

# Simple Linear Regression

# Models

- Representation of some phenomenon
- Mathematical model is a mathematical expression of some phenomenon
- Often describe relationships between variables
- Types
  - Deterministic models
  - Probabilistic models

# Deterministic Models

- Hypothesize exact relationships
- Suitable when prediction error is negligible
- Example: force is exactly mass times acceleration
  - $F = m \cdot a$

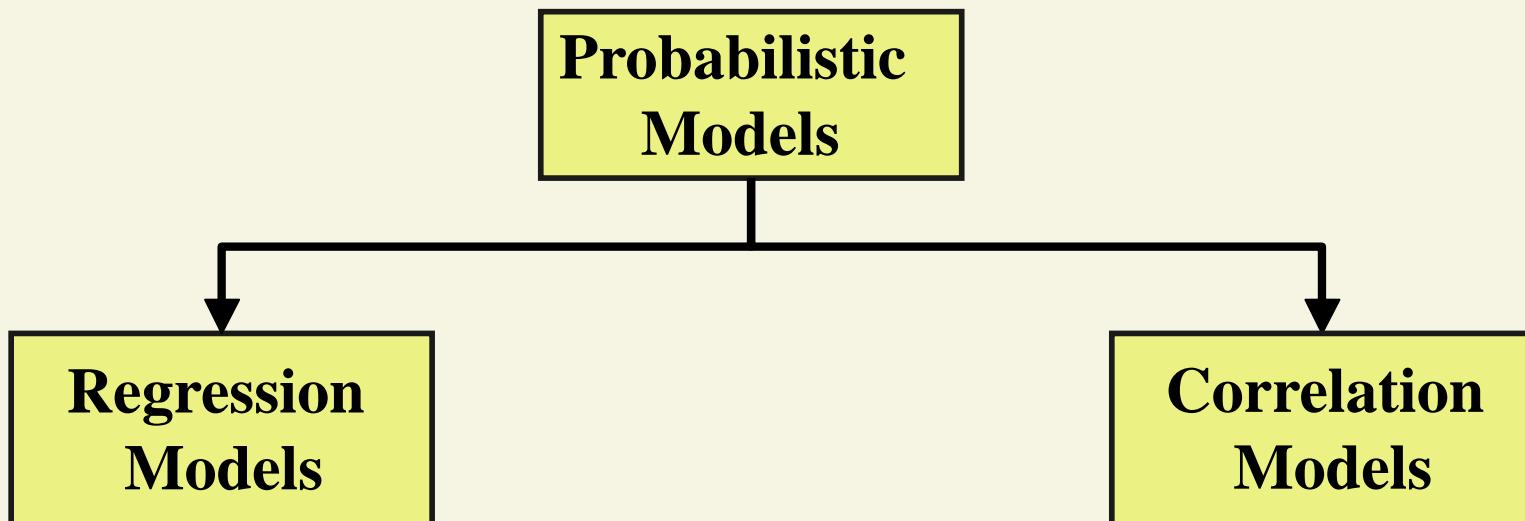


© 1984-1994 T/Maker Co.

# Probabilistic Models

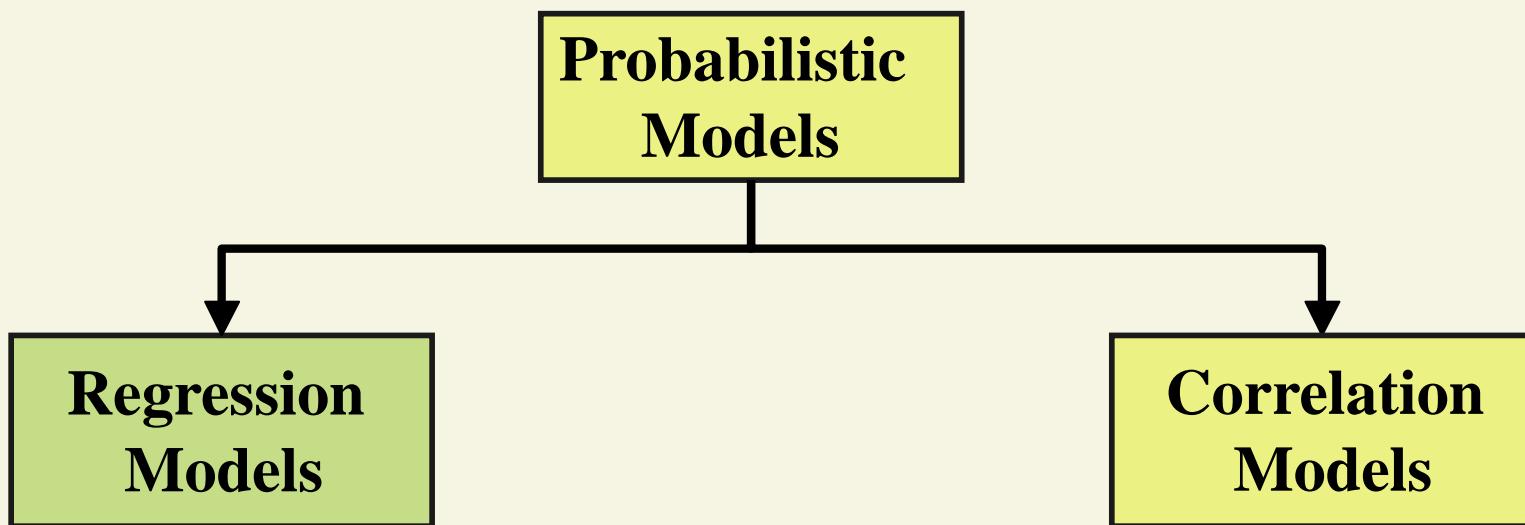
- Hypothesize two components
  - Deterministic
  - Random error
- Example: sales volume ( $y$ ) is 10 times advertising spending ( $x$ ) + random error
  - $y = 10x + \varepsilon$
  - Random error may be due to factors other than advertising

# Types of Probabilistic Models



# **Regression Models**

# Types of Probabilistic Models



# Regression Models

- Answers ‘What is the relationship between the variables?’
- Equation used
  - One numerical dependent (response) variable
    - What is to be predicted
  - One or more numerical or categorical independent (explanatory) variables
- Used mainly for prediction and estimation

# Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
  - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

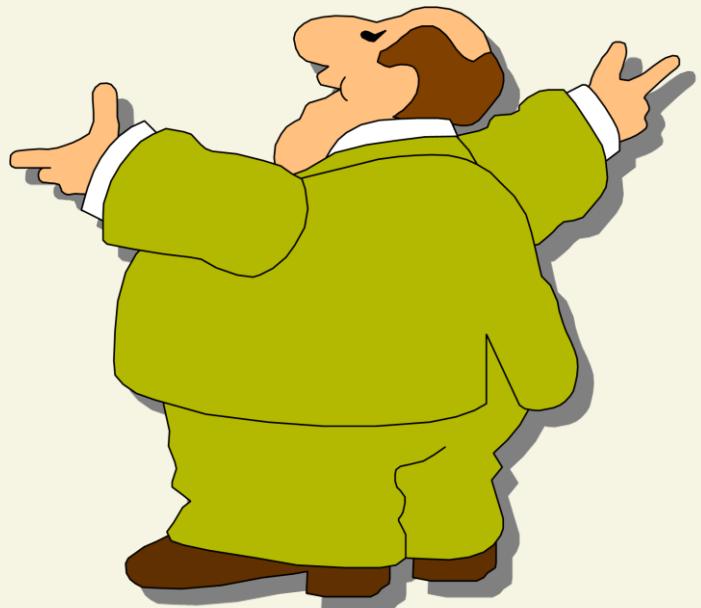
# **Model Specification**

# Regression Modeling Steps

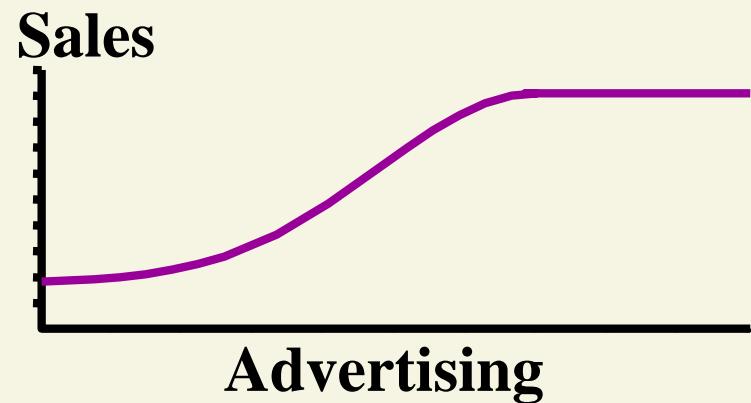
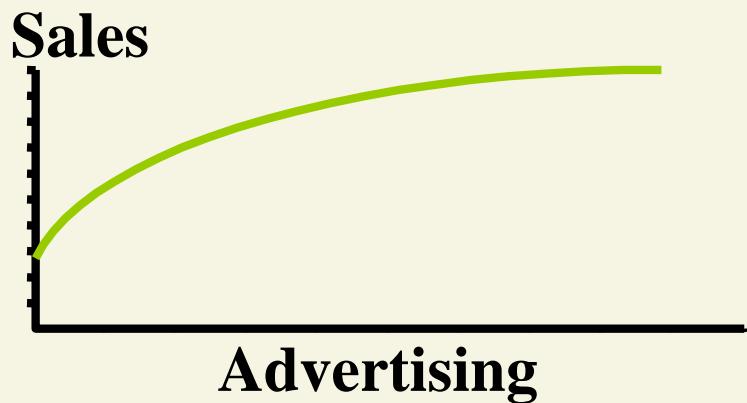
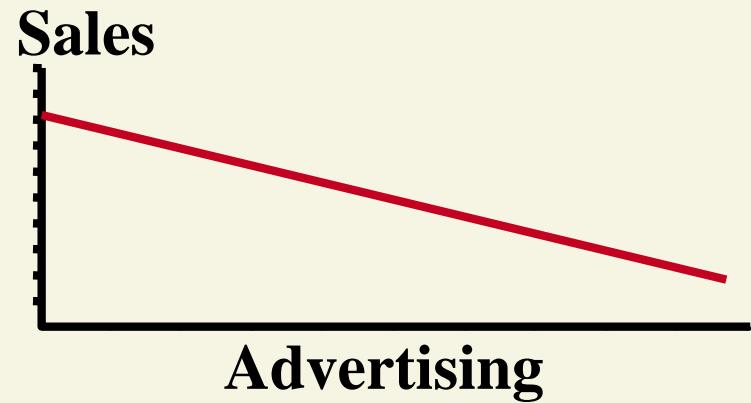
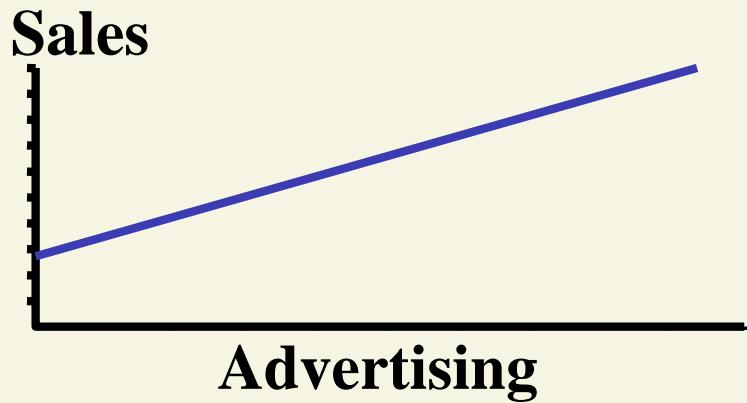
1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
  - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

# Model Specification Is Based on Theory

- Theory of field (e.g., Sociology)
- Mathematical theory
- Previous research
- ‘Common sense’



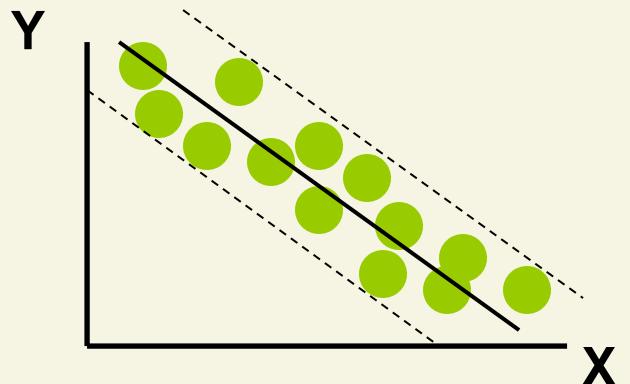
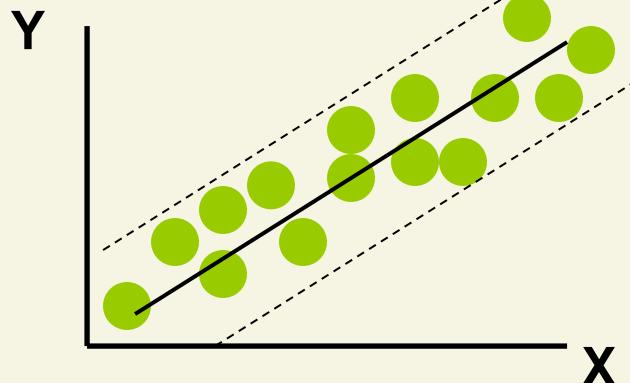
# Thinking Challenge: Which Is More Logical?



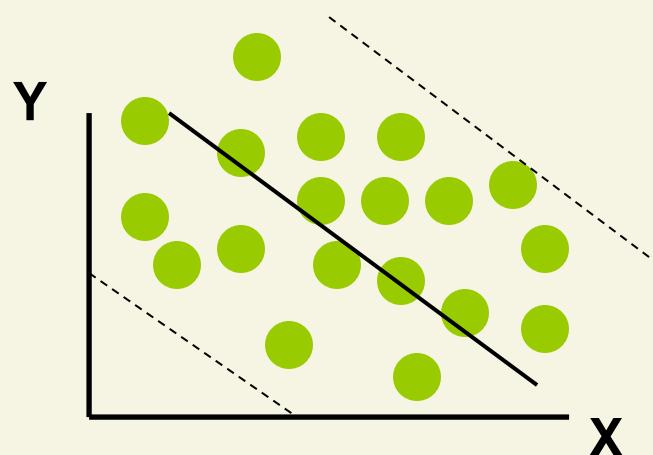
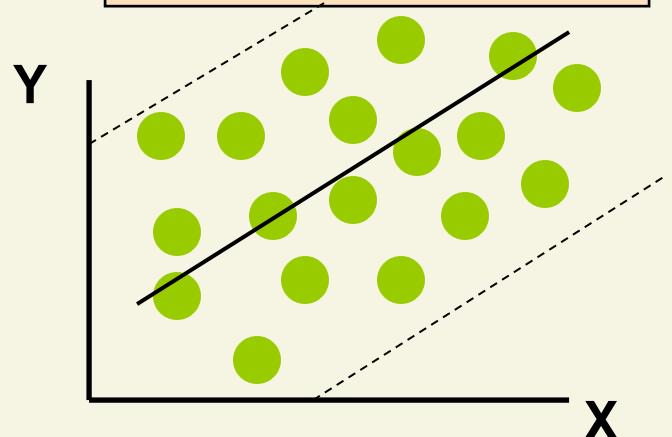
# Types of Relationships

*(continued)*

## Strong relationships



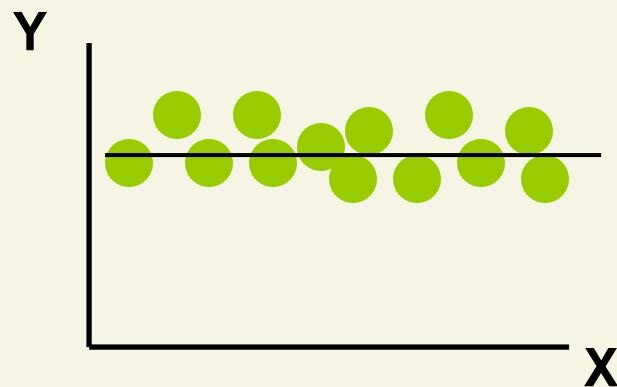
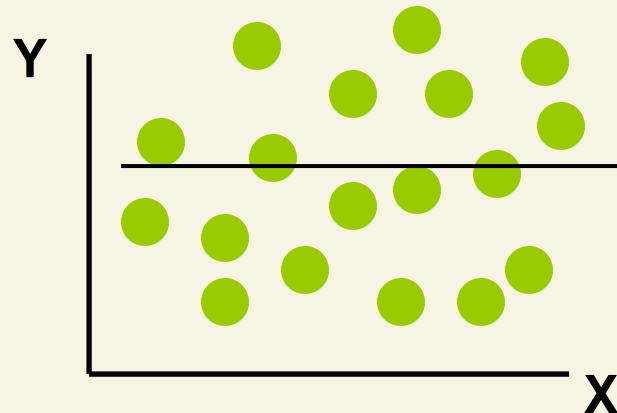
## Weak relationships



# Types of Relationships

*(continued)*

**No relationship**

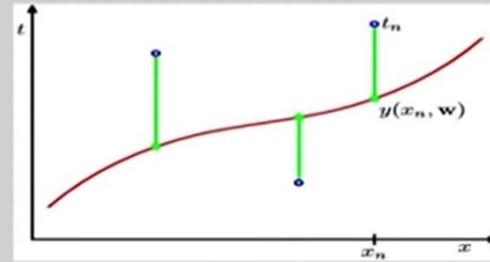
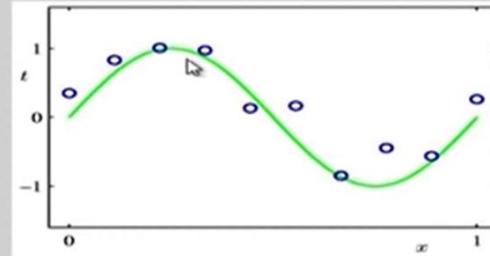




# Example: To measure how good the line is ?

## A Simple Example: Fitting a Polynomial

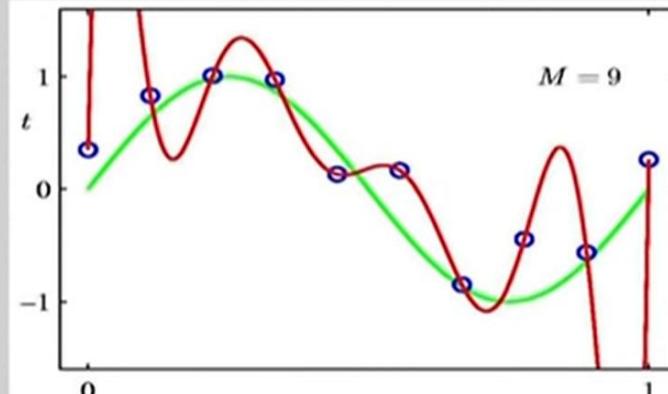
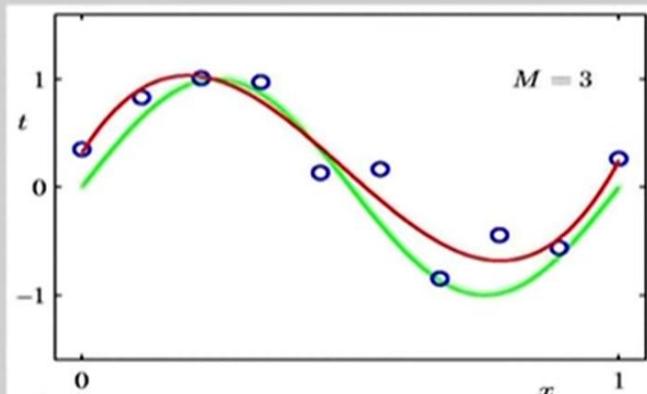
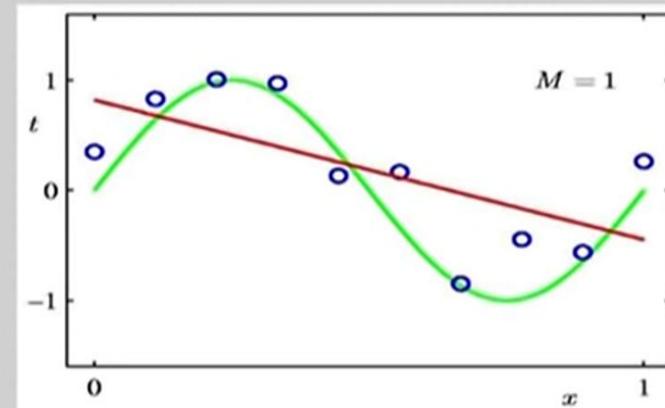
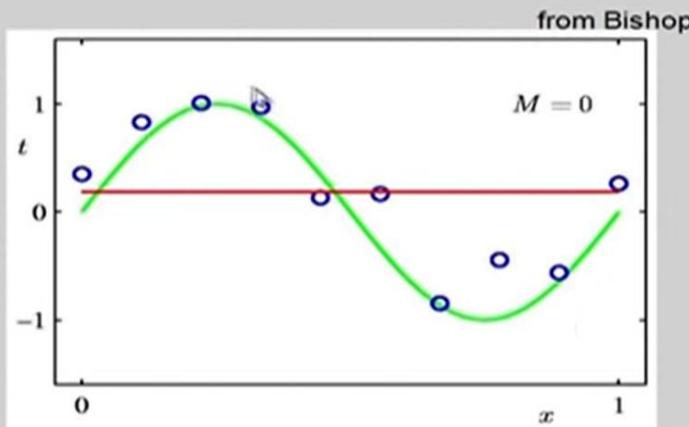
- The green curve is the true function (which is not a polynomial)
- We may use a loss function that measures the squared error in the prediction of  $y(x)$  from  $x$ .



from Bishop's book on Machine Learning

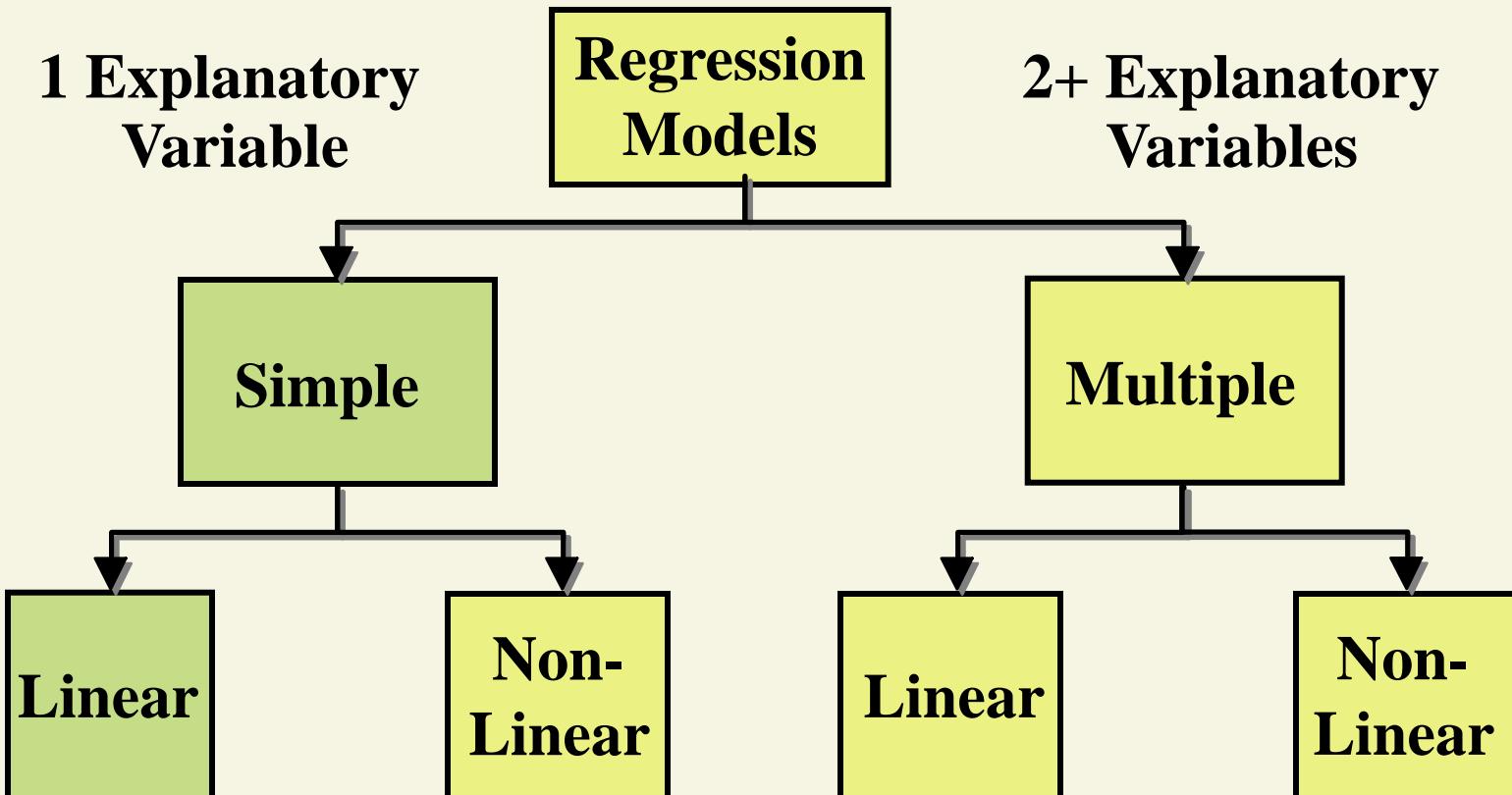
- Sum of Square error is one of the popular measures of the Error.
- $M$ = Degree of polynomial

## Some fits to the data: which is best?



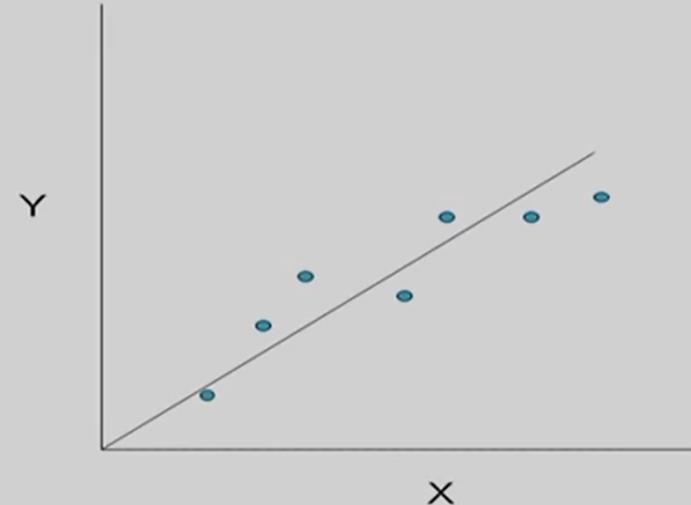
# **Linear Regression Model**

# Types of Regression Models



## Linear regression

- Given an input  $x$  compute an output  $y$
- For example:
  - Predict height from age
  - Predict house price from house area
  - Predict distance from wall from sensors



# Linear Regression Model

Relationship between variables is a linear function

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

Population  
y-intercept

Population  
Slope

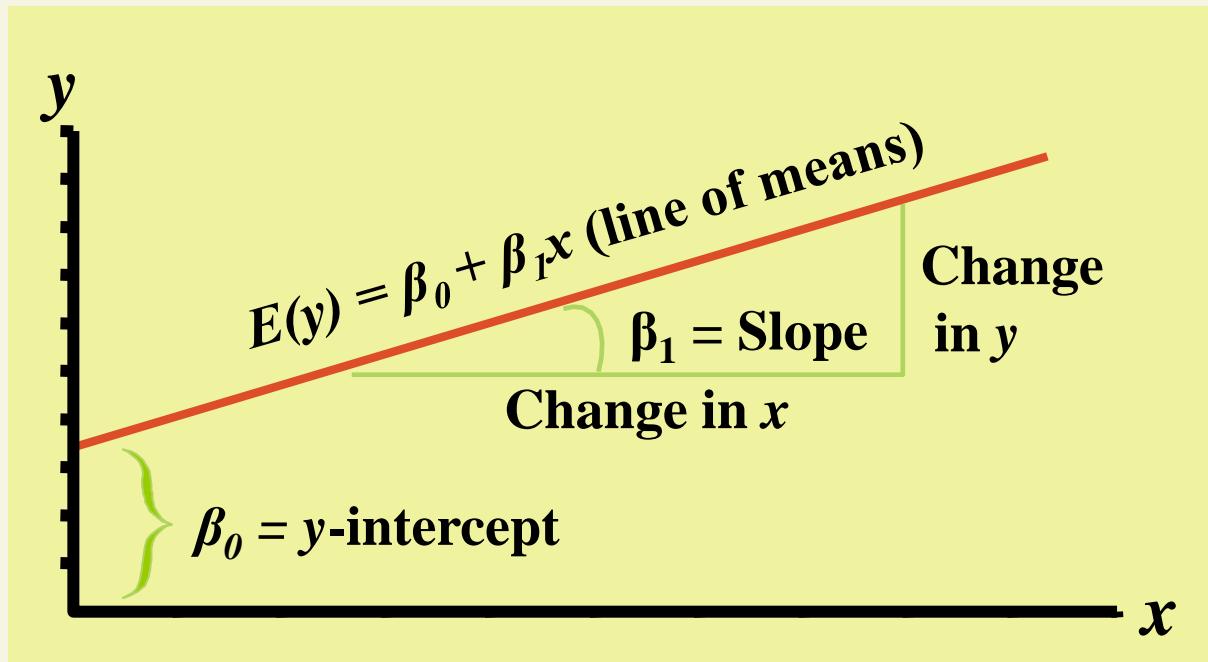
Random  
Error

Dependent  
(Response)  
Variable

Independent  
(Explanatory)  
Variable

```
graph LR; A[Population y-intercept] --> B["y = β₀ + β₁x + ε"]; C[Population Slope] --> B; D[Random Error] --> B; E[Independent (Explanatory) Variable] --> B; F[Dependent (Response) Variable] --> B;
```

# Line of Means



# Example:

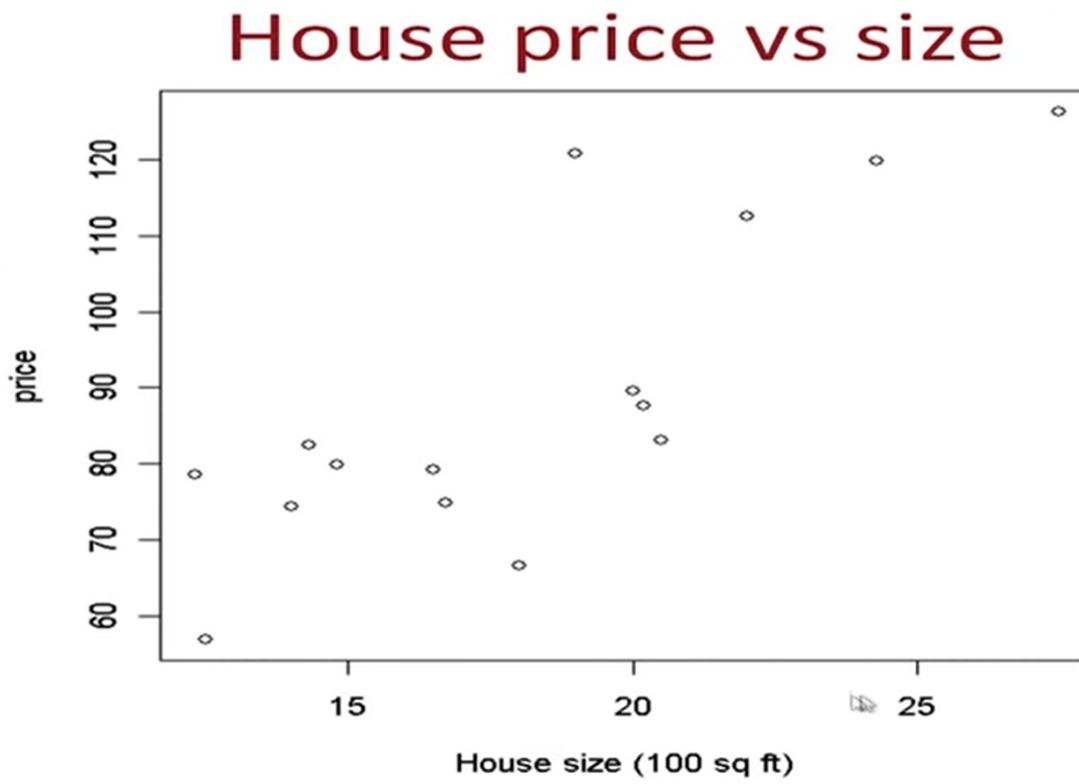
House price prediction given the area of house as input.

House Number	Y: Actual Selling Price	X: House Size (100s ft2)
1	89.5	20.0
2	79.9	14.8
3	83.1	20.5
4	56.9	12.5
5	66.6	18.0
6	82.5	14.3
7	126.3	27.5
8	79.3	16.5
9	119.9	24.3
10	87.6	20.2
11	112.6	22.0
12	120.8	.019
13	78.5	12.3
14	74.3	14.0
15	74.8	16.7
Averages	88.84	18.17

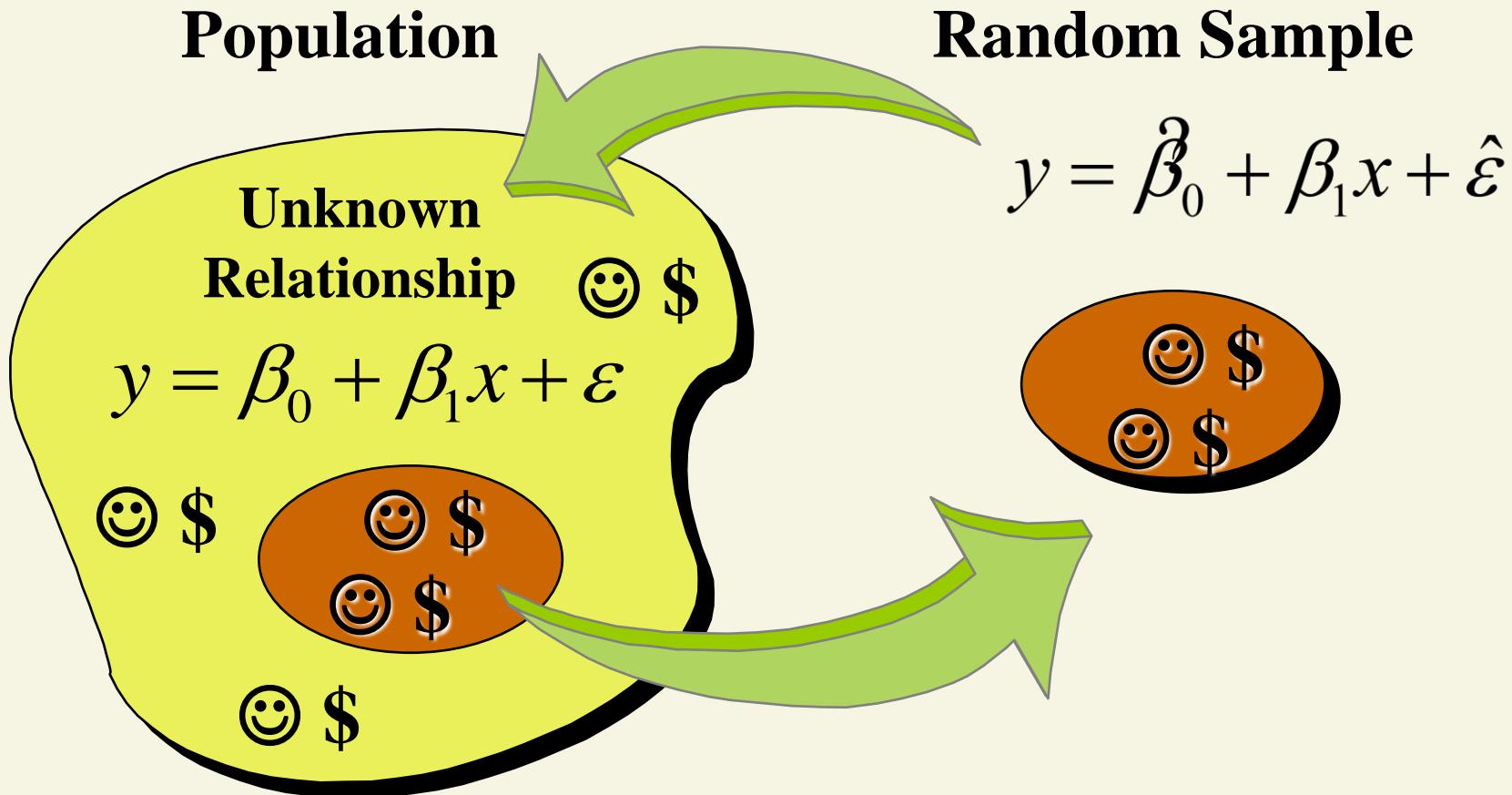
Sample 15 houses from the region.

# Plotting:

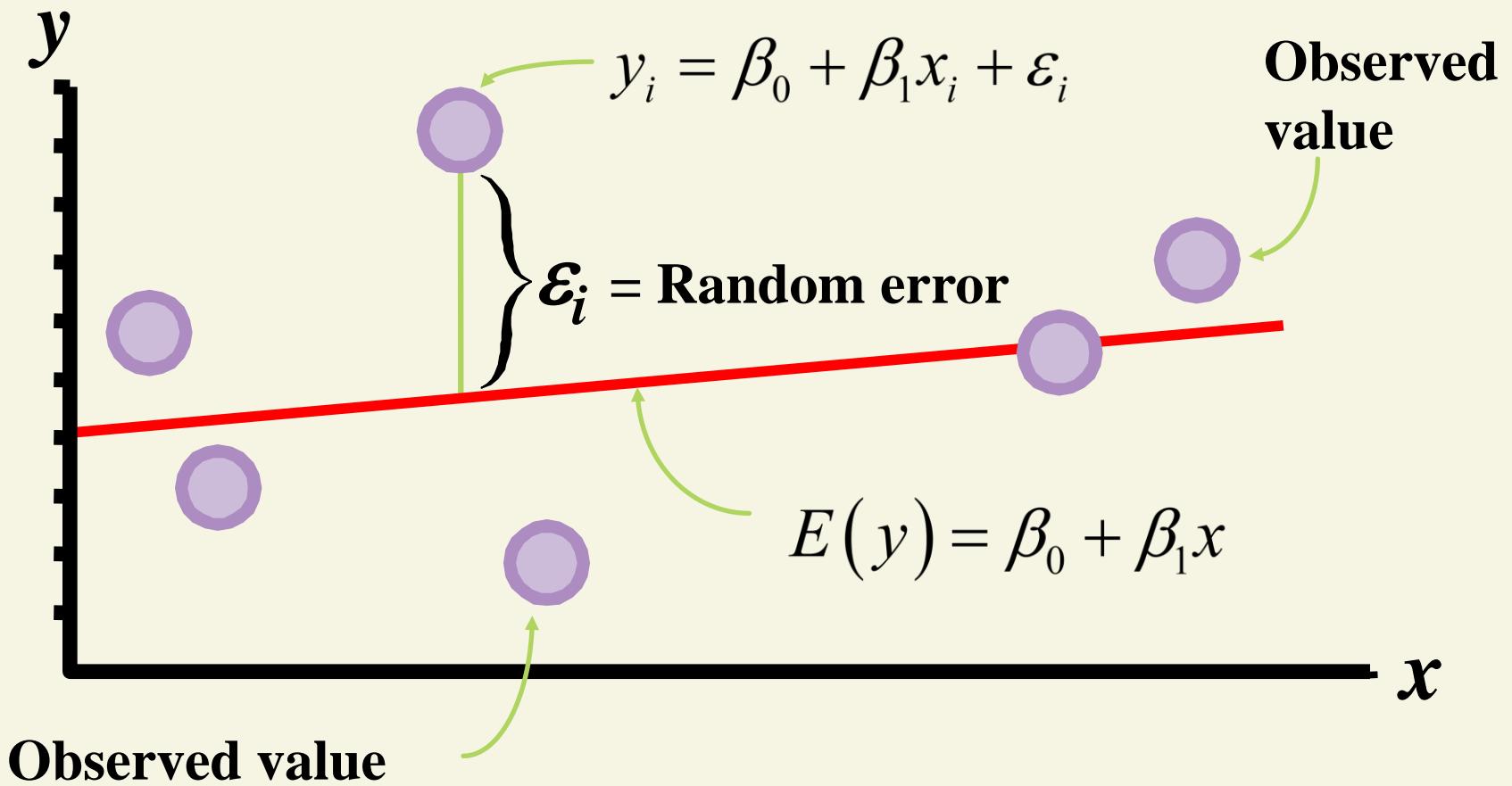
Given these points, to find Eqn of line.



# Population & Sample Regression Models

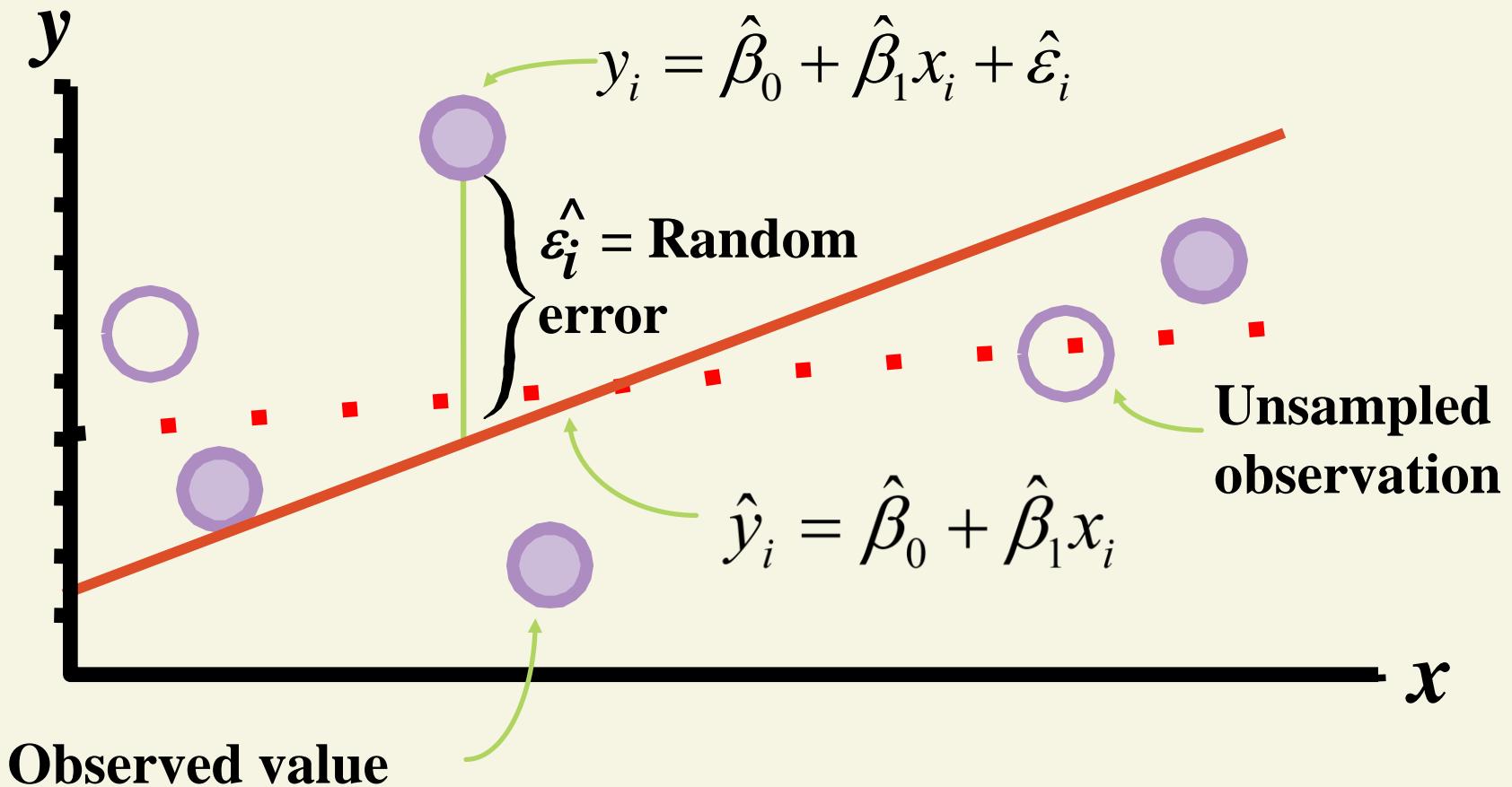


# Population Linear Regression Model



# Sample Linear Regression Model:

$\hat{\beta}$ : Estimated value of  $\beta$



# Multiple Linear Regression

## Linear Regression – Multiple Variables

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- $\beta_0$  is the intercept (i.e. the average value for Y if all the X's are zero),  $\beta_j$  is the slope for the  $j$ th variable  $X_j$

## Assumptions about the Error

- $E(\varepsilon_i) = 0$  for  $i = 1, 2, \dots, n$ .
- $\sigma(\varepsilon_i) = \sigma_\varepsilon$  where  $\sigma_\varepsilon$  is unknown.
- The errors are independent.
- The  $\varepsilon_i$  are normally distributed (with mean 0 and standard deviation  $\sigma_\varepsilon$ ).



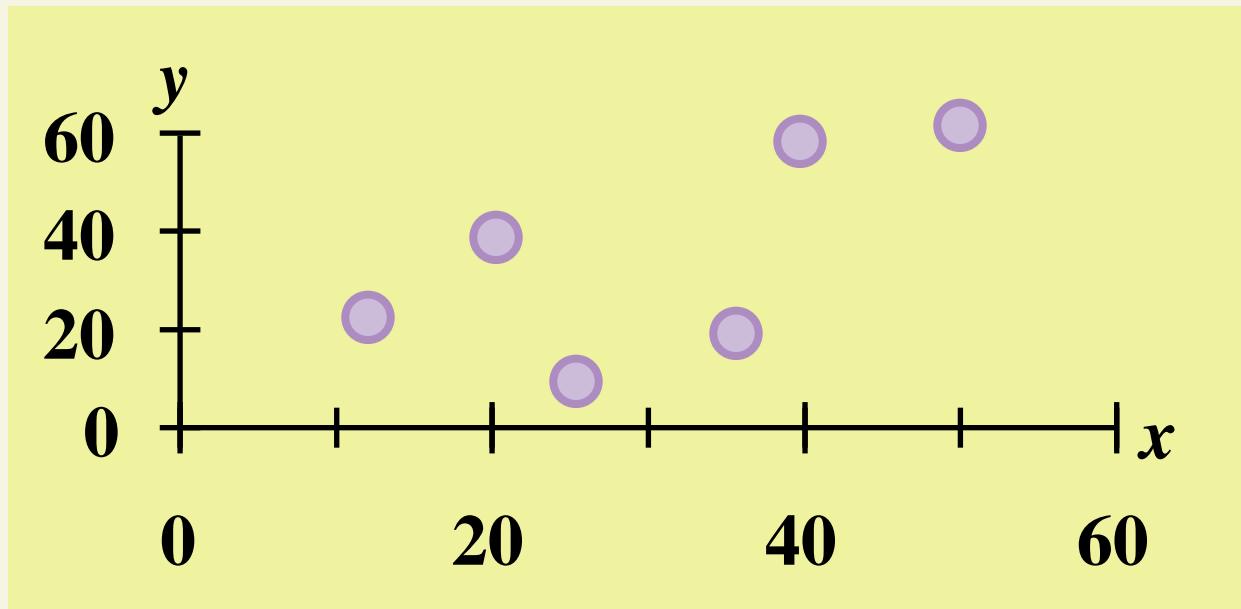
# **Estimating Parameters: Least Squares Method**

# Regression Modeling Steps

1. Hypothesize deterministic component
2. **Estimate unknown model parameters**
3. Specify probability distribution of random error term
  - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

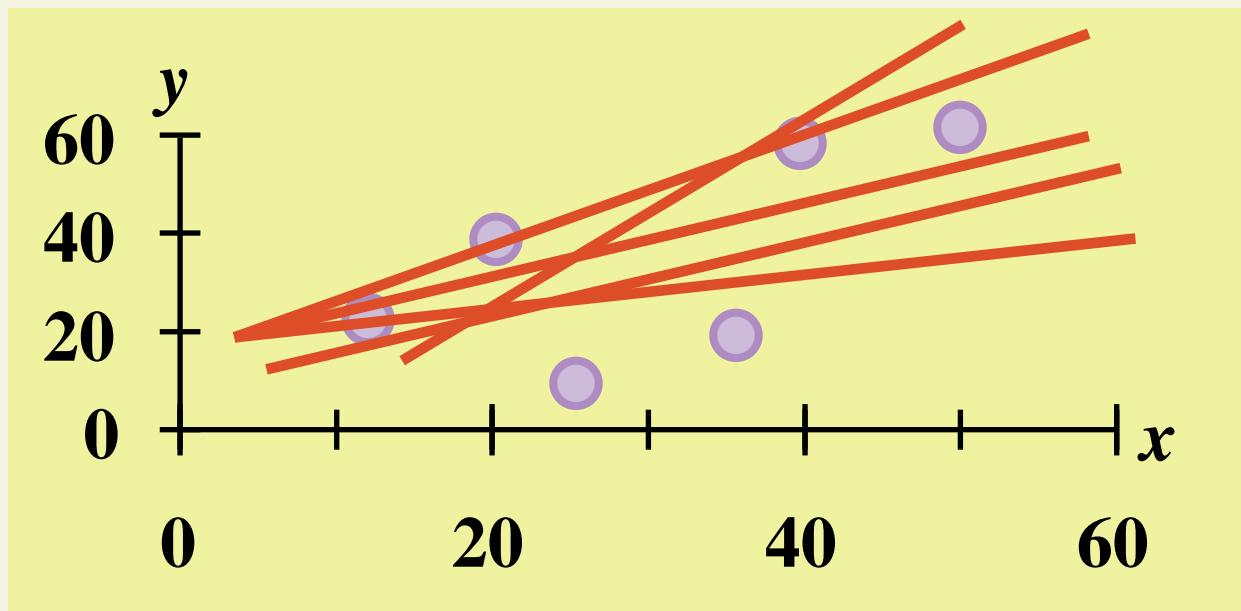
# Scattergram

1. Plot of all  $(x_i, y_i)$  pairs
2. Suggests how well model will fit



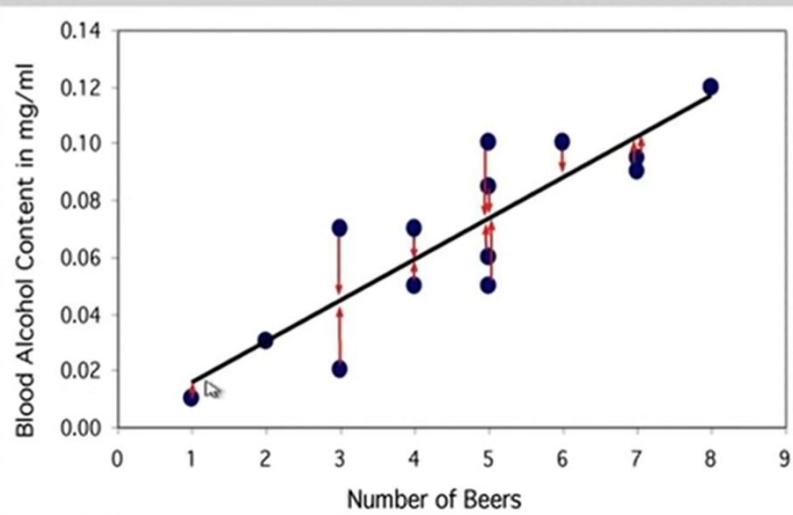
# Thinking Challenge

- How would you draw a line through the points?
- How do you determine which line ‘fits best’?



## The regression line

The least-squares regression line is the unique line such that the sum of the squared vertical ( $y$ ) distances between the data points and the line is the smallest possible.



# Least Squares

- 'Best fit' means difference between actual  $y$  values and predicted  $y$  values are a minimum
  - *But* positive differences off-set negative

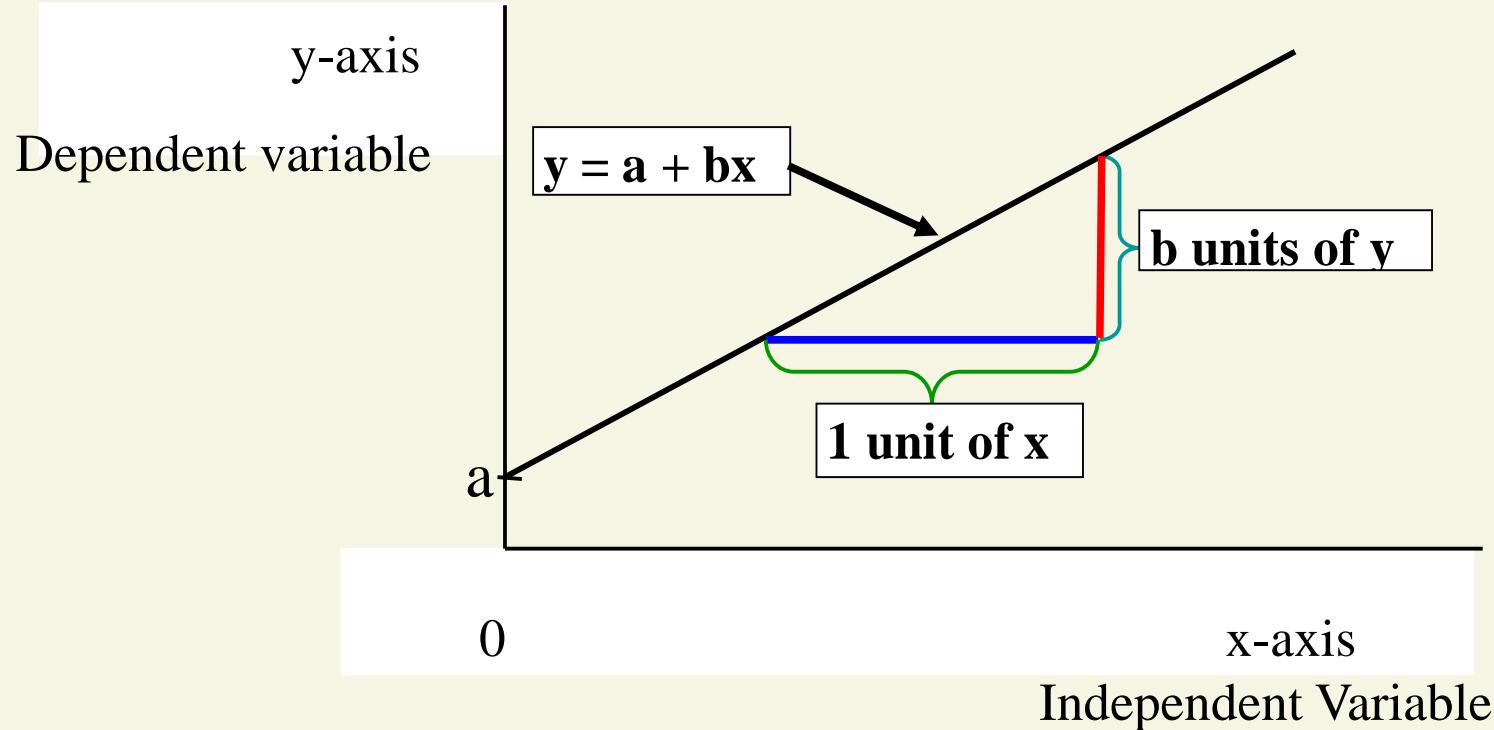
$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

- Least Squares minimizes the Sum of the Squared Differences (SSE)

# Simple Linear Regression Concepts

---

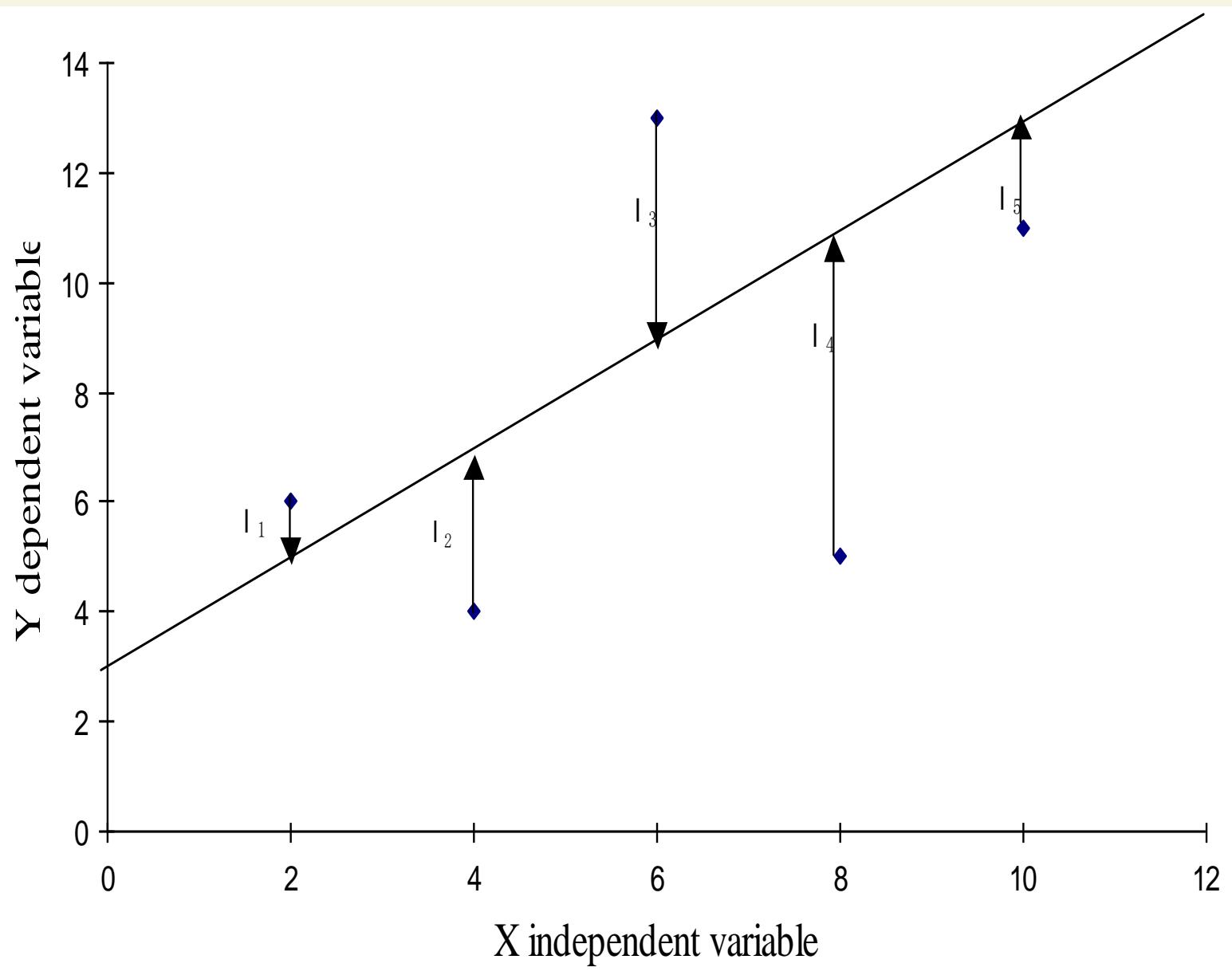
In general, simple linear regression finds the best straight line for describing the relationship between two variables. In its simplest form, which is what we consider here, it does not do a very good job of assessing how well the line describes the data, but nevertheless provides useful information.

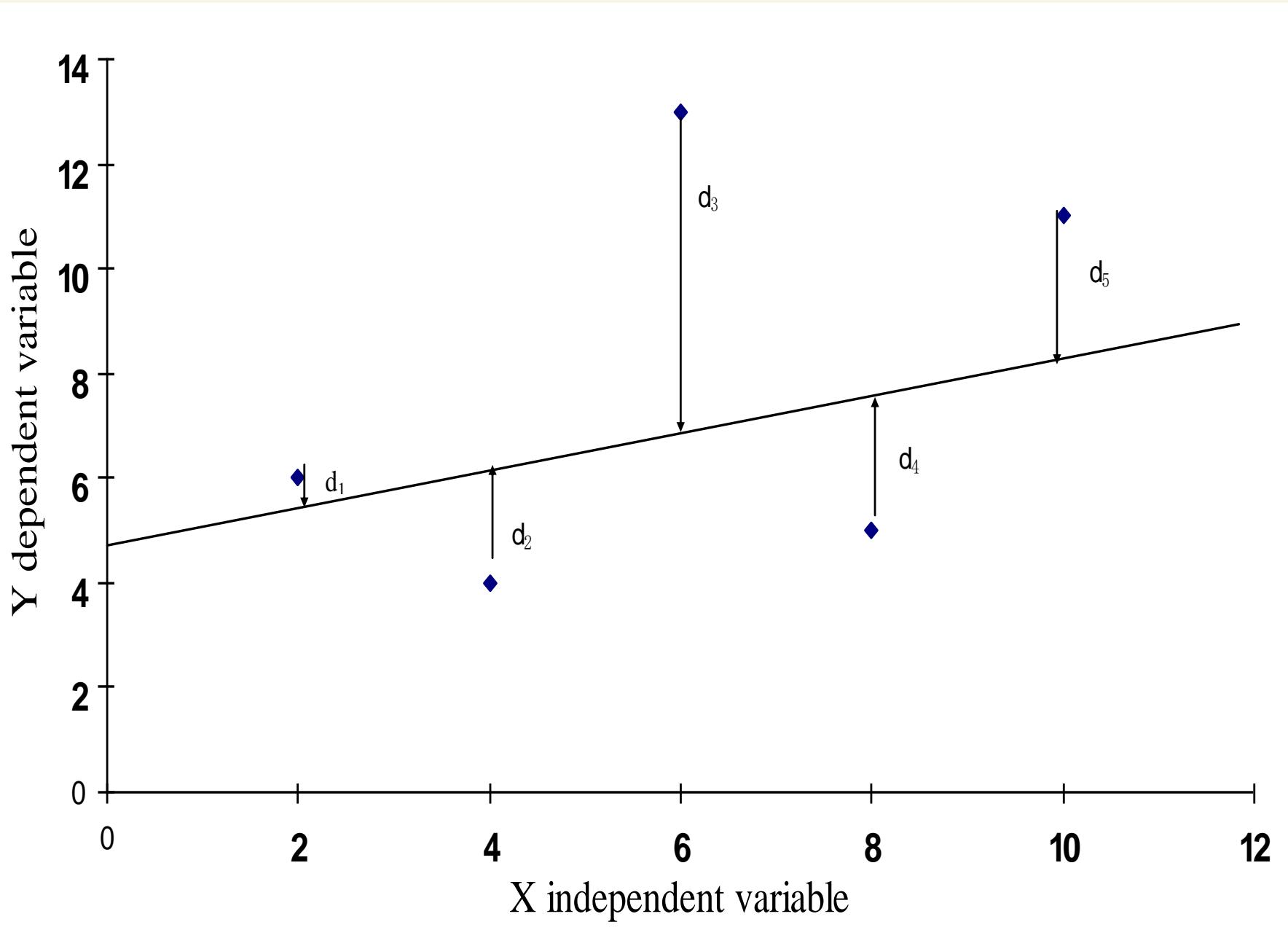


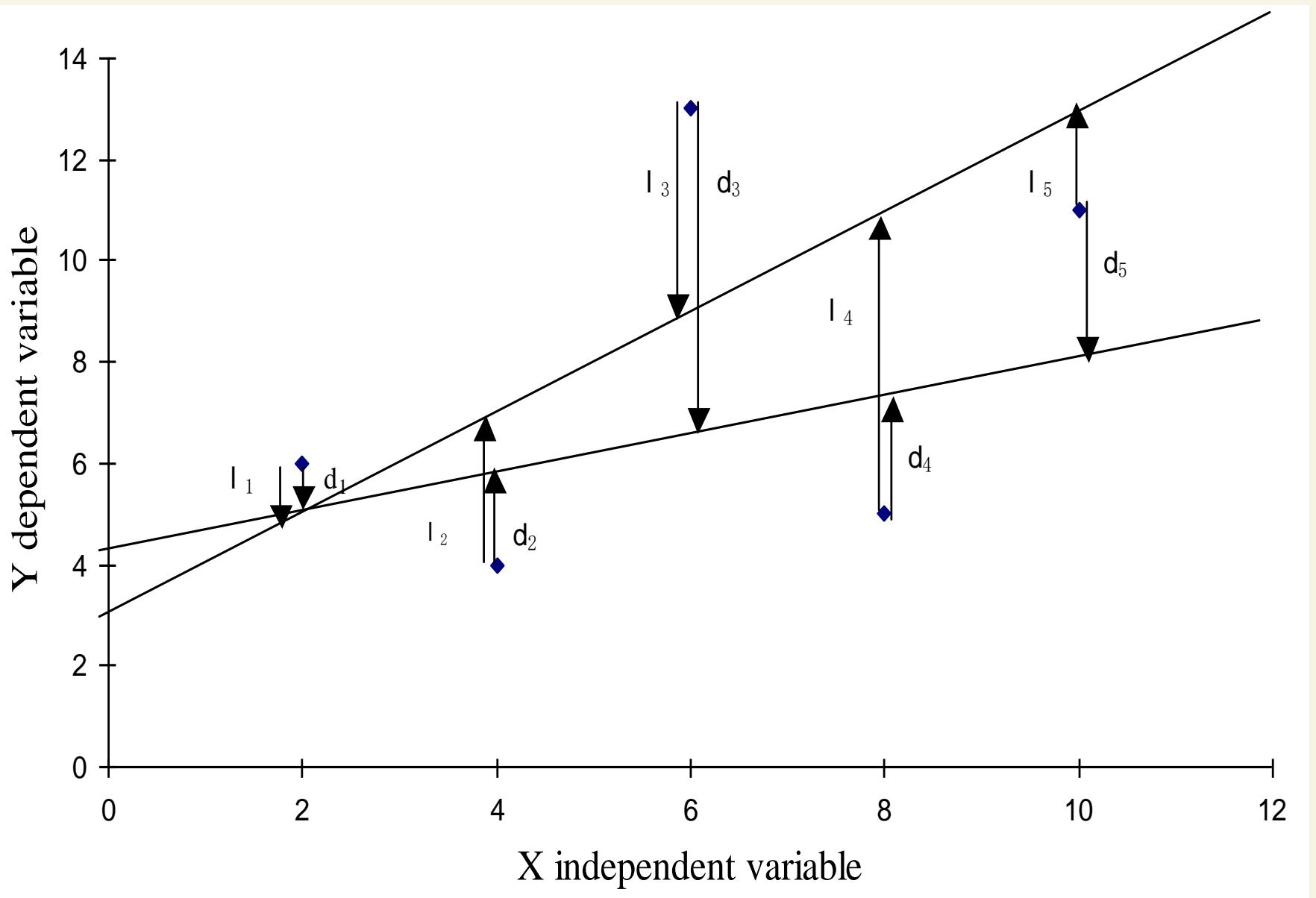
**a** = Intercept, that is, the point where the line crosses the y-axis, which is the value of y at  $x = 0$ .

**b** = Slope of the regression line, that is, the number of units of increase (positive slope) or decrease (negative slope) in y

# The Regression Line







The context for simple linear regression is that we have a random sample of persons from a set of well-defined populations, each defined by a specific value for x-variable. We have measurements of another variable, the y-variable so that we have two variables for each person. For simple linear regression, we focus on a straight line that depicts the relationship between these two variables. The best straight line is the one for which the sum of the squared vertical distances of each point from the line is the least. This "least squares" line has *slope*

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} = \frac{SS(xy)}{SS(x)},$$

and *intercept*  $a = \bar{y} - b\bar{x}$ .

For this situation, the sample line

$$y = a + bx$$

is an estimate of the population line

and a and b are estimates of  $\alpha$  and  $\beta$  respectively. For a specific value of  $x$ , such as  $x = 10$ , the value for  $y$  calculated from the regression equation is

# Simple Regression Example

---

The following data are diastolic blood pressure (DBP) measurements taken at different times after an intervention for  $n = 5$  persons. For each person, the data available include the time of the measurement and the DBP level. Of interest is the relationship between these two variables.

Patient	Time		DPB		
	x	x <sup>2</sup>	y	y <sup>2</sup>	xy
1	0	0	72	5,184	0
2	5	25	66	4,356	330
3	10	100	70	4,900	700
4	15	225	64	4,096	960
5	20	400	66	4,356	1,320
Sum	50	750	338	22,892	3,310
Mean	10		67.6		
n	5		5		

For the blood pressure data,

$$\bar{x} = 50/5 = 10,$$

$$\bar{y} = 338/5 = 67.6,$$

the slope is

$$b = \frac{\sum xy - \sum x \sum y/n}{\sum x^2 - (\sum x)^2/n} = \frac{SS(xy)}{SS(x)},$$

$$b = \frac{3,310 - (50)(338)/5}{750 - (50)^2/5} = -0.28$$

and the intercept is

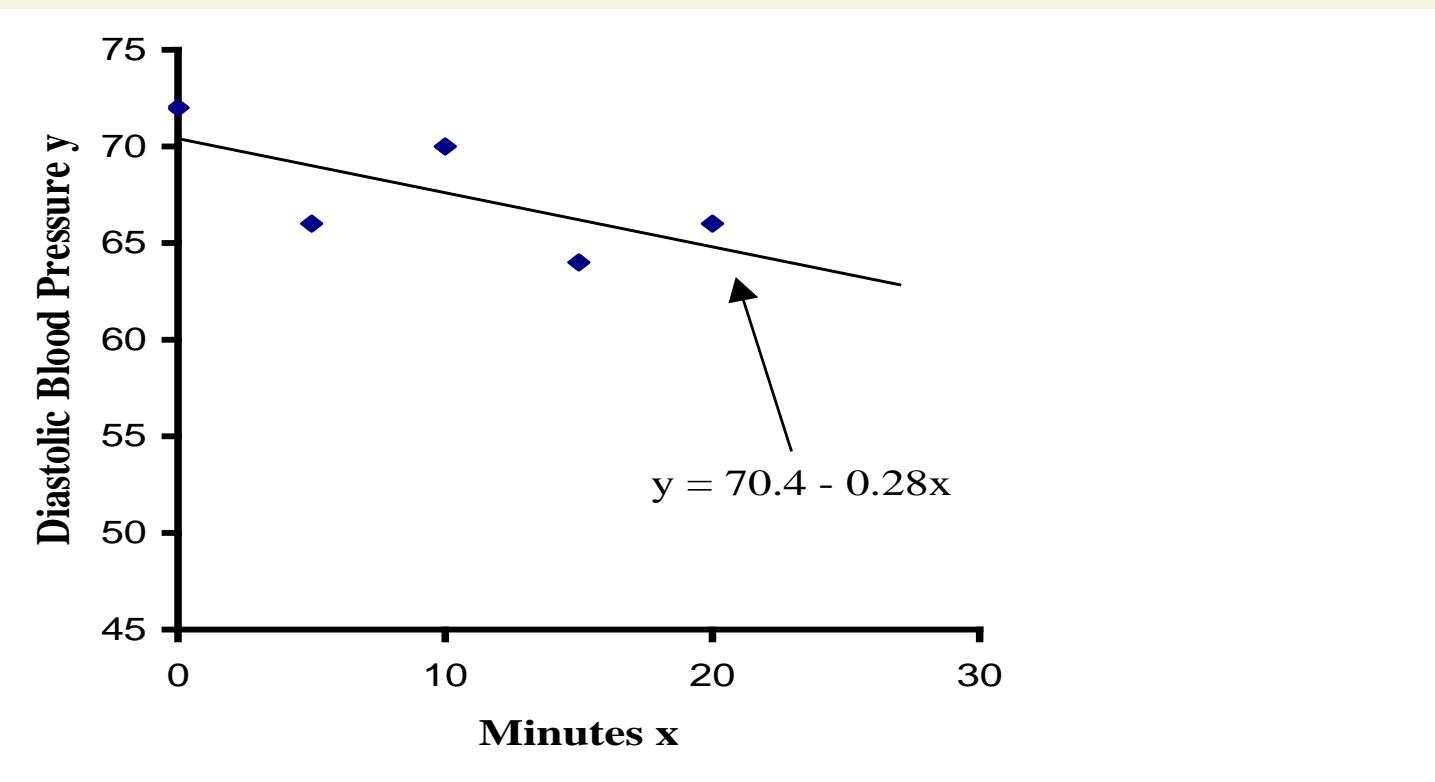
$$a = \bar{y} - b\bar{x},$$

$$a = 67.6 - (-0.28)10 = 70.4$$

The best line is

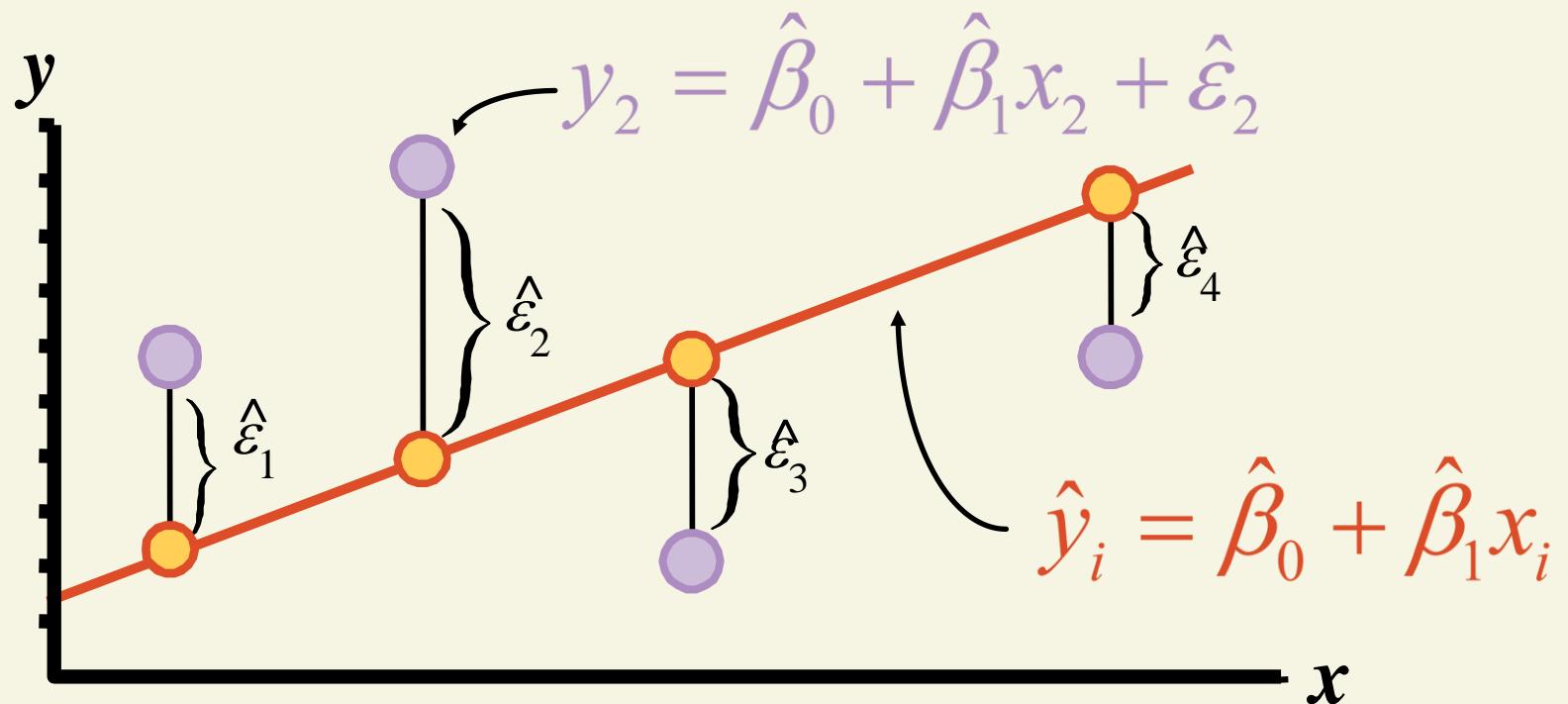
$$y = a + bx = 70.4 - 0.28x$$

Patient	Time x	DBP y
1	0	72
2	5	66
3	10	70
4	15	64
5	20	66



# Least Squares Graphically

LS minimizes  $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



# Coefficient Equations

**Prediction Equation**  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

**Slope** 
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n}$$
$$\sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n}$$

**y-intercept** 
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Computation Table

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
$x_1$	$y_1$	$x_1^2$	$y_1^2$	$x_1 y_1$
$x_2$	$y_2$	$x_2^2$	$y_2^2$	$x_2 y_2$
:	:	:	:	:
$x_n$	$y_n$	$x_n^2$	$y_n^2$	$x_n y_n$
$\Sigma x_i$	$\Sigma y_i$	$\Sigma x_i^2$	$\Sigma y_i^2$	$\Sigma x_i y_i$

# Interpretation of Coefficients

## 1. Slope ( $\hat{\beta}_1$ )

- Estimated  $y$  changes by  $\hat{\beta}_1$  for each 1unit increase in  $x$ 
  - If  $\hat{\beta}_1 = 2$ , then Sales ( $y$ ) is expected to increase by 2 for each 1 unit increase in Advertising ( $x$ )

## 2. Y-Intercept ( $\hat{\beta}_0$ )

- Average value of  $y$  when  $x = 0$ 
  - If  $\hat{\beta}_0 = 4$ , then Average Sales ( $y$ ) is expected to be 4 when Advertising ( $x$ ) is 0

# Least Squares Example

You're a marketing analyst for Hasbro Toys.  
You gather the following data:

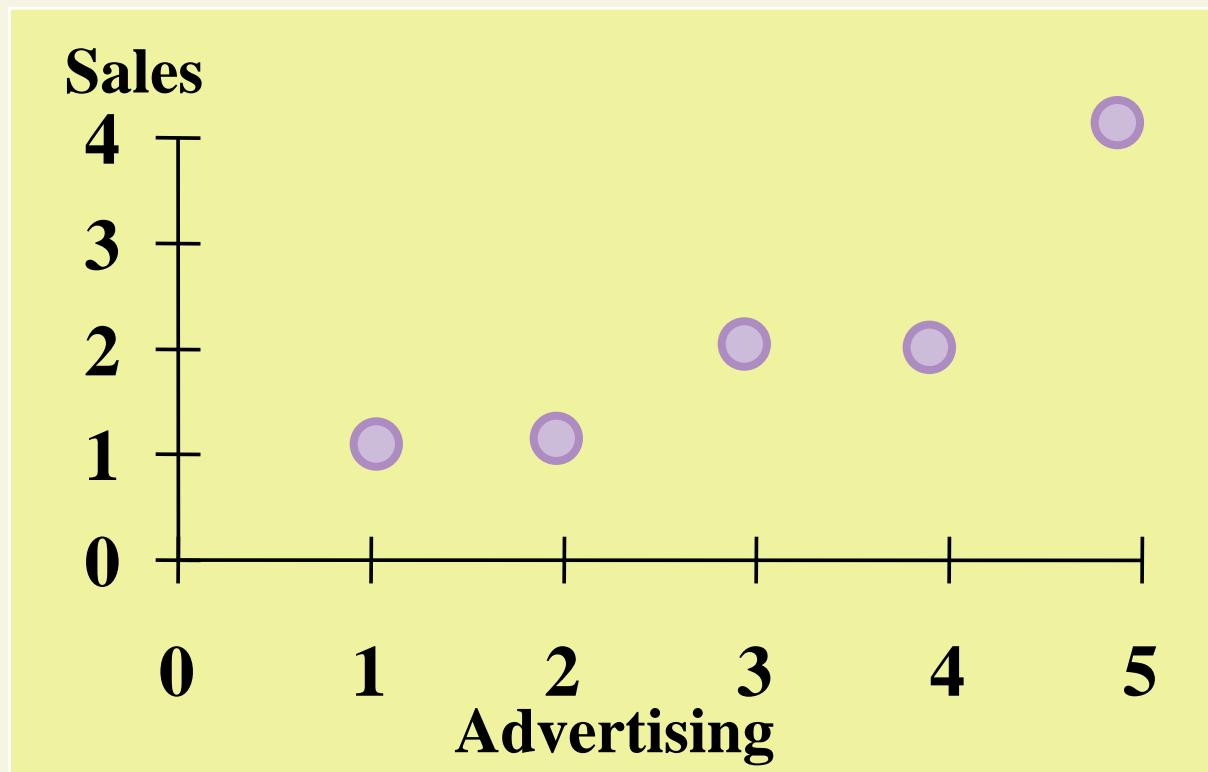
<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

Find the **least squares line** relating sales and advertising.



# Scattergram

## Sales vs. Advertising



# Parameter Estimation

## Solution Table

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

# Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = .70$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = 2 - (.70)(3) = -.10$$

$$\hat{y} = -.1 + .7x$$

# Parameter Estimation

## Computer Output

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Param=0	Prob> T
INTERCEP	1	-0.1000	0.6350	-0.157	0.8849
ADVERT	1	0.7000	0.1914	3.656	0.0354

$$\hat{y} = -.1 + .7x$$

# Coefficient Interpretation Solution

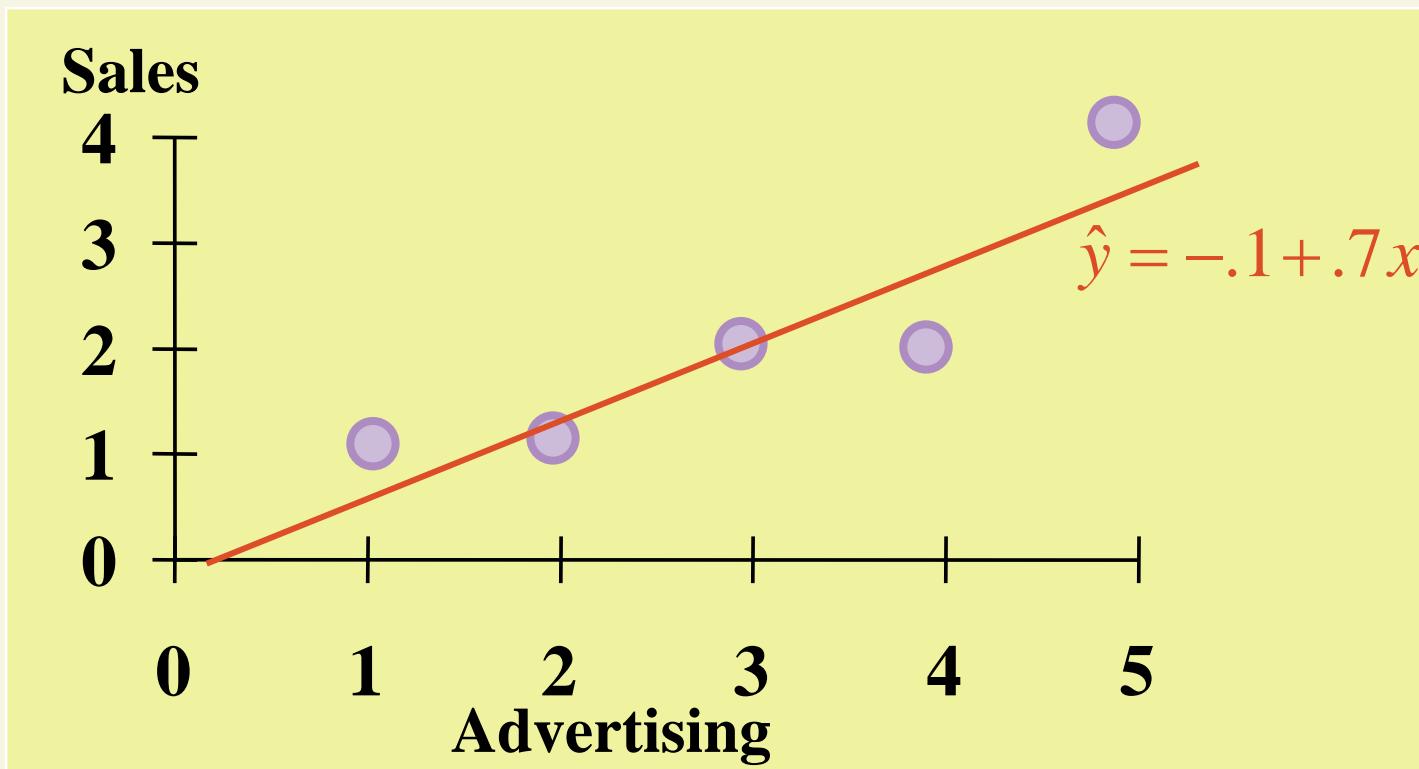
## 1. Slope ( $\hat{\beta}_1$ )

- Sales Volume ( $y$ ) is expected to increase by .7 units for each \$1 increase in Advertising ( $x$ )

## 2. $Y$ -Intercept ( $\hat{\beta}_0$ )

- Average value of Sales Volume ( $y$ ) is -.10 units when Advertising ( $x$ ) is 0
  - Difficult to explain to marketing manager
  - Expect some sales without advertising

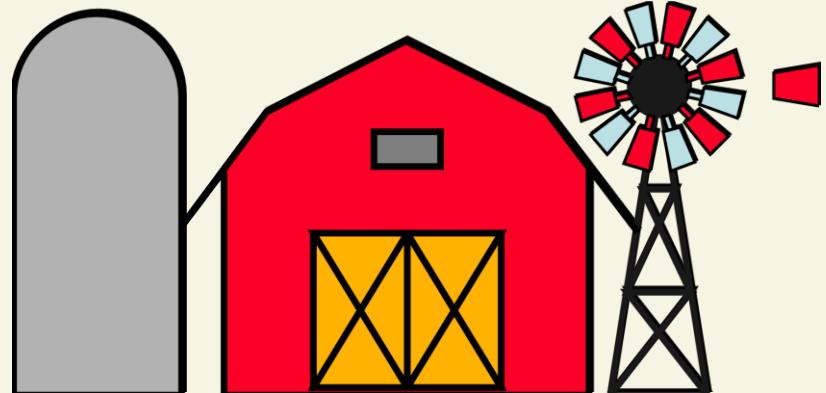
# Regression Line Fitted to the Data



# Least Squares Thinking Challenge

You're an economist for the county cooperative.  
You gather the following data:

<u>Fertilizer (lb.)</u>	<u>Yield (lb.)</u>
4	3.0
6	5.5
10	6.5
12	9.0

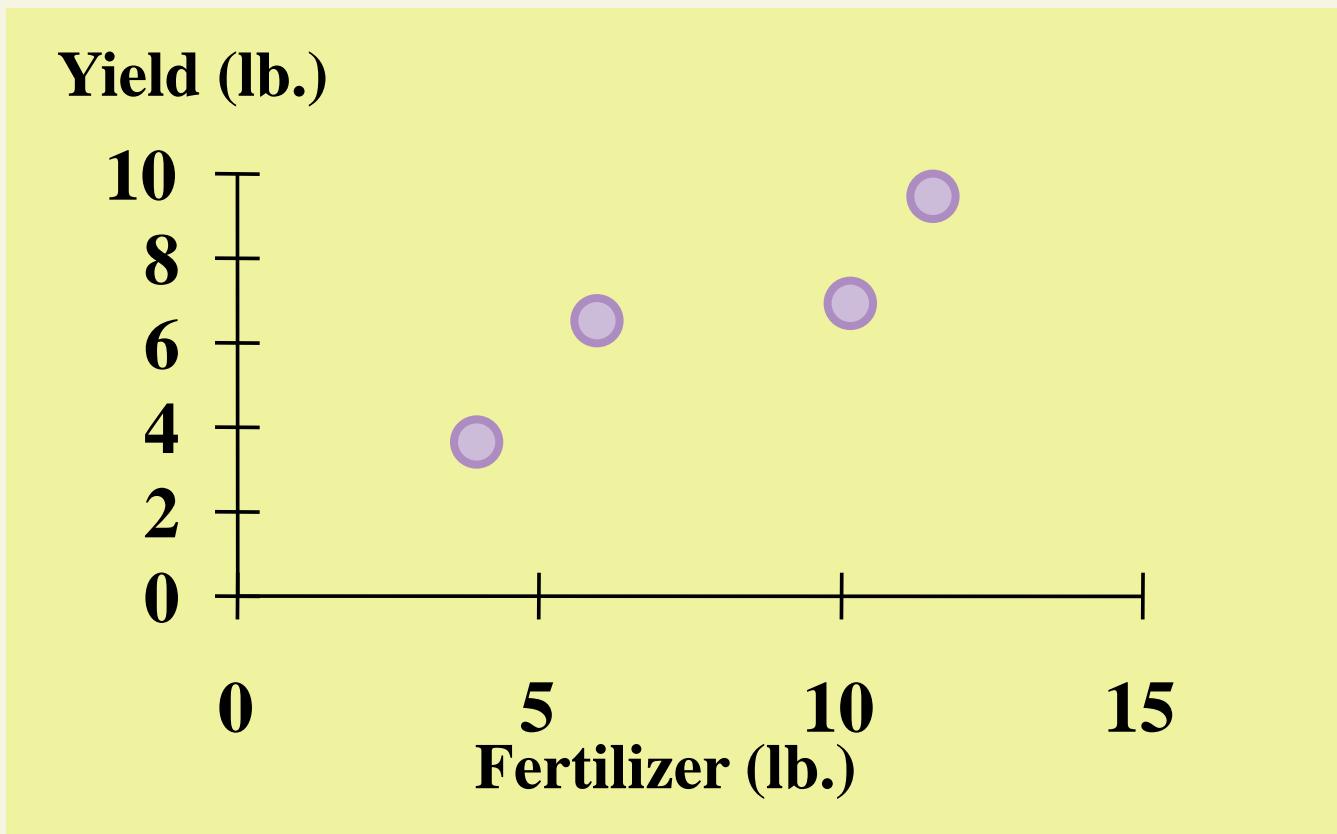


Find the **least squares line** relating crop yield and fertilizer.

© 1984-1994 T/Maker Co.

# Scattergram

## Crop Yield vs. Fertilizer\*



# Parameter Estimation

## Solution Table\*

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
4	3.0	16	9.00	12
6	5.5	36	30.25	33
10	6.5	100	42.25	65
12	9.0	144	81.00	108
32	24.0	296	162.50	218

# Parameter Estimation Solution\*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = .65$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 6 - (.65)(8) = .80$$

$$\hat{y} = .8 + .65x$$

# Coefficient Interpretation Solution\*

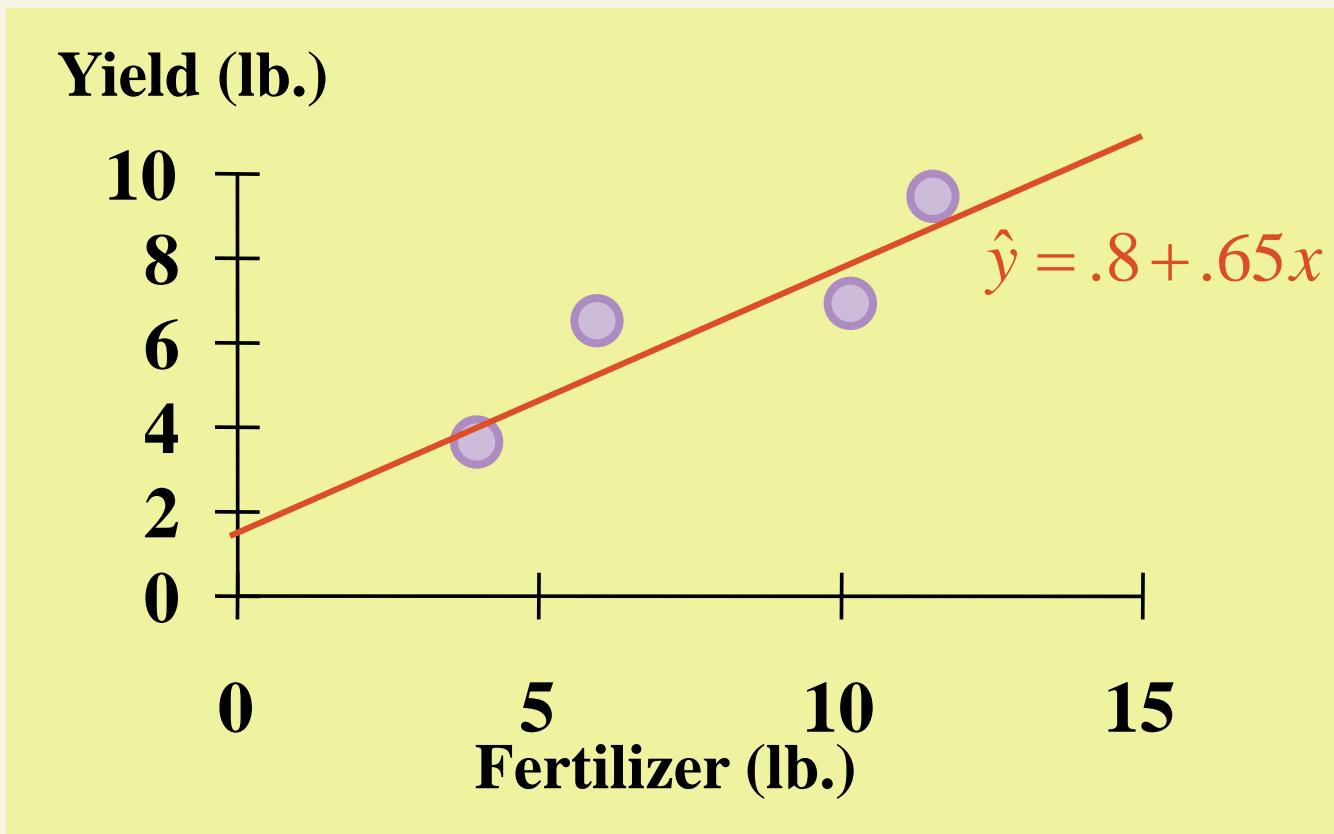
## 1. Slope ( $\hat{\beta}_1$ )

- Crop Yield ( $y$ ) is expected to increase by .65 lb. for each 1 lb. increase in Fertilizer ( $x$ )

## 2. Y-Intercept ( $\hat{\beta}_0$ )

- Average Crop Yield ( $y$ ) is expected to be 0.8 lb. when no Fertilizer ( $x$ ) is used

# Regression Line Fitted to the Data\*





# How do we "learn" parameters

- For the 2-d problem

$$Y = \beta_0 + \beta_1 X$$

- To find the values for the coefficients which minimize the objective function we take the partial derivates of the objective function (SSE) with respect to the coefficients. Set these to 0, and solve.

$$\beta_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$



# LMS Method or Delta Method

## Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

$$h(x) = \sum_{i=0}^n \beta_i x_i$$

- There is a closed form which requires matrix inversion, etc.
- There are iterative techniques to find weights
  - delta rule (also called LMS method) which will update towards the objective of minimizing the SSE.

# Linear Regression

$$h(x) = \sum_{i=0}^n \beta_i x_i$$

To learn the parameters  $\theta$  ( $\beta_i$ ) ?

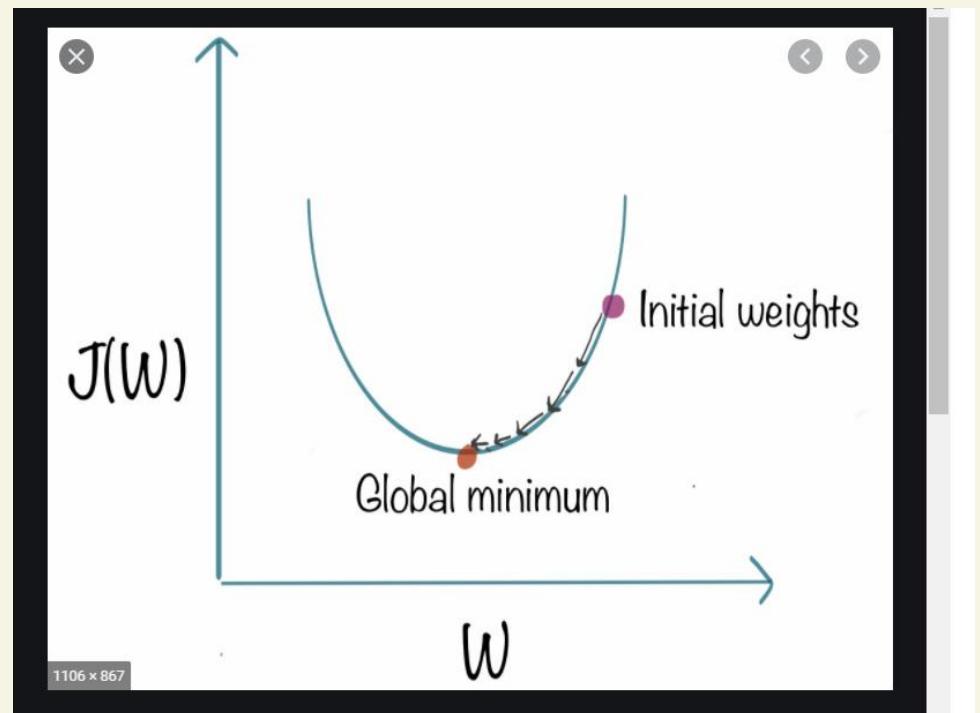
- Make  $h(x)$  close to  $y$ , for the available *training examples*.
- Define a cost function  $J(\theta)$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h(x)^{(i)} - (y)^{(i)})^2$$

- Find  $\theta$  that minimizes  $J(\theta)$ .

# Gradient Descent Method

- Here  $J(\theta)$  is quadratic and hence is a convex function, So it has a single optimal (minimal) solution.
- Here we use Gradient Descent Method



Here  $\alpha$ = step size

Press  to exit full screen

## LMS Algorithm

- Start a search algorithm (e.g. gradient descent algorithm,) with initial guess of  $\theta$ .
- Repeatedly update  $\theta$  to make  $J(\theta)$  smaller, until it converges to minima.

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\theta)$$

- $J$  is a convex quadratic function, so has a single global minima. gradient descent eventually converges at the global minima.
- At each iteration this algorithm takes a step in the direction of steepest descent(-ve direction of gradient).

## LMS Update Rule

- If you have only one training example  $(x, y)$

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h(x) - y)^2 \\ &= 2 \cdot \frac{1}{2} (h(x) - y) \frac{\partial}{\partial \theta_j} (h(x) - y) \\ &= (h(x) - y) \cdot \frac{\partial}{\partial \theta_j} \left( \sum_{i=0}^n \theta_i x_i - y \right) \\ &= (h(x) - y)x_j\end{aligned}$$

- For a single training example, this gives the update rule:

$$\beta_j = \beta_j + \alpha(y^{(i)} - h(x^i))x_j^{(i)}$$

## Linear Regression

$$h(x) = \sum_{i=0}^n \beta_i x_i$$

To learn the parameters  $\theta$  ( $\beta_i$ ) ?

- Make  $h(x)$  close to  $y$ , for the available *training examples*.
- Define a cost function  $J(\theta)$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h(x)^{(i)} - (y)^{(i)})^2$$

- Find  $\theta$  that minimizes  $J(\theta)$ .

# **Probability Distribution of Random Error**

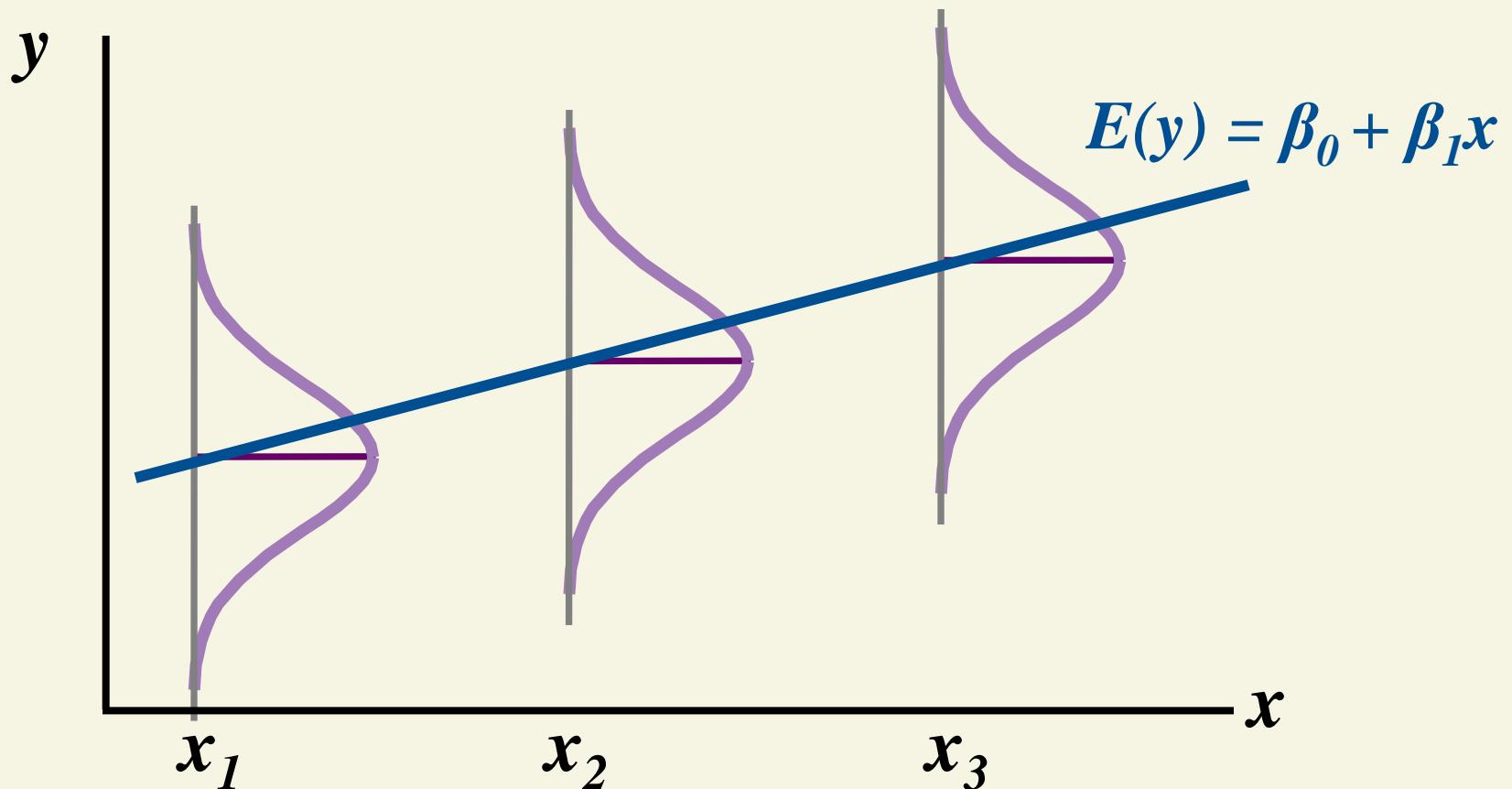
# Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. **Specify probability distribution of random error term**
  - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

# Linear Regression Assumptions

1. Mean of probability distribution of error,  $\varepsilon$ , is 0
2. Probability distribution of error has constant variance
3. Probability distribution of error,  $\varepsilon$ , is normal
4. Errors are independent

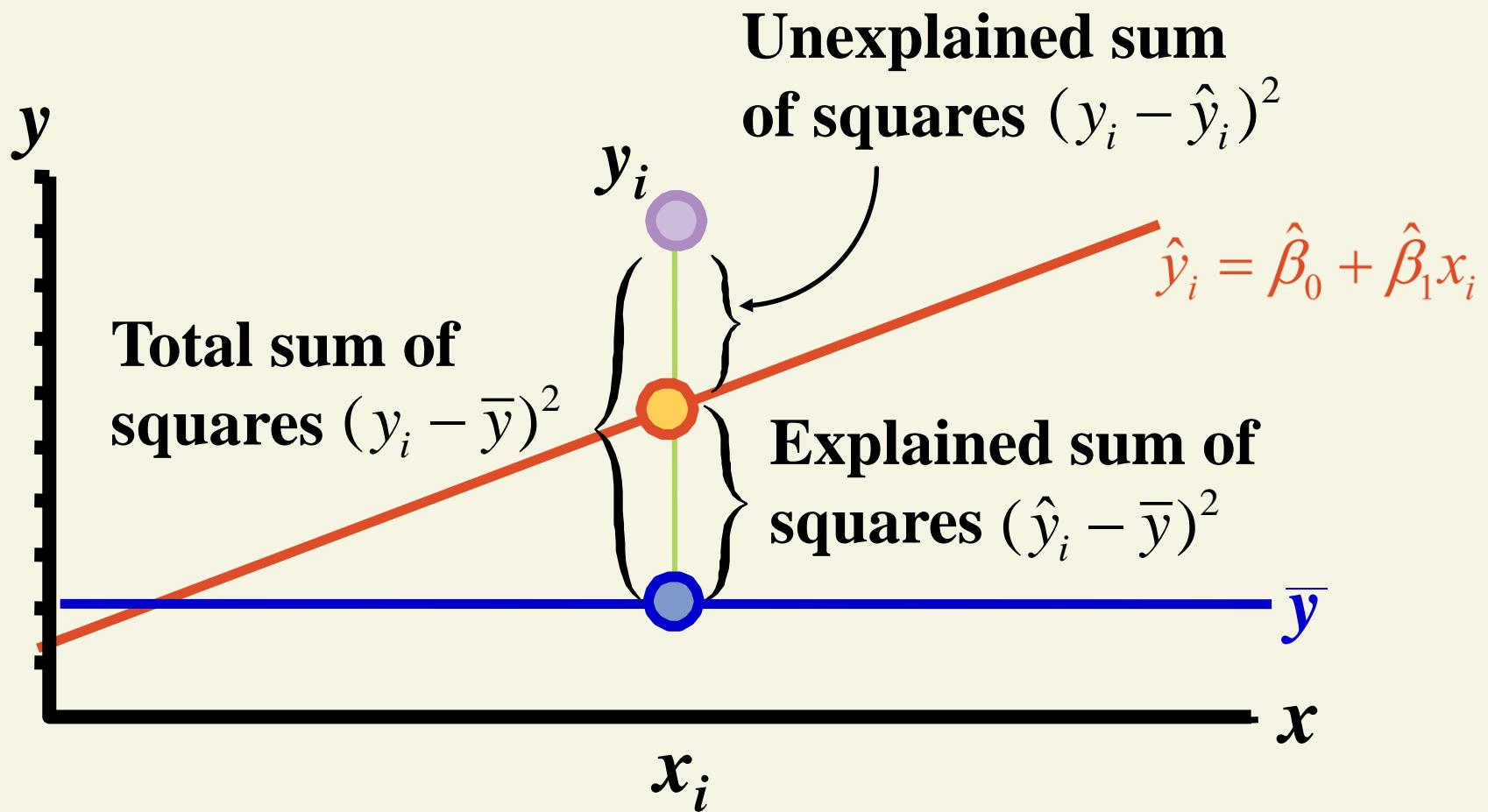
# Error Probability Distribution



# Random Error Variation

- Variation of actual  $y$  from predicted  $\hat{y}$
- Measured by standard error of regression model
  - Sample standard deviation of  $\hat{\varepsilon}$ :  $s$
- Affects several factors
  - Parameter significance
  - Prediction accuracy

# Variation Measures



# Estimation of $\sigma^2$

$$s^2 = \frac{SSE}{n-2} \quad \text{where} \quad SSE = \sum (y_i - \hat{y}_i)^2$$

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

# Calculating SSE, $s^2$ , $s$

## Example

You're a marketing analyst for Hasbro Toys.  
You gather the following data:

<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

Find **SSE**,  $s^2$ , and **s**.



# Calculating SSE Solution

$x_i$	$y_i$	$\hat{y} = -.1 + .7x$	$y - \hat{y}$	$(y - \hat{y})^2$
1	1	.6	.4	.16
2	1	1.3	-.3	.09
3	2	2	0	0
4	2	2.7	-.7	.49
5	4	3.4	.6	.36
				SSE=1.1

# Calculating $s^2$ and $s$ Solution

$$s^2 = \frac{SSE}{n-2} = \frac{1.1}{5-2} = .36667$$

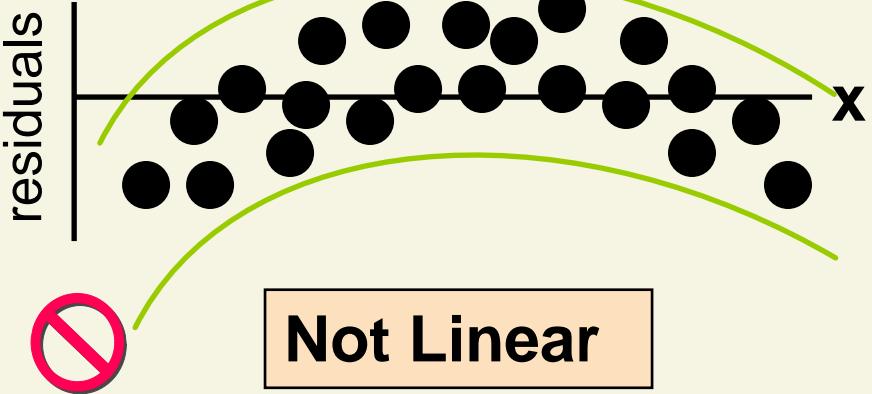
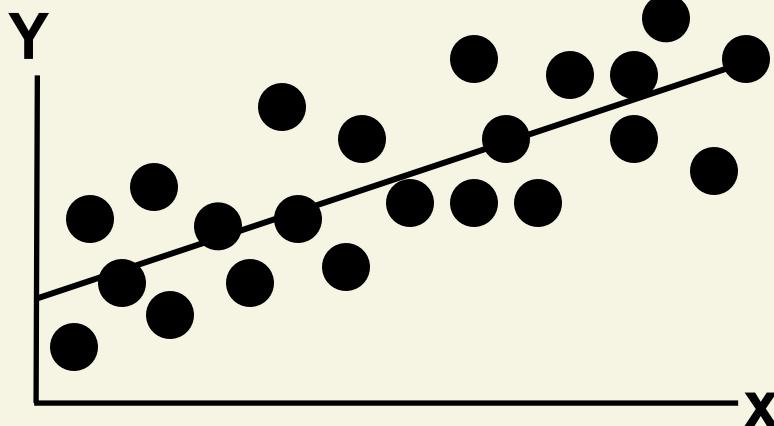
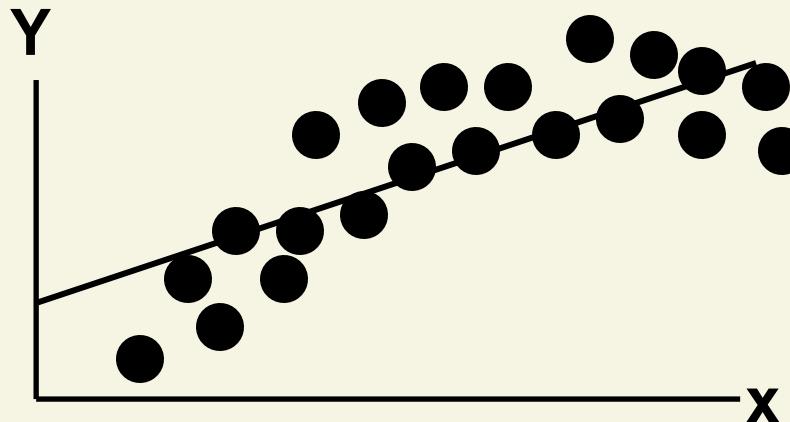
$$s = \sqrt{.36667} = .6055$$

# Residual Analysis

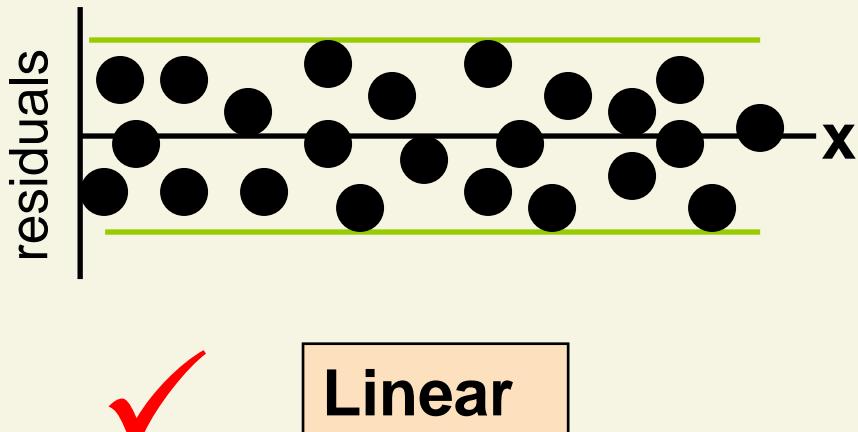
$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation  $i$ ,  $e_i$ , is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of  $X$  (homoscedasticity)

# Residual Analysis for Linearity



**Not Linear**

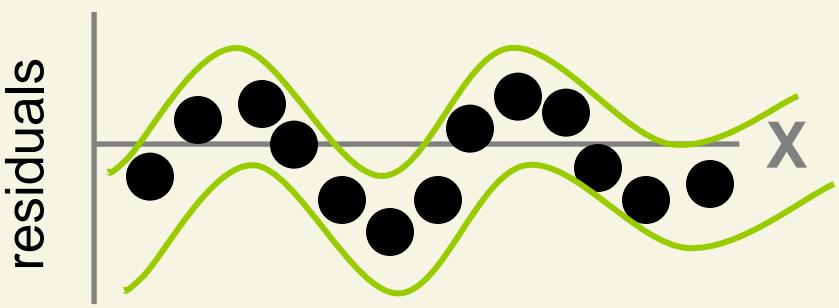
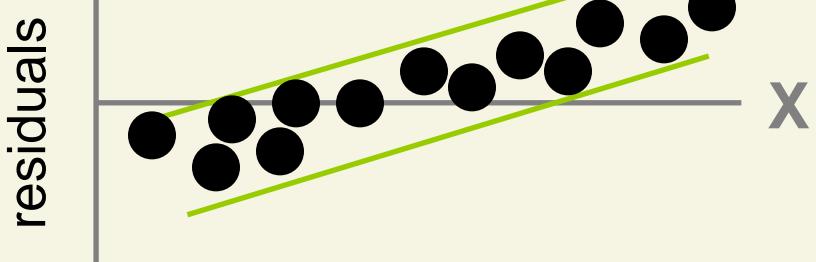


**Linear**

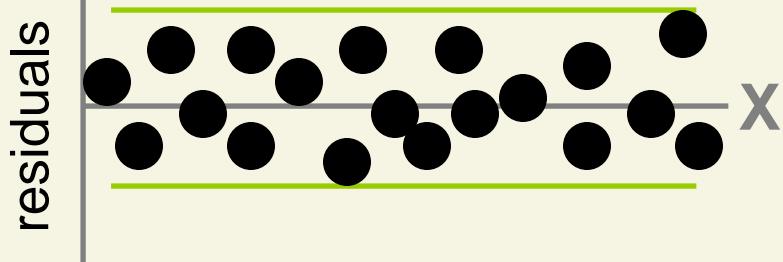
# Residual Analysis for Independence



Not Independent



Independent

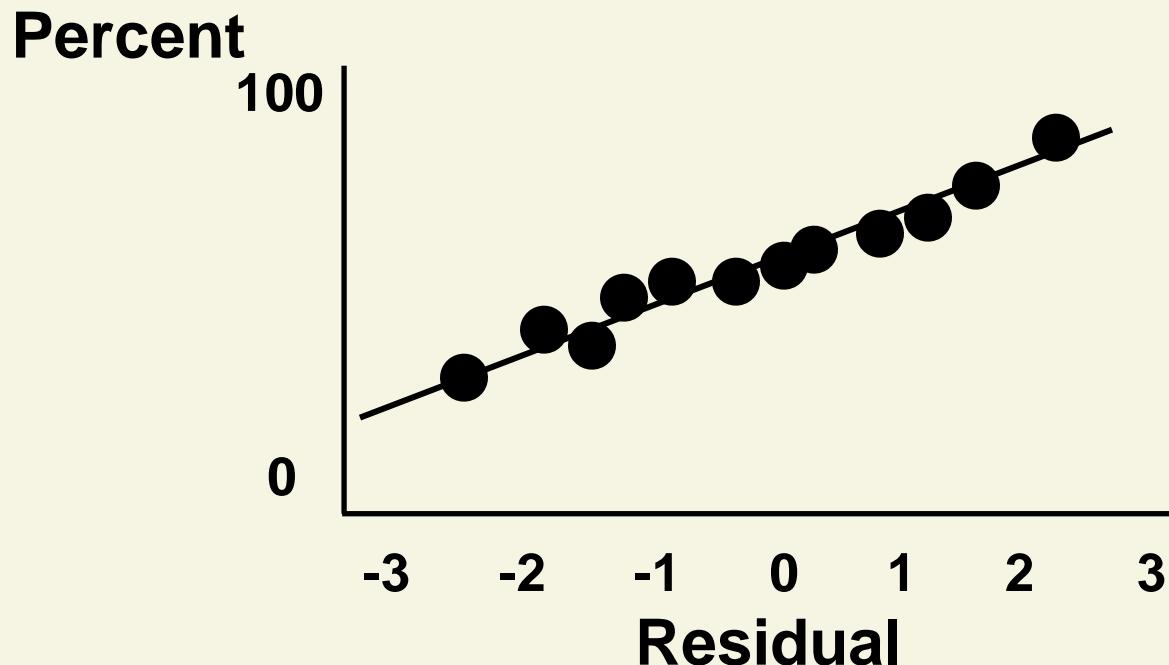


# Checking for Normality

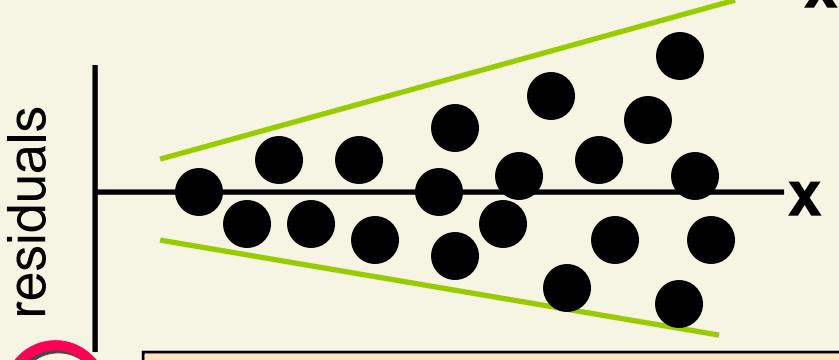
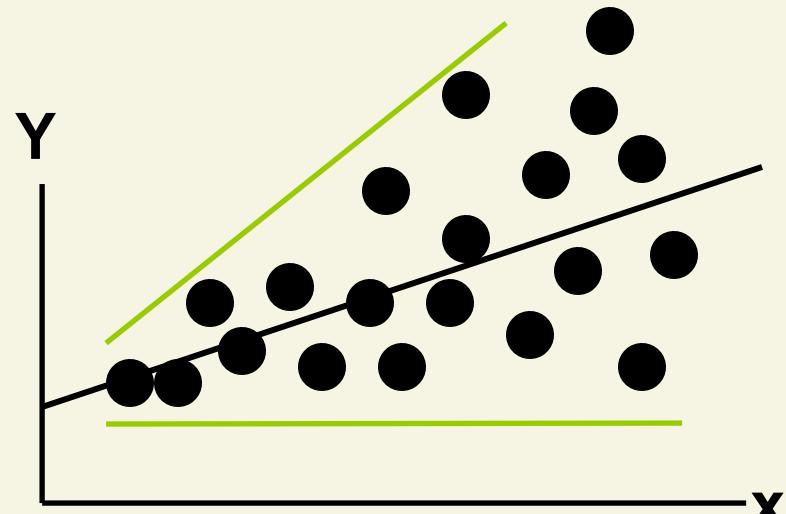
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

# Residual Analysis for Normality

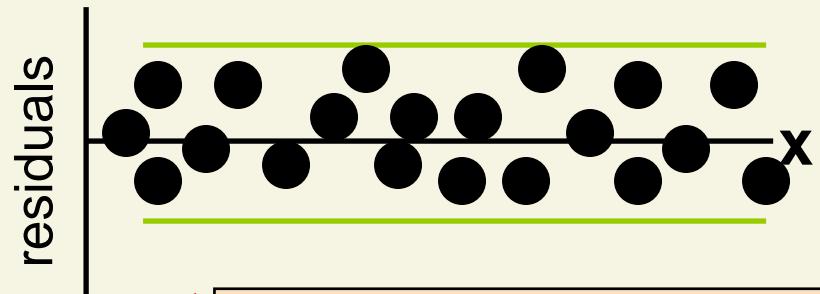
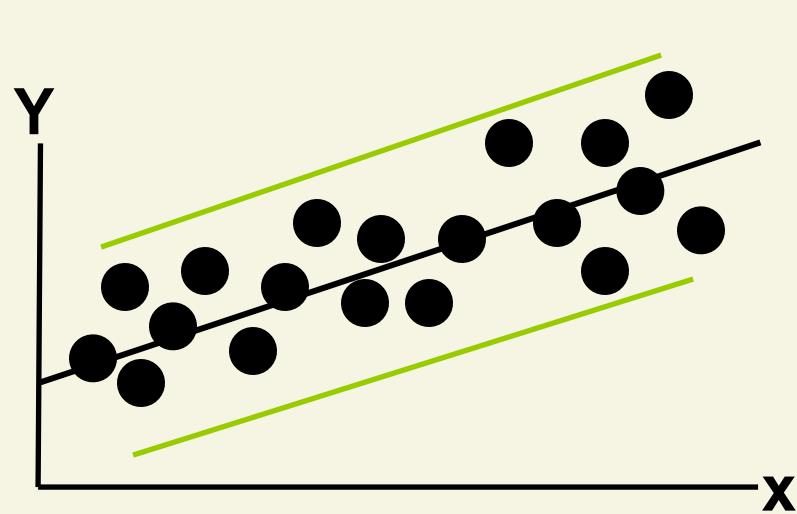
When using a normal probability plot, normal errors will approximately display in a straight line



# Residual Analysis for Equal Variance



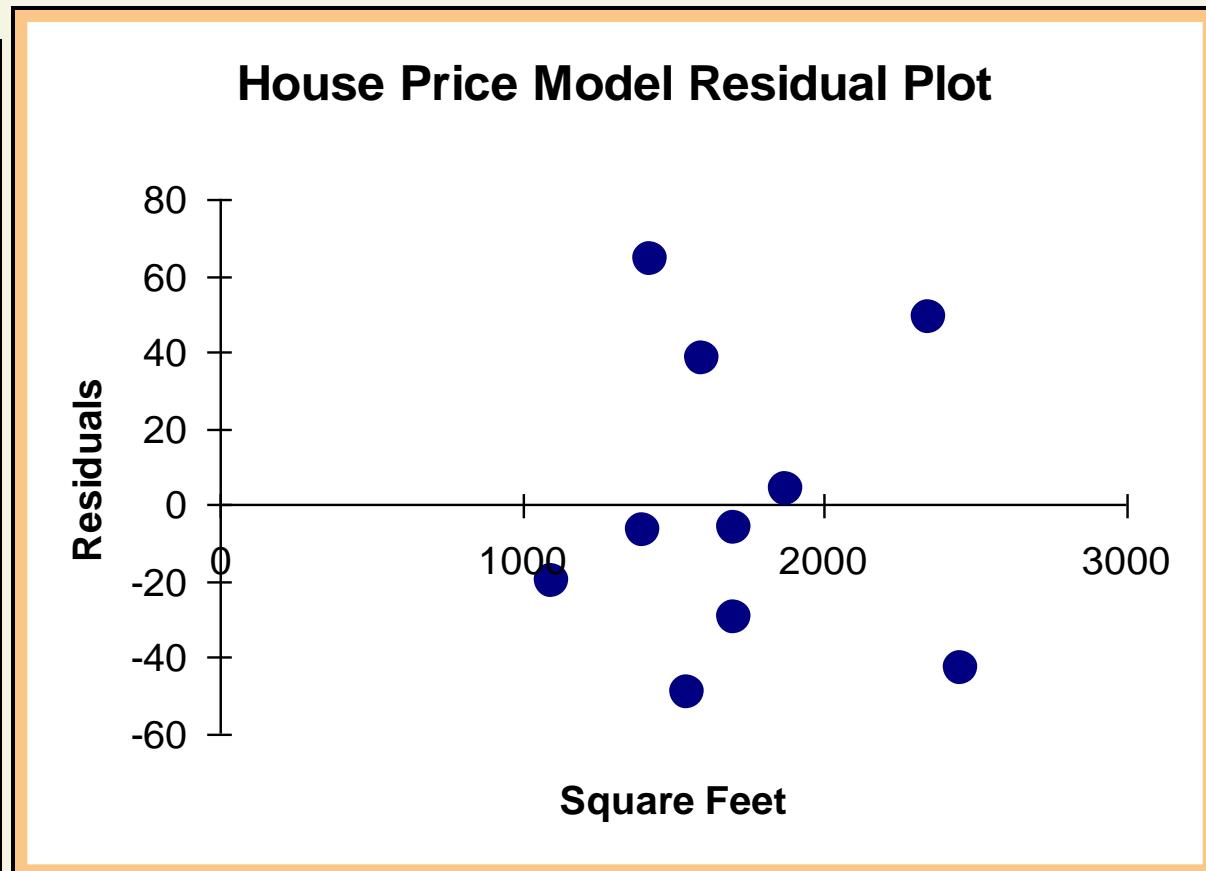
Non-constant variance



Constant variance

# Simple Linear Regression Example: Excel Residual Output

RESIDUAL OUTPUT		
	<i>Predicted House Price</i>	<i>Residuals</i>
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate  
any regression assumptions

# Evaluating the Model

Testing for Significance

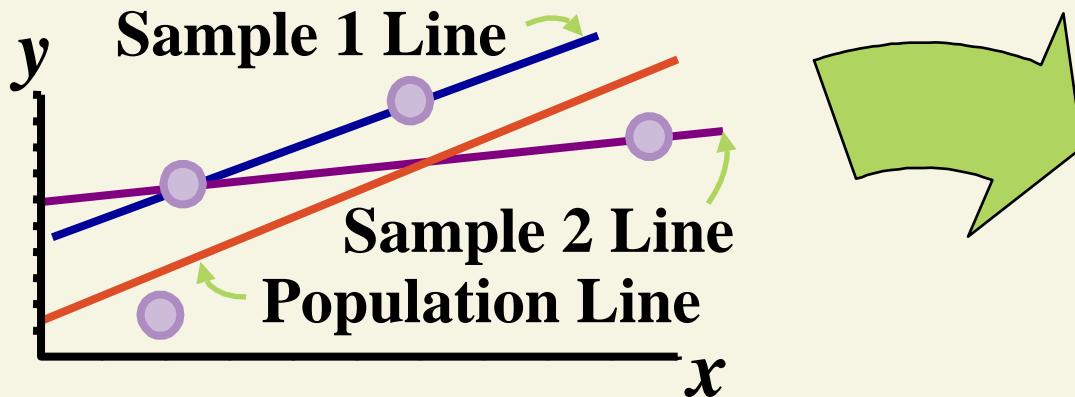
# Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
  - Estimate standard deviation of error
4. **Evaluate model**
5. Use model for prediction and estimation

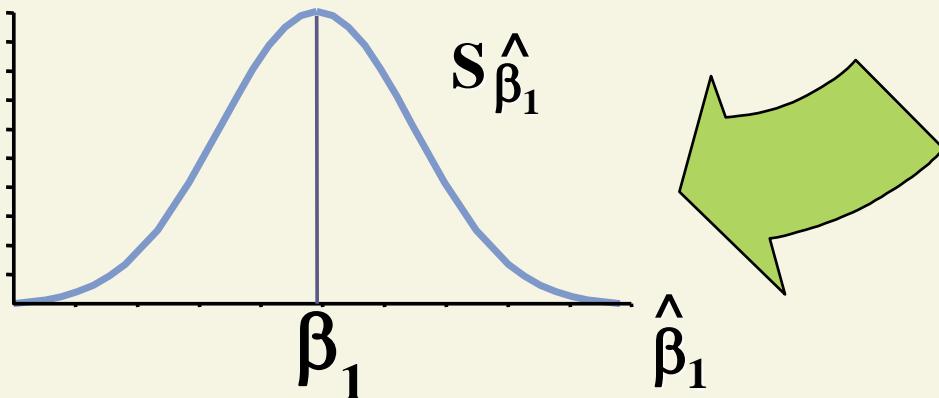
# Test of Slope Coefficient

- Shows if there is a linear relationship between  $x$  and  $y$
- Involves population slope  $\beta_1$
- Hypotheses
  - $H_0: \beta_1 = 0$  (No Linear Relationship)
  - $H_a: \beta_1 \neq 0$  (Linear Relationship)
- Theoretical basis is sampling distribution of slope

# Sampling Distribution of Sample Slopes



Sampling Distribution



All Possible  
Sample Slopes

Sample 1: 2.5

Sample 2: 1.6

Sample 3: 1.8

Sample 4: 2.1

: :

Very large number of  
sample slopes

# Slope Coefficient Test Statistic

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} \quad df = n - 2$$

where

$$SS_{xx} = \sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n}$$

# Test of Slope Coefficient Example

You're a marketing analyst for Hasbro Toys.

You find  $\hat{\beta}_0 = -.1$ ,  $\hat{\beta}_1 = .7$  and  $s = .6055$ .

<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

Is the relationship **significant** at the **.05** level of significance?



# Solution Table

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

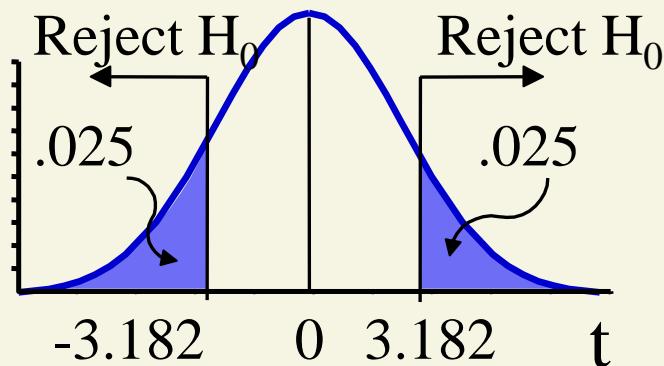
# Test Statistic Solution

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{SS_{xx}}} = \frac{.6055}{\sqrt{55 - \frac{(15)^2}{5}}} = .1914$$

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{.70}{.1914} = 3.657$$

# Test of Slope Coefficient Solution

- $H_0: \beta_1 = 0$
- $H_a: \beta_1 \neq 0$
- $\alpha = .05$
- $df = 5 - 2 = 3$
- Critical Value(s):



# Test of Slope Coefficient Solution

Test Statistic:

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{.70}{.1914} = 3.657$$

Decision:

Reject at  $\alpha = .05$

Conclusion:

There is evidence of a relationship

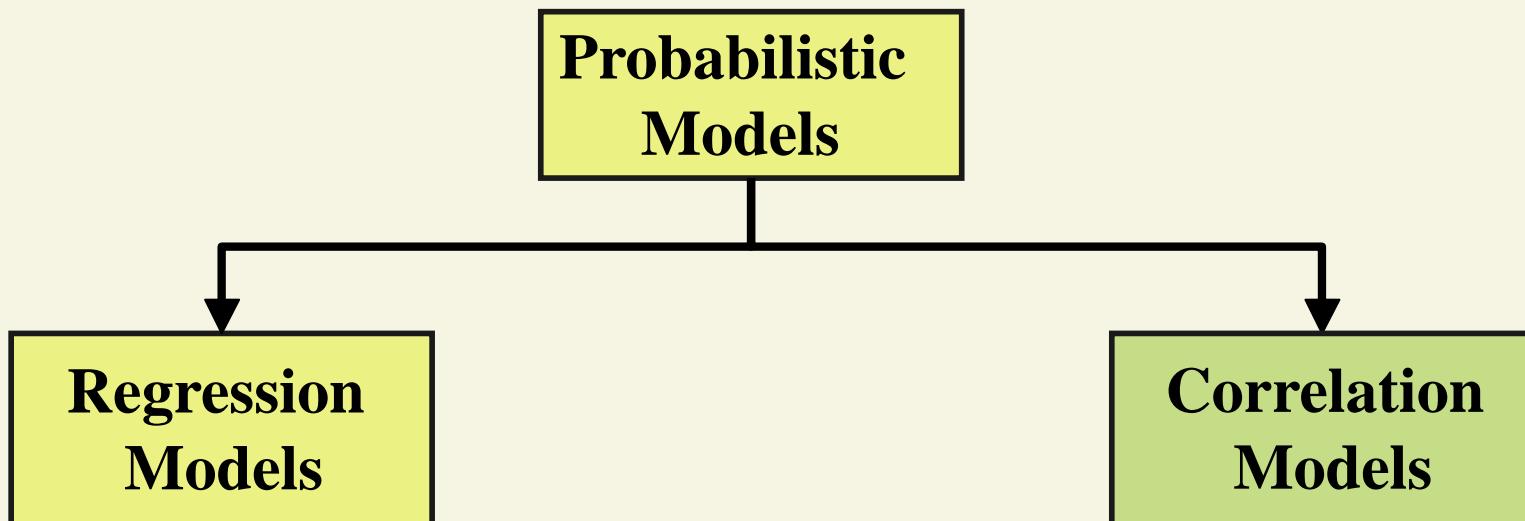
# Test of Slope Coefficient Computer Output

Parameter Estimates						
	Variable	DF	Parameter Estimate	Standard Error	T for H0:	Prob> T
	INTERCEP	1	-0.1000	0.6350	-0.157	0.8849
	ADVERT	1	0.7000	0.1914	3.656	0.0354

The diagram illustrates the calculation of the t-statistic and its corresponding P-value. Arrows point from the 'Parameter Estimate' and 'Standard Error' values for the 'ADVERT' variable in the table to the formula  $t = \hat{\beta}_1 / S_{\hat{\beta}_1}$ . Another arrow points from the calculated t-value (3.656) to the 'P-Value' (0.0354) listed in the table.

# **Correlation Models**

# Types of Probabilistic Models



# Correlation Models

- Answers ‘How strong is the **linear** relationship between two variables?’
- Coefficient of correlation
  - Sample correlation coefficient denoted  $r$
  - Values range from  $-1$  to  $+1$
  - Measures degree of association
  - Does not indicate cause–effect relationship

# Coefficient of Correlation

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

# Coefficient of Correlation Values

Perfect  
Negative  
Correlation



-1.0

-.5

No Linear  
Correlation



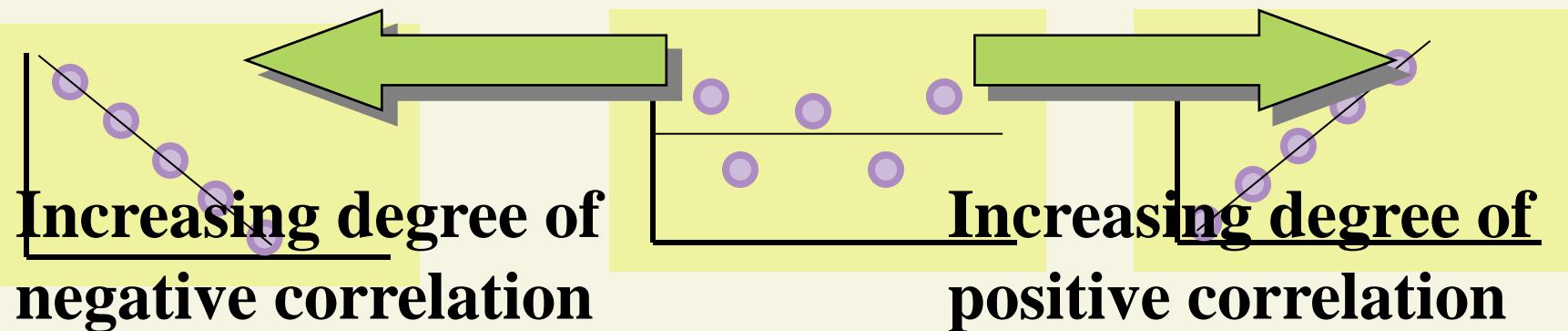
0

Perfect  
Positive  
Correlation



+.5

+1.0



# Coefficient of Correlation

## Example

You're a marketing analyst for Hasbro Toys.

<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

Calculate the **coefficient of correlation**.



# Solution Table

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

# Coefficient of Correlation Solution

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 55 - \frac{(15)^2}{5} = 10$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 26 - \frac{(10)^2}{5} = 6$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 37 - \frac{(15)(10)}{5} = 7$$

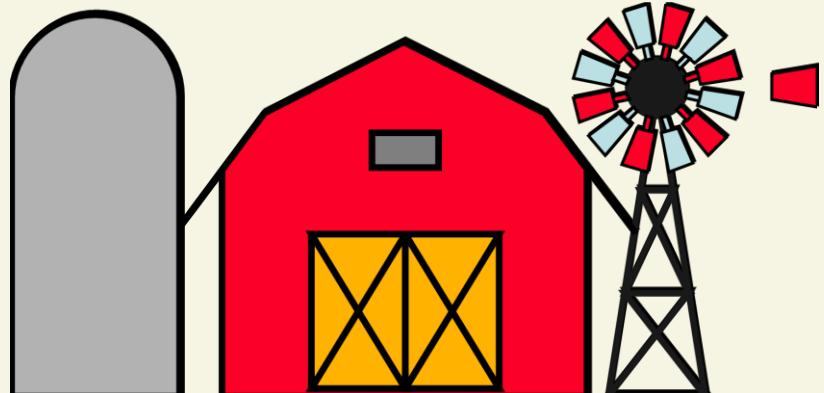
$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{7}{\sqrt{10 \cdot 6}} = .904$$

# Coefficient of Correlation

## Thinking Challenge

You're an economist for the county cooperative.  
You gather the following data:

<u>Fertilizer (lb.)</u>	<u>Yield (lb.)</u>
4	3.0
6	5.5
10	6.5
12	9.0



© 1984-1994 T/Maker Co.

Find the **coefficient of correlation**.

# Solution Table\*

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
4	3.0	16	9.00	12
6	5.5	36	30.25	33
10	6.5	100	42.25	65
12	9.0	144	81.00	108
32	24.0	296	162.50	218

# Coefficient of Correlation Solution\*

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 296 - \frac{(32)^2}{4} = 40$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 162.5 - \frac{(24)^2}{4} = 18.5$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 218 - \frac{(32)(24)}{4} = 26$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{26}{\sqrt{40 \cdot 18.5}} = .956$$

# Coefficient of Determination

**Proportion** of variation ‘explained’ by relationship between  $x$  and  $y$

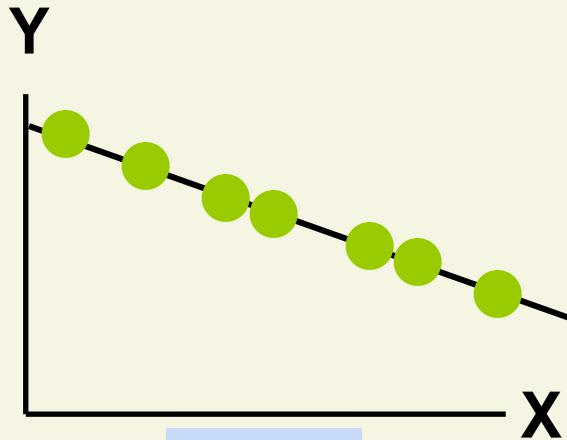
$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_{yy} - SSE}{SS_{yy}}$$

$$r^2 = (\text{coefficient of correlation})^2$$

$$0 \leq r^2 \leq 1$$

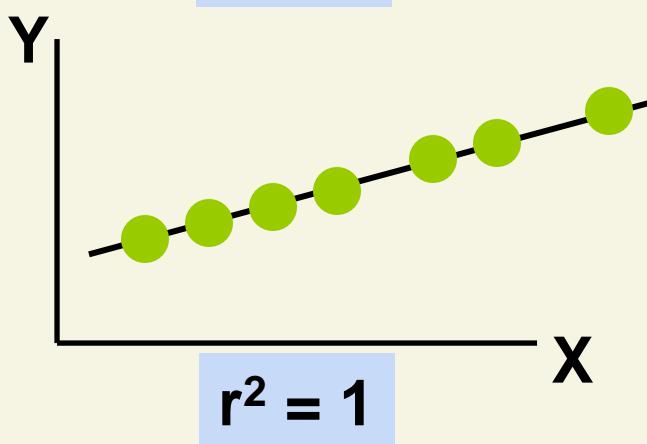


# Examples of Approximate $r^2$ Values



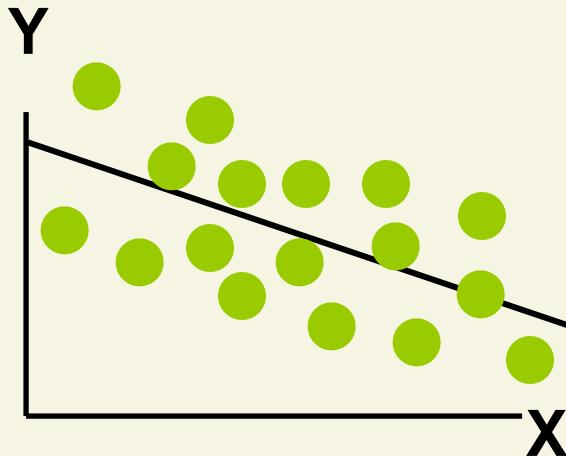
$$r^2 = 1$$

**Perfect linear relationship between X and Y:**



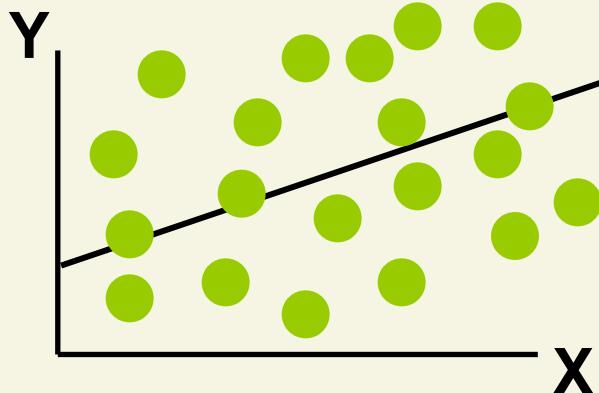
**100% of the variation in Y is explained by variation in X**

# Examples of Approximate $r^2$ Values



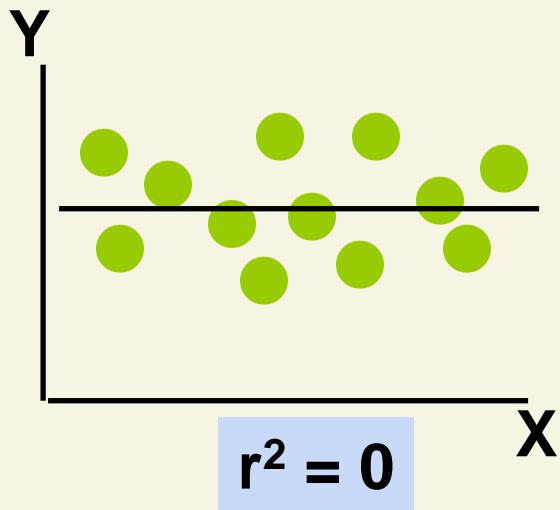
$$0 < r^2 < 1$$

**Weaker linear relationships between X and Y:**



**Some but not all of the variation in Y is explained by variation in X**

# Examples of Approximate $r^2$ Values



$$r^2 = 0$$

**No linear relationship between X and Y:**

**The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)**

# Coefficient of Determination Example

You're a marketing analyst for Hasbro Toys.

You know  $r = .904$ .

<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

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



# Coefficient of Determination Solution

$$r^2 = (\text{coefficient of correlation})^2$$

$$r^2 = (.904)^2$$

$$r^2 = .817$$

**Interpretation:** About 81.7% of the sample variation in Sales ( $y$ ) can be explained by using Ad \$ ( $x$ ) to predict Sales ( $y$ ) in the linear model.

# $r^2$ Computer Output

Root MSE	0 . 60553	R-square	0 . 8167
Dep Mean	2 . 00000	Adj R-sq	0 . 7556
C.V.	30 . 27650		

$$r^2$$

$r^2$  adjusted for number of explanatory variables & sample size

# **Using the Model for Prediction & Estimation**

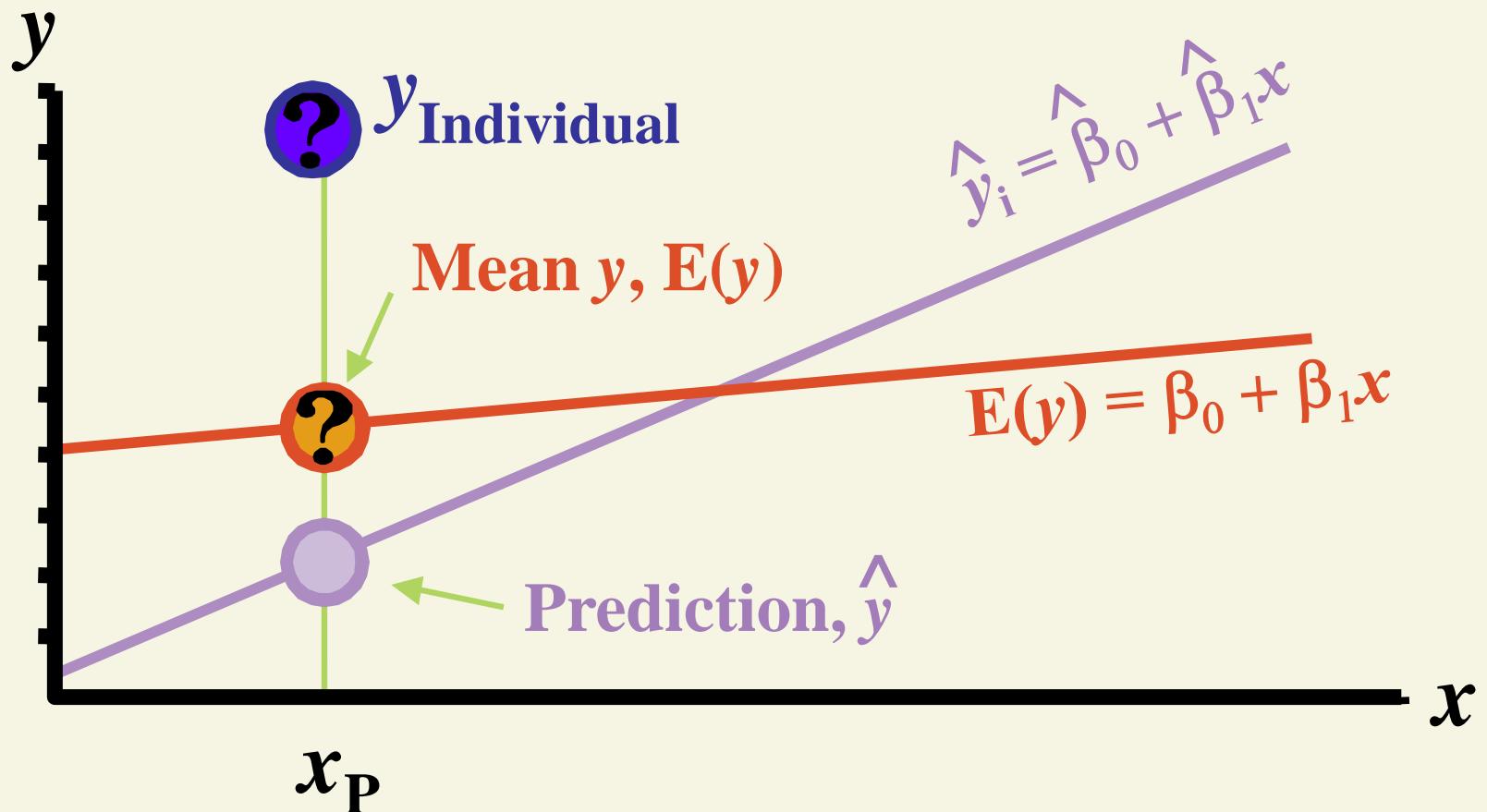
# Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
  - Estimate standard deviation of error
4. Evaluate model
5. **Use model for prediction and estimation**

# Prediction With Regression Models

- Types of predictions
  - Point estimates
  - Interval estimates
- What is predicted
  - Population mean response  $E(y)$  for given  $x$ 
    - Point on population regression line
  - Individual response ( $y_i$ ) for given  $x$

# What Is Predicted



# Confidence Interval Estimate for Mean Value of $y$ at $x = x_p$

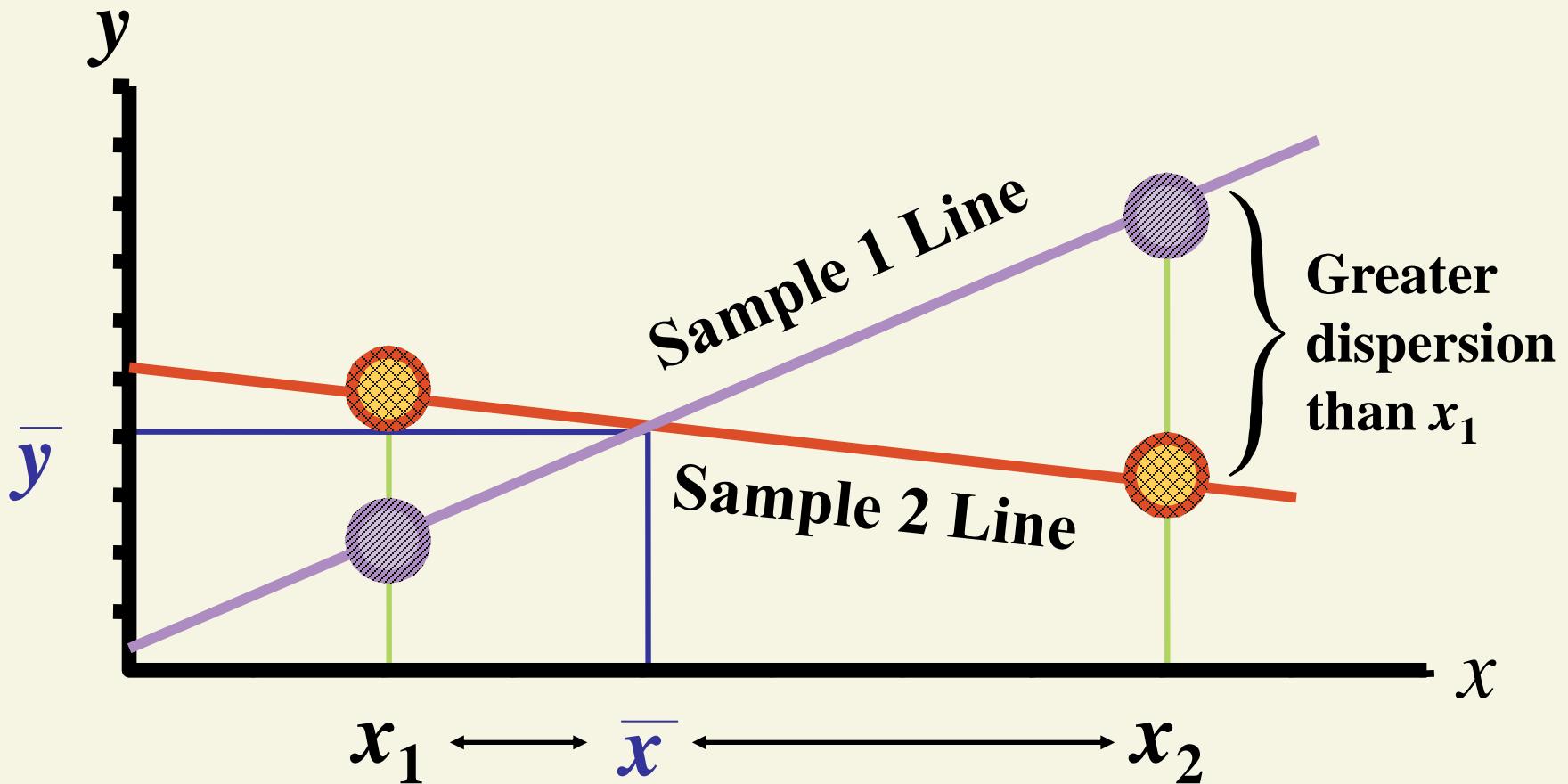
$$\hat{y} \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

$$\text{df} = n - 2$$

# Factors Affecting Interval Width

1. Level of confidence ( $1 - \alpha$ )
  - Width increases as confidence increases
2. Data dispersion ( $s$ )
  - Width increases as variation increases
3. Sample size
  - Width decreases as sample size increases
4. Distance of  $x_p$  from mean  $\bar{x}$ 
  - Width increases as distance increases

# Why Distance from Mean?



# Confidence Interval Estimate Example

You're a marketing analyst for Hasbro Toys.

You find  $\hat{\beta}_0 = -.1$ ,  $\hat{\beta}_1 = .7$  and  $s = .6055$ .

<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4



Find a **95%** confidence interval for the **mean** sales when advertising is **\$4**.

# Solution Table

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
<b>15</b>	<b>10</b>	<b>55</b>	<b>26</b>	<b>37</b>

# Confidence Interval Estimate Solution

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

$$\hat{y} = -.1 + (.7)(4) = 2.7$$

$$2.7 \pm (3.182)(.6055) \sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$1.645 \leq E(Y) \leq 3.755$$

# Prediction Interval of Individual Value of $y$ at $x = x_p$

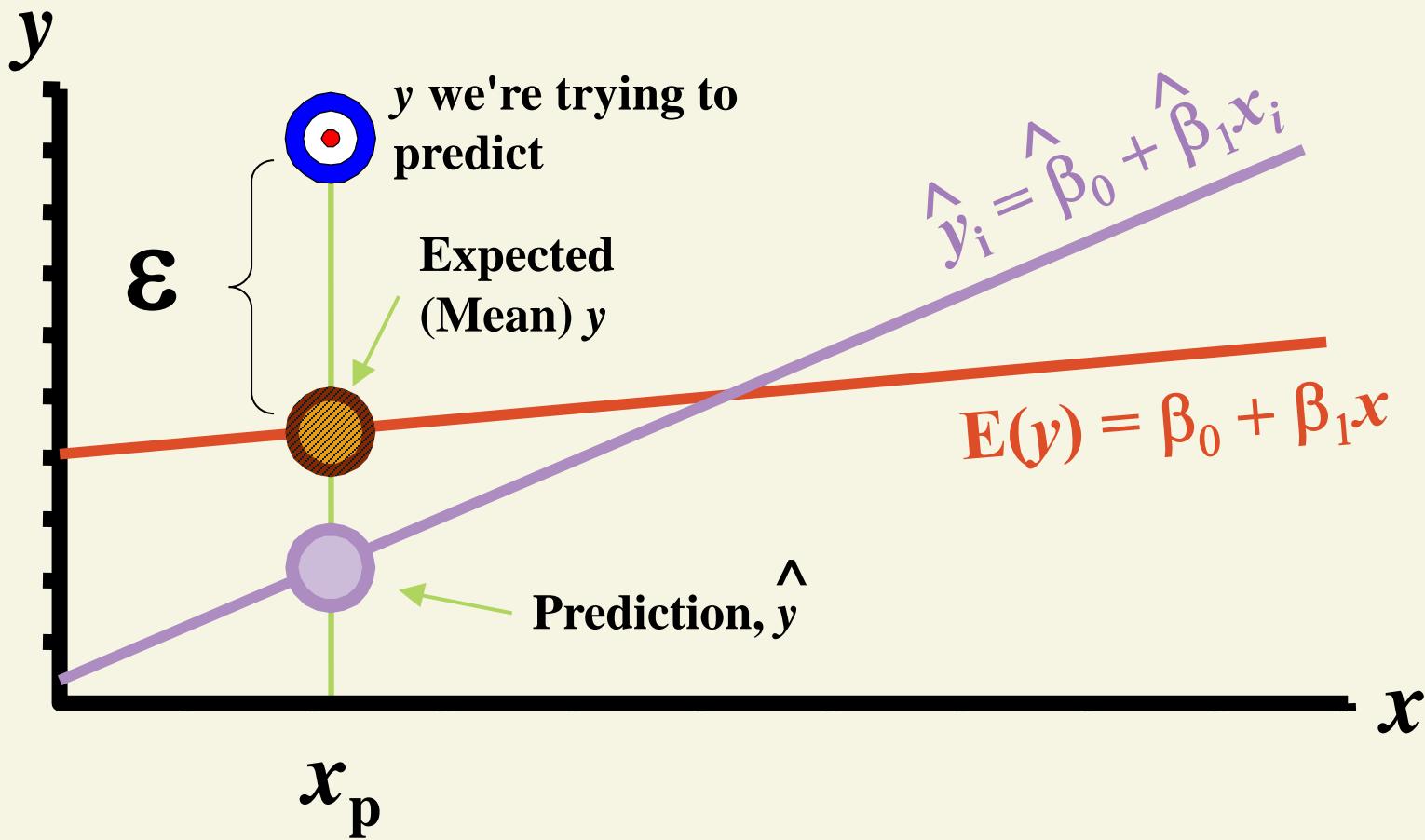
$$\hat{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

Note!



$$df = n - 2$$

# Why the Extra ‘S’?



# Prediction Interval Example

You're a marketing analyst for Hasbro Toys.

You find  $\hat{\beta}_0 = -.1$ ,  $\hat{\beta}_1 = .7$  and  $s = .6055$ .

<u>Ad \$</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4



Predict the sales when advertising is **\$4**. Use a **95% prediction** interval.

# Solution Table

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
<b>15</b>	<b>10</b>	<b>55</b>	<b>26</b>	<b>37</b>

# Prediction Interval Solution

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

*x to be predicted*

$$\hat{y} = -.1 + (.7)(4) = 2.7$$

$$2.7 \pm (3.182)(.6055) \sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$.503 \leq y_4 \leq 4.897$$

# Interval Estimate Computer Output

Obs	Dep Var	Pred Value	Std Err Predict	Low95% Mean	Upp95% Mean	Low95% Predict	Upp95% Predict	
1	SALES	1.000	0.600	0.469	-0.892	2.092	-1.837	3.037
2		1.000	1.300	0.332	0.244	2.355	-0.897	3.497
3		2.000	2.000	0.271	1.138	2.861	-0.111	4.111
4		2.000	2.700	0.332	1.644	3.755	0.502	4.897
5		4.000	3.400	0.469	1.907	4.892	0.962	5.837

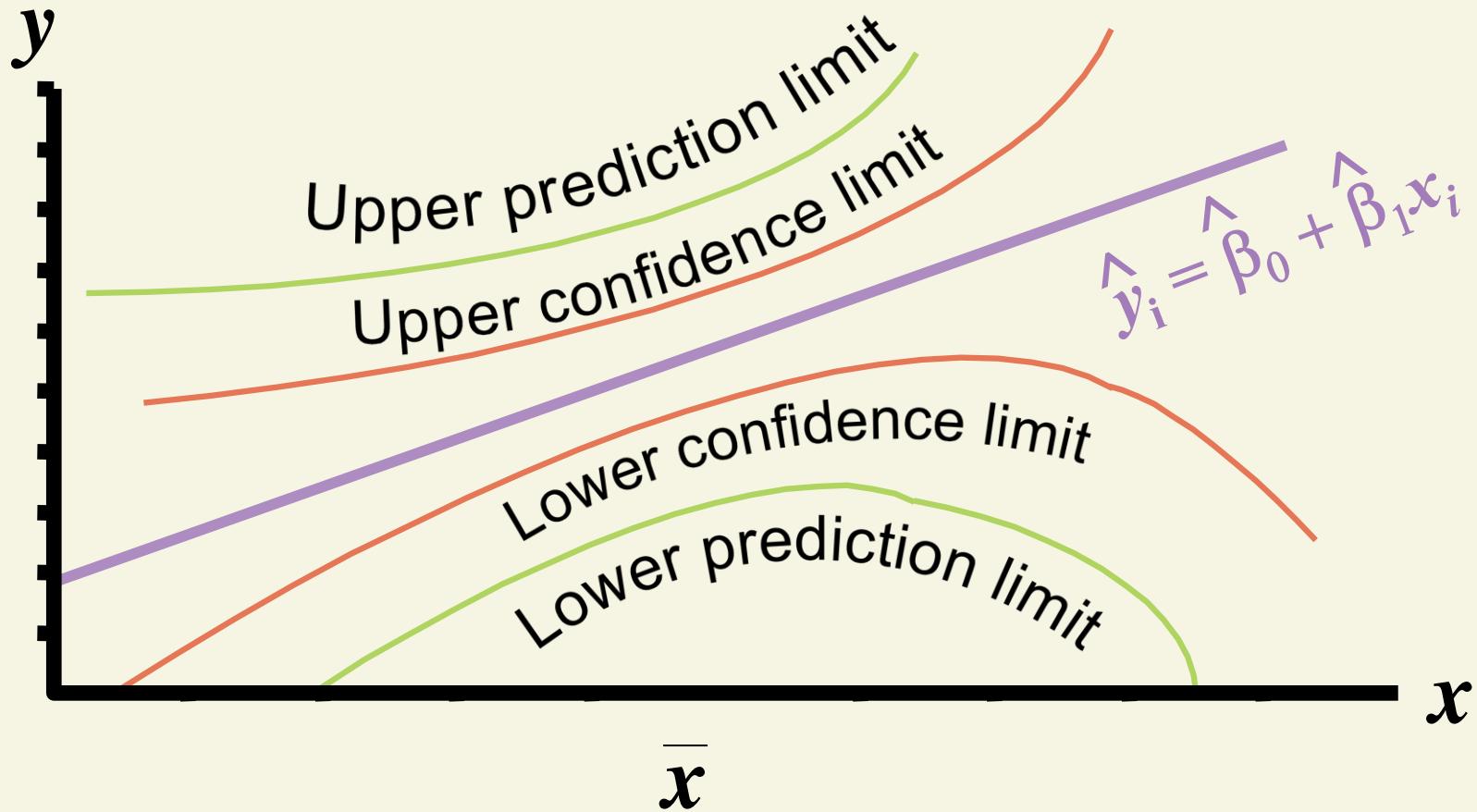
**Predicted  $y$  when  $x = 4$**

$\hat{S}_Y$

**Confidence Interval**

**Prediction Interval**

# Confidence Intervals v. Prediction Intervals



# Conclusion

1. Described the Linear Regression Model
2. Stated the Regression Modeling Steps
3. Explained Least Squares
4. Computed Regression Coefficients
5. Explained Correlation
6. Predicted Response Variable