データマイニング
Data Mining

4: 回帰① Regression 1

土居 裕和 Hirokazu Doi
長岡技術科学大学 Nagaoka University of Technology

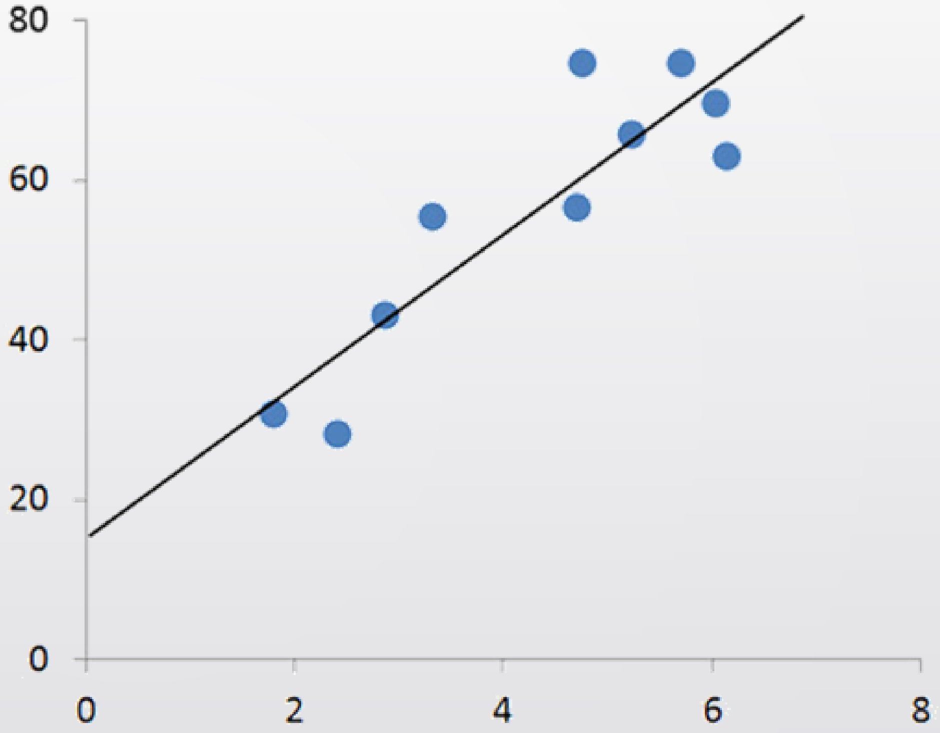
回帰 Regression

$$(x_i,\ y_i)$$

x_i: 家族の人数
Number of
Family Member

y_i: 購入数
Number of purchased
Items

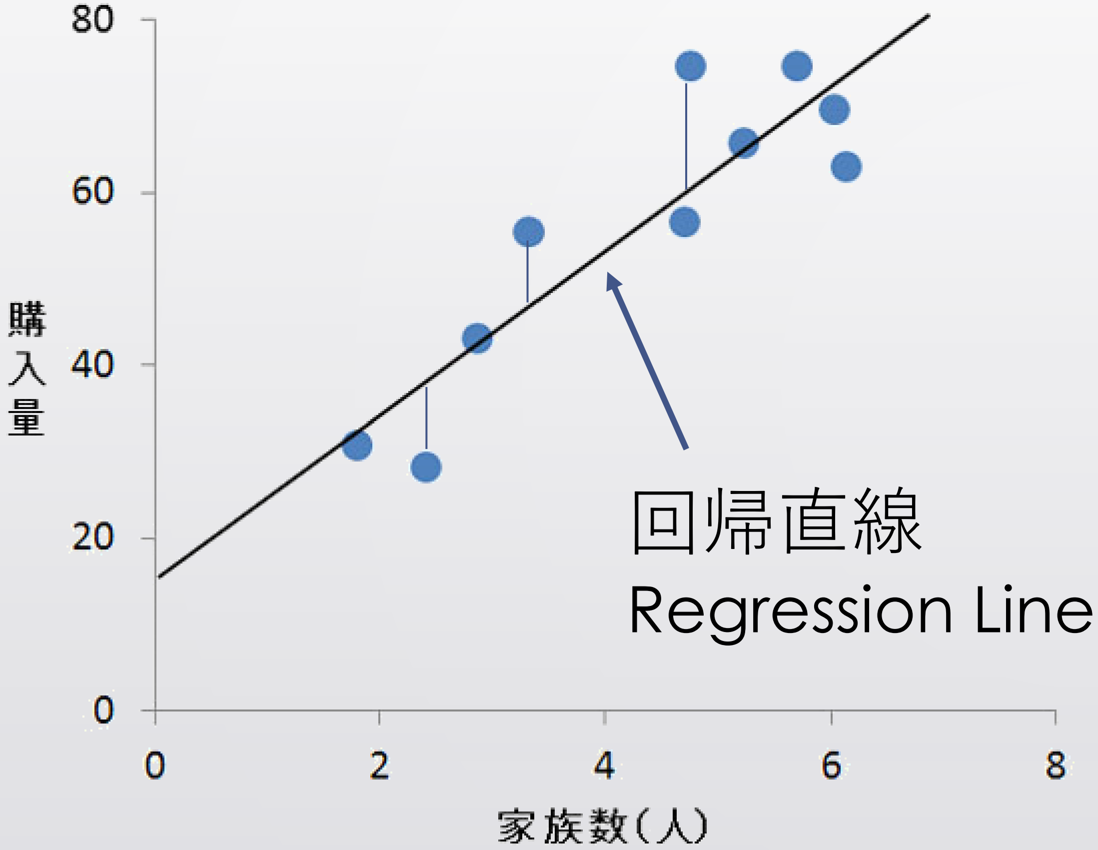
y_i



x_i

https://www.intage.co.jp/glossary/051/

線型回帰直線 Linear Regression Line



https://www.intage.co.jp/glossary/051/

$$y = kx + y_0$$

傾き
Slope

切片
Intercept

カリフォルニア住宅データセット
California Housing Dataset

```
California Housing dataset
--------------------------

**Data Set Characteristics:**

    :Number of Instances: 20640

    :Number of Attributes: 8 numeric, predictive attributes and the target

    :Attribute Information:
        - MedInc        median income in block
        - HouseAge      median house age in block
        - AveRooms      average number of rooms
        - AveBedrms     average number of bedrooms
        - Population    block population
        - AveOccup      average house occupancy
        - Latitude      house block latitude
        - Longitude     house block longitude

    :Missing Attribute Values: None

This dataset was obtained from the StatLib repository.
http://lib.stat.cmu.edu/datasets/

The target variable is the median house value for California districts.
```

複数の情報に基づいて、住宅価格を予測する

Predict housing price based on multiple information

予測変数 Predictors

ターゲット（目的）変数
Target Variable

重回帰分析 Multiple Linear Regression

複数の変数の線型和によって、ターゲット変数を予測
Predicts target variable by linear combination of multiple variables

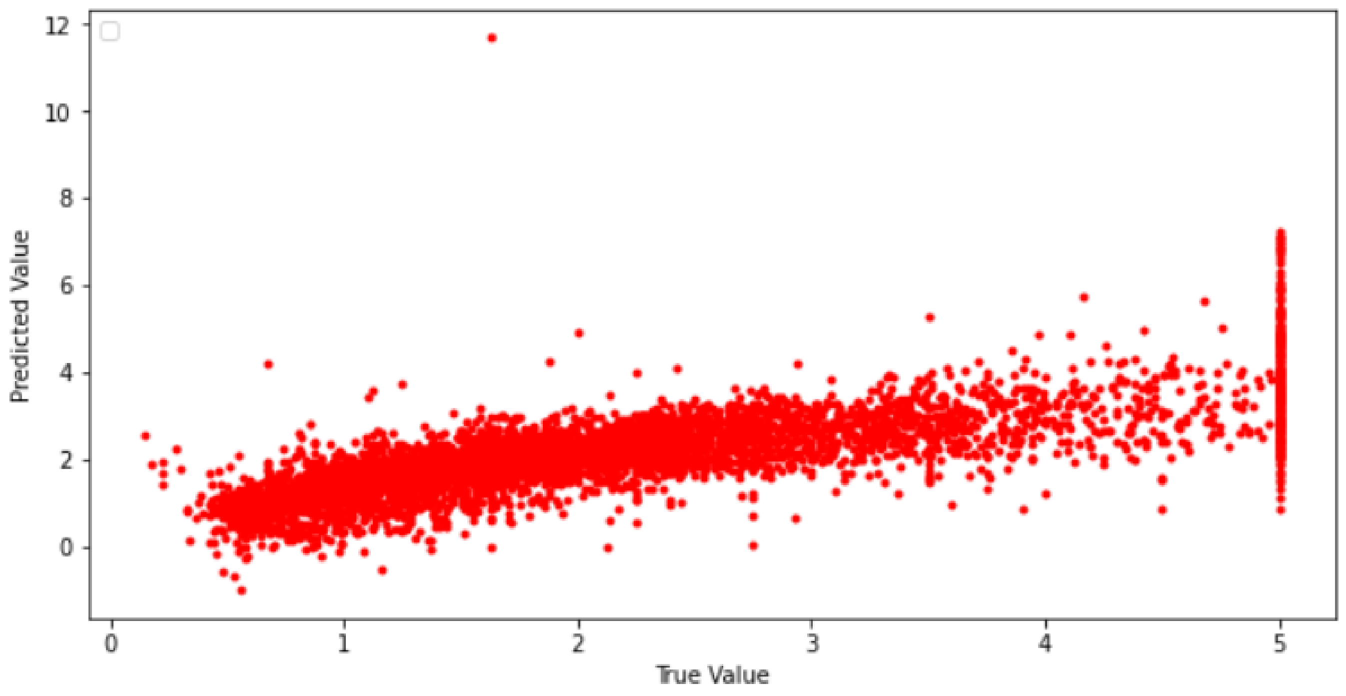
Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

https://www.i2tutorials.com/difference-between-simple-linear-regression-and-multi-linear-regression-and-polynomial-regression/

相関係数による性能評価

```
Out[33]: array([[1.        , 0.76907401],
                [0.76907401, 1.        ]])
```

予測値



正解値

重回帰分析 Multiple Linear Regression

```
California Housing dataset
--------------------------

**Data Set Characteristics:**

    :Number of Instances: 20640

    :Number of Attributes: 8 numeric, predictive attributes and the target

    :Attribute Information:
        - MedInc        median income in block
        - HouseAge      median house age in block
        - AveRooms      average number of rooms
        - AveBedrms     average number of bedrooms
        - Population    block population
        - AveOccup      average house occupancy
        - Latitude      house block latitude
        - Longitude     house block longitude

    :Missing Attribute Values: None

This dataset was obtained from the StatLib repository.
http://lib.stat.cmu.edu/datasets/

The target variable is the median house value for California districts.
```

$$House\ Value = \beta_1 MedInc + \beta_2 HouseAge + \cdots \beta_8 Longitude + \beta_0 + \varepsilon$$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M + \beta_0 + \varepsilon$$

誤差 Error

重回帰分析 Multiple Linear Regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M + \beta_0 + \varepsilon$$

$$= \sum_1^M \beta_i x_i + \beta_0 + \varepsilon \;\; = [\beta_1 \; \beta_2 \;\; \dots \;\; \beta_M] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} + \beta_0 + \varepsilon$$

$$= \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x} + \beta_0 + \varepsilon$$

誤差 Error

最小二乗法 Ordinary Least Squares (OLS) Method

データはM次元で N個の観測値（データ）がある

Data is M-dimensional and there are in total of N observations (Data points)

$$\boldsymbol{x_n} = \begin{bmatrix} x_{n,1} \; x_{n,2} \; \dots x_{n,M} \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \\ \vdots \\ \boldsymbol{x_N} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

それぞれのデータ $\boldsymbol{x_n}$ に対応するターゲットは $\boldsymbol{y_n}$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Target value for each data $\boldsymbol{x_n}$ is $\boldsymbol{y_n}$

$\boldsymbol{X}$と$\boldsymbol{y}$は中心化されている

Note $\boldsymbol{X}$ and $\boldsymbol{y}$ are centered

最小二乗法 Ordinary Least Squares (OLS) Method

$\boldsymbol{X}$から$\boldsymbol{y}$を精度よく予測できる重回帰モデルの $\boldsymbol{\beta}$を求める

Find $\boldsymbol{\beta}$ so that multiple regression model can predict $\boldsymbol{y}$ based on $\boldsymbol{X}$ with good precision

$$\boldsymbol{y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} = \boldsymbol{X\beta}$$

重回帰モデルによる$\boldsymbol{y}$ の推定値 Prediction of $\boldsymbol{y}$ based on multiple regression

観測値 $\boldsymbol{y}$と予測値 $\boldsymbol{y}'$との差を最小化する$\boldsymbol{\beta}$を求める

Find $\boldsymbol{\beta}$ that minimizes the difference between observed vector $\boldsymbol{y}$ and predicted vector $\boldsymbol{y}'$

最小二乗法 Ordinary Least Squares (OLS) Method

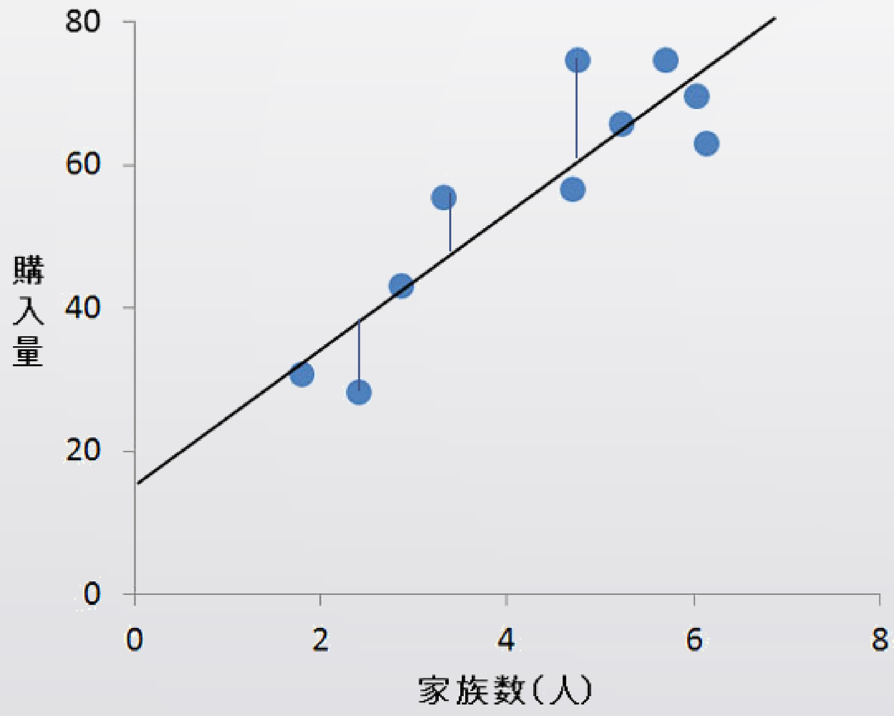
観測値 $\boldsymbol{y}$と予測値 $\boldsymbol{y}'$ との差を最小化する$\boldsymbol{\beta}$を求める

Find $\boldsymbol{\beta}$ that minimizes the difference between observed vector $\boldsymbol{y}$ and predicted vector $\boldsymbol{y}'$

残差二乗和 Residual Sum of Squares (RSS)

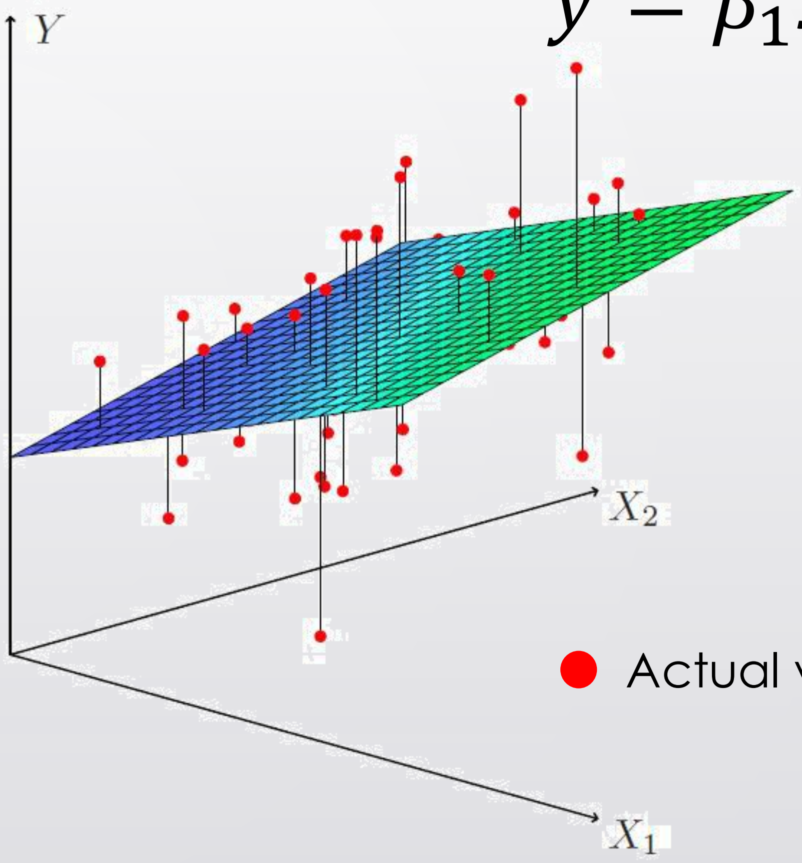
$$RSS = \sum_{1}^{N}(y_n - y_n')^2 = [y_1 - y_1' \;\; y_2 - y_2' \;\; \dots y_N - y_N'] \begin{bmatrix} y_1 - y_1' \\ y_2 - y_2' \\ \vdots \\ y_N - y_N' \end{bmatrix} = (\boldsymbol{y} - \boldsymbol{y}')^T(\boldsymbol{y} - \boldsymbol{y}')$$

残差二乗和 Residual Sum of Suqares



https://www.intage.co.jp/glossary/051/

$$y' = \beta_1 x_1 + \beta_2 x_2$$



https://medium.com/analytics-vidhya/multiple-linear-regression-an-intuitive-approach-f874f7a6a7f9

最小二乗法 Ordinary Least Squares (OLS) Method

$$RSS = \sum_{1}^{N} (y_n - y_n')^2 = (\boldsymbol{y} - \boldsymbol{y}')^T (\boldsymbol{y} - \boldsymbol{y}')$$

RSSを最小にする$\boldsymbol{\beta}$は次の条件を満たす

$\boldsymbol{\beta}$ that minimizes RSS satisfies the condition below

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

正規方程式 Normal Equation

RSSを最小にする$\boldsymbol{\beta}$は次の条件を満たす

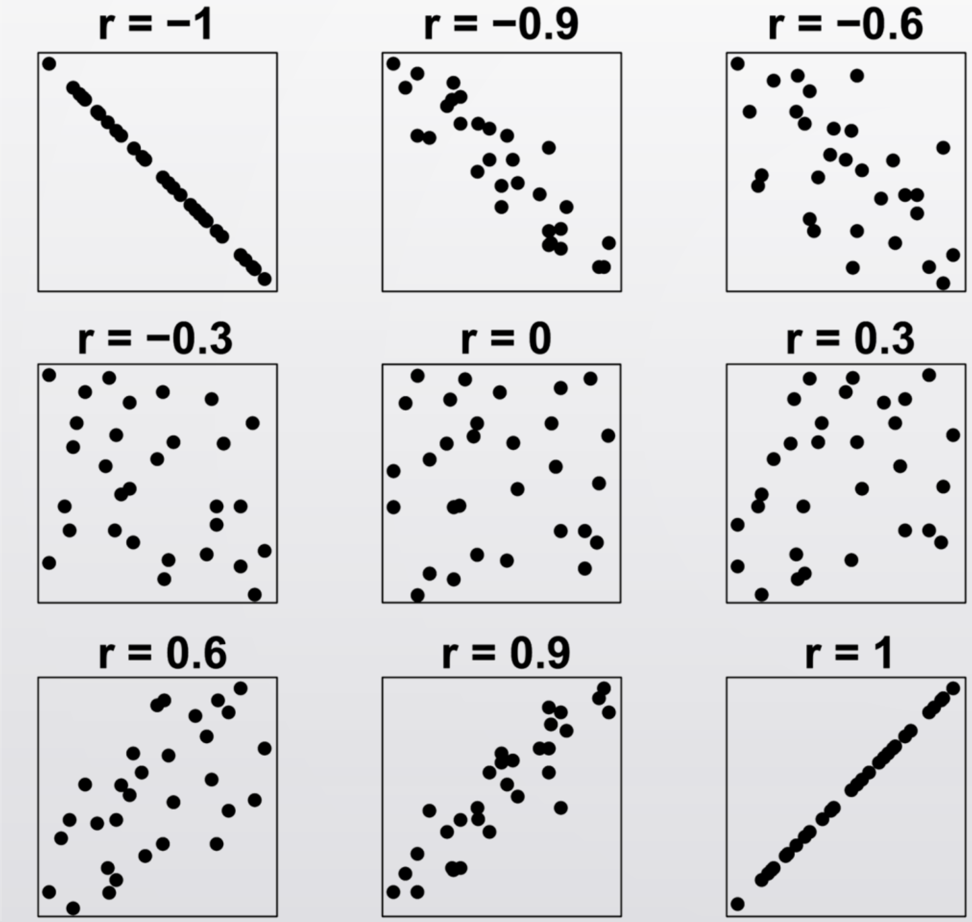
$\boldsymbol{\beta}$ that minimizes RSS satisfies the condition below

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

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \longleftarrow \text{正規方程式 Normal Equation}$$

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

相関係数 Correlational Coefficients



２つの変数の間の関連性の強さを表す

Quantifies the strength of association between two variables

[-1, 1]の間で変動するよう標準化されている

standardized between -1 to 1

共分散 Covariance

$$s_{xy} = \frac{1}{N}\sum_{1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

If

$x_i - \mu_x$と$y_i - \mu_y$が共に正

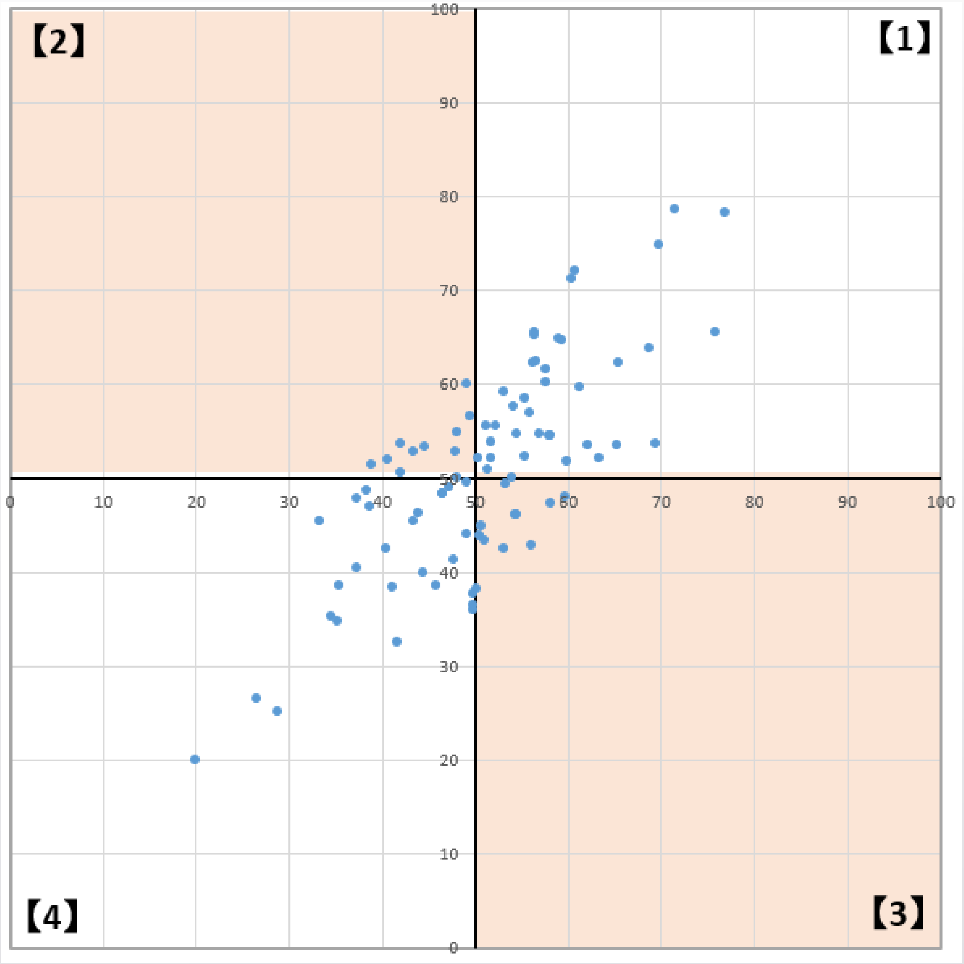
Or

$x_i - \mu_x$と$y_i - \mu_y$が共に負

Then

(x_i, y_i)は【1】か【4】に

(x_i, y_i) belongs to【1】or【4】



https://datasciencehenomiti.com/post-161/

相関係数 Correlational Coefficients

$$r_{xy} = \frac{1}{N}\sum\nolimits_{1}^{N} \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\frac{1}{N}\sum_1^N (x_i - \mu_x)^2}\sqrt{\frac{1}{N}\sum_1^N (y_i - \mu_y)^2}}$$

$$= \frac{1}{N}\sum\nolimits_{1}^{N} \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x\, \sigma_y} = \frac{1}{N}\sum\nolimits_{1}^{N} \frac{x_i - \mu_x}{\sigma_x} \times \frac{y_i - \mu_y}{\sigma_y}$$

$$= \frac{1}{N}\sum\nolimits_{1}^{N} \underset{\text{Z-scored } x_i}{z\text{スコア化された } x_i} \times \underset{\text{Z-scored } y_i}{z\text{スコア化された } y_i}$$

相関係数r_{xy}はzスコア化されたx_iと y_iの共分散

Correlational Coefficient r_{xy} is covariance between z-scored x_i and y_i

正規方程式と相関係数
Normal Equation and Correlational Coefficient

$$\boldsymbol{X} = \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

x_iと y_iが共にzスコア化されているとする

Both x_i and y_i are z-scored

$\|\boldsymbol{x}\|^2 = N, \mu_x = 0, \sigma_x = 1 \quad \|\boldsymbol{y}\|^2 = N, \mu_y = 0, \sigma_y = 1$

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = [x_1\ x_2 \ldots x_N] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \|\boldsymbol{x}\|^2 = N \qquad \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1} = \frac{1}{N}$$

正規方程式と相関係数
Normal Equation and Correlational Coefficient

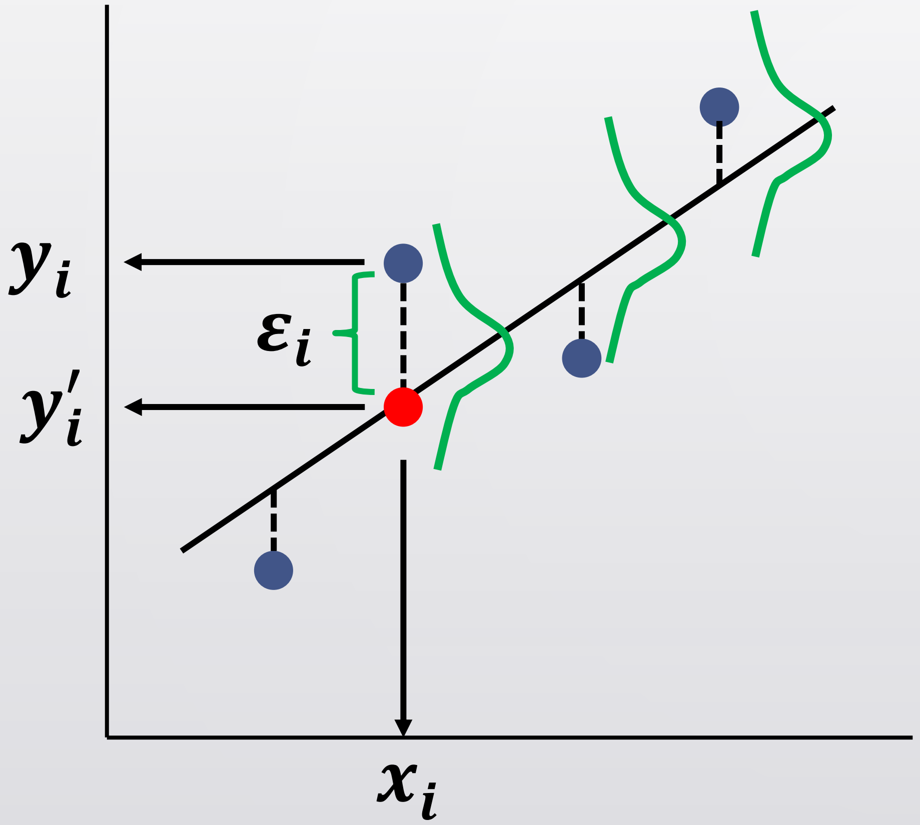
$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{N}\left[x_1\ x_2\ \ldots\ x_N\right]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \frac{1}{N}\left[\frac{x_1-\mu_x}{\sigma_x}\ \frac{x_2-\mu_x}{\sigma_x}\ \ldots\ \frac{x_N-\mu_x}{\sigma_x}\right]\begin{bmatrix} \frac{y_1-\mu_y}{\sigma_y} \\ \frac{y_2-\mu_y}{\sigma_y} \\ \vdots \\ \frac{y_2-\mu_y}{\sigma_y} \end{bmatrix}$$

$$= \frac{1}{N}\sum_1^N \frac{x_i-\mu_x}{\sigma_x} \times \frac{y_i-\mu_y}{\sigma_y} = r_{xy}$$

Zスコア化されたx_iとy_iを正規方程式に投入すると、相関係数r_{xy}が得られる

Correlational coefficient r_{xy} is obtained by entering z-scored x_i and y_i into normal equation

最尤推定による回帰分析
Linear Regression by Maximum Likelihood Estimation

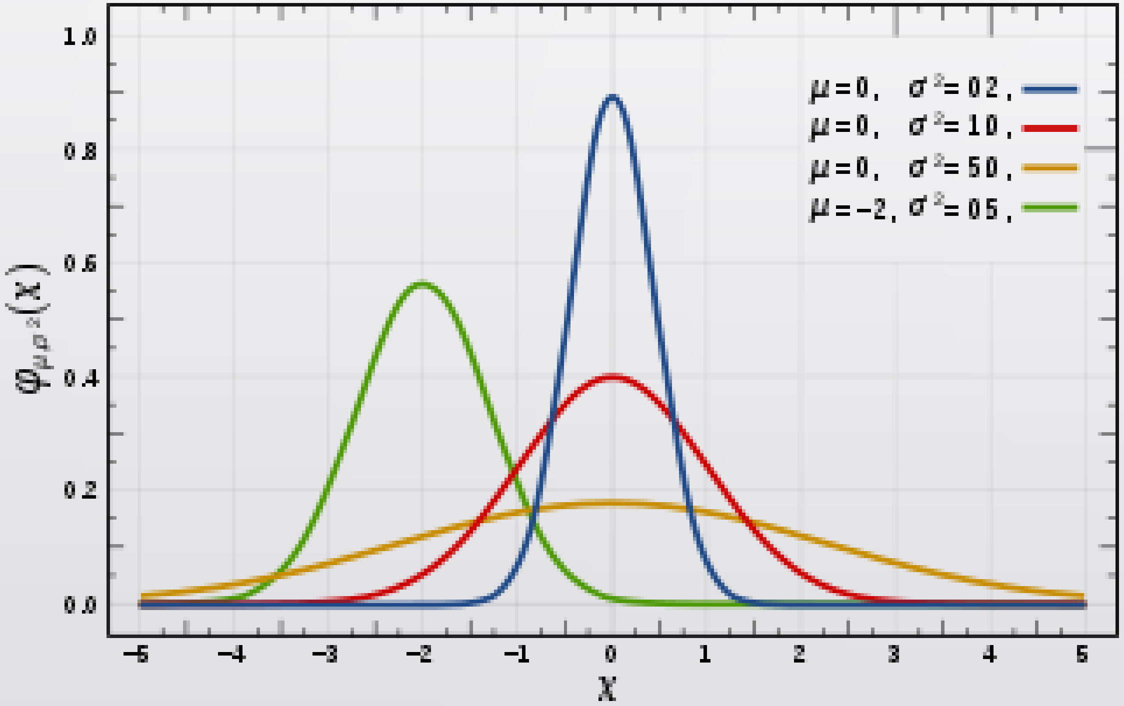


$$y'_i = \boldsymbol{\beta x}_i \quad \varepsilon_i = y_i - y'_i$$

予測誤差が正規分布に従うという前提で、回帰係数$\boldsymbol{\beta}$を推定する

Estimate regression coefficients on the assumption that prediction error conforms to the normal distribution

正規分布 Normal Distribution



https://ja.wikipedia.org/wiki/%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83

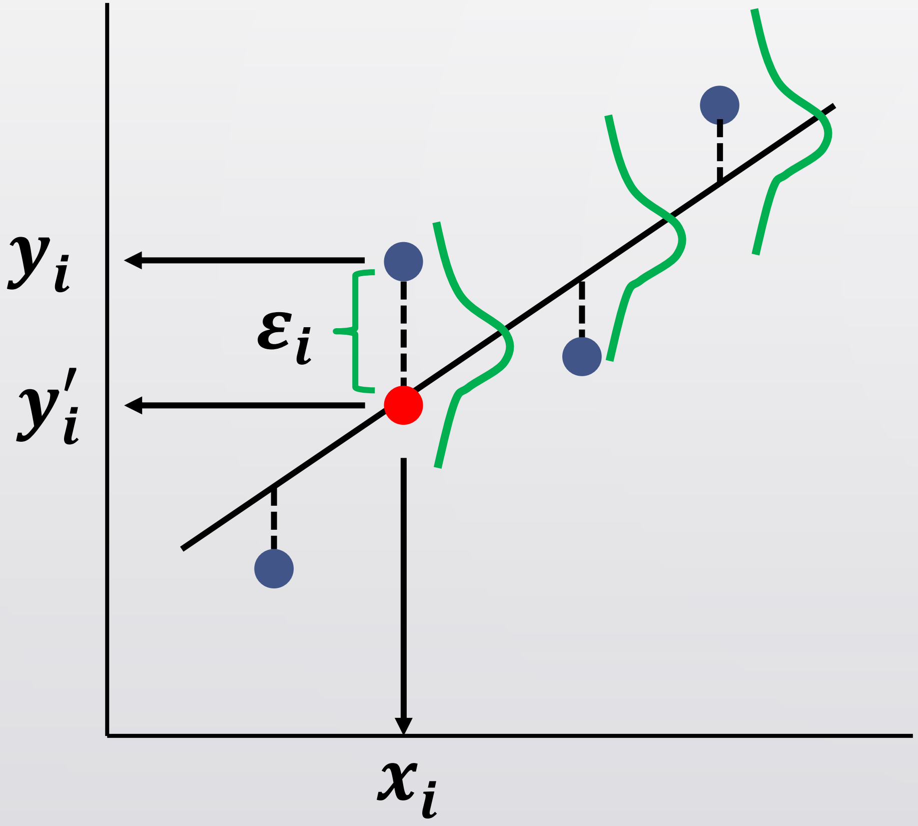
確率密度関数 Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

$\mu = 0, \sigma = 1$の時は、標準正規分布

Standard normal distribution when $\mu = 0, \sigma = 1$

最尤推定による回帰分析
Linear Regression by Maximum Likelihood Estimation



$$y'_i = \beta x_i \quad \varepsilon_i = y_i - y'_i$$

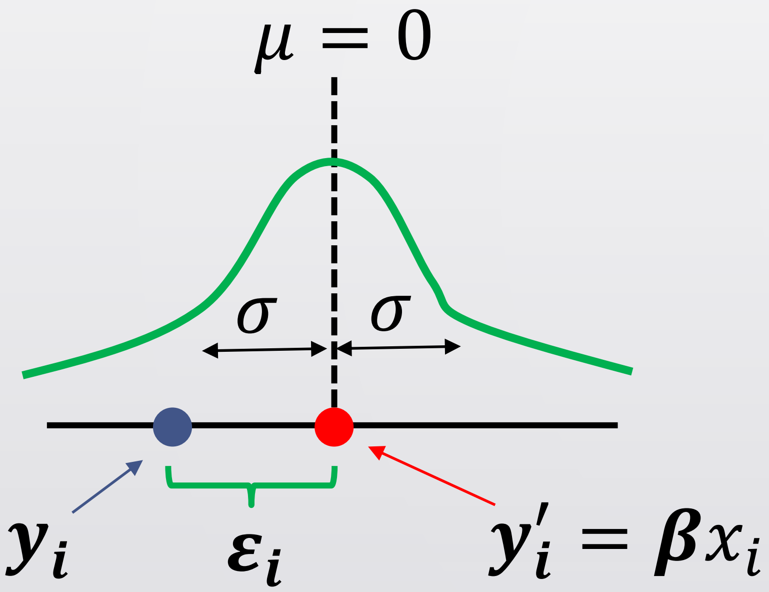
$$\varepsilon_i \sim \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - y'_i)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right)$$

誤差ε_iが0を平均とする正規分布に従うと仮定する

Assume that error ε_i conforms to the normal distribution with mean = 0

最尤推定による回帰分析
Linear Regression by Maximum Likelihood Estimation



$$P(y_i|\boldsymbol{\beta}, \sigma, \boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \boldsymbol{\beta x}_i)^2}{2\sigma^2}\right)$$

$P(y_i|\boldsymbol{\beta}, \sigma, \boldsymbol{x}_i)$:

回帰係数が$\boldsymbol{\beta}$で標準偏差が$\boldsymbol{\sigma}$の時、
x_iに対して, データ$\boldsymbol{y_i}$が観測される確率

Probability that data y_i is observed for x_i under the condition that regression coefficients are $\boldsymbol{\beta}$ and standard deviation is σ

最尤推定 Maximum Likelihood Estimation

データ列$\{y_1, y_2 \ldots y_{N-1}, y_N\}$が観測される同時確率は以下のように書ける

The joint probability that data $\{y_1, y_2 \ldots y_{N-1}, y_N\}$ is observed can be written as follows

$$L = \prod_1^N P(y_i|\boldsymbol{\beta}, \sigma, x_i) = \prod_1^N \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \boldsymbol{\beta}x_i)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_1 - \boldsymbol{\beta}x_1)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_2 - \boldsymbol{\beta}x_2)^2}{2\sigma^2}\right) \times \ \ldots \times \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_N - \boldsymbol{\beta}x_N)^2}{2\sigma^2}\right)$$

最尤推定 Maximum Likelihood Estimation

Lは$\{x_1, x_2 \dots x_{N-1}, x_N\}$ に対して$\{y_1, y_2 \dots y_{N-1}, y_N\}$が観測される同時確率

L is the joint probability that data $\{y_1, y_2 \dots y_{N-1}, y_N\}$ is observed for $\{x_1, x_2 \dots x_{N-1}, x_N\}$

最尤推定ではLが最大化されるような$\boldsymbol{\beta}, \boldsymbol{\sigma}$を求める

In maximum likelihood estimation,$\boldsymbol{\beta}, \sigma$ are determined so that L is maximized

$Log(L)$を最大化する

Maximize $Log(L)$

最尤推定 Maximum Likelihood Estimation

$$L = \prod_{1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \boldsymbol{\beta} x_i)^2}{2\sigma^2}\right)$$

$$Log(L) = \sum_{1}^{N} log\left(\frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \boldsymbol{\beta} x_i)^2}{2\sigma^2}\right)\right)$$

$$= -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{1}^{N} (y_i - \boldsymbol{\beta} x_i)^2$$

最尤推定 Maximum Likelihood Estimation

$$Log(L) = -\frac{N}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_1^N (y_i - \boldsymbol{\beta} x_i)^2$$

$$\frac{\partial Log(L)}{\partial \boldsymbol{\beta}} = 0 \qquad RSS = \sum_1^N (y_n - y_n')^2 \qquad y_i' = \boldsymbol{\beta} x_i$$

$Log(L)$を最大化する$\boldsymbol{\beta}$は, RSSを最小化する

$\boldsymbol{\beta}$ that maximizes $Log(L)$ minimizes RSS

２つの重回帰分析 Two Types of Multiple Regressions

最小二乗法
Ordinary Least Squares Method

RSSを最小化 Minimize RSS

最尤推定
Maximum Likelihood Estimation

誤差εが正規分布すると仮定
Assume that error ε conforms to normal distribution

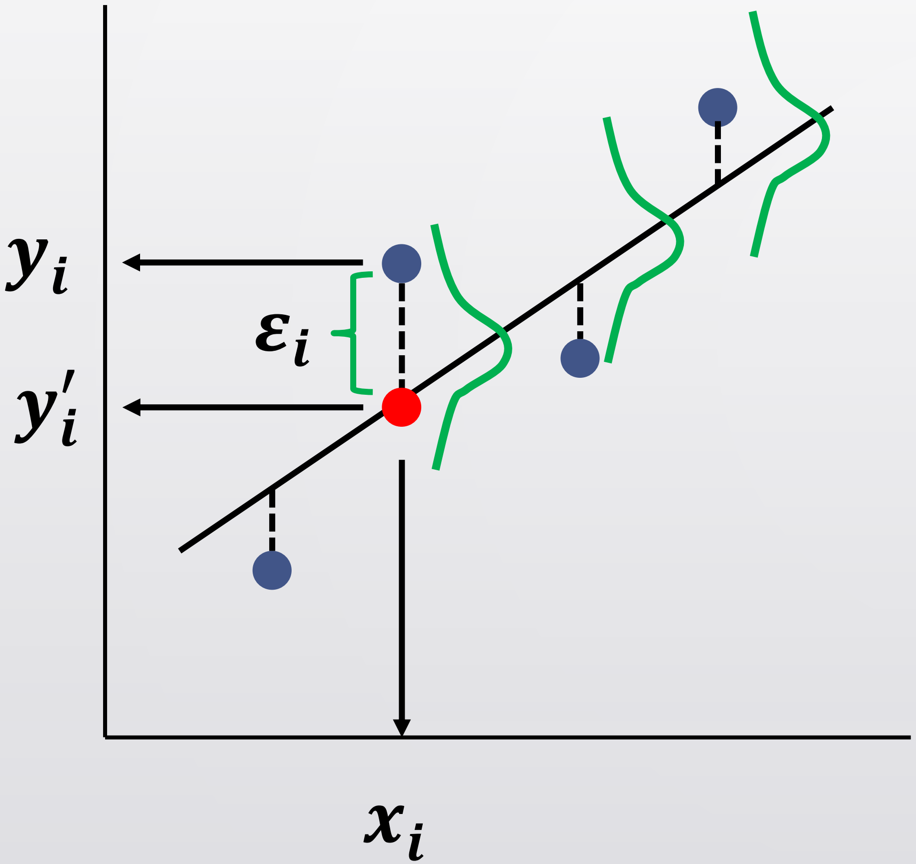
$Log(L)$を最大化
Maximize $Log(L)$

$$\sigma = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_n - y'_n)^2}$$

RSSを最小化 Minimize RSS

$\boldsymbol{\beta}$

一般化線型モデル Generalized Linear Model (GLM)



誤差が平均0の正規分布に従う

Error conforms to the normal distribution with $\mu = 0$

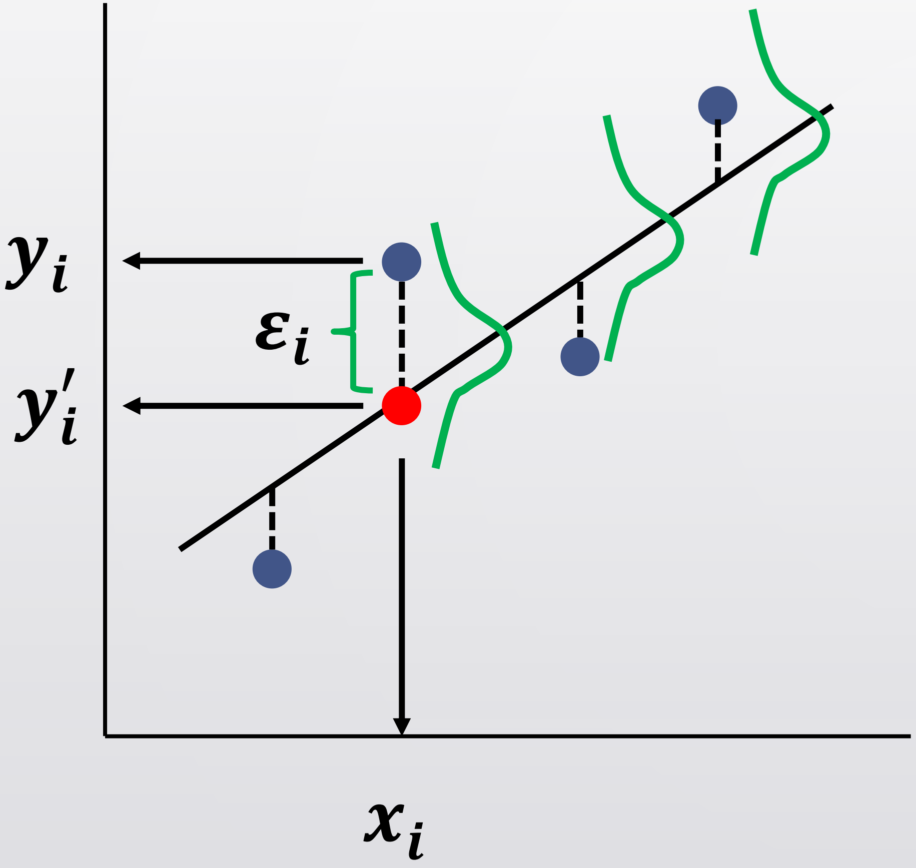
y_i'の期待値は$\boldsymbol{\beta}x_i$になる

Expected value of y_i' is $\boldsymbol{\beta}x_i$

$$g(E[y_i']) = \boldsymbol{\beta}x_i$$

$$g(\mu) = \mu$$

一般化線型モデル Generalized Linear Model (GLM)



$$g(E[y_i]) = \boldsymbol{\beta} x_i$$

$$g(\mu) = \mu$$

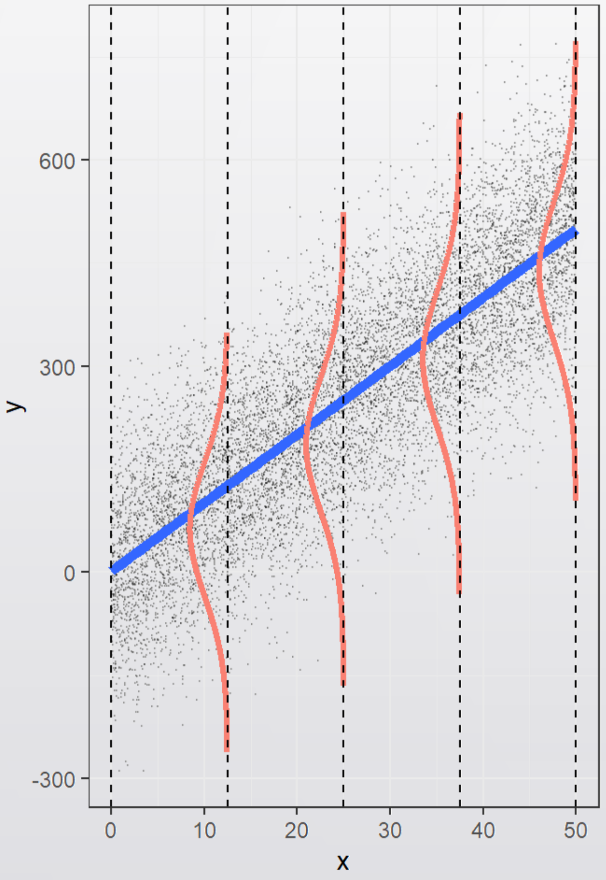
$\boldsymbol{\beta} x_i$: 線型予測子 Linear Predictor

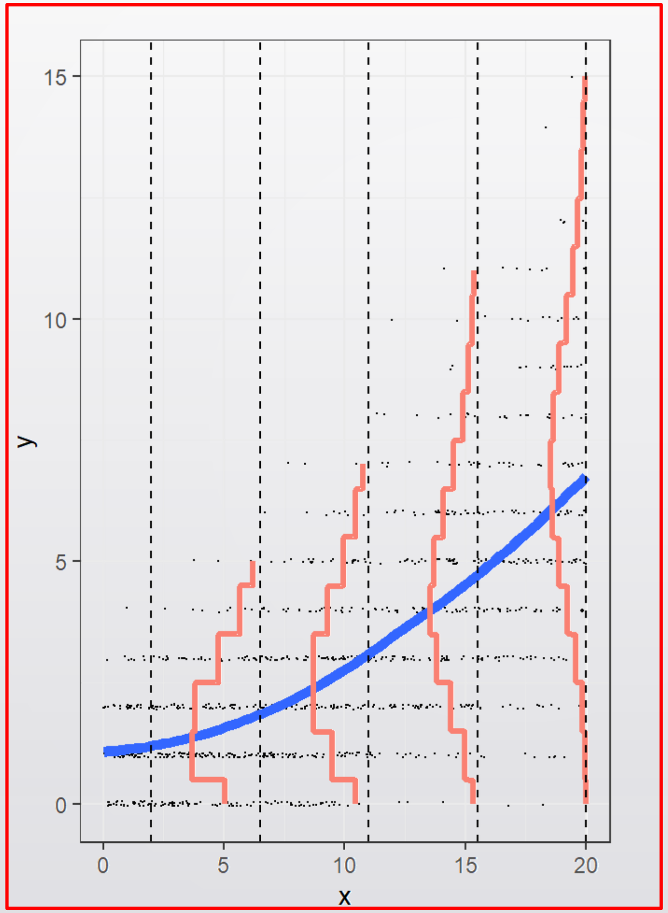
g: リンク関数 Link Function

重回帰分析の誤差構造は正規分布である

Error structure of multiple linear regression is normal distribution

ポアソン回帰 Poisson Regression





x_iが大きくなる程, y_iの期待値・分散が大きくなる

As x_i gets larger, so do expected value and variance of y_i

$$g(E[y_i]) = \boldsymbol{\beta} x_i$$

$$g(\mu) = log(\mu)$$

$$E[y_i] = V[y_i] = e^{\boldsymbol{\beta} x_i}$$

https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html