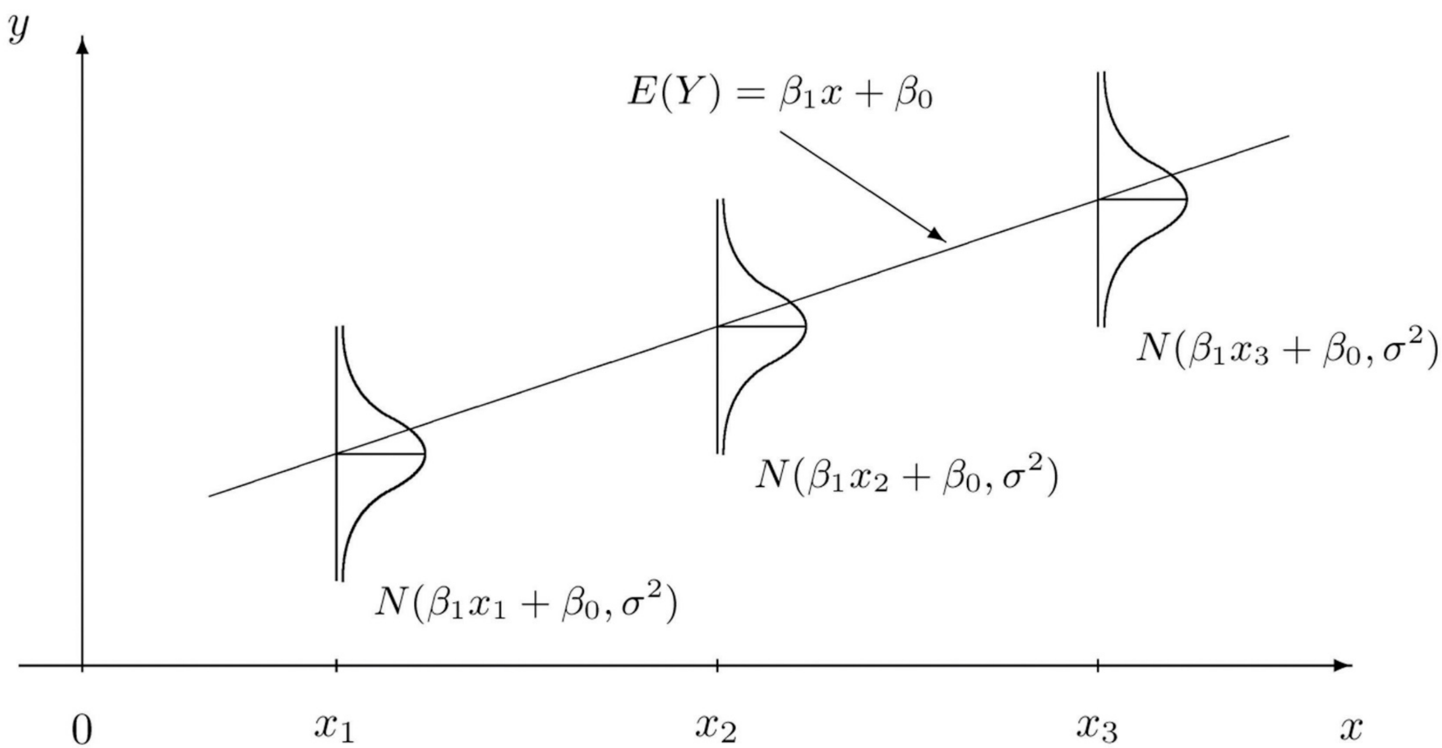
LINEAR REGRESSION/LOGISTIC REGRESSION

Week05

Linear Regression

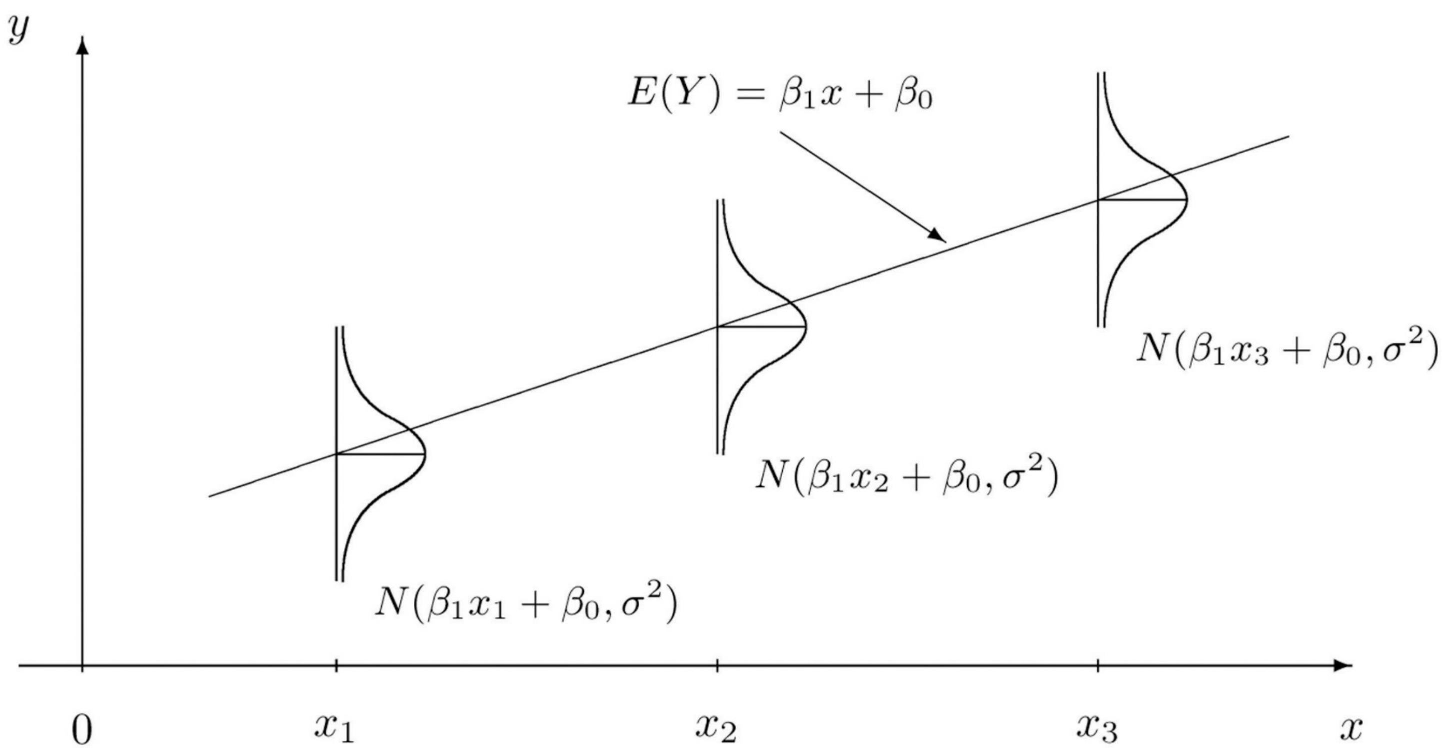
Check Appropriateness of Linear Regression

- Do you remember main assumptions of linear regression?



Main Assumption of Linear Regression

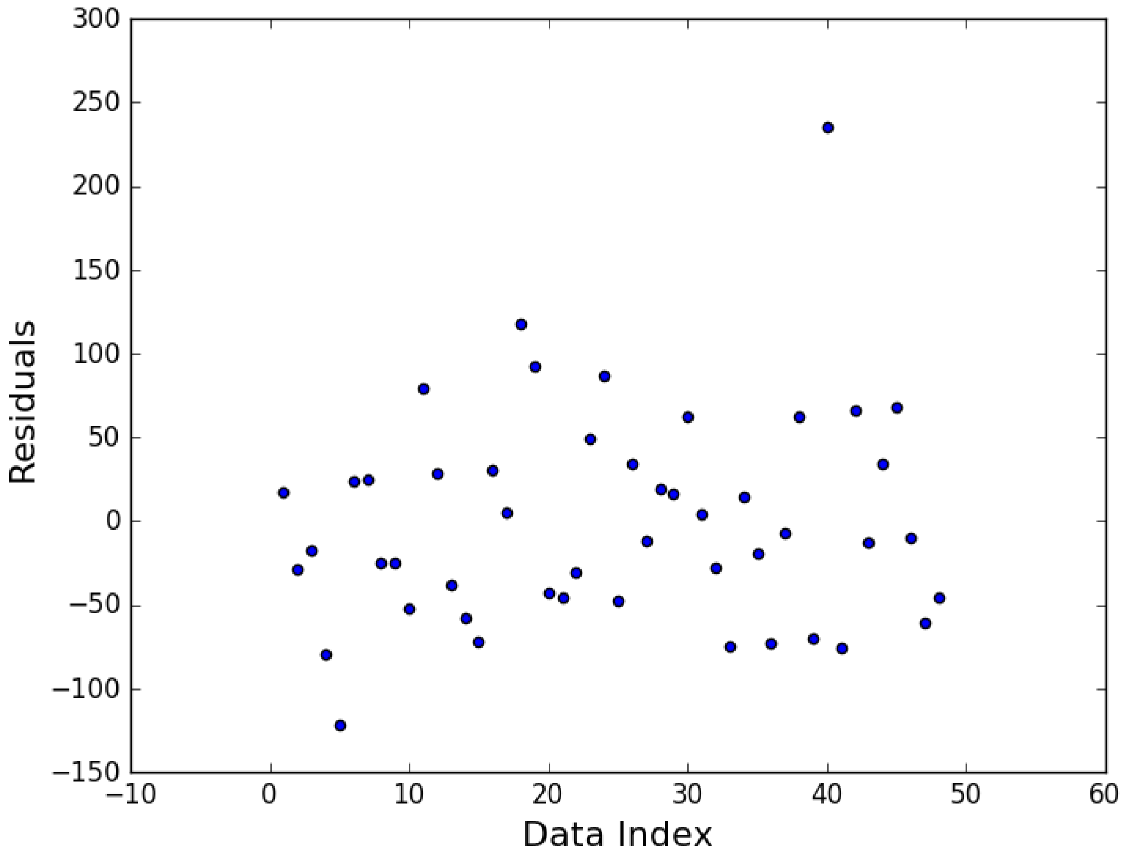
- Linear regression analysis makes several key assumptions
  - Linear relationship
  - Homoscedasticity
  - Normality
  - No or little multicollinearity



Check Appropriateness of Linear Regression

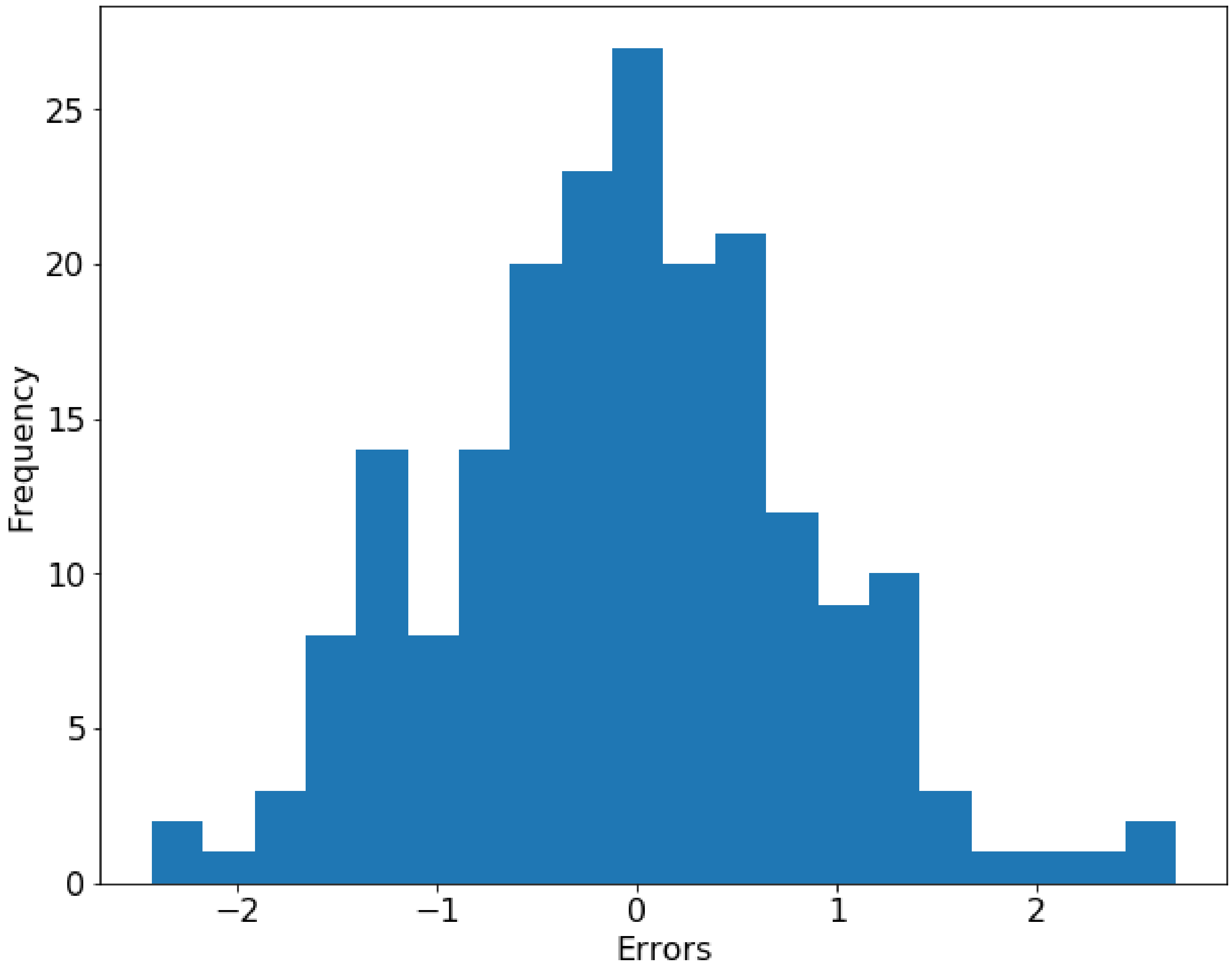
- Normality
  - Errors should follow normal distribution
  - Calculate errors (residuals) and check normality

$$e_i = y_i - \hat{y}_i$$



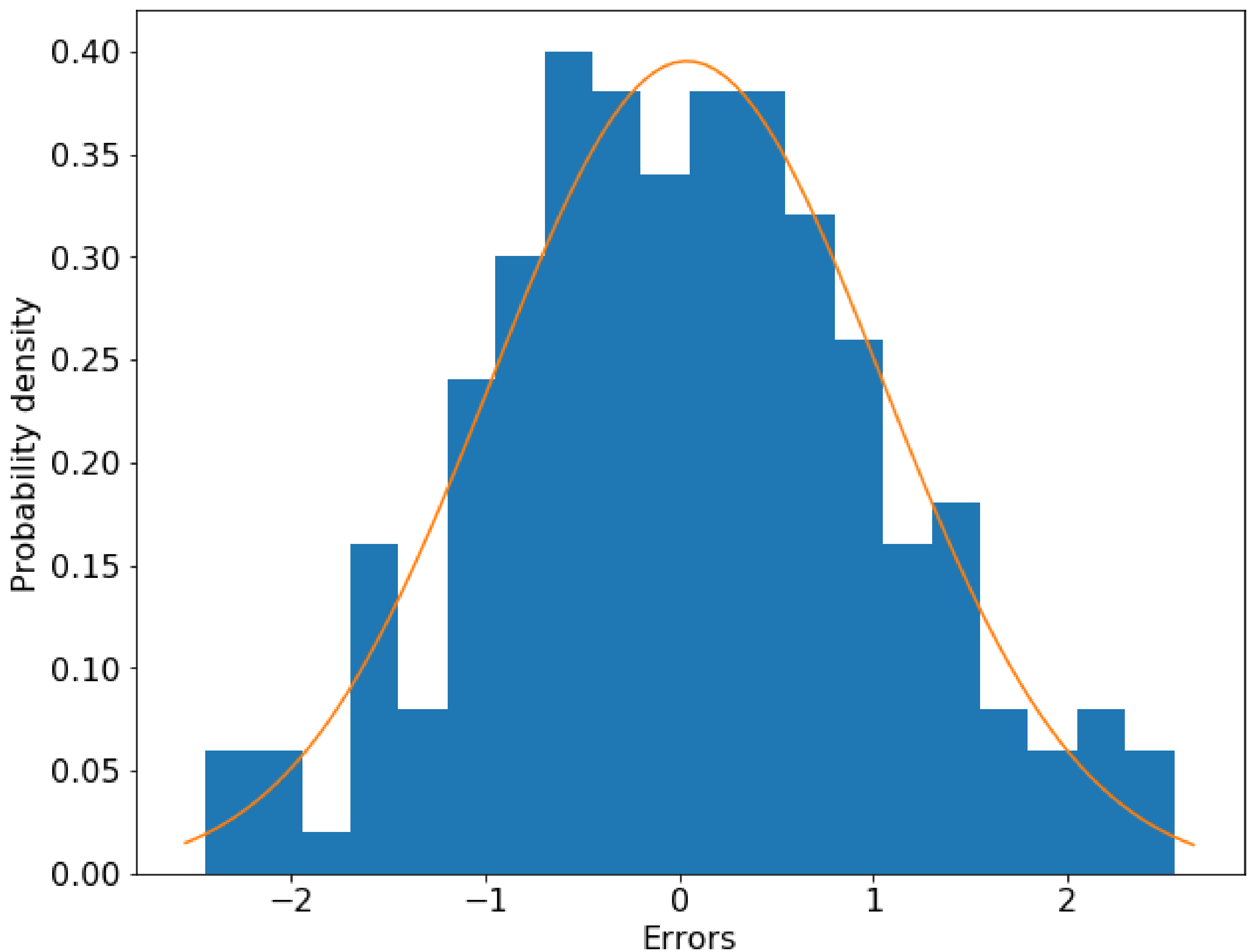
Check Appropriateness of Linear Regression

- Histogram



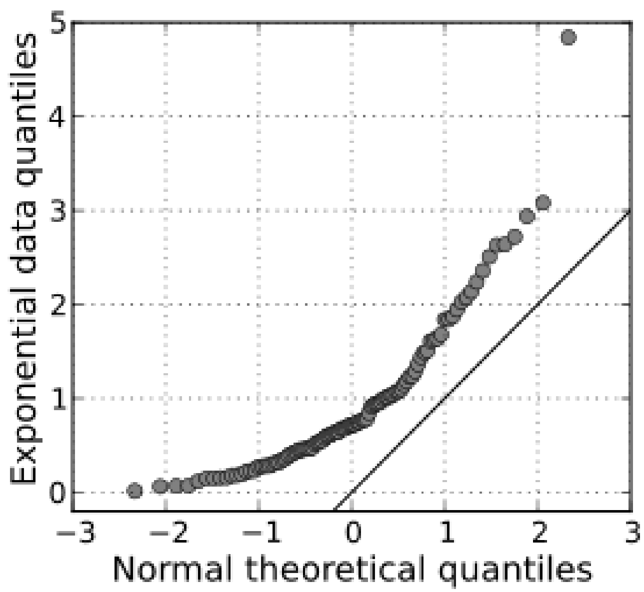
Check Appropriateness of Linear Regression

- Histogram

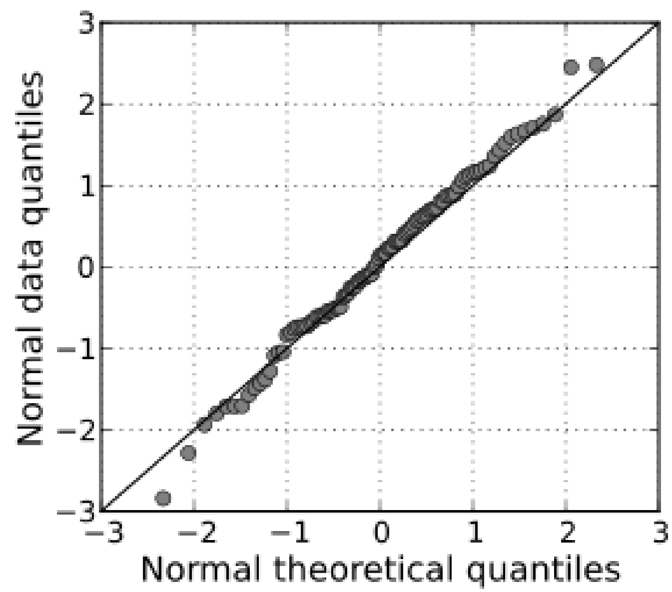


Check Appropriateness of Linear Regression

- Q-Q plot
  - A probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other
  - Quantiles are cutpoints dividing a set of observations into equal sized groups
    - q-Quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes
    - Median is 2-quartile, 0.5 quantile and 50 percentile

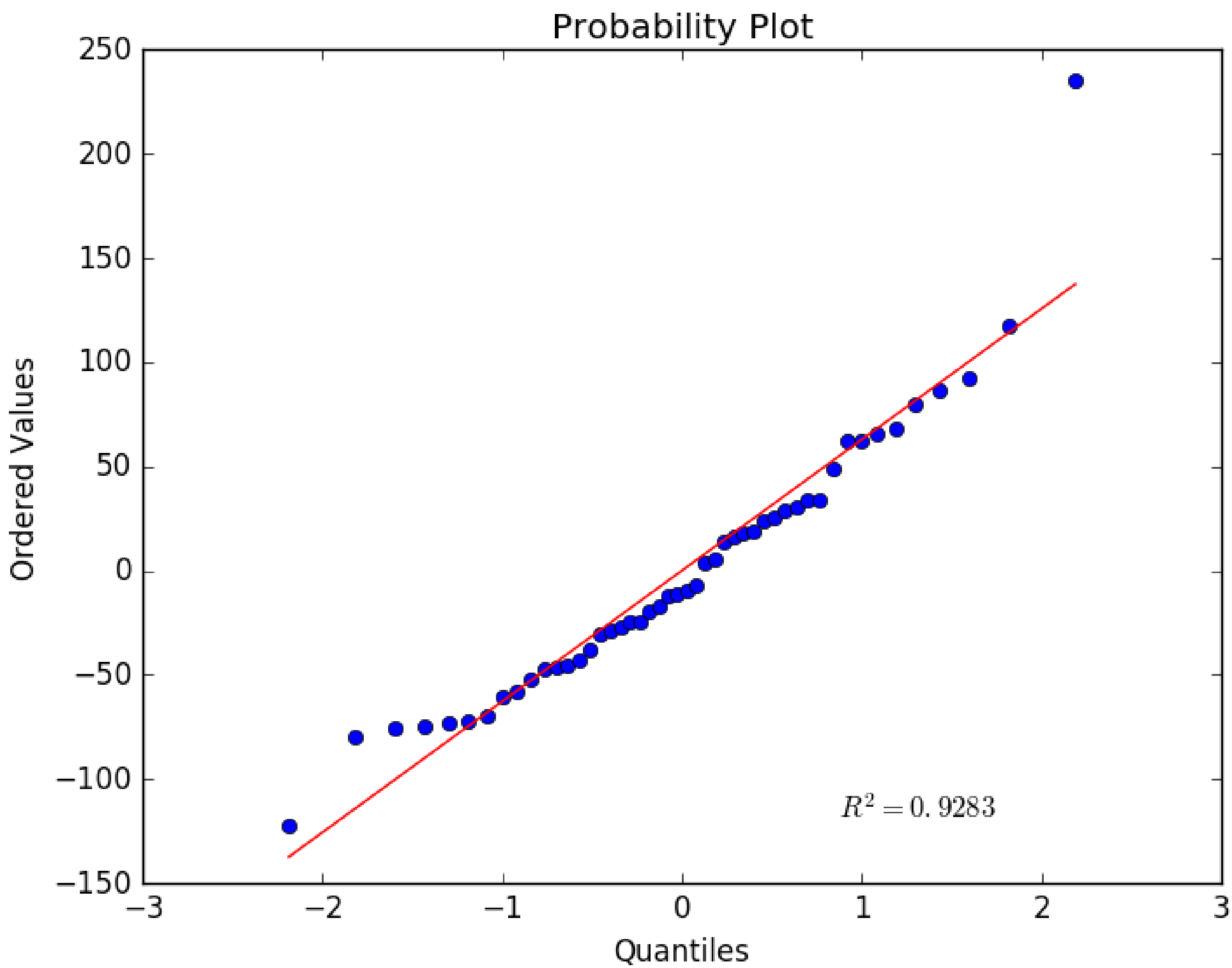


$X \sim Exp(1)$



$X \sim N(0,1)$

Q-Q Plot


Probability Plot
250
200
150
100
50
0
−50
−100
−150
Ordered Values
−3
−2
−1
0
1
2
3
Quantiles
$R^2 = 0.9283$

Jarque-Bera test

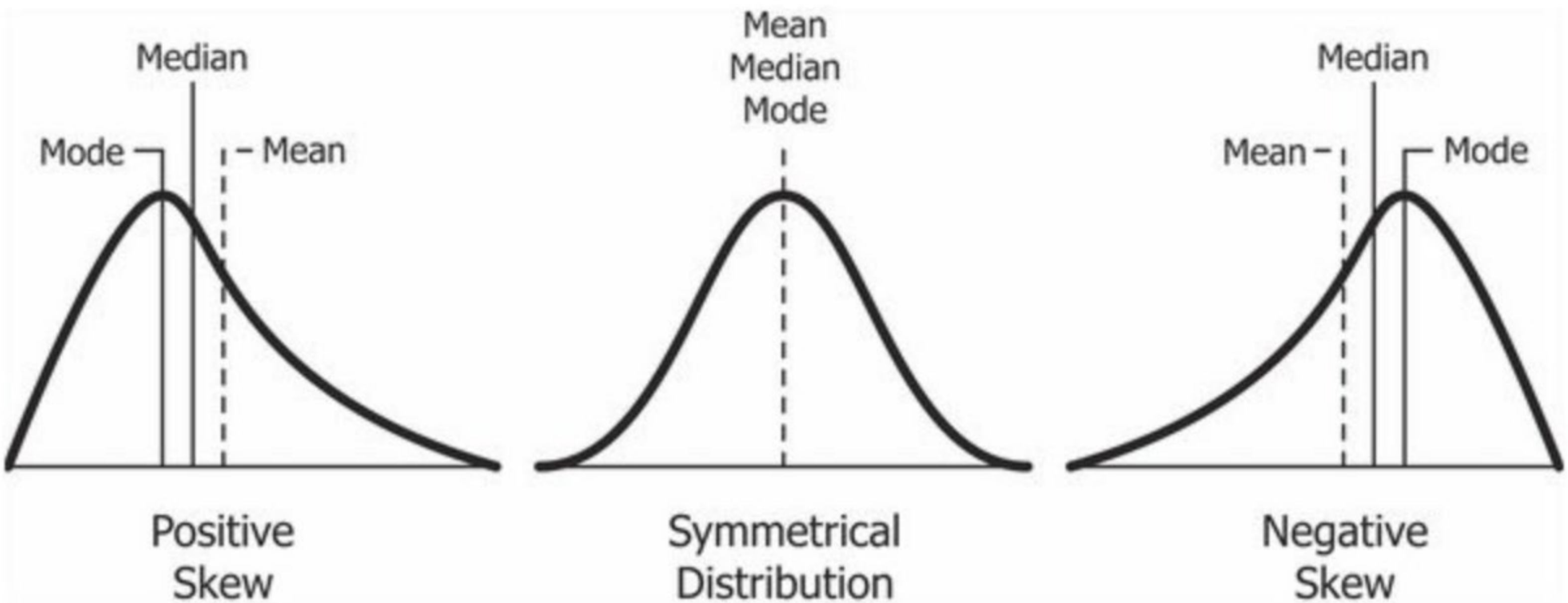
- Jarque-Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution
  - Test statistic

$$JB = \frac{n-k}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right)$$

    - $S = \frac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}$: sample skewness
    - $C = \frac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{4}{2}}}$: sample kurtosis
    - k: the number of input variables
  - If the data comes from a normal distribution, JB statistic asymptotically has a chi-squared distribution with two degrees of freedom

$$H_0\colon S = C - 3 = 0$$

※ Skewness

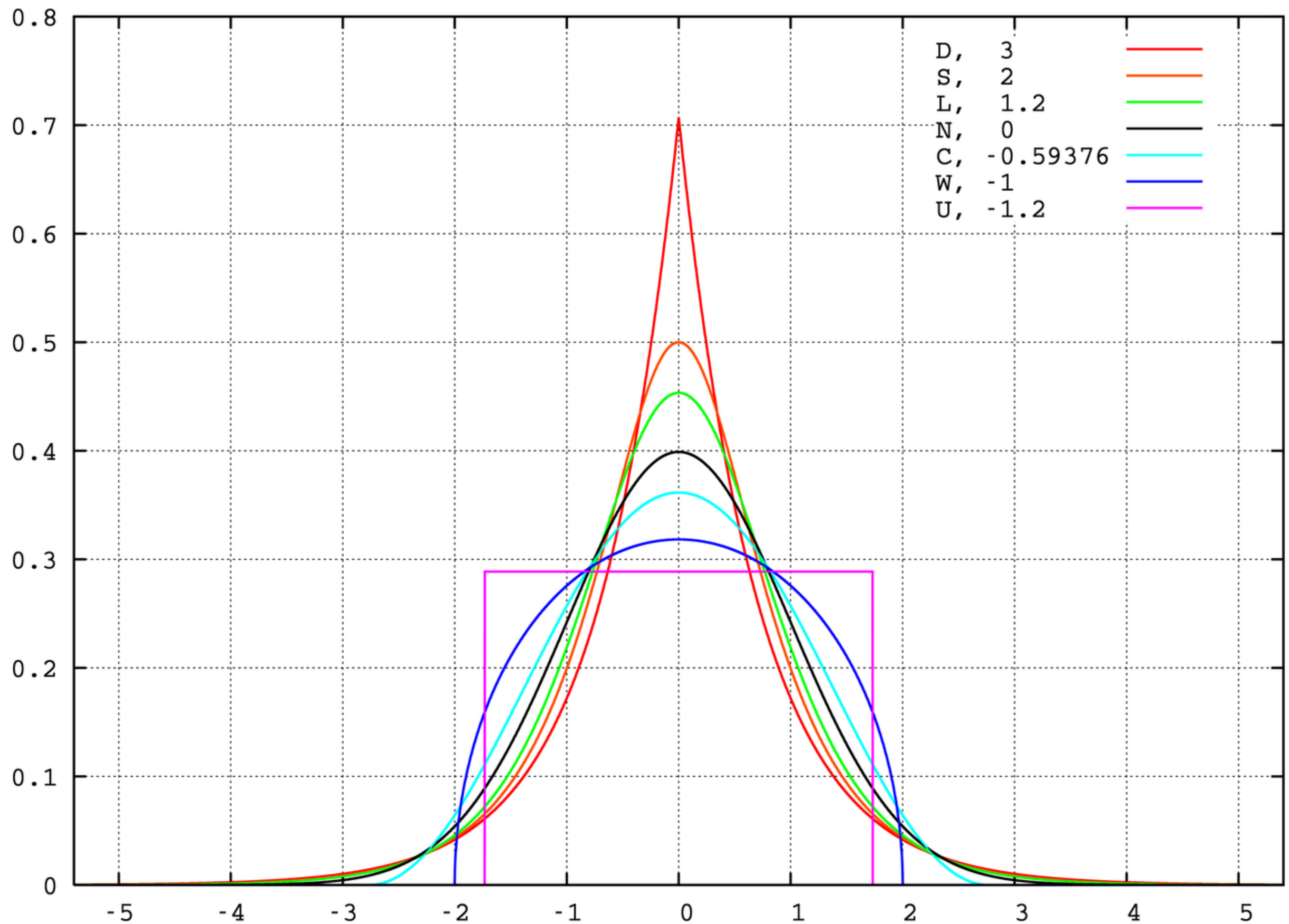


right-skewed, right-tailed,
or skewed to the right,

left-skewed, left-tailed,
or *skewed to the left*

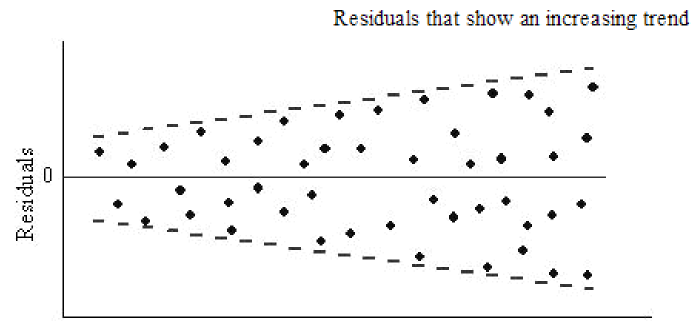
※ Kurtosis

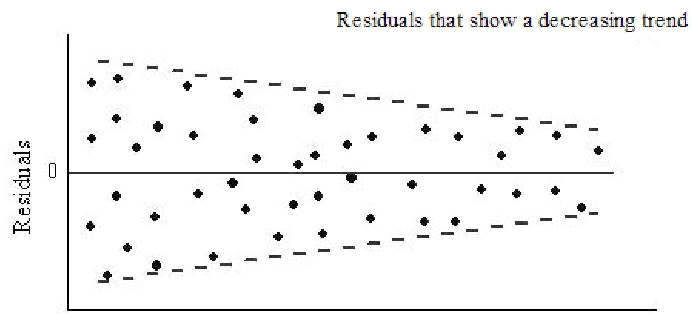
- Probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis

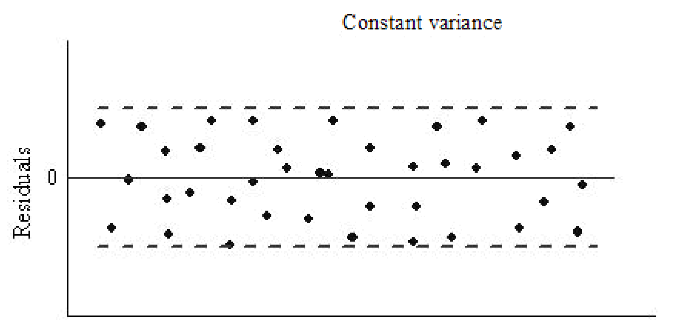


Check Appropriateness of Linear Regression

- Homoscedasticity↔Heteoscedasticity
  - Check whether all random variables in the sequence or vector have the same finite variance







Check Appropriateness of Linear Regression

- Breusch-Pagan test
  - Test of the hypothesis that the independent variables have no explanatory power on the squared errors
- Procedure of Breusch-Pagan test

① Apply linear regression in the model and compute the regression residuals

$$y = X\beta + \epsilon$$

② Perform the auxiliary regression

$$e^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_p x_p + \eta$$

③ Apply F-test on auxiliary regression

$$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$$
$$H_a: all\ \gamma_i\ is\ not\ 0$$

Alternative test statistics

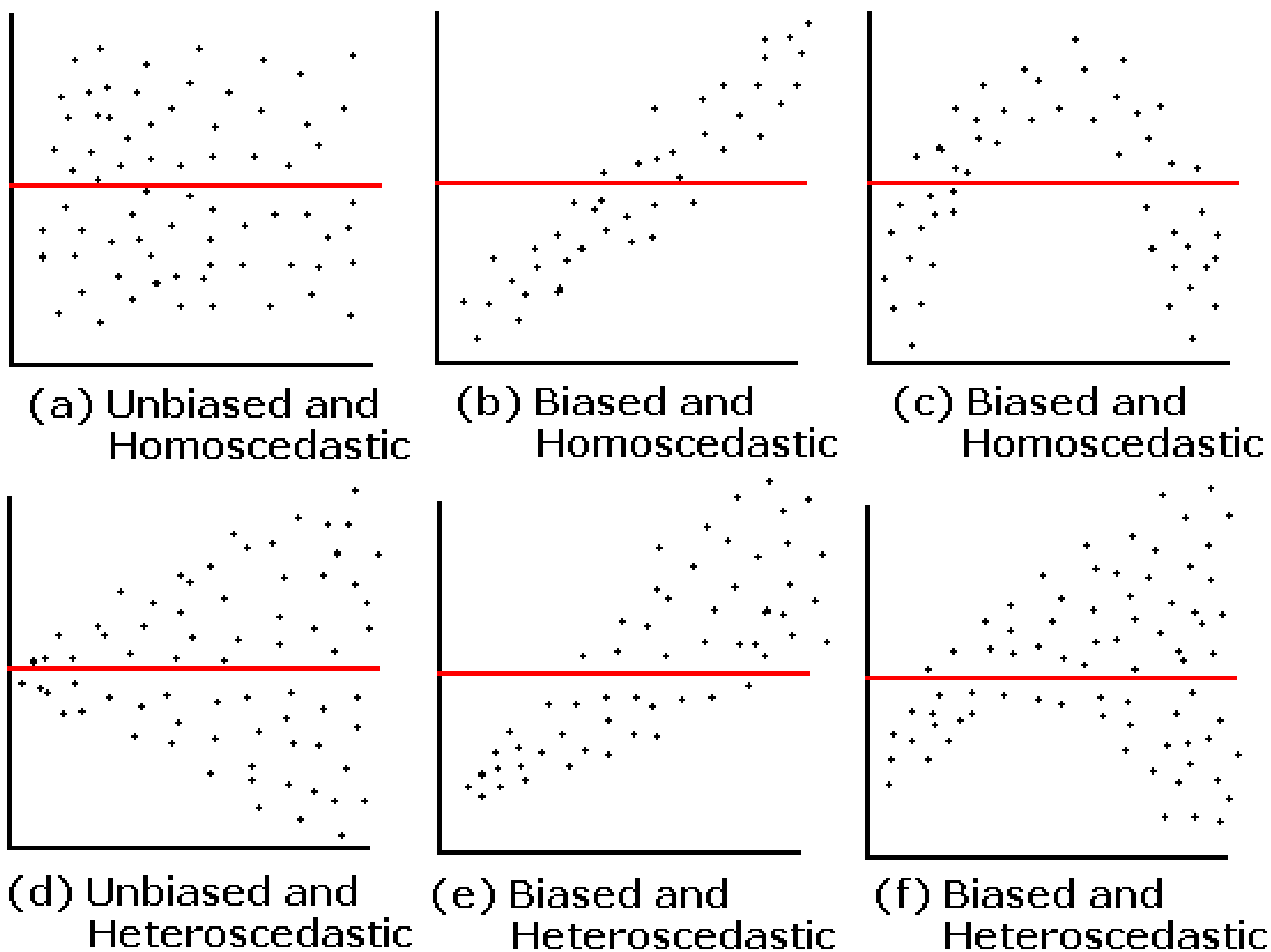
③ Use Lagrange multiplier statistic

$$LM = nR^2$$

- The test statistic is asymptotically distributed as χ_p^2 under the null hypothesis of homoskedasticity

Check Appropriateness of Linear Regression

- Residual plot



Interpretation & Prediction

- If the fitted regression model is appropriate and significant you can use the model for future use
  - Linear regression models have strength in interpretation
    - Each coefficient explains relationship between each explanatory variable and the target variable
  - Based on the fitted model, predict the target on test samples

Overall Process for Linear Regression

Validity of Linear Regression Model

Before learning

- Linearity
- Multicolinearity

After learning

- Normality
- Homoscedasticity
- Significance of model
- Goodness-of-fit

Model Utilization

- Figure out significant input variables
- Figure out relations between input variables and target variable → Interpretation
- Predict target with respect to new observations

Feature Scaling

- Predict consumption of petrol
  - Linear model by least square method

$$y = -34.8x_1 - 0.0666x_2 - 0.002x_3 + 1336x_4 + 377.3$$

Petrol Tax($)	Average Income ($)	Paved Highways (miles)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	1976	0.525	541
9	4092	1250	0.572	524
9	3865	1586	0.58	561
7.5	4870	2351	0.529	414
...	...	...	...	...

How about changing scale of variable?

Feature Scaling

- Change unit of paved highways from mile to cm

1mile=160934.4cm

Petrol Tax($)	Average Income ($)	Paved Highways (cm)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	31683974.4	0.525	541
9	4092	20043000	0.572	524
9	3865	25430558.4	0.58	561
7.5	4870	37696874.4	0.529	414
…	…	…	…	…

- Linear regression on new data

$$y = -34.8x_1 - 0.0666x_2 - 1.5 \times 10^{-7}x_3 + 1336x_4 + 377.3$$

Feature Scaling

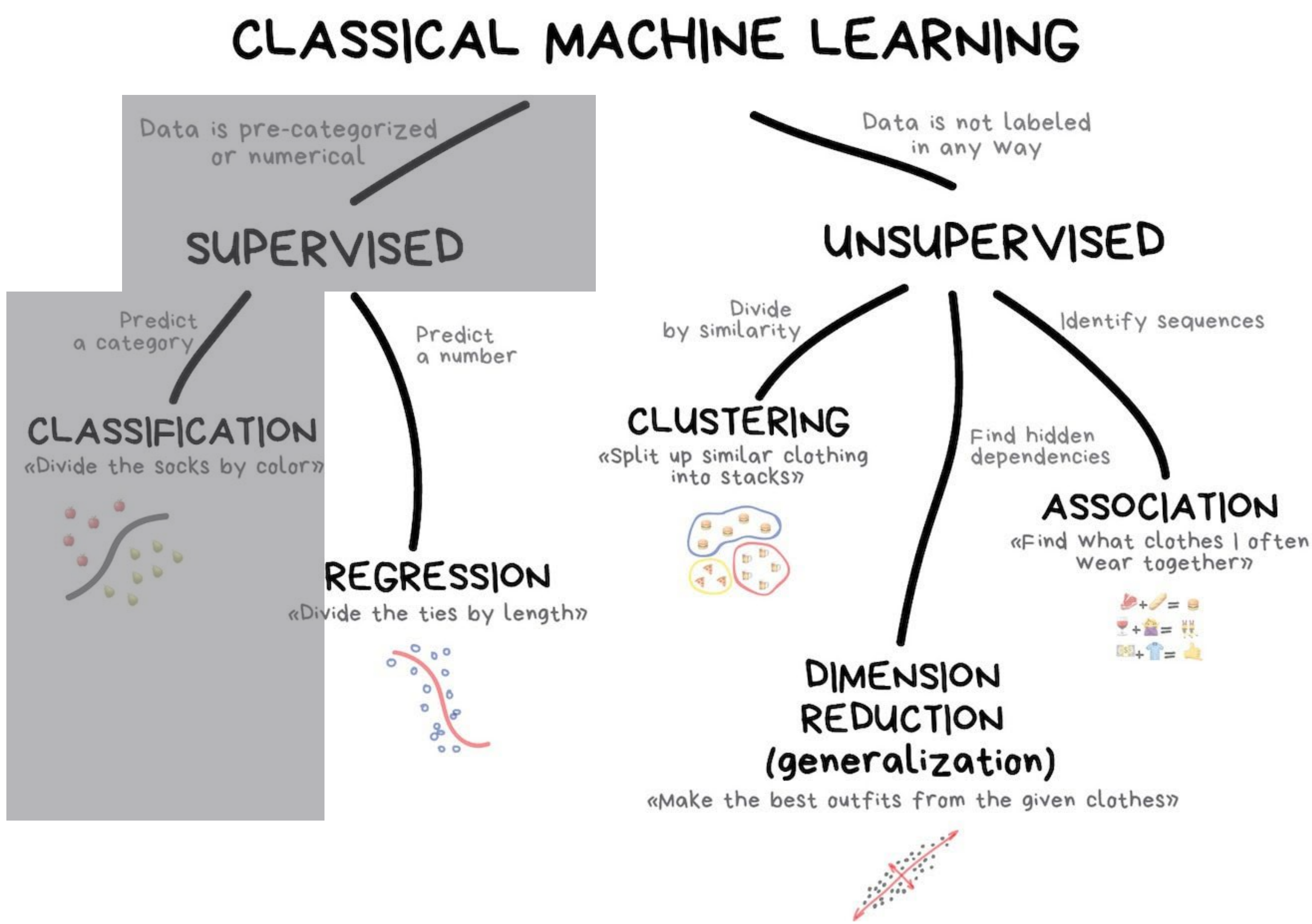
- Scale change only affects on the changed variable
  - Coefficients of other variables are not changed
  - If variable x is replaced with ax, coefficient of x, β by linear regression is changed to β/a
  - If scale of certain variable is too large, coefficient of the variable might be too small
    → It is better to change scale

Variable Transformation

- Linear regression algorithm is quite simple, but it can be extended using transformation
  - $x \to x^2$
  - $x \to \log x$
  - $x \to \sqrt{x}$

Logistic Regression

Topics Covered in This Class



Supervised: Classification

- Classification problem
  - Output is categorical variable
    - Spam/Non Spam
    - Male/Female
    - Long/Medium/Short
    - O/X

- Binary classification problem
  - The number of categories is 2
  - Generally, these two categories are denoted as 0 and 1
    - 0 and 1 are not integer in this case

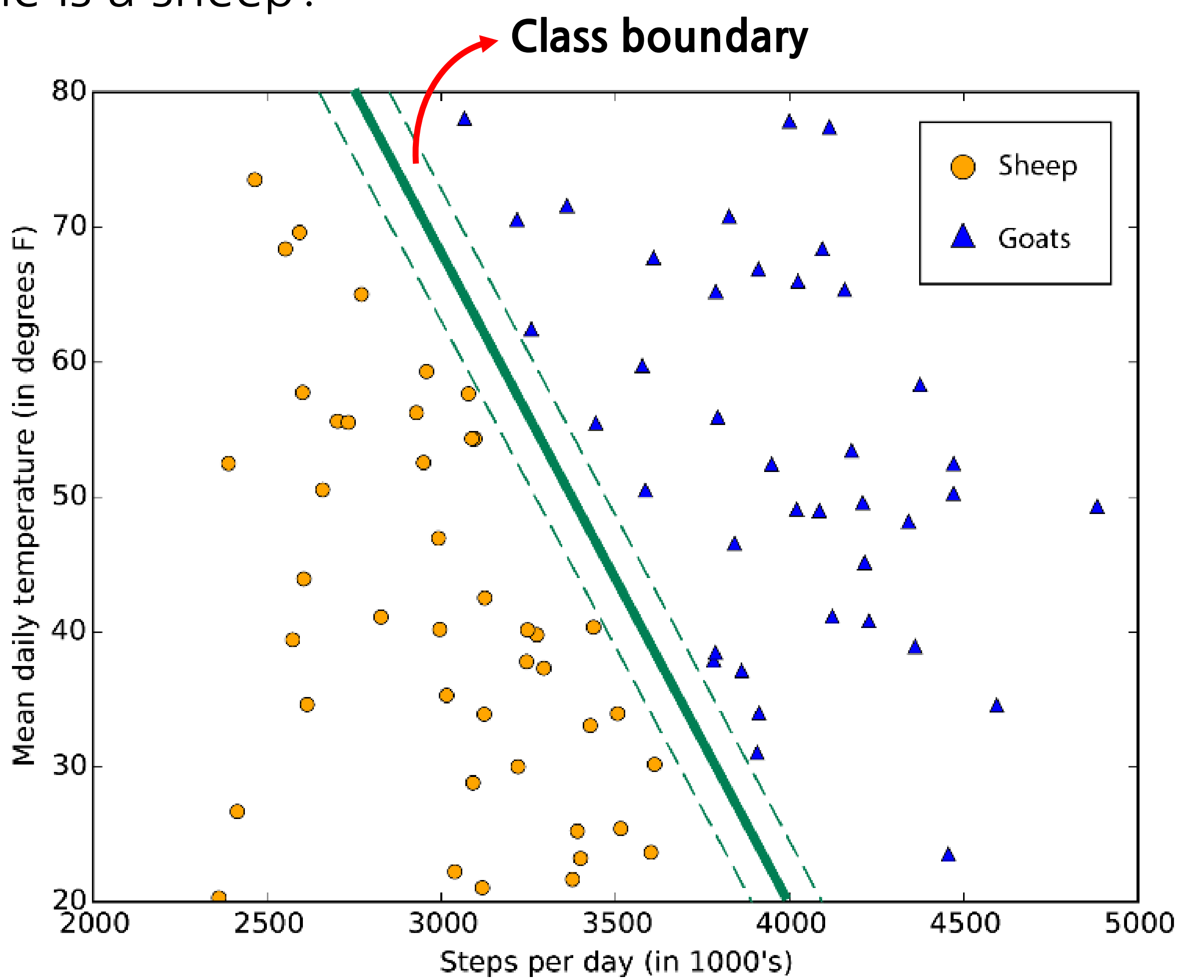
$$y \in \{0,1\}$$

- Multi-class classification problem
  - More than two classes

$$y \in \{1,2,\ldots,C\}, \qquad C > 2$$

Supervised: Classification

- Which one is a sheep?



Types of Classifiers

- A classifier is a function that assigns to a sample, $\mathbf{x}$ a class label $\hat{y}$

$$\hat{y} = f(\mathbf{x})$$

- A probabilistic classifier obtains conditional distributions $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \boldsymbol{X}$, they assign probabilities to all $y \in Y$
  - Hard classification

$$\hat{y} = \arg\max_{y} \Pr(Y = y|\,\mathbf{x})$$

The Decision Boundary of Classifiers

- Decision boundary

$$y = f(X), \qquad y \in \{1,2,\ldots,C\}$$

Logistic Regression

- Logistic regression
  - Regression model where the dependent variable is categorical
  - The probabilities describing the possible outcomes is modeled as explanatory variables

$$f(x) = P(Y|X)$$

  - Logistic regression is a linear classification algorithm

$$f\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = f(\boldsymbol{\beta} \cdot \mathbf{x})$$

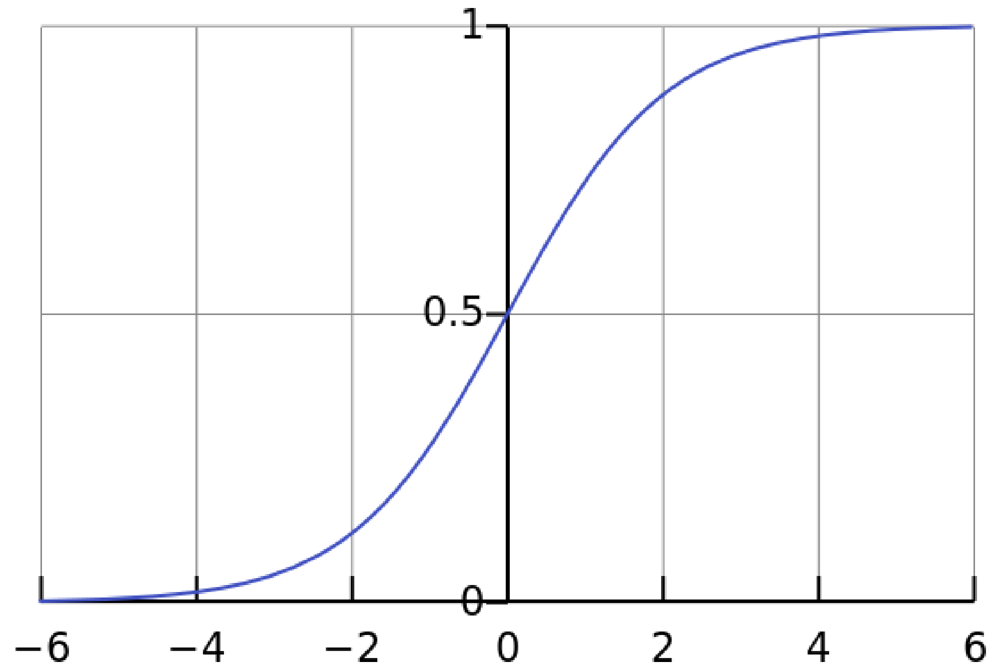
    - $f(x)\ should\ be\ 0 \leq f(x) \leq 1$

How to confine outcome of $f(x)$ within [0,1]?

Logistic Regression: Logistic function

- Logistic function is the function that can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one

$$\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$$



- In logistic regression, t is determined by explanatory variables

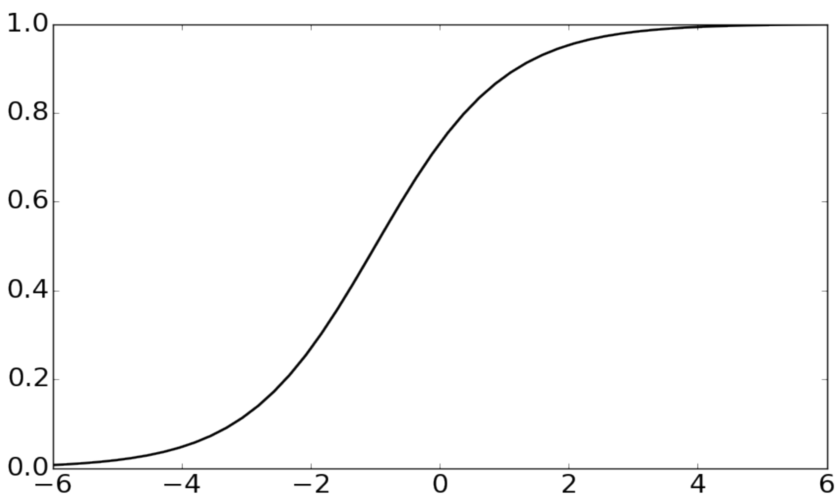
Logistic Regression

- t is determined by linear combination of explanatory variables

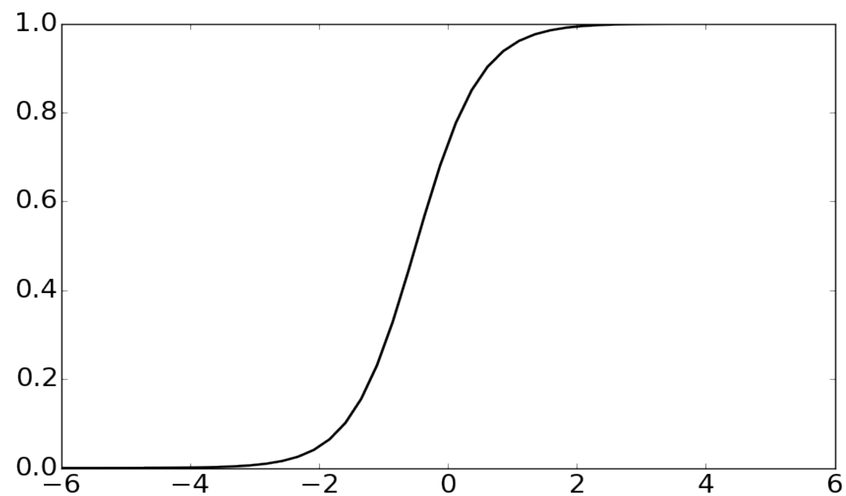
$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$f(x) = P(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p}}$$

$t = 1 + x$

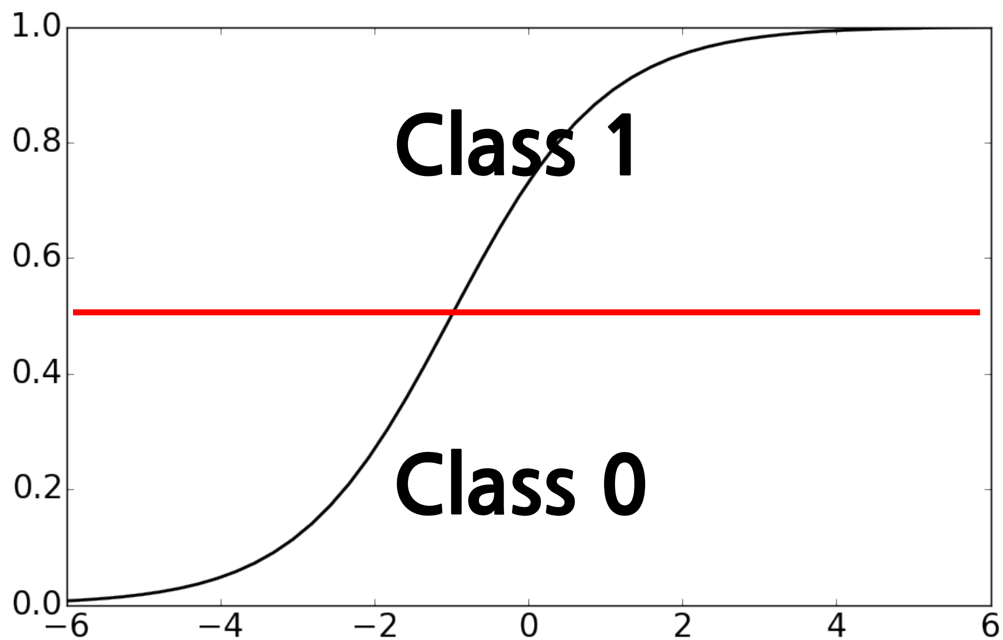


$t = 1 + 2x$



Logistic Regression

- Determine class
  - Set class boundary
    - Without any prior knowledge about class, set 0.5



    - If you have some knowledge about class distribution, class boundary can be determined based on the knowledge