TA:MSC B315 (B in the front of the number indicates basement of MSC) 1:15pm-2:45pm, Mondays (Central Time); 1:15pm-2:45pm Fridays

54. personalized rank

> based on user's history of favorite articles and other user-specific attributes
> ###build the engine (use `personalized PageRank algorithm`) I think the key point lies on the weight of edge..
>
> - the time the user spend on browsing a article; (to set weight for each edge)
> - apple news has a feature called *shared with you*, if there is any share between two person, we will assign an edge between them.(in personal rank algorithm, the is no difference between nodes that represents user and item), for this kind of edge, we can also use time spent on reading the article shared by a certain person as weight.
>   ###collect data?
> - once user open an article, automatically set a mark as start to track time.
> - we will have access to get the data of shared with you for a certain user
>   ###compare
> - randomly select a group of users, track the average time a user spend on Apple News a week(track for 1 months and get a mean) before and after the application of new engine and do two sample t-test
> - the number of clicks of this app, do the same as above.

55. video latency(x) and user's watch time(y)

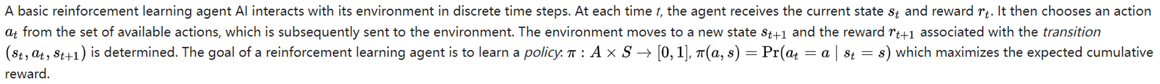
> the analysis just use a simple linear regression
>
> ## extra analysis you suggest doing
>
> - find if there is any `confounding variables`, if so, add the potential confounders in your regression model to control the effect of confounder to see the significance of coefficient

> - some confounding variables.. 想不到 要寄了
>
> ## recommendations for backend engineers
>
> -

56. law of large number and central limit theorem

> ## law of large number
>
> - when performing the same experiment a large number of times, the average of the results obtained from a large number of trials tends to become closer to the expected value as more trials are performed.(means the random variables should be 独立同分布)
> - strong version(converge almost surely) and weak version(converge in probability)
>
> ## CLT
>
> - For large enough n, the distribution of $\bar{X}_n$ gets arbitrarily close to the normal distribution with mean μ and variance σ^2/n (needs iid)

57. reinforcement learning

> ## definition
>
> 
>
> A basic reinforcement learning agent AI interacts with its environment in discrete time steps. At each time *t*, the agent receives the current state s_t and reward r_t. It then chooses an action a_t from the set of available actions, which is subsequently sent to the environment. The environment moves to a new state s_{t+1} and the reward r_{t+1} associated with the *transition* (s_t, a_t, s_{t+1}) is determined. The goal of a reinforcement learning agent is to learn a *policy*: $\pi : A \times S \rightarrow [0,1]$, $\pi(a,s) = \Pr(a_t = a \mid s_t = s)$ which maximizes the expected cumulative reward.
>
> ## (agent, environment, policy, signal, value function(long run reward from this state) and model)

> At each time step, the agent implements a mapping from states to probabilities of selecting each possible action. The map is called policy. Reinforcement learning methods specify how the agent changes its policy as a result of its experience
> what is agent ai? An agent is anything that can be viewed as : perceiving its environment through **sensors** and acting upon that environment through **actuators**

> ### a typical algorithm
>
> ### difference with supervised learning, semi-supervised learning and unsupervised learning
>
> supervised learning is based on a training set of labeled examples provided by a knowledgeable external supervisor, not adequate for learning from interaction;(to predict the output from input data)
> unsupervised learning is to find the hidden pattern/structure in the collections of unlabeled data, reinforcement learning is trying to maximum the reward signal instead of finding a hidden structure.
> semi-supervised models use both labeled and unlabeled data for training.

50. PCA and ICA(separate information)

> three key points of ICA:
> - The number of inputs equals number of outputs; (PCA will compress information, dimension deduction)
> - Assumes independent components are statistically independent(Two variates A and *B are statistically independent iff the conditional probability P(A|B) of A given B satisfies P(A|B)=P(A)*);
> - Assumes independent components are non-Gaussian;
> - about PCA, take the first k largest eigenvalues, and the related eigenvectors where k is less than the dimension of data we have, in this way, we can do data decomposition

为了解决最小二乘的问题，大矩阵求逆（数据点的个数比要求的参数的个数少，不能求逆之类的）

49. QR decomposition and SVD; implementation; strength; diff/similarity

> QR, a decomposition of matrix(real square matrix) into one orthogonal matrix times a upper triangle matrix; computing it by using Gram-Schmidt process; ease of implementation, numerically unstable(why?)
> SVD is a generalization of eigendecompostion, decompose it to two orthogonal matrix and a diagonal matrix, the elements in diagonal matrix is is the singular value of A(the matrix being decomposed) , the sqrt relationship between singular value and eigen value.. SVD has a

> more efficient numerical computing method.. based on QR algorithm, 迭代算法
>
> ## what are they useful for?
>
> - Inverse; for A to be a singular matrix, we can still compute its pseudoinverse
> - solving system of linear equations$Ax = b$ (好像也是求逆)
> - matrix similarity, find a matrix that is the best rank-k to A(set all but the largest k singular values to zero)
>
> ## similarity and difference
>
> - those decompositions are not unique
> - https://math.stackexchange.com/questions/2348807/is-there-any-connection-between-qr-and-svd-of-a-matrix ?

50. clustering algorithms and their strength and weakness(是指具体的算法吗)

    > clustering: group unlabeled examples
    >
    > - centroid based clustering: kmeans pro and cons
    > - density based clustering :?

51. random forests, how to implement, when to use random forest instead of other methods.

    > combines the output of multiple decision trees, so we should start with decision tree, ask multiple questions(the decision nodes in the tree) to determine an answer, 'yes' or 'no' follows the branches.
    > `decision tree can be prone to problems, such as bias and overfitting`
    > random forest is an ensemble of these decision trees, can be more accurate, especially when the decision trees are independent.
    >
    > ## how to implement:
    >
    > - Select random samples from a given dataset.
    > - Construct a decision tree for each sample and get a prediction result from each decision tree.
    > - Perform a vote for each predicted result.

> - Select the prediction result with the most votes as the final prediction.
>
> ## when to use:
>
> - the interpretability is not within our concern
> - random forest are less influenced by outliners(the vote mechanic?); well equipped to deal with noisy data
> - a good way to do feature selection since the if you fit the algorithm with features that are not useful, the algorithm simply won't use them to split on the data.(这点不是很懂)

52. resampling methods and how can they be used? What are their strengths and limitations(是说这四种重采样方法的优势和局限吗)

> ### resampling methods(when we are fitting a model, that means repeatedly drawing samples from a training set and refitting a model on each sample): There are four main types of resampling methods: `randomization`, `Monte Carlo`, `bootstrap`, and `jackknife`.
>
> ### how can they be used:
>
> - jackknife:
>   - remove one data point, calculate the statistic and the pseudovalue
>   - repeat this process, leaving out one data point at a time to build a set of n pseudovalues
>   - use the pseudovalues to estimate the parameter and the uncertainty

53. multiple testing, how to use in A/B testing

> Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses.(compare the treatment effects in a experiment with k treatment groups)
> if we want to conduct many A/B test simultaneously, we c

54. p-value and properties

> follow uniform distribution when the null hypothesis

55. statistical power

> power: in hypothesis test, reject the null hypothesis when any alternative hypothesis is actually true;
> why it is important?::
>
> - we can calculate power before running the test to help us decide the sample size of experiment design..
> - Having enough statistical power is necessary to draw accurate conclusions about a population using sample data.

56. testing the relationship between two category variables(contingency table)

> since we can not calculate the mean or stand deviation of categorical variables, instead we have frequencies, we can use chi-square test to measure the relationship between them.
> `chi-square test` can be used for test the independence of two or more categorical variables. we use contingency table to analysis the data (chi-square(large sample based statistics) test assumes that each cell has an expected frequency of five or more, the approximation is inadequate when sample sizes are small)
> `fisher exact test` used for contingency table analysis when the sample size is small, don't need the large sample assumption

57. test the difference between two vaccines

> we can randomly choose two groups of people, they have the same health conditions(which means we control other variables the same as much as possible), and give them different vaccines, then use the proportion of infected people in each vaccine group to compare the difference(by build a contingency table for these two variables), and do the chi-square test.

58. model to predict X-quarter sales based on past quarterly sales data

> since a time series analysis model involves using historical data to forecast the future, we can use this model to predict.
> like moving average, exponential moving average(the weight is less if the data is further from now), and ARIMA(Auto-regressive Integrated Moving Average), specify the parameters of ARIMA(set the window range), split the data into testing and training set, fit the model and plot the predict value and real value to see if they are in the same shape trend(goodness of fit)

59. shrinkage and penalization in regression

> when we do a multivariate linear regression, the columns of X may be linearly dependent, for example, the education level can be correlated with income level, this increases the variance of $\hat{\beta}$, in this case(in order to reduce variance), we can use a shrinkage estimator, such as ridge regression(Or LASSO), λ controls the amount of shrinkage, penalized the large value of β_j , need to pick value of λ; generalized cross-validation ("GCV") is a common method.
> we actually use penalization to shrink the big value of β, shrinkage mostly refer to ridge regression and LASSO, which measures β with L_1 and L_2 norm, penalization can be any arbitrary norm(I think)..

60. fixed effect and random effect for ANOVA based model

> - the treatment effect is random variable when it is treated as random effects, if it is fixed effect, than they are just fixed unknown parameters. The variance of Y_{ij} is different in two models, has another source of variance in the random effect ANOVA.
> - the error terms follows the iid assumptions in both models.

61. ANOVA with two factors vs two sample t-test

> The t-test of looks at quantitative outcomes with a categorical explanatory variable that has only two levels. The one-way Analysis of Variance (ANOVA) can be used for the case of a quantitative outcome with a categorical explanatory variable that has two or more levels of

> treatment, so when doing ANOVA analysis with the categorical variable has two levels, it's the same as two sample t-test.

62. interpret the coefficient of logistic regression

> the logistic regression equation $\frac{1}{1-exp(-(\beta_0+\beta_1 x)}$ implies the linear model for the log odds, $log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x$, β_1 means how much the log odds change given a unit change in the independent variable, β_0 means the value of log odds when the independent variable is 0

63. how to fit a logistic regression

> choose your independent and dependent variables, we use maximum likelihood to estimate β, when the training data is perfectly separated, the maximum estimation does not exist, we might need to use some iteration methods to estimate.

64. interpret the coefficient of multiple linear regression

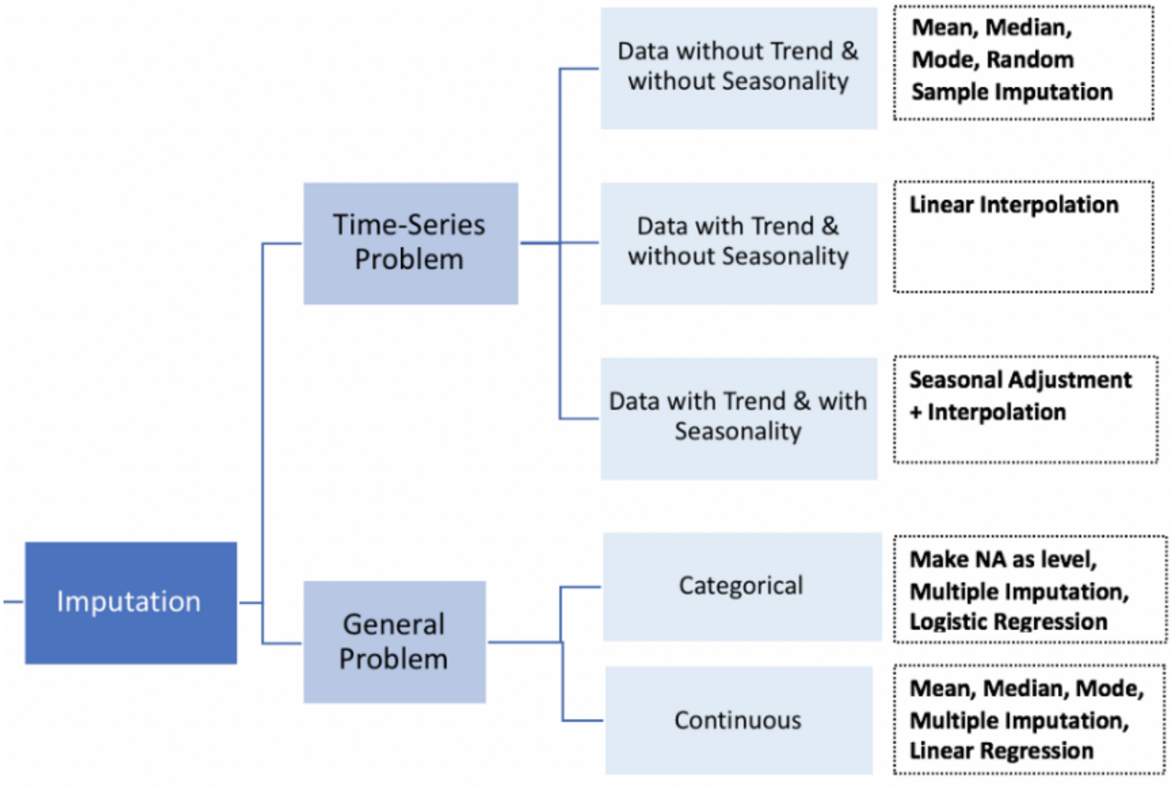
> How much the dependent variables changes given a unit change in the independent when control other independent variables constant. for β_0, it means the average value of Y_i for observations with other independent variables are 0.

65. . imputation methods
    `Imputation methods` are **those where the missing data are filled in to create a complete data matrix that can be analyzed using standard methods**.

> - use mean to fill missing value
> - regression imputation
> - Cold deck imputation: A systematically chosen value from an individual who has similar values on other variables.
> - k-nearest neighbors (dataset is small?)

66. `*missing value*` in outcomes for multiple linear regression

> imputation methods
> if the percentage of missing value is small, we can simply drop it


Imputation
Time-Series Problem
General Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

67. missing value of predictors in multiple linear regression

> matrix approximation
> imputation methods,
> simply drop it when the percentage is small
> for categorical variables, we can make NA as a level
> for a certain predictor, we can use the value of other samples which shares the same or close value in in other predictors.

68. multiple linear regression model :assumption and diagnose methods

> - The regression model is linear in regression parameters (*β-values*).
> - The residuals (ε*i*) follow a normal distribution and expected value (mean) of residuals, *E*(ε*i*|*Xi*), is zero.
> - In time series data, residuals are uncorrelated i.e. Cov(*εi, εj*)=0 ∀ *i≠j*.
> - The variance of the residuals, *Var*(ε*i*|*Xi*), is constant for all values of *Xi*. When the variance of the residuals is constant for different values of *Xi*, it is called **homoscedasticity** where as non-constant variance of residuals is called **heteroscedasticity**.
> - `residual plot, a plot between standardized residual value and standardized predicted value.`

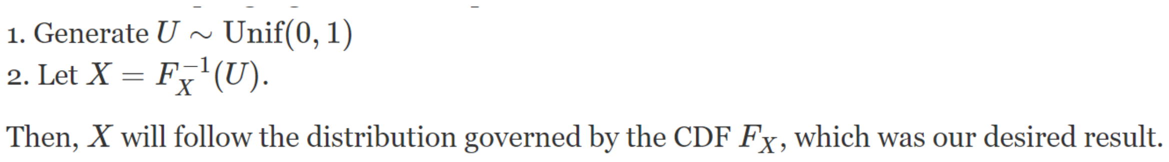
69. estimation for interval $[a, b]$

> min and max

70. estimation for {1,2,...M}

> max = M

71. generate sample for a specific distribution

> inverse method
>
> 
>
> 1. Generate $U \sim \text{Unif}(0, 1)$
> 2. Let $X = F_X^{-1}(U)$.
>
> Then, X will follow the distribution governed by the CDF F_X, which was our desired result.

72. let Uber driver know the demand

> since the demand is affected by time and historical data, we would use a time series model which incorporate the historical data to predict the demand. Such as moving average, exponential moving average and ARIMA, the features that we would consider to put in the model are time, blocks, weather, temperature and whether there are events or not.

73. anomaly detection algorithm for potentially false Uber requests

> we would use the requests as a binary outcome, and user history data, distance, type of car and location, time(such as the time an user use Uber most) and weather as potential features and then build a multiple linear regression model. we would assume to get an ideal dataset that contains the variables, and split them into two groups randomly as training and test dataset separately.

20-22是一样的，都是比较两种产品，用A/B test或者两样本t-test
22. design a way to test a new UX design.

> randomly pick two groups of people(by using the user graph to avoid the dependence between two closed users); randomly assign users with new

> and old buttons, tracking their interaction with the platform such as time spent, number of clicks on the button and use these data to build a metric to measure their degree of interaction with the button

19. get user-related data based on user graph, there might be dependency between them, so when we want skip the users who had a connection(the edge between two people) to get the related data.

20. Metropolis Hastings and Gibbs Sampling

> # Metropolis Hastings
>
> the distribution of X(t+1) depends only on the previous draw, we call this sequence a Markov Chain;
> given a current position x, move to y with the probability $q(y|x)$;
> calculate the ratio $r = \frac{p(y)q(x|y)}{p(x)q(y|x)}$;
> accept the proposed move with probability $\alpha = min\{1, r\}$;
> otherwise, remain at x(i.e. $X^{t+1} = X^t$
>
> # Gibbs Sampling

21. how would you know one algorithm is better than another?

> time taken by the algorithm to run
> space or memory it consumes while it runs
> accuracy of answers:
>
> - fit data (use goodness of fit metric)
> - predict (the performance on testing dataset)

22. concern with p-hacking A\B test and independency

> If you run enough t-tests over time, you're going to get a result that is statistically significant (p-hacking). P-hacking can result in dishonest results. Data may be the time series data, so it may be serially correlated, which is contrary to the independence assumption. Statistical tests commonly used for AB testing, like the two-sample z-test, `rely on the assumption that the experimental observations (i.e. samples) are independent`. If this assumption

> is not met, the test becomes unreliable in the sense that it may not achieve the desired false discovery rate.

23. test the ranking of users on the Amazon platform

24. In an *observational study*, values of the explanatory variable occur naturally.

> Difference:
>
> 1. In an observational study the data is gathered without the interference of the experiment to manipulate the environment
> 2. In randomized experiment the experimenter is randomly assigning treatments to subjects to attempt to avoid bias of confounding variables.

25. a metric to compare two users' rankings of products on Amazon's platform?

> t-test compare the average score of an user

26. voting preference within a specific state

> yes, we should pay attention to the clustering problem. for example, the education level could potentially be a confounding variable, a state of which the education level is low can be more likely to support republican party(which tend to be more conservative) which affects the outcome

27. confounding variables

> correlated to independent variable, and casually related to dependent variable
> Example: when we analyze the relationship between life span and the assumption of vitamin, the healthy habit can be a confounding variable 然后解释一下..
> how do we adjust the effect:
>
> - add potential confounding variables into our regression model
> - restrict it when choose our group, matching as much features the same as possible

28. selection bias

> when we decide who is to be studied in an experiment, the sample we choose may not be able to represent the situation of the whole population,

???

- anything about A/B test
  - Splitting traffic into more than three or four segments would make it hard to finish a test.
  - running cycles of tests one after another rather than more complex multivariate tests.
- There is no high correlation between independent variables (called **multicollinearity**). Multi-collinearity can result in an incorrect estimation of the regression parameters.

`By tracking the way visitors interact with the page they are shown – the videos they watch, the buttons they click, or whether or not they sign up for a newsletter – you can determine which version of the page is most effective.`

#todo # no stupid questions..

- □ 如何获取用户使用时间之类的数据，我感觉这些数据都会被软件本身追踪的。。难道不是本身就有的吗。。
- □ a/b test and t-test(continuous, independent, compare mean) 有啥关系。。