医療データ科学実習

R による解析のデモ

はじめに

- Rを用いることで、医療データを活用した多様な統計解析が可能

- 以下の解析を例示
  - ロジスティック回帰（2値分類）
  - 平滑化曲線の推定（非線形な関係性の推測）

デモ1：ロジスティック回帰による糖尿病予測

ロジスティック回帰モデル

- 目的変数 Y が2値（例：0/1、Yes/No）の場合に適用できる統計モデル
- 説明変数 X（例：年齢、血糖値、BMI など）を用いてある事象Y=1（例：糖尿病陽性の診断結果）の確率 p を説明
- ロジット変換により、0から1の確率を実数全体に対応

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

- ロジスティック回帰モデル

$$\mathrm{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

データセット：PimaIndiansDiabetes

- 糖尿病データセット（mlbenchパッケージ）
- National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)によって収集されたピマ・インディアンの血を引く21歳以上の女性のデータ

変数名	説明	型
pregnant	妊娠回数	数値
glucose	血糖値（経口耐糖試験結果）	数値
pressure	拡張期血圧（mm Hg）	数値
triceps	上腕三頭筋皮膚厚（mm）	数値
insulin	2時間後血清インスリン濃度（μU/ml）	数値
mass	BMI（体重kg/身長m^2）	数値
pedigree	糖尿病家系機能（遺伝的要因の指標）	数値
age	年齢（年）	数値
diabetes	糖尿病の有無（pos/neg）	因子

データの確認

```
data(PimaIndiansDiabetes2)
str(PimaIndiansDiabetes2)
```

```
'data.frame':  768 obs. of  9 variables:
 $ pregnant: num  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose : num  148 85 183 89 137 116 78 115 197 125 ...
 $ pressure: num  72 66 64 66 40 74 50 NA 70 96 ...
 $ triceps : num  35 29 NA 23 35 NA 32 NA 45 NA ...
 $ insulin : num  NA NA NA 94 168 NA 88 NA 543 NA ...
 $ mass    : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 NA ...
 $ pedigree: num  0.627 0.351 0.672 0.167 2.288 ...
 $ age     : num  50 31 32 21 33 30 26 29 53 54 ...
 $ diabetes: Factor w/ 2 levels "neg","pos": 2 1 2 1 2 1 2 1 2 2 ...
```

```
summary(PimaIndiansDiabetes2)
```

```
    pregnant         glucose         pressure         triceps         insulin           mass          pedigree           age         diabetes
 Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00   Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00   neg:500
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00   1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00   pos:268
 Median : 3.000   Median :117.0   Median : 72.00   Median :29.00   Median :125.00   Median :32.30   Median :0.3725   Median :29.00
 Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15   Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
 3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00   3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
                  NA's   :5       NA's   :35       NA's   :227     NA's   :374      NA's   :11
```

- 欠測の多い指標は今回の解析では使用しない

モデルの推定

```
fitted_model <- glm(diabetes ~ pregnant + glucose + mass
+ pedigree + age, data = PimaIndiansDiabetes2, family =
binomial)
summary(fitted_model)
```

```
Call:
glm(formula = diabetes ~ pregnant + glucose + mass + pedigree +
    age, family = binomial, data = PimaIndiansDiabetes2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8093  -0.7287  -0.4011   0.7275   2.4449

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.322789   0.737279 -12.645  < 2e-16 ***
pregnant     0.115058   0.032341   3.558 0.000374 ***
glucose      0.035941   0.003555  10.110  < 2e-16 ***
mass         0.087529   0.014722   5.945 2.76e-09 ***
pedigree     0.920583   0.300832   3.060 0.002212 **
age          0.011366   0.009315   1.220 0.222405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

係数推定値 $\hat{\beta}_1, \ldots, \hat{\beta}_p$

モデルの予測結果

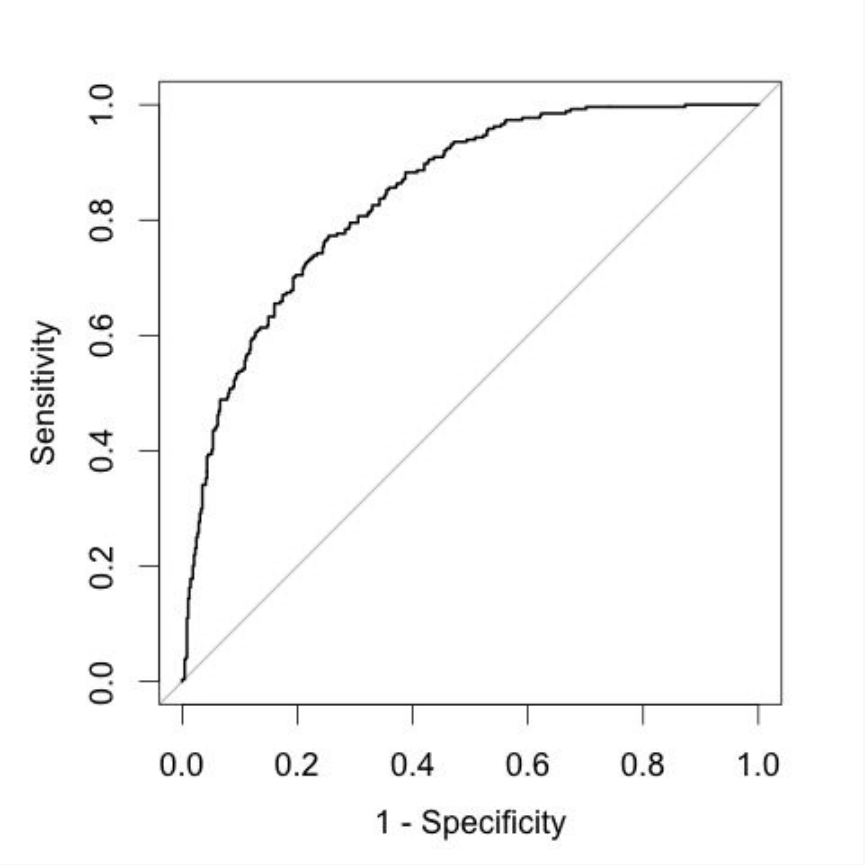
```
predicted <- ifelse(predict(fitted_model, type =
"response") > 0.5, "pos", "neg")
table(Predicted = predicted, Actual =
fitted_model$model$diabetes)
```

```
          Actual
Predicted neg pos
      neg 431 114
      pos  57 150
```

- 確率の予測値が0.5以上を陽性と予測した場合の
  今回のデータでの予測精度を示す表
- 感度（Sensitivity），真陽性率（True Positive Rate，TPR）：
  実際に糖尿病の患者をモデルが正しく陽性と予測する割合
  150/(114+150) = 0.568
- 特異度（Specificity），真陰性率（True Negative Rate，TNR）：
  実際に糖尿病でない患者をモデルが正しく陰性と予測する割合
  431/(431+57) = 0.883

モデルの予測結果

```
library(pROC)
roc_curve <-
roc(PimaIndiansDiabetes2$diabetes,fitted(fitted_model))
plot(roc_curve)
auc(roc_curve)
```



- ROC曲線:
  モデルから陽性と予測する閾値を動かした場合の感度、特異度の関係を示す図
- AUC（Area Under Curve):
  ROC曲線の下の面積で、モデルの予測性能を評価する指標
  1が完全な予測
  0.5がランダムな予測 を表す
  今回のデータではAUC=0.843

医学データへのロジスティック回帰の適用

2つの異なる解析目的が考えられる

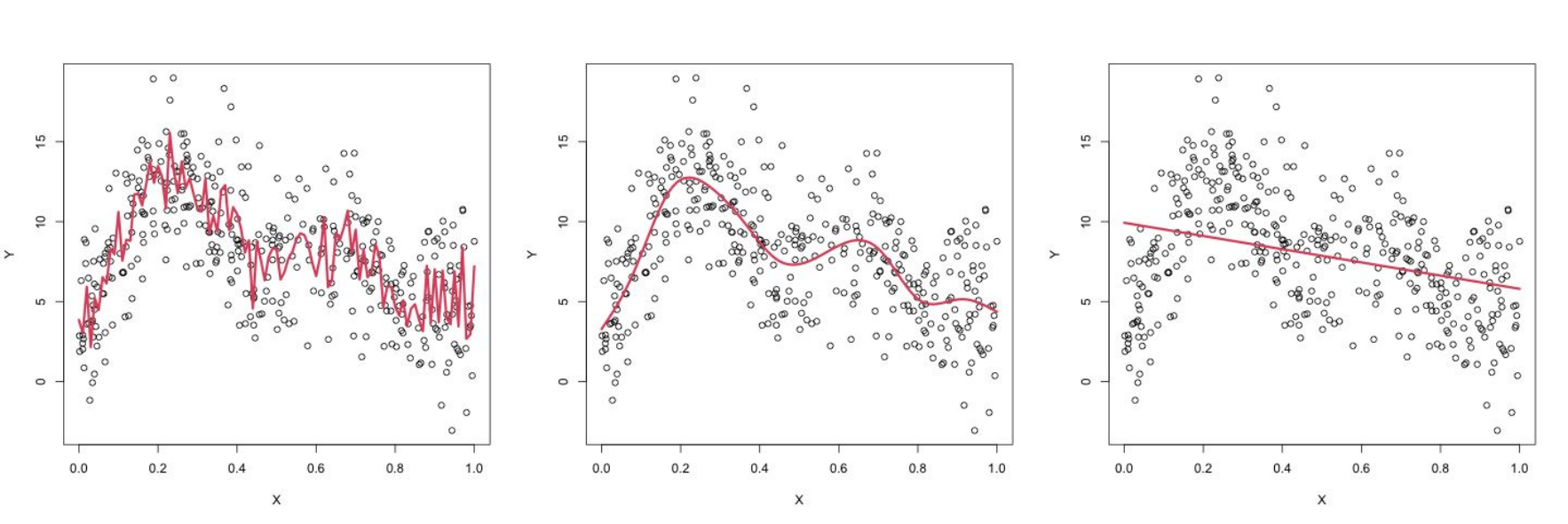
- 目的変数の予測:
  説明変数を用いた特定の目的変数の確率の予測に興味がある場合
  どの程度予測を正しくできるかの評価が重要

- 説明変数に関する推測:
  各説明変数が目的変数に与える影響の大きさに興味がある場合
  回帰係数（β）の推定とその解釈が重要
  説明変数が目的変数と関連しているか（β=0か）を判断する
  統計的仮説検定の枠組みにより統計的有意性を確認

デモ2：スプライン平滑化による BMIの非線形効果の推定

スプライン平滑化

- 説明変数と目的変数の非線形な関係を推定する方法
  - 例: BMIが高いとリスクであるだけでなく、低すぎてもリスク
- スプライン平滑化は、なめらかな曲線でその関係を表現する
- データにフィットしすぎず、全体的な傾向を捉えるよう調整



データセット：PimaIndiansDiabetes

- 糖尿病データセットにおける予測でのBMIの効果

```
Call:
glm(formula = diabetes ~ pregnant + glucose + mass + pedigree +
    age, family = binomial, data = PimaIndiansDiabetes2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8093  -0.7287  -0.4011   0.7275   2.4449

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.322789   0.737279 -12.645  < 2e-16 ***
pregnant     0.115058   0.032341   3.558 0.000374 ***
glucose      0.035941   0.003555  10.110  < 2e-16 ***
mass         0.087529   0.014722   5.945 2.76e-09 ***
pedigree     0.920583   0.300832   3.060 0.002212 **
age          0.011366   0.009315   1.220 0.222405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

BMIの効果の推定値

- BMIが1上昇すると対数オッズ比が0.08上昇

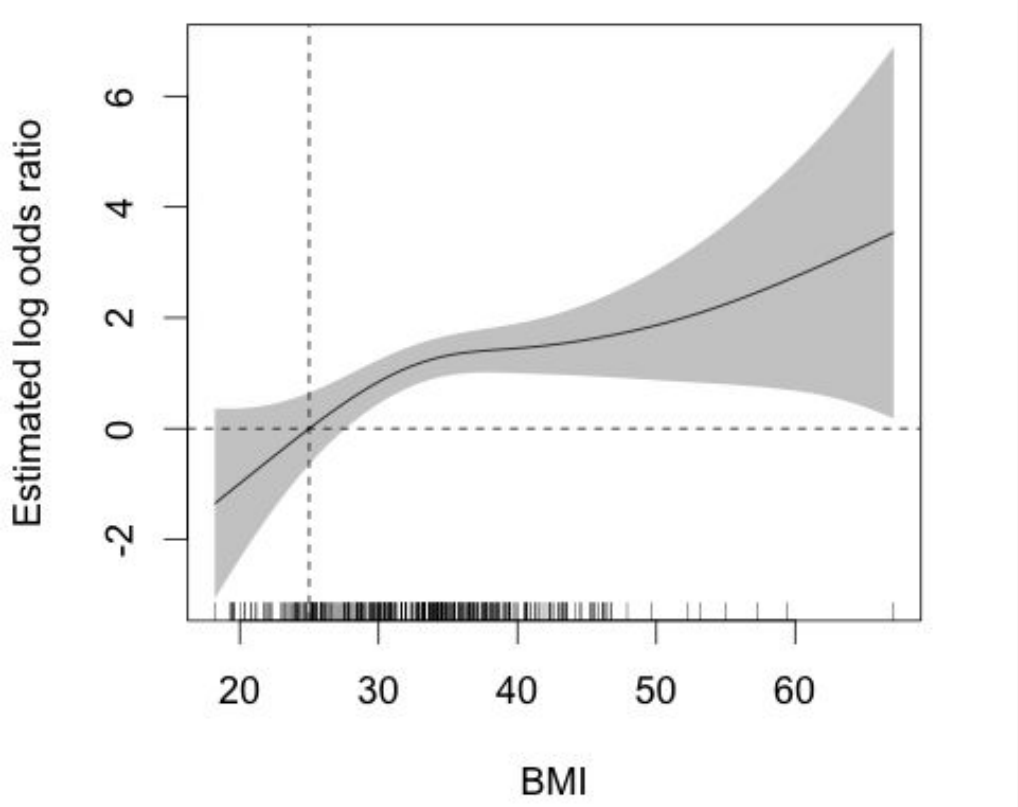
非線形の関係の推定

```
library(mgcv)
Diabetesdata <- na.omit(PimaIndiansDiabetes2)
model_gam <- gam(diabetes ~ pregnant + glucose +
s(mass,pc=25) + pedigree + age, family = binomial, data
= Diabetesdata)
```

- s(bmi)：BMIの効果をスプライン平滑化で推定
- family = binomial：
  目的変数が2値（糖尿病あり / なし）であることを指定
- gam() 関数:
  スムージング関数を含む回帰モデルを構築するための関数
- スプライン平滑化は splines パッケージなど他の方法でも実装可能

非線形の関係の推定

```
plot(model_gam, shade = TRUE)
```



- 横軸：BMI
- 縦軸：BMIが糖尿病の発生に与える非線形な影響（基準値BMI=25）
- グレーの帯：95%信頼区間（推定の不確かさ）

医学データでのスプライン曲線の使用例

- 用量・反応関係の推定
  薬物投与量と治療効果の関係を、直線ではなく曲線で評価
  最適な投与量や、有害作用の発生が増加する閾値の特定に使用

- バイオマーカーの閾値探索
  BMI・血糖値・血圧などと疾患リスクの非線形関係を視覚化
  リスクが急増するポイント（閾値）を探索的に明らかにする

- 経時的なリスク変化の解析
  年齢や時間経過とともに変化する疾患リスクを曲線で把握

- 仮説探索型データ解析（探索的データ解析）
  予測因子間の関係性について、新しい仮説や医学的洞察を得るために使用
  線形モデルに対する感度解析としての使用