# Lecture 10: Modelling Proportions — Logistic Regression

## MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

# Reading

**Course notes:** Chapter 5, Sections 5.1–5.2

www.richardpmann.com/MATH3823

# Binary and Binomial Responses

**Bernoulli trials:**

$$B = \begin{cases} 1 & \text{if event occurs (``success'')} \\ 0 & \text{otherwise (``failure'')} \end{cases}$$

with $\mathbb{P}(B = 1) = p$.

**Binomial distribution:**
Sum of $m$ independent Bernoulli trials with same $p$:

$$Y = \sum_{j=1}^{m} B_j \sim \text{Binomial}(m, p)$$

**Special case:** $m = 1$ gives Bernoulli (binary) data.

# The Logistic Regression Model

**Model specification:**

$$Y_i \sim \text{Binomial}(m_i, p_i)$$

with the logit link:

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \mathbf{x}_i' \boldsymbol{\beta}$$

**Equivalently:**

$$p_i = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{x}_i' \boldsymbol{\beta})}$$

This ensures $0 < p_i < 1$ for all $\boldsymbol{\beta}$.

# Interpreting Coefficients: Odds

**Odds of success:**

$$\text{Odds} = \frac{p}{1-p}$$

**Example:**

- $p = 0.8$: Odds $= 0.8/0.2 = 4$ (“4 to 1”)
- $p = 0.5$: Odds $= 1$ (“even odds”)
- $p = 0.2$: Odds $= 0.25$ (“1 to 4”)

**The logit is the log-odds:**

$$\text{logit}(p) = \log(\text{Odds})$$

# Interpreting Coefficients: Odds Ratios

**Model:** $\text{logit}(p) = \alpha + \beta x$

**For a unit increase in $x$:**

$$\begin{aligned}
\text{logit}(p_{x+1}) - \text{logit}(p_x) &= \beta \\
\log \frac{\text{Odds}_{x+1}}{\text{Odds}_x} &= \beta \\
\frac{\text{Odds}_{x+1}}{\text{Odds}_x} &= e^{\beta}
\end{aligned}$$

**Interpretation:**

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

A unit increase in $x$ multiplies the odds by $e^{\beta}$.

# Odds Ratio Examples

| $\beta$ | $e^{\beta}$ | Interpretation |
|---:|---:|:---|
| 0 | 1.00 | No effect |
| 0.5 | 1.65 | Odds increase by 65% |
| 1.0 | 2.72 | Odds nearly triple |
| $-0.5$ | 0.61 | Odds decrease by 39% |
| $-1.0$ | 0.37 | Odds reduced to 37% |

**Key point:**

- $e^{\beta} > 1$: Higher $x \Rightarrow$ higher probability of success
- $e^{\beta} < 1$: Higher $x \Rightarrow$ lower probability of success
- $e^{\beta} = 1$ $(\beta = 0)$: $x$ has no effect

# Maximum Likelihood Estimation

**Log-likelihood:**

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log p_i + (m_i - y_i) \log(1 - p_i) + \log \binom{m_i}{y_i} \right\}$$

where $p_i = \operatorname{logit}^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$.

**No closed form solution!**
Solved iteratively using Fisher Scoring / IRLS.

**Fitted values:**

$$\hat{p}_i = \frac{\exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}})}, \qquad \hat{y}_i = m_i \hat{p}_i$$

# Residuals for Binomial GLMs

**Pearson residuals:**

$$e_i^P = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}}$$

**Deviance residuals:**

$$e_i^D = \operatorname{sign}(y_i - m_i \hat{p}_i) \sqrt{d_i}$$

where

$$d_i = 2 \left\{ y_i \log \frac{y_i}{m_i \hat{p}_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i (1 - \hat{p}_i)} \right\}$$

For large $m_i$, residuals should be approximately $\mathcal{N}(0, 1)$.

# Deviance for Logistic Regression

**Model deviance:**

$$D = 2 \sum_{i=1}^{n} \left\{ y_i \log \frac{y_i}{\hat{y}_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{y}_i} \right\}$$

**Goodness of fit:** Under correct model, $D \sim \chi^2_{n-r}$.

**Alternative:** Pearson $\chi^2$ statistic

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - m_i \hat{p}_i)^2}{m_i \hat{p}_i (1 - \hat{p}_i)}$$

Both $D$ and $X^2$ are asymptotically $\chi^2_{n-r}$.

# Fitting Logistic Regression in R

## For grouped binomial data:

```
# y = number of successes, m = number of trials
y <- cbind(successes, failures)
model <- glm(y ~ x1 + x2, family = binomial)

# Or equivalently
model <- glm(cbind(successes, total - successes) ~ x1 + x2,
             family = binomial)
```

## For binary (0/1) data:

```
# y is 0 or 1 for each observation
model <- glm(y ~ x1 + x2, family = binomial)
```

# Example: R Output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -60.7175     5.1805  -11.72   <2e-16 ***
dose         34.2703     2.9122   11.77   <2e-16 ***

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.43
```

## Interpretation:

- $\hat{\beta} = 34.27$: log-odds ratio for unit dose increase
- $e^{34.27}$: odds ratio (very large)
- Residual deviance: 11.12 on 6 df — reasonable fit

# Hypothesis Testing

**Three types of tests:**

1. **Wald test** (from summary output):

$$z = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim \mathcal{N}(0,1) \text{ under } H_0 : \beta_j = 0$$

2. **Likelihood ratio test** (deviance difference):

$$D_0 - D_1 \sim \chi^2_{r_1 - r_0} \text{ under } H_0$$

3. **Goodness-of-fit test:**

$$D \sim \chi^2_{n-r} \text{ under correct model}$$

## Summary

**Key points:**

- Logistic regression models binomial/binary responses
- Logit link: $\text{logit}(p) = \log \frac{p}{1-p} = \mathbf{x}'\boldsymbol{\beta}$
- Coefficients are log-odds ratios
- $e^{\beta_j}$ = multiplicative effect on odds per unit increase in $x_j$
- MLE via iterative methods (Fisher Scoring)
- Deviance and Pearson $\chi^2$ for goodness of fit
- Use `family = binomial` in R

**Next lecture:** Overdispersion and odds ratios for $2 \times 2$ tables.