INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo
Dept. of Bio-industrial Mechatronics Engineering
National Taiwan University

Today:

- Linear regression

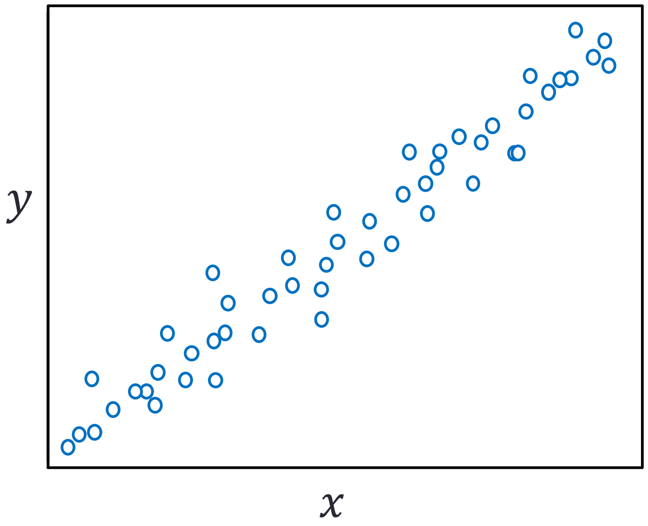
Outline

- Goal of the lecture
- Data dependency
- Simple linear regression
- Least squares
- Coefficient of determination
- Residual analysis
- Multiple regression

Goals

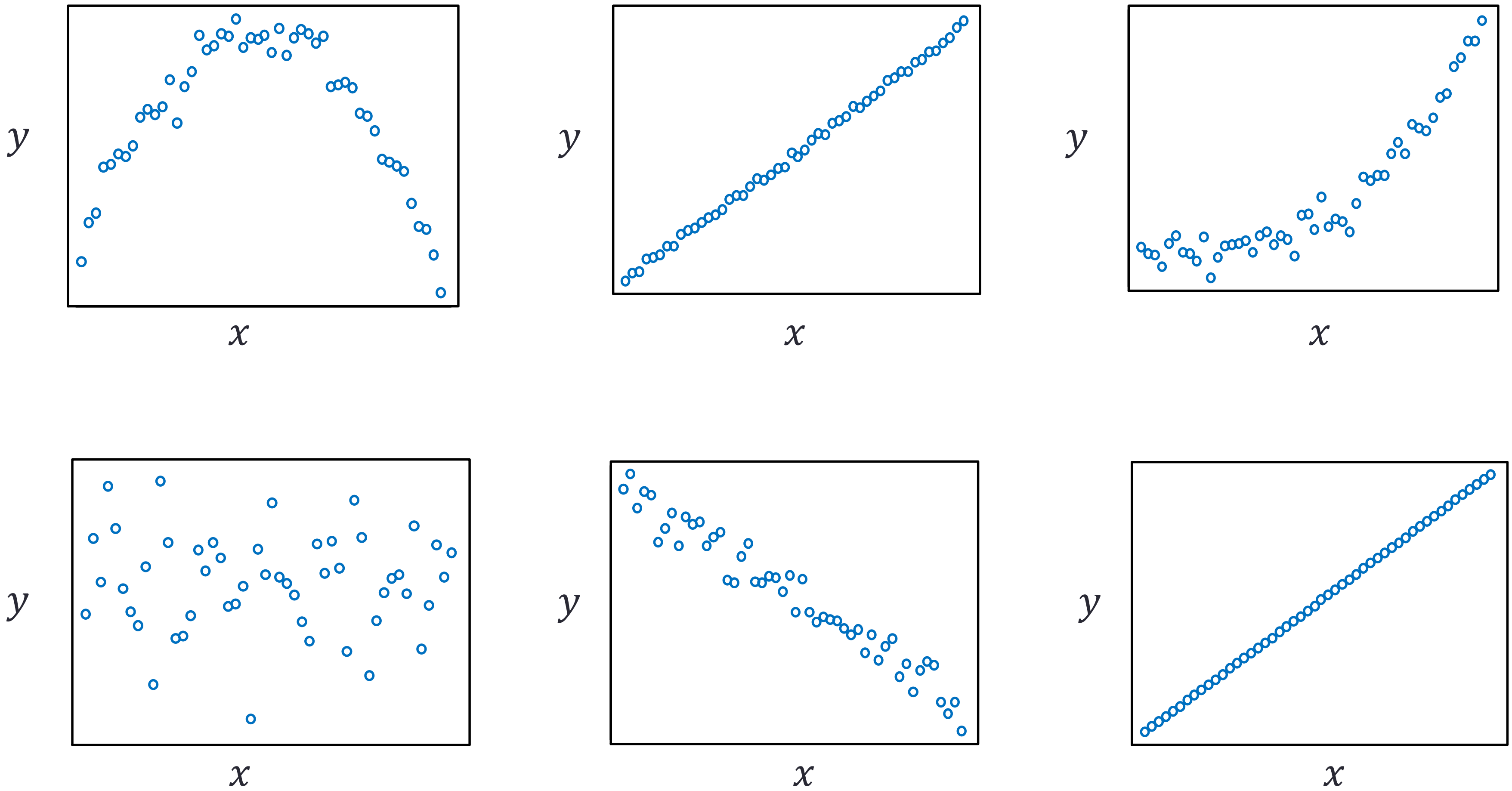
- After this, you should be able to:
  - Calculate and interpret the simple correlation between two variables
  - Calculate and interpret the simple linear regression equation for a set of data
  - Understand the assumptions behind regression analysis
  - Calculate the confidence interval for regression slope
  - Recognize some potential problems if regression analysis is used incorrectly

Scatter Plot



- The best way to view the relationship between two variables
- In some situations, we want to measure the dependency of one variable against another
- In other situations, we want to assess how the observed property matches the predicted property
- In all cases we will measure multiple samples or work with a population of subjects

Example Scatter Plots

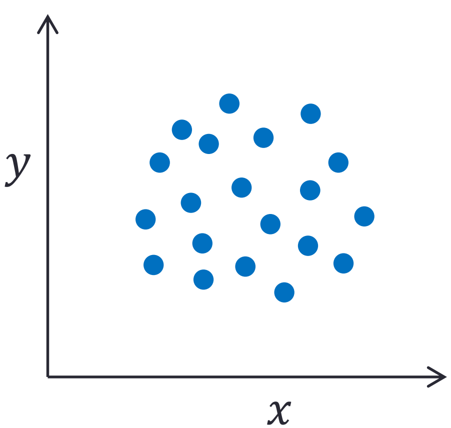


Correlation Analysis

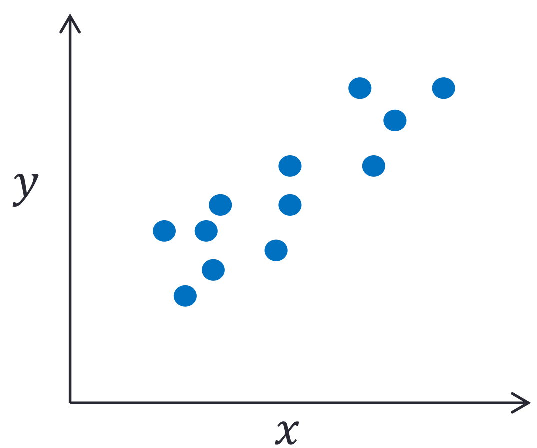
- Linear relationship between two variables x and y
- Correlation coefficient:

$$-1 \leq r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \leq 1$$

Low Correlation



Medium Correlation



High Correlation

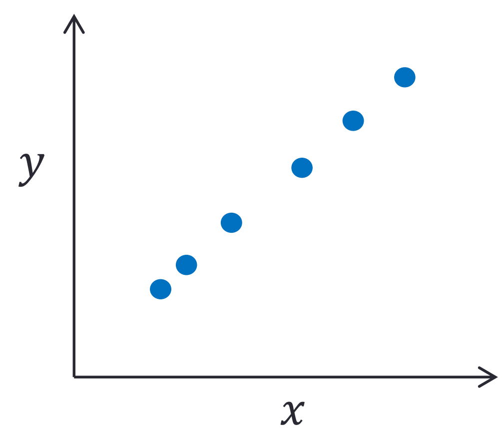
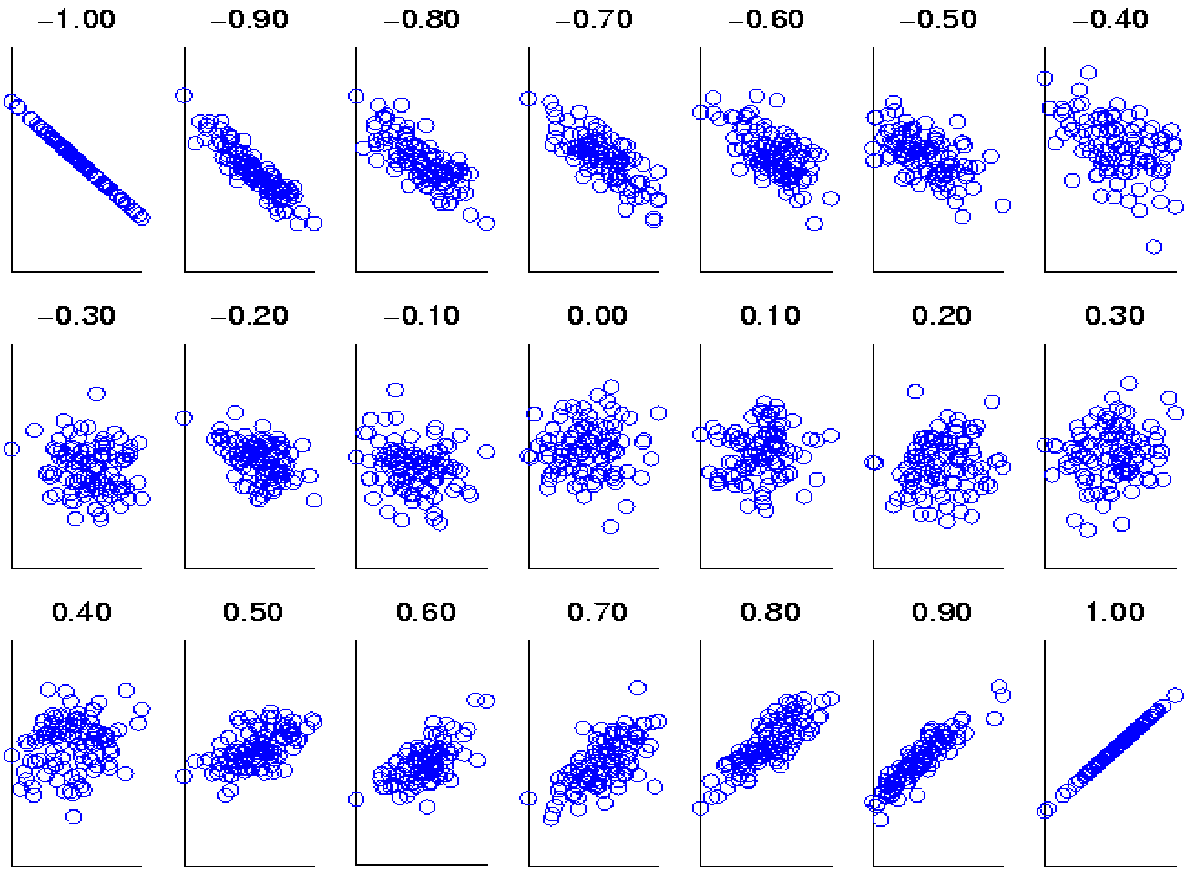


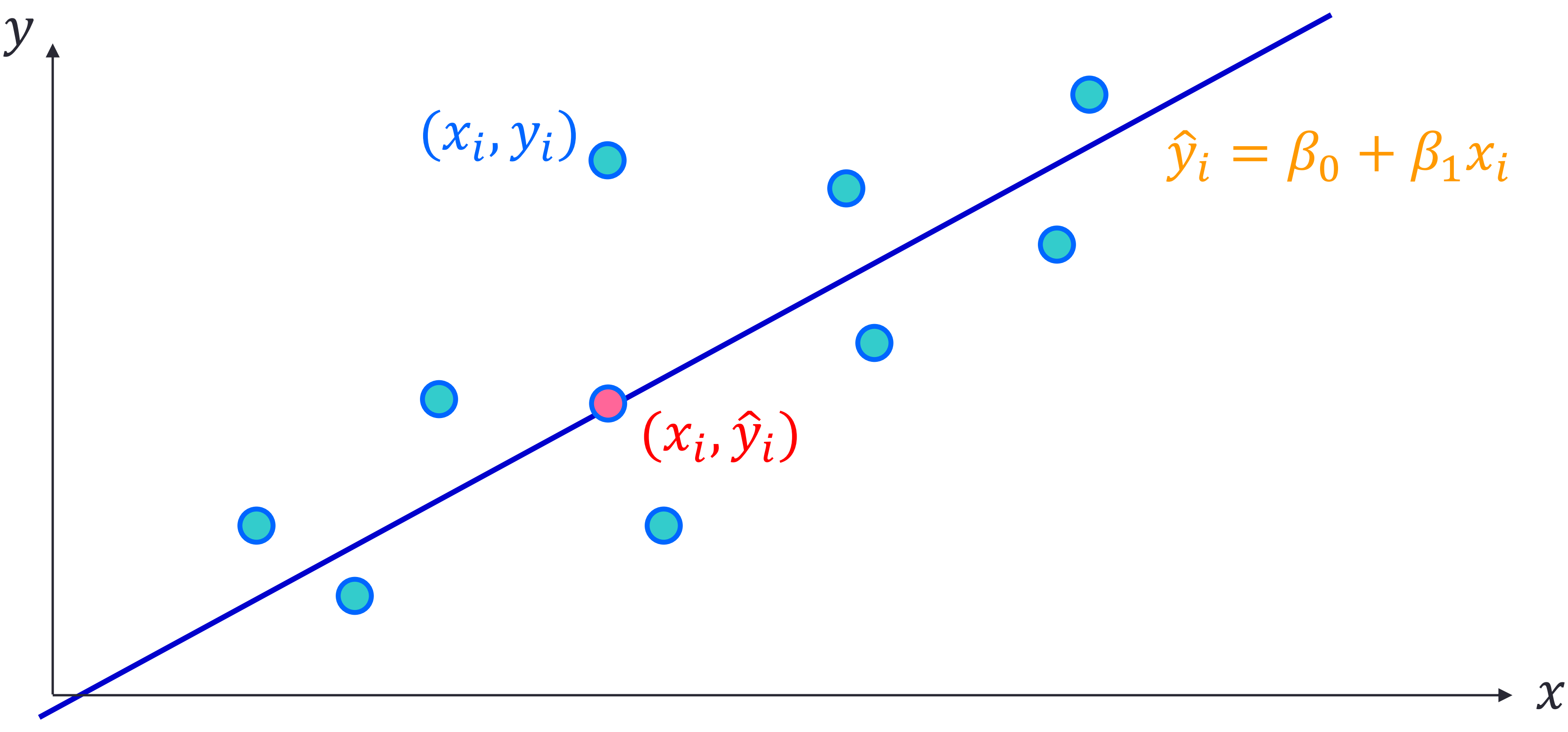
Illustration of Correlation Coefficient



Scatter plots showing the correlation coefficients ranged from –1 to 1

Simple Linear Regression

- A bunch of data points (x_i, y_i) are collected
- Assume x and y are linearly correlated



Simple Linear Regression Analysis

- Regression analysis is used to:
  - Predict the value of a response variable y based on the value of at least one explanatory variable x
  - Explain the impact of changes in an explanatory variable x on the response variable y

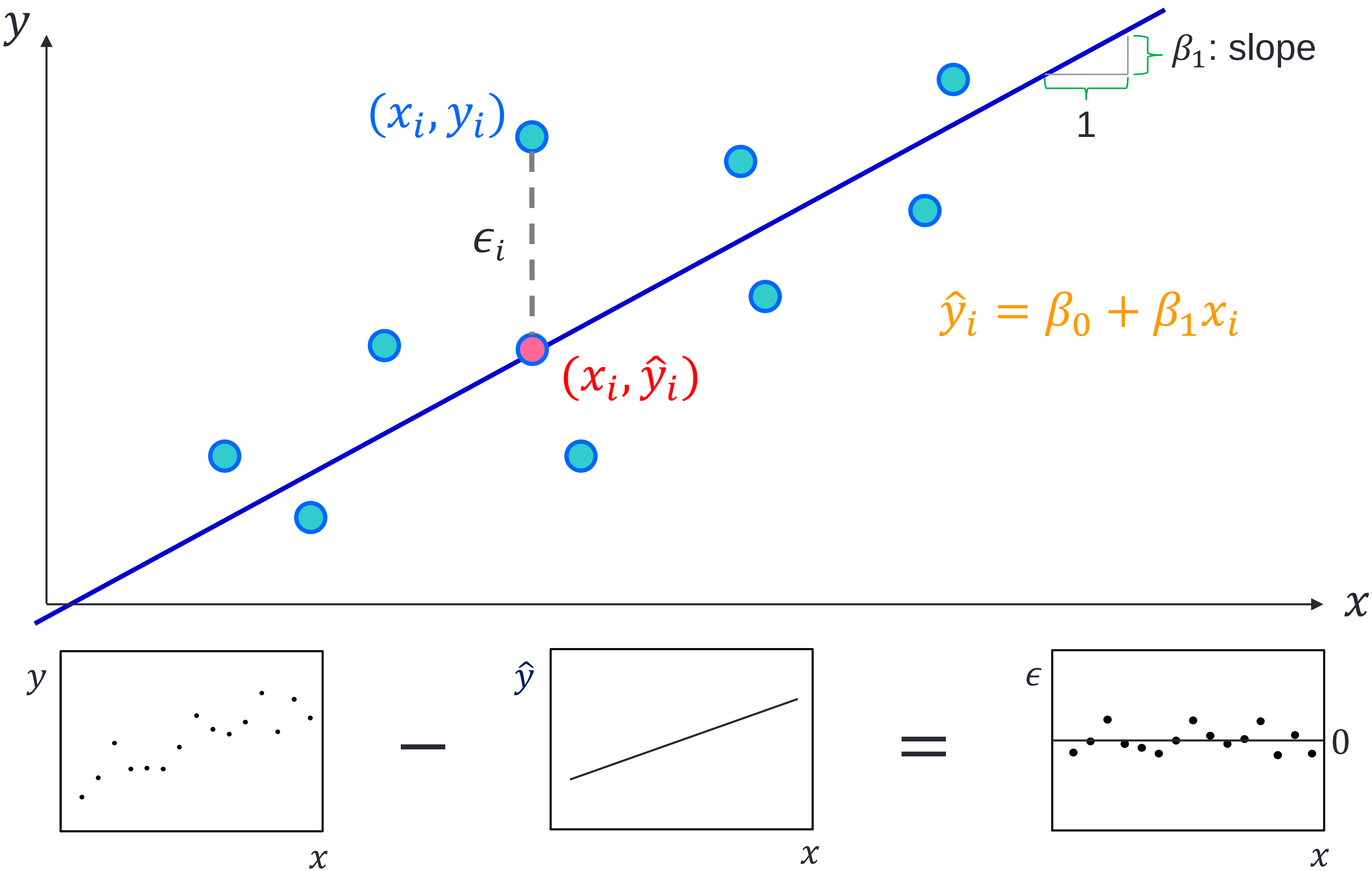
- Nominal relationship between x and y :

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

- Actual relationship between x and y :

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

The Error Term ϵ



Assumption – i.i.d.

- Relationship between x and y :

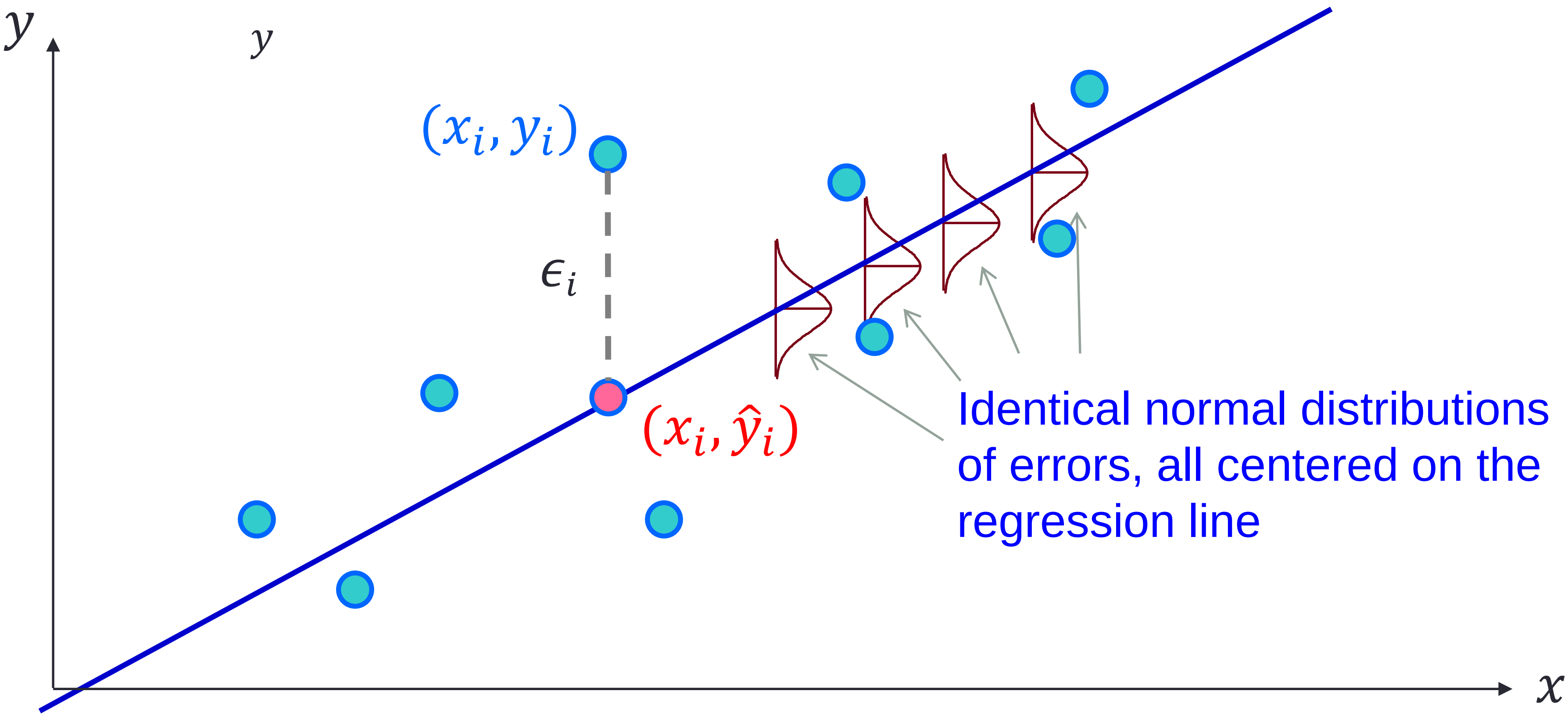
$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is assumed to be independently and identically distributed (i.i.d.)

- More specifically,

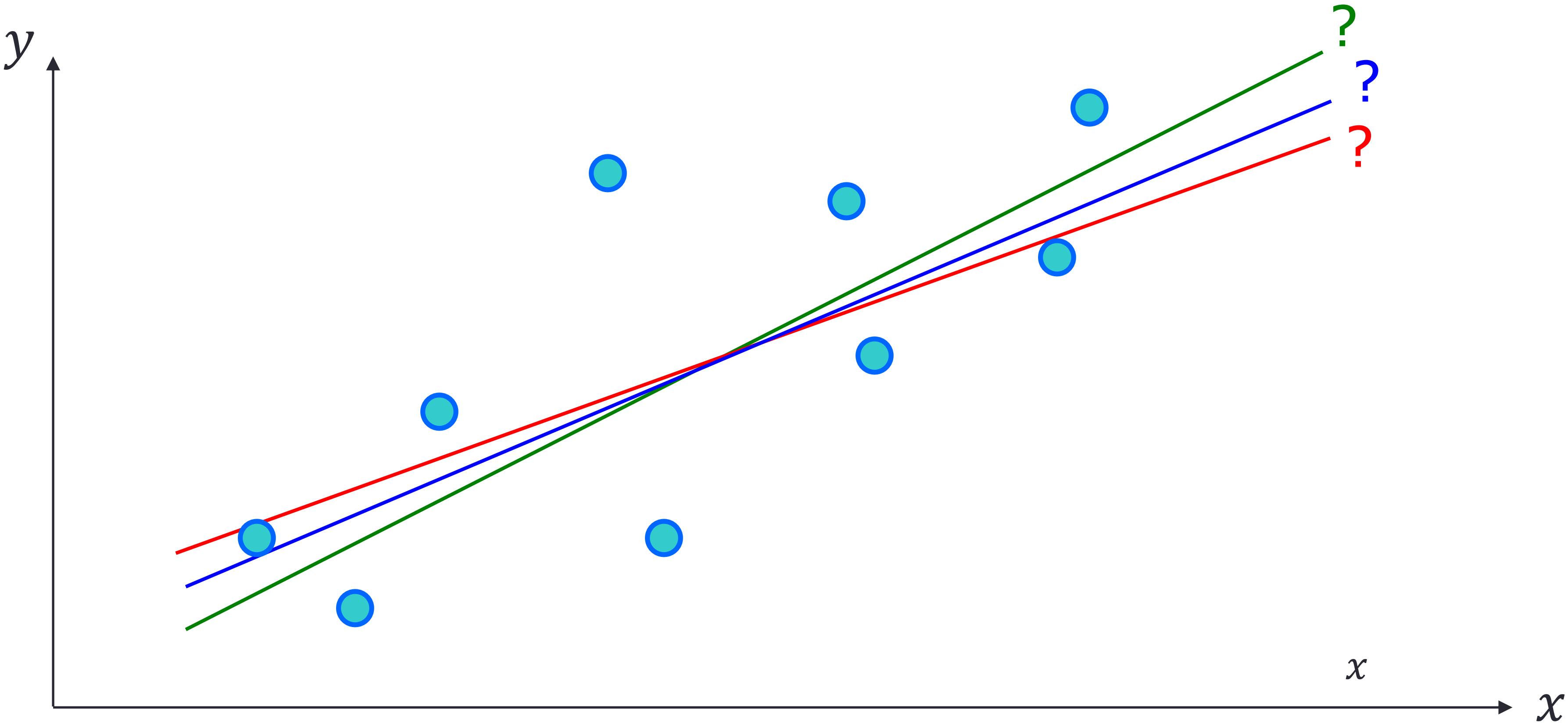
1. ϵ and x are independent, i.e., $P_{x,\epsilon}(x, \epsilon) = P_x(x) P_\epsilon(\epsilon)$
2. The probability distribution of the errors ϵ is normal

Observations



- The errors ϵ are uncorrelated in successive observations
- The errors ϵ are normally distributed, i.e., $\epsilon \sim N(0, \sigma^2)$

Which Line Best Fit to the Data?



- Through optimization!

Loss Function

- How does one mathematically define “best”?
- One has to define the “error” (or “loss”) first
- The loss may be, for example, the squared loss

$$loss(y, \hat{y}) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i {\epsilon_i}^2$$

- The goal is to minimize the error/loss on data points

Least-squares Method

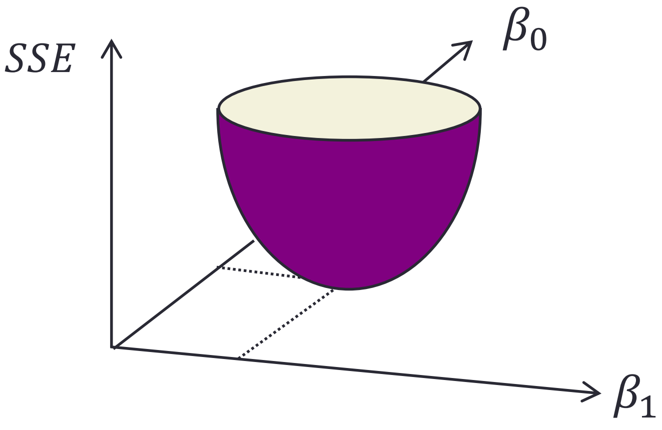
- Estimation of a simple linear regression relationship involves finding estimated values of the intercept β_0 and slope β_1 of the model

$$y = \beta_0 + \beta_1 x + \epsilon \quad \leftrightarrow \quad \hat{y} = \beta_0 + \beta_1 x$$

- Define sum of squared errors (SSE):

$$SSE = \sum_i {\epsilon_i}^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- Identify the β_0 and β_1 that minimize the SSE



Solving Least-squares Regression

- SSE is minimized when its gradient with respect to each parameter is equal to zero:

$$SSE = \sum_i {\epsilon_i}^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Solving Least-squares Regression

- Suppose there exists N data points:

$$\sum_{i=1}^{N} y_i = \beta_0 \cdot N + \beta_1 \sum_{i=1}^{N} x_i$$

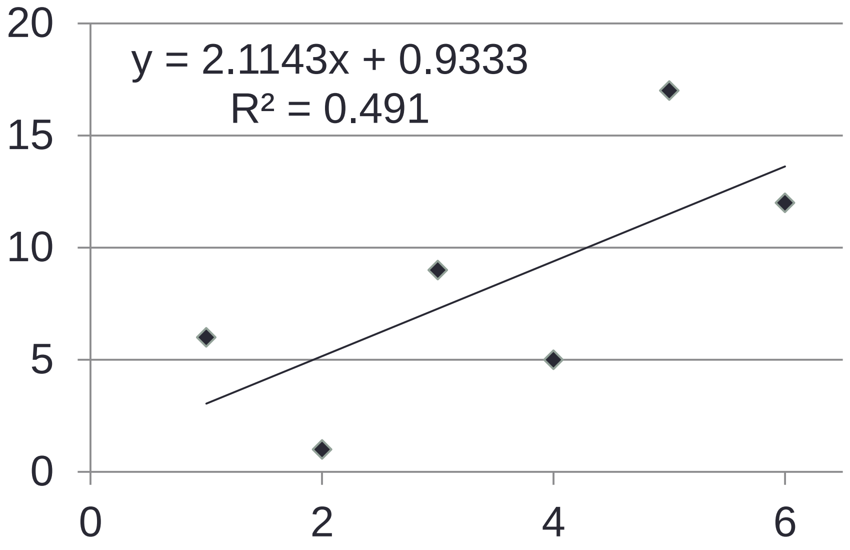
$$\sum_{i=1}^{N} y_i x_i = \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} {x_i}^2$$

$$\Rightarrow \begin{bmatrix} \sum y_i \\ \sum y_i x_i \end{bmatrix} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Simple Regression Example

- Suppose we have:

x	y
1	6
2	1
3	9
4	5
5	17
6	12



Least Squares Regression Properties

- The sum of the squared residuals is a minimum, i.e. minimal $\sum(y_i - \hat{y}_i)^2$
- The sum of the residuals from the least squares regression line is zero, i.e. $\sum(y_i - \hat{y}_i) = \sum \epsilon_i = 0$
- The simple regression line always passes through the mean of the response variable $\bar{y}$ and the mean of the explanatory variable $\bar{x}$
- The least squares coefficients are unbiased estimates of β_0 and β_1

Assessing the Model

- The least squares method will always produce a straight line
- Determining regression model coefficients is easy, but…

1. How does one access the model and know how well it fits the data?
    - One common method – check how much variation is explained by the model
2. Particularly, what if the relationship between the variables is NOT linear?

Explained and Unexplained Variation

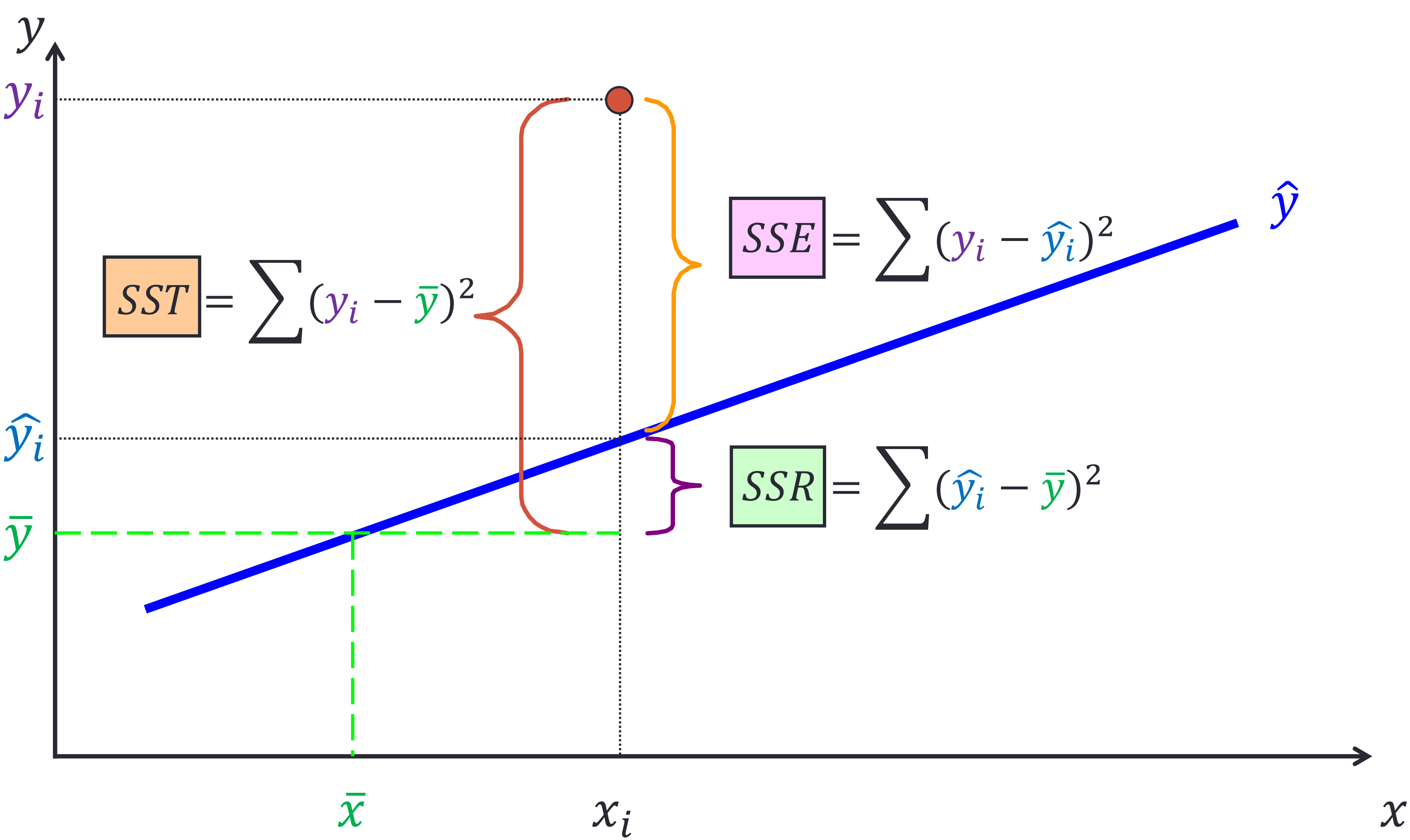
- Total variation is made up of two parts:

$$SST \quad = \quad SSE \quad + \quad SSR$$

Sum of squares total	Sum of squares error	Sum of squares regression
$SST = \sum (y_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$

- SST measures the variation of the y_i values around their mean $\bar{y}$
- SSE represents the variation attributable to factors other than the relationship between x and y
- SSR explains variation attributable to the relationship between x and y

Explained and Unexplained Variation (Cont’d)



Proof of $SST = SSE + SSR$

- Starting from $(y_i - \bar{y}) = [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]$:

$$\sum_i (y_i - \bar{y})^2 = \sum_i [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$= \sum_i (y_i - \hat{y}_i)^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_i (\hat{y}_i - \bar{y})^2$$

- The middle expression:

$$2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2\sum_i [\hat{y}_i(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)]$$

$$= 2\sum_i \hat{y}_i \epsilon_i - 2\sum_i \bar{y}\epsilon_i = 2\sum_i (\beta_0 + \beta_1 x_i)\epsilon_i - 2\bar{y}\sum_i \epsilon_i$$

$$= 2\beta_0 \sum_i \epsilon_i + 2\beta_1 \sum_i x_i \epsilon_i - 2\bar{y}\sum_i \epsilon_i = 0$$

Coefficient of Determination

- The coefficient of determination R^2 is a measure of how well the regression line fits the data
- The coefficient of determination is the portion of the total variation in the response variable that is explained by variation in the explanatory variable:

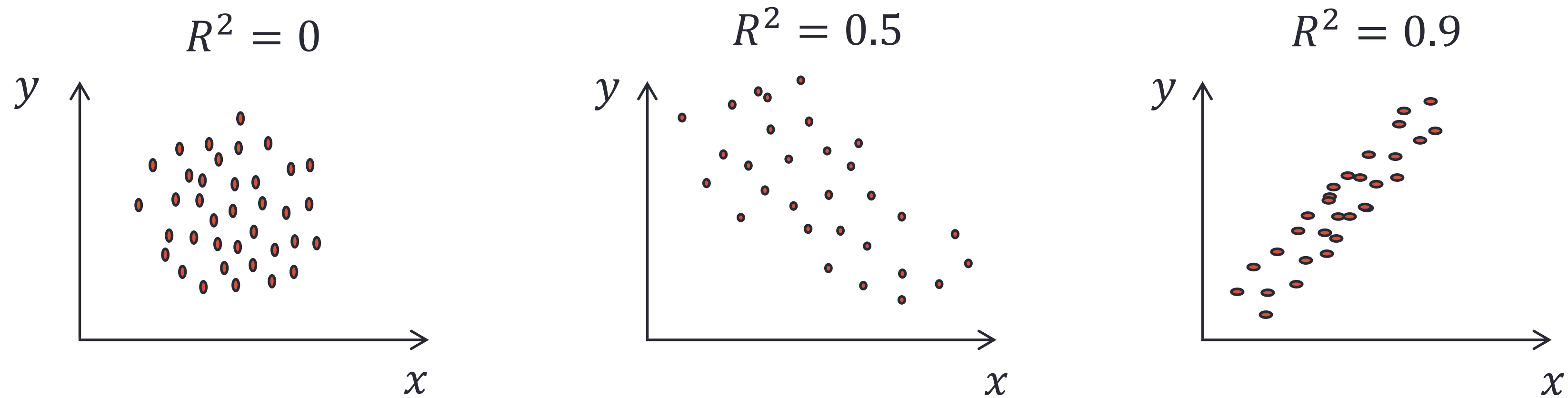
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\text{SS explained by regression}}{\text{total SS}}$$

where $0 \leq R^2 \leq 1$

- In the simple regression, the coefficient of determination is equal to the square of correlation coefficients, i.e., $R^2 = r^2$

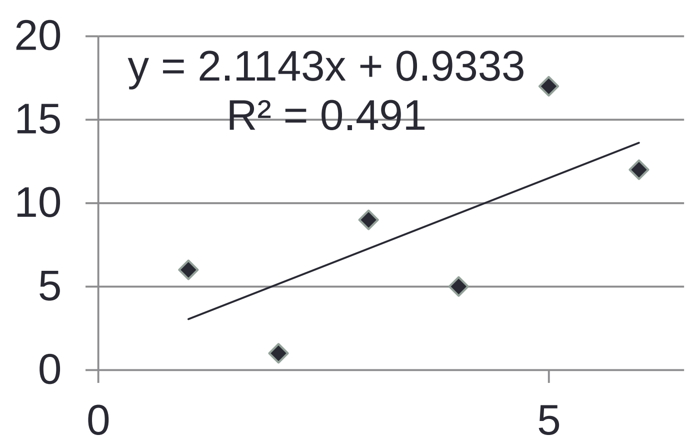
Example Coefficient of Determination

- Illustration of R^2

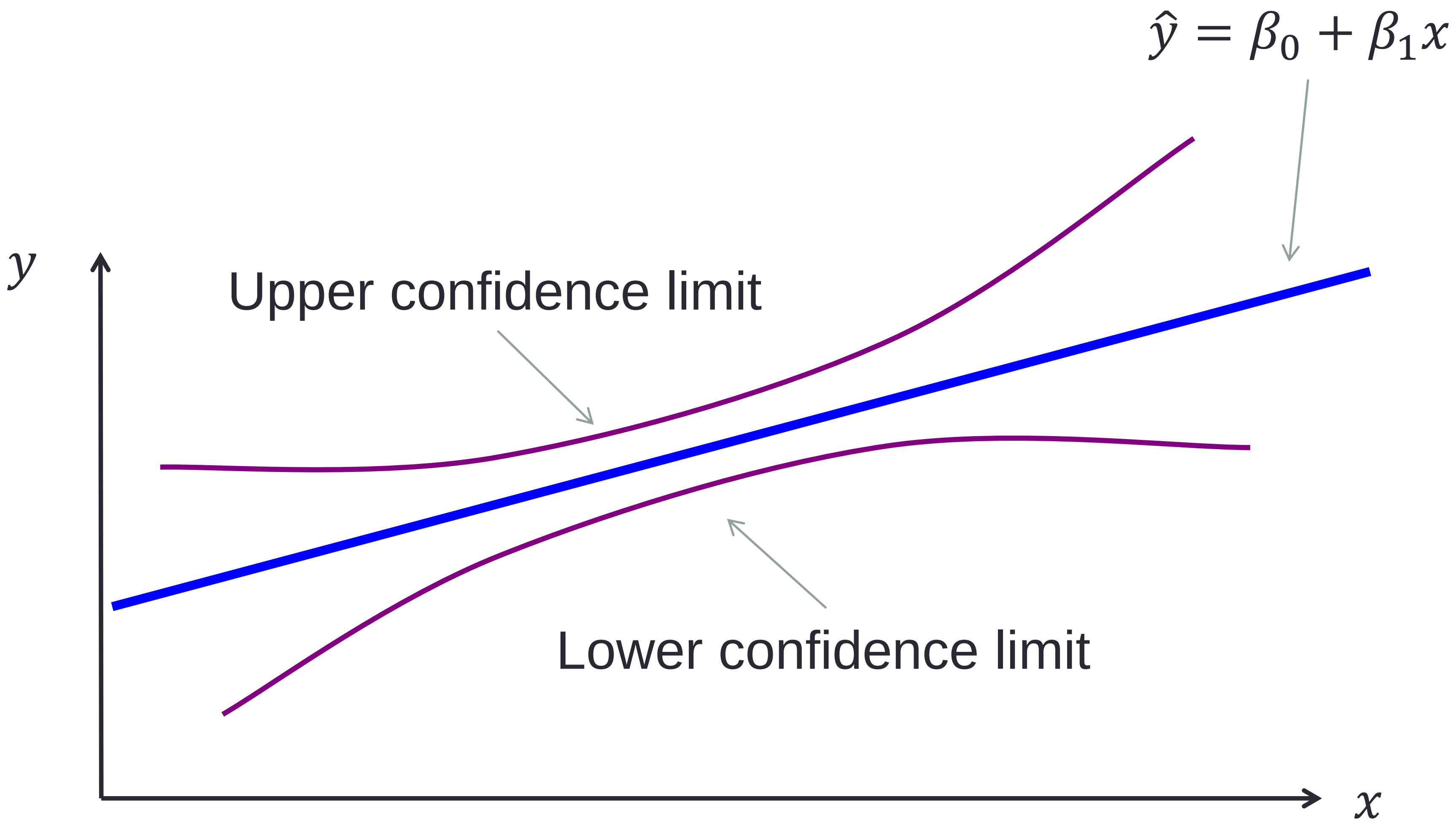


- Examples of coefficient of determination:

x	y
1	6
2	1
3	9
4	5
5	17
6	12



Interval Estimates



Standard Error and Confidence Interval of β_1

- Standard error of the regression line slope is defined as:

$$s_{\beta_1} = \frac{s_e}{\sqrt{\sum {x_i}^2 - \frac{(\sum x_i)^2}{N}}}$$

- The $(1-\alpha)\%$ confidence interval of the regression line slope is defined as:

$$s_{\beta_1} \cdot t_{\frac{\alpha}{2},N-2} \leq \beta_1 \leq s_{\beta_1} \cdot t_{\frac{\alpha}{2},N-2}$$

where $t_{\frac{\alpha}{2},N-2}$ is the $\frac{\alpha}{2}$th percentile t-distribution for $N-2$ degrees of freedom

Standard Deviation of the Residuals

- Sample statistics are point estimates for the population parameters, which is unknown
- The standard deviation of the residuals s_e, for all points in the population, is estimated by the standard deviation of the residuals:

$$s_e = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum {y_i}^2 - \beta_0 \sum y_i - \beta_1 \sum x_i y_i}{n-2}},$$

where $n-2$ is the degree of freedom

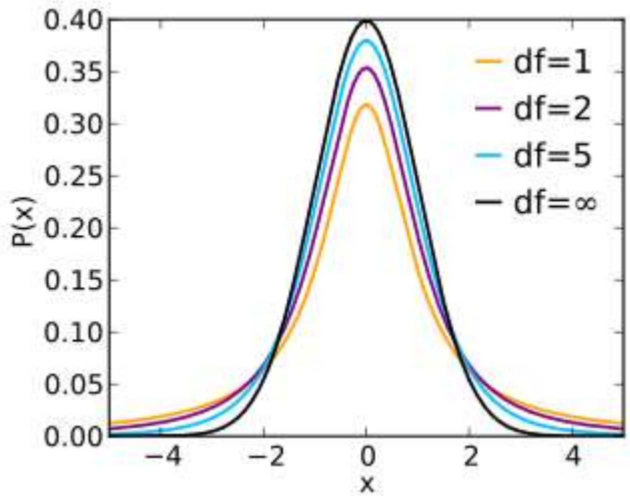
Student's t-distribution

- A continuous probability distribution for estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown
- Published in 1908 by William Sealy Gosset using the pseudonym "student"

- The shape of the t-distribution is similar to that of the normal distribution

Student's t-distribution (Cont'd)

- There are many different t-distributions, one for each degree of freedom
- For small degrees of freedom, the t-distribution is very dispersed
- The limiting distribution for the t distribution is the normal distribution



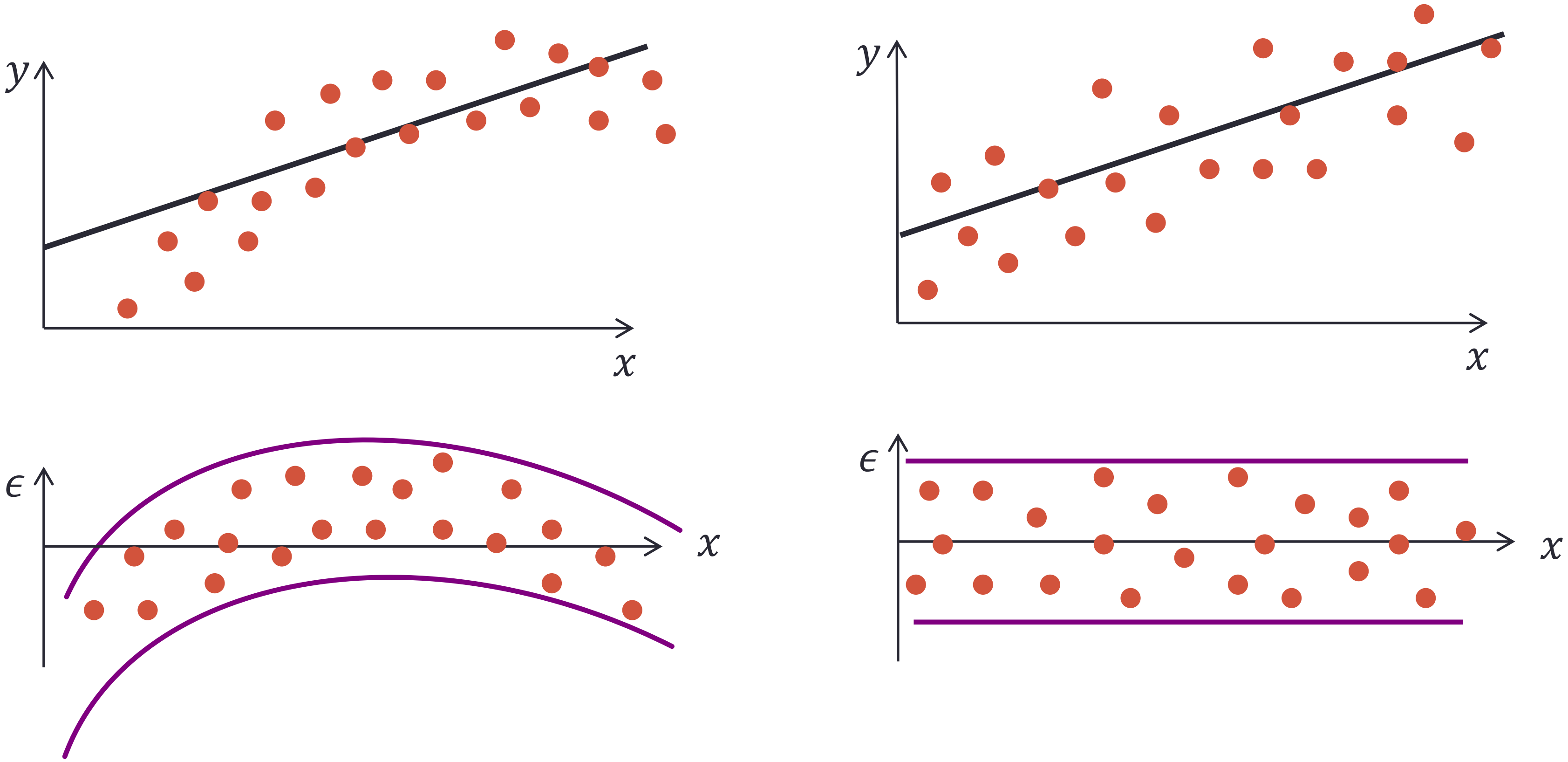
Linear Regression Assumptions

1. Error values ϵ are independent to the explanatory variable x
2. The probability distribution of the errors ϵ is normal
3. The probability distribution of the errors ϵ has constant variance
4. The underlying relationship between the explanatory variable x and the response variable y is linear

Residual Analysis

- Perform residual analysis to check any violation of the assumption
- The residual is the difference between its observed and predicted value, i.e. $\epsilon_i = y_i - \widehat{y}_i$
- Check the following assumptions:
  1. Linearity
  2. Homoscedasticity (constant variance)
  3. Normal distribution
  4. Independence

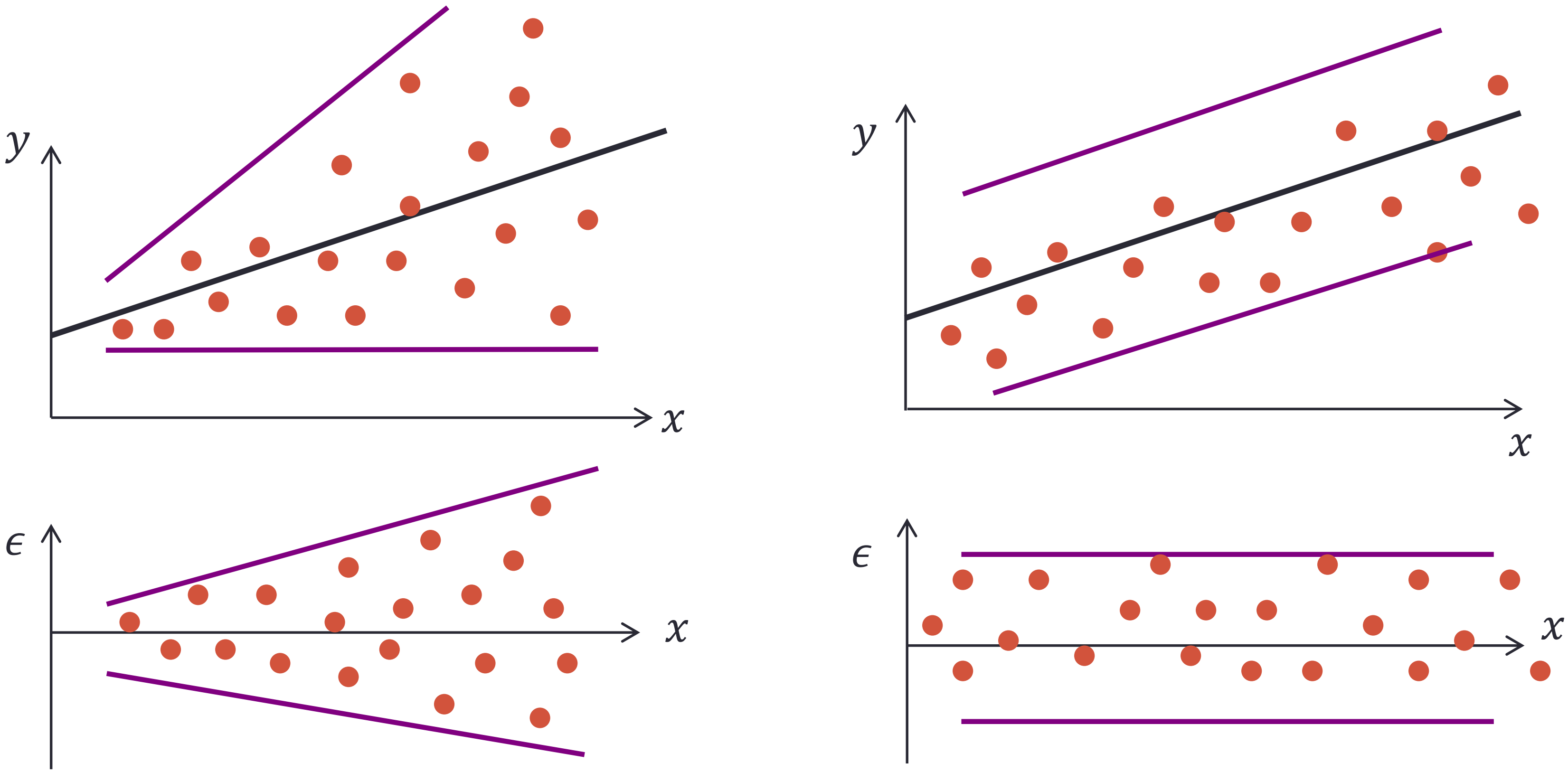
Residual Analysis for Linearity



Remedy for Violating Linearity

1. Piecewise linear regression:
   - Break down (x_i, y_i) into j sets, i.e., $\left({x_i}^{(j)}, {y_i}^{(j)}\right)$ where $\{x_i\} = \sum_j \{{x_i}^{(j)}\}$ and $\{y_i\} = \sum_j \{{y_i}^{(j)}\}$
   - Perform regression for each set ${y_i}^{(j)} = \beta_0 + \beta_1 {x_i}^{(j)}$
2. Variable transformation
   - Polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots$
   - Logarithm: $y = \beta_0 + \beta_1 \log(x)$ or $\log(y) = \beta_0 + \beta_1 x$
   - Exponential: $y = \beta_0 + \beta_1 e^x$
   - Inverse: plus $y = \beta_0 + \beta_1 \frac{1}{x}$

Residual Analysis for Homoscedasticity

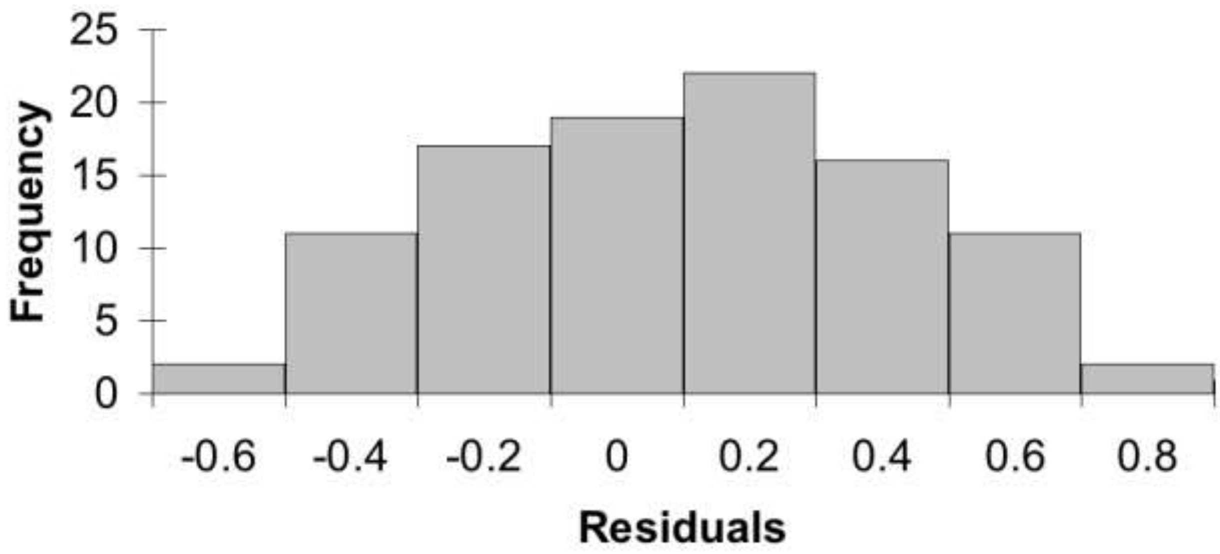


Remedy for Violating Homoscedasticity

- Divide the entire equation by x, e.g., . $y = \beta_0 + \beta_1 x$ will become $\frac{y}{x} = \frac{\beta_0}{x} + \beta_1$
- Notice that for large values of x the new error (ϵ/x) will be smaller

Normality

- Residual histogram

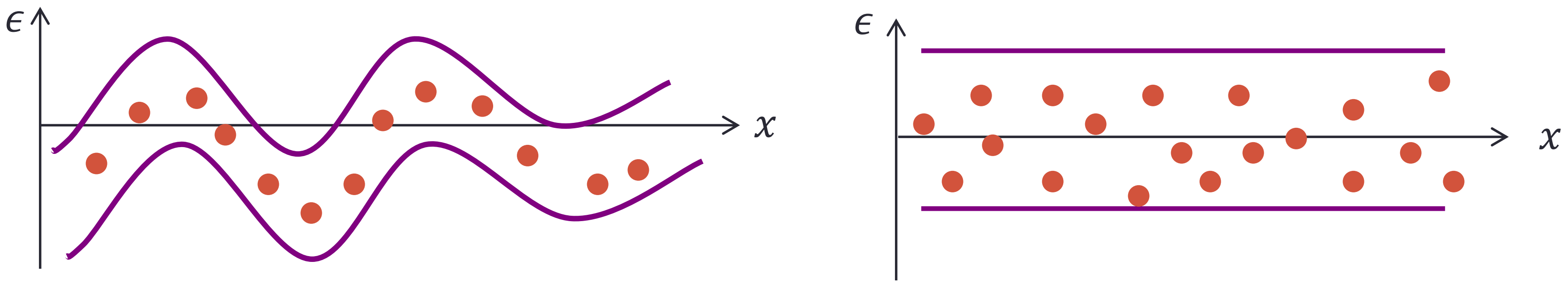


- Traditional way to test for normality – looking for a bell shaped histogram with the mean close to zero

Remedy for Violating Normality

- Transform the response variable to make the distribution of the random errors approximately normal
- Three transformations that are often effective for making the distribution of the random errors approximately normal:
  - $\sqrt{y}$
  - $\ln(y)$
  - $\frac{1}{y}$

Residual Analysis for Independence



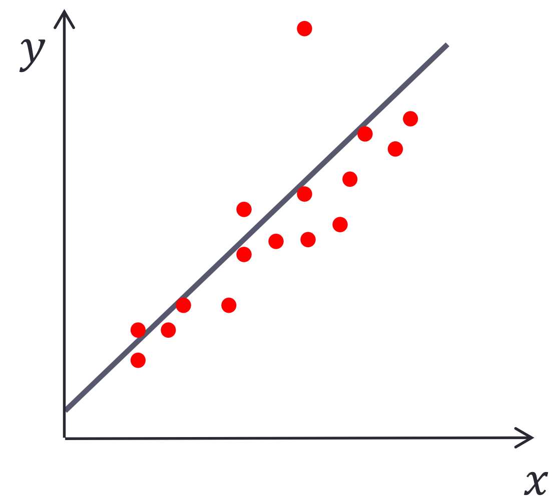
- A mathematical representation of the degree of periodical similarity between variable x over successive intervals
- This is called autocorrelation
- If a pattern emerges, it is likely that the independence requirement is violated

Remedy for Violating Independence

- For serial (temporal) correlation, include new variables in the equation, e.g., the value of y at moment $t - 1$ as an independent variable
- For spatial correlation, model the relationships by introducing an weighting matrix

Outlier

- An outlier is a data point that is unusually small or large
- Possible reasons for the existence of outliers include:
  - There was an error in recording the value
  - The point should not have been included in the sample
- Outliers can be easily identified from a scatter plot
- Outliers need to be dealt with since they can easily influence the regression model



Procedure for Regression Analysis

1. Gather data for the two variables in the model
2. Draw the scatter diagram to determine whether a linear model appears to be appropriate
3. Determine the regression equation
4. Assess the model’s fit
5. Calculate the residuals and check the required conditions

Multiple Regression

- What if we have multiple explanatory variables $\boldsymbol{x} = [x_1, x_2, \dots, x_M]$ and one response variable y?
- The multiple regression model:

$$y = f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M$$

- For example: polynomial model
- Denote the i^{th} observation as $\boldsymbol{x}_i = [x_{i1}, \dots, x_{iM}]^{\mathrm{T}}$

- Observation

$$\text{Observation}\left\{\overbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NM} \end{bmatrix}}^{\text{Parameter}}\right. = \boldsymbol{X} \quad \Rightarrow \quad \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \boldsymbol{y}$$

Formulating Multiple Regression

- Let $\boldsymbol{\beta} = [\beta_0 \quad \cdots \quad \beta_M]^{\mathrm{T}}$
- Data: $y_i = [1 \quad \boldsymbol{x}_i]^{\mathrm{T}}\boldsymbol{\beta} + \epsilon,\ \ \epsilon \sim N(0, \sigma^2)$
- Matrix form: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NM} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_M \end{bmatrix}$$

- Problem statement:

  Given the vector $\boldsymbol{y}$ and matrix $\boldsymbol{X}$ above, find the coefficient $\boldsymbol{\beta}$ of the regression model $\hat{y} = [1 \quad \boldsymbol{x}]^{\mathrm{T}}\boldsymbol{\beta}$ that most accurately predicts y

Solving Multiple Regression

- Sum of squares error:

$$SSE = \sum_{i=1}^{N} {\epsilon_i}^2 = \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X\beta})$$

$$= \left(\boldsymbol{y}^{\mathrm{T}} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\right)(\boldsymbol{y} - \boldsymbol{X\beta})$$

$$= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\beta} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}$$

$$= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}$$

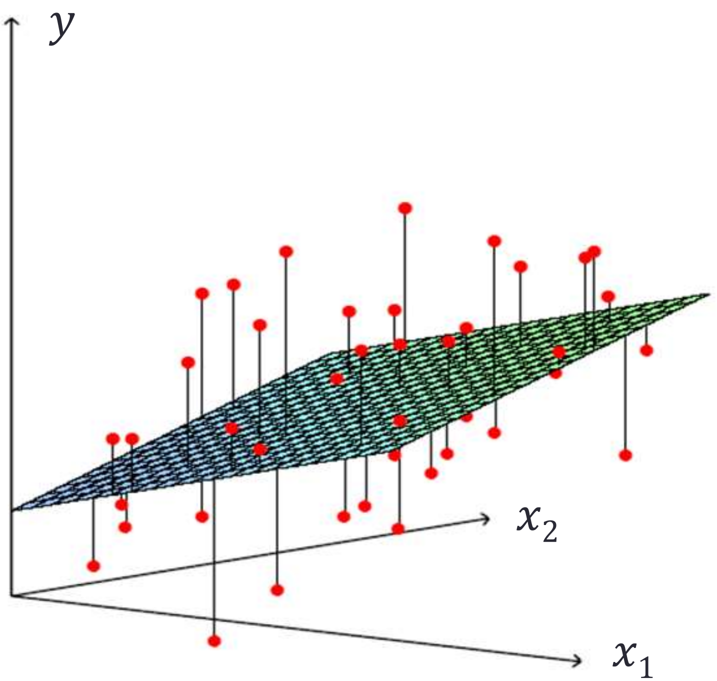
$$\frac{\partial SSE}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta} = 0$$

Multiple Regression

- Model coefficient $\boldsymbol{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$
- Recall regression model $\hat{y} = [1 \quad \boldsymbol{x}]^{\mathrm{T}}\boldsymbol{\beta}$
- Geometric explanation
- What about multivariate regression?

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B},$$

where $\boldsymbol{B}$ is the coefficient matrix



Summary

- Scatter plot is a useful diagnostic tool for determining association between variables
- Coefficient of determination provides a measure of how well future outcomes are likely to be predicted by the model
- There is an confidence interval for every statistic estimation
- The less the data samples, the smaller the degree of freedom, and the larger the confidence interval
- Always check the residual after performing regression analysis

References

- T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*
- M. Dodge and C. Stinson, *Microsoft® Office Excel® 2007 Inside Out*