



データマイニング

Data Mining

5: 回帰② Regression

土居 裕和 Hirokazu Doi

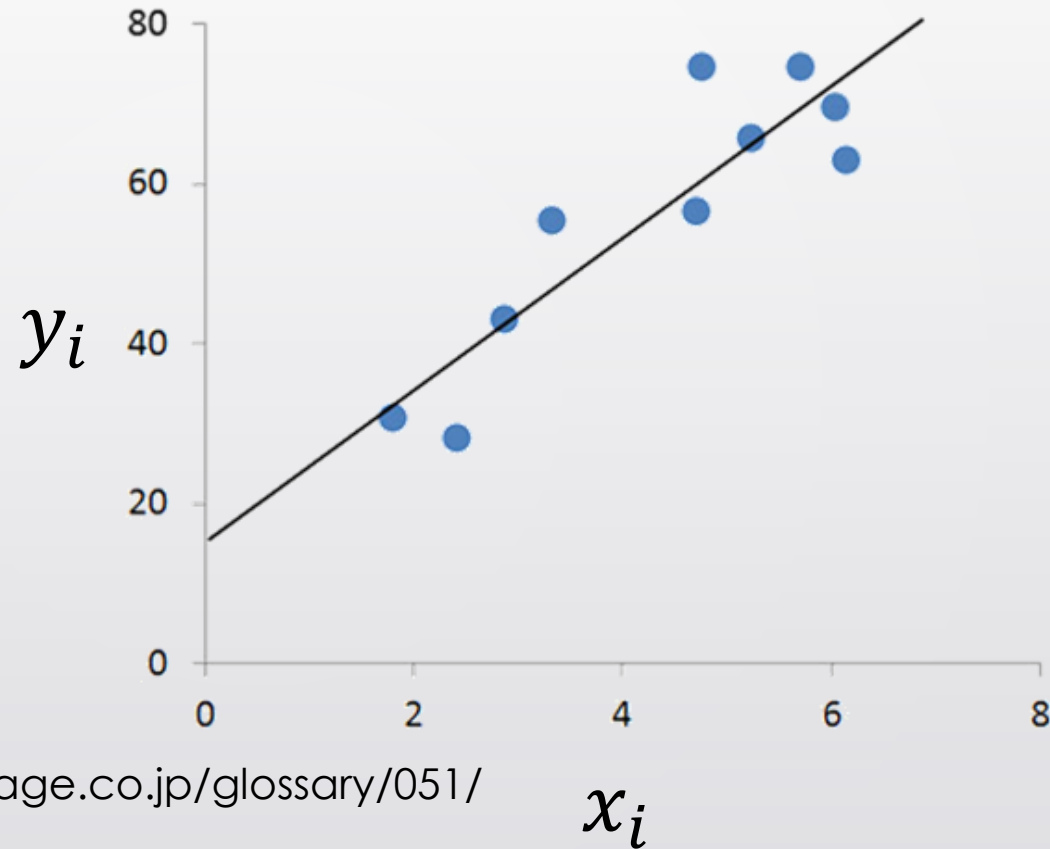
長岡技術科学大学 Nagaoka University of Technology

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



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

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


$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

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

最小二乗法 Ordinary Least Squares (OLS) Method

\mathbf{X} から \mathbf{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 \mathbf{y} based on \mathbf{X} with good precision

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


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

観測値 \mathbf{y} と予測値 \mathbf{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 \mathbf{y} and predicted vector \mathbf{y}'

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

β

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

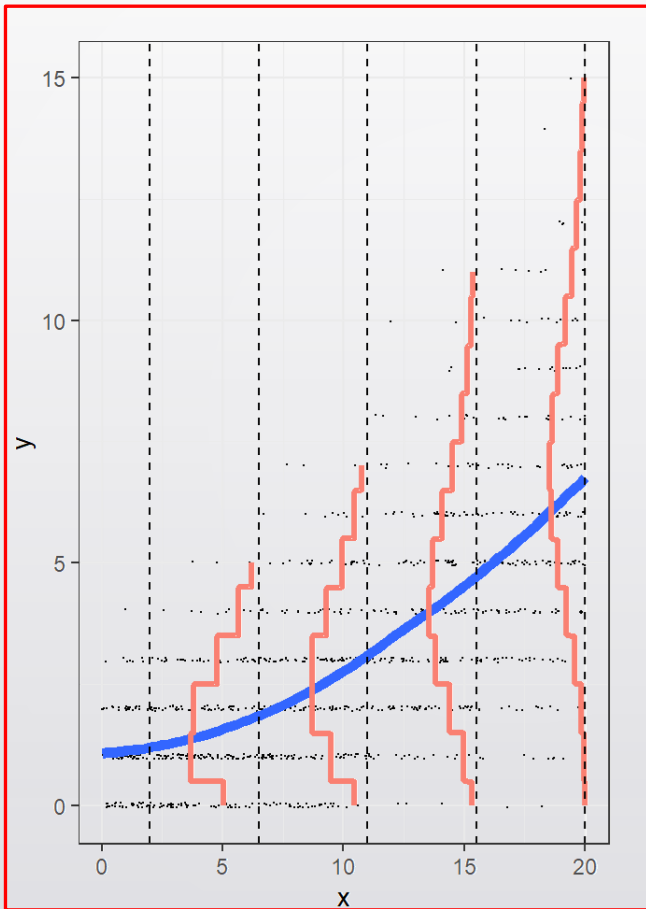
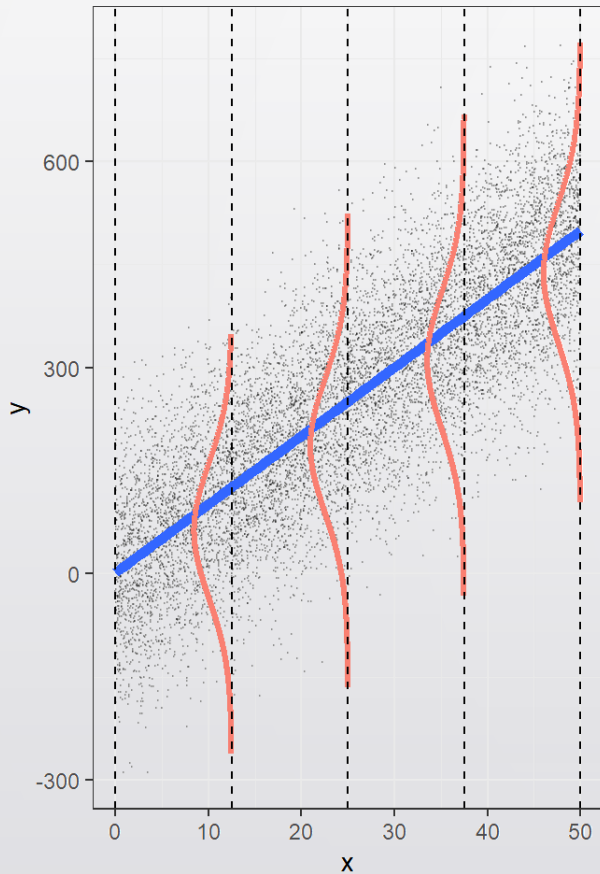
$\text{Log}(L)$ を最大化

Maximize $\text{Log}(L)$

RSS を最小化 Minimize RSS

$$\sigma = \sqrt{\frac{1}{N} \sum_1^N (y_n - y'_n)^2}$$

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

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

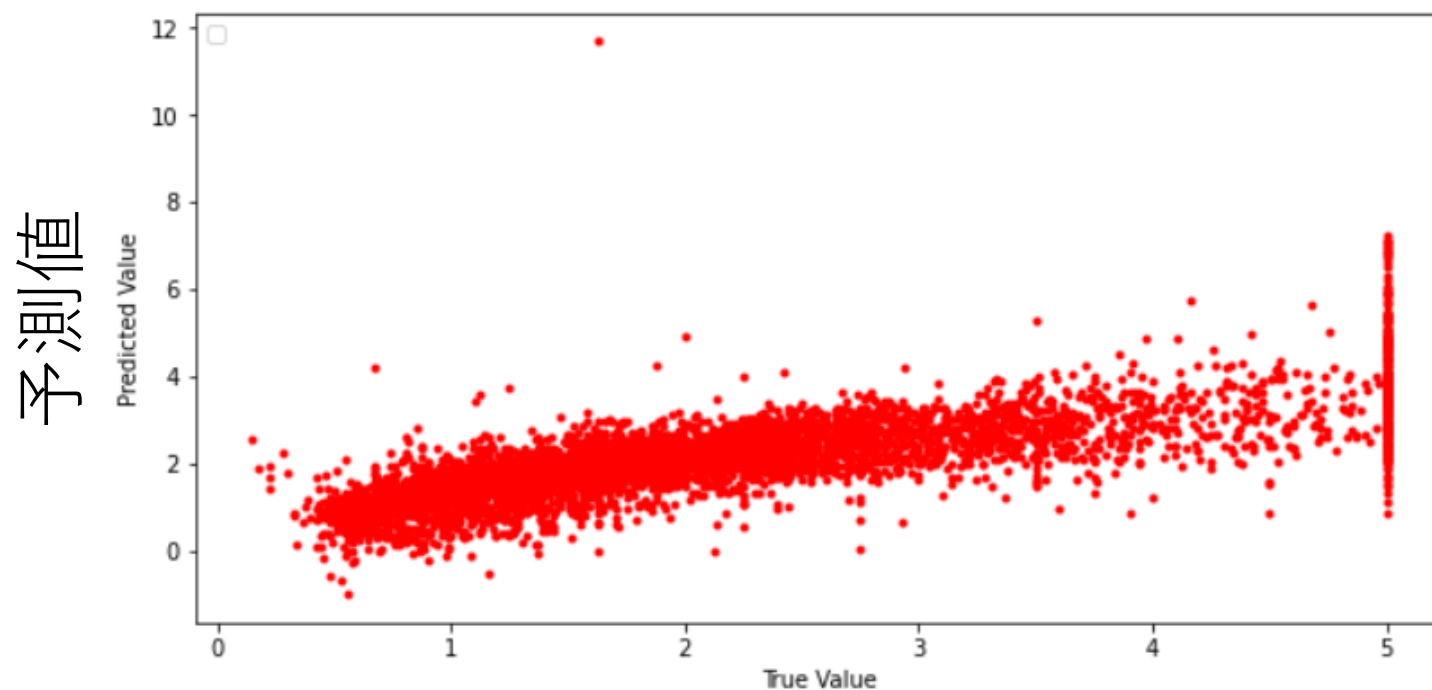
$$g(\mu) = \log(\mu)$$

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

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

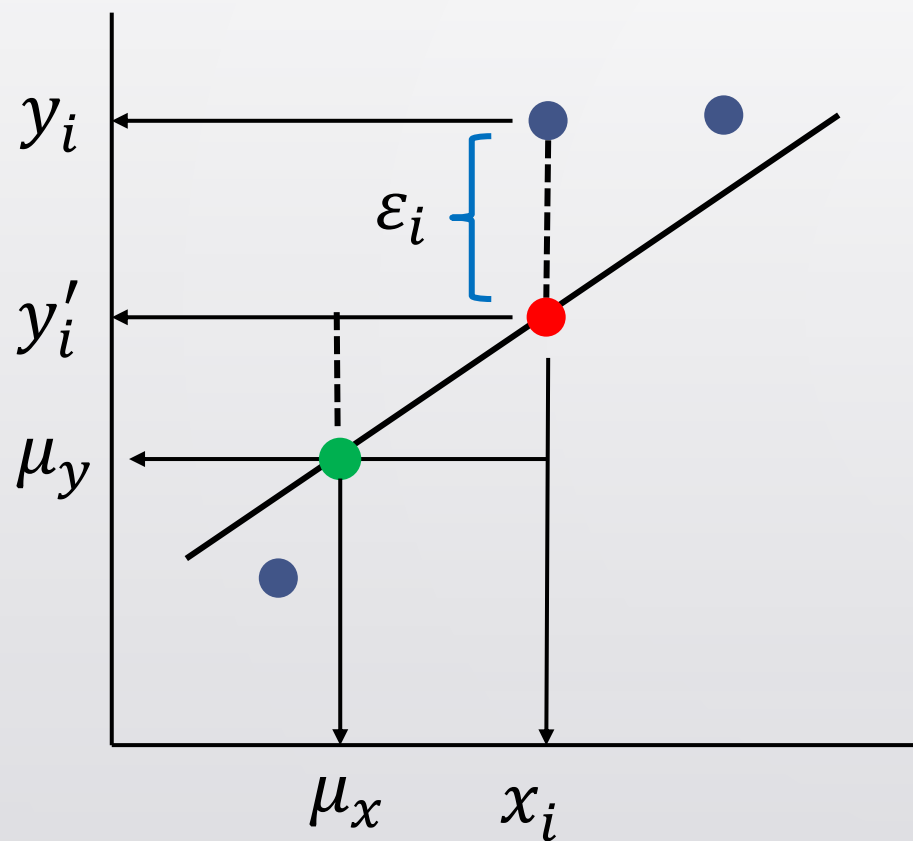
相関係数による性能評価

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



正解値

決定係数 R^2 Coefficient of Determination



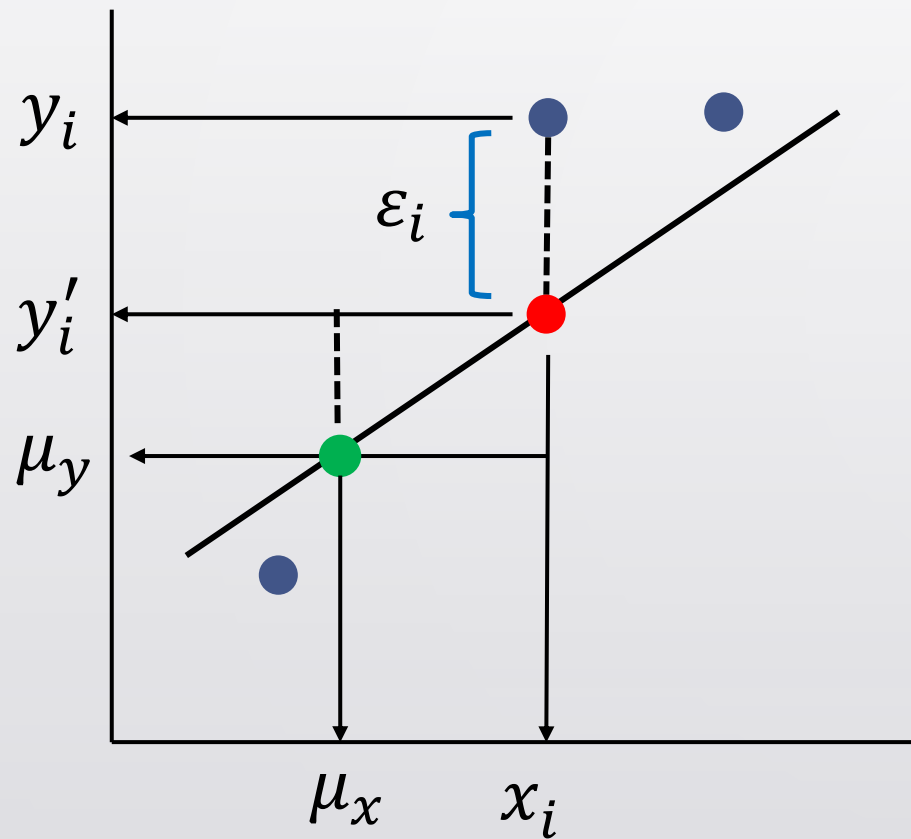
$$S_{total} = \sum_{i=0}^N (y_i - \mu_y)^2$$

$$S_{model} = \sum_{i=0}^N (y'_i - \mu_y)^2$$

$$RSS = \sum_{i=0}^N (\epsilon_i)^2 = \sum_{i=0}^N (y_i - y'_i)^2$$

残差二乗和 Squared Sum of Residuals

決定係数 R^2 Coefficient of Determination



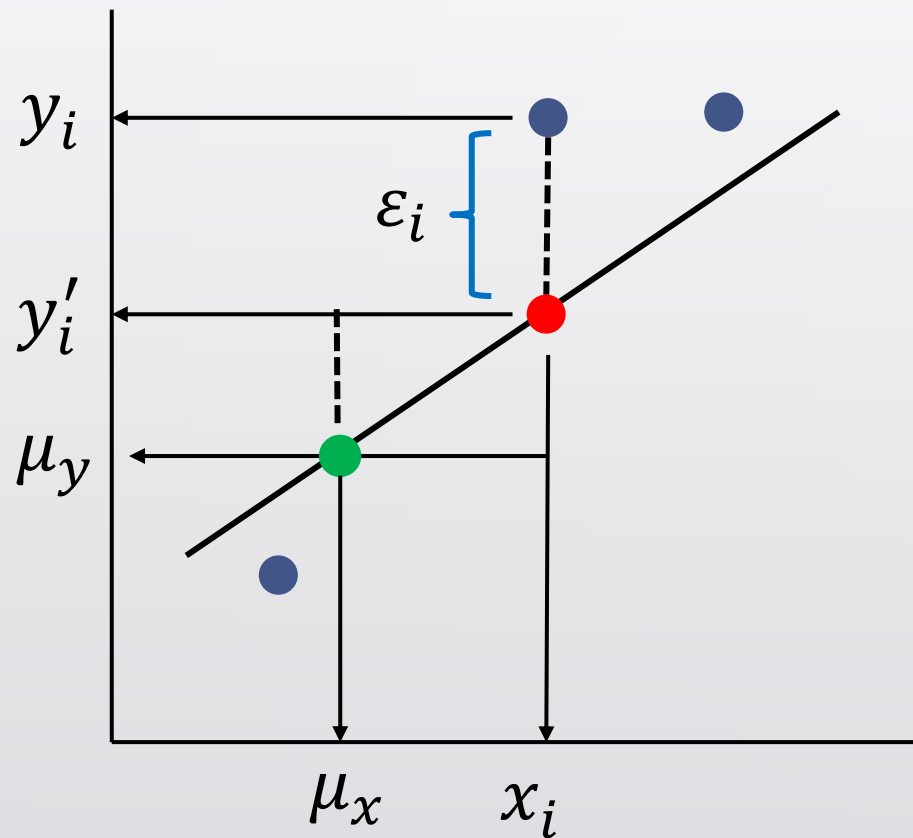
$$y_i - \mu_y = (y'_i - \mu_y) + \varepsilon_i$$

誤差 ε_i が平均0の正規分布に従うとき

When error ε_i conforms to the normal distribution $N(0, \sigma^2)$

$$S_{total} = S_{model} + RSS$$

決定係数 R^2 Coefficient of Determination



$$S_{total} = S_{model} + RSS$$

$$R^2 = 1 - \frac{RSS}{S_{total}} = \frac{S_{model}}{S_{total}}$$

単回帰の場合、決定係数は相関係数の二乗に一致する

In the case of simple regression, coefficient of determination equals to squared coefficient of regression

決定係数 R^2 Coefficient of Determination

調整可能なモデルパラメータが多い程,決定係数は大きくなる

The larger the number of adjustable model parameters, the larger the coefficient of determination becomes

重回帰分析の場合は、自由度調整済み決定係数を用いる

Adjusted coefficient of determination as described below is used in the case of multiple regression

$$\text{Adjusted } R^2 = 1 - \frac{RSS}{S_{total}} \frac{n - 1}{n - p - 1}$$

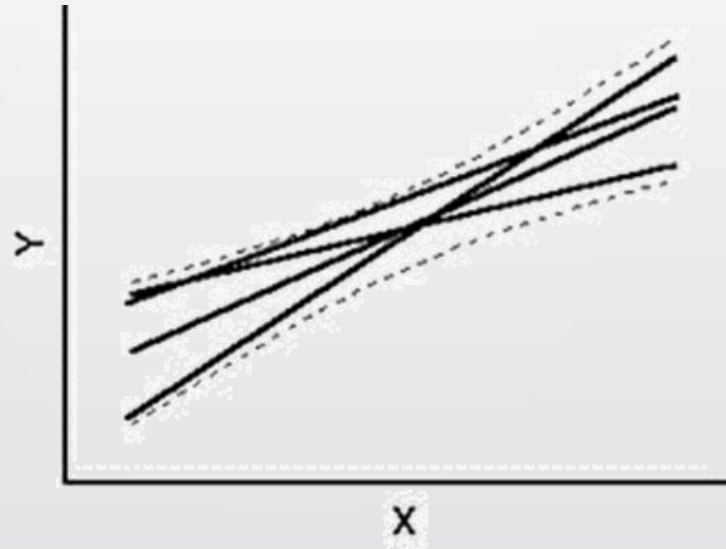
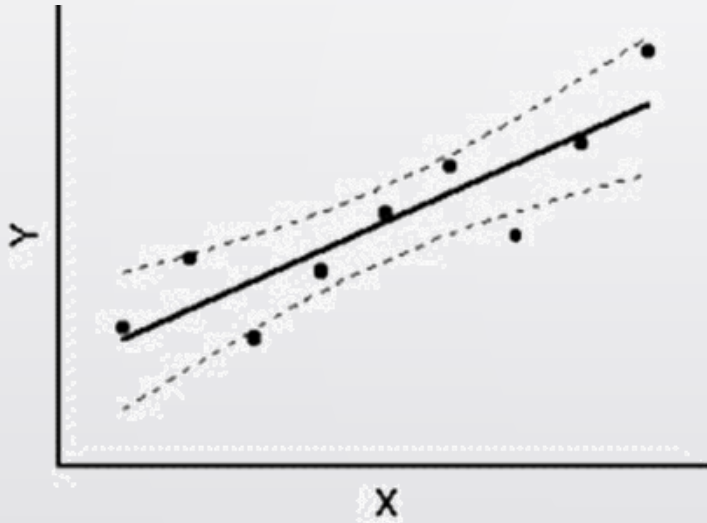
n : サンプル数
Sample Size

p : 予測変数の数
The number of predictors

信頼区間 Confidence Interval

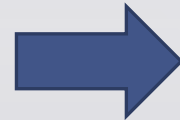
信頼区間は、回帰モデルの不確実性を表す

Confidence interval represents uncertainty about regression model



<https://real-statistics.com/regression/confidence-and-prediction-intervals/>

点線は95%信頼区間
Dotted lines represent 95% confidence interval

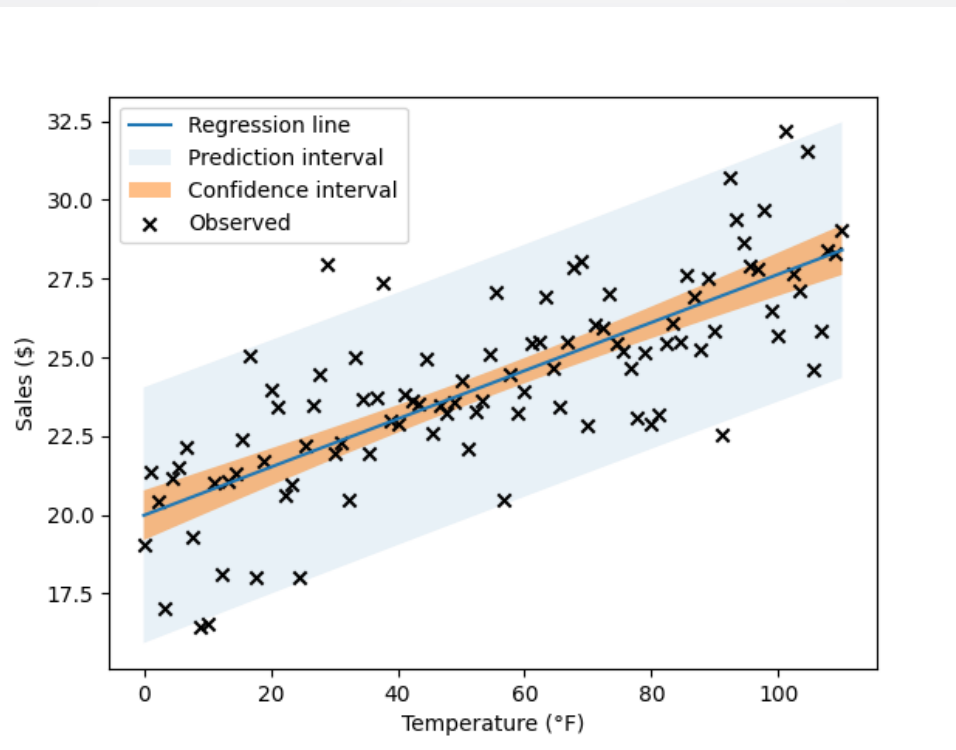


回帰直線は95%の確率で点線内のどこかに引かれる
Regression line falls somewhere inside the region defined by dotted lines with 95% probability

予測区間 Prediction Interval

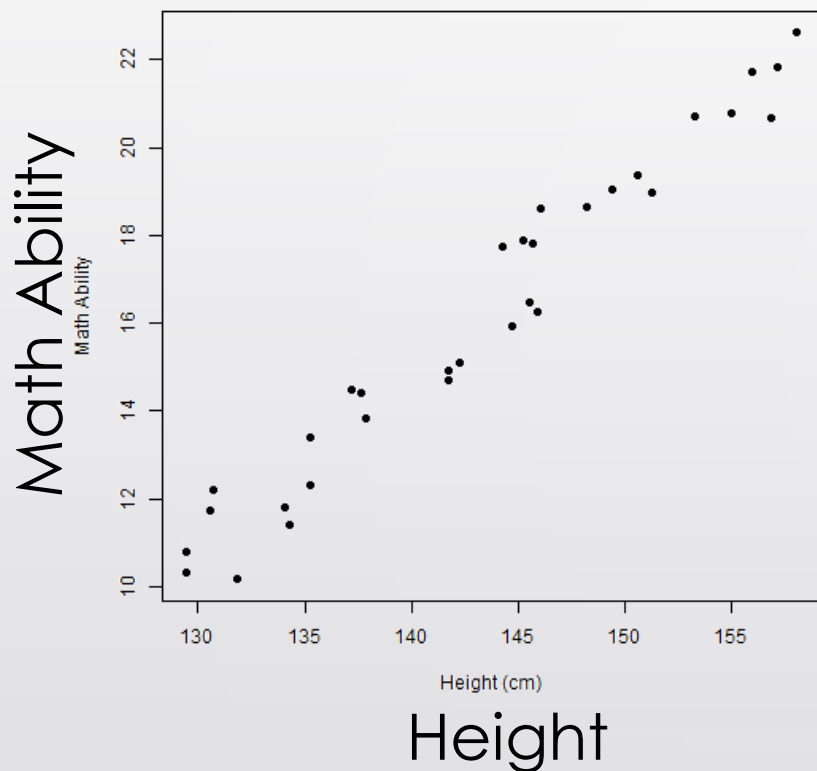
予測区間は、 y の予測値 y' の不確実性を表す

Prediction interval represents uncertainty about estimation of y based on given x

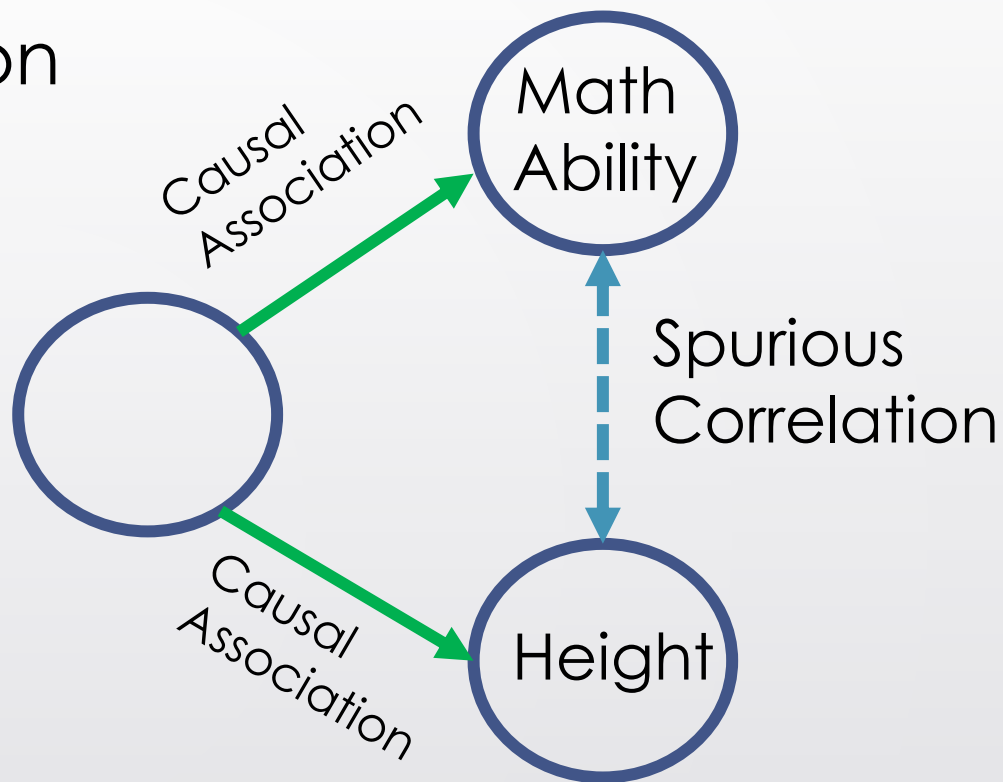


https://lmc2179.github.io/posts/confidence_prediction.html

疑似相関 Spurious Correlation



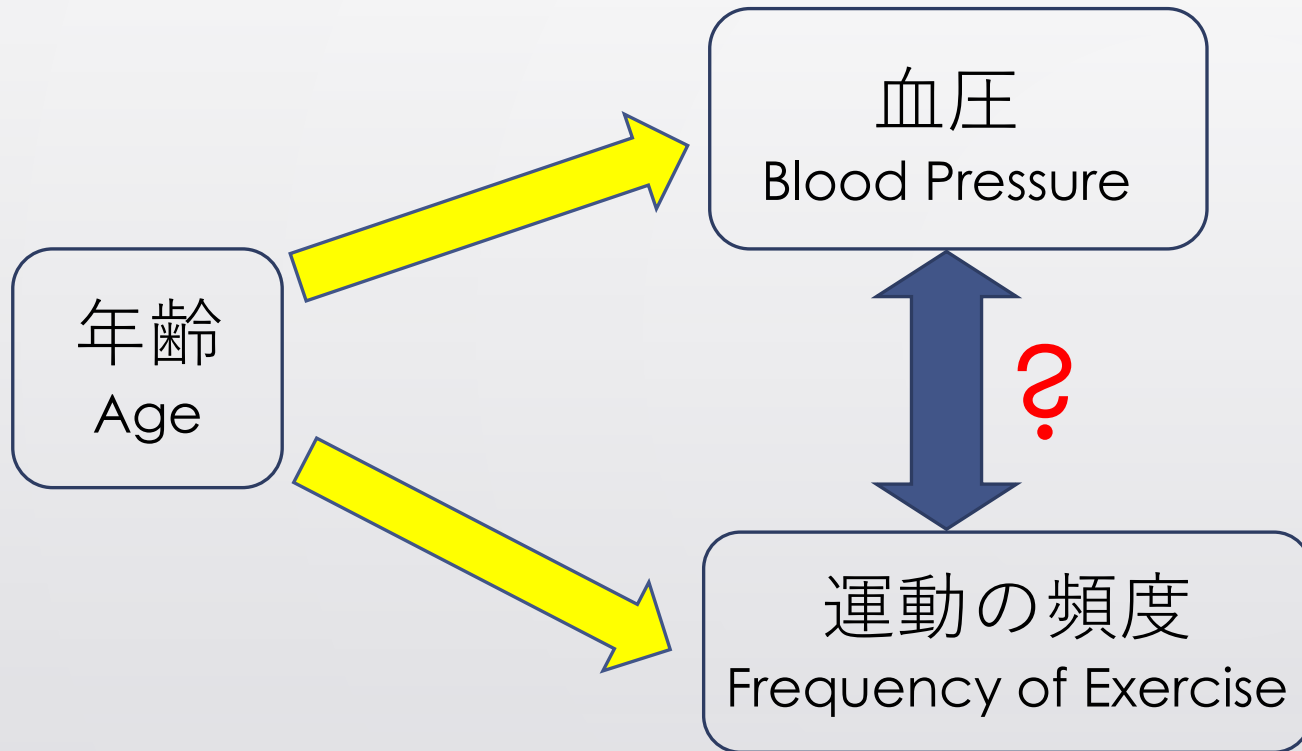
<https://hoxo-m.hatenablog.com/entry/20130711/p1>



データ分析では、見かけの関係性に注意

Be aware of spurious association in any kinds of data analysis

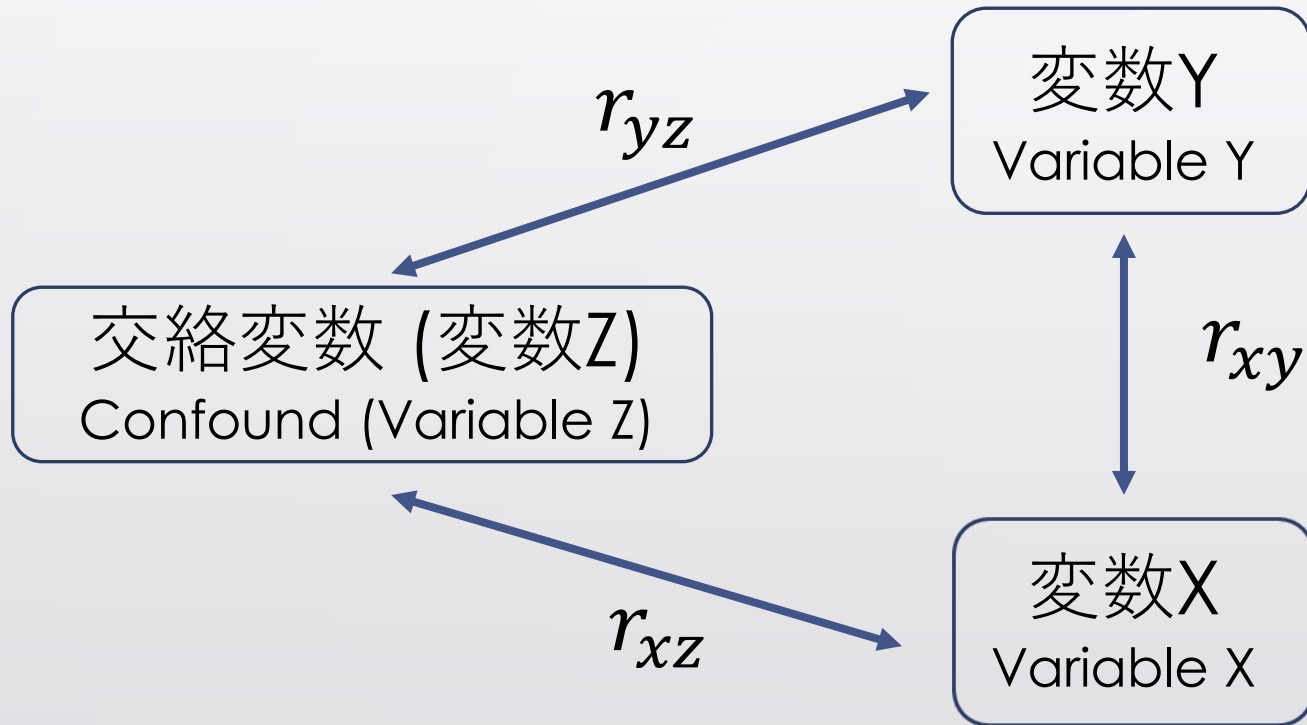
交絡変数 Confound



血圧と運動の頻度の関係には、
年齢が交絡している

Association between blood pressure and
exercise frequency is confounded by age

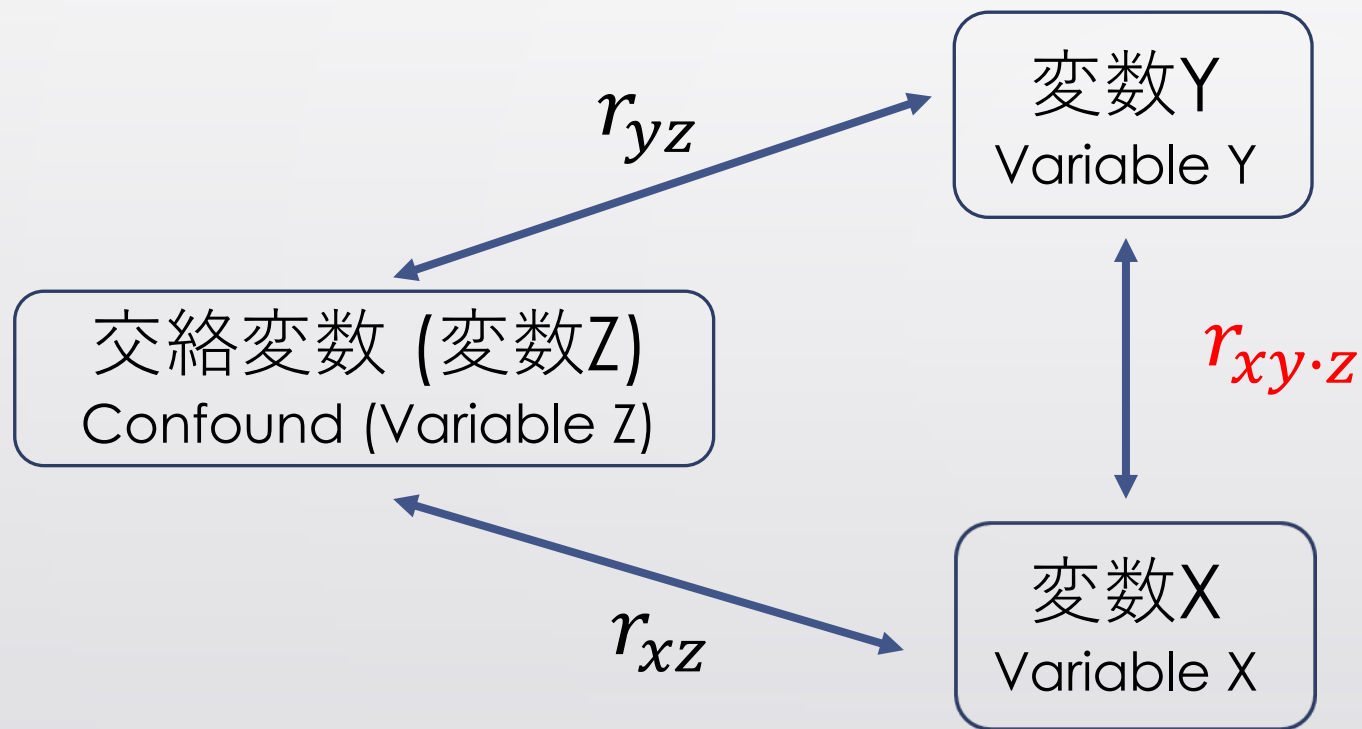
交絡変数 Confound



変数xとyの相関を分析するには、
変数zの影響を除く必要がある

To analyze the association between
variable x and y, the influence (confound)
of variable x should be eliminated

交絡変数 Confound



$$y'_z = Z\beta_{yz}$$

y'_z は変数 Z で説明できる Y の情報

y'_z is information of Y explainable by Z

$$x'_z = Z\beta_{xz}$$

x'_z は変数 Z で説明できる X の情報

x'_z is information of X explainable by Z

偏回帰係数 Partial Correlation Coefficient

$y - y'_z$: 変数 Z で説明できない Y の情報 Information of Y unexplainable by Z

$x - x'_z$: 変数 Z で説明できない X の情報 Information of X unexplainable by Z

X と Y の偏相関係数 $r_{xy \cdot z}$ は、 $x - x'_z$ と $y - y'_z$ の相関係数

Partial correlational coefficient $r_{xy \cdot z}$ between X and Y is correlational coefficient between $x - x'_z$ and $y - y'_z$

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

正規方程式 Normal Equation

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

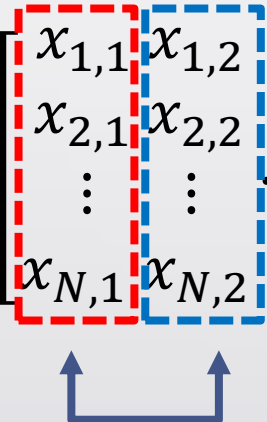
$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y} \leftarrow \text{正規方程式 Normal Equation}$$

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

多重共線性 Multicollinearity

予測変数の間に強い相関があると、回帰モデルが不安定化する

Regression model becomes unstable when there is strong correlation among predictors

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$


強い相関がある

Strongly correlates with each other

多重共線性 Multicollinearity

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$


極端なケース Extreme Case

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 6 \\ 1 & 2 & 1 \end{bmatrix} \quad \det \mathbf{X}^T \mathbf{X} = 0$$

行列式が0なので $\mathbf{X}^T \mathbf{X}$ の逆行列がない
Cannot find inverse matrix of $\mathbf{X}^T \mathbf{X}$ because its determinant is zero

多重共線性 Multicollinearity

現実的な例 More realistic example

$$\mathbf{X}_1 = \begin{bmatrix} -1.12 & -0.51 & 0.69 \\ -0.43 & -1.12 & 1.02 \\ 0.37 & 1.10 & -0.98 \\ 1.19 & 0.53 & -0.73 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} -1.12 & -0.51 & 0.70 \\ -0.43 & -1.12 & 1.02 \\ 0.36 & 1.10 & -0.98 \\ 1.20 & 0.53 & -0.73 \end{bmatrix}$$


とてもよく似た行列 Quite similar matrices

$$0.25 \times (1\text{列目}) - 0.8 \times (2\text{列目}) = (3\text{列目})$$

多重共線性 Multicollinearity

$$\mathbf{y} = \begin{bmatrix} 0.40 \\ 1.17 \\ -1.14 \\ -0.42 \end{bmatrix} \quad \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \boldsymbol{\beta}_1 = \begin{bmatrix} 0.54 \\ 0.24 \\ 1.64 \end{bmatrix} \quad \boldsymbol{\beta}_2 = \begin{bmatrix} -0.42 \\ -2.86 \\ -2.20 \end{bmatrix}$$

データ \mathbf{X}_1 は \mathbf{X}_2 よく似ているのに、回帰係数 $\boldsymbol{\beta}_1$ と $\boldsymbol{\beta}_2$ は全く異なる

Correlational coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are completely different from each other whereas data matrices \mathbf{X}_1 and \mathbf{X}_2 are quite similar

多重共線性があると、データのわずかな違いで、回帰分析の結果が大きく変化する

When there is multicollinearity, slight difference in data results in great change in regression result

分散拡大係数 Variance Inflation Factor (VIF)

VIFをチェックすることで、多重共線性が生じていないかを確認する

See if there is multicollinearity by checking VIF

$$\mathbf{x}'_j = \begin{bmatrix} x'_{1,j} \\ x'_{2,j} \\ \vdots \\ x'_{N,j} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & \mathbf{x_{1,j-1}} & \mathbf{x_{1,j+1}} & \dots & x_{1,M} \\ x_{2,1} & \dots & \mathbf{x_{2,j-1}} & \mathbf{x_{2,j+1}} & \dots & x_{2,M} \\ & & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & \dots & \mathbf{x_{N,j-1}} & \mathbf{x_{N,j+1}} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \mathbf{\beta_{j-1}} \\ \mathbf{\beta_{j+1}} \\ \vdots \\ \beta_M \end{bmatrix} \quad \mathbf{x}_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{N,j} \end{bmatrix}$$

R_j^2 = \mathbf{x}_j と \mathbf{x}'_j の決定係数



分散拡大係数 Variance Inflation Factor (VIF)

R_j^2 = x_j と x_j' の決定係数

$$VIF_j = \frac{1}{1 - R_j^2}$$

$VIF_j > 10$ を多重共線性の基準とすることが多い

$VIF_j > 10$ is usually taken as threshold of multicollinearity

第 k 主成分 k -th Principal Component

$$\mathbf{t}_k = \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ \vdots \\ t_{N,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,k} \\ p_{2,k} \\ \vdots \\ p_{M,k} \end{bmatrix} = \mathbf{X} \mathbf{p}_k$$

主成分同士は直交している Principal components are orthogonal

$$\mathbf{p}_i^T \mathbf{p}_j = \begin{cases} 1(i = j) \\ 0(i \neq j) \end{cases}$$

主成分回帰 Principal Component Regression

主成分回帰では、予測変数を主成分分析にかけたのち、主成分得点を回帰分析にかける

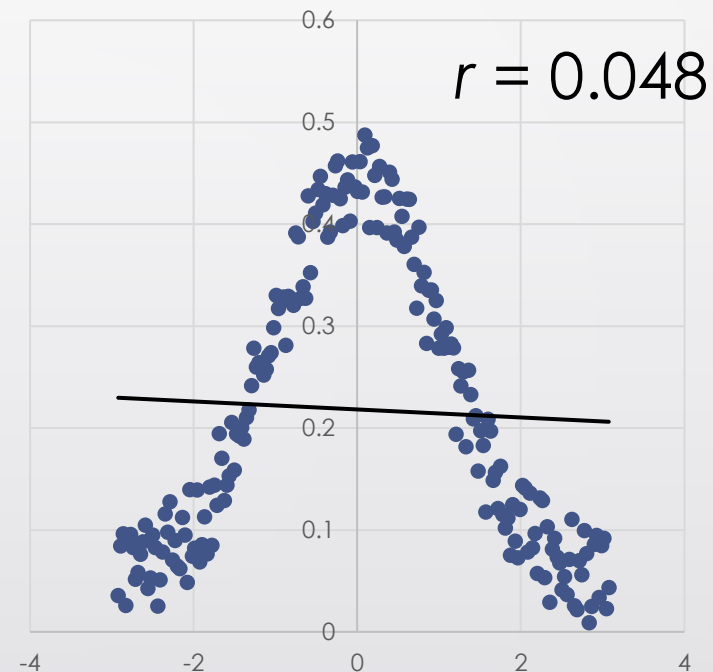
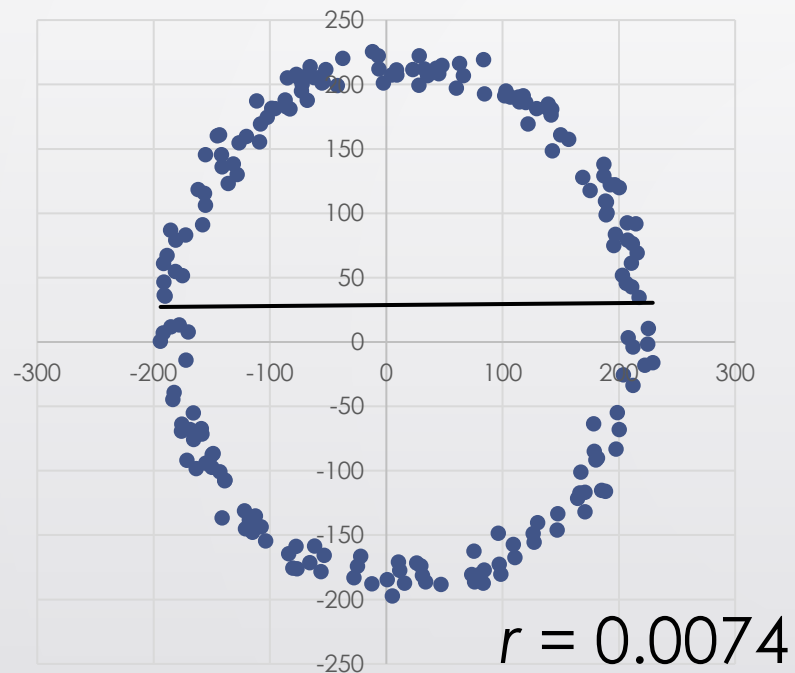
In principal component regression, factor scores are entered in to multiple regression after predictors are submitted to PCA

$$\begin{matrix} T' = \underbrace{[t_1 \ t_2 \ \dots \ t_H]}_{\substack{\text{主成分得点} \\ \text{Factor Scores}}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,H} \\ p_{2,1} & p_{2,2} & \dots & p_{2,H} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,H} \end{bmatrix} = X[p_1 \ p_2 \ \dots \ p_H] \\ \begin{matrix} (N, H) & & (N, M) & & (M, H) \end{matrix} \end{matrix}$$



重回帰分析

非線形的な関係性 Nonlinear Association



変数間の関係は、適切なモデルで評価する必要がある
Association between variables should be estimated by appropriate model

多項式回帰 Polynomial Regression

Simple
Linear
Regression

$$y = b_0 + b_1 x_1$$

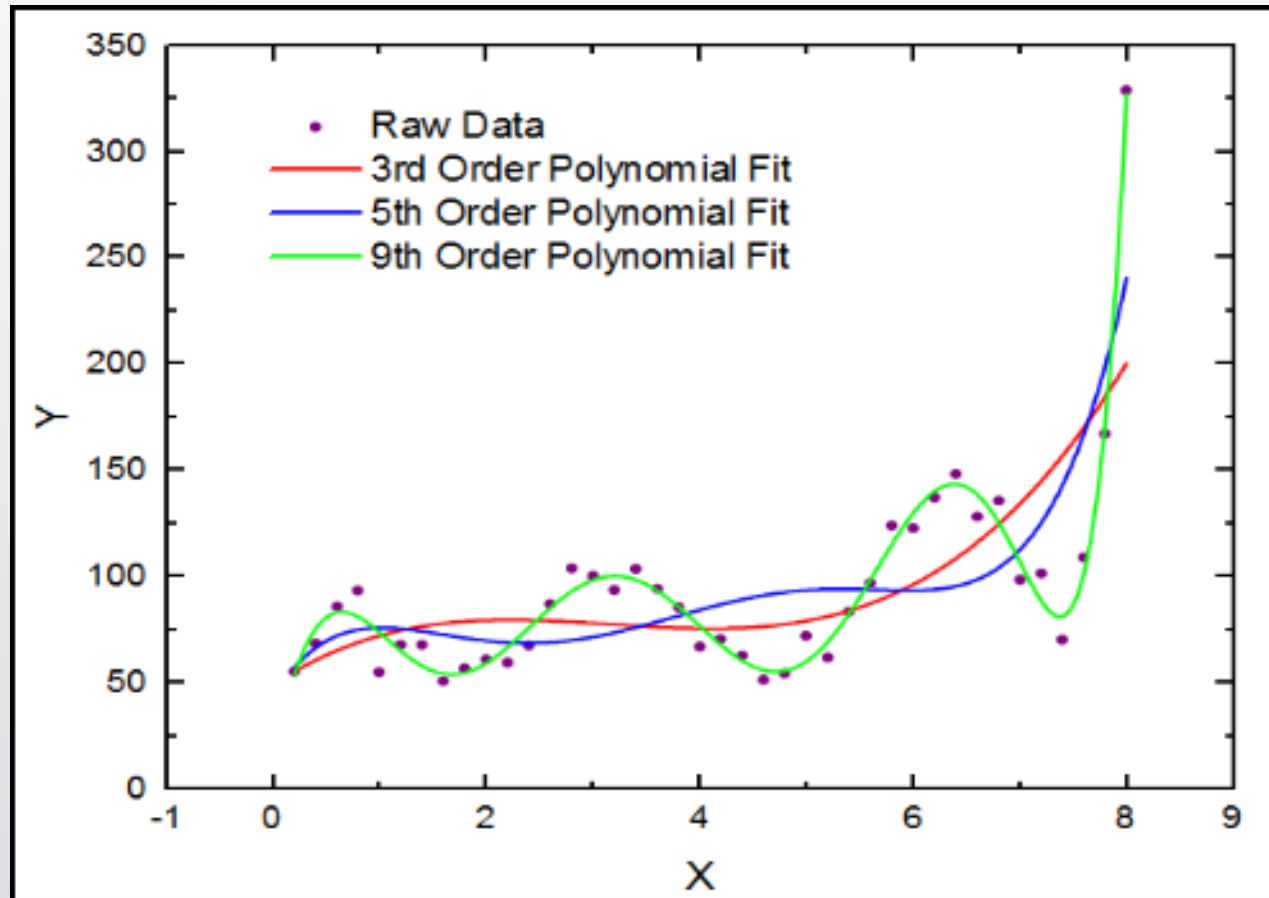
Multiple
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

Polynomial
Linear
Regression

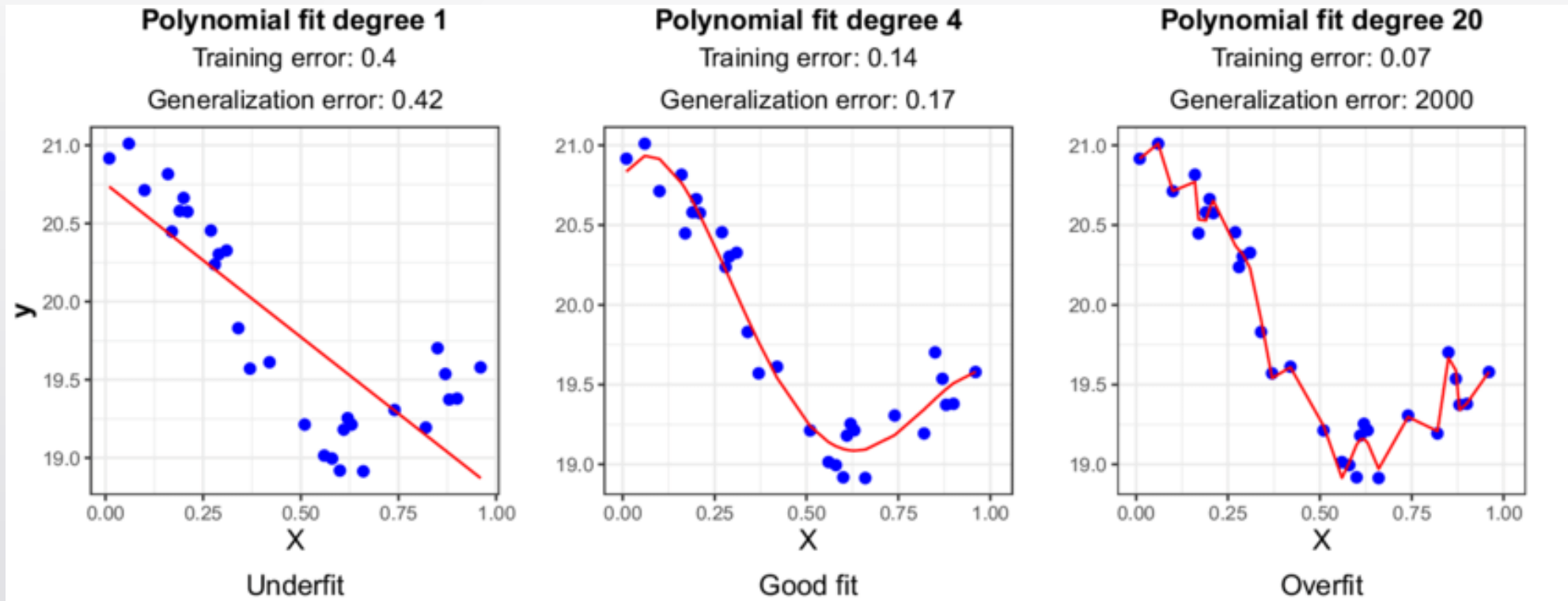
$$y = b_0 + b_1 x_1 + \underline{b_2 x_1^2} + \dots + b_n x_1^n$$

多項式回歸 Polynomial Regression



<https://towardsdatascience.com/polynomial-regression-an-alternative-for-neural-networks-c4bd30fa6cf6>

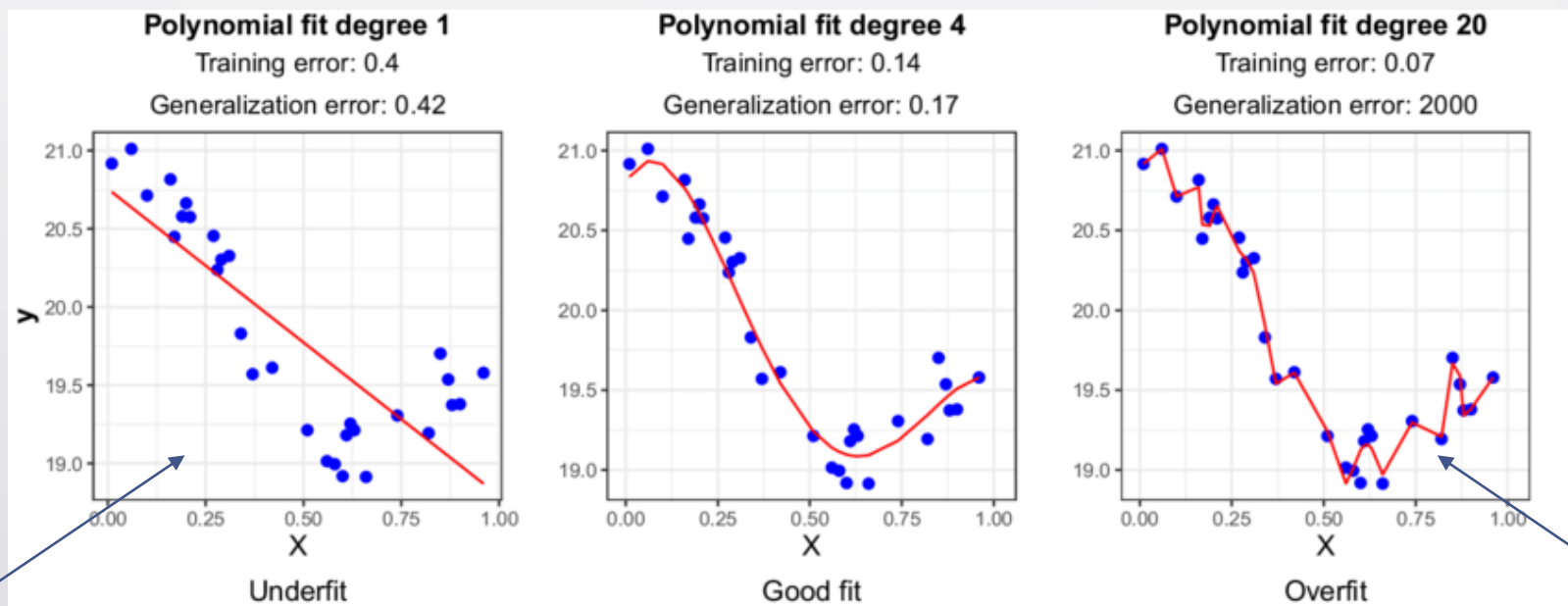
過学習 Overfitting



Badillo et al, 2020

良い回帰モデルとは？ What is good regression model?

観測されたデータはノイズを含む Observed data contains random noise



単純すぎてはダメ
Should not be too simplistic

複雑すぎてはダメ
Should not be too complex



赤池情報量基準 Akaike Information Criterion (AIC)

$$AIC = -2\ln(L) + 2k$$

L: 最大尤度 Maximum Likelihood

モデルのもとで、実際に観測されたデータが得られるもっともらしさ
(\approx モデルのデータへの当てはまりの良さ)

Likelihood that actually-observed dataset is obtained under the model
(\approx Goodness of fit of the model to the data)

k: モデルのパラメータ数

Number of parameters in the model

赤池情報量規準 Akaike Information Criterion (AIC)

$$AIC = -2\ln(L) + 2k$$

パラメータの数が少ない

The smaller the number of parameters

データへの当てはまりがいい

The better the model fits to the data



AICが小さくなる

The smaller AIC gets



変数選択 Feature Selection

$$AIC = -2\ln(L) + 2k$$

良い回帰モデルは、AICが小さい

AIC of good regression model is small

AICを基準としてモデルに含める予測変数を選択する

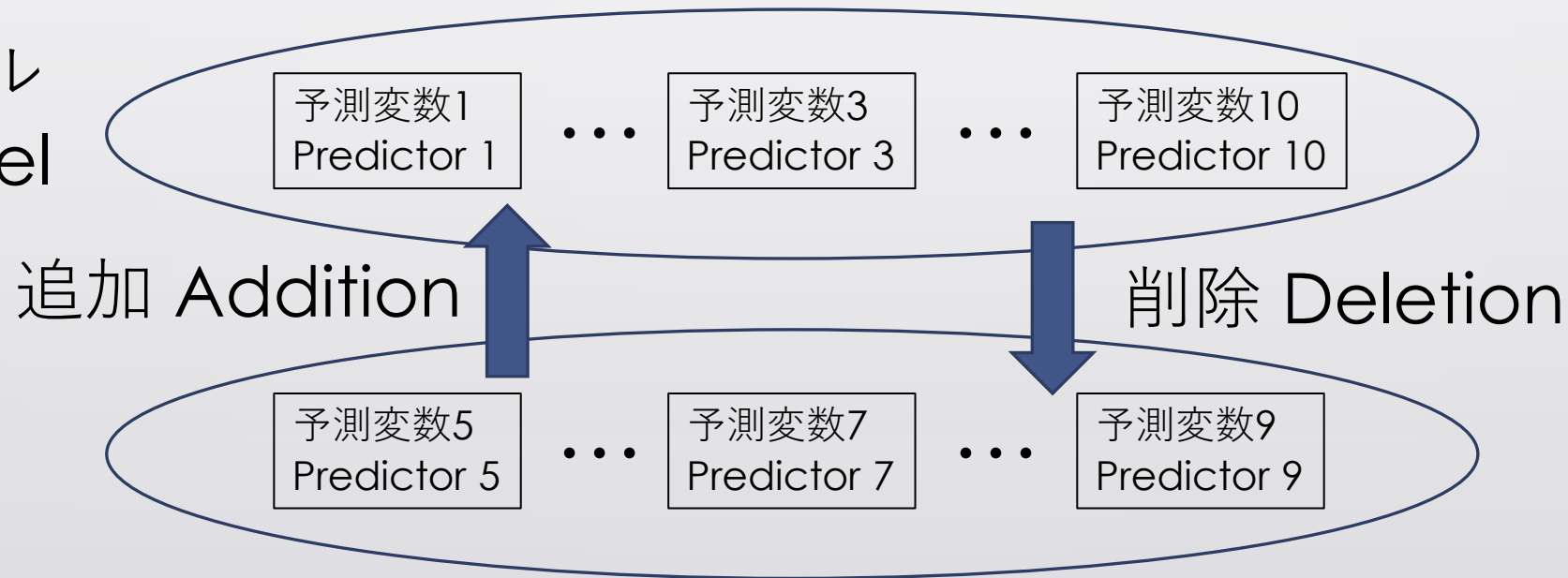
Select predictor variables to be included in the regression model by using AIC as criteria

ステップワイズ法 Stepwise feature Selection

予測変数を追加したり、削除したりしながら、AICを最小にする予測変数の組み合わせを探す

Find the best combination of predictors that minimizes AIC by adding and deleting variables from the model

モデル
Model



ベイズ情報量規準

Bayesian Information Criterion (BIC)

$$AIC = -2\ln(L) + 2k$$

$$BIC = -2\ln(L) + 2k\ln(N)$$

N : サンプルサイズ Sample Size

正則化回帰モデル Regularized Regression Model

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

残差二乗和 Residual Sum of Squares (RSS)

$$RSS = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}')$$

正則化回帰モデルでは、RSSに正則化項を加えることで、過学習を抑制する
Regularized regression model suppresses overfitting by adding regularization term to RSS

$$L = RSS + \text{Regularization term} = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}') + \text{Regularization term}$$

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0$$

リッジ回帰 Ridge Regression

$\|\boldsymbol{\beta}\|_2 \leq t$ という制約のもとでRSSを最小化する $\boldsymbol{\beta}$ を見つける

Find $\boldsymbol{\beta}$ that minimizes RSS under the constraint $\|\boldsymbol{\beta}\|_2 \leq t$

$$L = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}') + \alpha \|\boldsymbol{\beta}\|_2$$

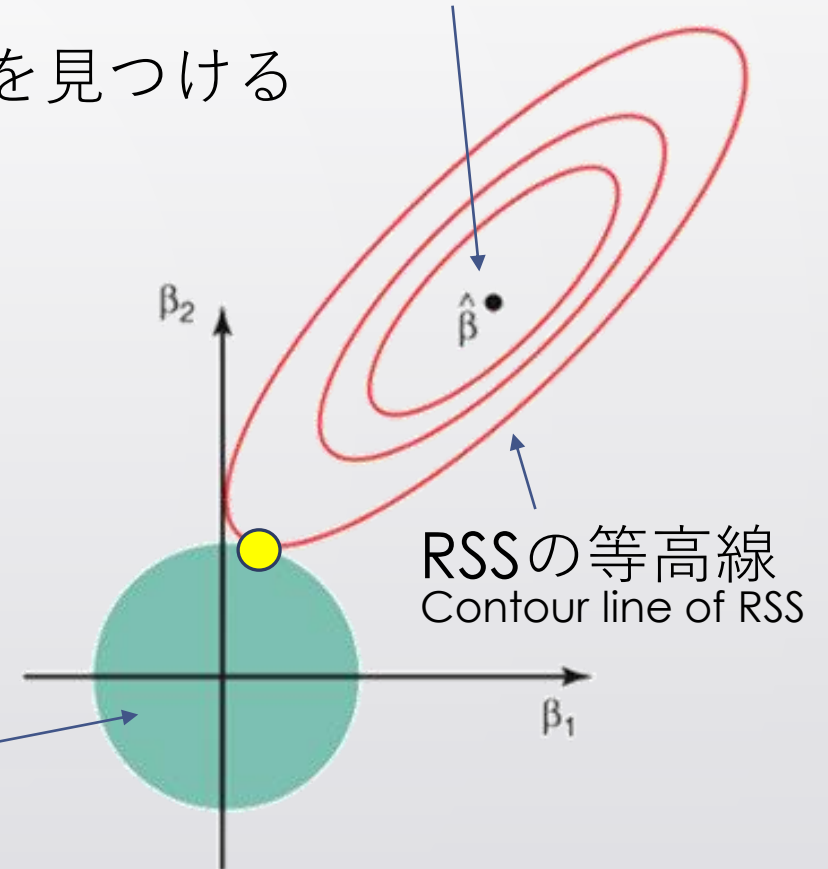
L2ノルム
L2 Norm

回帰係数の絶対値が小さくなる

Absolute value of regression coefficients tend to get smaller

この領域内の $\boldsymbol{\beta}$ は制約を満たす
 $\boldsymbol{\beta}$ inside this region satisfies the constraint

RSSを最小にする $\boldsymbol{\beta}$
 $\boldsymbol{\beta}$ minimizing RSS



Lasso回帰 Lasso Regression

$\|\boldsymbol{\beta}\|_1 \leq t$ という制約のもとでRSSを最小化する $\boldsymbol{\beta}$ を見つける

Find $\boldsymbol{\beta}$ that minimizes RSS under the constraint $\|\boldsymbol{\beta}\|_1 \leq t$

$$L = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}') + \alpha \|\boldsymbol{\beta}\|_1$$

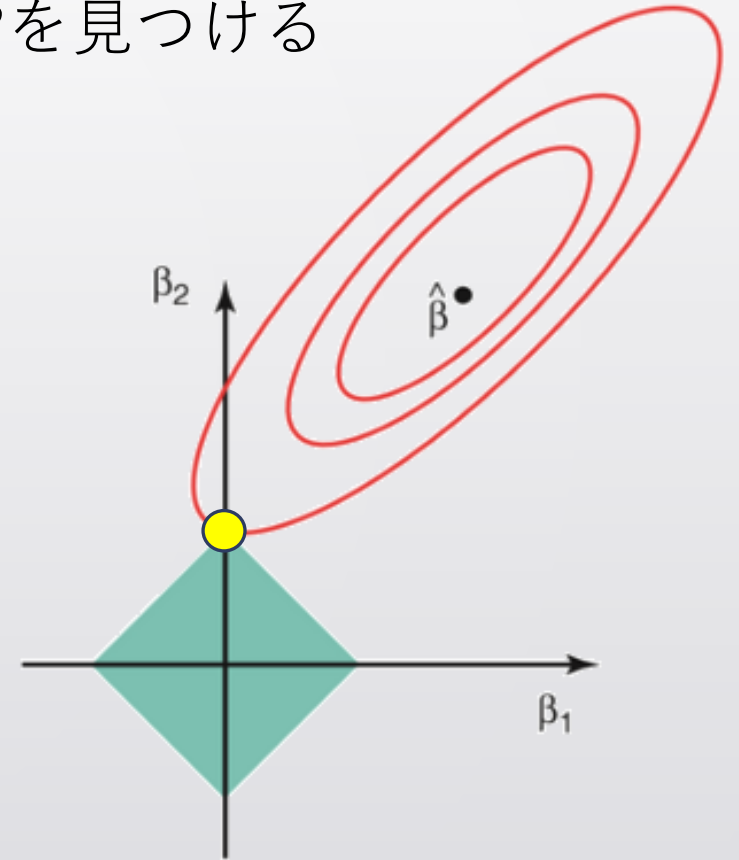
$$\|\boldsymbol{\beta}\|_1 = |\beta_1| + |\beta_2| + \cdots |\beta_n|$$

L1 ノルム
L1 Norm

影響が小さな変数の回帰係数は0になる

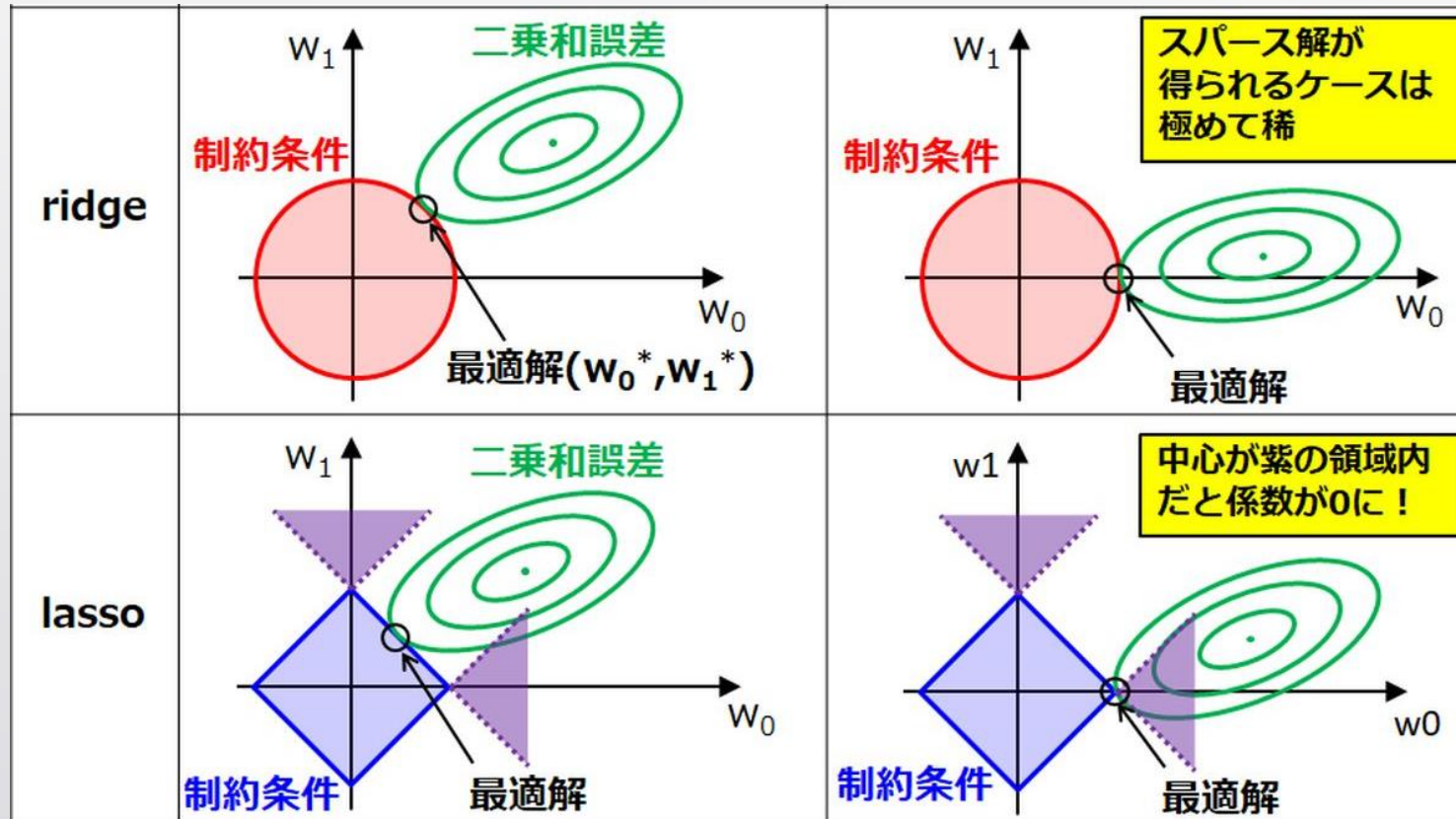
Regression coefficient of variable with little influence becomes zero

➡ スパースな（疎な）回帰モデル
Sparse regression model



リッジ回帰とLasso回帰

Ridge Regression and Lasso Regression



Lasso回帰では、回帰係数 β の真の値が紫色の領域にあれば、推定した回帰係数が0になる

In Lasso regression, estimated regression coefficient becomes zero when true values of β falls within purple-colored regions

<https://yuyumoyuyu.com/2021/01/03/regularized-leastsquares/>

Elastic Net

リッジ回帰の特徴とLasso回帰の特徴を組み合わせたもの

Elastic net has characteristics of both Ridge regression and Lasso regression

$$L = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}') + \alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \|\boldsymbol{\beta}\|_2$$

Lasso回帰の正則化項

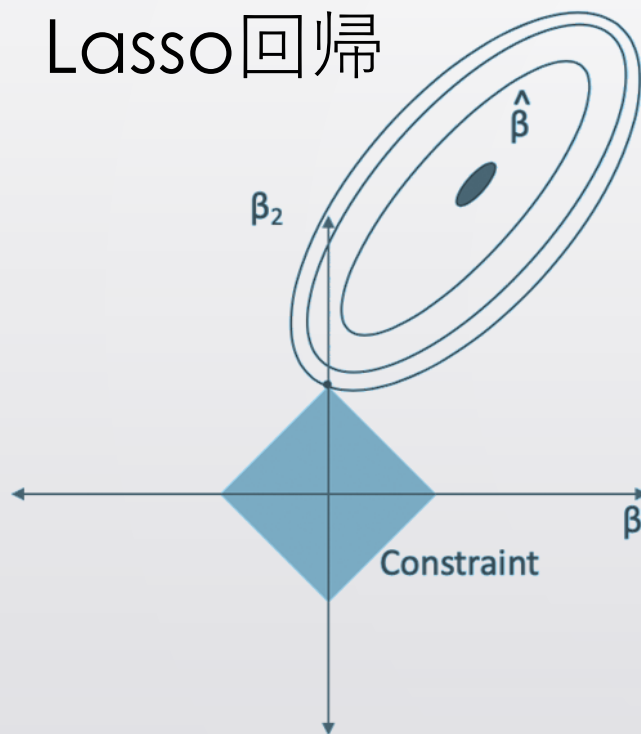
Regularization term of
Lasso regression

リッジ回帰の正則化項

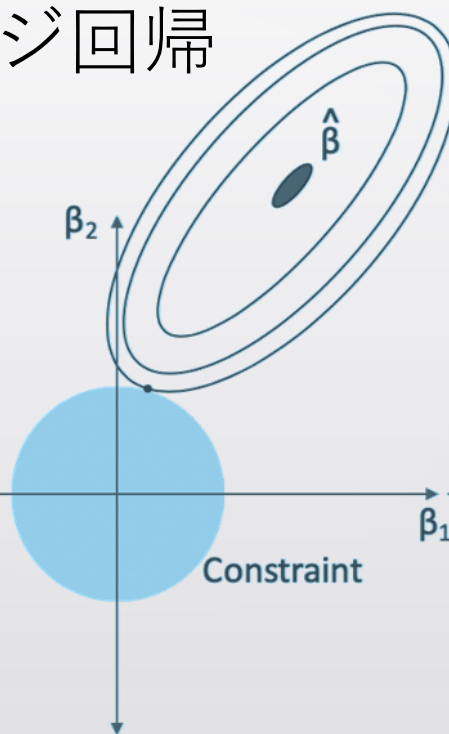
Regularization term of
Ridge regression

Elastic Net

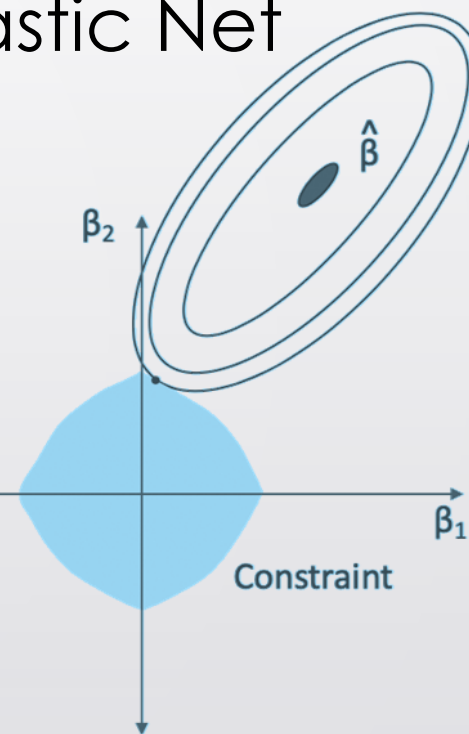
Lasso回帰



リッジ回帰

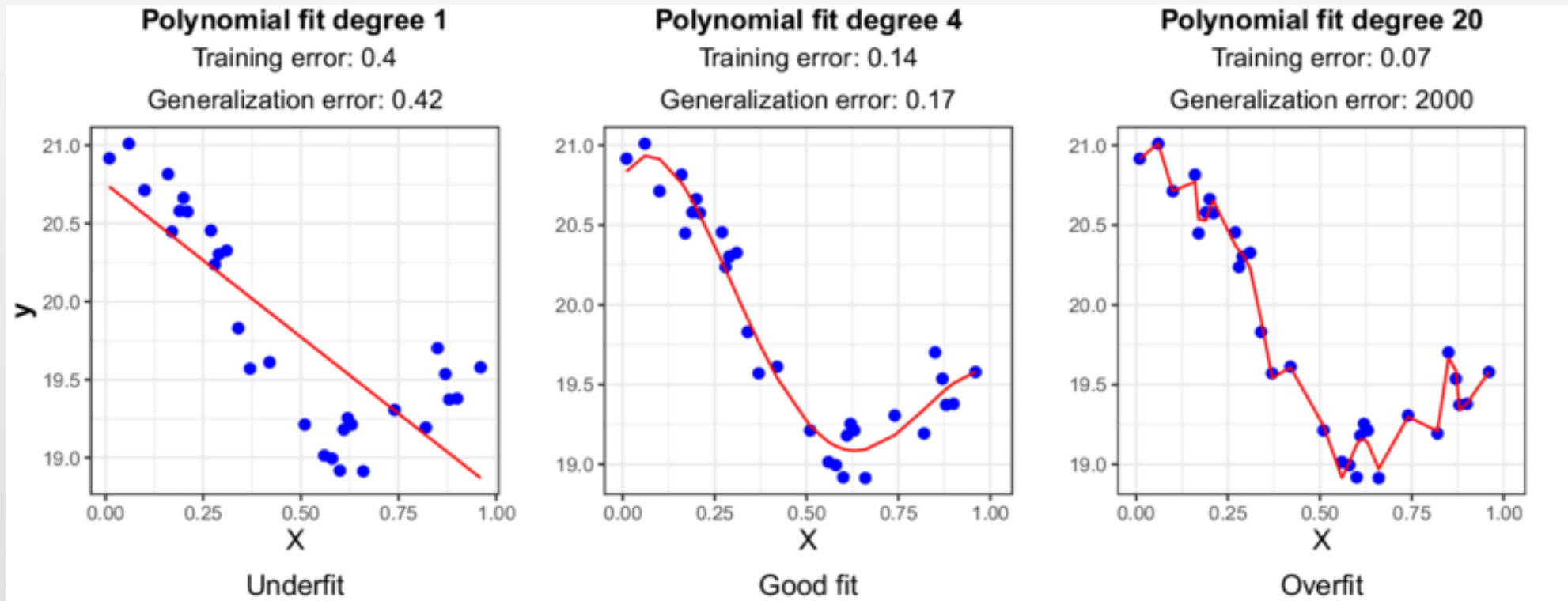


Elastic Net



<https://www.datasklr.com/extensions-of-ols-regression/regularization-and-shrinkage-ridge-lasso-and-elastic-net-regression>

過学習 Overfitting



Badillo et al, 2020



交差検証 Cross validation

1. データを学習(訓練)データとテストデータに分割する
Splitting data into training and test data
2. 学習(訓練)データを使って回帰モデルを作る
Create regression model based on training data
3. 回帰モデルの予測性能をテストデータで検証する
Evaluate prediction performance of regression model using test data

予測性能の指標 Indicator of Prediction Performance

正解値と予測値の相関係数

Correlational coefficient between predicted and actual values

根平均二乗誤差 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N (y_n - y'_n)^2} = \sqrt{\frac{1}{N} (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y})}$$

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

β

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

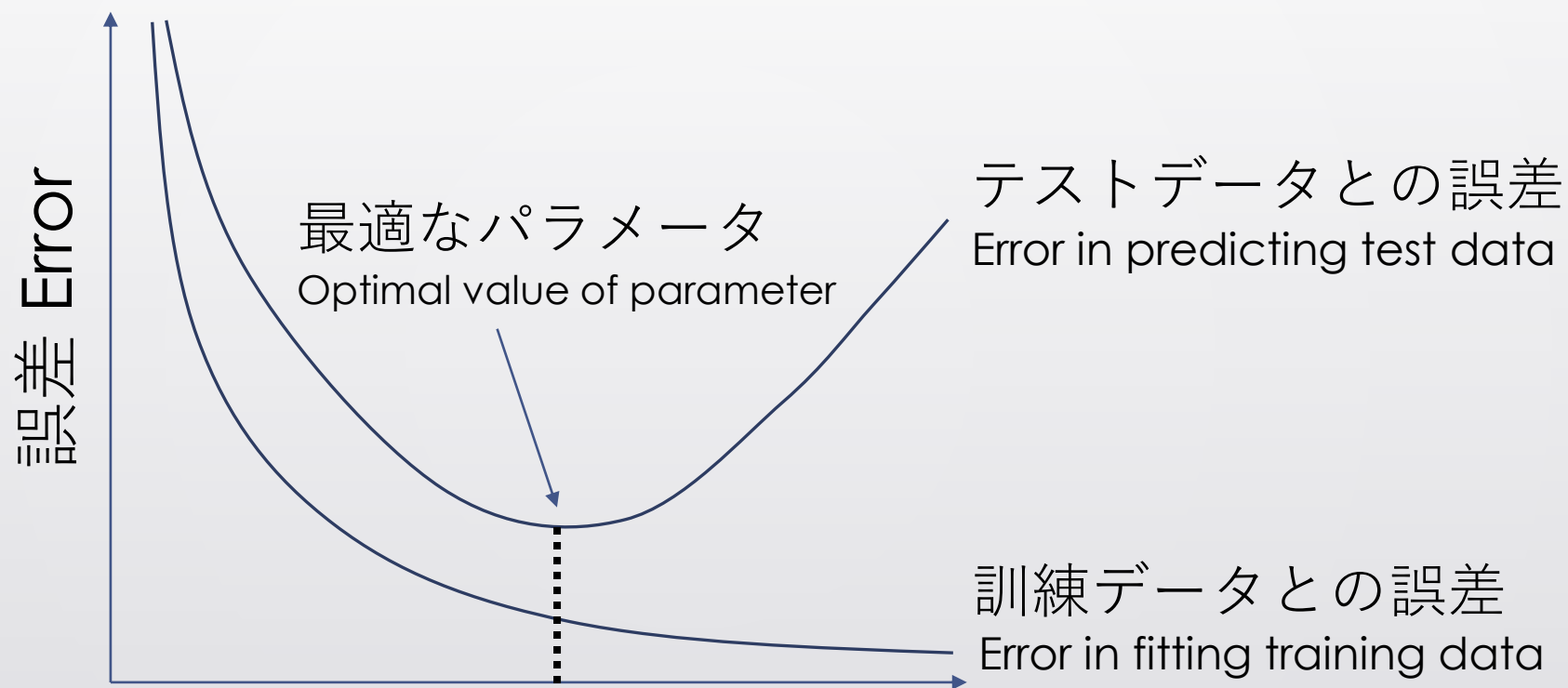
$\text{Log}(L)$ を最大化

Maximize $\text{Log}(L)$

RSS を最小化 Minimize RSS

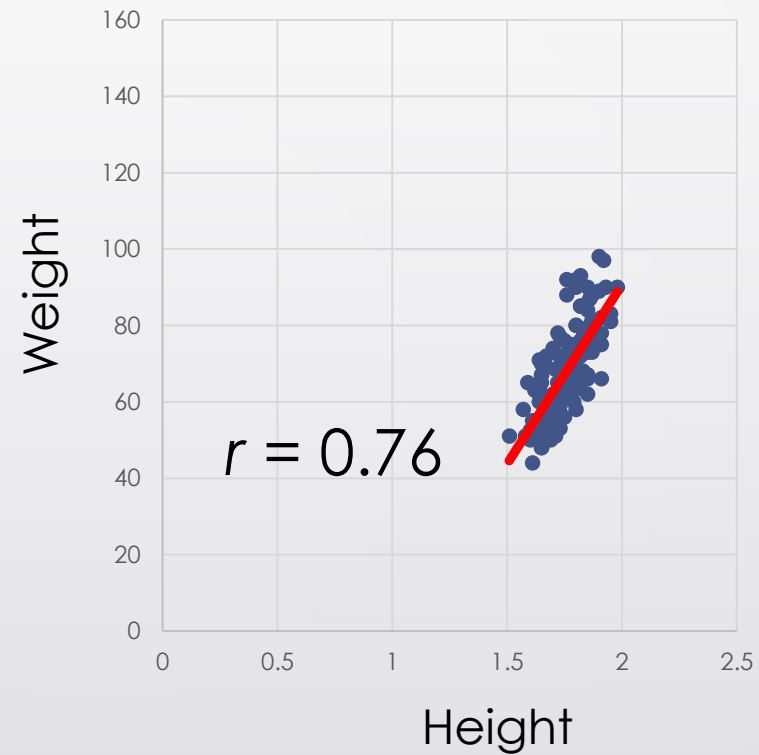
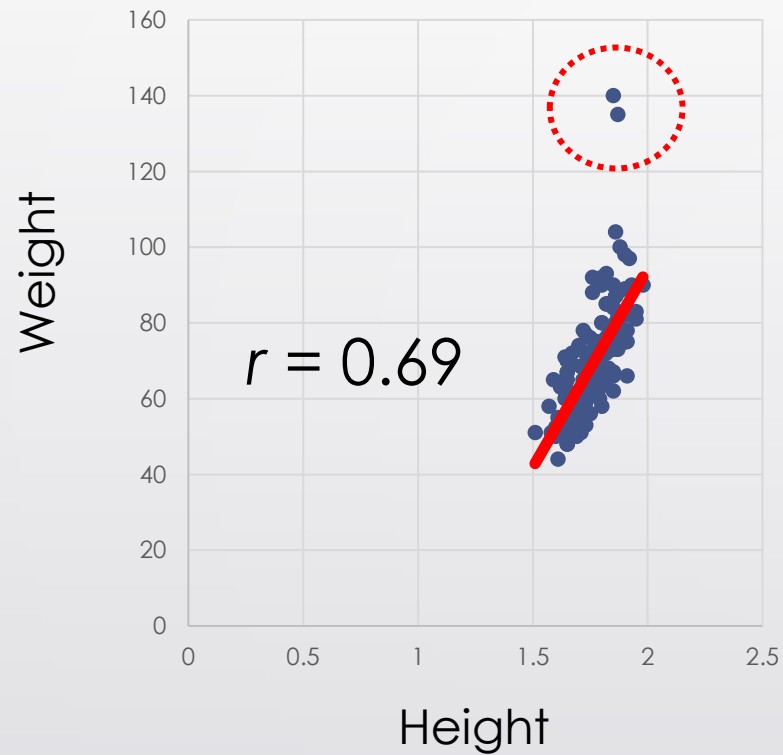
$$\sigma = \sqrt{\frac{1}{N} \sum_1^N (y_n - y'_n)^2}$$

過学習のサイン Signs of Overfitting



調整できるパラメータ : 選択された変数、ハイパーパラメータ 等
Adjustable Parameter Selected variable, hyperparameter etc

外れ値 Outlier

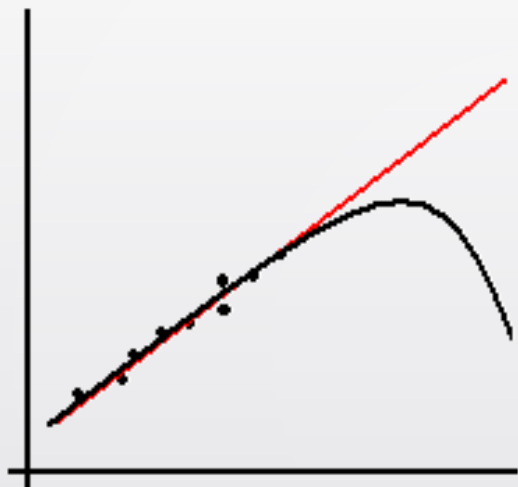


結果に影響を与える可能性がある外れ値の存在をチェックする必要がある
Data set should be checked for the existence of outliers that can influence the results

外挿の危険性 Peril of Extrapolation

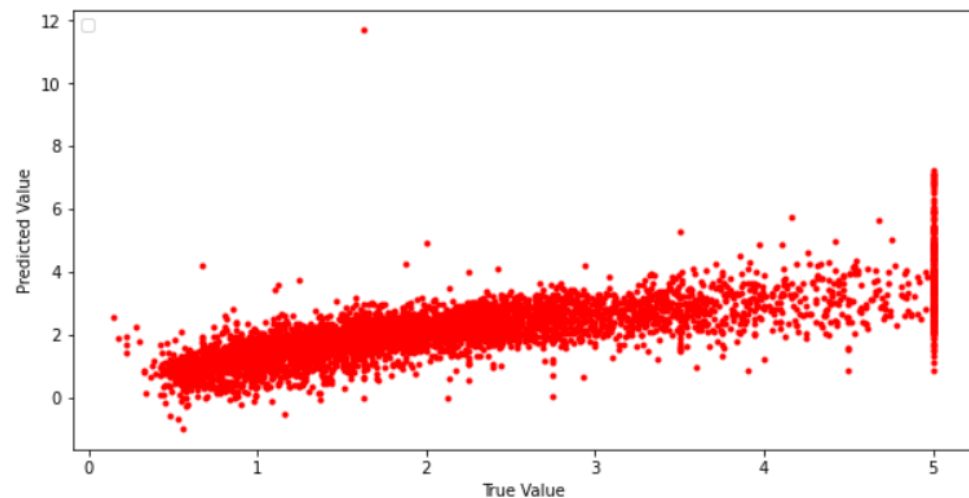


(a) Beware of extrapolation past the end of the data.



(b) Extrapolated line is red, actual response curve is black.

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



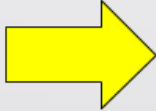
カテゴリー変数の扱い方

Treatment of Categorical variable

ダミー変数 Dummy Variable 男 \Rightarrow 1, 女 \Rightarrow 2

One-hot Encoding

Color	
Red	
Red	
Yellow	
Green	
Yellow	



Red	Yellow
1	0
1	0
0	1
0	0

<https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook>