Chapter 3

Introduction to generalized linear models

Contents

3 Introduction to generalized linear models **1**

1 Introduction . . . 3
1.1 Motivating examples . . . 3
2 The exponential family . . . 5
2.1 The exponential family form and maximum likelihood . . . 5
3 Components of a generalized linear model . . . 6
4 Estimating Generalized Linear Models . . . 7
4.1 General case . . . 7
4.2 Estimation of the dispersion parameter . . . 9
4.3 Example 1 . . . 10
5 Inference . . . 10
6 Diagnostics for GLMs . . . 12
6.1 Residuals . . . 12

1 Introduction

Generalized linear models (GLMs) expand the the well known linear model to accommodate non-normal response variables in a single unified approach. It is common to find response variables which do not fit the standard assumptions of the linear models (normally distributed errors, constant variance, etc.), for example: count data, dichotomous variables, truncated data, etc. The GLMs is based on well developed theory, starting with Nelder and Wedderburn (1972) and McCullagh and Nelder (1989), since then, and with the advances in statistical software, these models have become a basic tool for most researchers.
There are two fundamental issues in the notion of generalized linear models: the distribution of the response (as we mentioned above), but also the model that relates the mean response to the regression variables.

1.1 Motivating examples

We will motivate the need for GLMs, using two data sets where the distribution of the response variable has the property that the mean response, which is the expected response at each data point, and the variance of the response are related.

Toxicity experiment

The experiment tried to establish the relationship between the concentration of a toxic agent (nicotine) and the number of insects (common fruit fly) killed, the data can be found in Myers et al. (2002): The data follow a Binomial distribution

x Concentration (g/100cc)	n Number of insects	y Number Killed	Percent Killed
0.1	47	8	17.0
0.15	53	14	26.4
0.20	55	24	43.6
0.30	52	32	61.5
0.50	46	38	82.6
0.70	54	50	92.6
0.95	52	50	96.2

Table 1: Toxicity experiment data

$$y_i \sim B(n_i, p_i) \quad E[y_i] = n_i p_i(x_i) \quad Var[y_i] = n_i p_i(1 - p_i)$$

Clearly, the variance and the mean are a function of predictor variables, and mean and variance are related. Using ordinary linear regression to predict the percentage of flies killed would assume that the data are normally distributed, which is false, and with this type of data we have the constrain : $0 \leq p_i \leq 1$ which is not taken into account in linear regression. Fitting the model $p_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ we obtain

```
(lm(perc~x))$fitted
        1         2         3         4         5         6
0.3066231 0.3532890 0.3999550 0.4932869 0.6799507 0.8666145

    7
1.0999442
```

The fitted value of the last point is greater than 1!!!. Also, as we see in figure 1, the model is clearly non-linear

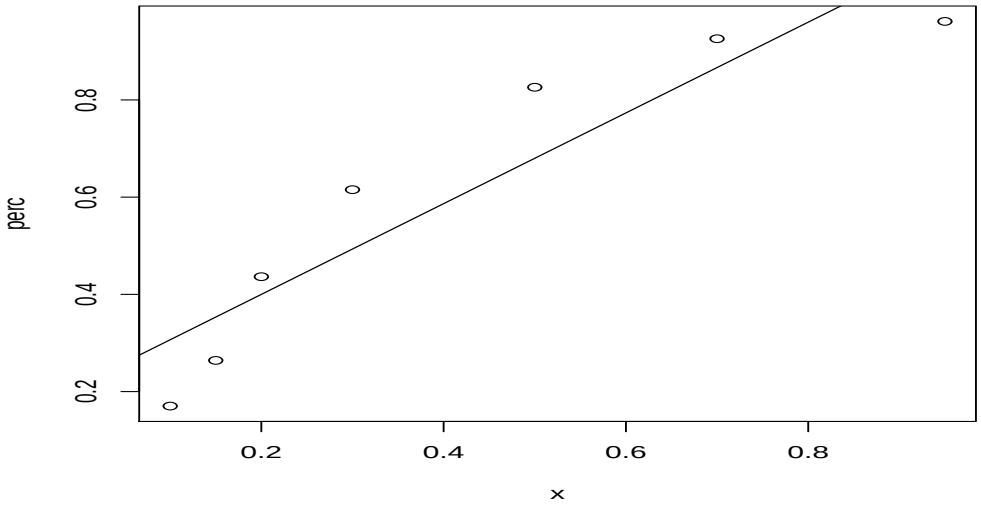

Figure 1: Plot of concentration versus percentage of animals killed and fitted line.

Fissures experiment

Nelson (1982) discusses an experiment to determine the relationship between the time in use and the number of fissures in turbines

```
      hours fissures
 [1,]   400        0
 [2,]  1000      212
 [3,]  1400       66
 [4,]  1800      511
 [5,]  2200      150
 [6,]  2600      351
 [7,]  3000      378
 [8,]  3400       78
 [9,]  3800      748
[10,]  4200      840
[11,]  4600      756
```

The data follow now a Poisson distribution, hence the mean equals the variance and we have the constraint that fitted values must be positive, fitting `lm(fissures ~ hours)` we get that the fitted value for the first observation is negative

```
 lm(fissures~hours)$fitted
       1         2         3         4         5         6         7         8
-1.47929 101.17751 169.61538 238.05325 306.49112 374.92899 443.36686 511.80473
       9        10        11
580.24260 648.68047 717.11834
```

The solution in both cases is to develop a proper statistical model, this is what generalized linear models do.

2 The exponential family

An important unifying concept underlying GLM is the **exponential family of distributions**. The exponential family form originates with Fisher (1934).
Members of the exponential family of distributions all have probability density (or probability mass) functions that can be expressed in the form

$$f(y;\boldsymbol{\theta},\phi) = exp\left\{\frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi)\right\} \tag{3.1}$$

where, in each case, $a(.)$, $b(.)$ and $c(.)$ are specific functions. The parameter $\boldsymbol{\theta}$ is a *canonical location parameter* and ϕ is a *dispersion parameter*. The Binomial, Poisson and Normal distribution (among others) are members of the exponential family. The most important case is the Normal distribution, its density function is:

$$\begin{aligned} f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-[y-\mu]^2/2\sigma^2\right\} \quad \text{which we can rewrite as} \\ &= exp\left\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}\left[y^2/\sigma^2 + \ln(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Therefore, $\boldsymbol{\theta} = \mu$, $b(\boldsymbol{\theta}) = \mu^2/2$, $a(\phi) = \phi$, $\phi = \sigma^2$ and

$$c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$$

Exercise 1

Find $\boldsymbol{\theta}$, ϕ, $a(.)$, $b(.)$ and $c(.)$ for the Poisson, binomial, exponential and gamma distributions.

2.1 The exponential family form and maximum likelihood

In order to obtain estimates of the unknown vector of parameters $\boldsymbol{\theta}$, given data values, we can employ maximum likelihood. It is more convenient to work with the ln of the likelihood

function, and this that not change any of the resulting parameter estimates. Using (3.1), the log-likelihood function for an exponential family distribution is:

$$l(\boldsymbol{\theta}, \phi|\boldsymbol{y}) = \ln\left(exp\left\{\frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi)\right\}\right) = \frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi) \tag{3.2}$$

The score function is the first derivative of the log-likelihood function with respect to the parameters of interest. For simplicity we will treat ϕ as a *nuisance parameter* (not of primary interest) and take derivatives with respect to $\boldsymbol{\theta}$, the resulting score function is

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{y - \frac{\partial}{\partial \boldsymbol{\theta}} b(\theta)}{a(\phi)} = \frac{y - b'(\boldsymbol{\theta})}{a(\phi)} \tag{3.3}$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = -\frac{b''(\boldsymbol{\theta})}{a(\phi)} \tag{3.4}$$

Exercise 2

Using the following results (see for example Rice (1995, chap. 8))

$$E\left(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = 0 \quad E\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right) + E\left(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2 = 0$$

show that

$$E(y) = b'(\boldsymbol{\theta}) \quad Var(y) = b''(\boldsymbol{\theta})a(\phi) \tag{3.5}$$

3 Components of a generalized linear model

Let start with the standard linear regression model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}) \quad E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$$

where $\boldsymbol{X\beta}$ is a linear combination of predictor variables called *linear predictor* (which is represented as $\boldsymbol{\eta}$), in this case the mean $\boldsymbol{\mu}$ is directly *linked* to the linear predictor, i.e., in this case $\boldsymbol{\mu} = \boldsymbol{\eta}$. From this simple model we can see easily two components of the model: the probability distribution of the response and the linear structure. The generalization of the linear model has the following components:

1. **Random component**: $\boldsymbol{y}$ is a vector of random components i.i.d. according to a specific exponential family distribution with mean $\boldsymbol{\mu}$.

2. **Systematic component**: is the linear predictor $\boldsymbol{\eta} = \boldsymbol{X\beta}$. This describes how the location of the response distribution changes with the explanatory variables.

3. **Link function**: It is a monotonic differentiable function which links the mean and the linear predictor

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) \quad E(\boldsymbol{y}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) \tag{3.6}$$

In the case of a linear regression model $\boldsymbol{\mu} = \boldsymbol{\eta}$, and thus the link function is the identity link.

There are many choices of link function. The **canonical link** function is a function which transforms the mean to the canonical location parameter $\boldsymbol{\theta}$

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\theta} \Rightarrow g \quad \text{is a canonical link}$$

Using the results from Exercise 1, the following table shows the canonical links for the most common distributions used in GLMs:

Distribution	Canonical Link
Normal	$\boldsymbol{\eta} = \boldsymbol{\mu}$ (identity link)
Binomial	$\boldsymbol{\eta} = \ln\left(\frac{P}{1-P}\right)$ (logistic link)
Poisson	$\boldsymbol{\eta} = \ln(\boldsymbol{\mu})$ (log link)
Exponential	$\boldsymbol{\eta} = \frac{1}{\boldsymbol{\mu}}$ (reciprocal link)
Gamma	$\boldsymbol{\eta} = \frac{1}{\boldsymbol{\mu}}$ (reciprocal link)

Table 2: Canonical links for the generalized linear model

We can view the selection of the link function as similar to the choice of a transformation on the response. However, the link function is a transformation on the *population mean*, not the data. More details on link functions can be found in McCullagh and Nelder (1989, chap. 2)

4 Estimating Generalized Linear Models

We saw in section 2 how to estimate $\boldsymbol{\theta}$ using maximum likelihood. However, this is not useful in practice, because, $\boldsymbol{\theta}$ will depend on the exploratory variables in general, and when using a canonical link, $\boldsymbol{\theta} = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, therefore $\boldsymbol{\beta}$ are now our parameters of interest.

4.1 General case

Given a vector of observations $\boldsymbol{y} = (y_i, y_2, \ldots, y_n)'$. The log likelihood is

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} ((y_i\theta_i - b(\theta_i)/a(\phi) + c(y_i, \phi)) \tag{3.7}$$

Therefore the score function

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \tag{3.8}$$

$$= \sum_{i=1}^{n} \frac{(y_i - b'(\theta_i))}{a(\phi)}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \tag{3.9}$$

but $\eta_i = g(\mu_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$ and because of (3.5) we have that

$$\begin{aligned} g(b'(\theta_i)) &= \boldsymbol{\beta}'\boldsymbol{x}_i \\ g'(b'(\theta_i))b''(\theta_i)\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} &= \boldsymbol{x}_i \\ \text{hence } g'(\mu_i)b''(\theta_i)\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} &= \boldsymbol{x}_i \\ \text{and so } \frac{\partial l}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{a(\phi)g'(\mu_i)b''(\theta_i)}\boldsymbol{x}_i \\ \text{i.e. } \frac{\partial l}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{g'(\mu_i)V_i}\boldsymbol{x}_i \end{aligned}$$

where $V_i = Var(y_i) = a(\phi)b''(\theta_i)$ as seen in (3.5).

$\hat{\boldsymbol{\beta}}$ is found as the solution of $\frac{\partial l}{\partial\boldsymbol{\beta}} = 0$. In general, this set of equations needs to be solved iteratively, if we use the Newton-Raphson algorithm

$$\left.\frac{\partial l}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_1} \approx \left.\frac{\partial l}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} + \left.\frac{\partial^2 l}{\partial\boldsymbol{\beta}\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) \tag{3.10}$$

$$0 = \left.\frac{\partial l}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} + \left.\frac{\partial^2 l}{\partial\boldsymbol{\beta}\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) \tag{3.11}$$

find $\hat{\boldsymbol{\beta}}_1$ from $\boldsymbol{\beta}_0$ and so on; this process gives $\boldsymbol{\beta}_\gamma \to \hat{\boldsymbol{\beta}}$.

In practice, the value of the second derivative of the score function is replaced by its expected value, this is called **Iteratively re-weighted least squares (IRLS)**. Using

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}'}\right) = -E\left(\frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^2$$

we find that

$$\begin{aligned} E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}'}\right) &= -E\left(\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{(g'(\mu_i)V_i)^2}\boldsymbol{x}_i\boldsymbol{x}_i'\right) \\ &= -\sum_{i=1}^{n}\frac{V_i}{(g'(\mu_i))^2 V_i^2}\boldsymbol{x}_i\boldsymbol{x}_i' \\ &= -\sum_{i=1}^{n} w_i\boldsymbol{x}_i\boldsymbol{x}_i' \end{aligned}$$

where $w_i = 1/V_i\,(g'(\mu_i))^2$. we can write the expression above in matrix form:

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}'}\right) = -\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

where $\boldsymbol{W}$ is a diagonal matrix with diagonal elements w_i. Hence, we can say that if $\hat{\boldsymbol{\beta}}$ is a solution of $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, then $\hat{\boldsymbol{\beta}}$ is asymptotically Normal with mean $\boldsymbol{\beta}$ and covariance matrix $(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$.

Equation (3.11) is now,

$$\begin{aligned}
\boldsymbol{\beta}_{new} &= \boldsymbol{\beta}_{old} - \left(E\left[\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right]\right)^{-1}_{\boldsymbol{\beta}_{old}} \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_{old}} \\
&= \boldsymbol{\beta}_{old} + (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old}) \\
&= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z}
\end{aligned}$$

where $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta}_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old}) = \boldsymbol{\eta}_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old})$, and $\boldsymbol{z}$ is called *working vector*.

Therefore, IRLS based on the Newton-Raphson method can be summarized as follows

1. Obtain an initial value of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}_{old}$.
2. Use $\hat{\boldsymbol{\beta}}_{old}$ to estimate $\boldsymbol{W}$ and $\boldsymbol{\mu}_{old}$.
3. Let $\hat{\boldsymbol{\eta}}_{old} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{old}$. Get $\boldsymbol{z}_{new}$.
4. Obtain a new estimate $\hat{\boldsymbol{\beta}}_{new}$, and repeat steps 2 to 5 until convergence.

Canonical link case

In general, in glms the value of the second derivative of the log-likelihood and its expectation may be different. However, if the canonical link is used, both matrices are the same, and $w_i = 1/g'(\mu_i)$. The proof of that result is left to the reader (**Exercise 3**).
Hint: Use the results on top of page 8 to prove that $g'(\mu_i)b''(\theta_i) = 1$, then then use this results after taking derivatives in equation (3.9)

What happens when the assumptions are not satisfied?

There are a vast number of modelling problems where: 1) responses are correlated, and 2) responses are independent but do not belong to an exponential family. Wedderburn (1974) developed the notion of quasi-likelihood which exploits the fact that the score function involves the distribution of the response only through the first two moments, i.e., to define a quasi-likelihood function we need only specify a relation between the mean and the variance of the observations.
For a one parameter exponential family, the log-likelihood is equal to the quasi-likelihood.

4.2 Estimation of the dispersion parameter

With the exception of the Binomial and Poisson distributions, the dispersion parameter is not necessarily known, and it will have to be estimated. Note that for the estimation of $\boldsymbol{\beta}$ lacking knowledge of ϕ is not a problem, since the likelihood score equations are independent of ϕ. In

the case of the Normal distribution, the dispersion parameter $\phi = \sigma^2$ is estimated so that the scaled residual deviance equals the degrees of freedom, i.e., $\hat{\sigma}^2 = RSS/(n-p) = D/(n-p)$, and so $d/\hat{\phi} = n - p$. The extension of this to GLMs would be to estimate ϕ as the mean squared Pearson residual

$$\hat{\phi} = \frac{\sum_{i=1}^{n} r_{iP}^2}{n-p}$$

4.3 Example 1

We will use the simple linear regression through the origin as example. Therefore

$$y_i \sim N(\underbrace{\beta x_i}_{\mu_i}, \sigma^2)$$

In section 2 we saw that the link function is the identity, and $\theta_i = \mu_i = x_i\beta$, $b(\theta) = \mu^2/2$, and $\phi = \sigma^2$. Hence, $V_i = \sigma^2$, $g(\mu) = \mu \Rightarrow g'(\mu) = 1$, and $w_i = 1/V_i(g'(\mu))^2 = 1/\sigma^2$. Then,

$$\begin{aligned}
\frac{dl}{d\beta} &= \sum_{i=1}^{n} \frac{y_i - \mu_i)x_i}{g'(\mu_i)V_i} = \sum_{i=1}^{n} \frac{(y_i - x_i\beta)x_i}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{x}'(\boldsymbol{y} - \boldsymbol{x}\beta) \\
E\left(\frac{d^2l}{d\beta^2}\right) &= -\sum_{i=1}^{n} w_i x_i^2 = -\sum_{i=1}^{n} x_i^2/\sigma^2 = -\boldsymbol{x}'\boldsymbol{x}/\sigma^2 \\
\boldsymbol{z} &= \boldsymbol{x}\beta_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old}) = \boldsymbol{x}\beta_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old}) = \boldsymbol{x}\beta_{old} + (\boldsymbol{y} - \boldsymbol{x}\beta_{old}) = \boldsymbol{y}.
\end{aligned}$$

Therefore, only one iteration is necessary and

$$\hat{\beta} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}\boldsymbol{y}$$

which is the least squares estimate of β.

Exercise 4: The second-hand store

The data set `lumber.txt` contains data on number of customers visiting a store in a certain region and 5 independent variables: number of housing units in the region, average household income, average housing unit age in region, distance to the nearest competitor and distance to store in miles.
The objective is to compare the fitting of the glm() function in R :

```
fit<-glm(Customers~Housing+Income+Age+Competitor+Store,family=poisson)
summary(fit)
```

with the *homemade* fitting using IRLS directly, and check the importance of the starting values on the number of iterations till convergence.

5 Inference

Returning to our original glm model, with log-likelihood given in (3.7), with $E(\boldsymbol{y}) = \boldsymbol{\mu}$, $g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$. There are several ways of testing hypothesis about the components of $\boldsymbol{\beta}$.

1. If we wish to test, e.g., $\beta_1 = 0$ (the first component of $\boldsymbol{\beta}$), the procedure is to find $\hat{\beta}_1$ and $se(\hat{\beta}_1)$ and refer $|\hat{\beta}_1|/se(\hat{\beta}_1)$ to $N(0,1)$, and reject $\beta_1 = 0$ if this is too large. Remember that $se(\hat{\beta}_1)$ is obtained as the squared root of the $(1,1)$ element of the inverse of the matrix

$$\left.\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}}$$

So here we are using the asymptotic normality of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ and the formula of its asymptotic variance. Therefore we need a large sample!.

2. If we have a small sample, the normal approximation may not work well. An alternative is to follow the *Analysis of Variance* method. The basis of this is to have a measure of the model fit which measures the discrepancy between the model and the data. For Normal data this is the *Residual sum of squares*, for non-Normal data this measure is called the **Deviance**. In general, the deviance is based on the value of the (maximized) log-likelihood. The test is then based on the reduction of this measure of fit. A good approximation to the distribution of this reduction is the Chi-squared distribution. Thus, we can test the significance of parameters using a Chi-squared test.
In general, suppose that we want to test $\boldsymbol{\beta} \subset \omega_r$ (the "reduced" model, e.g. $\boldsymbol{\beta}_1 = 0$) against $\boldsymbol{\beta} \subset \omega_f$ (the "full" model) where $\omega_r \subset \omega_f$. Consider the difference in the maximum of the log-likelihood for the full and the reduced model. Let L_f denote the maximized value of the log-likelihood under the full model (with β_1 present) and L_r denote the maximized value of the log-likelihood under the reduced model (with $\beta_1 = 0$). Note that L_f is at least as large as L_r (why?). The test statistic is:

$$S(\omega_r, \omega_f) = -2(L_r - L_f) = 2\sum_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right) \right] / \phi$$

where $\tilde{\theta}_i$=mle under ω_r and $\hat{\theta}_i$=mle under ω_f.

The distribution of $S(\omega_r, \omega_f)$ is approximated by a χ^2 with degrees of freedom equal to the difference between the number of parameter in the full and the reduced model (if we are testing $\beta_1 = 0$, then, χ^2_1). Confidence intervals for the parameters are derived from the χ^2 distribution.

Scaled Deviance: A saturated model is one that perfectly fits the data (i.e. there are as many parameters as observations, and the fitted values are equal to the observed values). Let l_s be the log-likelihood of the saturated model and let l_m be the maximized value of the log-likelihood of the model of interest. The *Scaled deviance* of the model of interest is

$$ScaledDeviance = S(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2(L(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - L(\boldsymbol{y}, \phi, \boldsymbol{y}) = -2(l_m - l_s)$$

and the deviance is

$$Deviance = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = S(\boldsymbol{y}, \hat{\boldsymbol{\mu}})\phi$$

Therefore, the scaled deviance can be seen as the test statistic for testing the hypothesis that all parameters in the saturated model, which are not in the model of interest, are equal to zero. In the case of Binomial or Poisson data, $\phi = 1$, and so both deviances are the same. In the case of Normal data $\phi = \sigma^2$, then, the Scaled deviance follows and F distribution.

Exercise 5:

Show that the deviance for data $y_i \sim N(\mu_i, \sigma^2)$ is $\sum_{i=1}^{n}(y_i - \mu_i)^2$

6 Diagnostics for GLMs

Diagnostics for the GLM are mostly linear regression diagnostics based on the last iteration of the IRLS algorithm.

6.1 Residuals

In ordinary linear regression, the residuals $y_i - \hat{\mu}_i$ are used to detect the violation of the assumptions, such as the non-homogeneous variance, etc. In GLMs there are three types of residuals

1. **Response residuals:** $y_i - \hat{\mu}_i$, they are not appropriate since $Var(y_i)$ is not constant.

2. **Pearson residuals:**
$$r_{i,P} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var(y_i)}}}$$
They have constant variance and mean zero. Mostly useful for detecting variance misspecification.

3. **Deviance residuals:**
$$sign(y_i - \hat{\mu}_i)\sqrt{d_i^2}$$
where d_i is the contribution to the model deviance of the ith observation. These residuals have the property that they carry the same sign as $y_i - \hat{\mu}_i$, and sum of they squares is the deviance.
For many models the deviance residuals are closer to a Normal distribution than the Pearson residuals, and so they are more appropriate for constructing diagnostic plots.

4. **Standardized residuals:** Both the Pearson and the deviance residuals can be variance-standardized and corrected for the effects of leverage by dividing them by $\sqrt{\phi(1 - h_{ii})}$, in most cases $\phi = 1$, and when it is not, it is replaced by an estimate. These residuals should be approximately $N(0, 1)$ for Poisson and binomial models with large counts and should lie within -2 and +2. McCullagh and Nelder (1989, chap. 12) recommend doing several checks:

(a) **Informal checks using residuals:** Plot standardized deviance residuals against the linear predictor $\hat{\boldsymbol{\eta}}$ or the fitted values $\hat{\boldsymbol{\mu}}$ transformed to the constant information scale of the error distribution, this plot should have no pattern. Typical deviations are:

- Appearance of curvature in the mean
- Systematic change of the range with the fitted values

The curvature may arise because of: bad choice of the link function, wrong choice of the scale or covariates.

(b) **Checking the variance function:** Plot absolute values of residuals against fitted values, an ill chosen variance function will result in a trend in the mean. For example, we could choose the variance to be a linear function of the mean, but it might be a quadratic function instead.

(c) **Checking the link function:** A plot of the working vector $\boldsymbol{z}$ described in page 9 against the linear predictor $\boldsymbol{\eta}$ gives an informal check for the link function. If the link function is correct this pattern is a straight line (why?).

(d) **Measure of leverage:** In ordinary regression we used the diagonal elements of the hat matrix as a measure of leverage. The hat matrix is obtained from the last iteration of IRLS. What is the form of this matrix?

(e) **Measure of influence:** The equivalent of the Cook's distance for a GLM is

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\phi}}$$

Bibliography

Fisher, R. (1934). Thow new properties of mathematical likelihood. *Proceedings of the Royal Statistical Society of London, A*, 144:285–307.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

Myers, R., Montgomery, D., and Vining, G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences*. John Wiley & sons, New York.

Nelder, J. and Wedderburn, R. (1972). Generalized linera models. *Journal of the Royal Statistical Society, Series A*, 135:370–385.

Nelson, W. (1982). *Applied Life Data Analysis*. John Wiley, New York.

Rice, J. (1995). *Mathematical statistics and data analysis*. Duxbury Press.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models and the gauss newton method. *Biometrika*, 61:439–447.