Outcome variable: salary earned. Treatment: Employment position: taking a job in Finance or Human resources position.
2. Counterfactual outcome (Y0i) - salary of a graduate when i is working in a Human Resources position. Counterfactual outcome (Y1i) - salary of a graduate when i is working in a Finance position. 0 - HR, 1 - Finance position.
3. A plausible channel is: an economics education provides students with math skills that are relevant for Finance positions, such as financial management. So, an economics student is more likely to secure a better-paid position in this sector than in Human Resources, which does not match the courses they studied during university (they may have to start with training, which is paid little). Moreover, the average labor market is known to pay more for the tremendous quantitative aptitude skills linked to the Finance sector than for human management and recruitment skills needed for Human Resources.
4. A possible source of selection bias:

* Labour market values math skills more, so a Finance position is likely to have better salary for graduates from whichever background, not just economics.
* Only researching starting salaries. Having position-specific knowledge may pay more in the beginning. However, in the future, having intersectional knowledge usually has higher demand as firms' sections need to understand each other. Hence, if economics student who already has math skills chooses to work in Human Resources to get human management skills, they will in the future have the ability to work in bigger companies or in higher positions in current firms where all sections cooperate and intersectional knowledge is valued and well-paid. So, intersectional knowledge from an HR position will likely pay more than not gaining new skills continuing in Finance due to the potential of landing a big firm or higher position. So, research conducted only on starting salaries may underestimate the salary of an economics graduate going to HR.

**3. Mortgage interest rates -> new housing starts.**

1. Outcome variable: the quantity of new housing starts in a given period. Treatment: mortgage interest rates: low mortgage interest rates or high mortgage interest rates.
2. Counterfactual outcome (Y0i) - quantity of new housing starts in period i when mortgage interest rates are low. Counterfactual outcome (Y1i) - quantity of new housing starts in period i when mortgage interest rates are high. 0 - low, 1 - high mortgage interest rates.
3. A plausible channel is:

* High mortgage interest rates mean high monthly mortgage requirements, so there are higher opportunity costs of building or buying a house and as people won’t be able to spend this money in other investments. So when the mortgage interest rate is higher, there fewer people may commit to housing starts because they won’t want to waste money on interest and may want to spend it more differently.
* Increased monthly income, which makes high mortgage interest rates more affordable and increases the quantity of houses built even with high monthly mortgage payment requirements.

1. A possible source of selection bias: Economical condition.

* Downward bias

The central bank lowers interest rates, including mortgage interest rates if the economy experiences economic recession. But, during the recession, people lack income and are pessimistic regarding their ability to afford to pay off their mortgages. So, even with low interest rates due to financial uncertainty, people will be unlikely to build houses. As a result, it will lead to a downward bias, underestimating the effect of low mortgage interest rates on the number of housing starts.

* Upward bias

The central bank raises interest rates, including mortgage interest rates if the economy experiences economic growth and boom. However, because firms pay higher salaries to maintain workers’ motivation at high levels to meet the extra demand, people have higher incomes. As a result, they are optimistic about their ability to pay off the mortgage. Hence, they are more likely to build houses despite higher opportunity costs associated with higher rates. This will lead to an upward bias, an overly optimistic attitude toward high mortgage interest rates.

**2. Coming back to this example we had dicussed in class:**

**(a) What are the variance and standard deviation of the population mean Yi?**

* σ^2 = (1-3.6)^2 + (2-3.6)^2 + (2-3.6)^2 + (3-3.6)^2 + (4-3.6)^2 + (4-3.6)^2 + (5-3.6)^2 + (5-3.6)^2 + (5-3.6)^2 + (5-3.6)^2 + (5-3.6)^2 = 2.04
* 𝜎 = sqrt(σ^2) = sqrt(2.04) = 1.42828569
* **Using R**

# for sample

> var(pop)

[1] 2.266667

> #determine length of data

> n <- length(pop)

> #calculate population variance

**> var(pop) \* (n-1)/n**

**[1] 2.04**

# calculate population standard deviation

**> sd(pop) \* (sqrt((n-1)/n))**

**[1] 1.428286**

**(b) What is the standard error of the means of Samples 1 and Sample 2?**

* 𝑆𝐸𝑥 (1) = 1.41421356/ sqrt(4) = 0.707106781
* > std.error(sample1)
* [1] 0.7071068
* 𝑆𝐸𝑥 (2) = 1/ sqrt(3) = 0.57735
* > std.error(sample2)
* [1] 0.5773503

Pischke (LSE) Standard Error (Mean) September 24, 2013 5 / 22

(c) Suppose now we want to examine if the sample means of Sample 1 and Sample 2 are significant different from each other, what kind of test we shall use?

We can use a two sample **t-test (inferential statistic),** which uses the t-statistic, the [t-distribution](https://www.investopedia.com/terms/t/tdistribution.asp) values, and the degrees of freedom to find out statistical significance of the difference between the means of two samples and their relationship. We should also set the significance level and looking at p-value, if it is less that that level, conclude that the differences are similar.

Overall, t-tests are used if data sets follow a normal distribution and have unknown variances, which is the case for us.

**(d) What is the exact statisics number and what does it tell us about the difference?**

|  |  |
| --- | --- |
| >t.test(sample1, sample2, var.equal = TRUE)  Two Sample t-test  data: sample1 and sample2  t = 1.0351, df = 5, p-value = 0.3481  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -1.483418 3.483418  sample estimates:  mean of x mean of y  4 3 | > t.test(sample1, sample2, var.equal = FALSE)  Welch Two Sample t-test  data: sample1 and sample2  t = 1.0954, df = 5, p-value = 0.3233  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -1.346609 3.346609  sample estimates:  mean of x mean of y  4 3 |

The significance level, fixed probability of inaccurate elimination of null hypothesis which is in fact true (Type I Error) for a study is chosen before data collection. Here, we set the significance level to 5%. The value significant at 5% refers to [p-value](https://byjus.com/maths/p-value/), which is less than 0.05. Using both Two Sample and Welch Two Sample t-test, we see that the p-value is greater than 0.05, which means that the differences between recognised values (means of Sample 1 and Sample 2) are insignificant because there is not enough evidence to reject the null hypothesis that the Sample means are all equal/similar.

**3. Suppose we are interested in find out whether increasing the availability of computer for student could help to improve their test scores. We start from two variables: the number of computers per student and test scores for 5th graders, comp stu and testscr.**

Download the data set caschool.dta and read it into R. You could use the code in “ps2.R” to read in the dta dataset to R. It contains observatiosn on 420 California school districts in 1999 on 14 variables. Description on these variables is as the following:

* 􏰡 district: District code.
* 􏰡 school: School name.
* 􏰡 county: County name.
* 􏰡 gr span: Grade span of district.
* 􏰡 enrl tot: Total enrollment.
* 􏰡 teachers: Number of teachers.
* 􏰡 calw pct: Percent qualifying for CalWorks (income assistance).
* 􏰡 meal pct: Percent qualifying for free lunch.
* 􏰡 computer: Number of computers.
* 􏰡 testscr: Average test scores.
* 􏰡 comp stu: Number of computers per student.
* 􏰡 expn stu: Number of expenditure per student.
* 􏰡 str: Student-teacher ratio.
* 􏰡 avginc: Average income.
* 􏰡 el pct: Percent of English learners.
* 􏰡 read scr: Average reading score.
* 􏰡 math scr: Average math score.

**(a) Draw a distribution of test scores per student. Describe in words what you see.**

We observe that test scores are almost normally distributed with a slight skew in one direction, meaning  **the majority of data points are relatively similar, so most students in schools across California receive similar test scores,** within a small range of **640 and 650** with fewer outliers on the high and low ends of the data range. Moreover, mean, median and mode are close to each other.

> mean(ClassSize$testscr)

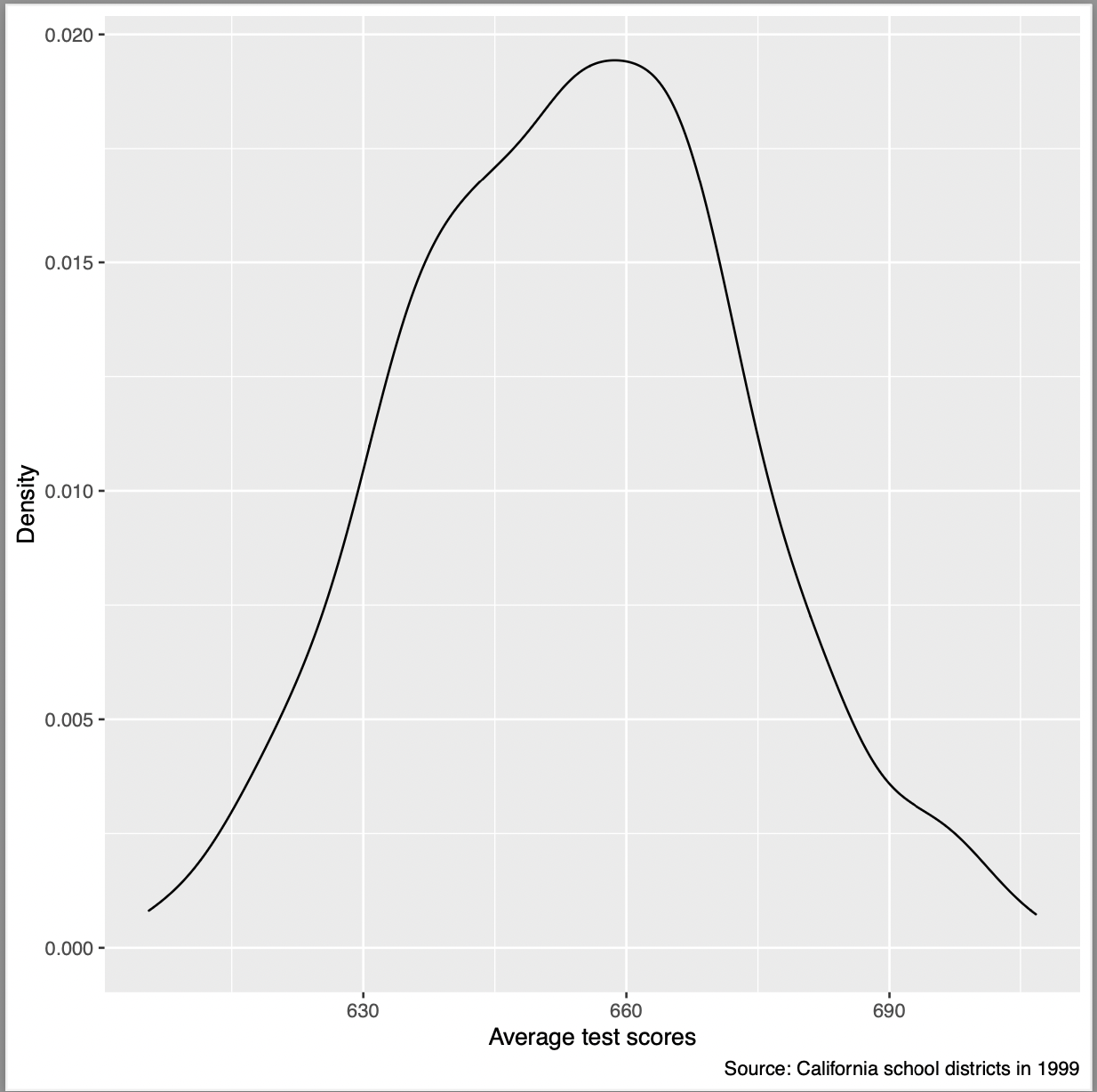
[1] 654.1565

> mfv(ClassSize$testscr)

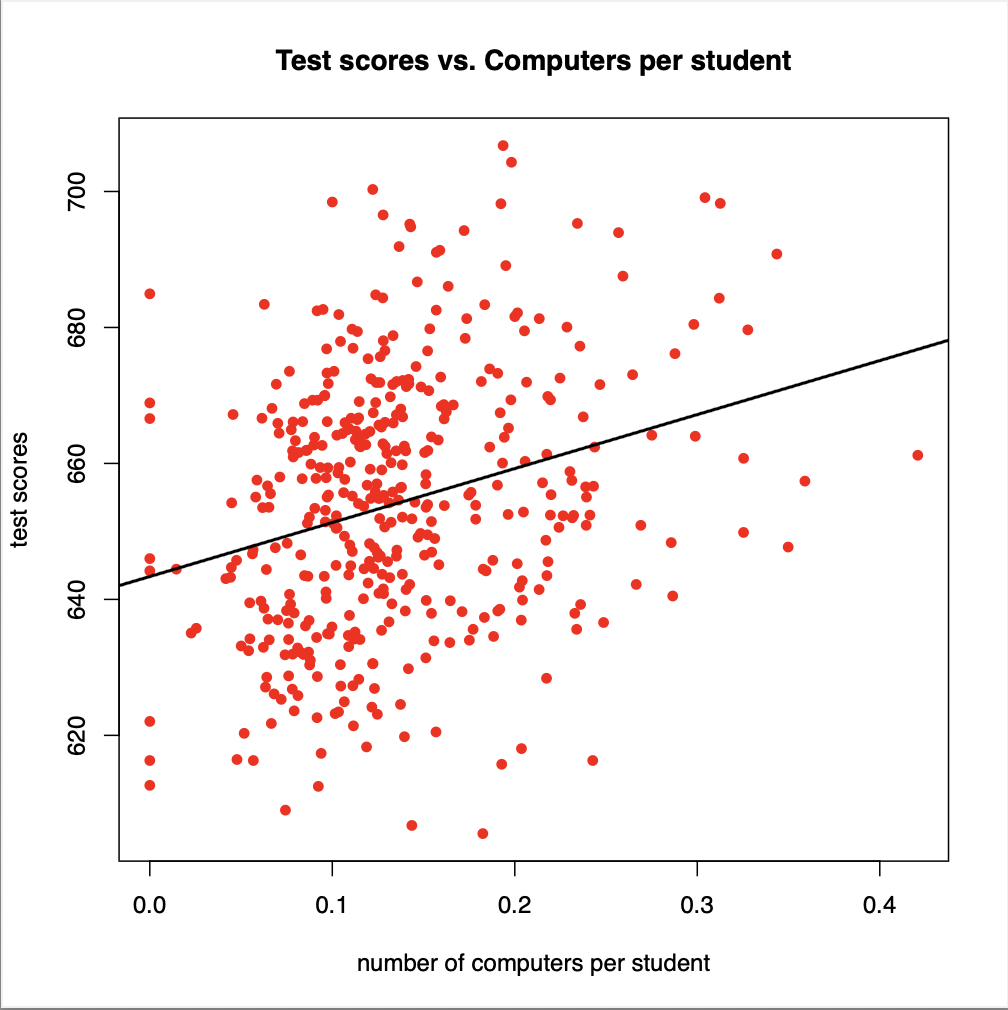
[1] 616.30 661.85

> median(ClassSize$testscr)

[1] 654.45



**(b) Draw a scatterplot of test scores versus number of computers per student. Describe in words what you see.**



According to the scatterplot, overall, there is a positive association/correlation between the independent variable ‘compstu’ - number of computers per student and the dependent variable ‘testscr’ - test scores of student, but it is not strong. However, one of the highest scores (more than 700) are achieved by students who have access to average fraction of a computer (0.2), not by students with the highest (0.4). Still, these are outliers because students with the same access also got one of the lowest marks (less than 620).Populations