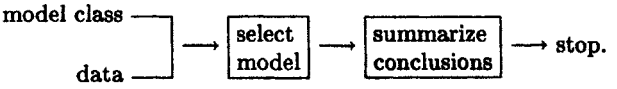
CHAPTER 12

Model checking

12.1 Introduction

The process of statistical analysis as presented in many textbook examples appears to take the form



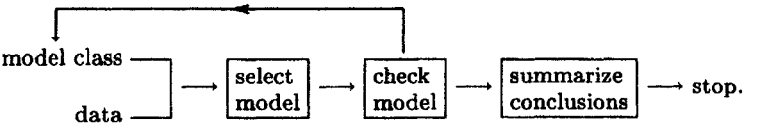
Sometimes the model class has only one member, as, for example, in the standard analysis of a randomized blocks experiment with an unstructured treatment factor. At other times model selection may involve, say, selecting a subset of terms from a full factorial model, using differences in deviance to decide which terms should be included. Whether or not the prior model is unique the process effectively assumes that at least one model from the class is the right one, so that, after fitting, all that remains is to summarize the analysis in terms of parameter estimates, standard errors covariance matrix, and so on.

A good statistician selects his model class carefully, paying attention to the type and structure of the data. Thus in modelling counts, fitted values from a model should be confined to non-negative values, because counts are. Similarly if it is known a priori that a response to a stimulus variable x tails off beyond a certain level that is well within the range of x in the data, then a linear term only in x will not be adequate for the model.

However, even after a careful selection of model class, the data themselves may indicate that the particular model selected is unsuitable. Such indications can take two forms. It may be that the data as a whole show some systematic departure from the fitted values, or it may be that a few data values are discrepant from the

rest. The detection of both systematic and isolated discrepancies is part of the technique of model checking. To exemplify a systematic discrepancy consider a plot of the residuals r against x, one of the covariates in the linear predictor. If the fit is good the pattern expected is null, i.e. no relation between r and x. However, if the plot shows residuals of one sign concentrated at the ends of the x scale and residuals of the other sign at the centre then this may be evidence that x^2 has been omitted as a covariate in the linear predictor, or for wrong choice of link function. By contrast, an isolated discrepancy would occur if a few points have residuals very far from the rest. This latter pattern indicates something unusual about those particular points; they may be at the extremes of the x range in a region where the model does not apply, or, more mundanely, the values may simply be wrong, the result of misrecording or errors in transcription.

The effect of model checking is to introduce a loop into the analysis process as follows:



The introduction of this loop changes profoundly the process of analysis and the reliability of the final models found. In this chapter we extend, where possible, techniques originally devised for regression models to the whole class of generalized linear models, and develop new ones for aspects of the general class that have no analogues in the restricted one.

12.2 Techniques in model checking

Model-checking techniques may be either informal or formal. Informal techniques rely upon the human mind and eye to detect pattern. Such methods take a successful model to be one that, among other things, leaves a patternless set of residuals. The argument is that if we can detect pattern in the residuals we can find a better model; the practical problem is that any finite set of residuals can be made to yield some kind of pattern if we look hard enough,

so that we have to guard against over-interpretation. Nonetheless informal methods are an important component in model checking.

Formal methods rely on embedding the current model in a wider class that includes extra parameters. If θ is such a parameter, and θ_0 its value in the current model, then a formal method would find $\hat{\theta}$, the estimate of θ giving the best fit in the wider class, and compare the fit at $\hat{\theta}$ with that at θ_0. The current model passes the check if the inclusion of θ as an extra parameter does not markedly improve the fit. Extra parameters might arise from including an additional covariate, from embedding a covariate x in a family $h(x;\theta)$ indexed by θ, from embedding a link function $g(\eta)$ in a similar family $g(\eta;\theta)$, or from including a constructed variate, say $\hat{\eta}^2$, obtained from the original fit. Formal methods thus look for deviations from the fit in certain definite directions thought likely to be important a priori.

Formal methods for dealing with isolated discrepancies include adding dummy variates taking the value 1 for the discrepant unit and zero elsewhere. The change in deviance then measures the effect of that unit on the fit. The addition of such a dummy variate has an effect on the fit equivalent to deleting that unit from the data matrix. In assessing the significance of that change, due allowance must be made for the effect of having picked the most discrepant unit.

12.3 Score tests for extra parameters

Many procedures used in model checking can be shown to be special cases of the class of score tests (Rao, 1973, Chapter 6). Consider two models, one (M_0) with p parameters and a second (extended) model (M_1) with $p + k$ parameters. The deviance test is based on the reduction in deviance for M_1 relative to M_0. The score test, on the other hand, is based on the log likelihood derivatives with respect to the extra parameters: both the derivatives and the Fisher information are computed under M_0. For generalized linear models the score statistic can be computed by first fitting M_0, followed by one step of the iteration for M_1. The reduction in X^2 in this first step is the score statistic, sometimes also called the quadratic score statistic. Pregibon (1982) gives the details.

The computing advantage of the score test over the deviance

test is that it requires only a single iteration for the extended model, compared with iteration to convergence for the deviance (or likelihood-ratio) test. Note that the two tests give identical results for linear models with constant variance. For $k = 1$ and no nuisance parameters the statistics are interpreted geometrically in Fig. 12.1a, in which the log likelihood derivative is plotted against the extra parameter λ.

Figure 12.1b shows a typical comparison of the two statistics over a range of λ-values. The solid curve gives the minimum deviance for the model M_1 for varying values of the extra parameter λ, M_0 itself being defined by $\lambda = \lambda_0$. The deviance statistic is thus the difference in ordinates $D_0 - D_1$. The score statistic at λ_0 has the form $S(\lambda_0) = U(\lambda_0)^T i^{-1}(\lambda_0 \mid \cdot) U(\lambda_0)$, where $U(\lambda_0)$ is the log-likelihood derivative with respect to λ at λ_0, and $i(\lambda_0 \mid \cdot)$ is the Fisher information for λ as defined in Appendix A, treating the original parameters as nuisance parameters. Its value for varying λ is shown by the dashed line in Fig 12.1b. At $\lambda = \hat{\lambda}$ both statistics are minimized and $S(\hat{\lambda}) = 0$.

Neither the score statistic nor the deviance statistic is affected by re-parameterization of λ. The difference between the two statistics, which is typically small, arises chiefly from two sources, (i) the difference between the observed and expected Fisher information, and (ii) third-order properties of the log-likelihood function. Wald's statistic, which is a quadratic form based on $\hat{\lambda} - \lambda$, is not similarly invariant.

12.4 Smoothing as an aid to informal checks

Some informal checks involve assessing a scatter plot for approximate linearity (or other relation) in the underlying trend. This exercise may be difficult, particularly if the density of points on the x scale varies widely over the range of x values observed. The problem is that where the density of x-values is high, the expected range of y values is larger than in a neighbourhood where the x-values are less dense. The eye finds it hard to make the necessary adjustments, and may be greatly helped if the scatter-plot is augmented by an empirical curve produced by a suitable smoothing algorithm (see, e.g. Cleveland, 1979). Such smoothed curves must be treated with some caution, however, since the algorithm is quite

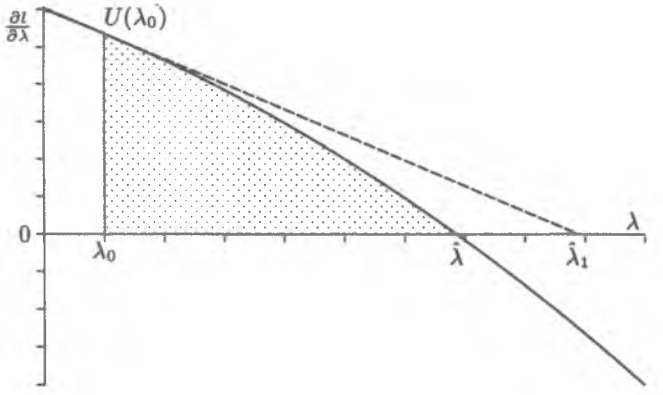


Fig. 12.1a. *The geometry of the score test for one parameter. The solid line is the graph of the log likelihood derivative. The shaded area is one half of the likelihood ratio statistic: the score statistic is twice the area of the triangle* λ_0, $U(\lambda_0)$, $\hat{\lambda}_1$.

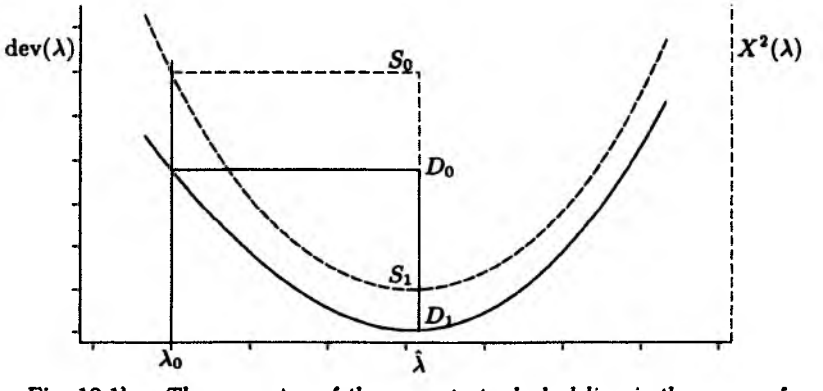


Fig. 12.1b. *The geometry of the score test: dashed line is the curve of* X^2 *using the adjusted dependent variate and weights from the fit at* λ*; solid line is the deviance for varying* λ.

capable of producing convincing-looking curves from entirely random configurations. Nonetheless, smoothing is undoubtedly useful as an aid to informal checks.

12.5 The raw materials of model checking

We consider first linear regression models with supposedly constant variance, where model checking uses mainly the following statistics derived from a fit:

The fitted values $\hat{\boldsymbol{\mu}}$,

The residual variance s^2,

The diagonal elements $\mathbf{h}$ of the projection ('hat') matrix,

$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, which maps $\mathbf{y}$ into $\hat{\boldsymbol{\mu}}$.

An important idea is that of *case deletion*, whereby the fit using all the cases (points) is compared with that obtained when one point is deleted. We shall denote statistics obtained by deleting point i by a suffix in brackets, so that, for example, $s^2_{(i)}$ is the residual variance for the model fitted omitting point i.

A central role is played by the residuals from a fit, and several forms have been proposed in addition to the basic $r = y - \hat{\mu}$. We shall call a residual *standardized* if it has been divided by a factor that makes its variance constant. Standardization produces the form

$$\frac{y_i - \hat{\mu}_i}{\sqrt{(1 - h_i)}},$$

where $\{h_i\}$ are the components of $\mathbf{h}$. If in addition the residual is scaled by dividing by s, we call it a Studentized standardized residual, and write

$$r'_i = \frac{y_i - \hat{\mu}_i}{s\sqrt{(1 - h_i)}}. \tag{12.1}$$

Note that r'^2_i is just the reduction in the residual sum of squares caused by omitting the point i, scaled by the residual mean square for all the points.

Finally there is the important *deletion residual* defined by

$$r^*_i = \frac{y_i - \hat{\mu}_{(i)}}{s_{(i)}\sqrt{(1 + h_{(i)})}} = \frac{y_i - \hat{\mu}_i}{s_{(i)}\sqrt{(1 - h_i)}}. \tag{12.2}$$

in which $\mathbf{x}_i^T$ is the ith row vector, $\mathbf{X}_{(i)}$ is the model matrix with the ith row deleted, and

$$h_{(i)} = \mathbf{x}_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{x}_i = h_i/(1 - h_i).$$

The two residuals r' and r^* are related by

$$r_i^* = r_i' s / s_{(i)},$$

(see Atkinson, 1985), so that r_i^{*2} is again a reduction in the residual sum of squares, but this time scaled by $s_{(i)}^2$ instead of s^2. The deletion residual r_i^* measures the deviation of y_i from the value predicted by the model fitted to the remaining points, standardized and Studentized like r'. (The difference in sign in the terms in the denominators of r' and r^* arises from the fact that y_i and $\hat{\mu}_{(i)}$ are independent, whereas y_i and $\hat{\mu}_i$ are positively correlated.)

For generalized linear models some extensions and modifications are needed to the above definitions. First, where checks on linearity are involved the vectors $\mathbf{y}$ and $\hat{\boldsymbol{\mu}}$ are ordinarily replaced by $\mathbf{z}$, the adjusted dependent variate, and $\hat{\boldsymbol{\eta}}$, the linear predictor. The residual variance is replaced by an estimate of the dispersion parameter ϕ, and the $\mathbf{H}$ matrix becomes

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{\frac{1}{2}}, \tag{12.3}$$

equivalent to replacing $\mathbf{X}$ by $\mathbf{W}^{\frac{1}{2}}\mathbf{X}$ in the regression version. It can be shown that, to a close approximation,

$$\mathbf{V}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \simeq \mathbf{H}\mathbf{V}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}),$$

where $\mathbf{V} = \text{diag}(V(\mu_i))$. Thus $\mathbf{H}$ measures the influence in Studentized units of changes in $\mathbf{Y}$ on $\hat{\boldsymbol{\mu}}$. The corresponding matrix in unstandardized units is $\mathbf{V}^{1/2}\mathbf{H}\mathbf{V}^{-1/2}$, which is an asymmetric projection matrix.

In section 2.4 we defined three forms of residual for generalized linear models, and two of these, the Pearson and deviance residuals, have been widely used in model checking. For the Pearson residual the analogous form of (12.1) is given by

$$r_{\text{P}}' = \frac{y - \hat{\mu}}{\sqrt{\hat{\phi}V(\hat{\mu})(1-h)}}. \tag{12.4}$$

The calculations of Cox and Snell (1968) support a similar standardization for the deviance residual giving

$$r_{\text{D}}' = \frac{r_{\text{D}}}{\sqrt{\hat{\phi}(1-h)}}. \tag{12.5}$$

Exact calculations for deletion residuals may become expensive when iteration is required for every fit. It has become common practice to approximate the quantities involved in the deletion residuals by using one-step approximations. This involves doing one cycle of the fit without point i, starting from the fitted values, weights, etc. from the complete fit. Use of one-step approximations allows certain shortcuts to be made, analogous to those used for regression. We write ${}_1r'_{\mathrm{P}}$ and ${}_1r'_{\mathrm{D}}$ for the one-step approximations to (12.4) and (12.5); thus ${}_1r^2_{\mathrm{P}}$ is the one-step approximation to r^2_{P}, measuring the change in the Pearson χ^2 caused by omitting a point. The analogous approximation for the change in deviance has been shown by Williams (1987) to be given by

$$r^2_{\mathrm{D}} = h\ {}_1r'^2_{\mathrm{P}} + (1-h)\ {}_1r'^2_{\mathrm{D}}. \tag{12.6}$$

An equivalent formula is given by Pregibon (1981, p.720). In general the deviance residual, either unstandardized or standardized, is preferred to the Pearson residual for model checking procedures because its distributional properties are closer to the residuals arising in linear regression models (Pierce and Schafer, 1986).

12.6 Checks for systematic departure from model

We consider first checks for systematic departure from the model, beginning with three residual plots.

12.6.1 *Informal checks using residuals*

If the data are extensive, no analysis can be considered complete without inspecting the residuals plotted against some function of the fitted values. Standardized deviance residuals are recommended, plotted either against $\hat{\eta}$ or against the fitted values transformed to the constant-information scale of the error distribution. Thus we use

$\hat{\mu}$	for Normal errors,
$2\sqrt{\hat{\mu}}$	for Poisson errors,
$2\sin^{-1}\sqrt{\hat{\mu}}$	for binomial errors,
$2\log\hat{\mu}$	for gamma errors,
$-2\hat{\mu}^{-\frac{1}{2}}$	for inverse Gaussian errors.

The argument for the constant-information scale is as follows: for Normal errors if we plot $y - \hat{\mu}$ against $\hat{\mu}$ then the contours of fixed y are parallel straight lines with a slope of -1. With other distributions the contours are curves but the constant-information scale gives a slope of -1 at $r = 0$ to match the Normal case and makes the curvature generally slight. For data with binomial errors, note that $\hat{\mu}$ is interpreted as $\hat{\pi}$ rather than $m\hat{\pi}$.

The null pattern of this plot is a distribution of residuals for varying $\hat{\mu}$ with mean zero and constant range. Typical systematic deviations are (i) the appearance of curvature in the mean and (ii) a systematic change of range with fitted value. Smoothing may be useful in judging whether curvature is present, but cannot help with assessing variable range (see section 12.6.2). Note that this plot is generally uninformative for binary data because all the points lie on one of two curves according as $y = 0$ or 1. Furthermore for $\hat{\mu}$ near zero almost all the points have $y = 0$, and conversely for $\hat{\mu}$ near one.

Curvature may arise from several causes, including the wrong choice of link function, wrong choice of scale of one or more covariates, or omission of a quadratic term in a covariate. Ways of distinguishing between these will be discussed further in sections 12.6.3–4.

A second informal check plots the residuals against an explanatory variable in the linear predictor. The null pattern is the same as that for residuals *vs* fitted values. Again the appearance of systematic trend may indicate the wrong choice of link function or scale of the explanatory variable, or point to a missing quadratic term. Such a trend may also be an artefact caused by a faulty scale in another explanatory variable closely correlated with the one under investigation. Smoothing may help in overcoming the effect of variable density of points.

A third residual plot, known as an added-variable plot, gives a check on whether an omitted covariate, $\mathbf{u}$, say, should be included in the linear predictor. It is not adequate to plot the residuals against $\mathbf{u}$ itself for this purpose. First we must obtain the unstandardized residuals for $\mathbf{u}$ as response, using the same linear predictor and quadratic weights as for $\mathbf{y}$. The unstandardized residuals for $\mathbf{y}$ are then plotted against the residuals for $\mathbf{u}$. If $\mathbf{u}$ is correctly omitted no trend should be apparent.

12.6.2 *Checking the variance function*

A plot of the absolute residuals against fitted values gives an informal check on the adequacy of the assumed variance function. The constant-information scale for the fitted values is usually helpful in spreading out the points on the horizontal scale. The null pattern shows no trend, but an ill-chosen variance function will result in a trend in the mean. Again smoothing may help to see the trend more clearly. A positive trend indicates that the current variance function is increasing too slowly with the mean, so that, for example, an original choice of $V(\mu) \propto \mu$ may need to be replaced by $V(\mu) \propto \mu^2$. A negative trend indicates the reverse effect.

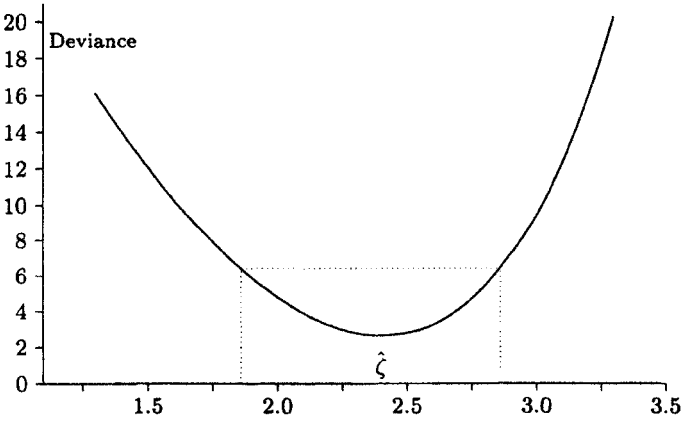


Fig. 12.2. *The profile extended quasi-likelihood curve plotted against ζ for the power-family of variance functions applied to the car-insurance data.*

To check the variance function formally we embed the current one in a suitable family, usually $V(\zeta) = \mu^{\zeta}$, indexed by a parameter ζ, and observe how the fit improves as ζ varies. For this comparison we need the extended quasi-likelihood discussed in section 9.6, which allows the comparison of different variance functions. We compute the deviance for a range of ζ, producing a profile quasi-likelihood curve; approximate likelihood limits are given by the χ^2_1 values for a chosen significance level and the prior value ζ_0 is evaluated in respect of the interval so produced. Fig. 12.2 shows

the curve obtained for the car insurance data (the example in section 8.4.1); the 95% limits for ζ are about $(1.87, 2.85)$ showing that our original choice of $\zeta_0 = 2$ was satisfactory.

12.6.3 *Checking the link function*

An informal check involves examining the plot of the adjusted dependent variable z against $\hat{\eta}$, the estimated linear predictor. The null pattern is a straight line. For link functions of the power family an upwards curvature in the plot points to a link with higher power than that used, and downwards curvature to a lower power. Smoothing may be helpful in interpreting the plot. For binary data this plot is uninformative and formal methods must be used.

There are two formal checks in common use. The simpler (Hinkley, 1985) involves adding $\hat{\eta}^2$ as an extra covariate and assessing the fall in deviance. (A score test may be used as an alternative.) The other formal check involves embedding the link function in a family indexed by a parameter λ and testing the prior value λ_0 in the usual way. Uncertainty about the link function is probably commonest with continuous data having gamma errors, and with proportions having binomial errors. For the former the power family $\eta = \mu^\lambda$ is the most useful. section 11.3 describes the techniques for estimating λ and assessing the adequacy of λ_0.

For binomial data various families have been constructed that include the logistic link (the canonical link) and the complementary log-log link as special cases. Some of these are discussed in section 11.3.2

Checks on the link functions are inevitably affected by failure to establish the correct scales for the explanatory variables in the linear predictor. In particular, if the formal test constructed by adding $\hat{\eta}^2$ to the linear predictor indicates deviation from the model this may point either to a wrong link function, or to wrong scales for explanatory variables or both. The methods described in the next section may help in distinguishing the various alternatives.

12.6.4 *Checking the scales of covariates*

The *partial residual* plot is an important tool for checking whether a term βx in the linear predictor might be better expressed as $\beta h(x; \theta)$ for some monotone function $h(\cdot; \theta)$. In its generalized form

the partial residual is defined by

$$u = z - \hat{\eta} + \hat{\gamma}x$$

where z is the adjusted dependent variable, $\hat{\eta}$ the fitted linear predictor and $\hat{\gamma}$ the parameter estimate for the explanatory variable x.

The plot of u against x provides an informal check. If the scale of x is satisfactory the plot should be approximately linear. If not its form may suggest a suitable alternative. The scatter about any trend may not be uniform, in which case smoothing may help interpretation. The partial residual plot, if smoothed, can be remarkably informative even for binary data. However, distortions will occur if the scales of other explanatory variables are wrong, so that iteration may be necessary in looking at the partial residual plots for several xs. This problem may be less severe with the following formal check which allows simultaneous transformation of several xs to be tested.

As usual the formal check involves embedding the current scale x in a family $h(x;\theta)$ indexed by θ; we then calculate the deviance for a suitable grid of values of θ to find the position of the minimum, which gives $\hat{\theta}$. The fit at $\hat{\theta}$ can then be compared with that at our initial choice of θ_0, which is usually 1. The method is equivalent to the use of a maximum profile-likelihood estimator. Clearly this procedure can be used for several xs simultaneously. This is particularly useful when several xs have the same physical dimensions, so that a simultaneous transformation is likely to be required. By far the commonest family of transformations is the power family given by

$$h(x;\theta) = \begin{cases} \dfrac{x^\theta - 1}{\theta} & \text{for } \theta \neq 0, \\ \log(\theta) & \text{for } \theta = 0. \end{cases}$$

An informal check for a single covariate takes the form of a constructed-variable plot for $v = \partial h/\partial\theta_0$; we first fit a model with v as dependent variable, with the linear predictor and quadratic weight as for y, and form the residuals. We then plot the residuals of y against the residuals of v; a linear trend indicates a value of $\theta \neq \theta_0$, while a null plot would indicate no evidence against $\theta = \theta_0$.

12.6.5 *Checks for compound systematic discrepancies*

So far we have mainly considered checks for a single cause of discrepancy, e.g. a covariate on the wrong scale, a covariate omitted, or a faulty link function. Each of these discrepancies can be tested formally by including an extra variable in the linear predictor and calculating either the deviance reduction or the score statistic for its inclusion; the process is thus analogous to forward selection (Section 3.9). The danger, as usual, is that correlations between the extra variables can lead to each mimicking the effect of the others. The use of backward-selection, where possible, gives a means of avoiding the danger; we now fit all the extra variables, giving the joint effect of all the causes of discrepancy to be tested, and then find the effect of omitting each one in turn from the joint fit. Again either the deviance reduction or the score statistic may be used. Davison and Tsai (1988) give examples of this technique.

12.7 Checks for isolated departures from the model

In this section we look at model-checking procedures associated with particular points in the data, especially those that appear in some way at variance with the pattern set by the remainder. We deal first with the case of a single possibly discrepant point.

For simplicity we consider first data with a response variable y and one explanatory variable x. We assume that an identity link is relevant. The scatter plot of y against x may show an isolated extreme point, which we define loosely as one well apart from the main cluster. There are three types of configuration that are worth distinguishing, and these are shown in Fig. 12.3 with the extreme point indicated by a circle. In Fig. 12.3(a) the x-value of the extreme point is close to the mean. Exclusion of this point has only a small effect on the estimate of the slope, but it substantially reduces the intercept. Its exclusion also produces a big improvement in the goodness of fit.

In Fig. 12.3(b) the extreme point is consistent with the rest, in that a straight line fitted through the rest passes near the extreme point. Inclusion of the extreme point will increase the accuracy of $\hat{\beta}$ without affecting its estimate greatly.

In Fig. 12.3(c) the straight line fitted through the non-extreme points does not pass close to the extreme point, so that if it is

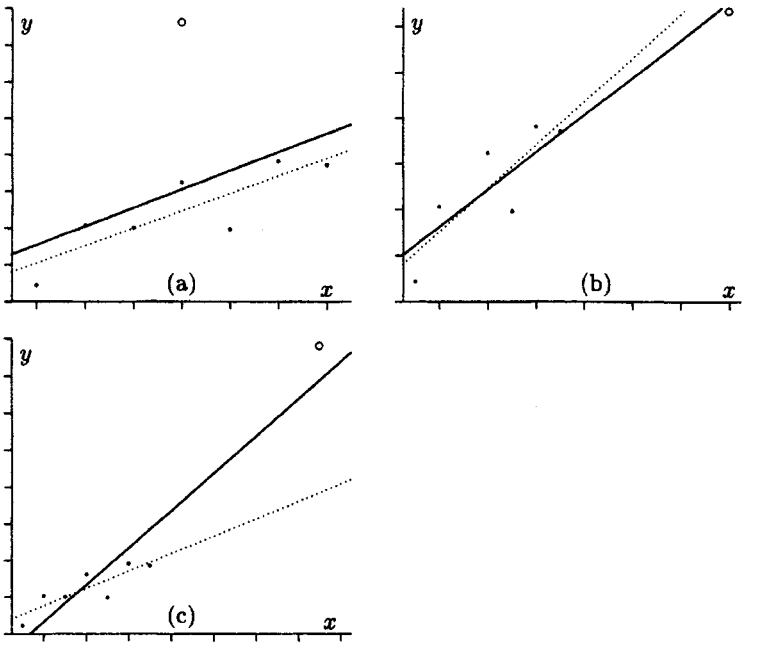


Fig. 12.3. *Scatter plots showing the effect of an extreme point in relation to the configuration of the remaining points. The solid lines are the least squares fits with all points included: the dotted lines are the fits with the extreme points excluded.*

now included in the fit the value of $\hat{\beta}$ will change sharply, and the deviance also.

Three ideas are useful in thinking about the configurations in Fig. 12.3. The first is that of *leverage*, which distinguishes (a) from (b) and (c). The inclusion of the extreme point in (b) and (c) greatly increases the information about $\hat{\beta}$, i.e. the point has high leverage whereas in (a) it has low leverage. The second idea is that of *consistency* which distinguishes (b) from (a) and (c). In (b) the (x, y) values of the extreme point are consistent with the trend suggested by the remainder, while in (a) and (c) they are not. The third idea, termed *influence*, distinguishes (c) from (a) and (b). The extreme point has high influence if the estimate of the slope is greatly changed by its omission, as in (c), and low influence if it is little changed, as in (a) or (b).

We now develop statistics for measuring leverage, consistency and influence quite generally.

12.7.1 *Measure of leverage*

For linear regression models the well-known measure of leverage is given by the diagonal elements of the 'hat' matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{12.7}$$

the ith element of which is

$$h_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i.$$

Note that if the columns of $\mathbf{X}$ are orthogonalized with the first column constant, $\mathbf{H}$ is unaffected. The diagonal elements are then expressible in the form

$$h_i = 1/n + \frac{x_{i2}^2}{\sum x_{j2}^2} + \ldots + \frac{x_{ip}^2}{\sum x_{jp}^2}.$$

Thus $h_i - 1/n$ is an invariant measure of squared distance between $\mathbf{x}_i$ and the centroid of all n points in the x-space.

The general form of $\mathbf{H}$, given in (12.3), has $\mathbf{X}$ replaced by $\mathbf{W}^{\frac{1}{2}}\mathbf{X}$, which effectively allows for the change in variance with the mean. It can also be thought of in an informal sense as the ratio of the covariance matrix of $\hat{\boldsymbol{\mu}}$ to that of $\mathbf{Y}$. Now

$$\sum h_i = \text{trace}\,\mathbf{H} = p,$$

and there is some advantage in working with a standardized form

$$h_i' = nh_i/p$$

so that $\sum h_i' = n$. Hoaglin and Welsch (1978) suggest using $h > 2p/n$, i.e. $h' > 2$, to indicate points of high leverage. An isolated point of high leverage may have a value of h approaching unity. An index plot of h' with the limit $h' = 2$ marked is a useful informal tool for looking at leverage.

Note that for GLMs a point at the extreme of the x-range will not necessarily have high leverage if its weight is very small.

12.7.2 *Measure of consistency*

An inconsistent point is one with a large residual from the curve fitted to the remaining points. Thus the deletion residual introduced in section 12.5 is a natural measure of inconsistency, i.e. small deletion residuals denote consistent points. For generalized linear models the one-step approximation given by (12.6) is appropriate.

12.7.3 *Measure of influence*

Influence can be measured as a suitably weighted combination of the changes $\hat{\beta}_{(i)} - \hat{\beta}$, where $\hat{\beta}_{(i)}$ denotes the estimates without the extreme point, and $\hat{\beta}$ those with it. Cook (1977) first proposed a statistic, which for regression models takes the form

$$D_i = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})/ps^2, \tag{12.8}$$

as a measure of influence of the ith point, when s^2 is an estimate of the dispersion parameter. As shown in Fig. 12.2(a) not all parameters are equally affected by an extreme point, and D_i is intended to provide a suitably weighted combined measure.

From the relation

$$\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}} = -(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i r_i/(1 - h_i)$$

(Atkinson, 1985, p.21), it follows that

$$D_i = \frac{r_i'^2 h_i}{p(1 - h_i)} \tag{12.9}$$

showing that D_i is a function of the quantities involved in the measurement of leverage and consistency.

Atkinson (1981) suggests modifications to D which have advantages in standardizing it for different configurations of $\mathbf{X}$ and making extreme points stand out more sharply. First he replaces r'^2 in (12.9) by r^{*2}, which is equivalent to the use of $s^2_{(i)}$ in place of s^2. Secondly he scales by a factor $(n - p)/p$; this has the effect of making the modified D_i equal to r^{*2} when all points have equal leverage. Finally he takes the square root, producing the modified Cook statistic

$$C_i = \left\{\frac{n - p}{p} \cdot \frac{h_i}{1 - h_i}\right\}^{\frac{1}{2}} |r_i^*|. \tag{12.10}$$

To adapt these statistics for use with generalized linear models is straightforward. D_i is now defined by

$$D_i = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})(\mathbf{X}^T\mathbf{W}\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})/p\hat{\phi},$$

where the $\hat{\boldsymbol{\beta}}_{(i)}$ will usually be the one-step approximations discussed in section 12.5. The modified Cook statistic C_i can be adapted by simply replacing r^* by ${}_1r^*_D$, the one-step approximation to the deletion deviance residual.

12.7.4 *Informal assessment of extreme values*

The three statistics h, r^* and C, introduced above for the measurement of leverage, consistency and influence respectively, each yield a vector of n values. The interesting values for model-checking purposes are the large ones (of either sign in the case of r^*). To interpret these we need plots that allow for the fact that we have chosen the most extreme values to examine. We thus need some measure of how large the extreme values would be in a sample of a given size even if no unusual points were present.

The simple index plot of the statistic against case number does not have this property, but it has value, particularly when a few points are far from the rest. Normal plots, which make allowance for selection effects, come in two forms, the half-Normal plot and the full Normal plot. The former is appropriate for non-negative quantities like h and C; for a statistic like r^*, there are two options, either a half-Normal plot of $|r^*|$ or a full Normal plot of r^* itself. For either plot the ordered values of the statistic are plotted against the expected order statistics of a Normal sample. The latter may be generated with sufficient accuracy for practical purposes by

$$\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right) \qquad i = 1, \ldots, n$$

for the full Normal plot, and by

$$\Phi^{-1}\left(\frac{n + i + \frac{1}{2}}{2n + \frac{9}{8}}\right) \qquad i = 1, \ldots, n$$

for the half Normal plot. Note that the use of either of these scales is to some extent conventional, for, while the plot for r^* may, in

the absence of unusual points, be approximately linear, there is usually no reason to expect the h or C plots to be so. Extreme points will appear at the extremes of the plot, possibly with values that deviate from the trend indicated by the remainder.

To aid the interpretation of Normal plots Atkinson (1981) developed the useful idea of an envelope constructed by simulation. For generalized linear models with a fully specified error distribution this is constructed as follows: for each simulation form pseudo-data $\mathbf{y}^*$ by generating random variables from the appropriate error distribution with mean $\hat{\mu}$ and dispersion $\hat{\phi}$. Refit the model and calculate the statistic. Order its values. Do k simulations and for each ordered position select the extreme values from the k simulations. Plot these with the original points to give the envelope. More stable envelopes can be obtained, if simulation is cheap, by using larger values of k and less extreme values of the ordered samples.

Atkinson (1985) gives detailed examples on the interpretation of plots with envelopes for regression models. Simulation is particularly simple here, because the variance is independent of the mean, so that $\hat{\mu}$ can be disregarded, and y^* requires just $N(0,1)$ variables. Note that simulation for models with non-Normal errors can be speeded up using a one or two-step approximation to the full iteration beginning with $\hat{\mu}$ as the initial estimate of the fitted values.

Residuals from data in the form of counts or proportions will show distortions if there are many zeros (counts) or zeros and ones (proportions). These produce a concentration of small residuals near zero, which may appear as a plateau in the Normal plot.

12.7.5 *Extreme points and checks for systematic discrepancies*

Up to now we have divided model-checking techniques into those for systematic and those for isolated discrepancies. However it is possible to formulate questions that involve both kinds of discrepancy. For example we might ask 'does the evidence for the inclusion of a covariate depend largely on the influence of a few isolated points?'. One way of answering such a question has been given by Williams (1987); see also Davison and Tsai (1988).

Consider a test by backward selection (Section 12.6.5) for a systematic discrepancy as measured by the extra variable $\mathbf{u}$. Suppose that the full linear predictor gives squared residuals $r^2_{\mathrm{G}1}$, and that

without $\mathbf{u}$ the residuals are r^2_{G0}. Then the differences $r^2_{G0} - r^2_{G1}$ can be used in an index plot to show the influence of each point on this test for systematic discrepancy. Such a plot might reveal, for example, that most of the evidence for the effect in question comes from one or two points that have previously been identified as possible outliers. If required, the analysis may be repeated with suspected outliers omitted.

12.8 Examples

12.8.1 *Damaged carrots in an insecticide experiment*

The data shown in Table 12.1, taken from Phelps (1982), are discussed by Williams (1987). They give the proportion of carrots showing insect damage in a trial with three blocks and eight dose levels of insecticide. With a logit link function and simple additive linear predictor $block + x$, where x is the log dose, we find a deviance of 40.0 with 20 d.f., rather too large for binomial variation.

Table 12.1 *Proportion of carrots damaged in an insecticide experiment*

Dose		*Block*		
level j	*log dose* x_j	1	2	3
1	1.52	10/35	17/38	10/34
2	1.64	16/42	10/40	10/38
3	1.76	8/50	8/33	5/36
4	1.88	6/42	8/39	3/35
5	2.00	9/35	5/47	2/49
6	2.12	9/42	17/42	1/40
7	2.24	1/32	6/35	3/22
8	2.36	2/28	4/35	2/31

Source: Phelps (1982).

There may be general over-dispersion or perhaps isolated extreme points. Fig. 12.4 shows an index plot (with the data ordered by columns) of the one-step deletion residual r^*. This plot quickly decides the issue; point 14 (dose level 6 and block 2) is far away from the rest. The fit omitting this point gives a deviance of 25.3 with 19 d.f. Though somewhat above the baseline of 19, it is clearly

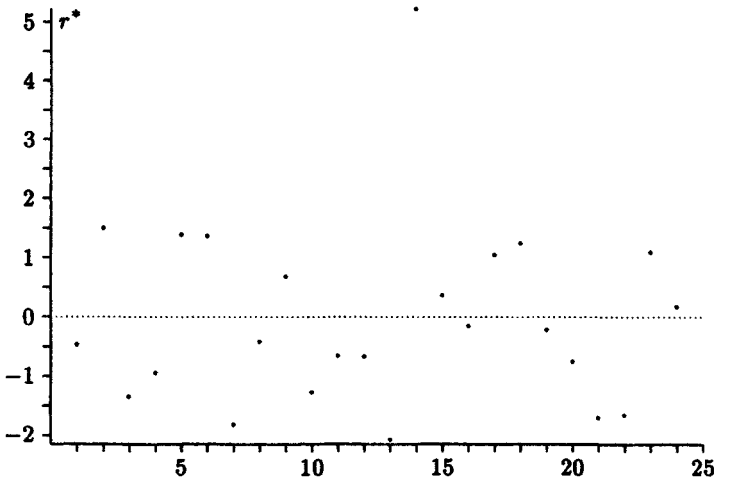


Fig. 12.4 *Index plot of one-step deletion residuals r^* for carrot data, showing a single outlier.*

a great improvement on the original fit. The fitted value for point 14 is much closer to 7 than to the 17 recorded.

Inclusion of the constructed variable $\hat{\eta}^2$ after omitting point 14 gives an insignificant reduction (0.2) in the deviance, so that our choice of link function is not contradicted. (Phelps used the complementary log-log link, but the difference in fit between it and the logistic is small.) This example thus illustrates the effect of an isolated extreme point having an anomalous y-value.

12.8.2 *Minitab tree data*

This famous set of data on the volume, diameter (at 4′ 6″ above ground level), and height of black cherry trees was given by Ryan *et al.* (1976). Interest attaches to deriving a formula to predict tree volume v from measurements of diameter d and height h. If all the trees were the same shape we would expect to find

$$v = c \times d^2 \times h \tag{12.11}$$

for some constant c. Thus we might expect that a successful linear model would involve $\log v$ as response variable with $\log d$ and $\log h$ as explanatory variables.

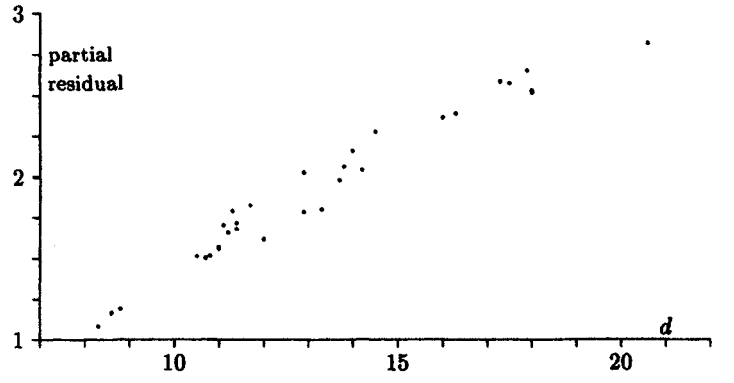


Fig. 12.5a. *Partial residual plot for d in the joint fit of d and h to* $\log(v)$.

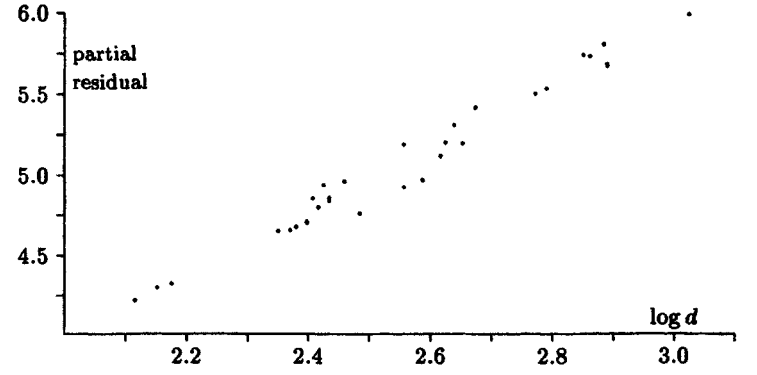


Fig. 12.5b. *Partial residual plot for* $\log d$ *in the joint fit of* $\log d$ *and* $\log h$ *to* $\log(v)$.

Suppose, however, that someone with no understanding of dimensionality begins by regressing v linearly on d and h with Normal errors. The deviance is 421.9 with 28 d.f. The $(y, \hat{\mu})$ plot is curved upwards and the $(r, \hat{\mu})$ plot is quadratic. Addition of $\hat{\eta}^2$ to the linear predictor decreases the deviance by an enormous 242.6, about 57.5%.

We can now either transform y to a lower power, or choose an equivalent link function; for simplicity we shall follow the first path, and examine the effect of using $\log v$ in place of v. The result of adding $\hat{\eta}^2$ is still appreciable, the deviance falling from 0.262

to 0.181, though proportionally much less than before, while the absolute residual plot shows no obvious pattern. Remembering that the test based on adding $\hat{\eta}^2$ to the linear predictor may reflect either a faulty link or a faulty covariate scale, we next look at the partial residual plot of d, the more important of the two explanatory variables. This is shown in Fig. 12.5(a); it is curved downwards. That for h is more scattered and does not deviate obviously from linearity. The partial residual plot for d suggests a lower power for d and so we try $\log d$. Given that the dimensions of d and h are identical, external considerations suggest that we transform h to $\log h$ at the same time. The deviance is now 0.185, very similar to that given by the $\hat{\eta}^2$ test with d and h, and the addition of $\hat{\eta}^2$ does not further improve the fit. Both partial residual plots look linear and that for $\log d$ is shown in Fig. 12.5(b). There is no monotone trend in the absolute-residual plot, though all the big residuals are for points in the intermediate range. The formal test for the joint power transformation of d and h to d^θ and h^θ gives the deviance curve shown in Fig. 12.6. The minimum is at about $\hat{\theta} = 0.15$ where the deviance is 0.1829 with 27 d.f. giving $s^2 = 0.006530$. This gives 95% limits for the deviance of $0.1829 + 4s^2 = 0.2088$, corresponding to limits for θ of $(-0.32, 0.63)$. This excludes the original $\theta = 1$ and includes the final $\theta = 0$, corresponding to the log transformation.

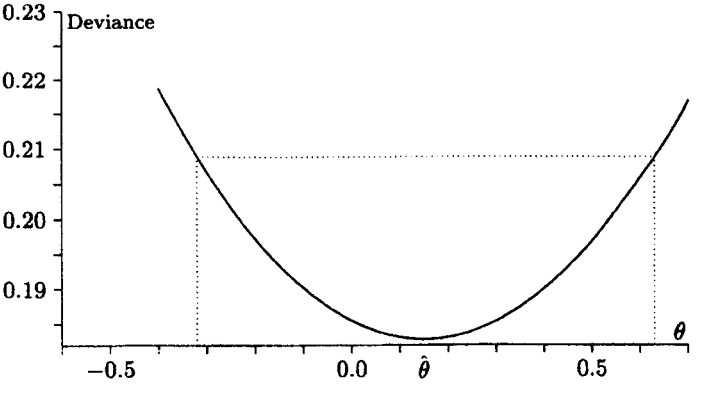


Fig. 12.6. *Minitab tree data: deviance for joint transformation of d and h to d^θ and h^θ*

Note that our original guess at the relation, equation (12.11)

predicts coefficient for $\log d$ and $\log h$ of 2 and 1 respectively. Use of these in the fit gives a deviance of 0.1877 with 30 d.f., which is trivially different from 0.185 found from the model with the parameters estimated from the data.

Checks on isolated discrepancies show that:
trees 20 and 31 have high leverage;
trees 15 and 18 have the largest negative deletion residuals;
trees 11 and 17 have the largest positive deletion residuals;
tree 18 has a large modified Cook statistic.

The fit omitting point 18 does substantially reduce the deviance (to 0.154), but affects markedly only the intercept among the parameters. There is a curious cluster of extreme residuals for trees 15–18, which may be accidental. On the whole these latter checks do not suggest the rejection of any of the trees from the final model.

12.8.3 *Insurance claims (continued)*

In the initial analysis of these data in Chapter 8 a model was fitted using gamma errors and the inverse link. Subsequently the link function was embedded in the family $g(\mu;\lambda) = \mu^\lambda$ (Section 11.3.1), and separately the variance function was embedded in the same power family $V(\mu;\zeta) = \mu^\zeta$ (Section 12.6.2). The original choice of $\lambda_0 = -1$ and $\zeta_0 = 2$ was thus compared with the best fitting $\hat{\lambda}$ with $\zeta = \zeta_0$ fixed, and also with $\hat{\zeta}$ for $\lambda = \lambda_0$ held constant. We now consider a formal check on the joint settings (λ_0, ζ_0) against the best fitting $(\hat{\lambda}, \hat{\zeta})$ when both parameters are allowed to vary. The criterion is the extended (quasi) deviance when ϕ is also estimated, namely

$$\sum_i \log(\hat{\phi}\, V(y_i;\zeta))$$

where $\hat{\phi}$ is estimated by the mean deviance.

The contours are shown in Fig. 12.7 for the χ^2_2 values for $p = 0.50, 0.80, 0.95$ and 0.99. The minimum occurs at $\zeta = 2.4$ and $\lambda = 0.75$, with the original choice of $(2, -1)$ lying comfortably inside the 95% contour. The fit for $(2, 0)$, i.e. with a log link, is less good, but again lies within the 95% contour. Note that the axes of the contours are closely aligned to the axes of (ζ, λ), showing that the parameters are effectively orthogonal. Thus conservative 95% limits can be obtained by projecting onto the axes.

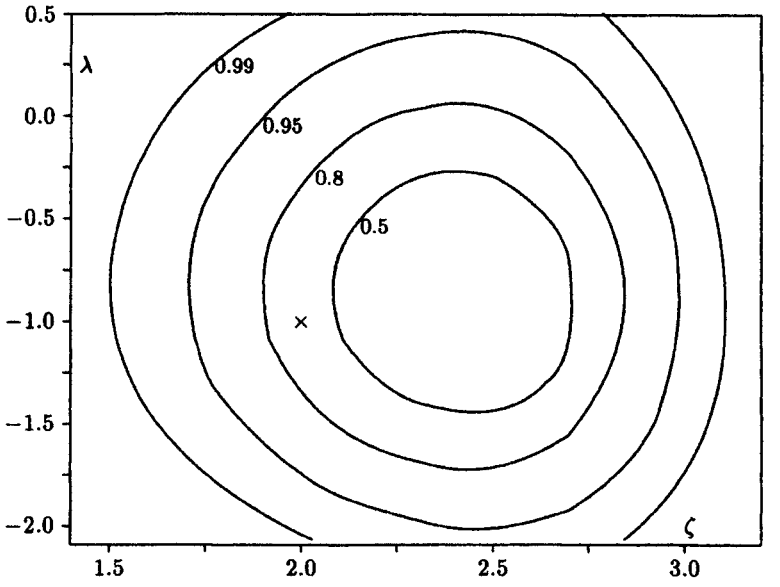


Fig. 12.7 *Contour plot of the extended profile quasi-deviance for* (ζ, λ) *for the car-insurance data. The contours correspond to nominal confidence levels of* 0.5, 0.8, 0.95 *and* 0.99.

12.9 A strategy for model checking?

This chapter has presented a number of techniques, both formal and informal, for checking the internal consistency of models, looking for both systematic deviation and for isolated unusual points. In principle we might hope to develop a strategy in which the various techniques are applied in an algorithmic fashion to give a complete check with accurate diagnoses of any deviations present. In practice such a strategy seems a long way off and model checking remains almost as much art as science. A major problem lies in the complex way that different deviations can interact. Thus a goodness-of-link test may give a significant result because of a faulty choice of link function; but it may also fail because one or more covariates is mis-scaled, or an interaction term is missing, or because of the presence of a few 'bad' points. Similarly an inconsistent point of high leverage may reflect the fact that the chosen model is breaking down at the edge of the treatment space,

or that someone made a recording error or a mis-transcription. ('All interesting points are wrong' is one cynic's view). Again many model-checking methods for extreme points are heavily dependent on the extreme points being fairly isolated. The occurrence of a small clump of such points may be much harder to identify. One promising possibility, for regression models, is the method of least median of squares, in which a very robust fit is used to identify such clumps of extreme points: see Atkinson (1986) and Rousseeuw and Leroy (1988). The action to be taken after identifying an inconsistent point is itself heavily dependent on the context of the problem and the special knowledge of the analyst. These considerations all lead to the presence of the question mark in the section heading.

12.10 Bibliographic notes

For an early account of methods for the examination of residuals, see Anscombe (1961) and Anscombe and Tukey (1963).

Most recent published work is for linear regression models, usually under the heading of regression diagnostics. Atkinson (1985) gives a very readable account, with some references to generalized linear models, though he mostly prefers the route via data transformation rather than the GLM specification involving link functions for the mean μ. See also the books by Cook and Weisberg (1982), Belsley *et al.* (1980) and Hawkins (1980).

Cox and Snell (1968) discuss residuals in a very general context including topics such as standardization and non-linear transformations. Goodness-of-link tests are discussed by Pregibon (1980) and by Atkinson (1982). The technique of using a constructed variable for detecting departures of a specific nature goes back to Tukey (1949). Other useful references are Andrews and Pregibon (1978) and Pregibon (1979).

Cook (1977, 1979) introduced the notion of influential observations in regression. Williams (1987) deals explicitly with methods for GLMs and introduces the modified form of deviance deletion residual. Chatterjee and Hadi (1986) review work in regression, while Kay and Little (1987) and Fowlkes (1987) deal specifically with the awkward case of binary data.

12.11 Further results and exercises 12

12.1 Schreiner Gregoire and Lawrie (1962) conducted an experiment to examine the effect of supposedly inert gases on fungal growth. Their data were presented in graphical form only: this report and the data are taken from Bliss (1967).

The fungus *Neurospora crassa* was grown at 30°C on an agar medium in tubes filled with an inert gas containing approximately 5% oxygen. The following growth rates in millimetres per hour, each the mean of 5 or 6 tests, were thought to be related to a suitable function of the molecular weight (MW) of the inert gas.

Table 12.2 *Rate of growth of the fungus* Neurospora crassa

Gas	He	Ne	N_2	N_2	Ar	Ar	Kr	Kr	Xe	Xe
MW	4.0	20.2	28.2	28.2	39.9	39.9	83.8	83.8	131.3	131.3
mm/hr	3.51	3.14	3.03	2.83	2.71	2.76	2.27	2.17	1.88	1.85

Source: Bliss (1967), p.471.

1. Plot the growth rate against MW and $(\text{MW})^{1/2}$. Comment.
2. Taking the growth rate, R, as the response fit linear models using identity, log and reciprocal links, combined with various power transformations (identity, $x^{2/3}$, $x^{1/2}$, $x^{1/3}$ and log) of MW. For which combinations is the residual deviance smallest?
3. Interpret in biological terms the combination of powers obtained in part 1. What possible physical interpretation could be given to the $x^{2/3}$ transformation?
5. Examine the deviance residuals, first for pattern in the plot against MW, and second for conformity marginally with the Normal distribution.
5. Is the estimate of residual variance consistent with that obtained from the four replicate pairs?
6. Schreiner *et al.* report that

$$R = 3.88 - 0.1785(\text{MW})^{1/2}.$$

Is this summary consistent with your findings?

Table 12.3 *Relation in male cats of heart weight in gm. to body weight in kg.*

Body wt.	*Heart weight (gm.)*												
1.7	6.5	7.0											
1.8	5.8	7.3	6.1	7.1	7.7	7.4							
1.9	8.1	9.1	8.0	7.2	7.3	8.0							
2.0	6.5	6.5	6.7	7.5	7.8	8.1	8.6	7.7					
2.1	10.1	7.0	7.2	8.1	8.3								
2.2	7.2	7.6	10.7	9.6	9.1	7.9	8.5	9.6	8.9				
2.3	9.6	9.6	8.5	8.8	8.2	9.2	8.7	8.9					
2.4	9.3	9.1	7.3	7.9	7.9	9.6	9.1	9.0	10.8	9.6			
2.5	8.8	12.7	8.6	12.7	9.3	7.9	11.0	8.8	9.3	8.2	8.7	10.4	9.6
2.6	10.5	8.3	9.4	7.7	11.5	9.4	13.6	10.1	10.9	9.6	9.9		
2.7	12.0	10.4	8.0	9.6	9.6	9.8	12.5	9.0	11.1	10.5	11.6	11.9	
2.8	10.0	12.0	13.5	13.3	9.1	10.2	11.4	10.1	10.9				
2.9	9.4	11.3	10.1	10.6	11.8								
3.0	13.3	10.0	13.8	10.6	12.4	12.7	10.4	11.6	12.2				
3.1	9.9	12.1	14.3	12.5	11.5	13.0							
3.2	11.6	13.6	12.3	13.0	13.5	11.9							
3.3	11.5	14.9	14.1	15.4	12.0								
3.4	14.4	12.2	12.8	11.2	12.4								
3.5	15.6	11.7	15.7	12.9	17.2								
3.6	14.8	13.3	15.0	11.8									
3.7	11.0												
3.8	14.8	16.8											
3.9	14.4	20.5											

Source: Chen, Bliss and Robbins (1942)

12.2 Repeat the analysis of the data in Table 12.1, using an index plot of the deviance residuals to detect outliers. Compare this plot with that in Fig. 12.4. Comment on the differences and similarities.

12.3 Chen, Bliss and Robbins (1942) obtained the data shown in Table 12.3 as part of an assay experiment comparing the effect of *calotropin* with other cardiac substances such as uscarin and ouabain. Following the experiment the animals' hearts were weighed to see whether the cardiac effect might be more closely related to heart weight than to body weight. Table 12.3 shows the relationship between heart mass and body mass for 149 male cats used in the experiment.

1. Plot heart weight against body weight. Comment.

2. Fit the regression model of heart weight against body weight. Are the data consistent with a straight-line model passing through the origin? Fit the model passing through the origin.
3. Plot the residuals against body weight for the models fitted in part 2. Comment.
4. Regress log(heart weight) against log(body weight). Are the data consistent with the hypothesis that the slope is unity? What is the physiological interpretation of a unit slope? Fit the model in which the slope is unity and examine the residuals graphically.
5. Compute the mean of log(heart weight) for each of the 23 distinct values of body weight. Regress these sample means on log(body weight) using a weighted linear regression model. Compare the parameter estimates and standard errors in this weighted regression with those obtained in part 4. Explain the similarities and discrepancies observed.
6. For the model fitted in part 5, plot the residuals against the fitted values, taking care to use an appropriate standardization. Comment on this plot and on its relation to the plot in part 3.
7. Test the adequacy of the linear regression model in part 4 by including a non-linear term in log(body weight).
8. Give a brief summary of your findings.