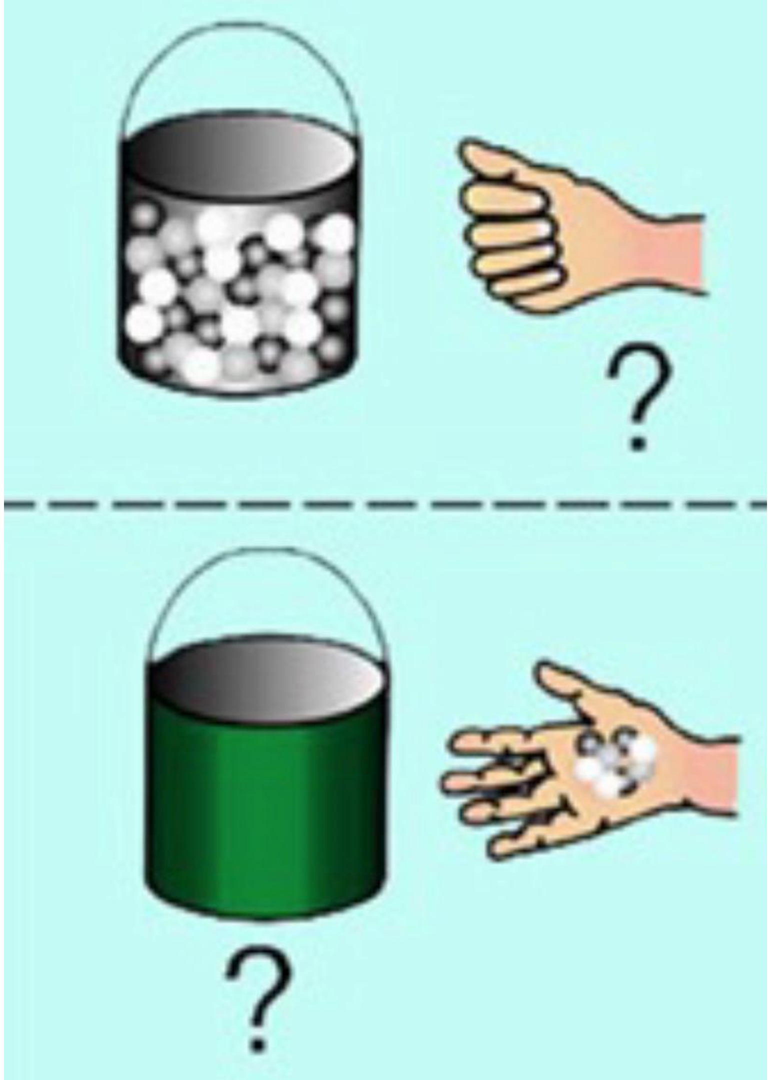
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Introduction to Data Science

Lecture 11: Statistics
Point Estimation: Regression Analysis

Zicheng Wang

Recap



Probability

- **Knowing** the true **model**, you can **predict** the **samples** outcome.

Statistics

- **Knowing** the generated **samples**, you can **estimate** the true **model**.

Statistics

- Target: **extract some useful information** about the data.

- What you need?
  - Samples
  - A set of possible models
  - Criterion for quantifying model performance

Maximum likelihood estimate (MLE)

- Target: estimate θ of a model

- Samples: $X_1,\ X_2 \dots,\ X_n$
- Possible models: $\theta \in \Theta$ (additional information)
- Model performance (probability): L(θ)= P ($X_1,\ X_2 \dots,\ X_n | \theta$)

- Estimator: $\hat{\theta}$ such that L(θ) is maximized at θ= $\hat{\theta}$.

Formal Definition

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$
- The probability that the models generates the samples is called **likelihood**. L(θ)= P ($X_1,\ X_2 \ldots,\ X_n | \theta$)
- To determine the best θ, we choose $\hat{\theta}$ such that L(θ) is maximized at $\theta = \hat{\theta}$. **Maximum likelihood estimate (MLE)**

Likelihood function

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$

Continuous RV model:

- Likelihood: L(θ)= Π_i f($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log($f$(X_i\,|\,\theta))$

f: the model's probability density function (PDF)

Discrete RV model:

- Likelihood: L(θ)= Π_iP ($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\mathrm{P}(X_i\,|\,\theta))$

P: the model's probability mass function (PMF)

Likelihood function

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$

<u>Continuous RV model:</u>

- Likelihood: L(θ)= $\boxed{\Pi_i}$ f($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)= \sum_i \log(\mathrm{f}(X_i\,|\,\theta))$

<u>Discrete RV model:</u>

- Likelihood: L(θ)= $\boxed{\Pi_i}$P ($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)= \sum_i \log(\mathrm{P}(X_i\,|\,\theta))$

Why do we multiply the pdfs or pmfs here?

Because we assume those samples are independent observations

Example: no additional information

- For drug experiments,
    - **N** experiments
    - **M** successes

- Possible models: Bernoulli distribution with probability p.

- Given one model, the probability of generating such samples:

  L(p) = $p^{\boldsymbol{M}}(1-p)^{\boldsymbol{N}-\boldsymbol{M}}$

- Most likely: the model **maximizes** the probability L(p)

Example: no additional information

- Given one model, the probability of generating such samples:
  $$L(p) = p^{\boldsymbol{M}}(1-p)^{\boldsymbol{N}-\boldsymbol{M}}$$

- Maximize the probability

We often maximize the **logarithm of** the probability (usually easier to maximize) of generating such samples

$$l(p) = \log L(p) = \mathbf{M} * \log p + (\mathbf{N}-\mathbf{M}) * \log(1-p)$$

Example: no additional information

- To maximize a continuous function l(p)
  - Find the point ***l'*(p) = 0** and l''(p)<0
  - $p \in [0,1]$

l(p)=log L(p) = **M** * log p + (**N**-**M**) * log (1-p)

- $l'(p)$= $\frac{M}{p} + \frac{N-M}{1-p}$ = 0 $\Rightarrow$ p = **M**/**N** (sample cure rate)

Example: with additional information

- You may have additional information.
  - Drug 1: p = 0.8
  - Drug 2: p = 0.9
  - Samples: M=81 successes from $\mathrm{N} = 100$ experiments.

- Potential models: Bernoulli with $p \in \{0.8, 0.9\}$

l = Log L = M*log p+(N-M)*log (1-p)

- Drug 1: -48.6539
- Drug 2: -52.2833

Conclusion: drug 1 is more likely to be experimented.

An Additional Example of MLE
Baseball Team

- The weights for a baseball team players are

{150, 143, 132, 160, 175, 190, 123, 154}

- Assume their weights are **uniformly** distributed over an interval [a, b]

- What are good estimators for a? for b?

This example will show that the **MLE could be complicated to solve**, e.g., the equation $l'(\theta) = 0$ may be difficult to solve, or it may **not** always be possible to use calculus methods **directly** to find the maximum of L(θ).

MLE: Uniform

Let *X* be a Uniform random variable on the interval [*0, θ*]

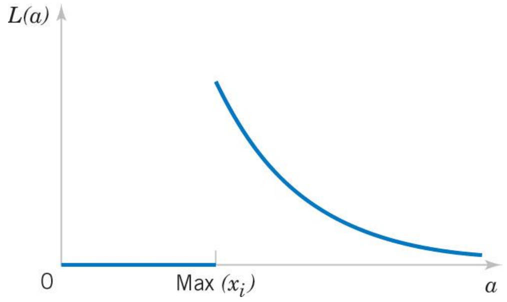
$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \le x \le \theta, \\ 0, & \text{otherwise,} \end{cases} = \frac{1}{\theta}\mathbf{1}_{\{0 \le x \le \theta\}}$$

indicator function $\mathbf{1}_A(x)$

$$\mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise} \end{cases}$$

The likelihood function of a random sample of size *n* is:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}_{\{0 \le x_i \le \theta\}} = \begin{cases} \frac{1}{\theta^n}, & \text{if } \theta \ge max\{X_1, X_2, \ldots, X_n\} \\ 0, & \text{if } \theta < max\{X_1, X_2, \ldots, X_n\} \end{cases}$$



$$\hat{\theta} = max\{X_1, X_2, \ldots, X_n\}$$

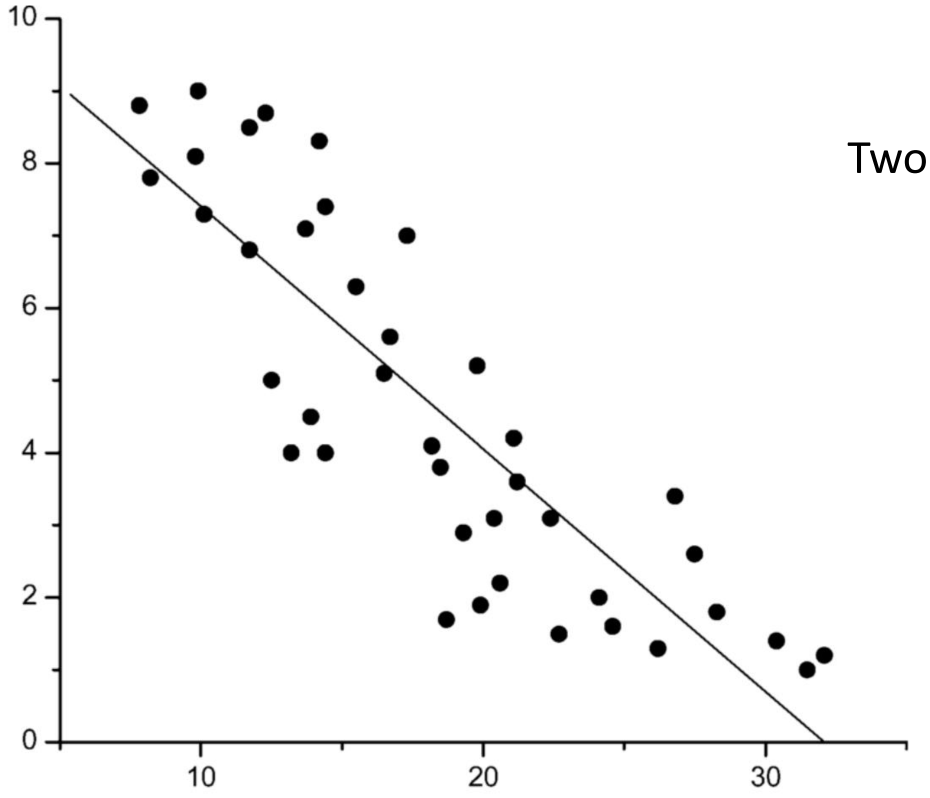
Calculus methods don't work here because $L(\theta)$ is maximized at the **discontinuity**.
Clearly, θ cannot be smaller than max(x_i), thus the MLE is $max\{X_1, X_2, \ldots, X_n\}$.

More discussion on MLE

- Is MLE of the mean of a distribution always the sample average?
- Not always!
- Ture for Normal distribution, Bernoulli distribution…
- False for uniform distribution…

Linear regression

Find the relationship between X and Y



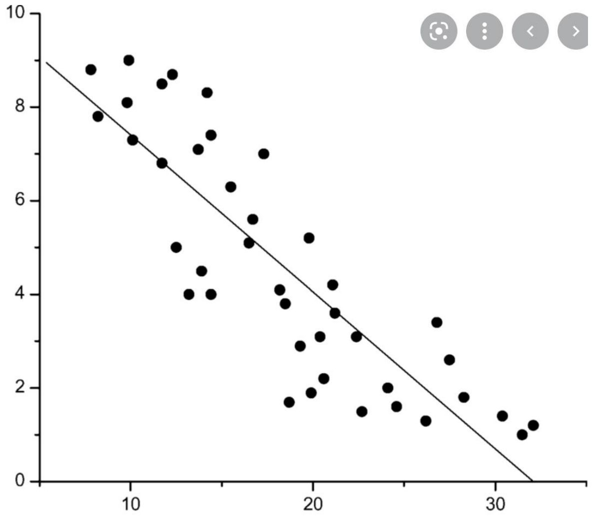
Two Stocks (X,Y)

Correlation

- With a positive correlation (positively correlated),
  - Larger x implies larger y (vice versa)

- With zero correlation (uncorrelated),
  - No clear relationship between x and y

- With a negative correlation (negatively correlated),
  - Larger x implies smaller y (vice versa)



- Negative: larger x implies smaller y.

- **Question:** when x increases by a certain quantity, what's the reduction in y?

- Use a line to approximate the relationship:
  - Regression analysis.

Collect past data

Propose some models

Probability: quantify uncertainty.

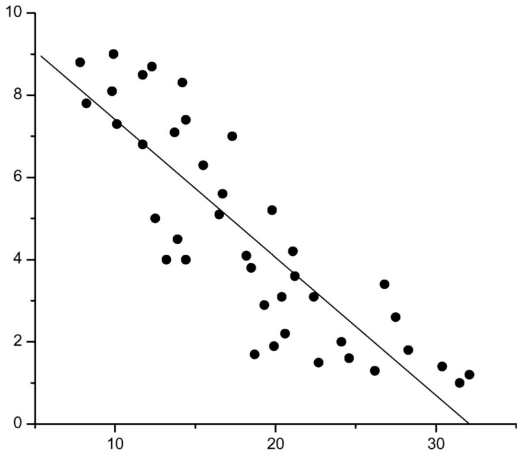
Choose the best model

Statistics: test credibility.

Prediction for given input

Optimize input

Make a decision



What model to propose?

Collect past data

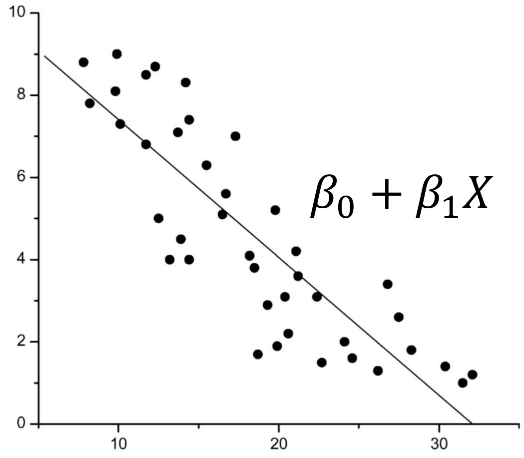
Propose some models

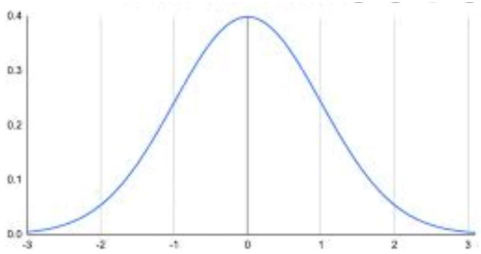
$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Choose the best model

Prediction for given input

Optimize input





Make a decision

- Larger x implies smaller y.

- But when x increases by a certain quantity, what's the reduction in y?

- Regression analysis: knowing $\beta_0 + \beta_1 X$, you can answer.

Collect past data

Propose some models

$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

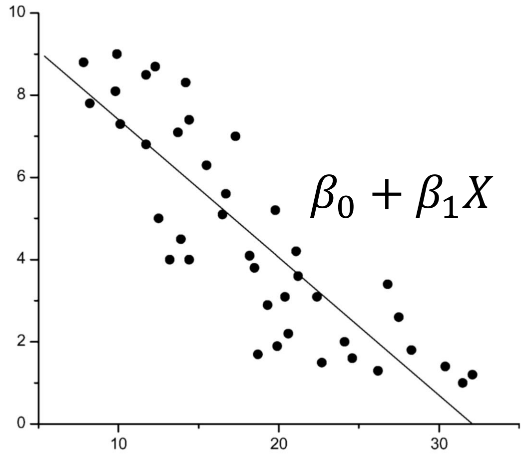
Choose the best model

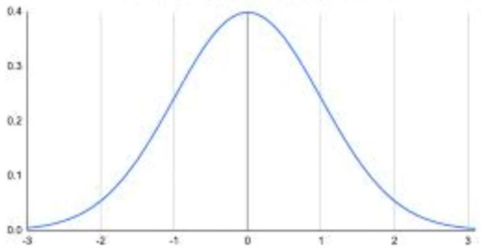
MLE: choose best $\beta_0, \beta_1, \sigma^2$

Prediction for given input

Optimize input

Make a decision





- PDF for normal

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$
- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2} \frac{\Sigma_i (Y_i - \boxed{\beta_1 X_i - \beta_0})^2}{\sigma^2}\right]$$

- PDF for normal
$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

- PDF for normal

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

- Samples: $(X_1, Y_1), \dots, (X_N, Y_N)$

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left[-\frac{1}{2}\frac{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\boxed{\Sigma_i (Y_i - \beta_1 X_i - \beta_0)^2}$$

From High School:
Least square regression
最小二乘法

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) = 0$$

$$\Sigma_i (Y_i - \beta_1 X_i - \beta_0) X_i = 0$$

$$\Sigma_i(Y_i - \beta_1 X_i - \beta_0)X_i = 0 \text{ AND } \Sigma_i(Y_i - \beta_1 X_i - \beta_0) = 0$$

Eliminate β_0 first:

$$\Sigma_i(Y_i - \beta_1 X_i - \beta_0) = 0 \quad \text{->} \quad \beta_0 = \frac{1}{N}\Sigma_i(Y_i - \beta_1 X_i) = \bar{Y} - \beta_1 \bar{X}$$

MLE:

$$\begin{cases} \widehat{\beta_1} = \dfrac{\Sigma_i(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma_i(X_i - \bar{X})^2} \\ \widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X} \end{cases}$$

When simple regression is invalid?

- The model we propose is not correct.

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2) \text{ or } Y - \beta_0 - \beta_1 X \sim N(0, \sigma^2)$$

Linear regression assumes that…

1. The relationship between X and Y is **linear**

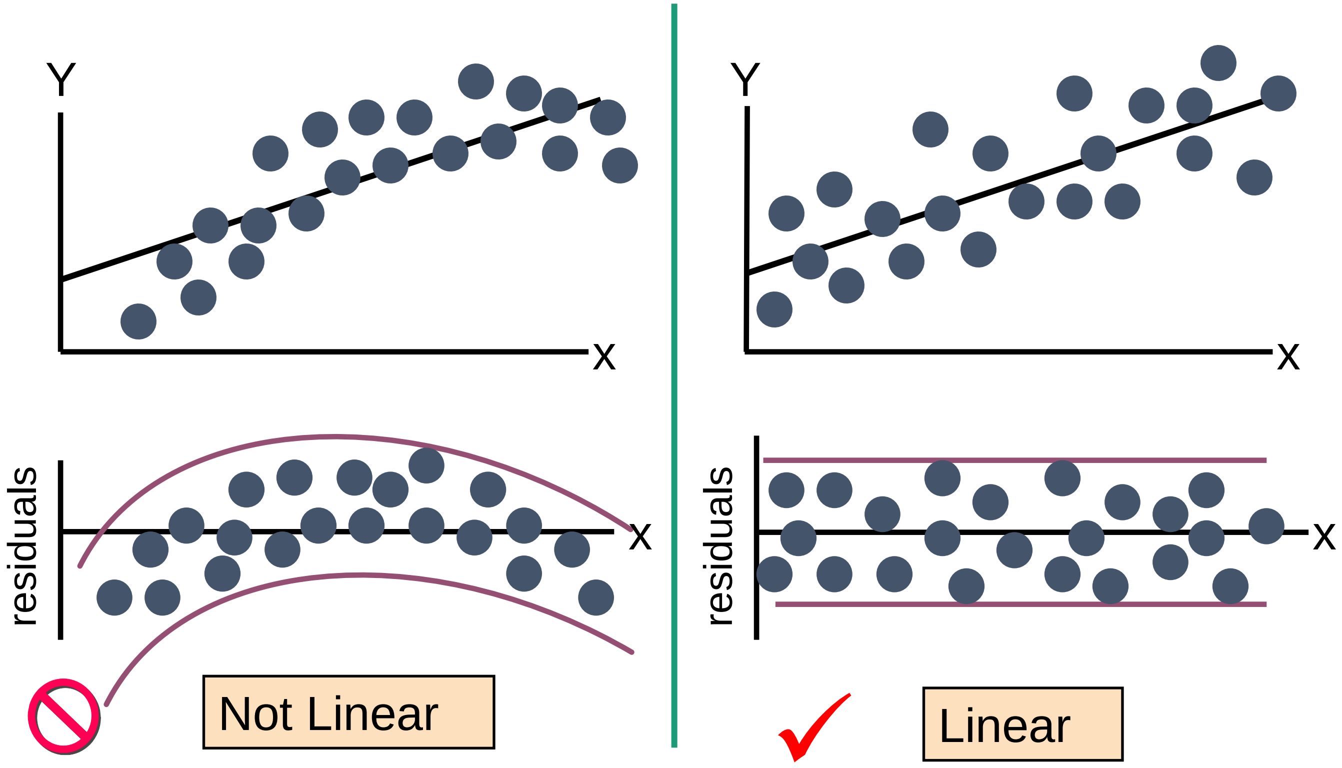
2. The variance of $Y - \beta_0 - \beta_1 X$ at every value of X is the **same** (homogeneity of variances)

Residual Analysis: check assumptions

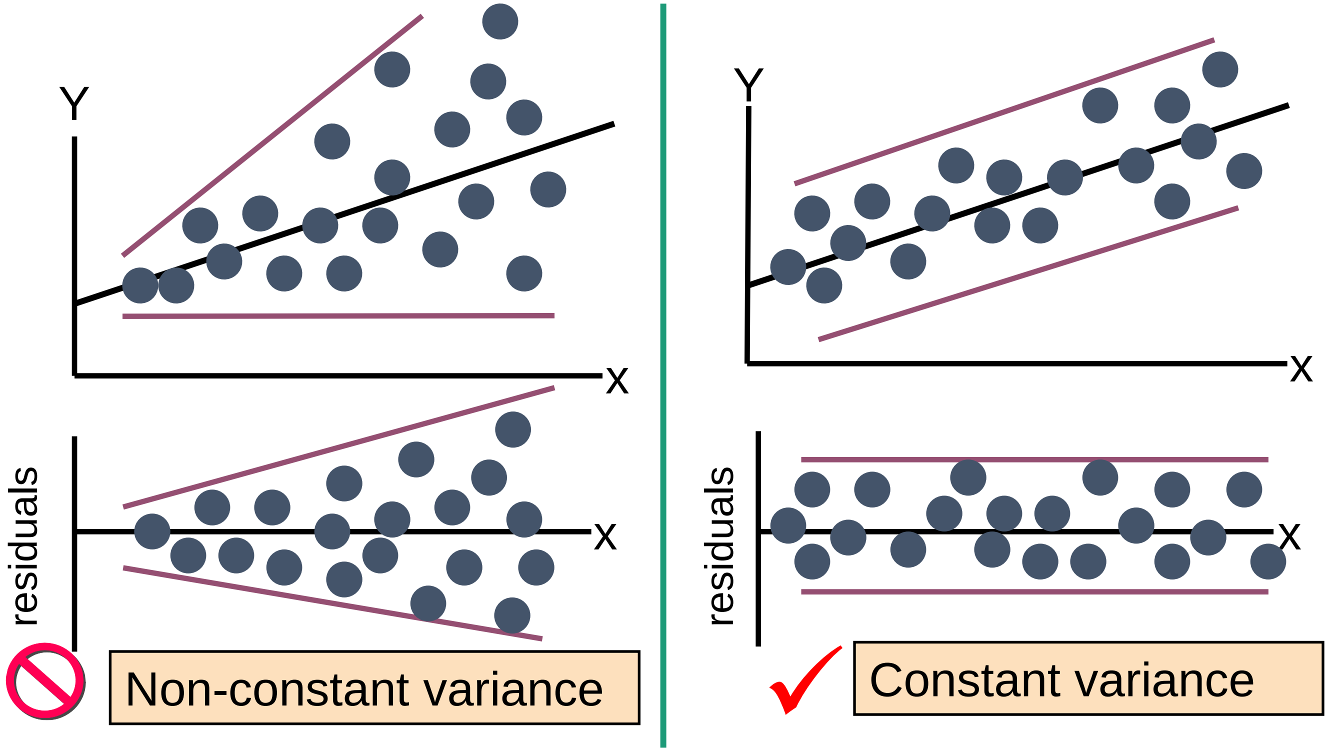
Residual: $e_i := Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$

- Check the assumptions by examining the residuals
  - Examine for linearity assumption:
    - e_i *does not depend on* X_i
  - Evaluate constant-variance assumption:
    - variance of e_i *does not depend on* X_i
- Graphical Analysis of Residuals: Can plot residuals vs. X

Residual Analysis for Linearity



Residual Analysis for constant-variance



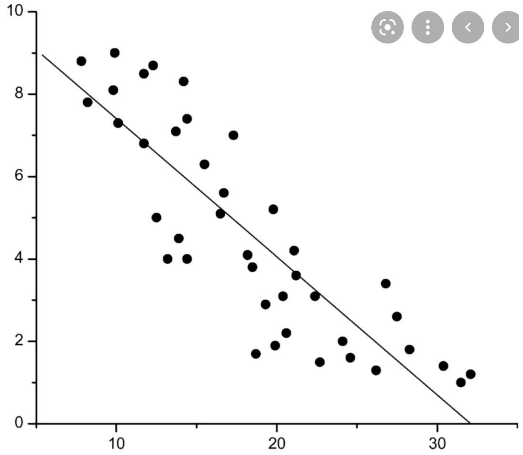
Summary of Regression Analysis:

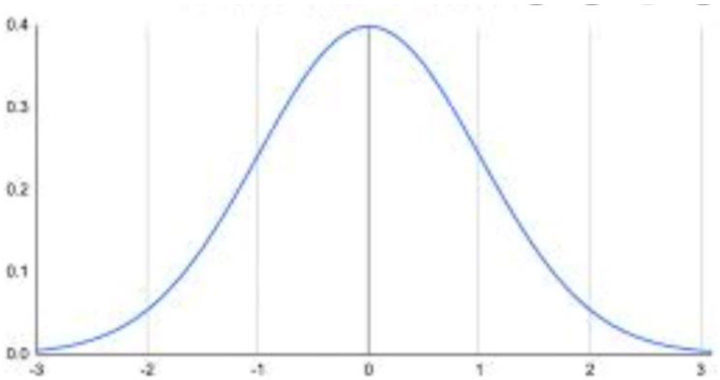
- Target: find a relationship between X and Y.
  - Proposed models: $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
  - Choose the best parameters: MLE.
  - Check whether the model is acceptable: Residual analysis.

1. Why we propose $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$?

From data, we observe that

- They are more likely to be linearly dependent with each other.
- Y is centralized at some value $\beta_0 + \beta_1 X$.





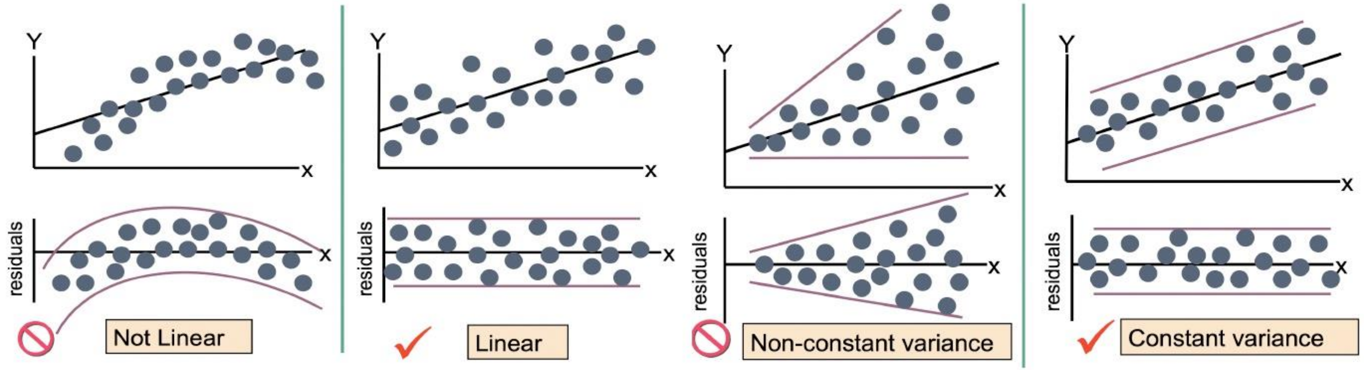
2. Why we use MLE to estimate parameters?

- Among the proposed models, choose the one with the **largest probability** that the samples are generated.

- Recall example
  - Drug 1: p = 0.8
  - Drug 2: p = 0.9
  - Samples: M = 81 successes from N=100 experiments
  - Which drug do you think is experimented?

3. Whether the model is acceptable?

- The model assumption may be incorrect.
  - The relationship between X and Y is linear
  - The variance of $Y - \beta_0 - \beta_1 X$ at every value of X is the same (homogeneity of variances)
- Residual analysis



(a) (b)