Regression Methods

Myrto Limnios

Autumn 2025 - Week 7

General Models - Generalized Linear Models

Generalised Linear Models (GLM)

Back to the big picture:

$$\boxed{Y \text{ (random output)} \overset{\text{whose law is influenced by}}{\longleftarrow} x \text{ (non-random input)}}$$

General formulation:

$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

Distribution / Function g	$g(x_i^\top) = x_i^\top \beta$	g **nonparametric**
Gaussian	Linear Regression ✓	Smoothing
Exponential Family	GLM ←	GAM

Motivation

- Need to generalise linear model beyond normal responses, e.g. to data with $y \in \{0, 1, \ldots, m\}$, or $y \in \{0, 1, \ldots\}$, or $y > 0$.
- Consider **exponential family** response distributions (binomial, Poisson, ...), which have an elegant unifying theory, and encompass many possibilities (in addition to the normal)
- Basic idea is to build models such that

$$\mathrm{E}(y) = \mu, \quad g(\mu) = \eta = x^{\mathrm{T}}\beta,$$

  where g is a suitable function, and $y \sim$ exponential family (almost).
- **Warnings**:
  - **Don't** confuse Generalized Linear Model (GLM) with General Linear Model (GLM, in older books, the latter is $y = X\beta + \varepsilon$, with $\mathrm{cov}(\varepsilon) = \sigma^2 V$ not diagonal);
  - **Don't** write $y = \mu + \varepsilon$, since in a GLM the distribution of ε usually depends on μ.

GLM

- Normal linear model has three key aspects:
  - structure for covariates: **linear predictor**, $\eta = x^{\mathrm{T}}\beta$;
  - response distribution: $y \sim N(\mu, \sigma^2)$;
  - linear relation $\eta = \mu$ between $\mu = \mathrm{E}(y)$ and η.
- GLM extends last two to
  - Y has density/mass function

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Omega_\theta \text{ open}, \phi > 0, \tag{1}$$

    where
    - $\mathcal{Y}$ is the support of Y,
    - Ω_θ is the parameter space of valid values for $\theta \equiv \theta(\eta)$, and
    - the **dispersion parameter** ϕ is often known;
  - $\eta = g(\mu)$, where g is monotone **link function**
    - the **canonical link** function giving $\eta = \theta = b'^{-1}(\mu)$ has nice statistical properties;
    - but a range of link functions are possible for each distribution of Y.

Interpretation of $\eta_j = x_j^\top \beta$?

- In key cases η_j is directly interpretable. If not, can switch perspective using the mean μ_j as defining parameter, connected to the **linear predictor** $x_j^\top \beta$ via
$$[b']^{-1}(\mu_j) = x_j^\top \beta = \eta_j$$
- Instead of $[b']^{-1}$ can use other **link functions** $g(\cdot)$ and postulate
$$g(\mu_j) = x_j^\top \beta.$$
This will also yield a GLM, but now the canonical parameter will not be equal to the linear predictor but to some function $u(x_j^\top \beta)$ of it.
- In summary, the nomenclature is:
  - $g(\cdot)$ is the **link function**
  - $h = g^{-1}$ is the **inverse link function**
  - $g(\cdot) = [b']^{-1}(\cdot)$ is the **canonical link function**

Examples

Example 14 (GLM density)

Show that the moment-generating function of $f(y;\theta,\phi)$ is $M_Y(t) = \exp[\{b(\theta+t\phi) - b(\theta)\}/\phi]$, and deduce that

$$\mathrm{E}(Y) = b'(\theta) = \mu, \quad \mathrm{var}(Y) = \phi b''(\theta) = \phi b''\{b'^{-1}(\mu)\} = \phi V(\mu);$$

the function $\mu \mapsto V(\mu)$ is known as the **variance function**.

Example 15 (Poisson distribution)

Write the Poisson mass function as a GLM density, and find its canonical link function.

Example 16 (Normal distribution)

Write the normal density function as a GLM density, and find its canonical link function.

Example 17 (IWLS algorithm - Estimation of β)

Find the components of the IWLS algorithm for a GLM.

Example 14

- Suppose that Y has a continuous density; if not the argument below is the same, except that integrals are replaced by summations.
- Let $\Omega_\theta = \{\theta : b(\theta) < \infty\}$. Then

$$\begin{aligned} M_Y(t) &= \mathrm{E}\{\exp(tY)\} \\ &= \int e^{ty} \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\} \mathrm{d}y \\ &= \int \exp\left\{\frac{y(\theta + t\phi) - b(\theta)}{\phi} + c(y;\phi)\right\} \mathrm{d}y. \end{aligned}$$

If $\theta + t\phi \in \Omega_\theta$, then

$$\int \exp\left\{\frac{y(\theta + t\phi) - b(\theta + t\phi)}{\phi} + c(y;\phi)\right\} \mathrm{d}y = 1,$$

so

$$M_Y(t) = \mathrm{E}\{\exp(tY)\} = \exp\left[\{b(\theta + t\phi) - b(\theta)\}/\phi\right].$$

- Hence the cumulant-generating function of Y is

$$K_Y(t) = \log M_Y(t) = \{b(\theta + t\phi) - b(\theta)\}/\phi,$$

and differentiating twice with respect to t and setting $t = 0$ yields

$$\mathrm{E}(Y) = K_Y'(t)\big|_{t=0} = b'(\theta), \quad \mathrm{var}(Y) = K_Y''(t)\big|_{t=0} = \phi b''(\theta).$$

- One can show that $b(\theta)$ is strictly convex on Ω_θ. Thus $b'(\theta)$ is a monotonic increasing function of θ, so $b'^{-1}(\cdot)$ exists and is itself monotonic, so $V(\mu) = b''\{b'^{-1}(\mu)\}$ is well-defined.

Example 15

The Poisson density may be written as

$$f(y;\mu) = \exp\left(y \log \mu - \mu - \log y!\right), \quad y = 0, 1, \ldots, \quad \mu > 0,$$

which has GLM form (1) with $\theta = \log \mu$, $b(\theta) = e^{\theta}$, $\phi = 1$, and $c(y;\phi) = -\log y!$. The mean of y is $\mu = b'(\theta) = e^{\theta} = \mu$, and its variance is $b''(\theta) = e^{\theta} = \mu$, so the variance function is linear: $V(\mu) = \mu$.

Example 16

The normal density with mean μ and variance σ^2 may be written

$$f(y;\mu,\sigma^2) = \exp\left\{-\frac{(y^2-2y\mu+\mu^2)}{2\sigma^2} - \tfrac{1}{2}\log(2\pi\sigma^2)\right\},$$

so

$$\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \tfrac{1}{2}\theta^2, \quad c(y;\phi) = -\tfrac{1}{2\phi}y^2 - \tfrac{1}{2}\log(2\pi\phi).$$

As the first and second derivatives of $b(\theta)$ are θ and 1, we have $V(\mu) = 1$; the variance function is constant.

Example 17

- If canonical link is used then $\theta_j = x_j^{\mathrm{T}}\beta$, so if ϕ is known, then

$$\begin{aligned} \ell(\beta) &= \sum_{j=1}^{n} \left\{ \frac{y_j x_j^{\mathrm{T}}\beta - b(x_j^{\mathrm{T}}\beta)}{\phi} + c(y_j;\phi) \right\} \\ &= \{y^{\mathrm{T}} X\beta - K(\beta)\}/\phi + C(y;\phi), \end{aligned}$$

  say, which in terms of β is a linear exponential family with

  - **canonical parameter** $\beta_{p\times 1}$
  - **canonical statistic** $(X^{\mathrm{T}}y)_{p\times 1}$,

  and many nice properties then hold.

- If X is full rank, then $\ell(\beta)$ is (almost always) strictly concave and has a unique maximum in terms of β.
- Problem: the maximum may be at infinity in certain (rare) cases—this can arise with binomial responses.

Example 17

- To compute the quantities needed for the IWLS step $\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W(X\beta + W^{-1}u)$, we need

$$X_{n\times p} = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}, \quad W_{n\times n} = \mathrm{diag}\{\mathrm{E}(-\partial^2 \ell_j/\partial \eta_j^2)\}, \quad u_{n\times 1} = \{\partial \ell_j/\partial \eta_j\},$$

where (with ϕ_j instead of ϕ for generality, see the next slide),

$$\ell_j(\beta) = \left\{ \frac{y_j\theta_j - b(\theta_j)}{\phi_j} + c(y_j;\phi_j) \right\}, \quad b'(\theta_j) = \mu_j, \quad \eta_j = g(\mu_j) = x_j^{\mathrm{T}}\beta.$$

- First note that $\partial \eta_j/\partial \beta_r = x_{jr}$, so $X = \partial \eta/\partial \beta^{\mathrm{T}}$ is just a matrix of constants.
- We need the first and second derivatives of ℓ_j with respect to η_j, so we write

$$\frac{\partial \ell_j}{\partial \eta_j} = \frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \theta_j}{\partial \mu_j}\frac{\partial \ell_j}{\partial \theta_j},$$

with

$$\frac{\partial \eta_j}{\partial \mu_j} = g'(\mu_j), \quad \frac{\partial \mu_j}{\partial \theta_j} = b''(\theta_j) = V(\mu_j), \quad \frac{\partial \ell_j}{\partial \theta_j} = \frac{y_j - b'(\theta_j)}{\phi_j},$$

which yields

$$u_j = \frac{\partial \ell_j}{\partial \eta_j} = \frac{y_j - b'(\theta_j)}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{y_j - \mu_j}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{A(\theta_j)}{B(\theta_j)},$$

say, where $\mathrm{E}(A) = 0$.

Example 17

- For the second derivative, we note that

$$\frac{\partial^2 \ell_j}{\partial \eta_j^2} = \frac{\partial}{\partial \eta_j}\frac{\partial \ell_j}{\partial \eta_j} = \left(\frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \theta_j}{\partial \mu_j}\frac{\partial}{\partial \theta_j}\right)\frac{\partial \ell_j}{\partial \eta_j} = \frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \theta_j}{\partial \mu_j}\left\{\frac{A'(\theta_j)}{B(\theta_j)} - \frac{A(\theta_j)B'(\theta_j)}{B(\theta_j)^2}\right\},$$

and on noting that $B(\theta_j)$ is non-random and $A'(\theta_j) = -b''(\theta_j) = -V(\mu_j)$, we obtain

$$w_j = \mathrm{E}\left(-\frac{\partial^2 \ell_j}{\partial \eta_j^2}\right) = \frac{1}{g'(\mu_j)}\frac{1}{V(\mu_j)}\frac{V(\mu_j)}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{1}{g'(\mu_j)^2 \phi_j V(\mu_j)}.$$

- From above we see that the components of the score statistic $u(\beta)$ and the weight matrix $W(\beta)$ may be expressed in terms of components μ_j of the mean vector μ as

$$\begin{aligned} u_j &= \frac{\partial \theta_j}{\partial \eta_j}\frac{\partial \ell_j(\theta_j)}{\partial \theta_j} = \frac{y_j - \mu_j}{g'(\mu_j)\phi_j V(\mu_j)}, \\ w_j &= \left(\frac{\partial \theta_j}{\partial \eta_j}\right)^2 \frac{\partial^2 \ell_j(\theta_j)}{\partial \theta_j^2} = \frac{1}{g'(\mu_j)^2 \phi_j V(\mu_j)}, \end{aligned} \qquad (2)$$

where $g'(\mu_j) = dg(\mu_j)/d\mu_j$. Thus $\widehat{\beta}$ is obtained by iterative weighted least squares regression of response

$$z = X\beta + g'(\mu)(y - \mu) = \eta + g'(\mu)(y - \mu)$$

on the columns of X using weights (2).

Example 17

- By using y as an initial value for μ and $g(y)$ as an initial value for $\eta = X\beta$, we avoid needing an initial value for β.
- It may be necessary to modify y slightly for this initial step. For example if we use the log link for Poisson data, and some y_j equal zero, then we may need to replace them with some small positive value to avoid taking $\log 0$ for some components of the initial $\eta = \log y$.

Estimation of ϕ

- When ϕ unknown, it is often replaced by $\phi_j = \phi a_j$, with known a_j and a_j^{-1} treated as a weight. Then we replace the scaled deviance by the **deviance** ϕD.
- If the model is correct and ϕ is known, then **Pearson's statistic**

$$P = \frac{1}{\phi} \sum_{j=1}^{n} \frac{(y_j - \widehat{\mu}_j)^2}{a_j V(\widehat{\mu}_j)} \stackrel{\cdot}{\sim} \chi^2_{n-p},$$

  analogously to the sum of squares in a linear model, with $\mathrm{E}(P) \doteq n - p$.
- The MLE of ϕ can be badly behaved, so usually we prefer the method of moments estimator

$$\widehat{\phi} = \frac{1}{n-p} \sum_{j=1}^{n} (y_j - \widehat{\mu}_j)^2 / \{a_j V(\widehat{\mu}_j)\},$$

  which is obtained by solving the equation $P = n - p$, based on noting that $\mathrm{E}(\chi^2_{n-p}) = n - p$.
- If the data are sparse (e.g., many small binomial or Poisson counts), then standard asymptotic results are suspect.

Summary

- Generalized linear models extend the classical linear model in two ways:
  - the response distribution is (almost) exponential family, so includes binomial, Poisson, gamma and other distributions in addition to the normal;
  - the relation between the linear predictor $\eta = x^{\mathrm{T}}\beta$ and the mean μ is determined by a wide range of possible link functions.
- Canonical link functions give particularly simple models and are widely used.
- Estimates of β are obtained by IWLS, which has a simple form, with no need for initial values.
- A simple estimate of the dispersion parameter ϕ is available using the method of moments.
- Models are compared using the analysis of deviance, which generalises the analysis of variance in the classical linear model.
- Standard likelihood theory results are used for inference (standard errors, confidence intervals, etc.)
- Standard diagnostics (residuals, ...) extend in a natural way to this setting.

General Models - Proportion Data

Binary response

- Response Y has Bernoulli distribution with

$$\Pr(Y=1)=\pi, \quad \Pr(Y=0)=1-\pi, \quad 0<\pi<1.$$

  and $\mathrm{E}(Y)=\mu=\pi$, $\mathrm{var}(Y)=\pi(1-\pi)$.
- Linear link function $\pi=\eta=x^{\mathrm{T}}\beta$ can give $\pi \notin [0,1]$, so not usually a good idea.
- Y can be interpreted in terms of a hidden variable/tolerance distribution: let $Z=x^{\mathrm{T}}\gamma+\sigma\varepsilon$, where $\varepsilon \sim F$. Set $Y=I(Z>0)$, and note that

$$\pi=\Pr(Y=1)=\Pr(x^{\mathrm{T}}\gamma+\sigma\varepsilon>0)=\Pr(\varepsilon>-x^{\mathrm{T}}\gamma/\sigma)=1-F(-x^{\mathrm{T}}\beta),$$

  say. Note that $\beta=\gamma/\sigma$ is estimable, but γ and σ are not.
- The corresponding link function is given by

$$\eta=x^{\mathrm{T}}\beta=-F^{-1}(1-\pi)=g(\pi),$$

  so different choices of F yield different possible link functions.

Link functions

Tolerance distributions and corresponding link functions for binary data.

Distribution F		Link function	
Logistic	$e^u/(1+e^u)$	Logit	$\eta = \log\{\pi/(1-\pi)\}$
Normal	$\Phi(u)$	Probit	$\eta = \Phi^{-1}(\pi)$
Log Weibull	$1-\exp(-\exp(u)\}$	Log-log	$\eta = -\log\{-\log(\pi)\}$
Gumbel	$\exp\{-\exp(-u)\}$	Complementary log-log	$\eta = \log\{-\log(1-\pi)\}$

- The logit and probit links are symmetric.
- Logit (canonical link) is usual choice, good for medical studies (later), with nice interpretation, but the probit is very similar to it and may be preferred in some cases, for its relation to the normal distribution.
- The log-log and complementary log-log links are asymmetric.

Logistic regression

- Common choice of link function for proportion data is the **logit**, which gives

$$\Pr(Y=1)=\pi=\frac{\exp(x^{\mathrm{T}}\beta)}{1+\exp(x^{\mathrm{T}}\beta)},\quad \Pr(Y=0)=1-\pi=\frac{1}{1+\exp(x^{\mathrm{T}}\beta)},$$

leading to a linear model for the **log odds** of success,

$$\log\left\{\frac{\Pr(Y=1)}{\Pr(Y=0)}\right\}=\log\left(\frac{\pi}{1-\pi}\right)=x^{\mathrm{T}}\beta,\quad \beta\in\mathbb{R}^p.$$

- The likelihood for β based on independent responses $y_1,\ldots,y_n$ with covariate vectors $x_1,\ldots,x_n$ and corresponding probabilities $\pi_1,\ldots,\pi_n$ is

$$L(\beta)=\prod_{j=1}^{n}\pi_j^{y_j}(1-\pi_j)^{1-y_j}=\cdots=\frac{\exp\left(\sum_{j=1}^{n}y_jx_j^{\mathrm{T}}\beta\right)}{\prod_{j=1}^{n}\left\{1+\exp\left(x_j^{\mathrm{T}}\beta\right)\right\}},$$

which is a regular exponential family with $s(y)=X^{\mathrm{T}}y$ and log likelihood

$$\ell(\beta)=(X^{\mathrm{T}}y)^{\mathrm{T}}\beta-\sum_{j=1}^{n}\log\left\{1+\exp\left(x_j^{\mathrm{T}}\beta\right)\right\},\quad \beta\in\mathbb{R}^p,$$

known as the **logistic regression model**.

Nodal involvement data

Data on nodal involvement: 53 patients with prostate cancer have nodal involvement (r), with five binary covariates age, stage, etc.

m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
1	1	0	0	1	0	1
1	0	0	0	0	1	1
1	0	0	0	0	1	0

Deviances for nodal involvement models

Scaled deviances D for 32 logistic regression models for nodal involvement data. + denotes a term included in the model.

age	st	gr	xr	ac	df	D	age	st	gr	xr	ac	df	D
					52	40.71	+	+	+			49	29.76
+					51	39.32	+	+		+		49	23.67
	+				51	33.01	+	+			+	49	25.54
		+			51	35.13	+		+	+		49	27.50
			+		51	31.39	+		+		+	49	26.70
				+	51	33.17	+			+	+	49	24.92
+	+				50	30.90		+	+	+		49	23.98
+		+			50	34.54		+	+		+	49	23.62
+			+		50	30.48		+		+	+	49	19.64
+				+	50	32.67			+	+	+	49	21.28
	+	+			50	31.00	+	+	+	+		48	23.12
	+		+		50	24.92	+	+	+		+	48	23.38
	+			+	50	26.37	+	+		+	+	48	19.22
		+	+		50	27.91	+		+	+	+	48	21.27
		+		+	50	26.72		+	+	+	+	48	18.22
			+	+	50	25.25	+	+	+	+	+	47	18.07

Model selection

- We have 32 competing models, and would like to select the 'best', or a few 'near-best'.
- In general we have 2^p models, so automatic selection of some sort is helpful.
- Could use likelihood ratio tests (differences of deviances) to compare competing models, but this involves many correlated tests, so may lead to spurious results.
- Usually minimise an information criterion, which accounts for the number of parameters in each model, such as

$$\text{AIC} \equiv D + 2p, \quad \text{BIC} \equiv D + p \log n,$$

  where D is the deviance.

- Recall their properties, with p fixed and as $n \to \infty$:
  - AIC tends to overfit, i.e., it has a positive probability of choosing a model that is too complex,;
  - BIC applies a stronger penalty, so *if the true model is among those fitted*, it will choose it with probability one;
  - BIC usually yields less complex models than AIC, but they may predict less well.
- There are many other information criteria, but these are most used in practice.

Example: Nodal involvement

- Model with lowest AIC has `stage`, `xray`, `acid`:

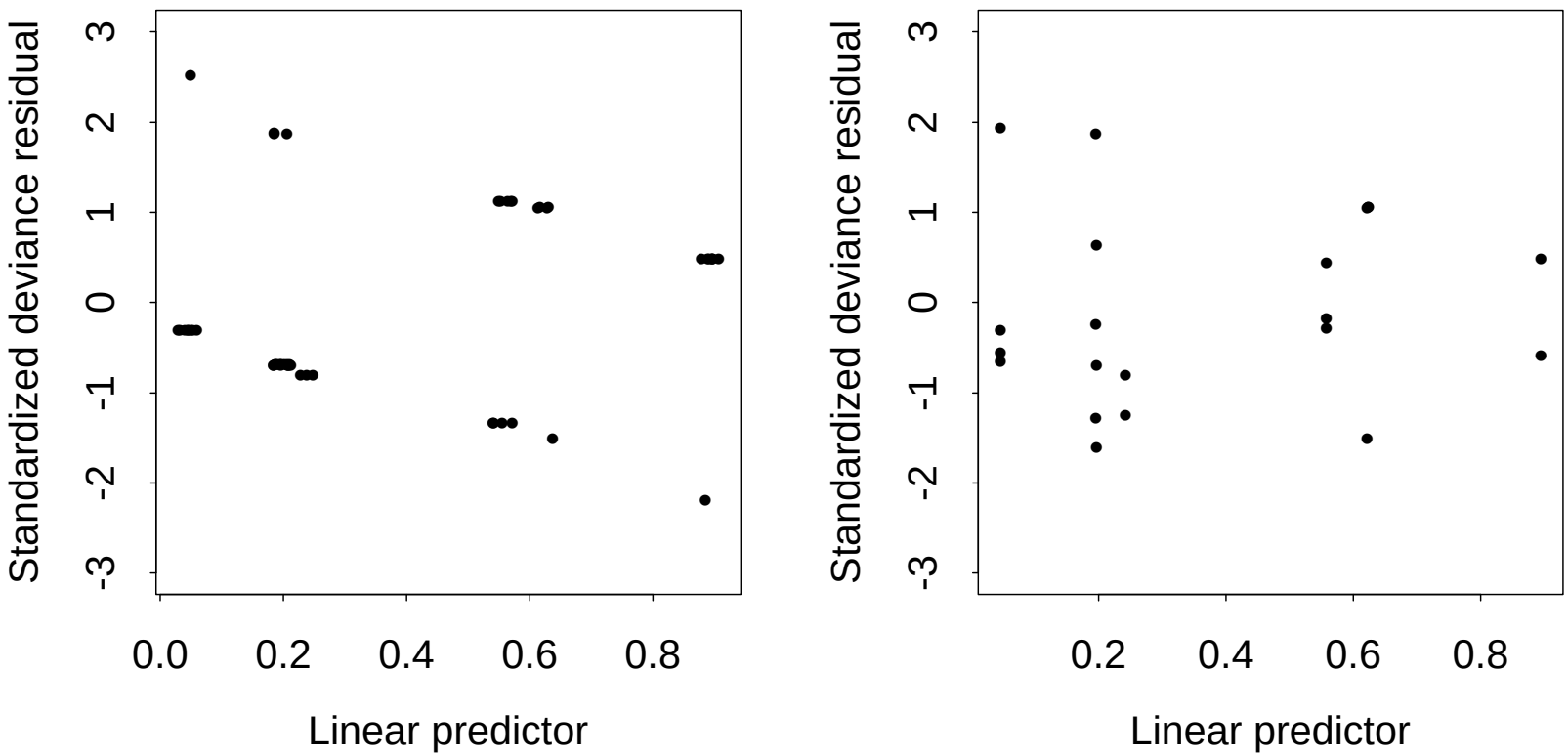
$$x^{\mathrm{T}}\widehat{\beta} = -3.05 + 1.65 I_{\texttt{stage}} + 1.91 I_{\texttt{xray}} + 1.64 I_{\texttt{acid}},$$

  where $I_{\texttt{stage}} = 1$ indicates that `stage` takes its higher level, etc.

- Interpretation of this model:
  - for an individual with `stage`, `xray` and `acid` at their lowest levels, the fitted probability of nodal involvement is $e^{-3.05}/(1 + e^{-3.05}) \doteq 0.045$ (though there are no such people in the data, so this involves extrapolation);
  - for someone with only $I_{\texttt{stage}} = 1$, the odds of nodal involvement are $e^{-3.05+1.65} = e^{-1.4} \doteq 0.25$, a probability of 0.2;
  - for someone with $I_{\texttt{stage}} = I_{\texttt{xray}} = I_{\texttt{acid}} = 1$, the odds of nodal involvement are $e^{-3.05+1.65+1.91+1.64} \doteq 8.6$, a probability of 0.9;
- Problems with interpretation of residual deviance of 19.64: how many df? — can amalgamate independent binary responses with same covariates.
- Likewise problems with residuals ...

Nodal involvement residuals

Figure: Standardized deviance residuals for nodal involvement data, for ungrouped responses (left) and grouped responses (right).



Summary

- Proportion data are often modelled using the Bernoulli/binomial response distributions.
- Link functions (logit, probit, ...) have interpretations in terms of underlying continuous variables that have been dichotomized.
- The canonical and most commonly-used link is the logit, and fitting using this yields logistic regression, in which
  - the canonical parameter is the log odds;
  - classical data structures (e.g., the 2×2 table) have nice interpretations.
- The deviance can be used to compare models (so can AIC, BIC, ...), but using its absolute value to assess fit can be dangerous (exercise).
- Residuals for binary data are not very informative.