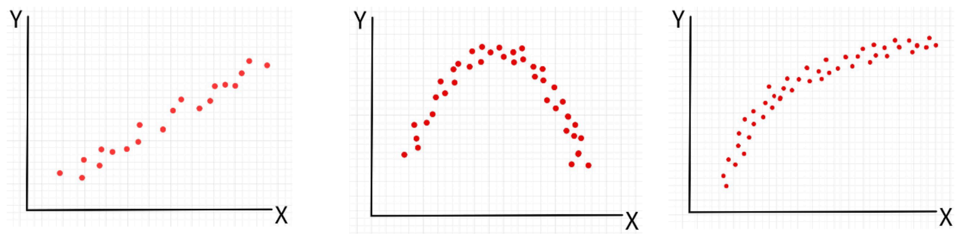
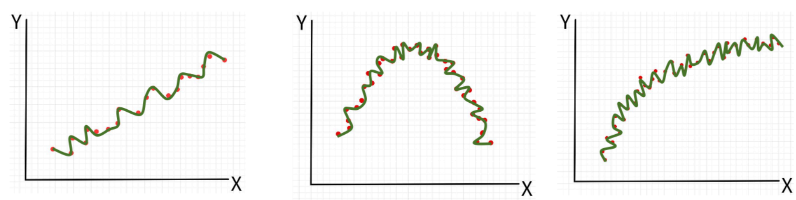
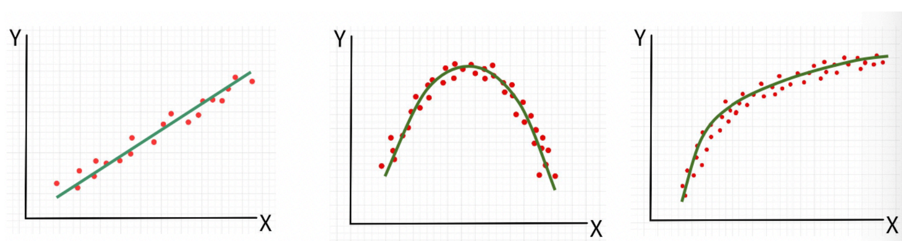
CAS CS 506
Lec 16

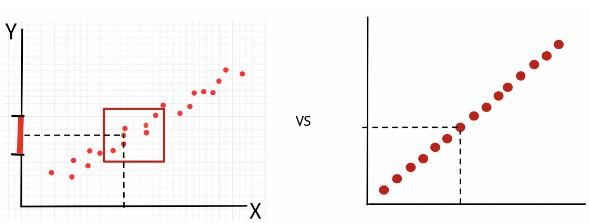
Linear Regression

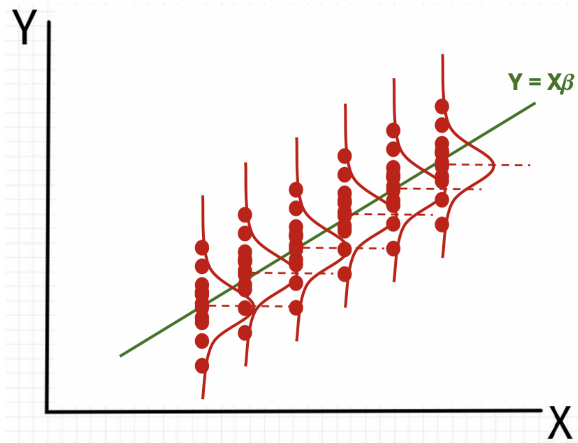
1. Motivation
    a. Given n samples / data points (y_i, x_i). Y is a continuous variable (as opposed to classification)
    b. 

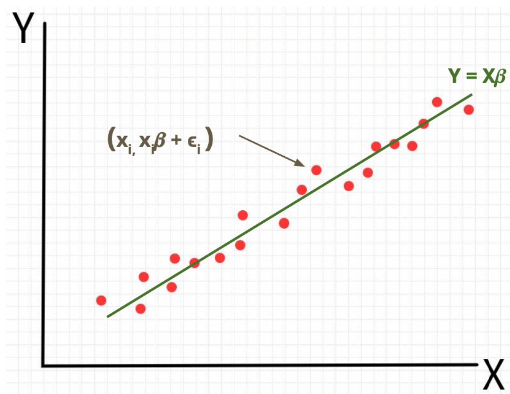
    c. Understand/explain how y varies as a function of x (i.e. find a function y = h(x) that best fits our data)
    d. Should h be the curve that goes through the most samples? I.e. do we want $h(x_i) = y_i$ for the maximum number of i?
        i. 

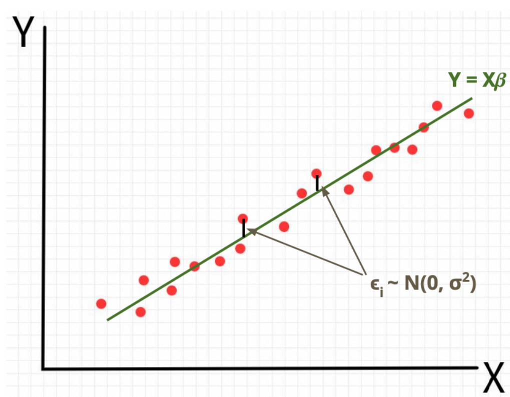
        ii. h may be too complex
        iii. overfitting - may not perform well on unseen data
    e. The following curves seem the most intuitive "best fit" to our samples. How can we define this best fit mathematically? Is it just about finding the right distance function?

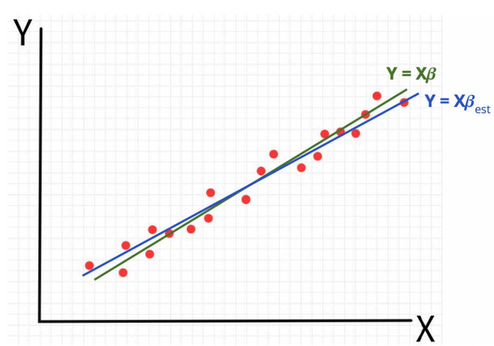


2. Assumptions
    a. 


Y
Y = Xβ
X
b.


Y
Y = Xβ
(x_i, $x_i\beta + \epsilon_i$)
X
c.


Y
Y = Xβ
$\epsilon_i \sim N(0, \sigma^2)$
X
d.


Y
Y = Xβ
Y = Xβ_{est}
X
e.

f. Our data was generated by a linear function plus some noise:
    i. $y = h_x(B) + e$
    where h is liner in a parameter B
    where e are independent N(0, o^2) distribution

3. Cost Function
    a. Given our data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
    b. Suppose we are given a curve y = h(x), how can we evaluate whether it is a good fit to our data?
    c. Compare $h(x_i)$ to y_i for all i
    d. Goal: For a given distance function d, find h where L is smallest
    e. $$L(h) = \sum_i d(h(x_i), y_i)$$
4. Assumptions
    a. The relation between x (indepdent variable) and y (dependent variable) is linear in a parameter B
    b. e are independent, identically distributed random variables following a N(0, o^2) distribution
5. Goal
    a. Given these assumptions, let's try to minimize the cost function denied earlier
    b. What parameter(s) are we trying to learn / estimate?
        i. A: B
6. Least Squares
    a. $$\beta_{LS} = \arg\min_{\beta} \sum_i d(h_\beta(x_i), y_i)$$
    (x_i, y_i) are from our dataset
    b. $$\begin{aligned} \beta_{LS} &= \arg\min_{\beta} \sum_i d(h_\beta(x_i), y_i) \\ &= \arg\min_{\beta} \|\vec{y} - h_\beta(\mathrm{X})\|_2^2 \\ &= \arg\min_{\beta} \|y - X\beta\|_2^2 \end{aligned}$$

c.
$$\frac{\partial}{\partial \beta}(y - X\beta)^T(y - X\beta) = 0$$
$$\frac{\partial}{\partial \beta}(y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta) = 0$$
$$\frac{\partial}{\partial \beta}(y^T y - 2\beta^T X^T y - \beta^T X^T X\beta) = 0$$
$$-2X^T y - X^T X\beta = 0$$
$$X^T X\beta = X^T y$$
$$\boxed{\beta_{LS} = (X^T X)^{-1} X^T y}$$

7. Assumptions
    a. Our data was generated by a linear function plus some noise:
    $$\vec{y} = h_X(\beta) + \vec{\epsilon}$$
    b. Where h is linear in a parameter B.
    c. Which functions below are linear in B?
        i. $h(B) = B_1 * x$
        ii. $h(B) = B_0 + B_1 x$
        iii. $h(B) = B_0 + B_1 x + B_2 * x^2$
        iv. $h(B) = B_1 \log(x) + B_2 * x^2$
        v. $h(B) = B_0 + B_1 x + B_1^2 * x$
8. Maximum Likelihood
    a. Another way to define this problem is in terms of probability
    b. Define P(Y|h) as the probability of observing Y given that it was sampled from h
    c. Goal: Find h that maximizes the probability of having observed our data
    d. Maximize L(h) = P(Y|h)
    e. Since e ~ N(0, o^2) and Y = XB + e then Y ~ N(XB, o^2)
    f.
$$\begin{aligned}\beta_{MLE} &= \arg\max_{\boldsymbol{\beta}} \frac{1}{\sqrt{(2\pi)^n \sigma^n}} \exp(-\frac{\|y - X\beta\|_2^2}{2\sigma^2}) \\ &= \arg\max_{\boldsymbol{\beta}} \exp(-\|y - X\beta\|_2^2) \\ &= \arg\max_{\boldsymbol{\beta}} -\|y - X\beta\|_2^2 \\ &= \arg\min_{\boldsymbol{\beta}} \|y - X\beta\|_2^2 \\ &= \beta_{LS} = (X^T X)^{-1} X^T y\end{aligned}$$

9. An Unbiased Estimator
    a. BLS is an unbiased estimator of the true B. That is E[BLS] = B

$$
\begin{aligned}
E[\beta_{LS}] &= E[(X^T X)^{-1} X^T y] \\
&= (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T E[X\beta + \epsilon] \\
&= (X^T X)^{-1} X^T X\beta + E[\epsilon] \\
&= \beta
\end{aligned}
$$

    b.