

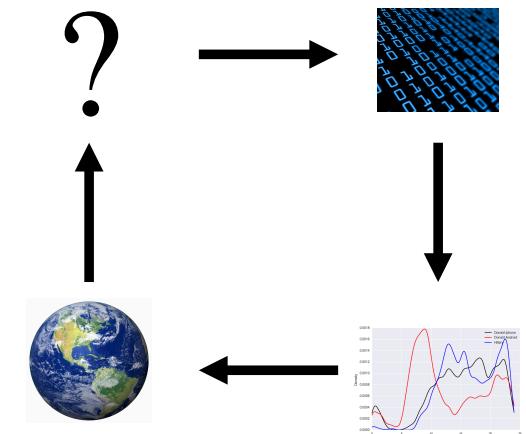
Data Science 100

Review

Problem formulation, experimental design,
Clustering, LS, and linear models

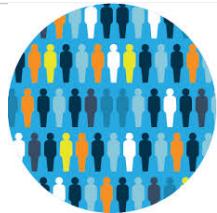
Slides by Bin Yu (some from Joe Gonzalez)

binyu@stat.berkeley.edu



Data analytics cycle through PQR+S

- Population



- Question



- Representative data collection (data neutral from the ethics angle)



- +... Prediction, inference, ... Least Squares, Linear Models, Clustering (K-means, EM, Hierarchical)...

- Scrutiny



P. “Population” of interest

- Population is the relevant group of people (objects, units) that the data-driven answer to the question will be applied to
 - Example: all patients who need a kidney transplant on May 1, 2017 (is this enough a description to define “P”?)
-

Q ? Question, question, question

- Can we predict acute kidney rejection ahead of time? -- clearly a prediction question
 - What are the important predictors for the acute rejection? -- an inference question since we need to address the uncertainty in the data
-

Predictors in the organ transplant project (dim=49)

- 43 genes are identified by the Sarwal Group at UCSF and collaborators (underlying kSORT, a blood test, to predict a rejection three months earlier than an invasive biopsy)
 - Recipient age, race and gender
 - Donor age
 - Donor/recipient gender match
 - Donor status (living or deceased)
-

Data collection

experimental design principle

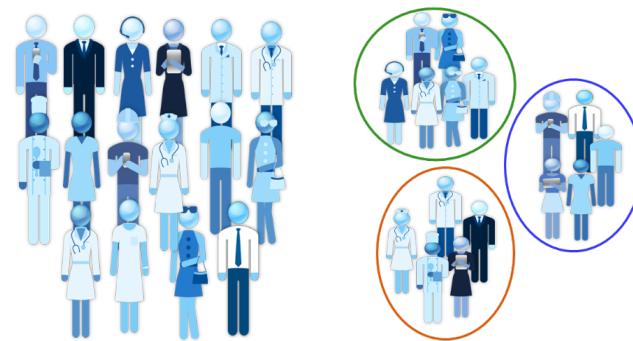
- Representativeness: data are similar to population in key aspects of interest
 - R- through randomization
 - Simple Random Sampling (randomized experiment vs. observational study)
 - Confounding factors (in both randomized and obs. studies)
 - Association is not causation in general esp. in obs. studies
-

Understanding data before modeling

- Visualization of raw data if possible
 - **Clustering** and then visualization
 - the clusters in low-dim (after PCA say)
 - look at the clusters, separately
 - Try different clustering methods and compare results
 - **Scrutiny**
-

Clustering is a form of information reduction/organization

- To store in human finite memory (or computer's finite memory) and facilitate understanding



- To communicate between people (or processors) for more effective understanding between people and collective decisions

Effective decision-making is impossible based on raw big data

Three clustering methods

- **K-means**
- **Hierarchical clustering:**
- EM: basic ideas (e.g. complete and incomplete data concepts)

K-means, given the number K of clusters

- Algorithm: alternating minimization between finding centers and partitions
- Starting values: random data points
- Selection of K: search for elbow in loss function, CV, Silhouette (graphic)
- Remarks:
 - K-means is sensitive to outliers if we use squared Euclidean distance
 - It converges to a local minimum of the loss function: good idea to start from multiple starting values and compare loss function values to find the best.

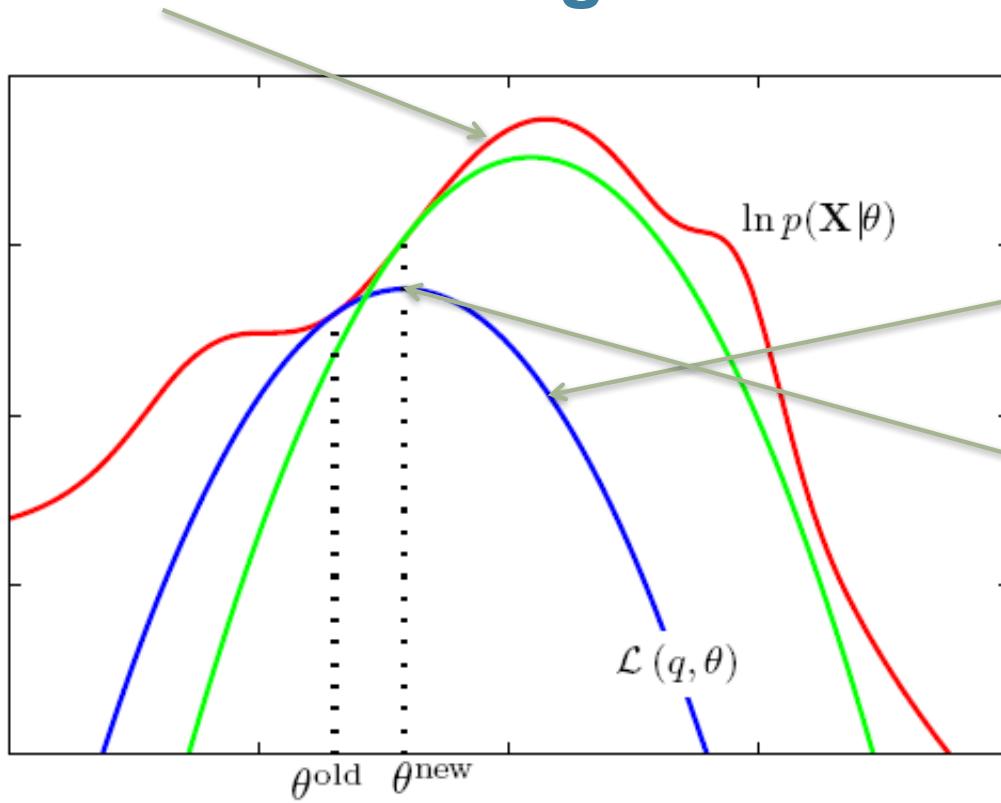
EM for model based clustering

- Applications: speech recognition with the hidden Markov Models (HMMs)
- Given iid data, no closed form for maximum likelihood.
- Numerical optimization routines can be used or the EM algorithm can be used.

The EM (Expectation-Maximization) algorithm is an iterative algorithm as gradient descent, however, the E- and M-steps have statistical meanings. The EM framework utilizes some missing data notation.

The EM algorithm is effective, when the maximum likelihood (M-step) is easy to solve with complete data (the missing data plus the observed data), and when the missing data is easy to impute (E-step). Modern generalizations of EM use approximations of various forms in the M- and E- steps.

EM (Dempster et al , 1977) to maximize the red curve – log likelihood function



- At a current estimate point,
- E-step makes a (lower bound blue) curve (in a systematic way) on the log likelihood function
 - M-step maximizes the blue curve for the next estimate
 - now on the green curve
 - ..

EM for two component Gaussian mixture

Data: incomplete **iid** data X_1, \dots, X_n without the labels

- Start with initial estimates of the 5 parameters
- Iterate between the E- and M-steps from previous slide until convergence: estimate the missing labels to impute the complete data; do maximum likelihood to estimate the parameters using the imputed complete data.
- It is guaranteed to converge to a local maximum of the likelihood function of the observed or incomplete data X_1, \dots, X_n

Hierarchical clustering

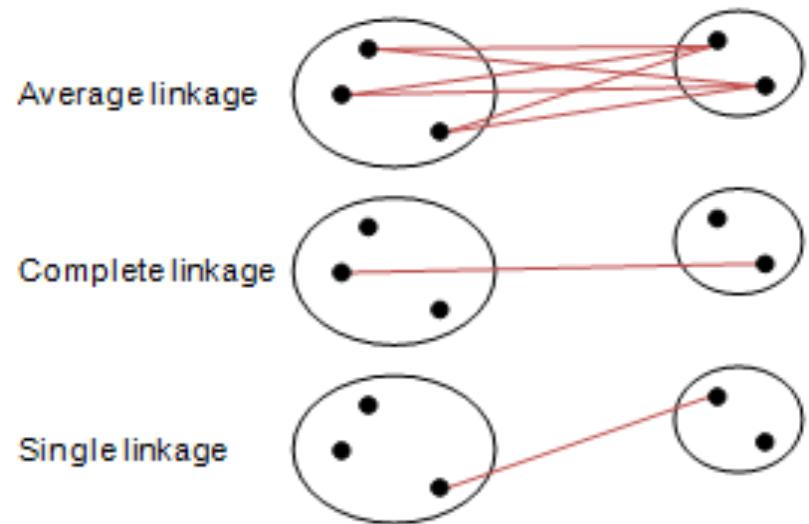
For a hierarchical clustering algorithm, we need

- A. a distance or dissimilarity measure between any two points.
- B. a rule to calculate the distances/dissimilarities between disjoint clusters of objects.

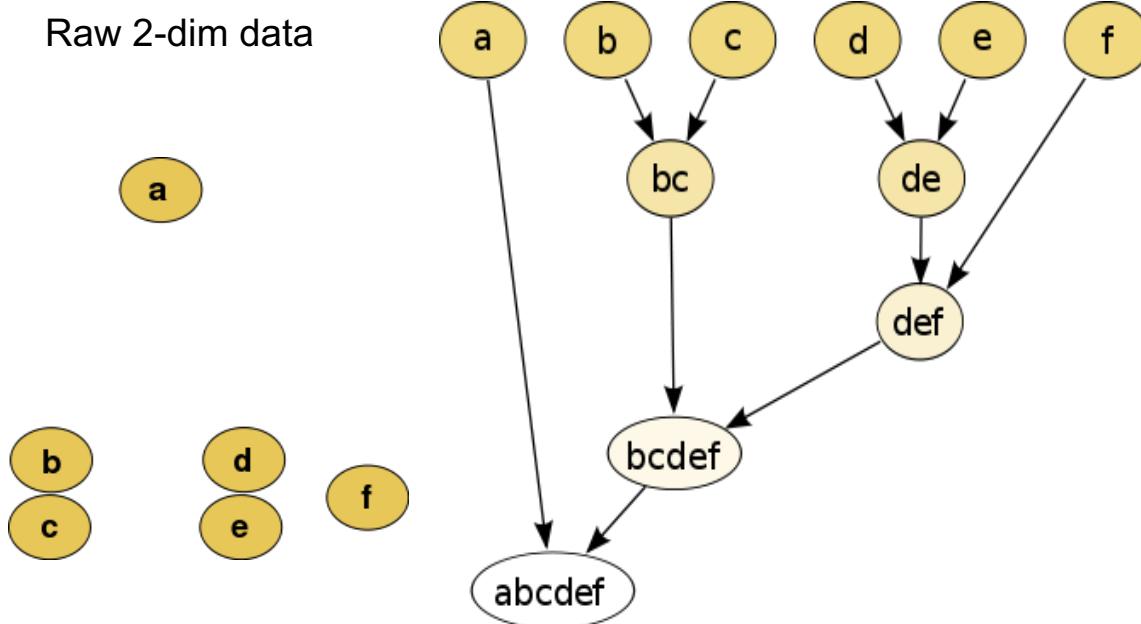
This between cluster distance can generally be calculated directly from the distances of the various elements involved in the clustering.

Three B. rules

- **Average linkage:**
average dissimilarities between two clusters
- **Complete-linkage:**
farthest points between the two clusters
- **Single-linkage:**
smallest dissimilarity between the two clusters



Wiki example

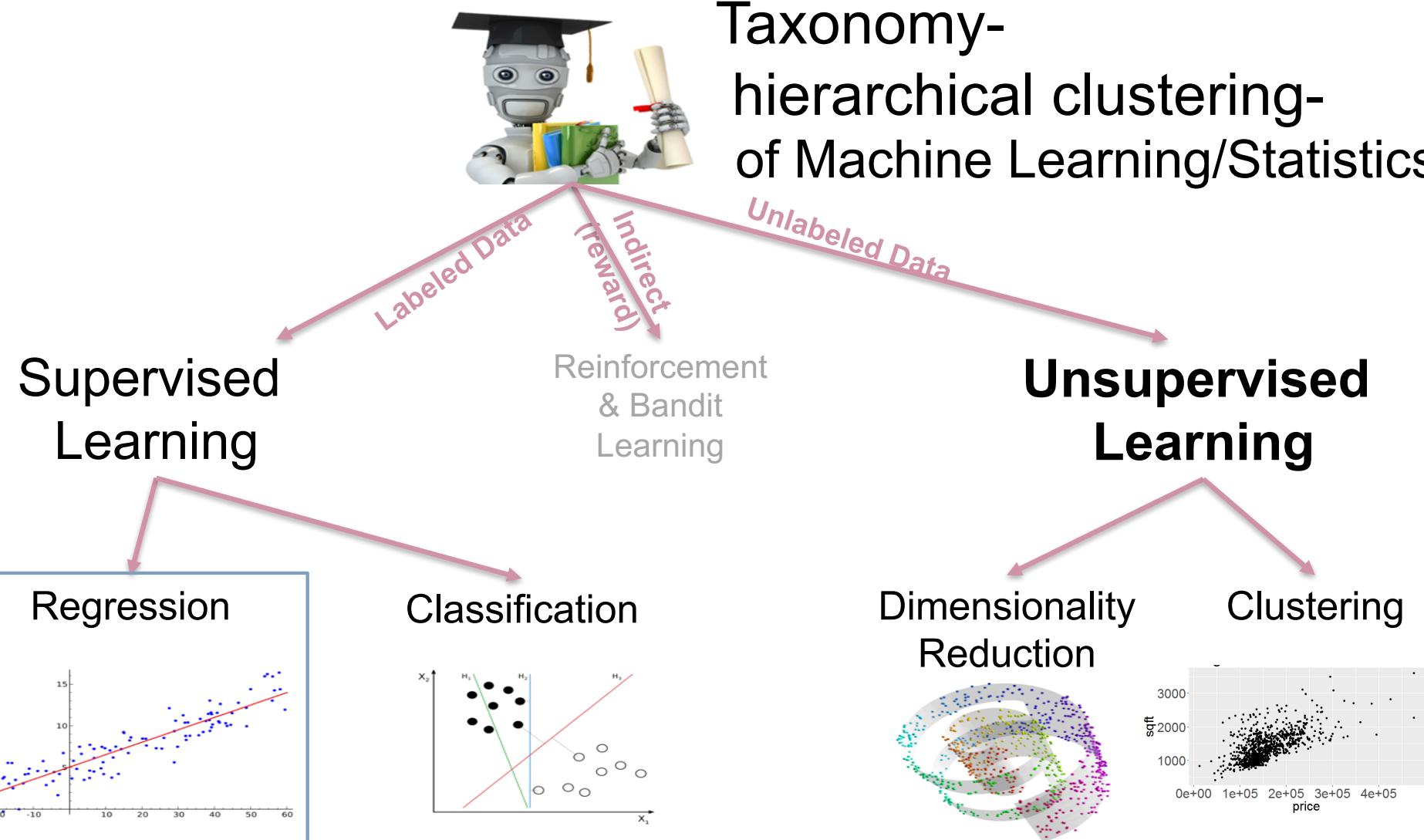


1. Take the two most similar
2. points and merge them into a new (super) point – we now have n-1 points

B-rule allows us to have a dissimilarity measure for the n-1 points

3. Continue until all points are grouped.

Taxonomy-hierarchical clustering-of Machine Learning/Statistics



LS: Least Squares as a fitting or prediction method

- Represent data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ as:

Covariate (Design) Matrix

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{np}$$

Response Vector

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

The diagram illustrates the representation of data for Least Squares. It shows two vectors, X and Y, representing the covariate (design) matrix and response vector respectively. The covariate matrix X is a column vector of length n, containing p columns labeled x1, x2, ..., xn. The response vector Y is a column vector of length n, containing elements y1, y2, ..., yn. Orange brackets indicate the dimensions: n for the number of observations and p for the number of predictors. The labels 'n' and 'p' are placed near the corresponding brackets.

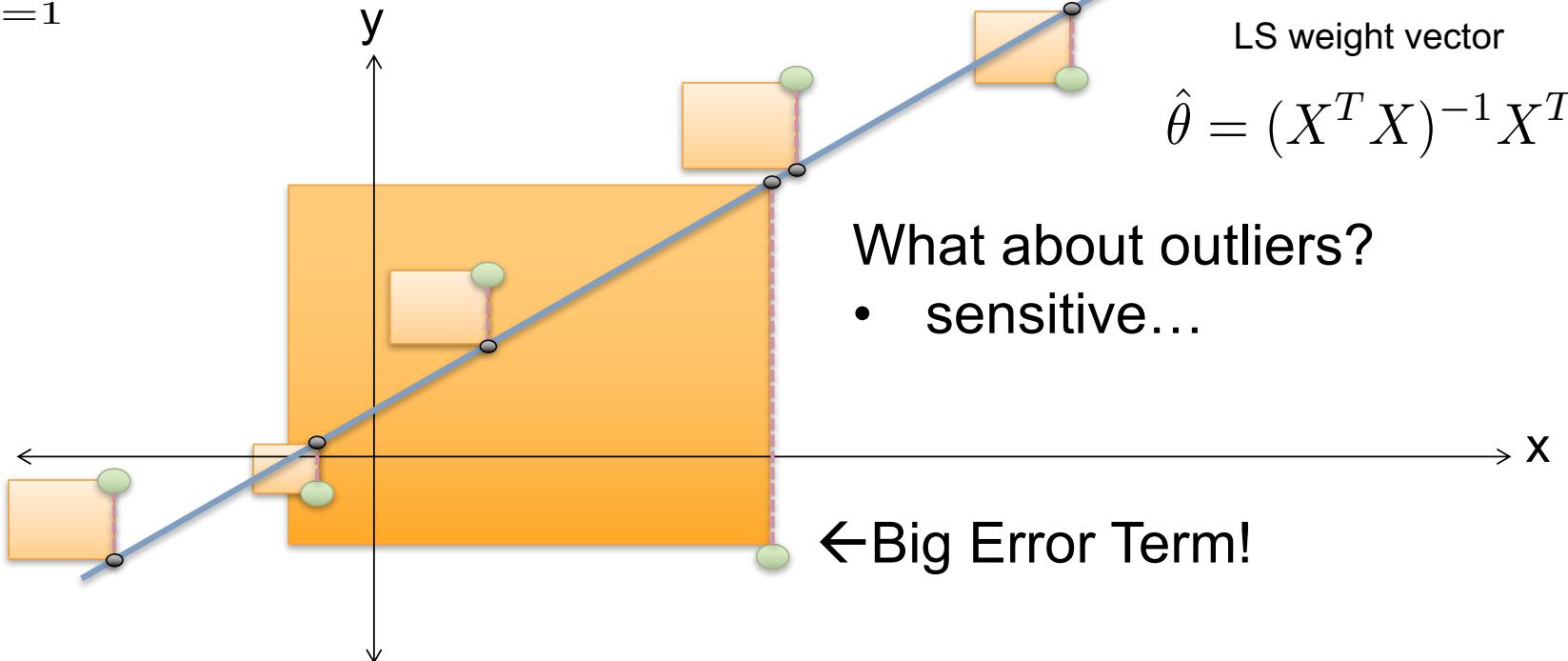
Least Squares (LS) predictor: $\hat{\theta}^T x$

$$\frac{1}{n} \sum_{i=1}^n (y_i - \theta^T x_i)^2$$

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (y_i - \theta^T x_i)^2$$

LS weight vector

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$



Basic understanding of LS estimator derivation

- Taking derivatives of the loss function and setting them to zero
- Basic matrix manipulations at the HW problem levels

e.g. Residuals

$$e_i = y_i - \hat{\theta}^T x_i$$

$$e = Y - \hat{Y}$$

where

$$\hat{Y} = X\hat{\theta} = X(X^T X)^{-1} X^T Y$$

Model assumption checking

- Residual plots (e against predictors, against fitted value) (
- Checking constant variance
- Checking dependence between the noise term and predictors

Normal (Gaussian) Linear Regression Model – idealized but useful

All models are wrong, but some are useful – George Box

$$y_i = \underbrace{\theta^T x_i}_{\begin{array}{c} \text{Vector of} \\ \text{Parameters} \\ \hline p \\ \parallel \end{array}} + \epsilon_i \begin{array}{c} \text{Vector of} \\ \text{Features} \\ \hline \text{Real Value} \\ \text{Noise} \end{array}$$

Real Valued Observations

Linear Combination of Covariates

$$\sum_{j=1}^p \theta_j x_{ij} \quad \theta, x \in \mathbb{R}^p$$

ϵ are independent, and independent of X , and with dist. $N(0, \sigma^2)$

$$\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$$

Maximum likelihood Est. under the Normal Linear Regression Model is the same as LS

- Since $y_i \sim N(\theta^T x_i, \sigma^2)$ and they are indep, the likelihood function is

$$\prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-(y_i - \theta^T x_i)^2 / (2\sigma^2) \right] \right\}$$

- The log likelihood function is

$$-\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n (y_i - \theta^T x_i)^2 / (2\sigma^2)$$

- Maximizing the above for θ is the same as minimizing the squared sum term or LS regardless of the value of variance σ^2

Estimating variance σ^2

- The maximum likelihood estimator (MLE) of σ^2 is

$$\hat{\sigma}_1^2 = \sum_{i=1}^n (y_i - \hat{\theta}^T x_i)^2 / n$$

where $\hat{\theta} = (X^T X)^{-1} X^T Y$ is the MLE or LS (or OLS) estimator of θ
OLS= Ordinary Least Squares (a statistics term)

- In fact, we always use another variance estimator

$$\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{\theta}^T x_i)^2 / (n - p)$$

Linear Regression Model – idealized but useful

All models are wrong, but some are useful – George Box

$$y_i = \theta^T x_i + \epsilon_i$$

Real Valued Observations

Vector of Parameters

Vector of Features

Real Value Noise

Linear Combination of Covariates

$\sum_{j=1}^p \theta_j x_{ij}$

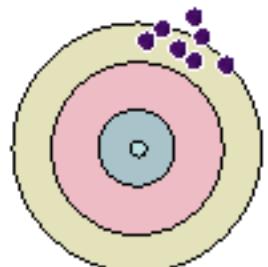
$\theta, x \in \mathbb{R}^p$

ϵ are independent, and independent of X , and with mean 0 and var. σ^2

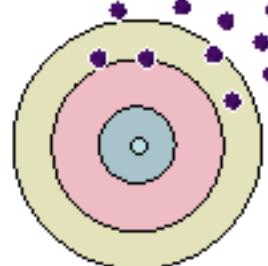
$$\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \quad (\epsilon \text{ 's distribution's tail is not too heavy})$$

Good properties of estimators: unbiasedness under linear model

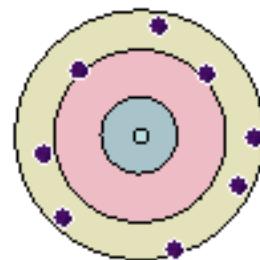
- $\hat{\sigma}^2$ is an unbiased estimator of σ^2
- $\hat{\theta}$ is an unbiased estimator of θ



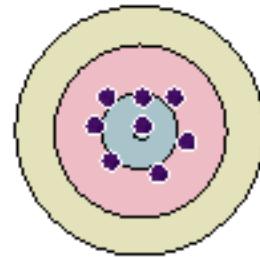
Bias is large
variation is small



Bias is large
variation is large



Bias is small
variation is large



Bias is small
variation is small

Accuracy versus Quality of an Estimator Using Bias and Variation as Measurable Quantities Respectively

Facts under Linear Regression Model

- LS estimator is still unbiased and so is $\hat{\sigma}^2$
- When **n is large**, LS estimator has an **approximate normal distribution**

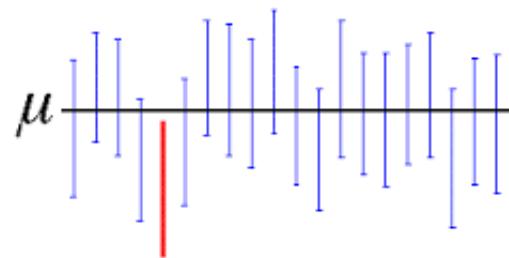
$$N(\theta, \sigma^2(X^T X)^{-1})$$

provided that $X^T X/n$ is approximately a positive definite matrix.

We can plug in $\hat{\sigma}^2$ for data-driven confidence intervals

What does a 95% confidence interval for θ_1 mean?

- Given 100 (**oracle**) sets of n random samples (or 100 replicates) following the Linear Regression Model, we get 100 confidence interval for θ_1 using approximate normality, about 95 of them should cover the true θ_1 in the model.



For one data set in practice, one can also use bootstrap for confidence interval construction of θ_1

- We need 100 or more bootstrap replicates from the data set to do the job
- use approximate normality or $N(\hat{\theta}, \hat{\sigma}^2(X^T X)^{-1})$

Hypothesis testing

- Null hypothesis H_0 : slope $\theta_1 = 0$
- Alternative hypothesis H_1 : $\theta_1 \neq 0$ (or $\theta_1 > 0$)

Remarks:

- Rationale: null hypothesis is favored unless there is strong evidence to refute it. Hence **accepting the null is not the same as proving null**.
- With one toss of a coin, the null (“coin is fair”) will be always accepted, which is apparently no proof that the coin is fair.

Testing hypothesis via confidence interval construction

- Use a 95% confidence interval construction to test the null:

if the interval contains the 0, the null is accepted;
otherwise rejected – all at level 5%

Two confidence interval constructions

- (Full-blown) Bootstrap
- (Approximate) normality

Scrutiny – vetting or validation

- Stability over perturbed data and methods (e.g. do different clustering methods give similar clusters? If not, which to take? Domain knowledge can help prioritize or believe only the stable clusters)
- Residual analysis for LS and linear models
- Domain matter interpretation (e.g. do the important predictors make sense medically?)
- Prediction on new patients: does it work better than old prediction method
- Down-stream impacts (e.g. If they develop new matching policy based on the important predictors found, does the rejection rate go down?)

Extra slide: terms used in hypothesis testing

- Type I error = Probability that the null rejected when that null is true
- Type II error = Probability that the alternative is rejected when it is true
- Power = $1 - \text{Type II error}$ = Prob. that the alternative is correctly accepted

In medical sciences (these terms below sound better or more positive)

- Sensitivity = $1 - \text{Type II error}$ = Power
- Specificity (or selectivity) = $1 - \text{Type I error}$
- Setting type I error at 5% -- statistically significant
- Setting type I error at 1% -- highly statistically significant

- Thank you for being a wonderful class – I had fun co-teaching it...
- Good luck with your finals