<u>Random Variable</u>

- a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes
- a variable that doesn't have a value until some experiment is performed

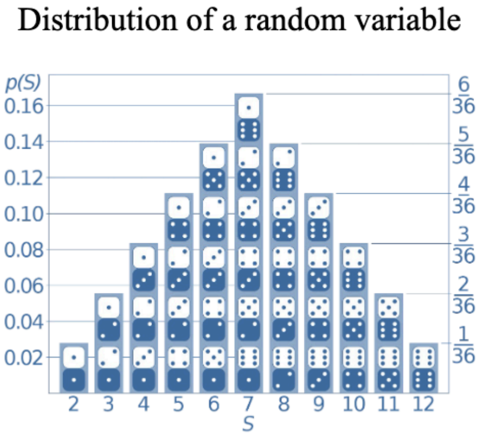
<u>Sample Space (outcome space)</u>

- set of all possible outcomes
- are mutually exclusive (no overlaps)
- are collectively exhaustive (the sum of all groups covers all possible options)

<u>Probability</u>

- relative frequency of an outcome of a set of outcomes
- probabilities of all possible outcomes add up to 1

<u>Distribution of a random variable</u>

- description of probabilities of all possible outcomes for the random variable
- example, the distribution of the sum of two dice
- 

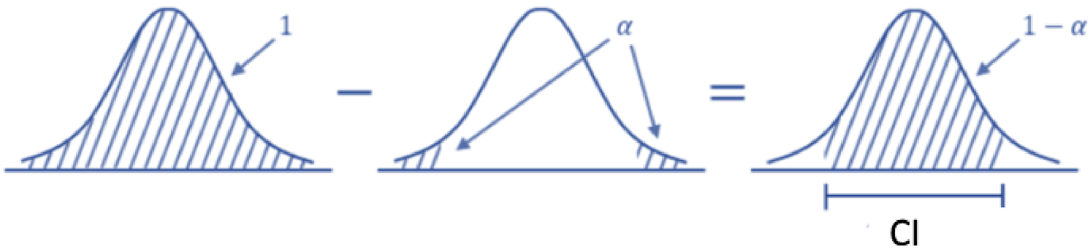
<u>Statistical Inference</u>

- the theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically based on random sampling.

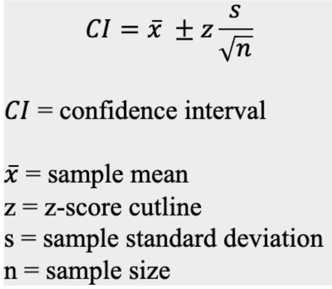
- <u>Population</u>
    - group or collection of all possible entities of interests
    - typically thought of as being infinitely large
    - population parameters(characteristics of a population)
        - mean
        - probability
        - median
        - min, max

Population	Sample
entire set of items	subset of the population
characteristics are called parameters	characteristics are called statistics
parameters are generally unknown	statistic values are known based on the sample collected
parameters are fixed	statistics can vary depending on the sample

- <u>Confidence Interval</u>
    - range of values that is likely to include a population value with a certain degree of confidence
    - 95% confidence interval for the population mean is a range of values that contains the population (true value) mean in 95% of repeated samples.
    - probability that a population value does include within the confidence interval = ɑ%
    - the probability that the population value is included within a certain degree of confidence (100 - ɑ)%
    - each side of the curve is ɑ/2%



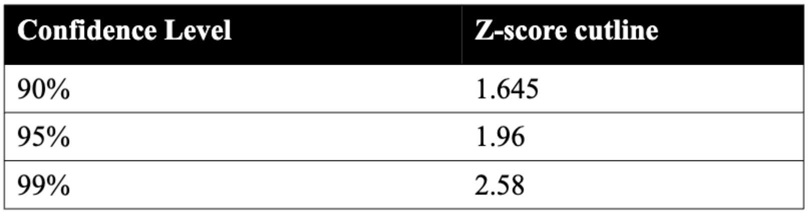
    -
    - the narrower the interval(upper and lower values), the more precise the estimate but the lower the confidence level
    - as sample size increases,
        - the confidence interval should become narrower
        - tighter distribution of sample means
        - decrease in standard error, sample mean is closer to the population mean
    - 1. compute the mean and standard error of the sample
    - 2. calculate the upper and low score for the confidence interval using the z-score cutline for the chosen confidence level
        - z-score cutline means cutoff or critical values of standard normal distribution
        - it is the ɑ/2-quantile of the standard normal distribution

        - 

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean
z = z-score cutline
s = sample standard deviation
n = sample size

        - 

Confidence Level	Z-score cutline
90%	1.645
95%	1.96
99%	2.58

- Central Limit Theorem
    - states that the distribution of sample means approximates a Normal distribution as the sample size gets larger, regardless of the population's distribution
    - sample size >= 30 is considered sufficient for CLT to hold and the CLT tends to provide more reliable and predictable results
    - Law of Large Numbers
        - as the sample size increases, the distribution of sample means will become tighter (sample mean = true mean)
        - fluctuations in sample means tend to cancel out as the sample size gets bigger
        - this justifies the practice of using sample statistics to estimate population parameters

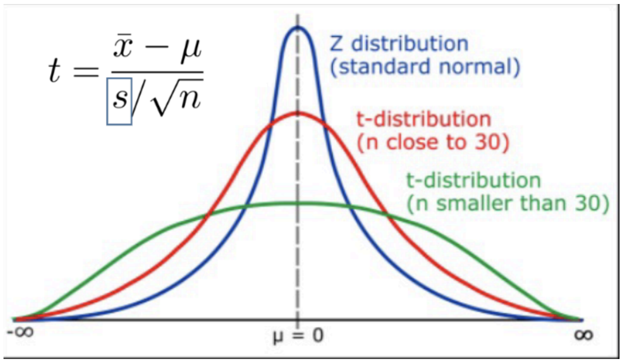
Hypothesis

- assumption about a population parameter
- examine random samples from a population due to constraints of examining the entire population(time, money, resources)
- there is no predetermined outcome, it has to be an idea that can be supported or rejected through carefully crafted experimentation or observation
- Alternative Hypothesis (H1/Ha)
    - what you hope to prove true
    - states that a population parameter is smaller, greater, or different than the hypothesized value in the null hypothesis
- Null Hypothesis(H0)
    - the statement being tested, by rejecting it, you conclude that the alternative hypothesis is true
    - statement of "no effect" or "no difference"
    - states that a population parameter is equal to a hypothesized values

- example, average patient satisfaction scores for hospital use AI are greater than the US hospital satisfaction average(65/100)
    - H1: mean > 65
    - H0: mean <= 65
- 1. State the hypothesis
- 2. Set the significance level
    - type 1 error, α: the probability that a hypothesis test leads to rejection of the null hypothesis when Null is true
    - α: the probability of a type 1 error and the rejection region for the null hypothesis
    - significance level cannot be zero as it is not reasonable as there is always a chance that we make an error
    - when null is true, the distribution follows a normal
    - significance level = probability of type 1 error
    - high significance level -> low validity
    - low significance level -> high validity
    - example, significance level of 5%, null hypothesis is true when we reject, will not happen more than 5% of the time
- 3. Compute the test statistic
    - one-tailed test
        - left-tail test
            - Ha: mean < value
            - H0: mean >= value
        - right-tail test
            - Ha: mean > value
            - H0: mean <= value
        - to detect an effect, when the hypothesis is about the direction of an effect
        - example, to test if the new drug is an improvement over an existing drug, but will fail to test that the new drug is less effective than the existing one
    - two-tailed test
        - Ha: mean ≠ value
        - H0: mean = value
        - to determine if there is any difference between the groups that you are comparing
        - example, to see if A scored higher or lower than B
    - Using the cutline approach (z-score)

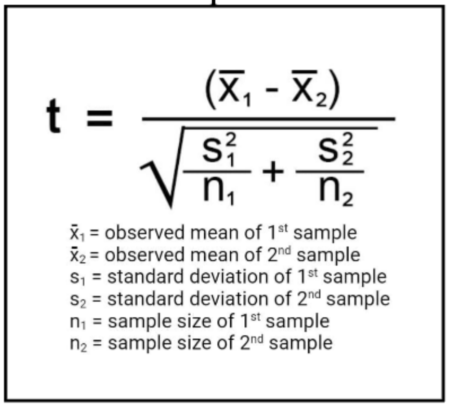
$$Z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$$

        -
        - looks at a sample size of more than 30, uses population standard deviation
    - Using region approach (p-value)
        - p-value: if the null hypothesis were true, what is the probability of observing values at least as extreme as actually observed

    - p-value: the probability of observing a test statistic just as or more extreme than one we obtained, assuming null hypothesis H0 is true.
    - p-value is calculated from Z value
    - if p-value > significance level, cannot reject the null hypothesis, the observed difference is not statistically significant, there is not enough evidence to reject the null hypothesis.
    - if p-value < significance level, it suggests that the observed data is unlikely to have occurred under the assumption that the null hypothesis is true, enough evidence to reject the null, the observed difference is statistically significant
    - the lower the p-value, the less consistent the null hypothesis is with the observed statistic.
    - example, if p-value is 0.007, we would only witness the null hypothesis or more extreme only 0.7% of the time
  - T-test
    - used for limited sample sizes
    - 

      $$t = \frac{\bar{x} - \mu}{\boxed{s}/\sqrt{n}}$$

    - often used to determine whether two groups are different from one another
    - only used to compare the means of 2 groups
    - Two-sample t-test

      $$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

      $\bar{x}_1$ = observed mean of 1st sample
      $\bar{x}_2$ = observed mean of 2nd sample
      s_1 = standard deviation of 1st sample
      s_2 = standard deviation of 2nd sample
      n_1 = sample size of 1st sample
      n_2 = sample size of 2nd sample

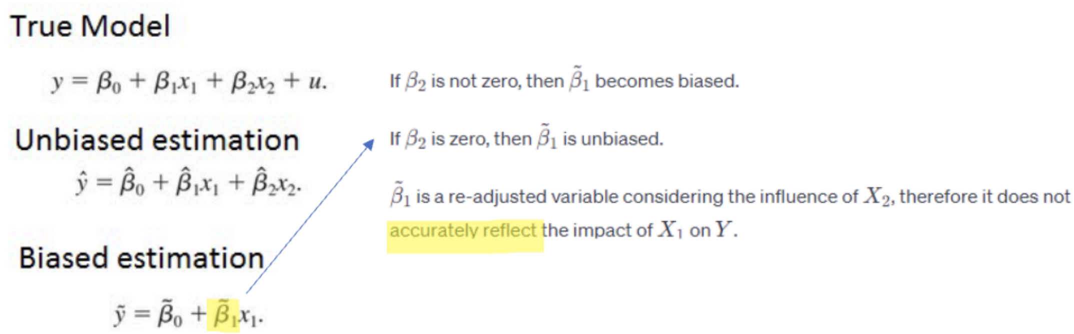
<u>Causal Inference</u>

- X causes Y, the change of value of X, without changing anything else, will lead to a change in Y.

- does not mean that X is the only factor that causes Y.
- Correlation does not equal causation
  - correlation indicated a relationship or a patter between two variables while causation establishes that one variable actually causes the change in the other.
- if it is a correlation, we cannot measure the impact and cannot predict the future.
- if it is a causation, we can measure the impact of C and we can predict the future, we can make better decisions
- correlation only tells us that causation MAY exist

- <u>Treatment Effects</u>
  - Y(0): the potential outcome if individual does not receive treatment
  - Y(1): the potential outcome if individual receives treatment
  - Broadest possible average effect (ATE)
    - measures the average individual treatment effects over the whole population
    - E(Y1) - E(Y0)
  - it is not constant across people
  - is not independent of potential outcomes
  - can overcome with randomization

- <u>Endogeneity</u>
  - arise when the focal variables are endogenously determined within a system of interest.
  - leads to a spurious causation
  - in actual there is no causal relation but yet we observe it
  - <u>Omitted variable bias</u>
    - occurs when a statistical model fails to include one or more relevant variables
    - X2 variable must be a determinant of the dependent variable Y and must be correlated to X1



True Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

If β_2 is not zero, then $\tilde{\beta}_1$ becomes biased.

Unbiased estimation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

If β_2 is zero, then $\tilde{\beta}_1$ is unbiased.

$\tilde{\beta}_1$ is a re-adjusted variable considering the influence of X_2, therefore it does not accurately reflect the impact of X_1 on Y.

Biased estimation

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1.$$

    -
    - also referred to as a confounder
    - sample selection is a sub-form of the omitted-variable bias as the selection process represents an excluded variable that manifests in the error term and correlates with the endogenous choice construct and the outcome variable

    - example, when measuring the effect of AI robot adoption on consumer satisfaction, if we only select larger hospitals, it could lead to omitted variable bias as larger hospitals tend to have more well-organised services, increasing consumer satisfaction and they are more likely to adopt AI since they have more capital.
    - we can add control variables to the model
        - multiple linear regression
    - we can do randomised experiments by ensuring that the treatment and control groups are similar with respect to both observed and unobserved factors
    - create pairs of treated and control participants with similar pre-treatment characteristics
        - Nearest-nieghbour matching, Mahalanobis distance matching, optimal matching
        - propensity score matching aims to balance the propensity score between treatment and control groups by accounting for various factors such as demographic and socioeconomic factors
        - propensity scores are probabilities that reflect the likelihood of an individual or unit being assigned to the treatment group. Usually calculated using logistic regression or other statistical methods.

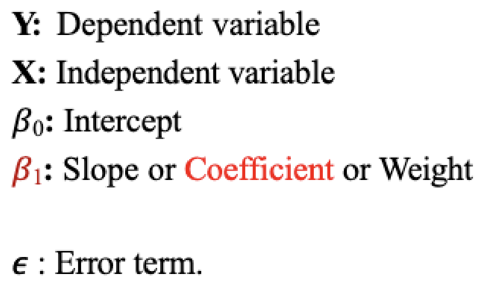
- <u>reverse causality</u>
    - the cause and effect are reversed. the cause is said to be the effect and vice versa.
    - X and Y are associated, instead of X causing Y, it is the other way around
    - simultaneity bias occurs when in time series analysis, two variables affect each other at the same time. It is common where certain variables are interdependent.
    - through randomized experiments
        - participants are exposed to the treatment or control conditions at the same time. it helps to ensure that any effects observed are due to the treatment itself, rather than some other variable causing both the treatment and the outcome
        - it establishes a clear temporal order between the cause and the effect, ruling out the possibility that the outcome is causing the treatment
    - having a clear shock
        - example, earthquakes can increase in medication consumption but medical consumption cannot lead to an earthquake
    - instrument variables
        - variable Z that is correlated with the independent variable X and uncorrelated with the error term in the linear equation
    - use lagged values

            - explanatory variables as predictors to help establish the direction of causality, especially in time series analysis.
            - assumes that past values influence the present but not vice versa
            - instead of looking at both variables at the same time, look at how past values of one variable affect the current value of the other variable, to establish a time-ordered sequence of events
            - in the regression model, the regressors are lagged by one period of time

    - errors-in-variables bias
        - occurs when the independent variables in the model are measured with error, estimation based on the standard assumption leads to inconsistent estimates, meaning that the parameter estimates do not tend to the true values even in very large samples.
        - arise fom using incorrectly measured variables or proxy variables in regression models.
        - correcting it requires additional information or assumptions about the measurement error.
        - survey and data collection process
            - data entry errors should be carefully discussed before the analyses
            - researchers shold consider ambiguity
            - false answers and omitted values should be considered in a careful manner
            - use well-developed, theory driven questions and measures
        - use observational data
            - compare the effect size before and after adding the value that is omitted
            - use zero value, exclude them, or use an average value
            - develop math model to correct the errors(such as instrument variable)

Linear Regression

- Use a line scatter plot to visually summarise the relationship
- to have an equation for the relationship between 2 variables
- outcome variable (dependent variable) affected by the explanatory variable(independent variable/regressor), assuming all other variables remain constant
- it is not possible to have all of the plots on the line as in most cases, the dependent variable is not only explained by the chosen independent variable
- it allows us to estimate and make inferences about effect on T of a unit change in X
- $y = \beta_0 + \beta_1 \times x + \epsilon$



Y: Dependent variable

X: Independent variable

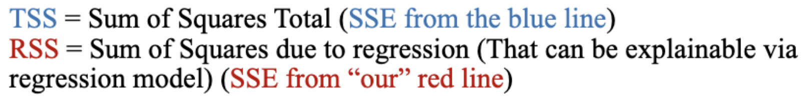
β_0**:** Intercept

β_1**:** Slope or Coefficient or Weight

ϵ : Error term.

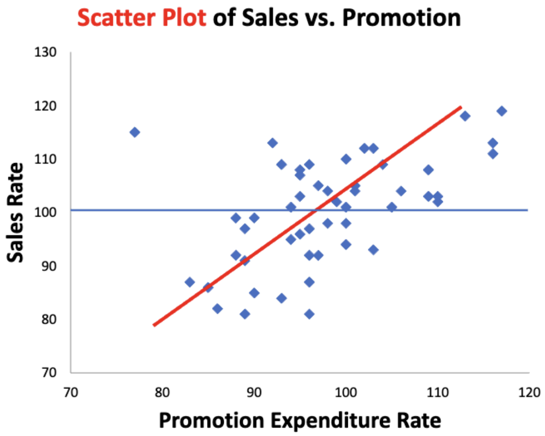
-
    - error term: there are variables that have an impact on the dependent variable
    - the coefficient: it is a quantitative measure of the relationship between the dependent and independent variable, that is the impact of the independent variable, it measures the how the change in the independent variable results in the change in the dependent variable
- find the best line with the lowest error, which minimizes the sum of squared vertical differences between the points and the line, this is the sum of squares error
    - the smaller the sum of squared differences, the better the fit of the line to the data
- Ordinary least squares(OLS) finds the coefficient term that minimises the errors between the actual values and the predicted values based on the model
    - it is intuitive and easier to calculate than other estimate methods such as least absolute deviation estimation, which minimizes the sum of absolute errors.
- coefficient significance test
    - check the p-value of the coefficient term
    - if it is less than the significant level, we can reject the null and accept the alternate hypothesis and the coefficient term is statistically significant
- model fit using r-squared
    - statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable in a regression model

$$R^2 = 1 - \frac{RSS}{TSS}$$



TSS = Sum of Squares Total (SSE from the blue line)
RSS = Sum of Squares due to regression (That can be explainable via regression model) (SSE from “our” red line)

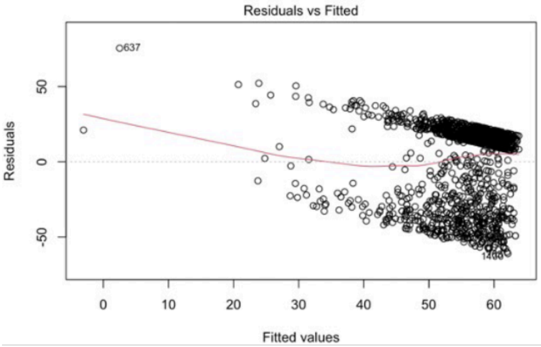
    -



    -
    - if RSS is close to zero, r-sqaured goes to zero, which is what we want
    - r-sqaured close to 1, indicates a very high explanatory power of the model
    - r-sqaured close to 0, indicates a very low explanatory power fo the model
    - it is not a an indicator of whether a relationship is causal
    - if its > 0.5, it is usually a good model fit, depending on the context
    - when more variables are added, r-squared values typically increase.
- adjusted r-squared
    - compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance.
- even with low r-squared, the model is still acceptable as it can help us understand the relation between X and Y.
- it is only problematic when we want to produce predictions that are reasonably accurate and precise.
- Underlying Assumptions for OLS
    - a good estimator is efficient, unbiased and consistent
        - efficient(minimum variance): among all linear unbiased estimators, it has the lowest variance
        - unbiased(correctness): its expected value is equal to the true value of the parameter being estimated. the estimator accurately reflects the parameter without systematic error
        - consistent(asymptotical convergence): if, as the sample size increases indefinitely, it converges in probability towards the true value of the parameter. with a large sample, the estimator gets closer and closer to the actual parameter value.

- <u>Gauss-Markov Assumptions</u>
    - 1. linear relationship
        - variables estimating using the OLS method must be linear
        - we can log the variable to stabilise the variance and normalise the distribution of the variables especially if the variables are skewed.
        - it retains the relationships within the data while reducing the range

- 2. no perfect multicollinearity
    - perfect multicollinearity is when two or more independent variables have an exact linear relationship
    - if there is, there is no unique solution, and we cannot measure the impact of the independent variables on the dependent variable as we cannot identify that the effect is due to the specific independent variable
    - use Variance Inflation Factor to test
        - less than 10 generally indicated an absence of multicollinearity
    - plotting data and calculating correlation can also provide an assessment
        - if they are highly correlated, there may be multicollinearity
        - if correlation is > 0.9, use the variable that is more representative
- 3. Homoscedasticity
    - guaranteeing equal variance of residuals
    - a situation in which the error term is the same across all values of the independent variables.
    - heteroscedasticity: variability of the independent variable is unequal across the range of values of a dependent variable.
    - residual-fitted plot
        - residuals vs predicted plot
        - residuals are randomly scattered around the horizontal axis
        - heteroscedasticity exist when the residual plot has a trumpet/funnel shape
    - robust standard errors
        - also known as heteroscedasticity-consistent standard errors are adjusted to account for the possibility of heteroscedasticity.
        - provide more reliable standard error estimates, which in turn lead to more accurate confidence intervals and hypothesis tests about the regression coefficients
        - it ensures that the inference remains valid even in the presence of scedasticity
    - use Goldfeld-Quandt test
    - use Breush-Pagan test
    - use likelihood ratio test
- 4. No autocorrelation
    - there is no serial correlation between the error terms
    - degree of correlation of a variable’s value over time
    - use residual fitted plot

            - 

Autocorrelation between the residuals and fitted value ŷ or implicitly some of the independent variables X's as well.

        - use Durbin-Watson test
            - if p is statistically significant, then there is autocorrelation
            - remove autocorrelation by using estimated p by differencing
            - That is $Y_i - \rho Y_{i-1} = \beta_0(1-\rho) + \beta_1(X_i - \rho X_{i-1}) + \nu_i$.
        - common to test for it in time-series data
        - is present in cross-sectional data as neighboring units tend to be similar with respect to the characteristics under study
        - use cluster-adjust standard errors
        - use Prais-Winsten Estimation Procedure
    - 5. Exogeneity
        - endogeneity leads to spurious correlation, raising concerns about causality

Multiple Linear Regression

- $$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 X_3} + \cdots + \boldsymbol{\beta_k X_k} + \varepsilon_k$$
- all the coefficient terms are values that minimise the sum fo squared residuals
    - example, B1 = 0.07, keeping all other variables constant, the increase in one unit of x1 will lead to an increase in 0.07 units of Y.
- intercept: the value of Y if all regressors are equals to 0.
- r-sqaured
    - a measure of how much variance in the Y is explained by the model
    - more regressrs will lead to increase in r-sqaured, which means the model is more explainable but not all regressors are meaningful
- adjusted R-sqaured
    - a measure of how much variance in Y is explained by the model after controlling for the numbr of regressors

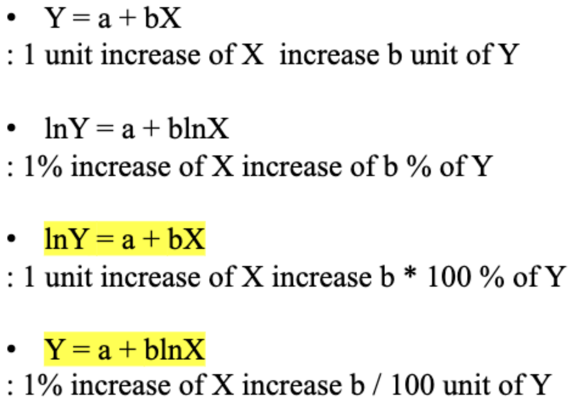
Dummy Variables

- used in regression analysis to represent categorical variables that have more than two levels
- allow is to include categorical variables in our analysis
- can help to control for confounding factors and improve the validity of our results

- can be used as independent or dependent variable
- in regression models that use dummy variables, the baseline group, the category against which all other categories are compare, should be omitted from the regression equation to void the situation where the independent variables are multicollinear
- example, when examining the impact of an AI robot on customer satisfaction by dividing customer group inot three random groups
    - B0 is the average effect of the no-interaction group on customer satisfaction
    - B1 is the coefficient for the AI group, which represents the change in customer satisfaction when comparing the AI group to the No-interaction group
    - B2 is the coefficient foe the human group, which represents the change in customer satisfaction when comparing the AI group to the No-interaction group.
- seasonal dummy variables
    - one is excluded to act as the baseline, where all other dummies refer to differences between themselves and the reference month
    - for example, 12 months, only include 11 dummy variables

Log transformation

- transformed data follows a normal or near-normal distribution
- transform skewed data to increase the validity of the associated statistical analyses.
- can reduce data variability in datasets that include outlying observatons



- Y = a + bX

: 1 unit increase of X increase b unit of Y

- lnY = a + blnX

: 1% increase of X increase of b % of Y

- lnY = a + bX

: 1 unit increase of X increase b * 100 % of Y

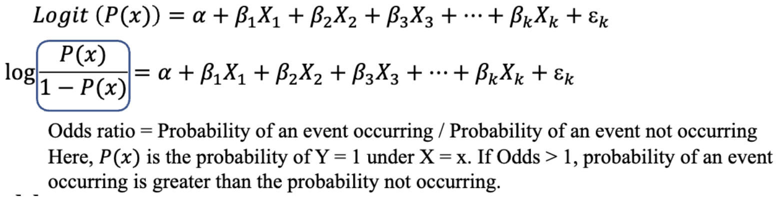
- Y = a + blnX

: 1% increase of X increase b / 100 unit of Y

-
- make trends more apparent and interpretable
- simplifies interpretation and makes it more practical and convenient

Logit Model

- used for binary output variable



$$Logit\ (P(x)) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \cdots + \beta_kX_k + \varepsilon_k$$

$$\log\boxed{\frac{P(x)}{1-P(x)}} = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \cdots + \beta_kX_k + \varepsilon_k$$

Odds ratio = Probability of an event occurring / Probability of an event not occurring
Here, $P(x)$ is the probability of Y = 1 under X = x. If Odds > 1, probability of an event occurring is greater than the probability not occurring.

-
- fibnds coefffiecients that minimises logit loss

- other dependent variables remain fixed, as the value of X increases by one unit, the odds of Y increase by exp(B1).
- failure to converge: it is possible foe the expected likelihood estimation process that learns the coefficients to fail to converge, when there are many highly correlated inputs in the data or the data is very sparse(high percentage of missing data or data set to 0)
- adding fixed effects or interaction terms with logit is trickier

Interaction term

- variable in the model that captures the effect of 2 or more variables acting together that is different from the effect of each of the variables on their own.
- it is to understand how the change in one independent variable modifies the influence of another independent variable on the dependent variable Y.

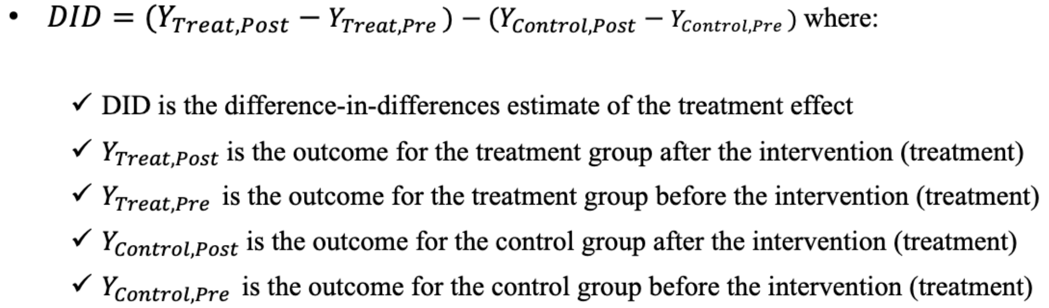
Panel Data

- observe the same individual(or state, county, or business) over multiple periods
- allows for the analysis of dynamics that cannot be observed in a single cross-sectional snapshot.
- enables the control for individual heterogeneity that is not observed or cannot be measured, especially importan when these unobserved factors could be correlated with other explanatory variables, poteitonally, biasingthe results in a cross-sectional analysis
- example, if cross-sectional data has 11 states in each panel and 3 time periods, there is 11*3 observations in the panel set
- Fixed Effect
    - there exist within sample time variation and between sample variation
        - between variation is looking at the relationship between the means of each county, things that are fixed overtimea
        - within variation looks at the variation within the county from year to year
    - time-invariant variables
        - variables with between variation only
        - such as geography and landmarks
    - time-variant variables
        - variables with both between and within variation
        - such as police budgets and poverty level
    - controls time-invariant variables
    - time-variant controls should be added to the model
    - time-invariant terms such as individual characteristics are not in the model but in the error term
    - to remove between variation via de-meaning
        - 1. for each variable, gett he mean value of the variable for each individual
        - 2. subtriact out that mean to get residual, (x - xmean), (y- y mean)
        - 3. use the residuals as the independent and dependent variables
    - dummy to remove between variation
        - treat individual as a categorical variable and add it as a control.
    - as adding all time-invariant controls is impossible, we use Fixed Effect(FE)

  - time-invariant characteristics are controlled
  - for example, gender, education, living area,
  - time-varaint characteristics like budget, mental status should still be included in the model
  - it does not reduce autocorrelation concerns
  - adding location to FE does not remove year-to-year autocorrelation
  - adding year does not remove county-to-county autocorrelation

Difference-in-Differences

- refers to an empirical strategy or technique or approach used to estimate the causal effects of a treatment of invervention in observational studies
- cna be used when there is a clear shock or natural treatment
- it is a quasi-experimental approach, it does not guarantee randomisation
- the true treatment effect is the difference in the change in outcomes between the treatment group and the control group that is directly attributable to the intervention or treatment being studied
- changes not derived from the intervention should be removed



- $DID = (Y_{Treat,Post} - Y_{Treat,Pre}) - (Y_{Control,Post} - Y_{Control,Pre})$ where:

✓ DID is the difference-in-differences estimate of the treatment effect

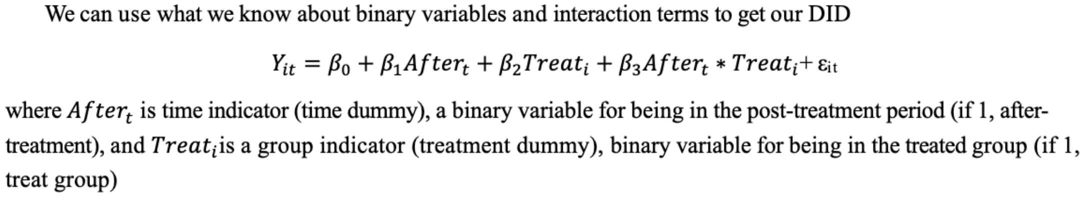
✓ $Y_{Treat,Post}$ is the outcome for the treatment group after the intervention (treatment)

✓ $Y_{Treat,Pre}$ is the outcome for the treatment group before the intervention (treatment)

✓ $Y_{Control,Post}$ is the outcome for the control group after the intervention (treatment)

✓ $Y_{Control,Pre}$ is the outcome for the control group before the intervention (treatment)

-
- the model compared the change in outcomes for the treatment group between the pre and post-intervention periods with the change in outcomes for the control group over the same time period
- difference between these two changes represents the estimated treatment effect



We can use what we know about binary variables and interaction terms to get our DID

$$Y_{it} = \beta_0 + \beta_1 After_t + \beta_2 Treat_i + \beta_3 After_t * Treat_i + \varepsilon_{it}$$

where $After_t$ is time indicator (time dummy), a binary variable for being in the post-treatment period (if 1, after-treatment), and $Treat_i$ is a group indicator (treatment dummy), binary variable for being in the treated group (if 1, treat group)

-
  - B0: the intercept term, which is the predicted value of Y when After and Treat are both 0. Control group mean in before treatment
  - B1: average change in the outcome variable in the control group in the pre and post-treatment period
  - B2: average change in the outcome variable between the treatment and the control group in the pre-treatment period
  - B3: the average difference in the change in the outcome variable between the treatment and control groups from the pre-treatment period to the post-treatment period

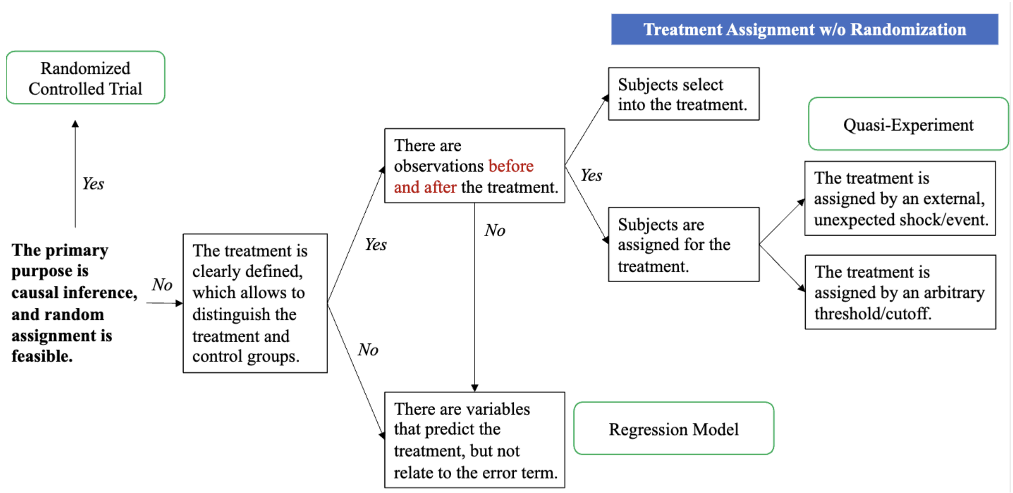
- can add individual fixed effects to control individual heterogeneity
- if so, all time-invariant individual characteristics will be controlled via individual dummies
- controls include time-variant variables that could affect Y.
- if B3 is statistically significant, there is a statistically significant treatment effect on the outcome variable, keeping other variables unchanged.
- after the treatment, there will be a B3 unit increase or decrease in the outcome variable when all other variables are the same
- Parallel Trends
    - this is the most critical assumption of DiD
    - it suggests that the outcome variable would have to follow a similar trajectory for both groups if the treatment had not been introduced
    - it allows us to attribute any difference in outcomes between the treatment and control groups to the treatment itself, rather than other factors that may have affected the outcome over time
    - compare the graphical trends of outcome variables before treatment
        - if the trends are parallel, then the assumption holds, even if the levels differ.
- As-good-as-random treatment
    - in the absence of actual randomization, the treatment and control groups should be similar in all aspects except for the treatment itself
    - compare pre treatment characteristics between the treatment and control groups to ensure that they are not significantly different.
    - compare the mean difference of outcome variables between treatment and control during the pre-treatment period
    - if the t-test shows a significant difference between the treatment and control groups, we can not argue that our data satisfies the parallel trends
- Matching between treatment and control
    - to construct the control group that is as possible to the treatment group on the observed characteristics
        - Coarsened exact matching: matching for all variables
        - propensity score matching: reducing various variables into one dimension,
            - allows us to have control and treat groups to have similar determinants before treatment.
            - create a statistical equivalence between the treat and control groups, we identify whether the control and treat groups were similar in demographics, consumer characteristics, and spending history
            - it does not eliminate selection bias as it only adjusts for observed and measured confounders. Any unknown remains a source of bias
            - effectiveness is dependent on the quality of matching, if there are no appropriate matches, it can lead to biased results

        - during matching, some data are discarded and this can lead to a loss of valuable information and reduce sample size, affecting the study's power
        - there needs to be a sufficient overlap in the distribution of propensity scores between the treatment and control groups. Without the overlap, it cannot effectively adjust for differences.
- 1. Identify a treatment or intervention that can be measured before and after the treatment
- 2. Identify the treatment and control groups
- 3. check for parallel trend assumption or as-good-as-random assumption
- 4. conduct matching if necessary
- 5. structure the model, consider the fixed effect and the time-variant variables within the model
- 6. analyse the data
- if we include a fixed effect, the treatment gorup indicator will be omitted automatically as fixed effects can cause perfect multicollinearity. introducing fixed effect, we are giving each unit(person/business etc) different intercept. the treatment group indicator would br included in the fixed effect as individual variation already covers the variation brought by the individual being in the treatment or not. it will also net out all the fixed characteristics' impact and the model will not need to predict the effect brought by treatment group indicator and will need to omit it as adding redundant information will cause perfect multicollinearity

Experiments

- an experiment constitutes a methodologically structured procedure designed to either corroborate or invalidate a specific hypothesis, or to ascertain the effectiveness or probability of phenomena that have not been previously tested
- experiments furnish empirical evidence that elucidates the causal relationships governing the observed outcomes, thereby advancing out understanding of the underlying principles
- the baseline group is the control group and the effect should be calculated by comparing it with the control group.
- Gold standard of causal inference, Random Assignment
    - ensures that the subjects with various characteristics are distributed evenly across the treatment and control groups so that they are comparable.
    - assumes that the randomisation of observables can guarantee the randomisation of unobservables
    - the probability of any one person or group being assigned to a particular treatment group is equal o the likelihood of any other person or group being assigned to that group. P(T) = P(C)
    - helps to eliminate influence of confounding variables and simultaneous bias on the treatment effect.
    - is done by literally assigning them randomly,

  - predetermining characteristics such as gender ratio in the treatment and control group is known as stratified sampling
  - randomisation check (t-test)
    - process of verifying whether the random assignment of samples to the different groups was properly implemented
    - it is to reduce the risk of bias and ensure that differences between the groups are due to the intervention and not to other factots\
  - endogeneity issues: multiple or repeated experiments
    - impact the outcome by causing lingering effects of a previous treatment from carrying over inot the next treatment phase
    - a washout period help us prevent any lingering effects of a previous treatment
    - washout period ensures that any adverse events associated with a previous treatment have resolved before the next treatment phase begins
    - during the washout period, the impact of the initial exposure may diminish as the memory of the experience fades. it allows the users' behaviours to return to their baseline levels, ensuring that the second exposure is assessed on a more neutral basis.
  - endogeneity issues: treatment that requires multiple actions
    - can lead to biases
    - one treatment shoudl induce only one response change, if not it is necessary to determine how the first response affects the second response.
    - heckman correction is a statistical technique to correct bias from non-randomly selected samples
- Random Assignment is not always feasible
  - quasi-experiments mimic randomised experiment by assigning the treatment and control groups based on an exogenous shock, self-selection or an arbitrary cutoff, instead of random assignment
  - it is to seek the comparable control group and infer the counterfactual from the control group.
  - how similar is the control group to the treatment group in the absence of treatment?
  - the level of difficulty to prove ceteris paribus increases from assigning randomly, assigning by exogenous shock, assigning by themselves. self selection is harder to prove than exogenous shock.


Treatment Assignment w/o Randomization
Randomized Controlled Trial
Yes
The primary purpose is causal inference, and random assignment is feasible.
No
The treatment is clearly defined, which allows to distinguish the treatment and control groups.
Yes
No
There are observations before and after the treatment.
No
There are variables that predict the treatment, but not relate to the error term.
Regression Model
Yes
Subjects select into the treatment.
Subjects are assigned for the treatment.
Quasi-Experiment
The treatment is assigned by an external, unexpected shock/event.
The treatment is assigned by an arbitrary threshold/cutoff.

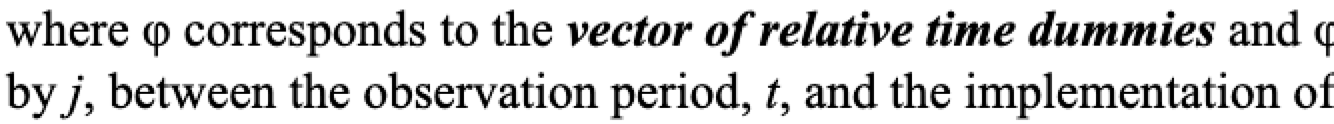
- Validity of Shock
    - random implementation test
        - ensure robustness of the shock
        - if the treatment is a valid shock, we may not observe significance from pseudo-treatment, where shock is randomly created without any meaning
        - helps to assess the spuriousness of any significant results obtained in the main analyses arising from autocorrelation in the dependent variable.
        - random shuffle test
            - 1. reassign treatment indicators at random in the data, thereby creating a pseudo treatment
            - 2. estimate a standard difference-in-diffrence model ad store the coefficient of this pseudo treatment
            - 3. repeat for 1000 times
            - 4. exmaine the pseudo presence dummy.
                - check its significance and size: if the pseudo presence dummy is not significant and nearly zero, it shows the validity of real genuine shock
                - compare the genuine coefficient with the pseudo presence dummy: if they are totally different, it shows that the shock is valid
            - concern that the possibility of the observed effects arised coincidentally
                - randomly assign the treatment statis then estimate DiD and repeat for 1000 times, results are insignificant, the shock is valid.

Relative Time Model

- extended version of DiD
- it can determine if a significant difference between the treatment and control group before the treatment exists, in order to determine if the untreated is an acceptable control group. if there is a significant difference, it violates the assumption of the model.

- the model can also examine how the impact of a shock evolves over time, thereby understanding its long-term effect.

$$Y_{ct} = \sum_{j} \beta'_{j} \ (\varphi_{ct}) + \lambda'_{c} + \varepsilon_{ct},$$



where φ corresponds to the ***vector of relative time dummies*** and φ
by *j*, between the observation period, *t*, and the implementation of

-
- relative time dummies create a series of indicators that reflects the relative chronological distance between time and when the treatment was implemented
- (rel+1) signifies the estimated effect of the treatment one time period after it is administered and (rel-1) signifies the estimated effect of the treatment one period before it is administered.
- relative time before the treatment are expected to be non-significant
- 1. create the relative time indicator which shoes the distance between time and the treatment period
- 2. make relative time model indicator as dummy and create relative time dummy by interacting treat
- 3. conduct analysis
    - due to multicollinearity, it is essential to omit a baseline category when including dummies and rel-1 is commonly used
    - coefficients for the pre-treatment time dummies are statistically not significant, indicating that the control and the treatment groups were statistically comaprable in the absence of the treatment, satisfying parallel trend assumption
    - effect of treatment more potent on day 0 and diminishes the next day suggests that there is the presence of novelty effect, which is the effect of introducing a new element.