一般化線形モデル(GLM)入門

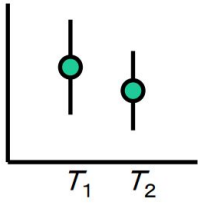
2018/01/14 篠田昌和

1. 一般線形モデル
2. 一般化線形モデルの概観
3. 一般化線形モデルの定義
4. GLMのモデル選択
5. 尤度比検定

一般線形モデルとは

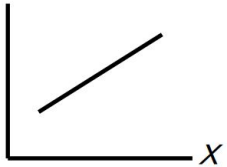
分散分析 ANOVA

カテゴリー要因の影響を分析



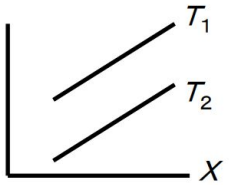
回帰分析 Regression

量的要因の影響を分析



共分散分析 ANCOVA

量的要因とカテゴリー要因の影響を分析（一つずつの場合）



線型モデル
Linear model
一般線形モデル
General Linear Model

anova= analysis of variance

単回帰分析

式で書くと

$$y_i = a + bx_i + e_i \quad (i = 1,2,\ldots,n)$$

ただし e_i は等分散の正規分布にしたがう

これは次のようにも書ける

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

→ひとまとめにして書くと

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

重回帰分析

2変量の場合，式で書くと

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + e_i \quad (i = 1,2,\ldots,n)$$

ただし e_i は等分散の正規分布にしたがう

これは次のようにも書ける

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

→ひとまとめにして書くと

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

一元配置分散分析

（例）三つの水準を設けて，完全無作為法で
それぞれの処理を2回ずつ反復した場合

$$y_{ij} = T_i + e_{ij} \quad (i = 1,2,3; j = 1,2)$$

ただしe_{ij}は等分散の正規分布にしたがう

全体の基準値μをくくり出して表現すると

$$y_{ij} = \mu + T_i + e_{ij} \quad (i = 1,2,3; j = 1,2)$$

これは

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

→ひとまとめにして書くと

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

一般線形モデルの定義

以上のすべては次のような式に書けた

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

y：従属変数のベクトル
X：デザイン行列（実験計画などに依存して決まる）
b：パラメーターのベクトル
e：誤差のベクトル（等分散正規分布にしたがう）

これを**線型モデル**，あるいは**一般線型モデル**と呼ぶ。

同じ式で書けるため，「推定・検定の手順は基本的に同じ」
→回帰分析と分散分析は区別する必要がない。
→回帰分析と分散分析が混合した分析も可能。

1. 一般線形モデル
2. **一般化線形モデルの概観**
3. 一般化線形モデルの定義
4. GLMのモデル選択
5. 尤度比検定

一般化線形モデルとは

Nelder and Wedderburn (1972)

線型モデル
Linear model
一般線形モデル
General Linear Model

一般化線形モデル
Generalized Linear Model

誤差が等分散正規分布と仮定

ロジスティック回帰 Logistic regression

誤差が二項分布と仮定

対数線形モデル loglinear model

誤差が多項分布と仮定

何を拡張するのか

線型モデル　　$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$　　**e は等分散の正規分布にしたがう**

これを言い換えると

1. 観測値の分散
   観測値$\mathbf{y}$ は期待値$E(\mathbf{y})$の周りに等分散正規分布に従って分布
2. 期待値$E(\mathbf{y})$ に関して$E(\mathbf{y}) = \mathbf{X}\mathrm{b}$

この二つを拡張する。

何を拡張するのか

拡張ポイント1:観測値の分散

正規分布の必要はなく、指数分布族に従えば良い。
（ポアソン分布，二項分布，ガンマ分布など）

拡張ポイント2:線形でなく、線形を関数で処理する

期待値$E(\mathbf{y})$ に関して$E(\mathbf{y}) = g(\mathbf{X}\mathrm{b})$
$g(x)$は指数関数やロジスティック関数などの「単調関数」
$g(x)$の逆関数をリンク関数と呼ぶ。

これをfと記すと$f[\mathrm{E}(\mathrm{y})] = \mathbf{X}\mathrm{b}$
期待値$E(\mathbf{y})$ をリンク関数で変換したものが線形関数。

一般化線形モデルの例1

(1) 観測値yは$E(\mathbf{y})$の周りに二項分布にしたがって分布

(2) 期待値$E(\mathbf{y})$に関して $$E(\mathbf{y}) = \frac{\exp(\mathbf{Xb})}{1 + \exp(\mathbf{Xb})}$$

Xbをロジスティック関数で変換したもの

ロジスティック関数の逆関数はロジット関数：
→ リンク関数がロジット関数

$$\log_e\left(\frac{E(\mathbf{y})}{1 - E(\mathbf{y})}\right) = \mathbf{Xb}$$

→これはロジスティック回帰

一般化線形モデルの例2

(1) 観測値yは$E(\mathbf{y})$の周りにポアソン分布にしたがって分布

(2) 観測値$E(\mathbf{y})$に関して$E(\mathbf{y}) = \exp(\mathbf{Xb})$

$$\log_e[E(\mathbf{y})] = \mathbf{Xb}$$

リンク関数は対数関数

→これはポアソン回帰の一種

尤度の最大値を見つける方法

誤差に等分散正規分布を仮定する場合（線形モデル）は

行列計算で求まった $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

一方、正規分布以外の誤差の場合はそう簡単には求まらない。

1. 試行錯誤法（ニュートン・ラフソン法，スコア法など）
2. 対数尤度が最大となるパラメーターを探す。
3. 対数尤度をパラメーターで微分してゼロとなるパラメー ターを探す。

1. 一般線形モデル
2. 一般化線形モデルの概観
3. **一般化線形モデルの定義**
4. GLMのモデル選択
5. 尤度比検定

一般化線形モデル（GLM）の3要素

一般化線形モデル（GLM）は、
・誤差構造（error structure）
・線形予測子（linear predictor）
・リンク関数（link function）
の 3 つの要素によって構成されている。

一般化線形モデル - 誤差構造

誤差構造は、従属変数が従う確率分布のこと。

例

・従属変数が身長ならば、その誤差構造は正規分布

・従属変数が種子数を数えたカウントデータならば、その誤差構造はポアソン分布

このように、一般化線形モデル構築時、どのような誤差構造を選ぶのかを、従属変数の性質に応じて決める必要がある。

一般化線形モデル - 誤差構造

一般化線形モデルで用いる誤差構造は、**指数型分布族**と呼ばれている。

指数型分布族の密度関数・確率質量は、$f(x;\theta)=\exp(\eta(\theta)T(x)-A(\theta)+B(x))$

の形で表せる。

正規分布、指数分布、ガンマ分布、ポアソン分布、二項分布など。

一般化線形モデル - 線形予測子

線形予測子は、パラメーター β とデザイン行列 **x** の積で表される。

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}^T \beta$$

x は**デザイン行列**と呼ばれ、確率変数とモデルのパラメーター間の関係を制御する。

一般化線形モデル - 線形予測子

例えば、以下の **Y** = **xβ** におけるデザイン行列 **x** において、確率変数は Y_1 は β_1 と β_2 の2 つのパラメーターに影響されるが、Y_2 は β_2 のみに影響されることを意味する。

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

一般化線形モデル - 線形予測子

デザイン行列は 0 または 1 のように因子で構成されることも、1.2 や 2.3 などのようにパラメーターの加重として構成されることもできる。
-> デザイン行列を因子とするか、加重とするかは解析するデータの特徴に合わせて決める必要がある。
同じデータに対して、どんな結果を期待したいかによってもデザイン行列の作り方が異なってくる。
-> どんな結果を期待しているのかを予め決定した上で実験を行うべき。

リンク関数

従属変数が正規分布に従わないとき、E[Y] = Xβ の式でモデル化すると、正確さが失われる。
-> そこで、従属変数をある関数 G で変換してから、モデル化することで、モデルの正確さが向上する。

G(E[Y]) = Xβ
関数 G はリンク関数と呼ばれている。
一般に、誤差構造が決まれば、リンク関数も自動的に決まることが多い。

リンク関数

誤差構造	リンク関数
正規分布	$g(\mu) = \mu$
指数分布	$g(\mu) = -\mu$
ポアソン分布	$g(\mu) = log\mu$
二項分布	$g(\mu) = log\frac{\mu}{1-\mu}$

リンク関数の例

植物の種子 個体$_i$ のサイズx_i 平均種子数λ_i

$\lambda_i=\exp(\beta_1+\beta_2 x_i)$

とする。

$\log\lambda_i=\beta_1+\beta_2 x_i$

$\beta_1+\beta_2 x_i$は線形予測子。β_1,β_2の線型結合になっているから。

(λ_iの関数) = (線形予測子)　なので左辺の関数はリンク関数

この場合は対数関数が指定されていることから、対数リンク関数と呼ぶ。ポアソン回帰をする場合は対数を指定。

一般線形モデルと一般化線形モデル

$\log\lambda_i=\beta_1+\beta_2 x_i$
$\rightarrow\lambda_i=\beta_1+\beta_2 x_i$

のように、平均が線形予測子に等しい＝リンク関数がない場合はこの状態を恒等リンク関数(identity link function)と呼ぶ。
→ これは一般化(generalized)ではない線形モデル、
一般線形モデル(general linear model)と呼ぶ。

1. 一般線形モデル
2. 一般化線形モデルの概観
3. 一般化線形モデルの定義
4. GLMのモデル選択
5. 尤度比検定

最尤推定の復習

ポアソン分布 $$p(y|\lambda) = \frac{\lambda^y exp(-\lambda)}{y!}$$

尤度は全ての個体の確率分布の積なので

$$L(\lambda) = \prod_i p(y_i|\lambda)$$

対数尤度とると

$$logL(\lambda) = \sum_i (y_i log\lambda - \lambda - \sum_k^{y_i} logk)$$

対数尤度`logL`を最大化するパラメータを探すのが最尤推定法。

逸脱度と残差逸脱度

逸脱度　D

$$D = -2logL^*$$

はあてはまりの悪さを表現する指標。最大対数尤度*(−2)

最小逸脱度 フルモデルをあてはめたときのD

残差逸脱度 D ー 最小のD

パラメータを増やせば、残差逸脱度はどんどん小さくなる。

AICは予測の良さを重視

モデル選択基準にAIC(Akaike's information criterion)を考える。AICはモデルの当てはまりの良さ(goodness of fit)ではなく予測の良さ(goodness of prediction)を重視するモデル選択基準。

パラメータ数kとすると

AIC= -2{(最大対数尤度)-(最尤推定したパラメータ数)}

= -2(log L^* - k)

= D+2k →AICが最小のモデルが良いモデル。

AICは予測の良さを重視

最大対数尤度 $\log L^*$

・たまたま得られた観測データへの当てはまりの良さ。

平均対数尤度 E(logL)

・繰り返し観測データをサンプリングして真のモデルへの予測の良さを評価したもの。

真の統計モデルは不明なので平均対数尤度は現実には分からないが、一定の
バイアス **$b = \log L^* - E(\log L)$** があり、補正の必要がある。
→ **$b = \log L^* - E(\log L) - k$** (k は最尤推定するパラメータ数)
だと分かっているので、
一定モデル $\log\lambda_i=\beta_1$ では、k=1なので、-2をかけた
$AIC=-2(\log L^* - 1)$ となる。

AICは予測の良さを重視

・モデルを複雑化するだけで、観測データへの当てはまりの良さを表す
対数尤度$\log L^*$ は改善されるので、モデルの複雑さを考慮した
AICでモデル選択する必要がある。

・モデル選択基準AICは、統計モデルの予測の良さである平均対数尤度の推定値
であり、これは最大対数尤度 $\mathrm{og}L^*$ のバイアス補正によって評価される。

1. 一般線形モデル
2. 一般化線形モデルの概観
3. 一般化線形モデルの定義
4. GLMのモデル選択
5. **尤度比検定**

AICによるモデル選択と統計モデル検定の比較

統計モデルの検定

解析対象のデータを確定

↓

統計モデル設計

(帰無仮説・対立仮説)

↓

帰無仮説棄却の危険率を評価

↓

帰無仮説棄却の可否を判断

AICによるモデル選択

解析対象のデータを確定

↓

統計モデル設計

(単純モデル・複雑モデル)

↓

モデル選択基準AICの評価

↓

予測の良いモデルを選ぶ

ネストした統計モデルたちのパラメータの最尤推定計算

ネストとは

$$log\lambda_i = \beta_1$$ 一定モデル

$$log\lambda_i = \beta_1 + \beta_2 x_1$$ xモデル

一方のモデルが他方のモデルに含まれる場合、ネストしている
という。ここで一定モデルとxモデルでは、xモデルでβ_2=0 とおくと
一定モデルになる。なので一定モデルとxモデルはネストしている。

このようにネスト関係にあるモデルでAICの推定値を比較して、
どちらかより望ましいかを評価する。

尤度比(likelihood ratio)検定

$$log\lambda_i = \beta_1$$ 一定モデル

$$log\lambda_i = \beta_1 + \beta_2 x_1$$ xモデル

尤度比＝L_1^*/L_2^*= (一定モデルの最尤尤度)/(xモデルの最尤尤度)

さらに尤度の対数をとり、-2 をかけて逸脱度の差を計算。

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

$\Delta D_{1,2} = D_1 - D_2$なので二つのモデルの逸脱度の差となっています。
これを検定統計量として使用する。

帰無仮説・対立仮説と2種類の過誤

帰無仮説：一定モデル $k=1, \beta_2=0$
対立仮説：xモデル $k=2, \beta_2 \neq 0$

↓帰無仮説は	観察された逸脱度の差は めったにない差 (帰無仮説を棄却)	よくある差 (帰無仮説を棄却できない)
真のモデルである	第一種の過誤	(問題なし)
真のモデルでない	(問題なし)	第二種の過誤

・帰無仮説が真のモデルの場合

データが一定モデルから生成されたのに「逸脱度の差が大きいのでxモデルの方が良い、帰無仮説は正しくない」と判断する第一種の過誤(type Ⅰ error)

帰無仮説・対立仮説と2種類の過誤

・帰無仮説が真のモデルでない場合

　データがxモデルから生成されたのに「逸脱度の差が小さいのでxモデルは意味なく複雑。一定モデルで観察されたパターンを説明できるから、帰無仮説は正しい」と判断する
第二種の過誤(typeⅡ error)

→実際には第一種の過誤のみ検討する。
→二つのモデルの逸脱度の差が大きいので、帰無仮説は正しくないとする。

第一種の過誤の重視は、検定の非対称性という。

参考文献

国立研究開発法人農業・食品産業技術総合研究機構
農業環境変動研究センター　山村光司
数理統計短期集合研修　一般化線形モデル
http://cse.naro.affrc.go.jp/yamamura/Images/kenshuu_slide_glm_2015_applied.pdf

データ解析のための統計モデリング入門
https://www.amazon.co.jp/dp/400006973X