## 3.3 Inference

As in the inference section for linear models, section 2.3, we discuss how to construct confidence intervals and perform hypothesis tests. In particular, for hypothesis testing we shall discuss how to compare two related models. For GLMs, we say that two models are related if

1. the distribution of the response $\boldsymbol{Y}$ is the same; and
2. the same link functions is used.

The models differ in the number of parameters i.e. the dimensionality of $\boldsymbol{\beta}$. To compare models we require a measure of their **goodness of fit**. We shall present goodness of fit statistics based on the log-likelihood function.

### 3.3.1 Sampling Distributions for GLMs

In this section, we discuss the sampling distributions[†] relevant to GLMs. We shall use the following notion: under appropriate regularity conditions, which are satisfied for generalized linear models, if $S$ is a statistic of interest, then

$$\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{var}(S)}} \dot{\sim} N(0,1)$$

or equivalently

$$\frac{(S - \mathrm{E}(S))^2}{\mathrm{var}(S)} \dot{\sim} \chi^2_1,$$

where we shall use $\dot{\sim}$ to denote "approximately distributed as".

If there is a vector of statistics of interest

$$\boldsymbol{S} = \begin{pmatrix} S_1 \\ \vdots \\ S_p \end{pmatrix}$$

with asymptotic expectation $\mathrm{E}(\boldsymbol{S})$ and asymptotic variance-covariance matrix $Q$, then asymptotically

$$\boldsymbol{S} - \mathrm{E}(\boldsymbol{S}) \dot{\sim} N(0, Q) \tag{3.14}$$

and further, asymptotically

$$(\boldsymbol{S} - \mathrm{E}(\boldsymbol{S}))^T Q^{-1} (\boldsymbol{S} - \mathrm{E}(\boldsymbol{S})) \dot{\sim} \chi^2_p \tag{3.15}$$

provided $Q$ is non-singular so that $Q^{-1}$ exists and is unique.

[†]The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample

To be clear, these approximations follow from the central limit theorems:

**Central Limit Theorem**

Suppose $A_1, \ldots, A_n$ are iid random variables with $\mathrm{E}(A_i) = \mu$ and $\mathrm{var}(A_i) = \sigma^2$ (both finite). Then

$$\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} A_i - \mu}{\sqrt{\sigma^2}}\right) \xrightarrow{\mathrm{d}} N(0,1) \quad \text{as } n \to \infty$$

where $\xrightarrow{\mathrm{d}}$ means convergence in distribution.

**Multivariate Central Limit Theorem**

If $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ are iid random vectors with $\mathrm{E}(\boldsymbol{A}_i) = \boldsymbol{\mu} \in \mathbb{R}^p$ and finite, positive definite, symmetric covariance matrix $Q$ then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i - \boldsymbol{\mu}\right) \xrightarrow{\mathrm{d}} N(0, Q) \quad \text{as } n \to \infty$$

The asymptotic chi-squared result (3.15) follows from the fact: If $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, Q)$ where $\mu \in \mathbb{R}^p$ and $Q \in \mathbb{R}^{p \times p}$ is a positive definite, symmetric matrix, then

$$(\boldsymbol{Z} - \boldsymbol{\mu})^T Q^{-1} (\boldsymbol{Z} - \boldsymbol{\mu}) \sim \chi^2_p.$$

**Sampling distribution for the score**

We can apply the ideas above to the score vector $\boldsymbol{U}$. Recall from (3.6) that

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right], \quad j = 1, \ldots, p.$$

Since $E(\boldsymbol{U}) = \boldsymbol{0}$ and $\mathrm{cov}(\boldsymbol{U}) = \mathcal{J}$ we have, from (3.14) asymptotically

$$\boldsymbol{U} \dot{\sim} N(\boldsymbol{0}, \mathcal{J}) \tag{3.16}$$

and, from (3.15), asymptotically,

$$\boldsymbol{U}^T \mathcal{J}^{-1} \boldsymbol{U} \dot{\sim} \chi^2_p \tag{3.17}$$

**Example** Suppose $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu, \sigma^2)$ where $\mu$ is the parameter of interest and $\sigma^2 > 0$ is a known constant. The log-likelihood function is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) =$$

Then the score is

$$U =$$

and the Fisher's information matrix is

$$\mathcal{J} =$$

Then by rearranging (3.16), we get (asymptotically)

$$\overline{Y} \sim N(\mu, \sigma^2/n).$$

We can use this result to construct a confidence interval for $\mu$. For example, a 95% confidence interval for $\mu$ is $\overline{y} \pm 1.96\sigma/\sqrt{n}$ approximately due to asymptotic normality[‡]. ■

**Sampling distribution for MLEs**

Let $\widehat{\boldsymbol{\beta}}$ be the MLE for $\boldsymbol{\beta}$ in a GLM i.e.

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\text{argmax}}\, \ell(\boldsymbol{\beta})$$

and so $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = \frac{\partial \ell(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \mathbf{0}$.

Now consider the 1st order Taylor expansion of $\boldsymbol{U}(\boldsymbol{\beta})$ about the MLE $\widehat{\boldsymbol{\beta}}$:

$$\boldsymbol{U}(\boldsymbol{\beta}) \approx$$

where we approximated $\boldsymbol{U}'$ by $\mathrm{E}(\boldsymbol{U}') = -\mathcal{J}$. Consequently, we find

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left\{\mathcal{J}(\widehat{\boldsymbol{\beta}})\right\}^{-1} \boldsymbol{U}(\boldsymbol{\beta}),$$

assuming that $\mathcal{J}$ is invertible. If $\mathcal{J}$ is regarded as a constant then $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. It turns out that this is at least asymptotically true.

The variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right] =$$

[‡]Actually, this is exact and not approximate as the observations are Normal in this example. This coincides with the results presented in Chapter 2

where we have used that $\mathcal{J} = \mathcal{J}^T$ and treated $\mathcal{J}$ as a constant.

Thus, by (3.14), we have

$$\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \mathcal{J}^{-1}). \tag{3.18}$$

and also, using (3.15), we have

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathcal{J}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \dot{\sim} \chi^2_p. \tag{3.19}$$

This is called the **Wald statistic**.

We can use (3.18) to construct confidence intervals and conduct hypothesis tests involving the components of $\boldsymbol{\beta}$. For instance, from (3.18), we have

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\text{var}(\widehat{\beta}_1)}} \dot{\sim} N(0,1), \quad \text{where} \quad \text{var}(\widehat{\beta}_1) = \underbrace{(X^T W X)^{-1}_{11}}_{\text{the 1,1 entry of } (X^T W X)^{-1}}$$

which we can use to construct confidence intervals for $\beta_1$. Further, under the null hypothesis $H_0 : \beta_1 = 0$ we have $\frac{\widehat{\beta}_1}{\sqrt{\text{var}(\widehat{\beta}_1)}} \dot{\sim} N(0,1)$.

## 3.4 Prediction

For given covariates $\boldsymbol{x}_\star$, we have[§] $\widehat{\eta}_\star = \boldsymbol{x}_\star^T \widehat{\boldsymbol{\beta}}$ with variance $\boldsymbol{x}_\star^T (X^T W X)^{-1} \boldsymbol{x}_\star$. Therefore, an approximate confidence interval can be constructed using the normal distribution. To obtain an approximate confidence interval in terms of $\mu$ we can use the inverse of the link function to transform the end points. As with the linear models, we can use the `predict` function in `R` for GLMs. More precisely, an approximate 95% confidence interval for the mean response of a prediction with covariates $\boldsymbol{x}_\star$ is

$$\left( g^{-1}\left( \widehat{\eta}_\star - 1.96 \sqrt{\boldsymbol{x}_\star^T (X^T W X)^{-1} \boldsymbol{x}_\star} \right), g^{-1}\left( \widehat{\eta}_\star + 1.96 \sqrt{\boldsymbol{x}_\star^T (X^T W X)^{-1} \boldsymbol{x}_\star} \right) \right),$$

where $g^{-1}$ is the inverse of the link function $g$.

[§] $\boldsymbol{x}_\star$ is a column vector containing covariates and maybe an intercept. Also, note that $\widehat{\eta}_\star$ is just a number.

### 3.4.1 Measuring the Goodness of Fit

One measure of the goodness of fit of a GLM is to compare it with a more general model with the maximum number of parameters that can be estimated. This model is called the saturated model. It is a GLM with exactly the same response distribution and link function as the model of interest.

For $n$ observations, $y_1, \ldots, y_n$, all with potentially different values for the linear predictors $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, a saturated model can be specified with $n$ parameters. It turns out that the saturated model forces $\mu_i \equiv \mathrm{E}(Y_i) = y_i$ for $i = 1, \ldots, n$ which achieves the maximum attainable log-likelihood.

For $n$ observations from a GLM with $n$ parameters, the log-likelihood is

$$\ell(\boldsymbol{\theta}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \left( \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

Then, for $i = 1, \ldots, n$,

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$

For this it is clear that log-likelihood is maximised if we force $\mu_i = y_i$ for $i = 1, \ldots, n$.

Now reconsider the log-likelihood in terms of the mean response $\boldsymbol{\mu}$:

$$\ell(\boldsymbol{\mu}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \left( \frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{a(\phi)} + c(y_i, \phi) \right).$$

Thus, the maximum achievable log-likelihood is $\ell(\boldsymbol{y}, \phi; \boldsymbol{y})$.

For the model of interest, a GLM with typically less parameters than observations ($p < n$), we can maximise the log-likelihood to get $\widehat{\boldsymbol{\beta}}$, then $\widehat{\boldsymbol{\eta}} = X\widehat{\boldsymbol{\beta}}$ and so the estimated mean response $\widehat{\boldsymbol{\mu}} = g^{-1}(\widehat{\boldsymbol{\eta}})$ — leading the maximised log-likelihood $\ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$.

**Deviance**

A measure of the goodness of fit of a GLM is the deviance.

**Definition.** The deviance for a model with estimated mean response $\widehat{\boldsymbol{\mu}}$ is defined as

$$D = 2\phi \left\{ \ell(\boldsymbol{y}, \phi; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y}) \right\},$$

and the scaled deviance is $D^* = D/\phi$.

To make things concrete: Suppose we have a GLM of interest and consider the following extreme models with the same response distribution and link function:

**Saturated Model:** The GLM with number of parameters, $p$, equal to the number of (distinct) observations. In this case, $\mu_i \equiv \mathrm{E}(Y_i)$ is equal to the observed response $y_i$, so there is no variation in the random component.

**Null Model:** The GLM with only one parameter ($p = 1$) representing a common mean response $\mu$ for all $y$s.

In practice, the null model will be too simple and the saturated model is uninformative as it just repeats the observed response. We aim for a model with a likelihood close to the likelihood of the saturated model, but with fewer parameters.

**Deviance — constrained, unconstrained likelihood?**

Consider a model where $\boldsymbol{\beta} \in \mathbb{R}^n$ and (as usual) $\boldsymbol{\eta} = X\boldsymbol{\beta}$, with $X \in \mathbb{R}^{n \times n}$. Assuming that $X$ is invertible, the model imposes no constraints on the linear predictor $\boldsymbol{\eta}$ — it can take any value on $\mathbb{R}^n$. In turn, this means $\boldsymbol{\mu}$ and also $\theta$ are unconstrained. Thus, for the saturated model ; the maximised log-likelihood is

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \ell(\boldsymbol{\mu}; \boldsymbol{y}) = \ell(\boldsymbol{y}; \boldsymbol{y}).$$

For models with $p < n$ parameters, the maximised log-likelihood is

$$\max_{\substack{\boldsymbol{\mu} \in \mathbb{R}^n \text{ s.t. } \boldsymbol{\mu} = g^{-1}(X\boldsymbol{\beta}) \\ \text{for some } \boldsymbol{\beta} \in \mathbb{R}^p}} \ell(\boldsymbol{\mu}; \boldsymbol{y}) = \ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y})$$

For instance, suppose we have a GLM with $n = 3$, $p = 1$ with $X = (1\ 1\ 1)^T$ and identity link function $g(z) = z$. Then

$$\eta_i = \beta \in \mathbb{R} \text{ for } i = 1, \ldots, n,$$

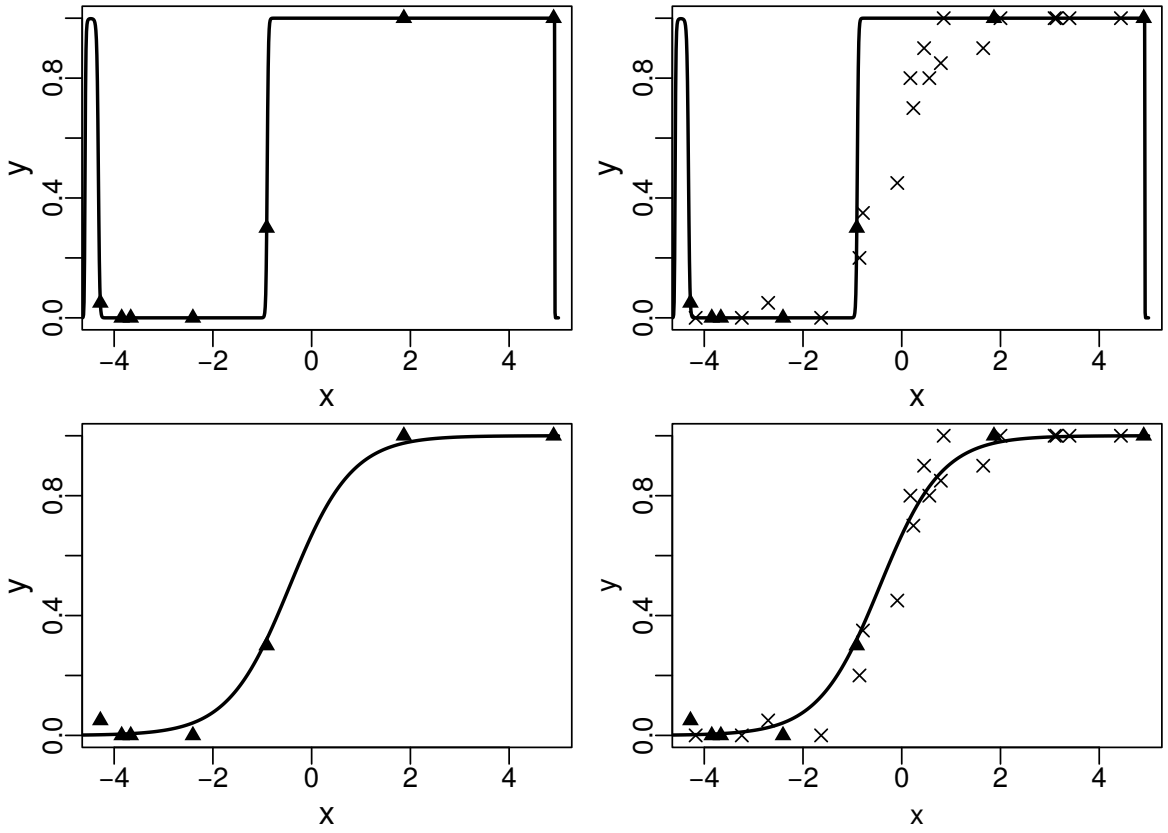
and

$$\boldsymbol{\mu} = \begin{pmatrix} g^{-1}(\beta) \\ g^{-1}(\beta) \\ g^{-1}(\beta) \end{pmatrix} = \beta \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Thus $\boldsymbol{\mu}$ is constrained to lie line in $\mathbb{R}^3$.

**Why we do not use the saturated model.**

We can make the fit as close as possible, in terms of minimising the deviance, by including sufficiently many parameters. Below are some fits for a GLM based on the 7 data points represented as triangles. The first row corresponds to a saturated model — note how the fit goes through every data point. The second row is another fit using just 2 parameters. The second column is a repeat of the first, but with additional data points — note how the saturated model would give poor predictions for the additional data points. Thus simplicity, represented by parsimony of parameters, is a desirable feature for models; we do not include parameters that are unnecessary.




Let us now switch back, considering the log-likelihood in terms of $\boldsymbol{\beta}$ — write the scaled deviance as

$$D^* = 2\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y})\right\},$$

where $\widehat{\boldsymbol{\beta}}_{sat}$ is the MLE of $\boldsymbol{\beta}_{sat}$ for the saturated model and $\widehat{\boldsymbol{\beta}}$ is still the MLE of $\boldsymbol{\beta}$ in the model of interest.

**Example** Consider the with $Y_1, \ldots, Y_n$ independent where $Y_i \sim N(\mu_i, \sigma^2)$ and

$$\mathrm{E}(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

The log-likelihood function is

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^T(\boldsymbol{y} - \boldsymbol{\mu}).$$

Setting $\boldsymbol{\mu} = \boldsymbol{y}$ yields the maximum achievable log-likelihood:

$$\ell(\boldsymbol{y}; \boldsymbol{y}) =$$

Then for any other model with $p < n$ (not a saturated model) we find the maximum likelihood estimate $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ — which in turn gives the estimated mean response $\widehat{\boldsymbol{\mu}} = g^{-1}(\widehat{\boldsymbol{\eta}}) = X\widehat{\boldsymbol{\beta}}$. So

$$\ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}) =$$

Therefore the deviance is

$$D = 2\phi \left\{ \ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}) \right\}$$

■

### 3.4.2 Pearson's $X^2$ statistic

Another important measure of the discrepancy of a GLM is Pearson's $X^2$ statistics.

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

Pearson's $X^2$ statistic shares similar asymptotic distribution properties with the deviance. However, in this course we shall focus on the deviance as a measure of goodness-of-fit as

- the deviance has a general advantage as a discrepancy measure in that it is additive for nested sets of models (see next section).
- the deviance leads to better normalised residuals.

### 3.4.3 Hypothesis Testing

We now discuss how to compare GLMs by hypothesis testing. These tests will involve the deviance and scaled deviance, therefore we first need their sampling distributions.

#### Model Comparison with known $\phi$

First, consider the 2nd order Taylor expansion of the log-likelihood around the MLE $\widehat{\boldsymbol{\beta}}$:

$$\ell(\boldsymbol{\beta}) \approx$$

where we approximated $\boldsymbol{U}'$ by $\mathrm{E}(\boldsymbol{U}') = -\mathcal{J}$. Consequently, we find

$$2\left\{\ell(\widehat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta})\right\} = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \mathcal{J}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \tag{3.20}$$

which is approximately $\chi^2_p$ distributed by (3.19). Rewriting the deviance and using (3.20) gives

$$D^* =$$

Therefore, we find that the scaled deviance

$$D^* \dot{\sim} \chi^2_{n-p}(v),$$

where $v$ is the non-centrality parameter. If the model with $p$ parameters is correct, then

$$D^* \dot{\sim} \chi^2_{n-p},$$

that is, the non-centrality parameter is approximately 0 under the null.

Now consider comparing two nested models as follows: Consider the null hypothesis

$$H_0 : \ \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix},$$

corresponding to model $M_0$ and a more general hypothesis, the alternative,

$$H_1 : \ \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

corresponding to model $M_1$ with $q < p < n$. We can test $H_0$ against $H_1$ using the difference of the scaled deviances

$$\begin{aligned} D_0^* - D_1^* &= 2\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{y})\right\} - 2\left\{\ell(\widehat{\boldsymbol{\beta}}_{sat}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y})\right\} \\ &= 2\left\{\ell(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{y})\right\}. \end{aligned}$$

The statistic $D_0^* - D_1^*$ is approximately $\chi^2_{p-q}$ distributed under the null hypothesis.

**Model Comparison with unknown $\phi$**

Under the null hypothesis we have

$$D_1^* \dot{\sim} \chi^2_{n-p} \quad \text{and} \quad D_0^* - D_1^* \dot{\sim} \chi^2_{p-q}.$$

If we consider $D_1^*$ and $D_0^* - D_1^*$ as asymptotically independent, then

$$\frac{(D_0^* - D_1^*)/(p-q)}{D_1^*/(n-p)} \dot{\sim} F_{(p-q),(n-p)}.$$

The advantage of this test statistic is that we can multiply top and bottom by $\phi$ to get a test statistic based on the deviance:

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} \dot{\sim} F_{(p-q),(n-p)}.$$

The advantage of this hypothesis test for model comparison is that it does not depend on $\phi$.

**Warning**

There are certain assumptions made when proving that the deviance, $D$, is approximately or asymptotically $\chi^2_{n-p}$ distributed which we should be wary of

- The observations are independent and distributed according to some member of the exponential family.
- The approximation relies on the number of parameters in the model staying fixed, while the sample size tends to infinity. But the saturated model has as many parameters as number of data. For instance, in the Binomial case the $\chi^2_{n-p}$ approximation (or asymptotics) for the deviance are based upon the following assumptions (McCullagh & Nelder, 1989, p. 118)
  - the observations are truly distributed independently according to the binomial distribution;
  - The approximation is based on a limiting operation in which $n$ is fixed and $n_i \to \infty$ for each $i$ (and in fact $n_i\pi_i(1-\pi_i) \to \infty$).

However, the $\chi^2$ approximation is sound when comparing two nested models, as the deviance for the saturated model cancels out.

### 3.4.4 Estimating the Dispersion

Recall that the MLE for $\boldsymbol{\beta}$ does not depend on the dispersion $\phi$. However, in cases where the dispersion is unknown, it must be estimated. There are two commonly used estimators:

First is based on the deviance:

$$\widehat{\phi}_D = \frac{D}{n-p},$$

where $D$ is the deviance of the model. This follows from the expected value of $D/\phi = D^* \dot{\sim} \chi^2_{n-p}$.

Second is

$$\widehat{\phi}_P = \frac{X^2}{n-p}$$

where $X^2$ is Pearson's statistic — this is based on the approximation $X^2/\phi \dot{\sim} \chi^2_{n-p}$.

**Note**

If $Z \sim \chi^2_d$ then $\mathrm{E}(Z) = d$; that is the expected value of a $\chi^2$ distribution is equal to its degrees of freedom.

We can then plug in an estimator for the dispersion parameter $\phi$ to estimate $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$:

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}) \approx \widehat{\phi}(X^T \widetilde{W} X)^{-1} \quad \text{(see page 60)}.$$

### 3.4.5 Akaike's Information Criteria (AIC)

An alternative statistic to compare two models was suggested by Akaike: pick whichever model minimises

$$\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\beta}}) + 2p,$$

for a given set of data. Notice, that it is similar to using the log-likelihood of the data, but with a penalisation term for the number of parameters i.e. the more parameters included, the higher the AIC. Therefore the AIC involves a trade-off between goodness-of-fit of the model and its complexity (in terms of its number of parameters).

## 3.5 Diagnostics

As with the normal linear models, we can use residuals to explore the fit of GLMs. For GLMs we require extenstions of residual definitions to accommodate for all non-Normal distributions.

### 3.5.1 Residuals

**Pearson's Residuals**

**Definition.** For a single observation $y$, Pearson's residual is defined as

$$r_p = \frac{y - \mu}{\sqrt{V(\mu)}}.$$

**Deviance Residuals**

Suppose the deviance, $D$, is used as a measure of discrepancy of a GLM. Then each observation contributes a quantity $d_i$ to $D$ so that "$D = \sum_{i=1}^{n} d_i$". Thus, it makes sense to define a deviance based residual.

**Definition.** For a single observation, $y_i$, the deviance residual is defined as

$$r_D = \text{sign}(y_i - \mu_i)\sqrt{d_i},$$

thus the deviance is $D = \sum_i r_D^2$.[¶]

**Example** For the Poisson distribution, we have

Pearson's residual: $r_p =$

Deviance residual: $r_D =$

■

Similar with the residuals for linear models, we can standardise to account for each observations leverage. To this end, we require the corresponding hat matrix for GLMs:

$$P = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}.$$

The standardised Pearson's and deviance residuals are obtained by dividing by $\sqrt{(1 - P_{ii})\widehat{\phi}}$.

[¶] $\text{sign}(x) = \begin{cases} +1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$

**Note**

These residuals are approximately normal in general. However, the deviance residuals are the closest, for $n$ large and $n_i$ (for Binomial) large. In practice, one should not expect them to lie on a straight line in a QQ plot, but rather on a smooth curve — departure from this curve may indicate an outlier.

### 3.5.2 Cook's Distance

Similar to the normal linear model case, we can examine the Cook's distance for the observations.

$$C_i = \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^T (X^T W X) \left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{p\widehat{\phi}}$$

where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the estimator calculated without using the $i$th observation. Again we look for large $C_i$ close to 1.

## 3.6 Worked Examples

### 3.6.1 Worked Example 1

`seeds-data.RData` `seeds.R`

The data presented in Table 3.2 shows the number of Orobanche seeds germinated for two genotypes and two treatments.

| # | Germinated $y$ | Total tested $n$ | Genotype $x_1$ | Treatment $x_2$ |
|---|---|---|---|---|
| 1 | 10 | 39 | 0 | 0 |
| 2 | 23 | 62 | 0 | 0 |
| 3 | 23 | 81 | 0 | 0 |
| 4 | 26 | 51 | 0 | 0 |
| 5 | 17 | 39 | 0 | 0 |
| 6 | 5 | 6 | 0 | 1 |
| 7 | 53 | 74 | 0 | 1 |
| 8 | 55 | 72 | 0 | 1 |
| 9 | 32 | 51 | 0 | 1 |
| 10 | 46 | 79 | 0 | 1 |
| 11 | 10 | 13 | 0 | 1 |
| 12 | 8 | 16 | 1 | 0 |
| 13 | 10 | 30 | 1 | 0 |
| 14 | 8 | 28 | 1 | 0 |
| 15 | 23 | 45 | 1 | 0 |
| 16 | 0 | 4 | 1 | 0 |
| 17 | 3 | 12 | 1 | 1 |
| 18 | 22 | 41 | 1 | 1 |
| 19 | 15 | 30 | 1 | 1 |
| 20 | 32 | 51 | 1 | 1 |
| 21 | 3 | 7 | 1 | 1 |

Table 3.2: Orobanche seeds Dataset

We are interested in, $y_i/n_i$, the proportion, so we want the fit to remain between 0 and 1. Suppose $Y_1, \ldots, Y_n$ are independent Binomial$(n_i, \pi_i)$ random variables and take the logit link function. Take the design matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{21,1} & x_{21,2} & x_{21,1}x_{21,2} \end{pmatrix}$$

Using the binomial distribution we would have $\mu_i \equiv \mathrm{E}(Y_i) = n_i\pi_i$ with pdf

$$\prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} \left(1 - \pi_i\right)^{n_i - y_i}.$$

Now we fit the model using the IWLS algorithm (Algorithm 3.1)

```
> beta <- c(0.5,0.5,0,0) #initial guess
> #inverse logit function
> inv.link <- function(u)
+     n*(1/(1+exp(-u)))
> #deviance function
> D <- function(p){
+     a <- y*log(y/p)
+     b <- (n-y)*log((n-y)/(n-p))
+     a[y==0] <- 0
+     2*sum(a+b)
+ }
> oldD <- D(inv.link(as.numeric(X%*%beta)))
> jj <- 0
> while(jj==0){
+     eta <- X%*%beta #estimated linear predictor
+     mu <- inv.link(eta)  #estimated mean response
+     detadmu <- n/(mu*(n-mu))
+     z <- eta+ (y-mu)*detadmu #adjusted dependent variable
+     w <- mu*(n-mu)/n #weights
+     lmod <- lm(z~x, weights=w) #regress z on x with weights w
+     beta <- as.vector(lmod$coeff) #new beta
+     newD <- D(inv.link(X%*%beta))
+     control <- abs(newD-oldD)/(abs(newD)+0.1)
+     if(control<1e-8)
+         jj <- 1
+     oldD <- newD
+ }
> beta #final estimate
```

```
[1] -0.5581717  0.1459269  1.3181819 -0.7781037
```

```
> newD #last deviance calculated
```

```
[1] 33.27779
```

Notice that this time we have used the change in the deviance as the convergence criterion; this is what R uses: If

$$\frac{|D^{\text{new}} - D^{\text{old}}|}{|D^{\text{new}}| + 0.1}$$

is less than $1 \times 10^{-8}$ then the algorithm is deemed to have converged.

By (3.13), the standard errors for these estimates are

```
> J <- t(X)%*%diag(as.vector(w))%*%X
> invJ <- solve(J)
> beta.sd <- sqrt(as.vector(diag(invJ)))
> beta.sd
```

```
[1] 0.1260213 0.2231657 0.1774677 0.3064330
```

The deviance residuals (3.5.1) are

```
> p <- as.vector(inv.link(X%*%beta))
> a <- y*log(y/p)
> b <- (n-y)*log((n-y)/(n-p))
> a[y==0] <- 0
> d <- sign(y-mu)*sqrt(2*(a+b))
> summary(d)
```

```
       V1
 Min.   :-2.01617
 1st Qu.:-1.24398
 Median : 0.05995
 Mean   :-0.08655
 3rd Qu.: 0.84695
 Max.   : 2.12122
```

We can test individual parameters using (3.18). The corresponding $p$-values are:

```
> z <- beta/beta.sd
> z
```

```
[1] -4.429187  0.653895  7.427729 -2.539229
```

```
> 2*(1-pnorm(abs(z),lower.tail=TRUE))
```

```
[1] 9.458885e-06 5.131795e-01 1.105782e-13 1.110970e-02
```

The AIC (section 3.4.5) is

```
> -2*sum(dbinom(y,n,as.vector(mu/n),log=TRUE)) + 2*length(beta)
```

```
[1] 117.874
```

Of course, R can do this for us:

```
> dat2 <- seeds
> y <- cbind(dat2$r, dat2$n - dat2$r)
> my.bin.glm <- glm(y ~ seed + extract + seed * extract,
+     family = binomial(link = "logit"), data = dat2)
> summary(my.bin.glm)
```

```
Call:
glm(formula = y ~ seed + extract + seed * extract, family = binomial(link = "logit"),
    data = dat2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.5582     0.1260  -4.429 9.46e-06 ***
seed           0.1459     0.2232   0.654   0.5132
extract        1.3182     0.1775   7.428 1.10e-13 ***
seed:extract  -0.7781     0.3064  -2.539   0.0111 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: 117.87

Number of Fisher Scoring iterations: 4
```

### 3.6.2 Worked Example 2

The data used in this example can be directly loaded in `R` using `infert`

Consider the following data relating to the study of abortions. Table 3.3 presents the first 6 data only — in total there are 248 cases.

| | education | age | parity | induced | case | spontaneous | stratum | pooled.stratum |
|---|---|---|---|---|---|---|---|---|
| 1 | 0-5yrs | 26 | 6 | 1 | 1 | 2 | 1 | 3 |
| 2 | 0-5yrs | 42 | 1 | 1 | 1 | 0 | 2 | 1 |
| 3 | 0-5yrs | 39 | 6 | 2 | 1 | 0 | 3 | 4 |
| 4 | 0-5yrs | 34 | 4 | 2 | 1 | 0 | 4 | 2 |
| 5 | 6-11yrs | 35 | 3 | 1 | 1 | 1 | 5 | 32 |
| 6 | 6-11yrs | 36 | 4 | 2 | 1 | 1 | 6 | 36 |

Table 3.3: Infertility Dataset

Let us fit a binomial model to see if `case`, a binary observation, can be predicted using the other measures as covariates. Note that `~.` means that we include all other data columns singly (useful for large datasets).

```
> myglm0 <- glm(case ~.,data=infert,family=binomial)
> summary(myglm0)
```

```
Call:
glm(formula = case ~ ., family = binomial, data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7975  -0.7836  -0.4599   0.8556   2.8998

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.039297   2.135797  -1.891   0.0586 .
education6-11yrs  1.320471   1.565614   0.843   0.3990
education12+ yrs  3.489701   2.965837   1.177   0.2393
age               0.078590   0.038060   2.065   0.0389 *
parity           -0.451423   0.276877  -1.630   0.1030
induced           1.435629   0.320870   4.474 7.67e-06 ***
spontaneous       2.191282   0.329069   6.659 2.76e-11 ***
stratum          -0.002842   0.014621  -0.194   0.8459
pooled.stratum   -0.078768   0.043800  -1.798   0.0721 .
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 254.53  on 239  degrees of freedom
AIC: 272.53

Number of Fisher Scoring iterations: 4
```

At first glance, the residual deviance looks reasonable for the degrees of freedom.

We can look at sequentially adding terms using (somewhat confusingly the `anova` function again):

```
> anova(myglm0,test="Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: case

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            247     316.17
education      2    0.002       245     316.17   0.99886
age            1    0.006       244     316.16   0.94012
parity         1    0.026       243     316.14   0.87088
induced        1    0.056       242     316.08   0.81372
spontaneous    1   58.284       241     257.80 2.269e-14 ***
stratum        1    0.003       240     257.79   0.95346
pooled.stratum 1    3.263       239     254.53   0.07085 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The `anova` function recognises that a GLM was fitted and produces an analysis of deviance table. The `test="Chisq"` option reports the $p$-values on the right.

Model comparisons and selection can be done automatically in R. First, consider similar models where just one parameter is dropped. This is achieved using the drop1 function:

```
> drop1(myglm0,test="Chisq")
```

```
Single term deletions

Model:
case ~ education + age + parity + induced + spontaneous + stratum +
    pooled.stratum
              Df Deviance    AIC    LRT  Pr(>Chi)
<none>             254.53 272.53
education      2   256.76 270.76  2.230   0.32791
age            1   258.85 274.85  4.315   0.03778 *
parity         1   257.26 273.26  2.728   0.09861 .
induced        1   277.42 293.42 22.890 1.716e-06 ***
spontaneous    1   316.03 332.03 61.504 4.419e-15 ***
stratum        1   254.57 270.57  0.038   0.84591
pooled.stratum 1   257.79 273.79  3.263   0.07085 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This suggests dropping the education and stratum parameters (and perhaps more) from the model.

The drop1 function has a brother function — the add1 function, which adds individual terms to the given model:

```
> add1(myglm0,~.^2,test="Chisq")
```

```
Single term additions

Model:
case ~ education + age + parity + induced + spontaneous + stratum +
    pooled.stratum
                         Df Deviance    AIC    LRT Pr(>Chi)
<none>                        254.53 272.53
education:age             2   254.51 276.51 0.0198  0.99013
```

```
education:parity            2   254.33 276.33 0.1977  0.90589
education:induced           2   247.58 269.58 6.9513  0.03094 *
education:spontaneous       2   252.28 274.28 2.2478  0.32500
education:stratum           2   254.41 276.41 0.1219  0.94089
education:pooled.stratum    2   254.01 276.01 0.5227  0.77003
age:parity                  1   254.19 274.19 0.3444  0.55728
age:induced                 1   254.20 274.20 0.3282  0.56670
age:spontaneous             1   251.39 271.39 3.1409  0.07635 .
age:stratum                 1   254.36 274.36 0.1723  0.67805
age:pooled.stratum          1   254.51 274.51 0.0205  0.88626
parity:induced              1   252.90 272.90 1.6332  0.20126
parity:spontaneous          1   252.57 272.57 1.9616  0.16134
parity:stratum              1   254.48 274.48 0.0521  0.81942
parity:pooled.stratum       1   254.42 274.42 0.1089  0.74145
induced:spontaneous         1   254.50 274.50 0.0323  0.85736
induced:stratum             1   248.77 268.77 5.7651  0.01635 *
induced:pooled.stratum      1   250.36 270.36 4.1678  0.04120 *
spontaneous:stratum         1   251.27 271.27 3.2615  0.07093 .
spontaneous:pooled.stratum  1   253.63 273.63 0.9052  0.34140
stratum:pooled.stratum      1   254.42 274.42 0.1120  0.73790
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The `~.^2` informs `R` to consider all possible two-factor interactions. The result above suggests including an `education:induced`, `induced:stratum` and `induced:pooled.stratum` terms.

The `drop1` and `add1` perform many individual $\chi^2$ tests between models — it does not select any particular model.

The `step` function will automatically search through the models for us.

```
> stepsearch <- step(myglm0,~.^2,test="Chisq")
```

... this is produce a lot of output! To summarise the step search, we can use the `anova` component:

```
> stepsearch$anova
```

```
               Step Df Deviance Resid. Df Resid. Dev      AIC
1                    NA       NA       239   254.5310 272.5310
2 + induced:stratum -1 5.765135       238   248.7658 268.7658
3       - education  2 1.749457       240   250.5153 266.5153
4 + age:spontaneous -1 4.072592       239   246.4427 264.4427
5  - pooled.stratum  1 1.888343       240   248.3310 264.3310
```

Reading from top to bottom: a `induced:stratum` parameter was added, then the `education` parameter removed, etc. The criteria to add or remove parameters is based on the AIC. The step search continues until the AIC cannot be reduced any further.

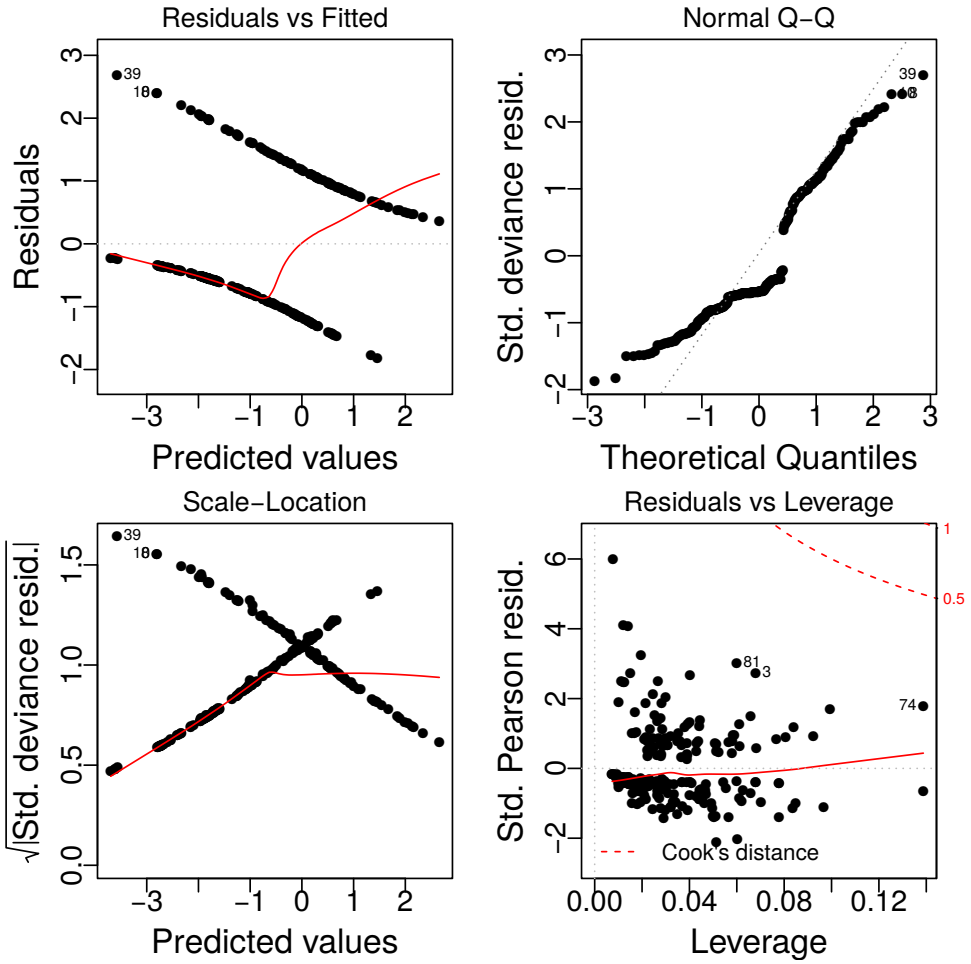
Finally, a summary oxf the final chosen model can be called (no need to fit again):

```
> summary(stepsearch)
```

```
Call:
glm(formula = case ~ age + parity + induced + spontaneous + stratum +
    induced:stratum + age:spontaneous, family = binomial, data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8253  -0.7699  -0.5257   0.8536   2.6835

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.086896   1.463758  -0.743  0.45776
age              0.000298   0.041023   0.007  0.99420
parity          -0.880102   0.193553  -4.547 5.44e-06 ***
induced          2.278841   0.487873   4.671 3.00e-06 ***
spontaneous     -0.503841   1.395070  -0.361  0.71798
stratum          0.003730   0.008352   0.447  0.65519
induced:stratum -0.025723   0.009161  -2.808  0.00498 **
age:spontaneous  0.080038   0.044768   1.788  0.07380 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 248.33  on 240  degrees of freedom
AIC: 264.33
```

```
Number of Fisher Scoring iterations: 4
```

Finally, let us inspect the diagnostic plots.




The residual plots are less informative than they were for linear models. The response contains less information than a continuous one. Nevertheless, the issue of outliers and influential observations are just as prevalent in logistic regression as for linear models — so look at the Cook's distance plot and investigate any highly influential observations.[‖]

Finally, for convenience, the residual values as follows:

```
> residuals(stepsearch,type="pearson")
> residuals(stepsearch,type="deviance")
```

and the Cook's distances:

[‖]but this is not the end of the investigation...

```
> cooks.distance(stepsearch)
```

and the standardised residuals:

```
> rstandard(stepsearch,type="pearson")
> rstandard(stepsearch,type="deviance")
```

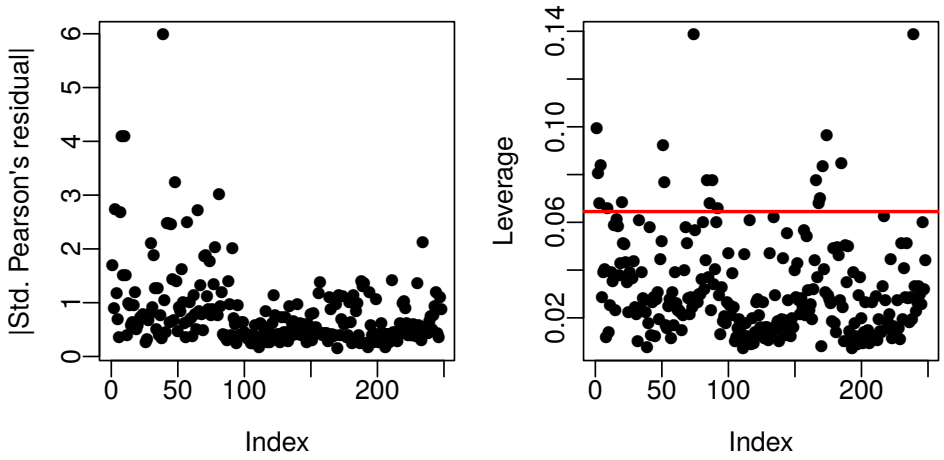
The commands above can be executed upon any `glm` model fitted in `R`.

We can proceed to check for large (in magnitude) residuals and observations with high leverages:

```
> plot(abs(rstandard(stepsearch,type="pearson")))
> plot(influence(stepsearch)$hat) #extract hat matrix values
> l.threshold <- 2*8/248 #rule of thumb
> l.threshold
```

```
[1] 0.06451613
```

```
> abline(h=l.threshold) #adds a horizontal line
```




We can mark any suspicious points and investigate if removing them significantly influences the fit.

A handy way of looking for these: First find the largest, say 5, standardised residuals:

```
> order(abs(rstandard(stepsearch,type="pearson")),decreasing=TRUE)[1:5]
```

```
[1] 39  8 10 48 81
```

These are the corresponding index numbers.

Next, find the observations 5 with the largest leverages:

```
> order(influence(stepsearch)$hat,decreasing=TRUE)[1:5]
```

```
[1]  74 239   1 174  51
```

If we see any reoccuring indices, we may want to investigate further, as they will have both high leverage and high magnitude residual.