母分布のモデルの推定

Data visualization

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

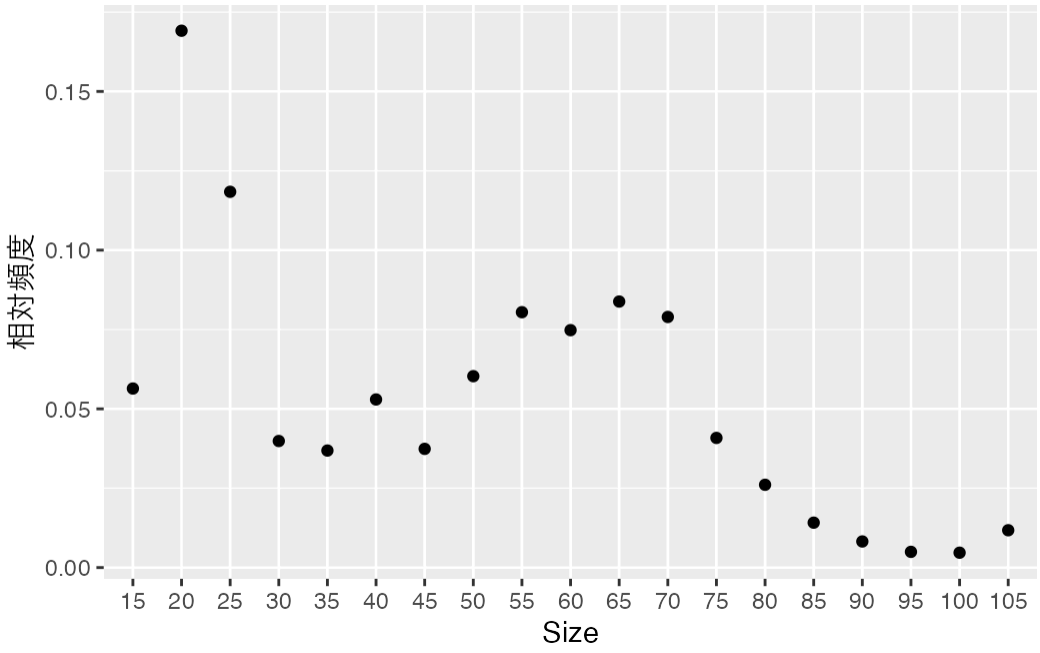
2025-07-31

1 分布のモデル

1.1 分布のモデル

- 母分布の推定が難しい場合の、代替案は、分布を単純(モデル)化し推定
  - 分布の特徴の推定(OLS や平均値)と大枠では同じ議論が適用できる

1.2 例: Size の分布

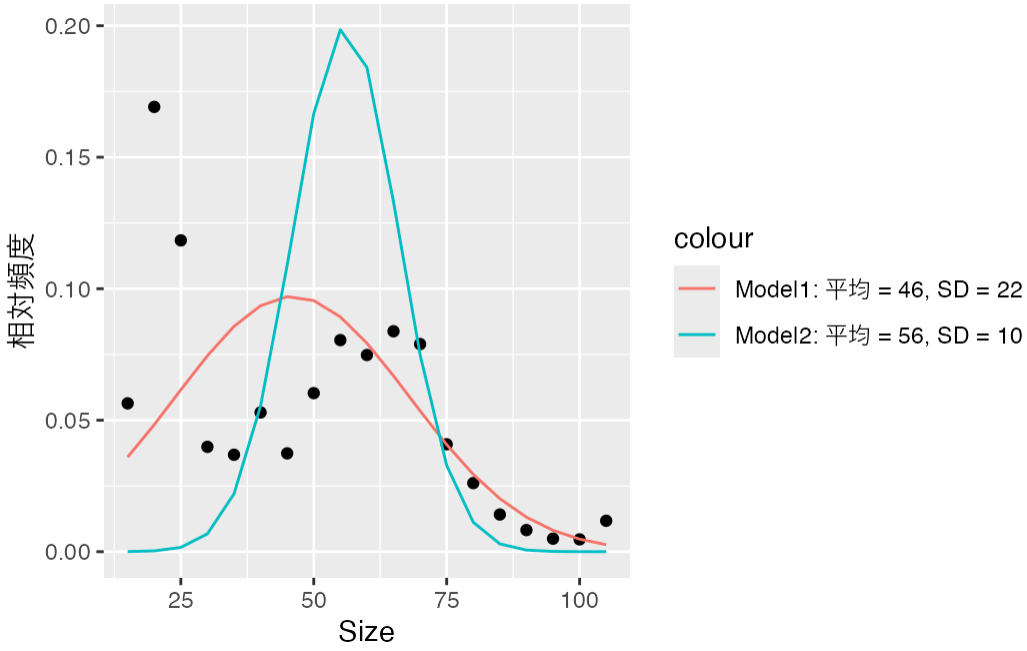


1.3 例: 正規分布によるモデル化

- 代表的な分布のモデル wiki

- ベル(富士山)型の分布
- 二つのパラメタ(平均値,分散)が決まれば、分散が決まる

1.4 例: 正規分布によるモデル化



1.5 分布の距離

- どのモデルがデータへの当てはまり最も良いか?
- 代表的指標は、カルバック・ライブラー情報量 (KL divergence)

$$\underbrace{\sum_{Y} \text{データ上の}Y\text{割合} \times \log \frac{\text{データ上の}Y\text{割合}}{\text{モデル上の}Y\text{割合}}}_{\text{分布の乖離度}}$$

1.6 例

Size	model	model_bias	freq	KL: Model1	KL: Model2
15	0.036	0.000045	0.056	0.02537	0.403
20	0.048	0.000306	0.169	0.21199	1.068
25	0.062	0.001633	0.118	0.07741	0.507
30	0.075	0.006792	0.04	−0.02494	0.071
35	0.086	0.021992	0.037	−0.03109	0.019
40	0.094	0.055461	0.053	−0.03013	−0.002

Size	model	model_bias	freq	KL: Model1	KL: Model2
45	0.097	0.108927	0.037	−0.03564	−0.04
50	0.095	0.166613	0.06	−0.02772	−0.061
55	0.089	0.198477	0.08	−0.00838	−0.073
60	0.079	0.184136	0.075	−0.00436	−0.067
65	0.067	0.133043	0.084	0.01894	−0.039
70	0.054	0.074864	0.079	0.03066	0.004
75	0.041	0.032808	0.041	0.00012	0.009
80	0.029	0.011197	0.026	−0.00313	0.022
85	0.02	0.002976	0.014	−0.00502	0.022
90	0.013	0.000616	0.008	−0.00385	0.021
95	0.008	0.000099	0.005	−0.00245	0.019
100	0.005	0.000012	0.005	−0.00009	0.028
105	0.003	0.000001	0.012	0.01746	0.108

1.7 最尤法

- カルバック・ライブラー情報量を最小にするように、モデルのパラメタを推定する
  - 他の推定方法として、ベイズ法が有力

1.8 最尤法の実際

- カルバック・ライブラー情報量を最小 = 以下を最大化

$$\sum_{Y} \text{データ上の}Y\text{割合} \times \log\left(\underbrace{\text{モデル上の}Y\text{割合}}_{=\text{尤度}}\right)$$

- “データが実現する確率を最大にするように推定する方法”としても**解釈できる**

1.9 複数変数の分布

- 複数の変数の(同時)分布も、モデル化できる
- 代表的なモデルは、古典的線形モデル

$$Y = \beta_0 + \beta_1 X_1 + .. + \underbrace{u}_{\text{正規分布}}$$

- 他には、Y が 2 値変数 $(Y = 0/1)$ のケースでよく使われる logit モデル/probit モデル

1.10 実装

- 代表的なモデルについては、容易に実装可能

```r
library(tidyverse)

data <- read_csv("Data/example.csv")

glm(Price ~ Size, data, family = "gaussian") # 古典的線型モデル
```

```
Call:  glm(formula = Price ~ Size, family = "gaussian", data = data)

Coefficients:
(Intercept)         Size
     -6.463        1.133

Degrees of Freedom: 11310 Total (i.e. Null);  11309 Residual
Null Deviance:      23130000
Residual Deviance: 15830000     AIC: 114000
```

1.11 実装

```r
glm(year_2024 ~ Size, data, family = "binomial") # ロジットモデル
```

```
Call:  glm(formula = year_2024 ~ Size, family = "binomial", data = data)

Coefficients:
(Intercept)         Size
 -0.0909797    0.0005865

Degrees of Freedom: 11310 Total (i.e. Null);  11309 Residual
Null Deviance:      15670
Residual Deviance: 15670    AIC: 15670
```

2 母分布の推定

2.1 アイディア

- OLS と類似した解釈が可能

1. 母分布上で、仮想的に最尤推定を行った結果を推定目標として定義
2. データ上で行った最尤推定の結果を、推定値として定義
3. サンプリング誤差を測定

2.2 古典的方法

- 正しいモデル化できていれば、以下の方法で信頼区間が計算可能
  - 母分布 = 母集団上での最尤推定の結果

```
library(tidyverse)

data <- read_csv("Data/example.csv")

model <-glm(Price ~ Size, data, family = "gaussian") # 古典的線型モデル

confint(model)
```

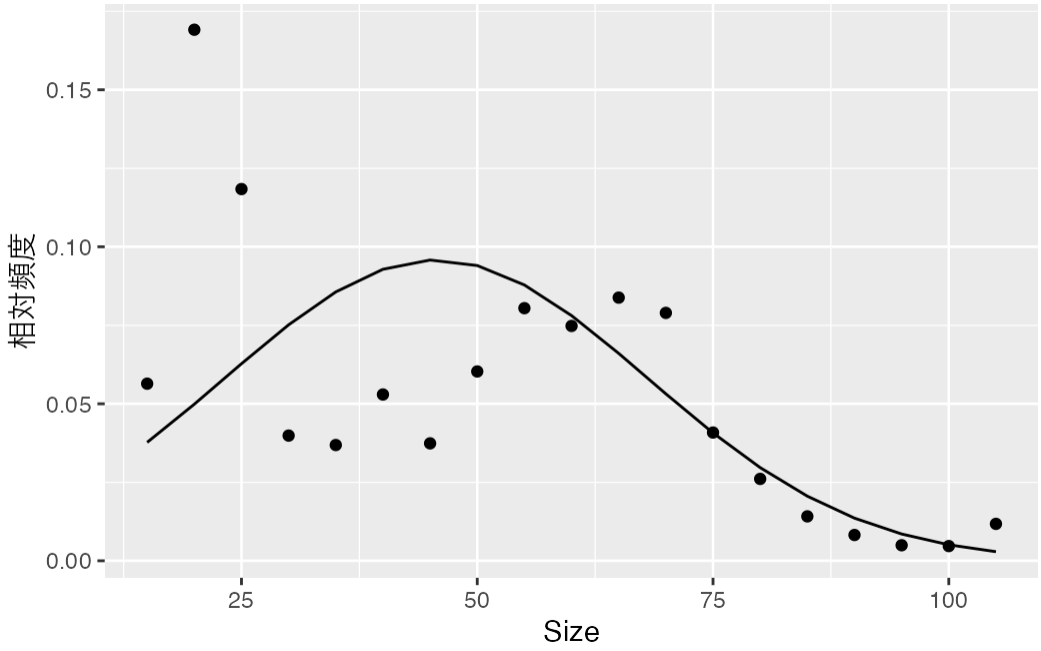
```
                2.5 %    97.5 %
(Intercept) -8.025177 -4.900060
Size         1.102097  1.163541
```

3 誤定式化

3.1 実践的な解釈

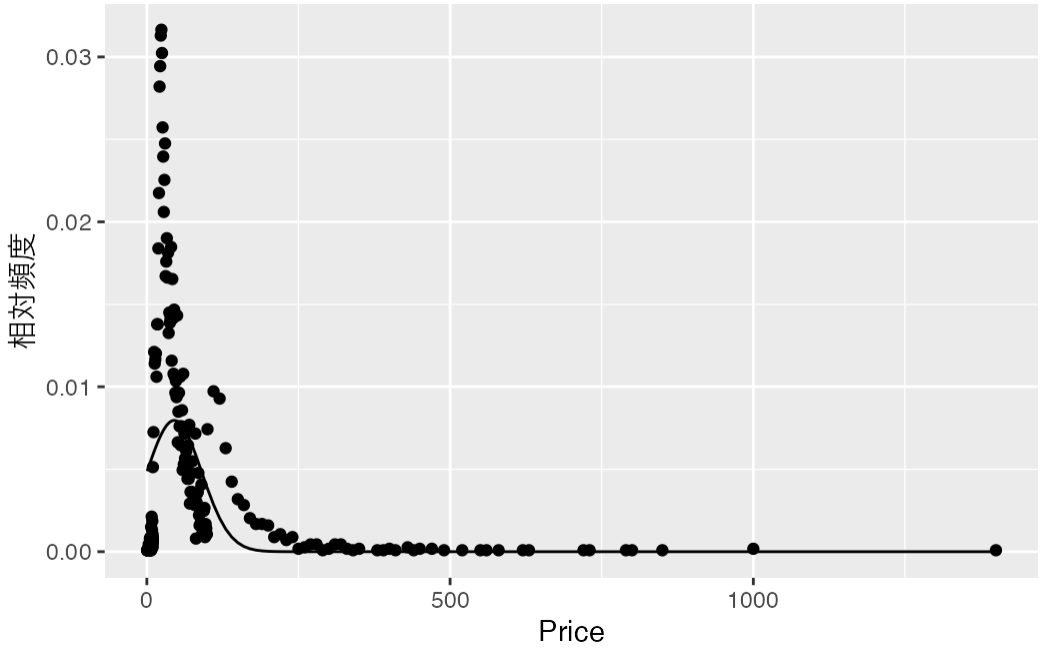
- 実践的な解釈は、「母分布をある程度近似するモデル」をデータから推定する
  - モデルの大枠は研究者が設定する
    - 誤定式化を犯していることを前提とする
- 特殊なケースは、母分布を完璧に近似するモデル (誤定式化がないモデル)をデータから推定する

3.2 例: Size



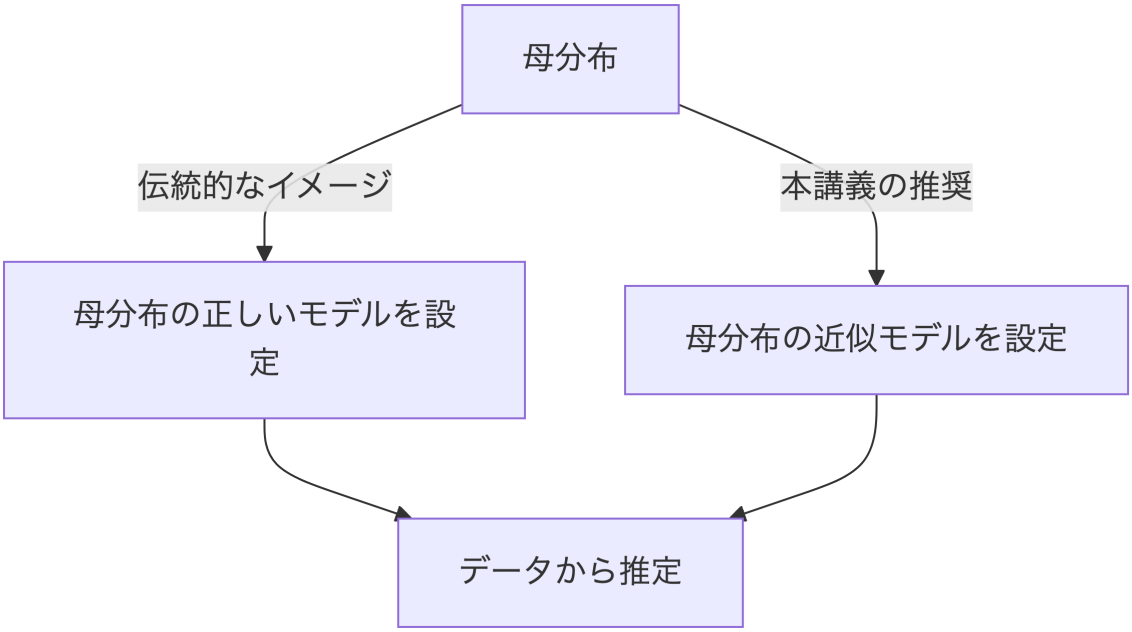
- 正規(富士山)分布モデルでは、“連山”的な分布を捉えられない

3.3 例: Price



- 裾野の長い分布を捉えられない

3.4 伝統的な教科書との対比



3.5 Takeaway

- OLS: 推定目標 = “母平均の母集団上でのモデル (Population OLS)”
- 最尤法: 推定目標 = “母分布の母集団上でのモデル (Population OLS)”
  - どちらも、データ上でのモデル ≃ 母集団上でのモデル、が基本アイディア
  - どちらも、誤定式化があることを前提に推定目標を定義できる

Bibliography