第4回：ロジスティック回帰

Code

1. 前準備

これまで解説した線形回帰や非線形回帰は、目的変数が**量的**（numeric）の場合の統計モデルです。しかし、目的変数として**質的**（categorical）を考えたいこともあります。以下は、ある会社の訪問時刻・担当者年齢と契約結果の関係を記録したデータです。

Hide

```
# keiyakuデータセットの確認
dat <- read.csv("./data/keiyaku.csv",
                row.names = "会社",
                fileEncoding = "cp932")
head(dat, n = 5)
```

	訪問時刻 <chr>	担当者年齢 <int>	契約結果 <int>
A	午前	42	1
B	午前	22	0
C	午後	24	0
D	午前	36	0
E	午前	35	0

5 rows

このデータを用いて、訪問時刻・担当者年齢を用いて契約結果を説明するモデルを作ることを考えてみましょう。説明変数は訪問時刻と担当者年齢、目的変数は契約結果です。契約結果は「契約成立」の場合 １、「契約不成立」の場合 ０ が記録された2値の質的変数になっています。そこで今回は、このような2値の質的変数に対する統計モデルとして有名な**ロジスティック回帰**（logistic regression）を紹介します。

標本サイズと次元、変数の名前を確認しておきましょう。

Hide

```
# 標本サイズ、次元、変数の名前の確認
str(dat)
```

```
'data.frame':   22 obs. of  3 variables:
 $ 訪問時刻  : chr  "午前" "午前" "午後" "午前" ...
 $ 担当者年齢: int  42 22 24 36 35 42 36 39 36 42 ...
 $ 契約結果  : int  1 0 0 0 0 1 0 1 0 1 ...
```

2. ロジスティック回帰

この節では、ロジスティック回帰の概要を説明します。

2.1 ロジスティック回帰の概要

A. 課題設定

データに D 個の変数 $x_1, \cdots, x_D$ と y が記録されているとします。ここで変数 y は 0 か 1 のいずれかをとる2値の質的変数とします。変数 y が 1 になる確率を変数 $x_1, \cdots, x_D$ を用いて

$$\mathbb{P}[y = 1] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D)}}$$

で表現することで説明・予測できると仮定します。例えば、訪問時刻・担当者年齢を用いて契約結果を説明するロジスティック回帰は

$$\mathbb{P}[\text{契約結果} = 1] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{訪問時刻} + \beta_2 \times \text{担当者年齢})}}$$

という式を検討していることになります。

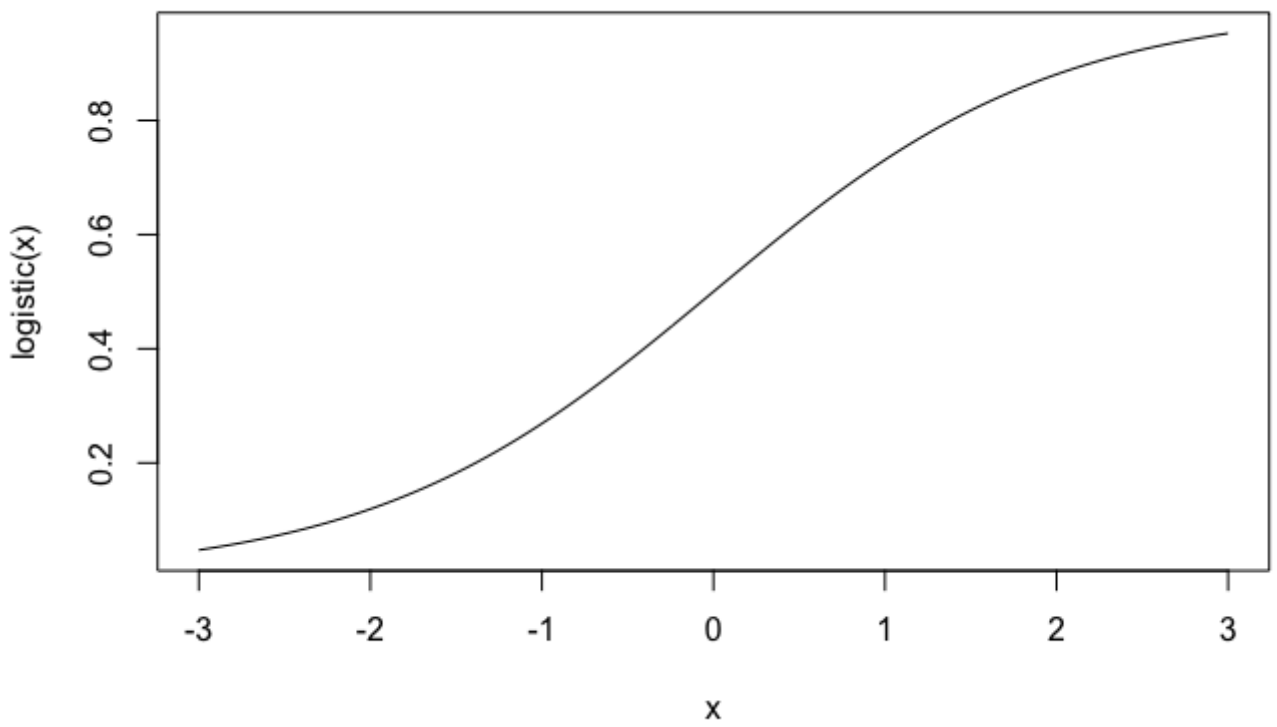
ここで、ロジスティック回帰を理解する上で役に立つ**ロジスティック関数**（logistic function）について紹介します。関数

$$f(x) = \frac{1}{1 + e^{-x}}$$

をロジスティック関数といいます。ロジスティック関数のグラフは次のようになります。

Hide

```
# ロジスティック関数
logistic <- function(x){return(1.0/(1.0+exp(-x)))}
curve(logistic, from = -3.0, to = 3.0)
```



ロジスティック関数の大きな特徴は、以下の3点です。

1. 0 より大きく 1 より小さい値をとる。
2. x に対して単調増加な関数である。
3. $f(0) = 0.5$

ロジスティック回帰のなかで特に大切になるのは、1番目と2番目の性質です。1番目の性質は、説明変数の1次式によって得られた値がロジスティック関数によって確率の値に変換できることを意味しています。2番目の性質は、偏回帰係数の値が正（resp. 負）の説明変数の値が大きくなればなるほど、確率が大きく（resp. 小さく）なるような統計モデルになっていることを意味しています。これを**線形性**といいます。

B. 出来ること

モデルに含まれる値が未知のパラメータ $\beta_1, \cdots, \beta_D$ を**偏回帰係数**（coefficient）といいます。これらの係数を推定することで、各説明変数の値が決まっているとき、目的変数の値はいくらと予測できるかを推定することができます。また、偏回帰係数の区間推定や統計的仮説検定を行うことができます。これについてはデモで例を交えながら解説します。

ここで、偏回帰係数の解釈の仕方に役立つ**オッズ**（odds）を紹介します。確率 p に対して、

$$\frac{p}{1-p}$$

をオッズといいます。

問題 : 以下の確率について、オッズを計算してください。

1. $p = 0.5$
2. $p = 0.4$
3. $p = 0.6$

解答 : (1) $p/(1-p) = 1$, (2) $p/(1-p) = 2/3$, (3) $p/(1-p) = 3/2$■

確率 $p = 0.4$ と確率 $p = 0.5$ との間でオッズは $3/2$ 倍、また確率 $p = 0.5$ と確率 $p = 0.6$ との間でもオッズは $3/2$ 倍になるところに注目すると、オッズという概念を受け入れやすいかもしれません。ところで、ロジスティック回帰は $y = 1$ の確率 p をオッズで表すとき

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D}$$

と仮定していることに対応しています。このことから、オッズを用いると説明変数 x_d の偏回帰係数 β_d は「説明変数 x_d の値のみが 1 増えたとき、$y = 1$ のオッズが e^{β_d} 倍増える」と解釈できます。

C. 偏回帰係数の推定の仕組み

偏回帰係数は、**最尤推定**（maximum likelihood estimation）という方法で推定されます。最尤法とは何か、考え方を単純な例で紹介しましょう。

問題 : 表が確率 p で出るコインを5回投げ、その結果から p の値を推定することを考えます。仮に、コインを投げて「表裏表表裏」という結果を得た場合、p の値はいくらと推定できるでしょうか。

解答 : 「表裏表表裏」という結果は確率 $p^3(1-p)^2$ で得られます。そこで、この確率が最も大きくなるような p の値を推定値として用いるのが最尤推定です。本来は**微分**によってわかることですが、今回は関数のグラフをかくことで $p = 3/5$ とわかります。

Hide

```
# 最尤推定
lik <- function(x){return(x^3*(1-x)^2)}
curve(lik, from = 0.0, to = 1.0)
```

ゆえに、p の最尤推定値は $\hat{p} = 3/5$ です。■

さて、契約結果を説明するロジスティック回帰の場合、次のような関数が最大になるような $\beta_0, \beta_1, \beta_2$ の値を偏回帰係数の推定値に用います。

$$\frac{1}{1+e^{-(\beta_0+\beta_1\times 0+\beta_2\times 42)}}\times\frac{1}{1+e^{-(\beta_0+\beta_1\times 0+\beta_2\times 22)}}\times\cdots\times\frac{1}{1+e^{-(\beta_0+\beta_1\times 0+\beta_2\times 21)}}$$

2.2 ロジスティック回帰のデモ

訪問時刻・担当者年齢を用いて契約結果を説明するモデル、つまり

$$\mathbb{P}[\text{契約結果}=1]=\frac{1}{1+e^{-(\beta_0+\beta_1\times\text{訪問時刻}+\beta_2\times\text{担当者年齢})}}$$

を考えます。

A. 線形性の確認

ロジスティック回帰は、説明変数と目的変数との間に線形性が仮定されていました。説明変数と目的変数との間の関係を可視化・数値要約することで、これを確認しましょう。

Hide

```
# 訪問時刻と契約結果の間の関係
table(dat$訪問時刻, dat$契約結果)
```

Hide

```
# 担当者年齢と契約結果の間の関係：線形性が確認できる。
par(family = "ヒラギノ角ゴシック W3")
plot(dat$担当者年齢, dat$契約結果)
```

B. ロジスティック回帰の計算

R 言語では次のようにロジスティック回帰を計算することができます。

Hide

```
# ロジスティック回帰
result <- glm(契約結果 ~ 訪問時刻 + 担当者年齢,
              data = dat)
summary(result)
```

問題 : 以下の問いに答えてください。

1. 推定によって得られた式を答えてください。
2. 担当者年齢の偏回帰係数の値が意味することを「オッズ」という言葉を用いて説明してください。

解答 :

1. 得られた式は以下の通りです。

$$\mathbb{P}[\text{契約結果}=1]=\frac{1}{1+e^{-(-0.34-0.45\times\text{訪問時刻}+0.03\times\text{担当者年齢})}}$$

2. 担当者年齢が1歳増えると、契約結果のオッズが$e^{0.03}\sim 1.03$倍増える。■

2.3 モデル選択

A. 赤池情報量規準

今回は、訪問時刻・担当者年齢を用いて契約結果を説明する式を作りました。しかし、次のように訪問時刻を含まないモデルも考えることができます。

Hide

```
# 訪問時刻を抜いたモデル
result2 <- glm(契約結果 ~ 担当者年齢, data = dat)
summary(result2)
```

複数のモデルを作ったら、どちらのモデルを採用するかを検討しましょう。今回も、モデル選択の際にヒントになる指標の一つとして、未知のデータへの予測の精度を推定する方法だった赤池情報量規準を用いることができます。（第2回資料も参考にしてみてください。）

Hide

```
# モデル比較
AIC(result); AIC(result2)
```

```
[1] 23.31996
[1] 28.80944
```

今回は、赤池情報量規準がより小さい訪問時刻を含んだロジスティック回帰を採用するほうが良いのではないかと推定できます。