第２章　Binominal data

上野真由美

2.1 チャレンジャー号の悲劇 Challenger Disaster Example

チャレンジャー事故の原因：発射時の温度が低かったことによる O リングの破損

問いかけ：O リングの破損は予測できたのか？

まず、O リングと発射温度の散布図かいてみた。
次に、ある温度で、O リングが破損する割合を予測するために、回帰直線をいれてみた。
しかし、これらは間違い！

なぜならば、
・これだと、1 以上や 0 以下の割合も予測できることになる（そんなことありえない）。
・誤差は正規分布ではなさそう。
・2 項データは等分散性を仮定できない。

<u>ということで、線形モデルにあてはめることはできない。</u>

2.2 二項回帰モデル Binominal Regression Model

Yi：レスポンス（従属変数）。それぞれの試行 Y は独立であると仮定。
xi1－xiq：プレディクター（独立変数）
covariate class：ある:試行グループ
linear predictor:線形予測式（モデル）

＊線形予測式では、カテゴリ変数・量的変数とも扱える。また説明変数を変換したり、合体することもできる。また、それぞれの独立変数が従属変数に与える影響をそれぞれ抽出できる（これは他の GLM モデルにもいえること、第 6 章参照）。

3 つのリンクファンクション（link function）
リンクファンクションとは：比較的広いクラスのモデルの中で、独立変数を従属変数の平均にリンクさせる。
p は 0-1 の間でしか動かないから、η =p はよくなくて、p を使った関数を、η にする

1. Logit: $\eta = \log(p/(1-p))$
2. Probit: $\eta = \Phi^{-1}(p)$

3. Complementary log-log: η = log(-log(1-*p*))

パラメータ推定は、さいゆう法(アルゴリズムは 6 章で)。
例）チャレンジャーデータ

二項データでは、反応は YES か NO（y—n）
この情報を引き出すには、2 列行列にして示すのがよい。
デフォルトは logit（＝ロジスティック回帰）

他のリンクファンクションが適切な場合もあるが、この例では probit でもほぼ同じ結果

2.3 推論 Inference

the likelihood statistics: 尤度比検定とは？
2 つのモデルを比較して、どちらがもっともらしいかを評価する方法

$$2\log\frac{L_L}{L_S}$$

‘（たくさんのパラメータを含むモデルをデータにあてはめたときの最大尤度／少ないパラメータのモデルの最大尤度）の自然対数×2’
両者の最大尤度に差がない（少ないパラメータモデルでも十分尤度を最大化することができる）ならば、尤度比はカイ二乗分布に従う。
→「尤度に差がない」というきむ仮説の下でモデルのあてはまりを評価する。

いま、与えられたモデルが（サンプル数について？）十分に満たされた場合、p = y/n となるから、上記の尤度比は、

$$D = 2\sum_{i=1}^{n}\left\{y_i\log y_i/\hat{y}i + (n_i - y_i)\log(n_i - y_i)/(n_i - \hat{y}_i)\right\}$$

デビエンス（D）：現在のモデルが、完全モデルに比べて、どれだけ完璧に近いかを検定する手法。当てはまりがよければ deviance は自由度 n-1 のカイ二乗分布に従うので、検定ができる。

Residual deviance: 現在のモデルの（完璧モデルに対する）あてはまり度
Null deviance: 切片だけの（完璧モデルに対する）モデルのあてはまり度

☆サンプル数 ni が増えるほど、カイ二乗分布と仮定することができる。
サンプル数が少ないと、どんどんデビエンスは減っていく。
デビエンスのためには、$\hat{p}$ のモデル値と、データ y を比べなくてはいけないが、我々歯に $\hat{p}$ の関数の形でしか知らないので、あてはまりを評価することはできないし、ましてやカイ二乗分布はしていないだろう。

☆この検定は、少なくともサンプル数が 5 以上であることが好ましい。
代替法→Permutation や Bootstrap

☆２つのネストモデルの比較も可能

この場合は、$D_S - D_L$ で検定。

小さなモデルが正しいならば、分布の想定が妥当であれば、$D_S - D_L$ は自由度 l-s のカイ二乗分布になるはずである。
このテストの代替法は、Z バリューであるが、スパースデータで SE を過大評価してしまい、Z 値が小さく出てしまう可能性があるから（Hauck-Donner effect）、デビエンスベイスト検定の方がいい。

☆蛇足ですけど！
デビエンスの違いを含むテストのほうが、単一のデビエンスを指す goodness of fit よりも、一般的に正しい。

・パラメータ値の信頼区間の出しかた：2 通りある
・ここではロジットしか使ってないけど、他のリンクファンクションでも使えるよ。

2.4 トレランス分布 Tolerance Distribution
＞ここでは、何がいいたいのか？
複数のリンク関数（ロジット・プロビット・コンプリ）は、独立変数の分布の違いから自然と派生してきたものである。

従属変数：ある質問に答えられない確率
独立変数：学生の学力

学力分布が正規分布→プロビット型回帰モデルになる。をしたトレランス分布から導かれ
ロジスティック分布→ロジットモデル
極値分布（*）→コンプリモデル

（＊）極値分布：正規分布を含む多様な分布から抽出された大きい標本における最小あるいは最大値に限った分布。

2.5 オッズ比の解釈 Interpreting Odds

オッズ比とは：賭け事で使われる物差し＝もうけた分/支払った分

オッズ比：起こる確率(p)／起こらない確率(1-p)

オッズ比の便利なところ：

- unbounded (p: 0-1, whereas o: 1-∞)
- p:は対称性とか長期的な頻度を考慮することで決定されるが、こういった考慮する情報はしばしば分からないから、オッズ比を見ることで、本質的な確率を判断することができる。このようにして、損をしないようなお得な査定を行うことができる。

・オッズ比の他の概念として、相対リスク relative risk：

あるコンディションがあったときの成功率（$p1$）／ない場合の成功率（$p2$）

（これって、話題から脱線・・？）

例）小児呼吸器病に関する調査

問い：粉ミルク・母乳と粉ミルク・母乳のみという､3 種類のミルクの与え方によって、小児呼吸器病の発生は予測できるのか？

結果：性差あり・食べ物（粉ミルクと母乳間）で有意さあり

交互作用の考慮は→残差が小さいので、考えなくていいだろう。

主要因の効果（因子分析）*using **drop1** function*

2 つの要因に共通性はないかを検定する。

結果：どちらの要因も有意に病気の発生に関わっている

母乳の効果を見ると

＞exp（-0.669）

[1] 0.51222

つまり、母乳によって、粉ミルクで起こる確率を 51%減らすことができる。

2.6 前向きサンプリングと後向きサンプリング Prospective and Retrospective Sampling

前向き: 予測をもとに、それを検証する実験デザインを作る
後ろ向き: 得られた結果を引き起こした要因を探る。

例）小児呼吸器病と授乳の関係
前向き：とりあえず 2 グループに分けて別別の授乳を行う→病気の発生チェック
後ろ向き：病気の人をまずは対象に、どんな要因がひそんでいたのか問診＋病気じゃなかった人にも同じようなこと聞いてみよう。

後ろ向きサンプリングの方が、安い・早い・効果的

両者の違いは？
Logit 変換した式には、推定したい病気になる確率 p(x)を含む項と、それを含まない項がある。この、p(x)を含まない項にしか、先ほどの病気を持っている人の事前分布（π0 と π1）は含まれていません。だから、Prospective な研究と、Retrospective な研究の違いは、切片に出てくる。（by 飯島さん）

2.7 リンクファンクションの選択 Choice of Link Function

・中間くらいの p の範囲では、どのリンクファンクションを利用しても変わらない。
・大きな違いは、極端な p 値のところで生じるが（Fig.2.3）、とても小さい p 値の場合、リンクファンクション自体の違いを知るにはたくさんのデータが必要で大変である。
・結局は、、、物理的知識や単純な便利さから得られた想定に基づいてリンクファンクションは選ばれる。
・デフォルトチョイスはロジットである。3 つの利点：シンプルな数式・オッズ比を用いることで解釈が簡単・後ろ向きデータの解析が簡単・

Fig.2.3
p 値の極端なところで、リンクファンクションによる予測結果の違いあり。

例）カルシノゲンと他の物質が、人に有害であるかをテストするための検定。
→低量でも人に有害な影響をあたえる確率を推定するために、適切なリンクファンクションを設定することが必要。高いドーズのデータはあまり必要ではない（cf アスベスト）

2.8 推定に関わる問題 Estimating Problems

パラメータ推定ができない、収束しない。これはデータのせいである。
データが線形乖離 linearly separable している。

Residual Deviance みたら、とってもフィットしているのに、なぜ！？
こういうのを‘embarrassment of riches の苦悩’と呼ぶ。

こういった場合の対処法
正確ロジスティック回帰が適しているが、R ではできない。
R での代替法：Firth1993
O（1/n）を、推定係数の漸近的バイアスから除いてやる。

我々は要因が有意にも関わらず、この結果を見ることができる。あてはまっているようだが、係数から判断すると、glm の結果から、異なっているように思える。このように、パラメータの不安定性は、線形乖離に近づくデータセットの場合に、発生する。

（これって、estrogen と androgen の比率を変数にしたら、うまくいかないのかな？）

2.9 当てはまり度 Goodness of Fit

デビアンス以外にも、統計モデルがデータにどの程度あてはまるかを示す尺度がある。

・Pearson カイ二乗値

Pearson カイ二乗値は正規分布でいう残差平方和に相当する。
Pearson の X2 統計量は、通常の線形モデルにおける残差平方和として近似可能。
（つまりこれ自体は当てはまりの指標ではない）
カイ二乗値とデビエンスの違いはそんなに大きくない。

・R2 値の二項分布バージョン

いわゆる自由度調整済み R2 値を、GLM でも使えるようにしたもの。

2.10 予測と効果的な投与量 Prediction and Effective Doses

ある共変量に対する、結果の値の出し方

CI はこうやって出す

この方法でも出せるよ。

線形回帰と違って、将来の観察に対する CI と、平均の反応に対する CI に違いはない。う？つまり、がいそうもないそうも変わらないとかそういうこと？

ロジスティック回帰は、50%を基準に使われてきた。でも、かならずしも、対象性ではないならば、0.5 じゃなくてもいい。

2.11 過分散 Overdispersion

過分散とは？

本来予想される分散よりも分散がばらついている状態

なぜ Overdispersion が起こるの？

1. まず、モデルの問題（みかけの overdispersion）
   - パラメータの欠如・モデルの構造
   - いくつかの外れ値（外れ値の除き方は 6.4 で）
   - ちらばりデータの存在
   - モデルのランダムパートにおける欠陥

2．モデル以外の問題(これがいわゆる overdispersion)

二項分布（あるいはポアソン分布）は、分散が期待値によってのみ定義されているため、実際には分散が予測値から予想される以上にばらついてしまうことがあります。

例）グループ内で p の変異を生む要因（O リングの位置や、個体群内のサブ個体群構造）

Overdispersion の克服方法：分散パラメータを推定して、モデルに考慮する（？）。

過分散が起こっていなければ、分散パラメータは、1 になるはず（前の方のページでもそうなってる）。

2.12 マッチングを行った症例対照研究 Matched Case-Control Studies

背景

症例対照研究では、ある要因の発症リスクを知りたい。しかし、その要因以外にも複合要因がある。

さて、どうするか？

関わっている要因全てを説明要因にいれて、ロジスティック回帰したら？

NO！複合要因を正しい関数でモデル化しないと、正しい答えはみつからない

そこで！

ある要因を固定して、他の要因の効果を見る。

例）56 歳のヒスパニック系の人間、もう少し大雑把にするなら、50 代男性でデータを整理してみることで、それ以外の要因がある疾病に与える影響を評価することができる→このようにすると、年代と人種がこの症例に与える影響については考えずに、他の要因に注目して分析できる。

マッチングの利点：直接測定できない影響について考慮することができる。

例）自分が調べたい要因以外に、その人の生活環境もある疾病に関与しているとしたら。

→住んでいる場所や職場でマッチングすれば、環境要因を考慮しつつ（つまりこの要因を除去したなかで）、他要因の疾病への影響を調べられる。

この方法の欠点

・要因間の関係が分からなければ、何でマッチングさせるかが困難である。

・マッチングで固定させてしまった要因が症例に与える影響は見れない。たとえば性の違いによる影響は、性でマッチングさせてしまったら見れない。

例）X 線と骨髄発血病との関係

今回はダウン症以外の要因について考えたいから、

ダウン症を発症している症例と、それらの症例と（age で）マッチングした症例であるデータを除去した。

strata ( ): ( )でマッチングする

マッチングしたものと、しなかったもので、分析結果の違いは何？（p.51）
→マッチングに関わらず、unclass (CnRay)の要因は有意であったが、子供の X 線照射と発症因果があるかどうかは分からない。

・・でも、それって、マッチングしなかったことと関係あるのか？
？

マッチングデータは、Section4.3 でもありますので、そちらで・・

素朴な疑問
共変量（covariate）、因子、predictor って、全て同じことなのに、なぜ呼び方変える？

Oh, what's a devience？

数学的に考えると、デビアンスは特定の推定されたモデルと「フルモデル」とを比較する尤度比統計量の対数値となる。「フルモデル」においては、観察データを理論値として用いているので、すべての変動は、ランダム成分ではなく系統的成分に起因するものとみなすことになる。逆に、共通の平均をすべてのデータにあてはめるのが、いわゆる「ヌルモデル」である。ヌルモデルには変量は含まれないので、すべての変動は系統的成分ではなくランダム成分に起因する。統計解析の目的は、フルモデルのようにデータをモデル中で再び使用するのではなく、ヌルモデルよりもデータをより良く説明するモデル（そのようなモデルがあるとすれば）を見つけることである。モデルの推定は、幾つかのネスティッド・モデル間でデビアンス差を調べ、データに一番よくあてはまりがよく、できるだけ少ない数の母数をもつモデルをみつけることである。というのは、データが多種の変量に依存しているか、もしそうならその依存がどのようなものかを研究者が知りたいと考えているためである。