Chapter 9

Generalised Linear Models

To motivate the GLM approach let us briefly overview linear models.

9.1 An overview of linear models

Let us consider the two competing linear nested models

$$\begin{aligned}\text{Restricted model:} \qquad & Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j} + \varepsilon_i, \\ \text{Full model:} \qquad & Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j} + \sum_{j=q+1}^{p} \beta_j x_{i,j} + \varepsilon_i, \qquad (9.1)\end{aligned}$$

where $\{\varepsilon_i\}$ are iid random variables with mean zero and variance σ^2. Let us suppose that we observe $\{(Y_i, x_{i,j})\}_{i=1}^{n}$, where $\{Y_i\}$ are normal. The classical method for testing H_0 : Model 0 against H_A : Model 1 is to use the F-test (ANOVA). That is, let $\hat{\sigma}_R^2$ be the residual sum of squares under the null and $\hat{\sigma}_F^2$ be the residual sum of squares under the alternative. Then the F-statistic is

$$F = \frac{\left(S_R^2 - S_F^2\right)/(p-q)}{\hat{\sigma}_F^2},$$

where

$$\begin{aligned} S_F^2 &= \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} \hat{\beta}_j^F x_{i,j})^2 \quad S_R^2 = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{q} \hat{\beta}_j^R x_{i,j})^2 \\ \sigma_F^2 &= \frac{1}{n-p} \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} \hat{\beta}_j^F x_{i,j})^2. \end{aligned}$$

and under the null $F \sim F_{p-q,n-p}$. Moreover, if the sample size is large $(p-q)F \xrightarrow{\mathcal{D}} \chi^2_{p-q}$. We recall that the residuals of the full model are $r_i = Y_i - \hat{\beta}_0 - \sum_{j=1}^{q} \hat{\beta}_j x_{i,j} - \sum_{j=q+1}^{p} \hat{\beta}_j x_{i,j}$ and the residual sum of squares S_F^2 is used to measure how well the linear model fits the data (see STAT612 notes).

The F-test and ANOVA are designed specifically for linear models. In this chapter the aim is to generalise

- Model specification.
- Estimation
- Testing.
- Residuals.

to a larger class of models.

To generalise we will be in using a log-likelihood framework. To see how this fits in with the linear regression, let us now see how ANOVA and the log-likelihood ratio test are related. Suppose that σ^2 is known, then the log-likelihood ratio test for the above hypothesis is

$$\frac{1}{\sigma^2}\big(S_R^2 - S_F^2\big) \sim \chi^2_{p-q},$$

where we note that since $\{\varepsilon_i\}$ is Gaussian, this is the exact distribution and not an asymptotic result. In the case that σ^2 is unknown and has to be replaced by its estimator $\hat{\sigma}_F^2$, then we can either use the approximation

$$\frac{1}{\hat{\sigma}_F^2}\big(S_R^2 - S_F^2\big) \xrightarrow{\mathcal{D}} \chi^2_{p-q}, \quad n \to \infty,$$

or the exact distribution

$$\frac{\big(S_R^2 - S_F^2\big)/(p-q)}{\hat{\sigma}_F^2} \sim F_{p-q,n-p},$$

which returns us to the F-statistic.

On the other hand, if the variance σ^2 is unknown we return to the log-likelihood ratio statistic. In this case, the log-likelihood ratio statistic is

$$\log \frac{S_R^2}{S_F^2} = \log\left(1 + \frac{(S_F^2 - S_R^2)}{\hat{\sigma}_F^2}\right) \xrightarrow{\mathcal{D}} \chi^2_{p-q},$$

recalling that $\frac{1}{\hat{\sigma}}\sum_{i=1}^{n}(Y_i - \widehat{\beta}x_i) = n$. We recall that by using the expansion $\log(1+x) = x + O(x^2)$ we obtain

$$\begin{aligned}\log\frac{S_R^2}{S_F^2} &= \log\left(1 + \frac{(S_R^2 - S_F^2)}{S_F^2}\right)\\ &= \frac{S_R^2 - S_F^2}{S_F^2} + o_p(1).\end{aligned}$$

Now we know the above is approximately χ^2_{p-q}. But it is straightforward to see that by dividing by $(p-q)$ and multiplying by $(n-p)$ we have

$$\begin{aligned}\frac{(n-p)}{(p-q)}\log\frac{S_R^2}{S_F^2} &= \frac{(n-p)}{(p-q)}\log\left(1 + \frac{(S_R^2 - S_F^2)}{S_F^2}\right)\\ &= \frac{(S_R^2 - S_F^2)/(p-q)}{\hat{\sigma}_F^2} + o_p(1) = F + o_p(1).\end{aligned}$$

Hence we have transformed the log-likelihood ratio test into the F-test, which we discussed at the start of this section. The ANOVA and log-likelihood methods are asymptotically equivalent.

In the case that $\{\varepsilon_i\}$ are non-Gaussian, but the model is linear with iid random variables, the above results also hold. However, in the case that the regressors have a nonlinear influence on the response and/or the response is not normal we need to take an alternative approach. Through out this section we will encounter such models. We will start by focussing on the following two problems:

(i) How to model the relationship between the response and the regressors when the reponse is non-Gaussian, and the model is nonlinear.

(ii) Generalise ANOVA for nonlinear models.

9.2 Motivation

Let us suppose $\{Y_i\}$ are independent random variables where it is believed that the regressors x_i (x_i is a p-dimensional vector) has an influence on $\{Y_i\}$. Let us suppose that Y_i is a binary random variable taking either zero or one and $\mathrm{E}(Y_i) = P(Y_i = 1) = \pi_i$.

How to model the relationship between Y_i and x_i? A simple approach, is to use a linear model, ie. let $\mathrm{E}(Y_i) = \beta' x_i$, But a major problem with this approach is that $\mathrm{E}(Y_i)$, is a probability, and for many values of β, $\beta' x_i$ will lie outside the unit interval - hence a linear model is not meaningful. However, we can make a nonlinear transformation which transforms the a linear combination of the regressors to the unit interval. Such a meaningful transformation forms an important component in statistical modelling. For example let

$$\mathrm{E}(Y_i) = \pi_i = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} = \mu(\beta' x_i),$$

this transformation lies between zero and one. Hence we could just use nonlinear regression to estimate the parameters. That is rewrite the model as

$$Y_i = \mu(\beta' x_i) + \underbrace{\varepsilon_i}_{Y_i - \mu(\beta' x_i)}$$

and use the estimator $\widehat{\beta}_i$, where

$$\widehat{\beta}_n = \arg\min_{\beta} \sum_{i=1}^{n} \Big(Y_i - \mu(\beta' x_i) \Big)^2, \tag{9.2}$$

as an estimator of β. This method consistently estimates the parameter β, but there are drawbacks. We observe that Y_i are not iid random variables and

$$Y_i = \mu(\beta' x_i) + \sigma_i \epsilon_i$$

where $\{\epsilon_i = \frac{Y_i - \mu(\beta' x_i)}{\sqrt{Y_i}}\}$ are iid random variables and $\sigma_i = \sqrt{\mathrm{var} Y_i}$. Hence Y_i has a heterogeneous variance. However, the estimator in (9.2) gives each observation the same weight, without taking into account the variability between observations (which will result in a large variance in the estimator). To account for this one can use the weighted leasts squares estimator

$$\widehat{\beta}_n = \arg\min_{\beta} \sum_{i=1}^{n} \frac{(Y_i - \mu(\beta' x_i))^2}{\mu(\beta' x_i)(1 - \mu(\beta' x_i))}, \tag{9.3}$$

but there is no guarantee that such an estimator is even consistent (the only way to be sure is to investigate the corresponding estimating equation).

An alternative approach is to use directly use estimating equations (refer to Section 8.2). The the simplest one solves

$$\sum_{i=1}^{n} (Y_i - \mu(\beta' x_i)) = 0,$$

where $\mu(\beta' x_i)$. However, this solution does not lead to an estimator with the smallest "variance". Instead we can use the "optimal estimation equation" given in Section 8.3 (see equation 8.12). Using (8.12) the optimal estimating equation is

$$\begin{aligned} & \sum_{i=1}^{n} \frac{\mu_i'(\theta)}{V_i(\theta)} (Y_i - \mu_i(\theta)) \\ = \quad & \sum_{i=1}^{n} \frac{(Y_i - \mu(\beta' x_i))}{\mu(\beta' x_i)[1 - \mu(\beta' x_i)]} \frac{\partial \mu(\beta' x_i)}{\partial \beta} = \sum_{i=1}^{n} \frac{(Y_i - \mu(\beta' x_i))}{\mu(\beta' x_i)[1 - \mu(\beta' x_i)]} \mu'(\beta' x_i) x_i = 0, \end{aligned}$$

where we use the notation $\mu'(\theta) = \frac{d\mu(\theta)}{d\theta}$ (recall $var[Y_i] = \mu(\beta' x_i)(1 - \mu(\beta' x_i)))$. We show below (using the GLM machinery) that this corresponds to the score function of the log-likelihood function.

The GLM approach is a general framework for a wide class of distributions. We recall that in Section 1.6 we considered maximum likelihood estimation for iid random variables which come from the natural exponential family. Distributions in this family include the normal, binary, binomial and Poisson, amongst others. We recall that the natural exponential family has the form

$$f(y; \theta) = \exp\Big(y\theta - \kappa(\theta) + c(y)\Big),$$

where $\kappa(\theta) = b(\eta^{-1}(\theta))$. To be a little more general we will suppose that the distribution can be written as

$$f(y; \theta) = \exp\left(\frac{y\theta - \kappa(\theta)}{\phi} + c(y, \phi)\right), \tag{9.4}$$

where ϕ is a nuisance parameter (called the disperson parameter, it plays the role of the variance in linear models) and θ is the parameter of interest. We recall that examples of exponential models include

(i) The exponential distribution is already in natural exponential form with $\theta = \lambda$ and $\phi = 1$. The log density is

$$\log f(y; \theta) = -\lambda y + \log \lambda.$$

(ii) For the binomial distribution we let $\theta = \log(\frac{\pi}{1-\pi})$ and $\phi = 1$, since $\log(\frac{\pi}{1-\pi})$ is invertible this gives

$$\log f(y; \theta) = \log f(y; \log \frac{\pi}{1-\pi}) = \big(y\theta - n \log\big(\frac{\exp(\theta)}{1 + \exp(\theta)}\big) + \log \binom{n}{y}\big).$$

(iii) For the normal distribution we have that

$$\begin{aligned} \log f(y;\mu,\sigma^2) &= \Big(-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi)\Big) \\ &= \frac{-y^2+2\mu y-\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi). \end{aligned}$$

Suppose $\mu = \mu(\beta' x_i)$, whereas the variance σ^2 is constant for all i, then σ^2 is the scale parameter and we can rewrite the above as

$$\log f(y;\mu,\sigma^2) \;=\; \frac{\underbrace{\mu}_{\theta}\, y - \underbrace{\mu^2/2}_{\kappa(\theta)}}{\sigma^2} - \underbrace{\Big(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi)\Big)}_{=c(y,\phi)}.$$

(iv) The Poisson log distribution can be written as

$$\log f(y;\mu) = y\log\mu - \mu + \log y!,$$

Hence $\theta = \log\mu$, $\kappa(\theta) = -\exp(\theta)$ and $c(y) = \log y!$.

(v) Other members in this family include, Gamma, Beta, Multinomial and inverse Gaussian to name but a few.

Remark 9.2.1 (Properties of the exponential family (see Chapter 1 for details)) *(i) Using Lemma 1.6.3 (see Section 1.6) we have* $\mathrm{E}(Y) = \kappa'(\theta)$ *and* $\mathrm{var}(Y) = \kappa''(\theta)\phi$.

(ii) If the distribution has a "full rank parameter space" (number of parameters is equal to the number of sufficient statistics) and $\theta(\eta)$ (where η is the parameter of interest) is a diffeomorphism then the second derivative of the log-likelihood is non-negative. To see why we recall for a one-dimensional exponential family distribution of the form

$$f(y;\theta) = \exp\Big(y\theta - \kappa(\theta) + c(y)\Big),$$

the second derivative of the log-likelihood is

$$\frac{\partial^2 \log f(y;\theta)}{\partial\theta^2} = -\kappa''(\theta) = -\mathrm{var}[Y].$$

If we reparameterize the likelihood in terms of η, such that $\theta(\eta)$ then

$$\frac{\partial^2 \log f(y;\theta(\eta))}{\partial \eta^2} = y\theta''(\eta) - \kappa'(\theta)\theta''(\eta) - \kappa''(\theta)[\theta'(\eta)]^2.$$

Since $\theta(\eta)$ is a diffeomorphism between the space spanned by η to the space spanned by θ, $\log f(y;\theta(\eta))$ will be a deformed version of $\log f(y;\theta)$ but it will retain properties such concavity of the likelihood with respect to η.

GLM is a method which generalises the methods in linear models to the exponential family (recall that the normal model is a subclass of the exponential family). In the GLM setting it is usually assumed that the response variables $\{Y_i\}$ are independent random variables (but not identically distributed) with log density

$$\log f(y_i;\theta_i) = \left(\frac{y_i\theta_i - \kappa(\theta_i)}{\phi} + c(y_i,\phi)\right), \tag{9.5}$$

where the parameter θ_i depends on the regressors. The regressors influence the response through a linear predictor $\eta_i = \beta' x_i$ and a link function, which connects $\beta' x_i$ to the mean $\mathrm{E}(Y_i) = \mu(\theta_i) = \kappa'(\theta_i)$.

Remark 9.2.2 (Modelling the mean) *The main "philosophy/insight" of GLM is connecting the mean $\mu(\theta_i)$ of the random variable (or sufficient statistic) to a linear transform of the regressor $\beta' x_i$. The "link" function g is a monotonic (bijection) such that $\mu(\theta_i) = g^{-1}(\beta' x_i)$, and usually needs to be selected. The main features of the link function depends on the distribution. For example*

(i) If Y_i are positive then the link function g^{-1} should be positive (since the mean is positive).

(i) If Y_i take binary values the link function g^{-1} should lie between zero and one (it should be probability).

Let $g : \mathbb{R} \to \mathbb{R}$ be a bijection such that $g(\mu(\theta_i)) = \eta_i = \beta' x_i$. If we ignore the scale parameter, then by using Lemma 1.6.3 (which relates the mean and variance of sufficient statistics to $\kappa(\theta_i)$) we have

$$\begin{aligned}
\frac{d\kappa(\theta_i)}{d\theta_i} &= g^{-1}(\eta_i) = \mu(\theta_i) = \mathrm{E}(Y_i) \\
\theta_i &= \mu^{-1}(g^{-1}(\eta_i)) = \theta(\eta_i) \\
\mathrm{var}(Y_i) &= \frac{d^2\kappa(\theta_i)}{d\theta_i^2} = \frac{d\mu(\theta_i)}{d\theta_i}.
\end{aligned} \tag{9.6}$$

Based on the above and (9.5) the log likelihood function of $\{Y_i\}$ is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \bigg(\frac{Y_i\theta(\eta_i) - \kappa(\theta(\eta_i))}{\phi} + c(Y_i, \phi) \bigg).$$

Remark 9.2.3 (Concavity of the likelihood with regressors) *We mentioned in Remark 9.2.1 that natural exponential family has full rank and $\theta(\eta)$ is a reparameterisation in terms of η, then $\frac{\partial^2 logf(y;\eta)}{\partial\eta^2}$ is non-positive definite, thus $\log f(y;\theta)$ is a concave function. We now show that the likelihood in the presence of regressors in also concave.*

We recall that

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \bigg(Y_i\theta(\eta_i) - \kappa(\theta(\eta_i)) + c(Y_i) \bigg).$$

where $\eta_i = \beta' x_i$. Differentiating twice with respect to β gives

$$\nabla^2_{\beta}\mathcal{L}_n(\beta) = \mathbf{X}' \sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i;\theta(\eta_i))}{\partial\eta_i^2}\mathbf{X},$$

where $\mathbf{X}$ is the design matrix corresponding to the regressors. We mentioned above that $\frac{\partial^2 \log f(Y_i;\eta_i)}{\partial\eta_i^2}$ is non-positive definite for all i which in turn implies that its sum is non-positive definite. Thus $\mathcal{L}_n(\beta)$ is concave in terms of β, hence it is simple to maximise.

Example: Suppose the link function is in canonical form i.e. $\theta(\eta_i) = \beta' x_i$ (see the following example), the log-likelihood is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \bigg(Y_i\beta' x_i - \kappa(\beta' x_i) + c(Y_i) \bigg).$$

which has second derivative

$$\nabla^2_{\beta}\mathcal{L}_n(\beta) = -\mathbf{X}' \sum_{i=1}^{n} \kappa''(\beta' x_i)\mathbf{X}$$

which is clearly non-positive definite.

The choice of link function is rather subjective. One of the most popular is the canonical link which we define below.

Definition 9.2.1 (The canonical link function) *Every distribution within the exponential family has a* canonical link *function, this is where $\eta_i = \theta_i$. This immediately implies that $\mu_i = \kappa'(\eta_i)$ and $g(\kappa'(\theta_i)) = g(\kappa'(\eta_i)) = \eta_i$ (hence g is inverse function of κ').*

The canonical link is often used because it make the calculations simple (it also saves one from "choosing a link function"). We observe with the canonical link the log-likelihood of $\{Y_i\}$ is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \left(\frac{Y_i \beta' x_i - \kappa(\beta' x_i)}{\phi} + c(Y_i, \phi) \right).$$

Example 9.2.1 (The log-likelihood and canonical link function)

(i) The canonical link for the exponential $f(y_i; \lambda_i) = \lambda_i \exp(-\lambda_i y_i)$ is $\theta_i = -\lambda_i = \beta' x_i$, and $\lambda = -\beta' x_i$. The log-likelihood is

$$\sum_{i=1}^{n} \left(Y_i \beta' x_i - \log(\beta' x_i) \right).$$

(ii) The canonical link for the binomial is $\theta_i = \beta' x_i = \log(\frac{\pi_i}{1-\pi_i})$, hence $\pi_i = \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}$. The log-likelihood is

$$\sum_{i=1}^{n} \left(Y_i \beta' x_i + n_i \log\big(\frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}\big) + \log \binom{n_i}{Y_i} \right).$$

(iii) The canonical link for the normal is $\theta_i = \beta' x_i = \mu_i$. The log-likelihood is

$$\left(-\frac{(Y_i - \beta' x_i)^2}{2\sigma^2} + \frac{1}{2}\log \sigma^2 + \frac{1}{2}\log(2\pi) \right),$$

which is the usual least squared criterion. If the canonical link were not used, we would be in the nonlinear least squares setting, with log-likelihood

$$\left(-\frac{(Y_i - g^{-1}(\beta' x_i))^2}{2\sigma^2} + \frac{1}{2}\log \sigma^2 + \frac{1}{2}\log(2\pi) \right),$$

(iv) The canonical link for the Poisson is $\theta_i = \beta' x_i = \log \lambda_i$, hence $\lambda_i = \exp(\beta' x_i)$. The log-likelihood is

$$\sum_{i=1}^{n} \left(Y_i \beta' x_i - \exp(\beta' x_i) + \log Y_i! \right).$$

However, as mentioned above, the canonical link is simply used for its mathematical simplicity. There exists other links, which can often be more suitable.

Remark 9.2.4 (Link functions for the Binomial) *We recall that the link function is defined as a monotonic function g, where $\eta_i = \beta' x_i = g(\mu_i)$. The choice of link function is up to the practitioner. For the binomial distribution it is common to let g^{-1} = a well known distribution function. The motivation for this is that for the Binomial distribution $\mu_i = n_i \pi_i$ (where π_i is the probability of a 'success'). Clearly $0 \leq \pi_i \leq 1$, hence using g^{-1} = distribution function (or survival function) makes sense. Examples include*

(i) The Logistic link, this is the canonical link function, where $\beta' x_i = g(\mu_i) = \log(\frac{\pi_i}{1-\pi_i}) = \log(\frac{\mu_i}{n_i - \mu_i})$.

(i) The Probit link, where $\pi_i = \Phi(\beta' x_i)$, Φ is the standard normal distribution function and the link function is $\beta' x_i = g(\mu_i) = \Phi^{-1}(\mu_i / n_i)$.

(ii) The extreme value link, where the distribution function is $F(x) = 1 - \exp(-exp(x))$. Hence in this case the link function is $\beta' x_i = g(\mu_i) = \log(-\log(1 - \mu_i / n_i))$.

Remark 9.2.5 *GLM is the motivation behind single index models where $\mathrm{E}[Y_i | X_i] = \mu(\sum_{j=1}^{p} \beta_j x_{ij})$, where both the parameters $\{\beta_j\}$ and the link function $\mu(\cdot)$ is unknown.*

9.3 Estimating the parameters in a GLM

9.3.1 The score function for GLM

The score function for generalised linear models has a very interesting form, which we will now derive.

From now on, we will suppose that $\phi_i \equiv \phi$ for all t, and that ϕ is known. Much of the theory remains true without this restriction, but this makes the derivations a bit cleaner, and is enough for all the models we will encounter.

With this substitution, recall that the log-likelihood is

$$\mathcal{L}_n(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - \kappa(\theta_i)}{\phi} + c(Y_i, \phi) \right\} = \sum_{i=1}^{n} \ell_i(\beta, \phi),$$

where

$$\ell_i(\beta, \phi) = \left\{ \frac{Y_i \theta_i - \kappa(\theta_i)}{\phi} + c(Y_i, \phi) \right\}$$

and $\theta_i = \theta(\eta_i)$.

For the MLE of β, we need to solve the *likelihood equations*

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{for } j = 1, \ldots, p.$$

Observe that

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{(Y_i - \kappa'(\theta_i))}{\phi}\theta'(\eta_i)x_{ij}.$$

Thus the score equation is

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{[Y_i - \kappa'(\theta_i)]}{\phi}\theta'(\eta_i)x_{ij} = 0 \quad \text{for } j = 1, \ldots, p. \tag{9.7}$$

Remark 9.3.1 (Connection to optimal estimating equations) *Recall from (8.12) the optimal estimating equation is*

$$G_n(\beta) = \sum_{i=1}^{n} \frac{1}{V_i(\beta)}\left(Y_i - \mu_i(\beta)\right)\frac{\partial}{\partial \beta_j}\mu_i(\beta), \tag{9.8}$$

we now show this is equivalent to (9.7). Using classical results on the exponential family (see chapter 1) we have

$$\begin{aligned} \mathrm{E}[Y_i] &= \kappa'(\theta) = \mu_i(\beta) \\ \mathrm{var}[Y_i] &= \kappa''(\theta) = V_i(\beta). \end{aligned}$$

We observe that

$$\frac{\partial}{\partial \beta_j}\mu_i(\beta) = \underbrace{\frac{\partial \mu(\theta_i)}{\partial \theta_i}}_{=V_i(\beta)}\frac{\partial \theta_i}{\partial \beta_j} = V_i(\beta)\theta'(\eta_i)x_{ij},$$

substituting this into (9.8) gives

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^{n} \frac{[Y_i - \kappa'(\theta_i)]}{\phi}\theta'(\eta_i)x_{ij} = 0$$

which we see corresponds to the score of the likelihood.

To obtain an interesting expression for the above, recall that

$$\mathrm{var}(Y_i) = \phi\mu'(\theta_i) \text{ and } \eta_i = g(\mu_i),$$

and let $\mu'(\theta_i) = V(\mu_i)$. Since $V(\mu_i) = \frac{d\mu_i}{d\theta_i}$, inverting the derivative we have $\frac{d\theta_i}{d\mu_i} = 1/V(\mu_i)$. Furthermore, since $\frac{d\eta_i}{d\mu_i} = g'(\mu_i)$, inverting the derivative we have $\frac{d\mu_i}{d\eta_i} = 1/g'(\mu_i)$. By the chain rule for differentiation and using the above we have

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{d\ell_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{d\ell_i}{d\eta_i}\frac{\partial \eta_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} \qquad (9.9)\\ &= \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}\\ &= \frac{d\ell_i}{d\theta_i}\left(\frac{d\mu_i}{d\theta_i}\right)^{-1}\left(\frac{d\eta_i}{d\mu_i}\right)^{-1}\frac{\partial \eta_i}{\partial \beta_j}\\ &= \frac{(Y_i - \kappa'(\theta_i))}{\phi}(\kappa''(\theta_i))^{-1}(g'(\mu_i))^{-1}x_{ij}\\ &= \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)}\end{aligned}$$

Thus the likelihood equations we have to solve for the MLE of β are

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = \sum_{i=1}^{n} \frac{(Y_i - g^{-1}(\beta' x_i))x_{ij}}{\phi V(g^{-1}(\beta' x_i))g'(\mu_i)} = 0, \qquad 1 \le j \le p, \qquad (9.10)$$

(since $\mu_i = g^{-1}(\beta' x_i)$).

(9.10) has a very similar structure to the Normal equations arising in ordinary least squares.

Example 9.3.1 *(i) Normal $\{Y_i\}$ with mean $\mu_i = \beta' x_i$.*

Here, we have $g(\mu_i) = \mu_i = \beta' x_i$ so $g'(\mu_i) = \frac{dg(\mu_i)}{d\mu_i} \equiv 1$; also $V(\mu_i) \equiv 1$, $\phi = \sigma^2$, so the equations become

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta' x_i)x_{ij} = 0.$$

Ignoring the factor σ^2, the LHS is the jth element of the vector $X^T(Y - X\beta')$, so the equations reduce to the Normal equations of least squares:

$$X^T(Y - X\beta') = 0 \quad \textit{or equivalently} \quad X^T X\beta' = X^T Y.$$

(ii) Poisson $\{Y_i\}$ with log-link function, hence mean $\mu_i = \exp(\beta' x_i)$ (hence $g(\mu_i) = \log \mu_i$). This time, $g'(\mu_i) = 1/\mu_i$, $\mathrm{var}(Y_i) = V(\mu_i) = \mu_i$ and $\phi = 1$. Substituting $\mu_i = \exp(\beta' x_i)$, into (9.10) gives

$$\sum_{i=1}^{n}(Y_i - e^{\beta' x_i})x_{ij} = 0.$$

9.3.2 The GLM score function and weighted least squares

The GLM score has a very interesting relationship with weighted least squares. First recall that the GLM takes the form

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = \sum_{i=1}^{n} \frac{(Y_i - g^{-1}(\beta' x_i))x_{ij}}{\phi V_i g'(\mu_i)} = 0, \qquad 1 \leq j \leq p. \tag{9.11}$$

Next let us construct the weighted least squares criterion. Since $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \phi V_i$, the weighted least squares criterion corresponding to $\{Y_i\}$ is

$$S_i(\beta) = \sum_{i=1}^{n} \frac{(Y_i - \mu(\theta_i))^2}{\phi V_i} = \sum_{i=1}^{n} \frac{(Y_i - g^{-1}(\beta' x_i))^2}{\phi V_i}.$$

The weighted least squares criterion $\mathcal{S}_i$ is independent of the underlying distribution and has been constructed using the first two moments of the random variable. Returning to the weighted least squares estimator, we observe that this is the solution of

$$\frac{\partial S_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} + \sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial V_i} \frac{\partial V_i}{\partial \beta_j} = 0 \quad 1 \leq j \leq p,$$

where $s_i(\beta) = \frac{(Y_i - \mu(\theta_i))^2}{\phi V_i}$. Now let us compare $\frac{\partial S_i}{\partial \beta_j}$ with the estimating equations corresponding to the GLM (those in (9.11)). We observe that (9.11) and the first part of the RHS of the above are the same, that is

$$\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = 0.$$

In other words, the GLM estimating equations corresponding to the exponential family and the weighted least squares estimating equations are closely related (as are the corresponding estimators). However, it is simpler to solve $\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i} \frac{\partial \mu_i}{\beta_j} = 0$ than $\frac{\partial S_i}{\partial \beta_j} = 0$.

As an aside, note that since at the true β the derivatives are

$$\mathrm{E}\left(\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \right) = 0 \text{ and } \mathrm{E}\left(\frac{\partial S_i}{\partial \beta_j} \right) = 0,$$

then this implies that the other quantity in the partial sum, $\mathrm{E}\big(\frac{\partial S_i}{\partial \beta_j}\big)$ is also zero, i.e.

$$\mathrm{E}\left(\sum_{i=1}^{n} \frac{\partial s_i(\beta)}{\partial V_i} \frac{\partial V_i}{\partial \beta_j} \right) = 0.$$

9.3.3 Numerical schemes

The Newton-Raphson scheme

It is clear from the examples above that usually there does not exist a simple solution for the likelihood estimator of β. However, we can use the Newton-Raphson scheme to estimate β (and thanks to the concavity of the likelihood it is guaranteed to converge to the maximum). We will derive an interesting expression for the iterative scheme. Other than the expression being useful for implementation, it also highlights the estimators connection to weighted least squares.

We recall that the Newton Raphson scheme is

$$(\beta^{(m+1)})' = (\beta^{(m)})' - (H^{(m)})^{-1}u^{(m)}$$

where the $p \times 1$ gradient vector $u^{(m)}$ is

$$u^{(m)} = \left(\frac{\partial \mathcal{L}_n}{\partial \beta_1}, \ldots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}\right)' \rfloor_{\beta=\beta^{(m)}}$$

and the $p \times p$ Hessian matrix $H^{(m)}$ is given by

$$H^{(m)}_{jk} = \frac{\partial^2 \mathcal{L}_n(\beta)}{\partial \beta_j \partial \beta_k} \rfloor_{\beta=\beta^{(m)}},$$

for $j, k = 1, 2, \ldots, p$, both $u^{(m)}$ and $H^{(m)}$ being evaluated at the current estimate $\beta^{(m)}$.

By using (9.9), the score function at the mth iteration is

$$\begin{aligned} u^{(m)}_j &= \frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_j} \rfloor_{\beta=\beta^{(m)}} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} \rfloor_{\beta=\beta^{(m)}} \\ &= \sum_{i=1}^{n} \frac{d\ell_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \rfloor_{\beta=\beta^{(m)}} = \sum_{i=1}^{n} \frac{d\ell_i}{d\eta_i} \rfloor_{\beta=\beta^{(m)}} x_{ij}. \end{aligned}$$

The Hessian at the ith iteration is

$$
\begin{aligned}
H_{jk}^{(m)} &= \frac{\partial^2 \mathcal{L}_i(\beta)}{\partial\beta_j\partial\beta_k}\rfloor_{\beta=\beta^{(m)}} = \sum_{i=1}^{n} \frac{\partial^2 \ell_i}{\partial\beta_j\partial\beta_k}\rfloor_{\beta=\beta^{(m)}} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k}\left(\frac{\partial\ell_i}{\partial\beta_j}\right)\rfloor_{\beta=\beta^{(m)}} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k}\left(\frac{\partial\ell_i}{\partial\eta_i}x_{ij}\right)\rfloor_{\beta=\beta^{(m)}} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\eta_i}\left(\frac{\partial\ell_i}{\partial\eta_i}x_{ij}\right)x_{ik} \\
&= \sum_{i=1}^{n} \frac{\partial^2\ell_i}{\partial\eta_i^2}\rfloor_{\beta=\beta^{(m)}}x_{ij}x_{ik} \qquad (9.12)
\end{aligned}
$$

Let $s^{(m)}$ be an $n \times 1$ vector with

$$s_i^{(m)} = \frac{\partial\ell_i}{\partial\eta_i}\rfloor_{\beta=\beta^{(m)}}$$

and define the $n \times n$ diagonal matrix $\widetilde{W}^{(m)}$ with entries

$$\widetilde{W}_{ii} = -\frac{d^2\ell_i}{d\eta_i^2}.$$

Then we have $u^{(m)} = X^T s^{(m)}$ and $H = -X^T\widetilde{W}^{(m)}X$ and the Newton-Raphson iteration can succinctly be written as

$$
\begin{aligned}
(\beta^{(m+1)})' &= (\beta^{(m)})' - (H^{(m)})^{-1}u^{(m)} \\
&= (\beta^{(m)})' + (X^T\widetilde{W}^{(m)}X)^{-1}X^T s^{(m)}.
\end{aligned}
$$

Fisher scoring for GLMs

Typically, partly for reasons of tradition, we use a modification of this in fitting statistical models. The matrix $\widetilde{W}$ is replaced by W, another diagonal $n \times n$ matrix with

$$W_{ii}^{(m)} = \mathrm{E}(\widetilde{W}_{ii}^{(m)}|\beta^{(m)}) = \mathrm{E}\left(-\frac{d^2\ell_i}{d\eta_i^2}|\beta^{(m)}\right).$$

Using the results in Section 1.6 we have

$$W_{ii}^{(m)} = \mathrm{E}\left(-\frac{d^2\ell_i}{d\eta_i^2}|\beta^{(m)}\right) = \mathrm{var}\left(\frac{d\ell_i}{d\eta_i}|\beta^{(m)}\right)$$

so that $W = \text{var}(s^{(m)}|\beta^{(m)})$, and the matrix is non-negative-definite.

Using the Fisher score function the iteration becomes

$$(\beta^{(i+1)})' = (\beta^{(m)})' + (X^T W^{(m)} X)^{-1} X^T s^{(m)}.$$

Iteratively reweighted least squares

The iteration

$$(\beta^{(i+1)})' = (\beta^{(m)})' + (X^T W^{(m)} X)^{-1} X^T s^{(m)} \tag{9.13}$$

is similar to the solution for least squares estimates in linear models

$$\beta = (X^T X)^{-1} X^T Y$$

or more particularly the related *weighted least squares* estimates:

$$\beta = (X^T W X)^{-1} X^T W Y$$

In fact, (9.13) can be re-arranged to have exactly this form. Algebraic manipulation gives

$$\begin{aligned} (\beta^{(m)})' &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} X (\beta^{(m)})' \\ (X^T W^{(m)} X)^{-1} X^T s^{(m)} &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} (W^{(m)})^{-1} s^{(m)}. \end{aligned}$$

Therefore substituting the above into (9.13) gives

$$\begin{aligned} (\beta^{(m+1)})' &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} X (\beta^{(m)})' + (X^T W^{(m)} X)^{-1} X^T W^{(m)} (W^{(m)})^{-1} s^{(m)} \\ &= (X^T W^{(m)} X)^{-1} X^T W^{(m)} \big(X (\beta^{(m)})' + (W^{(m)})^{-1} s^{(m)} \big) \\ &:= (X^T W^{(m)} X)^{-1} X^T W^{(m)} Z^{(m)}. \end{aligned}$$

One reason that the above equation is of interest is that it has the 'form' of weighted least squares. More precisely, it has the form of a weighted least squares regression of $Z^{(m)}$ on X with the diagonal weight matrix $W^{(m)}$. That is let $z_i^{(m)}$ denote the ith element of the vector $Z^{(m)}$, then $\beta^{(m+1)}$ minimises the following weighted least squares criterion

$$\sum_{i=1}^{n} W_i^{(m)} \big(z_i^{(m)} - \beta' x_i \big)^2.$$

Of course, in reality $W_i^{(m)}$ and $z_i^{(m)}$ are functions of $\beta^{(m)}$, hence the above is often called *iteratively reweighted least squares.*

9.3.4 Estimating of the dispersion parameter ϕ

Recall that in the linear model case, the variance σ^2 did not affect the estimation of β.

In the general GLM case, continuing to assume that $\phi_i = \phi$, we have

$$s_i = \frac{d\ell_i}{d\eta_i} = \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i} = \frac{Y_i - \mu_i}{\phi V(\mu_i) g'(\mu_i)}$$

and

$$\begin{aligned} W_{ii} &= \mathrm{var}(s_i) = \frac{\mathrm{var}(Y_i)}{\{\phi V(\mu_i) g'(\mu_i)\}^2} = \frac{\phi V(\mu_i)}{\{\phi V(\mu_i) g'(\mu_i)\}^2} \\ &= \frac{1}{\phi V(\mu_i)(g'(\mu_i))^2} \end{aligned}$$

so that $1/\phi$ appears as a scale factor in W and s, but otherwise does not appear in the estimating equations or iteration for $\widehat{\beta}$. Hence ϕ does not play a role in the estimation of β.

As in the Normal/linear case, (a) we are less interested in ϕ, and (b) we see that ϕ can be separately estimated from β.

Recall that $\mathrm{var}(Y_i) = \phi V(\mu_i)$, thus

$$\frac{E((Y_i - \mu_i)^2)}{V(\mu_i)} = \phi$$

We can use this to suggest a simple estimator for ϕ:

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} = \frac{1}{n-p}\sum_{i=1}^{T} \frac{(Y_i - \widehat{\mu}_i)^2}{\mu'(\widehat{\theta}_i)}$$

where $\widehat{\mu_i} = g^{-1}(\widehat{\beta}' x_i)$ and $\widehat{\theta}_i = \mu^{-1} g^{-1}(\widehat{\beta}' x_i)$. Recall that the above resembles estimators of the residual variance. Indeed, it can be argued that the distribution of the above is close to χ^2_{n-p}.

Remark 9.3.2 *We mention that a slight generalisation of the above is when the dispersion parameter satisfies $\phi_i = a_i \phi$, where a_i is known. In this case, an estimator of the ϕ is*

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{a_i V(\widehat{\mu}_i)} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{a_i \mu'(\widehat{\theta}_i)}$$

9.3.5 Deviance, scaled deviance and residual deviance

Scaled deviance

Instead of *minimising* the sum of squares (which is done for linear models) we have been *maximising* a log-likelihood $\mathcal{L}_i(\beta)$. Furthermore, we recall

$$S(\hat{\beta}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \big(Y_i - \hat{\beta}_0 - \sum_{j=1}^{q} \hat{\beta}_j x_{i,j} - \sum_{j=q+1}^{p} \hat{\beta}_j x_{i,j}\big)^2$$

is a numerical summary of how well the linear model fitted, $S(\widehat{\beta}) = 0$ means a perfect fit. A perfect fit corresponds to the Gaussian log-likelihood $-\frac{n}{2}\log\sigma^2$ (the likelihood cannot be larger than this).

In this section we will define the equivalent of residuals and square residuals for GLM.

What is the *best* we can do in fitting a GLM? Recall

$$\ell_i = \frac{Y_i\theta_i - \kappa(\theta_i)}{\phi} + c(Y_i, \phi)$$

so

$$\frac{d\ell_i}{d\theta_i} = 0 \iff Y_i - \kappa'(\theta_i) = 0$$

A model that achieves this equality for all i is called *saturated* (the same terminology is used for linear models). In other words, will need one free parameter for each observation. Denote the corresponding θ_i by $\widetilde{\theta}_i$, i.e. the solution of $\kappa'(\widetilde{\theta}_i) = Y_i$.

Consider the differences

$$\begin{aligned} 2\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\} &= \frac{2}{\phi}\{Y_i(\widetilde{\theta}_i - \theta_t) - \kappa(\widetilde{\theta}_i) + \kappa(\theta_i)\} \geq 0 \\ \text{and } 2\sum_{i=1}^{n}\Big\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\Big\} &= \frac{2}{\phi}\{Y_i(\widetilde{\theta}_i - \theta_t) - \kappa(\widetilde{\theta}_i) + \kappa(\theta_i)\}. \end{aligned}$$

Maximising the likelihood is the same as *minimising* the above quantity, which is always non-negative, and is 0 only if there is a perfect fit for all the i^{th} observations. This is analogous to linear models, where maximising the normal likelihood is the same as minimising least squares criterion (which is zero when the fit is best). Thus $\mathcal{L}_n(\widetilde{\theta}) = \sum_{i=1}^{n} \ell_i(\tilde{\theta}_i)$ provides a baseline value for the log-likelihood in much the same way that $-\frac{n}{2}\log\sigma^2$ provides a baseline in least squares (Gaussian set-up).

Example 9.3.2 (The normal linear model) *$\kappa(\theta_i) = \frac{1}{2}\theta_i^2$, $\kappa'(\theta_i) = \theta_i = \mu_i$, $\tilde{\theta}_i = Y_t$ and $\phi = \sigma^2$ so*

$$2\{\ell_i(\widetilde{\theta}_n) - \ell_i(\theta_i)\} = \frac{2}{\sigma^2}\{Y_i(Y_i - \mu_i) - \frac{1}{2}Y_i^2 + \frac{1}{2}\mu_i^2\} = (Y_i - \mu_i)^2/\sigma^2.$$

Hence for Gaussian observations $2\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\}$ corresponds to the classical residual squared. □

In general, let

$$D_i = 2\{Y_i(\widetilde{\theta}_i - \hat{\theta}_i) - \kappa(\widetilde{\theta}_i) + \kappa(\hat{\theta}_i)\}$$

We call $D = \sum_{i=1}^n D_i$ the *deviance* of the model. If ϕ is present, let

$$\frac{D}{\phi} = 2\{\mathcal{L}_n(\widetilde{\theta}) - \mathcal{L}_n(\widehat{\theta})\}.$$

$\phi^{-1}D$ is the *scaled deviance*. Thus the residual deviance plays the same role for GLM's as does the residual sum of squares for linear models.

Interpreting D_i

We will now show that

$$D_i \;=\; 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \kappa(\tilde{\theta}_i) + \kappa(\hat{\theta}_i)\} \approx \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

To show the above we require expression for $Y_i(\tilde{\theta}_i - \hat{\theta}_i)$ and $\kappa(\tilde{\theta}_i) - \kappa(\hat{\theta}_i)$. We use Taylor's theorem to expand κ and κ' about $\hat{\theta}_i$ to obtain

$$\kappa(\tilde{\theta}_i) \approx \kappa(\hat{\theta}_i) + (\tilde{\theta}_i - \hat{\theta}_i)\kappa'(\hat{\theta}_i) + \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2\kappa''(\hat{\theta}_i) \tag{9.14}$$

and

$$\kappa'(\tilde{\theta}_i) \approx \kappa'(\hat{\theta}_i) + (\tilde{\theta}_i - \hat{\theta}_i)\kappa''(\hat{\theta}_i) \tag{9.15}$$

But $\kappa'(\tilde{\theta}_i) = Y_i$, $\kappa'(\hat{\theta}_i) = \widehat{\mu}_i$ and $\kappa''(\hat{\theta}_i) = V(\widehat{\mu}_i)$, so (9.14) becomes

$$\begin{aligned}\kappa(\tilde{\theta}_i) &\approx \kappa(\hat{\theta}_i) + (\tilde{\theta}_i - \hat{\theta}_i)\widehat{\mu}_i + \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i) \\ \Rightarrow \kappa(\tilde{\theta}_i) - \kappa(\hat{\theta}_i) &\approx (\tilde{\theta}_i - \hat{\theta}_i)\widehat{\mu}_i + \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i), \qquad (9.16)\end{aligned}$$

and (9.15) becomes

$$\begin{aligned} Y_i &\approx \widehat{\mu}_i + (\tilde{\theta}_i - \hat{\theta}_i)V(\widehat{\mu}_i) \\ \Rightarrow Y_i - \widehat{\mu}_i &\approx (\tilde{\theta}_i - \hat{\theta}_i)V(\widehat{\mu}_i) \end{aligned} \tag{9.17}$$

Now substituting (9.16) and (9.17) into D_i gives

$$\begin{aligned} D_i &= 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \kappa(\tilde{\theta}_i) + \kappa(\hat{\theta}_i)\} \\ &\approx 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (\tilde{\theta}_i - \hat{\theta}_i)\widehat{\mu}_i - \frac{1}{2}(\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i)\} \\ &\approx (\tilde{\theta}_i - \hat{\theta}_i)^2 V(\widehat{\mu}_i) \approx \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}. \end{aligned}$$

Recalling that $\mathrm{var}(Y_i) = \phi V(\mu_i)$ and $\mathrm{E}(Y_i) = \mu_i$, $\phi^{-1}D_i$ behaves like a standardised squared residual. The signed square root of this approximation is called the *Pearson residual*. In other words

$$\mathrm{sign}(Y_i - \widehat{\mu}_i) \times \sqrt{\frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}} \tag{9.18}$$

is called a Pearson residual. The distribution theory for this is very approximate, but a rule of thumb is that *if the model fits*, the scaled deviance $\phi^{-1}D$ (or in practice $\widehat{\phi}^{-1}D$) $\approx \chi^2_{n-p}$.

Deviance residuals

The analogy with the normal example can be taken further. The square roots of the individual terms in the residual sum of squares are the residuals, $Y_i - \beta' x_i$.

We use the square roots of the individual terms in the deviances residual in the same way. However, the classical residuals can be both negative and positive, and the deviances residuals should behave in a similar way. But what sign should be used? The most obvious solution is to use

$$r_i = \begin{cases} -\sqrt{D}_i & \text{if } Y_i - \widehat{\mu}_i < 0 \\ \sqrt{D}_i & \text{if } Y_i - \widehat{\mu}_i \geq 0 \end{cases}$$

Thus we call the quantities $\{r_i\}$ the *deviance residuals*. Observe that the deviance residuals and Pearson residuals (defined in (9.18)) are the same up to the standardisation $\sqrt{V(\widehat{\mu}_i)}$.

Diagnostic plots

We recall that for linear models we would often plot the residuals against the regressors to check to see whether a linear model is appropriate or not. One can make similar diagnostics plots which have exactly the same form as linear models, except that deviance residuals are used instead of ordinary residuals, and linear predictor values instead of fitted values.

9.4 Limiting distributions and standard errors of estimators

In the majority of examples we have considered in the previous sections (see, for example, Section 2.2) we observed iid $\{Y_i\}$ with distribution $f(\cdot;\theta)$. We showed that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \approx \mathcal{N}\left(0, I(\theta)^{-1}\right),$$

where $I(\theta) = \int -\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)dx$ ($I(\theta)$ is Fisher's information). However this result was based on the observations being iid. In the more general setting where $\{Y_i\}$ are independent but not identically distributed it can be shown that

$$\left(\widehat{\beta} - \beta\right) \approx \mathcal{N}_p(0, (I(\beta))^{-1})$$

where now $I(\beta)$ is a $p \times p$ matrix (of the entire sample), where (using equation (9.12)) we have

$$(I(\beta))_{jk} = E\left(-\frac{\partial^2 \mathcal{L}_n(\beta)}{\partial\beta_j \partial\beta_k}\right) = \mathrm{E}\left(-\sum_{i=1}^{n} \frac{d^2\ell_i}{d\eta_i^2} x_{ij}x_{ik}\right) = (X^T W X)_{jk}.$$

Thus for large samples we have

$$\left(\widehat{\beta} - \beta\right) \approx \mathcal{N}_p(0, (X^T W X)^{-1}),$$

where W is evaluated at the MLE $\widehat{\beta}$.

Analysis of deviance

How can we test hypotheses about models, and in particular decide which explanatory variables to include? The two close related methods we will consider below are the log-likelihood ratio test and an analogue of the analysis of variance (ANOVA), called the analysis of deviance.

Let us concentrate on the simplest case, of testing a full vs. a reduced model. Partition the model matrix X and the parameter vector β as

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where X_1 is $n \times q$ and β_1 is $q \times 1$ (this is analogous to equation (9.1) for linear models). The full model is $\eta = X\beta' = X_1\beta_1 + X_2\beta_2$ and the reduced model is $\eta = X_1\beta_1'$. We wish to test $H_0 : \beta_2 = 0$, i.e. that the reduced model is adequate for the data.

Define the rescaled deviances for the full and reduced models

$$\frac{D_R}{\phi} = 2\{\mathcal{L}_n(\tilde{\theta}) - \sup_{\beta_2=0,\beta_1} \mathcal{L}_n(\theta)\}$$

and

$$\frac{D_F}{\phi} = 2\{\mathcal{L}_n(\tilde{\theta}) - \sup_{\beta_1,\beta_2} \mathcal{L}_n(\beta)\}$$

where we recall that $\mathcal{L}_n(\tilde{\theta}) = \sum_{i=1}^{T} \ell_t(\tilde{\theta}_i)$ is likelihood of the saturated model defined in Section 9.3.5. Taking differences we have

$$\frac{D_R - D_F}{\phi} = 2\big\{ \sup_{\beta_1,\beta_2} \mathcal{L}_n(\beta) - \sup_{\beta_2=0,\beta_1} \mathcal{L}_n(\theta)\big\}$$

which is the likelihood ratio statistic.

The results in Theorem 3.1.1, equation (3.7) (the log likelihood ratio test for composite hypothesis) also hold for observations which are not identically distributed. Hence using a generalised version of Theorem 3.1.1 we have

$$\frac{D_R - D_F}{\phi} = 2\big\{ \sup_{\beta_1,\beta_2} \mathcal{L}_n(\beta) - \sup_{\beta_2=0,\beta_1} \mathcal{L}_n(\theta)\big\} \xrightarrow{\mathcal{D}} \chi^2_{p-q}.$$

So we can conduct a test of the adequacy of the reduced model $\frac{D_R - D_F}{\phi}$ by referring to a χ^2_{p-q}, and rejecting H_0 if the statistic is too large (p-value too small). If ϕ is not present in the model, then we are good to go.

If ϕ is present (and unknown), we estimate ϕ with

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_t)} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\mu'(\widehat{\theta}_n)}$$

(see Section 9.3.4). Now we consider $\frac{D_R - D_F}{\widehat{\phi}}$, we can then continue to use the χ^2_{p-q} distribution, but since we are estimating ϕ we can use the statistic

$$\frac{D_R - D_F}{p-q} \div \frac{D_F}{n-p} \quad \text{against} \quad F_{(p-q),(n-p)},$$

as in the normal case (compare with Section 9.1).

9.5 Examples

Example 9.5.1 *Question Suppose that $\{Y_i\}$ are independent random variables with the canonical exponential family, whose logarithm satisfies*

$$\log f(y;\theta_i) = \frac{y\theta_i - \kappa(\theta_i)}{\phi} + c(y;\phi),$$

where ϕ is the dispersion parameter. Let $E(Y_i) = \mu_i$. Let $\eta_i = \beta' x_i = \theta_i$ (hence the canonical link is used), where x_i are regressors which influence Y_i. *[14]*

(a) (m) Obtain the log-likelihood of $\{(Y_i, x_i)\}_{i=1}^n$.

(ii) Denote the log-likelihood of $\{(Y_i, x_i)\}_{i=1}^n$ as $\mathcal{L}_n(\beta)$. Show that

$$\frac{\partial \mathcal{L}_n}{\partial \beta_j} = \sum_{i=1}^n \frac{(Y_i - \mu_i)x_{i,j}}{\phi} \quad \text{and} \quad \frac{\partial^2 \mathcal{L}_n}{\partial \beta_k \partial \beta_j} = -\sum_{i=1}^n \frac{\kappa''(\theta_i)x_{i,j}x_{i,k}}{\phi}.$$

(b) Let Y_i have Gamma distribution, where the log density has the form

$$\log f(Y_i;\mu_i) = \frac{-Y_i/\mu_i - \log \mu_i}{\nu^{-1}} + \left\{ -\frac{1}{\nu^{-1}} \log \nu^{-1} + \log \Gamma(\nu^{-1}) \right\} + \left\{ \nu^{-1} - 1 \right\} \log Y_i$$

$E(Y_i) = \mu_i$, $var(Y_i) = \mu_i^2/\nu$ and $\nu_i = \beta' x_i = g(\mu_i)$.

(m) What is the canonical link function for the Gamma distribution and write down the corresponding likelihood of $\{(Y_i, x_i)\}_{i=1}^n$.

(ii) Suppose that $\eta_i = \beta' x_i = \beta_0 + \beta_1 x_{i,1}$. Denote the likelihood as $\mathcal{L}_n(\beta_0, \beta_1)$. What are the first and second derivatives of $\mathcal{L}_n(\beta_0, \beta_1)$?

(iii) Evaluate the Fisher information matrix at β_0 and $\beta_1 = 0$.

(iv) Using your answers in (ii,iii) and the mle of β_0 with $\beta_1 = 0$, derive the score test for testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$.

Solution

(a) (m) The general log-likelihood for $\{(Y_i, x_i)\}$ with the canonical link function is

$$\mathcal{L}_n(\beta, \phi) = \sum_{i=1}^n \left(\frac{Y_i(\beta' x_i - \kappa(\beta' x_i))}{\phi} + c(Y_i, \phi) \right).$$

(ii) In the differentiation use that $\kappa'(\theta_i) = \kappa'(\beta' x_i) = \mu_i$.

(b) (m) For the gamma distribution the canonical link is $\theta_i = \eta_i = -1/\mu_i = -1/beta' x_i$. Thus the log-likelihood is

$$\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \frac{1}{\nu}\Big(Y_i(\beta' x_i) - \log(-1/\beta' x_i) \Big) + c(\nu_1, Y_i),$$

where $c(\cdot)$ can be evaluated.

(ii) Use part (ii) above to give

$$\begin{aligned} \frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_j} &= \nu^{-1} \sum_{i=1}^{n} \Big(Y_i + 1/(\beta' x_i) \Big) x_{i,j} \\ \frac{\partial \mathcal{L}_n(\beta)}{\partial \beta_i \partial \beta_j} &= -\nu^{-1} \sum_{i=1}^{n} \frac{1}{(\beta' x_i)} \Big) x_{i,i} x_{i,j} \end{aligned}$$

(iii) Take the expectation of the above at a general β_0 and $\beta_1 = 0$.

(iv) Using the above information, use the Wald test, Score test or log-likelihood ratio test.

Example 9.5.2 *Question: It is a belief amongst farmers that the age of a hen has a negative influence on the number of eggs she lays and the quality the eggs. To investigate this, m hens were randomly sampled. On a given day, the total number of eggs and the number of bad eggs that each of the hens lays is recorded. Let N_i denote the total number of eggs hen i lays, Y_i denote the number of bad eggs the hen lays and x_i denote the age of hen i.*

It is known that the number of eggs a hen lays follows a Poisson distribution and the quality (whether it is good or bad) of a given egg is an independent event (independent of the other eggs).

Let N_i be a Poisson random variable with mean λ_i, where we model $\lambda_i = \exp(\alpha_0 + \gamma_1 x_i)$ and π_i denote the probability that hen i lays a bad egg, where we model π_i with

$$\pi_i = \frac{\exp(\beta_0 + \gamma_1 x_i)}{1 + \exp(\beta_0 + \gamma_1 x_i)}.$$

Suppose that $(\alpha_0, \beta_0, \gamma_1)$ are unknown parameters.

(a) Obtain the likelihood of $\{(N_i, Y_i)\}_{i=1}^{m}$.

(b) Obtain the estimating function (score) of the likelihood and the Information matrix.

(c) Obtain an iterative algorithm for estimating the unknown parameters.

(d) For a given $\alpha_0, \beta_0, \gamma_1$, evaluate the average number of bad eggs a 4 year old hen will lay in one day.

(e) Describe in detail a method for testing $H_0 : \gamma_1 = 0$ against $H_A : \gamma_1 \neq 0$.

Solution

(a) Since the canonical links are being used the log-likelihood function is

$$\begin{aligned}
\mathcal{L}_m(\alpha_0, \beta_0, \gamma_1) &= \mathcal{L}_m(\underline{Y}|\underline{N}) + \mathcal{L}_m(\underline{N}) \\
&= \sum_{i=1}^{m} \Big(Y_i \underline{\beta}\underline{x}_i - N_i \log(1 + \exp(\underline{\beta}\underline{x}_i)) + N_i \underline{\alpha}\underline{x}_i - \underline{\alpha}\underline{x}_i + \log \binom{N_i}{Y_i} + \log N_i! \Big) \\
&\propto \sum_{i=1}^{m} \Big(Y_i \underline{\beta}\underline{x}_i - N_i \log(1 + \exp(\underline{\beta}\underline{x}_i)) + N_i \underline{\alpha}\underline{x}_i - \underline{\alpha}\underline{x}_i \Big).
\end{aligned}$$

where $\underline{\alpha} = (\alpha_0, \gamma_1)'$, $\underline{\beta} = (\beta_0, \gamma_1)'$ and $\underline{x}_i = (1, x_i)$.

(b) We know that if the canonical link is used the score is

$$\nabla \mathcal{L} = \sum_{i=1}^{m} \phi^{-1} \big(Y_i - \kappa'(\beta' x_i) \big) = \sum_{i=1}^{m} \big(Y_i - \mu_i \big)$$

and the second derivative is

$$\nabla^2 \mathcal{L} = -\sum_{i=1}^{m} \phi^{-1} \kappa''(\beta' x_i) = -\sum_{i=1}^{m} \mathrm{var}(Y_i).$$

Using the above we have for this question the score is

$$\begin{aligned}
\frac{\partial \mathcal{L}_m}{\partial \alpha_0} &= \sum_{i=1}^{m} \big(N_i - \lambda_i \big) \\
\frac{\partial \mathcal{L}_m}{\partial \beta_0} &= \sum_{i=1}^{m} \big(Y_i - N_i \pi_i \big) \\
\frac{\partial \mathcal{L}_m}{\partial \gamma_1} &= \sum_{i=1}^{m} \Big(\big(N_i - \lambda_i \big) + \big(Y_i - N_i \pi_i \big) \Big) x_i.
\end{aligned}$$

The second derivative is

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_m}{\partial \alpha_0^2} &= -\sum_{i=1}^{m} \lambda_i \quad \frac{\partial^2 \mathcal{L}_m}{\partial \alpha_0 \partial \gamma_1} = -\sum_{i=1}^{m} \lambda_i x_i \\
\frac{\partial^2 \mathcal{L}_m}{\partial \beta_0^2} &= -\sum_{i=1}^{m} N_i \pi_i (1-\pi_i) \quad \frac{\partial^2 \mathcal{L}_m}{\partial \beta_0 \partial \gamma_1} = -\sum_{i=1}^{m} N_i \pi_i (1-\pi_i) x_i \\
\frac{\partial^2 \mathcal{L}_m}{\partial \gamma_1 2} &= -\sum_{i=1}^{m} \Big(\lambda_i + N_i \pi_i (1-\pi_i) \Big) x_i^2.
\end{aligned}$$

Observing that $E(N_i) = \lambda_i$ the information matrix is

$$I(\theta) = \left(\begin{array}{ccc} \sum_{i=1}^{m} \lambda_i & 0 & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i \\ 0 & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i \\ \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i & \sum_{i=1}^{m} \lambda_i \pi_i (1-\pi_i) x_i & \sum_{i=1}^{m} \Big(\lambda_i + \lambda_i \pi_i (1-\pi_i) \Big) x_i^2 \end{array} \right).$$

(c) *We can estimate $\theta_0 = (\alpha_0, \beta_0, \gamma_1)$ using Newton-Raphson with Fisher scoring, that is*

$$\theta_i = \theta_i + I(\theta_i)^{-1} S_{i-1}$$

where

$$S_{i-1} = \left(\begin{array}{c} \sum_{i=1}^{m} \big(N_i - \lambda_i \big) \\ \sum_{i=1}^{m} \big(Y_i - N_i \pi_i \big) \\ \sum_{i=1}^{m} \Big(\big(N_i - \lambda_i \big) + \big(Y_i - N_i \pi_i \big) \Big) x_i. \end{array} \right).$$

(d) *We note that given the regressor $x_i = 4$, the average number of bad eggs will be*

$$\begin{aligned}
E(Y_i) &= E(E(Y_i|N_i)) = E(N_i \pi_i) = \lambda_i \pi_i \\
&= \frac{\exp(\alpha_0 + \gamma_1 x_i) \exp(\beta_0 + \gamma_1 x_i)}{1 + \exp(\beta_0 + \gamma_1 x_i)}.
\end{aligned}$$

(e) *Give either the log likelihood ratio test, score test or Wald test.*