Repeated Binary Data Analysis

0. Reference

- Fitting linear mixed-effects models using lme4

1. Introduction

binary 데이터, 즉 0, 1의 값이 연속적으로 발생한 데이터에 대해 모델링을 할 때, 어떤 방법이 있는지 알아본다. 기본적으로 로지스틱 회귀를 사용하지만, subject index가 반영된다. 아래와 같이 R과 SAS에서 해볼 수 있다.

- subject index을 random effect로 설정하는 mixed model
  - log odds의 covarianc structure을 compund symmetry로 가정한다.
  - R의 glmer을 통해서 수행한다.
- log odds의 시점간에 dependency을 설정하는 markov model
  - log odds의 covarianc structure을 직접 정한다. 즉, covariance pattern model 느낌이다. random effect을 지정하지 않을 수도 있다.
  - R의 bild package을 통해서 수행한다.
  - SAS에서 glimmix을 통해서 수행한다.

여기서는 R에서 모델링을 수행한다.

2. Data Description

데이터는 배고픔이 메뚜기의 움직임에 미치는 영향을 연구하기 위해 기록되었다. 총 24마리의 메뚜기가 30초 간격으로 움직였는지 (move 변수) 여부가 관측되었다. 이때, 성별과 함께 먹이를 주었는지 (feed 변수) 도 함께 기록되었다. 아래는 데이터의 일부이다.

```
library(bild)
head(locust)
```

```
##   id move sex        time feed
## 1  1    0   1 0.008333333    1
## 2  1    0   1 0.016666667    1
## 3  1    0   1 0.025000000    1
## 4  1    0   1 0.033333333    1
## 5  1    0   1 0.041666667    1
## 6  1    0   1 0.050000000    1
```

3. Modeling

Markov Model

가정한 모델은 아래와 같다.

$$logit(\theta_{it}) = \beta_0 + \beta_1 times + \beta_2 feed + \beta_3 times \times feed \tag{1}$$

```
locust1 = bild(move ~ time*feed, data=locust, aggregate = feed, dependence='MC1', start=NULL)
locust2 = bild(move ~ time*feed, data=locust, aggregate = feed, dependence='MC2', start=NULL)
```

bild 함수의 입력 데이터는 subject, time 변수가 이름 그대로 존재해야 한다. 각각 반복 측정치에서 subject을 구분해주고, 반복 측정된 시점을 의미한다. locust 데이터에 대해서 makrov order 1, 2 모델을 적합하였다. 아래는 각 모델의 AIC, log likelihood 값이다.

	markov order 1	markov order 2
AIC	3251	3157
log likelihood	-1620	-1572

markov order 2 모델이 order 1 모델을 nest 하므로 $2(L_{order\ 2} - L_{order\ 1}) = 88 >> \chi^2_1$이고 귀무가설을 기각한다. 즉, markov oder 2 모델이 더 적합하다고 결론지을 수 있다. markov order 2 모델의 결과를 살펴보자.

```
summary(locust2)
```

```
##
## Call:
## bild(formula = move ~ time * feed, data = locust, aggregate = feed,
##     start = NULL, dependence = "MC2")
##
## Number of profiles in the dataset:  24
##
## Number of profiles used in the fit:  24
##
## Log likelihood:  -1572.706
##
## AIC:  3157.411
##
```

```
## Coefficients:
##             Label      Value Std. Error t value  p-value
## (Intercept)     1 -0.7551136 0.16637070  -4.539 0.000006
## time            2  1.1145083 0.21115822   5.278 0.000000
## feed1           3 -3.6770819 0.35826616 -10.264 0.000000
## time:feed1      4  1.1936595 0.39965067   2.987 0.002820
## log.psi1        5  1.5111245 0.11199667  13.493 0.000000
## log.psi2        6  0.9494884 0.09618394   9.872 0.000000
##
## Message:  0
```

모든 계수와 log.psi1, log.psi2의 p-value가 유의한 결과를 얻었다. 결과를 아래와 같이 해석해보자.

- log.psi1, log.psi2는 markov order의 한 시점, 두 시점 간의 log odds이다. 두 시점까지 log odds를 가정하는 것이 유의하다는 뜻이다.
- time, feed 변수가 유의하다. 즉, 두 변수가 $logit(\theta_{it})$에 통계적으로 유의한 영향, 즉 계수만큼 영향을 미친다.
- time, feed의 interaction term이 유의한 것으로 보아, 두 변수가 $logit(\theta_{it})$에 미치는 영향은 각 변수의 수준에 따라서 달라진다.

Mixed Models

다음으로는 glmer 함수를 이용하여 mixed model을 적합한다. glmer 함수는 glm 함수의 mixed model 버전이다. family을 통해서 y가 지수족인 데이터에 대한 모델을 세울 수 있고 여기에 random effect을 추가할 수 있다.

우선 glmer의 syntax에 대해서 확실히 짚고 넘어가겠다. glmer 하나만 잘해두면 continous, binary, count data 모두에 대해서 mixed model을 세울 수 있고 intercept 뿐만 아니라 slope까지 random으로 가정하는 random coefficient model로 모델을 확장할 수도 있어서 꼼꼼하게 살펴보겠다. 아래 내용은 이곳을 참조하여 정리하였다.

먼저 기본적으로 아래의 사항을 숙지하자.

- syntax는 dependent ~ independent | grouping 구조이며 grouping은 일반적으로 random effect을 넣는다.
- random effect에 대해서는 아래와 같이 세 가지로 나뉜다.
  - random intercept term만 있음: (1 | random effect)
  - random slope만 있음: (0 + fix effect | random effect)

- random intercept, slope 모두 있고 이 둘의 correlation을 가정함: (1 + fix effect | random effect)
- random intercept, slope 모두 있는데 이 둘이 독립적으로 계산 됨: fix effect + (fix effect || random effect)

이제 직접 model equation을 적어가며 R의 syntax와 비교해보자. random effect로는 S, I를 넣는다고 가정한다. (s는 S의 index, i는 I의 index, 각 데이터 index까지 추가하려면 k도 있어야 함.)

1. $Y_{si} = \beta_0 + \beta_1 X_i + e_{si} \longleftrightarrow$ *not mixed model*
2. $Y_{si} = \beta_0 + S_{0s} + \beta_1 X_i + e_{si} \longleftrightarrow$Y ~ X + (1 | Subject)
3. $Y_{si} = \beta_0 + S_{0s} + (\beta_1 + S_{1s}) X_i + e_{si} \longleftrightarrow$Y ~ X + (1 + X | Subject)
4. $Y_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s}) X_i + e_{si} \longleftrightarrow$Y ~ X + (1 + X | Subject) + (1 | Item)
5. $Y_{si} = \beta_0 + S_{0s} + I_{0i} + \beta_1 X_i + e_{si} \longleftrightarrow$Y ~ X + (1 | Subject) + (1 | Item)
6. 4.와 동일하지만 S_{0s}, S_{1s}가 독립인 경우$\longleftrightarrow$Y ~ X+(1 Subject)+(0 + X Subject)+(1 Item)
7. $Y_{si} = \beta_0 + I_{0i} + (\beta_1 + S_{1s}) X_i + e_{si} \longleftrightarrow$Y ~ X+(0 + X Subject)+(1 Item)

locust 데이터에 대해서, 총 세 가지 mixed model을 적합할 것이다: 1) Y ~ X + (1 | Subject) 2) Y ~ X + (1 + X | Subject) 3) Y ~ X + (X || Subject)

1) Mixed model with random intercept only

random effect을 추가하여 mixed model을 세우는 경우에 적합하려는 model equation, random effect을 넣음으로써 fixed effect을 넣었을 때와는 다르게 어떤 변화가 생기는지, random effect을 넣는 것이 유의미한지 등을 살펴보아야 한다.

- model equation

$$logit(\theta_{it}) = \beta_0 + S_{0i} + \beta_1 sex + \beta_2 time + \beta_3 feed + \beta_4 time \times feed \tag{2}$$

  수식 (2)는 $P(move = 1) = \theta_{it}$의 logit에 대한 GLMMs이다. random effect가 intercept항에만 추가 되었고(S_{0i}) 나머지는 fixed effect이다.

- random effect을 넣음으로써 얻는 효과
  subject를 random effect에 넣음으로써 subject별로 관측된 데이터 간에는 correlation이 있음을 가정한다. 또한 수식 (2)에서 절편 $\beta_0 + S_{0i}$에는 첨자 i가 있는데, 이는 곧 subject 별로 절편이 다른 logistic 회귀 식을 적합할 것임을 의미한다. 즉, subject 별로 $sex, time, feed, time \times feed$가 $logit(\theta_{it})$에 미치는 영향은 동일하지만 baseline이 다르다는 것을 가정한다. 물론 이것이 truely 맞을지는 모르지만 우선 이렇게 가정하는 것이다.

```r
library(lme4)
```

```
## Loading required package: Matrix
```

```r
locust_glmer = glmer(move ~ time*feed + sex +  + (1 | id), family=binomial, data=locust,
```

우선 random effect의 추정된 standard deviance을 살펴봄으로써 random effect을 추가하는 것이 적절한지를 알아보자. 만약 추정된 분산이 0에 가깝다면 이는 random effect의 변동이 거의 존재하지 않는다는 뜻이므로 차라리 제외하는 것이 낫다는 뜻이다.

```r
ranef(locust_glmer)
```

```
## $id
##    (Intercept)
## 1   -1.3969992
## 2    1.7109201
## 3   -0.2120935
## 4    0.1187238
## 5   -0.5057247
## 6   -0.9418926
## 7    1.0894491
## 8   -0.6981046
## 9   -0.2120935
## 10   1.6648253
## 11   1.2490315
## 12  -1.2354779
## 13   0.1411729
## 14  -0.2025071
## 15   1.1396411
## 16   1.4352300
## 17   0.2997149
## 18  -1.4794363
## 19   0.2199954
## 20  -0.4703020
## 21  -0.7235152
## 22   0.9161123
## 23  -0.8072186
```

```
## 24  -0.4703020
##
## with conditional variances for "id"
```

standard deviance가 1인 것으로 보아 random effect을 추가하는 것이 적절함을 알 수 있다. (이에 대한 유의성 검정은 찾는 중이다.)
fixed effect는 모수를 가정하므로 이에 대해서 추정(estimate)을 하여 추정치에 대한 p-value를 계산할 수 있지만 random effect는 그 자체가 모수가 아닌 random variable이다. random variable은 추정이 아니라 예측(predict)의 대상이다. 따라서 위의 random effect인 Subject에 대한 직접적인 p-value는 나오지 않고 이 random variable의 분산을 봐야하는 것이다.

다음으로는 fixed effect의 계수 추정치와 p-value을 살펴보자.

```
summary(locust_glmer)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
##  Family: binomial  ( logit )
## Formula: move ~ time * feed + sex + +(1 | id)
##    Data: locust
##
##      AIC      BIC   logLik deviance df.resid
##   3185.1   3222.7  -1586.6   3173.1     3858
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0109 -0.5098 -0.1633  0.5102 17.1384
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 1.046    1.023
## Number of obs: 3864, groups:  id, 24
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9758     0.3824  -2.552   0.0107 *
## time          1.3007     0.1337   9.731  < 2e-16 ***
## feed1        -4.3652     0.5335  -8.183 2.78e-16 ***
```

```
## sex1          0.2318     0.4383   0.529   0.5969
## time:feed1    1.3810     0.3246   4.255 2.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) time   feed1  sex1
## time        -0.238
## feed1       -0.475  0.166
## sex1        -0.575  0.005 -0.007
## time:feed1  0.098 -0.411 -0.557 -0.003
```

sex 변수를 제외하고 모두 유의한 결과를 얻었다. 계수에 대한 해석을 하기 이전에, 예측된 random intercept을 살펴보자.

```
ranef(locust_glmer)


## $id
##    (Intercept)
## 1   -1.3969992
## 2    1.7109201
## 3   -0.2120935
## 4    0.1187238
## 5   -0.5057247
## 6   -0.9418926
## 7    1.0894491
## 8   -0.6981046
## 9   -0.2120935
## 10   1.6648253
## 11   1.2490315
## 12  -1.2354779
## 13   0.1411729
## 14  -0.2025071
## 15   1.1396411
## 16   1.4352300
## 17   0.2997149
## 18  -1.4794363
```

```
## 19   0.2199954
## 20  -0.4703020
## 21  -0.7235152
## 22   0.9161123
## 23  -0.8072186
## 24  -0.4703020
##
## with conditional variances for "id"
```

24개의 subject에 대한 24개의 예측된 random intercept이다. 수식 (2)에서 회귀식의 절편은 random effect와 overall effect로 구성되어 있음을 알 수 있다. 위 24개의 결과는 random effect이므로 이를 overall effect와 더해야 한다. overall effect는 추정된 절편 항이다. 이 둘을 섞어서 subject 별로 적합된 로지스틱 회귀식은 아래와 같다.

```
coef(locust_glmer)


## $id
##    (Intercept)     time     feed1      sex1 time:feed1
## 1  -2.37280804 1.300665 -4.365229 0.2317706   1.380971
## 2   0.73511129 1.300665 -4.365229 0.2317706   1.380971
## 3  -1.18790228 1.300665 -4.365229 0.2317706   1.380971
## 4  -0.85708500 1.300665 -4.365229 0.2317706   1.380971
## 5  -1.48153357 1.300665 -4.365229 0.2317706   1.380971
## 6  -1.91770143 1.300665 -4.365229 0.2317706   1.380971
## 7   0.11364022 1.300665 -4.365229 0.2317706   1.380971
## 8  -1.67391346 1.300665 -4.365229 0.2317706   1.380971
## 9  -1.18790228 1.300665 -4.365229 0.2317706   1.380971
## 10  0.68901643 1.300665 -4.365229 0.2317706   1.380971
## 11  0.27322270 1.300665 -4.365229 0.2317706   1.380971
## 12 -2.21128676 1.300665 -4.365229 0.2317706   1.380971
## 13 -0.83463594 1.300665 -4.365229 0.2317706   1.380971
## 14 -1.17831589 1.300665 -4.365229 0.2317706   1.380971
## 15  0.16383231 1.300665 -4.365229 0.2317706   1.380971
## 16  0.45942115 1.300665 -4.365229 0.2317706   1.380971
## 17 -0.67609389 1.300665 -4.365229 0.2317706   1.380971
## 18 -2.45524509 1.300665 -4.365229 0.2317706   1.380971
## 19 -0.75581339 1.300665 -4.365229 0.2317706   1.380971
```

```
## 20 -1.44611082 1.300665 -4.365229 0.2317706   1.380971
## 21 -1.69932399 1.300665 -4.365229 0.2317706   1.380971
## 22 -0.05969655 1.300665 -4.365229 0.2317706   1.380971
## 23 -1.78302743 1.300665 -4.365229 0.2317706   1.380971
## 24 -1.44611082 1.300665 -4.365229 0.2317706   1.380971
##
## attr(,"class")
## [1] "coef.mer"
```

절편만 상이하고 나머지 fixed effect에 대한 계수 추정치는 모두 동일함을 확인할 수 있다.

2) Mixed model with correlated random intercept and slope

- model equation

$$logit(\theta_{it}) = \beta_0 + S_{0i} + \beta_1 sex + (\beta_2 + S_{2i})time + \beta_3 feed + \beta_4 time \times feed \qquad (2)$$

- random effect을 넣음으로써 얻는 효과
  subject별로 상이한 intercept, time 계수를 가지는 로지스틱 회귀식.

```
locust_glmer_slope = glmer(move ~ time*feed + sex +  (1 + time | id), family=binomial, data=locust
VarCorr(locust_glmer_slope)


##  Groups Name        Std.Dev. Corr
##  id     (Intercept) 1.1482
##         time        1.4634   -0.629
```

random effect의 분산이 0과 가깝게 나오지 않았다. 또한 두 random effect간에 독립을 가정하지 않았기 때문에 correlation이 -0.629로 존재함을 확인할 수 있다.

```
coef(locust_glmer_slope)


## $id
##    (Intercept)        time     feed1      sex1 time:feed1
## 1  -1.69303972  0.57837575 -3.971469 0.3349745  0.8213679
## 2  -1.17198827  3.49176287 -3.971469 0.3349745  0.8213679
## 3  -0.70898800  0.77018752 -3.971469 0.3349745  0.8213679
```

```
## 4   0.03455375  0.34929864 -3.971469 0.3349745  0.8213679
## 5  -1.66505882  1.56756740 -3.971469 0.3349745  0.8213679
## 6  -1.07653250  0.41092526 -3.971469 0.3349745  0.8213679
## 7  -0.13675677  1.57543540 -3.971469 0.3349745  0.8213679
## 8  -0.47942491 -0.08054621 -3.971469 0.3349745  0.8213679
## 9  -2.75930195  3.03392087 -3.971469 0.3349745  0.8213679
## 10 -2.04639434  4.34656067 -3.971469 0.3349745  0.8213679
## 11  0.71083771  0.75471801 -3.971469 0.3349745  0.8213679
## 12 -1.01748622 -0.09348024 -3.971469 0.3349745  0.8213679
## 13 -0.99835539  1.39505637 -3.971469 0.3349745  0.8213679
## 14 -0.36177483  0.10564120 -3.971469 0.3349745  0.8213679
## 15  0.43273673  0.65681639 -3.971469 0.3349745  0.8213679
## 16  0.12608146  1.87740399 -3.971469 0.3349745  0.8213679
## 17 -0.09538744  0.24227523 -3.971469 0.3349745  0.8213679
## 18 -1.68533309  0.27432445 -3.971469 0.3349745  0.8213679
## 19 -2.20690493  3.48905983 -3.971469 0.3349745  0.8213679
## 20 -2.31662323  2.48745345 -3.971469 0.3349745  0.8213679
## 21 -1.29272694  0.58819917 -3.971469 0.3349745  0.8213679
## 22 -0.21772727  1.54878592 -3.971469 0.3349745  0.8213679
## 23 -1.48773642  0.75357131 -3.971469 0.3349745  0.8213679
## 24 -2.63714467  2.91095548 -3.971469 0.3349745  0.8213679
##
## attr(,"class")
## [1] "coef.mer"
```

subject별로 intercept와 time의 계수가 다르게 나온다. 이 두 변수에 대해서 subject random effect을 추가했기 때문이다.

이제 random intercept만 추가한 mixed model과 random slope까지 추가한 mixed model 중 어느 것이 더 좋은지 판단해야 한다. anova() 함수를 이용하여 LRT를 통해 판단한다. 먼저 귀무가설과 대립가설을 살펴보면

$$H_0 : \ Reduced\ Model(random\ intercept\ only)\ vs\ H_1 : \ Full\ Model(random\ intercept,\ slope)$$

귀무 가설을 기각한다면 Full Model, 즉 random slope까지 포함한 mixed model을 선택한다.

```
anova(locust_glmer, locust_glmer_slope)
```

```
## Data: locust
## Models:
## locust_glmer: move ~ time * feed + sex + +(1 | id)
## locust_glmer_slope: move ~ time * feed + sex + (1 + time | id)
##                    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## locust_glmer        6 3185.1 3222.7 -1586.6   3173.1
## locust_glmer_slope  8 3124.6 3174.6 -1554.3   3108.6 64.546      2  9.641e-15
##
## locust_glmer
## locust_glmer_slope ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value가 매우 유의하므로 귀무가설을 기각한다. 즉, random slope까지 포함한 mixed model이 데이터를 더 잘 설명한다는 뜻이다.

3) Mixed model with independent random intercept and slope

```
locust_glmer_slope_indep = glmer(move ~ time*feed + sex + (1 | id)  + (0+time | id), family=binomi
VarCorr(locust_glmer_slope_indep)
```

```
##  Groups Name        Std.Dev.
##  id     (Intercept) 0.94285
##  id.1   time        1.19502
```

두 random effect간에 독립을 가정하였기 때문에 correlation이 존재하지 않는다.

```
anova(locust_glmer_slope,locust_glmer_slope_indep)
```

```
## Data: locust
## Models:
## locust_glmer_slope_indep: move ~ time * feed + sex + (1 | id) + (0 + time | id)
## locust_glmer_slope: move ~ time * feed + sex + (1 + time | id)
##                           Df    AIC    BIC  logLik deviance  Chisq Chi Df
## locust_glmer_slope_indep  7 3129.2 3173.0 -1557.6   3115.2
```

```
## locust_glmer_slope        8 3124.6 3174.6 -1554.3   3108.6 6.6027      1
##                           Pr(>Chisq)
## locust_glmer_slope_indep
## locust_glmer_slope          0.01018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

귀무가설과 대립가설은 다음과 같다.

$$H_0 : \ Reduced\ Model(indep)\ vs\ H_1 : \ Full\ Model(not\ indep)$$

p-value가 유의하므로 귀무가설을 기각한다. 즉, 독립을 가정하지 않은 mixed model이 더 적절하다.

4) Conclusion

위의 논의들을 종합할 때, 데이터를 가장 잘 설명하는 모형은 random slope, intercept을 포함하는 mixed model이다.