Introduction と Likelihood theory

飯島勇人[*†‡§]

2007 年 4 月 23 日

目次

1 Introduction

第 1 章では、本書の中核概念である線形モデルのうち、誤差分布に正規分布を仮定する、いわゆる重回帰を用いてのデータ解析例を示している。`faraway` の中には、`gavote` という、2000 年のアメリカ大統領選挙のときの Georgia 州でのデータが含まれている。列は、左から、投票設備、経済

* 北海道大学大学院農学院専門研究員
† 連絡先: hayato-i@for.agr.hokudai.ac.jp または `http://www.geocities.jp/iijima_web/index.html`
‡ 本ゼミのサポートページ: `http://www7.atwiki.jp/hayatoiijima/pages/33.html`
§ この文章は LaTeX で作成しました。わーどなどではけっしてありません。

状態、アフリカ系アメリカ人の割合、田舎か都会か、州都 Atlanta の町かどうか、Gore に投票した人の数、Bush に投票した人の数、他の候補に投票した人の数、有効投票数、総投票数である。

このデータフレームについて、右の二列（votes と ballots）の差は、無効票の数を示している。本章では、この、無協票数が総投票数に占める割合が、どのような要因に影響されているかを検討する。

1.1 データの見方（すぐになんでも検定すんな！）

すべてのデータ解析において、最初にすべきことは、図や要約数で**データを見る**ことである。

まず、gavote には、主たる解析対象である、無効票の割合、という列がないので作る。ついでに、Gore に投票した人の割合も算出する。

```
> gavote$undercount <- (gavote$ballots - gavote$votes)/gavote$ballots
> gavote$pergore <- (gavote$gore + gavote$bush)/gavote$votes
```

は、このデータフレームから何を読み取るか？

- summary(gavote)：データフレームの各列に関し、基礎統計量を示す。
- hist(gavote$undercount)：ヒストグラムを描く。データの分布を見る基本中の基本。
- pie(table(gavote$equip))：円グラフを描く。データの相対的関係を見るときに使う。
- plot(pergore ~ perAA, gavote)：散布図を描く。2 変量の関係を把握するのに使う。
- boxplot(undercount ~ equip, gavote)：箱ひげ図。カテゴリーごとのデータの関係を見るのに使う。
- xtabs(~ atlanta + rural, gavote)：2 つのカテゴリカルデータに含まれるデータ数を調べるときに使う。1 つのカテゴリカルデータの場合は table() で。
- cor(gavote[, c(3, 10, 11, 12)])：変数間の相関係数を示す。

1.2 線形モデルの構築

線形モデルとは、

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

のように、独立変数が線形に結合したモデルのことである。undercount を説明するモデルとして、今回は、

$$\text{undercount} = \beta_0 + \beta_1 \text{pergore} + \beta_2 \text{perAA} + \varepsilon$$

というモデルを作成した。R においては、

```
> lmod <- lm(undercount ~ pergore + perAA, gavote)
> summary(lmod)
```

とすることで、モデルの当てはめ、および結果の出力を行うことができる。

- モデルの結果予測される undercount の値は、> predict(lmod) によって取り出すことができる。
- モデルの当てはまりを見るもう一つの方法として、predict の値と、従属変数の関係を検討する方法もある。例えば、> plot(predict(lmod), gavote$undercount) など。

なお、summary() で出力できるデータのうち、モデルの当てはまりに関する R-Squared は通常の R^2 値、Adjusted R-squared は自由度調整済み R^2 値であり、R^2 は、

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y}_i)^2} = 1 - \frac{RSS}{TSS}$$

自由度調整済み R^2 値は、

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

で定義されている。

1.3 モデルの解釈

唐突だが、以下のようなモデルを作成したとする。

```
> gavote$cpergore <- gavote$pergore - mean(gavote$pergore)
> gavote$cperAA <- gavote$perAA - mean(gavote$perAA)
> lmodi <- lm(undercount ~ cperAA + cpergore*rural + equip, gavote)
```

このモデルの解釈をしてみよう。

都市部に向かうほど、undercount は少なくなる傾向がある。また、線形モデルにカテゴリー変数を投入した場合、通常アルファベット順で最初となるカテゴリー名の係数が 0 となり、それに対する関係として、他の変数の係数が推定される。例えば、equipOS-PC の係数は 0.01564 であり、これは、他の変数が固定されている場合、equipLEVER よりも undercount が 1.56% 高くなると考えられる。

1.4 信頼区間

信頼区間とは、母数が存在すると考えられる範囲であり、設ける水準の程度により、95% 信頼区間、99% 信頼区間などがある。

各変数の係数の 95% 信頼区間は、R では、confint() で容易に取り出すことができる（もちろん引数で level=0.99 とすれば、99% 信頼区間で取り出すことができる）。

1.5 モデルの妥当性

通常の線形回帰では、誤差の正規性や等分散性が仮定されている。これらが本当に当てはまっているかを確認するために、R では便利な関数が用意されている。

```
> plot(lmodi)
```

のように、線形回帰の結果を`> plot()` することで、回帰診断図を作成することができる。

1.5.1 残差–予測値プロット

残差の傾向をチェックするのに使う。残差の分布が曲線形になっていれば、何らかの変数変換が必要である。誤差の等分散性をチェックするのにも使える。

1.5.2 調整残差–予測値プロット

これも残差の傾向をチェックするのに使う。

1.5.3 Q–Q プロット

残差の順位と標準化された残差の関係を見る。正規性が満たされているなら、点は一直線上に並ぶはずである。

1.5.4 Cook の距離プロット

Cook の距離は、全てのデータを用いた場合と、1 つだけデータを除いた場合の回帰の予測値の変動を示す。そのため、Cook の距離が大きい値は異常値である可能性がある。一般に Cook の距離が 0.5 以上の場合、異常値であることを疑う必要がある。`cooks.distance()` を使うと、Cook の距離を取り出すことができる。

1.6 より信頼できる回帰

はずれ値と思われるデータをどうするか？

- そのデータを除く。
- そのデータのエラーの大きさを考慮して回帰を行う。

1 番目に関しては、除く明確な理由がない場合、行いにくい。2 番目の方法は、エラー（すなわち予測値からのズレ）が大きい値の重みを低くし、回帰させる方法である。R の中では、`> library(MASS)` で呼び出せる、`rlm()` という方法で行える（あんまし聞いたことはないが）。

1.7 重みつき最小二乗法

今回の undercount のような変数は、選挙権所持者数 ballots に依存してばらつきが大きくなると考えられる。そのため、誤差の等分散性を満たすため、従属変数のばらつきに影響を与える要因で重み付けをして回帰する方法がある。これを、重みつき最小二乗法という。

R の中では、lm() の引数に weights をとることで実行できる。

```
> wlmodi <- lm(undercount ~ cperAA + pergore*rural + equip, gavote, weights=ballots)
```

1.8 変換

普通の線形回帰の場合、回帰の仮定（誤差の等分散性、正規性）を満たすために、変数変換を行うことがある。変数変換を行うかどうか判断する基準として、

- Box–Cox 変換：library(MASS) に収録されている boxcox() 関数を使い、データを正規分布に最も近づける λ を探索する。そして y^λ のように従属変数を変換する。λ が 0 の場合は、$\log y$ と変換する。
- 回帰診断図を作成し、どのような変換方法がよいか探索する（log など）。

などがあげられる。

1.9 変数選択

1.9.1 AIC による選択

線形モデルのモデル選択方法として一般に用いられるのは、Akaike Information Criterion（AIC）である。AIC は以下のように定義される。

$$\mathrm{AIC} = -2\mathrm{maximumloglikelihood} + 2p$$

p は変数の数である。R の中では、線形モデルの実行結果を stepAIC() に投入することで、実行できる。

1.9.2 F 検定による選択

もう一つのモデル選択方法は、F 検定である。検定による方法は、大きいモデル（すなわち多くの独立変数を含んだモデル）から、変数を一つ抜いてよいかを検討することができる。この方法は AIC による方法よりも推奨されていないが、

- 変数の階層性を考慮したい（respect restrictions of hierarchy）
- ある変数は必ずモデルの中に含めたい。

という状況では有効である。R の中では、

```
> drop1(model, test="F")
```

によって実行することができる。

1.10 結論

最終的に決定されたモデルに関して、モデルの解釈をしてみよう。最終的に決定されたモデルは以下のようである。

```
> finalm <- lm(undercount ~ equip + econ + perAA + equip:econ + equip:perAA, gavote)
```

交互作用を解釈するためには、予測値を算出すことが有効である。ここでは、perAA が 0.233（中央値）の場合の、equip と econ の全組み合わせにおける undercount を算出してみよう。

- 総じて、裕福な都市では undercount が小さいようだ。

また、perAA と equip の組み合わせから undercount を見てみると、

- African American の割合が多い場合は、OS-CC や PUNCH で undercount が多く、割合が少ない場合は、LEVEROS-PC で undercount が多い傾向にあった。

結論として、undercount は都市の経済状態に大きく依存していることがわかった。African American の割合や投票方法も多少の影響は与えるが、それほど重要ではない。

2 Likelihood Theory

ぶっちゃけこの部分の内容は、ぴんく本（「生物学を学ぶための統計のはなし」）の 6 章を読んだ方がわかりやすいと思いますが、一応やらせていただきます。

2.1 最尤法

尤度とは、あるデータが得られたときに、そのデータが得られる確率を示すものである。例えば、じゃんけんをして勝った、負けた（ベルヌーイ試行）を繰り返すときに、勝つ確率は以下のように定義することができる。

$$L(p|y) = \binom{\mathrm{n}}{y} p^{y}(1-p)^{\mathrm{n-y}}$$

これはいわゆる二項分布である。尤度は通常 L と表記する。

最尤推定（*maximum likelihood* estimate, MLE）は、得られたデータが得られる確率を最も高くするように、確率関数のパラメータを推定する方法である。そのためには、推定しようとするパラメータで確率関数を微分して、0 となるようにパラメータを決定すればよい。

ただし、通常、尤度はそのままでは計算が困難なため（微分できない）、対数尤度を最大化する。対数にすると、微分によるパラメータ推定が容易になるケースが多いためである。対数は単調増加という性質を持っているため、対数尤度を最大化するパラメータは、尤度を最大化するパラメータと同じである。この、尤度を最大化するパラメータを、**最尤推定値**という。二項分布の場合、y/n が最尤推定値である。二項分布の対数尤度は、以下のようになる。

$$l(p|y) = \log \binom{\mathrm{n}}{y} + y \log p + (n-y) \log(1-p)$$

先ほど述べたように、（対数）尤度を最大化するためには、目的とするパラメータで微分して、その結果が 0 になるようにパラメータを決定すればよい。以上の式を p で微分すると、

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

となり、これが 0 となる p は y/n である。これが最尤推定値である。この最尤推定が正しいかどうか、p を様々な値に変化させた場合の、対数尤度の変化を、図に示してみよう。

```
> loglik <- function(p, y, n) {lchoose(n, y) + y*log(p) + (n - y)*log(1 - p)}
# これは二項分布の対数尤度の式
> nloglik <- function(p, y, n) {loglik(p, y, n) - loglik(y/n, y, n)}
# 標準化した対数尤度
> pr <- seq(0.05, 0.95, by=0.01)
# p の値のレンジを指定。
> matplot(pr, cbind(nloglik(pr, 10, 25), nloglik(pr, 20, 50)), type="l",
          xlab="p", ylab="log-likelihood")
```

しかし、このように単純に最尤推定値が求まるケースはまれである。通常は、*Newton–Raphson* 法、あるいは、Fisher scoring 法（iteratively reweighted least squares 法に等しい。GLM の係数推定ではこちらが用いられる）を用いた数値探索によって、最尤推定値を決定する（数学的な詳細は割愛）。

2.2 仮説検定

尤度を用いることで、モデル選択を行うことができる。また、最尤推定値の 1 標本検定を行うことができる。

2.2.1 尤度比検定

尤度を用いることで、より大きい（制約のない）モデル Ω とより小さい（大きいモデルに制約を加えたもの）モデル ω が異なるモデルかどうかを検討することができる。これを尤度比検定といい、検定統計量は以下のように構築される。

$$2 \log(L(\hat{\theta}_\omega)/L(\hat{\theta}_\Omega)) = 2(l(\hat{\theta}_\omega) - l(\hat{\theta}_\Omega))$$

いくつかの条件（大標本である、etc）を満たしていれば、この検定統計量は、自由度がパラメータ数の χ^2 分布に漸近的に従う。

2.2.2 Wald 検定

Wald 検定は、最尤推定値がある値と有意に異なっているかを検定する方法である。帰無仮説は $H_0 : \theta = \theta_0$ で表され、検定統計量は、

$$(\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0)$$

で表される。この検定統計量は、自由度がパラメータ数の χ^2 分布に正確に従う。

また、複数のパラメータを持つモデル間で、ある単一のパラメータだけが異なるかどうかを検定するのにも用いることができる。その場合、検定統計量は、

$$z = \frac{\hat{\theta}_i - \theta_{i0}}{se(\hat{\theta}_i)}$$

で表される。

2.2.3 Score 検定

Wald 検定と同じく最尤推定値の検定方法であり、検定統計量は、

$$u(\theta_0)^T I^{-1}(\theta_0) u(\theta_0)$$

で表される。この検定統計量は、自由度が、検定しようとするパラメータ数の自由度の χ^2 分布に漸近的に従う。

どの検定が特別優れているということはないが、（検出力が高いという意味で？）尤度比検定は他の検定よりも優れている。そのため、尤度比検定の条件が満たせない条件下でのみ、他の検定法を用いるべきである。

2.2.4 最尤推定値の信頼区間

Wald 検定法による信頼区間は、

$$\hat{\theta}_i \pm z^{1-\alpha/2} se(\hat{\theta}_i)$$

のようにして計算することができる。

尤度比検定の場合は以下のようになる。

$$2(l(\hat{\theta}_i|y) - l(\theta_i|y)) < \chi_1^{1-\alpha}$$

カッコ内は、対数尤度から $l(\theta_i|y)$ を引くことで標準化を行っている。これらの方法で算出した信頼区間の値はほとんど同じであり、実用的にはどちらを使ってもよい。これらの方法の差が生じるのは、サンプルサイズが少ないときである。