Lecture 11: Modelling Proportions — Overdispersion and Odds Ratios

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 5, Sections 5.3–5.4

`www.richardpmann.com/MATH3823`

Overdispersion: The Problem

Binomial assumption:

$$\mathrm{Var}[Y_i] = m_i p_i (1 - p_i)$$

In practice: The observed variance often exceeds this.

Detection:

- Residual deviance $\gg$ residual degrees of freedom
- Ratio $D/(n-r) \gg 1$

Example: $D = 45.2$ on 10 df suggests overdispersion.

Causes of Overdispersion

Possible sources:

1. **Missing covariates:**
    - Important variables not in the model
    - Unmeasured heterogeneity
2. **Incorrect link function:**
    - Logit may not be appropriate
3. **Lack of independence:**
    - Trials within groups may be correlated
    - Clustering effects

Modelling Overdispersion

Introduce a dispersion parameter $\tau > 1$:

$$\mathrm{Var}[Y_i] = \tau \cdot m_i p_i (1 - p_i)$$

Estimation:

$$\hat{\tau} = \frac{D}{n - r} \quad \text{or} \quad \hat{\tau} = \frac{X^2}{n - r}$$

Effect:

- Parameter estimates $\hat{\boldsymbol{\beta}}$ unchanged
- Standard errors multiplied by $\sqrt{\hat{\tau}}$
- Confidence intervals become wider
- p-values increase (more conservative)

Quasi-Binomial in R

```r
# Standard binomial (assumes phi = 1)
model1 <- glm(y ~ x, family = binomial)

# Quasi-binomial (estimates phi)
model2 <- glm(y ~ x, family = quasibinomial)

# Compare summaries
summary(model1)  # SE assuming no overdispersion
summary(model2)  # SE adjusted for overdispersion
```

Key difference:

- `binomial`: Fixed $\phi = 1$
- `quasibinomial`: Estimates ϕ from data

2×2 Contingency Tables

Setup:

	Success	Failure	Total
Group 1	y_1	$m_1 - y_1$	m_1
Group 2	y_2	$m_2 - y_2$	m_2

Probabilities:

- Group 1: $\mathbb{P}(\text{success}) = \pi_1$
- Group 2: $\mathbb{P}(\text{success}) = \pi_2$

Question: Is there an association between group and outcome?

Odds and Odds Ratio

Odds of success in each group:

$$O_1 = \frac{\pi_1}{1-\pi_1}, \qquad O_2 = \frac{\pi_2}{1-\pi_2}$$

Odds ratio:

$$\boxed{\psi = \frac{O_1}{O_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}}$$

Interpretation:

- $\psi = 1$: No association (same odds in both groups)
- $\psi > 1$: Group 1 has higher odds of success
- $\psi < 1$: Group 1 has lower odds of success

Estimating the Odds Ratio

Sample estimates:

$$\hat{\pi}_1 = \frac{y_1}{m_1}, \qquad \hat{\pi}_2 = \frac{y_2}{m_2}$$

Estimated odds ratio:

$$\hat{\psi} = \frac{y_1(m_2 - y_2)}{y_2(m_1 - y_1)}$$

Log odds ratio:

$$\log \hat{\psi} = \log y_1 - \log(m_1 - y_1) - \log y_2 + \log(m_2 - y_2)$$

Confidence Interval for Odds Ratio

Standard error of log odds ratio:

$$\mathrm{SE}(\log \hat{\psi}) = \sqrt{\frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2}}$$

95% CI for $\log \psi$**:**

$$\log \hat{\psi} \pm 1.96 \cdot \mathrm{SE}(\log \hat{\psi})$$

95% CI for ψ**:**

$$\left(\hat{\psi} \cdot e^{-1.96 \cdot \mathrm{SE}}, \quad \hat{\psi} \cdot e^{+1.96 \cdot \mathrm{SE}}\right)$$

If CI includes 1: No significant association.

Connection to Logistic Regression

Model: $\text{logit}(\pi_i) = \alpha + \beta \cdot \text{Group}_i$
where $\text{Group}_i = 0$ for group 2 (reference), $\text{Group}_i = 1$ for group 1.

Then:

$$\text{logit}(\pi_2) = \alpha$$
$$\text{logit}(\pi_1) = \alpha + \beta$$

Therefore:

$$\beta = \text{logit}(\pi_1) - \text{logit}(\pi_2) = \log \psi$$

And:

$$\boxed{e^{\beta} = \psi = \text{Odds Ratio}}$$

Example: Odds Ratio Calculation

Data:

	Disease	No Disease	Total
Exposed	30	70	100
Not Exposed	10	90	100

Odds ratio:

$$\hat{\psi} = \frac{30 \times 90}{10 \times 70} = \frac{2700}{700} = 3.86$$

Interpretation: Exposed individuals have 3.86 times the odds of disease compared to unexposed.

Relative Risk vs. Odds Ratio

Relative Risk (Risk Ratio):

$$\mathrm{RR} = \frac{\pi_1}{\pi_2}$$

Odds Ratio:

$$\mathrm{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Relationship:

- When π_1, π_2 are small: $\mathrm{OR} \approx \mathrm{RR}$
- OR always more extreme than RR (farther from 1)
- OR has nicer mathematical properties
- OR is what logistic regression estimates

Summary

Key points:

- Overdispersion: variance > binomial assumption
- Detect via $D/(n-r) \gg 1$
- Use `quasibinomial` to adjust standard errors
- Odds ratio: $\psi = O_1/O_2$
- $\psi = 1$ means no association
- From logistic regression: $\psi = e^{\beta}$
- CI for ψ via log transformation
- OR $\approx$ RR when probabilities are small

Next lecture: Dose-response experiments and applications.