Statistics in Geophysics: Generalized Linear Regression

Steffen Unkel

Department of Statistics
Ludwig-Maximilians-University Munich, Germany

Components of the classical linear model

- Generalized linear models (GLMs) are an extension of classical linear models.

- Recall the classical linear regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

- The systematic part of the model is a specification for the (conditional) mean of $\mathbf{y}$, which takes the form $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$.

- For the random part we assume $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. A further specialization of the model involves the assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

- Then, $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and the ith component of $\boldsymbol{\mu} \in \mathbb{R}^{n\times 1}$ is $\mu_i = \mathbf{x}_i^\top\boldsymbol{\beta}$ $(i = 1\ldots, n)$.

Components of a generalized linear model II

Three-part specification of the classical linear model:

1. The random component: $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$.

2. The systematic component: The p predictor variables produce a linear predictor $\boldsymbol{\eta} = (\eta_i, \ldots, \eta_n)^\top$, where

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} \ , \quad (i = 1, \ldots, n) \ .$$

3. The link between the random and systematic components:

$$\boldsymbol{\mu} = \boldsymbol{\eta} \ .$$

This specification introduces a new symbol η for the linear predictor and the 3rd component then specifies that μ and η are identical.

The generalization

- If we write

$$\eta_i = g(\mu_i) \quad \text{or} \quad \mu_i = h(\eta_i) \ ,$$

  then $g(\cdot)$ will be called the link function and $h(\cdot)$ the response function with $g = h^{-1}$.

- Classical linear models have a Gaussian distribution in component 1 and the identity function for the link in component 3.

- GLMs allow two extensions:
  1. The distribution in component 1 may come from an exponential family other than the Gaussian.
  2. The link function in component 3 may become any monotonic differentiable function.

Exponential family

- We assume that each component of $\mathbf{y}$ has a distribution in the (univariate) exponential family, taking the form

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

  for some specific functions $b(\cdot)$ and $c(\cdot)$.

- The parameter θ is called the natural or canonical parameter.

- The second parameter ϕ is a dispersion parameter.

- It can be shown that $\mathrm{E}(y) = \mu = b'(\theta)$ and $\mathrm{Var}(y) = \phi b''(\theta)$.

Exponential family parameters, expectation and variance

Distribution		$\theta(\mu)$	$b(\theta)$	ϕ
Normal	$\mathcal{N}(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2
Bernoulli	$\mathcal{B}(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Poisson	$\mathcal{P}(\lambda)$	$\log(\lambda)$	$\exp(\theta)$	1

Distribution	$E(y) = b'(\theta)$	$b''(\theta)$	$Var(y) = b''(\theta)\phi$
Normal	$\mu = \theta$	1	σ^2
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)$
Poisson	$\lambda = \exp(\theta)$	λ	λ

Maximum likelihood estimation in GLMs

- The ML estimator $\hat{\beta}$ is obtained in form of iteratively weighted least squares estimates

$$\hat{\beta}^{(t+1)} = (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \tilde{\mathbf{y}}^{(t)} \quad , \quad t = 0, 1, 2, \ldots$$

where $\mathbf{W}^{(t)} = \mathrm{diag}\left(\tilde{w}_1(\hat{\eta}_1^{(t)}), \ldots, \tilde{w}_n(\hat{\eta}_n^{(t)})\right)$ is a matrix of "working weights"

$$\tilde{w}_i(\hat{\eta}_i^{(t)}) = \frac{(h'(\hat{\eta}_i^{(t)}))^2}{\sigma_i^2(\hat{\eta}_i^{(t)})}$$

and $\tilde{\mathbf{y}}^{(t)} = \left(\tilde{y}_1(\hat{\eta}_1^{(t)}), \ldots, \tilde{y}_n(\hat{\eta}_n^{(t)})\right)^\top$ is a vector of "working observations" with elements

$$\tilde{y}_i(\hat{\eta}_i^{(t)}) = \hat{\eta}_i^{(t)} \frac{\left(y_i - h(\hat{\eta}_i^{(t)})\right)}{h'(\hat{\eta}_i^{(t)})} \quad .$$

Maximum likelihood estimation in GLMs II

- A key role in the iterations plays the matrix $\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}$.
- Invertibility of $\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}$ does not follow from the invertibility of $\mathbf{X}^\top \mathbf{X}$.
- However, usually all of the weights are positive such that $\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}$ is invertible.
- Then, the algorithm typically converges after a number of iterative steps.

Maximum likelihood estimation in GLMs III

Asymptotic properties of the ML estimator

- Let $\hat{\boldsymbol{\beta}}_n$ denote the ML estimator based on a sample of size n. Under regularity conditions:

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

  where $\mathbf{F}(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{W} \mathbf{X}$ is the expected Fisher information matrix.

- The expected Fisher information matrix is $\mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right)$, where $-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \mathbf{F}_{obs}$ is the observed Fisher information matrix and $l(\boldsymbol{\beta})$ is the log-likelihood.

Estimation of the scale parameter

- Denote by $v(\mu_i) = b''(\theta_i)$ the so-called variance function and note that $b''(\theta_i)$ implicitly depends on μ_i through the relation $b'(\theta_i) = \mu_i$.

- The dispersion parameter is estimated by

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \quad ,$$

where p denotes the number of regression parameters, $\hat{\mu}_i = h(\mathbf{x}_i^\top \hat{\beta})$ is the estimated expectation and $v(\mu_i)$ is the estimated variance function.

Testing linear hypotheses

Hypotheses H_0: $\mathbf{C}\beta = \mathbf{d}$ versus H_1: $\mathbf{C}\beta \neq \mathbf{d}$:

Let $\tilde{\beta}$ be the ML estimator under H_0.

- Test statistics:
  1. Likelihood ratio statistic: $lr = -2\left\{l(\tilde{\beta}) - l(\hat{\beta})\right\}$
  2. Wald statistic: $w = (\mathbf{C}\hat{\beta} - \mathbf{d})^\top [\mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^\top]^{-1}(\mathbf{C}\hat{\beta} - \mathbf{d})$
  3. Score statistic: $u = \mathbf{s}^\top(\tilde{\beta})\mathbf{F}^{-1}(\tilde{\beta})\mathbf{s}(\tilde{\beta})$
- Test decision: For large n and under H_0, it holds that

  $$lr, w, u \overset{a}{\sim} \chi^2_r \quad ,$$

  where r is the (full) row rank of the $r \times p$ matrix $\mathbf{C}$.
  We reject H_0 when

  $$lr, w, u > \chi^2_r(1-\alpha) \quad .$$

Criteria for model fit

- The most frequently used goodness-of-fit statistics are the Pearson statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

and the deviance

$$D = 2\left\{(l(\mathbf{y}) - l(\hat{\boldsymbol{\mu}})\right\} \quad ,$$

where $l(\hat{\boldsymbol{\mu}})$ and $l(\mathbf{y})$ represent the log-likelihood for the estimated and the saturated model, respectively.

- Both statistics are approximately χ^2_{n-p}-distributed.

Criteria for model selection

- The Akaike information criterion (AIC) for model selection is defined generally as

$$\mathrm{AIC} = -2l(\hat{\beta}) + 2p \ .$$

- The Bayesian information criterion (BIC) is defined generally as

$$\mathrm{BIC} = -2l(\hat{\beta}) + \log(n)p \ .$$

- If the model contains a dispersion parameter ϕ, its ML estimator should be substituted into the respective model and the total number of parameters should be increased to $p + 1$.

Binary regression models

- Suppose that the response variable y is binary and can take only two possible values, denoted by 0 and 1.

- We may write $\pi_i = \mathrm{P}(y_i = 1)$ and $1 - \pi_i = \mathrm{P}(y_i = 0)$ for the probabilities of 'success' and 'failure', respectively $(i = 1, \ldots, n)$.

- We want to model and and estimate the effects of the covariates on the (conditional) probability

$$\pi_i = \mathrm{P}(y_i = 1) = \mathrm{E}(y_i) \quad ,$$

  for the outcome $y_i = 1$ and given values of the covariates $x_{i1}, \ldots, x_{ik}$.

- In this specification, the response variables are assumed to be (conditionally) independent.

Binary regression models III

- We combine the probability π_i with the linear predictor η_i through a relation of the form

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \ ,$$

where the response function h is a strictly monotonically increasing cdf on the real line.

- This ensures $h(\eta) \in [0, 1]$ and the relation above can always be expressed in the form

$$\eta_i = g(\pi_i) \ ,$$

with the inverse link function $g = h^{-1}$.

- Logit and probit models are the most widely used binary regression models.

Logit model

- The logit model results from the choice of the logistic response function:

$$\pi = h(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$$

or (equivalently) the logit link function

$$g(\pi) = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \quad .$$

- This yields a linear model for the logarithmic odds (log-odds) $\log(\pi/(1-\pi))$.

- The effects of the covariates affect the odds $\pi/(1-\pi)$ in an exponential-multiplicative form.

Probit model

- In the probit model we use for h the standard normal cumulative distribution function $\Phi(\cdot)$, that is,

$$\pi = \Phi(\eta) = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \ .$$

  or (equivalently) the probit link function

$$g(\pi) = \text{probit}(\pi) = \Phi^{-1}(\pi) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \ .$$

- A (minor) disadvantage is the required numerical evaluation of Φ in the maximum likelihood estimation of $\boldsymbol{\beta}$.

Interpretation of the logit model

Summary:

- The odds $\pi_i/(1-\pi_i) = \mathrm{P}(y_i = 1|\mathbf{x}_i)/\mathrm{P}(y_i = 0|\mathbf{x}_i)$ follow the multiplicative model

$$\frac{\mathrm{P}(y_i = 1|\mathbf{x}_i)}{\mathrm{P}(y_i = 0|\mathbf{x}_i)} = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \cdot \ldots \cdot \exp(x_{ik}\beta_k) \quad .$$

- If, for example, x_{i1} increases by one unit to $x_{i1} + 1$, the following applies to the odds ratio:

$$\frac{\mathrm{P}(y_i = 1|x_{i1}+1, \ldots)}{\mathrm{P}(y_i = 0|x_{i1}+1, \ldots)} \Big/ \frac{\mathrm{P}(y_i = 1|x_{i1}, \ldots)}{\mathrm{P}(y_i = 0|x_{i1}, \ldots)} = \exp(\beta_1) \quad .$$

$\beta_1 > 0$: odds ratio > 1,
$\beta_1 < 0$: odds ratio < 1,
$\beta_1 = 0$: odds ratio $= 1$.

Fitting the logit model

- The parameters of the logistic regression model are estimated by using the method of maximum likelihood.

- Once $\hat{\beta}$ has been obtained, the relationship between the estimated response probability and values $x_1, x_2, \ldots, x_k$ can be expressed as

$$\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \quad ,$$

or equivalently,

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)} \quad .$$

Fitting the logit model II

- The estimated value of the linear systematic component of the model for the *i*th observation is

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \quad .$$

- From this, the fitted probabilities, $\hat{\pi}_i$, can be found from

$$\hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} \quad .$$

Standard errors of parameter estimates

- Following the estimation of the β-parameters in a logistic linear model, information about their precision will generally be needed.

- Such information is conveyed in the standard error of an estimate, $\mathrm{se}(\hat{\beta}_j)$, for $j = 0, \ldots, k$.

- From the standard error of $\hat{\beta}_j$, $100(1-\alpha)\%$ confidence limits for the corresponding true value, β_j, are $\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \times \mathrm{se}(\hat{\beta}_j)$.

- These interval estimates throw light on the likely range of values of the parameter.

Count data

- Count data are frequently observed when the number of events within a fixed time frame or frequencies in a contingency table have to be analyzed.
- Sometimes, a normal approximation can be sufficient.
- In general, however, discrete distributions recognizing the specific properties of count data are most appropriate.
- The Poisson distribution is the simplest and most widely used choice.

Log-linear Poisson model

- The most widely used model for count data connects the rate $\lambda_i = \mathrm{E}(y_i)$ of the Poisson distribution with the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ via

$$\lambda_i = \exp(\eta_i) = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})$$

  or in log-linear form through

$$\log(\lambda_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \quad .$$

- The effect of covariates on the rate λ is, thus, exponentially multiplicative similar to the effect on the odds $\pi/(1-\pi)$ in the logit model.

- The effect on the logarithm of the rate is linear.

Overdispersion

- The assumption of a Poisson distribution for the responses implies
$$\lambda_i = \mathrm{E}(y_i) = \mathrm{Var}(y_i) \ .$$
- For similar reasons as in case with binomial data, a significantly higher empirical variance is frequently observed in applications of Poisson regression.
- This phenomenon is known as overdispersion.
- For this reason, it is often useful to introduce an overdispersion parameter ϕ by assuming
$$\mathrm{Var}(y_i) = \phi\lambda_i \ .$$

Overdispersion II

- The overdispersion parameter ϕ can be estimated as the average Pearson statistic or the average deviance:

$$\hat{\phi}_P = \frac{1}{n-p}\chi^2 \quad \text{or} \quad \hat{\phi}_D = \frac{1}{n-p}D \ .$$

- We then have to multiply the estimated covariance matrix with $\hat{\phi}$, i.e., $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})$.

- This approach to the estimation of overdispersion does not correspond to a true likelihood method, but rather to a quasi-likelihood model.