STAT3015/4030/7030:
Generalised Linear Modelling
GLMS - Theory

Semester 2 2017

Originally prepared by Bronwyn Loong

References

Ch 8 - Faraway, Extending Linear models with R
(chapter 6 in the first edition of Faraway)
Ch 6 - Gelman and Hill

Components of a GLM

Linear regression and logistic regression are special cases of generalised linear models. A GLM is composed of:

1. A set of (independent) response data $y_1, ..., y_n$, and $E[y_i] = \mu_i$
2. A set of (vector) predictors $x_1, ..., x_p$ which form the $n \times p$ matrix X.
3. A **link** function g, relating the mean of the responses μ_i to the covariates $g(\mu_i) = x_i^T \beta$, given coefficients β. The quantity $\eta_i = x_i^T \beta$ is sometimes referred to as the **linear predictor**.
4. A data distribution to describe $Pr(y_i|\eta_i)$, that is the error component of the model.

In a GLM

- The response does not need to be normally distributed
- The relationship between the response and the predictors does not need to be linear.

Components of a GLM

Describe the GLM components (as listed above) for:

1. Linear regression
2. Logistic regression

Components of a GLM

Linear regression and logistic regression are special cases of generalised linear models. A GLM is composed of:

1. A set of (independent) response data $y_1, ..., y_n$, and $E[y_i] = \mu_i$
2. A set of (vector) predictors $x_1, ..., x_p$ which form the $n \times p$ matrix X.
3. A **link** function g, relating the mean of the responses μ_i to the covariates $g(\mu_i) = x_i^T \beta$, given coefficients β. The quantity $\eta_i = x_i^T \beta$ is sometimes referred to as the **linear predictor**.
4. **A data distribution to describe $Pr(y_i|\eta_i)$, that is the error component of the model**.

Exponential families

Generally we choose the data distribution to be a member of the **exponential family** (EF) of distributions. The EF class of distributions has many nice properties. A member of the exponential family has general form:

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

for some specified functions b and c. $b(\theta_i)$ is the cumulant generating function.
θ: natural parameter, represents location
ϕ: dispersion parameter, represents scale

Varying the forms of the functions b and c, we can define various members of the exponential family.

Exponential families

If you can write the density of y in the form

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

then the density is a member of the exponential family and is in its natural parameterization.

Exponential Families

1. Normal or Gaussian $(\mu, \sigma^2) \rightarrow E[Y] = \mu$, $Var[Y] = \sigma^2$

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

So $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \theta^2/2$ and
$c(y, \phi) = -(y^2/\phi + \log(2\pi\phi))/2$

Exponential Families

2. **Poisson**$(\lambda) \rightarrow E[Y] = \mu = \lambda,\ Var[Y] = \lambda$

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

So $\theta = \log(\mu)$, $\phi = 1$, $b(\theta) = \exp(\theta)$ and $c(y, \phi) = -\log y!$

Exponential Families

3. **Binomial**$(n, p) \rightarrow E[Y] = \mu = np,\ Var[Y] = np(1-p)$

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

So $\theta = \log \frac{p}{1-p} = \log \frac{\mu}{n-\mu}$, $\phi = 1$,

$b(\theta) = -n\log(1-p) = n\log(1+\exp\theta)$ and $c(y,\phi) = \log\binom{n}{y}$

Exponential Families

The exponential family has the following properties:

$$E[Y] = \mu = b'(\theta)$$

$$\mathrm{Var}[Y] = b''(\theta)\phi = V(\mu)\phi$$

The mean is a function of the location parameter θ only. The variance is a function of both location and scale parameters. $V(\mu)$ is called the variance function and describes how the variance relates to the mean.

Exercise: Use the above results for the mean and variance of an EF to verify the mean and variance for the Normal, Poisson and Binomial distributions.

GLM Components

1. A set of (independent) response data $y_1, ..., y_n$, and $E[y_i] = \mu_i$
2. A set of (vector) predictors $x_1, ..., x_p$ which form the $n \times p$ matrix X.
3. **A link function g, relating the mean of the responses μ_i to the covariates $g(\mu_i) = x_i^T \beta$, given coefficients β. The quantity $\eta_i = x_i^T \beta$ is sometimes referred to as the linear predictor**.
4. A data distribution to describe $Pr(y_i|\eta_i)$, that is the error component of the model

Link function

We assume:

$$y_1, \ldots, y_n \overset{\text{indep.}}{\sim} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

And the log-likelihood is

$$l(\theta; y) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i, \phi)$$

In general, we do not wish to estimate as many parameters as there are data $\theta_1, ..., \theta_n$.

We would like to model how the mean response $E[Y_i]$ is related to a set of explanatory variables $x_i = (x_{i1}, ..., x_{ik})$ (for $k < n$). Specifically let $\beta = (\beta_0, ..., \beta_k)$ be a set of associated parameters, and we want to define a relationship between $E[Y_i]$ and the linear function $x_i^T \beta$.

Link function

Recall: the link function g describes how the mean response $\mu_i = E[Y_i]$ is linked to the covariates through the linear predictor

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} = x_i^T \beta = \eta_i \quad i = 1, .., n$$

Various choices of link function may be examined, provided the link function is monotone continuous and differentiable.

However, there are some convenient and common choices for each exponential family distribution.

Link function

The Gaussian linear model is described by the equation

$$y_i = x_i^T \beta + \epsilon_i$$

What is the link function?

Link function

Consider data generated by the Poisson distribution. That is $Y \sim Pois(\lambda)$ (and $E[Y] = \mu = \lambda$).

Is the identity link appropriate? That is, the model $\mu_i = x_i^T \beta$?

What may be a better choice?

Link function

Consider data generated by the Poisson distribution. That is $Y \sim Pois(\lambda)$ (and $E[Y] = \mu = \lambda$).

Is the identity link appropriate? That is, the model $\mu_i = x_i^T \beta$?

What may be a better choice? The standard choice is $\mu_i = e^{\eta_i}$ so that $\eta_i = \log \mu_i$, which ensures $\mu_i > 0$ (where $\eta_i = x_i^T \beta$)

Link function

The canonical link is g such that $\eta = g(\mu) = \theta$, the canonical parameter of the EF distribution. This means that $g(b'(\theta)) = \theta$.

Choosing the canonical link is mathematically convenient for parameter estimation.

(note: if the canonical link is chosen, then $X^T Y$ is sufficient for estimation of β)

Distribution	Canonical link function		$E[y_i]$
Normal(μ_i, σ^2)	identity	$\theta_i = \mu_i = \eta_i$	$\theta_i = \mu_i$
Poisson (μ_i)	log	$\theta_i = log(\mu_i) = \eta_i$	$exp(\theta_i) = \mu_i$
Binomial(n, p_i)	logit	$\theta_i = \log(\frac{p_i}{1-p_i}) = \log(\frac{\mu_i}{n-\mu_i}) = \eta_i$	$n\left(\frac{exp(\theta_i)}{1+\exp(\theta_i}\right) = np_i = \mu_i$
Gamma(α, β_i)	reciprocal	$\theta_i = -\frac{1}{\mu_i} = \eta_i$	$-\frac{1}{\theta_i} = \frac{\alpha}{\beta_i} = \mu_i$

Table : Canonical links for GLMs

Note: the canonical link is not necessarily the most appropriate choice for a given set of data.

Link function

Family	$Var(\mu_i)$	ϕ
Normal(μ_i, σ^2)	1	σ^2
Poisson(μ_i)	μ_i	1
Binomial(n, π_i)	$\frac{\mu_i(n-\mu_i)}{n}$	1
Gamma(α, β_i)	μ_i^2	$1/\alpha$

Table : Canonical links for GLMs

$$Var(Y_i) = Var(\mu_i)\phi$$

Link function

Exercise: show that the canonical link for the Gamma distribution is $\theta = -1/\mu$.

Note, the probability density function for $Y \sim Gamma(\alpha, \beta)$ is $f(y) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha} y^{\alpha-1} e^{-\beta y}$ for $y > 0$, and $E[Y] = \mu = \alpha/\beta$ and $Var(Y) = \alpha/\beta^2$.

Reparametize the model by setting $\beta = \alpha/\mu$ so that

$$f(y) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^{\alpha} y^{\alpha-1} e^{-\frac{\alpha}{\mu} y}$$

Write $f(y)$ in exponential form to show the canonical link is $\theta = -1/\mu$. Also show the dispersion parameter is $\frac{1}{\alpha}$. Find $b(\theta)$ and verify that $E[Y] = b'(\theta)$ and $Var[Y] = b''(\theta)\phi$ for the Gamma distribution.

Maximum Likelihood Estimation

(Recall)

The likelihood function for the data is simply another name for the probability density or probability mass function of the data, regarded now as a function of the parameters instead of as a function of the data values.

Maximum likelihood estimation: estimate the parameters by those values which make the likelihood function as large as possible for our particular set of observed data. These estimates are the ones which have the maximum likelihood of having produced the observed data.

Maximum Likelihood Estimation

The log-likelihood for a single observation in a GLM:

$$\log L(\theta_i, \phi | y_i) = \left\{ \frac{y_i \theta - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

and $\theta_i = g(\mu_i) = \eta_i = x_i^T \beta$

Then, for n independent observations

$$\log L(\theta, \phi | Y) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} - (1)$$

A closed-form solution to (1) is available only for the Normal GLM. Numerical optimization methods are required instead.

The R statistical software package uses the optimization method *Iteratively Reweighted Least Squares*

Weighted least squares (WLS)

Under classical linear regression $Var(\epsilon) = \sigma^2\Sigma$, where $\Sigma = \mathrm{I}$ (the identity matrix).

What if the errors are uncorrelated (non-diagonal elements of $\Sigma = 0$), but we have non-constant variance (diagonal elements of $\Sigma \neq 1$)?

Weighted least squares can be used in this situation.

What data points do we want to apply high weight to? low weight to? $\rightarrow$

Weighted least squares (WLS)

Under classical linear regression $Var(\epsilon) = \sigma^2\Sigma$, where $\Sigma = \mathrm{I}$ (the identity matrix).

What if the errors are uncorrelated (non-diagonal elements of $\Sigma = 0$), but we have non-constant variance (diagonal elements of $\Sigma \neq 1$)?

Weighted least squares can be used in this situation.

What data points do we want to apply high weight to? low weight to? $\rightarrow$ In linear regression, our goal is to minimise the prediction error. This is an optimization problem, and we want to apply larger weight to those points with low variability; high variability points are assigned a low weight.

Weighted least squares (WLS)

Regress $\sqrt{w_i}y_i$ on $\sqrt{w_i}x_i$. The weighted sum of squares is

$$WSS = \sum_{i=1}^{n} w_i^2(y_i - X_i\hat{\beta})^2$$

And

$$\hat{\beta}_{WLS} = (X^T WX)^{-1}X^T WY$$

Weighted least squares is equivalent to the maximum likelihood estimation of β in the normal regression model

$$y_i \sim N(X_i\beta, \sigma^2/w_i)$$

Weighted least squares (WLS)

Examples of weighted linear regression

- Errors proportional to a predictor: $var(\epsilon) \propto x_i$ suggests $w_i = 1/x_i$
- When Y_i are averages of n_i observations, $var(Y_i) = var(\epsilon_i) = \sigma^2/n_i$, so set $w_i = n_i$. But take care that the variance in the response is really proportional to group size (and not to some other factor).

When weights are used, use $\sqrt{w_i}\hat{\epsilon}_i$ for diagnostics.

Iteratively Reweighted Least Squares (IRWLS)

If we knew μ, then we could estimate β using a linear regression of $g(\mu)$ on X. But we don't know the μ's.

Perhaps we can try to regress $g(y)$ on X. A one-step Taylor series expansion of $g(y)$:

$$\begin{aligned} g(y) &\approx g(\mu) + (y - \mu)g'(\mu) \\ &= \eta + (y - \mu)\frac{d\eta}{d\mu} \\ &\equiv z \end{aligned}$$

z is the linearised response or adjusted dependent variable. Using the *Delta Method* we can obtain the variance estimate :

$$\widehat{Var[g(y)]} \approx \left(\frac{d\eta}{d\mu}\right)^2 V(Y)$$

Hence, $\widehat{Var[g(y)]}$'s is not constant. How to deal with this?

Iteratively Reweighted Least Squares (IRWLS)

If we knew μ, then we could estimate β using a linear regression of $g(\mu)$ on X. But we don't know the μ's.

Perhaps we can try to regress $g(y)$ on X. A one-step Taylor series expansion of $g(y)$:

$$\begin{aligned} g(y) &\approx g(\mu) + (y - \mu)g'(\mu) \\ &= \eta + (y - \mu)\frac{d\eta}{d\mu} \\ &\equiv z \end{aligned}$$

z is the linearised response or adjusted dependent variable. Using the *Delta Method* we can obtain the variance estimate :

$$\widehat{Var[g(y)]} \approx \left(\frac{d\eta}{d\mu}\right)^2 V(Y)$$

Hence, $\widehat{Var[g(y)]}$'s is not constant. How to deal with this?

Iteratively Reweighted Least Squares (IRWLS)

IRWLS procedure:

1. Set initial estimates $\hat{\eta}_0$ and $\hat{\mu}_0$
2. Form the "adjusted dependent variable" $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{d\eta}{d\mu}\big|_{\hat{\eta}_0}$
3. Form the weights $w_0^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 \big|_{\hat{\eta}_0} V(\hat{y}_0)$
4. Regress $\sqrt{w_0}Z$ on $\sqrt{w_0}X$, Re-estimate β using weighted least squares to get $\hat{\eta}_1$
5. Iterate steps 2-3-4 until convergence

That is, at iteration t, $\beta_t = [(X^T W X)^{-1} X^T W Z]_{t-1}$, where X is the design matrix, Z is the vector of linearised response values, and W is the diagonal weight matrix with $W_{ii} = \frac{1}{V(y_i)g'(\mu_i)^2}$

Iteratively Reweighted Least Squares (IRWLS)

Note: the fitting procedure uses only $\eta = g(\mu)$ and $V(Y)$, but requires no further knowledge of the distribution of Y.
Also

$$\widehat{var}(\hat{\beta}) = \phi(X^T W X)^{-1}$$

Based on the values for $\hat{\beta}$ at the final interation
(similar to weighted least squares for regression but weights are now a function of the mean response for a GLM).

Precision of parameter estimates and confidence intervals

For any linear combination of the maximum likelihood estimators:

$$E[c^T \hat{\beta}] = c^T \beta$$

$$Var[c^T \hat{\beta}] = \phi c^T (X^T W X)^{-1} c$$

Hence, we can construct confidence interval estimates and perform hypotheses tests on any linear combination of the β's.

What about for non-linear combinations of the parameters?

Precision of parameter estimates and confidence intervals

Suppose we want a confidence interval for $g^{-1}(x_0^T \beta)$??

- Construct a confidence interval for $x_0^T \beta$, say (l, u)
- Compute the interval $(g^{-1}(l), g^{-1}(u))$

Other methods exist based on likelihood theory (outside scope of this course).

Important: make sure your confidence interval contains values inside the allowable range for the quantity in question.

Precision of parameter estimates and confidence intervals

The potency of an anaesthetic agent is measured in terms of the minimum concentration at which at least 50% of patients exhibit no response or stimulation. Thirty patients are administered a particular anaesthetic at various predetermined concentrations for 15 minutes before a stimulus was applied. The response variable was simply an indicator as to whether the patient responded or not.

Fit a glm model to predict the probability of response given the level of anaesthetic. Provide a 95% confidence interval estimate of the response probability at an anaesthetic concentration of 1.5 atmospheres.

(see R code - example from page 42 of "brick")