

第4章 分割表

カテゴリカルなデータの分析

要因間の独立じゃないっぴりをどうモデルに組み込むか？

4.1 2×2 表を分析する

データのとり方によって解析の方向はいろいろ。

poisson、multinomial、binomial、hypergeometric model

1、poisson モデル

ウエハーの quality, particle の頻度に差があるか？

$$\rightarrow \log \mu = \gamma + \alpha i + \beta j$$

予測値の有意性をチェック→z よりも LRT がいい？

drop1(test="Chi")で検討→2 要因とも重要

2 要因間の関係は interaction term について検当が必要だがモデルが飽和する

→deviance を検討

*test="F"でないの？

2、multinomial モデル

n (=450) 回観測したうちそれぞれのセルについて y_{ij} 回観測→p_{ij}

p_{ij}=p_ip_j? (独立)

P_{ij}=P_iP_j なモデルの p_i, p_j の MLE から LR を計算

→saturated model ($\mu_{ij}=y_{ij}$)と尤度比を求める

deviance の代わりとしては χ^2 検定があるが deviance-based なほうがいい？

3、binomial モデル

particle を説明、quality を response

no particle を 400 個、particle を 50 個調べるとして、quality を調査

null モデル (particle によって出来に差がない)を当てはめてみる。

* $y \sim 1$ な binomial モデルは何者か？

4、Hypergeometric model

すべての marginal total を固定、各セルの期待値と観察された値を検定

→Fisher's exact test

4.2 2×2 よりでかい表を分析する

変数間が独立かどうか見てみる Pearson's χ^2 test →独立ではない

データを視覚的にとらえる (figure 4.1)

Dotchart

Mosaicplot

Poisson GLM で分析してみる

→単に毛の色や目の色に頻度の違いがあったり、なかったり。

毛の色と目の色の関係を知りたい

Residual deviance 146.44 on 9 DF

毛の色と目の色は独立でない

どのように独立でないか→correspondence analysis

残差をUとDとVとに分解

UとVとで散布図を描くと変数間の関係がよくわかる (figure 4.2)

*主成分分析の親戚っぽい？

4.3 Matched Pairs

右目と左目の視力は関係があるか？

とりあえず $y \sim \text{right} + \text{left}$ として独立性のチェック→独立性は当然ない

$p_{ij} = p_{ji}$ か？

→ 左右の視力の組み合わせを $p_{ij} = p_{ji}$ な感じにまとめてモデルをフィット

残差がいっぱい (19.249 on 6) → 対称でない

残差に注目→右目と左目の関係

周辺度数の均一性がなさそう

↓

$p_{ij} = \alpha_i \beta_j \gamma_{ij}$ ($\gamma_{ij} = \gamma_{ji}$): quasi-symmetry model

周辺度数の不均一性と対称性をあわせてモデルに組み込む

4.4 Three-Way Contingency Tables

タバコすう・すわない、生死、年齢の関係

Simpson paradox : marginal な関係と conditional な関係が異なる

→ タバコを吸う人は若い人に集中 → タバコを吸う人ほど長生き

K 層からなる 2×2 table における独立性の検定 → Mantel-Haenszel test

線型モデルによる解析: p_{ijk} をどう表現するか?

Mutual Independence なモデル

$P_{ijk} = p_i * p_j * p_k$: $\log E y_{ijk} = \log n + \log p_i + \log p_j + \log p_k$

Smoker, dead, age を独立にモデルに入れる → フィットしない

Joint Independence

$P_{ijk} = p_{ij} * p_k$: $\log E y_{ijk} = \log n + \log p_{ij} + \log p_k$

Interaction を一個含んだ log-linear モデル

どの説明変数のペアが独立かわかる

今回は三つのモデルともダメ

Conditional Independence

$P_{ij|k} = p_{i|k} * p_{j|k} \rightarrow p_{ijk} = p_{i|k} * p_{j|k} / p_k$

$\log E y_{ijk} = \log n + \log p_{i|k} + \log p_{j|k} - \log p_k$

今回のデータでは $y \sim \text{smoker} * \text{age} + \text{age} * \text{dead}$ # R の glm でやるならこれでいい。

Uniform Association

すべての two-way interaction を含む

$\log E y_{ijk} = \log n + \log p_i + \log p_j + \log p_k + \log p_{ij} + \log p_{ik} + \log p_{jk}$

解釈はよくわからん!

Model Selection

uniform association model が選択される

Binomial Model

1 つを従属変数、2 つを説明変数とする

モデル選択すると。。。uniform association と一緒

おなじオッズ比も取りだせる

poisson モデルと binomial モデルの違い

binomial の null モデルは $y \sim \text{smoker} * \text{age} + \text{dead}$ なモデルと一緒に

つまり、binomial はあらかじめ $\text{smoker} * \text{age}$ な関係を仮定してモデリング

こういう仮定が無いだけ poisson のほうがいい？

4. Ordinal variable

複雑なカテゴリーを単純化

変数間の順序を情報として含める

カテゴリカル変数に

$u_1 < u_2 < \dots < u_I$, $v_1 < v_2 < \dots < v_J$ のようなスコアを割り振る

$$\log E Y_{ij} = \log \mu_{ij} = \log \pi_{ij} = \log \pi + \alpha_i + \beta_j + \gamma_{uij}$$

順序付けのもっともらしさを確認するために

$u_1 < u_2 < \dots < u_I$ を満たす別な数字を割り振ってみたい

順序の幅を変えて見てモデルの頑健さを確認してみるのもいい

γ の予測値は u_i, u_{i+1} と u_j, u_{j+1} の行列間の log-オッズ比である

column/row effect model という方法もある。

Colum/row の一方のみを順序化する

$$\text{Log} E Y_{ij} = \log \pi + \alpha_i + \beta_j + u_i \gamma_j$$

例) 学歴によって (educ) 政党支持の傾向が一樣ではない (PID)

政党支持 (PID) の順序には解析の興味がある場合

α_i : 政党支持 (PID), β_j : 学歴 (educ)、

$$y \sim \text{PID} + \text{educ} + \text{educ} : \text{PID}$$