Lecture 7: GLM Theory — Link Functions

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 3, Sections 3.5–3.6

`www.richardpmann.com/MATH3823`

Recall: The Role of the Link Function

Problem: The mean μ may be constrained:

- Poisson: $\mu > 0$
- Binomial: $0 < \mu/m < 1$

But: The linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ is unconstrained.

Solution: Use a link function g such that:

$$\eta = g(\mu) \in \mathbb{R}$$

The inverse $h = g^{-1}$ maps η back to the valid range of μ.

Canonical Links Revisited

Recall: The canonical link sets $\eta = \theta$ (natural parameter).

Distribution	Canonical Link	$g(\mu)$
Normal	Identity	μ
Poisson	Log	$\log(\mu)$
Binomial	Logit	$\log \frac{p}{1-p}$
Gamma	Reciprocal	$-1/\mu$

Property: With canonical link,

$$g'(\mu) = \frac{1}{b''(\theta)} = \frac{1}{V(\mu)}$$

Advantages of Canonical Links

1. **Sufficient statistics:**

$$S_j = \sum_{i=1}^{n} y_i x_{ij}$$

   contain all information about β_j.

2. **Simple score equations:** MLE satisfies

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i) x_{ij} = 0$$

3. **Computational efficiency:** Algorithms converge reliably.

However: Other links may be more interpretable or provide a better fit.

Link Functions for Binomial Data

For proportions/probabilities $p \in (0, 1)$:

Name	Link $g(p)$	Inverse $h(\eta)$
Logit	$\log \frac{p}{1-p}$	$\frac{e^\eta}{1+e^\eta}$
Probit	$\Phi^{-1}(p)$	$\Phi(\eta)$
Cloglog	$\log(-\log(1-p))$	$1 - e^{-e^\eta}$
Cauchit	$\tan(\pi(p - 0.5))$	$0.5 + \frac{1}{\pi}\arctan(\eta)$

where Φ is the standard normal CDF.

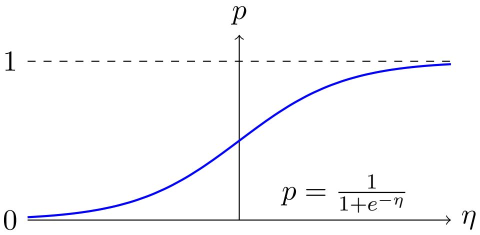
The Logit Link

Definition:

$$\operatorname{logit}(p) = \log \frac{p}{1-p} = \log(\text{odds})$$

Inverse (logistic function):

$$p = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$



The Probit Link

Definition:

$$\text{probit}(p) = \Phi^{-1}(p)$$

where Φ is the standard normal CDF.

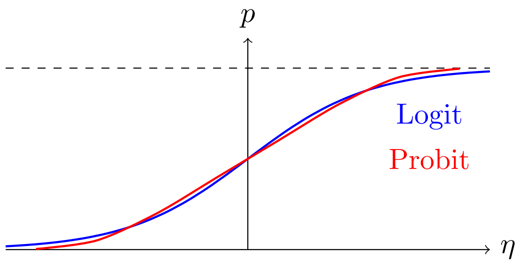
Inverse:

$$p = \Phi(\eta)$$

Interpretation:

- If there's a latent normal variable $Z = \eta + \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$
- Then $\mathbb{P}(Z > 0) = \Phi(\eta)$
- Probit link arises naturally from latent variable models

Comparing Logit and Probit



Key points:

- Very similar for moderate probabilities
- Logit has slightly heavier tails
- In practice, results are often nearly identical
- Logit preferred for interpretability (odds ratios)

The Complementary Log-Log Link

Definition:

$$\text{cloglog}(p) = \log(-\log(1 - p))$$

Inverse:

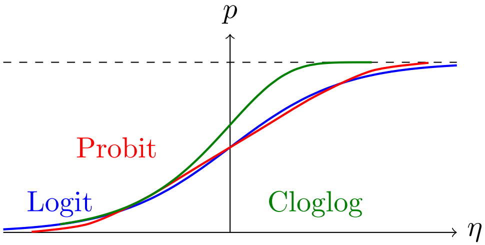
$$p = 1 - \exp(-e^{\eta})$$

Key property: Asymmetric — behaves differently near 0 and 1.

Use when:

- Probability of an event increases steeply then levels off
- Complementary to the log-log link: $\log(-\log(p))$
- Arises from extreme value distributions

Comparing Link Functions



Note: Cloglog approaches 1 faster but 0 slower than logit/probit.

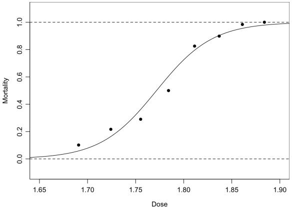
Beetle Data: Comparing Links

Fit logistic regression with different link functions:

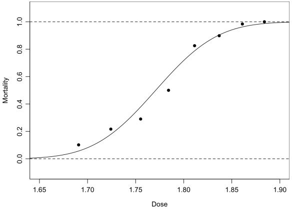
```r
# Logit link (default)
glm(y ~ dose, family = binomial(link = "logit"))

# Probit link
glm(y ~ dose, family = binomial(link = "probit"))

# Complementary log-log link
glm(y ~ dose, family = binomial(link = "cloglog"))
```

Compare using residual deviance, AIC, and residual plots.

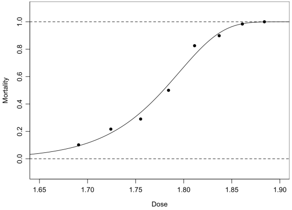
Beetle Data: Fitted Curves with Different Links



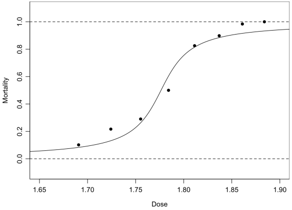
Logit



Probit



Cloglog



Cauchit

Link Functions for Poisson Data

For count data with $\mu > 0$:

Name	Link $g(\mu)$	Ensures $\mu > 0$?
Log (canonical)	$\log(\mu)$	Yes, $\mu = e^\eta > 0$
Square root	$\sqrt{\mu}$	Need $\eta > 0$
Identity	μ	Need $\eta > 0$

Log link is almost always used:

- Ensures positive fitted values
- Coefficients have multiplicative interpretation
- e^{β_j} = rate ratio for unit increase in x_j

Choosing a Link Function

Considerations:

1. **Theoretical:** Does a particular link arise from the science?
   - Probit from latent normal models
   - Cloglog from survival/extreme value theory

2. **Interpretability:**
   - Logit gives odds ratios
   - Log gives rate ratios

3. **Empirical fit:**
   - Compare deviances/AIC
   - Check residual patterns

R: Specifying Link Functions

```r
# Binomial family with different links
glm(y ~ x, family = binomial(link = "logit"))
glm(y ~ x, family = binomial(link = "probit"))
glm(y ~ x, family = binomial(link = "cloglog"))

# Poisson family with different links
glm(y ~ x, family = poisson(link = "log"))
glm(y ~ x, family = poisson(link = "sqrt"))

# Gaussian family
glm(y ~ x, family = gaussian(link = "identity"))
glm(y ~ x, family = gaussian(link = "log"))
```

Summary

Key points:

- Link functions map constrained μ to unconstrained η
- Canonical links have computational advantages
- For binomial: logit, probit, cloglog are common choices
- Logit and probit give similar results; logit is more interpretable
- Cloglog is asymmetric — useful for specific applications
- For Poisson: log link is standard
- Choose based on theory, interpretability, and fit

Next lecture: Maximum likelihood estimation for GLMs.