36-720: Log-Linear Models

Brian Junker

August 29, 2007

- Two-way ANOVA
- Two-way Log-Linear Model
- Odds Ratios, Independence, Interaction Plots
- Example 1: Husbands' & Wives' Heights
- Example 2: Politics by College
- Extending the Notation to Three-Way Tables

Two-way ANOVA

For continuous response data recall the two-way ANOVA model

$$y_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + \varepsilon_{ijk}\,, \quad \text{where } \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

($i = 1, \ldots, I;\ j = 1, \ldots, J$). This is equivalent to

$$y_{ijk} \overset{indep}{\sim} N(m_{ij}, \sigma^2)\,, \quad \text{where } m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

If the design is balanced with K observations per cell then the cell-means MLE's $\hat{m}_{ij} = \bar{y}_{ij}$ satisfy

$$\hat{m}_{ij} \sim N(m_{ij}, \sigma^2/K)$$

and we can learn everything about the table of means m_{ij} from the table of MLE's $\hat{m}_{ij}$ (except for estimating σ^2, which is essentially the MSE of the residuals $y_{ijk} - \hat{m}_{ij}$).

- A feature of the model

$$m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{1}$$

  is that it is over-parametrized: $u_{12(ij)}$ already contains $I \times J$ parameters corresponding to the cell means m_{ij} and we don't really need the additional $1 + I + J$ parameters $u + u_{1(i)} + u_{2(j)}$; for any choice of these we can compensate with $u_{12(ij)}$ to exactly match $\hat{m}_{ij}$.
- Constraints such as $\sum_i u_{1(i)} = 0$; $\sum_i \sum_j u_{12(ij)} = 0$ are a way to deal with this overparametrization.
- If the model changes, e.g. to

$$H_0 : \; m_{ij} = u + u_{1(i)} + u_{2(j)} \; , \tag{2}$$

  then the MLE's change from $\hat{m}_{ij} = \overline{y}_{ij}$ to

$$\hat{m}^0_{ij} = \overline{y}_i + \overline{y}_j - \overline{y} = \overline{y} + (\overline{y}_i - \overline{y}) + (\overline{y}_j - \overline{y}) \; ,$$

  at least in the balanced case of K observations per cell.
- We learn about the adequacy of a model like (2) by comparing the fit of its MLE's $\hat{m}^0_{ij}$ to the unconstrained MLE's $\hat{m}_{ij}$.

Two-way Log-Linear Model

Now let m_{ij} be the expected counts in an $I \times J$ table. An analogous model is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{3}$$

- *Why write in terms of* $\log m_{ij}$*?*
  - The observed counts n_{ij} and their expected values $m_{ij} = E[n_{ij}]$ are bounded below by zero (and above by n_{++}); this places awkward limits on the u-terms in (1); taking logs removes these limits in (3).
  - This is especially true when considering $n_{++} = 1$, which is useful for thinking about the jount distribution of the row and column variables!
  - Log-linear modeling is natural for the Poisson, Multinomial and Product-Multinomial sampling models.
  - There is a good asymptotic theory for (3).
- *What about the error term "+*ε_{ijk}*"?*
  - This is where the sampling models come in.
  - $n_{ij} \sim$ (sampling model with mean m_{ij}).

Once again the model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \, , \tag{3}$$

is over-parametrized. We will talk more about constraints to identify the model, but one common set is just as in ANOVA:

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i \sum_j u_{12(ij)} = 0 \, .$$

A more interesting model is the *additive log-linear model*:

- Under *multinomial sampling*, the $I \times J$ table satisfies independence iff $p_{ij} = p_{i+}p_{+j}$, so that

  $$m_{ij} = n_{++}p_{i+}p_{+j}$$

  or equivalently

  $$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{4}$$

- Under *product multinomial sampling*, the columns are independent of the rows iff $p_{1j} = \cdots = p_{Ij} \equiv \pi_j \; \forall \; j$; therefore

  $$m_{ij} = n_{i+}\pi_j$$

  and once again

  $$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{4}$$

- *The converses are also true*:
  - Under *multinomial sampling*, (4) implies $p_{ij} = p_{i+}p_{+j}$ and hence $m_{ij} = m_{i+}m_{+j}/m_{++}$;
  - Under *product-multinomial* sampling, (4) implies $p_{1j} = \cdots = p_{Ij} \equiv \pi_j \; \forall \; j$; and hence $m_{ij} = n_{i+}\pi_j$.

  The proofs are just careful bookkeeping.

Thus, *the additive log-linear model corresponds exactly to independence*.

Odds Ratios, Independence, Interaction Plots

By cancelling the n's in either the multinomial cell means $m_{ij} = n_{++}p_{ij}$ or the product multinomial cell means $m_{ij} = n_{i+}p_{ij}$, we see that the odds ratio

$$OR(i, j, i', j') = \frac{p_{ij}p_{i'j'}}{p_{i'j}p_{ij'}} = \frac{m_{ij}m_{i'j'}}{m_{i'j}m_{ij'}}$$

Taking logs, we see

$$\begin{aligned} \log OR(i, j, i', j') &= \log m_{ij} + \log m_{i'j'} - \log m_{i'j} - \log m_{ij'} \\ &= u_{12(ij)} + u_{12(i'j')} - u_{12(i'j)} - u_{12(ij')} \end{aligned}$$

where in the last line we have substituted $\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$. Thus

$$OR(i, j, i', j') = 1 \Leftrightarrow \sum_{r=1}^{I}\sum_{s=1}^{J} q_{rs}u_{12(rs)} = 0$$

where $q_{ij} = q_{i'j'} = 1$, $q_{i'j} = q_{ij'} = -1$ and otherwise $q_{rs} = 0$.
Since $\sum_r \sum_s q_{rs} = 0$, then $\sum_r \sum_s q_{rs}u_{12(ij)}$ is a *contrast*.

What does the null hypothesis

$$H_0 : u_{12(ij)} + u_{12(i'j')} - u_{12(i'j)} - u_{12(ij')} = 0$$

mean, in the model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \ ?$$

Let a, b, c, d be constants, such that

$$\begin{aligned} u_{12(ij)} &= a \\ u_{12(ij')} &= a + b \\ u_{12(i'j)} &= a + c \\ u_{12(ij)} &= a + d \end{aligned}$$

Then H_0 above is equivalent to

$$a + (a + d) - (a + b) - (a + c) = 0$$

or $d = b + c$. Taking

$$\begin{aligned} \alpha_i &= a \qquad \beta_j = 0 \\ \alpha_{i'} &= a + c \qquad \beta_{j'} = b \end{aligned}$$

we see that H_0 is equivalent to $u_{12(rs)} = \alpha_r + \beta_s$, for $r = i, i'$ and $s = j, j'$.
Since α_r and β_s can be subsumed into $u_{1(r)}$ and $u_{2(s)}$ respectively, H_0 above is equivalent to

$$H_0 : u_{12(rs)} = 0 \ , \ \ r = i, i' \ , \ \ s = j, j'$$

Independence of rows and columns

We know that

$$\begin{aligned}
(rows) \perp\!\!\!\perp (columns) \quad &\Leftrightarrow \quad OR(i, j, i', j') = 1\ , \quad \forall i, j, i', j' \\
&\Leftrightarrow \quad u_{12(ij)} + u_{12(i'j')} - u_{12(i'j)} - u_{12(ij')} = 0 \\
&\Leftrightarrow \quad OR(1, 1, i, j) = 1\ , \quad \forall i, j \\
&\Leftrightarrow \quad u_{12(11)} + u_{12(ij)} - u_{12(i1)} - u_{12(1j)} = 0
\end{aligned}$$

and there are clearly just $(I-1)(J-1)$ of the latter contrasts.

Applying the result of the previous slide to these contrasts we know that the

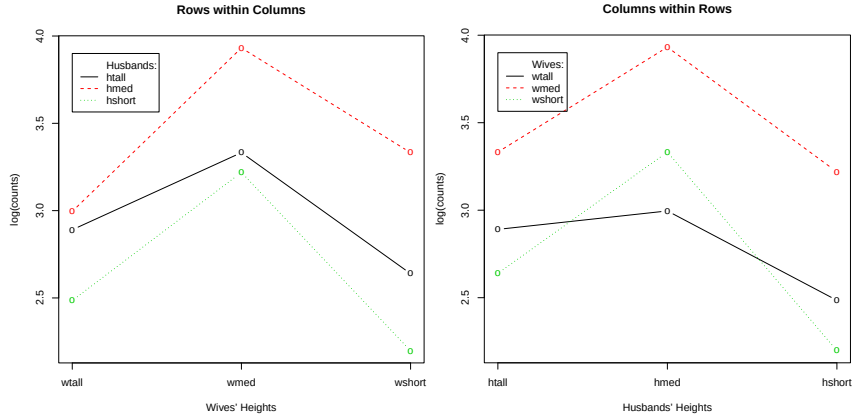
$$(rows) \perp\!\!\!\perp (columns) \Leftrightarrow u_{12(ij)} \equiv 0\ , \quad \forall i, j$$

Thus we can explore for independence by making a *log-linear interaction plot*, very much like an interaction plot for ANOVA models.

Example 1: Husbands' & Wives' Heights

OBS	W. Tall	W. Med	W. Short
H. Tall	18	28	14
H. Med	20	51	28
H. Short	12	25	9

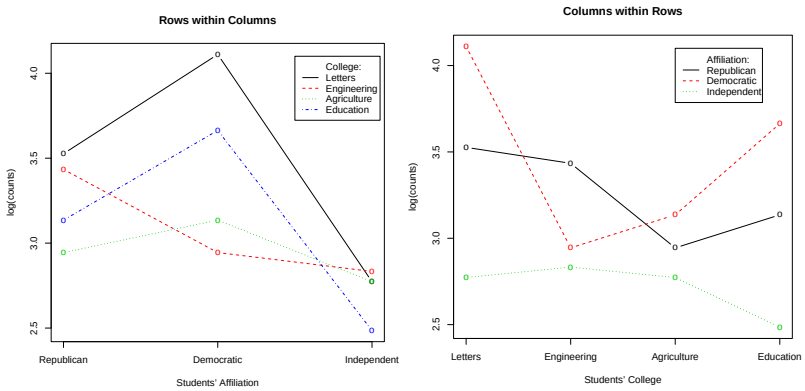
log(OBS)	W. Tall	W. Med	W. Short
H. Tall	2.89	3.33	2.64
H. Med	3.00	3.93	3.33
H. Short	2.48	3.22	2.20



$G^2 = 2.92$ on $(3-1)(3-1) = 4$ d.f.; $p = 0.57$; $\log m_{ij} = u + u_{1(i)} + u_{2(j)}$ seems OK.

Example 2: Politics by College

OBS	Rep	Dem	Indep	log(OBS)	Rep	Dem	Indep
L.Arts	34	61	16	L.Arts	3.53	4.11	2.77
Eng	31	19	17	Eng	3.43	2.94	2.83
Agr	19	23	16	Agr	2.94	3.14	2.77
Educ	23	39	12	Educ	3.14	3.66	2.48



$G^2 = 16.39$ on $(4-1)(3-1) = 6$ d.f.; $p = 0.01$; $\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$ probably needed.

Extending the Notation to Three-Way Tables

		Adversity of school (k)						
		Low		Med		High		
	Family Risk (j)	N	R	N	R	N	R	Total
Classrooom	Nondeviant	16	7	15	34	5	3	80
Behavior (i)	Deviant	1	1	3	8	1	3	17
	Total	17	8	18	42	6	6	97

- There are a variety of two-way tables here:
  - Conditional on low School Adversity we could examine

	N	R
Non	16	7
Dev	1	1

    to see if there is a relationship between Class. Beh. and Family Risk.
  - Conditional on deviant Classroom Behavior we could examine

	Low	Med	High
N	1	3	1
R	1	8	3

    to see if there is a relationship between Family Risk and School Adv.
- Any of these 2-way tables can be analyzed with $\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$.

Analysis of the two-way subtables is limited to questions of independence or dependence between pairs of variables.

If we expand to analysis of the full 3-way table, we can ask more interesting questions:

- Is classroom behavior independent of school adversity, given family risk factors?
- How does the relationship between classroom behavior and school adversity change, for boys from high-risk families vs. boys from low-risk families?
- etc.

Expanding the log-linear model *notation* to 3-way tables is not difficult:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} + u_{13(ik)} + u_{123(ijk)}$$

and the three main sampling models (Poisson, Multinomial, Product Multinomial) generalize as well.

The main questions for the next several lectures are:

- What do the u-terms mean in this model? What hypotheses on them correspond to conditional independence, etc.?
- What is a more efficient way to organize, specify, and interpret these models (and tables)?
- What is a more efficient way to fit them and select among competing models?