The Wrong Model

Part I: how to spot it

J N S Matthews
Biostatistics Research Group,
Newcastle University

1 Some well-known background

This document focuses on the linear model, and as such is largely relevant to continuous data. We will be concerned with what model checks are needed and how best to carry these out. Many things in this document will be well-known and will refer to things that are very familiar when using standard software - especially when making choices for things like the types of residuals. The emphasis here will be on some of the theoretical and mathematical aspects that have been used in the development of these things, and which by now might be a slightly hazy memory.

Throughout it is supposed that there are n observations, $y_i, i = 1, \ldots, n$ and that these can be written as an n-dimensional vector $\boldsymbol{y}$. It is supposed that there are p covariates and that the covariates corresponding to observation i are written as a p-dimensional vector $\boldsymbol{x}_i$ and these can be grouped in an $n \times p$ matrix $\boldsymbol{X}$, where the $\boldsymbol{x}_i^T$ form the rows, i.e.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}$$

The number of covariates is p and it is tacitly assumed in much of the following that an intercept is included, i.e. the first element of each $\boldsymbol{x}_i$ is 1. In parts of the literature the number of covariates is sometimes written as $p+1$ to emphasise the inclusion of an intercept, with p now being the number of covariates other than the intercept.
Of course, the vast majority of models that we fit in practice include an intercept. The reason for drawing attention to the usual presence of an intercept is that quite a few of the very familiar results we will encounter are true only when an intercept is included, and this can easily be overlooked. For example, the residuals from a model fitting a single (non-constant) covariate without an intercept, do not sum to zero.
The usual model that we fit is

$$y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i, \tag{1}$$

where the *residual*, ϵ_i, has zero mean and variance σ^2 and $\boldsymbol{\beta}$ is the p-dimensional vector of the parameters associated with the p covariates. It is also assumed that the residual terms corresponding to different is are uncorrelated. These conditions are often supplemented with the assumption that the ϵ_i are Normally distributed (and hence independent), which permits certain hypothesis tests and interval estimates to be specified.

A more succinct description of key results is obtained if we write (1) in matrix terms as

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

where $\boldsymbol{\epsilon}$ is the n-dimensional random vector of the ϵ_is. If the residuals are assumed to be Normally distributed then the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is well-known to be

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}; \tag{2}$$

here and throughout we assume that $n \geq p$ and $\boldsymbol{X}$ is of full rank p. The *fitted* values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{y}$, where

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

is the projection matrix onto the column space of $\boldsymbol{X}$, sometimes called the 'hat' matrix (and so some authors use $\boldsymbol{H}$ in place of $\boldsymbol{P}$). It should be noted that $\boldsymbol{P} = \boldsymbol{P}^T$ and $\boldsymbol{P}$ is idempotent, i.e. $\boldsymbol{P} = \boldsymbol{P}^2$.

The estimated residuals, $\hat{\boldsymbol{\epsilon}}$ are defined as $\boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$, where $\boldsymbol{I}$ is the identity matrix of appropriate dimension. Note that $\mathsf{E}[\hat{\boldsymbol{\epsilon}}] = (\boldsymbol{I} - \boldsymbol{P})\mathsf{E}[\boldsymbol{y}] = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$, i.e. the estimated residuals have mean zero.

It should also be noted that it follows that $\mathsf{cov}[\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{y}}] = \mathsf{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{y}}^T]$, and this is $(\boldsymbol{I} - \boldsymbol{P})\mathsf{E}[\boldsymbol{y}\boldsymbol{y}^T]\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T + \sigma^2\boldsymbol{I})\boldsymbol{P} = 0$. That is, assuming the model it correct, the fitted values and estimated residuals are uncorrelated (so independent if Normality is assumed).

The estimated residuals have some constraints. If $\boldsymbol{x}_j^C$ denotes the jth column of $\boldsymbol{X}$, then $\hat{\boldsymbol{\epsilon}}^T\boldsymbol{x}_j^C = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}_j^C$. Now as $\boldsymbol{x}_j^C$ is in the column space of $\boldsymbol{X}$, then $\boldsymbol{P}\boldsymbol{x}_j^C = \boldsymbol{x}_j^C$, so $\hat{\boldsymbol{\epsilon}}^T\boldsymbol{x}_j^C = 0$. If the model contains an intercept, then $\boldsymbol{x}_1^C = \boldsymbol{1}$, where $\boldsymbol{1}$ is a vector of ones of suitable dimension, so it follows that $\sum \hat{\epsilon}_i = 0$, i.e. the residuals sum to zero and hence have a sample mean of 0.

The sum of squares of the estimated residuals, $\sum \hat{\epsilon}_i^2$, provides the basis for the estimation of σ^2. Dividing by n gives the MLE, but it is more usual to divide by $n - p$ to get an unbiased estimator of σ^2. To see this note [1]

$$\begin{aligned}\mathsf{E}\left[\sum \hat{\epsilon}_i^2\right] &= \mathsf{E}[\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}] = \mathsf{E}[\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}] = \mathsf{E}[\mathrm{tr}((\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)] \\ &= \mathrm{tr}((\boldsymbol{I} - \boldsymbol{P})\mathsf{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]) = \sigma^2\mathrm{tr}(\boldsymbol{I} - \boldsymbol{P}) = \sigma^2(n - p)\end{aligned}$$

(For an explanation why $\mathrm{tr}(\boldsymbol{P}) = p$, see 3.4

1.1 The Gauss-Markov theorem

The estimator (2) has the optimal properties that comes from maximum likelihood, provided the $\boldsymbol{\epsilon}$ have a Normal distribution. The form of the Normal density means that (2) is also the *least squares* estimator (LSE). While it is geometrically plausible that the LSE performs well, even if $\boldsymbol{\epsilon}$ is not Normal, can more be said?

The answer is provided by the *Gauss-Markov* theorem. If the ϵ_i are independent, with common variance, then (2) has the smallest variance among all estimators that are linear in $\boldsymbol{y}$ and unbiased for $\boldsymbol{\beta}$. To see this, note that

[1] We use the matrix identity $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$, where $\mathrm{tr}(\cdot)$ denotes the *trace*, or sum of diagonal elements, of the matrix

$\boldsymbol{Ay}$ is unbiased for $\boldsymbol{\beta}$ if $\boldsymbol{AX} = \boldsymbol{I}$, and has variance $\sigma^2 \boldsymbol{AA}^T$. Also

$$\boldsymbol{AA}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{AA}^T - \boldsymbol{AX}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^T = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{A}^T. \quad (3)$$

As $\boldsymbol{P}$ is idempotent and symmetric, so is $\boldsymbol{I} - \boldsymbol{P}$ and from this is follows that the final term in (3) is positive definite, so the variance of any contrast based on $\boldsymbol{Ay}$ exceeds that based on the LSE.

So, if our model checking reveals non-Normality, but no heteroscedasticity, then the estimator (2), still has much to commend it.

2 Model checking

2.1 Some general considerations

Ideally any statistical model should be chosen to:

1. enable the analyst to address the questions being posed;
2. take account of knowledge of the subject area;
3. take into account any lessons that may be available from previous analyses in the area.

What might these worthy aims actually mean in practice? This is likely to vary between areas of application. For clinical trials, items 1 and 2 would include things like ensuring that broadly the right covariates are specified, while restricting the range of models being considered to ensure that issues of multiplicity are kept in check. Items 2 and 3 might lead the investigator to choose appropriate scales for measurement. Some variables such as skin-fold thickness and bilirubin concentration, are known to be skewed, so logging them may be helpful, whether they are response variables, when transformation may provide better distributional properties, or are covariates, when transformation may prevent undue influence of larger values.

Item 1 also implies that there may be specific parameters, such as a treatment effect in a clinical trial, that are of overriding importance. In such cases, the extent to which interval estimates of the parameter vary between alternative, plausible models will determine the importance of the precise choice of model.

2.2 Model checking - diagnostics

One might hope that, for a carefully chosen model, most of the assumptions are going to be right, or close to right. Nevertheless, it will usually be prudent to conduct some checks on the fitted model. This will usually entail fitting a model, computing suitable quantities known as model *diagnostics* and then assessing whether these diagnostics have the properties that they should have if the fitted model is broadly correct.

The hope is that the form of any departures from the expected properties will indicate how the model might be amended so that its assumptions more nearly hold true. Sometimes the departures will indicate that general aspects of the model need to be changed, such as transforming the response or altering the form of a covariate. However, such diagnostics are often very good at identifying individual points in the data that are out of line with the rest of the data. Omitting such points is a possibility, but there may be circumstances where such a step might be difficult to justify. With such unusual cases it is helpful to assess whether they have played a significant role in the determination of the model parameters, or even on the choice of detailed form of the model - i.e. the *influence* of the point needs to be considered.

The assumed model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, has several components, each of which needs to be considered

1. The form, $\boldsymbol{X}\boldsymbol{\beta}$, of the assumed mean of $\boldsymbol{y}$ can be assessed by simple graphs. Another approach is to fit an extended model and see if this improves the fit, e.g. by adding a quadratic term to a linear term and testing if there is evidence that the coefficient of the quadratic term is not zero.

2. An alternative is to compute the *estimated residuals*[2], $\hat{\epsilon}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, and plot these against variables in the model, or variables which might have been included. Patterns in the plots can indicate if a term should have been included or if its form needs to be modified.

[2]Hereafter, for succinctness and when the context makes it clear, *residual* may mean either ϵ_i or $\hat{\epsilon}_i$.

3. The ϵ_i are assumed to be independent, have zero mean, constant variance and, perhaps, a Normal distribution. The estimated residuals are the natural quantities to assess these assumptions.

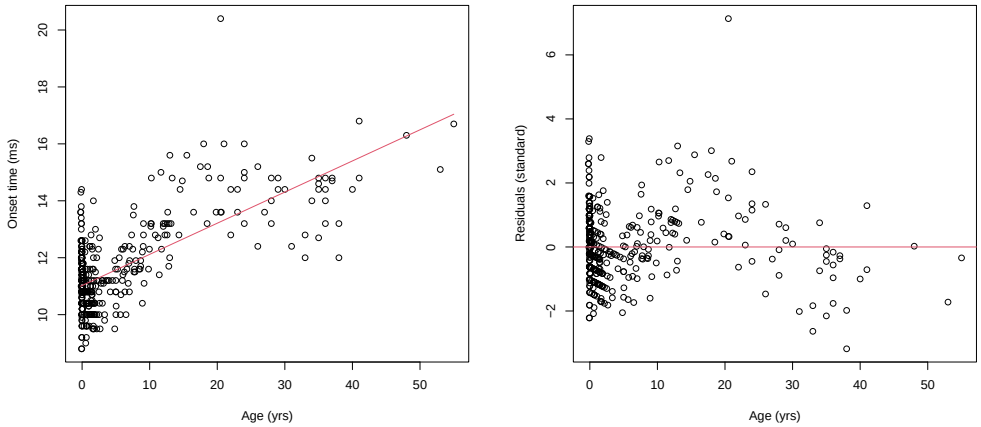
   (a) In many applications the assumption of independence of the ϵ_i is taken as read from the context. Common situations where this might be violated, such as repeated measures on an individual or observations within some group, such as a cluster in a cluster randomized trial, will usually have been noticed during the design of the study, with the analysis specified accordingly.

   (b) The fitted residuals will have zero mean, i.e. $\sum \hat{\epsilon}_i = 0$ because of the properties of the usual methods of estimation, *provided* that the model contains an intercept.

   (c) Assessing constancy of variance, and if needed, Normality, can be based on the estimated residuals, but a number of issues arise that complicate matters, at least in principle. Much of the next section will focus on these.

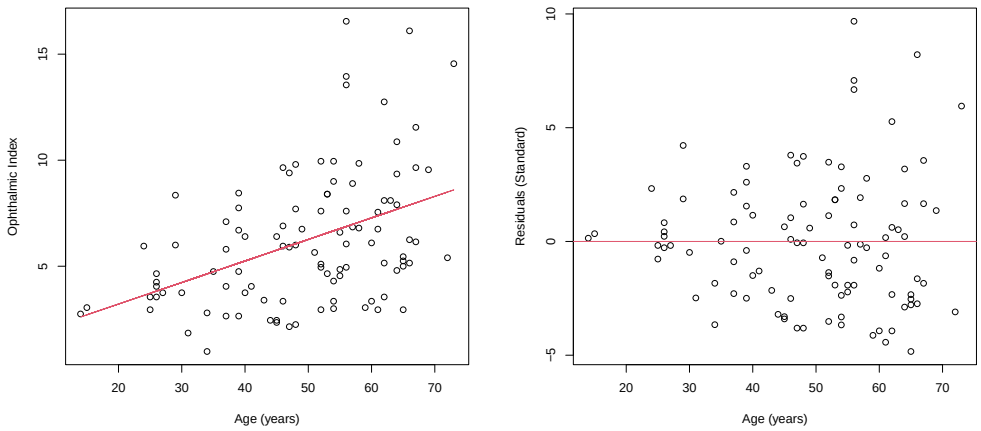
4. Common ways to use residuals to assess model fit are generally graphical. Plots of the $\hat{\epsilon}_i$ against covariates, or possible covariates can be useful. As with most of these residuals plots, no pattern is what is sought - if the model is correct then the ϵ_i are just independent noise. Patterns can suggest a missing covariate should be included, or an included covariate is not in the right form. Patterns might also suggest that the ϵ_i do not have constant variance. If the model is correct then we have seen that the estimated residuals, $\hat{\epsilon}_i$ and the fitted values, $\hat{y}_i$ are uncorrelated and a plot of these quantities should exhibit no relationship. Two example will illustrate some of these features.

   The first example concerns the onset times (in ms) for evoked muscle contractions against age (with term being 0 years and determinations in preterm infants using negative ages): the data are related to those in O'Sullivan et al. (1998). Figure 1(a) shows the data and a fitted regression line, whereas Figure 1(b) plots the residuals against age. This shows a clear pattern, with almost only positive residuals in late teenage to early twenties and largely negative values in the later ages. If the ϵ_i were genuinely uncorrelated, with zero mean and constant variance, then this would not happen. In fact, looking at Figure 1(a) and



(a) Linear fit to onset times (b) Residuals vs age for fit shown left

Figure 1: Data on onset times related to study in (O'Sullivan et al., 1998)



(a) Linear fit age to OI (b) Residuals vs age

Figure 2: Data on ophthalmic index (OI) versus age in 101 patients with Graves' ophthalmopathy (Perros et al., 1993)

appreciating the area of application, it is likely that a curve rising to an asymptote would be far more suitable than a straight line.

The second example concerns data on 101 patients with Graves' ophthalmopathy (Perros et al., 1993). The ophthalmic index (OI) is a measure of the ophthalmic health of the individual, with larger values corresponding to poorer ophthalmic performance. The issue here seems to be that the spread of the data about the line increases for older patients. Although this can be seen in Figure 2(a), it is clearer when residuals are plotted against age, as in Figure 2(b). This suggests that the assumption of constant variance of the ϵ_i is questionable.

5. If the analyst wishes to use the assumption of Normality of the ϵ_i, then Normal probability plots of the $\hat{\epsilon}_i$ might be useful.

3 Types of residual

The ideas rehearsed in the previous section essentially try to confirm or refute the properties of the ϵ_i on the basis of those of $\hat{\epsilon}_i$. However, some salient features of the ϵ_i, such as constant variance, are not reflected in the corresponding features of the $\hat{\epsilon}_i$ and, at least in principle, this can complicate the process of drawing conclusions about the residuals on the basis of their estimates.

An important quantity in this respect is the dispersion matrix of $\hat{\boldsymbol{\epsilon}}$. That of $\boldsymbol{\epsilon}$ is $\sigma^2\boldsymbol{I}$, whereas that of $\hat{\boldsymbol{\epsilon}}$ is:

$$\mathsf{var}(\hat{\boldsymbol{\epsilon}}) = \mathsf{var}((\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}) = (\boldsymbol{I}-\boldsymbol{P})\sigma^2\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{P})^T = \sigma^2(\boldsymbol{I}-\boldsymbol{P})$$

where we have used the symmetry of $\boldsymbol{P}$ and the fact that the idempotence of $\boldsymbol{P}$ implies that of $\boldsymbol{I}-\boldsymbol{P}$. Thus the variance of $\hat{\epsilon}_i$ is proportional to the ith diagonal element of $\boldsymbol{I}-\boldsymbol{P}$. As mentioned, while authors are split over whether to write $\boldsymbol{P}$ or $\boldsymbol{H}$ for the projection, or 'hat' matrix, there is unanimity on the use of h for the individual elements of this matrix. As the off-diagonal elements of $\boldsymbol{P}$ are of less interest, the diagonal elements are often shortened to h_i instead of h_{ii}. Consequently,

$$\mathsf{var}(\hat{\epsilon}_i) = \sigma^2(1-h_i). \tag{4}$$

The quantities h_i are known as the *leverages* and play an important role in the definition of more sophisticated forms of residuals. They are not, in

general, constant, so (4) indicates that the constant variance of the ϵ_i does not necessarily lead to the constant variance of the $\hat{\epsilon}_i$. In the following the h_i play an important role, so it is useful to consider their properties in more detail.

3.1 Leverages

Leverages measure the effect that a point with the given covariates can *potentially* have on the fitted line. They are functions of the covariates alone, not of the y_i. The larger the value, the greater may be the effect of an observation at that point.
In Figure 3, it is clear that at the location of the red point, i.e. with $x_i = 13.7$, then the effect of the observed y_i could have a much more marked effect on the parameter estimates, than is likely for the value of y_i observed at the blue point ($x_i = 9.79$). This is reflected in the larger leverage value for the red point, $h_i = 0.139$, compared with that for the blue point $h_i = 0.034$. In general, and certainly for models including an intercept, values further from the centroid of the data potentially have greater influence, which is reflected in larger leverages. Note that for the data in Figure 3, the mean of the x values is 9.99.

Assessing the value of a leverage is helped by knowing a few of their mathematical properties. In the following $\boldsymbol{e}_i$ denotes a vector with 1 in the ith place and 0 elsewhere.

3.2 For all i, $0 \leq h_i \leq 1$

Matrices that are idempotent and symmetric are non-negative definite. There are various way to see this - e.g. the eigenvalues of such matrices must be 0 or 1. A simple algebraic argument is to note that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^2\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}) \geq 0$. Consequently, we have that for any $\boldsymbol{x}$, both $\boldsymbol{x}^T\boldsymbol{P}\boldsymbol{x}$ and $\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}$ are non-negative. then setting $\boldsymbol{x} = \boldsymbol{e}_i$ in the above shows that $h_i \geq 0$ and $1 - h_i \geq 0$.
In general, these bounds can be attained. Consider the model without an intercept in which $\mathsf{E}(y_i) = \gamma x_i$. The leverages for this model are:

$$h_i = \frac{x_i^2}{\sum_k x_k^2}.$$

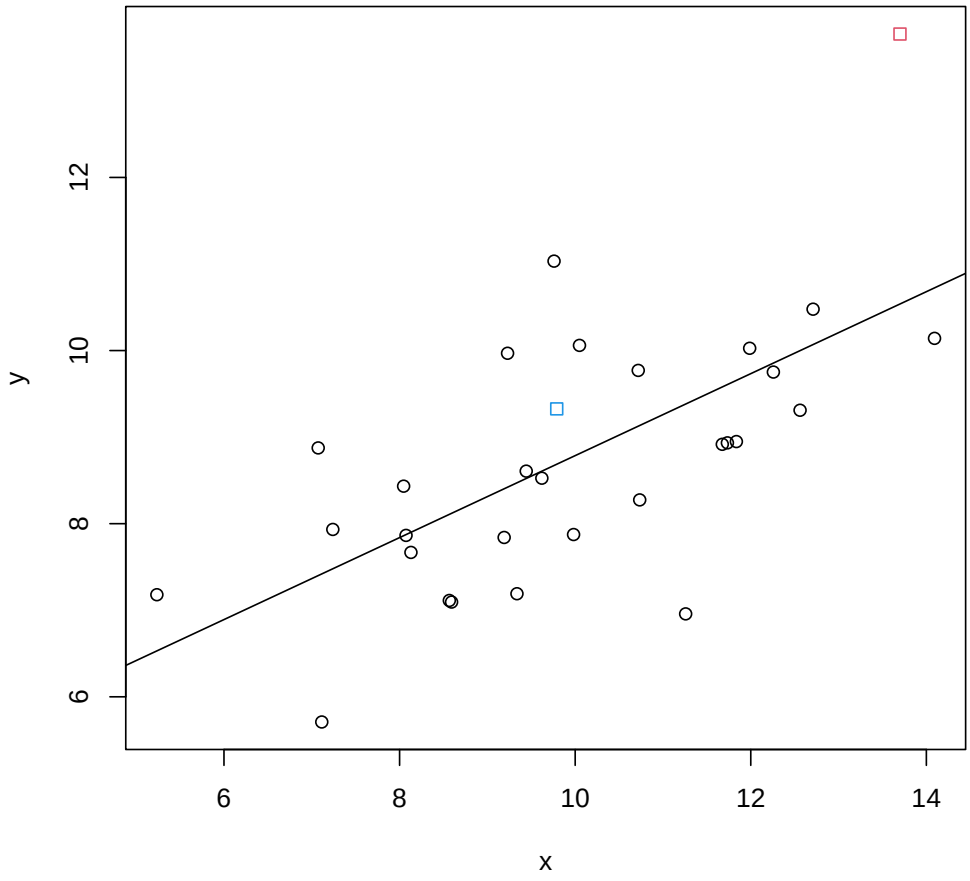


Figure 3: Leverages from a simple regression model with $n = 30$: for red square $x_i = 13.70, h_i = 0.139$; for blue square $x_i = 9.79, h_i = 0.034$.

Any point with $x_i = 0$ will have $h_i = 0$. This is no surprise - the effect on the fitted line of an observation with $x_i = 0$ is zero because the line has to pass through the origin. If all but one of the xs are small with, say, x_i being substantially larger than the other values, then this point will be very influential in determining the estimate of γ. This is reflected in the value of h_i, which will approach 1 arbitrarily closely as x_i increases in magnitude.

3.3 For models including an intercept, $\frac{1}{n} \leq h_i \leq 1$

If there is an intercept in the model, then the n-dimensional vector of ones, $\mathbf{1}$, is one of the columns of $\boldsymbol{X}$. As $\boldsymbol{P}$ is the projection onto the column space of $\boldsymbol{X}$, $\boldsymbol{P}\mathbf{1} = \mathbf{1}$. Define

$$\boldsymbol{f}_i = \boldsymbol{e}_i - \tfrac{1}{n}\mathbf{1},$$

then expanding $\boldsymbol{f}_i^T \boldsymbol{P} \boldsymbol{f}_i$ gives

$$\boldsymbol{e}_i^T \boldsymbol{P} \boldsymbol{e}_i - \frac{\boldsymbol{e}_i^T \mathbf{1}}{n} - \frac{\mathbf{1}^T \boldsymbol{e}_i}{n} + \frac{\mathbf{1}^T \mathbf{1}}{n^2} = h_i - \frac{1}{n}$$

and this must be non-negative. If there is a general mean in the model then all observations will have some effect on the estimated parameters, so zero values of h_i will not occur; this result indicates the value of the lower bound. If the model only has an intercept, then all the h_i are $1/n$.

3.4 The leverages sum to the number of parameters, $\sum h_i = p$

As $\sum h_i = \mathrm{tr}(\boldsymbol{P})$, this result can be seen algebraically by applying the identity $\mathrm{tr}(\boldsymbol{ABC}) = \mathrm{tr}(\boldsymbol{BCA}) = \mathrm{tr}(\boldsymbol{CAB})$ to the formula for $\boldsymbol{P}$ in terms of $\boldsymbol{X}$.

A more geometrical approach is to note that $\mathrm{tr}(\boldsymbol{P}) = \sum \lambda_j$, where the λs are the eigenvalues of $\boldsymbol{P}$. As $\boldsymbol{P}$ is idempotent, these are all 0 or 1. Moreover, as $\boldsymbol{P}$ is the projection onto the column space of $\boldsymbol{X}$, and as $\boldsymbol{X}$ is of full rank, then this space has dimension p, which must be the same as the number of eigenvalues equal to 1.

Either way, this result shows that the mean leverage is p/n, which can help in interpreting these quantities. For example, in Figure 3 the mean leverage is $2/30 \approx 0.0667$, so the red point has a leverage much larger than the mean, whereas that of the blue point is smaller.

Some idea of the size of the off-diagonal elements of $\boldsymbol{P}$ is available for models which include an intercept. The sum of all of the elements of $\boldsymbol{P}$ is $\mathbf{1}^T \boldsymbol{P} \mathbf{1} = \mathbf{1}^T \mathbf{1} = n$. So the mean of the off-diagonal elements of $\boldsymbol{P}$ is $(n-p)/[n(n-1)]$.

3.5 Standardized and deletion residuals

Consider the example in Figure 2(b). It appears that the spread of the residuals is not constant. This could be because the variance of the $\hat{\epsilon}_i$, namely $\sigma^2(1-h_i)$ is not constant, not because $\mathsf{var}(\epsilon_i)$ changes with age but because of the factor $1-h_i$. So it would be sensible to take steps to remove this potential source of confusion.

3.5.1 Standardized residuals

The changing variance of the simple residual, $\hat{\epsilon}_i$, can be removed by dividing by $\sqrt{1-h_i}$. However, most authors combine the step of removing variation in variance with putting the residuals on a dimensionless scale, giving the *standardized residual* as

$$r'_i = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_i}} = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_i}}, \tag{5}$$

where s^2 is the estimate of σ^2 provided by the residual mean square[3]. The advantage of including s in the standardization is that it produces values independent of the scale on which observations were made. Moreover, if the ϵ_i are at least approximately Normal, then the values of the r'_i might be expected to be similar to a standard Normal distribution, so these would be familiar to most analysts, so e.g. values outside the interval (-2,2) might attract attention, depending on sample size, of course.

It is perhaps worth making some comments about the distribution of the r'_i.

1. While it is reasonable to think that the r'_i will be approximately Normal, this cannot be exact. While the numerator is Normal, provided that the ϵ_i are, the division by the random variable s means that the ratio is not exactly Normal (although for many practical purposes the departure is not serious).

2. One might think that a quantity obtained by dividing a Normal numerator by an estimate of its variance will follow a t-distribution. However, this is not the case because the required independence between numerator and denominator does not obtain: while s and $\hat{y}_i$ are independent, s is not independent of y_i.

[3]The notation r'_i follows that in Atkinson (1985)

3. The r'_i cannot be t-distributed because, as will be seen shortly, $r'^2_i < n - p$, i.e. r'_i has bounded support.

3.5.2 Deletion residuals

In practice, model-checking is largely focused on two main areas of concern. The first concern is that the model is wrong globally: the form of the regression or the error terms might have been wrongly chosen. This sort of departure will usually be seen in patterns in various plots already discussed: for example, Figure 2(b) suggests that the residuals in the regression of OI on age do not have constant variance.

The second concern is that the model may be broadly satisfactory but for a few points that do not seem to fit the overall pattern - i.e there may be some outliers in the data. Here, model-checking techniques may be useful in identifying individual points that deserve more detailed scrutiny. What should be done with such points depends on the context and is beyond our current scope. Perhaps the main issues will turn on the effect of such points on the fitted model and whether they are especially influential on estimates of key parameters, such as treatment effects, and their associated standard errors.

A problem, at least in principle, is that including potential outliers might distort the fitted model in ways that make it more difficult to detect the unusual observations. For example, an unusually large outcome, y_i, might pull the fitted line towards itself, thereby reducing the size of the residual. A solution to one aspect of this is to assess the residual at a point using parameters estimated after omitting the point in question: such a residual is known as *deletion residual*, denoted by r^*_i. Some authors refer to this as the jack-knife residual, the cross-validatory residual or the (externally) studentized residual, but we will stick with deletion residual.

The definition of r^*_i starts with $y_i - x_i^T \hat{\boldsymbol{\beta}}_{(i)}$, where $\hat{\boldsymbol{\beta}}_{(i)}$ is the estimate of $\boldsymbol{\beta}$ based on the data with point i omitted (here and in the following, a subscript (i) refers to a statistic or other quantity obtained after omitting the ith point). This now needs to be standardised in a way analogous to (5). As y_i and $\hat{\boldsymbol{\beta}}_{(i)}$ are independent, the variance of the above difference is

$$\sigma^2 \left(1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{x}_i\right),$$

so the definition of the deletion residual is

$$r_i^* = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)} \sqrt{\left(1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{x}_i\right)}}. \tag{6}$$

Evaluation of this for all points in a dataset looks a bit awkward, as the deleted design matrix, the deleted parameter estimates and the deleted root mean square error are all needed. Computing this by looping through the data should not pose too much trouble, and these days the computational burden of this approach would be noticeable only for very large datasets.

However, a more elegant and efficient mathematical approach is available. The most awkward part of (6) is to evaluate $(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1}$ and this can be done using the so-called *Woodbury Identity* and a little determination. This approach also provides dividends in terms of gaining insight into and computing the value of other quantities, such as $\hat{\boldsymbol{\beta}}_{(i)}$.

Various forms of the of the Woodbury Identity exist and the one which suits our present purposes best is the following, where $\boldsymbol{A}$ is a $p \times p$ matrix and $\boldsymbol{U}, \boldsymbol{V}$ are $p \times m$ matrices

$$(\boldsymbol{A} - \boldsymbol{U}\boldsymbol{V}^T)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I} - \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}.$$

The key advantage of this formula is that the matrix in parentheses on the right hand side, which needs to be inverted, is an $m \times m$ matrix, and if m is much smaller than p, this inversion will be easier than the inversion on the left hand side. In our case, $m = 1$.

Three observations are useful at this point.

1. Note that $\boldsymbol{X}^T\boldsymbol{X} = \sum_j \boldsymbol{x}_j \boldsymbol{x}_j^T$,
2. and $\boldsymbol{X}^T\boldsymbol{y} = \sum_j \boldsymbol{x}_j y_j$.
3. From the definition of $\boldsymbol{P}$, $\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_i$.

Applying the above identity with $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{U} = \boldsymbol{V} = \boldsymbol{x}_i$ gives

$$\begin{aligned}(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} =& (\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1} \\ =& (\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i(1 - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i)^{-1}\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1},\end{aligned}$$

and thus

$$\boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i = h_i + \frac{h_i^2}{1-h_i} = \frac{h_i}{1-h_i}.$$

Similarly,

$$\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{(i)} = (\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}(\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_i y_i),$$

and substituting for $(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}$ gives

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\epsilon}_i}{1-h_i} \tag{7}$$

and hence the ith fitted value, based on parameters estimated from the data with the ith point omitted, is

$$\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \frac{h_i\hat{\epsilon}_i}{1-h_i},$$

so

$$y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} = \hat{\epsilon}_i + \frac{h_i\hat{\epsilon}_i}{1-h_i} = \frac{\hat{\epsilon}_i}{1-h_i}.$$

Putting all these results together shows that (6) becomes

$$r_i^* = \frac{\hat{\epsilon}_i}{s_{(i)}\sqrt{1-h_i}}.$$

Although this is showing progress, we need to roll up our sleeves one last time because the above includes the root mean square from the analysis with point i omitted. In the usual analysis, the mean square error, s^2, obeys $(n-p)s^2 = \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}$, which for present purposes is best written

$$(n-p)s^2 = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}.$$

Consequently

$$\begin{aligned}(n-p-1)s_{(i)}^2 =& \boldsymbol{y}_{(i)}^T\boldsymbol{y}_{(i)} - \hat{\boldsymbol{\beta}}_{(i)}^T\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{(i)} \\ =& \boldsymbol{y}^T\boldsymbol{y} - y_i^2 - \hat{\boldsymbol{\beta}}_{(i)}^T(\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_i y_i)\end{aligned} \tag{8}$$

Substituting for $\hat{\boldsymbol{\beta}}_{(i)}$ gives

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(i)}^T(\boldsymbol{X}^T\boldsymbol{y}-\boldsymbol{x}_i y_i) =& \left(\hat{\boldsymbol{\beta}}-\frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\epsilon}_i}{1-h_i}\right)^T(\boldsymbol{X}^T\boldsymbol{y}-\boldsymbol{x}_i y_i)\\ =&\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i y_i-\frac{\hat{\epsilon}_i\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{1-h_i}+\frac{h_i\hat{y}_i\epsilon_i}{1-h_i}\\ =&\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}-(y_i-\hat{\epsilon}_i)y_i-\frac{\hat{\epsilon}_i\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{1-h_i}+\frac{h_i\hat{y}_i\epsilon_i}{1-h_i}\\ =&\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}-y_i^2+\frac{\hat{\epsilon}_i^2}{1-h_i}\end{aligned}$$

Substituting this in (8) gives

$$\begin{aligned}(n-p-1)s_{(i)}^2 =&\boldsymbol{y}^T\boldsymbol{y}-\hat{\boldsymbol{\beta}}\boldsymbol{X}^T\boldsymbol{y}-\frac{\hat{\epsilon}_i^2}{1-h_i}\\ =&(n-p)s^2-\frac{\hat{\epsilon}_i^2}{1-h_i}\\ =&s^2(n-p-r_i'^2)\end{aligned}$$

(n.b. it is at this point we see that $r_i'^2 \leq n-p$). This allows the final step, namely

$$r_i^* = \frac{\hat{\epsilon}_i}{s_{(i)}\sqrt{1-h_i}} = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_i}}\sqrt{\frac{n-p-1}{n-p-r_i'^2}} = r_i'\sqrt{\frac{n-p-1}{n-p-r_i'^2}}. \qquad (9)$$

This a much more convenient form for the deletion residual. It also shows that for points with $\mid r_i' \mid < 1$, the deletion residual is shrunk towards zero, whereas for $\mid r_i' \mid > 1$, they are further from zero.

3.6 Some distributional properties of residuals[4]

[In the following, it is assumed that the model includes an intercept]

If the fitted model is correct, we expect estimated residuals to have mean zero and variance related to σ^2 or, if we have normalised them in some

[4]This sub-section provides some rarely presented properties of the different types of residual, but their practical moment is limited

way, variance close to 1. These properties are largely true. If the ϵ_i are independent and Normally distributed, we also expect the estimated residuals to be something close to Normal. This latter expectation is largely untrue, although deviations are not usually of practical importance.

3.6.1 Standard residuals, $\hat{\epsilon}_i$

The distributional properties of these have been adduced above, namely that they have expectation zero and variances $\sigma^2(1-h_i)$. They are not independent, as $\sum \hat{\epsilon}_i = 0$ (along with $p-1$ other constraints, one for each of the non-constant covariates). The residuals i and j have correlation $-h_{ij}/\sqrt{(1-h_i)(1-h_j)}$. These residuals are independent of the fitted values, $\hat{\boldsymbol{y}}$.

A simple normalisation, namely scaling by s, leads to residuals $\hat{\epsilon}_i/s$ that sum to zero and, as s is a function solely of the residuals, are independent of $\hat{\boldsymbol{y}}$. However, as already noted, these residuals do not have a t-distribution as numerator and denominator are not independent.

3.6.2 Deletion residuals

Consider the deletion residual

$$r_i^* = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)}\sqrt{\left(1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{x}_i\right)}} = \frac{(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)})\sqrt{1-h_i}}{s_{(i)}}.$$

As $\hat{\boldsymbol{\beta}}_{(i)}$ is Normally distributed with expectation $\boldsymbol{\beta}$, the numerator has zero mean and has a Normal distribution. As y_i and $\hat{\boldsymbol{\beta}}_{(i)}$ are independent, the numerator has variance σ^2. The square of the denominator has the same distribution as $\sigma^2 X^2/(n-p-1)$, where X^2 has a χ^2-distribution on $n-p-1$ df. As $s_{(i)}$ is a function of the residuals from the fit of the model with point i omitted, it is independent of both y_i and the fitted value $\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$, so dividing numerator and denominator by σ shows that r_i^* has a t-distribution on $n-p-1$ df. It follows that r_i^* has mean 0 and variance $(n-p-1)/(n-p-3)$.

It should be noticed that the deletion residuals will not sum to zero. Nor will they be independent of the fitted values, even if the 'deletion fitted values'

$\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}$ are used. The ordinary residuals are independent of the fitted values because $\mathsf{cov}(\hat{\epsilon}_i, \hat{y}_i) = 0$. With deletion residuals the corresponding quantity is:

$$\mathsf{cov}(r_i^*, \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}) \propto \mathsf{E}\left(\frac{1}{s_{(i)}}\right)\left(\mathsf{cov}(y_i, \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}) - \mathsf{var}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)})\right),$$

so, as the covariance on the right hand side vanishes, the covariance on the left hand side is inevitably negative.

3.6.3 Standardized residuals

Discussion of these residuals has been deferred to last as their properties are slightly more delicate than the other residuals, largely because s in the denominator of r_i' is dependent on the numerator, $\hat{\epsilon}_i$.
The presence of the factor $\sqrt{1-h_i}$ means that, in general, the r_i' will not sum to zero. However, r_i' is a function only of the residuals being, apart from factors depending on n and p,

$$r_i' \propto \frac{\hat{\epsilon}_i}{\sqrt{\sum_k \hat{\epsilon}_k^2}}, \tag{10}$$

so is independent of the fitted value $\hat{y}_i$.

The dependence of numerator and denominator in r_i' means that evaluation of $\mathsf{E}(r_i')$ requires some care. If the model is true, then $\hat{\boldsymbol{\epsilon}}$ has a multivariate Normal distribution with mean 0 and dispersion $\sigma^2(\boldsymbol{I}-\boldsymbol{P})$. This is the same distribution as that of $-\hat{\boldsymbol{\epsilon}}$. Consequently, from (10)

$$\mathsf{E}\left(\frac{\hat{\epsilon}_i}{\sqrt{\sum_j \hat{\epsilon}_j^2}}\right) = \mathsf{E}\left(\frac{-\hat{\epsilon}_i}{\sqrt{\sum_j (-\hat{\epsilon}_j)^2}}\right) = -\mathsf{E}\left(\frac{\hat{\epsilon}_i}{\sqrt{\sum_j \hat{\epsilon}_j^2}}\right),$$

and so $\mathsf{E}(r_i') = 0$.

The variance of r_i' is found by evaluating its distribution. Starting from (9) we have

$$\frac{r_i'^2}{n-p} = \frac{r_i^{*2}}{n-p-1+r_i^{*2}}$$

and r_i^* has a t-distribution with $n-p-1$df. Such a distribution is that of a random variable $Z/\sqrt{X^2/(n-p-1)}$, where Z and X^2 are independent

random variables with, respectively, a standard Normal distribution and a χ^2-distribution on $n-p-1$ df. Substituting in the above shows that $r_i'^2/(n-p)$ has the same distribution as $Z^2/(X^2+Z^2)$. Now recall the following:

1. Z^2 has a χ^2-distribution on 1 df;
2. a χ^2-distribution on ν df is a Gamma distribution with shape parameter $\frac{1}{2}\nu$ and scale parameter 2;
3. If U and V are independent Gamma variables with common scale parameter and shapes parameters k_1 and k_2, then $U/(U+V)$ has a Beta distribution with parameters k_1 and k_2.

Consequently $r_i'^2$ is distributed as $n-p$ times a Beta variable with parameters $\frac{1}{2}$ and $\frac{1}{2}(n-p-1)$. Using the formula for the expectation of a Beta variable, namely $k_1/(k_1+k_2)$ shows that

$$\mathsf{E}(r_i'^2) = (n-p) \times \frac{1}{1+n-p-1} = 1,$$

showing that the standardized residuals do indeed have variance one.

As with the $\hat{\epsilon}_i$, the standardized residuals are not independent and the correlation of r_i' and r_j' is $-h_{ij}/\sqrt{(1-h_i)(1-h)j)}$, the same as for $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$: see Cook and Weisberg (1982, p.19), quoting Ellenberg (1973).

Although a Beta distribution may sound rather different from anything to do with a Normal distribution, things are not that different. If r_i' is normal with mean 0 and variance 1, then $r_i'^2$ would be expected to follow a χ^2-distribution, which has mean 1 and variance 2. Using the scaled Beta distribution, the mean and variance would be 1 and $2(n-p-1)/(n-p+2)$, respectively, and in most applications $n-p$ will be large compared with 2. A plot of the scaled Beta density and a χ^2_1 density is shown in Figure 4, for $n=11, p=2$, showing that the two densities are very similar. Only in practically unimportant cases in which the Beta parameters are similar, i.e. $n-p$ is small, will the two densities diverge noticeably.

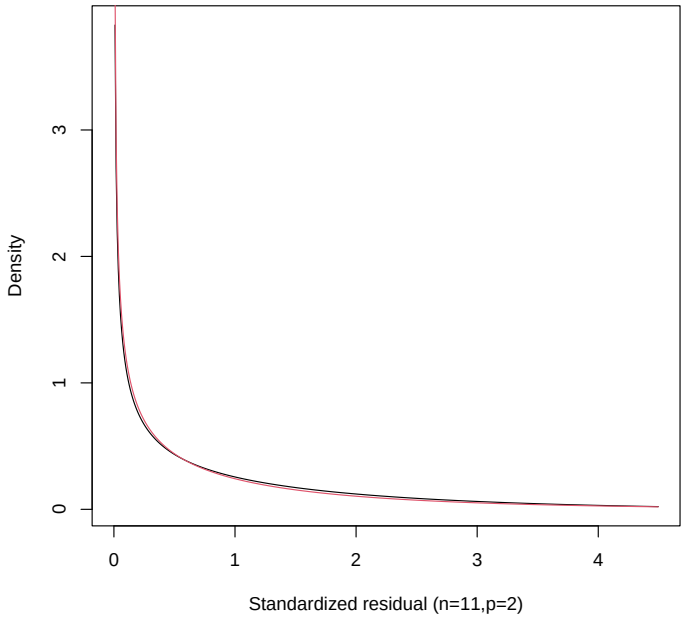


Figure 4: Approximate and exact densities for $r_i'^2$: red is the χ^2_1 density and black is the $n-p$ scaled Beta density with parameters $\frac{1}{2}$ and $\frac{1}{2}(n-p-1)$, with $n=11$ and $p=2$ [shown over range 0 to 5].

4 Does the type of residual matter and what to do anyway

4.1 Does it matter?

In comparing the different types of residuals, the first thing to deal with is that the 'usual' residual, $\hat{\epsilon}_i$ is on the scale of the observations and cannot be compared directly with the standardized and deletion residuals, which are dimensionless. As such, when comparing the types of residuals, it is necessary to replace the usual residuals by what will be referred to as *scaled residuals*, namely $\hat{\epsilon}_i/s$.

Once this has been done, it might be argued that the relationship between the residuals is already apparent from the mathematics. For example, the

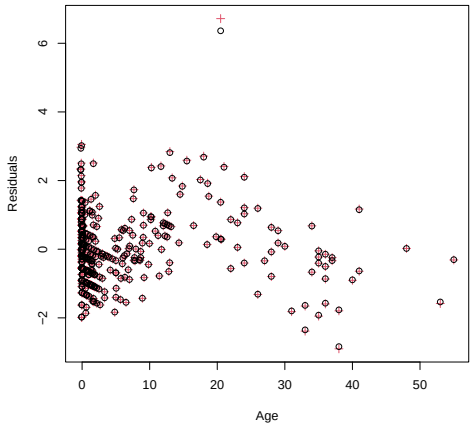
presence of $\sqrt{1-h_i}$ in the definition of standardized residuals mean they are always further from 0 than the scaled residuals, while (9) also provides pertinent information.

However, the formulae alone do not give the sense of the practical differences between the types that can be found from illustrations based on real data. The following plots compare type of residuals using the data displayed in Figures 1 and 2. While this exercise only uses two datasets, there does not seem any reason to think that there is anything untypical about these data.

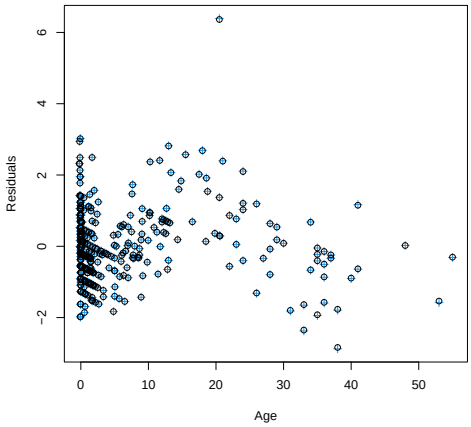
Perhaps the most striking feature of Figure 5 is that the differences between the different types of residual are very minor. The extreme points are slightly more extreme for deletion residuals, as shown in the largest positive residual in Figure 5(a) and the larger positive residuals at the top right of Figure 5(c). The larger deletion residuals probably reflect the influence these points had on the fitted lines. Figure 5(b) shows that the standardized and scaled residuals are very similar indeed.

For the Graves' data, Figure 5(d) compares the deletion and standardized residuals. For the largest residuals, the deletion residuals are more extreme than the standardized residuals, closely mirroring the comparison in Figure 5(c). Careful inspection shows that deletion residuals are always more extreme outside the interval [-1,1] on the vertical axis. Within this interval (9) indicates that the standardized residuals are more extreme. Very careful inspection shows that this is the case, although, in the parts of the interval close to 0, the effect is not really visible because, also following (9), a standardized residual vanishes if, and only if the same is true of the corresponding deletion residual.

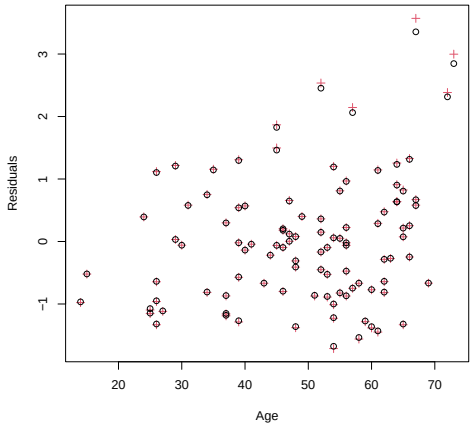
In these examples, and probably more widely, the perceived and theoretical shortcomings of the $\hat{\epsilon}_i$, or scaled residuals, do not seem to be important. Nevertheless, as the different forms of residual are now widely available, any analyst will have to choose one form or other when assessing model fit. In R , if the regression is stored in `mod`, then scaled residuals can be found as `resid(mod)/sigma(mod)`, standardized residuals as `rstandard(mod)` and deletion residuals as `rstudent(mod)`, so all types of residual are equally readily available. My own practice tends to prefer deletion residuals if there is concern about outliers, whereas the distributional properties of other forms,



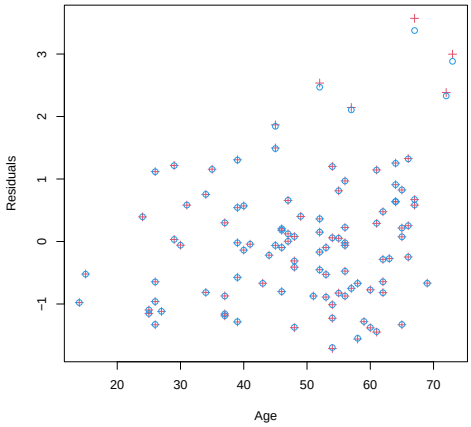
(a) Scaled and deletion for onsets



(b) Scaled and standardized for onsets



(c) Scaled and deletion for OI



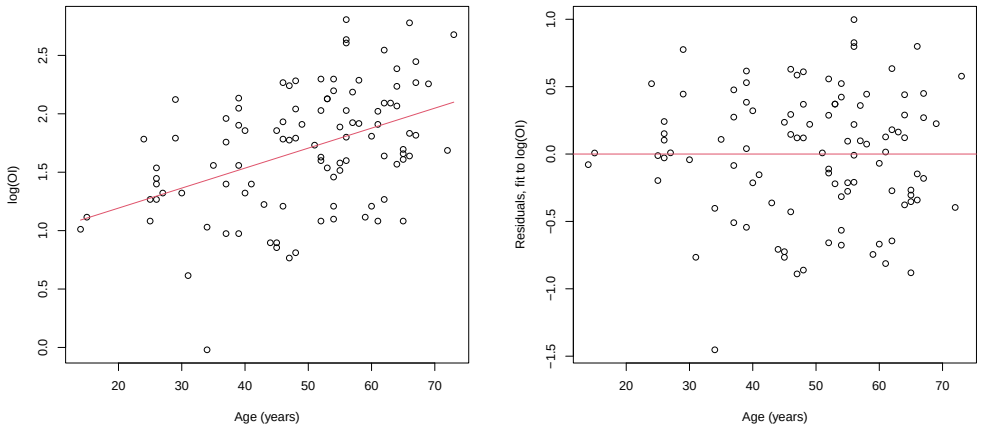
(d) Standardized and deletion for OI

Figure 5: Residuals for onset times (O'Sullivan et al., 1998) and ophthalmic index (OI)(Perros et al., 1993). Black circles are scaled residuals, blue crosses are standardized residuals and red crosses denote deletion residuals.

such as independence of residuals and fitted values, make them more suitable for assessment of global departures in model fit.

4.2 What to do next

What has been conspicuous by its absence from this article is any discussion about how model inadequacy revealed by the techniques considered here can be remedied. This will largely continue to be the case. There are techniques, such as added variable plots (Atkinson, 1985, p.67), which try to guide the process of model amendment, but most changes will come from a more fundamental rethinking of the form of the model, based on the context of the application and the nature of the problem revealed.
A good example of this comes from the problem with the OI data seen in Figure 2(b). Here the variance of the residuals seems to increase with the age of the patient, as does the size the OI itself (Figure 2(a)). On the basis of this it seems plausible that the variance of the OI acts not additively, but proportionally - the variance may be a constant percentage of the OI. If this is the case, then modelling not OI but log(OI) may be preferable.



(a) Linear fit of age to log(OI)

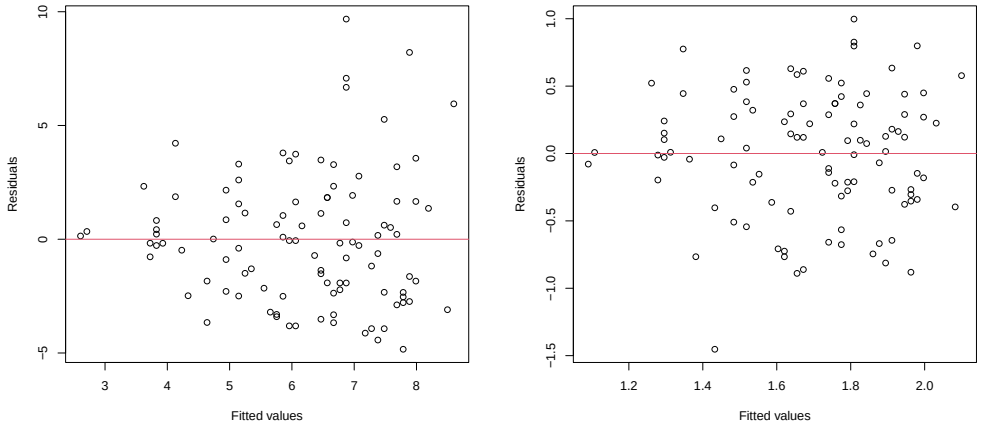
(b) Residuals vs age

Figure 6: Data on log(OI) versus age in 101 patients with Graves' ophthalmopathy (Perros et al., 1993)

Figure 6 presents the same plots as in Figure 2 but now applied to the log of the OI. The increasing spread of the residuals with increasing age is no longer present. This seems to offer an improved fit. Plots of the usual residuals, $\hat{\epsilon}_i$ against the fitted values for both OI and log(OI) are shown in Figure 7.

Again, there is no suggestion of a change in the variance of the $\hat{\epsilon}_i$ with fitted values in Figure 7(b), whereas there is a clear pattern in Figure 7(a), with a clear indication that in this model a constant σ^2 is not appropriate.

In general, using routine model-checking to identify shortcomings in a fitted model is sensible. However, if a questionable fit is revealed then context-dependent methods for model amendment, rather than an algorithmic approach, is probably to be preferred. It ought to lead to a final model that is better understood and meaningful.



(a) Residuals vs fitted, OI

(b) Residuals vs fitted, log(OI)

Figure 7: Plots of the usual residuals versus fitted values for the Graves' ophthalmopathy data, untransformed OI and log-transformed OI

5 Things not in this document

Books have been written on this material and even just touching on the things not covered above would greatly lengthen what is already a long document. However, two topics should be mentioned.

5.1 Cook's statistic

The purpose of Cook's statistic is to gauge the effect of each observation on the estimated parameters. It is defined as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2} = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})^T (\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{ps^2} = \frac{r_i'^2 h_i}{p(1-h_i)},$$

where (7) has been used to obtain the final form. The presence of $\boldsymbol{X}$ in the definition is to put all the changes on a common scale, namely that of the outcome, and division by s^2 leads to a dimensionless quantity, which might be thought to make interpretation easier.
Cook's statistic can readily be found in most statistical packages, and is doubtless of use in some applications, but it is probably less often useful than might be thought.

1. Assessing the effect of deletion of cases on all parameters at the same time is, in many applications, probably less pertinent than assessing the effect on certain subsets of the parameters. While an appropriate amendment of the definition would clearly be possible, some of the convenience would be lost because it would need special calculation in most packages.

2. Indeed, the number of parameters of interest may be quite small. In many medical applications, attention would often be focused on one or two treatment or effect parameters. In these cases, they would be assessed best on the scale of the outcome, not in the dimensionless form in D_i.

3. While investigating the effect of a few questionable on key parameters is often a very important exercise, making the assessment by deleting all observations in turn is not usually appropriate and, indeed, could be rather unhelpful.

5.2 Assessing Normality of the residuals

There has been no mention of the assessment of the Normality of the residuals in this article. It should be remembered that from the Gauss-Markov theorem, homoscedasticity is all that is needed for optimal estimation of the $\boldsymbol{\beta}$ in the sense of minimising the error of estimation. Nevertheless, in a

model-based approach, validity of hypothesis tests and confidence intervals does rest on the Normality assumption, so some attention should be paid to this aspect. Moreover, some methods of checking Normality can be very good at detecting outliers.

While simple methods, such as plotting histograms of residuals, are useful, the Normal probability plot is probably the tool of choice. A thorough discussion of these plots raises quite a few issues and will be the subject of a separate article.

References

A C Atkinson. *Plots, Transformations, and Regression*. Oxford University Press, Oxford, 1985.

R Dennis Cook and Sandford Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, London, 1982.

J H Ellenberg. Joint distribution of standardized least-squares residuals from a general linear-regression. *Journal of the American Statistical Association*, 68:941–943, 1973. doi: 10.2307/2284526.

M. C. O'Sullivan, S. Miller, V. Ramesh, E. Conway, K. Gilfillan, S. McDonough, and J. A. Eyre. Abnormal development of biceps brachii phasic stretch reflex and persistence of short latency heteronymous reflexes from biceps to triceps brachii in spastic cerebral palsy. *Brain*, 121:2381–2395, 1998. doi: 10.1093/brain/121.12.2381.

P. Perros, A. L. Crombie, J. N. S. Matthews, and P. Kendall-Taylor. Age and gender influence the severity of thyroid-associated ophthalmopathy: a study of 101 patients attending a combined thyroid-eye clinic. *Clinical Endocrinology*, 38:367–372, 1993. doi: 10.1111/j.1365-2265.1993.tb00516.x.