GENERALIZED LINEAR MODELS: A SUMMARY

SEMINAR IN STATISTICS, SPRING TERM 2010, MYRIAM RIEK

1. Introduction and Definition

1.1. **Generalization of the Linear Model.** The concept of generalized linear models (GLMs) unifies different approaches to explaining variation in data in terms of a linear combination of covariates. Examples of approaches to non-normal data that fall into this category of models are the logistic regression and the log-Poisson regression.

In what follows we are going to consider a sample of n independent observations for some response variable Y.

Recall the *linear model* which can be formulated as follows for the i-th observation:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

with

$$\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \beta = \eta_i = g^{-1}(\eta_i) \qquad \eta_i = g(\mu_i) = \mu_i$$

where η_i, $g^{-1}(\eta_i)$ and $g(\mu_i)$ are typically called linear predictor, inverse link function and link function, respectively.

The generalization of the linear model is twofold by relaxing the following two crucial assumptions:

(1) $\mathrm{E}[Y_i] = \mu_i$ can be a function of the linear predictor η_i other than the identity
(2) Distribution of Y_i can be any distribution from the exponential family of distributions

The exponential family of distributions is very varied and encompasses distributions such as the normal, Bernoulli, binomial, Poisson, exponential and gamma distributions for instance.

Logistic regression: $Y_i \sim \mathrm{Bern}(\mu_i)$ or $Y_i = \frac{Z_i}{m_i} \sim \mathcal{B}(m_i, \mu_i)$ (where Z_i is the sum of m_i *i.i.d.* copies of a $\mathrm{Bern}(\mu_i)$-distributed random variable)

$$\mathrm{E}[Y_i] = \mu_i = g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \qquad \eta_i = g(\mu_i) = \log \frac{\mu_i}{1 - \mu_i}$$

where the link function $g(\mu_i)$ is the so called logit function and the inverse link function $g^{-1}(\eta_i)$ is the logistic function, hence the name logistic regression.

Log-Poisson regression: $Y_i \sim \mathcal{P}(\mu_i)$

$$\mathrm{E}[Y_i] = \mu_i = g^{-1}(\eta_i) = e^{\eta_i} \qquad \eta_i = g(\mu_i) = \log \mu_i$$

and the name stems from the fact that the link function used is the log.

1.2. **Canonical Representation of the Probability Density (Mass) Function (pdf/pmf) and the Link Function.** The pdf (or pmf) of Y_i following a distribution from the exponential family of distributions can be written in canonical form as,

$$f(y_i|\theta_i, \tau^2, \omega_i) = \exp\left(\frac{\theta_i y_i - d(\theta_i)}{\tau^2}\omega_i\right)h(y_i, \tau^2, \omega_i)$$

where

θ_i	canonical parameter of the distribution
τ^2	dispersion parameter
ω_i	some number
	usually 1, in case of binomially distributed data equal to m_i
$d(\theta_i)$	function characterizing the type of distribution
$h(y_i, \tau, \omega_i)$	normalizing function

and everything known except θ_i.

It holds that

$\mathrm{E}[Y_i] = \mu_i = \frac{\partial d(\theta_i)}{\partial \theta_i} = d'(\theta_i)$ and $\mathrm{var}(Y_i) = \frac{\partial^2 d(\theta_i)}{\partial^2 \theta_i}\frac{\tau^2}{\omega_i} = d''(\theta_i)\frac{\tau^2}{\omega_i} = \nu(\mu_i)\frac{\tau^2}{\omega_i}$

For normal data:
$\theta_i = \mu_i \quad d(\theta_i) = \frac{\theta_i^2}{2} \quad \tau^2 = \sigma^2 \quad \omega_i = 1 \quad \nu(\mu_i) = 1$

For Bernoulli data:
$\theta_i = \log\frac{\mu_i}{1-\mu_i} \quad d(\theta_i) = \log(1 + e^{\theta_i}) \quad \tau^2 = 1 \quad \omega_i = 1 \quad \nu(\mu_i) = \mu_i(1 - \mu_i)$

For Poisson data:
$\theta_i = \log\mu_i \quad d(\theta_i) = e^{\theta_i} \quad \tau^2 = 1 \quad \omega_i = 1 \quad \nu(\mu_i) = \mu_i$

In principle any link function $g(\mu_i)$ could be used but the most useful and mathematically convenient link functions are the canonical link functions. Canonical because these link functions relate μ_i to the canonical parameter θ_i, i.e., $\theta_i = \eta_i = \mathbf{x}_i^T\beta$. For the normal, Bernoulli/binomial and Poisson distributions the canonical link functions are the identity, the logit and the log function, respectively. In what follows we are only considering canonical link functions.

2. Parameter Estimation

2.1. **Maximum Likelihood Estimates.** Assuming canonical link is used, we solve the following set of equations to obtain the maximum likelihood estimate (MLE) of β, $\hat{\beta}$,

$$\frac{\partial \ell(\mathbf{y}, \beta)}{\partial \beta}\Big|_{\beta=\hat{\beta}} = \sum_{i=1}^{n} \frac{\partial \log f(y_i, \eta_i)}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta}\Big|_{\beta=\hat{\beta}} = \sum_{i=1}^{n} \frac{(y_i - d'(\eta_i))}{\tau^2}\omega_i x_i\Big|_{\beta=\hat{\beta}} = 0$$

using a modification of the iterative Newton-Raphson algorithm called *Fisher's method of scoring.*

Recall Newton-Raphson at the $(k+1)$-th iteration:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \mathbf{H}^{-1}(\hat{\beta}^{(k)}) \frac{\partial \ell(\mathbf{y}, \beta)}{\partial \beta}\Big|_{\beta=\hat{\beta}^{(k)}}$$

with (for canonical link function)

$$\mathbf{H}(\hat{\beta}^{(k)}) = \frac{\partial^2 \ell(\mathbf{y}, \beta)}{\partial \beta \partial \beta^T}\Big|_{\beta=\hat{\beta}^{(k)}} = -\sum_{i=1}^{n} \frac{d''(\eta_i)}{\tau^2} \omega_i x_i x_i^T \Big|_{\beta=\hat{\beta}^{(k)}}$$

Fisher's method of scoring replaces the Hessian $\mathbf{H}(\hat{\beta}^{(k)})$ in the Newton-Raphson algorithm with its expectation, i.e., $\mathrm{E}[\frac{\partial^2 \ell(\mathbf{y},\beta)}{\partial \beta \partial \beta^T}|_{\beta=\hat{\beta}^{(k)}}]$, which is the negative Fisher information $\mathbf{I}(\hat{\beta}^{(k)})$. If the canonical link is used, Fisher's method of scoring and Newton-Raphson are identical since $\frac{\partial^2 \ell(\mathbf{y},\beta)}{\partial \beta \partial \beta^T}$ do not involve any y_i-values and, thus, $\mathbf{I}(\hat{\beta}^{(k)}) = -\mathbf{H}(\hat{\beta}^{(k)})$.

2.2. **Iteratively Weighted Least Squares (IWLS) Estimates.** The iterative ML approach using Fisher scoring is equivalent to an IWLS approach.

3. Parameter Inference

3.1. **Distribution of MLE.** MLEs are asymptotically unbiased and normally distributed, i.e., if n is sufficiently large, it holds approximately that

$$\hat{\beta} \sim \mathcal{N}_p(\beta, I^{-1}(\hat{\beta}_{MLE}))$$

3.2. **Tests and Confidence Intervals.** There are mainly two ways in which hypotheses on β (or a linear transformation or subset of it) can be tested and confidence intervals (CIs) can be derived: Wald tests and CIs or likelihood based tests and CIs.

3.2.1. *Wald Tests and CIs.* Assuming that normality of the MLE holds, we can derive the Wald test statistic W for any linear transformation $\mathbf{B}\beta$ of β with $\mathbf{B}$ being a (q x p) matrix,

$$W = (\mathbf{B}\hat{\beta} - \mathbf{B}\beta)^T V^{-1} (\mathbf{B}\hat{\beta} - \mathbf{B}\beta) \sim \chi^2_q \qquad V = \mathrm{cov}(\mathbf{B}\hat{\beta}) = \mathbf{B} I^{-1}(\hat{\beta}) \mathbf{B}^T$$

$(1-\alpha)100\%$-CIs for $\mathbf{B}\beta$ based on the Wald test statistic are of the form

$$\{b : W = (\mathbf{B}\hat{\beta} - b)^T V^{-1} (\mathbf{B}\hat{\beta} - b) \leq \chi^2_{q,1-\alpha}\}$$

Note that these CIs are symmetric around $\mathbf{B}\hat{\beta}$.

3.2.2. *Likelihood Based Tests and CIs.* A likelihood ratio test (LRT) compares the maximum likelihood under some null hypothesis, $\mathcal{L}_{H_0}$, with the maximum likelihood under an alternative hypothesis, $\mathcal{L}_{H_A \supset H_0}$, that encompasses the null hypothesis, i.e., is 'larger'. The LRT-statistic follows asymptotically a χ^2-distribution, i.e.,

$$LRT - \text{statistic} = -2(\ell_{H_0} - \ell_{H_A \supset H_0}) \sim^{app} \chi^2_q \text{ (if } H_0 \text{ is true)}$$

where q is the difference in df.

A $(1-\alpha)100\%$-CI for e.g. β_j based on the LRT statistic is

$$\{b : LRT - \text{statistic} = -2(\ell_{\hat{\beta}|\beta_j=b} - \ell_{\hat{\beta}}) \leq \chi^2_{1,1-\alpha}\}$$

Note that these CIs must not be symmetric around MLE.

The approximation for the LRT is better than the approximation for the Wald test statistic. On the other hand, LRT comes with the major drawback of having to fit the model repeatedly under H_0 to construct CIs.

4. Model Fit and Diagnostics

4.1. **Assessing Model Fit.** A way of examining the goodness of fit of a GLM is by looking at its deviance. The deviance D is defined as

$$-2(\ell_M - \ell_F)(\text{up to the factor } \tau^2)$$

where ℓ_M is the log likelihood of the fitted model and ℓ_F is the maximized log likelihood under the full or saturated model, i.e., when each observation is fitted with a separate parameter.

The distribution of D can be approximated by a χ^2_{n-p} distribution if the chosen model is correct. The approximation can, however, be poor or not hold at all as in the case of binary data ($Y_i \in \{0, 1\}$).

Note that in case of normal data, the deviance is equal to $\Sigma_{i=1}^n (y_i - \hat{\mu}_i)^2$, i.e., it is equal to the sum of squared residuals (or residual SS) and exactly $\tau^2\chi^2_{n-p}$-distributed.

4.2. **Diagnostics - Assessing Adequacy of Fitted Model.** Similar to linear models, GLMs assume that the specified model is correct, i.e., that the Y_i values are independent and follow the specified type of distribution with $E[Y_i] = \mu_i = g^{-1}(\eta_i)$ and $\eta_i = \mathbf{x}_i^T\beta$. Fulfillment of these assumptions should therefore be investigated.

Measures of agreement between the individual observed response y_i and its corresponding fitted value $\hat{\mu}_i$ are known as residuals and form the basis of many diagnostic techniques. In GLMs, there are several different definitions of residuals with different properties. Some are

Raw or response residuals $R_i^{(R)}$	$= y_i - \hat{\mu}_i$
Working residuals $R_i^{(W)}$	$= R_i^{(R)}/\nu(\hat{\mu}_i)$
Pearson residuals $R_i^{(P)}$	$= R_i^{(R)}/\sqrt{\frac{\nu(\hat{\mu}_i)}{\omega_i}}$
Standardized Pearson residuals $R_i^{(SP)}$	$= R_i^{(R)}/\sqrt{\frac{\hat{\text{var}}(R_i^{(R)})}{\tau^2}}$
Deviance residual $R_i^{(D)}$	$= \text{sgn}(R_i^{(R)})\sqrt{d_i}$ (with $\Sigma_{i=1}^n d_i = D$)
Standardized deviance residual $R_i^{(SD)}$	$=$ standardized $R_i^{(D)}$

The basic diagnostics approach is as for linear models. E.g., use of Tukey-Anscombe plots of residuals versus $\hat{\mu}_i$ or $\hat{\eta}_i$ (add a smoothing graph (not too robust) to the plot for ease of interpretation).

There are also special methods to investigate whether a chosen link function or distribution is satisfactory. So called *overdispersion*, i.e., if the variance of Y_i is greater than expected given the distribution, can be an indication for the distribution not being

correct. A mean deviance, i.e., the deviance divided by the degrees of freedom of the approximate χ^2 distribution, which is considerably larger than 1 is indicative of overdispersion given the model. This can, for instance, occur as a result of a misspecification of the linear predictor, e.g, a relevant covariate is missing in the model, or an inappropriate link function or outliers in the data. If overdispersion cannot be remedied by changes in the linear predictor or link function, or by excluding outlying observations, a model can be fitted using a quasi-likelihood approach. For this approach, instead of specifying a likelihood (and, thus, a specific distribution) one can fix only the relationship between μ_i, the expectation of Y_i, and the variance of Y_i. If we model count data that exhibit overdispersion, instead of assuming a Poisson distribution, we can assume quasi-Poisson with the variance of Y_i modeled as $\tau^2\mu_i$ and estimate τ^2. Note that in such cases, comparisons between nested models cannot be done with the LRT, i.e., a deviance difference, but must be based upon comparing the deviance difference divided by the difference in degrees of freedom (*df*) with the deviance of the larger model divided by its *df*. This results in an approximate F-test. As a consequence of estimating τ^2 from the data, estimated variances for $\hat{\beta}$ are larger and CIs for β wider.

For Bernoulli data, assessing the adequacy of the fitted model is much more difficult due to the nature of the residuals. Also overdispersion cannot be assessed for such data.

5. Fitting GLMs in R

Fitting GLMs in R can be done with the function glm(). Important arguments of the glm() function are:

formula: as for e.g., function lm, the right hand side is a symbolic description of the linear predictor

family: specifies the distribution of the response variable and the link function to be used

...: several arguments for supplying starting values for the parameters, must not be specified (especially if canonical links are used)

offset: to specify an a priori known component to be included in the linear predictor during fitting

6. References

(1) Collett, David. *Modelling Binary Data.* CRC Press LLC. 2002: ch 3, 5, 6
(2) Stahel, Werner. *Statistische Regressionsmodelle, Teil II.* 2008: ch 12, 13
(3) McCulloch, Charles E and Searle, Shayle R. *Generalized, Linear, and Mixed Models.* John Wiley & Sons. 2001: ch 5
(4) ...