Module recap and Q&A
PSM 2

Bennett Kleinberg

12 March 2019

Probability, Statistics & Modeling II

Today

Module recap and Q&A

- Model comparison
- The loglinear model
- Interpretation

Model comparison

When do we need it?

- you can model an outcome variable in many ways
  - $income \sim age + gender$
  - $income \sim age + gender + education$
  - $income \sim ethnicity + familystatus$
- Which model explains the data (i.e. the outcome) better?

How to do that comparison?

Nested vs unnested models

> *One model is nested in another if you can always obtain the first model by constraining some of the parameters of the second model.*

Nice explanation in this SO answer

Nested vs unnested models

Model 1: $Y \sim x1 + x2 + x1 : x2 + x3$

Model 2: $Y \sim x1 + x2$

Can we constrain the parameters of Model 1 to obtain Model 2?

Model parameters

Model 1: $Y \sim \beta_1 x1 + \beta_2 x2 + \beta_3 (x1 : x2) + \beta_4 x3$

Model 2: $Y \sim \beta_1 x1 + \beta_2 x2$

Can we constrain the parameters of Model 1 to obtain Model 2?

–> Yes: set $\beta_3 = \beta_4 = 0$ so that $Model1 = Model2$

Nested vs unnested models

Model 1: $Y \sim x1 + x2$

Model 2: $Y \sim x1$

Nested?

Nested vs unnested models

Model 1: $Y \sim x1 + x2$

Model 2: $Y \sim x1 + x3$

Nested?

Nested vs unnested models

Model 1: $Y \sim x1 + x2 + x3 + x4$

Model 2: $Y \sim x5$

Nested?

Nested models

$$income \sim age + gender$$

$$income \sim age + gender + education$$

Nested?

Nested models

M1: $income \sim age + gender$

M2: $income \sim age + gender + education$

M3: $income \sim ethnicity + familystatus$

Nested models

$$income \sim age + gender$$

$$income \sim age + gender + education$$

In essence: do we really need the additional predictor $education$?

Nested structure allows for formal statistical tests!

Formal model comparison logic

- if nested, we can test whether a simpler model is significantly worse than a more complex model
- if the model comparison is sign., then choose the more complex model
- if the test is not sign., choose the simpler one (Ockham's razor principle)

Non-nested models

$$income \sim age + gender + education$$

vs.

$$income \sim ethnicity + familystatus$$

No formal test possible?

Non-nested models

- for non-nested models, compare goodness of fit indices
- e.g. sum of squared residuals, mean absolute error, …
- other fit indices: AIC, Log-likelihood, BIC

In essence: you have to make a judgment without formal statistical test.

The loglinear model

A step back:

	No anti-virus software	**Anti-virus software**	**Sum**
Hacked	300	250	550
Not hacked	200	250	450
Sum	500	500	1000

For r-by-c tables

Idea of the Chi-square test:

- Observed values O
- Expected values (if there were no association) E
- rows: i
- columns: j

$$E_{i,j} = \frac{(total_i * total_j)}{total}$$

Chi-square test for r-by-c

$$\chi^2 = \sum \frac{(O-E^2)}{E}$$

- Null-hypothesis: there is no association between the two factors
- Alt. hypothesis: there is a significant association

Thus: if sign. –> there is a sign. association between r and c

More dimensions?

```
##                               vandalised yes  no
## area   natural surveillance
## urban  yes                               911  44
##        no                                538 456
## suburb yes                                 3   2
##        no                                 43 279
```

The Log-Linear Model

- GLM with link function for count data
- count data aptly modelled as a Poisson distrubuted variable

Stepwise

1. we build the “independence” model
   - no relationships between variables
2. we assess the H_0 of model adequacy
   - if significant: model not adequate for the data
   - if non-sign.: model is considered adequate
3. we build more complex models
   - e.g. with dependencies (i.e. interactions) between variables

Stepwise

- remember, we're modelling counts that come about due to a combination of factors
- thus: the saturated (= full) model will explain the data perfectly

Example

```r
example = array(c(40, 70, 80, 30), dim=c(2,2))
dimnames(example) = list('gender' = c('male', 'female')
                        , 'UK' = c('yes', 'no')
                        )
ftable(example)
```

```
##          UK yes no
## gender
## male         40 80
## female       70 30
```

Example

```
(exampledata = as.data.frame(as.table(example)))
```

```
##   gender  UK Freq
## 1   male yes   40
## 2 female yes   70
## 3   male  no   80
## 4 female  no   30
```

Example

```
indep_model = glm(formula = Freq ~ gender + UK
                  , data = exampledata
                  , family = poisson)
```

Next:

- look at H_0 of model adequacy
- look at predicted values

Model adequacy hyp.

```r
summary(indep_model)
```

```
##
## Call:
## glm(formula = Freq ~ gender + UK, family = poisson, data = exampledata
##
## Deviance Residuals:
##      1       2       3       4
## -2.750   2.666   2.455  -3.058
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.094e+00  1.135e-01  36.078   <2e-16 ***
## genderfemale -1.823e-01  1.354e-01  -1.347    0.178
## UKno          7.856e-12  1.348e-01   0.000    1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

Model adequacy hyp.

```r
pchisq(30.048, 1, lower.tail = F)
```

```
## [1] 4.21483e-08
```

Predicted values

gender	UK	Freq	round(fitted(indep_model), 2)
male	yes	40	60
female	yes	70	50
male	no	80	60
female	no	30	50

Example

Next step: full model

```
full_model = glm(formula = Freq ~ gender*UK
                , data = exampledata
                , family = poisson)
```

Example

gender	UK	Freq	round(fitted(full_model), 2)
male	yes	40	40
female	yes	70	70
male	no	80	80
female	no	30	30

Conclusion

```
##         UK yes no
## gender
## male        40 80
## female      70 30
```

There is an association between gender and "UK".

Making sense of the coefficients

```
coefficients(full_model)
```

```
##          (Intercept)        genderfemale                    UKno genderfemale:UKn
##            3.6888795           0.5596158               0.6931472          -1.54044
```

```
exp(coefficients(full_model))
```

```
##          (Intercept)        genderfemale                    UKno genderfemale:UKn
##           40.0000000           1.7500000               2.0000000           0.21428
```

Remeber: we're modelling the log (hence log-linear model)

- UK_no: 2.00
  - The odds of a person being from the UK are 1:2.00, regardless of their gender.
- gender_female: 1.75
  - The odds of a person being female are 1.75:1, regardless of their UK status.
- gender_female:UK_no: 0.21
  - People that are female have estimated odds of not being from the UK is 0.21 times the odds for males of not being from the UK.

Log-linear model

- extends Chisquare idea to mutliple dimensions
- brings in the modelling aspect
- aim: find a model that is simpler than the full model
- core: simplest model to explain the data

Interpretation

Interpretation of results

General strategy:

- there's always a hypothesis
- make the hypothesis explicit
- every RQ must come down to one or multiple hypotheses

Hypotheses

- difference in means 2 groups (t-test, rank sum test)
- difference in means 2+ groups (ANOVA, Kruskal-Wallis test)
- predictor combinations to explain an outcome (model comparison tests)
  - linear models
  - logistic regression models
  - log linear models

Interpretation strategy

When you ran your test/model:

- ask yourself: what did I test?
- which hypothesis was behind the test?
- what does the hyp. testing result reveal?
- how does this feed back to my RQ?

Pitfalls

- forgetting to re-transform coefficients in logistic regression or loglinear models
- forgetting the unit of interpretation of coefficients
- forgetting the direction of effects
- attributing causality to correlational data

Your interpretation becomes very difficult if you do not know the question you want to answer.

Easiest trick

Always start with the question!

Open Q&A session

Next week

CLASS TEST

- Tuesday, 19 March 2019
- 10am-12pm
- 60 min
- 10 questions (5 MC, 5 open)

END