Lecture 13: Loglinear Models

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 6

`www.richardpmann.com/MATH3823`

Count Data and Contingency Tables

Setting:

- Response: counts Y_{ij} (non-negative integers)
- All explanatory variables are categorical (factors)
- Data often displayed in contingency tables

Examples:

- Disease cases by region and age group
- Customer purchases by product and store
- Survey responses by demographic factors

Example: Melanoma Data

400 malignant melanoma patients:

Tumor Type	Head/Neck	Trunk	Extremities	Total
Hutchinson’s	22	2	10	34
Superficial	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

Question: Is tumor type associated with body site?

The Poisson Model for Counts

Assume:

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij})$$

independently for each cell.

Loglinear model:

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Components:

- μ: overall mean (log scale)
- α_i: row effect (tumor type)
- β_j: column effect (body site)
- $(\alpha\beta)_{ij}$: interaction (association)

Parameter Constraints

Problem: The model is overparameterized.

R uses corner constraints:

- $\alpha_1 = 0$ (first row is reference)
- $\beta_1 = 0$ (first column is reference)
- $(\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$

Interpretation:

- α_i: log ratio of row i to row 1 (column effects averaged out)
- β_j: log ratio of column j to column 1 (row effects averaged out)
- $(\alpha\beta)_{ij}$: departure from additive model

The Independence Model

No interaction:

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j$$

Equivalently:

$$\lambda_{ij} = e^{\mu} \cdot e^{\alpha_i} \cdot e^{\beta_j}$$

Key property: Expected counts factor into row and column effects.

Under independence:

$$\hat{\lambda}_{ij} = \frac{y_{i+} \cdot y_{+j}}{y_{++}}$$

(row total × column total / grand total)

Maximum Likelihood for Independence Model

Fitted marginal totals equal observed marginal totals:

$$\hat{y}_{i+} = y_{i+}, \qquad \hat{y}_{+j} = y_{+j}$$

Fitted cell values:

$$\hat{\lambda}_{ij} = \frac{y_{i+} \cdot y_{+j}}{y_{++}}$$

This is the “expected count under independence” familiar from χ^2 tests.

Testing Independence

Deviance (likelihood ratio) test:

$$G^2 = 2 \sum_{i,j} y_{ij} \log \frac{y_{ij}}{\hat{\lambda}_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Pearson chi-squared:

$$X^2 = \sum_{i,j} \frac{(y_{ij} - \hat{\lambda}_{ij})^2}{\hat{\lambda}_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Both test H_0: Row and column variables are independent.

Degrees of freedom: $(r-1)(c-1)$ for $r \times c$ table.

Fitting in R

```r
# Create data frame
melanoma <- data.frame(
  Type = factor(rep(c("Hutchinson", "Superficial",
                      "Nodular", "Indeterminate"),
                    each = 3)),
  Site = factor(rep(c("Head", "Trunk", "Extrem"), 4)),
  Count = c(22, 2, 10, 16, 54, 115,
            19, 33, 73, 11, 17, 28)
)

# Independence model
model_ind <- glm(Count ~ Type + Site,
                 family = poisson, data = melanoma)

# Saturated model (with interaction)
model_sat <- glm(Count ~ Type * Site,
                 family = poisson, data = melanoma)
```

R Output: Independence Model

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.1517     0.1878  11.459  < 2e-16 ***
TypeIndeterminate  -0.4855     0.1731  -2.805  0.00503 **
TypeNodular         0.2993     0.1445   2.071  0.03837 *
TypeSuperficial     0.6946     0.1318   5.271 1.36e-07 ***
SiteHead           -1.2004     0.1421  -8.449  < 2e-16 ***
SiteTrunk          -0.7568     0.1173  -6.450 1.12e-10 ***

    Null deviance: 216.130  on 11  degrees of freedom
Residual deviance:  51.795  on  6  degrees of freedom
```

Test of independence:

- Deviance: 51.8 on 6 df
- p-value $< 0.001 \Rightarrow$ Strong evidence against independence

Examining Residuals

Large Pearson residuals indicate departure from model:

```
# Pearson residuals
residuals(model_ind, type = "pearson")
```

For melanoma data:

Type	Head	Trunk	Extrem
Hutchinson’s	+4.06	−2.02	−1.53
Superficial	−1.55	+0.49	+0.73
Nodular	+0.09	+0.31	−0.26
Indeterminate	+0.39	+0.64	−0.71

Hutchinson’s melanoma on head/neck is over-represented.

Multi-way Tables

Three-way contingency table: Factors A, B, C

Saturated model:

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

Hierarchy of models:

- Mutual independence: $\mu + \alpha_i + \beta_j + \gamma_k$
- Conditional independence: Add some two-way interactions
- Full model: Include three-way interaction

Hierarchical Models

Principle: If a higher-order term is included, all lower-order terms involving those factors must also be included.

Example: If $(\alpha\beta\gamma)_{ijk}$ is in the model, then must include:

- $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$
- α_i, β_j, γ_k
- μ

Why? Interpretability — effects should be measured relative to properly defined baselines.

Model Selection

Compare nested models via deviance:

$$\Delta G^2 = G^2_{\text{simpler}} - G^2_{\text{complex}} \sim \chi^2_{\Delta df}$$

Strategy:

1. Start with independence model
2. Add interactions one at a time
3. Test each addition using ΔG^2
4. Choose simplest adequate model

Or: Use AIC/BIC for automatic selection.

Summary

Key points:

- Loglinear models: Poisson GLM with categorical predictors
- $\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
- Independence: No interaction term
- Test independence via deviance or Pearson χ^2
- Residuals identify cells with poor fit
- Multi-way tables: hierarchy of interactions
- Hierarchical principle: include lower-order terms

Next lecture: Extensions — fixed marginals and product-multinomial models.