Data Analysis

Part I
Correlation Models:
Covariance & Correlation

Part II
Regression Models
(Simple & Multiple) Linear Regression

Data Analysis – Covariance & Correlation

Classroom Hands-on

- Open RStudio
- Work with the "**`CorrCoeff.Rmd`**" file

Line fitting:

- Example of Method of Least Squares
- Revise the material from Bilinear Algebra (**homework**)

Data Analysis – Covariance & Correlation

Anscombe's quartet

- But just because fitting a line is easy doesn't mean that it always makes sense.
- Let's take another look at **Anscombe's quartet** to underscore this point.
- Anscombe's Quartet was developed by statistician Francis Anscombe.
- It comprises four datasets, each containing eleven (x, y) −pairs.
- The essential thing to note about these datasets is that they share the **same descriptive statistics**.
- But things change completely, and I must emphasize **COMPLETELY,** when they are graphed.
- Each graph tells a different story irrespective of their similar summary statistics.

Data Analysis – Covariance & Correlation

Anscombe's quartet

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

- mean of x = 9
- mean of y = 7.50

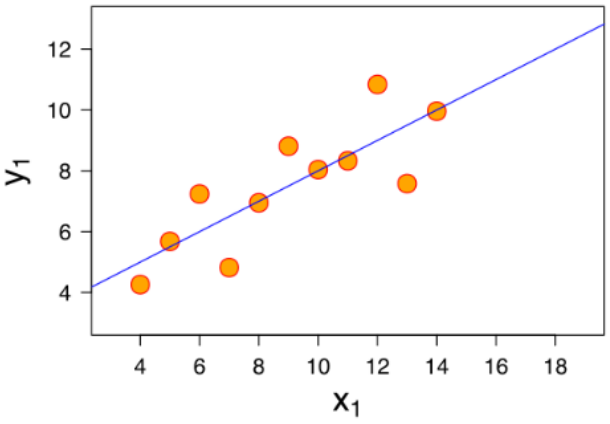
- variance of x = 11
- variance of y = 4.13

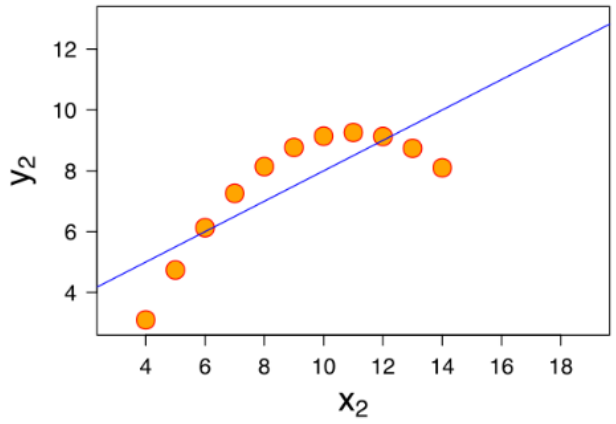
- The **correlation coefficient** (how strong a relationship is between 2 variables) between x and y is 0.816 for each dataset.

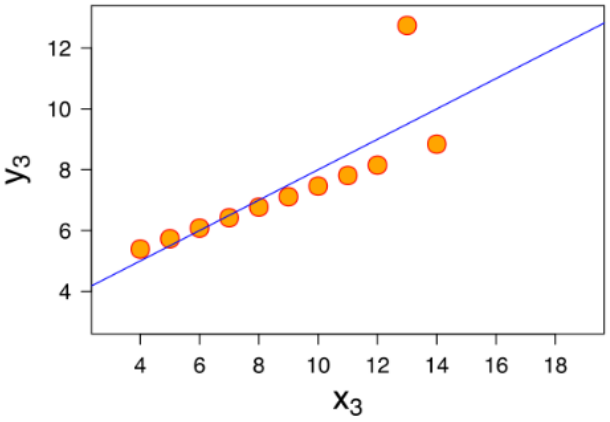
Data Analysis – Covariance & Correlation

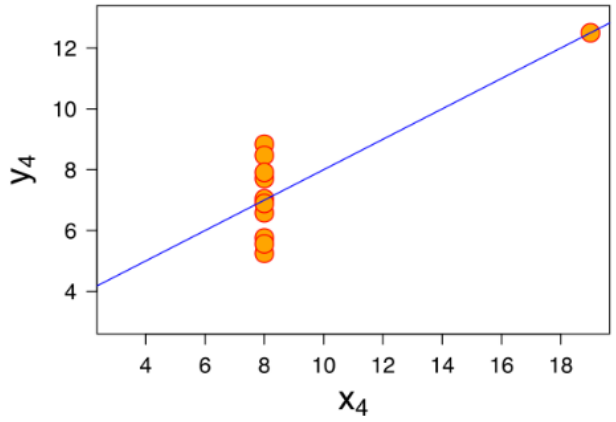
Anscombe's quartet

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same **regression lines** as well, but each dataset is telling a different story:









- **Dataset I** appears to have clean and well-fitting linear model.
- **Dataset II** is not distributed normally.
- In **Dataset III** the distribution is linear, but the calculated regression is thrown off by an outlier.
- **Dataset IV** shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of **visualization** in Data Analysis. **Looking** at the data reveals a lot of the structure and a clear picture of the dataset.

Data Analysis – Covariance & Correlation

Classroom Hands-on

- Open RStudio
- Type "**`Anscombe or knitr::kable(anscombe)`**" (it's a data set)
- Let's do the simple descriptive statistics on each data set
- Calculate the mean of x and y (there are 4 sets)
- Calculate the SD of x and y (there are 4 sets)
- Calculate the correlation between x and y (there are 4 sets)
- What do you observe?
- Plot the different pairs of x vs y

Data Analysis

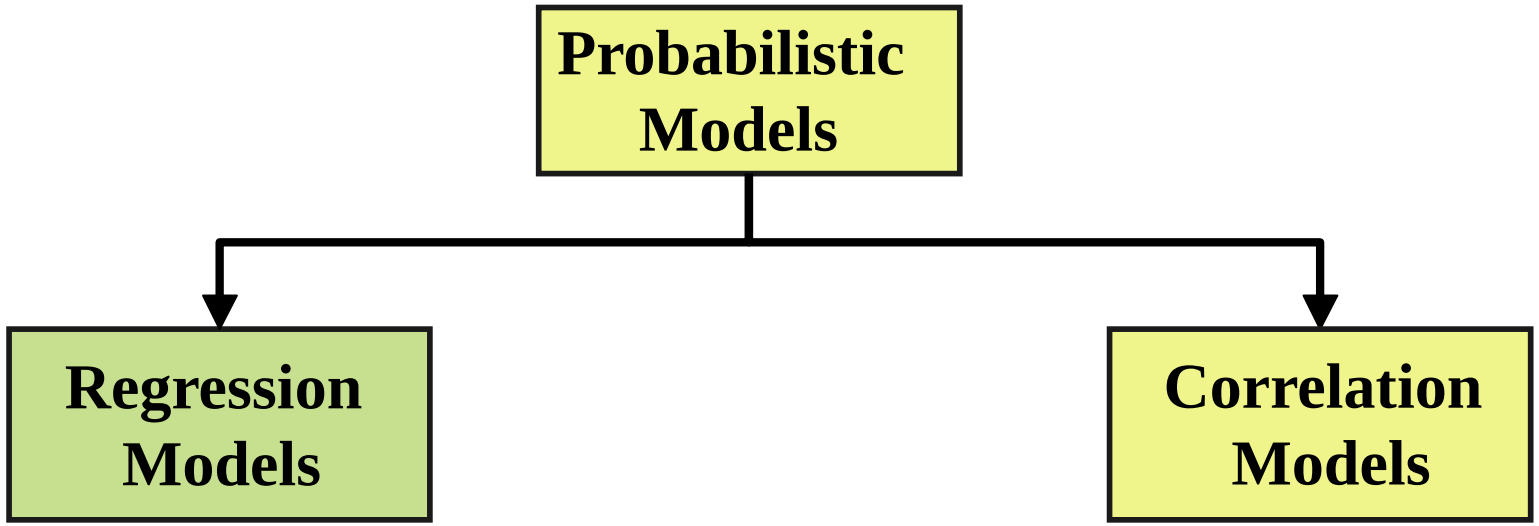
Part II

Regression Models

(Simple & Multiple) Linear Regression

Data Analysis – Simple Linear Regression

Types of Probabilistic Models



Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

Regression Modeling Steps

Model Specification

Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

Specifying the Model

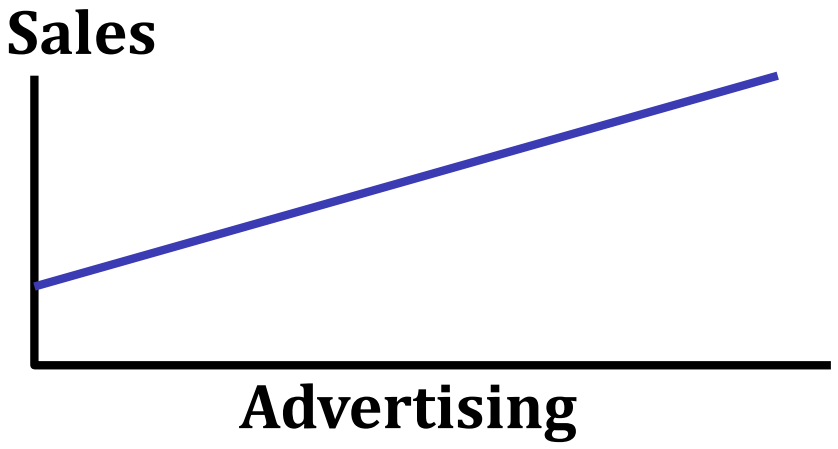
1. Define variables
    - Conceptual (e.g., Advertising, price)
    - Empirical (e.g., List price, regular price)
    - Measurement (e.g., $, Units)
2. Hypothesize nature of relationship
    - Expected effects (i.e., Coefficients' signs)
    - Functional form (linear or non-linear)
    - Interactions

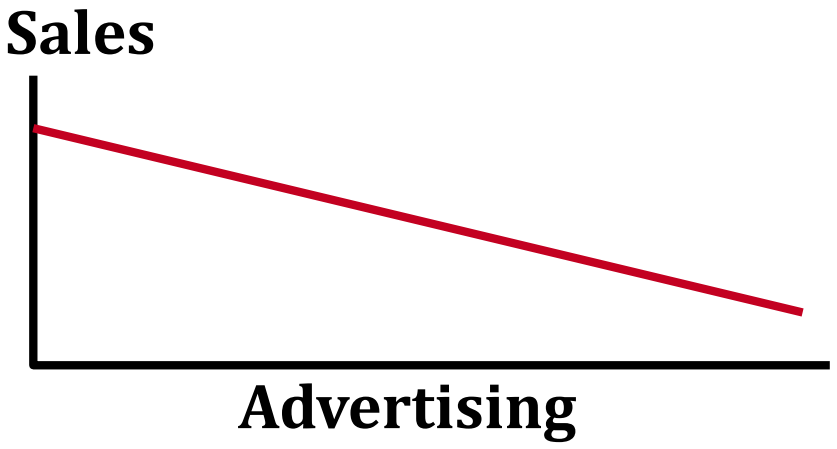
Based on:

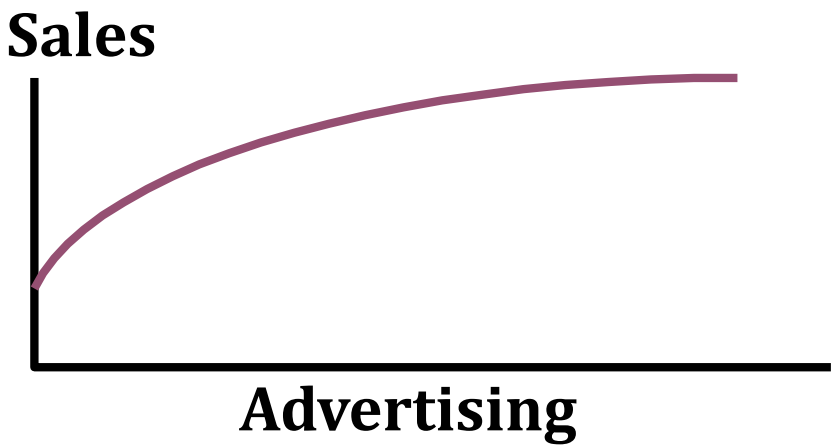
- Theory of field (e.g., Sociology)
- Mathematical theory
- Previous research
- 'Common sense'

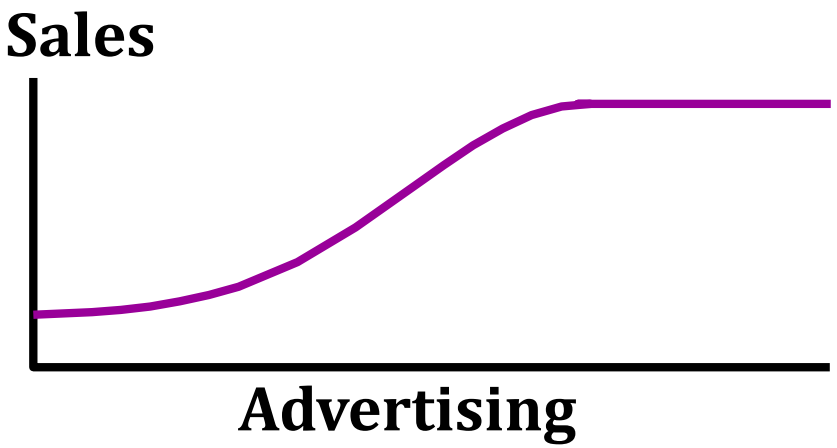
Data Analysis – Simple Linear Regression

Which Is More Logical?





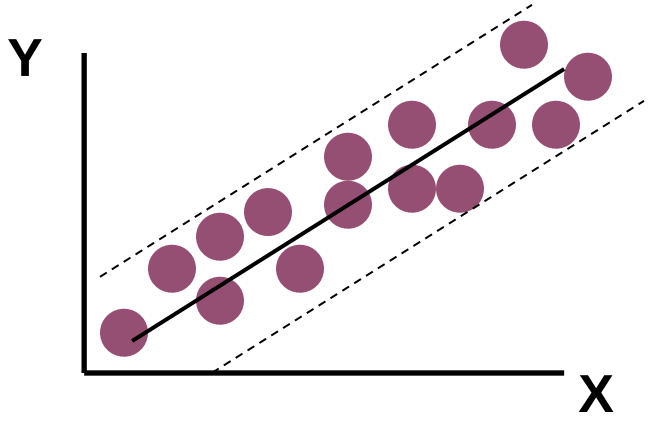


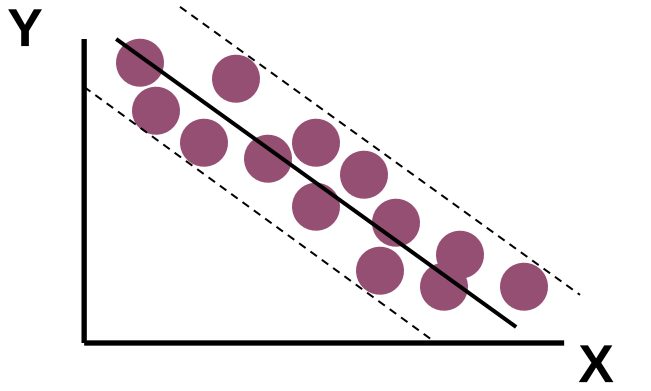


Data Analysis – Simple Linear Regression

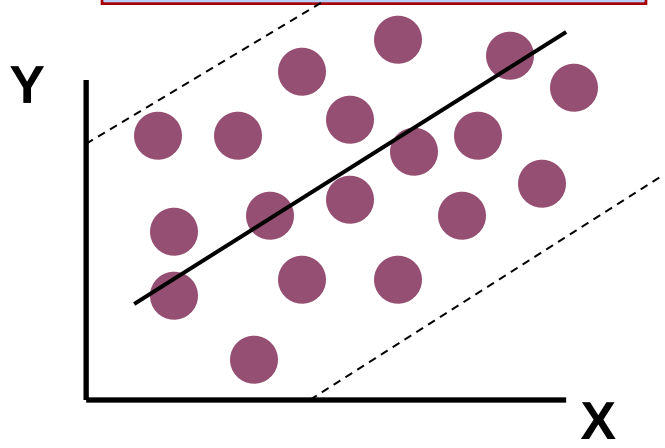
Types of Relationships

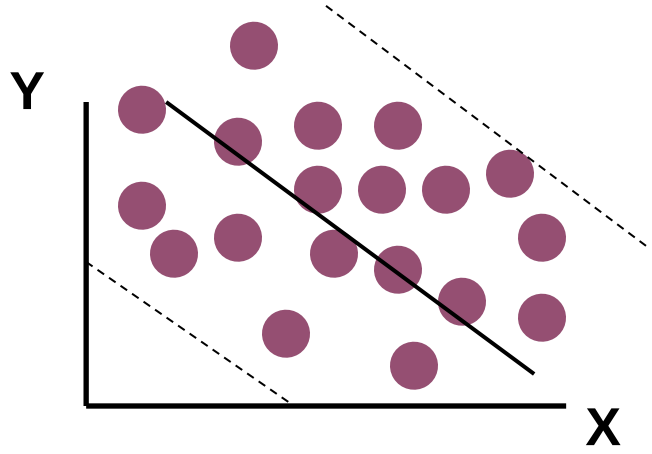
Strong relationships



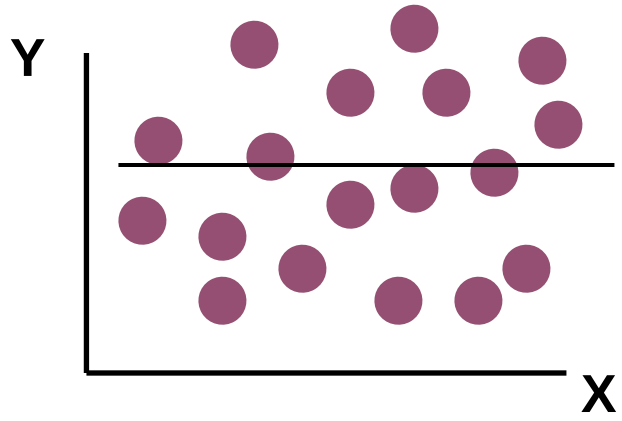


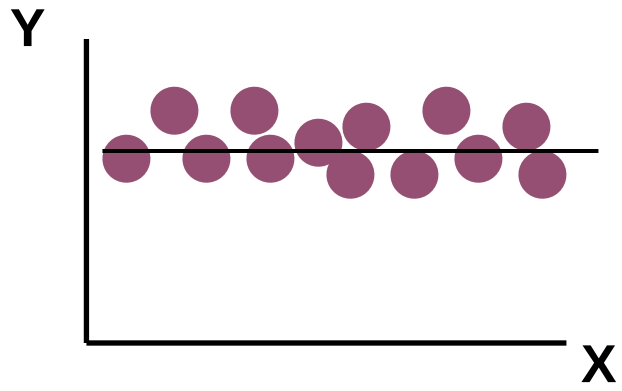
Weak relationships





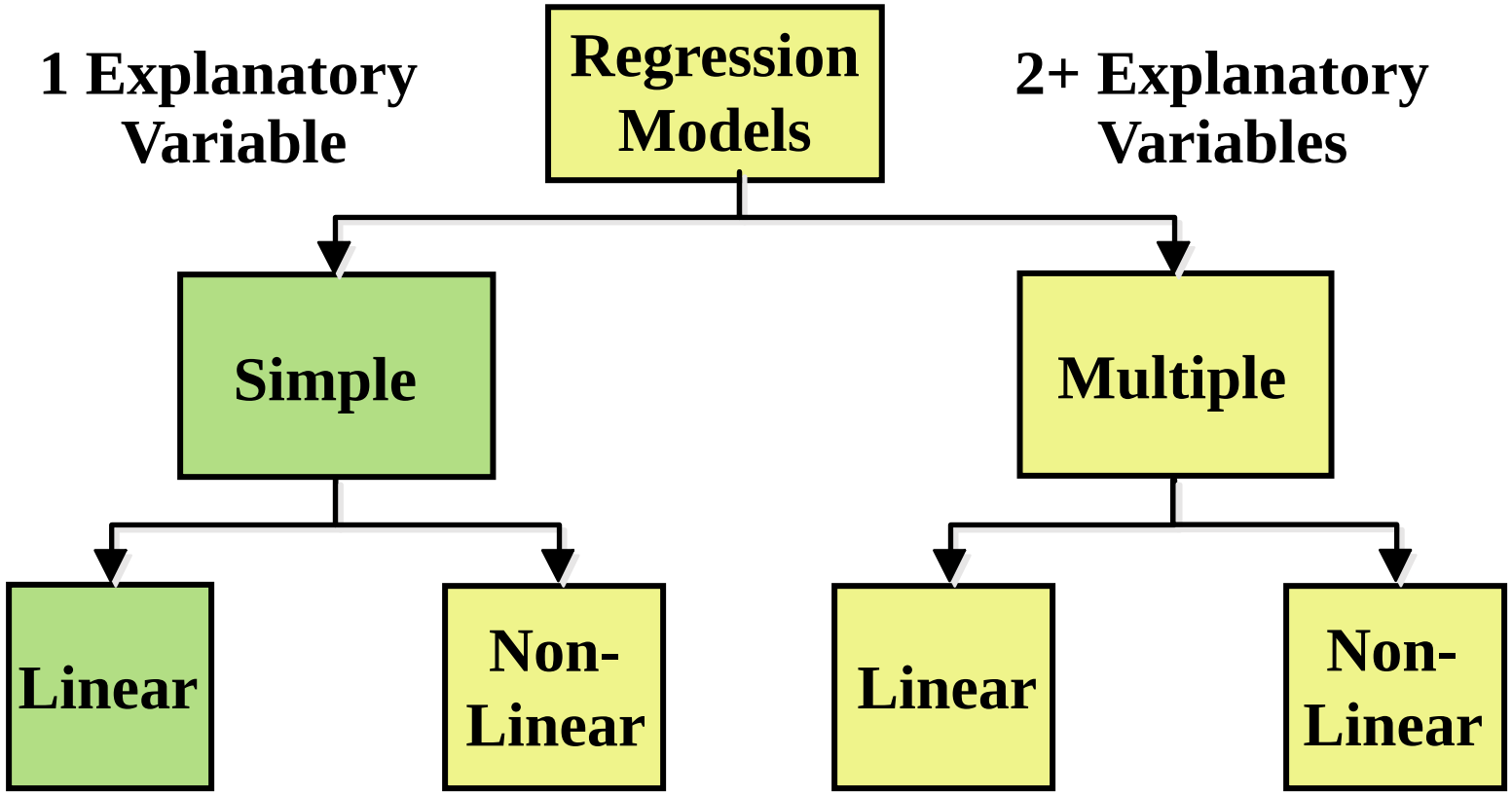
No relationship





Data Analysis – Simple Linear Regression

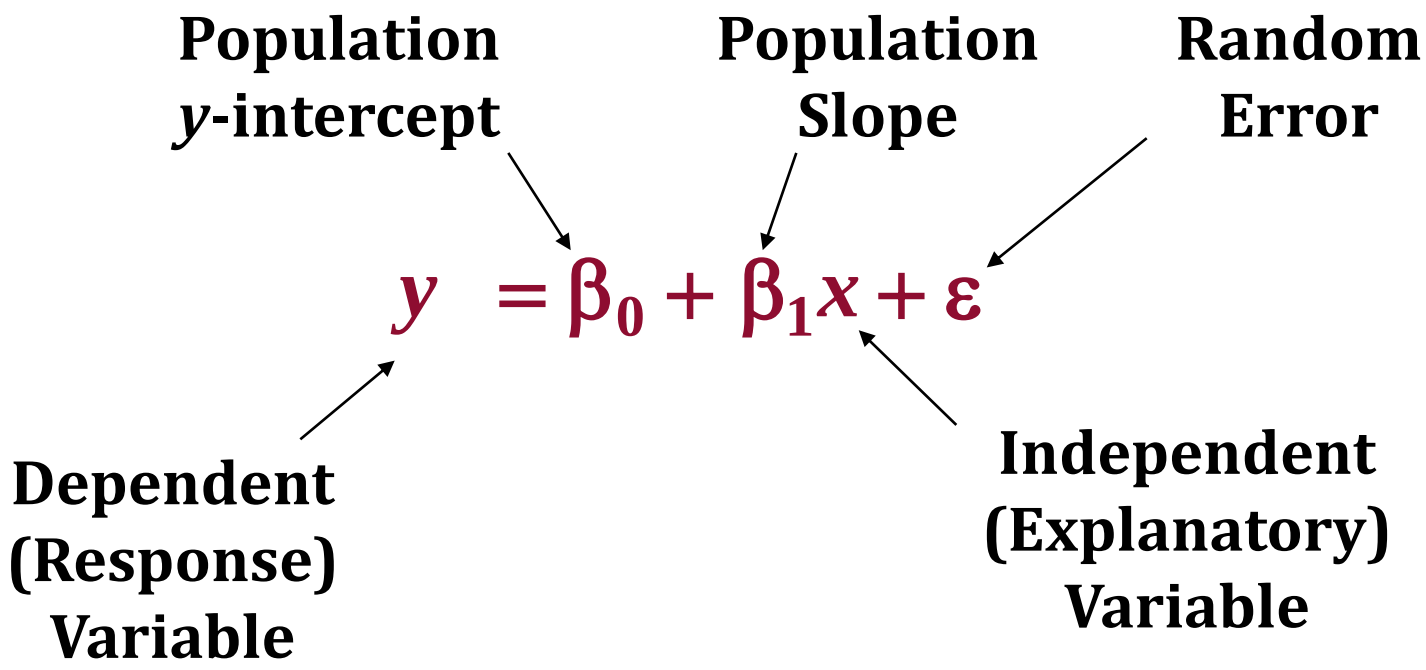
Types of Regression



Data Analysis – Simple Linear Regression

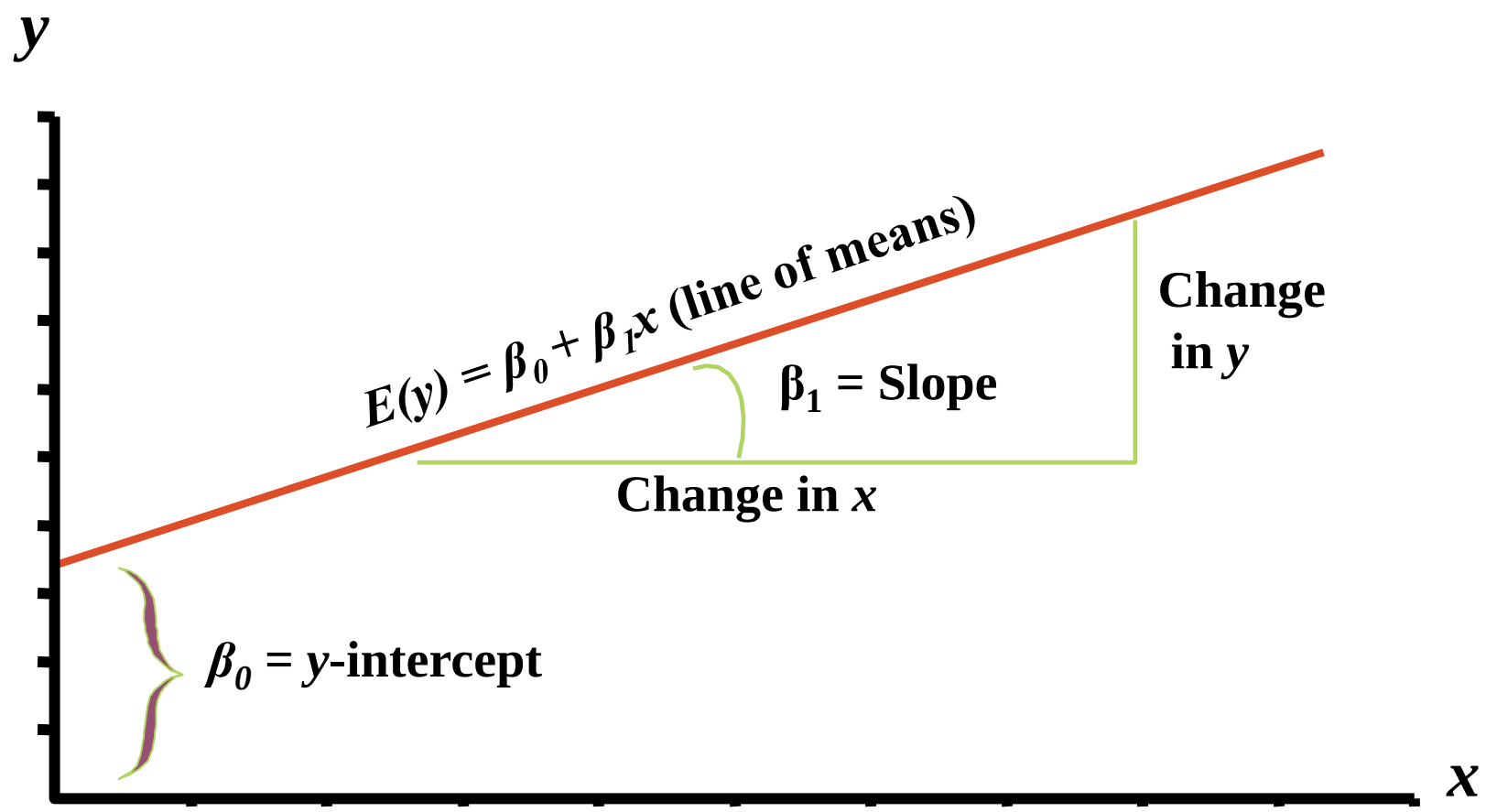
Linear Regression Model

Relationship between variables is a linear function



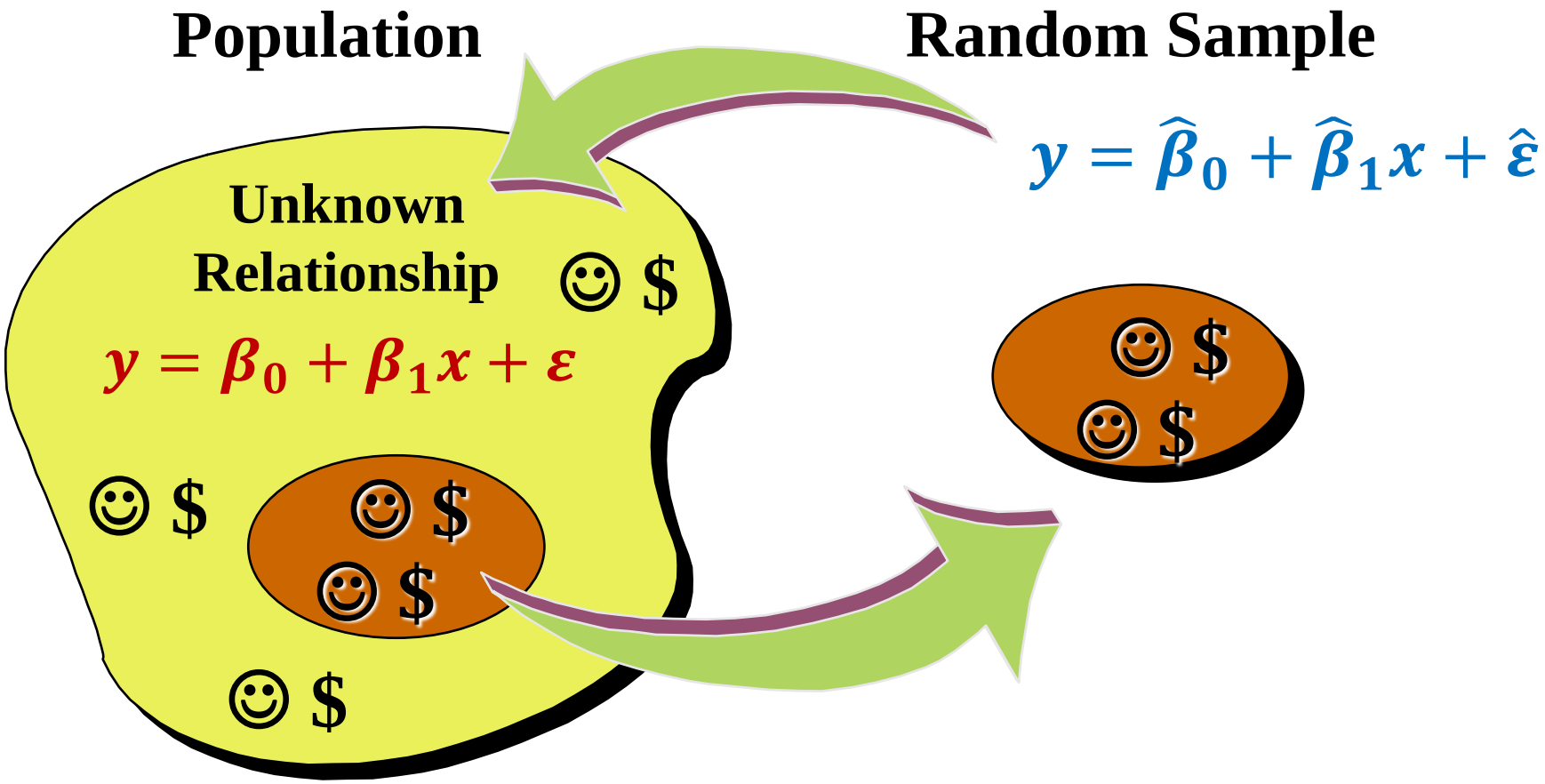
Data Analysis – Simple Linear Regression

Line of Means



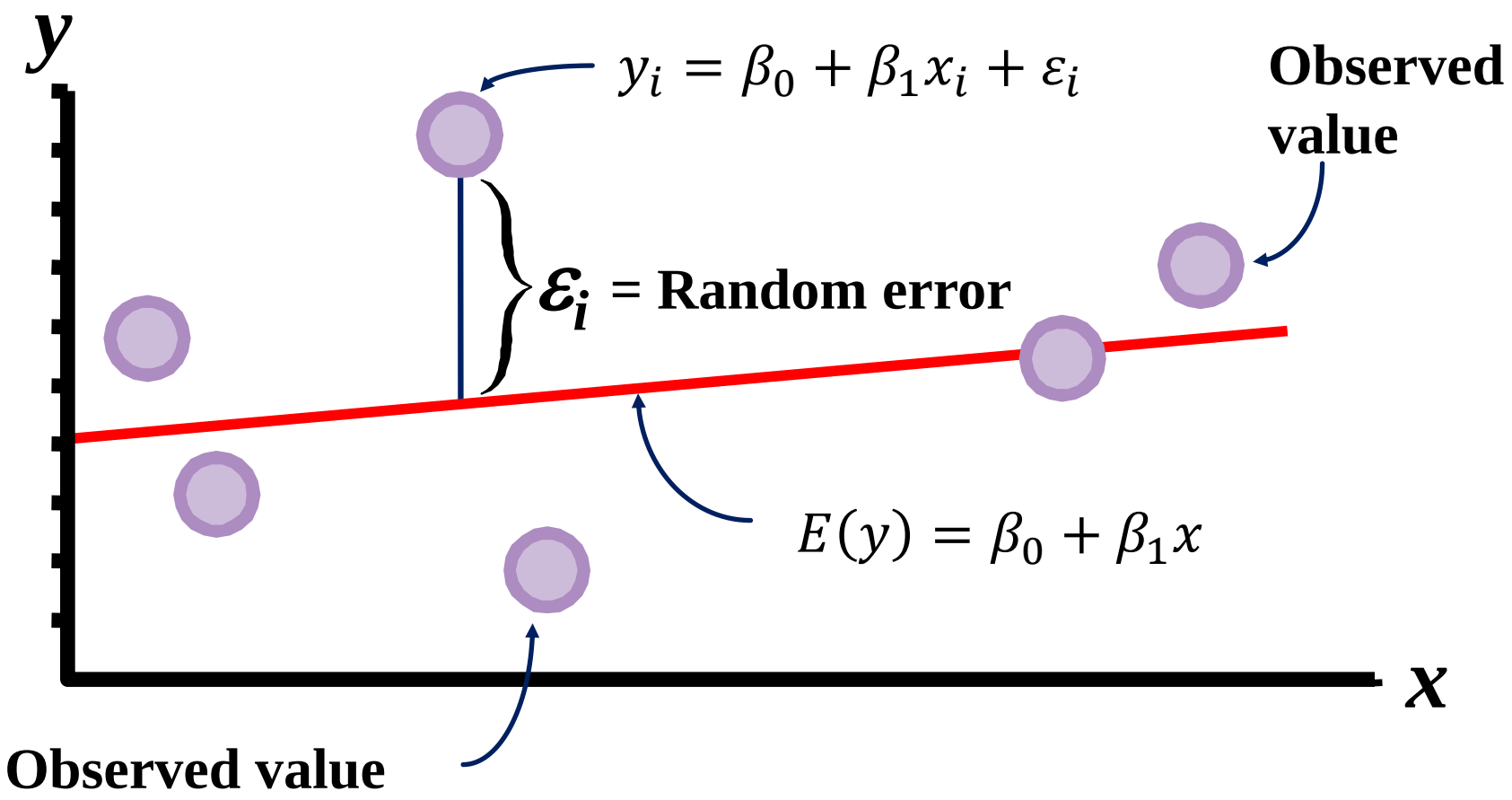
Data Analysis – Simple Linear Regression

Population & Sample Regression Models



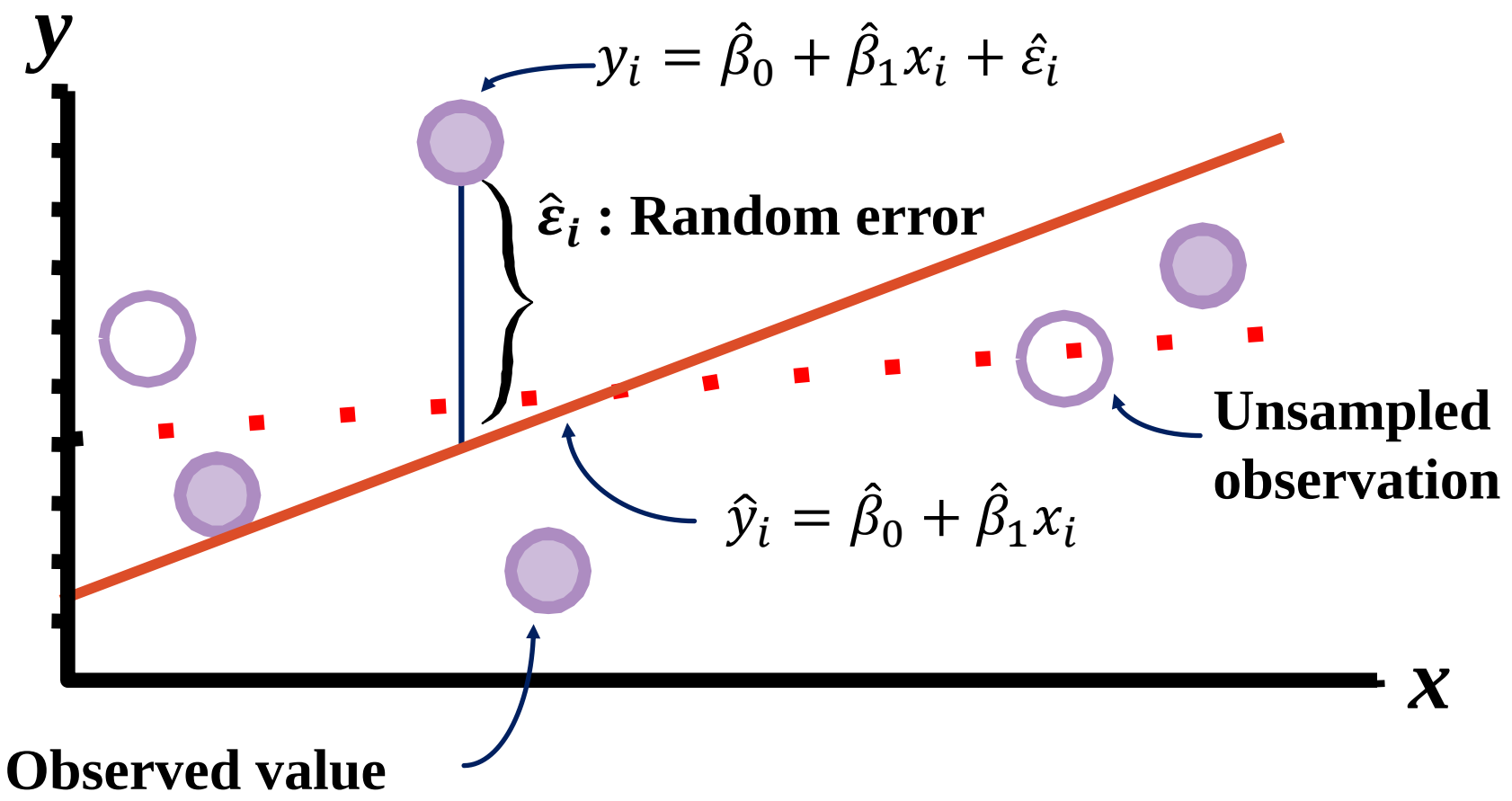
Data Analysis – Simple Linear Regression

Population Linear Regression Model



Data Analysis – Simple Linear Regression

Sample Linear Regression Model



Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

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

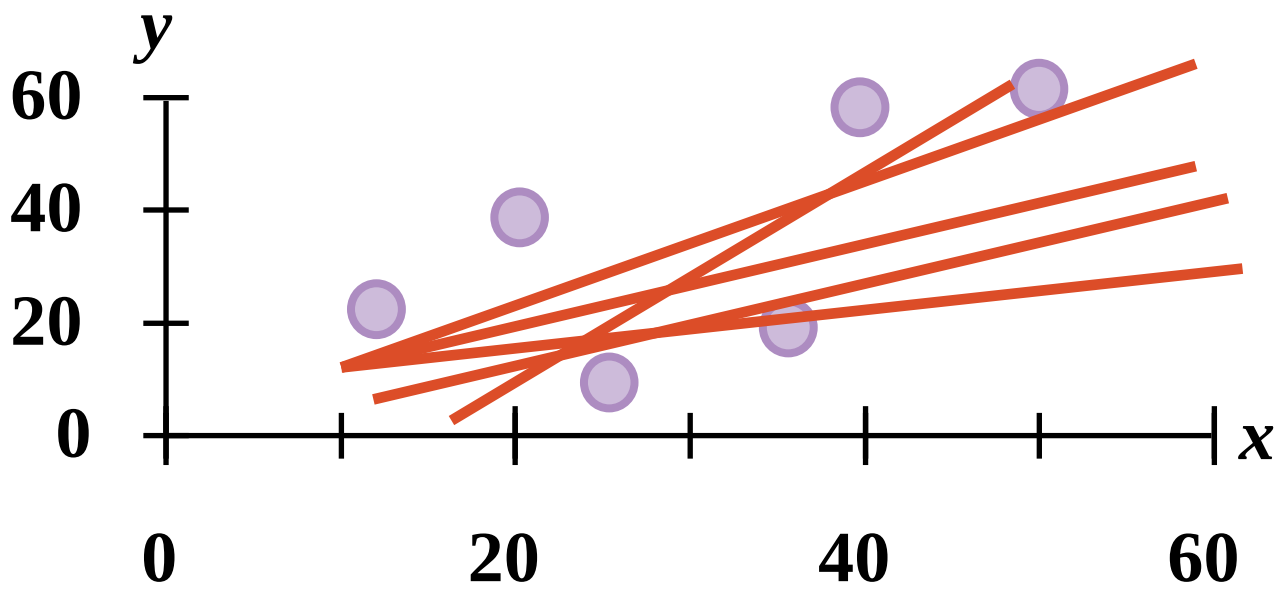
4. Evaluate model

5. Use model for prediction and estimation

Data Analysis – Simple Linear Regression

Thinking challenge (we saw it in Bilinear Algebra too)

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



Data Analysis – Simple Linear Regression

Least Squares

- Best fit means that the difference between **actual** *y* values and **predicted** *y* values are a **minimum**.
  - *But* positive differences off-set negative:

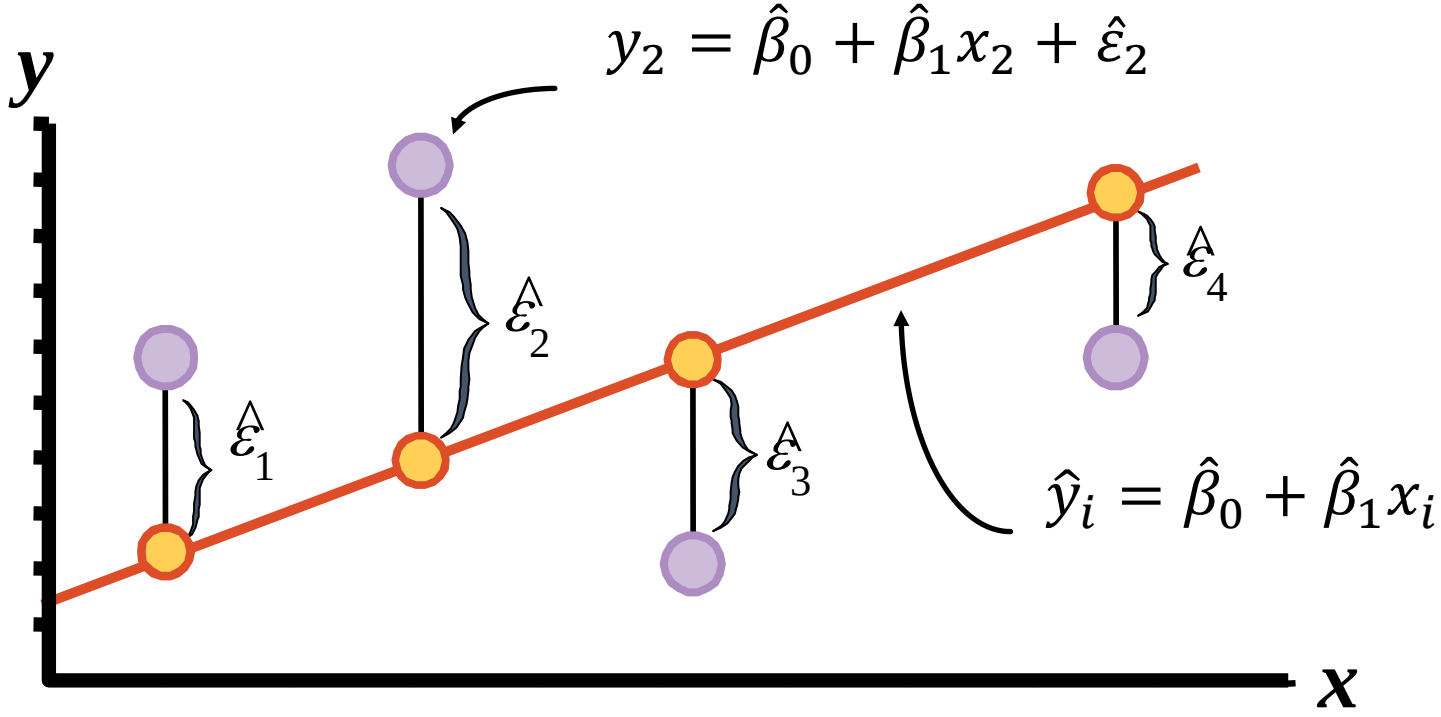
$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

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

Data Analysis – Simple Linear Regression

Least Squares graphically

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



Data Analysis – Simple Linear Regression

Coefficient Equations

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

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

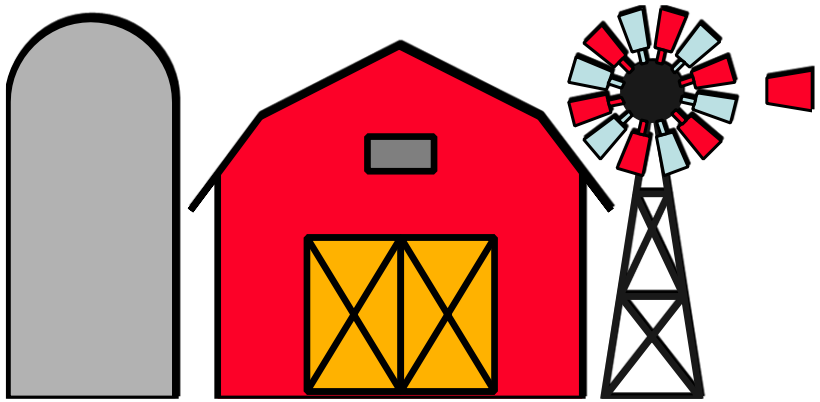
***y*-intercept:** $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Data Analysis – Simple Linear Regression

Least Squares - Thinking Challenge

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

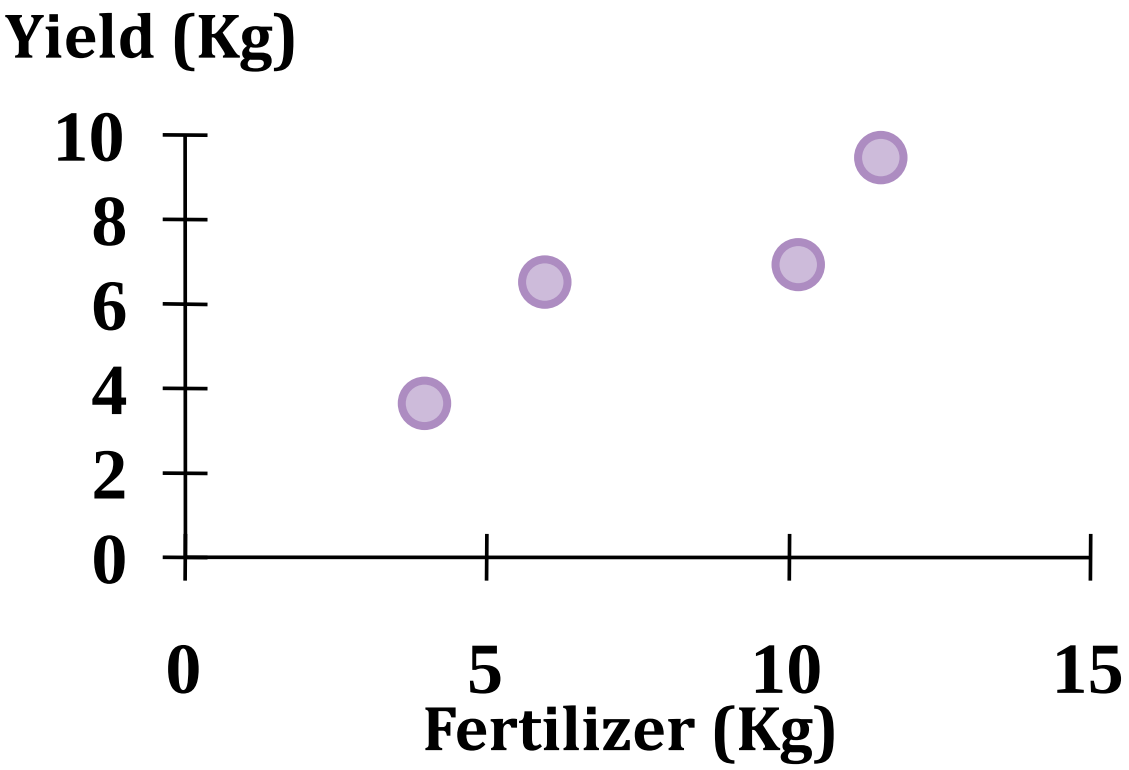
Fertilizer (Kg)	Yield (Kg)
4	3.0
6	5.5
10	6.5
12	9.0



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

Data Analysis – Simple Linear Regression

Least Squares - Thinking Challenge: Scattergram Crop Yield vs. Fertilizer



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

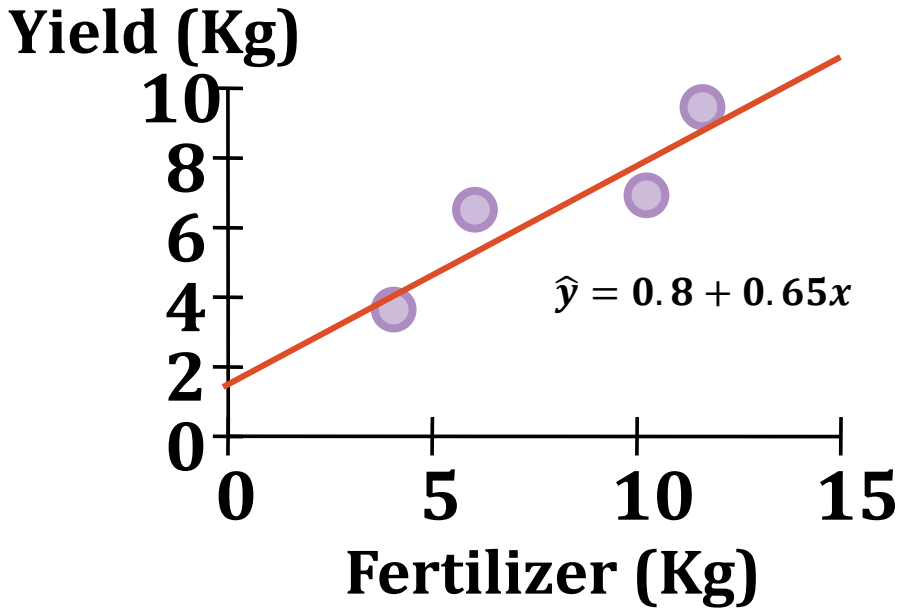
Data Analysis – Simple Linear Regression

Least Squares - Thinking Challenge: Parameter Estimation Solution

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

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

$$\hat{y} = 0.8 + 0.65x$$



1. Slope ($\hat{\beta}_1$)
   - Crop Yield (y) is expected to increase by 0.65 Kg for each 1 Kg increase in Fertilizer (x).
2. Y-Intercept ($\hat{\beta}_0$)
   - Average Crop Yield (y) is expected to be 0.8 Kg when no Fertilizer (x) is used.

Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

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

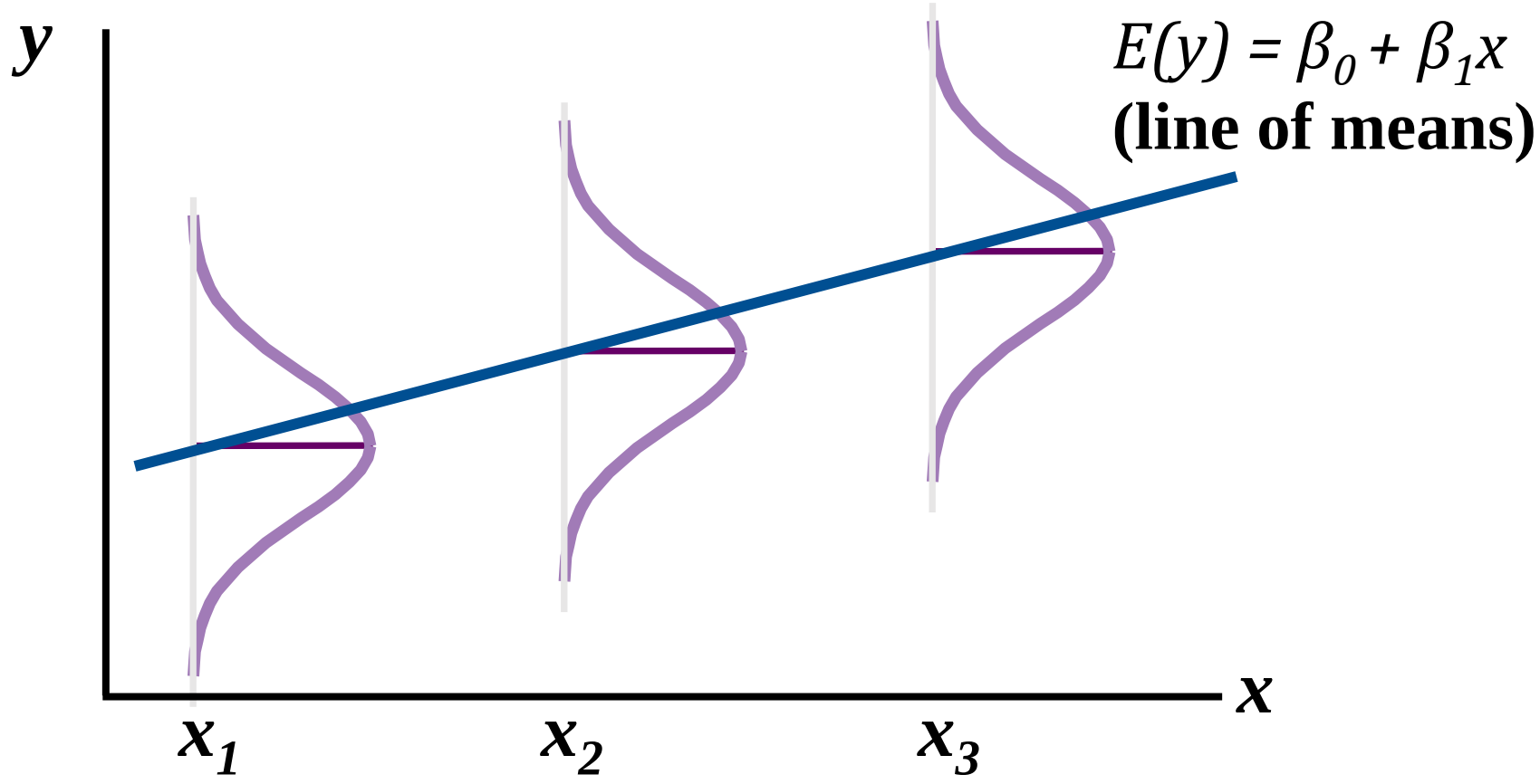
Data Analysis – Linear Regression

Linear Regression Assumptions

I. Mean of probability distribution of error, ε, is 0.
II. Probability distribution of error has constant variance.
III. Probability distribution of error, ε, is normal.
IV. Errors are independent.

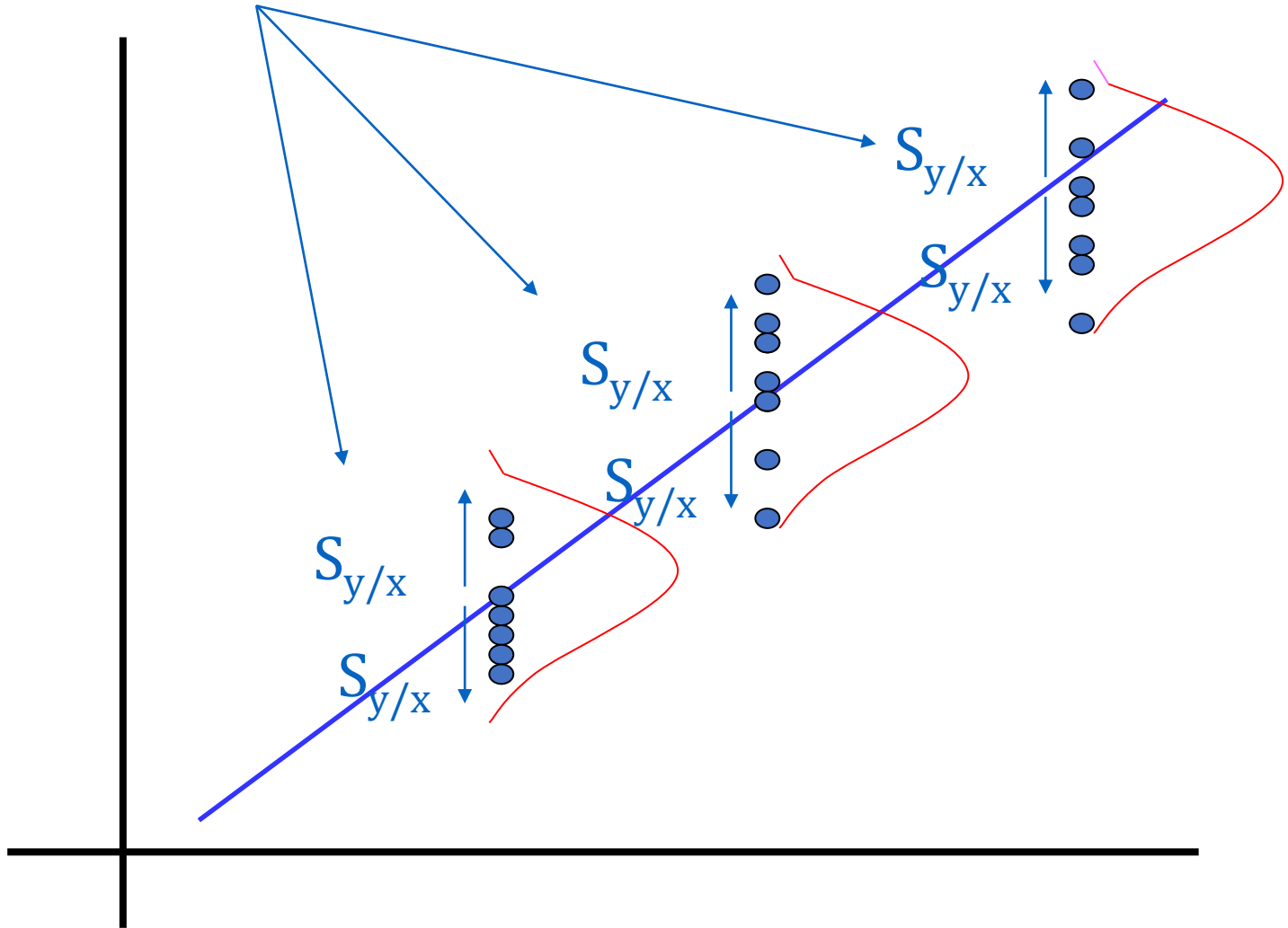
Data Analysis – Simple Linear Regression

Linear Regression Assumptions: Error Probability Distribution



Data Analysis – Simple Linear Regression

The **standard error** of y given x is the
average variability around the regression line at any given value of x.



Data Analysis – Simple Linear Regression

Linear Regression Assumptions: Random Error Variation

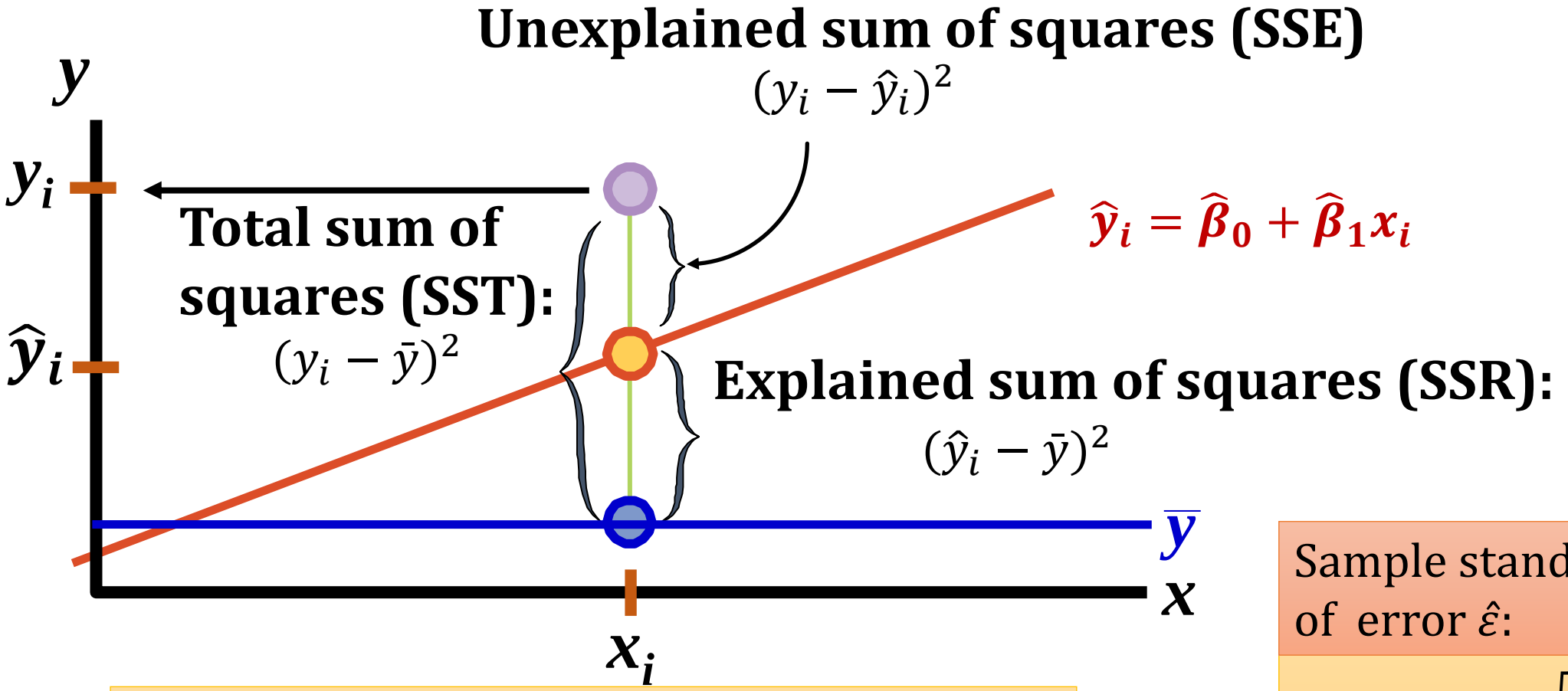
- Variation of actual y from predicted y, $\hat{y}$

- Measured by standard error of regression model
  - Sample standard deviation of $\hat{\varepsilon}$: s

- Affects several factors
  - Parameter significance
  - Prediction accuracy

Data Analysis – Simple Linear Regression

Linear Regression Assumptions: Variation Measures



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

Sample standard deviation of error $\hat{\varepsilon}$:

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

Data Analysis – Simple Linear Regression

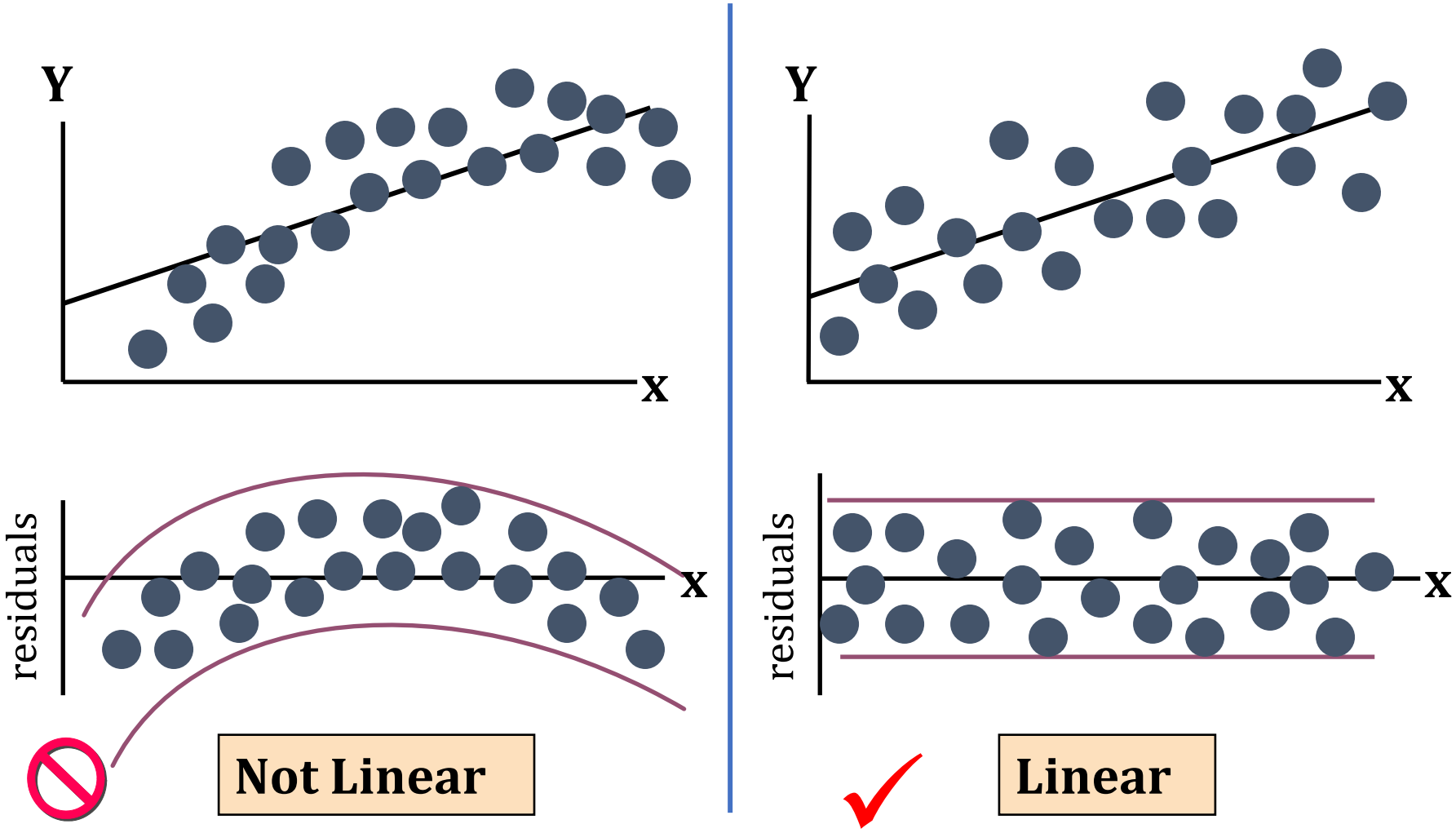
Linear Regression Assumptions: Residual Analysis

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

- 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**)

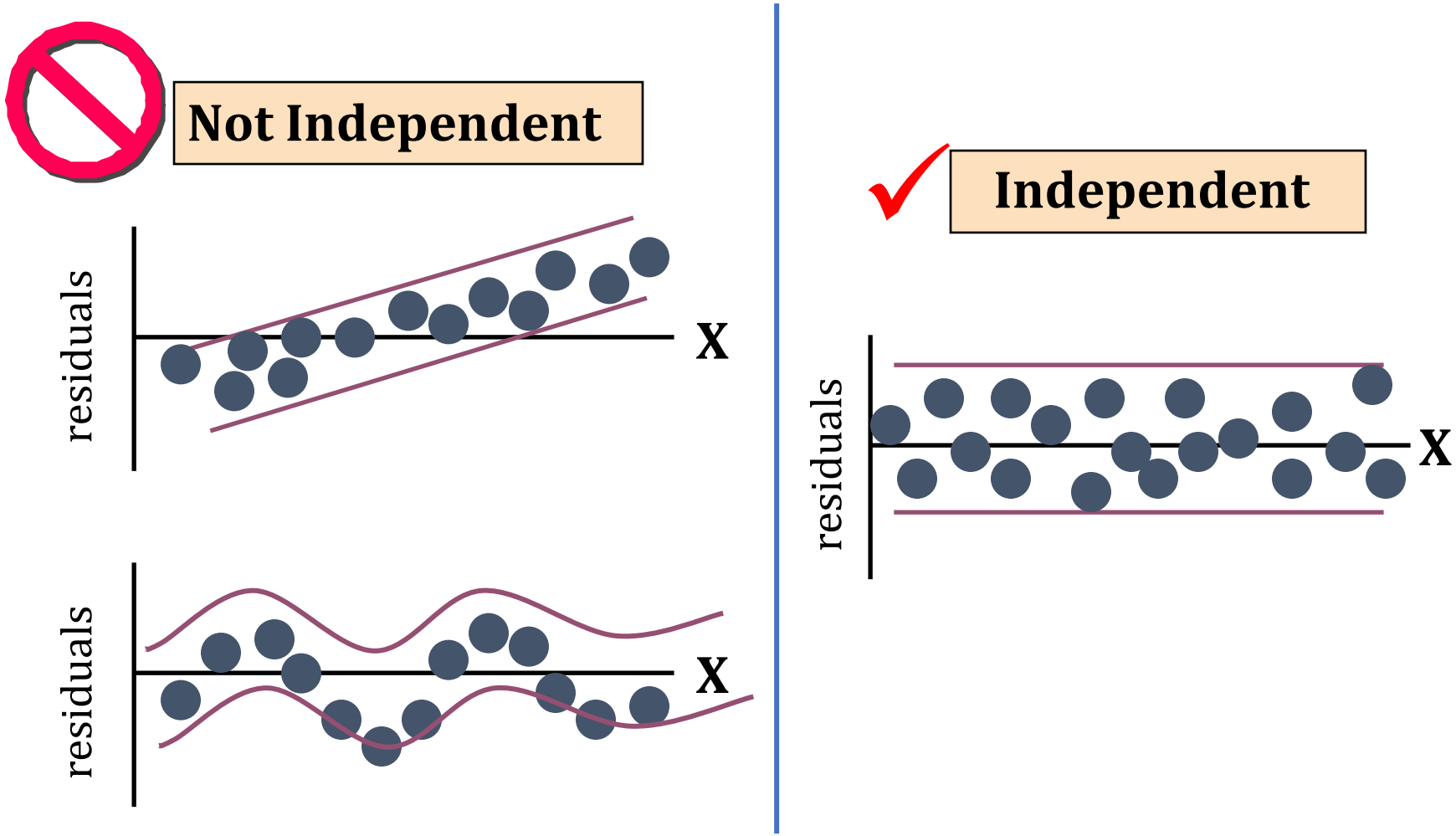
Data Analysis – Simple Linear Regression

Linear Regression Assumptions: Residual Analysis for Linearity



Data Analysis – Linear Regression

Linear Regression Assumptions: Residual Analysis for Independence



Data Analysis – Simple Linear Regression

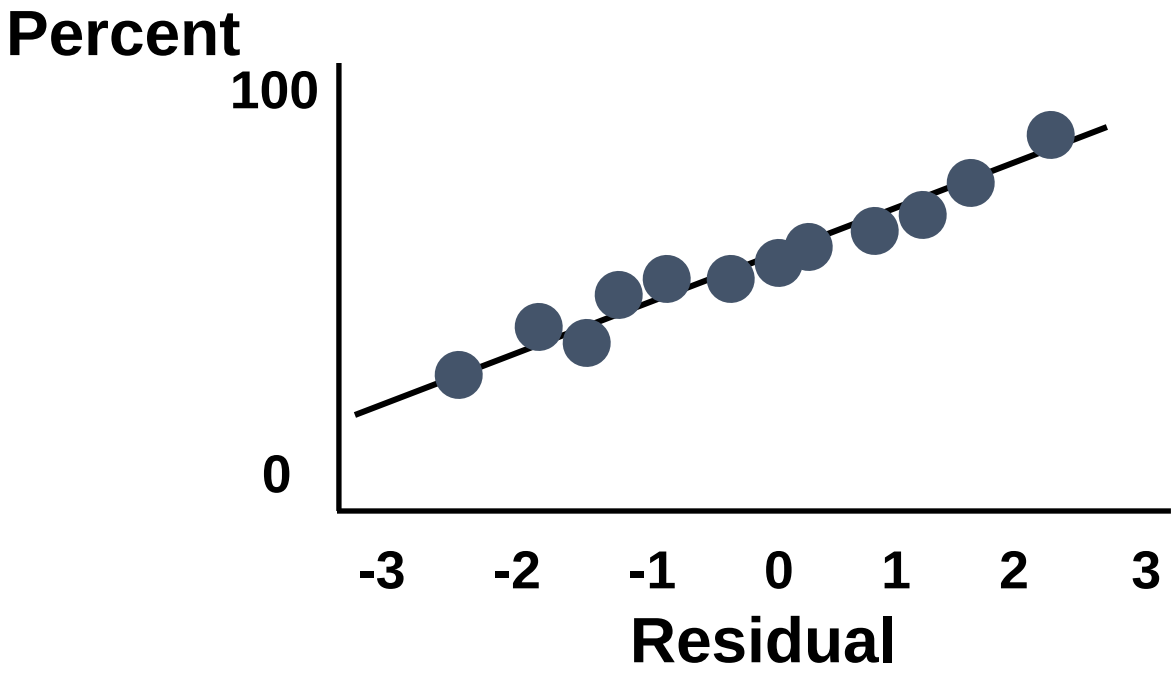
Linear Regression Assumptions: Checking for Normality

- 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

Data Analysis – Simple Linear Regression

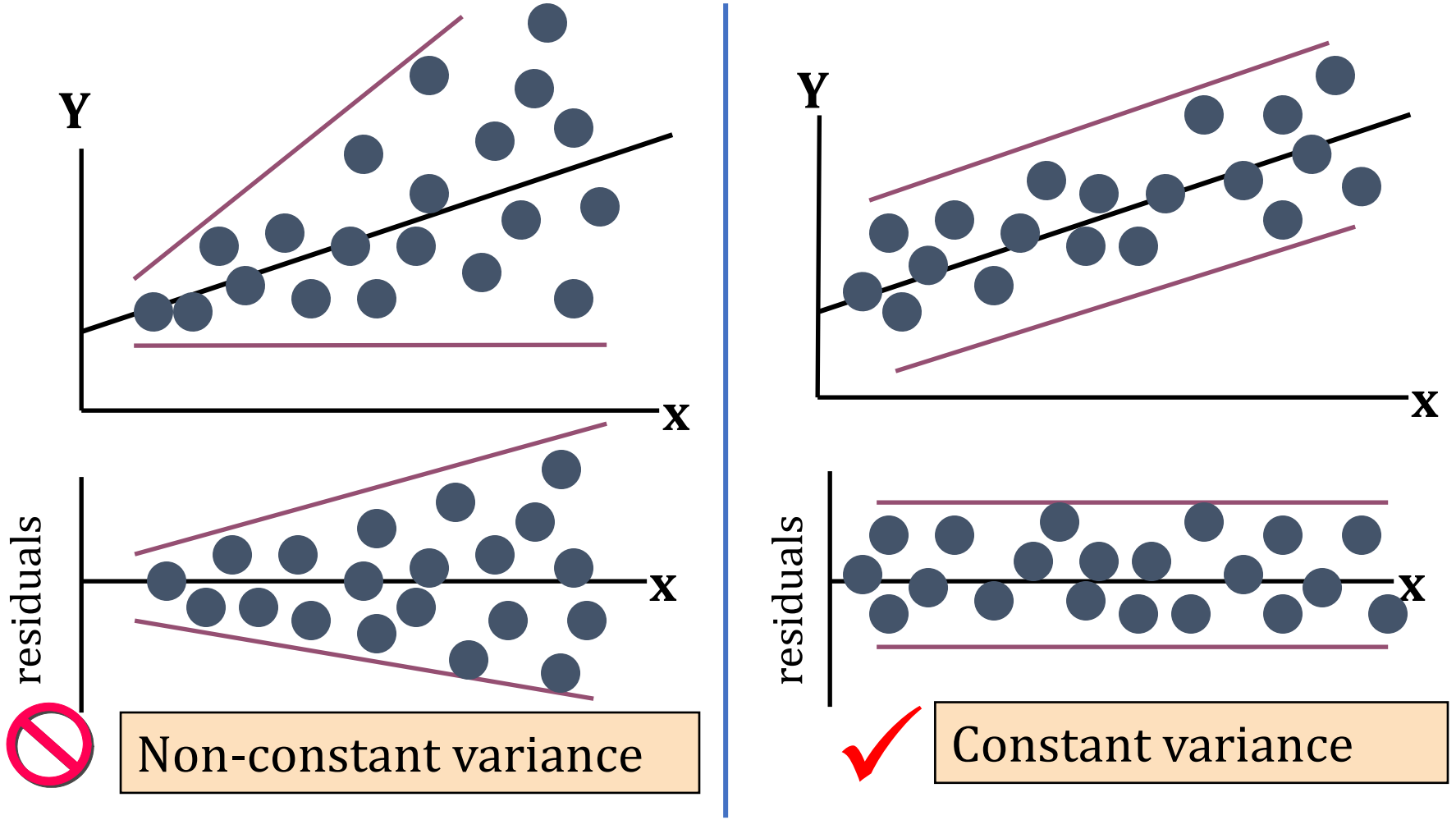
Linear Regression Assumptions: Residual Analysis for Normality

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



Data Analysis – Simple Linear Regression

Linear Regression Assumptions: Residual Analysis for Equal Variance



Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

Linear Regression: Evaluate the Model – Testing for Significance

Test of Slope Coefficient

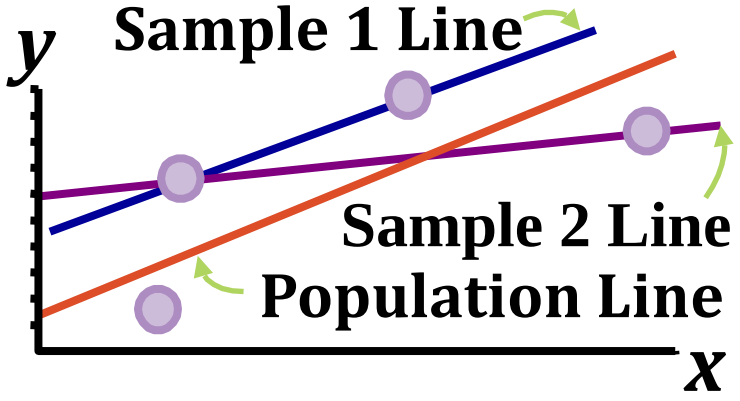
- Shows if there is a linear relationship between x and y
- Involves population slope β_1
- Hypotheses
  - $\mathbf{H_0}$: $\beta_1 = 0$ (**No Linear Relationship**)
  - $\mathbf{H_a}$: $\beta_1 \neq 0$ (**Linear Relationship**)
- Theoretical basis is sampling distribution of slope

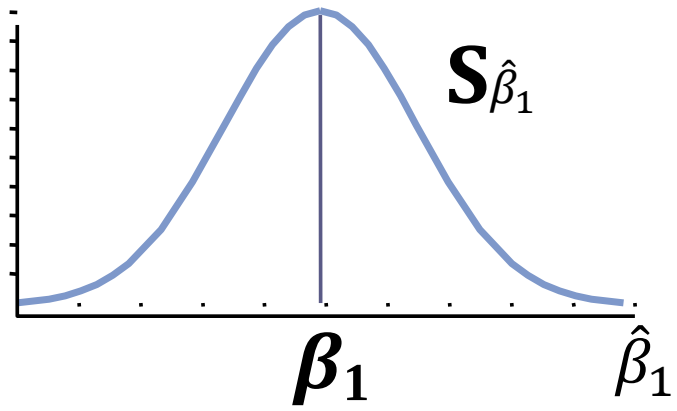
- ❑ The **p-values** help determine whether the relationships that you observe in your sample also exist in the larger **population**.
- ❑ The p-value for each independent variable tests the **null hypothesis** that the variable has **no correlation** with the dependent variable.
- ❑ If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is insufficient evidence to conclude that there is an effect at the population level.

Data Analysis – Simple Linear Regression

Linear Regression: Evaluate the Model – Testing for Significance

Sampling Distribution of Sample Slopes





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

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{SS_{xx}}}},$$

$$df = n - 2$$

where

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

Data Analysis – Simple Linear Regression

Linear Regression: Evaluate the Model – Testing for Significance

Example

You're a marketing analyst for a toy company.

You find $\hat{\beta}_0$ **= –0.1**, $\hat{\beta}_1$ **=0 .7** and ***s*** **=0 .6055**.

Is the relationship **significant** at the **0.05** level of significance?

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

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

$$S_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}} = \frac{0.6055}{\sqrt{55 - \frac{(15)^2}{5}}} = 0.1914$$

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0.70}{0.1914} = 3.657$$

Data Analysis – Simple Linear Regression

Linear Regression: Evaluate the Model – Testing for Significance

Example

Sample standard deviation of $\hat{\varepsilon}$

You're a marketing analyst for a toy company.

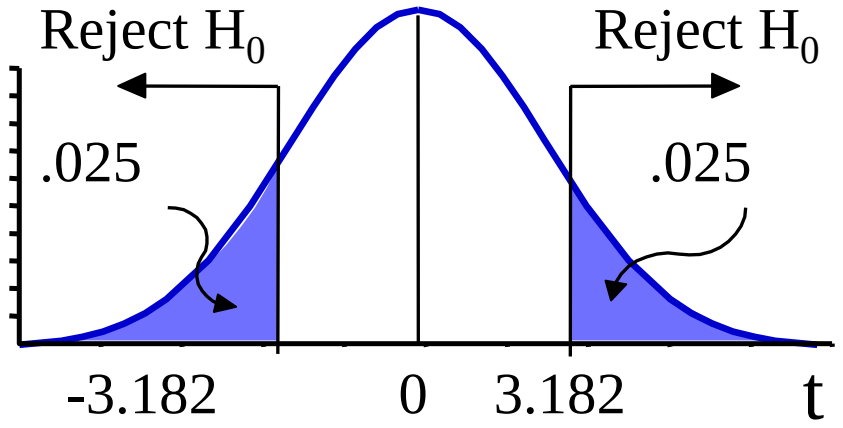
You find $\hat{\beta}_0$ **= –0.1**, $\hat{\beta}_1$ **= 0.7** and ***s* = 0.6055**.

Is the relationship **significant** at the **0.05** level of significance?

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

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0.70}{0.1914} = 3.657$$

- **H0:** $\beta_1 = 0$ **/ Ha:** $\beta_1 \neq 0$
- $\alpha = 0.05$, **df** $= 5 - 2 = 3$
- **Critical Value(s):**



Decision:

Reject at α = 0.05.

Conclusion:

There is evidence of a relationship.

H_0: $\beta_1 = 0$ (**No Linear Relationship**)

H_a: $\beta_1 \neq 0$ (**Linear Relationship**)

Data Analysis – Simple Linear Regression

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

Data Analysis – Simple Linear Regression

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**

Data Analysis – Simple Linear Regression

Prediction With Regression Models

- **Types of predictions**
  - Point estimates
  - Interval estimates

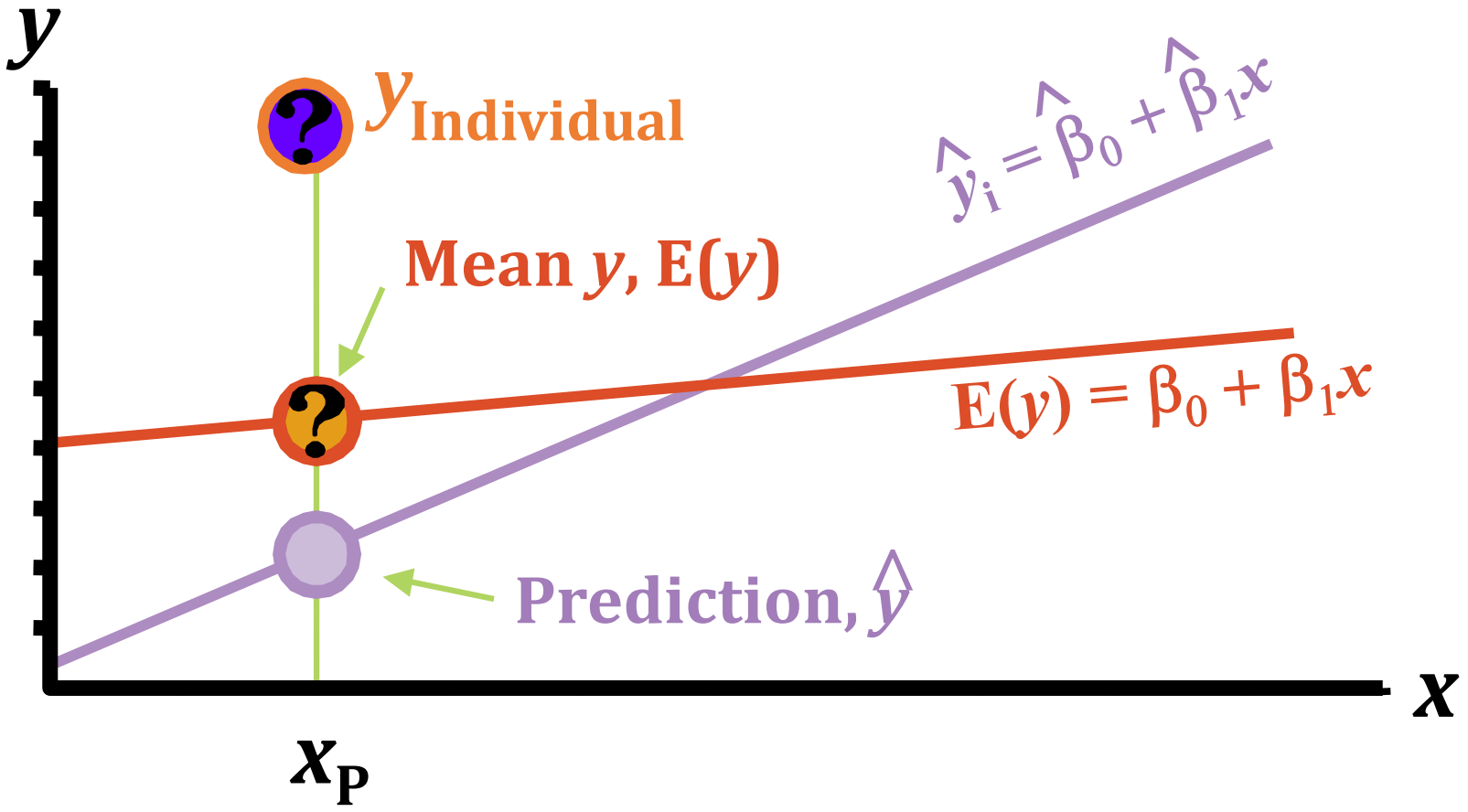
- **What is predicted**
  - Population mean response $E(y)$ for given x
    - Point on population regression line
  - Individual response (y_i) for given x

Data Analysis – Simple Linear Regression

Prediction With Regression Models - What is predicted

❑ **What is predicted**

- Population mean response $E(y)$ for given x
  - Point on population regression line
- Individual response (y_i) for given x



Data Analysis – Simple Linear Regression

Prediction With Regression Models - What is predicted

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

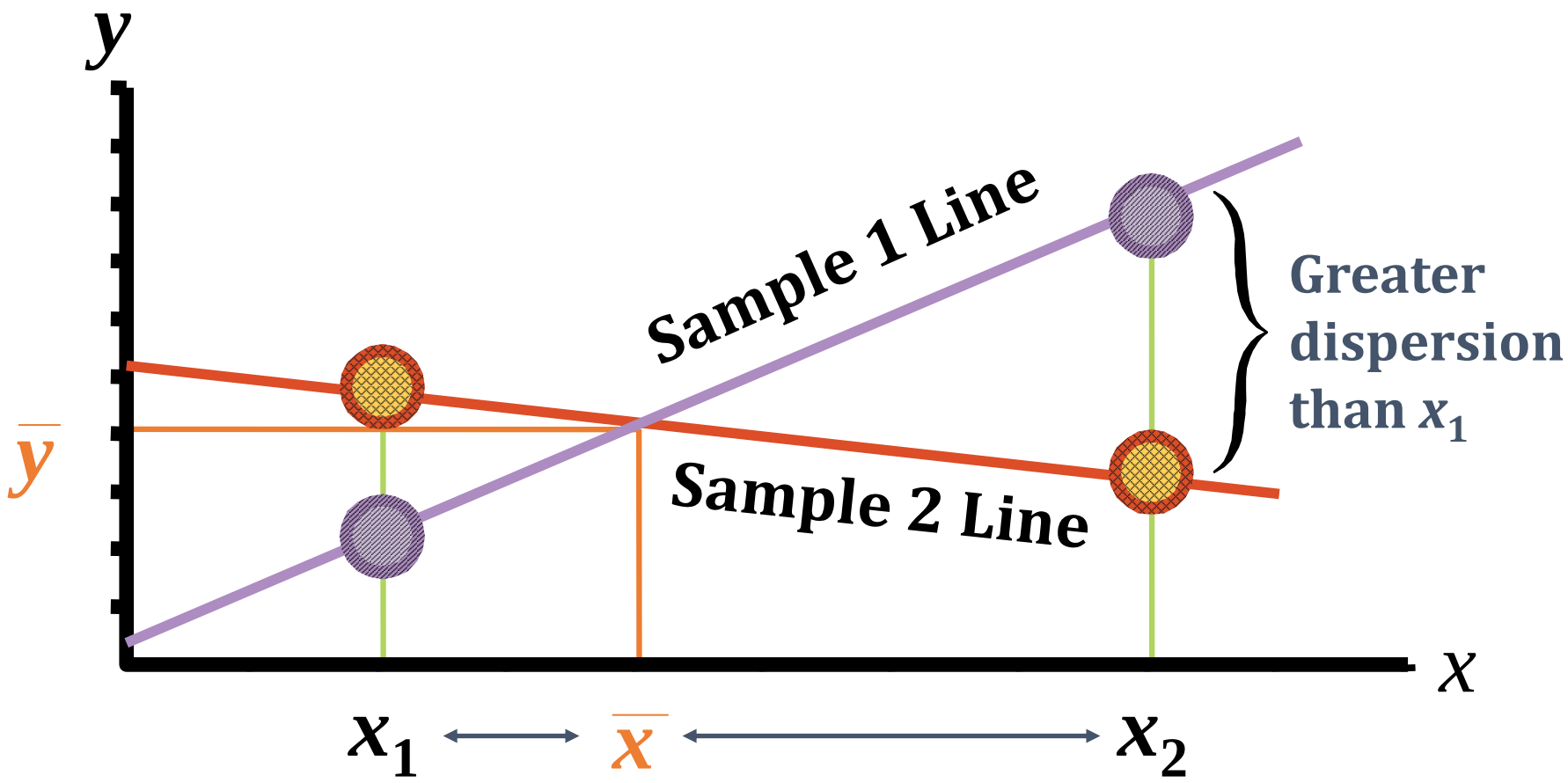
$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} \qquad \text{df} = n - 2$$

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

Data Analysis – Simple Linear Regression

Prediction With Regression Models - Why Distance from Mean?



The closer to the mean, the less variability.
This is due to the variability in estimated slope parameters.

Data Analysis – Linear Regression

Prediction With Regression Models

Confidence Interval Estimate - Example

You're a marketing analyst for a toy company.

You find $\hat{\beta}_0$ **= -0.1**, $\hat{\beta}_1$ **= 0.7** and ***s* = 0.6055**.

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

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

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

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

***x* to be predicted**

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

$$2.7 \pm (3.182)(0.6055)\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}}$$

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

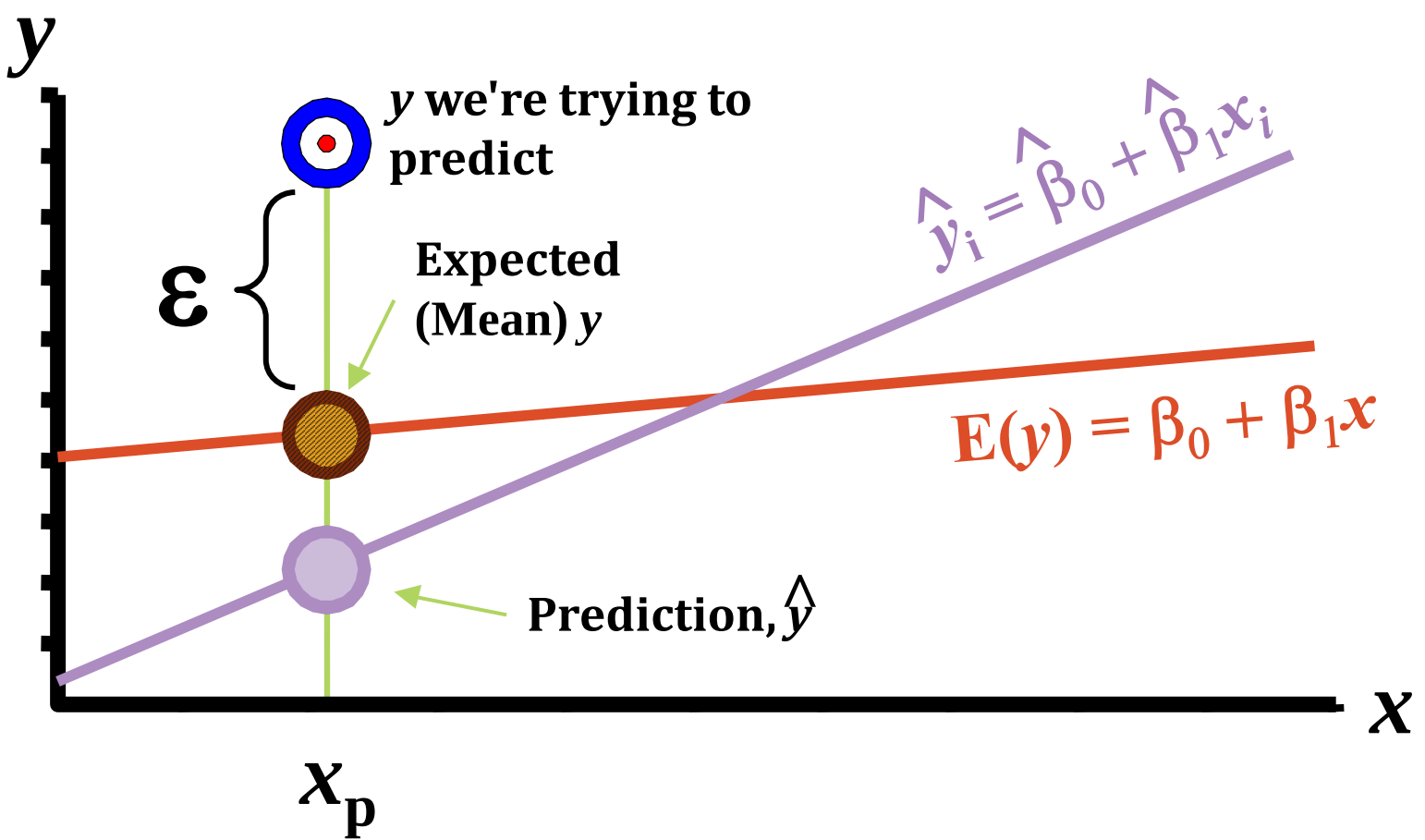
Data Analysis – Simple Linear Regression

Prediction With Regression Models - What is predicted

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

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

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



Data Analysis – Simple Linear Regression

Prediction With Regression Models

Prediction Interval Estimate - Example

You're a marketing analyst for a toy company.

You find $\hat{\beta}_0$ **= -0.1**, $\hat{\beta}_1$ **=0.7** and ***s* = 0.6055**.

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

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

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

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

***x* to be predicted**

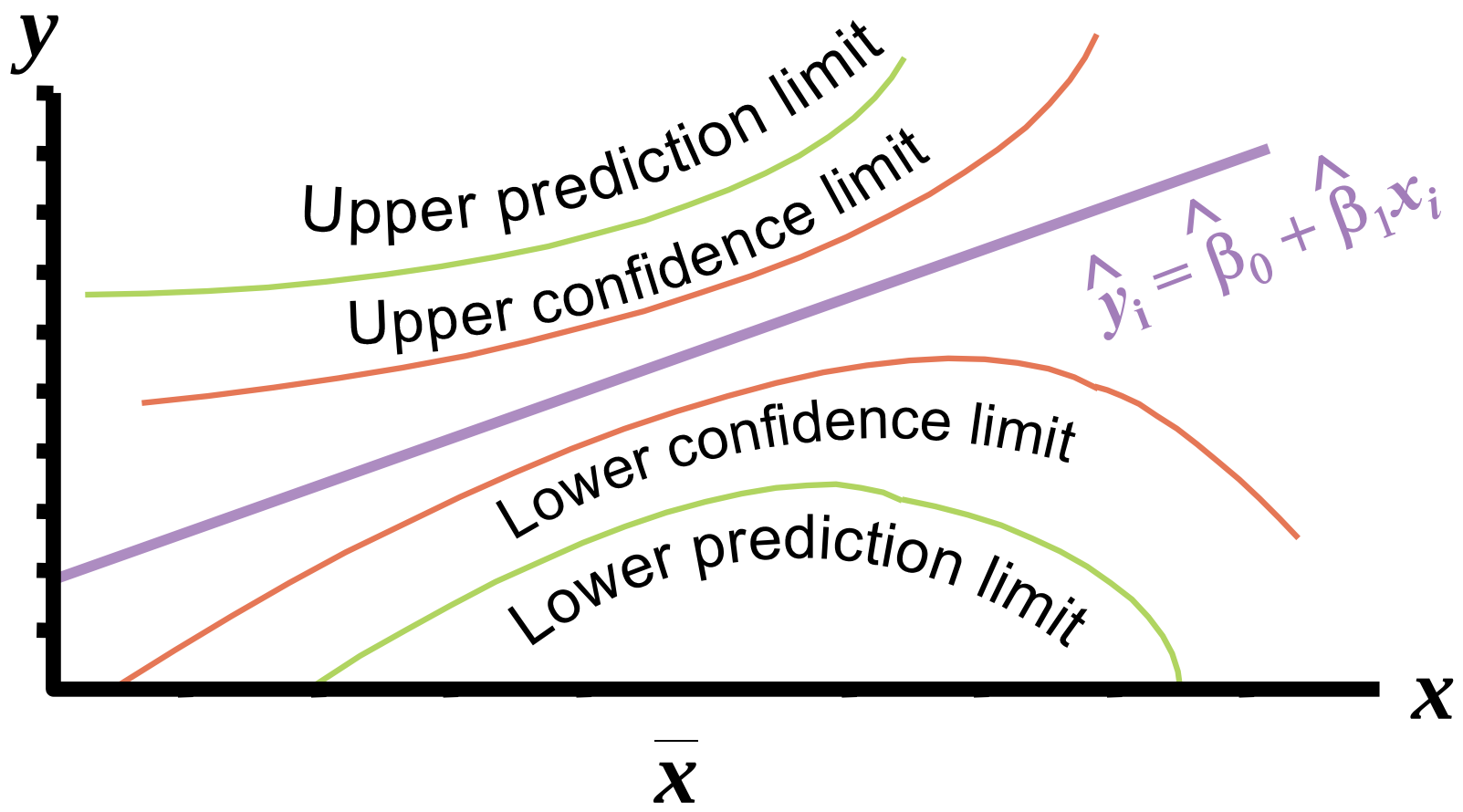
$$\hat{y} = -0.1 + (0.7)(4) = 2.7$$

$$2.7 \pm (3.182)(0.6055)\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$0.503 \leq y_4 \leq 4.897$$

Data Analysis – Simple Linear Regression

Confidence vs Prediction Intervals



Data Analysis – Simple Linear Regression

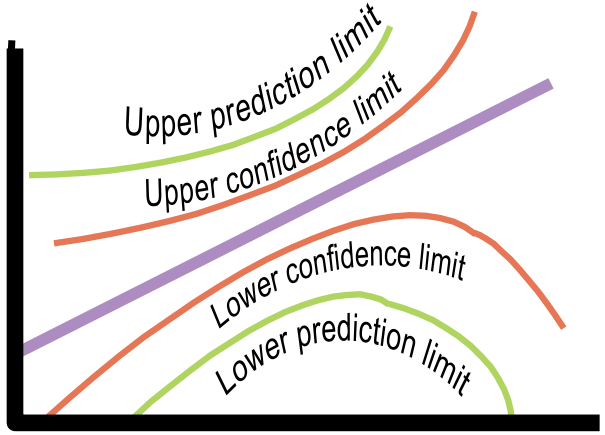
Confidence vs Prediction Intervals

Confidence intervals → **how well you have determined the mean.**

Assume that the data really are randomly sampled from a Gaussian distribution. If you do this many times, and calculate a confidence interval of the mean from each sample, you'd expect about 95% of those intervals to include the true value of the population mean. The key point is that the **confidence interval** tells you about the ***likely location of the true population parameter***.

Prediction intervals → **where you can expect to see the next data point sampled.**

Assume that the data really are randomly sampled from a Gaussian distribution. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you would expect that next value to lie within that prediction interval in 95% of the samples. The key point is that the **prediction interval** tells you about the ***distribution of values, not the uncertainty in determining the population mean***.



Data Analysis – Multiple Linear Regression

Multiple Linear Regression

More than one independent variable can be used to explain variance in the dependent variable, as long as they are not linearly related. A multiple regression takes the form:

$$\boldsymbol{y = A + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon}$$

where k is the number of variables, or parameters.
The parameters in the linear regression model are very easy to interpret.

- β_0 is the intercept (i.e. the average value for y if all the X's are zero),

- β_j is the slope for the j-th variable X_j

- β_j is the average increase in y when X_j is increased by one and **all other X's are held constant**.

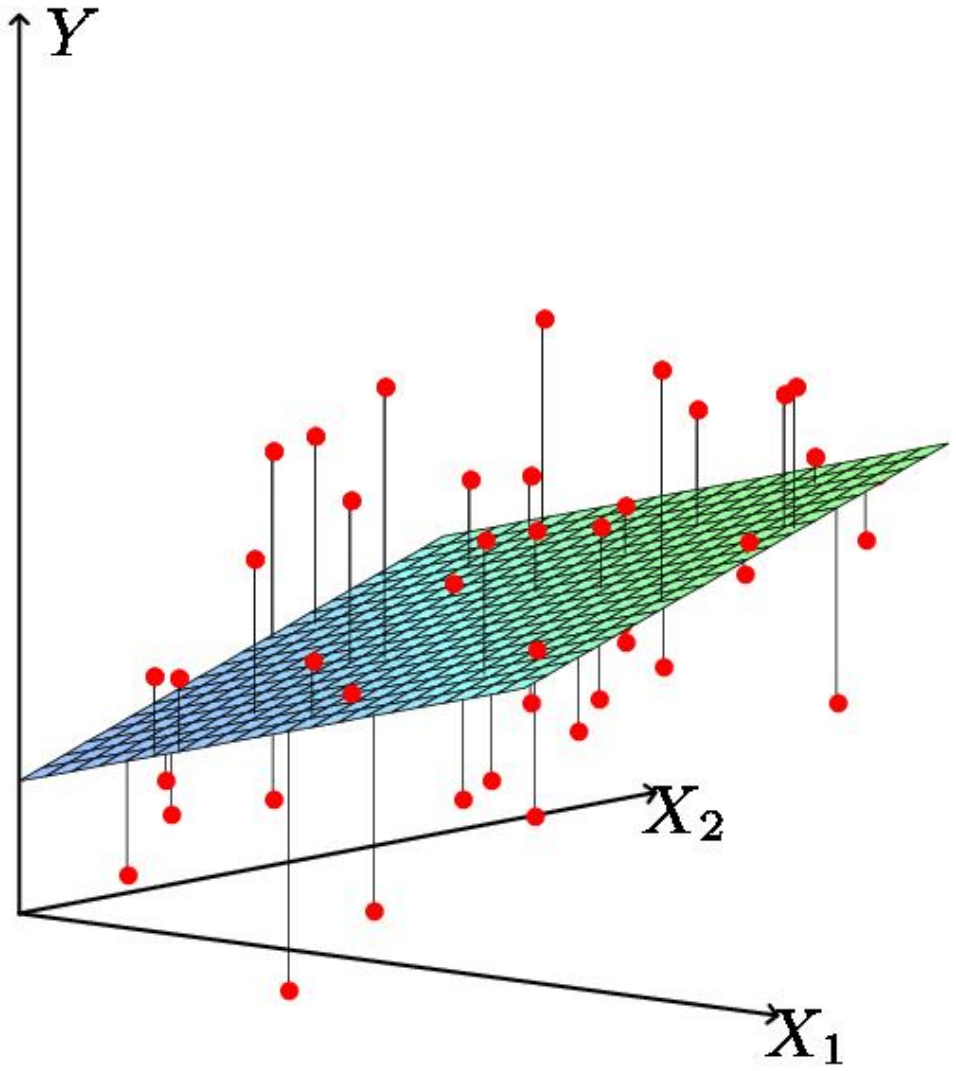
Data Analysis – Multiple Linear Regression

Least Squares Fit

- We estimate the parameters using least squares i.e. minimize:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_1 - \cdots - \hat{b}_p X_p\right)^2$$



Data Analysis – Multiple Linear Regression

Relationship between population and least squares lines

Population line

$$Y_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + \epsilon$$

Least Squares line

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots + \hat{b}_p X_p$$

- We would like to know β_0 through β_p i.e. the population line. Instead we know $\hat{\beta}_0$ through $\hat{\beta}_p$ i.e. the least squares line.
- Hence we use $\hat{\beta}_0$ through $\hat{\beta}_p$ as guesses for β_0 through β_p and $\hat{Y}_i$ as a guess for Y_i. The guesses will not be perfect just as $\hat{X}$ s not a perfect guess for μ.

Data Analysis –Correlation & Linear Regression

Blog article: **Introduction to Linear Regression and Polynomial Regression**
https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

Supporting material: Video Session

- **P Values, clearly explained:** https://www.youtube.com/watch?v=5Z9OIYA8He8
- **Covariance:** https://www.youtube.com/watch?v=qtaqvPAeEJY
- **Pearson's Correlation:** https://www.youtube.com/watch?v=xZ_z8KWkhXE
- **R-squared explained:** https://www.youtube.com/watch?v=2AQKmw14mHM

- **Fitting a line to data,:** https://www.youtube.com/watch?v=PaFPbb66DxQ

- **Linear Models Pt. 1 - Linear Regression:** https://www.youtube.com/watch?v=nk2CQITm_eo
- **Linear Models Pt. 2 - t-tests and ANOVA:** https://www.youtube.com/watch?v=NF5_btOaCig
  (Pt. 2 → also for next lecture where ANOVA is introduced)
- **Linear Models Pt.3 - Design Matrices:** https://www.youtube.com/watch?v=CqLGvwi-5Pc
- **Linear Models Pt.4 - Design Matrix in R :** https://www.youtube.com/watch?v=Hrr2anyK_5s

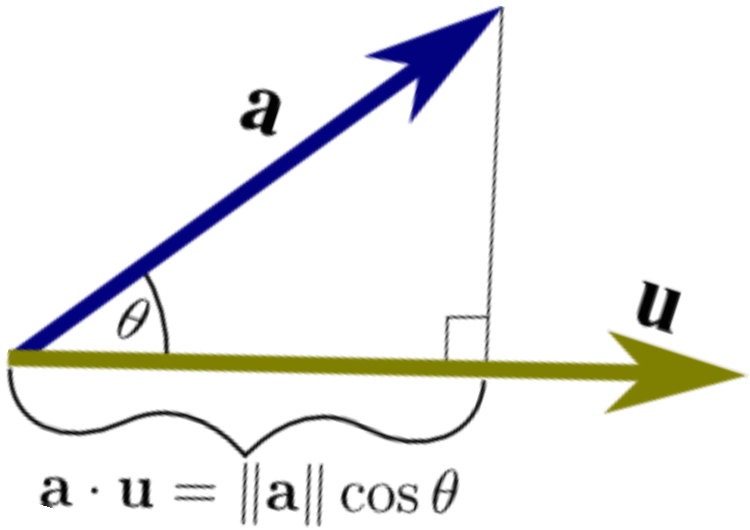
Data Analysis –Correlation & Linear Regression

Geometric Interpretation of the Correlation between Two Variables

Vectors and angles $a = (a_1, a_2, \dots a_n)$

Euclidean norm in R^n:

$$\|a\| = \sqrt{\sum_{i=1}^{n} a_i^2}.$$



Inner (scalar) product):

$$a \cdot b = \sum_{i=1}^{n} a_i b_i$$

$$a \cdot b = \|a\| \cdot \|b\| \cdot cos(a, b) \longrightarrow cos(a, b) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$

Data Analysis –Correlation & Linear Regression

Some statistics

Let X: random variable and n: size of a random sample

Expected value: $\mu = E(X) = \sum X_i p_i$ where p_i: probability of i −th event

average value of X: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Measure the scatter of the random variable → variance:

$$\sigma^2(X) = E[(X-\mu)^2] = \sum_i (X_i - \mu)^2 p_i$$

From variance → standard deviation: $\sigma = \sqrt{\sigma^2(X)}$

Data Analysis –Correlation & Linear Regression

Some statistics

"Standardization": $$x = \frac{X - \mu}{\sigma}$$

A measure of the relationship between two random variables is the covariance:

$$\sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)]$$

Correlation coefficient (normalized covariance):

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}} \rightarrow \rho_{X,Y} = \frac{\sum_{i=1}^{n}[(X_i - \bar{x})(Y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{y})^2}}$$

If we set $x_i = X_i - \bar{x}$ and $y_i = Y_i - \bar{y}$: $$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{x \cdot y}{\|x\| \cdot \|x\|} = cos(x, y)$$

Formal identity between the correlation coefficient and the cosine of the angle between 2 random vectors.