KEY QUESTIONS REVIEW

HOW RANDOMIZED CONTROLLED TRIALS WORK?

The problem for causal inference is that an individual cannot be treated and untreated at the same time. Therefore, instead of comparing individuals we compare groups. If the treatment group and the control group have similar characteristics, and if we only administer the treatment to the former, then we can attribute subsequent differences in outcomes to the treatment.

A randomized controlled trial works by allocating a large number of individuals randomly to either of the two groups. Random allocation and the law of large numbers imply that the two groups have the same characteristics on average. Therefore, if the randomization is successful and the two groups are large, we interpret differences in outcomes as causal effects of treatment.

If individuals are not randomly assigned to the treatment and control groups, there is the risk of systematic differences in (observed and unobserved) characteristics other than the treatment. In this case, we cannot know whether observed differences in outcomes is caused by the treatment or whether they represent other factors just as reverse causality, unobserved heterogeneity, or self-selection.

Random allocation can be difficult to implement in practice, both in the initial stage and during the trial (sample attrition). It is therefore important to check that the final samples are balanced.

WHAT IS AVERAGE TREATMENT EFFECT?

The framework defines potential outcomes with treatment, Y_{1i}, and potential outcome without treatment, Y_{0i}. The observed outcome is $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$, where D_i is an indicator of treatment. The counterfactual outcome is the potential outcomes which is not unobserved.

With this notation, the causal effect of treatment for entity i is $Y_{1i} - Y_{0i}$. The average treatment effect is

$$ATE = E(Y_{1i} - Y_{0i}) = E(Y_{1i}) - E(Y_{0i})$$

If individuals are randomly allocated to the treatment and control groups, then

$$E(Y_{0i}) = E(Y_{0i}|D_i = 1) = E(Y_{0i}|D_i = 0)$$

$$E(Y_{1i}) = E(Y_{1i}|D_i = 1) = E(Y_{1i}|D_i = 0)$$

so the average treatment effect can be written as:

$$\begin{aligned} ATE &= E(Y_{1i}) - E(Y_{0i}) \\ &= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \\ &= E(Y_i|D_i = 1) - E(Y_i|D_i = 0) \end{aligned}$$

The last equation is important, because Y_i is observed while Y_{1i} and Y_{0i} are not.

In words, the result means that the difference in average outcomes between the treatment and control groups is the average treatment effect.

WHAT IS THE LAW OF LARGE NUMBERS AND THE CENTRAL LIMIT THEOREM?

Khinchin's law of large numbers say that if observations are independently sampled from the same distribution and the sample size is large, then the sample average will be close to the population average with high probability. (Assuming the population mean exists.) Formally, the probability that the (absolute) difference between the sample mean and the population mean exceed a fixed constant will converge to zero as the sample size increases. Intuitively, if sampling was done randomly and the sample size is large, knowing the sample average is as good as knowing the population average.

The central limit theorem (CLT) concerns the sampling distribution of the sample average. The Lindeberg-Levy CLT says that if the observations in a sample are *iid* random variables and the sample size is large, then sampling distribution of the standardized sample average (obtained by subtracting the mean of the average and dividing by the standard deviation of the average) will be close to a standard normal distribution. (Assuming the population variance exists.) This means that we can approximate the sampling distribution of the sample mean using a normal distribution with appropriate mean and variance.

WHAT IS SELF-SELECTION BIAS?

When we estimate treatment effects we (usually) compare outcomes between a treatment group (who experience the treatment) and a control group (consisting of untreated individuals).

In general, differences in outcomes between the treatment and control group can be a causal effect of the treatment, or they can reflect the fact that the two

groups consist of different individuals who have different characteristics and different experiences, and who make different choices.

The term self-selection refer to phenomenon that when the treatment and control groups are not formed in a controlled manner (ie not controlled by an external experimenter) and individuals self-select into one of the groups, they have their self-interest in mind and so naturally the treatment group will consist of individuals who prefer or benefit from the treatment and the control group will consist of people who prefer or benefit from non-treatment.

This implies that individual characteristics are not identically distributed in the two groups, so difference in outcomes cannot be attributed to the treatment alone.

Example: Suppose we compare participants and non-participants in a job training program to identify a causal effect of the job training on labour market outcomes. If treatment is self-selected, more productive, more motivated, or younger people may be more willing to attend the job training. In this case, the treatment and control groups may differ substantially in productivity, motivation, or age. This implies the difference in labour market outcomes between treatment and control groups may be either due to the job training, or due to the difference in the aforementioned factors, or both.

Even in randomised controlled trials, self-selection issues can arise at the end of the trial if participants can choose to drop out or continue during the trial. Attrition from the trial leads to the same problem described above, where at the end of the trial the two groups are not comparable.

WHAT IS OMITTED VARIABLE BIAS?

Omitted variable bias and selection bias essentially refer to the same issue, namely that an estimate does not have causal interpretation. The bias arises when there are confounding factors (unobserved heterogeneity) that are not controlled for. That is, when we compare average outcomes for groups with different levels of treatment we are not holding other things equal (on average).

We tend to use the term selection bias in a non-regression context (e.g. when we analyse a randomised controlled trial) and when we want to emphasize that the confounding factors have to do with the choices individuals make. For example, people choosing different levels of health insurance based on their view of their own health condition and their general risk attitude.

The term omitted variable bias is used more generally, and also covers confounding factors that individuals have little control over. For example, the climate may play a role when comparing health outcomes between Sydney and Hobart.

As another example, suppose we seek to estimate the causal effect of education on labour market outcomes. Students with higher ability probably tend to invest more in education. Since students self-select their levels of education, comparison among different levels of education leads to selection bias. If we run a regression of labour outcomes on years of education without controlling for ability we are likely to see some omitted variable bias.

WHAT IS FULL RANK ASSUMPTION IN MULTIPLE REGRESSION?

The full rank assumption is needed in order to identify the population parameters of interest. If the full rank assumption is not satisfied, the parameters that determine the regression function are not uniquely determined.

The full rank assumption fails when there are linear dependencies among the regressors. For example, suppose we estimate gender differences in labour market outcomes. In estimation, including both male and female dummies and a constant term will introduce a linear dependency, because $female\_dummy + male\_dummy - 1 = 0$.

As another example, suppose we estimate effects of time allocation on academic achievement. If we include time sleeping, time awake, and also total time, there is a linear dependence if $time\_sleeping + time\_awake = total\_time$.

HOW DOES MULTIPLE REGRESSION REVEAL CAUSAL EFFECT?

The linear (in parameters) regression model allows estimation of the population regression function, so the question is really about whether the population regression function with control variables identify causal effects.

Let Y_i be the dependent variable, let T_i be treatment variable, and let $\boldsymbol{A}_i$ be a vector of control variables. Consider the modern causal model

$$Y_i = \gamma T_i + \boldsymbol{A}_i^T \boldsymbol{\delta} + \lambda + V_i$$

In this model, γ may represent the casual effect of T_i on Y_i, if we are able to hold all other things equal. Mathematically, the key condition is

$$E(V_i|T_i, \boldsymbol{A}_i) = E(V_i|\boldsymbol{A}_i)$$

That is, holding $\boldsymbol{A}_i$ constant, the average of V_i is the same for all values of T_i, as good as randomised.

But there is still the caveat that holding other things equal cannot distinguish between causal effects, reverse causality, and simultaneity. And the interpretation of $\boldsymbol{\delta}$ is still difficult – a mixture of causality, confounders, omitted variable bias.

HOW DOES THE DIFFERENCE IN DIFFERENCES APPROACH ESTIMATE CAUSAL EFFECTS?

DD estimates the difference between the changes of the treatment group and control group across time. In notation, let Y_i denote the outcome, G_i be the group dummy (1 if treated, 0 otherwise), and P_i be the time period dummy (1 if aftermath, 0 otherwise). The DD estimator of the average treatment effect is:

$$\begin{aligned} ATE = &\{E(Y_i|G_i = 1, P_i = 1) - E(Y_i|G_i = 1, P_i = 0)\} \\ &- \{E(Y_i|G_i = 0, P_i = 1) - E(Y_i|G_i = 0, P_i = 0)\} \end{aligned}$$

DD can be estimated using a regression framework,

$$E(Y_i|G_i, P_i) = \alpha G_i + \eta P_i + \delta G_i P_i + \kappa$$

where δ captures the DD effect.

If we can assume that the treatment group would experience the same change as the control group without the treatment (*counter-factual outcome*), then the difference in the changes of the two group must be attributed to the treatment effect. This is known as the common trends assumption. DD is only valid if the common trends assumption holds.

The *common trends* assumption essentially states, in the absence of intervention, the group-specific differences are constant over time. One way to provide support for the common trend assumption is to show that the trends are similar in some periods before treatment, and the treatment is the only new change.

WHAT IS HETEROGENEITY BIAS IN ESTIMATING A FIXED EFFECTS MODEL?

Heterogeneity bias is a kind of omitted variable bias that is caused by omitted variables that are fixed for an individual over time. In a fixed effects model,

$$Y_{it} = \boldsymbol{\beta}^T \boldsymbol{X}_{it} + \lambda_2 B2_{it} + \cdots + \lambda_T BT_{it} + V_{it}$$

the error term can be decomposed as:

$$V_{it} = A_i + U_{it}$$

A_i is called the *fixed effect*, which represents time-invariant unobserved factors; U_{it} is called the *idiosyncratic error term*, which represents time-varying unobserved factors.

In a regression, if the aggressor $\boldsymbol{X}_{it}$ is correlated with A_i, then the OLS estimator is bias. This bias induced by failing to control the time-invariant factors in analyzing panel data is called heterogeneity bias.

The heterogeneity bias in this case can be avoided by using First-Difference (FD), Fixed-effects (FE), or Least Square Dummy Variable (LSDV) techiques.

HOW DOES AN INSTRUMENT VARIABLE WORK?

A good instrument variable effects assignment to treatment group, but does *not* have indirect effects through unobservables on potential outcomes, nor any direct effects on potential outcomes. If an instrument variable satisfies these requirements, then using the instrument automatically sorts people into two groups where the treatment effect is different across groups, but the unobserved variables have the same means.In this way, it works like a proxy for random assignment. So any difference in outcomes must be attributable to the treatment.

Mathematically, suppose there is a *structural* model:

$$Y_i = \beta X_i + \gamma Z_i + U_i$$

where $E(U_i|X_i) \neq 0$. If this model is estimated directly by OLS, $\hat{\beta}$ will surely be biased.
Suppose we have an instrument Z_i that satisfies:

- Instrument relevance: $Cov(Z_i, X_i) \neq 0$;
- Instrument exogeneity: $Cov(Z_i, U_i) = 0$;
- Exclusion restriction: $\gamma = 0$;

Since $Cov(Z_i, X_i) \neq 0$, we can construct a linear projection:

$$X_i = \pi Z_i + V_i$$

where $\pi \neq 0$ and $E(V_i|Z_i) = 0$. And the reduced form model for Y_i:

$$Y_i = \beta\pi Z_i + (U_i + \beta V_i) = \delta Z_i + R_i$$

Since Z_i is uncorrelated with U_i, OLS is consistent for δ and π. And we can recover the true β by the ratio δ/π.

WHAT IS MAXIMUM LIKELIHOOD ESTIMATION?

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ whose probability distribution depends on some unknown parameter θ. A maximum likelihood estimator $\hat{\theta}$ would be the value that maximizes the probability of getting the data observed.

In practice, suppose the probability of observing $X_i = x_i$ is given by

$$P(X_i = x_i) = f(x_i; \theta)$$

Then the *likelihood function*

$$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n) = \Pi_{i=1}^{n} f(x_i; \theta)$$

represents the probability of getting all the data points observed. A good estimator $\hat{\theta}$ would be the one that maximizes $L(\theta)$.

If the following assumptions are satisfied:

- The model is correctly specified (linear or nonlinear);
- The first-order conditions have a unique solution (full rank);
- Very large observations are unlikely;
- Random sampling;

Then the ML estimator $\hat{\theta}$ is unbiased and has approximately normal distribution.

WHAT IS STATIONARY TIME SERIES?

A time series $\{Y_t\}$ is *stationary* if its probability distribution (mean, variance, etc.) does not change over time. If a time series is stationary, historical relationships in the series can be generalized to the future.

For example, white noise is stationary. Let V be any scalar random variable, then the time series

$$Y_t = V \text{ for all } t$$

is stationary.

The *linear deterministic trend* model

$$Y_t = \beta_1 t + \beta_0 + U_t$$

is non-stationary. Since $E(Y_t) = \beta_1 t$ is not the same for all t.

The *random walk*

$$Y_t = Y_{t-1} + U_t, \qquad E(U_t | Y_{t-1}, Y_{t-2}, \ldots) = 0$$

is non-stationary. Since the variance $Var(Y_t) = Var(Y_{t-1}) + Var(U_t)$ increases over time.

If a time series is not stationary, then the asymptotic sampling distribution is not normal, and the OLS standard errors are invalid are misleading. This is known as *spurious regression*.

Stationarity can be tested by *Dickey-Fuller Test*. Consider the $AR(p)$ model:

$$\Delta Y_t = \alpha t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \cdots + \gamma_p \Delta Y_{t-p} + \beta_0 + U_t$$

Dickey-Fuller tests:

H_0: $\delta = 0$ (Y_t is a stochastic trend)

H_1: $\delta \neq 0$ (Y_t is stationary)

αt is included in the model for testing stationarity around a deterministic trend.

WHAT IS AN AUTOREGRESSIVE DISTRIBUTED LAG MODEL?

An autoregressive distributed lag model (ADL(p, q)) is defined as:

$$E(Y_t | Y_{t-1}, \ldots, Y_{t-p}, \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-q}) = \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \boldsymbol{X}_{t-1}^T \boldsymbol{\delta_1} + \cdots + \boldsymbol{X}_{t-q}^T \boldsymbol{\delta_q} + \beta_0$$

The OLS estimator is asymptotically normally distributed if:

- The model is *dynamically complete* i.e.

$$E(Y_t | Y_{t-1}, \ldots, Y_{t-p}, \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-q}) = E(Y_t | Y_{t-1}, Y_{t-2}, \ldots \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots)$$

- The distribution of $\{Y_t, \boldsymbol{X}_t\}$ is *stationary*;
- The series $\{Y_t, \boldsymbol{X}_t\}$ is *weakly dependent* i.e. $\{Y_t, \boldsymbol{X}_t\}$ and $\{Y_{t-j}, \boldsymbol{X}_{t-j}\}$ are independent for large j;
- Large values of $\{Y_t, \boldsymbol{X}_t\}$ are unlikely (forth moments exists);
- No perfect collinearity.

If all above assumptions hold, the LLN and CLT work for time series, and we can do statistical inference as usual.