APTS Statistical Modelling

Lecture Discussion 3

Helen Ogden

April 2021

Lecture discussion 3

- Please mute your microphone and turn off video, unless you are asking a question.
- Feel free to ask questions about any of the lecture content so far.
- Use the chat to ask questions directly, or to state that you have a question.

Return to toxoplasmosis example

Recall the dataset `toxo` in `SMPracticals` provides data on the number of people testing positive for toxoplasmosis (`r`) out of the number of people tested (`m`) in 34 cities in El Salvador, along with the annual rainfall in mm (`rain`) in those cities.

```
head(toxo)
```

```
##   rain  m  r city
## 1 1620 18  5    1
## 2 1650 30 15    2
## 3 1650  1  0    3
## 4 1735  4  2    4
## 5 1750  2  2    5
## 6 1750  8  2    6
```

Logistic regression

We can fit various logistic regression models for relating toxoplasmosis incidence to rainfall, of the form

$$R_i \sim \text{Binomial}(m_i, \mu_i), \quad \text{logit}(\mu_i) = \eta_i.$$

```
mod0 <- glm(r/m ~ 1, data = toxo, weights = m,
            family = "binomial")
mod3 <- glm(r/m ~ poly(rain, 3), data = toxo,
            weights = m, family = "binomial")
```

Logistic regression with random intercepts

Alternatively, we might want to include models with random intercepts

```
library(lme4)

toxo$city <- factor(1:nrow(toxo))
mod0_ri <- glmer(r/m ~ (1 | city),
                 data = toxo, weights = m,
                 family = "binomial")
mod3_ri <- glmer(r/m ~ poly(rain, 3) + (1 | city),
                 data = toxo, weights = m,
                 family = "binomial")
```

Comparing models via information criteria

```
AICs <- c(AIC(mod0), AIC(mod3),
          AIC(mod0_ri), AIC(mod3_ri))
AICs
```

```
## [1] 166.9044 161.3272 154.6499 154.6593
```

```
BICs <- c(BIC(mod0), BIC(mod3),
          BIC(mod0_ri), BIC(mod3_ri))
BICs
```

```
## [1] 168.4308 167.4326 157.7026 162.2911
```

Individual-level toxoplasmosis data

We could "expand out" the toxoplasmosis dataset, to give $\sum_{i=1}^{34} m_i = 697$ records on whether each individual in each city tests positive to toxoplasmosis:

```
head(toxo_ind)
```

```
##   rain y city
## 1 1620 1    1
## 2 1620 1    1
## 3 1620 1    1
## 4 1620 1    1
## 5 1620 1    1
## 6 1620 0    1
```

Models for individual-level data

We can fit various logistic regression models for relating toxoplasmosis incidence to rainfall, of the form

$$Y_{ij} \sim \text{Bernoulli}(\mu_i), \quad \text{logit}(\mu_i) = \eta_i.$$

```
mod0_ind <- glm(y ~ 1, data = toxo_ind,
                family = "binomial")
mod3_ind <- glm(y ~ poly(rain, 3), data = toxo_ind,
                family = "binomial")
mod0_ri_ind <- glmer(y ~ (1 | city), data = toxo_ind,
                     family = "binomial")
mod3_ri_ind <- glmer(y ~ poly(rain, 3) + (1 | city),
                     data = toxo_ind,
                     family = "binomial")
```

Parameter estimates from individual-level data

```
coef(mod0)
```

```
## (Intercept)
##  0.04304825
```

Is `coef(mod0_ind)` also 0.04304825?

- Yes, the two estimates are the same
- No, the two estimates are different
- Not sure or need more information

Parameter estimates from individual-level data

```
coef(mod0)
```

```
## (Intercept)
##  0.04304825
```

Is `coef(mod0_ind)` also 0.04304825?

- Yes, the two estimates are the same
- No, the two estimates are different
- Not sure or need more information

```
coef(mod0_ind)
```

```
## (Intercept)
##  0.04304825
```

AICs from individual-level data

```
AIC(mod0)
```

```
## [1] 166.9044
```

Recall AIC $= 2\left(p - \hat{\ell}\right)$. Is `AIC(mod0_ind)` also 166.9044?

- Yes, the two AICs are the same
- No, the two AICs are different
- Not sure or need more information

AICs from individual-level data

```
AIC(mod0)
```

```
## [1] 166.9044
```

Recall AIC $= 2\left(p - \hat{\ell}\right)$. Is `AIC(mod0_ind)` also 166.9044?

- Yes, the two AICs are the same
- No, the two AICs are different
- Not sure or need more information

```
AIC(mod0_ind)
```

```
## [1] 967.9243
```

Why is this happening?

For the original data, we have $R_i \sim \text{Binomial}(m_i, \mu_i(\beta))$ with likelihood

$$L(\beta) = \prod_{i=1}^{34} \binom{m_i}{r_i} \mu_i(\beta)^{r_i} (1 - \mu_i(\beta))^{m_i - r_i}.$$

For the individual-level data, we have $Y_{ij} \sim \text{Bernoulli}(\mu_i(\beta))$ with likelihood

$$L_{\text{ind}}(\beta) = \prod_{i=1}^{34} \prod_{j=1}^{m_i} \mu_i(\beta)^{y_{ij}} (1 - \mu_i(\beta))^{1 - y_{ij}} = \prod_{i=1}^{34} \mu_i(\beta)^{r_i} (1 - \mu_i(\beta))^{m_i - r_i},$$

so

$$\ell(\beta) = \sum_{i=1}^{34} \log \binom{m_i}{r_i} + \ell_{\text{ind}}(\beta).$$

The difference is constant in β, but does change the values of $\hat{\ell}$ in AIC.

Differences in AICs

```
AIC(mod3) - AIC(mod0)
```

```
## [1] -5.577275
```

Is `AIC(mod3_ind) - AIC(mod0_ind)` also -5.577275?

- Yes, the two differences in AICs are the same
- No, the two differences in AICs are different
- Not sure or need more information

Differences in AICs

```
AIC(mod3) - AIC(mod0)
```

```
## [1] -5.577275
```

Is `AIC(mod3_ind) - AIC(mod0_ind)` also -5.577275?

- Yes, the two differences in AICs are the same
- No, the two differences in AICs are different
- Not sure or need more information

```
AIC(mod3_ind) - AIC(mod0_ind)
```

```
## [1] -5.577275
```

Differences in AICs

```
AICs_ind <- c(AIC(mod0_ind), AIC(mod3_ind),
              AIC(mod0_ri_ind), AIC(mod3_ri_ind))
```

```
AICs - min(AICs)
```

```
## [1] 12.254594808  6.677319438  0.000000000  0.009471144
```

```
AICs_ind - min(AICs_ind)
```

```
## [1] 12.25459481  6.67731944  0.00000000  0.00947115
```

Differences in BICs

```
BIC(mod3) - BIC(mod0)
```

```
## [1] -0.9981938
```

Recall $\text{BIC} = 2\left(\frac{1}{2}p \log n - \hat{\ell}\right)$.

Is `BIC(mod3_ind) - BIC(mod0_ind)` also -0.9981938?

- Yes, the two differences in BICs are the same
- No, the two differences in BICs are different
- Not sure or need more information

Differences in BICs

```
BIC(mod3) - BIC(mod0)
```

```
## [1] -0.9981938
```

Recall $\text{BIC} = 2\left(\frac{1}{2}p \log n - \hat{\ell}\right)$.

Is `BIC(mod3_ind) - BIC(mod0_ind)` also -0.9981938?

- Yes, the two differences in BICs are the same
- No, the two differences in BICs are different
- Not sure or need more information

```
BIC(mod3_ind) - BIC(mod0_ind)
```

```
## [1] 8.063081
```

Comparing BICs

```
BICs_ind <- c(BIC(mod0_ind), BIC(mod3_ind),
              BIC(mod0_ri_ind), BIC(mod3_ri_ind))
```

```
BICs - min(BICs)
```

```
## [1] 10.728234  9.730040  0.000000  4.588553
```

```
BICs_ind - min(BICs_ind)
```

```
## [1]  7.707809 15.770890  0.000000 13.649827
```

Differences in BIC are not the same in the two equivalent forms of the model!

Why is this happening?

Recall

$$\text{BIC} = 2\left(\frac{1}{2}p\log n - \hat{\ell}\right).$$

How is n measured here?

```
nobs(mod0)
```

```
## [1] 34
```

```
nobs(mod0_ind)
```

```
## [1] 697
```

What should n be here?

The theoretical development of BIC (not covered in lectures) assumes n is the rate at which information grows, i.e. the Fisher information matrix $I(\theta)$ grows at rate n. See Neath and Cavanaugh (2011) for a nice review.

If we believe the logistic regression models, then should therefore take $n = 697$.

What should n be here?

The theoretical development of BIC (not covered in lectures) assumes n is the rate at which information grows, i.e. the Fisher information matrix $I(\theta)$ grows at rate n. See Neath and Cavanaugh (2011) for a nice review.

If we believe the logistic regression models, then should therefore take $n = 697$.

What if there is really clustering in the data, as in the mixed effects models?

Much more complex story: different components of the Fisher information matrix might grow at different rates!

See Delattre et. al. (2014) for some discussion.

Summary

- Be careful when applying information criteria: you should use the same form of response in all models being compared, and make sure the full loglikelihood (including all constant terms) is used.
- BIC requires particular care, and the defaults in R might give unexpected results.
- Exercise even more caution using BIC with mixed effects models.

Questions on lecture content?