GENERALIZED LINEAR MODEL

ARTHUR ALLIGNOL

INTRODUCTION

INTRODUCTION

Ordinary linear regression models assume the response variable to be (approximately) normal distributed. However, many experiments require an assessment of the relationship between covariates and a binary response variable, i.e., a variable measured at only two levels, or counts.

Generalised linear models provide a framework for the estimation of regression models with non-normal response variables. The regression relationship between the covariates and the response is modelled by a linear combination of the covariates.

INTRODUCTION

The ordinary multiple regression model is described as $Y \sim \mathcal{N}(\mu, \sigma^2)$ where

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

This makes it clear that this model is suitable for continuous response variables with, conditional on the values of the explanatory variables, a normal distribution with constant variance.

So clearly the model would not be suitable for, e.g., a binary response variable.

LOGISTIC REGRESSION FOR BINARY RESPONSE

For modelling the expected value of the response directly as a linear function of explanatory variables, a suitable transformation is modelled. In this case of a binary response, the most suitable transformation is the **logistic** or **logit** function of $\pi = P(y = 1)$ leading to the model

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

The logit of a probability is simply the log of the odds of the response taking the value one.

LOGISTIC REGRESSION FOR BINARY RESPONSE

The logit function can take any real value, but the associated probability always lies in the required $[0, 1]$ interval.

In a logistic regression model, the parameter β_j associated with explanatory variable x_j is such that $\exp(\beta_j)$ is the odds that the response variable takes the value one when x_j increases by one, conditional on the other explanatory variables remaining constant.

The parameters of the logistic regression model (the vector of regression coefficients β) are estimated by maximum likelihood.

POISSON REGRESSION FOR COUNT DATA

Suppose that $\mathbf{y}$ can be treated as realisations of independent Poisson r.v (e.g., count data).

A simple linear model of the form

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represents an expected count, has to be non-negative.

A solution is to consider $\eta = \log(\mu)$ and the generalized linear model

$$\log \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

GENERALIZED LINEAR MODEL

GLM DEFINITION

The linear, logistic, Poisson, ..., regression models have some common features that we can abstract of form the *generalized linear model*

A GLM is defined by specifying two components.

- The response should be a member of exponential family distribution
- The link function describes how the mean of the response and a linear combination of the predictors are related

EXPONENTIAL FAMILY

In a GLM, the distribution of Y is from the exponential family of distribution which takes the form

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

- θ is called the *canonical parameter* and represents the location
- ϕ is the *dispersion parameter* and represents the scale

We can define various members of the family by specifying the functions a, b, and c

EXPONENTIAL FAMILY — NORMAL

The density is

$$
\begin{aligned}
f(y|\theta,\phi) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\
&= \exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right)\right],
\end{aligned}
$$

such that

- $\theta = \mu$
- $\phi = \sigma^2$
- $a(\phi) = \phi$
- $b(\theta) = \theta^2/2$

EXPONENTIAL FAMILY — POISSON

$$
\begin{aligned}
f(y|\theta,\phi) &= e^{-y}\mu^y/y! \\
&= \exp(y\log\mu - \mu - \log y!)
\end{aligned}
$$

So we can write

- $\theta = \log\mu$
- $\phi = 1$
- $a(\phi) = 1$
- $b(\theta) = \exp(\theta)$

EXPONENTIAL FAMILY — BINOMIAL

$$
\begin{aligned}
f(y|\theta,\phi) &= \binom{n}{y}\mu^y(1-\mu)^{n-y} \\
&= \exp\left(y\log\mu + (n-y)\log(1-\mu) + \log\binom{n}{y}\right) \\
&= \exp\left(y\log\frac{\mu}{1-\mu} + n\log(1-\mu) + \log\binom{n}{y}\right)
\end{aligned}
$$

such that

- $\theta = \log\frac{\mu}{1-\mu}$
- $\phi = 1$
- $a(\phi) = \phi$
- $b(\theta) = -n\log(1-\mu) = n\log(1+\exp\theta)$

EXPONENTIAL FAMILY

The exponential family distributions have mean and variance

$$\begin{aligned} \mathrm{E}Y &= \mu = b'(\theta) \\ \mathrm{var}Y &= b''(\theta)a(\phi) = V(\mu)a(\phi) \end{aligned}$$

Gaussian:

$$\mathrm{E}Y = b'(\theta) = \theta$$

$$\mathrm{var}\ Y = b''(\theta)a(\phi) = \sigma^2$$

Note that the variance does not depend on the mean

Poisson:

$$\mathrm{E}Y = \mathrm{var}Y = \mu$$

Binomial:

$$EY = n\mu$$

$$\mathrm{var}Y = n\mu(1-\mu)$$

LINK FUNCTION

Suppose we express the effect of the predictors on the response through a *linear predictor*

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q = x^T \beta$$

The link function, g, describes how the mean response, $\mathrm{E}Y = \mu$, is linked to the covariates through the linear predictors

$$\eta = g(\mu)$$

When the link function makes the linear predictor η the same as the canonical parameter θ, we say that we have a *canonical link*. If a canonical link is used, $X^T Y$ is *sufficient* for β.

CANONICAL LINKS FOR GLM

Family	Link	Variance Function ($b''(\theta)$)
Normal	$\eta = \mu$	1
Poisson	$\eta = \log \mu$	μ
Binomial	$\eta = \log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	μ^2
Inverse Gaussian	$\eta = \mu^{-2}$	μ^3

One is not required to use the canonical link, but that simplifies estimation

ESTIMATION

ESTIMATION

The parameters β of a GLM can be estimated using maximum likelihood.

Assume $a_i(\phi) = \phi / w_i$, are known *prior weights*. The log-likelihood for the sample $y_1, \ldots, y_n$ is

$$\log L(\theta, \phi; y) = \sum_{i=1}^{n} w_i \left[\frac{y_i \theta_i - b(\theta_i)}{\phi} \right] + c(y_i, \phi)$$

Exact solution only for the Gaussian case. Otherwise use numerical methods.

ESTIMATION

The estimates $\hat{\beta}$ have the usual properties of MLE, in particular,

$$\hat{\beta} \sim \mathcal{N}(\beta, (X^T W X)^{-1} \phi)$$

An estimator for ϕ can be obtained by the method of moments

$$\hat{\phi} = \frac{1}{n-p} \sum \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

INFERENCE

WALD TEST

Test of

$$H_0 : \beta_j = 0$$

using the test statistic

$$z = \frac{\hat{\beta}_j}{\sqrt{\phi}(X^T W X)^{-1}_{jj}}$$

with $z \sim \mathcal{N}(0, 1)$

LIKELIHOOD RATIO TEST AND DEVIANCE

A statistical model describes how we partition the data into a systematic structure and random variation

- The null model represents the situation where the data is represented entirely as random variation
- The saturated model (full model) represents the data as being entirely systematic

Let's consider the difference between the log-likelihood for the full model, $l(y, \theta|y)$, and that of the model under consideration, $l(\hat{y}, \theta|y)$, expressed as a likelihood ratio statistic

$$2(l(y, \phi; y) - l(\hat{y}, \phi; y)).$$

LIKELIHOOD RATIO TEST AND DEVIANCE

For an exponential family distribution, and considering $a_i(\phi) = \phi/w_i$, the statistic simplifies to

$$\sum_i 2w_i \frac{(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))}{\phi},$$

where $\tilde{\theta}$ are the estimates under the full model

LIKELIHOOD RATIO TEST AND DEVIANCE

That can be written as

$$\frac{D(y, \hat{\mu})}{\phi},$$

where

- $D(y, \hat{\mu})$ is the *deviance*
- $D(y, \hat{\mu})/\phi$ is the *scaled deviance*

LIKELIHOOD RATIO TEST AND DEVIANCE

Comparison of two nested models ω_1, with p_1 parameters, and ω_2, with p_2 parameters, such that $\omega_1 \in \omega_2$ and $p_2 > p_1$.

The log of the ratio of maximised likelihoods under the two models can be written as a difference of deviances, since the maximised log-likelihood under the saturated model cancels out. Thus, we have

$$-2 \log \Lambda = \frac{D(\omega_1) - D(\omega_2)}{\phi}$$

The scale parameter is either known, or estimated from the larger model.

The criterion is approximately $\chi^2_{p_2 - p_1}$

GLM DIAGNOSTICS

RESIDUALS

Several kinds of residuals can be defined for GLMs

Response residuals

$$\hat{\varepsilon} = y - \hat{\mu}$$

Pearson residuals

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$

Deviance residuals

Deviance residuals r_D defined such that $\sum r_D^2 = \text{Deviance} = \sum d_i$.
Thus

$$r_D = \text{sign}(y - \hat{\mu})\sqrt{d_i}$$

DIAGNOSTIC PLOTS

- Residual plots are not really useful
- Plots that look for outliers should be looked. Again what to do with outliers depend on the subject matter

MODELS FOR BINOMIAL DATA

BINOMIAL REGRESSION MODEL

Suppose the response variable Y_i for $i = 1, \ldots, n_i$ is binomially distributed, with parameters n_i and π_i such that

$$P(Y_i = y_i) = \binom{n}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

With individual data, $n_i = 1$ for all i.

We need a link function g such that $\eta_i = g(\pi_i)$ and $0 \leq g^{-1}(\eta) \leq 1$ for any η. There is 3 common choices

1. **Logit**: $\eta = \log(\pi/(1 - \pi))$
2. **Probit**: $\eta = \Phi^{-1}(\pi)$, where Φ is the normal cumulative distribution function
3. **Complementary log-log**: $\eta = \log(-\log(1 - \pi))$

Logit is the most natural (canonical link)

LOGISTIC REGRESSION MODEL

$$\text{logit}\pi_i = x^T\beta$$

- The regression coefficients β can be interpreted along the same lines as in linear models.
- β_j represents the change in the logit of the probability associated with a unit change in the j-th predictor holding all other predictors constant.

Taking the exp of the equation above, we find

$$\frac{\pi_i}{1-\pi_i} = \exp(x^T\beta)$$

If we change the j-th predictor by one unit while holding the other variables constant, we multiply the odds by $\exp\beta_j$.

- $\exp\beta_j$ is an *odds-ratio*.

INTERPRETING ODDS

Let's say the probability of success is $.8$, thus

$$\pi = 0.8$$

The probability of failure is $1 - \pi = 0.2$

Odds are defined as the ratio of the probability of success and probability of failure

$$\text{odds}(\text{success}) = 4$$

Thus the odds of success are 4 to 1

Say a variable X ($X = 1$) reduces π to $.4$ thus the odds for $X = 1$ is 0.667.

The odds-ratio is the ratio of the odds. Thus a person with $X = 1$ has an odds 0.167 times smaller than that for $X = 0$.

THE OTHER LINK FUNCTIONS

Any transformation that maps probabilities into the real line could be used, as long as the transformation is one-to-one, continuous and differentiable.

Suppose $F(.)$ is the cumulative distribution function of a r.v. defined on the real line, and write

$$\pi_i = F(\eta_i)$$

for $-\infty < \eta_i < \infty$. Then

$$\eta_i = F^{-1}(\pi_i)$$

could be used as link function

LATENT VARIABLE FORMULATION

Let Y_i be a r.v. representing a binary response coded 0 or 1. Y_i is the *manifest* response.

Suppose that there is an unobservable continuous variable Y_i^* such that Y_i equals 1 iff Y_i^* exceeds a threshold ζ. Y_i^* is the *latent* response

Possible interpretation

Y_i a binary choice such as purchasing or renting a home, while Y_i^* is the difference in the utilities of purchasing and renting

Thus

$$\pi_i = P(Y_i = 1) = P(Y_i^* > \zeta)$$

LATENT VARIABLE FORMULATION

Write

$$Y_i^* = x_i^T \beta + U_i$$

where U_i is the error term with distribution with cdf $F(u)$. Under this model,

$$\begin{aligned} \pi_i &= P(Y_i = 1) \\ &= P(U_i > -\eta_i) \\ &= 1 - F(-\eta_i) \end{aligned}$$

If the distribution of the error terms is symmetric around 0, $\pi_i = F(\eta_i)$. This defines a GLM with Bernoulli response and link

$$\eta_i = F^{-1}(\pi_i)$$

PROBIT MODEL

Assume $U \sim \mathcal{N}(0, 1)$. This leads to

$$\pi_i = \Phi(\eta_i)$$

where Φ is the standard normal c.d.f. The inverse transformation, which gives the linear predictor as a function of the probabilities

$$\eta_i = \Phi^{-1}(\pi_i)$$

is called the *probit*.

The β's are interpreted in units of standard deviation of the latent variable

COMPLEMENTARY LOG-LOG TRANSFORMATION

$$\eta_i = \log(1 - \log(1 - \pi_i))$$

which is the inverse of the cdf of the log-Weibull distribution, with cdf

$$F(\eta_i) = 1 - e^{-e^{\eta_i}}$$

The complementary log-log transformation has a direct interpretation in terms of hazard ratios, and thus has practical applications in terms of hazard models (see Survival models)

ILLUSTRATION

Challenger disaster example

In January 1986, the space shuttle Challenger exploded shortly after lunch.

An investigation was launched shortly after and attention focused on rubber O-rings in the rocket boosters. At lower temperatures, rubber becomes more brittle and is a less effective sealant.

At the time of launch, the temperature was 31°F (-.5°C).

In the 23 previous missions, some evidence of damage was recorded for some O-rings.

Could the failure have been predicted?

ILLUSTRATION

plasma: Erythrocyte sedimentation rate (ESR)

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. If the ESR increases when the level of certain proteins in the blood plasma rise in association with conditions such as rheumatic diseases, chronic infections and malignant diseases, its determination might be useful in screening blood samples taken from people suspected of suffering from one of the conditions mentioned. The absolute value of the ESR is not of great importance, rather it is whether it is less than 20mm/hr since lower values indicate a 'healthy' individual.

The question of interest is whether there is any association between the probability of an ESR reading greater than 20mm/hr and the levels of the two plasma proteins. If there is not then the determination of ESR would not be useful for diagnostic purposes.

COUNT DATA

COUNT REGRESSION

When the response is a count (a positive integer), we can use a count regression model to explain this response in terms of the given predictors.

We consider two distributions for counts:

- The Poisson
- The negative binomial

POISSON REGRESSION

If Y is Poisson with mean μ then

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

with $\mu > 0$.

Then we can use the log link function such that

$$\log \mu = \eta = x^{T}\beta$$

In this model, $\exp \beta_j$ represents a multiplicative effect of the j-th predictor on the mean. Increasing x_j by one unit multiplies the mean by a factor of $\exp(\beta_j)$.

RATE MODELS

Let

- Y be a count (infections, violent acts)
- t index of the time / space (days, neighbourhood)

The sample rate of occurrence is Y/t The expected value of the rate is

$$\mathrm{E}(Y/t) = \frac{1}{t}\mathrm{E}(Y) = \frac{\mu}{t}$$

The Poisson regression model for the expected rate is

$$\log(\mu/t) = x^T\beta$$
$$\log(\mu) = \log(t) + x^T\beta$$

$\log(t)$ is known as *offset* and may be different for each i

OVERDISPERSION

One key feature of the Poisson distribution is that

$$\mathrm{var} Y = \mathrm{E} Y = \mu.$$

However, we often find data for which there is overdispersion, i.e., the variance is larger than the mean.

In the context of count data, we could assume

$$\mathrm{var} Y = \phi \mathrm{E} Y = \phi \mu.$$

One would just need to correct the standard errors of the Poisson model using an estimate of ϕ. Only meaningful is the model is correct

NEGATIVE BINOMIAL MODEL

An alternative approach is the negative binomial regression model.

Suppose that the conditional distribution of Y given an unobserved variable is Poisson with mean and variance $\delta\mu$.

In this model, the data would be Poisson if we could observe δ, but we don't.

- Make some assumption about the distribution of δ
- Integrate it out of the likelihood

NEGATIVE BINOMIAL MODEL

It is convenient to assume that δ is gamma distributed with parameters α and β. The distribution has mean α/β and variance α/β^2.

Consider further $\alpha = \beta = 1/\sigma^2$.

The unconditional distribution of the outcome is then the negative binomial distribution.

The negative binomial distribution with $\alpha = \beta = 1/\sigma^2$ has mean

$$\mathrm{E}Y = \mu$$

and variance

$$\mathrm{var}Y = \mu(1 + \sigma^2\mu)$$

If $\sigma^2 = 0$ there is no overdispersion

ILLUSTRATION

solder data

ATT ran an experiment varying 5 factors relevant to a wave-solderinr components on printed circuit boards.

The response variable `skips` is a count of how many solder skips appeared to a visual inspection.

tb_real

Dataset holding information on outbreaks of tuberculosis in dairy and beef cattle, cervids and bison in Canada between 1985 and 1994.

`reactors`
: Numbers of reactors (A reactor animal is one that has shown a significant response to the tuberculin skin test) in a group

`par`
: Animal days at risk in the group

MULTINOMIAL DATA

MULTINOMIAL DATA

Consider a random variable Y_i that may take one of several discrete values in $1, 2, \ldots, J$. Let

$$\pi_{ij} = P(Y_i = j),$$

and assume $\sum_j \pi ij = 1$. Thus we have $J - 1$ parameters

MULTINOMIAL LOGIT MODEL

The idea is to pick a reference category (baseline), calculate log-odds for all other categories relative to the baseline, and then let the log-odds be a linear function of the predictors.

Pick J as baseline category (as in SAS) and compute π_{i1}/π_{iJ}

The multinomial logit model assumes that the log-odds of each response follow a linear model

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x_i^T \beta_j$$

The $J-1$ multinomial logit equations contrast each category $1, 2, \ldots J-1$ with category J.

ILLUSTRATION

nes96 data set

The data contain a subset of the 1996 American National Election Study. We consider only the age, income and education of the respondants. The response variable is the party identification of the individual: Democrat, Independent or Republican.

MODELS FOR ORDINAL DATA

Let $\pi_{ij} = P(Y_i = j)$ and let γ_{ij} be the corresponding cumulative probability

$$\gamma_{ij} = P(Y_i \leq j).$$

Thus $\gamma_{ij} = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}$.

We consider models of the form

$$g(\gamma_{ij}) = \theta_j + x_i^T \beta,$$

where θ_j is a constant representing baseline value of the transformed cumulative probability for category j, and β represents the effect of the covariates on the transformed cumulative probability

PROPORTIONAL ODDS MODEL

Extension of the logistic model that applies the logit transformation to the cumulative response probabilities

$$\text{logit}(\gamma_{ij}) = \log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j + x_i^T \beta$$

Exponentiating, we find that

$$\frac{\gamma_{ij}}{1-\gamma_{ij}} = \exp(\theta_j)\exp(x_i\beta)$$

Thus $\exp(\theta_j)$ may be interpreted as a baseline odds of a response in category j or below.

The effect of the covariate x is to raise or lower the odds of a response in category j or below by the factor $\exp(\beta)$.

Note that if a certain combination of covariate values doubles the odds of being in category 1, it also doubles the odds of being in category 2 or below, or in category 3 or below

ORDERED PROBIT MODEL

The ordered probit model models the probit of the cumulative probabilities as a linear function of the covariates

$$\Phi^{-1}(\gamma_{ij}) = \theta_j + \boldsymbol{x}_i^T \beta.$$

ILLUSTRATION

CHFLS: Chinese Health and Family Life Survey

The Chinese Health and Family Life Survey sampled 60 villages and urban neighbourhoods chosen in such a way as to represent the full geographical and socioeconomic range of contemporary China excluding Hong Kong and Tibet.

Eighty-three individuals were chosen at random for each location from official registers of adults aged between 20 and 64 years to target a sample of 5000 individuals in total.

Here, we restrict our attention to women with current male partners and the following variables:

ILLUSTRATION

R_edu
: level of education of the responding woman,

R_income
: monthly income (in yuan) of the responding woman,

R_health
: health status of the responding woman in the last year,

R_happy
: how happy was the responding woman in the last year,

A_edu
: level of education of the woman's partner,

A_income
: monthly income (in yuan) of the woman's partner.

QUASI-LIKELIHOOD

QUASI-LIKELIHOOD

Suppose that we are able to specify the link and variance functions of the model for some data.

But we have no strong idea about the distributional form of the response.

We also know that the important part of the model specification is the link and variance; the outcome is less sensitive to the distribution of the response.

However, we need some distributional assumption to compute a likelihood, deviance for making inference.

→ We need a substitute of the likelihood that can be computed without distributional assumption

QUASI-LIKELIHOOD

Let Y_i have mean μ_i and variance $\phi V(\mu_i)$. Assume the Y_i to be independent.

Define a score U_i

$$U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}.$$

One can show that $\mathrm{E}U_i = 0$ and $\mathbf{var}U_i = 1/\phi V(\mu_i)$. Actually U_i and the derivative of the log-likelihood share the same properties, so we can use U instead.

Define

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$

and $Q = \sum_i Q_i$.

Q behaves like a log-likelihood and shares the same asymptotic properties.