マーケティング・リサーチ特論
～パラメータ推定の原理と手段～

2024年度1学期:　水曜3限
担当教員:　石垣　司

再掲: モデル・推定原理・推定手段

データを用いた情報の創出には3つの調和が必須

モデル

対象の特性の部分的な表現。(本授業では主に) 確率的構造の数学的表現

-本授業で扱うモデル-

線形回帰モデル,
因子分析モデル,
多項ロジットモデル,
階層回帰モデル, など

推定原理

未知のパラメータを含んでいるモデルとデータを整合させるための原理

-本授業で扱う原理-

① ズレの最小化
(最小2乗推定, 一般化モーメント法, LASSOなど)
② 尤度の最大化
③ 事後分布の推定

推定手段

モデルとデータを整合させるための関数やアルゴリズムなどの手段

-本授業で扱う手段-

推定量, ニュートン法,
EMアルゴリズム,
マルコフ連鎖モンテカルロ法, など

※3つは混同しやすいので注意。特に原理と手段に関しては"推定法"などとまとめて書かれている書籍等もある。ここでは混同を防ぐため, 推定原理と推定手段という言葉で区別する。

例えば、"階層ベイズモデル"は階層的モデルをベイズ推定の原理で扱うということ。モデルと原理を両方含んでいる言葉。経済学やマーケティング・リサーチで扱う階層ベイズモデルは回帰モデルや離散選択モデルがほとんどだが、他の分野ではそうとは限らないので注意

推定原理の種類

統計モデルのパラメータ推定の原理

– **統計モデルに含まれているパラメータとデータを整合させるための基準となる原則**

1. **モデルとデータのズレを最小化する原理**
   – **例: 最小二乗法, 一般化モーメント法など**
2. **最尤推定の原理**
   – **モデルとデータの尤度を最大化**
3. **ベイズ推定の原理**
   – **データが与えられた時のパラメータの条件付き分布を求める**

#原理: 他のものを規定するが、それ自身は他に依存しない根本的、根源的なもの (デジタル大辞泉)。原理の例:期待効用最大化原理、光速度不変の原理

最小二乗法によるパラメータの推定

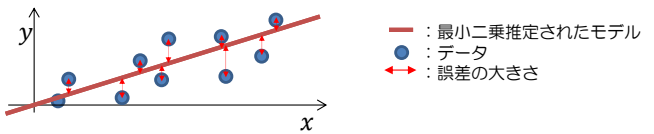
最小二乗法の原理

– **モデルとデータの誤差の二乗和(RSS)を最小化**
– **考え方: モデルとデータの全体的なズレは小さい方が良い**

線形回帰モデルの RSS: $(\boldsymbol{y} - X^T\boldsymbol{b})^T(\boldsymbol{y} - X^T\boldsymbol{b})$

– **パラメータ b に関する2次式 ⇒ 正則なら必ず最小解が見つかる**

線形回帰分析の最小二乗推定量: $\hat{\boldsymbol{b}} = (X^TX)^{-1}X^T\boldsymbol{y}$



– **RSSを適切な目的関数として設定できる場合は有効**

反例: ロジスティック回帰モデルを最小二乗法でパラメータ推定すると, 目的関数は非凸関数 & 推定係数の統計的性質を扱いにくい

#メモ:関数形を決めるだけなら、誤差項に正規分布の仮定は不必要。計量経済学の講義で学習する回帰係数の検定や推定量の不偏性・一致性などの性質は、誤差項に確率分布を仮定したときの論理展開

ズレを最小化する原理に基づく推定の例

Ridge **回帰**

$$\text{Minimize } (\boldsymbol{y} - X^T\boldsymbol{b})^T(\boldsymbol{y} - X^T\boldsymbol{b}) + \lambda\|\boldsymbol{b}\|^2$$

– **モデルとデータの誤差の二乗和＋L2正則化項**

LASSO **回帰**

$$\text{Minimize } (\boldsymbol{y} - X^T\boldsymbol{b})^T(\boldsymbol{y} - X^T\boldsymbol{b}) + \lambda|\boldsymbol{b}|$$

– **モデルとデータの誤差の二乗和＋L1正則化項**

一般化モーメント法

– $E[\boldsymbol{z}(\boldsymbol{y} - X^T\boldsymbol{b})]$**の2次形式最小化**

誤差ベクトル $(\boldsymbol{y} - X^T\boldsymbol{b})$ **と操作変数ベクトル** $\boldsymbol{z}$ **の内積（直交条件）**

– **両ベクトルの直交空間からのズレを最小化**

#メモ：正則化項について本授業では詳述しないが、現代的な機械学習の手法において重要な役割をもつ

最尤法によるパラメータ推定 #1

尤度：与えられたデータがある確率分布から発生していると考えたときの尤もらしさの度合い

– **データ** y_i **を発生させる確率分布：** $p(y_i;\boldsymbol{\theta})$

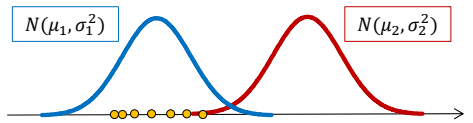
（パラメータ $\boldsymbol{\theta} = [\theta_1, \dots, \theta_P]^T$**）**

– **データ** y_i **の尤度：** $l(\boldsymbol{\theta}|y_i) = p(y_i;\boldsymbol{\theta})$

– **独立同一分布から発生したデータ** $\boldsymbol{y} = [y_1, \dots, y_N]^T$ **の尤度：**

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{N} p(y_i;\boldsymbol{\theta})$$

例：次のデータは青と赤のどちらの正規分布から発生していると考えるのが妥当か？



$$l_1(\mu_1, \sigma_1^2|\boldsymbol{y}) > l_2(\mu_2, \sigma_2^2|\boldsymbol{y})$$

確率分布や尤度の表記法

$p(y)$**：確率変数** y **の確率分布：**

– **この授業の範囲内では，具体的な確率関数か確率密度関数**

$p(y,z)$**：確率変数** y **と** z **の同時分布（結合分布）**

$p(y|z)$**：確率変数** z **の値が定まったときの** y **の条件付き分布**

– $p(y,z)$ **は2変量の分布で，**$p(y|z)$ **は1変量の分布**

$p(y;\theta)$**：パラメータ** θ **を持つ確率変数** y **の確率分布**

– **ここでの「;」は，パラメータは確率変数ではない点を強調している**

$l(\theta|y)$**：確率変数** y **の値が定まったときのパラメータ** θ **の関数**

– **具体的な数式は確率関数・確率密度関数** $p(y;\theta)$ **と同じ場合がほとんどである。ただし，引数が異なることに注意**

最尤推定によるパラメータ推定 #2

最尤推定の原理：尤度の最大化

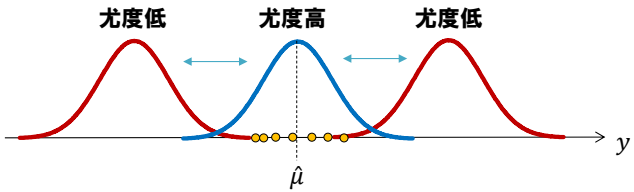
– **考え方：最も尤もらしいモデルからデータは発生する**

最尤推定

– **尤度関数を最大にするパラメータを推定値とする推定方法**

尤度関数：尤度 $l(\boldsymbol{\theta}|\boldsymbol{y})$ **はパラメータ** $\boldsymbol{\theta}$ **の関数**

例：分散 σ^2**の値を固定して正規分布** $N(\mu, \sigma^2)$ **の** μ **を動かすと尤度** $l(\mu|\boldsymbol{y})$ **の値は変化する。その中で尤度が最大となる** μ **を最尤推定値** $\hat{\mu}$ **とする。**



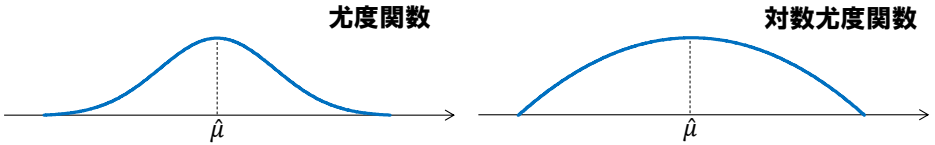
対数尤度関数

多くの場合，パラメータの推定には対数尤度関数を用いる

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \log\{l(\boldsymbol{\theta}|\boldsymbol{y})\} = \sum_{i=1}^{N} \log\{p(y_i;\boldsymbol{\theta})\}$$

- **対数尤度関数は上に凸な関数となることが多く使いやすい**

  例：正規分布の平均パラメータ μ の尤度関数と対数尤度関数



尤度関数と対数尤度関数の両者で最尤推定値は同じ

例：ロジスティック回帰モデルの対数尤度関数は上に凸な関数形となり数値最適化が容易。統計的性質として一致性をもつ

線形回帰係数の最尤推定

線形回帰モデルの尤度関数

- **線形重回帰モデル：** $\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e}, \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$

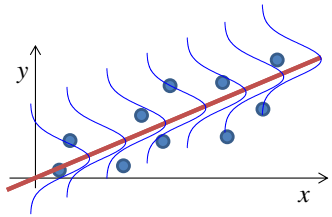
  $N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ **は** N **次元多変量正規分布**

  $\boldsymbol{x}_i = [1 \quad x_{i1} \quad \cdots \quad x_{iP}]^T$（**行列** X **の** i **番目の行**）
- **尤度関数：** $l(\boldsymbol{b}|\{\boldsymbol{x}\}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i\boldsymbol{b})^2\right\}$

線形回帰モデルの回帰係数の最尤推定量

$$\widehat{\boldsymbol{b}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

#メモ：最小2乗推定量と一致



ベイズ推定の理解のための数学的準備

データとパラメータの確率分布の表記方法

- $p(\boldsymbol{y})$**：データベクトル** $\boldsymbol{y} = [y_1, \ldots, y_N]^T$ **が生じる同時確率密度関数。** $p(\boldsymbol{y}) = p(y_1, \ldots, y_N)$
- $p(\boldsymbol{\theta})$**：パラメータベクトル** $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_P]^T$ **の各要素を確率変数とみなしたときに** $\boldsymbol{\theta}$ **が生じる同時確率密度関数.** $p(\boldsymbol{\theta}) = p(\theta_1, \ldots, \theta_P)$

  例：正規分布では $p(\boldsymbol{\theta}) = p(\mu, \sigma^2)$
- **確率の乗法定理：** $p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
- $p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}$
- $\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}$**：パラメータの集合で積分をする記号**

  例： $\theta_i \in \mathbb{R}\ (i = 1, \ldots, P)$ **なら,**

  $\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\theta_1 \cdots d\theta_P$

ベイズ統計学とは？

ベイズ統計学での統計モデル

- **パラメータ** θ **を確率変数とみなして事前情報をモデル化する**
- **事前情報＋データを用いて** θ **の値を推定**

データとパラメータを結びつけるベイズの定理

- **連続型確率変数のベイズの定理**

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

- $p(\boldsymbol{\theta}|\boldsymbol{y})$**：事後確率分布（事後分布）**
- $p(\boldsymbol{\theta})$**：事前確率分布（事前分布）**
- $p(\boldsymbol{y}|\boldsymbol{\theta})$**：尤度関数**
- $p(\boldsymbol{y})$**：周辺尤度**

事後分布が新しい情報

尤度の観点からみたベイズの定理の意味

- パラメータ θ を確率変数と見ている
- 独立同一分布から発生したデータ $\boldsymbol{y} = [y_1, \ldots, y_N]^T$ を考えると，$p(\boldsymbol{y}|\boldsymbol{\theta})$ は尤度関数 $\prod_{i=1}^{N} p(y_i; \boldsymbol{\theta})$
- 事後分布 $p(\boldsymbol{\theta}|\boldsymbol{y})$ はデータ観測後のパラメータの分布

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\prod_{i=1}^{N} p(y_i; \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto \prod_{i=1}^{N} p(y_i; \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

$\propto$：比例するという記号

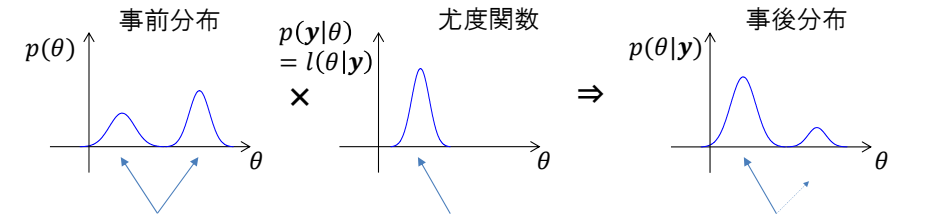
- ベイズの定理の構造

事後分布 ⇐ データ（尤度関数）× 事前分布

事後分布 $p(\boldsymbol{\theta}|\boldsymbol{y})$ のイメージ

データによって更新されたパラメータの情報

- 事前分布：データが得られる前のパラメータに関する情報を確率分布によって表現
- 尤度：データ自体が持つパラメータに関する情報



この辺りに本当の θ がありそう（データ取得前）

データはこの辺りだと言っている

この辺りに本当の θ がありそう（データ取得後）

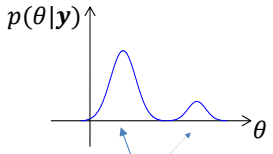
ベイズ推定によるパラメータの推定

ベイズ推定の原理：事後分布の導出

- 考え方：事前分布 $p(\theta)$ が所与のとき，事後分布 $p(\theta|\boldsymbol{y})$ はデータ所与後の統計モデルのパラメータの情報そのものである

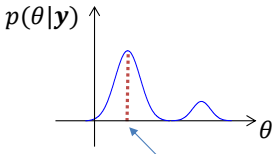
  パラメータ θ を確率変数と見なして，事後分布 $p(\theta|\boldsymbol{y})$ や事後分布の代表値の推定値 $\hat{\theta}$ を求める

推定された事後分布自体がパラメータに関する情報



この辺りに本当の θ がありそう（データ取得後）

事後分布の代表値の値はパラメータに関する情報



パラメータ θ の事後分布が最大となる点はここだ

推定の手段

統計モデルに含まれている未知パラメータの値や分布を求めるための手段

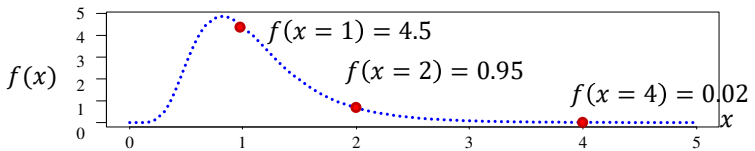
（その代表例）

1. 推定量
   - どの原理でも利用される。最小2乗推定量，最尤推定量，ベイズ推定量，一般化モーメント推定量など

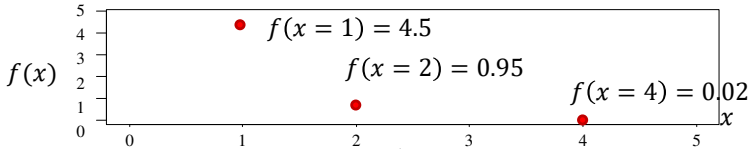
     計量経済学の主眼の一つ。本日の授業では省略
2. 数値的最適化
   - どの原理でも利用される。Newton法，EMアルゴリズムなど
3. 乱数を利用した事後分布の計算
   - ベイズ推定のみで利用。マルコフ連鎖モンテカルロ法など

数値的最適化の問題設定

1. 関数 $f(x)$ の形は分からない ⇒ 青の点線の形はわからない
2. 関数 $f(x)$ に最適解があることは分かっている ⇒ 範囲は不明
3. 関数 $f(x)$ の導関数 $f'(x)$ は計算可能
4. ある値 c における関数の値 $f(x=c)$ は計算可能 ⇒ 赤点の値



関数 $f(x)$ の形状

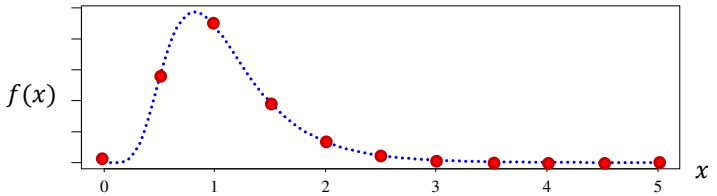


我々が知覚できる情報

グリッドサーチ

ある範囲で設定した格子点(グリッド)の値を調べる

- ○: 単純で直感的にも理解しやすい
- ×: 最適解を見つけられる保証はない
- ×: 変数の数が増えると探索点も指数的に増加
  - 1変数について10格子点を設定 ⇒ 10点の探索
  - 10変数についてそれぞれ10格子点を設定 ⇒ 10^{10} 点の探索
    - 前々回の授業のロジスティック回帰モデルの回帰係数ベクトル b の要素数は10
- 例: $x=[0,5]$ の範囲で0.5刻みで10個の格子点を設定

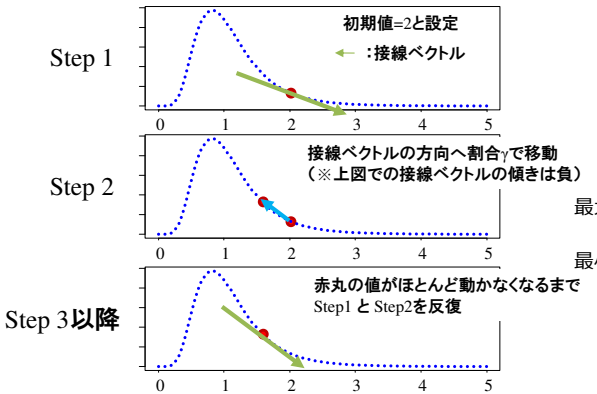


この場合、$x=1$ の値が最も大きいので、最適解として $x=1$ を採用

最急降下法(山登り法)

関数 $f(x)$ の1階導関数を利用した反復計算

- ○: 1回の計算量は少ない
- ×: 反復回数が大きくなる傾向
- ×: 初期値に反復回数が依存



【アルゴリズム】
初期設定
Set $k=0$
Set initial value of $x^{(k=0)}$
Set parameter γ

収束するまで反復計算

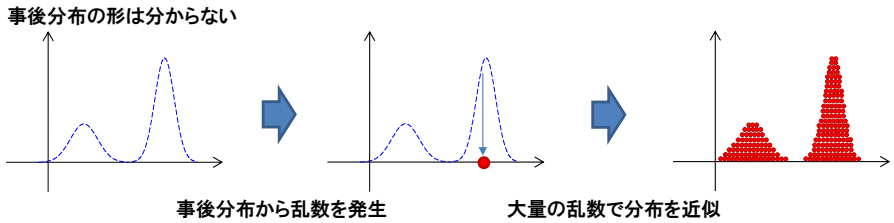
最大化問題 $x^{(k+1)}=x^{(k)}+\gamma\dfrac{\partial f(x^{(k)})}{\partial x^{(k)}}$

最小化問題 $x^{(k+1)}=x^{(k)}-\gamma\dfrac{\partial f(x^{(k)})}{\partial x^{(k)}}$

事後分布の計算のためのMCMC法

乱数により事後分布を近似する手法の総称

- 長所:実用的なほとんどの事後分布の推定に適用可能。乱数の数を増やすと, 真の事後分布と一致する性質
  - 現在では最もスタンダードなベイズ推定のための手法
- 短所:計算コストが高い。処理を並列化できないため, 本質的な高速化が困難



#メモ 1990年代以降, ベイズ統計学が普及してきた技術的な理由は計算機の高速化によってMCMC法が実用的になったため。現在でもより効率的な手法の開発が盛んに研究されている

統計的モデリングのまとめ

統計的方法の本質は，データを用いて必要な情報を創り出すことにある

- モデルは我々の持つ期待の構造の形式的な表現であり，より良いモデルの探究によって真理に迫る
- ズレの最小化による推定，最尤推定，ベイズ推定は推定原理自体が異なる
- モデルや推定原理の違いで推定手段も異なる
- 情報の創出にはこの3つの調和が必須

モデル

対象の特性の部分的な表現。（本授業では主に）確率的構造の数学的表現

推定原理

未知のパラメータを含んでいるモデルとデータを整合させるための原理

推定手段

モデルとデータを整合させるための関数やアルゴリズムなどの手段