Within-cases analysis of binary responses[1]

STA442/2101 Fall 2018

[1] This slide show is an open-source document. See last slide for copyright information.

The idea

- There are several binary responses for each case.
- Like was the person employed right after graduation, 6 months after, one year after ... Yes or No
- Or did the consumer purchase at least one computer in 2016, 2017, 2018 ...
- Binary choices in laboratory studies can be repeated measures.
- Model: Logistic regression with a random shock for case, pushing all the log odds values for that case up and down by the same amount.
- Random shock is added to the regression equation for the log odds.
- Usually the random shock is normal — what else?

A random intercept model

For $i = 1, \dots, n$ and $j = 1, \dots, m$

- $B_1, \dots, B_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$
- Conditionally on $B_i = b_i$ for $i = 1, \dots, n$, binary responses y_{ij} are independent with

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\beta_0 + b_i) + \beta_1 x_{ij1} + \ldots + \beta_{p-1} x_{ij,p-1}$$

where $\pi_{ij} = P\{y_{ij} = 1\}$.

Some of the x_{ij} could be dummy variables for time period or treatment, different for $j = 1, \dots, m$ within case i.

Likelihood function: $\boldsymbol{\theta} = (\boldsymbol{\theta}, \sigma^2)$

Denoting $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,m})^\top$, have

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} P_{\boldsymbol{\theta}}(\mathbf{Y}_i = \mathbf{y}_i) \\ &= \prod_{i=1}^{n} \int_{-\infty}^{\infty} P_{\boldsymbol{\theta}}(\mathbf{Y}_i = \mathbf{y}_i | B_i = b_i) f(b_i; \sigma^2)\, db_i \\ &= \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{m} \pi_{i,j}^{y_{i,j}} (1 - \pi_{i,j})^{1-y_{i,j}} f(b_i; \sigma^2)\, db_i \end{aligned}$$

where $\pi_{i,j} = \frac{e^{\mathbf{x}_{ij}^\top \boldsymbol{\beta} + b_i}}{1 + e^{\mathbf{x}_{ij}^\top \boldsymbol{\beta} + b_i}}$.

Maximum likelihood

Numerical, of course

- In principle, this is mostly straightforward.
- It's all classical likelihood stuff.
- In this class, we just have a random intercept.
- But the model can be extended to

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$$

- Where $\mathbf{w}$ is a vector of log odds.
- That's what the `glmer` function in the `lme4` package does.

There are problems

- Nobody can do the integral:
  $\int_{-\infty}^{\infty} \prod_{j=1}^{m} \pi_{i,j}^{y_{i,j}} (1-\pi_{i,j})^{1-y_{i,j}} f(b_i;\sigma^2)\, db_i$.
- It's really brutal for multivariate normal **b** and complicated designs.
- The approximate solutions are imperfect.
- There are numerical issues, even in our simple case.
- For the general case, it's easy to specify models whose parameters are not identifiable.
- Identifiability is not a problem for us, but there is massive confusion in the user community.

The `glmer` function in the `lme4` package

- Syntax is like `lmer` for linear models.
- And like `glm` for generalized linear models with fixed effects.
- We are going to keep it simple.
- Just add `+(1|Subject)` for the random shock (intercept).
- Use effect coding (`contr.sum`) if there are interactions between factors.
- `Anova(model,type='III')` from the `car` package to test each effect controlling for all others.
- For follow-up tests, fit a no-intercept model on a combination variable and test contrasts on the categories of the combination variable using the `linearHypothesis` function from the `car package`.

Copyright Information