Log-linear Part 5

Fixed Margins and Logit Models

Please read 6.1, 6.2, 6.4

Product Multinomial Models

- Separate multinomial distribution for each combination of explanatory variable values.
- Very natural for experimental studies in which the explanatory variables are controlled by the investigator (ideally, with random assignment to experimental conditions). Marginal totals are fixed by the design.
- Also reasonable for observational studies if you want to do the analysis conditionally upon the values of the explanatory variables (as in conditional probability).
- (Note that standard regression methods are conditional).

Marginal totals are fixed by the design

- Email fund-raising study
- 3,000 emails (spam) asking for donations to a children's charity
- Recipients have contributed in the past.
- They are randomly assigned to one of 6 pictures of a child (500 each).
- Two explanatory variables: Picture of girl *vs.* boy, and type of disability
- Response variable is whether they make an online donation.

Marginal table of Child's gender by Type of disability is fixed by the design

	Wheelchair	Down syndrome	No disability
Female	500	500	500
Male	500	500	500

```
spam <- numeric(12); dim(spam) <- c(2,3,2)
spamlabels <- list() # An empty list
spamlabels$Gender <- c("Female","Male")
spamlabels$Disability <- c("Wheelchair"," Down syndrome"," None Visible")
spamlabels$Donation <- c("Yes","No")
dimnames(spam) <- spamlabels
spam[,,1] <- rbind( c( 70,  29,  13),
                    c( 73,  19,  21) )
spam[,,2] <- rbind( c(430, 471, 487),
                    c(427, 481, 479) )
```

To fit a product multinomial model

- Use a standard multinomial log-linear model that includes all interactions among explanatory variables, whether or not they significantly help model fit.
- Theorem: This yields same MLEs as a log-linear model that directly incorporates the product multinomial structure.
- Idea: If some marginals are fixed by the design, they are known *exactly*, and if their terms are omitted, fit will (usually) be worse. Use the information we have!

Does it always matter if the interactions among explanatory variables are included?

```
> # 1=Gender,  2=Disability,  3=Donation
> loglin(spam,list(c(1,2),c(1,3),c(2,3)))$lrt
3 iterations: deviation 0.005521471
[1] 4.238558
> loglin(spam,list(c(1,3),c(2,3)))$lrt
2 iterations: deviation 0
[1] 4.238726
```

	Wheelchair	Down syndrome	No disability
Female	500	500	500
Male	500	500	500

Not if there is exactly zero relationship between explanatory variables (“Balanced Design”).

What about degrees of freedom
for the goodness of fit test?

- Df = number of terms in the saturated model minus number of terms in the model under consideration
- Null hypothesis is that the terms that are in the saturated model but not in the model being considered are all zero.
- Terms that are known to be zero are also zero in the saturated model.
- *Difference* in df is the same.
- So you can trust the model with all interactions among explanatory variables.

Or, if you really want to

- For balanced designs, you can fit the model with no interactions among explanatory variables, and count the degrees of freedom by hand.
- But remember, if sample sizes are all equal, main effects for the explanatory variables are missing too.
- For equal sample sizes, the correct no-interaction model is usually not hierarchical.

Situations where you might want a conditional model

- Experimental study that goes as planned – almost always balanced.
- Experimental study that goes wrong – drop a test tube, lose some data randomly with respect to response. Unbalanced design by accident.
- Observational study with clear distinction between explanatory and response vars.
  - Unbalanced (usually)
  - Balanced by selection (sometimes)

If the design is

- **Unbalanced**: Model with interactions among explanatory variables is mandatory (for a conditional model). Otherwise, estimates of parameters that are actually unknown partly compensate to fit the data, and are not as good as they could be.
- **Balanced**: Model with interactions among explanatory variables does the job and is usually less trouble.
- Note that conditional models for observational studies ignore information about relationships among explanatory variables. This is a weakness.

Developing Linear Logit Models

- Logit means log odds
- Includes logistic regression
- These are all conditional models

- In the following example, suppose
  - Variables 1 and 2 are explanatory, Variable 3 is response.
  - Variable 3 has just 2 categories (for now).
  - Conditional

Conditional model

$$\begin{aligned}\log m_{ijk} = \mu &+ \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} \\ &+ \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \\ &+ \mu_{123(ijk)}\end{aligned}$$

Response category One minus category Two

$$\begin{aligned}\log m_{ij1} - \log m_{ij2} &= \log\left(\frac{m_{ij1}}{m_{ij2}}\right) \\ &= \log\left(\frac{p_{ij1}}{p_{ij2}}\right) \\ &= \text{Conditional log odds}\end{aligned}$$

$$\begin{aligned}\log m_{ijk} = \mu &+ \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} \\ &+ \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \\ &+ \mu_{123(ijk)}\end{aligned}$$

$$\begin{aligned}\log\left(\frac{p_{ij1}}{p_{ij2}}\right) &= \log m_{ij1} - \log m_{ij2} \\ &= [\mu_{3(1)} - \mu_{3(2)}] \\ &\quad + [\mu_{13(i1)} - \mu_{13(i2)}] + [\mu_{23(j1)} - \mu_{23(j2)}] \\ &\quad + [\mu_{123(ijk)} - \mu_{123(ijk)}] \\ &\quad + [\mu_{123(ij1)} - \mu_{123(ij2)}] \\ &= 2\mu_{3(1)} + 2\mu_{13(i1)} + 2\mu_{23(j1)} + 2\mu_{123(ij1)} \\ &= 2\left(w + w_{1(i)} + w_{2(j)} + w_{12(ij)}\right)\end{aligned}$$

A linear model for the log odds

Roadmap

- Ordinary regression with dummy variables (review?)
- Logistic regression
  - With continuous variables
  - With dummy variables
- Poisson regression
- SAS (running under Linux)
  - Introduction
  - Log-linear models, Logistic regression
- Logistic regression with 2+ categories (SAS)
- Logistic regression with ordered categorical responses (SAS)