Statistical Methods for Causal Inference, Practical 1:

Aim: In the first practical of Statistical Methods for Causal Inference, we will delve into advanced techniques to enhance your understanding of statistical modelling and inference. In the first part, we will explore the fundamentals of data visualization, specifically focusing on running scatterplots and fitting regression lines to estimate the direction of the relationship between variables. We will also delve into running both simple and multiple linear regression models, allowing you to analyse and interpret the regression output, including coefficients (alpha, beta), t-statistics, p-values, and various sum of squares measures (TSS, ESS, RSS). Additionally, you will learn how to obtain residuals, test for multicollinearity, and identify potential heteroskedasticity issues. We will also cover the concept of heteroskedasticity-robust inference, enabling you to conduct reliable inference even in the presence of heteroskedasticity. Furthermore, we will introduce logistic regression, providing you with a powerful tool for analysing binary outcomes. You will also gain insight into obtaining marginal effects of the explanatory variables, allowing for a deeper understanding of the impact of each predictor.

In the second and third part of the practical, our focus will shift towards the critical concepts of causality and experimental design. We will emphasize the importance of recognizing that correlation does not imply causation, highlighting the need to consider confounding factors in causal inference. Furthermore, we will explore the benefits of randomization in experimental design, enabling unbiased estimation of treatment effects. To ensure the validity of such estimations, we will discuss and conduct baseline balance checks between treatment and control groups, ensuring that the groups are comparable before the intervention. Finally, we will guide you through various methods for estimating treatment effects, equipping you with the necessary tools to quantify the impact of interventions accurately.

Logistics: Please make sure that you have downloaded the .zip file containing all the necessary materials for Practical 1. There you will find the Practical 1 exercises and the corresponding data file we will work with (Alan2021_final). It is strongly recommended that you extract these files to your computer in a sensible place (e.g. create a folder for Statistical "Methods for Causal Inference" with a sub-folder for "Seminars" and then extract all the files in there).

Context: The data used in this seminar is derived from a research study conducted by Alan et al. (2021) that investigates the impact of an educational program on social cohesion in ethnically diverse schools, specifically those affected by an influx of Syrian refugee children in Turkish elementary schools. The study aims to evaluate the effectiveness of the program in developing perspective-taking ability among children and its subsequent effects on various indicators of a cohesive school environment. The researchers employ a randomized approach to program implementation, allowing for causal inference.

By analysing this research, you will gain valuable insights into the application of econometric and experimental techniques in studying the positive effects of an educational program on social cohesion in ethnically diverse schools. Through the examination of real-world data, you will develop a deeper understanding of the importance of perspective-taking and its implications for creating a cohesive and inclusive learning environment. These insights will enhance your econometric and experimental knowledge, enabling you to effectively study and evaluate similar interventions in the future.

Computer tutorial: Work through the following exercises during the class. You can also find all the questions in the do-file which will be provided. Make sure to download and place the dataset for this seminar (named Alan2021_final) in your Seminars folder.

Part 1: Review of parametric regressions

Exercise 1: We want to understand whether there is any relationship (and if so, what the direction of the relationship is) between the class size and the empathy (perspective taking) of the students in our available dataset. In order to do this, we will start with simple visualization of the relationship between the two variables.

- Q1a: In Stata, using a scatterplot, visualize the relationship between the variable "srefshare" and "b_schoolsize". Is this a good visual representation?
- Q1b: Add a line of best fit in the scatterplot. Is the direction of association between the two variables expected?

Exercise 2: Running a simple linear regression in Stata, uncover the relationship between the "level of empathy of the students" (Y) and the "class size" (X).

- Q2a: Comment on the significance and the meaning of the explanatory variable and interpret the intercept (β0) and slope (β1) coefficients.
- Q2b: Interpret the coefficient of determination (R-Squared).

Exercise 3: In this exercise, you will expand the analysis by running a multivariate regression model, incorporating additional explanatory variable for: age, gender, refugee, developmentally challenged students and school size.

- Q3a: Comment on the significance and the meaning of the explanatory variables and interpret the intercept (β0) and the slope coefficients. Note that we now have three dummy variables as regressors.
- Q3b: Interpret the coefficient of determination (R-Squared). How has it changed? Why?
- Q3c: Using Variance Inflation Factor (VIF) test, see if there is multicollinearity in the regression model. Interpret the results.
- Q3d: Performing a Breusch-Pagan test, see if there is heteroskedasticity in the regression model. Interpret the results. Note that the null hypothesis indicates that the error variance is constant (homoscedastic).
- Q3e: Estimate again the same regression model, this time accounting for the heteroskedasticity using robust standard errors.
- Q3f: Run a F-test for joint significance in Stata to test the null hypothesis that the coefficients of all the independent variables in the regression model are equal to zero, against the alternative hypothesis that at least one of the coefficients is not zero. Discuss your findings.

Exercise 4: Run a binary logistic regression using the command logit, where the binary dependent variable is "bullying in class", and the predicting variables are: age, gender, refugee, developmentally challenged students, school size and class size.

- Q4a: Interpret the obtained odds ratios.
- Q4b: Calculate the marginal effects of the continuous variable "age in months" in the model on the events of bullying.

Part 2: Randomised experiments – Simulations

In causal inference, we are generally interested in examining the effect of a key explanatory variable (which we often refer to as the "treatment") on an outcome of interest. We are in general attempting to identify the effect that this key explanatory variable has on the outcome (the so-called "treatment effect"), to the exclusion of any plausible competing explanations. If we cannot reasonably exclude all possible competing explanations for variation in an outcome that we want to attribute to the key explanatory variable, then we have a threat to inference known as a confounding explanation.

Formally, if D is an indicator of treatment group status (treatment or control) and Y is an outcome, a confounder, X, is any factor that is correlated with both D and Y . If we see a statistical relationship between D and Y , we will not know whether it is because D is causing variation in Y , or variation in either or both is a function of X.

Simulation: Now we will simulate some data to illustrate some general principles about experimental research within the potential outcomes framework.

Using Stata, we generate a population of size N = 5000, with potential outcomes Y0 and Y1 constructed as a linear function of the measurement X1, which is drawn from a normal distribution with a mean of 1 and standard deviation of 2. Y0 is a linear function of X1, parameters, and random error, while Y1 is an additive function of Y0, a constant, and random error.

```
* Sample size of 5000

set obs 5000

* Continuous variable, drawn from a normal distribution with mean 1 and s.d. 2

gen x1 = rnormal(1,2)

* Coefficients determining data generating process for outcome variable

scalar b0 = 5

scalar b1 = 0.5

* Outcome variable, linear function of coefficients, x1, and standard normal distributed error term

generate y0 = b0+b1*x1 + rnormal(0,1)

* Generate potential outcomes representing a positive treatment effect that is highly statistically significant

egen y0_mean = mean(y0)

gen y1 = y0 + y0_mean + rnormal(0,1)
```

Before moving on, recall that our ability to see both potential outcomes for a case only exists in the world of simulations. In real-world data, we will only see an outcome called Y , which is either Y0 or Y 1 depending on the value of the treatment indicator D.

Since we know both potential outcomes for each case, then we would be able to calculate the difference between groups as a simple difference in the means of Y1 and Y0.

```
ttest y1 == y0
```

We find that the true value of average treatment effect is 5.53

In reality, as you learned in the lecture, units are assigned a value of a treatment indicator that determines the treatment group to which a unit belongs. In other words, the value of D that a unit is assigned determines whether we consider an observed outcome to be Y0 or Y1.

Non-randomised treatment:

What if treatment D were assigned (by whatever mechanism did the "assigning") so that cases with lower values of X1 are more likely to receive treatment? If constructed this way, then X1 and D would be correlated by definition. Because we will deal with several treatment indicators, let us call this one D1.

```
egen x1_median = median(x1)

gen d1 = 0
```

Because both Y and D1 were produced as a function of X1, then we have a classic confounding problem; namely, if our goal is to study the causal relationship between D1 and Y , then we have to account for X1, lest it remain as a competing explanation.

What are the implications of assigning treatment in this way? To answer this question, we generate outcome variable determined by the non-randomised treatment D1.

```
gen outcome = cond(d1 == 0, y0, y1)

reg outcome d1
```

Estimated effect = 3.94. We see that when we estimate the relationship between Y and D1, while not accounting for the fact that X1 is correlated with both Y and D1, the treatment effect estimate is about 30% smaller than the true value of 5.53, a substantial difference.

Randomised treatment:

```
gen d2 = runiform() < 0.5

gen outcome2 = cond(d2 == 0, y0, y1)

reg outcome2 d2
```

Estimated effect = 5.50. With the randomized treatment indicator, on the other hand, the estimate is much closer to the true value (about 1% away from the true value).

Part 3: Randomised experiment

Exercise 1: In this exercise, we will conduct baseline balance checks between the control and treatment groups in the "Alan2021_final" dataset. To do so, we first need to identify the treatment variable and the baseline variables.

Q1a: Check for balance between control and treatment groups, for three baseline variables in the dataset that are measures of social ties between the student. Please note: the variable in the dataset are called "bnode_in_friend", "bnode_in_supportself" and "bnode_in_studyself", and they indicate friendship ties, emotional support by peers and academic support by peers, respectively.

Q1b: Again, check for balance for the previous three social ties baseline indicators, but this time, instead by treatment group, use the following endogenous variables "refugee" and "astudent", which indicate whether the students is a refugee, or is developmentally challenged. What insights can you provide and why?

Exercise 2: In this exercise, we will estimate the effects of the treatment on student and teacher reports of violence and antisocial behaviour. Our outcome variables will be "fsbully_c", "fsbully_s" and "tbully_f", which stand for in-class bullying, out-class bullying and teacher behavioural grade. In addition, we will control for the following background characteristics: age, gender, refugee, developmentally challenged student, raven test score, eyes test, school size, class size and district.

Exercise 3: In this exercise, in our analysis we will employ interaction terms. We will estimate the heterogenous effects of the treatment on student and teacher reports of violence and antisocial behaviour, both for refugees and Turkish students. Again, our outcome variables will be "fsbully_c", "fsbully_s" and "tbully_f", which stand for in-class bullying, out-class bullying and teacher behavioural grade. In addition, we will control for the following background characteristics: age, gender, refugee, developmentally challenged student, raven test score, eyes test, school size, class size and district.