データ解析のための統計モデリング入門

久保拓弥

2012 年（岩波書店）

1　データを理解するために統計モデルを作る

1.2　「ブラックボックスな統計解析」の悪夢

ブラックボックス統計学（理解しないままソフトウェアを使う作法）による誤用の例

- R^2 値は「説明力」なので、ひたすら 1 に近ければ良い。

2　確率分布と統計モデルの最尤推定

2.1　例題：種子数の統計モデリング

R によるヒストグラム　`breaks=` で、-0.5 から区間 1 ずつに区切ったヒストグラムを描く。

```
hist(data, breaks=seq(-0.5, 9.5, 1))
```

2.3　ポアソン分布とは何か？

どのようなデータに対してポアソン分布によるモデル化を行うか。

1. データが非負の整数である（カウントデータ）
2. 下限はゼロみたいだが上限はよくわからない
3. データでは平均と分散がだいたい等しい

上記に加えて、独立性や均質性も成り立っていること。

2.5　統計モデルの要点：乱数発生・推定・予測

統計モデルの良さを評価する上では、**予測の良さ**（推定されたモデルが新しく得られたデータにどれくらいよく当てはまるか）という考え方が重要になる。

予測の良さの検証（validation）

1. データは人間には見えない真の統計モデルから発生している
2. たまたま得られたデータからある統計モデルを仮定しパラメータを推定する
3. 推定されたモデル分布が新たに得られるデータ分布をどれくらい予測できるか評価する

2.6　確率分布の選び方

- 説明したい量は離散か連続か？
- 説明したい量の範囲は？
- 説明したい量の標本平均と標本分散の関係は？

3　一般化線形モデル（GLM）

3.4　ポアソン回帰の統計モデル

データ

- `y`：種子数
- `x`：個体の体サイズ[1]
- `f`：施肥処理（統制群 `C` or 実験群 `T`）

```
> d <- read.csv("data3a.csv")
> head(d,n=5)
   y     x f
1  6  8.31 C
2  6  9.44 C
3  6  9.50 C
4 12  9.07 C
5 10 10.16 C
```

3.4.1　線形予測子と対数リンク関数

平均種子数が個体ごとに異なり、それが体サイズだけに依存すると考える（施肥処理はとりあえず無視する）。
　その上で、モデルを

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!} \tag{3.1}$$

ただし、

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i) \tag{3.2}$$

とする。つまり、平均種子数は**リンク関数**と**線形予測子**を用いて

$$\log \lambda_i = \beta_1 + \beta_2 x_i \tag{3.3}$$

と表せる。

- 平均を exp でモデル化しておけば、非負となり都合が良い。
- 要因の効果が積で表される。

正準リンク関数

- 対数リンク関数：ポアソン回帰
- ロジットリンク関数：ロジスティック回帰

[1] x_i は測定誤差が全くないと仮定している。これを無視することで生じるバイアスを避けるには、説明変数 x_i も統計モデル化する必要がある。

3.4.2 あてはめとあてはまりの良さ

モデルの対数尤度は、

$$\log L(\beta_1, \beta_2) = \sum_i \log \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!} \tag{3.4}$$

となる。

R によるあてはめ `glm()` 関数でポアソン分布を指定[2]。正式には `family=poisson(link="log")` と指定しなくてはいけないが、ポアソン分布を指定した場合はデフォルトで対数リンク関数が用いられる。

```
> fit <- glm(y~x, data=d, family=poisson)
> summary(fit)

Call:
glm(formula = y ~ x, family = poisson, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3679  -0.7348  -0.1775   0.6987   2.3760

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.29172    0.36369   3.552 0.000383 ***
x            0.07566    0.03560   2.125 0.033580 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 89.507  on 99  degrees of freedom
Residual deviance: 84.993  on 98  degrees of freedom
AIC: 474.77

Number of Fisher Scoring iterations: 4
```

`fit` の中身は、`names(fit)` や `str(fit)` で見られる。後者の方が詳細。
`z value` は **Wald 統計量**で、最尤推定値 / SE。

最大対数尤度 パラメータ値が最尤推定値となっているときの対数尤度。

```
> logLik(fit)
'log Lik.' -235.3863 (df=2)
```

3.6.1 対数リンク関数のわかりやすさ：かけ算される効果

個体のサイズ `x` と施肥処理 f を説明変数とすると、モデルは

$$\lambda_i = \exp(1.26) \times \exp(0.08 x_i) \times \exp(-0.032)$$

と推測される。

- x_i が 1 増えると、λ_i は $\exp(0.08 \times 1) \approx 1.083$ 倍に増える。
- 肥料をやることで種子数の平均が $\exp(-0.032) \approx 0.9685$ 倍になる。

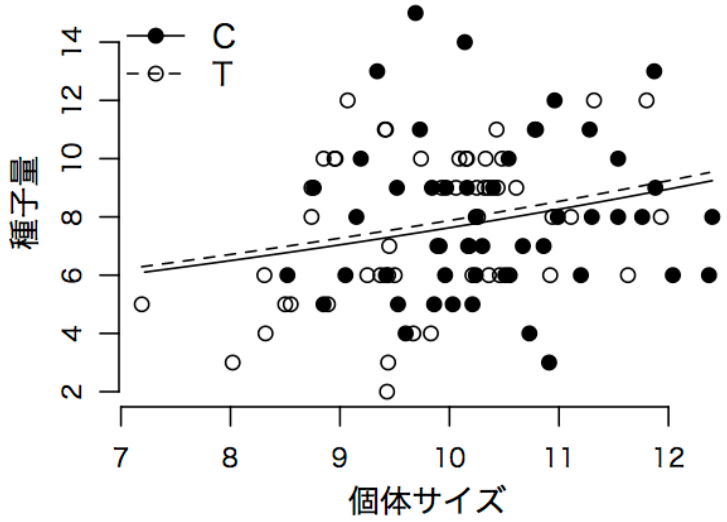

Figure 3.8: ポアソン回帰の当てはめ

データと予測値のプロットを表したものが図 3.8。この図では分かりにくいが、2 つの曲線の幅は、右に行くほど広くなっている。

本書は、変数変換によってではなく y の構造に合わせて適切な確率分布を選んでモデリングを行うという方針をとる。

4 GLM のモデル選択

4.1 データはひとつ、モデルはたくさん

「**最大対数尤度** = モデルの良さ」か？（そうではない。）

4.2 統計モデルのあてはまりの悪さ：逸脱度

逸脱度（deviance） 最大対数尤度を $\log L^*$ として、

$$D = -2 \log L^*$$

逸脱度は、フルモデルを当てはめた時に最小値、ゼロモデルを当てはめたときに最大値をとる。そこで、フルモデルの逸脱度を基準として、

- **残差逸脱度（residual deviance）** = 逸脱度 − フルモデルの逸脱度
- **ゼロ逸脱度（null deviance）** = ゼロモデルの逸脱度 − フルモデルの逸脱度

と定義する。R の `glm()` では、これら 2 つが報告される。

残差逸脱度は、パラメータ数 k さえ増やせばどんどん小さくなる（あてはまりが良くなる）。

[2]説明変数として施肥処理 `f` のような因子型を含める場合も、`y~x+f` のようにすればよい。ただし、`f` がちゃんと因子型になっているか、`class(d$f)` で確認しておくこと。

4.3 モデル選択規準 AIC

当てはまりの良さではなく、**予測の良さ**を重視する。

最大対数尤度を $\log L^*$、最尤推定したパラメータ数を k とすると、

$$\mathrm{AIC} = -2(\log L^* - k) \tag{4.1}$$
$$= D + 2k \tag{4.2}$$

で定義される。

4.5 なぜ AIC でモデル選択して良いのか？

4.5.1 統計モデルの予測の良さ：平均対数尤度

最大対数尤度は、推定された統計モデルがたまたま得られた観測データにどれくらい当てはまっているかを表す量。

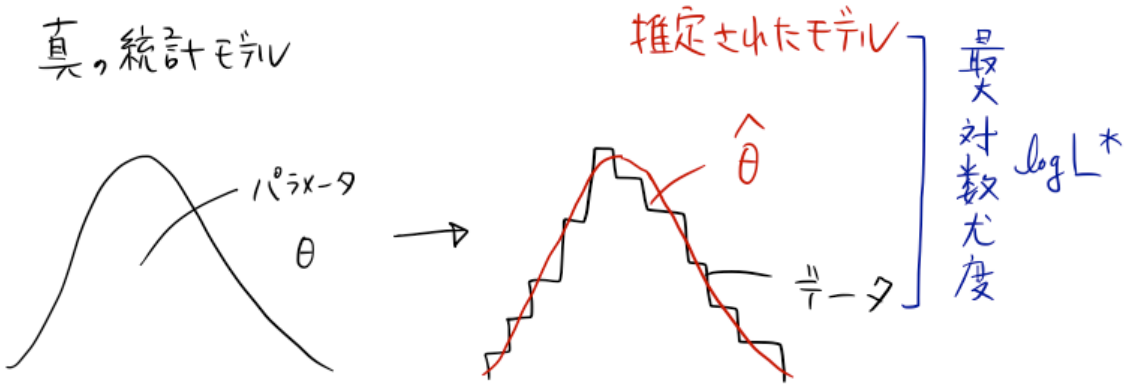

Figure 4.6: 最大対数尤度

今、真の統計モデル（パラメータ θ）から大きさ N の標本が得られ、

- パラメータの推定値 $\hat{\theta}$
- 最大対数尤度 $\log L^*$

が得られているとする。仮に同じ手続きで大きさ N の標本が多数回抽出され、$\theta = \hat{\theta}$ とした時の尤度を算出し、その平均（**平均対数尤度**：$E(\log L)$）を取ることができたとする。このとき、最大対数尤度と平均対数尤度の差（**バイアス**）

$$b = \log L^* - E(\log L) \tag{4.3}$$

の分布を考えると[3]、$E(b) = k$ となることが知られている[4]。

すると、データから計算された最大対数尤度とパラメータ数を用いて、平均対数尤度を

$$\widehat{E}(\log L) = \log L^* - k \tag{4.4}$$

と推定できる[5]。これを**バイアス補正**と呼ぶ。

つまり、

$$\mathrm{AIC} = -2\widehat{E}(\log L) \tag{4.5}$$

[3] b は $\hat{\theta}$ と $\log L^*$ を算出するために用いた当初データの標本変動によって左右される。

[4] 坂元・石黒・北川 (1983) P.52, 式 (4.37)

[5] 実際は、b の分布のばらつきが小さいとは限らないので（ネストされたモデルではばらつきは小さい）、点推定値で補正して良いのかという疑問も残る。

ということ。
　ここで、パラメータ数 k のモデルに l 個のパラメータを追加したモデルを当てはめることを考える。このとき、バイアスの期待値は

$$E(b_k) = k \tag{4.6}$$
$$E(b_{k+l}) = k + l \tag{4.7}$$

となる。つまり、パラメータを l 個追加することで、最大対数尤度と平均対数尤度の差は、平均で l 増加する。
　また、AIC は

$$\mathrm{AIC}_k = -2\widehat{E}(\log L_k) = -2(\log L_k^* - k) \tag{4.8}$$
$$\mathrm{AIC}_{k+l} = -2\widehat{E}(\log L_{k+l}) = -2(\log L_{k+l}^* - k - l) \tag{4.9}$$

なので、

$$\mathrm{AIC}_k - \mathrm{AIC}_{k+l} = -2(\log L_k^* - \log L_{k+1}^* + l) \tag{4.10}$$

より $\log L_{k+1}^* - \log L_k^* > l$ なら $\mathrm{AIC}_k > \mathrm{AIC}_{k+l}$ となる。つまり、パラメータを l 加えることによる最大対数尤度の増分が l を上回るのであれば、AIC は小さくなる。

5　GLM の尤度比検定と検定の非対称性

5.2　尤度比検定の例題：逸脱度の差を調べる

逸脱度の差から構成される検定統計量（尤度比検定統計量）

$$\Delta D_{1,2} = -2 \times (\log L_1^* - \log L_2^*) \tag{5.1}$$

を用いる。

5.2.1　方法 (1) 汎用性のあるパラメトリックブートストラップ法

以下のステップ 2-4 をたくさん繰り返す。

1. 得られたデータからモデル（モデル 1）とパラメータ推定値 $\hat{\theta}$ を決める
2. 上記のモデルとパラメタ値から大きさ N のデータを発生させる
3. 発生させたデータにモデル 1 とモデル 2 を当てはめる
4. モデル 1 とモデル 2 の逸脱度の差を計算する

すると、逸脱度の差（尤度比検定統計量）の分布が得られる。
　→ 得られたデータから計算された逸脱度の差は、この分布の上位何%点に位置するか？

5.2.2　方法 (2) χ^2 分布を使った近似計算法

尤度比検定統計量

$$\Delta D_{1,2} = -2 \times (\log L_1^* - \log L_2^*)$$

が帰無仮説のもとで近似的に χ^2 分布に従うことを利用する。

6　GLM の応用範囲をひろげる

6.1　さまざまな種類のデータで応用できる GLM

6.2　例題：上限のあるカウントデータ

二項分布　応答変数の上限が定まっている場合に利用される。

Table 6.1: R 内で GLM 構築に使える確率分布の一部

	確率分布	乱数発生	`glm()` の `family`	よく使うリンク関数
（離散）	二項分布	`rbinom()`	`binomial`	`logit`
	ポアソン分布	`rpois()`	`poisson`	`log`
	負の二項分布	`rnbinom()`	`glm.nb()`	`log`
（連続）	ガンマ分布	`rgamma()`	`gamma`	`log`
	正規分布	`rnorm()`	`gaussian`	`identity`

6.4 ロジスティック回帰とロジットリンク関数

二項分布を用いた GLM のひとつ。ロジットリンク関数以外に、プロビットリンク関数や complementary log-log リンク関数などが用いられる。

6.4.1 ロジットリンク関数

ロジスティック関数 種子の生存確率 q_i について、

$$q_i = \mathrm{logistic}(z_i) = \frac{1}{1+\exp(-z_i)} \tag{6.1}$$

で、説明変数に x_i と f_i を用いて

$$z_i = \beta_1 + \beta_2 x_i + \beta_3 f_i \tag{6.2}$$

とする。これは、

$$\mathrm{logit}(q_i) = \log \frac{q_i}{1-q_i} = z_i \tag{6.3}$$

と**ロジット関数**として表すこともできる。

つまり、ロジスティック関数とロジット関数はお互いに逆関数の関係にある。

6.4.2 パラメータ推定

尤度関数の対数を取り、対数尤度関数

$$\log L(\{\beta_j\}) = \sum_j \left\{ \log \binom{N_i}{y_i} + y_i \log(q_i) + (N_i - y_i)\log(1-q_i) \right\} \tag{6.4}$$

を最大化する βs を探し出す[6]。

R による分析 もし各個体について

- `y` 生存数
- `N-y` 死亡数

のようにデータが得られていれば、

```
glm(cbind(y, N-y)~x+f, data=d, family=binomial)
```

とする。

一方、各個体ごとに `y = 1`（生存）、`y = 0` （死亡）のようにデータが得られていれば、単純に

```
glm(y~x+f, data=d, family=binomial)
```

としてよい。

[6] q_i は βs の関数である。

6.4.3 ロジットリンク関数の意味・解釈

式 (6.3) より左辺をオッズに変換して、

$$\frac{q_i}{1-q_i} = \exp(\beta_1 + \beta_2 x_i + \beta_3 f_i) \tag{6.5}$$

$$= \exp(\beta_1)\exp(\beta_2 x_i)\exp(\beta_3 f_i) \tag{6.6}$$

と書ける。つまり、x_i が 1 増えるとオッズは $\exp(\beta_2)$ 倍増え、$f_i = 1$ であれば（$f_i = 0$ のときに比べて）オッズは $\exp(\beta_3)$ 倍増える。

6.4.4 ロジスティック回帰のモデル選択

R の MASS パッケージの関数 stepAIC() を用いると、ネストしているモデルの AIC を自動的に比較しながら、AIC 最小のモデルを選択できる。

R による分析とモデル選択

```
> d <- read.csv("data4a.csv")
> head(d,n=5)
  N y     x f
1 8 1  9.76 C
2 8 6 10.48 C
3 8 5 10.83 C
4 8 6 10.94 C
5 8 1  9.37 C
> fit01 <- glm(cbind(y,N-y)~x+f, data=d, family=binomial)
> fit01

Call:  glm(formula = cbind(y, N - y) ~ x + f, family = binomial, data = d)

Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:     499.2
Residual Deviance: 123  AIC: 272.2
> library(MASS)
> stepAIC(fit01)
Start:  AIC=272.21
cbind(y, N - y) ~ x + f

       Df Deviance    AIC
<none>      123.03 272.21
- f     1   217.17 364.35
- x     1   490.58 637.76

Call:  glm(formula = cbind(y, N - y) ~ x + f, family = binomial, data = d)

Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:     499.2
Residual Deviance: 123  AIC: 272.2
```

6.5 交互作用項の入った線形予測子

線形予測子

$$\mathrm{logit}(q_i) = \beta_1 + \beta_2 x_i + \beta_3 f_i + \beta_4 x_i f_i \tag{6.7}$$

を考える。

R では、

```
glm(cbind(y, N-y)~x*f, data=d, family=binomial)
```

のようにすれば OK。

交互作用を入れない場合

```
(Intercept)           x          fT
    -19.536       1.952       2.022
```

と入れた場合

```
(Intercept)           x          fT         x:fT
  -18.52332     1.85251    -0.06376     0.21634
```

とでは、f_i の効果は随分異なるように見えるが、結果を図示するとほとんど変わらないことがわかる。交互作用項の解釈は数値だけ見ていては結構難しい。

- むやみに交互作用は入れないこと
- 個体差や場所差をモデルに組み込めば、交互作用は消えることが多い（第 7 章）

6.6 割り算値の統計モデリングはやめよう

以下のような作法を慎もう。

- 観測値を割り算して分析用データにする
- 観測値を変数変換する
- 複数の観測値をひとつの平均値に直してしまう

その理由は、

- 情報が失われる（300/1000 と 3/10 はどちらも 3 割だが、確からしさの度合いが違う）
- 変換後の値はどう分布する？（分母分子に誤差あり、カウントデータに 1 を足して対数変換...）

オフセット項　線形予測子の中でパラメータがつかない項

例えば、

- i 調査地番号
- A_i 面積
- x_i 明るさ
- y_i 植物個体数

のようなデータで植物個体の密度が明るさにどう左右されているか知りたいとする。このとき、植物個体の密度は y_i/A_i であるが、割り算値を作るのではなく、

$$\mathrm{density}_i = \frac{\lambda_i}{A_i} \tag{6.8}$$

を考え、

$$\lambda_i = A_i \times \mathrm{density} = A_i \times \exp(\beta_1 + \beta_2 x_i) = \exp(\beta_1 + \beta_2 x_i + \log A_i) \tag{6.9}$$

とモデル化する。この $\log A_i$ がオフセット項。

R による実行

```
glm(y~x, offset=log(A), family=poisson, data=d)
```

6.8 ガンマ分布の GLM

ガンマ分布の確率密度関数は

$$p(y \mid s, r) = \frac{r^s}{\Gamma(s)} y^{s-1} \exp(-ry) \tag{6.10}$$

で、

- s：shape パラメータ
- r：rate パラメータ（$1/r$：scale パラメータ）
- 平均：s/r
- 分散：s/r^2

である。ここで、

- y_i：花の重量
- x_i：葉の重量

というデータについて、y_i が平均 μ_i のガンマ分布に従っているとする。さらに、

$$\mu_i = Ax_i^b = \exp(a)x_i^b = \exp(a + b \log x_i), \quad \text{where } \exp(a) = A \tag{6.11}$$

と仮定する。つまり、

$$\log \mu_i = a + b \log x_i \tag{6.12}$$

となる。

R による実行

```
glm(y~log(x), family=Gamma(link="log"), data=d)
```

7 一般化線形混合モデル（GLMM）

データにばらつきをもたらす個体間の差異は、全てをデータとして定量化することはできない[7]。

一般化線形混合モデル　人間が測定できない、or しなかった個体差を組み込んだ GLM であり、複数の確率分布を部品とする統計モデル。

7.1 例題：GLM では説明できないカウントデータ

- i：観測個体番号
- n：観察種子数（全個体統一）
- y_i：生存種子数
- x_i：葉の数

[7] 仮に説明変数が全て同じ値であったとしても、応答変数にはばらつきが発生する（原因不明の差異）。説明変数以外は全部均質というわけではないから。

をもとに、

$$\mathrm{logit}(q_i) = \beta_1 + \beta_2 x_i \tag{7.1}$$

$$p(y_i \mid \beta_1, \beta_2) = \binom{n}{y_i} q_i^{y_i} (1-q_i)^{n-y_i} \tag{7.2}$$

なるモデルを考える。
→ はたして妥当か？

7.2 過分散と個体差

全ての個体を（観測されていない説明変数については全て）均質と考え、どの個体の生存種子数についても同じ二項分布で説明できると仮定してモデルを当てはめると、データのばらつきはモデルから予測されたものよりも大きくなる。

観測されていない個体差として、

- 生物的（個体差）
- 非生物的（場所差）

な要因によるものが挙げられる。

すべて観測することが不可能であるなら、個体差や場所差を原因不明のまま統計モデルに取り込む必要がある。

7.3 一般化線形混合モデル

7.3.1 個体差を表すパラメータの追加

$$\mathrm{logi}(q_i) = \beta_1 + \beta_2 x_i + r_i, \quad r_i \sim N(0, s) \tag{7.3}$$

7.4 一般化線形混合モデルの最尤推定

式 (7.3) で最尤推定できるのは、β_1, β_2, s の 3 つ。r_i については、個体ごとの尤度 L_i において積分消去してしまう。

$$L_i = \int_{-\infty}^{\infty} p(y_i \mid \beta_1, \beta_2, r_i) p(r_i \mid s) dr_i \tag{7.4}$$

これは、モデル分布である二項分布と、個体差を表す正規分布とを混合していることに相当する。

かつては**準尤度（quasi likelihood）**を用いていたが、現在では利用する利点がないので使われていない。

7.5 現実のデータ解析には GLMM が必要

GLMM の考え方が必要になるかを決めるポイント

- 同じ個体・場所などから何度もサンプリングしているか
- 個体差や場所差が識別できてしまうようなデータの取り方をしているか

7.6 反復・擬似反復と統計モデルの関係

反復（独立した反復） 各個体や場所からひとつだけのデータを取る。
{A 校の B さんから 1 回、C 校の D さんから 1 回、E 校の F さんから 1 回、...}

擬似反復（pseudo replication） 同じ個体や場所から複数のデータを取る。
{A 校から B さん、C さん、D さん、... } {E 校から F さん、G さん、H さん、...}
擬似反復の場合には個体差が推定可能であり、かつその影響を考慮しなければ推定結果に偏りが生じる。

8　マルコフ連鎖モンテカルロ法とベイズ統計モデル

8.2　ふらふら試行錯誤による最尤推定

尤度を最大化する q を求めたいとする。このとき、

1. 初期値を設定する $q^{(0)}$
2. $q^{(t+1)}$ の候補として、$q^{(t)}$ の右隣か左隣の値をランダムに選び、更新値の候補 q' とする
3. 対数尤度を評価し、$\log L(q') > \log L(q^{(t)})$ なら、$q^{(t+1)} = q'$ とする

8.3　メトロポリス法

上記のアルゴリズムに、

4. 対数尤度を評価し、$\log L(q') < \log L(q^{(t)})$ であっても、$r = L(q')/L(q^{(t)})$ の確率で $q^{(t+1)} = q'$ とする

を追加する。

もし**詳細釣り合い条件**

$$p(q^{\text{new}} \mid \boldsymbol{Y})p(q^{\text{new}} \to q) = p(q \mid \boldsymbol{Y})p(q \to q^{\text{new}}) \tag{8.1}$$

が成り立っていれば、

$$\sum_q p(q^{\text{new}} \mid \boldsymbol{Y})p(q^{\text{new}} \to q) = \sum_q p(q \mid \boldsymbol{Y})p(q \to q^{\text{new}}) \tag{8.2}$$

$$p(q^{\text{new}} \mid \boldsymbol{Y}) = \sum_q p(q \mid \boldsymbol{Y})p(q \to q^{\text{new}}) \tag{8.3}$$

となる。ただし、

$$p(q \mid \boldsymbol{Y}) = \frac{L(q)}{\sum_q L(q)} \tag{8.4}$$

$$\propto L(q) \tag{8.5}$$

なる**定常分布**であるとする。

補足

更新によって尤度が改善されない場合について考えると、

$$p(q \to q^{\text{new}}) = 0.5 \times \frac{L(q^{\text{new}})}{L(q)} (= 0.5r) \tag{8.6}$$

$$p(q^{\text{new}} \to q) = 0.5 \times 1 \tag{8.7}$$

を整理すると、

$$L(q^{\text{new}})p(q^{\text{new}} \to q) = L(q)p(q \to q^{\text{new}}) \tag{8.8}$$

となる。

補足ここまで

初期値から一定期間を経て q のサンプリングが安定してくれば、それを**定常分布**からのサンプリングとみなすことができる。

メモ

更新値を受容する確率に従って q を更新し続けていくと、それによって得られた値自体が尤度 $L(q)$ に比例した q の事後分布からのサンプルに等しい。

メモここまで

9　GLM のベイズモデル化と事後分布の推定

9.2　GLM のベイズモデル化

個体 i の種子数 y_i が平均 λ_i のポアソン分布に従うとする。この平均は、体サイズ x_i を用いて $\lambda_i = \exp(\beta_1 + \beta_2 x_i)$ とする。

事後分布は、

$$p(\beta_1, \beta_2 \mid \boldsymbol{Y}) \propto p(\boldsymbol{Y} \mid \beta_1, \beta_2) p(\beta_1) p(\beta_2) \tag{9.1}$$

となる。

9.3　無情報事前分布

無限区間の一様分布は積分しても 1 にならないので、例えば

- $-10^9 < \beta_* < 10^9$ の一様分布
- 平均ゼロで標準偏差がとても大きい平たい正規分布

のいずれかを無情報的な事前分布として指定することを考える。以降は、「平べったい正規分布」を利用する。

9.4　ベイズ統計モデルの事後分布の推定

BUGS のコード

```
model{
  for(i in 1:N){
      y[i] ~ dpois(lambda[i]) # 種子数は平均 lambda[i] のポアソン分布に従う
      log(lambda[i]) <- beta1 + beta2 * (X[i]-Mean.X) # 対数リンク関数（X[i] は高速化のため中心化）
}
  beta1 ~ dnorm(0,1.0E-4) # 事前分布（dnorm(mean, tau) で tau は分散の逆数）
  beta2 ~ dnorm(0,1.0E-4) # 事前分布
}
```

メモ

x_i を中心化すると（標準化もすると）MCMC が高速化する、という理由がよくわからない。

また、本文中では中心化の前後で結果は本質的に同じであるとされているが、切片の値は本質的に変わるはず。

メモここまで

R で WinBUGS を操作する　R の `R2WinBUGS` パッケージを用いて、データ・初期値・MCMC サンプリングの回数・BUGS コードファイル名などを R から WinBUGS に渡してやり、結果を R に渡してもらう。

9.4.3　どれだけ長く MCMC サンプリングすればいいのか？

初期値の異なる複数のサンプリング（サンプル列：chain）を比較することが有用。

サンプル列が 3 本以上ある場合、$\hat{R}$ が**収束診断**の指標として用いられる。サンプル列ごとの分散の平均を W、周辺事後分布の分散を

$$\widehat{\mathrm{var}}^{+} = \frac{n-1}{n} W + \frac{1}{n} B \tag{9.2}$$

として、

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}}{W}} \tag{9.3}$$

と定義する。

9.6 複数パラメータの MCMC サンプリング

複数パラメータを同時に更新するのは簡単ではないので、β_1 と β_2 を交互に更新することを考える。

9.6.1 ギブスサンプリング

全条件付き分布（FCD）[8]からのランダムサンプルを利用する。

- 各 MCMC ステップにおいてもとの値と更新された値の相関がより小さい
- MCMC サンプリングの詳細を指定しなくても良い[9]

10 階層ベイズモデル

10.1 例題：個体差と生存種子数（個体差あり）

100 個体から 8 つの種子を採集し、生存種子数 y_i を調べる。

→ 二項分布が使えそうだが... 平均種子数から計算できる分散は、二項分布に従うと仮定した場合よりもはるかに大きい。

→ 種子生存確率 q が全個体で共通していると仮定することに問題がありそう。

10.2 GLMM の階層ベイズモデル化

各個体の種子生存確率 q_i について、

$$\mathrm{logit}(q_i) = \beta + r_i, \quad r_i \sim N(0, s) \tag{10.1}$$

とする。

切片 β の事前分布は無情報事前分布（平たい正規分布）にするとして、s はどのようにすればよいのだろう？

最尤推定では s の点推定値を特定したが、ベイズ統計モデルでは s にも分布を仮定する。そこで、$p(s)$ を $0 < s < 10^4$ の連続一様分布とする。$p(s)$ は、$p(r_i \mid s)$ のパラメータの事前分布であり、**階層事前分布**、**超事前分布**と呼ばれ、s は**超パラメータ**と呼ばれる。

改装事前分布を使うベイズ統計モデルを**階層ベイズモデル**という。

10.3 階層ベイズモデルの推定・予測

BUGS のコード

```
model{
  for(i in 1:N){
    Y[i] ~ dbin(q[i],8) # 二項分布
    logit(q[i]) <- beta + r[i] # 生存確率
  }
  beta ~ dnorm(0,1.0E-4) # 無情報事前分布
  for(i in 1:N){
    r[i] ~ dnorm(0, tau) # 階層事前分布
  }
  tau <- 1/(s*s) # tau の定義（分散の逆数）
  s ~ dunif(0, 1.0E-4) # 無情報事前分布
}
```

[8] 他の変量全てを定数とする一変量確率分布

[9] ?

10.4　ベイズモデルで使うさまざまな事前分布

無情報事前分布と階層事前分布のいずれを選ぶべきなのか？

- 大域的なパラメータ β：無情報事前分布を指定
- 局所的なパラメータ r_i：個々に無情報事前分布を指定するのではなく、$\{r_i\}$ 全体のばらつきを変えられる階層事前分布を指定

10.5　個体差 + 場所差の階層ベイズモデル

以下の条件で、得られたデータから植物の種子数をモデリングしたい。

- 10 個の植木鉢（5 つが無処理、5 つが施肥処理）
- 各植木鉢に 10 個体の植物

各植木鉢の種子数と分散を算出すると、「平均 < 分散」となっており、過分散が生じている。
→ 個体差と場所差（植木鉢の差）によって生じた過分散（双方が擬似反復）
個体 i の種子数 y_i を平均 λ_i のポアソン分布で表現するが、さらに

$$\log \lambda_i = \beta_1 + \beta_2 f_i + r_i + r_{j(i)} \tag{10.2}$$

$$r_i \sim N(0, s) \tag{10.3}$$

$$r_{j(i)} \sim N(0, s_p) \tag{10.4}$$

とする。i は j にネストしている。
各事前分布を、

- βs：無情報事前分布（平たい正規分布）
- s と s_p：無情報事前分布（$0 - 10^4$ までの一様分布）

とする。

BUGS のコード

```
model{
  for(i in 1:N.sample){
    Y[i] ~ dpois(lambda[i]) # モデル分布（ポアソン分布）
    log(lambda[i]) <- beta1 + beta2 * F[i] + r[i] + rp[Pot[i]] # 対数リンク関数と線形予測子
  }
  beta1 ~ dnorm(0, 1.0E-4) # 切片の事前分布
  beta2 ~ dnorm(0, 1.0E-4) # 傾きの事前分布
  for(i in 1:N.sample){
    r[i] ~ dnorm(0, tau[1]) # 個体差の分布
  }
  for(j in 1:N.pot){
    rp[j] ~ dnorm(0, tau[2]) # 場所差の分布
  }
  for(k in 1:Ntau){
    tau[k] <- 1.0 / (s[k]*s[k]) # 個体差・場所差の分布の分散の逆数
    s[k] ~ dunif(0, 1.0E+4) # 個体差・場所差の標準偏差の事前分布（超事前分布）
  }
}
```

- BUGS では因子型変数をそのままあつかえないので、施肥処理 `F[i]` はダミー変数化しておく必要がある
- 植木鉢の効果 `rp[j]` は、個体 i が 植木鉢 j のものであることを `Pot[i]` で指定しておく必要がある

11　空間構造のある階層ベイズモデル

ここまでの場所差は、それぞれ独立に決まると仮定してきた。しかし、植木鉢のように違いが影響を及ぼさないものではなく、区画を区切ったもののように、隣り合ったものについてはよく似た環境で独立とはいえないような場合はどうすればよいのか？

→ 場所差の**空間相関（spatial correlation）**を考慮

空間相関... 距離の遠近に依存して、場所の類似性が弱くなったり強くなったりする傾向

11.1　例題：一次元空間上の個体数分布

50 個の調査区画が 1 本の直線上に等間隔に配置されており、生物の個体数を記録したとする。

→ 調査区画ごとに個体数を並べると、なだらかにへんかしていることがわかる。

→ なだらかに変化する局所密度を考える。（隣り合う区画はよく似ている。）

11.2　階層ベイズモデルに空間構造を組み込む

区画 j における個体数を

$$\log \lambda_j = \beta + r_j \tag{11.1}$$

とおく。ただし、r_j は（特に隣り合った区画同士では）独立な分布に従うと仮定できそうにはない。

そこで、

- 区間の場所差は「近傍」区間の場所差にしか影響されない
- 区間 j の「近傍」は有限個 n_j であり、分析者が指定する
- 「近傍」の影響は $1/n_j$

なる仮定をおいてみる。

条件つき自己回帰（CAR）モデル　仮に $n_j = 2$ とし、

$$p(r_j \mid \mu_j, s) = \sqrt{\frac{n_j}{2\pi s^2}} \exp\left\{-\frac{(r_j - \mu)^2}{2s^2/n_j}\right\} \tag{11.2}$$

$$\mu_j = \frac{r_{j-1} + r_{j+1}}{2} \tag{11.3}$$

つまり

$$p(\{r_j\} \mid s) \propto \exp\left\{-\frac{1}{2s^2}\sum_{j\sim j'}(r_j - r'_j)^2\right\} \tag{11.4}$$

を考えてみる（**intrinsic Gaussian CAR モデル**）。

11.3　空間統計モデルをデータにあてはめる

BUGS のコード

```
model{
  for(j in 1:N.site){
    Y[j] ~ dpois(mean[j])
    log(mean[j]) <- beta + r[j]
  }
  r[r:N.site] ~ car.normal(Adj[], Weights[], Num[], tau)
  beta ~ dnorm(0, 1.0E-4)
  tau <- 1 / (s*s)
  s ~ dunif(0, 1.0E+4)
}
```

11.4　空間統計モデルが作り出す確率場

確率場　相互作用する確率変数で埋め尽くされた空間のこと

11.5　空間相関モデルと欠測のある観測データ

観測データの欠測部分を予測する用途にも空間相関モデルを用いることができる。