3 Analysis of Deviance

3.1 Introduction

On the basis of a set of observations on the dependent variable Y, and the explanatory variables $X_1, X_2, \ldots, X_p$, we now know how to fit a proposed GLM to the data using maximum likelihood.

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, $\beta_1, \ldots, \beta_n$, $p = n$, in the linear predictor as the size of the sample. However, this model is unlikely to provide the 'best' explanation of the underlying processes that are generating the data in comparison to a model with fewer parameters.

We wish to know whether a model with fewer parameters would fit the data "sufficiently well" - a model somewhere between the *null model* with only one parameter representing a common mean, and the saturated model with n parameters. To investigate this we shall use the *generalized ratio statistic* to define a measure of the discrepancy between the fitted values for the model under consideration (the current model with p parameters), and the actual values of the dependent variable - the fitted values under the saturated model. The measure is termed the *scaled deviance*, and, from standard maximum likelihood theory, its asymptotic distribution is known under the null hypothesis that the current model holds. We are thus able to identify those discrepancies that are statistically significant.

3.2 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). That is we test whether

$H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_n = 0$ (i.e. $\boldsymbol{\beta} \in \mathbb{R}^p$)

vs

$H_1 : H_0$ false (i.e. at least one of the β_i, $i = p+1, \ldots, n$, is not zero: $\boldsymbol{\beta} \notin \mathbb{R}^p \subseteq \mathbb{R}^n$)

In this situation we may use the generalized likelihood ratio criterion to test H_0 vs H_1.

Recall that θ_i represents the i-th canonical parameter. Let $\widehat{\theta}_i$ be the M.L. estimate of θ_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. Note that $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ can be calculated from the corresponding $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ and $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^n$ using the relations

$$g(\mu_i) \;=\; g(b'(\theta_i)) \;=\; \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, n.$$

As derived in Section 2.1, the likelihood function for the n observations is

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

The *generalized likelihood ratio test* statistic for the above hypotheses is given by:

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

where the numerator of (2) represents the maximum of the likelihood under H_0 (the *current* model with p parameters), while the denominator represents the unconstrained maximum (attainable under the *saturated* model with n parameters).

Observe that $0 \leq \Lambda \leq 1$. Clearly small values of Λ are evidence against H_0, the current model, while values near 1 give us no reason to reject H_0. But how small is small? Maximum likelihood theory tells us that, under H_0, $-2\log\Lambda \;=\; -2(l_c - l_s)$ is asymptotically distributed as χ^2_{n-p}. Since it is large values of $-2\log\Lambda$ that are evidence against H_0, upper critical values of the χ^2_{n-p} distribution give us the necessary statistical goodness-of fit criteria.

Developing (2) further,

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

Therefore

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

In the (general) case where $a_i(\phi) \;=\; \phi/w_i$, the RHS is written as $S(c,s) \;=\; D(c,s)/\phi$ (or abbreviated as $S \;=\; D/\phi$), where

$$D(c,s) \text{ (or } D) \;=\; 2\sum_{i=1}^{n} w_i\left[y_i(\widetilde{\theta}_i - \hat{\theta}_i) - (b(\widetilde{\theta}_i) - b(\hat{\theta}_i))\right]. \tag{4}$$

We say that D is the *deviance* of the p-parameter model (relative to the saturated model), and that $S \;=\; D/\phi$ is its *scaled deviance*.

Our decision procedure (goodness-of-fit test) becomes:

Reject H_0 at the $100\alpha\%$ level of significance if $D/\phi > \chi^2_{n-p}(\alpha)$.

It should be noted that the χ^2_{n-p} distribution is a large sample approximation and can sometimes be a poor approximation to the true distribution of D/ϕ under H_0.

Remarks 3.1
If ϕ is unknown, we cannot carry out a goodness-of-fit test, but can only estimate ϕ. (See later).

Example 3.2
Find the scaled deviance for the Normal distribution.

For the Normal distribution $N(\mu, \sigma^2)$, we have

$$\theta = \mu, \quad b(\theta) = \frac{1}{2}\theta^2 \text{ and } \phi = \sigma^2.$$

Thus

$$\begin{aligned}
\frac{D}{\phi} &= \frac{2}{\phi}\sum_{i=1}^{n}\left[y_i(\widetilde{\theta}_i - \hat{\theta}_i) - (b(\widetilde{\theta}_i) - b(\hat{\theta}_i))\right] \\
&= \frac{2}{\sigma^2}\sum_{i=1}^{n}\left[y_i(\widetilde{\mu}_i - \hat{\mu}_i) - \frac{1}{2}(\widetilde{\mu}_i^2 - \hat{\mu}_i^2)\right] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[2y_i(y_i - \hat{\mu}_i) - (y_i^2 - \hat{\mu}_i^2)\right] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2
\end{aligned}$$

Thus D is just the usual residual sum-of-squares, and we can write the scaled deviance $S(c, s)$ as

$$\frac{D}{\phi} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(O_i - E_i)^2,$$

where the O_i represent the *observed* values of Y, and the E_i represent the *expected* (or fitted) values of Y under the current model.

Example 3.3
Find the scaled deviance for the Poisson distribution.

For the Poisson distribution $Po(\lambda)$ we have

$$f(y) = \frac{\lambda^y\, e^{-\lambda}}{y!}$$

with

$$\theta = \ln\lambda, \quad b(\theta) = e^\theta\ (= \lambda) \text{ and } \phi = 1, \text{ (with } \mu = \lambda).$$

Thus

$$\begin{aligned}\frac{D}{\phi} \;=\; D \;&=\; 2\sum_{i=1}^{n}\Big[y_i(\log\widetilde{\lambda}_i - \log\hat{\lambda}_i) - (\widetilde{\lambda}_i - \hat{\lambda}_i)\Big] \\ &=\; 2\sum_{i=1}^{n}\left[y_i\,\log\frac{\widetilde{\lambda}_i}{\hat{\lambda}_i} \;-\; (\widetilde{\lambda}_i - \hat{\lambda}_i)\right] \\ &=\; 2\sum_{i=1}^{n}\left[O_i\,\log\frac{O_i}{E_i} \;-\; (O_i - E_i)\right].\end{aligned}$$

It can be shown (prove as an exercise), that if the current model uses the canonical link (the log link), and if the linear predictor contains a constant term, then

$$\sum_{i=1}^{n}(O_i - E_i) \;=\; 0,$$

and so the scaled deviance S, which in this case is equal to the deviance D, becomes

$$D \;=\; \sum_{i=1}^{n} O_i\,\log\frac{O_i}{E_i}.$$

Exercises

(a) Show that if the observations are independently drawn binomial proportions p_i from binomial distributions $Bin(n_i, \pi)$, then the scaled deviance is given by

$$\frac{D}{\phi} \;=\; D \;=\; 2\sum_{i=1}^{n}\left[O_{si}\,\log\frac{O_{si}}{E_{si}} \;+\; (n_i - O_{si})\,\log\frac{n_i - O_{si}}{n_i - E_{si}}\right],$$

where $O_{si} = n_i p_i$ is the observed number of successes and $E_{si} = n_i\widehat{\pi}_i$ is the fitted number of successes.

(b) Show that if the observations y_i are independently drawn from the Gamma distributions $G(\mu_i, \nu)$, then the scaled deviance is given by

$$S \;=\; \nu D \;=\; 2\nu\left[\sum_{i=1}^{n}\left(\frac{O_i - E_i}{E_i}\right) \;-\; \sum_{i=1}^{n}\log\frac{O_i}{E_i}\right].$$

In the case of the Exponential distribution $Exp(\lambda)$, $\nu = 1$ (and $\mu = 1/\lambda$), and we have

$$S \;=\; D \;=\; 2\left[\sum_{i=1}^{n}\left(\frac{O_i - E_i}{E_i}\right) \;-\; \sum_{i=1}^{n}\log\frac{O_i}{E_i}\right].$$

For single parameter distributions (e.g. the Poisson, Binomial and Exponential distributions), the scaled deviance $S = D$, and can be calculated for any current model using the observed and fitted values. Thus we have a goodness-of-fit statistic that is asymptotically distributed as χ^2_{n-p} under the null hypothesis that the current (p-parameter) model holds.

For two-parameter distributions (e.g. the Normal and Gamma distributions), the scale parameter ϕ is unknown, and although the deviance D can be calculated, the scaled deviance $S = D/\phi$ cannot; a goodness-of-fit test is not possible. In this case we may obtain an estimate of ϕ, as discussed in the next section.

3.3 Model comparison

Suppose that we have available q (generally $q < n$) explanatory variables, $X_1, \ldots, X_q$. The 'full' model is the model that uses all q explanatory variables. This is the largest model (in terms of parameters) that we are able to fit, and we assume that the 'full' model fits the data. We are interested in finding a more parsimonious p-parameter model ($p < q$). That is, we are interested in testing:

$$H_0 : \beta_{p+1} = \ldots = \beta_q = 0$$

versus

H_1 : H_0 false i.e. at least one of the β_i, $i = p+1, \ldots, q$ is non-zero.

(We can always permute parameters and explanatory variables so that the X_i's under scrutiny are the last $q - p$).

Let $\hat{L}_f$ be the maximized likelihood of the $\{y_i\}$ under the 'full' model, where all q 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 p 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

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

Under H_0 ($\boldsymbol{\beta} \in \mathbb{R}^p \subseteq \mathbb{R}^q$), we have:

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

Thus we reject H_0 if $W > \chi^2_{q-p}(\alpha)$ at the $100\alpha\%$ level of significance.

It is 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

$$W = -2\log\Lambda_0 = -2\left[\log\left(\frac{\hat{L}_r}{\hat{L}_s}\right) - \log\left(\frac{\hat{L}_f}{\hat{L}_s}\right)\right].$$

In fact,

$$W = (D_r - D_f)/\phi$$

where D_r/ϕ and D_f/ϕ are the scaled deviances for the reduced and full models respectively.

If ϕ is known:
then $W \sim \chi^2_{q-p}$ asymptotically under H_0, as remarked earlier, and we carry out the required hypothesis test as a χ^2 test.

If ϕ is **not** known:

$$D_f/\phi \sim \chi^2_{n-q}, \text{ since we have assumed that the full model fits,}$$

$$\Rightarrow \quad E\left[\frac{D_f}{\phi}\right] = n - q \quad \Rightarrow \quad E\left[\frac{D_f}{n-q}\right] = \phi,$$

and so we may use the deviance to obtain an *unbiased estimate* of the scale parameter ϕ as

$$\widehat{\phi} = \frac{D_f}{n-q}.$$

This enables us to approach our hypothesis test in another way. Under H_0,

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

asymptotically and independently.

Therefore

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

approximately, under H_0.

Remarks 3.4 (A summary)
It is required to test

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

When ϕ is **known**, we reject H_0 at the $100\alpha\%$ level of significance if

$$W = \frac{D_r - D_f}{\phi} \; > \; \chi^2_{q-p}(\alpha).$$

When ϕ is **unknown**, we reject H_0 at the $100\alpha\%$ level of significance if

$$W_1 = \frac{(D_r - D_f)/(q-p)}{D_f/(n-q)} \; > \; F_{q-p,n-q}(\alpha).$$

Note that, in general, the assumed distributional results under H_0 are approximate.

Remarks 3.5
Using the preceding results we can build an *Analysis of Deviance Table*, and use changes in deviance, as terms are added to or subtracted from the model, to test the importance of these terms. These are essentially Likelihood Ratio tests.

Remarks 3.6 (General Linear Model as Special Case)
In the case of the Normal distribution with the canonical (identity) link function, we recover the usual General Linear Model and the usual analysis, including the unbiased estimate of the unknown variance σ^2, and the usual analysis of variance table and associated (in this case exact) F-tests. To see this recall, from Example 3.3, that the scaled deviance for the Normal distribution is given by

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2 \; = \; \frac{SS_{res}}{\sigma^2} \; = \; \frac{1}{\sigma^2}(\mathbf{y} - \widehat{\boldsymbol{\mu}})'(\mathbf{y} - \widehat{\boldsymbol{\mu}}),$$

where, in the case of the identity link (the General Linear Model), $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$, and the deviance is the usual residual sum-of-squares.

For the full model we shall denote the residual sum-of-squares by $SS_{res}(q)$, and for the reduced model we denote the residual sum-of-squares by $SS_{res}(p)$. Thus the estimate for the scale parameter is the usual estimate

$$\widehat{\sigma^2} = \frac{SS_{res}(q)}{n-q}.$$

Moreover, under H_0,

$$W_1 \; = \; \frac{(D_r - D_f)/(q-p)}{D_f/(n-q)} \; = \; \frac{(SS_{res}(p) - SS_{res}(q))/(q-p)}{SS_{res}(q)/(n-q)} \; \sim \; F_{q-p,n-q} \;\; \text{exactly,}$$

leading to our usual ANOVA Table F-tests.

Example 3.7 (Toxoplasmosis Data: Analysis of Deviance Table)
Efron(1986): The table overleaf shows the number of subjects (Y) out of samples of size (N) that tested positive for toxoplasmosis in a certain country. Interest lies in investigating whether there is any relationship between the rainfall (X), and the proportion testing positive in each city.

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

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

The data presented have been entered into R and stored as a data.frame under the name `toxoplas`. The columns of `toxoplas` can be accessed by invoking the function `attach`.

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

```
> summary(toxoplas.glm)

Call:
glm(formula = prop ~ rain.cor + I(rain.cor^2) + I(rain.cor^3) +
    I(rain.cor^4) + I(rain.cor^5), family = binomial, data = toxoplas,
    weights = no.tested)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9829  -1.2096  -0.4572   0.4159   2.8846

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.00797    0.14373   0.055    0.956
rain.cor       -1.83177    1.38143  -1.326    0.185
I(rain.cor^2)   6.34843   11.73816   0.541    0.589
I(rain.cor^3) -11.34461   54.33894  -0.209    0.835
I(rain.cor^4) -164.88357  142.70037  -1.155    0.248
I(rain.cor^5)  539.13267  483.51560   1.115    0.265

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 61.196  on 28  degrees of freedom
AIC: 163.89

Number of Fisher Scoring iterations: 3

> anova(toxoplas.glm)

Analysis of Deviance Table

Model: binomial, link: logit

Response: prop

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev
NULL                             33     74.212
rain.cor      1   0.1244        32     74.087
I(rain.cor^2) 1   0.0000        31     74.087
I(rain.cor^3) 1  11.4529        30     62.635
I(rain.cor^4) 1   0.1882        29     62.446
I(rain.cor^5) 1   1.2502        28     61.196
```

A significant drop in the (scaled) deviance occurs when we fit the cubic term in the presence of all the lower order terms. 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, $74.212 - 62.635 = 11.577$, with values of the χ^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

```
 pchisq(11.577, 3, lower.tail=F)
[1] 0.008982002
```

Although this is very convincing, note that the assumption that the larger model (the cubic) fits the data is doubtful. This is discussed below.

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

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

> summary(toxoplas.glm)

Call:
glm(formula = prop ~ rain.cor + I(rain.cor^2) + I(rain.cor^3),
    family = binomial, data = toxoplas, weights = no.tested)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7620  -1.2166  -0.5079   0.3538   2.6204

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.09939    0.10197   0.975 0.329678
rain.cor      -2.55187    0.88276  -2.891 0.003843 **
I(rain.cor^2) -6.06369    2.96348  -2.046 0.040743 *
I(rain.cor^3) 39.32248   11.73606   3.351 0.000806 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 62.635  on 30  degrees of freedom
AIC: 161.33

Number of Fisher Scoring iterations: 3
```

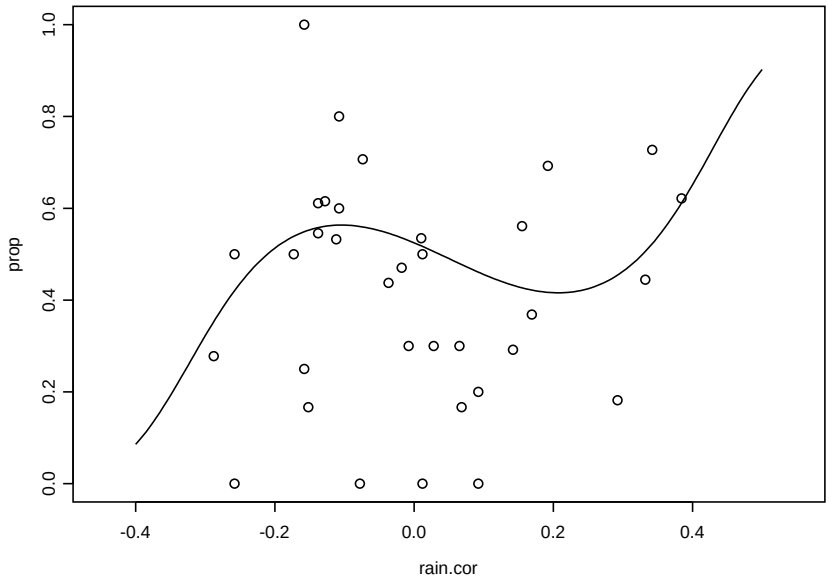
Thus

$$\hat{\beta}_0 = 0.09939, \quad \hat{\beta}_1 = -2.55187, \quad \hat{\beta}_2 = -6.06369, \quad \hat{\beta}_3 = 39.32248$$

Note again that we cannot claim here that this model fits the data adequately. Indeed, we see that the Residual Deviance is 62.635 on 30 degrees of freedom, which is very significant. It is likely that there are more explanatory variables that should be included. However it is also important to be aware that there are many reasons why a Binomial distribution might be *over-dispersed* and a rigid attitude to goodness-of-fit tests can be misleading. Again it is helpful to look at a graphical representation of the fitted model and the data and the following code

```
x <- seq(-0.4,0.5,0.005)
plot(rain.cor,prop,xlim=c(-0.45,0.55),ylim=c(0,1))
lines(x,predict(toxoplas.glm,data.frame(rain.cor=x),type="response"))
```

produces the associated plot



We will look at appropriate residual plots for this model in a later lecture (or exercise).