# 7 Analysis of Deviance

## 7.1 Introduction

On the basis of a set of observations on the dependendent variable $Y$, and explanatory variables $X_1, X_2, \ldots, X_p$, we seek to fit a proposed GLM to the data.

For a sample of size $n$, it is possible to fit a model, known as the *saturated model*, which fits the data exactly. This model has as many parameters, $\beta_1, \ldots, \beta_p$, $p = n$, in the linear predictor as the size of the sample. However, this model is perhaps unlikely to provide the 'best' explanation of the underlying processes that are generating the data in comparison to a model with fewer parameters.

It is of interest to know whether a model with fewer parameters would actually fit the data sufficiently well.

One way to proceed is to propose a measure of the discrepancy between the fitted values for the model under consideration, and the actual values of the dependent variable. If this measure has a distribution which is known under appropriate circumstances, then we can identify those discrepancies which are statistically significant.

**Example 7.1 (Toxoplasmosis Data)**
Efron(1986):
The table shows the number of subjects out of various samples that tested positive for toxoplasmosis in a certain country. Interests lies in investigating whether there is any relationship between the rainfall (in mm.), and the proportion testing positive in each city.

Let $y_i$, $n_i$, and $\pi_i$, be the number that test positive, the number tested, and the chance that a randomly selected individual is positive, for city $i$.

Then, clearly, the appropriate error structure should be that the $\{y_i\}$ are independent, and

$$y_i \sim \text{Bin}(n_i, \pi_i).$$

Let $x_i$ be the amount of rainfall for city $i$, **in metres**. We seek to fit

$$\log(\pi_i/(1-\pi_i) = \beta_0 + \beta_1(x_i - \overline{x}) + \beta_2(x_i - \overline{x})^2 + \ldots + \beta_p(x_i - \overline{x})^p$$

$i = 1, \ldots, n$, for appropriate $p$.

*Question: How do we decide on the appropriate value of $p$?*

| city | no.tested | no.pos | rain |
|---|---|---|---|
| 1 | 4 | 2 | 1735 |
| 2 | 10 | 3 | 1936 |
| 3 | 5 | 1 | 2000 |
| 4 | 10 | 3 | 1973 |
| 5 | 2 | 2 | 1750 |
| 6 | 5 | 3 | 1800 |
| 7 | 8 | 2 | 1750 |
| 8 | 19 | 7 | 2077 |
| 9 | 6 | 3 | 1920 |
| 10 | 10 | 8 | 1800 |
| 11 | 24 | 7 | 2050 |
| 12 | 1 | 0 | 1830 |
| 13 | 30 | 15 | 1650 |
| 14 | 22 | 4 | 2200 |
| 15 | 1 | 0 | 2000 |
| 16 | 11 | 6 | 1770 |
| 17 | 1 | 0 | 1920 |
| 18 | 54 | 33 | 1770 |
| 19 | 9 | 4 | 2240 |
| 20 | 18 | 5 | 1620 |
| 21 | 12 | 2 | 1756 |
| 22 | 1 | 0 | 1650 |
| 23 | 11 | 8 | 2250 |
| 24 | 77 | 41 | 1796 |
| 25 | 51 | 24 | 1890 |
| 26 | 16 | 7 | 1871 |
| 27 | 82 | 46 | 2063 |
| 28 | 13 | 9 | 2100 |
| 29 | 43 | 23 | 1918 |
| 30 | 75 | 53 | 1834 |
| 31 | 13 | 8 | 1780 |
| 32 | 10 | 3 | 1900 |
| 33 | 6 | 1 | 1976 |
| 34 | 37 | 23 | 2292 |

## 7.2 Goodness-of-fit

We test whether a model with $p$-parameters (and hence $p$ explanatory variables) provides a good fit to the data in comparison to a saturated model with $n$ parameters (which provides an exact fit).

Recall that $\theta_i$ represents the $i$-th canonical parameter. Let $\hat{\theta}_i$ be the M.L. estimate of $\theta_i$ under a $p$-parameter model, i.e. where $\eta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, and $\widetilde{\theta}_i$ be the M.L. estimate under the saturated $n$-parameter model.

The likelihood function for the $n$ observations is

$$L(\theta_1, \ldots, \theta_n; \phi) = \exp\left\{ \sum_{i=1}^{n} \frac{[y_i\theta_i - b(\theta_i)]}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \right\}. \tag{1}$$

Formally, we seek a test for

$$H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_n = 0 \text{ vs. } H_1 : H_0 \text{ false},$$

i.e. at least one of the $\beta_i,\ i{=}p+1, \ldots, n$, is not zero.

To formulate an appropriate decision rule, consider the *generalized likelihood ratio test* statistic for the above hypotheses:

$$\Lambda = \frac{L(\hat{\theta}_1, \ldots, \hat{\theta}_n)}{L(\widetilde{\theta}_1, \ldots, \widetilde{\theta}_n)}. \tag{2}$$

The numerator of (2) represents the maximum of the likelihood under $H_0$, while the denominator represents the unconstrained maximum (attainable under the *saturated model*). Further observe that $0 \leq \Lambda \leq 1$.

If $H_0$ were true, then we should expect $\Lambda$ to be close to 1; on the other hand, if $H_0$ were false (i.e. $H_1$ true), then we should expect $\Lambda$ to be relatively small. So our decision procedure could take the form:

Accept $H_0$ if $\Lambda \geq c_1$
Reject $H_0$ if $\Lambda < c_1$ for $0 < c_1 < 1$.
or, equivalently,
Accept $H_0$ if $\log \Lambda \geq c_2$
Reject $H_0$ if $\log \Lambda < c_2$
or,
Accept $H_0$ if $-2 \log \Lambda \leq c$
Reject $H_0$ if $-2 \log \Lambda > c$

In other words, large values of $-2 \log \Lambda$ lead us to doubt $H_0$.

Developing (2) further,

$$\Lambda = \frac{\exp\left\{\sum_{i=1}^n \{[y_i\hat{\theta}_i - b(\hat{\theta}_i)]/a_i(\phi) + c(y_i, \phi)\}\right\}}{\exp\left\{\sum_{i=1}^n \{[y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)]/a_i(\phi) + c(y_i, \phi)\}\right\}}$$

$$= \exp\left\{\sum_{i=1}^n [y_i(\hat{\theta}_i - \widetilde{\theta}_i) - b(\hat{\theta}_i) + b(\widetilde{\theta}_i)]/a_i(\phi)\right\}.$$

Therefore

$$-2 \log \Lambda = 2 \sum_{i=1}^n [y_i(\widetilde{\theta}_i - \hat{\theta}_i) - b(\widetilde{\theta}_i) + b(\hat{\theta}_i)]/a_i(\phi) \tag{3}$$

In the case where $a_i(\phi) = \phi$, the RHS is equal to $D/\phi$, where

$$D = 2 \sum_{i=1}^n [y_i(\widetilde{\theta}_i - \hat{\theta}_i) - b(\widetilde{\theta}_i) + b(\hat{\theta}_i)]. \tag{4}$$

We say that $D$ is the *deviance* of the $p$-parameter model, and that $D/\phi$ is its *scaled deviance*.

Relating these observations to our earlier comment, a large value of $D/\phi$ leads us to doubt $H_0$. Furthermore, it can be shown that under $H_0$

$$D/\phi \sim \chi^2_{n-p}$$

approximately.

It should be noted, however, that the $\chi^2_{n-p}$ distribution can sometimes be a poor approximation to the true distribution of $D/\phi$ under $H_0$. Our refined decision procedure becomes: Reject $H_0$ if $D/\phi > \chi^2_{n-p,\alpha}$ at the $100\alpha\%$ level of significance.

## 7.3 Model comparison

Suppose that we have two models under consideration: one with $q$ parameters, and the other with an additional $p-q$ parameters. A question of interest is whether the additional parameters provide a significantly better fit to the data.

More formally, suppose $p$ explanatory variables, $X_1, \ldots, X_p$, are available, and we are trying to decide whether we really need $X_{q+1}, \ldots, X_p$. (We can always permute parameters and explanatory variables so that the $X_i$'s under scrutiny are the last $p-q$).

We test

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0 \text{ vs. } H_1 : H_0 \text{ false}$$

i.e. at least one of the $\beta_i$, $i = q+1, \ldots, p$ is non-zero.

Let $\hat{L}_F$ be the maximized likelihood of the $\{y_i\}$ under the 'full' model, where all $p$ explanatory variables are being utilized.
Also, let $\hat{L}_R$ be the maximized likelihood of the $\{y_i\}$ under the 'reduced' model, where we only utilize the first $q$ explanatory variables.

Consider, yet again, the *generalized likelihood ratio test* statistic for this current set of hypotheses:

$$\Lambda_0 = \hat{L}_R / \hat{L}_F$$

and also define

$$W = -2 \log \Lambda_0.$$

If $H_0$ is true, then $\Lambda_0$ should be 'close' to 1, and so $W$ is 'close' to 0.

If $H_0$ is false (i.e. $H_1$ true), $\Lambda_0$ should be somewhat less than 1, and so $W$ should be somewhat larger than 0.

So we reject $H_0$ if $W > c'$.

However, it can be shown that under $H_0$:

$$W \sim \chi^2_{p-q} \qquad \text{approximately.}$$

So the decision procedure can be refined to:

Reject $H_0$ if $W > \chi^2_{p-q,\alpha}$ at the $100\alpha\%$ level of significance.

It will prove to be convenient to cast $W$ in terms of the deviances for the reduced and full models. To do this, observe that

$$\Lambda_0 = \frac{\hat{L}_R}{\hat{L}_S} \div \frac{\hat{L}_F}{\hat{L}_S}$$

where $\hat{L}_S$ is the maximized likelihood for the saturated model. Then

$$W = -2\log\Lambda_0 = -2\left[\log\left(\frac{\hat{L}_R}{\hat{L}_S}\right) - \log\left(\frac{\hat{L}_F}{\hat{L}_S}\right)\right].$$

Assuming that $a_i(\phi) = \phi$, then, in fact,

$$W = (D_R - D_F)/\phi$$

where $D_R/\phi$ and $D_F/\phi$ are the scaled deviances for the reduced and full models respectively.

If $\phi$ is known:
then $W \sim \chi^2_{p-q}$ approx. under $H_0$ (as remarked earlier).

If $\phi$ is **not** known:

$$D_F/\phi \sim \chi^2_{n-p}$$

$$\Rightarrow E\left[\frac{D_F}{\phi}\right] = n - p \Rightarrow E\left[\frac{D_F}{n-p}\right] = \phi$$

approx.

So, perhaps, an appropriate estimator of $\phi$ is $\widetilde{\phi} = \frac{D_F}{n-p}$ yielding a new statistic

$$W_2 = \frac{D_R - D_F}{\widetilde{\phi}} = \frac{D_R - D_F}{D_F/(n-p)}.$$

However, we find it convenient to consider a slightly modified version of this

$$W_1 = \frac{(D_R - D_F)/(p-q)}{D_F/(n-p)}.$$

Under $H_0$,

$$\frac{D_R - D_F}{\phi} \sim \chi^2_{p-q}, \qquad \frac{D_F}{\phi} \sim \chi^2_{n-p}$$

approx. and independently.

Therefore

$$W_1 = \frac{\left(\frac{D_R - D_F}{\phi}\right)/(p-q)}{\left(\frac{D_F}{\phi}\right)/(n-p)} = \frac{(D_R - D_F)/(p-q)}{D_F/(n-p)} \sim F_{p-q,n-p}$$

approximately, under $H_0$.

**Remarks 7.2 (A summary)**
To test
$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$
vs.
$H_1 : H_0$ false

when $\phi$ is **known**, reject $H_0$ if $W > X^2_{p-q,\alpha}$, and

when $\phi$ is **unknown**, reject $H_0$ if $W_1 > F_{p-q,n-p,\alpha}$

at the $100\alpha\%$ level of significance.

## 7.4 Recovering ANOVA for the General Linear Model

A motivating theme for the theory of the GLM is that it generalizes many of the key aspects and results for the General Linear Model.

Is it the case, then, that if we carry out an Analysis of Deviance procedure for the General Linear Model (i.e. Normal error structure, with its default, identity, link), that this really amounts to carrying out an Analysis of Variance?

The answer is yes! To see this, let us re-introduce the general linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is a $(k+1) \times 1$ vector, $\mathbf{y}$ is an $n \times 1$ vector, and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$, and the design matrix, $\mathbf{X}$, is $n \times (k+1)$. Here, the first column of $\mathbf{X}$ consists entirely of 1's, thus incorporating a constant term in the model.

Since the $\{Y_i\}$ are mutually independent, then the likelihood function for the $n$ observations can be written as:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{5}$$

This likelihood is maximized at the value of $\boldsymbol{\beta}$ that minimizes

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{6}$$

The value of $\boldsymbol{\beta}$ that achieves this is the the least squares estimator:

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

Thus, the maximized likelihood for this $(k+1)$-parameter model is

$$L(\hat{\boldsymbol{\beta}}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right\}. \tag{7}$$

Now, let $\widetilde{\boldsymbol{\beta}}$ be the parameter estimate for $\boldsymbol{\beta}$ in the corresponding *saturated* model. In this case, the fitted values are equal to the observed values, the $\{y_i\}$, i.e. $\mathbf{y} = \mathbf{X}\widetilde{\boldsymbol{\beta}}$, and so the maximized likelihood for the saturated model is:

$$L(\widetilde{\boldsymbol{\beta}}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}. \tag{8}$$

Hence, the *scaled deviance* of this (full) $(k+1)$-parameter model is:

$$\begin{aligned} D_F/\phi &= -2\log\Lambda = -2\log\left[\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right\}\right] \\ &= \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma^2}(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}). \end{aligned} \tag{9}$$

The bracketed term, $\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$, on the R.H.S. of (9), can be recognized as being $SS_R(k)$, the residual sum-of-squares for this model.
Hence

$$D_F/\phi = SS_R(k)/\sigma^2 \sim \chi^2_{n-k-1} \tag{10}$$

exactly.

Similarly, for a reduced model with $(m+1)$-parameters

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta}_1 + \epsilon$$

where $\mathbf{y}$ and $\boldsymbol{\epsilon}$ are $n \times 1$ vectors, $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, \ldots, \beta_m)'$ and $\mathbf{X}_1$ is a $n \times (m+1)$ matrix, then

$$D_R/\phi = \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1)'(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1)$$

$$= \frac{1}{\sigma^2}(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}_1'\mathbf{X}_1'\mathbf{y}) \sim \chi^2_{n-m-1}$$

where, again, the bracketed term on the RHS can be recognized to be the residual sum-of-squares, $SS_R(m)$, for this model.

Hence, it can be shown that

$$\frac{D_R - D_F}{\phi} = \frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}_1'\mathbf{X}_1'\mathbf{y}). \tag{11}$$

But the term in brackets on the RHS of (11) is $SS_{Reg}(k) - SS_{Reg}(m)$, the sum of squares due to $x_{m+1}, \ldots, x_k$, adjusted for the presence of $x_1, \ldots, x_m$; this is associated with $k - m$ degrees of freedom.

Under the hypothesis that the reduced $m + 1$ parameter model is adequate

$$\frac{D_R - D_F}{\phi} = \frac{SS_{Reg}(k) - SS_{Reg}(m)}{\sigma^2} \sim \chi^2_{k-m}$$

independently of (10).

Hence

$$\begin{aligned} W_1 &= \frac{(D_R - D_F)/(k-m)}{D_F/(n-k-1)} \\ &= \frac{\left(\frac{D_R - D_F}{\phi}\right)/(k-m)}{\left(\frac{D_F}{\phi}\right)/(n-k-1)} \\ &= \frac{\left(\frac{SS_{Reg}(k) - SS_{Reg}(m)}{\sigma^2}\right)/(k-m)}{\left(\frac{SS_R(k)}{\sigma^2}\right)/(n-k-1)} \\ &= \frac{(SS_{Reg}(k) - SS_{Reg}(m))/(k-m)}{SS_R(k)/(n-k-1)} \sim F_{k-m,n-k-1} \end{aligned}$$

exactly, under adequacy of the reduced model.

However, $W_1$ is the statistic used in the Analysis of Deviance for testing whether the reduced model is adequate (in comparison to the larger, full, model).

Thus, the asymptotic $F$-test in the Analysis of Deviance is actually **exact** for the general linear model.

We return to our earlier example. The data presented have been entered into S-PLUS as a *Data Set*, which has been stored under the name `toxoplas`. The columns of `toxoplas` can be accessed by invoking the function `attach`.

The alternative would be to construct the data frame in the usual manner, using the `data.frame` function.

```
> attach(toxoplas)
> rain.m <- rain/1000
> rain.cor <- rain.m - mean(rain.m)
> prop <- no.pos/no.tested
> toxoplas.glm <- glm(prop ~ rain.cor + rain.cor^2 + rain.cor^3 + rain.cor^4 + rain.cor^
    5, weights = no.tested, family = binomial, data = toxoplas)
> summary(toxoplas.glm)

Call: glm(formula = prop ~ rain.cor + rain.cor^2 + rain.cor^3 + rain.cor^4 + rain.cor^5, family
     = binomial, data = toxoplas, weights = no.tested)
Deviance Residuals:
       Min        1Q     Median        3Q      Max
 -2.982888 -1.209598 -0.4572224 0.4159548 2.884612

Coefficients:
                      Value  Std. Error     t value
  (Intercept)    0.007969961   0.1437323  0.05545005
     rain.cor   -1.831774141   1.3814359 -1.32599287
I(rain.cor^2)    6.348428465  11.7381399  0.54083769
I(rain.cor^3)  -11.344607602  54.3392089 -0.20877388
I(rain.cor^4) -164.883565581 142.6997465 -1.15545801
I(rain.cor^5)  539.132673276 483.5163183  1.11502477

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 74.21188 on 33 degrees of freedom

Residual Deviance: 61.19616 on 28 degrees of freedom

Number of Fisher Scoring Iterations: 3

Correlation of Coefficients:
              (Intercept)   rain.cor I(rain.cor^2) I(rain.cor^3) I(rain.cor^4)
     rain.cor  0.0879872
I(rain.cor^2) -0.7607631   0.1692942
I(rain.cor^3)  0.1625867  -0.8229070 -0.4857725
I(rain.cor^4)  0.6299306  -0.2560967 -0.9535043     0.6014199
I(rain.cor^5) -0.3259113   0.6327059  0.6795549    -0.9377777    -0.8103714
> anova(toxoplas.glm)
Analysis of Deviance Table

Binomial model

Response: prop

Terms added sequentially (first to last)
              Df Deviance Resid. Df Resid. Dev
         NULL                    33   74.21188
     rain.cor  1  0.12439        32   74.08749
I(rain.cor^2)  1  0.00001        31   74.08747
I(rain.cor^3)  1 11.45287        30   62.63460
I(rain.cor^4)  1  0.18819        29   62.44641
I(rain.cor^5)  1  1.25025        28   61.19616
```

A significant drop in the (scaled) deviance occurs when we fit the cubic term in the presence of all the lower order terms.[1]

Does the third order polynomial provide a significantly better fit than just fitting a constant in the linear predictor?
This question is answered by comparing the change in deviance, $74.21188 - 62.63460 = 11.57728$, with the $\chi^2_3$ distribution. Roughly speaking, if the change in deviance is much larger than the change in the d.o.f., then at least one of the additional terms is significant. More formally

```
> p.value.approx <- 1 - pchisq(11.57728, 3)
> p.value.approx
[1] 0.008980838
```

Thus, a third-order model provides a significantly better fit than one with just the constant term.

Let us re-fit the third-order polynomial, and extract the parameter estimates:

```
> toxoplas.glm <- glm(prop ~ rain.cor + rain.cor^2 + rain.cor^3, weights = no.tested, family
     = binomial, data = toxoplas)
> summary(toxoplas.glm)

Call: glm(formula = prop ~ rain.cor + rain.cor^2 + rain.cor^3, family = binomial, data =
    toxoplas, weights = no.tested)
Deviance Residuals:
       Min       1Q     Median        3Q      Max
 -2.761964 -1.21662 -0.5078691 0.3538459 2.620382

Coefficients:
                     Value Std. Error    t value
  (Intercept)   0.09939242  0.1019493  0.9749196
     rain.cor  -2.55186500  0.8826035 -2.8912926
I(rain.cor^2)  -6.06361025  2.9611658 -2.0477105
I(rain.cor^3)  39.32214289 11.7294136  3.3524389

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 74.21188 on 33 degrees of freedom

Residual Deviance: 62.6346 on 30 degrees of freedom
```

Thus

$$\hat{\beta}_0 = 0.09939242, \quad \hat{\beta}_1 = -2.55186500, \quad \hat{\beta}_2 = -6.06361025, \quad \hat{\beta}_3 = 39.32214289$$

Note that we are **not** claiming here that this model fits the data adequately. Indeed, we see that the Residual Deviance is 62.6346 on 30 degrees of freedom, which is very significant.

[1]Note: a constant term is always fitted by default unless explicitly removed by placing `-1` in the linear predictor.