電子情報通信実験IV－１
多変量解析と可視化

担当　村松、清水

１．目的

- モノのインターネット
  (IoT: Internet of Things)が普及
  - 多種多様なデータが取得されるようになった
- 本テーマでは，
  - 大量のデータから有益な情報を得るために必要な多変量解析，特に回帰分析の基礎理論を学ぶ
  - 実習を通して現象のモデル化と可視化の技法を身に付ける

２．解説

- 回帰分析(regression analysis)
  - ある変数yと変数xの間を

$$y = u(x)$$

のように関係づける未知の関数$u(\cdot)$を仮定
  - 変数x：　独立変数もしくは説明変数とよぶ
  - 変数y：　従属変数もしくは目的変数とよぶ

このような未知の関数を観測可能なデータから推論する統計的手法

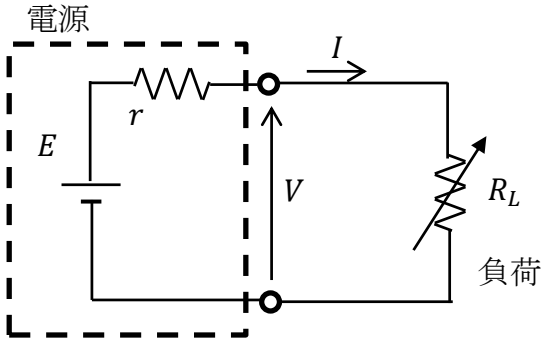
本テーマで扱う回帰モデル

回帰モデル	非線形	多変量	モデル選択	正則化
線形単回帰	×	×	不要	不要
線形重回帰	×	○	可能	可能
多項式単回帰	○	×	可能	可能
多項式重回帰	○	○	可能	可能
RBF回帰	○	○	可能	可能

２．１　線形回帰モデル

- ２変数間の関係

$$y = u(x) = \beta_0 + \beta_1 x$$



- ２組の観測データが得られれば

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} 1 & -I_1 \\ 1 & -I_2 \end{pmatrix} \begin{pmatrix} E \\ r \end{pmatrix}$$

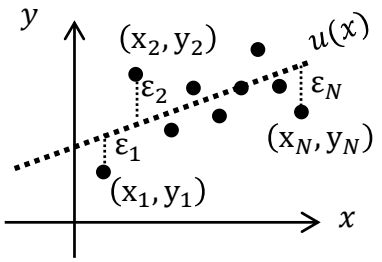
- 切片β_0と傾きβ_1を求められる

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \qquad \begin{pmatrix} E \\ r \end{pmatrix} = \begin{pmatrix} 1 & -I_1 \\ 1 & -I_2 \end{pmatrix}^{-1} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

２．１　線形回帰モデル

- 観測誤差が含まれる場合

$$y_i = u(x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



- 最小2乗法の問題設定

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{N} \varepsilon_i^2 = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

- 最小2乗法の解

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix}$$

２．１　線形回帰モデル

- 多変数間の関係

$$y = u(x_1, x_2, \cdots, x_p)$$

- 線形重回帰モデル

$$y = u(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \beta_0 + \sum_{k=1}^{p} \beta_k x_k$$

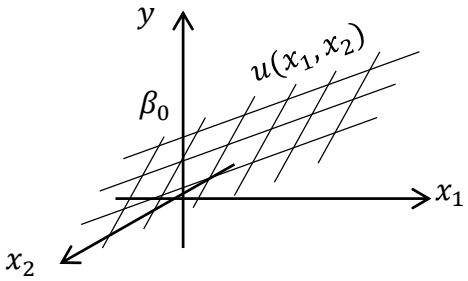
- 最小2乗法の解

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^+\mathbf{y}$$

ムーア・ペンローズの擬似逆行列

ただし、

$$\mathbf{X} = \begin{pmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1p} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{x}_{N1} & \mathrm{x}_{N2} & \cdots & \mathrm{x}_{Np} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_N \end{pmatrix}$$



$(\mathrm{x}_{11}, \mathrm{x}_{12}, \cdots, \mathrm{x}_{1p}, \mathrm{y}_1)$, $(\mathrm{x}_{21}, \mathrm{x}_{12}, \cdots, \mathrm{x}_{2p}, \mathrm{y}_2)$, ⋯ , $(\mathrm{x}_{N1}, \mathrm{x}_{N2}, \cdots, \mathrm{x}_{Np}, \mathrm{y}_N)$は観測データ$N$組

２．２　非線形回帰モデル

- 基底展開法の回帰モデル

$$u(\boldsymbol{x};\boldsymbol{w}) = w_0 + \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}), \boldsymbol{x} = \left(x_1, x_2, \cdots, x_p\right)^T$$

- 基底関数の例
  - 多項式単回帰モデル

$$\phi_j(x) = x^j, j = 1,2,\cdots,m$$

  - 多項式重回帰モデル（２変数）

$$\phi_j(x) = x_1^{q(j)} x_2^{r(j)}, \quad 1 \le q(j) + r(j) \le p, \quad j = 1,2,\cdots,m = \frac{(p+1)(p+2)}{2}$$

  - 動径基底関数(RBF)回帰モデル

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{\left\|\boldsymbol{x} - \boldsymbol{\mu}_j\right\|_2^2}{2h_j^2}\right), \quad j = 1,2,\cdots,m$$

２．２　非線形回帰モデル

- RFB回帰モデルの２段階推定法
  1. パラメータ
     $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m, h_1^2, h_2^2, \cdots, h_m^2\}$
     について，説明変数$\boldsymbol{x}$の観測データに対する
     クラスタリング手法で事前推定
     - クラスタリング手法：k-平均 (k-means)法など
     - h_j^2：関数の広がりを表すパラメータ
  2. 重みパラメータ$\boldsymbol{w}$を最小２乗法で推定

２．２　非線形回帰モデル

- 最小2乗法の解

$$\hat{\mathbf{w}} = \begin{pmatrix} \hat{w}_0 \\ \hat{w}_1 \\ \vdots \\ \hat{w}_m \end{pmatrix} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{\Phi}^T\mathbf{y} = \boldsymbol{\Phi}^+\mathbf{y}$$

ただし、

計画行列

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_m(\mathbf{x}_1) \\ 1 & \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_m(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathbf{x}_N) & \phi(\mathbf{x}_N) & \cdots & \phi_m(\mathbf{x}_N) \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

$(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, … , $(\mathbf{x}_N, y_N)$は観測データN組

２．３　モデル評価基準と正則化

- 情報量基準(AIC)
  観測データに対する当てはまりの良さを最大対数尤度で測り，自由パラメータ数で複雑さのペナルティを課す
  - 線形重回帰モデル／RBF回帰モデル
    $$\mathrm{AIC} = -2\left\{\log L\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) - (p+2)\right\}$$
    $$= N\log(2\pi\hat{\sigma}^2) + N + 2(p+2)$$
  - 多項式単回帰モデル
    $$\mathrm{AIC} = -2\{\log L(\hat{\mathbf{w}}, \hat{\sigma}^2) - (m+2)\}$$
    $$= N\log(2\pi\hat{\sigma}^2) + N + 2(m+2)$$
  - $\hat{\sigma}^2$は誤差分散の最尤推定値
- AIC最小化法
  - 過剰適合を回避するために，小さなAICを与えることをモデル採用の基準とする方法

２．３　モデル評価基準と正則化

- 正則化最小2乗法
  行列$(\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}$が計算不能となり最小2乗法が適用できない問題を回避

$$\hat{\mathbf{w}} = \arg\min_{\boldsymbol{w}} \left(\frac{1}{2} S(\boldsymbol{w}) + \lambda R(\boldsymbol{w}) \right)$$
$$S(\boldsymbol{w}) \coloneqq \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{w}\|_2^2 = (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{w})^T (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{w})$$

- 過剰適合を抑制する手法が正則化法の例
  - リッジ回帰

$$R(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2}\sum_{j=0}^{m} |w_j|^2$$

  - ラッソ回帰

$$R(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_{j=0}^{m} |w_j|$$

３．予習内容

- テキスト表1に掲載したデータに関して以下の課題に取り組むこと
  1. $\{(I_1, V_1), (I_2, V_2), \cdots, (I_8, V_8)\}$から任意の2組のデータを抽出し，逆行列を利用して電源電圧Eと内部抵抗rを推定せよ
  2. $\{(I_1, V_1), (I_2, V_2), \cdots, (I_8, V_8)\}$から全ての8組のデータを利用し，ムーア・ペンローズの擬似逆行列を利用して電源電圧Eと内部抵抗rを推定せよ

4. 実施内容

1. （個人ワーク） サンプルデータを配信する。多項式回帰分析を適用し回帰式を与えよ。
   a. 線形単回帰モデルにより回帰式を与えよ。
   b. 多項式回帰モデルの次数pをAIC最小化法により決定し回帰式を与えよ。

2. （個人ワーク） サンプルデータを配信する。~~k-平均法~~基底展開法を利用した~~RBF~~重回帰分析~~の2段階推定法~~を適用し回帰式を与えよ。
   a. 線形重回帰モデルにより回帰式を与えよ。
   b. ~~RBF~~基底展開重回帰モデルの~~次数p~~パラメータ数をAIC最小化法により決定し回帰式を与えよ。

RBF重回帰に限らない

3. （グループワーク） オープンデータ，既出論文からのデータ，もしくは研究室や個人的に取得した実測データのいずれかを用意し，任意の基底展開法に対してラッソ正則化を適用し~~単~~回帰分析を実施せよ。
   a. 線形~~重~~回帰モデル（$\lambda = 0$）により回帰式を与えよ。
   b. 基底関数の個数mと正則化パラメータλを適当に設定し回帰式を与えよ。

単回帰でも重回帰でもよい

実習項目1のサンプルデータ

- practice01_01.txtについて（清水研提供）

1. 本データは，平坦なガラス板の上に薄膜を約400nm堆積した物を，測定機により，ガラス板と薄膜部分の境目の段差を測定した結果
2. 横軸（一列目XDATA）の単位はmm，縦軸（二列目YDATA）の単位はnm
3. 本来であれば，直線的なステップ形状のグラフとなるはずだが，振動などの影響により，誤差が含まれている

実習項目2のサンプルデータ

- practice01_02.csvについて（清水研提供）

1. 本データは，薄膜を作製するためのスパッタ装置のアルゴンガスの放電特性を圧力計，電圧計，電流計を用いて測定した結果
2. 一行目の単位はmA，一列目の単位はTorr，それ以外の単位はV
3. 本来であれば，電流をI，電圧をV，圧力をP，a，bを定数とすると，関係

$$I = \frac{a}{1 - bPV^2}$$

を満たすグラフとなるが，測定機の接続や測定機の不安定性などの影響により，誤差が含まれている

サンプルコードとデータの配信

- サンプルコードについて
  - 以下のGitHubサイトにて公開
    https://github.com/msiplab/EicEngLabIV
- サンプルデータについて
  - 人工生成データを上記GitHubで公開
    （dataフォルダ内）
  - 実習用サンプルデータは7月３日正午に配信
    （学務情報システムレポート課題添付）

５．報告事項

1. （個人ワーク） 線形単回帰分析と多項式単回帰分析の結果を二次元散布図に重ねてプロットし，比較検討を行え。
2. （個人ワーク） 線形重回帰分析／~~RBF~~基底展開重回帰分析の結果を三次元散布図に重ねてプロットし，比較検討を行え。
3. （グループワーク） ラッソ正則化を利用した基底展開法の実習について以下の内容を報告せよ。
   a. 利用したデータに関して報告せよ。
   b. ~~二次元~~散布図と回帰曲線（面）を重ねてプロットし，比較検討を行え。

報告の際の注意

- 回帰分析について
  - 利用した回帰分析法を明記すること
    - モデルの種類
    - モデルパラメータ
    - 基底展開法の場合、採用したパラメータ数
    - 正則加法の場合、正則化パラメータ
- グラフについて
  - 内容に合わせたタイトル
  - 各軸のラベル（単位を忘れずに）
  - 必要に応じて凡例（legend）

（補足）グループワークについて

- 大学公式Zoom（村松か指導教員に相談）
  - 利点：全員が使い慣れている
  - 欠点：ホスト（教員）が必要で任意の時間の開始できない、チャット記録へのアクセスが不便
- 大学公式Microsoft Teams（村松か指導教員に相談）
  - 利点：任意の時間に開始できる、チャット記録へのアクセスが容易、学務情報アカウントで利用できる
  - 欠点：全員が使い慣れているとは限らない
- 大学公式Google Meet＋Slack（ダイレクトメッセージ）
  - 利点：任意の時間に開始できる、チャット記録へのアクセスが容易、Google Meetは学務情報アカウントで利用できる
  - 欠点：全員が使い慣れているとは限らない