Lecture 15: Course Revision and Summary

MATH3823 Generalised Linear Models

Richard P Mann

Reading

Course notes: All chapters

`www.richardpmann.com/MATH3823`

Course Overview

We have covered:

1. Normal linear models (revision)
2. GLM theory: exponential families, link functions
3. GLM estimation: MLE, deviance, residuals
4. Logistic regression for proportions
5. Loglinear models for counts
6. Extensions: fixed marginals

Unifying theme: GLMs provide a flexible framework for modelling non-normal responses.

The GLM Framework

Random	**Systematic**	**Link**
$Y \sim$ Exponential family $f(y;\theta,\phi)$	Linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$	$g(\mu) = \eta$ $\mu = h(\eta)$

Key equation:

$$g(\mathbb{E}[Y]) = \mathbf{x}'\boldsymbol{\beta}$$

Exponential Family: Key Results

General form:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

Moments:

$$\mathbb{E}[Y] = b'(\theta), \qquad \mathrm{Var}[Y] = b''(\theta)\cdot\phi$$

Common members:

Distribution	θ	$b(\theta)$	Canonical link
Normal	μ	$\theta^2/2$	Identity
Poisson	$\log\lambda$	e^θ	Log
Binomial	$\mathrm{logit}(p)$	$m\log(1+e^\theta)$	Logit

Link Functions Summary

Response	Range of μ	Common links	R syntax
Continuous	$(-\infty, \infty)$	Identity	`gaussian`
Counts	$(0, \infty)$	Log	`poisson`
Proportions	$(0, 1)$	Logit, probit, cloglog	`binomial`
Positive	$(0, \infty)$	Log, reciprocal	`Gamma`

Logit vs Probit:

- Similar results in most cases
- Logit: coefficients are log-odds ratios
- Probit: arises from latent normal model

Maximum Likelihood Estimation

Score function:

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0} \quad \text{at } \hat{\boldsymbol{\beta}}$$

Fisher information:

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]$$

Asymptotic distribution:

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \mathbf{J}^{-1}(\boldsymbol{\beta}))$$

Standard errors: $\mathrm{SE}(\hat{\beta}_j) = \sqrt{[\hat{\mathbf{J}}^{-1}]_{jj}}$

Deviance and Model Comparison

Deviance:

$$D = 2\phi[\ell(\text{saturated}) - \ell(\text{fitted})]$$

Goodness of fit:

$$D \overset{a}{\sim} \chi^2_{n-p} \quad \text{under correct model}$$

Comparing nested models $M_1 \subset M_2$:

$$D_1 - D_2 \overset{a}{\sim} \chi^2_{p_2-p_1} \quad \text{under } H_0 : M_1 \text{ adequate}$$

When ϕ unknown: Use F-test instead.

Residuals

Three types:

1. **Raw:** $e_i = y_i - \hat{\mu}_i$
2. **Pearson:** $e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$
3. **Deviance:** $e_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$

Diagnostic plots:

- Residuals vs fitted: check for patterns
- Q-Q plot: check approximate normality
- Residuals vs covariates: check for missed effects

Logistic Regression

Model:

$$Y_i \sim \text{Binomial}(m_i, p_i), \quad \text{logit}(p_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

Key interpretations:

- β_j = change in log-odds per unit change in x_j
- e^{β_j} = odds ratio

Overdispersion:

- Detected when $D/(n-p) \gg 1$
- Handle with `quasibinomial`

LD50: $x_{50} = -\alpha/\beta$

Loglinear Models

Model:

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad \log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Independence model: No interaction term

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

Under independence:

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

Fixed marginals: Include corresponding terms in model.

R Commands Summary

```r
# Fit GLM
model <- glm(y ~ x1 + x2, family = binomial)

# Summaries
summary(model)            # Coefficients, SE, z-values
anova(model)              # Analysis of deviance
anova(m1, m2, test="Chisq")  # Compare models

# Extraction
coef(model)               # Coefficients
fitted(model)             # Fitted values
residuals(model, "deviance")  # Residuals
deviance(model)           # Deviance
predict(model, type="response")  # Predictions
```

Families in R

```r
# Normal (Gaussian)
glm(y ~ x, family = gaussian)

# Poisson
glm(y ~ x, family = poisson)

# Binomial (grouped)
glm(cbind(success, fail) ~ x, family = binomial)

# Binomial (binary)
glm(y ~ x, family = binomial)  # y is 0/1

# Quasi families (for overdispersion)
glm(y ~ x, family = quasipoisson)
glm(y ~ x, family = quasibinomial)
```

Common Exam Topics

Conceptual:

- Identify appropriate distribution and link
- Interpret coefficients (log-odds, odds ratios)
- Explain deviance and its uses
- Describe exponential family properties

Computational:

- Derive exponential family form
- Calculate fitted values and residuals
- Perform hypothesis tests using deviance
- Find LD50 from logistic regression

Applied:

- Fit and interpret R output
- Choose between models
- Diagnose model problems

Key Formulas to Remember

Exponential family moments:

$$\mathbb{E}[Y] = b'(\theta), \quad \text{Var}[Y] = b''(\theta)\phi$$

Logit function:

$$\text{logit}(p) = \log \frac{p}{1-p}, \quad p = \frac{e^{\eta}}{1+e^{\eta}}$$

Odds ratio from logistic regression:

$$\text{OR} = e^{\beta}$$

LD50:

$$x_{50} = -\alpha/\beta$$

Independence expected counts:

$$E_{ij} = \frac{y_{i+} \cdot y_{+j}}{y_{++}}$$

Final Thoughts

The power of GLMs:

- Unified framework for many types of data
- Flexible: choose distribution and link
- Well-understood theory and diagnostics
- Widely implemented in statistical software

Beyond this course:

- Mixed models (random effects)
- Generalized additive models (GAMs)
- Bayesian approaches
- Machine learning extensions

Good luck with your exams.