Lecture 4

Generalised linear models

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be independent pairs of observations, where Y_i is real-valued and X_i are $\mathbb{R}^k$-valued random variables. Generalised linear models have the following three-part specification:

1. The **random component** (=response from an overdispersed exponential family)
   The data $Y_1, \ldots, Y_n$ are such that $Y_1|X_1, \ldots, Y_n|X_n$ are independent and $Y_i|X_i$ has the p.d.f.
   $$f_{\eta,\psi}(y_i|x_i) = \exp\left\{\frac{\eta_i y_i - \kappa(\eta_i)}{\psi_i}\right\} h(y_i, \psi_i), \quad i = 1, \ldots, n,$$
   where η_i is called **canonical parameter** and ψ_i is an unknown **scale or dispersion parameter**. Functions κ and h are known and $\kappa''(\eta) > 0$ is assumed. It is easy to see that
   $$\mu(\eta_i) := \mathrm{E}(Y_i|X_i) = \kappa'(\eta_i) \quad \text{and} \quad \mathrm{var}(Y_i|X_i) = \psi_i \kappa''(\eta_i), \quad i = 1, \ldots, n.$$

2. The **systematic component** (=linear predictor)
   Canonical parameter η_i is assumed to be related to X_i. The term $X_i^t\beta$ for unknown $\beta \in \mathbb{R}^d$ is called the linear predictor or systematic component.

3. The **link function** between random and systematic components
   The relationship between η_i and $X_i^t\beta$ is described through
   $$g\{\mu(\eta_i)\} = X_i^t\beta, \quad i = 1, \ldots, n,$$
   where g is called a link function. The link function g is assumed to be a known, one-to-one, third-order continuously differentiable function. If $g = \mu^{-1}$, then $\eta_i = X_i^t\beta$ and g is called the **canonical or natural link** function. If g is not canonical, then it is assumed that $d(g \circ \mu)(\eta)/d\eta \neq 0$ for all η.

In a GLM, the parameter of interest is β. Parameters ψ_i are considered to be *nuisance* parameters. It is often assumed that $\psi_i = \psi/t_i$, $i = 1, \ldots, n$ with an unknown ψ and known t_i's or, alternatively $\psi_i = a(\psi)$ for some known function a. Note that ψ_i enter $\mathrm{var}(Y_i|X_i) = \psi_i \kappa''(\eta_i)$, making it more flexible, that is allowing for *over- or under-dispersion*.

Example

Let $Y_i|X_i \sim Poi(\lambda_i)$. We can write the density

$$f_\eta(y_i) = \frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i)\mathbb{1}_{\{0,1,2,\ldots\}}(y_i) =: \exp\{y_i\log(\lambda_i) - \lambda_i\}\frac{1}{y_i!}\mathbb{1}_{\{0,1,2,\ldots\}}(y_i),$$

that is, the canonical parameter $\eta_i = \log(\lambda_i)$, $\kappa(\eta_i) = \lambda_i = \exp(\eta_i)$, $\psi_i = 1$ and $h(y_i) = (y_i)^{-1}\mathbb{1}_{\{0,1,2,\ldots\}}(y_i)$. Since $\mathrm{E}(Y_i|X_i) = \kappa'(\eta_i) = \exp(\eta_i) =: \mu(\eta_i)$, the canonical link is $g(x) = \mu^{-1}(x) = \log(x)$, which is called the **log-link** ($g(\mu(\eta_i)) = \eta_i$). Hence,

$$\log\{\mathrm{E}(Y_i|X_i = x_i)\} = x_i^t\beta,$$

where $x_i \in \mathbb{R}^k$, $i = 1, \ldots, n$.

Estimation

Let $\theta = (\beta, \psi)$ and $(g \circ \mu)^{-1} = \zeta$ (for a canonical link $\zeta(x) \equiv x$). Then

$$\ell(\theta) = \sum_{i=1}^n \left[\frac{\zeta(X_i^t\beta)Y_i - \kappa\left\{\zeta(X_i^t\beta)\right\}}{a(\psi)} + \log h(Y_i, \psi)\right].$$

Further, consider the canonical link. Taking derivatives w.r.t. β and ψ we get the following score equations

$$\begin{aligned}
\frac{\partial\ell(\theta)}{\partial\beta} &= \frac{1}{a(\psi)}\sum_{i=1}^n \left\{Y_i - \mu(X_i^t\beta)\right\}X_i = 0 \\
\frac{\partial\ell(\theta)}{\partial\psi} &= \sum_{i=1}^n \left[\frac{\partial\log h(y_i,\psi)}{\partial\psi} + \left\{a^{-1}(\psi)\right\}'\left\{X_i^t\beta Y_i - \kappa(X_i^t\beta)\right\}\right] = 0,
\end{aligned}$$

where $\kappa'(X_i^t\beta) = \mu(X_i^t\beta)$ was used. If MLE of β exists, then it can be found from the first equation without estimating ψ. Estimation of ψ from the second equation in many cases is a difficult task and depends on a particular distribution.

To estimate β and study its properties we also need

$$-\frac{\partial^2\ell(\theta)}{\partial\beta\partial\beta^t} = \frac{1}{a(\psi)}\sum_{i=1}^n \kappa''(X_i^t\beta)X_iX_i^t =: -\frac{F_n(\beta)}{a(\psi)}$$

With this, we can set up the Newton-Raphson algorithm as

$$\widehat{\beta}^{(j+1)} = \widehat{\beta}^{(j)} + \left\{F_n\left(\widehat{\beta}^{(j)}\right)\right\}^{-1} S_n\left(\widehat{\beta}^{(j)}\right), \quad j = 0, 1, 2\ldots,$$

where $S_n(\beta) = a(\psi)\ell(\theta)/\partial\beta$.

The estimator $\widehat{\beta}$ of β, defined as a solution to $S_n(\beta) = 0$ can be shown to be consistent and asymptotically normal However, the case of non-canonical link has to be treated with care.

After the model is estimated, one would like to assess how good the model *fits* the data, i.e., to measure the discrepancy between the data $Y_i|X_i$ and estimated $\mathrm{E}(Y_i|X_i) = \mu_i$. Measures of discrepancy or *goodness-of-fit* can be formed in various ways, we will consider the deviance and generalised Pearson statistics.

The simplest model is the **null model**, it has only one parameter, representing a common mean μ, say, for all $Y_i|X_i$. At the other extreme is the **full model**, which has n parameters, one for each observation. The full model gives a baseline for measuring the discrepancy for an intermediate model with k parameters. Assume for the moment that ψ is known and denote $\ell(\widehat{\mu}, \psi)$ the log-likelihood with $\widehat{\mu} = g^{-1}(X\widehat{\beta})$. The maximum likelihood in the full model is then $\ell(Y, \psi)$ ($=\mu_i$ are replaced by Y_i). Then the **deviance of the fitted model** is defined as

$$D(Y, \widehat{\mu}) = a(\psi)\ 2\{\ell(Y, \psi) - \ell(\widehat{\mu}, \psi)\}.$$

Note that $D(Y, \widehat{\mu})/a(\psi)$ is called the **scaled deviance**. Some authors swap the definitions and call $2\{\ell(Y, \psi) - \ell(\widehat{\mu}, \psi)\}$ the deviance.

The **generalised Pearson statistic** is defined via

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where $\widehat{\mu}_i$ and $V(\widehat{\mu}_i)$ are the estimated $\mathrm{E}(Y_i|X_i)$ and $\mathrm{var}(Y_i|X_i)$, respectively.

Example
Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be independent, $Y_i|X_i \sim Poi(\lambda)$. Consider GLM with the canonical log-link. In this case $\mu_i = \exp(X_i^t\beta) = \exp(\eta_i) = \kappa(\eta_i) = \kappa'(\eta_i)$ and we find

$$\begin{aligned} \ell(\widehat{\mu}) &= \sum_{i=1}^{n}\{\log(\widehat{\mu}_i) - \widehat{\mu}_i + \log h(Y_i)\}, \quad \ell(Y) = \sum_{i=1}^{n}\{\log(Y_i) - Y_i + \log h(Y_i)\} \\ D(Y, \widehat{\mu}) &= 2\sum_{i=1}^{n}\left\{\log\left(\frac{Y_i}{\widehat{\mu}_i}\right) - (Y_i - \widehat{\mu}_i)\right\}. \end{aligned}$$

Since $V(\mu_i) = \mu_i = \exp(X_i^t\beta)$, the Pearson statistics is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{\{Y_i - \exp(X_i^t\widehat{\beta})\}^2}{\exp(X_i^t\widehat{\beta})}.$$

Scaled deviance can be used to compare two **nested models**, i.e. the parameter space under one model is a subspace of that under the second model. Assume $\eta_i = X_i^t\beta$, $\beta \in \mathbb{R}^k$ corresponds to a larger model M_k, say, and $\eta_i = \widetilde{X}_i^t\widetilde{\beta}$ with $\widetilde{\beta} \in \mathbb{R}^q$, $q < k$ corresponds to a smaller model M_q, say, where $\widetilde{X}$ is obtained from X by deleting $k - q$ columns of X. Models M_k and M_q are nested. If we would like to test the null hypothesis that a smaller model is as good as a larger one, this is equivalent to testing that $k - q$ parameters in M_k are zero. Thus, the difference between scaled deviances of M_k and M_q is the twice difference between log-likelihoods of models M_k and M_q and is asymptotically χ^2_{k-q} distributed. The corresponding test is known as the **analysis of deviance**.

Since scaled deviance itself compares two nested models (the full one and a smaller one), it seems tempting to use the scaled deviance as a measure of goodness-of-fit. It is claimed that both the deviance and the Pearson statistics are approximately $a(\psi)\chi^2_{n-k}$ distributed, where k is the dimension of β in the model under consideration. Therefore, a widely used rule of thumb is that a good fit has the scaled deviance about $n - k$, which is the expectation of a χ^2_{n-k} distributed random variable. Large values of the scaled deviance are considered to indicate a bad fit. However, this has to be treated with care. For Poisson data with large λ_i and Binomial data with large m_i the approximation to χ^2_{n-k} works reasonable, but not in many other cases.

To assess goodness-of-fit graphical tools such as **residual analysis** are used as well. There are several types of residuals, which are used for the residual analysis in the generalised linear models. These residuals are expected to behave approximately as zero-mean normally distributed variables.

Pearson residuals are defined by

$$r_i^p = \frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}, \quad i = 1, \ldots, n.$$

Pearson residuals have the disadvantage of being skewed for non-normal responses. Anscombe proposed a residual that uses a function $A(Y_i)$ instead of Y_i, where transformation A is chosen to make the distribution of $A(Y_i)$ as normal as possible. McCullagh and Nelder (1983) in Section 2.4.2 give **Anscombe residulas** for important cases. For example, for the Poisson distribution

$$r_i^a = \frac{3\left(Y_i^{2/3} - \widehat{\mu}_i^{2/3}\right)}{2\,\widehat{\mu}_i^{1/6}}, \quad i = 1, \ldots, n.$$

Another type of residuals, which is most widely used in practice, is based on the deviance.

The **deviance residuals** are given by

$$r_i^d = \text{sign}(Y_i - \widehat{\mu}_i)\sqrt{2\{\ell_i(Y_i, \psi) - \ell_i(\widehat{\mu}_i, \psi)\}}, \quad i = 1, \ldots, n,$$

where ℓ_i is the log-likelihood corresponding to the i-th observation, so that $\sum_{i=1}^n (r_i^d)^2 = D(Y, \widehat{\mu})$.

Deviance residuals typically behave better than the Pearson ones and for most distributions are quite similar to the Anscombe residuals.
Similar to the normal response, a standardised version of the deviance (as well as Pearson) residuals are used:

$$\frac{r_i^d}{\sqrt{a(\widehat{\psi})(1 - h_i)}}, \quad i = 1, \ldots, n,$$

where $h_i = H_{i,i}$ with the hat matrix H taking now the form $H = W^{1/2}X(X^tWX)^{-1}X^tW^{1/2}$, where W is the weight matrix from the Fisher scoring. In an adequate model the plot of standardised residuals against $\widehat{\eta} = X\widehat{\beta}$ should show no patterns. The so-called *null pattern* is a distribution of residuals with mean zero and constant variance.

To compare models with different subset of parameters or even to compare two different models (e.g., with different link functions or a non-linear and with a linear model), Akaike information criterion (AIC) and Bayes information criterion (BIC) can be used. These two criteria are most popular examples of **penalised goodness-of-fit** criteria

$$\begin{aligned} AIC(M) &= -2\ell(M) + 2|M| \\ BIC(M) &= -2\ell(M) + \log(n)|M|, \end{aligned}$$

where $\ell(M)$ denotes the log-likelihood corresponding to a model M and $|M|$ is the number of parameters in that model M. The models, selected with these criteria are then

$$\begin{aligned} \widehat{M}_{AIC} &= \arg\min_{M \in \mathcal{M}} AIC(M) \\ \widehat{M}_{BIC} &= \arg\min_{M \in \mathcal{M}} BIC(M) \end{aligned}$$

In these criteria the term $-2\ell(M)$ describes the goodness-of-fit (e.g., residual sum of squares in the normal regression), while the second term penalises the number of the parameters in the model. Obviously, the larger the number of parameters in the model is, the smaller is the goodness-of-fit. Hence, the second term is the penalty for the large number of parameters (roughly speaking, the goal is the best possible fit with the smallest possible number of parameters). Note that BIC has a larger penalty. One can show that BIC based model selection procedures are consistent, while AIC based procedures are conservative.