|  |  |
| --- | --- |
| **Definition** | |
| Causal effects | An action is said to have cause an outcome if the outcome is the direct result, or consequence, of that action |
| Randomized controlled experiment | with double-blinded, it's the gold standard for investigating the causal effects |
| Treatment group | Group that undergo the treatment of interest |
| Control group | Having a control group permits measurements of the treatment effect |
| Randomized | Subject from the population are randomly assigned into treatment and control groups |
| Experimental Data | Comes from experiments designed to evaluate the policy/treatment effect |
| Observational Data | Obtained by observing actual behaviour outside of a experiment setting |
| Confounding factors | Other differences or factors contaminate estimation of the casual effects |
| Cross sectional data | data on multiple entities for a single period |
| Time series data | data for a single entity, collected at multiple time periods |
| Panel/ Longitude data | Data on multiple entities, in which entity is observed for two or more time periods |
| Random variable | A variable whose value is an outcome of a random phenomenon |
| Discrete variable | Only takes on a discrete set of values |
| Continuous variable | Takes on a continuum of possible values |
| Probability distribution | Lists the values and the probability that each value will occur |
| Normal distribution | ; 68-95-99.7 rule |
| Statistical inference | Using statistical methods to draw inferences from random sample about the full population |
| Estimation | Computing a best guess numerical value for an unknown characteristic of a population distribution, from a sample of data |
| Hypothesis testing | Formulating a specific hypothesis about the population, then using sample evidence to decide whether it is true |
| Confidence intervals | Computing an interval for an unknown population characteristic, using a sample of data |
| Estimator | An estimator is a procedure or a formula used to obtain an estimate of the parameter of interest.  It is a function of the randomly drawn sample of data  It is a Random Variable because it depends on a randomly selected sample |
| Estimate | An estimate is a numerical value of the estimator when it is computed using data from a specific sample  An estimate is just a number and so is non-random |
| Simple random sampling (SRS) | Every sample has the equal probability of being chosen |
| Independently and Identically distributed | i.i.d implies that distribution Yi shares the same distribution as the population |
| Bias |  |
| Central Limit Theorem |  |
| Law of large numbers |  |
| Consistency |  |
| Efficiency |  |
| Desirable Properties of Estimators | Unbiased, Consistent, Efficient  is the BLUE(best linear unbiased estimator) |
| Hypothesis Testing | Hypothesis tests are based on a statistics which estimates the parameter of interest |
| Null Hypothesis | A hypothesis to be tests, usually a statement of "no effect" or "no difference |
| Alternative Hypothesis | A hypothesis we test the null against, this is the statement we hope is true |
| Simple/ Composites | A hypothesis test is simple if it specifics a value for the parameter tested  , else it's composite. |
| Significance level | Fixed benchmark to reject H0 |
| Type I error | Rejecting the null hypothesis when it is true |
| Type II error | Not rejecting the null hypothesis when it is false |
| Confidence interval | An interval that contains the true value of a parameter with a certain prespecified probability |
| Linear Regression Model | Measures the average relationship between factors and outcomes |
| Ordinary Least Squares (OLS) | OLS chooses the estimators so that the estimated regression line comes as close as possible to the data points, where "closeness" is measured by the sum of the squared mistakes made in predicting Y given X |
| Measure of Fit | How well the OLS regression line describes the data depends on two regression statistics: R2 and Standard error of the regression |
| R2 | Measures the fraction of the variances of Y that is explained by X. It is unitless and ranges between zero (no fit) and one (perfect fit) |
| SER | Standard Error of the Regression is a measure of the spread of the observations around the regression line (measure in units of Y). SER is the sample standard deviation of the OLS residuals. |
| Least Squares Assumptions | 1. Given Xi, the conditional distribution of the population error term ui has a mean zero  2. Samples are independently and identically distributed (i.i.d)  3. Large outliers are rare |
| Report Regression Results |  |
| Regression when X is a Binary Variable | is the coefficient of Xi , which is the difference between mean outcome when X = 0 and X = 1  Note: Does not have graphical interpretation and is not the slope |
| Heteroskedasticity | Variance of error changes depending on the value of  In practice, always use heteroskedasiticity-robust standard errors (STATA: regress y x, robust) |
| Homoskedasticity | Variance of error does not depend on the value of , has the same spread regardless of |

|  |  |
| --- | --- |
| Calculation | |
| Mean | A measure of the centre of a distribution |
| Sample Average |  |
| Variance/standard deviation | Measures of the spread or dispersion of a distribution |
| Covariance | covariance is a measure of how much two random variables vary together |
| Standard normal distribution |  |
| Test statistic |  |
| Sample standard deviation |  |
| Standard Error |  |
| Linear Regression Model | ; |
| Sum of squared prediction mistakes | where and are some estimators of and |
| OLS estimators |  |
| R2 |  |
| SER | Since =0, given by the definition of OLS, min sum of squared prediction mistakes |
| Root Mean Squared Error | Given that n is large (Law of large numbers) |
| Unit Change in regression |  |
| LSA #1 | therefore |
| E( |  |
| Var( |  |
| Normal approximation of |  |
| SE( |  |
| Regression when X is a Binary Variable | where |
| Homoskedastic | Given  then |