11. Simple Linear Regression

Adapted From :

Probability & Statistics for Engineers & Scientists, 9^{th} Ed.

Walpole/Myers/Myers/Ye (c)2010

Introduction to Business Statistics, 5e

Kvanli/Guynes/Pavur (c)2000

South-Western College Publishing

Statistics for Managers

Using Microsoft® Excel 4th Edition

Learning Objectives

1. Describe the Linear Regression Model
2. State the Regression Modeling Steps
3. Explain Ordinary Least Squares (Understand and check model assumptions)
4. Compute Regression Coefficients
5. Predict Response Variable
6. Interpret Computer Output

Models

- Representation of Some Phenomenon
- Mathematical Model Is a Mathematical Expression of Some Phenomenon
- Often Describe Relationships between Variables
- Types:
  - Deterministic Models
  - Probabilistic Models

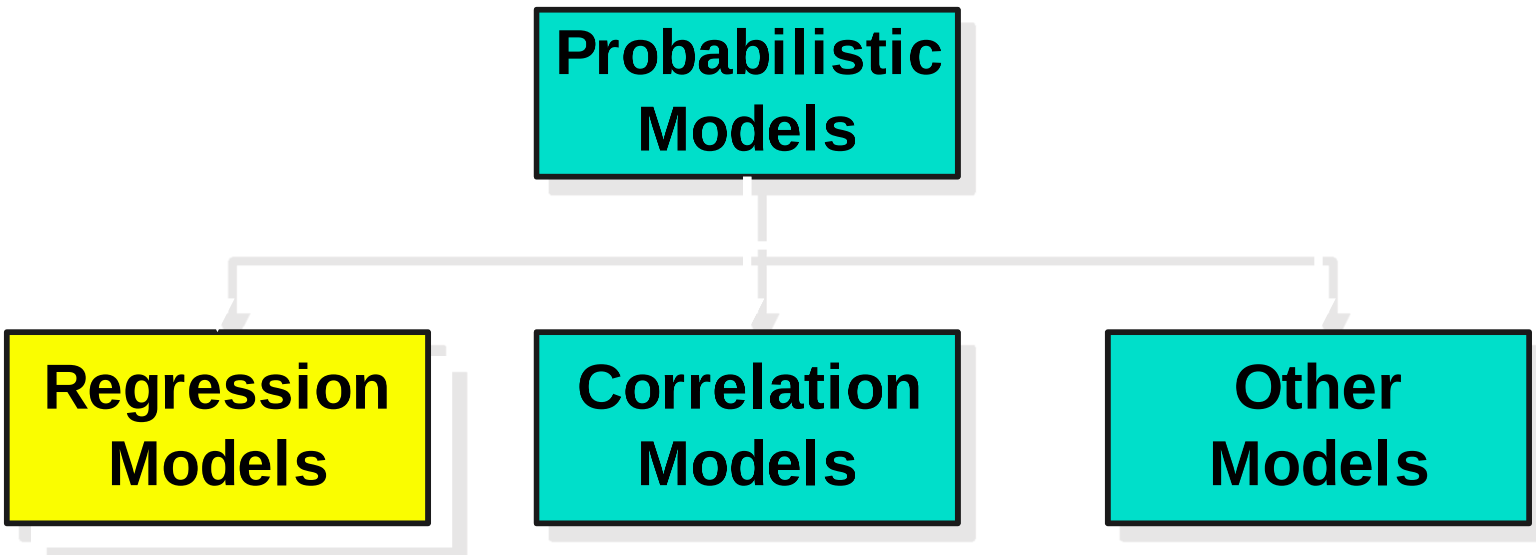
Deterministic Models

- Hypothesize Exact Relationships
- Suitable When Prediction Error is Negligible
- Example: Force Is Exactly Mass Times Acceleration
  - $F = m \cdot a$

Probabilistic Models

- Hypothesize 2 Components
  - Deterministic
  - Random Error
- Example: Sales Volume Is 10 Times Advertising Spending + Random Error
  - $Y = 10X + \varepsilon$
  - Random Error May Be Due to Factors Other Than Advertising

Types of Probabilistic Models



Regression Models

- Answer 'What Is the Relationship Between the Variables?'
- Equation Used
  - 1 Numerical Dependent (Response) Variable

    What Is to Be Predicted
  - 1 or More Numerical or Categorical Independent (Explanatory) Variables
- Used Mainly for Prediction & Estimation

Regression Modeling Steps

- Hypothesize Deterministic Component
- Estimate Unknown Model Parameters
- Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- Evaluate Model
- Use Model for Prediction & Estimation

Regression Modeling Steps

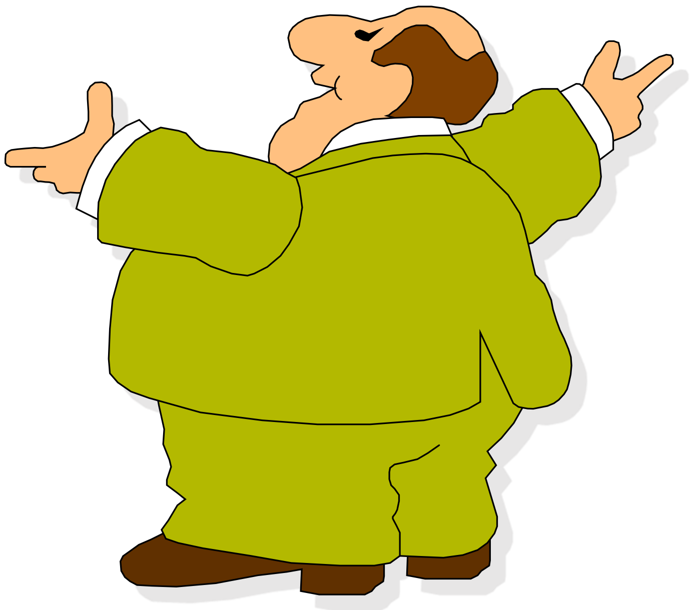
- **Hypothesize Deterministic Component**
- Estimate Unknown Model Parameters
- Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- Evaluate Model
- Use Model for Prediction & Estimation

Specifying the Model

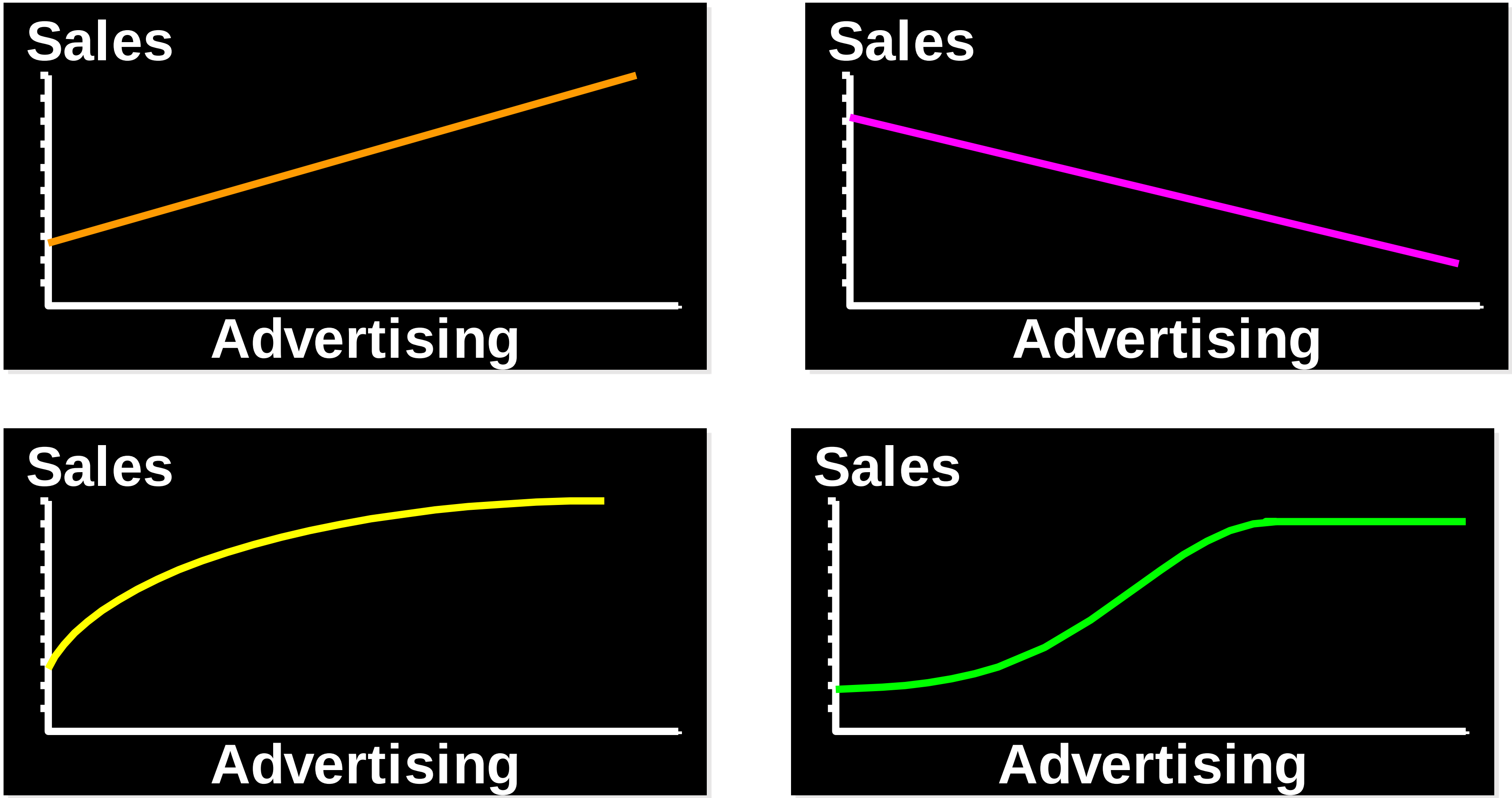
- Define Variables
- Hypothesize Nature of Relationship
  - Expected Effects (i.e., Coefficients' Signs)
  - Functional Form (Linear or Non-Linear)
  - Interactions

Model Specification Is Based on Theory

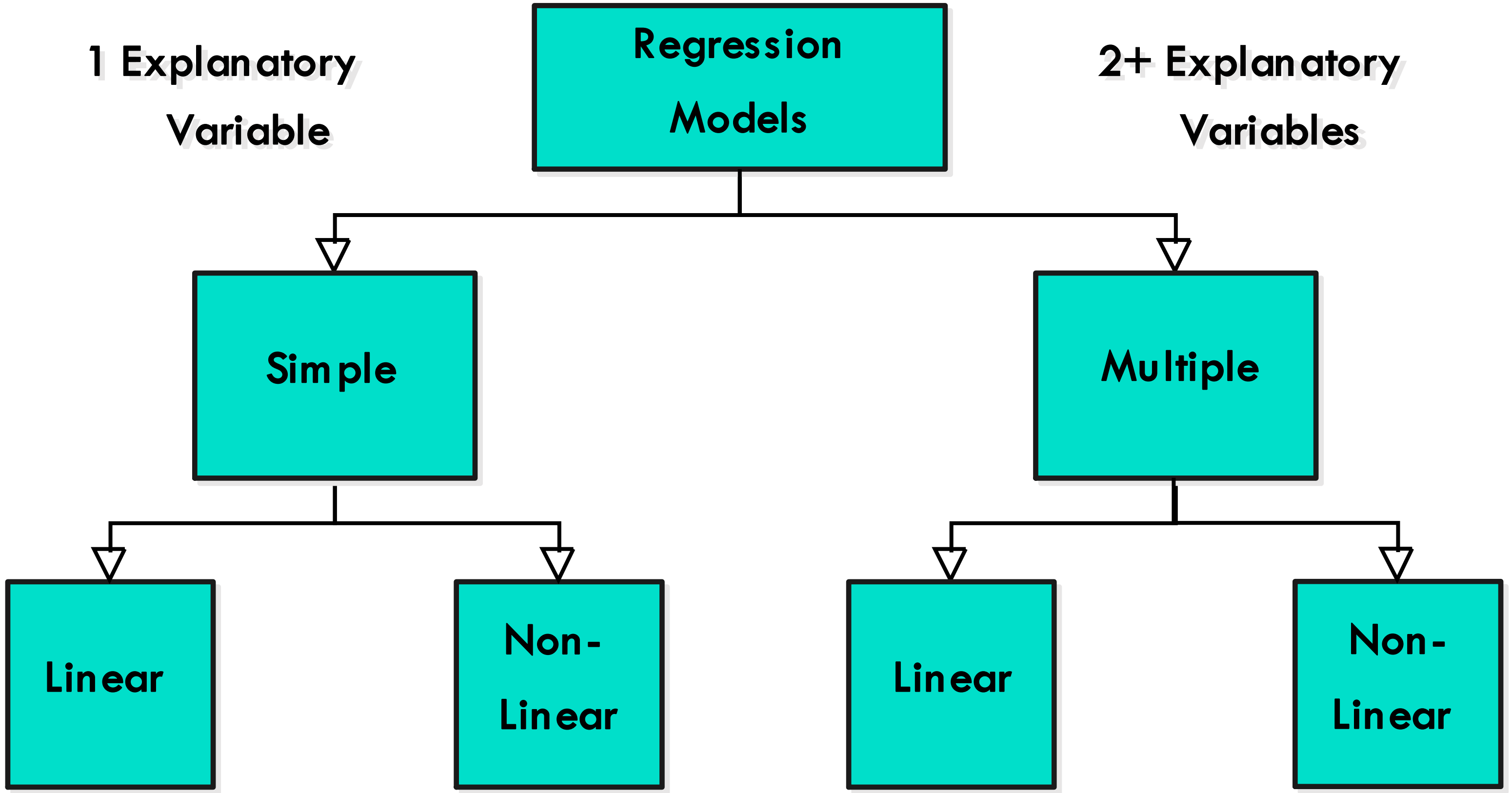
- Theory of Field (e.g., Sociology)
- Mathematical Theory
- Previous Research
- ‘Common Sense’



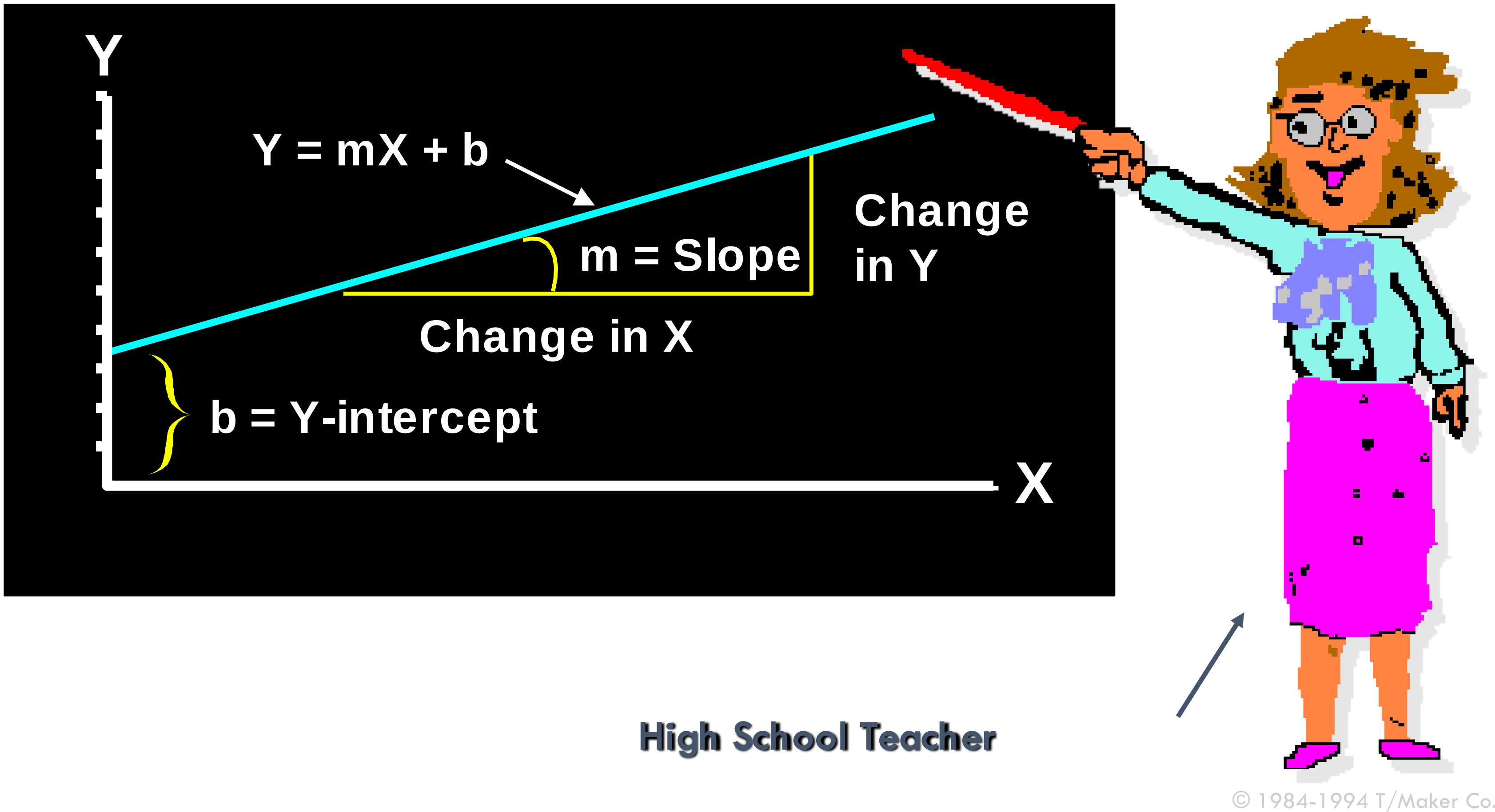
Thinking Challenge: Which Is More Logical?



Types of Regression Models



Linear Equations



Linear Regression Model

- Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

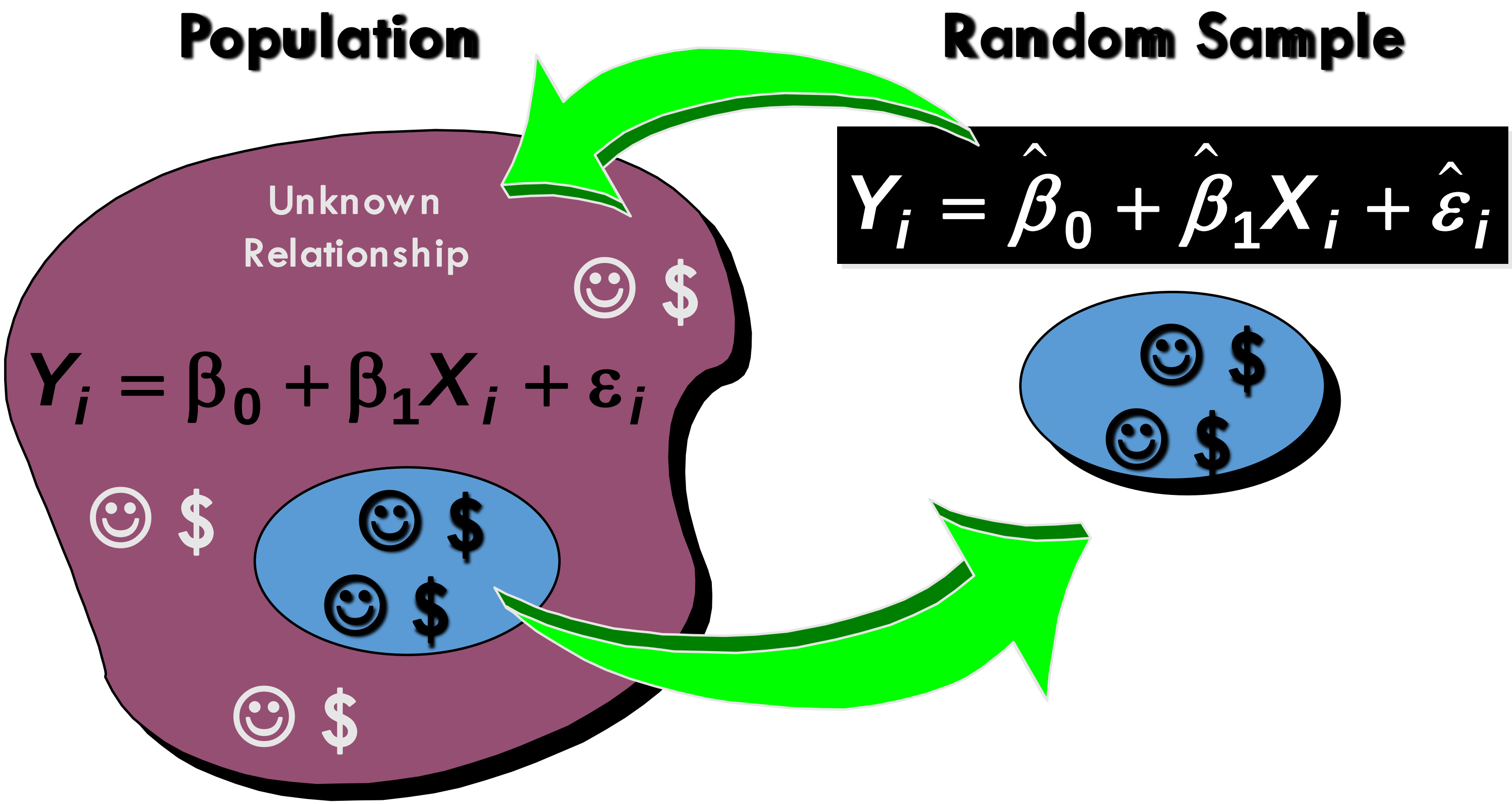
Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

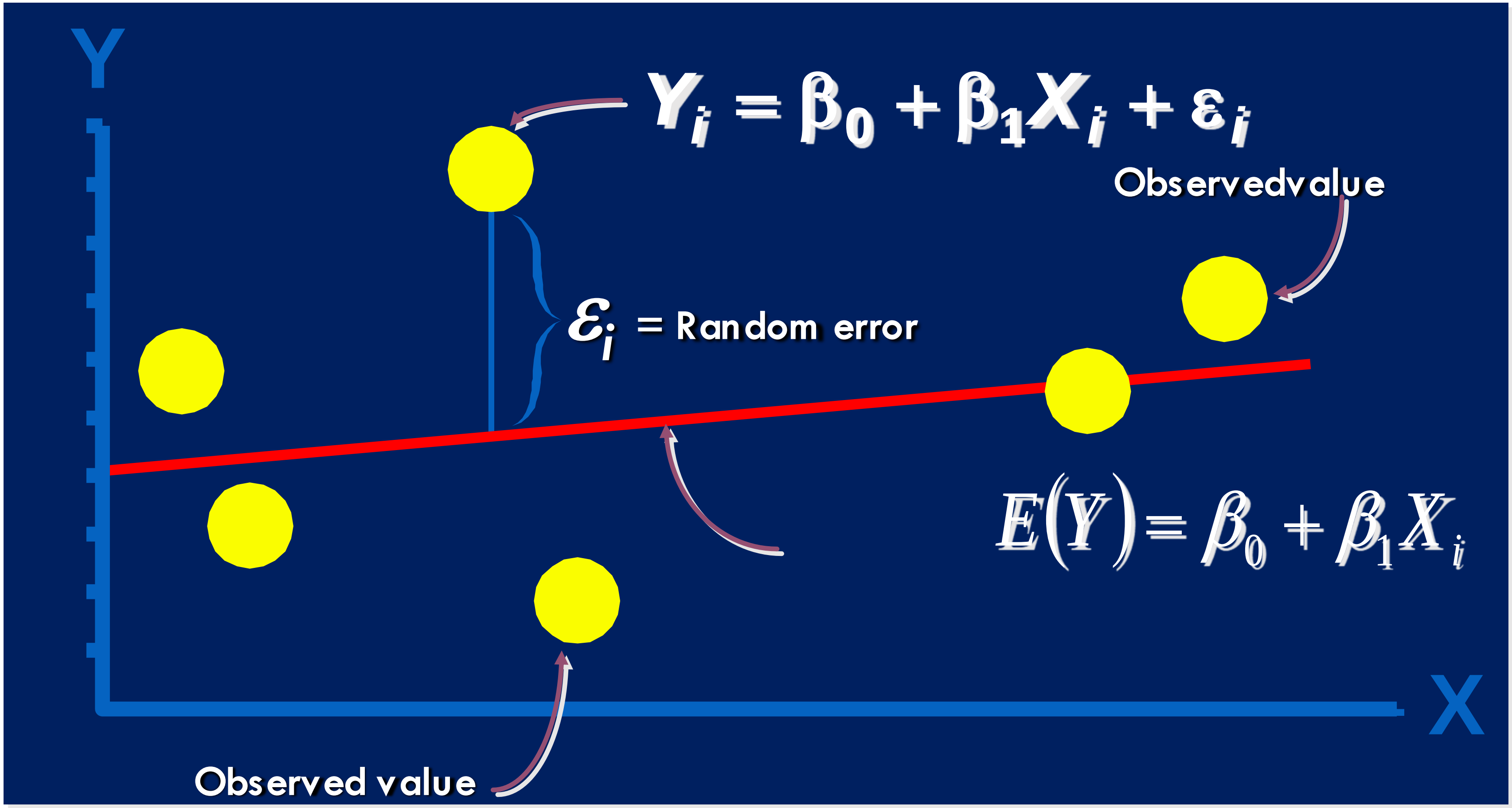
Dependent (Response) Variable (e.g., income)

Independent (Explanatory) Variable (e.g., education)

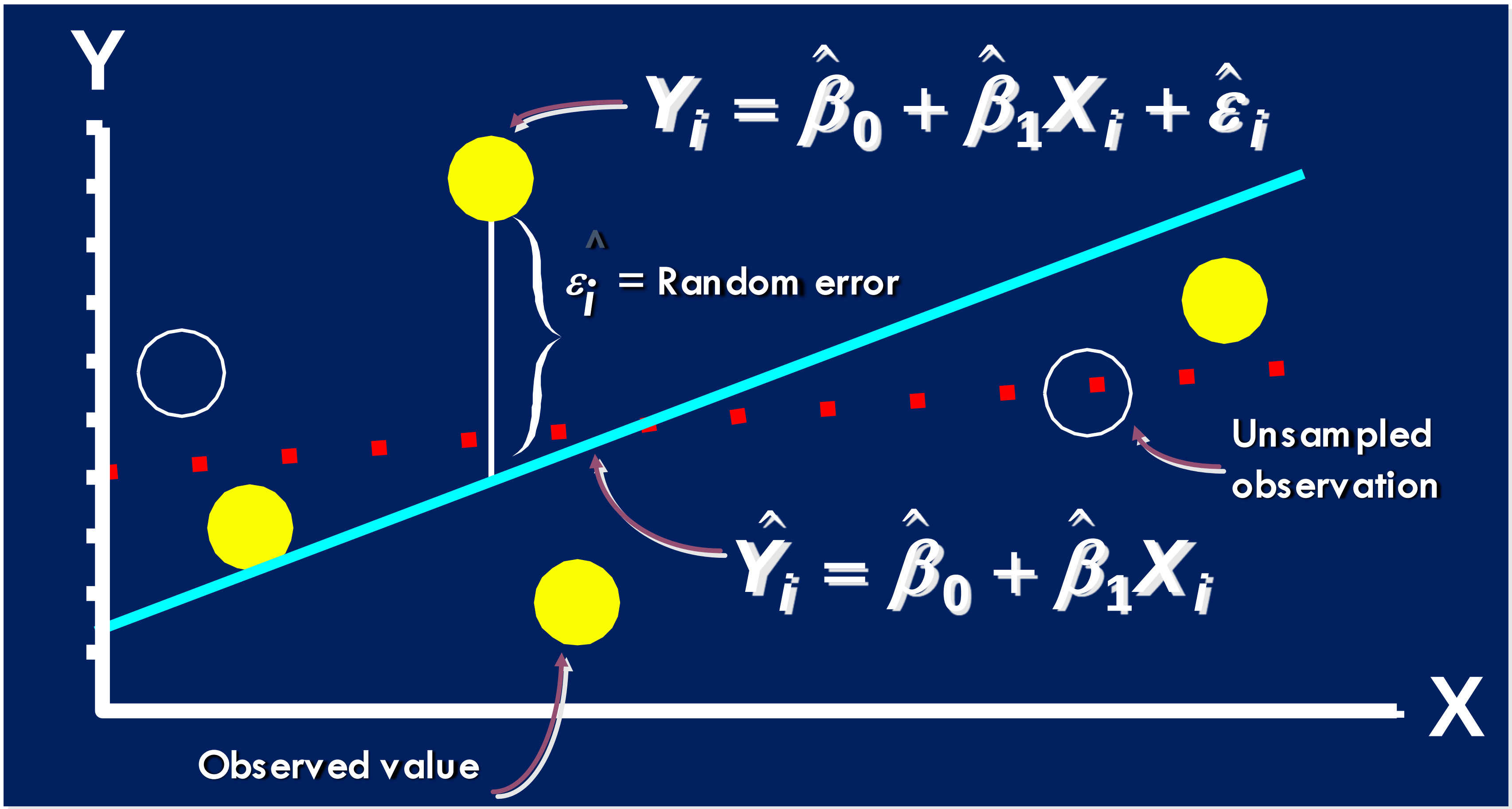
Population & Sample Regression Models



Population Linear Regression Model



Sample Linear Regression Model

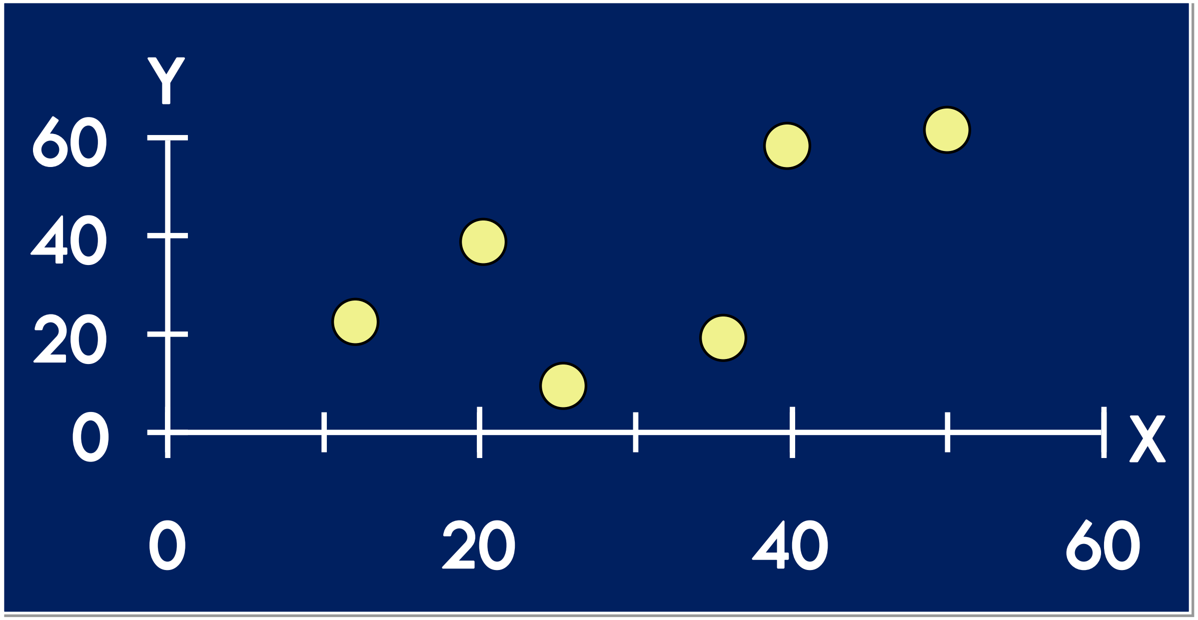


Regression Modeling Steps

- Hypothesize Deterministic Component
- **Estimate Unknown Model Parameters**
- Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- Evaluate Model
- Use Model for Prediction & Estimation

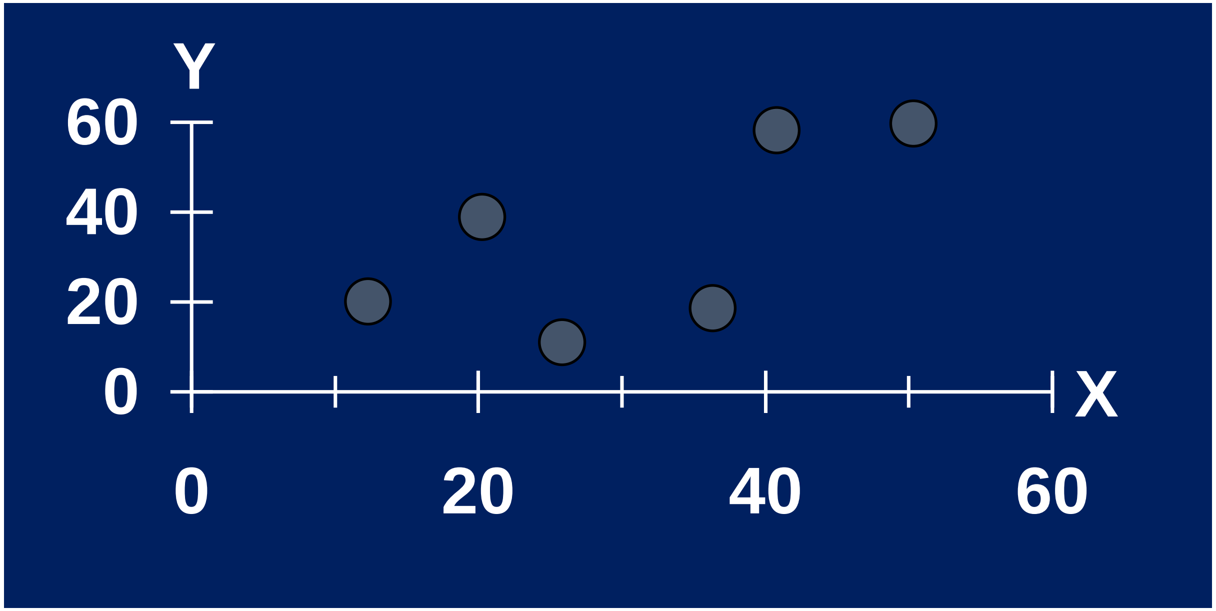
Scattergram

- Plot of All (X_i, Y_i) Pairs
- Suggests How Well Model Will Fit



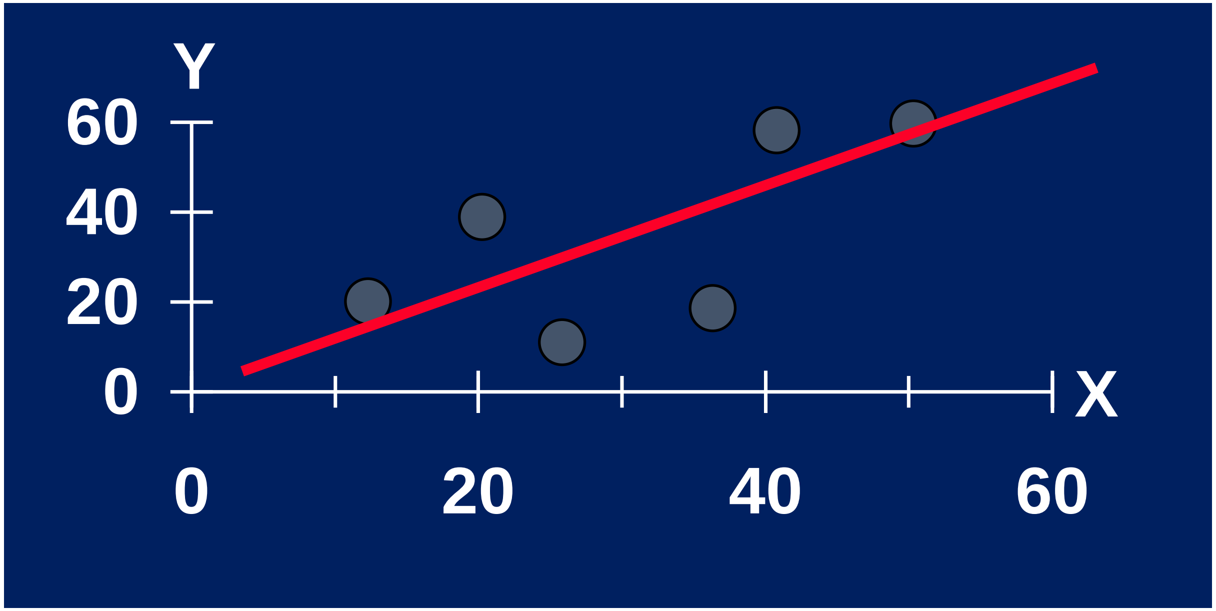
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



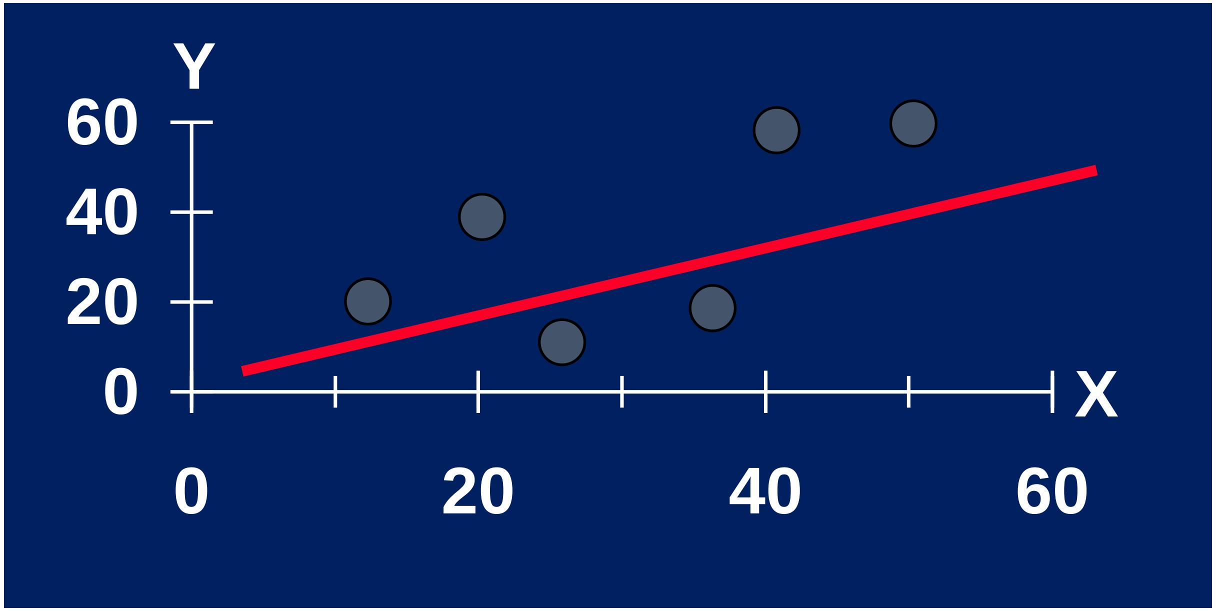
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



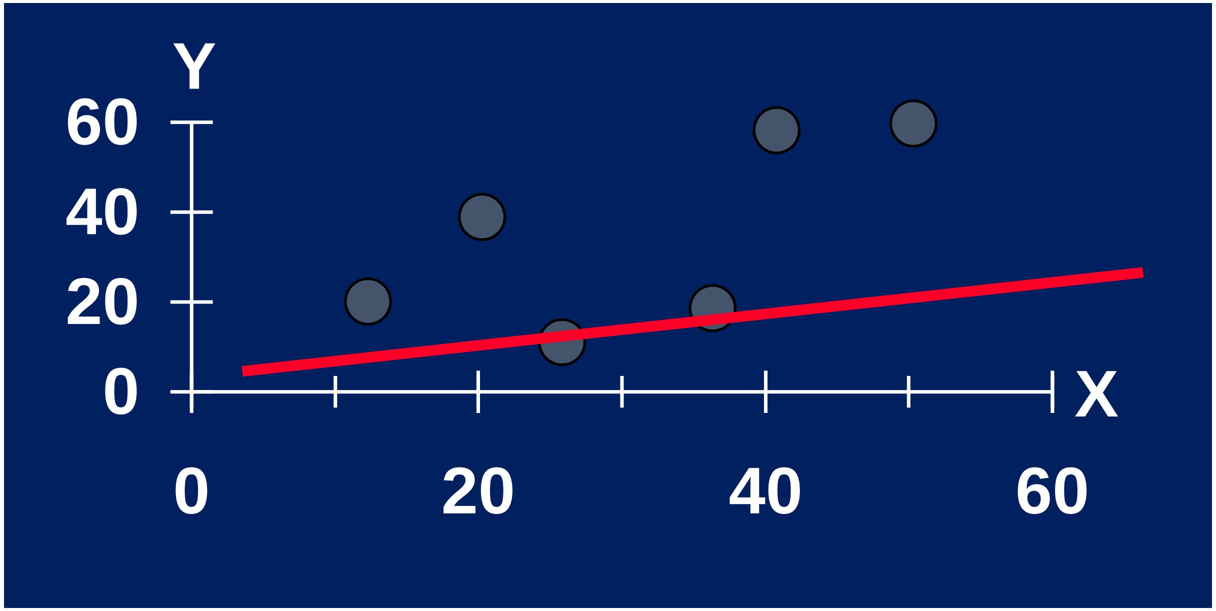
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



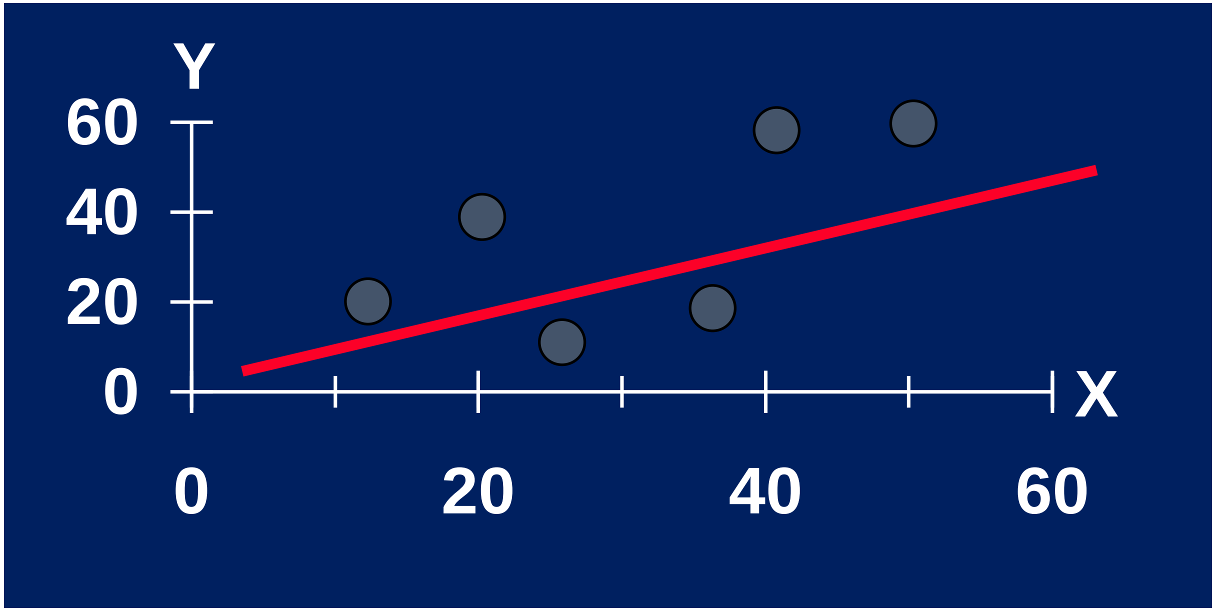
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



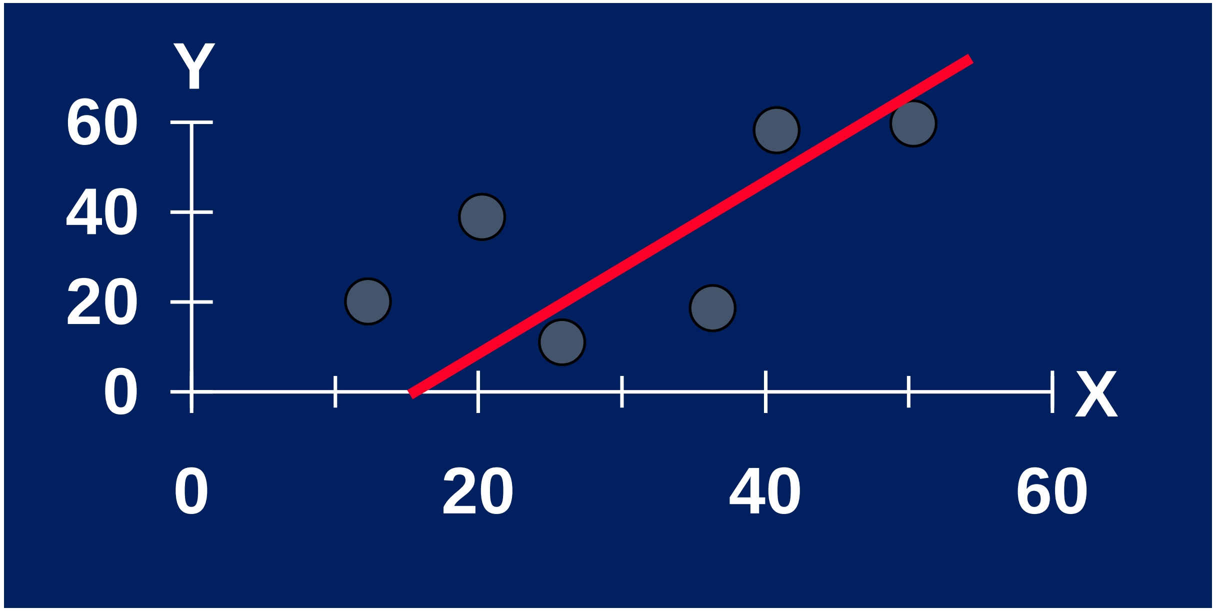
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



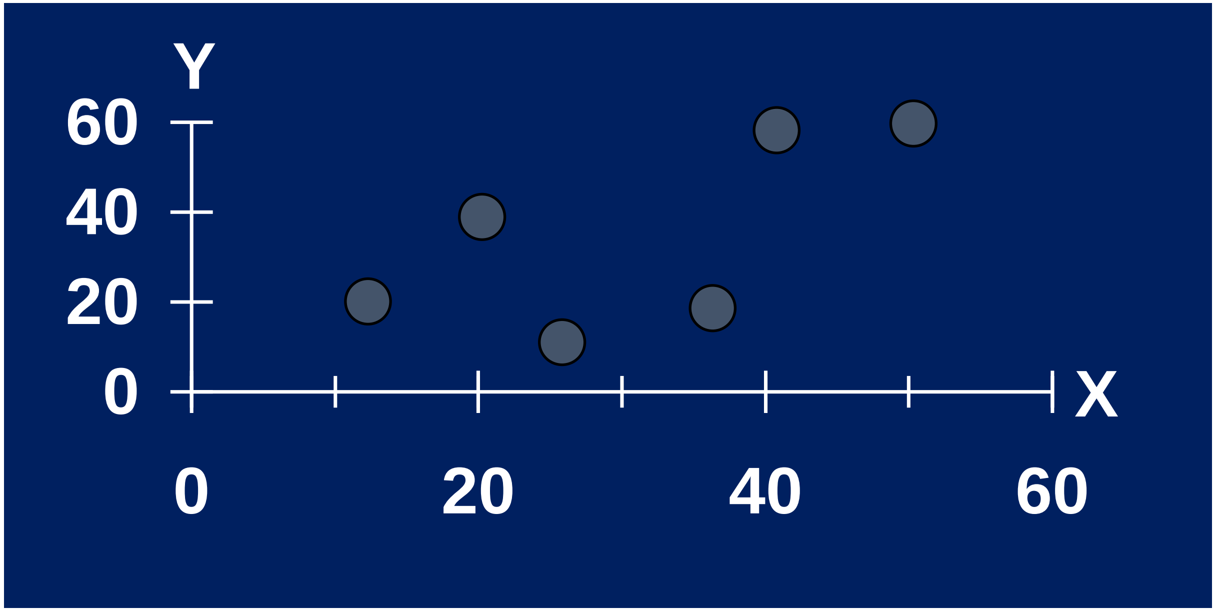
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Least Squares

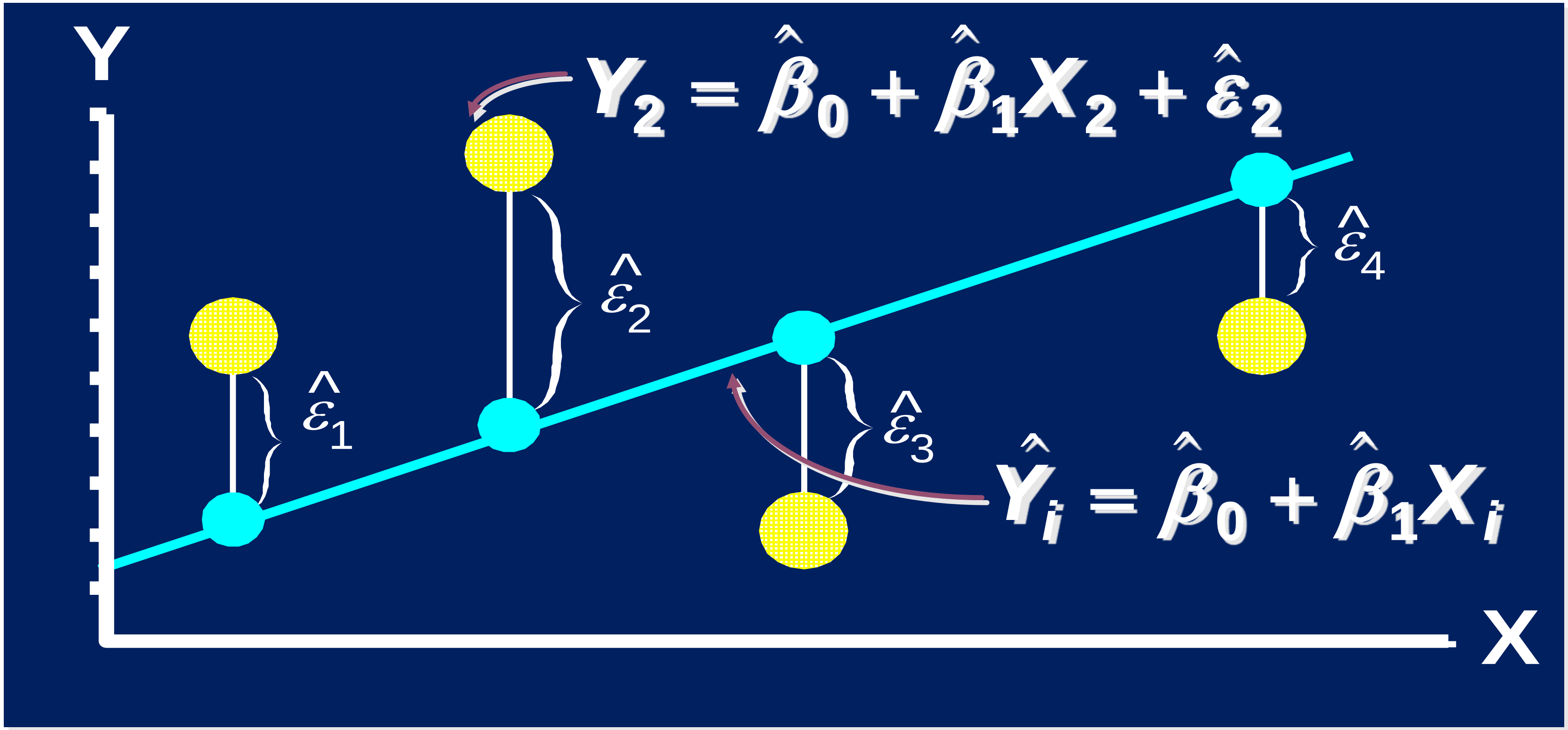
- 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum
  - *But* Positive Differences Off-Set Negative

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

LS Minimizes the Sum of the Squared Differences (SSE)

Least Squares Graphically

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



Computation Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

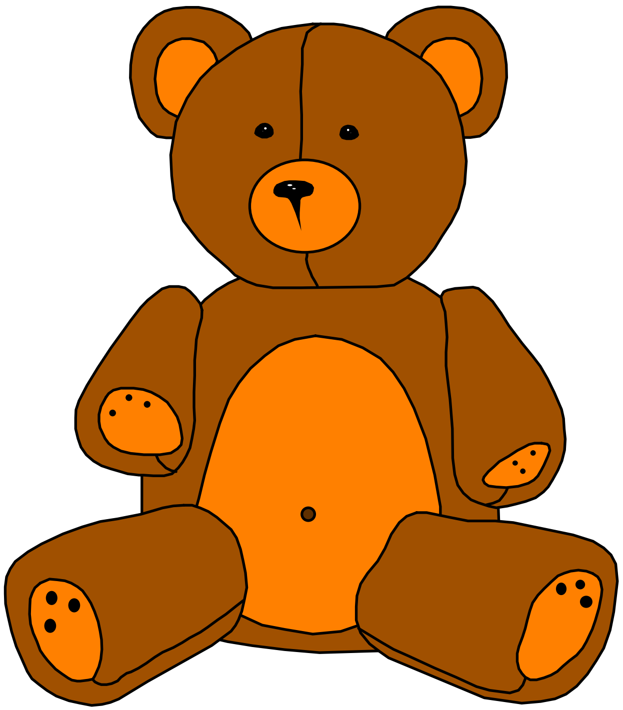
Interpretation of Coefficients

- 1. Slope ($\hat{\beta}_1$)
  - Estimated *Y* Changes by $\hat{\beta}_1$ for Each 1 Unit 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

Parameter Estimation Example

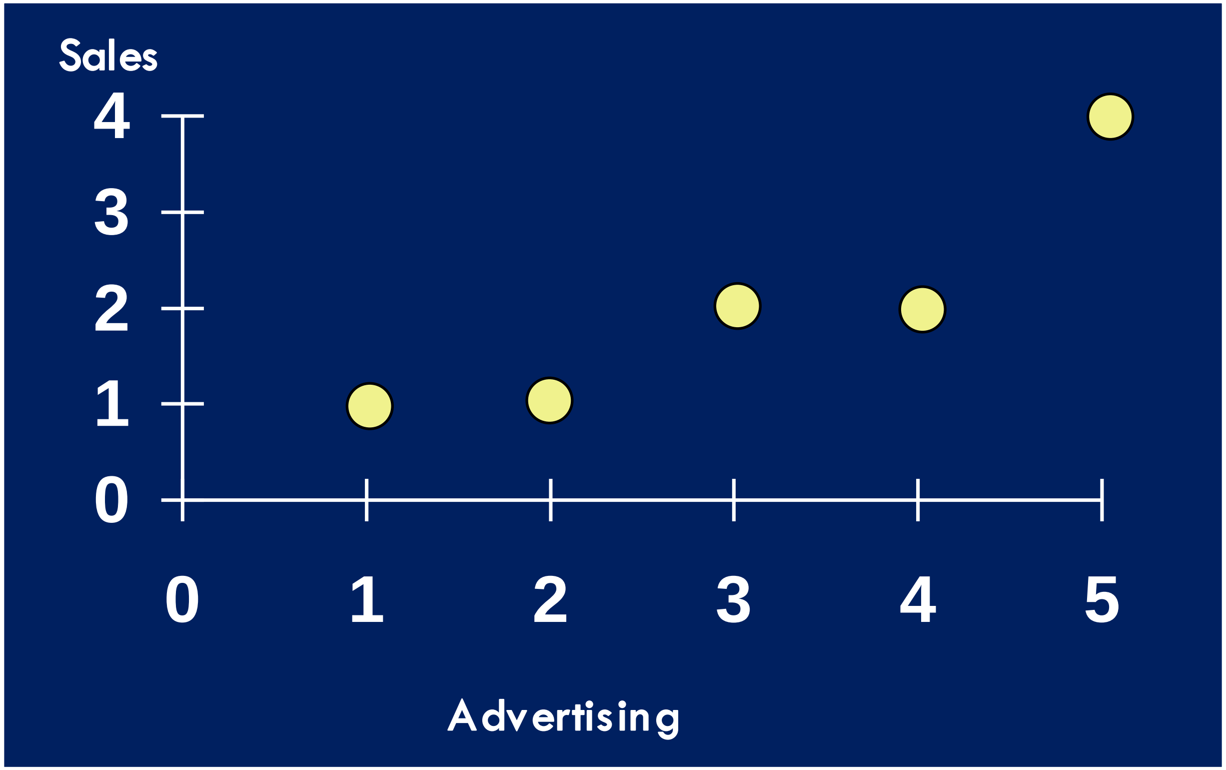
- You're a marketing analyst for Hasbro Toys. You gather the following data:

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



- What is the **relationship** between sales & advertising?

Scattergram
Sales vs. Advertising



Parameter Estimation Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = 0.70$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2 - (0.70)(3) = -0.10$$

Coefficient Interpretation Solution

- Slope ($\hat{\beta}_1$)
  - Sales Volume (Y) Is Expected to Increase by .7 Units for Each \$1 Increase in Advertising (X)
- 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

Parameter Estimation Computer Output

$\hat{\beta}_k$

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

$\hat{\beta}_0$ $\hat{\beta}_1$

Derivation of Parameter Equations

- Goal: Minimize squared error

$$0 = \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = \frac{\partial \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\partial \hat{\beta}_0}$$

$$= \sum -2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$$

$$= -2\left(n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x}\right)$$

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

Derivation of Parameter Equations

$$0 = \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = \frac{\partial \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2}{\partial \hat{\beta}_1}$$

$$= -2 \sum x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)$$

$$= -2 \sum x_i \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \right)$$

$$\hat{\beta}_1 \sum x_i (x_i - \bar{x}) = \sum x_i (y_i - \bar{y})$$

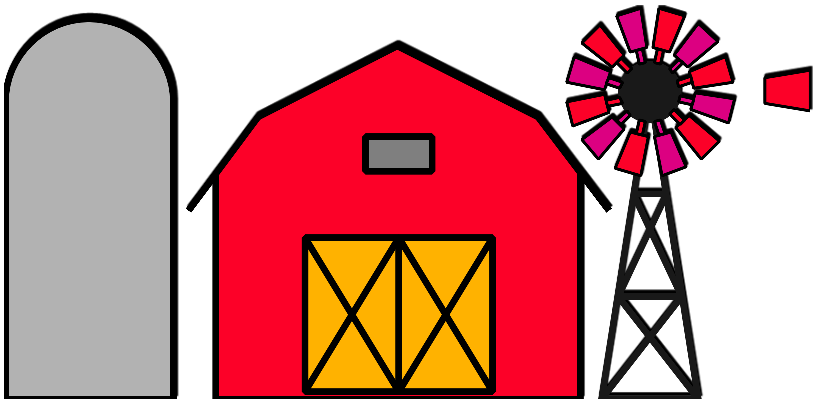
$$\hat{\beta}_1 \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Parameter Estimation Thinking Challenge

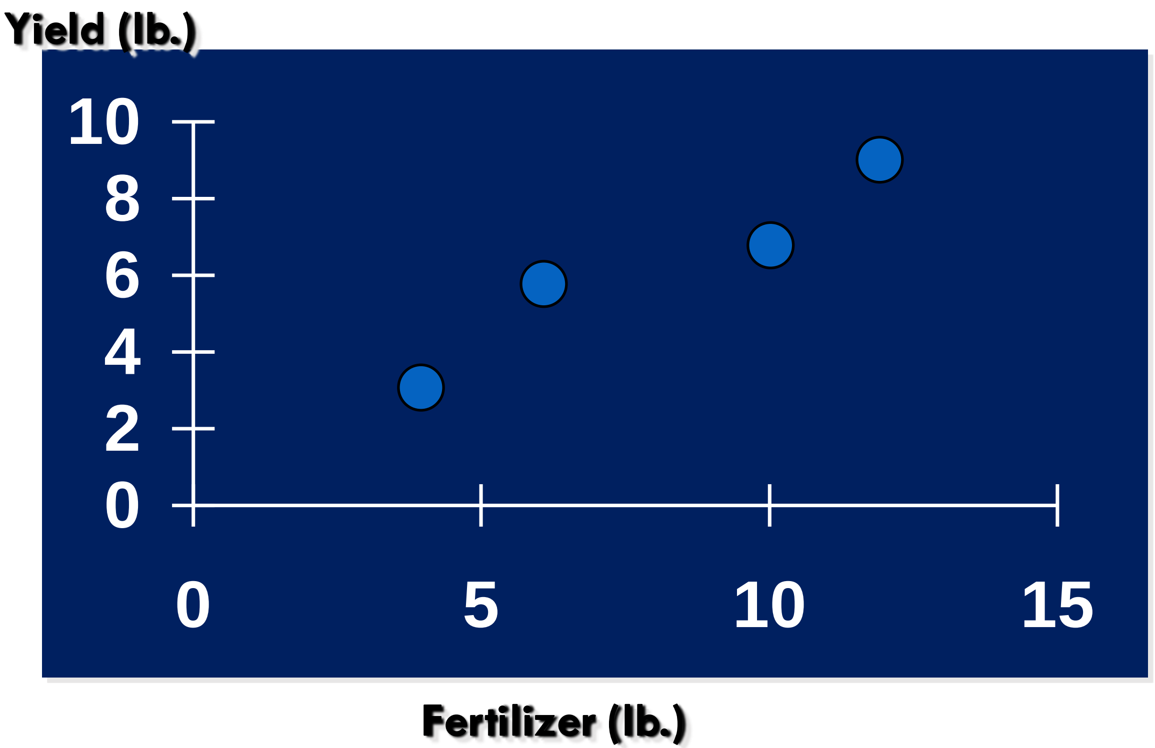
- You're an economist for the county cooperative. You gather the following data:

Fertilizer (lb.)	Yield (lb.)
4	3.0
6	5.5
10	6.5
12	9.0



- What is the **relationship** between fertilizer & crop yield?

Scattergram
Crop Yield vs. Fertilizer*



Parameter Estimation Solution Table*

X_i	Y_i	X_i^2	Y_i^2	X_iY_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*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = 0.65$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 6 - (0.65)(8) = 0.80$$

Coefficient Interpretation Solution*

- Slope ($\hat{\beta}_1$)
  - Crop Yield (*Y*) Is Expected to Increase by .65 lb. for Each 1 lb. Increase in Fertilizer (*X*)
- Y-Intercept ($\hat{\beta}_0$)
  - Average Crop Yield (*Y*) Is Expected to Be 0.8 lb. When No Fertilizer (*X*) Is Used

Regression Modeling Steps

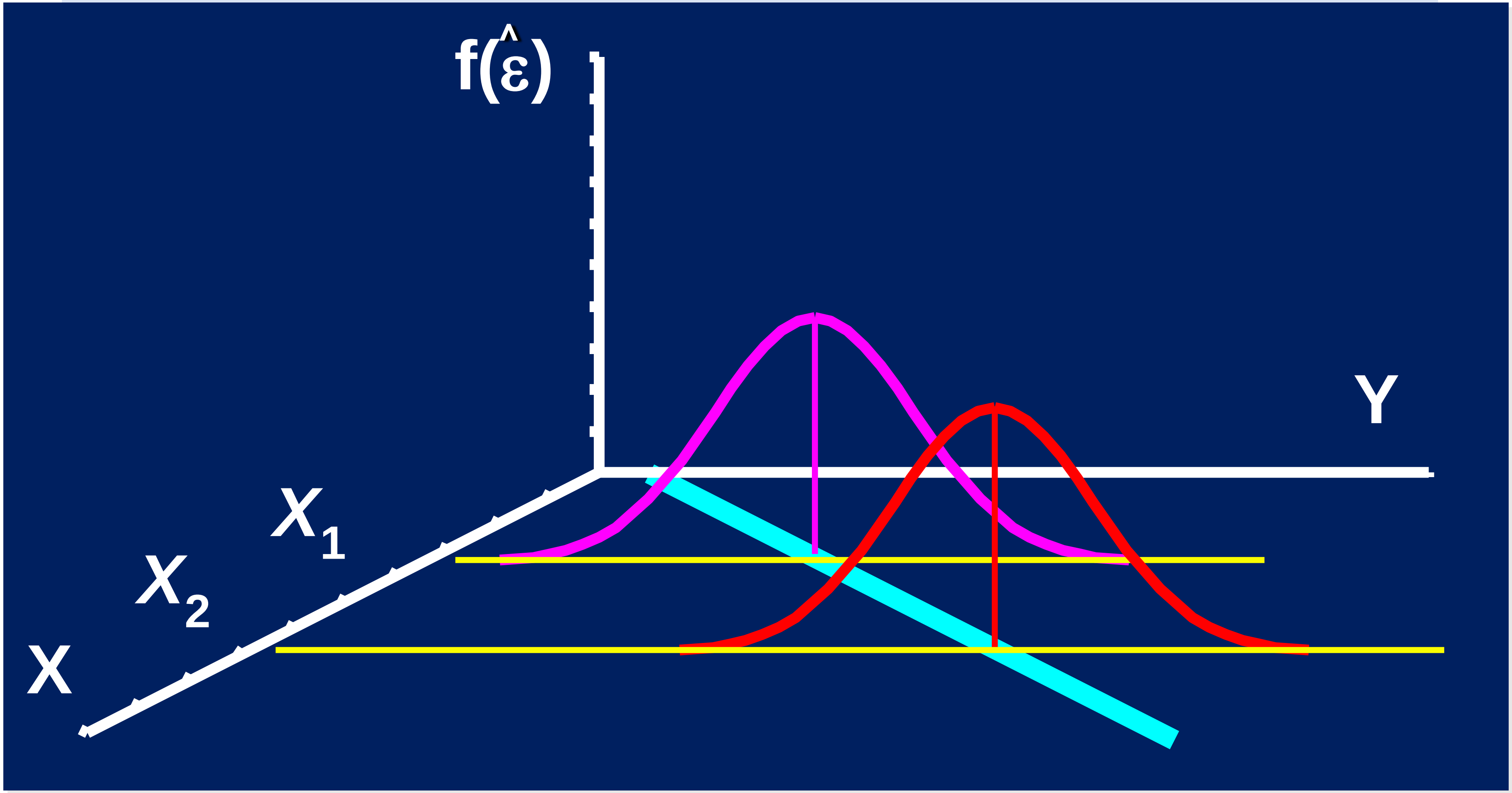
- Hypothesize Deterministic Component
- Estimate Unknown Model Parameters
- **Specify Probability Distribution of Random Error Term**
  - Estimate Standard Deviation of Error
- Evaluate Model
- Use Model for Prediction & Estimation

Linear Regression Assumptions

- Mean of Probability Distribution of Error Is 0
- Probability Distribution of Error Has Constant Variance

  Exercise: Constant across what?
- Probability Distribution of Error is Normal
- 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 ε, s
- Affects Several Factors
  - Parameter Significance
  - Prediction Accuracy

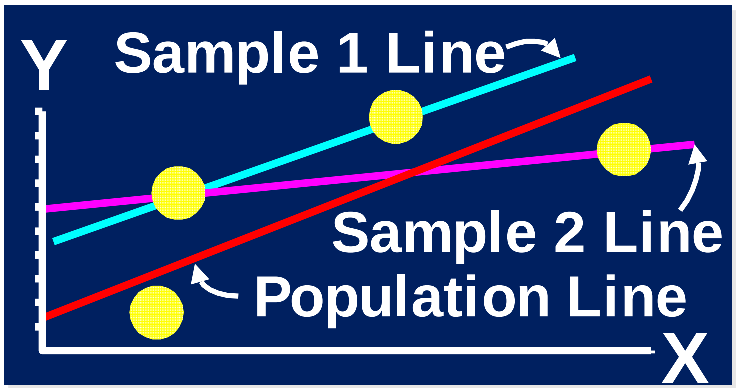
Regression Modeling Steps

- Hypothesize Deterministic Component
- Estimate Unknown Model Parameters
- Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- **Evaluate Model**
- Use Model for Prediction & Estimation

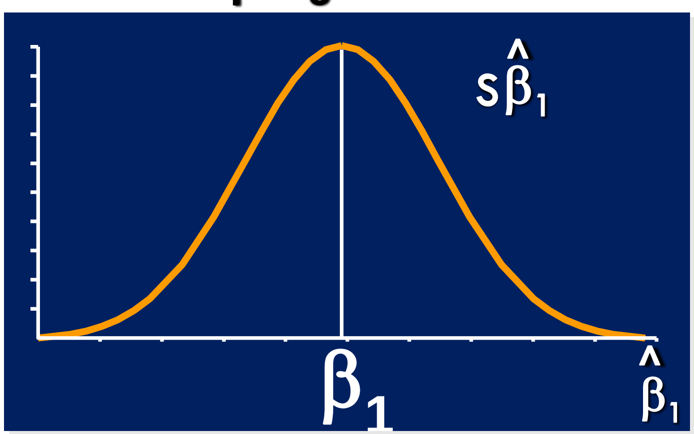
Test of Slope Coefficient

- Shows If There Is a Linear Relationship Between *X* & *Y*
- Involves Population Slope β_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

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

where

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}}$$

Test of Slope Coefficient Example

- You're a marketing analyst for Hasbro Toys. You find $\boldsymbol{b_0 = -.1}$, $\boldsymbol{b_1 = .7}$ & $\boldsymbol{s = .60553}$.

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

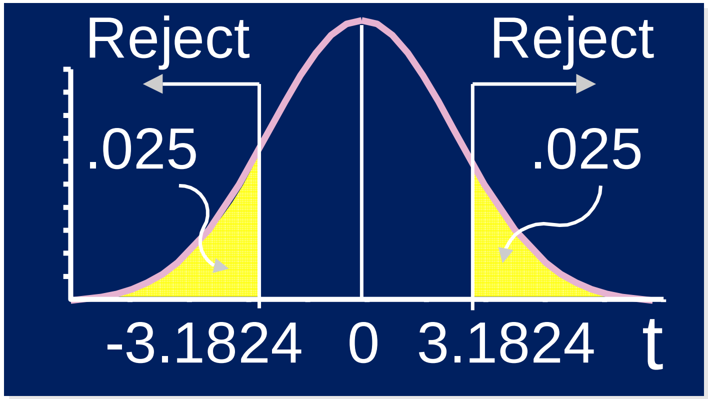
- Is the relationship **significant** at the **.05** level?

Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_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 of Slope Parameter Solution

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



Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{0.70 - 0}{0.1915} = +3.656$$

Decision:

Reject at $\alpha = .05$

Conclusion:

There is evidence of a relationship

Test Statistic Solution

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{0.70 - 0}{0.1915} = 3.656$$

where

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}} = \frac{0.60553}{\sqrt{55 - \frac{(15)^3}{5}}} = 0.1915$$

Test of Slope Parameter 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

$\hat{\beta}_k$

$S_{\hat{\beta}_k}$

$t = \hat{\beta}_k / S_{\hat{\beta}_k}$

P-Value

Measures of Variation in Regression

- Total Sum of Squares (SS_{yy})

  Measures Variation of Observed Y_i Around the Mean $\overline{Y}$

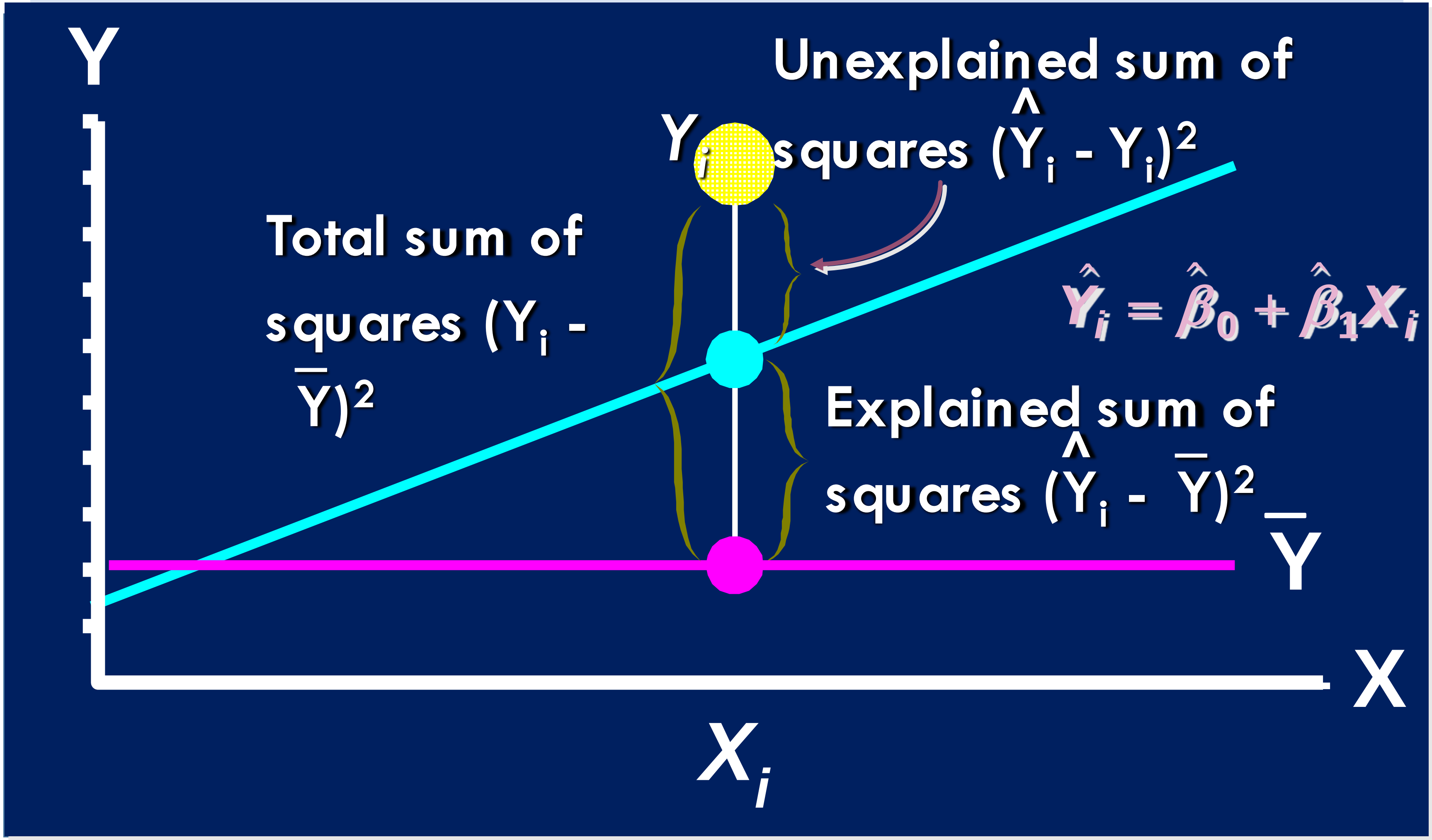
- Explained Variation (SSR)

  Variation Due to Relationship Between X & Y

- Unexplained Variation (SSE)

  Variation Due to Other Factors

Variation Measures

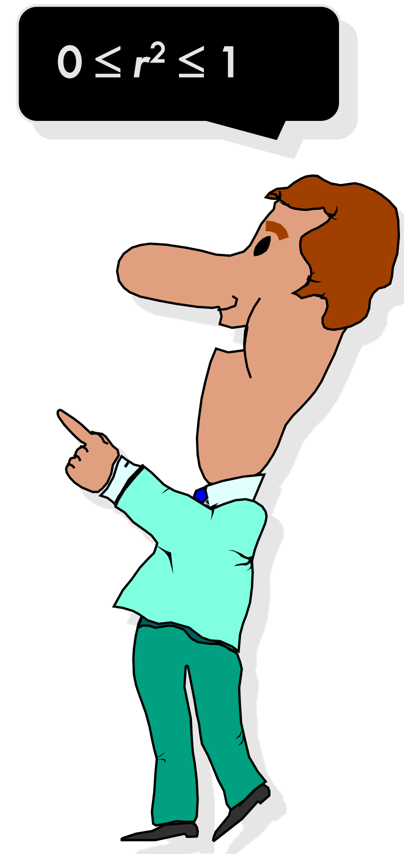


Coefficient of Determination

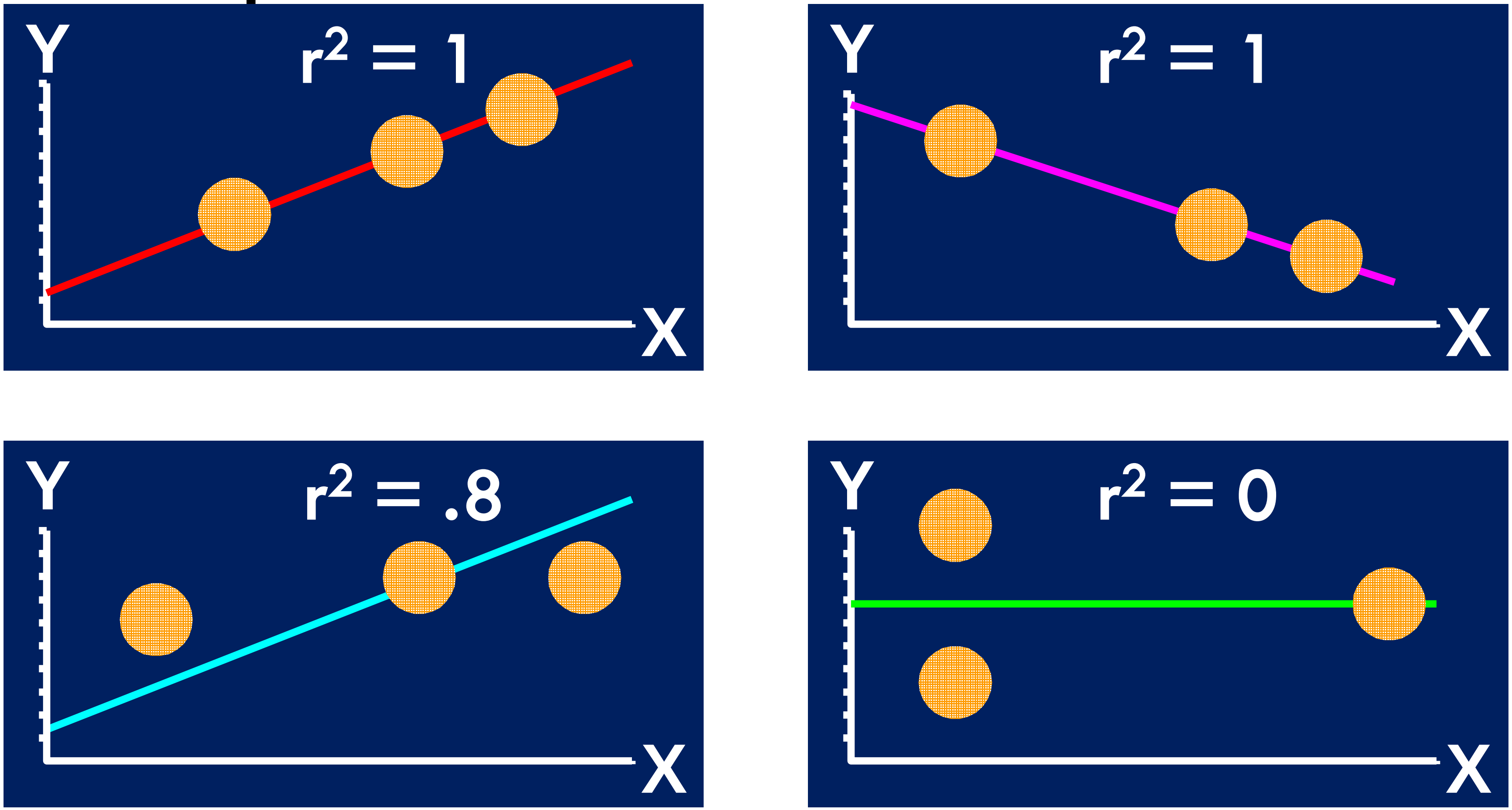
- **Proportion** of Variation 'Explained' by Relationship Between *X* & *Y*

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$= \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$



Coefficient of Determination Examples



Coefficient of Determination Example

You're a marketing analyst for Hasbro Toys.

You find $\hat{\beta}_0 = -0.1$ & $\hat{\beta}_1 = 0.7$.

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

Interpret a **coefficient of determination** of **0.8167.**

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

S

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

Regression Modeling Steps

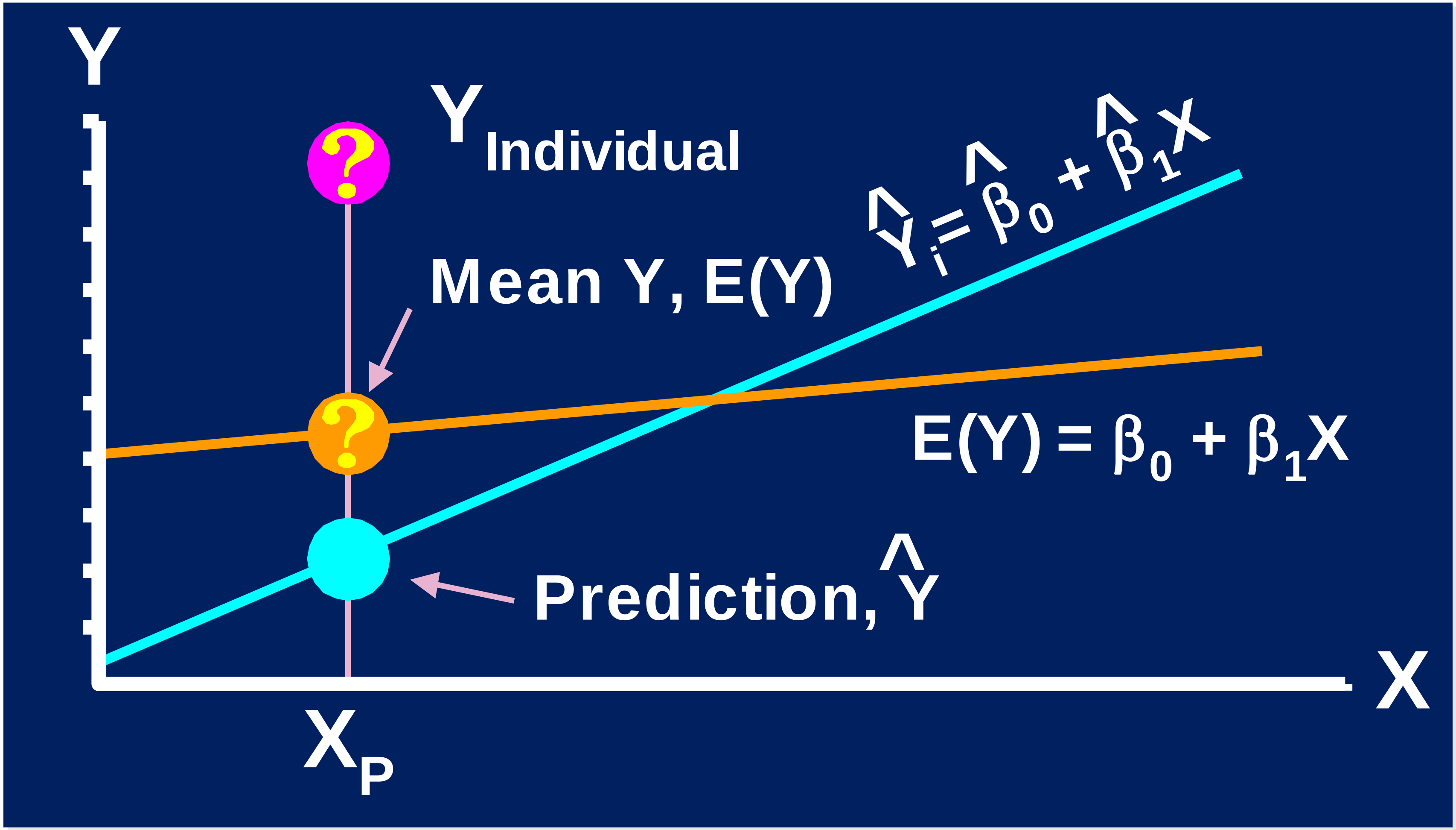
- Hypothesize Deterministic Component
- Estimate Unknown Model Parameters
- Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- Evaluate Model
- **Use Model for Prediction & Estimation**

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 of Mean *Y*

$$\hat{Y} - t_{n-2,\alpha/2} \cdot S_{\hat{Y}} \le E(Y) \le \hat{Y} + t_{n-2,\alpha/2} \cdot S_{\hat{Y}}$$

where

$$S_{\hat{Y}} = S\sqrt{\frac{1}{n} + \frac{\left(X_p - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}$$

Factors Affecting Interval Width

- Level of Confidence ($1 - \alpha$)

  Width Increases as Confidence Increases

- Data Dispersion (s)

  Width Increases as Variation Increases

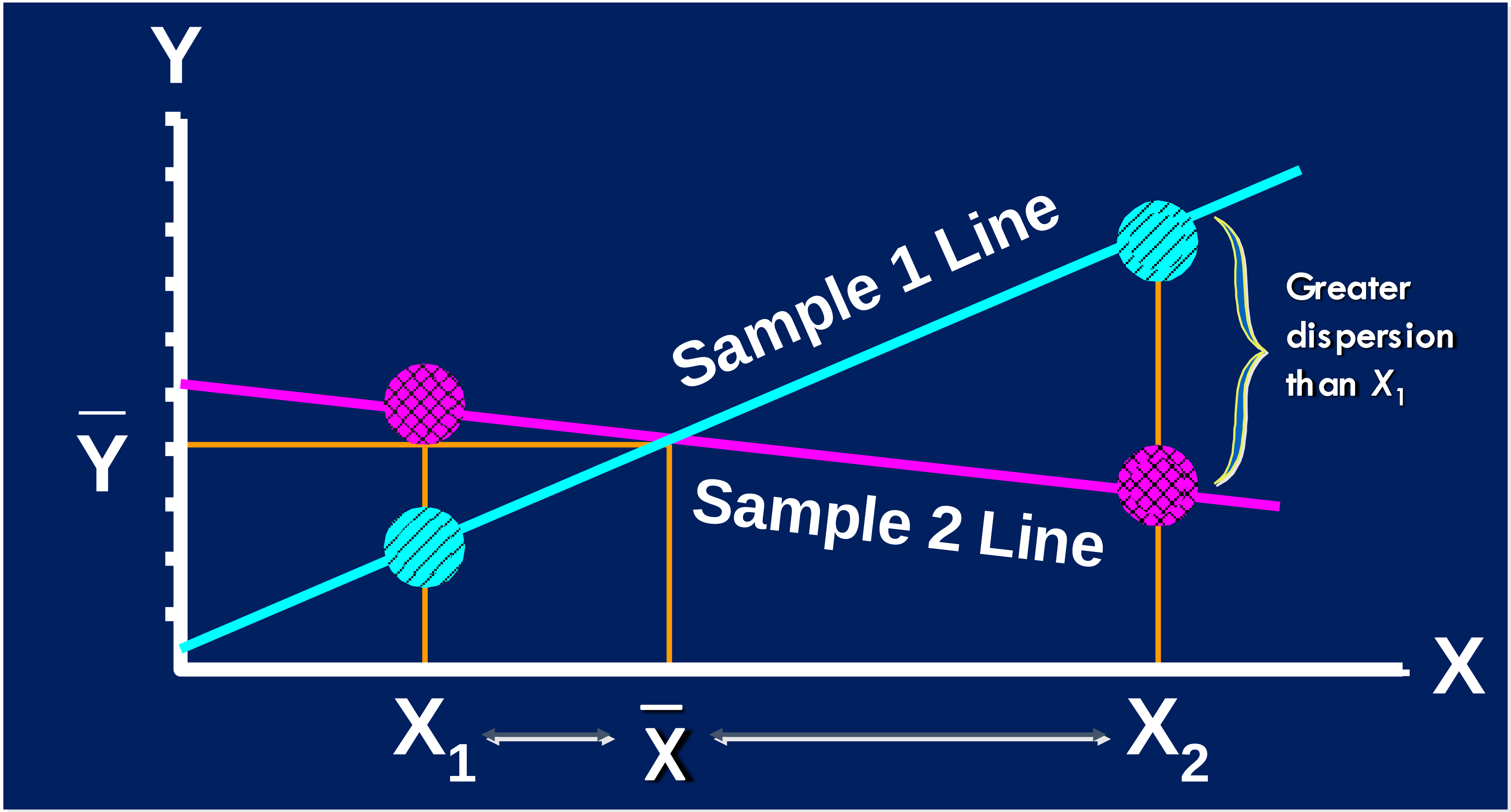
- Sample Size

  Width Decreases as Sample Size Increases

- 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 $\boldsymbol{b_0}$ **= -.1,** $\boldsymbol{b_1}$ **= .7** & $\boldsymbol{s}$ **= .60553**.

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

Estimate the **mean** sales when advertising is **$4** at the **.05** level.

Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_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**

Confidence Interval Estimate Solution

$$\hat{Y} - t_{n-2,\alpha/2} \cdot S_{\hat{Y}} \le E(Y) \le \hat{Y} + t_{n-2,\alpha/2} \cdot S_{\hat{Y}}$$

$$\hat{Y} = -0.1 + (0.7)(4) = 2.7$$

X to be predicted

$$S_{\hat{Y}} = .60553\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}} = 0.3316$$

$$2.7 - (3.1824)(0.3316) \le E(Y) \le 2.7 + (3.1824)(0.3316)$$

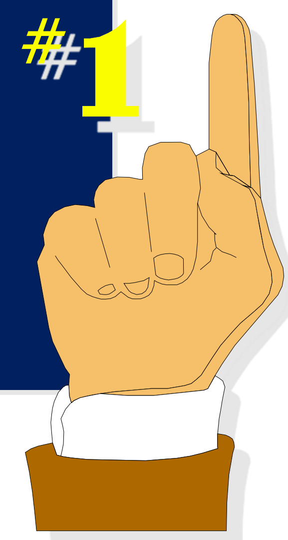
$$1.6445 \le E(Y) \le 3.7553$$

Prediction Interval of Individual Response

$$\hat{Y} - t_{n-2,\alpha/2} \cdot S_{(Y-\hat{Y})} \le Y_P \le \hat{Y} + t_{n-2,\alpha/2} \cdot S_{(Y-\hat{Y})}$$

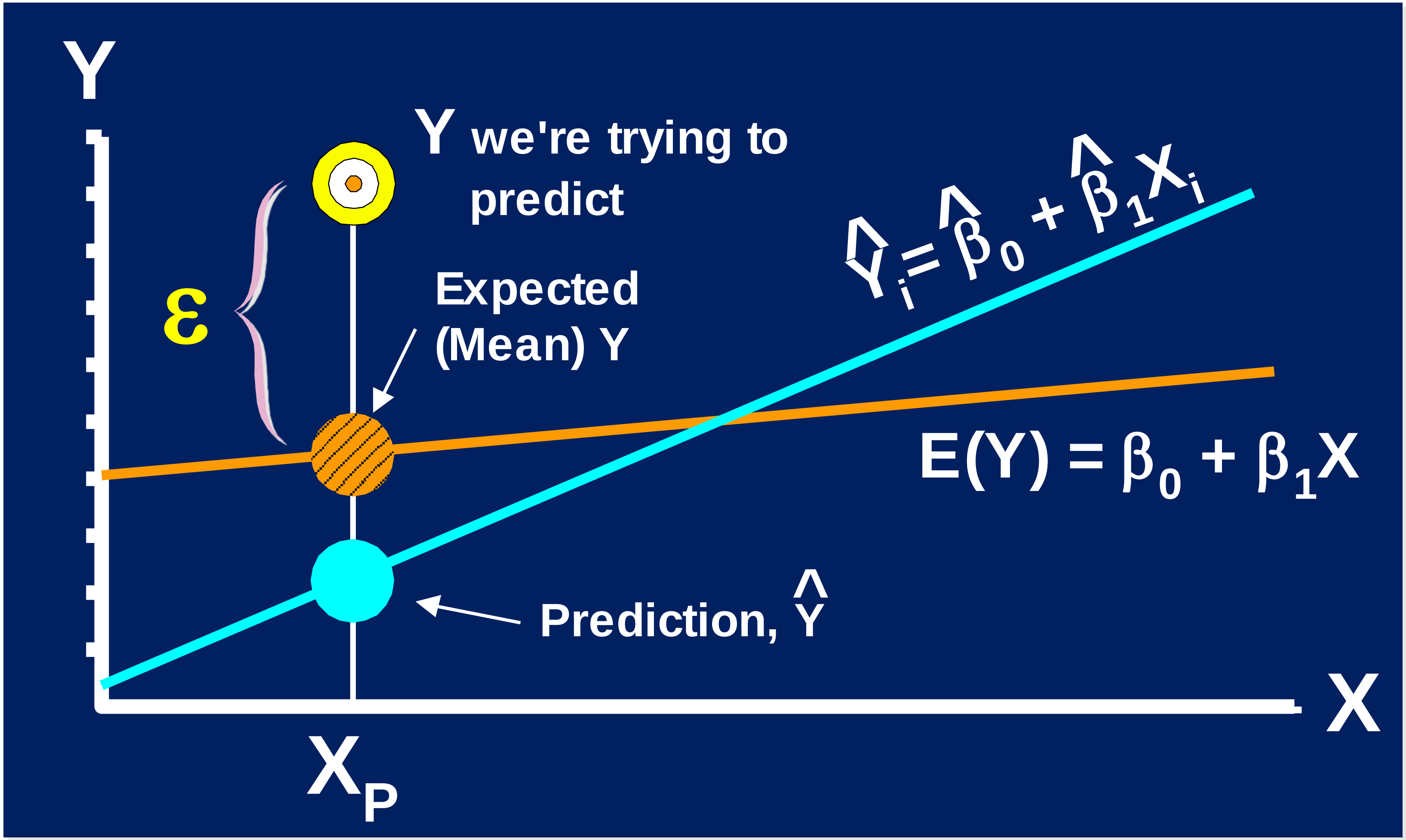
where

$$S_{(Y-\hat{Y})} = S\sqrt{1 + \frac{1}{n} + \frac{\left(X_P - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}$$



Note! The 1 under the radical in the standard error formula.
The effect of the extra Syx is to increase the width of the interval.
This will be seen in the interval bands.

Why the Extra 'S'?



Interval Estimate Computer Output

Obs	Dep Var SALES	Pred Value	Std Err Predict	Low95% Mean	Upp95% Mean	Low95% Predict	Upp95% Predict
1	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

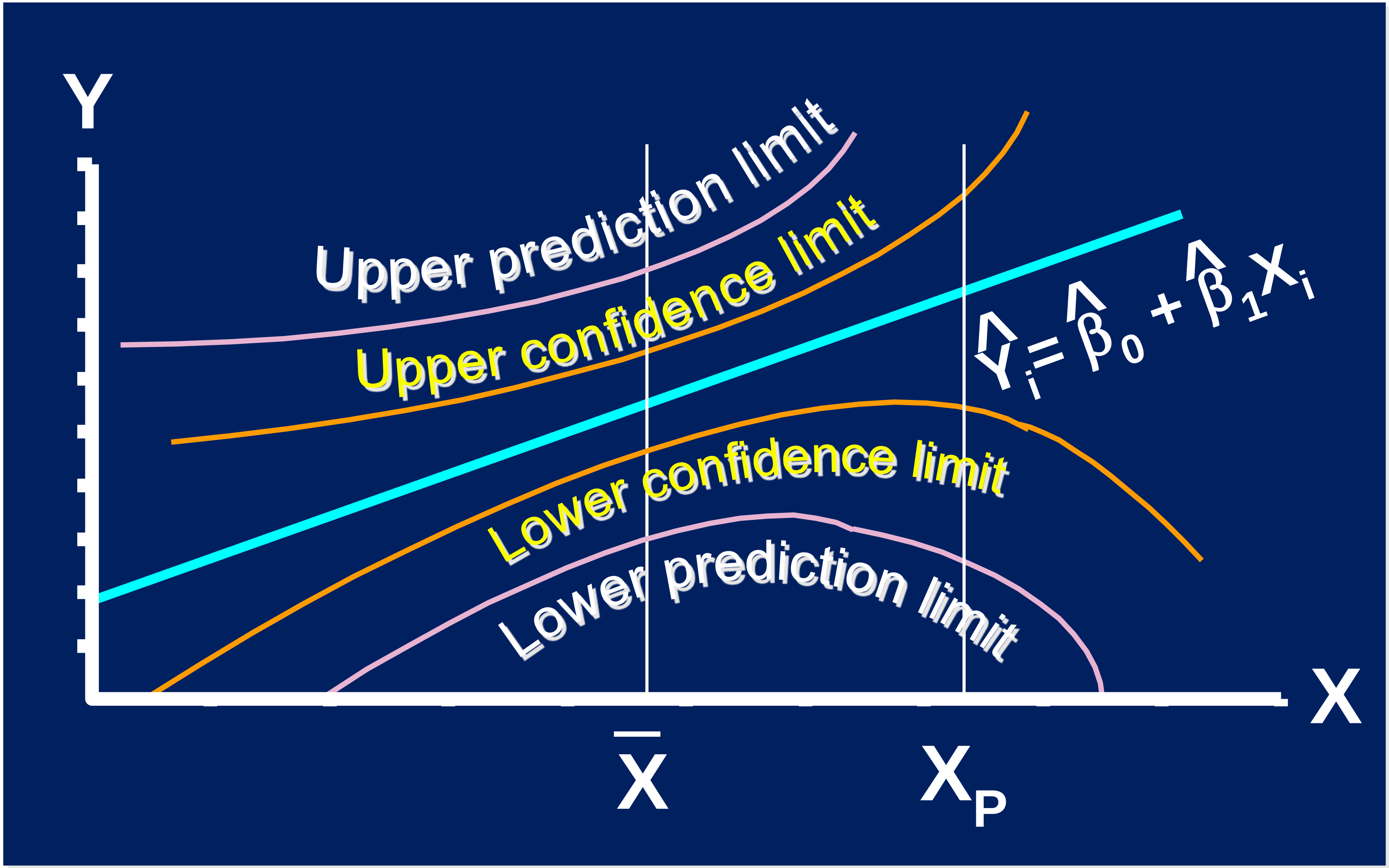
Predicted Y when X = 4

$S_{\hat{Y}}$

Confidence Interval

Prediction Interval

Hyperbolic Interval Bands



Correlation Models

- Answer '**How Strong** Is the Linear Relationship Between 2 Variables?'
- Coefficient of Correlation Used
  - Population Correlation Coefficient Denoted ρ (Rho)
  - Values Range from -1 to +1
  - Measures Degree of Association
- Used Mainly for Understanding

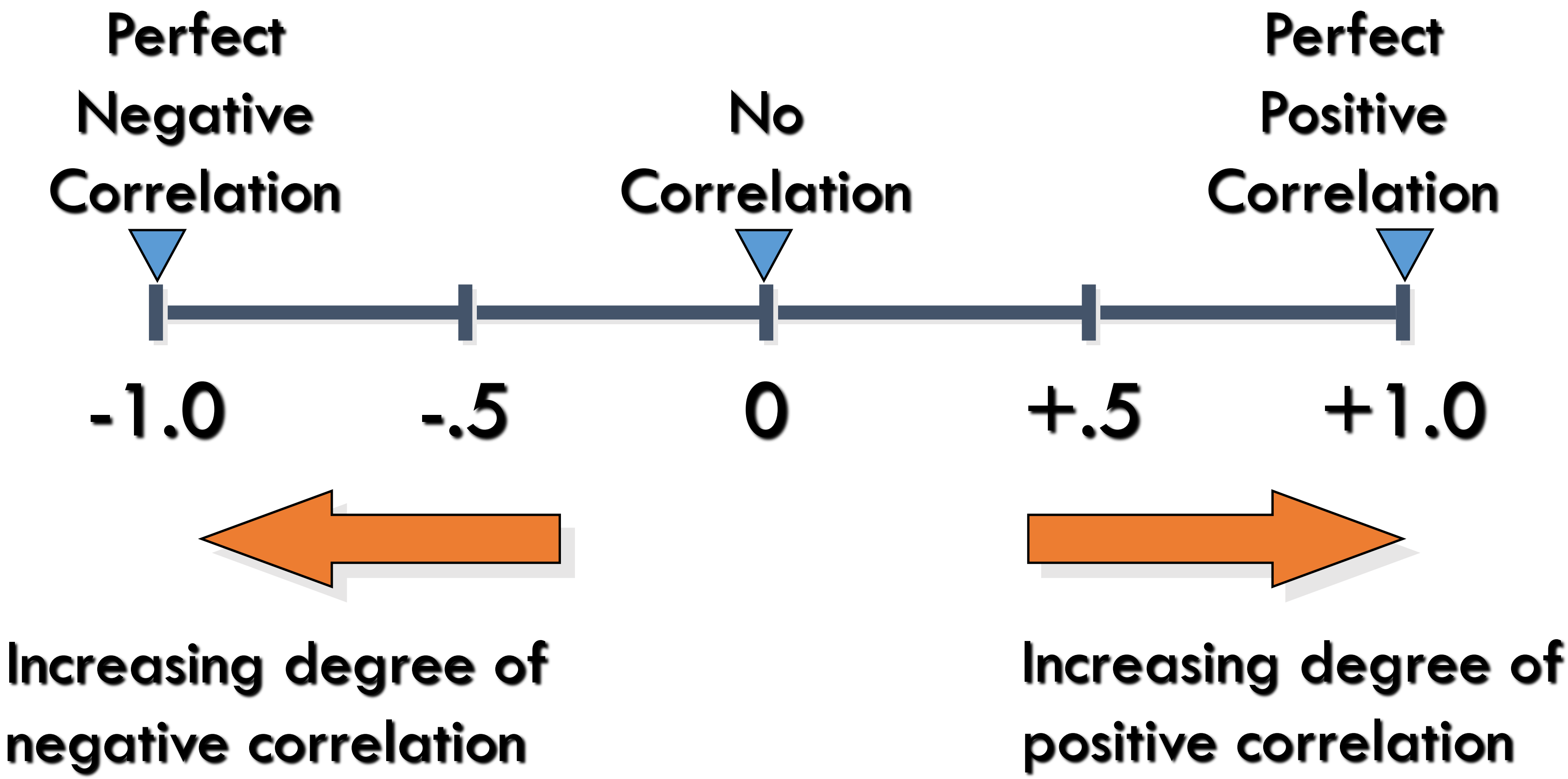
Sample Coefficient of Correlation

- Pearson Product Moment Coefficient of Correlation, *r*:

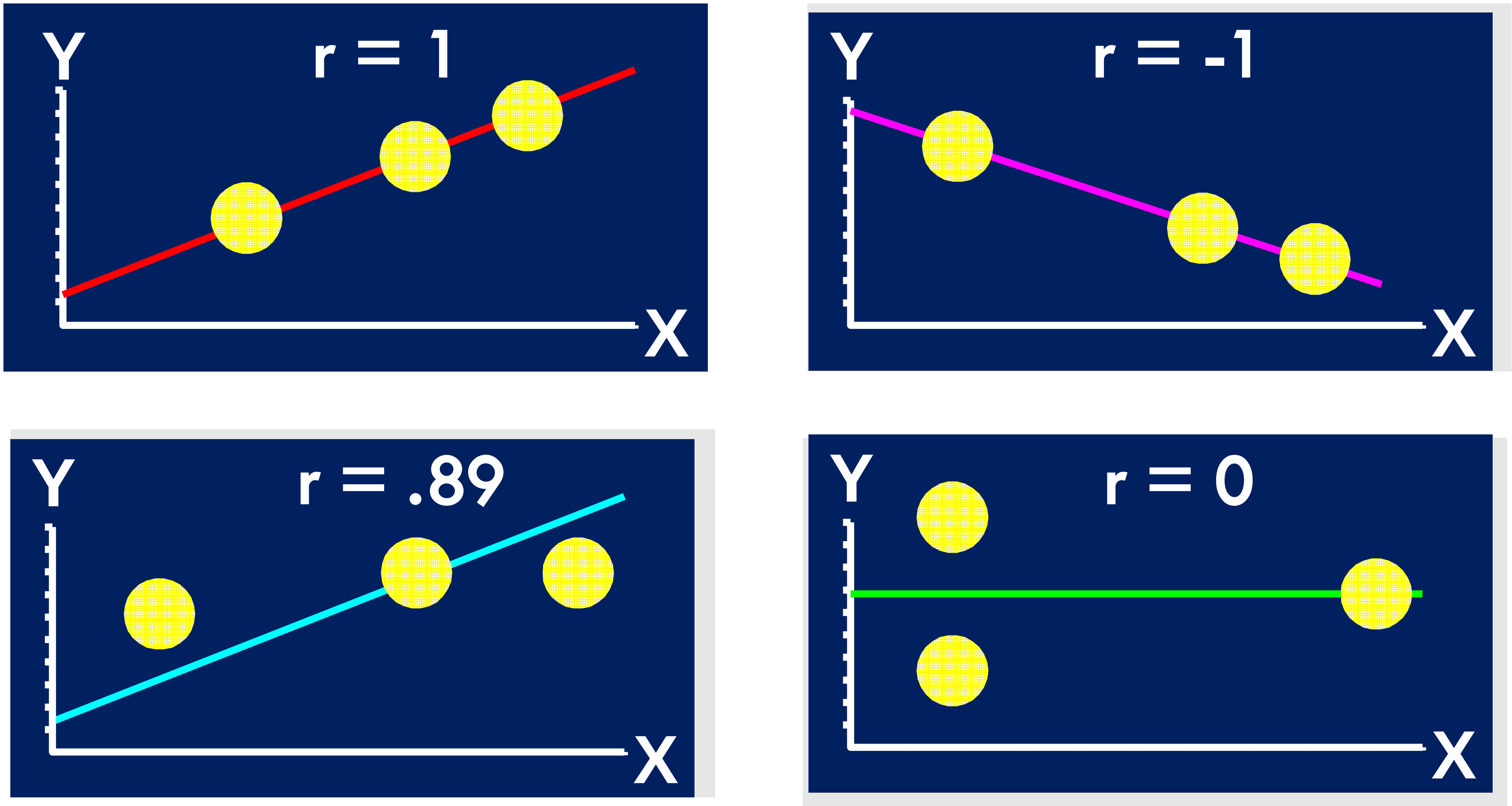
$$r = \sqrt{\text{Coefficien t of Determinat ion}}$$

$$= \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Coefficient of Correlation Values



Coefficient of Correlation Examples



Test of Coefficient of Correlation

- Shows If There Is a Linear Relationship Between 2 Numerical Variables
- Same Conclusion as Testing Population Slope β_1
- Hypotheses
  - H_0: $\rho = 0$ (No Correlation)
  - H_a: $\rho \neq 0$ (Correlation)

Summary

- Described the Linear Regression Model
- Stated the Regression Modeling Steps
- Explained Ordinary Least Squares
- Computed Regression Coefficients
- Predicted Response Variable
- Interpreted Computer Output

Thank You

"We trust in GOD, all others must bring data"