Generalised linear models (GLMs)

In the previous chapter, you saw how linear regression models can be used to assess the effects of covariates on a response variable. However, while this is a powerful approach, it is not suitable when the response variable does not take values along the whole of the real line, e.g., when the response variable is an indicator of success or failure, a probability with range $[0, 1]$, or a count. In this chapter, we will see how to fit models that are suitable for these cases.

Binomial data

We might have an outcome that takes one of two possible values. This could be a head or a tail on a coin toss, success or failure in an experiment, or more generally, the presence or absence of an attribute of interest.

Logistic regression models

Example: Asthma data. The table below shows part of a dataset collected during a study of asthma among 41 students at Newcastle.

asthma	sex	residence	smoker	fungal
0	1	1	0	0.46591

asthma	sex	residence	smoker	fungal
1	1	1	1	0.61690
0	1	1	0	1.45138
0	1	1	0	0.50357
1	1	1	1	1.67679
0	1	1	1	-1.18897

- `asthma` is a binary variable describing whether asthma is present or not.
- `sex` is 1 for males and 2 for females.
- `residence` is 1 for halls of residence, 2 for private accommodation and 3 for a university flat.
- `smoker` is 0 for non-smokers and 1 for smokers.
- `fungal` is a continuous measurement of fungal concentration.

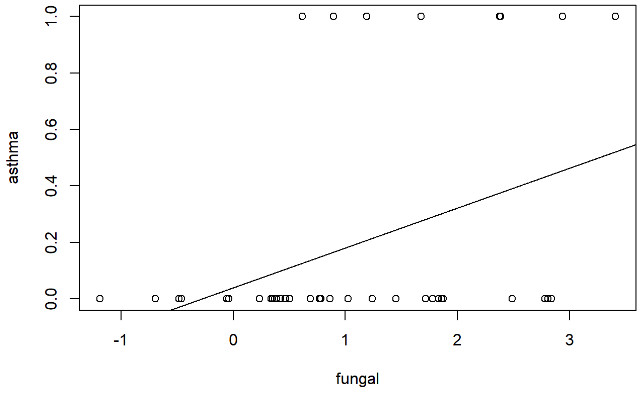
Suppose that we naively used a linear regression model to try to predict whether an individual had asthma. We might fit the model,

$$y_i = \beta_0 + \beta_1 x_4 + \varepsilon_i,\ \ i = 1, \ldots, 41.$$

```
asthma.mod<-lm(asthma~fungal,data=asthma.dat)
```

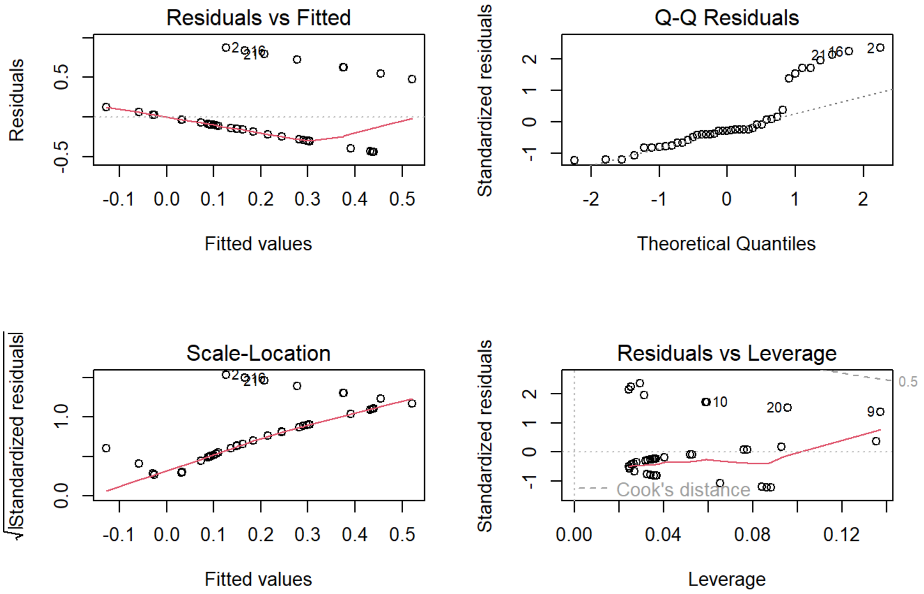
We can then plot the data and the fitted line from the regression,

```
plot(asthma.dat$fungal,asthma.dat$asthma,xlab="fun
abline(asthma.mod)
```



As always, once we have fitted a model, we should also check appropriate plots of the residuals.

```
par(mfrow=c(2,2))
plot(asthma.mod)
```



The scatterplot of the predictor against the response (i.e., `fungal` against `asthma`) with the fitted line and the residual plots both show that this model is not a good fit to the data. For example, the residuals are clearly not normal (see the "Q-Q Residuals" plot) and the fitted values cannot arise as observations (as y can only have values of 0 or 1). **A linear regression model is definitely not appropriate here!**

The logistic model for binary data

For binary data, the logistic model is an appropriate alternative to a linear regression model.

Let Y_i be a Bernoulli variable with parameter π_i. This is equivalent to

$$Y_i \sim Bin(m_i, \pi_i) \quad \text{with} \quad m_i = 1, \; i = 1, \dots, n,$$

where π_i is the probability that the i-th student has asthma. Thus

$$E(Y_i) = \pi_i \quad \text{and} \quad Var(Y_i) = \pi_i(1 - \pi_i).$$

If we were to fit a linear regression model, we would have

$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i.$$

However, the problem is that the **right-hand side** of this equation can take any value, while the **left-hand side** is a probability, which is restricted to $[0, 1]$.

The solution is to seek a function $h : \mathcal{R} \to [0, 1]$ such that

$$E(Y_i | x_i) = \pi_i = h(\beta_0 + \beta_1 x_i).$$

This is equivalent to the model

$$g(\mu_i) = g(\pi_i) = \beta_0 + \beta_1 x_i,$$

where $\mu_i = E(Y_i|x_i) = \pi_i$ and $g(\cdot) = h^{-1}(\cdot)$ is the inverse function of $h(\cdot)$. The function $g(\cdot)$ is called a **link function**.

For binomial data, the most widely used link function is

$$h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

which gives

$$g(\pi) = h^{-1}(\pi) = \log\left(\frac{\pi}{1-\pi}\right),$$

where $\eta = \beta_0 + \beta_1 x$ is known as the **linear predictor**.

Here, $g(\cdot)$ is the **logit** link function and the model is called the **logistic model**.

For the asthma data, we might fit a model where

$$Y_i \sim Bin(m_i, \pi_i) \quad \text{with} \quad m_i = 1$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i,$$

or equivalently,

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Note that while the logistic model is very widely used, other link functions are also suitable in this case and will be discussed later.

Fitting the logistic regression model

The unknown parameters (β_0, β_1) in the logistic regression model can be estimated using maximum likelihood estimation. The first step is to write down the likelihood, and since the y_i values are independent, we have

$$L = \prod_{i=1}^{n} f(y_i|\beta_0, \beta_1).$$

The log-likelihood is therefore

$$\ell = \log(L) = \sum_{i=1}^{n} \log[f(y_i|\beta_0, \beta_1)],$$

with

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

Therefore,

$$\ell = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\}.$$

After some algebraic manipulation, we find that,

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_i) - \log[1 + \exp(\beta_0 + \beta_1 x_i)]\}.$$

Taking the derivatives with respect to β_0 and β_1, we have

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left\{ y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\} = \sum_{i=1}^{n} (y_i - \pi_i),$$
$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left\{ y_i x_i - x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\} = \sum_{i=1}^{n} (y_i x_i - \pi x_i).$$

Setting these equations to zero, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions to the following equations,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \sum_{i=1}^{n} \pi_i,$$

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \sum_{i=1}^{n} \pi_i x_i.$$

We now encounter an important difference between logistic regression models and linear regression models: **there are no analytical solutions to these equations!** This means that we have to use a numerical method to obtain the maximum likelihood estimates. One option is to use the *Newton-Raphson* method, an iterative procedure for finding the root of a function. In practice, we will use statistical software that does this for us, such as through the `glm` command in `R`.

Example: Asthma data (continued). We can fit a logistic regression model to the asthma data in R as follows:

```
asthma.glm<-glm(asthma~fungal, family=binomial, da
summary(asthma.glm)

##
## Call:
## glm(formula = asthma ~ fungal, family = binomia
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|
## (Intercept)  -2.8280     0.8549  -3.308  0.0009
## fungal        0.9954     0.4393   2.266  0.0234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## (Dispersion parameter for binomial family taken
##
##     Null deviance: 40.472  on 40  degrees of fr
## Residual deviance: 34.225  on 39  degrees of fr
## AIC: 38.225
##
## Number of Fisher Scoring iterations: 5
```

Remarks:

- The R command for fitting GLMs, `glm`, has syntax similar to `lm`. However, we must add `family=binomial` so that R knows that the data follow a binomial distribution.
- The default link function for `glm` with `family=binomial` is `logit`. We only need to define the link function if another type of link function is to be used. We will do this later.

From the R output, we can see that the fitted model is

$$Y_i \sim Bin(m_i, \pi_i) \quad \text{with} \quad m_i = 1,$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -2.8280 + 0.9954x_i.$$

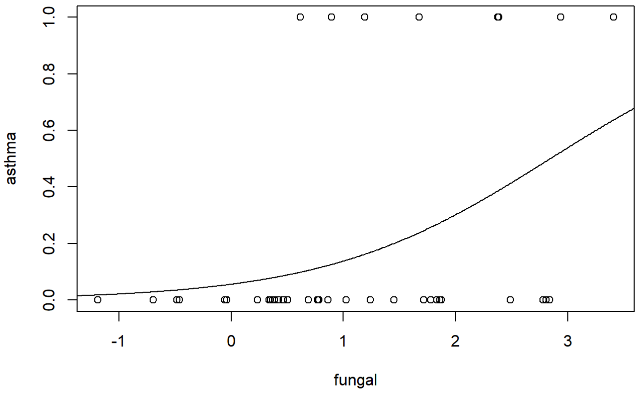
Therefore, the fitted probability of having asthma is

$$\pi_i = \frac{\exp(-2.8280 + 0.9954x_i)}{1 + \exp(-2.8280 + 0.9954x_i)}.$$

For example, when $x_i = -1$, the estimated probability of having asthma is 0.021. In contrast, when $x_i = 3$, the probability is 0.539.

The fitted model is shown in the plot below. Compare the fitted line here with the fitted line for linear regression that we saw earlier. In this case, the predicted values for `asthma` can only lie in the range [0, 1], as we require.

```
plot(asthma.dat$fungal,asthma.dat$asthma,xlab="fun
x<-seq(-2,4,0.01)
y<-predict(asthma.glm, list(fungal=x),type="respon
lines(x,y)
```



Binomial regression models

We have now seen how to model data where the response variable is a binary indicator. A related but distinct scenario arises when the response variable is binomial, i.e., counts of 'successes' or 'failures' in a number of trials.

Example: Weedkiller data. Suppose that we are testing the effectiveness of weedkiller at different dose levels. We collect 200 specimens of a weed that attacks corn fields, and divide the specimens into 5 groups of 40. The different groups receive weedkiller at different concentrations. Two weeks later, the number of weeds that have survived in each group are counted.

The data are shown in the table below.

Dose	Number of weeds	Surviving weeds
0.0028	40	35
0.0056	40	21
0.0112	40	9

Dose	Number of weeds	Surviving weeds
0.0225	40	6
0.4500	40	1

Examining the table, it appears that the proportion of weeds that survive decreases as the dose increases, as we might expect. However, we should carry out formal statistical analysis to investigate this relationship.

Let Y_i be the number of surviving weeds in each group i, π_i be the probability of survival, and x_i be the dose level.

We will assume that at each dose level, the survival of a particular weed is independent of the survival of another weed, and that they all have the same probability of survival. In other words, we assume that the number of surviving weeds follows a binomial distribution. If Y_1 is the total number surviving in the first group and π_1 is the probability of a weed surviving, then

$$Y_1 \sim Bin(40, \pi_1).$$

In this experiment, the observation is $y_1 = 35$.

In general, at each dose level x_i, we have

$$\begin{aligned} Y_i \sim Bin(40, \pi_i), \ \text{independently.} \\ E(Y_i) = \mu_i = 40\pi_i, \ i = 1, \dots, 5. \end{aligned}$$

As before, it would not be sensible to model μ_i or π_i as a linear function of x_i because π_i is a probability. Again, we need to use a link function and specify a generalised linear model,

$$g(\pi_i) = \beta_0 + \beta_1 x_i.$$

If we use the logit link function, the model is

$$\operatorname{logit}(\pi_i) \equiv \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i,$$

or equivalently,

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

A suitable model for the weedkiller data is therefore

$$Y_i \sim Bin(40, \pi_i), \quad \text{independently},$$

$$\mu_i = 40\pi = 40\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \quad \text{or}$$

$$\operatorname{logit}(\pi_i) = \beta_0 + \beta_1 x_i.$$

Example: Weedkiller data (continued). We can still use the `glm()` function in R, but we need to adapt the code slightly to fit a binomial regression model.

```
weed <- as.data.frame(cbind(
  c(0.0028,0.0056,0.0112,0.0225,0.0450),
  c(40,40,40,40,40),
  c(35,21,9,6,1)))
colnames(weed)<-c("dose","no.weed","no.survive")

logitmod <- glm(cbind(no.survive, no.weed-no.survi
family=binomial, data=weed)
summary(logitmod)

##
## Call:
## glm(formula = cbind(no.survive, no.weed - no.su
##     family = binomial, data = weed)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|
## (Intercept)  -9.1894     1.2551  -7.322 2.45e-1
## log(dose)    -1.8296     0.2545  -7.188 6.59e-1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## (Dispersion parameter for binomial family taken
##
##     Null deviance: 90.0516  on 4  degrees of fr
## Residual deviance:  2.8089  on 3  degrees of fr
## AIC: 23.577
##
## Number of Fisher Scoring iterations: 4
```

Remarks:

- When we have binomial response data, we have two pieces of information about the response values. In R, we can create a two-column matrix with the first column representing the number of *successes*, y (i.e., `no.survive` in the weedkiller example), and the second column representing the number of *failures*, $m - y$ (i.e., `no.weed-no.survive` in this example).
- `family=binomial` is also used to specify binomial data, and again the default link function is `logit`.
- `data=weed` specifies that the data are read from a table or matrix named `weed`, which is created in the first part of the code.

The fitted logistic regression model is:

$$Y_i \sim Bin(40, \pi_i) \quad \text{independently}$$

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

$$\eta_i = \beta_0 + \beta_1 x_i = -9.1894 - 1.8296 x_i,$$

where $x_i = \log(\text{dose})$.

We can then find the fitted values $\hat{\pi}_i,\ i = 1, \ldots, 5$. For the first group, which received `dose=0.0028`, we have,

$$\hat{\eta}_1 = -9.1894 - 1.8296 \times \log(0.0028) = 1.565,$$
$$\hat{\pi}_1 = \frac{\exp(1.565)}{1 + \exp(1.565)} = 0.827.$$

This means that approximately 83% of weeds survive when this weedkiller dose is used. This implies that

$$\hat{y}_1 = m_1 \hat{\pi}_1 = 40 \times 0.827 = 33.1.$$

We can carry out these calculations easily in R using the code:

```
round(predict(logitmod,type="response"),3)

##     1     2     3     4     5
## 0.827 0.574 0.275 0.096 0.029
```

Measuring goodness-of-fit

When we fitted multiple linear regression models, we could measure goodness-of-fit using measures such as R^2 and residual sum of squares. However, with GLMs we need a different approach.

General criteria

In general, the process of fitting a model to data can be regarded as replacing observed values $\boldsymbol{y} = (y_1, \dots, y_n)^T$ with fitted values $\boldsymbol{\mu}$ $= (\hat{\mu}_1, \dots, \hat{\mu}_n)^T$.

There are several possible outcomes when we do this:

- **Under-fitting**. For example, the simplest model, the *null* model, has only one parameter and fits the same value $\hat{\mu}$ to each value of y. This model is too simple to be useful.

- **Over-fitting**. At the other extreme, the *full* (saturated) model uses n parameters and gives the fit $\hat{\mu}_i = y_i$. This model is too specific to the particular data to make inference worthwhile.

- **Fitting well**. Hopefully there is a good fit somewhere between the first two cases! The objective is to find a *small* set of parameters that give good agreement between the observed and fitted values. If the discrepancies between the observed and fitted values lie within acceptable error limits, the model can be said to provide a good fit to the data.

Likelihood ratio and deviance

Likelihood ratio

Consider two models: a larger Model A with r parameters and a smaller Model B with q parameters. The smaller model is nested within the larger model, i.e., the smaller model is a special case of the larger model, with $q \leq r$.

The general theory for assessing the fit of two such nested models is based on the *likelihood ratio*. If we denote the maximum likelihood under Model A and Model B by L_A and L_B, respectively, L_A must be as large as L_B, so the likelihood ratio $L_A/L_B \geq 1$.

We will reject the smaller Model B in favour of the larger Model A if the likelihood ratio is larger than some critical value. If this is not the case, we accept the smaller model as having an adequate fit. This procedure is justified theoretically by the Neyman-Pearson lemma (which is beyond the scope of this module).

We will test the hypotheses:

$$H_0 : \text{Model } B \text{ provides a better fit} \quad \text{against} \quad H_1 : \text{Model } A \text{ pro}$$

Our test statistic is the *likelihood ratio test statistic*,

$$\Lambda = 2\log(L_A/L_B) = 2[\log(L_A) - \log(L_B)].$$

Asymptotically, this has a chi-squared distribution on $(r-q)$ degrees of freedom. We will reject H_0 in favour of H_1 if Λ exceeds the appropriate value of the $\chi^2_{(r-q)}$ distribution.

Deviance

The deviance arises as a special case when Model A is the **full** model. Applying the likelihood ratio test statistic, we have deviance D^*, where

$$D^* = 2(\ell_f - \ell_B).$$

For example, for the binomial model, the deviance is

$$D^* = 2(\ell_f - \ell_B) = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m}{m_i}\right.\right.$$

The deviance measures how close a given model is to a perfect fit, with a goodness-of-fit test based on the result:

$$D^* \sim \chi^2_{n-p-1}.$$

Example: Weedkiller data, model deviance. We can use the output from the `summary` function in R to assess the deviance of our fitted model.

```
summary(logitmod)

##
## Call:
## glm(formula = cbind(no.survive, no.weed - no.su
##     family = binomial, data = weed)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|
## (Intercept)  -9.1894     1.2551  -7.322 2.45e-1
## log(dose)    -1.8296     0.2545  -7.188 6.59e-1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## (Dispersion parameter for binomial family taken
##
##     Null deviance: 90.0516  on 4  degrees of fr
## Residual deviance:  2.8089  on 3  degrees of fr
## AIC: 23.577
##
## Number of Fisher Scoring iterations: 4
```

The `Residual deviance` is the deviance for the current model, while the `Null deviance` is the deviance for the null model that has no covariates (i.e., the model that just has an intercept term).

We have deviance $D^* = 2.8089$ and degrees of freedom $(n - p - 1)$ $= (5 - 1 - 1) = 3$. We can therefore calculate the corresponding p-value as:

$$P(\chi^2_3 > 2.8089) = 0.422.$$

This can be done easily using R code

```
round(1-pchisq(2.8089,3),3)

## [1] 0.422
```

Given that the p-value is well in excess of 0.05, we conclude that there is no statistically significant difference in fit between this model and the full model. In other words, this model fits the model as well as the full model, while using fewer variables.

For the null model, we have a deviance of 90.0516. We can calculate the p-value using the R code,

```
round(1-pchisq(90.0516,4),3)

## [1] 0
```

The p-value is (very close to) zero, so we reject the null hypothesis and conclude that the null model provides a (strongly) statistically significantly worse fit than the full model.

We can also use R to generate an **Analysis of Deviance Table**,

```
anova(logitmod, test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(no.survive, no.weed - no.surviv
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(
## NULL                          4     90.052
## log(dose)  1   87.243         3      2.809 < 2.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
```

We can see that the model that includes `log(dose)` fits the data far better than the null model.

Residuals

Our fitted values of y_i are

$$\hat{y}_i = \hat{\mu}_i = m_i\hat{\pi}_i.$$

The residuals are

$$r_i = y_i - \hat{y}_i.$$

Since $Y_i \sim Bin(m_i, \pi_i),\ Var(Y_i) = m_i\pi_i(1 - \pi_i)$. This means that the standardised (Pearson) residuals are defined by,

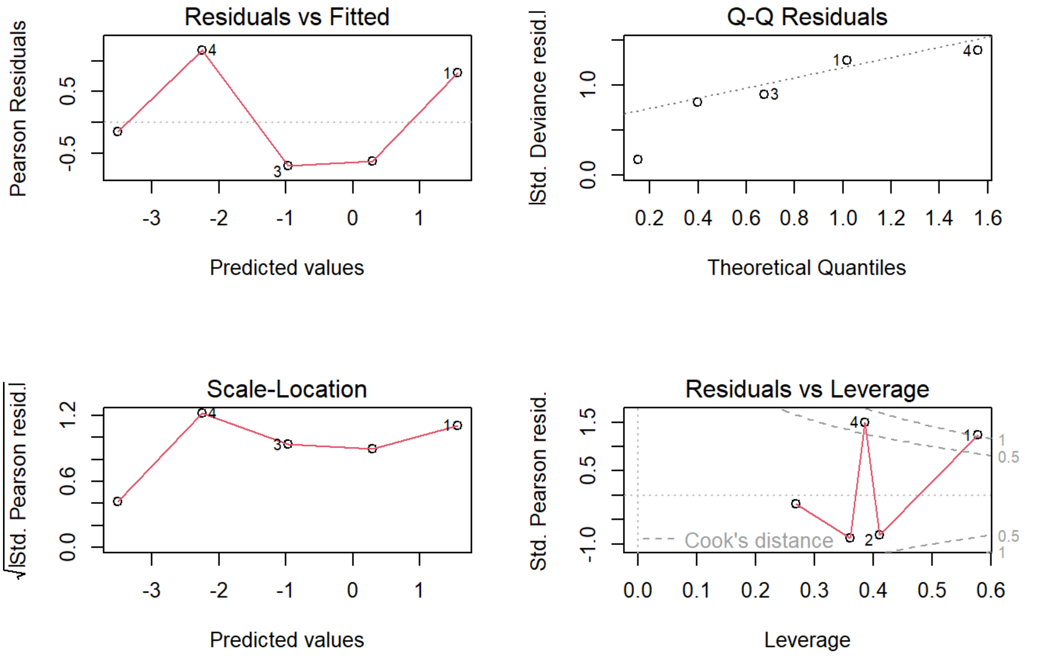
$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i (m_i - \hat{y}_i)/m_i}}.$$

For example, for the weedkiller data,

$$r_1^* = \frac{35 - 33.08}{\sqrt{33.08 \times (40 - 33.08)/40}} = 0.801.$$

If the model is correct, the Pearson residuals are independent observations from a standard normal distribution, i.e., $\mathcal{N}(0,1)$. Suitable residual plots can be produced in R as shown below.

```
par(mfrow=c(2,2))
plot(logitmod)
```



In this case, the plots give no cause for concern (e.g., there is no obvious sign of non-normality in the Q-Q plot), although we should be careful about drawing firm conclusions from a sample size this small.

Choice of link function

For binomial data, we need to use a link function $g(\pi) = \eta$ that maps $[0, 1]$ to the whole real line. While many functions are available that meet this requirement, in practice three link functions are commonly used:

- **Logit link**: $\eta = \log\{\pi/(1-\pi)\} \Leftrightarrow \pi = e^{\eta}/(1+e^{\eta})$.
- **Probit link**: $\eta = \Phi^{-1}(\pi) \Leftrightarrow \pi = \Phi(\eta)$.
- **Complementary log-log link**: $\eta = \log\{-\log(1-\pi)\} \Leftrightarrow \pi = 1 - \exp(-e^{\eta})$.

When using the `glm` function in R, the link functions can be applied using the following code:

- `family=binomial(link=logit)`
- `family=binomial(link=probit)`
- `family=binomial(link=cloglog)`

We have already fitted a model using the default `logit` link function, and can fit the equivalent models using `probit` and `cloglog`:

```
glm(cbind(no.survive, no.weed-no.survive)~log(dose
    family=binomial(link=probit), data=weed)

##
## Call:  glm(formula = cbind(no.survive, no.weed
##     family = binomial(link = probit), data = we
##
## Coefficients:
## (Intercept)    log(dose)
##      -5.277       -1.054
##
## Degrees of Freedom: 4 Total (i.e. Null);  3 Res
## Null Deviance:       90.05
## Residual Deviance: 3.193     AIC: 23.96
```

```
glm(cbind(no.survive, no.weed-no.survive)~log(dose
    family=binomial(link=cloglog), data=weed)

##
## Call:  glm(formula = cbind(no.survive, no.weed
##     family = binomial(link = cloglog), data = w
##
## Coefficients:
## (Intercept)    log(dose)
##      -7.512       -1.398
##
## Degrees of Freedom: 4 Total (i.e. Null);  3 Res
## Null Deviance:       90.05
## Residual Deviance: 1.31  AIC: 22.08
```

To choose a link function, we should fit models using each option. If a model has deviance that is clearly smaller than that for the alternative models, we should select that model. However, in this case, although the `cloglog` link has the smallest value, the differences in deviance are small and so we would probably choose the most widely used link function, which is `logit`.

Full analysis of asthma data

We now return to the asthma data. So far, we have fitted models using only one covariate, but we will now carry out a full analysis. Recall that the data have the following variables:

asthma	sex	residence	smoker	fungal
0	1	1	0	0.46591
1	1	1	1	0.61690
0	1	1	0	1.45138

asthma	sex	residence	smoker	fungal
0	1	1	0	0.50357
1	1	1	1	1.67679
0	1	1	1	-1.18897

- `asthma` is a binary variable describing whether asthma is present or not.
- `sex` is 1 for males and 2 for females.
- `residence` is 1 for halls of residence, 2 for private accommodation and 3 for a university flat.
- `smoker` is 0 for non-smokers and 1 for smokers.
- `fungal` is fungal concentration.

Since some of the covariates are categorical, we will define indicator variables to represent `sex`, `smoker` and `residence`:

$$
\begin{aligned}
g &= \begin{cases} 0, & \text{male} \\ 1, & \text{female} \end{cases} \\
s &= \begin{cases} 0, & \text{non-smoker} \\ 1, & \text{smoker} \end{cases} \\
r_1 &= \begin{cases} 1, & \text{private accommodation} \\ 0, & \text{otherwise} \end{cases} \\
r_2 &= \begin{cases} 1, & \text{university flat} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

We then define the linear predictor as

$$\eta_i = \beta_0 + \beta_1 g_i + \beta_2 r_{1i} + \beta_3 r_{2i} + \beta_4 s_i + \beta_5 x_{4i}.$$

We can fit this model in R as follows,

```
asthma.logit <- glm(asthma~factor(sex)+factor(resid
  +factor(smoker)+fungal, data=asthma.dat, family=b
summary(asthma.logit)

##
## Call:
## glm(formula = asthma ~ factor(sex) + factor(resi
##     fungal, family = binomial(link = logit), dat
##
## Coefficients:
##                       Estimate Std. Error z value
## (Intercept)          -3.48639    1.26878  -2.748
## factor(sex)2        -17.60157 3335.08084  -0.005
## factor(residence)2    0.28265    1.23696   0.229
## factor(residence)3    0.06719    1.38587   0.048
## factor(smoker)1       2.14779    1.02575   2.094
## fungal                0.86793    0.52075   1.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
##
## (Dispersion parameter for binomial family taken
##
##     Null deviance: 40.472  on 40  degrees of fre
## Residual deviance: 25.786  on 35  degrees of fre
## AIC: 37.786
##
## Number of Fisher Scoring iterations: 18
```

The R output suggests that the only estimates that are either significant or close to significance are:

$$\hat{\beta}_0 = -3.486,\ \hat{\beta}_4 = 2.148,\ \hat{\beta}_5 = 0.868.$$

We can use Analysis of Deviance to carry out more formal model selection:

```
asthma.order <- glm(asthma~factor(smoker)+fungal+fa
  +factor(residence), data=asthma.dat, family=binom
```

```
anova(asthma.order, test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: asthma
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid.
## NULL                                 40     40.4
## factor(smoker)     1   7.0921        39     33.3
## fungal             1   5.5279        38     27.8
## factor(sex)        1   2.0140        37     25.8
## factor(residence)  2   0.0519        35     25.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

We can draw the following conclusions:

- For the row corresponding to `factor(smoker)`, the corresponding model is

  $$\eta_i = \beta_0 + \beta_4 s_i.$$

  The p-value for testing $H_0 : \beta_4 = 0$ versus $H_1 : \beta_4 \neq 0$ is 0.008. We therefore reject H_0 and conclude that `asthma` depends on the variable `smoker`.

- For the row corresponding to `fungal`, the corresponding model is

  $$\eta_i = \beta_0 + \beta_4 s_i + \beta_5 x_{i4}.$$

  The p-value for testing $H_0 : \beta_5 = 0$ versus $H_1 : \beta_5 \neq 0$ is 0.019. We should therefore also include `fungal` in the model.

- For the row corresponding to `factor(sex)`, the corresponding model is

$$\eta_i = \beta_0 + \beta_4 s_i + \beta_5 x_{i4} + \beta_1 g_i.$$

For $H_0 : \beta_1 = 0$ versus $H_1 : \beta \neq 0$, the p-value is 0.156. We should not therefore reject H_0 as we do not have any evidence that `asthma` depends on `sex`, i.e., we should not include this variable in our model.

- We should draw similar conclusions about `residence`, which should not be included in our final model.

Our final model is therefore fitted using the following R code:

```
asthma.final = glm(asthma~factor(smoker)+fungal,
                   data=asthma.dat, family=binomia
summary(asthma.final)

##
## Call:
## glm(formula = asthma ~ factor(smoker) + fungal,
##     data = asthma.dat)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(
## (Intercept)      -3.9445     1.1509  -3.427 0.0
## factor(smoker)1   2.3341     1.0037   2.326 0.0
## fungal            0.9977     0.4830   2.066 0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## (Dispersion parameter for binomial family taken
##
##     Null deviance: 40.472  on 40  degrees of fr
## Residual deviance: 27.852  on 38  degrees of fr
## AIC: 33.852
##
## Number of Fisher Scoring iterations: 5
```

We have,

$$\hat{\eta}_i = -3.9445 + 2.3341 s_i + 0.9977 x_{i4}.$$

We can therefore conclude that being a smoker and living in an environment that has high levels of fungus increases the risk of having asthma, as we would expect. For example, if a smoker lives in an environment with a fungal concentration of 1.5, we have

$$\hat{\eta}^* = -3.9445 + 2.3341 \times 1 + 0.9977 \times 1.5 = -0.11385$$

and

$$\hat{\pi}^* = \frac{e^{-0.11385}}{1 + e^{-0.11385}} = 0.4716.$$

This means that the risk of having asthma is 47%!

Now adapt the code below to fit the same model but using each of the two alternative link functions. For an additional challenge, carry out the whole of the analysis of the asthma data with each alternative link function.

```r
asthma<-c(0,1,0,0,1,0,1,0,1,1,0,0,0,0,0,1,0,0,
sex<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
residence<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,
smoker<-c(0,1,0,0,1,1,1,0,1,0,0,1,1,0,1,0,0,0,
fungal<-c(0.46591,0.61690,1.45138,0.50357,1.67
 0.89422,0.34988,2.78088,0.76802,2.93615,1.189
 2.49040,0.38861,0.77466,-0.48308,0.23313,0.78
asthma.dat<-as.data.frame(cbind(asthma,sex,res
asthma.final = glm(asthma~factor(smoker)+funga
        family=binomial)
print(summary(asthma.final))
```

Poisson regression

Another type of response variable that occurs frequently in practice is count data. As with the binomial data that we have already seen, this type of data can also be analysed using GLMs. However, we will need to make different assumptions and use a different link function.

A random variable Y is said to follow a Poisson distribution with parameter μ if it takes integers $y = 0, 1, 2, \dots$, with probability

$$Pr(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!},$$

for $\mu > 0$. The mean and variance of this distribution have the same value and therefore this distribution has only one parameter, i.e.,

$$E(Y) = Var(Y) = \mu.$$

Example: Moth data. The number of moths caught overnight in a trap is recorded on 20 successive days. The average night-time temperature is also recorded. We want to model the relationship between Y_i, the number of moths caught overnight, and x_i, the average night-time temperature.

We assume that the number of moths caught overnight, Y_i, approximately follows a Poisson distribution,

$$Y_i \sim Po(\mu_i),$$

where $\mu_i = E(Y_i) > 0$.

As with binomial data, we cannot model μ_i directly using a linear function such as $\eta_i = \beta_0 + \beta_1 x_i$. In this case, this is because η_i can take any value, while μ_i is restricted to non-negative values.

However, it is sensible to model $\mu_i = \exp(\eta_i)$, i.e.,

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i.$$

Here we are using a *log* link function. Our model is called a *log-linear model* or a *Poisson regression model*.

Maximum likelihood estimation

The log-likelihood function for n independent observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is given by

$$\begin{aligned}
\ell(\beta_0, \beta_1 | \boldsymbol{y}) &= \sum_{i=1}^{n} \log Pr(Y_i = y_i) \\
&= \sum_{i=1}^{n} \log \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \\
&= \sum_{i=1}^{n} \{ y_i \log(\mu_i) - \mu_i \} + C \\
&= \sum_{i=1}^{n} \{ y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i) \} + C
\end{aligned}$$

The maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can then be found numerically.

Goodness-of-fit

For the full Poisson regression model, $\hat{\mu} = y_i$ and so

$$\ell_F(\boldsymbol{y}) = \sum_{i=1}^{n} \{ y_i \log(y_i) - y_i \} + C$$

For a general model,

$$\ell(\hat{\boldsymbol{\mu}} | \boldsymbol{y}) = \sum_{i=1}^{n} \{ y_i \log(\hat{\mu}_i) - \hat{\mu}_i \} + C$$

The deviance is therefore,

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2(\ell_F(\boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}} | \boldsymbol{y})) = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i \right.$$

For large samples, the distribution of the deviance approximately follows a chi-squared distribution,

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p-1}.$$

As for the binomial models seen earlier in the chapter, the deviance can be used to test the goodness-of-fit of the model.

Example: Moth data (continued). We can now fit a suitable GLM to the moth data.

```
moth<-read.table("https://www.mas.ncl.ac.uk/~ndw/m
head(moth)

##   no.caught temp
## 1         4    5
## 2         6   10
## 3         2    4
## 4         7    6
## 5         9    7
## 6         5    3

moth.glm <- glm(no.caught~temp, family=poisson, da
summary(moth.glm)

##
## Call:
## glm(formula = no.caught ~ temp, family = poisso
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|
## (Intercept)  1.02700    0.21545   4.767 1.87e-0
## temp         0.10282    0.02913   3.530 0.00041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## (Dispersion parameter for poisson family taken
##
##     Null deviance: 41.273  on 19  degrees of fr
## Residual deviance: 29.711  on 18  degrees of fr
## AIC: 98.25
##
## Number of Fisher Scoring iterations: 5
```

The fitted model is:

$$\begin{aligned} Y_i &\sim Po(\mu_i) \\ \log(\mu_i) &= \eta_i = 1.027 + 0.1028 \text{ temp}_i \\ \mu_i &= \exp(\eta_i) = \exp(1.027 + 0.1028 \text{ temp}_i) \end{aligned}$$

As temperature increases, the number of moths caught increases, as we would expect.

Example: Fish data. The table below shows the number of fish caught in one hour by various anglers in different rivers.

Angler	River 1	River 2	River 3	River 4
1	0	2	5	4
2	1	3	7	5
3	4	6	8	15
4	2	12	15	15
5	4	10	11	22
6	3	16	12	18
7	5	10	18	19
8	7	8	15	22
9	4	7	14	24
10	8	12	11	25

We are interested in knowing whether there are any differences between rivers, or any differences between anglers.

If Y_{ij} is the number of fish caught by angler i in river j, and we make the seemingly reasonable assumption that these values follow a Poisson distribution, we can specify a model,

$$\begin{gathered} Y_{ij} \sim Po(\mu_{ij}) \\ \log(\mu_{ij}) = \mu + \alpha_i + \beta_j \end{gathered}$$

for angler $i = 1, \dots, 10$ and river $j = 1, \dots, 4$.

In this case, we have covariates that are both **factors**. We can use the R function `gl` (short for **g**enerate **l**evels) to specify the patterns of their levels. For example, `angler` has 10 levels that need to be replicated 4 times each, while `river` has 4 levels with a total length of 40. Our data can be seen below.

```
fish <- read.table("https://www.mas.ncl.ac.uk/~ndw
fish$angler<-gl(10,4)
fish$river<-gl(4,1,40)
head(fish)

##   fish angler river
## 1    0      1     1
## 2    2      1     2
## 3    5      1     3
## 4    4      1     4
## 5    1      2     1
## 6    3      2     2
```

We can now fit a GLM and examine the Analysis of Deviance table.

```
fish.glm<-glm(fish~river+angler, family=poisson, d
anova(fish.glm,test="Chisq")

## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: fish
```

```
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Ch
## NULL                      39    182.304
## river   3   94.113        36     88.191 < 2.2e-
## angler  9   66.959        27     21.232 5.991e-
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
```

We see that we have overwhelming evidence against both null hypotheses. We can therefore conclude that there are significant differences between rivers and between anglers in the number of fish caught.

GLMs and the exponential family of distributions

We have now seen how to model data with response variables that follow the binomial and Poisson distributions. While these models have their own distinctive characteristics, they also share common features. In fact, theoretical results are available that link GLMs that have a much wider range of response variables.

A GLM is defined by specifying three components: a response, which should be a member of the *exponential family* of distributions; a *link function* that describes how the expectation of the response is related to the linear predictor; and a *linear predictor* that specifies a linear combination of explanatory variables.

Exponential family

A distribution belongs to the *exponential family* if its p.d.f. can be expressed in the form,

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

for specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. Here θ is called the *canonical parameter* and ϕ is called the *dispersion parameter*.

The exponential family includes many commonly used distributions such as the normal, binomial, Poisson and Gamma distributions. For example, suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then,

$$\begin{aligned} f(y|\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}. \end{aligned}$$

We have $\theta = \mu,\ \ \phi = \sigma^2$, and

$$\begin{aligned} a(\phi) &= \phi, \\ b(\theta) &= \theta^2/2, \\ c(y,\phi) &= -\frac{1}{2}[y^2/\phi + \log(2\pi\phi)]. \end{aligned}$$

Hence the normal distribution belongs to the exponential family.

As a second example, we can also show that this is the case for the Gamma distribution. Suppose that $Y \sim \Gamma(\alpha, \alpha/\mu)$, with this parameterisation chosen so that $E(Y) = \mu$. Then,

$$\begin{aligned} f(y|\theta,\phi) &= \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right)^{\alpha} y^{\alpha-1}\exp\left(-\frac{\alpha y}{\mu}\right) \\ &= \exp\left\{-\alpha\log\mu - \frac{\alpha y}{\mu} + \alpha\log\alpha + (\alpha-1)\log y - \log \right. \end{aligned}$$

We have $\theta = -1/\mu$, $\phi = 1/\alpha$, and

$$\begin{aligned} a(\phi) &= \phi, \\ b(\theta) &= -\log(-\theta), \\ c(y, \phi) &= -\phi^{-1} \log \phi + (\phi^{-1} - 1) \log y - \log \Gamma(\phi^{-1}). \end{aligned}$$

Thus the Gamma distribution also belongs to the exponential family. There are similar results for the binomial and Poisson distributions, among others.

In general, it can be shown that

$$\begin{aligned} \mu = E(Y) = b'(\theta), \quad \text{and} \\ Var(Y) = b''(\theta)a(\phi) \end{aligned}$$

The mean of Y is therefore a function of θ only. The variance of Y is a product of two functions: $b''(\theta)$ is called the *variance function*, which depends on the canonical parameter (and hence the mean μ) only; $a(\phi)$ depends on the dispersion parameter only.

Link functions

In general, the most commonly used link functions are the:

- Identity link: $g(\mu) = \mu$ for the normal distribution.
- Logit link: $g(\mu) = \log\left(\frac{\pi}{1-\pi}\right)$ for the binomial distribution.
- Log link: $g(\mu) = \log(\mu)$ for the Poisson distribution.
- Reciprocal link: $g(\mu) = 1/\mu$ for the Gamma distribution.
- Inverse square link: $g(\mu) = 1/\mu^2$ for the inverse Gaussian distribution.

Fitting a GLM

Maximum likelihood estimation can be used to fit this larger family of GLMs, although with the exception of the linear regression model, there are no analytical solutions. Fortunately, in general we can use software such as R to obtain MLEs numerically.

Deviance can be used to select between any models from the exponential family. We define the *scaled deviance* as

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2[\ell(\boldsymbol{y}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}|\boldsymbol{y})].$$

The (unscaled) *deviance* is defined as

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \phi D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}).$$

Recall that ϕ is the dispersion parameter in the exponential family. For the binomial and Poisson distributions, we have $\phi = 1$, so the scaled and unscaled deviance are the same.

Consider a multiple linear regression model with independent $\mathcal{N}(0,\sigma^2)$ errors, where σ^2 is known. The log-likelihood is

$$\ell(\boldsymbol{\mu}|\boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2.$$

Therefore

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2[\ell(\boldsymbol{y}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}|\boldsymbol{y})] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2.$$

In this model, the dispersion parameter $\phi = \sigma^2$ and therefore the deviance,

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2.$$

This is just the residual sum of squares for the fitted model.

The scaled deviance is a generalised likelihood test statistic and therefore for large n, we have approximately

$$D^* \sim \chi^2_{n-p-1}.$$

These results can be used to choose between models from any member of the exponential family.