Ling – EC3303

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	$X \sim N(\mu, \sigma^2)$; 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 Y_i shares the same distribution as the population
Bias	$bias = E(\hat{\theta}) - \theta$
Central Limit Theorem	$\lim_{n\to\infty} \bar{Y} \to N(\mu_Y, \frac{\sigma_Y{}^2}{n})$
Law of large numbers	$\lim_{n\to\infty} \bar{Y} \overset{p}{\to} \mu_Y$
Consistency	$\lim_{n\to\infty} \hat{\theta} \overset{p}{\to} \theta, bias = 0$ $\lim_{n\to\infty} var(\hat{\theta}) = 0$
Efficiency	$\tilde{\theta}$ is more efficient if $var(\tilde{\theta}) < var(\hat{\theta})$
Desirable Properties of Estimators	Unbiased, Consistent, Efficient $\bar{Y}$ is the BLUE(best linear unbiased estimator)

Hypothesis Testing	Hypothesis tests are based on a statistics which estimates the parameter of interest
Null Hypothesis H_0	A hypothesis to be tests, usually a statement of "no effect" or "no difference
Alternative Hypothesis H_a	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 $H_0 = a$, else it's composite. $H_0 \ > a$
Significance level	Fixed benchmark to reject H_0
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: R^2 and Standard error of the regression
R^2	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 X_i, the conditional distribution of the population error term u_i has a mean zero 2. Samples are independently and identically distributed (i.i.d) 3. Large outliers are rare
Report Regression Results	$$\hat{Y} = \frac{\widehat{\beta_0}}{SE(\widehat{\beta_0})} + \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})} x, R^2, SER$$
Regression when X is a Binary Variable	β_1 is the coefficient of X_i, which is the difference between mean outcome when X = 0 and X = 1 Note: Does not have graphical interpretation and β_1 is not the slope
Heteroskedasticity	Variance of error changes depending on the value of x_i $$Var(u_i \mid X_i = x) \neq {\sigma_u}^2$$ In practice, always use heteroskedasiticity-robust standard errors (STATA: regress y x, robust)
Homoskedasticity	Variance of error does not depend on the value of x_i, has the same spread regardless of x_i $$Var(u_i \mid X_i = x) = {\sigma_u}^2$$

Calculation	
Mean	A measure of the centre of a distribution $$E(X) = \mu_X = \sum_{i=1}^{k} x_i p_i$$ $$E(a) = a;\ E(a+bX) = a + bE(X);\ E(X+Y) = E(X) + E(Y)$$
Sample Average $\bar{Y}$	$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + .. + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$ $$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{n\mu_Y}{n} = \mu_Y$$ $$var(\bar{Y}) = \sigma_{\bar{Y}}{}^2 = \frac{1}{n^2}\sum_{i=1}^{n} var(Y_i) = \frac{\sigma_Y{}^2}{n}$$
Variance/standard deviation	Measures of the spread or dispersion of a distribution $$var(x) = \sigma_X{}^2 = E[(X-\mu_X)^2] = \sum_{i=1}^{k} (x_i - \mu_X)^2 p_i$$ $$sd = \sigma_X = \sqrt{var(x)}$$ $$var(a) = 0, var(aX - bY) = a^2 var(x) + b^2 var(Y) - 2ab cov(X,Y)$$
Covariance	covariance is a measure of how much two random variables vary together $$cov(X,Y) = \sum_{i=1}^{k} E((X_i - \bar{X})(Y_i - \bar{Y}))$$
Standard normal distribution	$$Z = \frac{X-\mu}{\sigma} \sim N(1,0)$$
Test statistic	$$\frac{X - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \frac{X - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$$
Sample standard deviation	$$s_y{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
Standard Error	$$\frac{s_y}{\sqrt{n}}$$
Linear Regression Model	$\beta y = \frac{\Delta x}{\Delta y}$; $y_i = \beta_0 + \beta_{x_i} x_i + u_i$
Sum of squared prediction mistakes	$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 = \sum \hat{u}^2$$ where b_0 and b_1 are some estimators of β_0 and β_1
OLS estimators	$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$ $$\beta_1 = \frac{cov(x,y)}{var(x)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
R^2	$$R^2 = \frac{Explained\ sum\ of\ squares}{Total\ sum\ of\ squares} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$ $$= 1 - \frac{Sum\ of\ Squared\ residuals}{Total\ sum\ of\ squares} = 1 - \frac{\sum(\hat{u}_i)^2}{\sum(y_i - \bar{y})^2}$$
SER	$$SER = \sqrt{\frac{1}{n-2}\sum(\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-2}\sum(\hat{u}_i)^2}$$ Since $\bar{\hat{u}}$=0, given by the definition of OLS, min sum of squared prediction mistakes
Root Mean Squared Error	$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{u}_i)^2}$$ Given that n is large (Law of large numbers)

Unit Change in regression	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_i x_i$ $\hat{y}_i \; to \; \widehat{ay}_i: \widehat{ay}_i = a\hat{\beta}_0 + a\hat{\beta}_i x_i$ $x_i \; to \; bx_i: \hat{y}_i = \hat{\beta}_0 + \frac{\hat{\beta}_i}{b}(bx_i)$
LSA #1	$E(u_i \vert X_i = x) = 0$ $Cov(u_i, X_i) = 0$ therefore $E(\widehat{\beta_1}) = \beta_1 \; and \; E(\widehat{\beta_0}) = \beta_0$
E($\widehat{\beta_1}$)	$E(\widehat{\beta_1}) \; = \; E(\frac{cov(x,y)}{var(x)}) = \beta_1$
Var($\widehat{\beta_1}$)	$Var(\widehat{\beta_1}) \; = \; Var(\frac{cov(x,y)}{var(x)}) = \frac{1}{n} \times \frac{Var[(X_i - \mu_x)u_i]}{[Var(X_i)]^2}$
Normal approximation of $\widehat{\beta_1}$	$\widehat{\beta_1} \sim N(\beta_1, \frac{1}{n} \times \frac{Var[(X_i - \mu_x)u_i]}{[Var(X_i)]^2})$
SE($\widehat{\beta_1}$)	$SE(\widehat{\beta_1}) = \sqrt{var(\widehat{\beta_1})} = \frac{1}{\sqrt{n}} \times \frac{\sqrt{Var[(X_i - \mu_x)u_i]}}{Var(X_i)}$
Regression when X is a Binary Variable	$\beta_1 \; = \; E(Y_i \vert X_i = 1) \; - \; E(Y_i \vert X_i = 0)$ where $Y_i = \beta_0 + \beta_1 X_i + u_i$ $E(Y_i \vert X_i = 1) \; = \; \beta_0 + \beta_1 + u_i$ $E(Y_i \vert X_i = 0) \; = \; \beta_0 + u_i$ $E(Y_i \vert X_i = 1) \; - \; E(Y_i \vert X_i = 0) \; = \; \beta_1$
Homoskedastic	Given $Var(u_i \vert X_i = x) = {\sigma_u}^2$ then $Var(\beta_1) = \frac{{\sigma_u}^2}{n{\sigma_x}^2}$ $SE(\widehat{\beta_1}) = \frac{1}{\sqrt{n}} \times \frac{\sqrt{Var(u_i)}}{Var(X_i)}$