INF552 Machine Learning

Liyue Fan
liyuefan@usc.edu
Integrated Media Systems Center
University of Southern California

Notes

- Presentation schedule on blackboard
  - double-check your assignment
  - 10/19 (5) -> 10/12 (3); 11/16(5) -> 11/02(3)
- Presentation folder on blackboard
  - upload your talk slides
- K-means example in W3 slides – prepared by Kien
- HW2 to give out on Wednesday

Regression Methods

Regression- Supervised Learning

- Training data includes response variable, or dependent variable:
  - predict housing price (continuous)
  - predict whether a patient has coronary heart disease (binary)
- Training consists of learning model parameters
- Linear Regression -> continuous output
- Logistic Regression -> binary classification

Recall Pearson's Correlation

- Measures the relative strength of the *linear* relationship between two variables

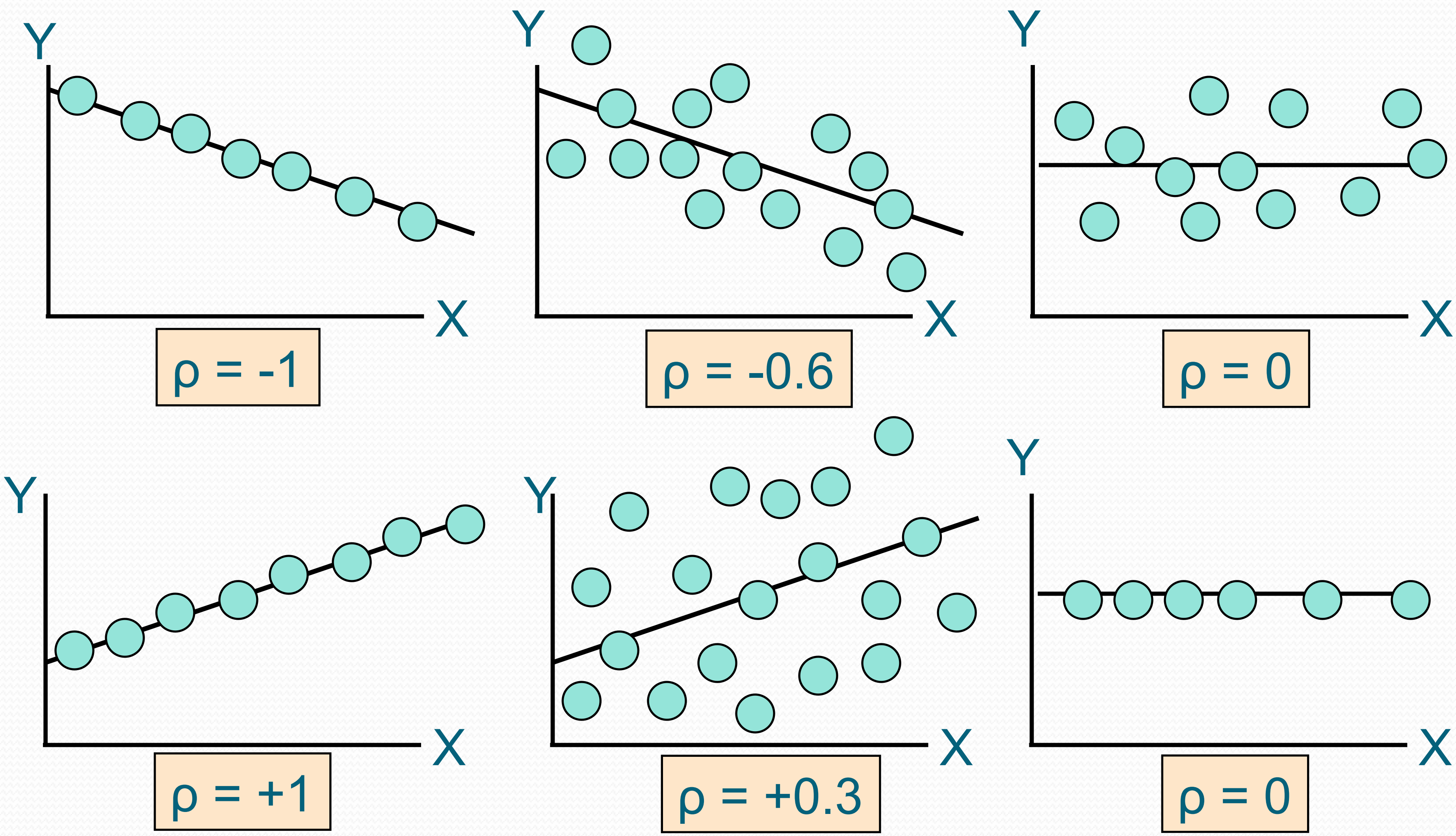
$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

$$\text{where } \mathrm{SS_x} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \text{and } \mathrm{SS_{xy}} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Correlation

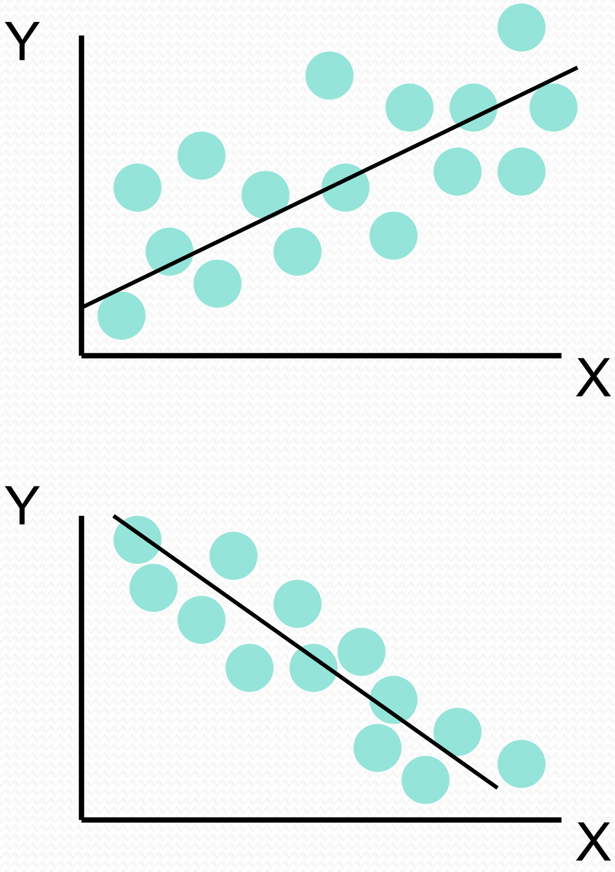
- Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

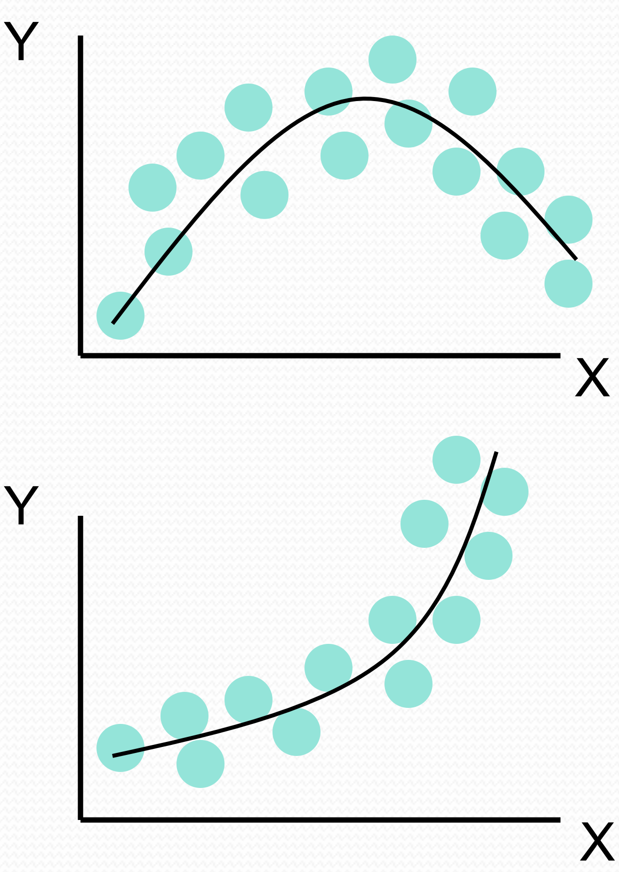


Linear Correlation

Linear relationships



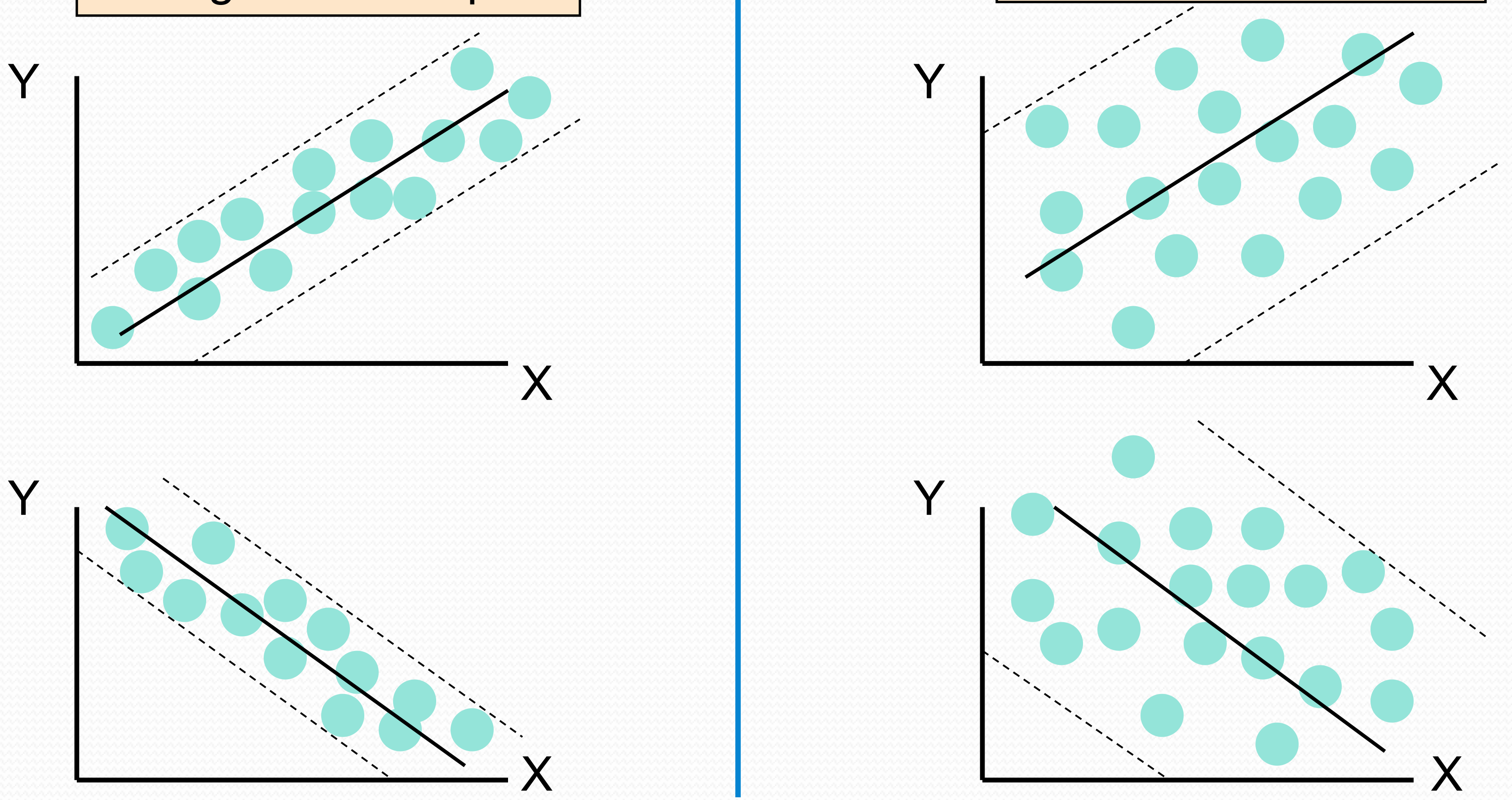
Curvilinear relationships



Linear Correlation

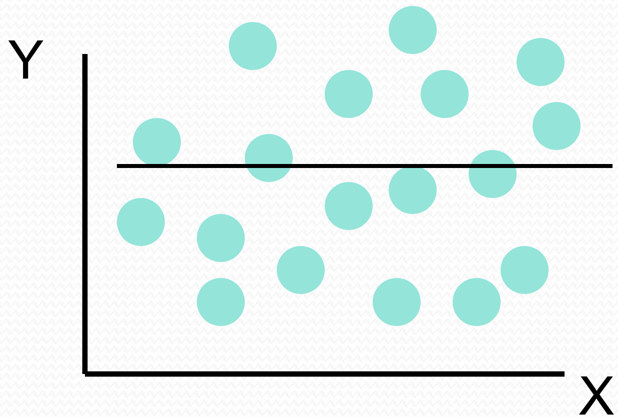
Strong relationships

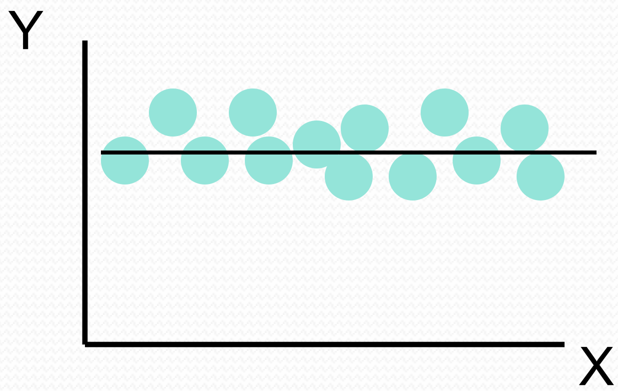
Weak relationships



Linear Correlation

No relationship



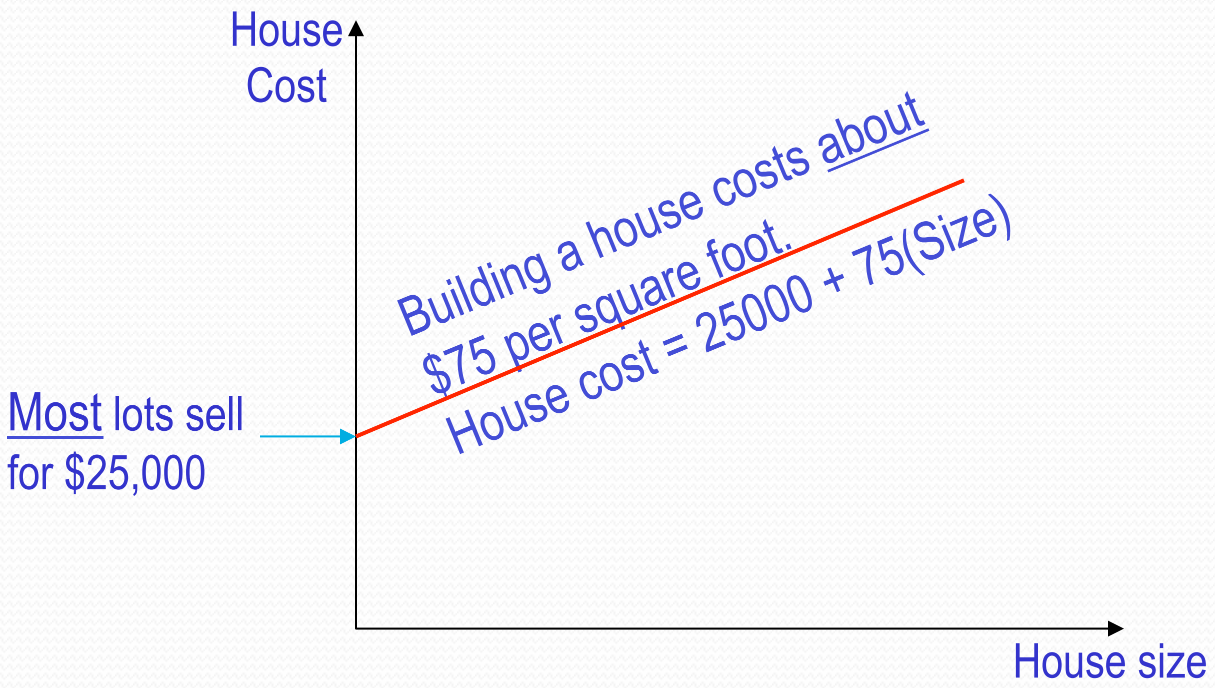


Linear Regression

- In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable(s) x and the other the dependent (=response) variable y.
- The motivation for using the technique:
  - Forecast the value of a dependent variable (y) from the value of independent variables (x_1, x_2,…x_k.).
  - Analyze the specific relationships between the independent variables and the dependent variable.

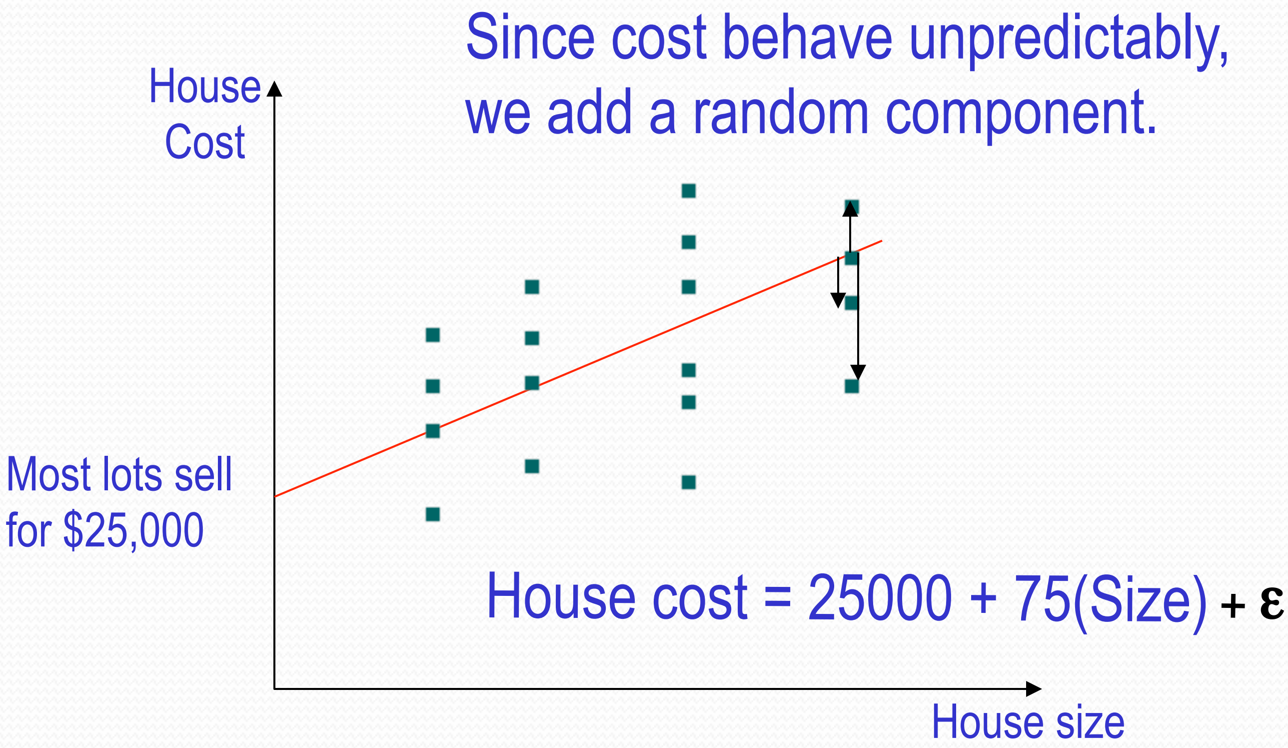
The Model

The model has a deterministic and a probabilistic components



The Model

However, house cost vary even among same size houses!



The Model

- The first order linear model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

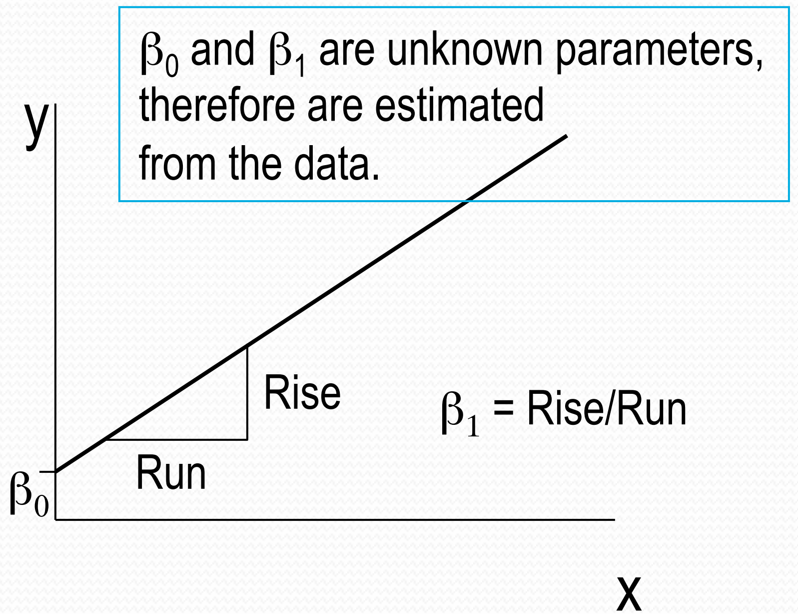
y = dependent variable

x = independent variable

β_0 = y-intercept

β_1 = slope of the line

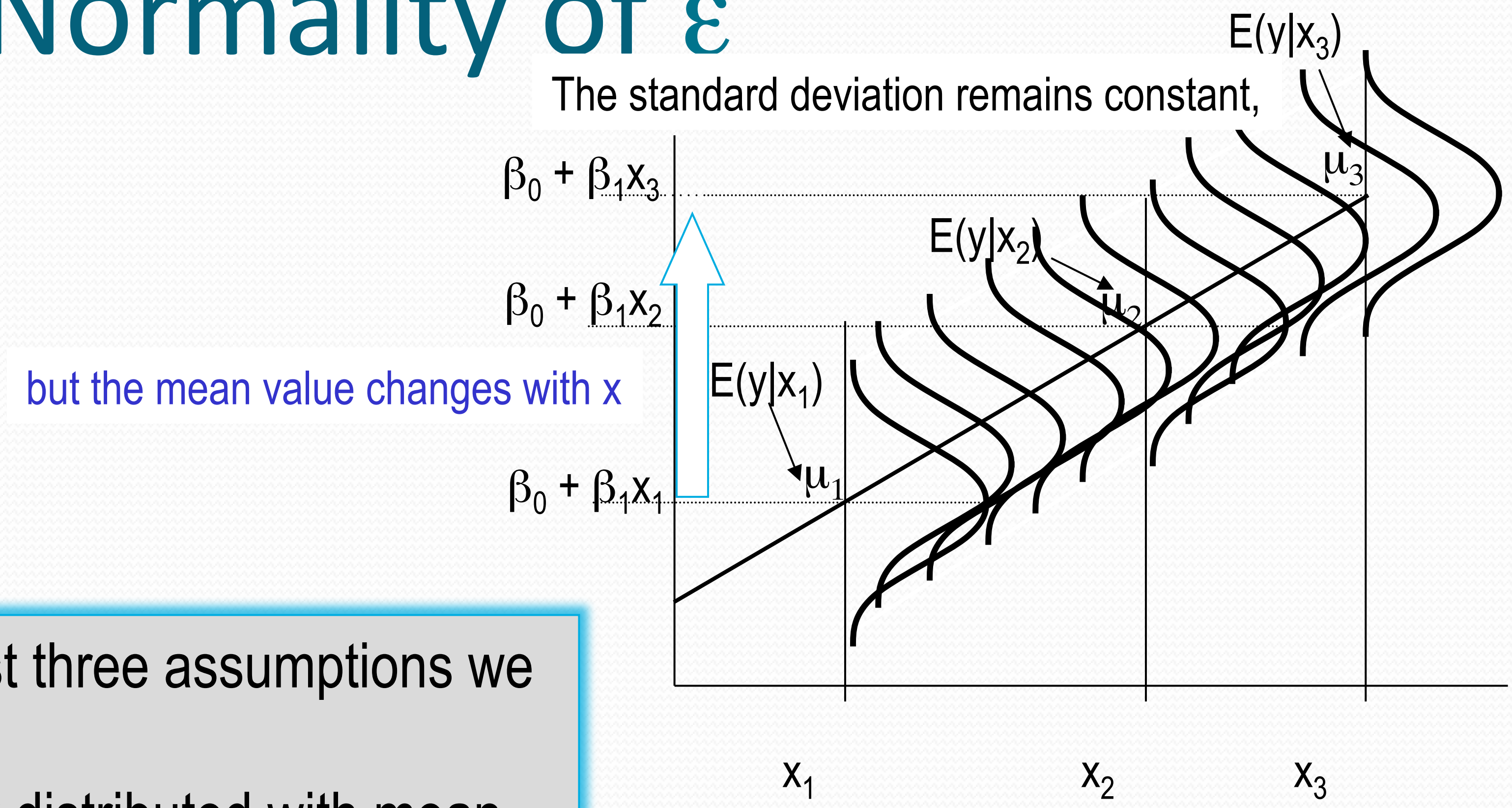
ε = error variable



Error Variable: Required Conditions

- The error ε is a critical part of the regression model.
- Four requirements involving the distribution of ε must be satisfied.
  - The probability distribution of ε is normal.
  - The mean of ε is zero: $E(\varepsilon) = 0$.
  - The standard deviation of ε is σ_ε for all values of x.
  - The set of errors associated with different values of y are all independent.

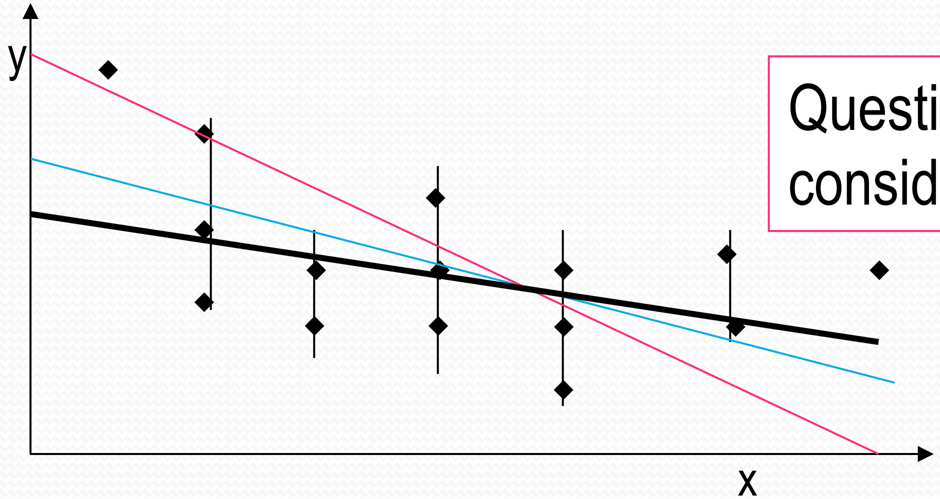
The Normality of ε



From the first three assumptions we have:
y is normally distributed with mean $E(y) = \beta_0 + \beta_1 x$, and a constant standard deviation σ_ε

Learning the Coefficients

- The estimates are determined by
  - training sampling/data drawn from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



Question: What should be considered a good line?

The Least Squares (Regression) Line

A good line is one that minimizes
the sum of squared differences between the points and the line.

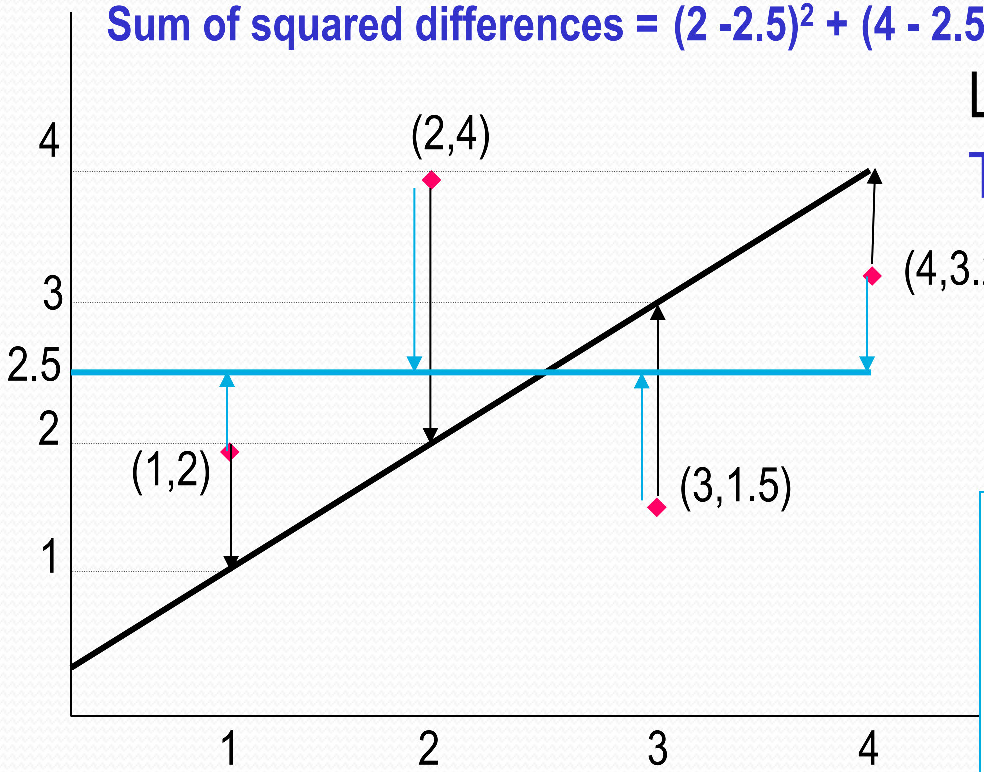
The Least Squares (Regression) Line

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$

Let us compare two lines

The second line is horizontal



The smaller the sum of squared differences the better the fit of the line to the data.

Minimize Sum of Squared Errors

- Regression model (expected value):

$$\hat{y} = \beta_0 + \beta_1 x$$

$$SSE\left(\beta_0, \beta_1 | \mathcal{X}\right) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$

- To minimize SSE:

$$\frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0} = 0$$

$$\frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_1} = 0$$

The Estimated Coefficients

To calculate the estimates of the slope and intercept of the least squares line , use the formulas:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

A shortcut for the slope

$$\beta_1 = \frac{SS_{xy}}{SS_x}$$

Alternate formula with Pearson Correlation for the slope

$$\beta_1 = \rho \frac{sd_y}{sd_x}$$

The Simple Linear Regression Line

- Example:
  - A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
  - A random sample of 6 cars is selected, and the data recorded.
  - Find the regression line.

Car	Odometer	Price
1	37388	14636
2	44758	14122
3	45833	14016
4	30862	15590
5	31705	15568
6	34010	14718

Independent variable x   Dependent variable y

The Simple Linear Regression Line

- Solving by hand:

$$\bar{x} = 37426; \qquad SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} =$$

$$\bar{y} = 14775; \qquad SS_{xy} = \sum (x_i y_i) - \frac{\sum x_i \sum y_i}{n} =$$

where n = 6.

Car	Odometer	Price
1	37388	14636
2	44758	14122
3	45833	14016
4	30862	15590
5	31705	15568
6	34010	14718
	x	y

$$\beta_1 = \frac{SS_{xy}}{SS_x} =$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} =$$

$$\boxed{\hat{y} = \beta_0 + \beta_1 x =}$$

Multiple Linear Regression

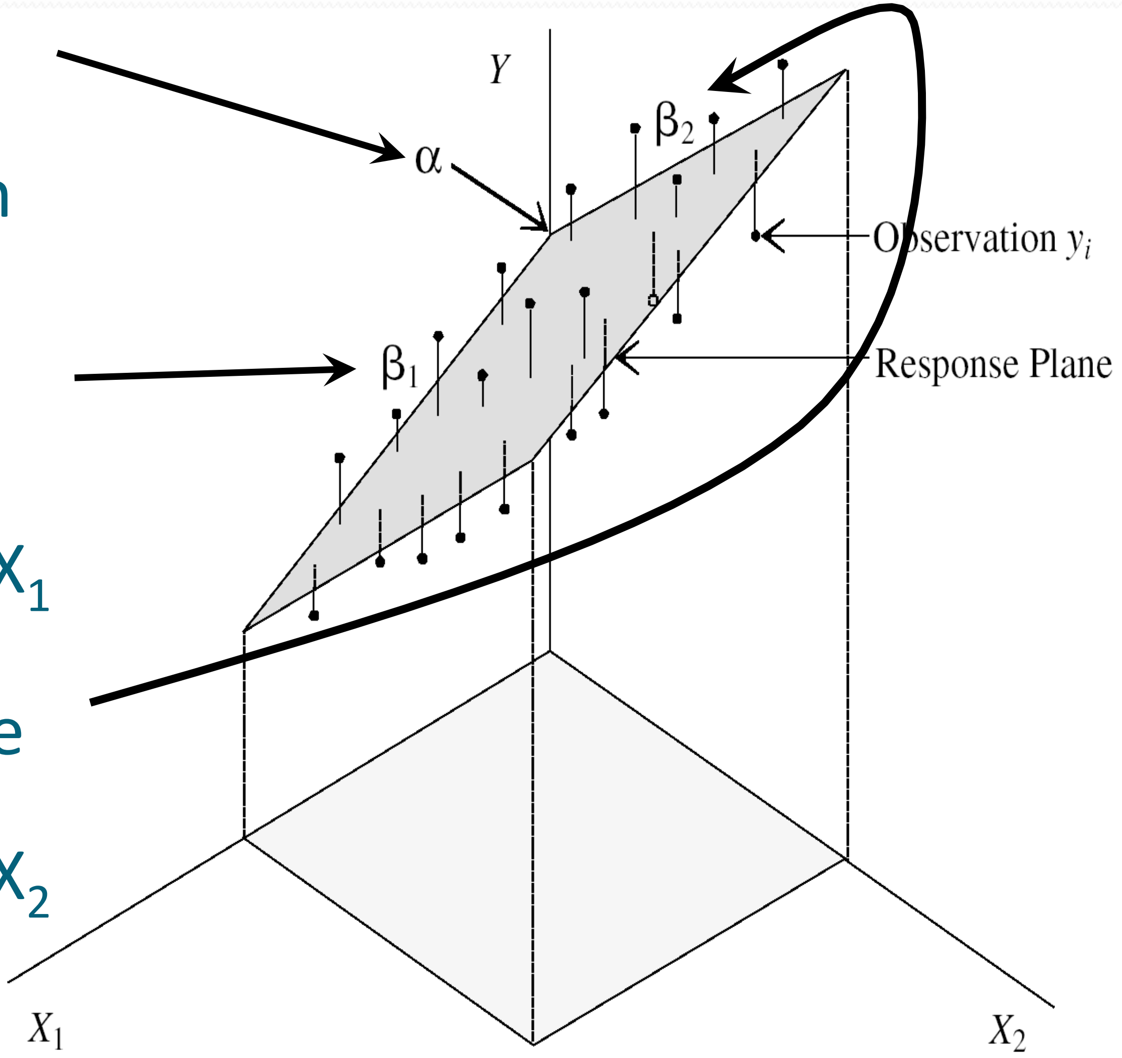
- More than one predictor

  $y = \beta_0 + \beta_1 * x + \beta_2 * w + \ldots$

- Learn β_0 ,β_1,β_2 given training data $\{x_i, w_i, y_i\}_{\mathrm{i=1\ldots n}}$
- Hint: Least Squares

Example: Multiple Regression Model

- Intercept α predicts where the regression *plane* crosses the Y axis
- Slope for variable X_1 (β_1) predicts the change in Y per unit X_1 holding X_2 constant
- The slope for variable X_2 (β_2) predicts the change in Y per unit X_2 holding X_1 constant



Model Evaluation – Linear Regression

- The least squares method will produces a regression line whether or not there is a linear relationship between x and y.
- Consequently, it is important to assess how well the linear model fits the data.

Sum of Squares for Errors

- This is the sum of differences between the points and the regression line.
- It can serve as a measure of how well the line fits the data.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

– A shortcut formula

$$SSE = \sum y_i^2 - \beta_0 \sum y_i - \beta_1 \sum x_i y_i$$

Note: R^2 = (Pearson's correlation)^2

Coefficient of determination

- R^2 measures the proportion of the variation in y that is explained by the variation in x.

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - SSE}{\sum (y_i - \bar{y})^2}$$

- R^2 takes on any value between zero and one.

  $R^2 = 1$: Perfect match between the line and the data points.

  $R^2 = 0$: There are no linear relationship between x and y.

Coefficient of Determination Example

You're a marketing analyst for Hasbro Toys. You know $\rho = .904$.

Ad Expenditure (100$)	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Calculate and interpret the **coefficient of determination.**

Coefficient of Determination Solution

r^2 = (Pearson's correlation)^2
$r^2 = (.904)^2$
$r^2 = .817$

Interpretation: About 81.7% of of the variation in Sales (y) is explained by the variation in Ad \$ (x). The rest (18.3%) remains unexplained by this model.

Logistic Regression

- Models relationship between set of independent variables X_i
  - binary (yes/no, smoker/nonsmoker,…)
  - categorical (social class, race, ...)
  - continuous (age, weight, gestational age, ...)

and

  - binary categorical response variable Y

    e.g. Success/Failure, Remission/No Remission Survived/Died, etc…

Logistic Regression

Example: Coronary Heart Disease (CD) and Age

In this study sampled individuals were examined for signs of CD (present = 1 / absent = 0) and the potential relationship between this outcome and their age (yrs.).

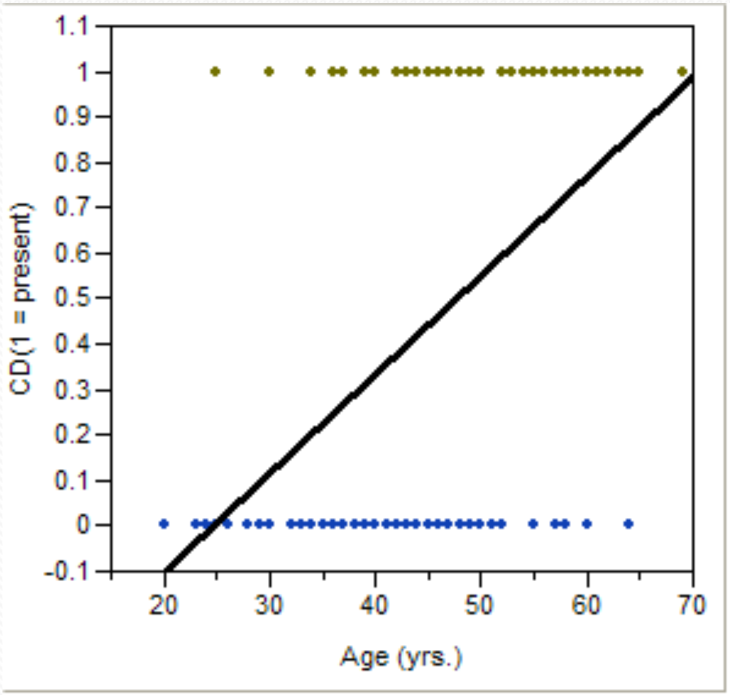
	Agegrp	Age	CD	Agegrp	Age	CD
1	1	20	0	2	30	0
2	1	23	0	2	30	0
3	1	24	0	2	30	0
4	1	25	0	2	30	0
5	1	25	1	2	30	1
6	1	26	0	2	32	0
7	1	26	0	2	32	0
8	1	28	0	2	33	0
9	1	28	0	2	33	0
10	1	29	0	2	34	0
11	2	30	0	2	34	0

. . .

Agegrp	Age	CD
8	60	0
8	60	1
8	61	1
8	62	1
8	62	1
8	63	1
8	64	0
8	64	1
8	65	1
8	69	1

Logistic Regression

Simple Linear Regression?

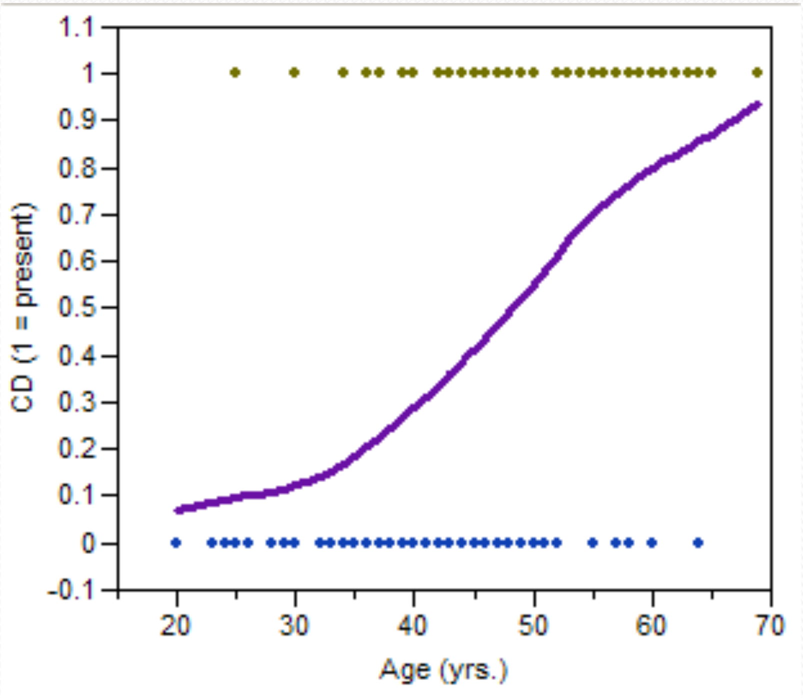


$E(CD \mid Age) = -.54 + .02 \cdot Age$

e.g. For an individual 50 years of age

$E(CD \mid Age = 50) = -.54 + .02 \cdot 50 = .46$??

Smooth Regression Estimate?



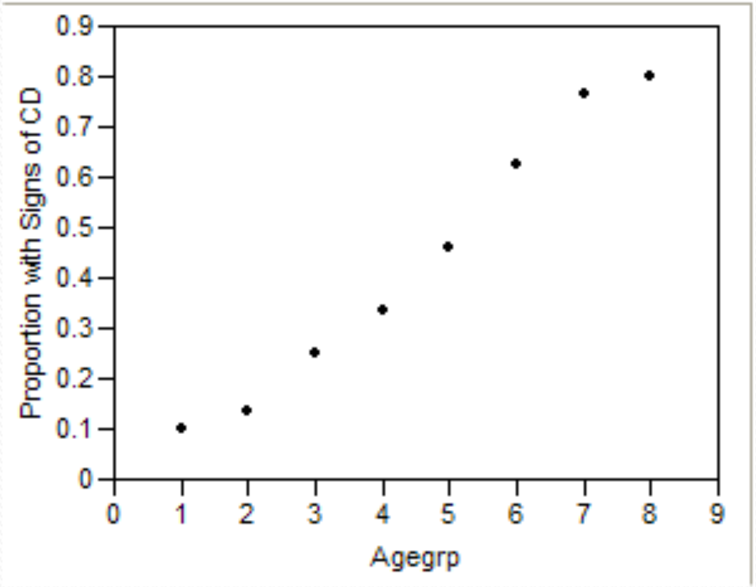
The smooth regression estimate is "S-shaped" but what does the estimated mean value represent?

Answer: P(CD | Age)!!!!

Logistic Regression

We can group individuals into age classes and look at the percentage/proportion showing signs of coronary heart disease.

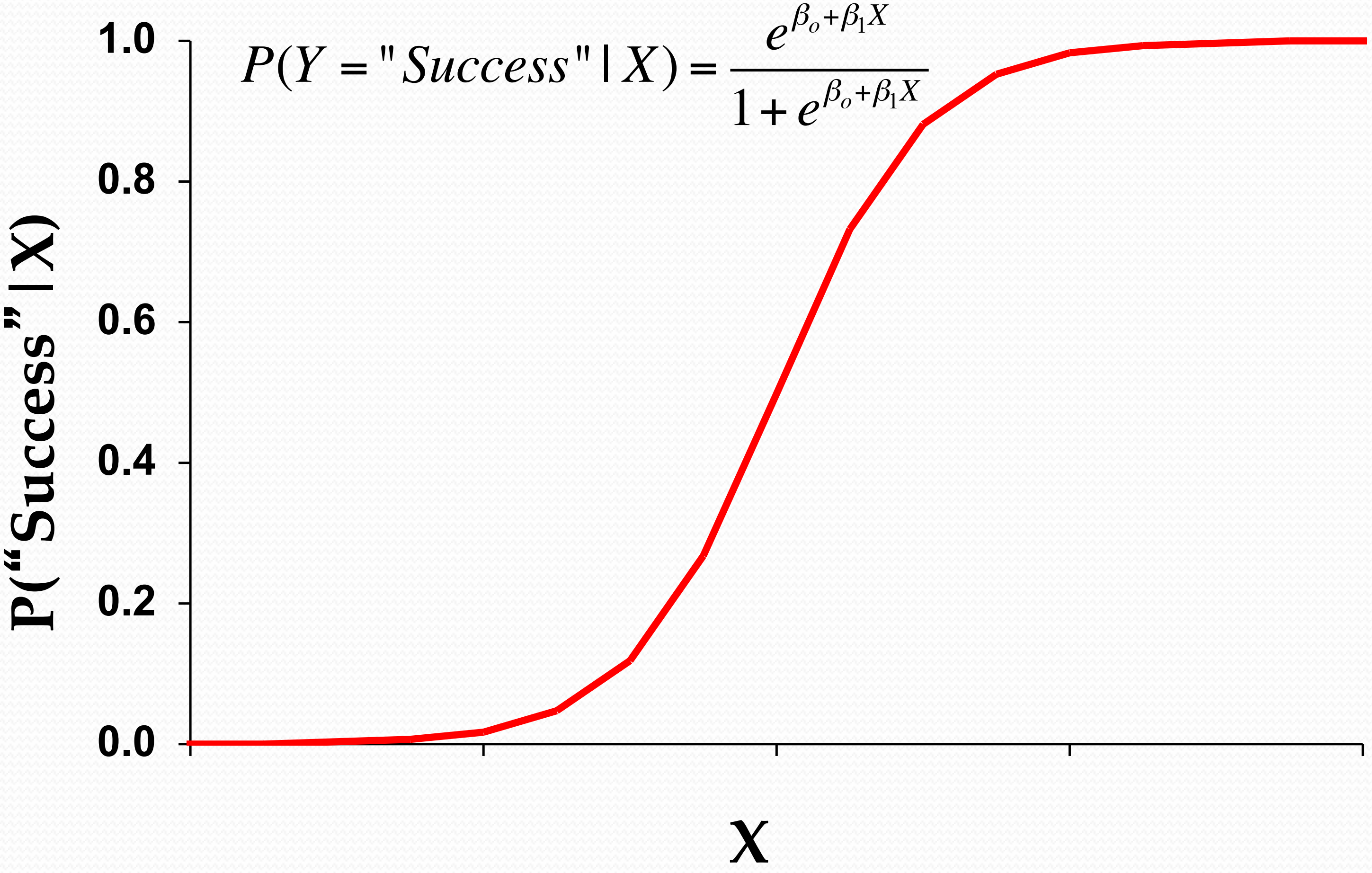
		Diseased	
Age group	**# in group**	**#**	**Proportion**
1) 20 - 29	10	1	.100
2) 30 - 34	15	2	.133
3) 35 - 39	12	3	.250
4) 40 - 44	15	5	.333
5) 45 - 49	13	6	.462
6) 50 - 54	8	5	.625
7) 55 - 59	17	13	.765
8) 60 – 64	10	8	.800



Notice the "S-shape" to the estimated proportions vs. age.

Logistic Function

$$P(Y = "Success" \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$



Logit Transformation

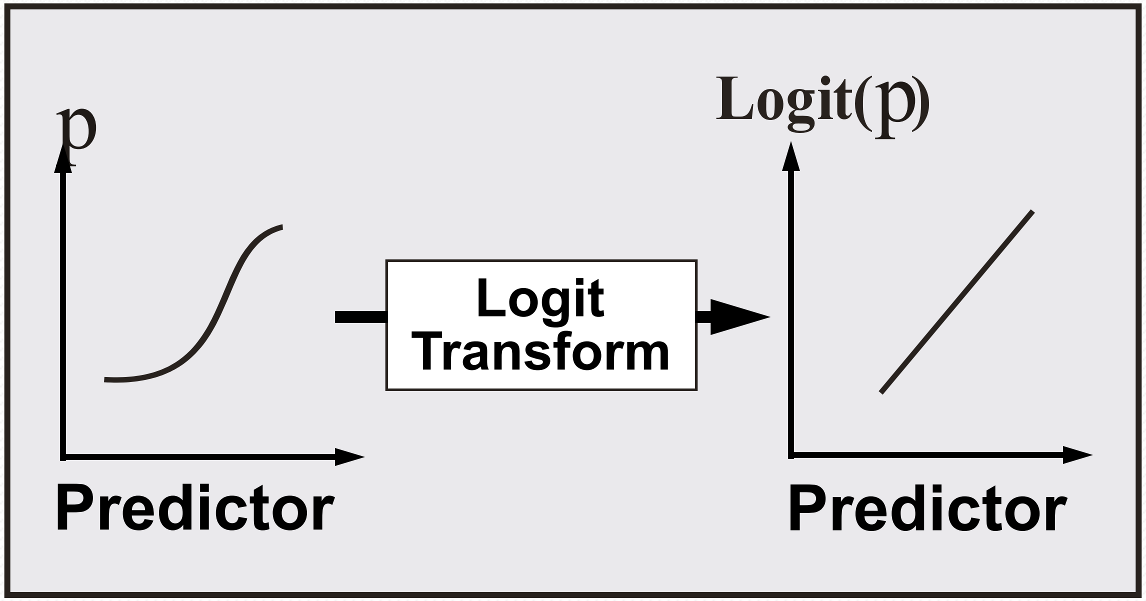
The logistic regression model is given by

$$P(Y \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

which is equivalent to

$$\underbrace{\ln\left(\frac{P(Y \mid X)}{1 - P(Y \mid X)}\right)}_{\text{This is called the Logit Transformation}} = \beta_o + \beta_1 X$$

Logit Transform



Learning Logistic Regression models

- Optimize parameters such that the model gives the best possible reproduction of the training set labels
  - usually done by numerical approximation of maximum likelyhood (vs. Lease Squares for Linear Regression – closed form maximum likelyhood, see Alpaydin Chapter 4)
  - on large data set, may use stochastic gradient descent
- If interested,
  - read Andrew Ng's notes on Logistic Regression, http://cs229.stanford.edu/notes/cs229-notes1.pdf
  - watch his tutorial on Newton's method for training, https://www.youtube.com/watch?v=TuttBDdbls8

Model Evaluation - Classification

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

	PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation…

	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS Class=Yes	a (TP)	b (FN)
ACTUAL CLASS Class=No	c (FP)	d (TN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F-measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)
- Examples to come…

Summary

- Linear Regression and Logistic Regression are nice tools for many simple situations
  - But both force us to fit the data with one shape (line or sigmoid) which will often underfit
- When problem includes more arbitrary non-linearity then we need more powerful models which we will introduce
  - Though non-linear data transformation can help in these cases while still using a linear model for learning.
- These models are commonly used in data mining applications and also as a "first attempt" at understanding data trends, indicators, etc.