# Chapter 5
# Multiple Linear Regression

In the multiple linear regression (MLR) model,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{5.1}$$

for $i = 1, \ldots, n$. In matrix notation, these $n$ equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{5.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{5.3}$$

Often the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, the $n \times 1$ vector of ones. The $i$th *case* $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the $i$th row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the $i$th element of $\boldsymbol{Y}$. If the $e_i$ are iid with zero mean and variance $\sigma^2$, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and $\sigma^2$.

**Definition 5.1.** Given an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted* or *fitted values* is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. The residual vector is $\boldsymbol{r} = \boldsymbol{r}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.

Most regression methods attempt to find an estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals where the $i$th residual $r_i(\boldsymbol{b}) = r_i = Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = Y_i - \hat{Y}_i$. The order statistics for the absolute residuals are denoted by

$$|r|_{(1)} \leq |r|_{(2)} \leq \cdots \leq |r|_{(n)}.$$

Two of the most used classical regression methods are ordinary least squares (OLS) and least absolute deviations ($L_1$).

**Definition 5.2.** The *ordinary least squares estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{5.4}$$

$$\text{and} \quad \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists.

**Definition 5.3.** The *least absolute deviations estimator* $\hat{\boldsymbol{\beta}}_{L_1}$ minimizes

$$Q_{L_1}(\boldsymbol{b}) = \sum_{i=1}^{n} |r_i(\boldsymbol{b})|. \tag{5.5}$$

**Definition 5.4.** The *Chebyshev* ($L_\infty$) *estimator* $\hat{\boldsymbol{\beta}}_{L_\infty}$ minimizes the maximum absolute residual $Q_{L_\infty}(\boldsymbol{b}) = |r(\boldsymbol{b})|_{(n)}$.

The location model is a special case of the multiple linear regression (MLR) model where $p = 1$, $\boldsymbol{X} = \boldsymbol{1}$ and $\boldsymbol{\beta} = \mu$. One very important change in the notation will be used. In the location model, $Y_1, ..., Y_n$ were assumed to be iid with cdf $F$. For regression, the *errors* $e_1, ..., e_n$ will be assumed to be iid with cdf $F$. For now, assume that the $\boldsymbol{x}_i^T\boldsymbol{\beta}$ are constants. Note that $Y_1, ..., Y_n$ are independent if the $e_i$ are independent, but they are not identically distributed since if $E(e_i) = 0$, then $E(Y_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ depends on $i$.

In the location model, $\hat{\boldsymbol{\beta}}_{OLS} = \overline{Y}$, $\hat{\boldsymbol{\beta}}_{L_1} = \text{MED}(n)$ and the Chebyshev estimator is the *midrange* $\hat{\boldsymbol{\beta}}_{L_\infty} = (Y_{(1)} + Y_{(n)})/2$. These estimators are simple to compute, but computation in the multiple linear regression case requires a computer. Most statistical software packages have OLS routines, and the $L_1$ and Chebyshev fits can be efficiently computed using linear programming. The $L_1$ fit can also be found by examining all

$$C(n, p) = \binom{n}{p} = \frac{n!}{p!(n-p)!}$$

subsets of size $p$ where $n! = n(n-1)(n-2)\cdots 1$ and $0! = 1$. The Chebyshev fit to a sample of size $n > p$ is also a Chebyshev fit to some subsample of size $h = p + 1$. Thus the Chebyshev fit can be found by examining all $C(n, p+1)$ subsets of size $p+1$. These two combinatorial facts will be useful for the high breakdown regression estimators LMS and LTA described in Sections 5.9 and 6.3.

## 5.1 Predictor Transformations

*As a general rule, inferring about the distribution of $Y|\boldsymbol{X}$ from a lower dimensional plot should be avoided when there are strong nonlinearities among the predictors.*
Cook and Weisberg (1999b, p. 34)

Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful for regression methods such as multiple linear regression, generalized linear models, generalized additive models, 1D regression, nonlinear regression, and nonparametric regression. Power transformations are particularly effective, and a power transformation has the form $x = t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ and $x = t_0(w) = \log(w)$ for $\lambda = 0$. Often $\lambda \in \Lambda_L$ where

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{5.6}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder", e.g. from $\lambda = 1$ to $\lambda = 0$, will be useful. If the transformation goes too far down the ladder, e.g. if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then it will be necessary to go back "up the ladder." Additional powers such as $\pm 2$ and $\pm 3$ can always be added.

**Definition 5.5.** A **scatterplot** of $x$ versus $Y$ is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors and the response variable $Y$.

In this section we will only make a scatterplot matrix of the predictors. Often nine or ten variables can be placed in a scatterplot matrix. The names of the variables appear on the diagonal of the scatterplot matrix. The $R$ software labels the values of each variable in two places, see Example 5.2 below. Let one of the variables be $W$. All of the marginal plots above and below $W$ have $W$ on the horizontal axis. All of the marginal plots to the left and the right of $W$ have $W$ on the vertical axis.

There are several rules of thumb that are useful for visually selecting a power transformation to remove nonlinearities from the predictors. Several of these rules need $p$ small, but the log rule can be used when $p$ is large. The rules are also useful for response transformations covered in Section 5.2. In this text, $\log(x) = \ln(x) = \log_e(x)$.

**Rule of thumb 5.1.** a) If strong nonlinearities are apparent in the scatterplot matrix of the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations.

b) Use theory if available.

c) Suppose that variable $X_2$ is on the vertical axis and $X_1$ is on the horizontal axis and that the plot of $X_1$ versus $X_2$ is nonlinear. The *unit rule* says that if $X_1$ and $X_2$ have the same units, then try the same transformation for both $X_1$ and $X_2$.

Assume that all values of $X_1$ and $X_2$ are positive. Then the following six rules are often used.

d) The **log rule** states that a positive predictor that has the ratio between the largest and smallest values greater than ten should be transformed to logs. So $X > 0$ and $\max(X)/\min(X) > 10$ suggests using $\log(X)$.

e) The **range rule** states that a positive predictor that has the ratio between the largest and smallest values less than two should not be transformed. So $X > 0$ and $\max(X)/\min(X) < 2$ suggests keeping $X$.

f) The *bulging rule* states that changes to the power of $X_2$ and the power of $X_1$ can be determined by the direction that the bulging side of the curve points. If the curve is hollow up (the bulge points down), decrease the power of $X_2$. If the curve is hollow down (the bulge points up), increase the power of $X_2$. If the curve bulges towards large values of $X_1$ increase the power of $X_1$. If the curve bulges towards small values of $X_1$ decrease the power of $X_1$. See Tukey (1977, p. 173–176).

g) The **ladder rule** appears in Cook and Weisberg (1999a, p. 86).
To spread *small* values of a variable, make $\lambda$ *smaller*.
To spread *large* values of a variable, make $\lambda$ *larger*.

h) If it is known that $X_2 \approx X_1^\lambda$ and the ranges of $X_1$ and $X_2$ are such that this relationship is one to one, then

$$X_1^\lambda \approx X_2 \text{ and } X_2^{1/\lambda} \approx X_1.$$

Hence either the transformation $X_1^\lambda$ or $X_2^{1/\lambda}$ will linearize the plot. Note that $\log(X_2) \approx \lambda \log(X_1)$, so taking logs of both variables will also linearize the plot. This relationship frequently occurs if there is a volume present. For example let $X_2$ be the volume of a sphere and let $X_1$ be the circumference of a sphere.

i) The *cube root rule* says that if $X$ is a volume measurement, then cube root transformation $X^{1/3}$ may be useful.

In the literature, it is sometimes stated that predictor transformations that are made without looking at the response are "free." The reasoning is that the conditional distribution of $Y|(x_2 = a_2, ..., x_p = a_p)$ is the same as the conditional distribution of $Y|[t_2(x_2) = t_2(a_2), ..., t_p(x_p) = t_p(a_p)]$: there is simply a change of labeling. Certainly if $Y|x = 9 \sim N(0, 1)$, then

$Y|\sqrt{x} = 3 \sim N(0,1)$. To see that Rule of thumb 5.1a does not always work, suppose that $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ where the $x_i$ are iid lognormal(0,1) random variables. Then $w_i = \log(x_i) \sim N(0,1)$ for $i = 2, ..., p$ and the scatterplot matrix of the $w_i$ will be linear while the scatterplot matrix of the $x_i$ will show strong nonlinearities if the sample size is large. However, there is an MLR relationship between $Y$ and the $x_i$ while the relationship between $Y$ and the $w_i$ is nonlinear: $Y = \beta_1 + \beta_2 e^{w_2} + \cdots + \beta_p e^{w_p} + e \neq \boldsymbol{\beta}^T \boldsymbol{w} + e$. Given $Y$ and the $w_i$ with no information of the relationship, it would be difficult to find the exponential transformation and to estimate the $\beta_i$. The moral is that predictor transformations, especially the log transformation, can and often do greatly simplify the MLR analysis, but predictor transformations can turn a simple MLR analysis into a very complex nonlinear analysis.

Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if $W$ = weight and $X_1$ = volume = $(X_2)(X_3)(X_4)$, then $W$ versus $X_1^{1/3}$ and $\log(W)$ versus $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ may both work. Also if $W$ is linearly related with $X_2, X_3, X_4$ and these three variables all have length units mm, say, then the units of $X_1$ are $(mm)^3$. Hence the units of $X_1^{1/3}$ are mm.

Suppose that all values of the variable $w$ to be transformed are positive. The log rule says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$. This rule often works wonders on the data and the log transformation is the most used (modified) power transformation. If the variable $w$ can take on the value of 0, use $\log(w + c)$ where $c$ is a small constant like 1, 1/2, or 3/8.

To use the ladder rule, suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$ where both $x_1 > 0$ and $x_2 > 0$. Also assume that the plotted points follow a nonlinear one to one function. Consider the ladder of powers

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1, \}.$$

To spread small values of the variable, make $\lambda_i$ smaller. To spread large values of the variable, make $\lambda_i$ larger.

For example, if both variables are **right skewed**, then there will be many more cases in the lower left of the plot than in the upper right. Hence small values of both variables need spreading.

Consider the ladder of powers. Often no transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal transformation.

**Example 5.1.** Examine Figure 5.1. Let $X_1 = w$ and $X_2 = x$. Since $w$ is on the horizontal axis, mentally add a narrow vertical slice to the plot. If a large amount of data falls in the slice at the left of the plot, then small values need spreading. Similarly, if a large amount of data falls in the slice at the right of the plot (compared to the middle and left of the plot), then large values need spreading. For the variable on the vertical axis, make a narrow horizontal slice. If the plot looks roughly like the northwest corner of a square then

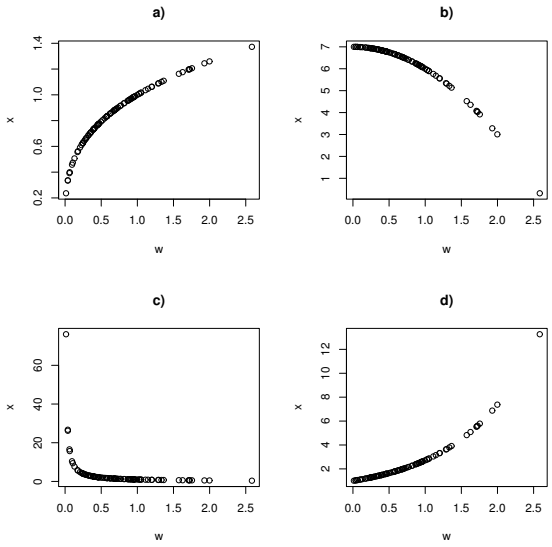


**Fig. 5.1** Plots to Illustrate the Bulging and Ladder Rules

small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 5.1a, small values of $w$ need spreading. Notice that the plotted points bulge up towards small values of the horizontal variable. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 5.1b, large values of $x$ need spreading. Notice that the plotted points bulge up towards large values of the horizontal variable. If the plot looks roughly like the southwest corner of a square, as in Figure 5.1c, then small values of both variables need spreading. Notice that the plotted points bulge down towards small values of the horizontal variable. If the plot looks roughly like the southeast corner of a square, then large values of the horizontal and small values of the vertical variable need spreading. Hence in Figure 5.1d, small values of $x$ need spreading. Notice that the plotted points bulge down towards large values of the horizontal variable.

**Example 5.2: Mussel Data.** Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The response is *muscle mass* $M$ in grams, and the predictors are a constant, the *length* $L$ and *height* $H$ of the shell in mm, the *shell width* $W$ and the *shell mass* $S$. Figure 5.2 shows the scatterplot matrix of the predictors $L$, $H$, $W$ and $S$. Examine the variable *length*. Length is on the vertical axis on the three top plots and the right of the scatterplot matrix labels this axis from 150 to 300. Length is on the horizontal axis on the three leftmost marginal plots, and this axis is labelled from 150 to 300 on the bottom of the scatterplot matrix. The

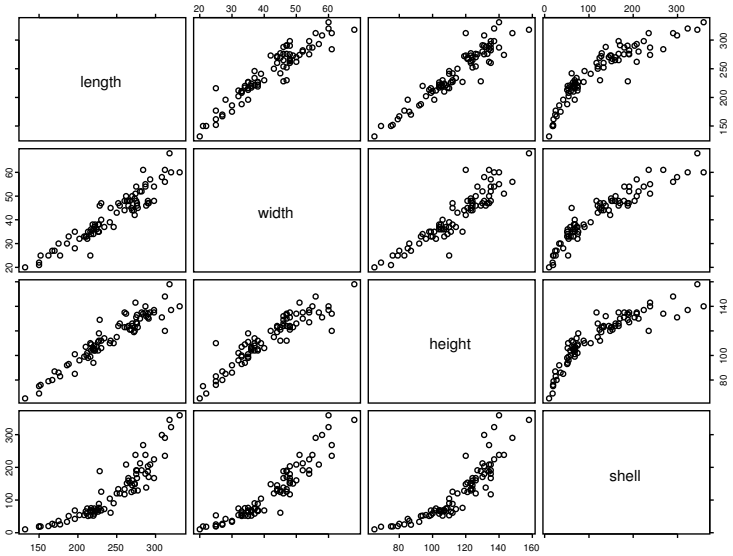


**Fig. 5.2** Scatterplot Matrix for Original Mussel Data Predictors

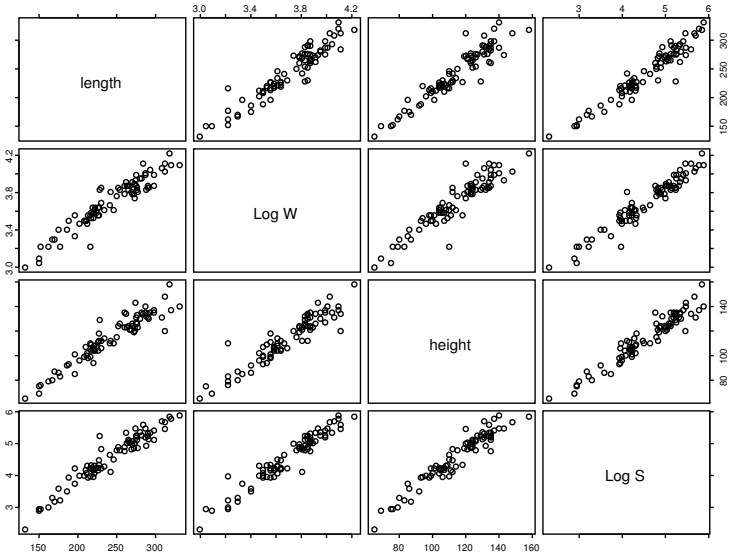


**Fig. 5.3** Scatterplot Matrix for Transformed Mussel Data Predictors

marginal plot in the bottom left corner has length on the horizontal and shell on the vertical axis. The marginal plot that is second from the top and second from the right has height on the horizontal and width on the vertical axis. If the data is stored in $x$, the plot can be made with the following command in *R*.

```
pairs(x,labels=c("length",'"width","height","shell"))
```

Nonlinearity is present in several of the plots. For example, width and length seem to be linearly related while length and shell have a nonlinear relationship. The minimum value of shell is 10 while the max is 350. Since $350/10 = 35 > 10$, the log rule suggests that $\log S$ may be useful. If $\log S$ replaces $S$ in the scatterplot matrix, then there may be some nonlinearity present in the plot of $\log S$ versus $W$ with small values of $W$ needing spreading. Hence the ladder rule suggests reducing $\lambda$ from 1 and we tried $\log(W)$. Figure 5.3 shows that taking the log transformations of $W$ and $S$ results in a scatterplot matrix that is much more linear than the scatterplot matrix of Figure 5.2. Notice that the plot of $W$ versus $L$ and the plot of $\log(W)$ versus $L$ both appear linear. This plot can be made with the following commands.

```
z <- x; z[,2] <- log(z[,2]); z[,4] <- log(z[,4])
pairs(z,labels=c("length","Log W","height","Log S"))
```

The plot of *shell* versus *height* in Figure 5.2 is nonlinear, and small values of *shell* need spreading since if the plotted points were projected on the horizontal axis, there would be too many points at values of *shell* near 0. Similarly, large values of *height* need spreading.

## 5.2 A Graphical Method for Response Transformations

*If the ratio of largest to smallest value of y is substantial, we usually begin by looking at log y.*
Mosteller and Tukey (1977, p. 91)

The applicability of the multiple linear regression model can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter $\lambda_o$, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i. \tag{5.7}$$

If $\lambda_o$ was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow a multiple linear regression model with $p$ predictors including the constant. Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients depending on $\lambda_o$, $\boldsymbol{x}$ is a $p \times 1$ vector of predictors that are assumed to be measured with negligible error, and the errors $e_i$ are assumed to be iid with zero mean.

**Definition 5.6.** Assume that **all** of the values of the "response" $Z_i$ are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

**Definition 5.7.** Assume that **all** of the values of the "response" $Z_i$ are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{5.8}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Often $Z_i^{(1)}$ is replaced by $Z_i$ for $\lambda = 1$. Generally $\lambda \in \Lambda$ where $\Lambda$ is some interval such as $[-1, 1]$ or a coarse subset such as $\Lambda_L$.

A graphical method for response transformations refits the model using the same fitting method: changing only the "response" from $Z$ to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus $W_i$ is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band if the unimodal MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$. See Definition 5.13. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

By adding the "response" $Z$ to the scatterplot matrix, the methods of the previous section can also be used to suggest good values of $\lambda$, and it is usually a good idea to use predictor transformations to remove nonlinearities from the predictors before selecting a response transformation. Check that the scatterplot matrix with the transformed variables is better than the scatterplot matrix of the original variables. Notice that the graphical method is equivalent to making "response plots" for the seven values of $W = t_\lambda(Z)$, and choosing the "best response plot" where the MLR model seems "most reasonable." The seven "response plots" are called transformation plots below. Our convention is that a plot of $X$ versus $Y$ means that $X$ is on the horizontal axis and $Y$ is on the vertical axis.

**Warning:** The Rule of thumb 5.1 does not always work. For example, the log rule may fail. If the relationships in the scatterplot matrix are already linear or if taking the transformation does not increase the linearity (especially in the row containing the response), then no transformation may be better than taking a transformation. For the Cook and Weisberg (1999a) *Arc* data

set `evaporat.lsp`, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

**Definition 5.8.** A *transformation plot* is a plot of $\hat{W}$ versus $W$ with the identity line added as a visual aid.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. According to Mosteller and Tukey (1977, p. 91), the **most commonly used power transformations** are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$ and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in $\Lambda_L$, then sometimes $\hat{\lambda}_n$ will converge (e.g. in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid $\Lambda_L$. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and $\pm 3$. Powers from numerical methods can also be added.

**Application 5.1.** This graphical method for selecting a response transformation is very simple. Let $W_i = t_\lambda(Z_i)$. Then for each of the seven values of $\lambda \in \Lambda_L$, perform OLS on $(W_i, \boldsymbol{x}_i)$ and make the transformation plot of $\hat{W}_i$ versus $W_i$. If the plotted points follow the identity line for $\lambda^*$, then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of $\lambda_o$ by adding $\hat{\lambda}$ to $\Lambda_L$. OLS can be replaced by other methods such as lasso.)

If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding "residual plots" of $\hat{W}$ versus $W - \hat{W}$ look reasonable. The values of $\lambda$ in decreasing order of importance are $1, 0, 1/2, -1$ and $1/3$. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good.

After selecting the transformation, the usual checks should be made. In particular, the transformation plot for the selected transformation is the response plot, and a residual plot should also be made. The following example illustrates the procedure, and the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the "fitted values" that result from using $t_\lambda(Z)$ as the "response" in the OLS software.

**Example 5.3: Textile Data.** In their pioneering paper on response transformations, Box and Cox (1964) analyze data from a $3^3$ experiment on the behavior of worsted yarn under cycles of repeated loadings. The "response" $Z$ is the *number of cycles to failure* and a constant is used along with the three

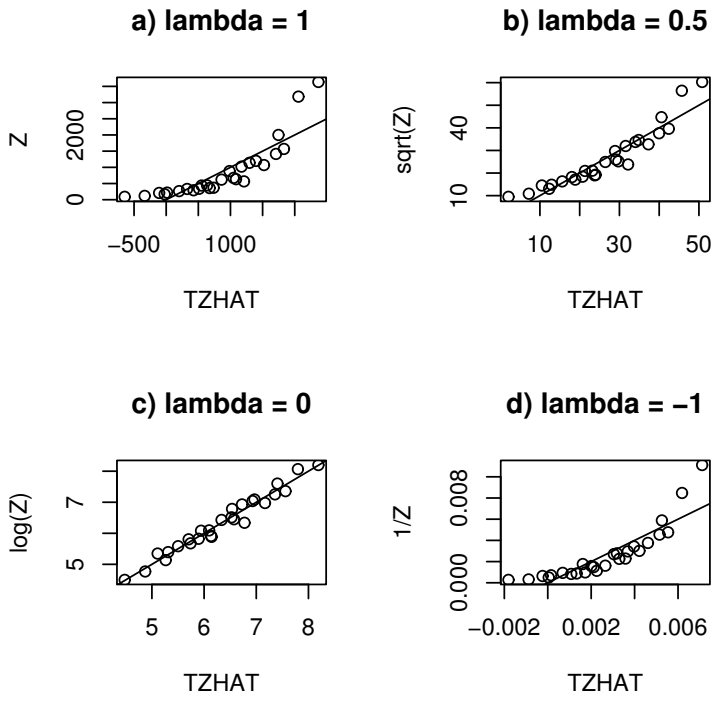


**Fig. 5.4** Four Transformation Plots for the Textile Data

predictors *length, amplitude* and *load.* Using the normal profile log likelihood for $\lambda_o$, Box and Cox determine $\hat{\lambda}_o = -0.06$ with approximate 95 percent confidence interval $-0.18$ to 0.06. These results give a strong indication that the log transformation may result in a relatively simple model, as argued by Box and Cox. Nevertheless, the numerical Box–Cox transformation method provides no direct way of judging the transformation against the data.

Shown in Figure 5.4 are transformation plots of $\hat{Z}$ versus $Z^{\lambda}$ for four values of $\lambda$ except $\log(Z)$ is used if $\lambda = 0$. The plots show how the transformations bend the data to achieve a homoscedastic linear trend. Perhaps more importantly, they indicate that the information on the transformation is spread throughout the data in the plot since changing $\lambda$ causes all points along the curvilinear scatter in Figure 5.4a to form along a linear scatter in Figure 5.4c. Dynamic plotting using $\lambda$ as a control seems quite effective for judging transformations against the data and the log response transformation does indeed seem reasonable.

Note the simplicity of the method: Figure 5.4a shows that a response transformation is needed since the plotted points follow a nonlinear curve while

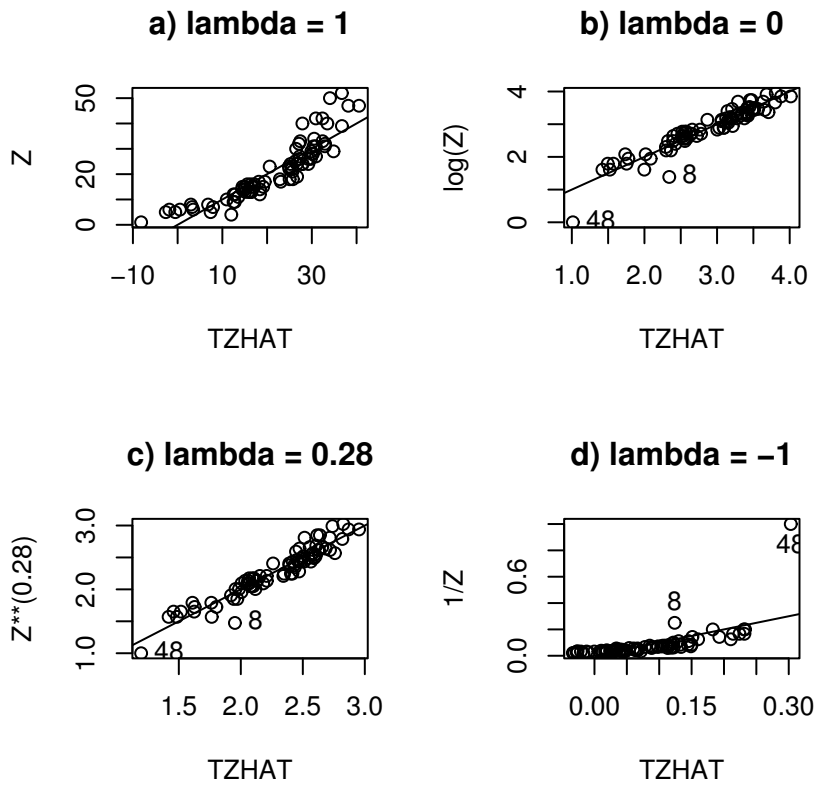


**Fig. 5.5** Transformation Plots for the Mussel Data

Figure 5.4c suggests that $Y = \log(Z)$ is the appropriate response transformation since the plotted points follow the identity line. If all 7 plots were made for $\lambda \in \Lambda_L$, then $\lambda = 0$ would be selected since this plot is linear. Also, Figure 5.4a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

The essential point of the next example is that observations that influence the choice of the usual Box–Cox numerical power transformation are often easily identified in the transformation plots. The transformation plots are especially useful if the bivariate relationships of the predictors, as seen in the scatterplot matrix of the predictors, are linear.

**Example 5.4: Mussel Data.** Consider the mussel data of Example 5.2 where the response is *muscle mass* $M$ in grams, and the predictors are the *length* $L$ and *height* $H$ of the shell in mm, the logarithm $\log W$ of the *shell width* $W$, the logarithm $\log S$ of the *shell mass* $S$ and a constant. With this

starting point, we might expect a log transformation of $M$ to be needed because $M$ and $S$ are both mass measurements and $\log S$ is being used as a predictor. Using $\log M$ would essentially reduce all measurements to the scale of length. The Box–Cox likelihood method gave $\hat{\lambda}_0 = 0.28$ with approximate 95 percent confidence interval 0.15 to 0.4. The log transformation is excluded under this inference leading to the possibility of using different transformations of the two mass measurements.

Shown in Figure 5.5 are transformation plots for four values of $\lambda$. A striking feature of these plots is the two points that stand out in three of the four plots (cases 8 and 48). The Box–Cox estimate $\hat{\lambda} = 0.28$ is evidently influenced by the two outlying points and, judging deviations from the identity line in Figure 5.5c, the mean function for the remaining points is curved. In other words, the Box–Cox estimate is allowing some visually evident curvature in the bulk of the data so it can accommodate the two outlying points. Recomputing the estimate of $\lambda_o$ without the highlighted points gives $\hat{\lambda}_o = -0.02$, which is in good agreement with the log transformation anticipated at the outset. Reconstruction of the transformation plots indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

Note that in addition to helping visualize $\hat{\lambda}$ against the data, the transformation plots can also be used to show the curvature and heteroscedasticity in the competing models indexed by $\lambda \in \Lambda_L$. Example 5.4 shows that the plot can also be used as a diagnostic to assess the success of numerical methods such as the Box–Cox procedure for estimating $\lambda_o$.

**Example 5.5: Mussel Data Again.** Return to the mussel data, this time considering the regression of $M$ on a constant and the four untransformed predictors $L$, $H$, $W$ and $S$. Figure 5.2 shows the scatterplot matrix of the predictors $L$, $H$, $W$ and $S$. Again nonlinearity is present. Figure 5.3 shows that taking the log transformations of $W$ and $S$ results in a linear scatterplot matrix for the new set of predictors $L$, $H$, $\log W$, and $\log S$. Then the search for the response transformation can be done as in Example 5.4.

## 5.3 A Review of Multiple Linear Regression

Good online references for multiple linear regression are Olive (2008, 2010). Good texts are Cook and Weisberg (1999a), Olive (2017a), Ryan (2009), and Weisberg (2005). The following review follows Olive (2017a: ch. 2) closely.

**Definition 5.9. Regression** is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable $Y$ given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

**Definition 5.10.** A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

**Definition 5.11.** Suppose that the response variable $Y$ and at least one predictor variable $x_i$ are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{5.9}$$

for $i = 1, \ldots, n$. Here $n$ is the *sample size* and the random variable $e_i$ is the $i$th *error*. Suppressing the subscript $i$, the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

See the beginning of this chapter for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ in matrix form. In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta}+e$, the $Y$ and $e$ are random variables, but we only have observed values $Y_i$ and $\boldsymbol{x}_i$. If the $e_i$ are **iid** (independent and identically distributed) with zero mean $E(e_i) = 0$ and variance $\mathrm{VAR}(e_i) = V(e_i) = \sigma^2$, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and $\sigma^2$.

**Definition 5.12.** The **constant variance MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with mean $E(e_i) = 0$ and variance $\mathrm{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)$ are independent for $i = 1, ..., n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $\boldsymbol{x}_i$. That is, observe the $\boldsymbol{x}_i$ and then act as if the observed $\boldsymbol{x}_i$ are fixed.

**Definition 5.13.** The **unimodal MLR model** has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. Note that $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$.

**Definition 5.14.** The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the $e_i$ are iid normal random variables with zero mean and variance $\sigma^2$.

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares (OLS).

**Notation.** The symbol $A \equiv B = f(c)$ means that $A$ and $B$ are equivalent and equal, and that $f(c)$ is the formula used to compute $A$ and $B$.

See Definitions 5.1 and 5.2 for fitted values, residuals, and the OLS estimator. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the $i$th fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p,$$

and the $i$th residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$. For the *ordinary least squares (OLS) estimator*, $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{HY}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists. Typically the subscript OLS is omitted, and the least squares *regression equation* is $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ where $x_1 \equiv 1$ if the model contains a constant.

**Definition 5.15.** For MLR, the *response plot* is a plot of the ESP = fitted values $= \hat{Y}_i$ versus the response variables $Y_i$, while the *residual plot* is a plot of the ESP $= \hat{Y}_i$ versus the residuals $r_i$.

**Theorem 5.1.** Suppose that the regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus $Y_i$, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

**Proof.** The identity line in the response plot is $Y = \boldsymbol{x}^T\boldsymbol{b}$. Hence the vertical deviation is $Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = r_i(\boldsymbol{b})$. □

The results in the following theorem are properties of least squares (OLS), not of the underlying MLR model. Parts f) and g) make residual plots useful. If the plotted points are linear with roughly constant variance and the correlation is zero, then the plotted points scatter about the $r = 0$ line with no other pattern. If the plotted points in a residual plot of $w$ versus $r$ do show a pattern such as a curve or a right opening megaphone, zero correlation will usually force symmetry about either the $r = 0$ line or the $w = \text{median}(w)$ line. Hence departures from the ideal plot of random scatter about the $r = 0$ line are often easy to detect.

Let the $n \times p$ design matrix of predictor variables be

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

**Warning:** If $n > p$, as is usually the case for the full rank linear model, $\boldsymbol{X}$ is not square, so $(\boldsymbol{X}^T\boldsymbol{X})^{-1} \neq \boldsymbol{X}^{-1}(\boldsymbol{X}^T)^{-1}$ since $\boldsymbol{X}^{-1}$ does not exist.

**Theorem 5.2.** Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of full rank $p$. Then

a) $\boldsymbol{H}$ is symmetric: $\boldsymbol{H} = \boldsymbol{H}^T$.

b) $\boldsymbol{H}$ is idempotent: $\boldsymbol{HH} = \boldsymbol{H}$.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$ so that $\boldsymbol{v}_j^T\boldsymbol{r} = 0$.

d) If there is a constant $\boldsymbol{v}_1 = \boldsymbol{1}$ in the model, then the sum of the residuals is zero: $\sum_{i=1}^n r_i = 0$.

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$.

f) If there is a constant in the model, then the sample correlation of the fitted values and the residuals is 0: $\text{corr}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = 0$.

g) If there is a constant in the model, then the sample correlation of the $j$th predictor with the residuals is 0: $\text{corr}(\boldsymbol{r}, \boldsymbol{v}_j) = 0$ for $j = 1, ..., p$.

**Proof.** a) $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric since $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T(\boldsymbol{X}^T)^T = \boldsymbol{X}^T\boldsymbol{X}$. Hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is symmetric since the inverse of a symmetric matrix is symmetric. (Recall that if $\boldsymbol{A}$ has an inverse then $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$.) Thus using $(\boldsymbol{A}^T)^T = \boldsymbol{A}$ and $(\boldsymbol{ABC})^T = \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T$ shows that

$$\boldsymbol{H}^T = \boldsymbol{X}^T[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^T(\boldsymbol{X}^T)^T = \boldsymbol{H}.$$

b) $\boldsymbol{HH} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$ since $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$, the $p \times p$ identity matrix.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I}_p - \boldsymbol{H})\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T]\boldsymbol{Y} = \boldsymbol{0}$. Since $\boldsymbol{v}_j$ is the $j$th column of $\boldsymbol{X}$, $\boldsymbol{v}_j^T$ is the $j$th row of $\boldsymbol{X}^T$ and $\boldsymbol{v}_j^T\boldsymbol{r} = 0$ for $j = 1, ..., p$.

d) Since $\boldsymbol{v}_1 = \boldsymbol{1}$, $\boldsymbol{v}_1^T\boldsymbol{r} = \sum_{i=1}^n r_i = 0$ by c).

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = [(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^T\boldsymbol{HY} = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{HY} = \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{Y} = 0$.

f) The sample correlation between $W$ and $Z$ is $\text{corr}(W, Z) =$

$$\frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{(n-1)s_w s_z} = \frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^n (w_i - \overline{w})^2 \sum_{i=1}^n (z_i - \overline{z})^2}}$$

where $s_m$ is the sample standard deviation of $m$ for $m = w, z$. So the result follows if $A = \sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}})(r_i - \overline{r}) = 0$. Now $\overline{r} = 0$ by d), and thus

$$A = \sum_{i=1}^n \hat{Y}_i r_i - \overline{\hat{Y}} \sum_{i=1}^n r_i = \sum_{i=1}^n \hat{Y}_i r_i$$

by d) again. But $\sum_{i=1}^n \hat{Y}_i r_i = \boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$ by e).

g) Following the argument in f), the result follows if $A = \sum_{i=1}^n (x_{i,j} - \overline{x}_j)(r_i - \overline{r}) = 0$ where $\overline{x}_j = \sum_{i=1}^n x_{i,j}/n$ is the sample mean of the $j$th predictor. Now $\overline{r} = \sum_{i=1}^n r_i/n = 0$ by d), and thus

$$A = \sum_{i=1}^n x_{i,j} r_i - \overline{x}_j \sum_{i=1}^n r_i = \sum_{i=1}^n x_{i,j} r_i$$

by d) again. But $\sum_{i=1}^n x_{i,j} r_i = \boldsymbol{v}_j^T\boldsymbol{r} = 0$ by c). □

### *5.3.1* The ANOVA F Test

After fitting least squares and checking the response and residual plots to see that an MLR model is reasonable, the next step is to check whether there is an MLR relationship between $Y$ and the nontrivial predictors $x_2$, ..., $x_p$. If at least one of these predictors is useful, then the OLS fitted values $\hat{Y}_i$ should be used. If none of the nontrivial predictors is useful, then $\overline{Y}$ will give as good predictions as $\hat{Y}_i$. Here the *sample mean* $\overline{Y}$ is given by Definition 2.2. In the definition below, $SSE$ is the sum of squared residuals and a residual $r_i = \hat{e}_i =$ "errorhat." In the literature "errorhat" is often rather misleadingly abbreviated as "error."

**Definition 5.16.** Assume that a constant is in the MLR model.
a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{5.10}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{5.11}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{5.12}$$

The result in the following theorem is a property of least squares (OLS), not of the underlying MLR model. An obvious application is that given any two of SSTO, SSE, and SSR, the 3rd sum of squares can be found using the formula $SSTO = SSE + SSR$.

**Theorem 5.3.** Assume that a constant is in the MLR model. Then $SSTO = SSE + SSR$.

**Proof.**

$$SSTO = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 = SSE + SSR + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}).$$

Hence the result follows if

$$A \equiv \sum_{i=1}^{n} r_i (\hat{Y}_i - \overline{Y}) = 0.$$

But

$$A = \sum_{i=1}^{n} r_i \hat{Y}_i - \overline{Y} \sum_{i=1}^{n} r_i = 0$$

by Theorem 5.2 d) and e). □

**Definition 5.17.** Assume that a constant is in the MLR model and that SSTO $\neq 0$. The **coefficient of multiple determination**

$$R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

where $\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})$ is the sample correlation of $Y_i$ and $\hat{Y}_i$.

**Warnings:** i) $0 \leq R^2 \leq 1$, but small $R^2$ does not imply that the MLR model is bad.

ii) If the MLR model contains a constant, then there are several equivalent formulas for $R^2$. If the model does not contain a constant, then $R^2$ depends on the software package.

iii) $R^2$ does not have much meaning unless the response plot and residual plot both look good.

iv) $R^2$ tends to be too high if $n$ is small.

v) $R^2$ tends to be too high if there are two or more separated clusters of data in the response plot.

vi) $R^2$ is too high if the number of predictors $p$ is close to $n$.

vii) In large samples $R^2$ will be large (close to one) if $\sigma^2$ is small compared to the sample variance $S_Y^2$ of the response variable $Y$. $R^2$ is also large if the sample variance of $\hat{Y}$ is close to $S_Y^2$. Thus $R^2$ is sometimes interpreted as the proportion of the variability of $Y$ explained by conditioning on $\boldsymbol{x}$, but warnings i) - v) suggest that $R^2$ may not have much meaning.

The following 2 theorems suggest that $R^2$ does not behave well when many predictors that are not needed in the model are included in the model. Such a variable is sometimes called a noise variable and the MLR model is "fitting noise." Theorem 5.5 appears, for example, in Cramér (1946, pp. 414-415), and suggests that $R^2$ should be considerably larger than $p/n$ if the predictors are useful. Note that if $n = 10p$ and $p \geq 2$, then under the conditions of Theorem 5.5, $E(R^2) \leq 0.1$.

**Theorem 5.4.** Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) $R^2$.

**Theorem 5.5.** Assume that a constant $\beta_1$ is in the MLR model, that $\beta_2 = \cdots = \beta_p = 0$ and that the $e_i$ are iid $N(0, \sigma^2)$. Hence the $Y_i$ are iid $N(\beta_1, \sigma^2)$. Then

a) $R^2$ follows a beta distribution: $R^2 \sim \mathrm{beta}(\frac{\mathrm{p}-1}{2}, \frac{\mathrm{n}-\mathrm{p}}{2})$.

b)

$$E(R^2) = \frac{p-1}{n-1}.$$

c)

$$\text{VAR}(R^2) = \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}.$$

Notice that each $SS/n$ estimates the variability of some quantity. $SSTO/n \approx S_Y^2$, $SSE/n \approx S_e^2 = \sigma^2$, and $SSR/n \approx S_{\hat{Y}}^2$.

**Definition 5.18.** Assume that a constant is in the MLR model. Associated with each SS in Definition 5.16 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSR, df $= p-1$ and $MSR = SSR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Under mild conditions, if the MLR model is appropriate, then MSE is a $\sqrt{n}$ consistent estimator of $\sigma^2$ by Su and Cook (2012).

The ANOVA $F$ test tests whether any of the nontrivial predictors $x_2, ..., x_p$ are needed in the OLS MLR model, that is, whether $Y_i$ should be predicted by the OLS fit $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p$ or with the sample mean $\overline{Y}$. ANOVA stands for analysis of variance, and the computer output needed to perform the test is contained in the ANOVA table. Below is an ANOVA table given in symbols. Sometimes "Regression" is replaced by "Model" and "Residual" by "Error."

Summary Analysis of Variance Table

| Source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| Regression | $p-1$ | SSR | MSR | $F_0$=MSR/MSE | for $H_0$: |
| Residual | $n-p$ | SSE | MSE | | $\beta_2 = \cdots = \beta_p = 0$ |

**Remark 5.1.** Recall that for a 4 step test of hypotheses, the p–value is the probability of getting a test statistic as extreme as the test statistic actually observed and that $H_0$ is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if $\delta$ is not given. The 4th step is the nontechnical conclusion which is crucial for presenting your results to people who are not familiar with MLR. Replace $Y$ and $x_2, ..., x_p$ by the actual variables used in the MLR model.

**Notation.** The p–value $\equiv$ pvalue given by output tends to only be correct for the normal MLR model. Hence the output is usually only giving an estimate of the pvalue, which will often be denoted by *pval*. So reject $H_0$ if pval $\leq \delta$. Often

$$\text{pval} - \text{pvalue} \xrightarrow{\text{P}} 0$$

(converges to 0 in probability, so pval is a consistent estimator of pvalue) as the sample size $n \to \infty$. Then the computer output pval is a good estimator of the unknown pvalue. We will use $Fo \equiv F_0$, $Ho \equiv H_0$, and $Ha \equiv H_A \equiv H_1$.

**The 4 step ANOVA F test of hypotheses is below**.
i) State the hypotheses $H_0 : \beta_2 = \cdots = \beta_p = 0$ $H_A$: not $H_0$.
ii) Find the test statistic $F_0 = MSR/MSE$ or obtain it from output.
iii) Find the pval from output or use the $F$–table: pval =

$$P(F_{p-1,n-p} > F_0).$$

iv) State whether you reject $H_0$ or fail to reject $H_0$. If $H_0$ is rejected, conclude that there is an MLR relationship between $Y$ and the predictors $x_2$, ..., $x_p$. If you fail to reject $H_0$, conclude that there is not an MLR relationship between $Y$ and the predictors $x_2$, ..., $x_p$. (Or there is not enough evidence to conclude that there is an MLR relationship between $Y$ and the predictors.)

Some assumptions are needed on the ANOVA $F$ test. Assume that both the response and residual plots look good. It is crucial that there are no outliers. Then a rule of thumb is that if $n - p$ is large, then the ANOVA $F$ test p–value is approximately correct. An analogy can be made with the central limit theorem, $\overline{Y}$ is a good estimator for $\mu$ if the $Y_i$ are iid $N(\mu, \sigma^2)$ and also a good estimator for $\mu$ if the data are iid with mean $\mu$ and variance $\sigma^2$ if $n$ is large enough.

If all of the $\boldsymbol{x}_i$ are different (no replication) and if the number of predictors $p = n$, then the OLS fit $\hat{Y}_i = Y_i$ and $R^2 = 1$. Notice that $H_0$ is rejected if the statistic $F_0$ is large. More precisely, reject $H_0$ if

$$F_0 > F_{p-1,n-p,1-\delta}$$

where

$$P(F \leq F_{p-1,n-p,1-\delta}) = 1 - \delta$$

when $F \sim F_{p-1,n-p}$. Since $R^2$ increases to 1 while $(n-p)/(p-1)$ decreases to 0 as $p$ increases to $n$, Theorem 5.6a below implies that if $p$ is large then the $F_0$ statistic may be small even if some of the predictors are very good. It is a good idea to use $n \geq 10p$ or at least $n \geq 5p$ if possible.

**Theorem 5.6.** Assume that the MLR model has a constant $\beta_1$.
a)

$$F_0 = \frac{MSR}{MSE} = \frac{R^2}{1 - R^2} \quad \frac{n-p}{p-1}.$$

b) If the errors $e_i$ are iid $N(0, \sigma^2)$, and if $H_0 : \beta_2 = \cdots = \beta_p = 0$ is true, then $F_0$ has an $F$ distribution with $p - 1$ numerator and $n - p$ denominator degrees of freedom: $F_0 \sim F_{p-1,n-p}$.

c) If the errors are iid with mean 0 and variance $\sigma^2$, if the error distribution is close to normal, and if $n - p$ is large enough, and if $H_0$ is true, then $F_0 \approx F_{p-1,n-p}$ in that the p-value from the software (pval) is approximately correct.

**Remark 5.2.** When a constant is not contained in the model (i.e. $x_{i,1}$ is not equal to 1 for all $i$), then the computer output still produces an ANOVA table with the test statistic and p–value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now $H_0 : \beta_1 = \cdots = \beta_p = 0$ $H_A$: not $H_0$, and you are testing whether or not there is an MLR relationship between $Y$ and $x_1, ..., x_p$. An MLR model without a constant (no intercept) is sometimes called a "regression through the origin."

### *5.3.2* The Partial F Test

Suppose that there is data on variables $Z$, $w_1, ..., w_r$ and that a useful MLR model has been made using $Y = t(Z)$, $x_1 \equiv 1$, $x_2, ..., x_p$ where each $x_i$ is some function of $w_1, ..., w_r$. This useful model will be called the full model. It is important to realize that the full model does not need to use every variable $w_j$ that was collected. For example, variables with outliers or missing values may not be used. Forming a useful full model is often very difficult, and it is often not reasonable to assume that the candidate full model is good based on a single data set, especially if the model is to be used for prediction.

Even if the full model is useful, the investigator will often be interested in checking whether a model that uses fewer predictors will work just as well. For example, perhaps $x_p$ is a very expensive predictor but is not needed given that $x_1, ..., x_{p-1}$ are in the model. Also a model with fewer predictors tends to be easier to understand.

**Definition 5.19.** Let the **full model** use $Y$, $x_1 \equiv 1$, $x_2, ..., x_p$ and let the **reduced model** use $Y$, $x_1$, $x_{i_2}, ..., x_{i_q}$ where $\{i_2, ..., i_q\} \subset \{2, ..., p\}$.

The partial $F$ test is used to test whether the reduced model is good in that it can be used instead of the full model. It is crucial that the reduced and full models be selected before looking at the data. If the reduced model is selected after looking at the full model output and discarding the worst variables, then the $p$–value for the partial $F$ test will be too high. If the data needs to be looked at to build the full model, as is often the case, data splitting is useful.

For (ordinary) least squares, usually a constant is used, and we are assuming that both the full model and the reduced model contain a constant. The partial $F$ test has null hypothesis $H_0 : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0$, and alternative hypothesis $H_A$ : at least one of the $\beta_{i_j} \neq 0$ for $j > q$. The null hypothesis is equivalent to $H_0$: "the reduced model is good." Since only the full model and

reduced model are being compared, the alternative hypothesis is equivalent to $H_A$: "the reduced model is not as good as the full model, so use the full model," or more simply, $H_A$ : "use the full model."

To perform the partial $F$ test, fit the full model and the reduced model and obtain the ANOVA table for each model. The quantities $df_F$, SSE(F) and MSE(F) are for the full model and the corresponding quantities from the reduced model use an $R$ instead of an $F$. Hence SSE(F) and SSE(R) are the residual sums of squares for the full and reduced models, respectively. Shown below is output only using symbols.
Full model

| Source df | SS | MS | $F_0$ and p-value |
|---|---|---|---|
| Regression $p-1$ | SSR | MSR | $F_0$=MSR/MSE |
| Residual $df_F = n-p$ | SSE(F) | MSE(F) | for $H_0 : \beta_2 = \cdots = \beta_p = 0$ |

Reduced model

| Source df | SS | MS | $F_0$ and p-value |
|---|---|---|---|
| Regression $q-1$ | SSR | MSR | $F_0$=MSR/MSE |
| Residual $df_R = n-q$ | SSE(R) | MSE(R) | for $H_0 : \beta_2 = \cdots = \beta_q = 0$ |

**The 4 step partial F test of hypotheses is below**. i) State the hypotheses. $H_0$: the reduced model is good $H_A$: use the full model
ii) Find the test statistic. $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the pval = P($F_{df_R - df_F, df_F} > F_R$). ( Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$, while pval is the estimated p–value.)
iv) State whether you reject $H_0$ or fail to reject $H_0$. Reject $H_0$ if the pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject $H_0$ and conclude that the reduced model is good.

Sometimes software has a shortcut. In particular, the $R$ software uses the `anova` command. As an example, assume that the full model uses $x_2$ and $x_3$ while the reduced model uses $x_2$. Both models contain a constant. Then the following commands will perform the partial $F$ test. (On the computer screen the second command looks more like
red <− lm(y∼x2).)

```
full <- lm(y~x2+x3)
red <- lm(y~x2)
anova(red,full)
```

For an $n \times 1$ vector $\boldsymbol{a}$, let

$$\|\boldsymbol{a}\| = \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{\boldsymbol{a}^T \boldsymbol{a}}$$

be the Euclidean norm of $\boldsymbol{a}$. If $\boldsymbol{r}$ and $\boldsymbol{r}_R$ are the vector of residuals from the full and reduced models, respectively, notice that $SSE(F) = \|\boldsymbol{r}\|^2$ and $SSE(R) = \|\boldsymbol{r}_R\|^2$.

The following theorem suggests that $H_0$ is rejected in the partial $F$ test if the change in residual sum of squares $SSE(R) - SSE(F)$ is large compared to $SSE(F)$. If the change is small, then $F_R$ is small and the test suggests that the reduced model can be used.

**Theorem 5.7.** Let $R^2$ and $R_R^2$ be the multiple coefficients of determination for the full and reduced models, respectively. Let $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_R$ be the vectors of fitted values for the full and reduced models, respectively. Then the test statistic in the partial $F$ test is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) =$$

$$\left[\frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\hat{\boldsymbol{Y}}_R\|^2}{df_R - df_F}\right] / MSE(F) =$$

$$\frac{SSE(R) - SSE(F)}{SSE(F)} \ \frac{n-p}{p-q} = \frac{R^2 - R_R^2}{1 - R^2} \ \frac{n-p}{p-q}.$$

**Definition 5.20.** An **FF plot** is a plot of fitted values from 2 different models or fitting methods. An **RR plot** is a plot of residuals from 2 different models or fitting methods.

Six plots are useful diagnostics for the partial $F$ test: the RR plot with the full model residuals on the vertical axis and the reduced model residuals on the horizontal axis, the FF plot with the full model fitted values on the vertical axis, and always make the response and residual plots for the full and reduced models. Suppose that the full model is a useful MLR model. If the reduced model is good, then the response plots from the full and reduced models should be very similar, visually. Similarly, the residual plots from the full and reduced models should be very similar, visually. Finally, the correlation of the plotted points in the RR and FF plots should be high, $\geq 0.95$, say, and the plotted points in the RR and FF plots should cluster tightly about the identity line. Add the identity line to both the RR and FF plots as a visual aid. Also add the OLS line from regressing $\boldsymbol{r}$ on $\boldsymbol{r}_R$ to the RR plot (the OLS line is the identity line in the FF plot). If the reduced model is good, then the OLS line should nearly coincide with the identity line in that it should be difficult to see that the two lines intersect at the origin. If the FF plot looks good but the RR plot does not, the reduced model may

be good if the main goal of the analysis is to predict $Y$. These plots are also useful for other methods such as lasso.

### *5.3.3* The Wald t Test

Often investigators hope to examine $\beta_k$ in order to determine the importance of the predictor $x_k$ in the model; however, $\beta_k$ is the coefficient for $x_k$ given that the other predictors are in the model. Hence $\beta_k$ depends strongly on the other predictors in the model. Suppose that the model has an intercept: $x_1 \equiv 1$. The predictor $x_k$ is highly correlated with the other predictors if the OLS regression of $x_k$ on $x_1, ..., x_{k-1}, x_{k+1}, ..., x_p$ has a high coefficient of determination $R_k^2$. If this is the case, then often $x_k$ is not needed in the model given that the other predictors are in the model. If at least one $R_k^2$ is high for $k \geq 2$, then there is multicollinearity among the predictors.

As an example, suppose that $Y = \textit{height}$, $x_1 \equiv 1$, $x_2 = \textit{left leg length}$, and $x_3 = \textit{right leg length}$. Then $x_2$ should not be needed given $x_3$ is in the model and $\beta_2 = 0$ is reasonable. Similarly $\beta_3 = 0$ is reasonable. On the other hand, if the model only contains $x_1$ and $x_2$, then $x_2$ is extremely important with $\beta_2$ near 2. If the model contains $x_1, x_2, x_3$, $x_4 = \textit{height at shoulder}$, $x_5 = \textit{right arm length}$, $x_6 = \textit{head length}$, and $x_7 = \textit{length of back}$, then $R_i^2$ may be high for each $i \geq 2$. Hence $x_i$ is not needed in the MLR model for $Y$ given that the other predictors are in the model.

**Definition 5.21.** The 100 $(1 - \delta)$ % CI for $\beta_k$ is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\, se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p \geq 30$, the N(0,1) cutoff $z_{1-\delta/2}$ may be used.

**Know how to do** the 4 step Wald $t$–test of hypotheses.
i) State the hypotheses $H_0 : \beta_k = 0$ $H_A : \beta_k \neq 0$.
ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find pval from output or use the $t$–table: pval =

$$2P(t_{n-p} < -|t_{o,k}|) = 2P(t_{n-p} > |t_{o,k}|).$$

Use the normal table or the $d = Z$ line in the $t$–table if the degrees of freedom $d = n - p \geq 30$. Again pval is the estimated p–value.
iv) State whether you reject $H_0$ or fail to reject $H_0$ and give a nontechnical sentence restating your conclusion in terms of the story problem.

Recall that $H_0$ is rejected if the pval $\leq \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if $\delta$ is not given. If $H_0$ is rejected, then conclude that $x_k$ is needed in the MLR model for $Y$ given that the other predictors are in the model. If you fail to reject $H_0$, then conclude that $x_k$ is not needed in the MLR model for $Y$ given that the other predictors are in the model. (Or there is

not enough evidence to conclude that $x_k$ is needed in the MLR model given that the other predictors are in the model.) Note that $x_k$ could be a very useful individual predictor, but may not be needed if other predictors are added to the model.

### *5.3.4* The OLS Criterion

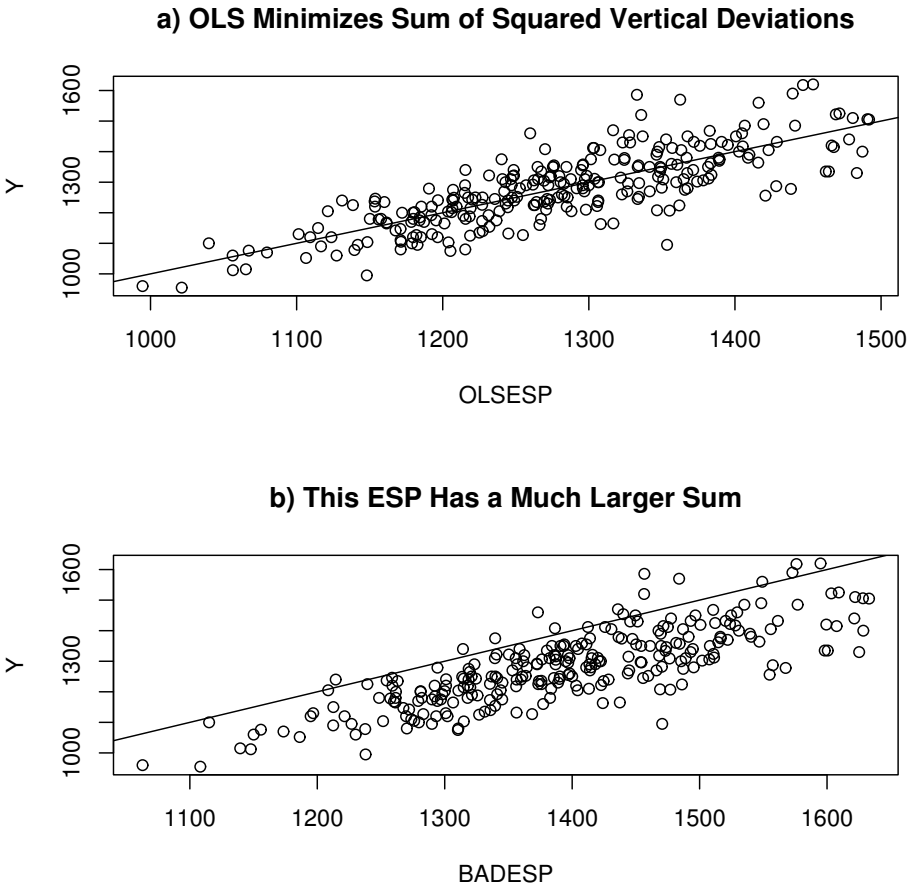


**Fig. 5.6** The OLS Fit Minimizes the Sum of Squared Residuals

The OLS estimator $\hat{\boldsymbol{\beta}}$ minimizes the OLS criterion

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$$

where the residual $r_i(\boldsymbol{\eta}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta}$. In other words, let $r_i = r_i(\hat{\boldsymbol{\beta}})$ be the OLS residuals. Then $\sum_{i=1}^n r_i^2 \leq \sum_{i=1}^n r_i^2(\boldsymbol{\eta})$ for any $p \times 1$ vector $\boldsymbol{\eta}$, and the equality holds (if and only if) iff $\boldsymbol{\eta} = \hat{\boldsymbol{\beta}}$ if the $n \times p$ design matrix $\boldsymbol{X}$ is of full rank $p \leq n$. In particular, if $\boldsymbol{X}$ has full rank $p$, then $\sum_{i=1}^n r_i^2 < \sum_{i=1}^n r_i^2(\boldsymbol{\beta}) = \sum_{i=1}^n e_i^2$ even if the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a good approximation to the data.

**Warning:** Often $\boldsymbol{\eta}$ is replaced by $\boldsymbol{\beta}$: $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^n r_i^2(\boldsymbol{\beta})$. This notation is often used in Statistics when there are estimating equations. For example, maximum likelihood estimation uses the log likelihood $\log(L(\boldsymbol{\theta}))$ where $\boldsymbol{\theta}$ is the vector of unknown parameters and the dummy variable in the log likelihood.

**Example 5.6.** When a model depends on the predictors $\boldsymbol{x}$ only through the linear combination $\boldsymbol{x}^T\boldsymbol{\beta}$, then $\boldsymbol{x}^T\boldsymbol{\beta}$ is called a sufficient predictor and $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is called an estimated sufficient predictor (ESP). For OLS the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, and the fitted value $\hat{Y} = ESP$. To illustrate the OLS criterion graphically, consider the Gladstone (1905) data where we used *brain weight* as the response. A constant, $x_2 = age$, $x_3 = sex$, and $x_4 = (size)^{1/3}$ were used as predictors after deleting five "infants" from the data set. In Figure 5.6a, the OLS response plot of the OLS ESP $= \hat{Y}$ versus $Y$ is shown. The vertical deviations from the identity line are the residuals, and OLS minimizes the sum of squared residuals. If any other ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ is plotted versus $Y$, then the vertical deviations from the identity line are the residuals $r_i(\boldsymbol{\eta})$. For this data, the OLS estimator $\hat{\boldsymbol{\beta}} = (498.726, -1.597, 30.462, 0.696)^T$. Figure 5.6b shows the response plot using the ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (498.726, -1.597, 30.462, 0.796)^T$. Hence only the coefficient for $x_4$ was changed; however, the residuals $r_i(\boldsymbol{\eta})$ in the resulting plot are much larger in magnitude on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting $\boldsymbol{\eta}$ will be such that the squared residuals are massive.

**Theorem 5.8.** The OLS estimator $\hat{\boldsymbol{\beta}}$ is the unique minimizer of the OLS criterion if $\boldsymbol{X}$ has full rank $p \leq n$.

**Proof: Seber and Lee (2003, pp. 36-37).** Recall that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and notice that $(\boldsymbol{I}-\boldsymbol{H})^T = \boldsymbol{I}-\boldsymbol{H}$, that $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$ and that $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. Let $\boldsymbol{\eta}$ be any $p \times 1$ vector. Then

$$(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y})^T(\boldsymbol{H}\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\eta}) =$$

$$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}) = \boldsymbol{0}.$$

$$\text{Thus } \ \mathrm{Q_{OLS}}(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2 =$$

$$\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2 + 2(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}).$$

Hence

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{5.13}$$

So

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

with equality iff

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\eta}) = \boldsymbol{0}$$

iff $\hat{\boldsymbol{\beta}} = \boldsymbol{\eta}$ since $\boldsymbol{X}$ is full rank. □

Alternatively calculus can be used. Notice that $r_i(\boldsymbol{\eta}) = Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p$. Recall that $\boldsymbol{x}_i^T$ is the $i$th row of $\boldsymbol{X}$ while $\boldsymbol{v}_j$ is the $j$th column. Since $Q_{OLS}(\boldsymbol{\eta}) =$

$$\sum_{i=1}^{n}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p)^2,$$

the $j$th partial derivative

$$\frac{\partial Q_{OLS}(\boldsymbol{\eta})}{\partial \eta_j} = -2\sum_{i=1}^{n} x_{i,j}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p) = -2(\boldsymbol{v}_j)^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})$$

for $j = 1, ..., p$. Combining these equations into matrix form, setting the derivative to zero and calling the solution $\hat{\boldsymbol{\beta}}$ gives

$$\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0},$$

or

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{5.14}$$

Equation (5.14) is known as the **normal equations**. If $\boldsymbol{X}$ has full rank then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. To show that $\hat{\boldsymbol{\beta}}$ is the global minimizer of the OLS criterion, use the argument following Equation (5.13).

## 5.4 Asymptotically Optimal Prediction Intervals

This section gives estimators for predicting a future or new value $Y_f$ of the response variable given the predictors $\boldsymbol{x}_f$, and for estimating the mean $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$. This mean is conditional on the values of the predictors $\boldsymbol{x}_f$, but the conditioning is often suppressed. See

**Warning:** All too often the MLR model seems to fit the data

$$(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$$

well, but when new data is collected, a very different MLR model is needed to fit the new data well. In particular, the MLR model seems to fit the data

$(Y_i, \boldsymbol{x}_i)$ well for $i = 1, ..., n$, but when the researcher tries to predict $Y_f$ for a new vector of predictors $\boldsymbol{x}_f$, the prediction is very poor in that $\hat{Y}_f$ is not close to the $Y_f$ actually observed. **Wait until after the MLR model has been shown to make good predictions before claiming that the model gives good predictions!**

There are several reasons why the MLR model may not fit new data well. i) The model building process is usually iterative. Data $Z, w_1, ..., w_k$ is collected. If the model is not linear, then functions of $Z$ are used as a potential response and functions of the $w_i$ as potential predictors. After trial and error, the functions are chosen, resulting in a final MLR model using $Y$ and $x_1, ..., x_p$. Since the same data set was used during the model building process, biases are introduced and the MLR model fits the "training data" better than it fits new data. Suppose that $Y$, $x_1, ..., x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

ii) If $(Y_f, \boldsymbol{x}_f)$ come from a different population than the population of $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$, then prediction for $Y_f$ can be arbitrarily bad.

iii) Even a good MLR model may not provide good predictions for an $\boldsymbol{x}_f$ that is far from the $\boldsymbol{x}_i$ (extrapolation).

iv) The MLR model may be missing important predictors (underfitting).

v) The MLR model may contain unnecessary predictors (overfitting).

Two remedies for i) are a) use previously published studies to select an MLR model before gathering data. b) Do a trial study. Collect some data, build an MLR model using the iterative process. Then use this model as the prespecified model and collect data for the main part of the study. Better yet, do a trial study, specify a model, collect more trial data, improve the specified model and repeat until the latest specified model works well. Unfortunately, trial studies are often too expensive or not possible because the data is difficult to collect. Also, often the population from a published study is quite different from the population of the data collected by the researcher. Then the MLR model from the published study is not adequate.

**Definition 5.22.** Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the $i$th diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then $h_i$ is called the $i$th **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

**Rule of thumb 5.3.** Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of $x_i$ versus $x_j$ should not have any strong nonlinearities. If there are strong nonlinearities

among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$.

**Example 5.7.** Consider predicting $Y = \textit{weight}$ from $x = \textit{height}$ and a constant from data collected on men between 18 and 24 where the minimum height was 57 and the maximum height was 79 inches. The OLS equation was $\hat{Y} = -167 + 4.7x$. If $x = 70$ then $\hat{Y} = -167 + 4.7(70) = 162$ pounds. If $x = 1$ inch, then $\hat{Y} = -167 + 4.7(1) = -162.3$ pounds. It is impossible to have negative weight, but it is also impossible to find a 1 inch man. This MLR model should not be used for $x$ far from the interval $(57, 79)$.

**Definition 5.23.** Consider the iid error MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$. Then **regression function** is the hyperplane

$$E(Y) \equiv E(Y|\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{5.15}$$

Assume OLS is used to find $\hat{\boldsymbol{\beta}}$. Then the **point estimator** of $Y_f$ given $\boldsymbol{x} = \boldsymbol{x}_f$ is

$$\hat{Y}_f = x_{f,1}\hat{\beta}_1 + \cdots + x_{f,p}\hat{\boldsymbol{\beta}}_p = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}. \tag{5.16}$$

The **point estimator** of $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$ given $\boldsymbol{x} = \boldsymbol{x}_f$ is also $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$. Assume that the MLR model contains a constant $\beta_1$ so that $x_1 \equiv 1$. The large sample 100 $(1 - \delta)$% confidence interval (CI) for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(\hat{Y}_f) \tag{5.17}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if $T$ has a $t$ distribution with $n - p$ degrees of freedom. Generally $se(\hat{Y}_f)$ will come from output, but

$$se(\hat{Y}_f) = \sqrt{MSE\ h_f} = \sqrt{MSE\ \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f}.$$

Recall the interpretation of a 100 $(1 - \delta)$% CI for a parameter $\mu$ is that if you collect data then form the CI, and repeat for a total of $k$ times where the $k$ trials are independent from the same population, then the probability that $m$ of the CIs will contain $\mu$ follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% CIs are made, $\rho = 0.95$ and about 95 of the CIs will contain $\mu$ while about 5 will not. Any given CI may (good sample) or may not (bad sample) contain $\mu$, but the probability of a "bad sample" is $\delta$.

The following theorem is analogous to the central limit theorem and the theory for the $t$–interval for $\mu$ based on $\overline{Y}$ and the sample standard deviation (SD) $S_Y$. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance $\sigma^2$, then $\overline{Y}$ is asymptotically normal and the $t$–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if

the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Convergence in probability, $Y_n \xrightarrow{P} c$, is similar to other types of convergence: $Y_n$ is likely to be close to $c$ if the sample size $n$ is large enough. Parts a) and b) of Theorem 5.2 are due to Huber and Ronchetti (2009, pp. 156-158). For c), see Sen and Singer (1993, p. 280). Part c) implies that $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

**Theorem 5.9:** Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

a) $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \to E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ in probability for $i = 1, ..., n$ as $n \to \infty$.

b) All of the least squares estimators $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ are asymptotically normal where $\boldsymbol{a}$ is any fixed constant $p \times 1$ vector.

c) OLS CLT: Suppose that the $e_i$ are iid and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}.$$

Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W}). \tag{5.18}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{5.19}$$

**Definition 5.24.** A large sample $100(1-\delta)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable $Y_f$ is independent of $Y_1, ..., Y_n$. Then the $100(1-\delta)\%$ PI for $Y_f$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred) \tag{5.20}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if $T$ has a $t$ distribution with $n - p$ degrees of freedom. Generally $se(pred)$ will come from output, but $se(pred) = \sqrt{MSE\ (1 + h_f)}$.

Often we want the coverage $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$. The interpretation of a 100 $(1 - \delta)\%$ PI for a random variable $Y_f$ is similar to that of a CI. Collect data, then form the PI, and repeat for a total of $k$ times where $k$ trials are independent from the same population. If $Y_{fi}$ is the $i$th random variable and $PI_i$ is the $i$th PI, then the probability that $Y_{fi} \in PI_i$ for $m$ of the PIs follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size $n$ goes to $\infty$ while the length of the PI converges to some nonzero number $L$, say. Secondly, the CI for $E(Y_f|\boldsymbol{x}_f)$ given in Definition 5.23 tends to work well for the iid error MLR model if the sample size is large while the PI in Definition 5.24 is made under the assumption that the $e_i$ are iid $N(0, \sigma^2)$ and may not perform well if the normality assumption is violated.

To see this, consider $\boldsymbol{x}_f$ such that the heights $Y$ of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for $E(Y|\boldsymbol{x}_f)$ should be centered at about 66 and the length should go to zero as $n$ gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval $66 \pm 1.96(3)$. This result follows because if $Y \sim N(66, 9)$ then $P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) = 0.025$. In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Assume that the iid error MLR model is valid so that $e$ is from some distribution with 0 mean and variance $\sigma^2$. Olive (2007) shows that if $1 - \gamma$ is the asymptotic coverage of the classical nominal $100(1 - \delta)\%$ PI (5.20), then

$$1 - \gamma = P(-\sigma z_{1-\delta/2} \leq e \leq \sigma z_{1-\delta/2}) \geq 1 - \frac{1}{z_{1-\delta/2}^2} \tag{5.21}$$

where the inequality follows from Chebyshev's inequality. Hence the asymptotic coverage of the nominal 95% PI is at least 73.9%. The 95% PI (5.20) was often quite accurate in that the asymptotic coverage was close to 95% for a wide variety of error distributions. The 99% and 90% PIs did not perform as well.

Let $\xi_\delta$ be the $\delta$ percentile of the error $e$, i.e., $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample $\delta$ percentile of the residuals. Then the results from Theorem 5.9 suggest that the residuals $r_i$ estimate the errors $e_i$, and that the sample percentiles of the residuals $\hat{\xi}_\delta$ estimate $\xi_\delta$. For many error distributions,

$$E(MSE) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-p}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right).$$

This result suggests that $\sqrt{\frac{n}{n-p}} r_i \approx e_i$. Using

$$a_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n}{n-p}} \sqrt{(1 + h_f)}, \tag{5.22}$$

a large sample semiparametric $100(1-\delta)\%$ PI for $Y_f$ is

$$[\hat{Y}_f + a_n\hat{\xi}_{\delta/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\delta/2}]. \tag{5.23}$$

This PI is very similar to the classical PI except that $\hat{\xi}_\delta$ is used instead of $\sigma z_\delta$ to estimate the error percentiles $\xi_\delta$.

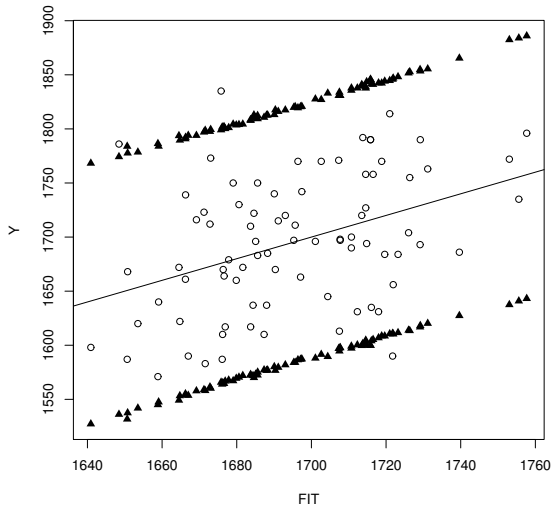


**Fig. 5.7** 95% PI Limits for Buxton Data

**Example 5.8.** For the Buxton (1920) data suppose that the response $Y =$ *height* and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five outliers were deleted leaving 82 cases. Figure 5.7 shows a response plot of the fitted values versus the response $Y$ with the identity line added as a visual aid. The plot suggests that the model is good since the plotted points scatter about the identity line in an evenly populated band although the relationship is rather weak since the correlation of the plotted points is not very high. The triangles represent the upper and lower limits of the semiparametric 95% PI (5.23). Notice that 79 (or 96%) of the $Y_i$ fell within their corresponding PI while 3 $Y_i$ did not. A plot using the classical PI (5.20) would be very similar for this data. The plot was made with the following $R$ commands, using the *rpack* function `piplot`.

```
x <- buxx[-c(61,62,63,64,65),]
Y <- buxy[-c(61,62,63,64,65)]
piplot(x,Y)
```

| Label | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Constant | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | $t_{o,1}$ | for Ho: $\beta_1 = 0$ |
| $x_2$ | $\hat{\beta}_2$ | $se(\hat{\beta}_2)$ | $t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$ | for Ho: $\beta_2 = 0$ |
| $\vdots$ | | | | |
| $x_p$ | $\hat{\beta}_p$ | $se(\hat{\beta}_p)$ | $t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$ | for Ho: $\beta_p = 0$ |

Given output showing $\hat{\boldsymbol{\beta}}_i$ and given $\boldsymbol{x}_f$, $se(pred)$ and $se(\hat{Y}_f)$, Example 5.9 shows how to find $\hat{Y}_f$, a CI for $E(Y_f|\boldsymbol{x}_f)$ and a PI for $Y_f$. Shown above is typical output in symbols.

**Example 5.9.** The Rouncefield (1995) data are female and male life expectancies from $n = 91$ countries. Suppose that it is desired to predict female life expectancy $Y$ from male life expectancy $X$. Suppose that if $X_f = 60$, then $se(\text{pred}) = 2.1285$, and $se(\hat{Y}_f) = 0.2241$. Below is some output.

```
Label       Estimate   Std. Error   t-value   p-value
Constant   -2.93739    1.42523       -2.061    0.0422
mlife       1.12359    0.0229362     48.988    0.0000
```

a) Find $\hat{Y}_f$ if $X_f = 60$.

Solution: In this example, $\boldsymbol{x}_f = (1, X_f)^T$ since a constant is in the output above. Thus $\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f = -2.93739 + 1.12359(60) = 64.478$.

b) If $X_f = 60$, find a 90% confidence interval for $E(Y) \equiv E(Y_f|\boldsymbol{x}_f)$.

Solution: The CI is $\hat{Y}_f \pm t_{1-\alpha/2,n-2} se(\hat{Y}_f) = 64.478 \pm 1.645(0.2241) = 64.478 \pm 0.3686 = (64.1094, 64.8466)$. To use the $t$–table on the last page of Chapter 14, use the 2nd to last row marked by $Z$ since $d = df = n - 2 = 90 > 30$. In the 3rd to last row find CI = 90% and intersect the 90% column and the Z row to get the value of $t_{0.95,90} \approx z_{.95} = 1.645$.

c) If $X_f = 60$, find a 90% prediction interval for $Y_f$.

Solution: The CI is $\hat{Y}_f \pm t_{1-\alpha/2,n-2} se(pred) = 64.478 \pm 1.645(2.1285) = 64.478 \pm 3.5014 = (60.9766, 67.9794)$.

An asymptotically conservative (ac) $100(1-\delta)\%$ PI has asymptotic coverage $1 - \gamma \geq 1 - \delta$. We used the (ac) $100(1-\delta)\%$ PI

$$\hat{Y}_f \pm \sqrt{\frac{n}{n-p}} \max(|\hat{\xi}_{\delta/2}|, |\hat{\xi}_{1-\delta/2}|) \sqrt{(1 + h_f)} \tag{5.24}$$

which has asymptotic coverage

$$1 - \gamma = P[-\max(|\xi_{\delta/2}|, |\xi_{1-\delta/2}|) < e < \max(|\xi_{\delta/2}|, |\xi_{1-\delta/2}|)]. \tag{5.25}$$

Notice that $1-\delta \leq 1-\gamma \leq 1-\delta/2$ and $1-\gamma = 1-\delta$ if the error distribution is symmetric with a pdf.

In the simulations described below, $\hat{\xi}_\delta$ will be the sample percentile for the PIs (5.23) and (5.24). A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth($c$) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $[r_{(d)}, r_{(d+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$ correspond to the interval with the smallest distance. Then the large sample 100 $(1-\delta)$% PI for $Y_f$ is

$$[\hat{Y}_f + a_n \tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}] \tag{5.26}$$

where $a_n$ is given by (5.22).

A small simulation study compares the PI lengths and coverages for sample sizes $n = 50, 100$ and 1000 for several error distributions. The value $n = \infty$ gives the asymptotic coverages and lengths. The MLR model with $E(Y_i) = 1 + x_{i2} + \cdots + x_{i8}$ was used. The vectors $(x_2, ..., x_8)^T$ were iid $N_7(\mathbf{0}, \boldsymbol{I}_7)$. The error distributions were N(0,1), $t_3$, and exponential(1) $-1$. Also, a small sensitivity study to examine the effects of changing $(1 + 15/n)$ to $(1 + k/n)$ on the 99% PIs (5.23) and (5.26) was performed. For $n = 50$ and $k$ between 10 and 20, the coverage increased by roughly 0.001 as $k$ increased by 1.

**Table 5.1** N(0,1) Errors

| $\delta$ | n | clen | slen | alen | olen | ccov | scov | acov | ocov |
|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 50 | 5.860 | 6.172 | 5.191 | 6.448 | .989 | .988 | .972 | .990 |
| 0.01 | 100 | 5.470 | 5.625 | 5.257 | 5.412 | .990 | .988 | .985 | .985 |
| 0.01 | 1000 | 5.182 | 5.181 | 5.263 | 5.097 | .992 | .993 | .994 | .992 |
| 0.01 | $\infty$ | 5.152 | 5.152 | 5.152 | 5.152 | .990 | .990 | .990 | .990 |
| 0.05 | 50 | 4.379 | 5.167 | 4.290 | 5.111 | .948 | .974 | .940 | .968 |
| 0.05 | 100 | 4.136 | 4.531 | 4.172 | 4.359 | .956 | .970 | .956 | .958 |
| 0.05 | 1000 | 3.938 | 3.977 | 4.001 | 3.927 | .952 | .952 | .954 | .948 |
| 0.05 | $\infty$ | 3.920 | 3.920 | 3.920 | 3.920 | .950 | .950 | .950 | .950 |
| 0.1 | 50 | 3.642 | 4.445 | 3.658 | 4.193 | .894 | .945 | .895 | .929 |
| 0.1 | 100 | 3.455 | 3.841 | 3.519 | 3.690 | .900 | .930 | .905 | .913 |
| 0.1 | 1000 | 3.304 | 3.343 | 3.352 | 3.304 | .901 | .903 | .907 | .901 |
| 0.1 | $\infty$ | 3.290 | 3.290 | 3.290 | 3.290 | .900 | .900 | .900 | .900 |

The simulation compared coverages and lengths of the classical (5.20), semiparametric (5.23), asymptotically conservative (5.24) and asymptotically optimal (5.26) PIs. The latter 3 intervals are asymptotically optimal for symmetric unimodal error distributions in that they have the shortest asymptotic length that gives the desired asymptotic coverage. The semiparametric PI

**Table 5.2** $t_3$ Errors

| $\delta$ | n | clen | slen | alen | olen | ccov | scov | acov | ocov |
|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 50 | 9.539 | 12.164 | 11.398 | 13.297 | .972 | .978 | .975 | .981 |
| 0.01 | 100 | 9.114 | 12.202 | 12.747 | 10.621 | .978 | .983 | .985 | .978 |
| 0.01 | 1000 | 8.840 | 11.614 | 12.411 | 11.142 | .975 | .990 | .992 | .988 |
| 0.01 | $\infty$ | 8.924 | 11.681 | 11.681 | 11.681 | .979 | .990 | .990 | .990 |
| 0.05 | 50 | 7.160 | 8.313 | 7.210 | 8.139 | .945 | .956 | .943 | .956 |
| 0.05 | 100 | 6.874 | 7.326 | 7.030 | 6.834 | .950 | .955 | .951 | .945 |
| 0.05 | 1000 | 6.732 | 6.452 | 6.599 | 6.317 | .951 | .947 | .950 | .945 |
| 0.05 | $\infty$ | 6.790 | 6.365 | 6.365 | 6.365 | .957 | .950 | .950 | .950 |
| 0.1 | 50 | 5.978 | 6.591 | 5.532 | 6.098 | .915 | .935 | .900 | .917 |
| 0.1 | 100 | 5.696 | 5.756 | 5.223 | 5.274 | .916 | .913 | .901 | .900 |
| 0.1 | 1000 | 5.648 | 4.784 | 4.842 | 4.706 | .929 | .901 | .904 | .898 |
| 0.1 | $\infty$ | 5.698 | 4.707 | 4.707 | 4.707 | .935 | .900 | .900 | .900 |

**Table 5.3** Exponential(1) $-1$ Errors

| $\delta$ | n | clen | slen | alen | olen | ccov | scov | acov | ocov |
|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 50 | 5.795 | 6.432 | 6.821 | 6.817 | .971 | .987 | .976 | .988 |
| 0.01 | 100 | 5.427 | 5.907 | 7.525 | 5.377 | .974 | .987 | .986 | .985 |
| 0.01 | 1000 | 5.182 | 5.387 | 8.432 | 4.807 | .972 | .987 | .992 | .987 |
| 0.01 | $\infty$ | 5.152 | 5.293 | 8.597 | 4.605 | .972 | .990 | .995 | .990 |
| 0.05 | 50 | 4.310 | 5.047 | 5.036 | 4.746 | .946 | .971 | .955 | .964 |
| 0.05 | 100 | 4.100 | 4.381 | 5.189 | 3.840 | .947 | .971 | .966 | .955 |
| 0.05 | 1000 | 3.932 | 3.745 | 5.354 | 3.175 | .945 | .954 | .972 | .947 |
| 0.05 | $\infty$ | 3.920 | 3.664 | 5.378 | 2.996 | .948 | .950 | .975 | .950 |
| 0.1 | 50 | 3.601 | 4.183 | 3.960 | 3.629 | .920 | .945 | .925 | .916 |
| 0.1 | 100 | 3.429 | 3.557 | 3.959 | 3.047 | .930 | .943 | .945 | .913 |
| 0.1 | 1000 | 3.303 | 3.005 | 3.989 | 2.460 | .931 | .906 | .951 | .901 |
| 0.1 | $\infty$ | 3.290 | 2.944 | 3.991 | 2.303 | .929 | .900 | .950 | .900 |

gives the correct asymptotic coverage if the unimodal errors are not symmetric while the PI (5.24) gives higher coverage (is conservative). The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where $Y_f$ fell within the nominal $100(1-\delta)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1-\gamma_n)$ distribution where $1-\gamma_n$ converges to the asymptotic coverage $(1-\gamma)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0014$, 0.0031 and 0.0042 for $p = 0.01, 0.05$ and 0.1, respectively. Hence an observed coverage $\hat{p} \in [0.986, 0.994]$ for 99%, $\hat{p} \in [0.941, 0.959]$ for 95% and $\hat{p} \in [0.887, 0.913]$ for 90% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Tables 5.1–5.3 show the results of the simulations for the 3 error distributions. The letters *c, s, a* and *o* refer to intervals (5.20), (5.23), (5.24) and (5.26) respectively. For the normal errors, the coverages were about right and the semiparametric interval tended to be rather long for $n = 50$ and 100. The classical PI asymptotic coverage $1-\gamma$ tended to be fairly close to the nominal coverage $1-\delta$ for all 3 distributions and $\delta = 0.01, 0.05$, and 0.1.

## 5.5 Numerical Diagnostics

*Using one or a few numerical summaries to characterize the relationship between x and y runs the risk of missing important features, or worse, of being misled.*
Chambers, Cleveland, Kleiner, and Tukey (1983, p. 76)

*Diagnostics* are used to check whether model assumptions are reasonable. Section 5.6 provides graphical diagnostics for assessing the unimodal MLR model adequacy while this section focuses on diagnostics for the unimodal MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the errors are iid from a unimodal distribution that is not highly skewed with $\mathrm{E}(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$. See Definition 5.13.

It is often useful to use notation to separate the constant from the nontrivial predictors. Assume that $\boldsymbol{x}_i = (1, x_{i,2}, ..., x_{i,p})^T \equiv (1, \boldsymbol{u}_i^T)^T$ where the $(p-1)\times 1$ vector of nontrivial predictors $\boldsymbol{u}_i = (x_{i,2}, ..., x_{i,p})^T$. In matrix form, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, $\boldsymbol{X} = [X_1, X_2, ..., X_p] = [\boldsymbol{1}, \boldsymbol{U}]$, $\boldsymbol{1}$ is an $n \times 1$ vector of ones, and $\boldsymbol{U} = [X_2, ..., X_p]$ is the $n \times (p-1)$ matrix of nontrivial predictors. The $k$th column of $\boldsymbol{U}$ is the $n \times 1$ vector of the $j$th predictor $X_j = (x_{1,j}, ..., x_{n,j})^T$ where $j = k + 1$. The sample mean and covariance matrix of the nontrivial predictors are

$$\overline{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}_i \tag{5.27}$$

and

$$\boldsymbol{C} = \mathrm{Cov}(\boldsymbol{U}) = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{5.28}$$

respectively.

Some important numerical quantities that are used as diagnostics measure the distance of $\boldsymbol{u}_i$ from $\overline{\boldsymbol{u}}$ and the *influence* of case $i$ on the OLS fit $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{OLS}$. The $i$th *residual* $r_i = Y_i - \widehat{Y}_i$, and the vector of fitted values is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H}$ is the *hat matrix*. *Case* (or *leave one out* or *deletion*) diagnostics are computed by omitting the $i$th case from the OLS regression. Let

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{5.29}$$

denote the $n \times 1$ vector of fitted values from estimating $\boldsymbol{\beta}$ with OLS without the $i$th case. Denote the $j$th element of $\widehat{\boldsymbol{Y}}_{(i)}$ by $\widehat{Y}_{(i),j}$. It can be shown that the variance of the $i$th residual $\mathrm{VAR}(r_i) = \sigma^2(1 - h_i)$. The usual estimator of the error variance is $\widehat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} r_i^2}{n-p}$. The (internally) *studentized residual* $\widehat{e}_i = \dfrac{r_i}{\widehat{\sigma}\sqrt{1-h_i}}$ has zero mean and approximately unit variance.

**Definition 5.25.** The $i$th *leverage* $h_i = \boldsymbol{H}_{ii}$ is the $i$th diagonal element of the hat matrix $\boldsymbol{H}$. The $i$th *squared (classical) Mahalanobis distance* $\mathrm{MD}_i^2 = (\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \boldsymbol{C}^{-1} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})$. The $i$th *Cook's distance*

$$\mathrm{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{5.30}$$

$$= \frac{1}{p\widehat{\sigma}^2} \sum_{j=1}^{n} (\widehat{Y}_{(i),j} - \widehat{Y}_j)^2.$$

**Theorem 5.10.** a) (Rousseeuw and Leroy 1987, p. 225)

$$h_i = \frac{1}{n-1} \mathrm{MD}_i^2 + \frac{1}{n}.$$

b) (Cook and Weisberg 1999a, p. 184)

$$h_i = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T (\boldsymbol{U}^T \boldsymbol{U})^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) + \frac{1}{n}.$$

c) (Cook and Weisberg 1999a, p. 360)

$$\mathrm{CD}_i = \frac{r_i^2}{p\widehat{\sigma}^2 (1 - h_i)} \frac{h_i}{1 - h_i} = \frac{\widehat{e}_i^2}{p} \frac{h_i}{1 - h_i}.$$

When the statistics $\mathrm{CD}_i$, $h_i$ and $\mathrm{MD}_i$ are large, case $i$ may be an outlier or *influential* case. Examining a dot plot of these three statistics for unusually large values can be useful for flagging influential cases. Cook and Weisberg (1999a, p. 358) suggest examining cases with $\mathrm{CD}_i > 0.5$ and that cases with $\mathrm{CD}_i > 1$ should always be studied. Since $\boldsymbol{H} = \boldsymbol{H}^T$ and $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$, the hat matrix is symmetric and idempotent. Hence the eigenvalues of $\boldsymbol{H}$ are zero or one and $\mathrm{trace}(\boldsymbol{H}) = \sum_{i=1}^{n} h_i = p$. Rousseeuw and Leroy (1987, p. 220 and p. 224) suggest using $h_i > 2p/n$ and $\mathrm{MD}_i^2 > \chi^2_{p-1,0.95}$ as benchmarks for leverages and Mahalanobis distances where $\chi^2_{p-1,0.95}$ is the 95th percentile of a chi–square distribution with $p-1$ degrees of freedom.

Note that Theorem 5.10c) implies that Cook's distance is the product of the squared residual and a quantity that becomes larger the farther $\boldsymbol{u}_i$ is from $\overline{\boldsymbol{u}}$. Hence influence is roughly the product of leverage and distance of $Y_i$ from $\widehat{Y}_i$ (see Fox 1991, p. 21). Mahalanobis distances and leverages both define ellipsoids based on a metric closely related to the sample covariance matrix of the nontrivial predictors. All points $\boldsymbol{u}_i$ on the same ellipsoidal contour are the same distance from $\overline{\boldsymbol{u}}$ and have the same leverage (or the same Mahalanobis distance).

Cook's distances, leverages, and Mahalanobis distances can be effective for finding influential cases when there is a single outlier, but can fail if there

are two or more outliers. Nevertheless, these numerical diagnostics combined with response and residual plots of the next section are probably the *most effective techniques* for detecting cases that effect the fitted values when the unimodal MLR model is a good approximation for the bulk of the data.

## 5.6 Graphical Diagnostics

*Automatic or blind use of regression models, especially in exploratory work, all too often leads to incorrect or meaningless results and to confusion rather than insight. At the very least, a user should be prepared to make and study a number of plots before, during, and after fitting the model.*
Chambers, Cleveland, Kleiner, and Tukey (1983, p. 306)

A scatterplot of $x$ versus $y$ (recall the convention that a plot of $x$ versus $y$ means that $x$ is on the horizontal axis and $y$ is on the vertical axis) is used to *visualize the conditional distribution* $y|x$ of $y$ given $x$ (see Cook and Weisberg 1999a, p. 31). For the simple linear regression model (with one nontrivial predictor $x_2$), an *effective* technique for checking the assumptions of the model is to make a scatterplot of $x_2$ versus $Y$ and a residual plot of $x_2$ versus $r_i$. Departures from linearity in the scatterplot suggest that the simple linear regression model is not adequate. The points in the residual plot should scatter about the line $r = 0$ with no pattern. If curvature is present or if the distribution of the residuals depends on the value of $x_2$, then the simple linear regression model is not adequate. The following two plots are **crucial for any multiple linear regression analysis,** regardless of the regression estimator (e.g. OLS, $L_1$, lasso, etc.).

**Definition 5.26.** A *residual plot* is a plot of a variable $w_i$ versus the residuals $r_i$. Typically $w_i$ is a linear combination of the predictors: $w_i = \boldsymbol{a}^T \boldsymbol{x}_i$ where $\boldsymbol{a}$ is a known $p \times 1$ vector. A *response plot* is a plot of the fitted values $\hat{Y}_i$ versus the response $Y_i$.

The most used residual plot takes $\boldsymbol{a} = \widehat{\boldsymbol{\beta}}$ with $w_i = \hat{Y}_i$. Plots against the individual predictors $x_j$ and potential predictors are also used. If the residual plot is not ellipsoidal with zero slope, then the *unimodal MLR model* (where the iid constant variance errors are from a unimodal distribution that is not highly skewed) *is not sustained*. In other words, if the variables in the residual plot show some type of dependency, e.g. increasing variance or a curved pattern, then the unimodal MLR model may be inadequate. Theorem 5.1 showed that the response plot simultaneously displays the fitted values, response, and residuals. The plotted points in the response plot should scatter about the identity line if the unimodal MLR model holds. Note that residual plots *magnify departures* from the model while the response plot emphasizes

*how well the model fits the data.* Cook and Weisberg (1997, 1999a ch. 17) call a plot that emphasizes model agreement a *model checking plot.*

One of the themes of this text is to use a several estimators to create plots and estimators. Many estimators $\boldsymbol{b}_j$ are consistent estimators of $\boldsymbol{\beta}$ when the multiple linear regression model holds.

**Definition 5.27.** Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_J$ be $J$ estimators of $\boldsymbol{\beta}$. Assume that $J \geq 2$ and that OLS is included. A *fit-fit* (FF) plot is a scatterplot matrix of the fitted values $\widehat{Y}(\boldsymbol{b}_1), ..., \widehat{Y}(\boldsymbol{b}_J)$. Often $Y$ is also included in the top or bottom row of the FF plot to see the response plots. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$. Often $\hat{Y}$ is also included in the top or bottom row of the RR plot to see the residual plots.

If the multiple linear regression model holds, if the predictors are bounded, and if all $J$ regression estimators are consistent estimators of $\boldsymbol{\beta}$, then the subplots in the FF and RR plots should be linear with a correlation tending to one as the sample size $n$ increases. To prove this claim, let the $i$th residual from the $j$th fit $\boldsymbol{b}_j$ be $r_i(\boldsymbol{b}_j) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_j$ where $(Y_i, \boldsymbol{x}_i^T)$ is the $i$th observation. Similarly, let the $i$th fitted value from the $j$th fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

$$\|r_i(\boldsymbol{b}_1) - r_i(\boldsymbol{b}_2)\| = \|\widehat{Y}_i(\boldsymbol{b}_1) - \widehat{Y}_i(\boldsymbol{b}_2)\| = \|\boldsymbol{x}_i^T(\boldsymbol{b}_1 - \boldsymbol{b}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \ (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{5.31}$$

The FF plot is a powerful way for comparing fits. The commonly suggested alternative is to look at a table of the estimated coefficients, but coefficients can differ greatly while yielding similar fits if some of the predictors are highly correlated or if several of the predictors are independent of the response.

To illustrate the RR plot, consider the four $R$ estimators: OLS, ALMS = the default version of `lmsreg`, ALTS = the default version of `ltsreg` and the MBA estimator described in Chapter 6. In the 2007 version of $R$, the last three estimators change with each call.

**Example 5.10.** Gladstone (1905) records the brain weight and various head measurements for 276 individuals. This data set, along with the Buxton data set in the following example, can be downloaded from the text's website. We'll predict *brain weight* using six head measurements (head *height, length, breadth, size, cephalic index* and *circumference*) as predictors, deleting cases 188 and 239 because of missing values. There are five infants (cases 238, and 263-266) of age less than 7 months that are $\boldsymbol{x}$-outliers. Nine toddlers were between 7 months and 3.5 years of age, four of whom appear to be $\boldsymbol{x}$-outliers (cases 241, 243, 267, and 269). (The points are not labeled on the plot, but the five infants are easy to recognize.)

Figure 1.1 shows the RR plot. The five infants seem to be "good leverage points" in that the fit to the bulk of the data passes through the infants. Hence the OLS fit may be best, followed by ALMS. Note that ALTS and MBA make

the absolute residuals for the infants large. The ALTS and MBA fits are not highly correlated for the remaining 265 points, but the remaining correlations are high. Thus the fits agree on these cases, focusing attention on the infants. The ALTS and ALMS estimators change frequently, and are implemented differently in *R* and *Splus*. Often the "new and improved" implementation is much worse than older implementations.

Figure 1.2 shows the residual plots for the Gladstone data when one observation, 119, had *head length* entered incorrectly as 109 instead of 199. This outlier is easier to detect with MBA and ALTS than with ALMS.

**Example 5.11.** Buxton (1920, p. 232-5) gives 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! This appears to be a clerical error; these individuals' stature was recorded as head length and the integer 18 or 19 given for stature, making the cases massive outliers with enormous leverage. These absurdly bad observations turned out to confound the standard high breakdown (HB) estimators. Figure 6.4 shows the RR plot for several estimators. The BB, MBA and MBALATA estimators, described in Chapter 6, give large absolute residuals for the outliers. Problem 5.9 shows how to create RR and FF plots.

## 5.7 MLR Outlier Detection

*Do not attempt to build a model on a set of poor data! In human surveys, one often finds 14–inch men, 1000–pound women, students with "no" lungs, and so on. In manufacturing data, one can find 10,000 pounds of material in a 100 pound capacity barrel, and similar obvious errors. All the planning, and training in the world will not eliminate these sorts of problems. ... In our decades of experience with "messy data," we have yet to find a large data set completely free of such quality problems.*

Draper and Smith (1981, p. 418)

There is an enormous literature on outlier detection in multiple linear regression. Typically a numerical measure such as Cook's distance or a residual plot based on resistant fits is used. The following terms are frequently encountered.

**Definition 5.28.** Suppose that some analysis to detect outliers is performed. *Masking* occurs if the analysis suggests that one or more outliers are in fact good cases. *Swamping* occurs if the analysis suggests that one or more good cases are outliers.

The following techniques are useful for detecting outliers when the multiple linear regression model is appropriate.
1) Find the OLS residuals and fitted values and make a response plot and a residual plot. Look for clusters of points that are separated from the bulk of the data and look for residuals that have large absolute values. Beginners frequently label too many points as outliers. Try to estimate the standard deviation of the residuals in both plots. In the residual plot, look for residuals that are more than 5 standard deviations away from the $r = 0$ line.
2) Make an RR plot. See Figures 1.1 and 6.4.
3) Make an FF plot. See Figure 6.3 and Problem 5.9.
4) Display the residual plots from several different estimators. See Figure 1.2.
5) Display the response plots from several different estimators. This can be done by adding $Y$ to the FF plot.
6) Make a DD plot of the continuous predictors.
7) Make a scatterplot matrix of several diagnostics such as leverages, Cook's distances and studentized residuals.

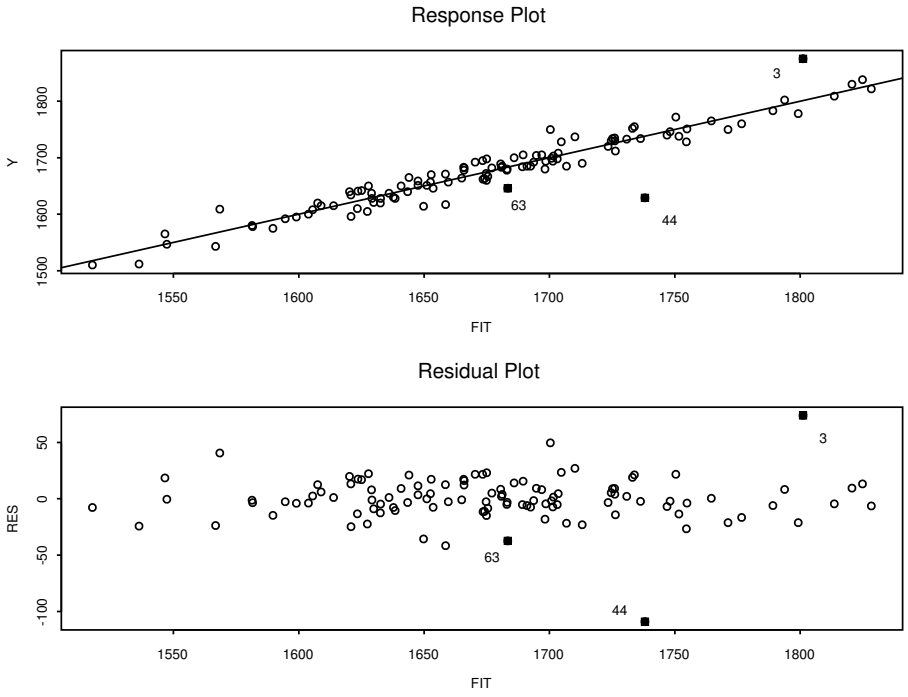


**Fig. 5.8** Residual and Response Plots for the Tremearne Data

**Example 5.12.** Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) because of missing values and used *height* as the response variable $Y$. The five predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 5.8 presents the OLS residual and response plots for this data set. Points corresponding to cases with Cook's distance $> \min(0.5, 2p/n)$ are shown as highlighted squares (cases 3, 44 and 63). The 3rd person was very tall while the 44th person was rather short. From the plots, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining, but are not necessarily outliers. Two other cases have residuals near fifty. The plots can be made with the following commands.

```
source("G:/rpack.txt")
#assume the data is stored in R matrix major
X<-major[,-6]; Y <- major[,6]; MLRplot(X,Y)
```

Data sets like this one are very common. The majority of the cases seem to follow a multiple linear regression model with iid Gaussian errors, but a small percentage of cases seem to come from an error distribution with heavier tails than a Gaussian distribution.

Detecting outliers is much easier than deciding what to do with them. After detection, the investigator should see whether the outliers are recording errors. The outliers may become good cases after they are corrected. But frequently there is no simple explanation for why the cases are outlying. Typical advice is that *outlying cases should never be blindly deleted* and that the investigator should *analyze the full data set including the outliers as well as the data set after the outliers have been removed* (either by deleting the cases or the variables that contain the outliers).

Typically two methods are used to find the cases (or variables) to delete. The investigator computes OLS diagnostics and subjectively deletes cases, or a resistant multiple linear regression estimator is used that automatically gives certain cases zero weight.

Suppose that the data has been examined, recording errors corrected, and impossible cases deleted. For example, in the Buxton (1920) data, 5 people with heights of 0.75 inches were recorded. For this data set, these heights could be corrected. If they could not be corrected, then these cases should be discarded since they are impossible. If outliers are present even after correcting recording errors and discarding impossible cases, then we can add two additional rough guidelines.

First, if the *purpose is to display the relationship between the predictors and the response*, make a response plot using the full data set (computing the fitted values by giving the outliers weight zero) and using the data set with the outliers removed. Both plots are needed if the relationship that holds for the bulk of the data is obscured by outliers. The outliers are removed from

the data set in order to get reliable estimates for the bulk of the data. The identity line should be added as a visual aid and the proportion of outliers should be given. Secondly, if the *purpose is to predict a future value of the response variable*, then a procedure such as that described in Example 1.5 may be useful. The prediction interval based on the shorth given by Equation (5.26) may also be useful.

For multiple linear regression, the OLS response and residual plots are very useful for detecting outliers. The DD plot of the continuous predictors is also useful. Use the *rpack* functions `MLRplot` and `ddplot4`. Response and residual plots from outlier resistant methods are also useful. See Chapter 6.

Huber and Ronchetti (2009, p. 154) noted that efficient methods for identifying leverage groups are needed. Such groups are often difficult to detect with regression diagnostics and residuals, but often have outlying fitted values and responses that can be detected with response and residual plots. The following *rules of thumb* are useful for finding influential cases and outliers. The trimmed views estimator of Section 6.1 is also useful. Dragging the plots, so that they are roughly square, can be useful.

When the bulk of the data follows the unimodal MLR model of Definition 5.13, the following *rules of thumb* are useful for finding influential cases and outliers. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit an MLR estimator such as OLS to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus $Y$ using the entire data set. If the identity line passes through the cluster, then the cases in the cluster may be good leverage points, otherwise they may be outliers.

To see why gaps are important, suppose that OLS was used to obtain $\hat{Y} = \hat{m}$. If the model contains a constant, then the squared correlation $(\text{corr}(Y, \hat{Y}))^2$ is equal to the coefficient of determination $R^2$. Even if an alternative MLR estimator is used, $R^2$ over emphasizes the strength of the MLR relationship when there are two clusters of data since much of the variability of $Y$ is due to the smaller cluster.

Assume that OLS is used to fit the model and to make the response plot $\hat{Y}$ versus $Y$. Then the $i$th Cook's distance $CD_i$ tends to be large if $\hat{Y}$ is far from the sample mean $\overline{Y}$ and if the corresponding absolute residual $|r_i|$ is not small. If $\hat{Y}$ is close to $\overline{Y}$ then $CD_i$ tends to be small unless $|r_i|$ is large. An exception to these rules of thumb occurs if a group of cases form a cluster and the OLS fit passes through the cluster. Then the $CD_i$'s corresponding to these cases tend to be small even if the cluster is far from $\overline{Y}$.

Influence diagnostics such as Cook's distances $CD_i$ from Cook (1977) and the weighted Cook's distances $WCD_i$ from Peña (2005) are sometimes useful. Although an index plot of Cook's distance $CD_i$ may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \text{median}(\text{WCD}_\text{i})| > 4.5\text{MAD}(\text{WCD}_\text{i})$ are highlighted with crosses, where the median absolute deviation $\text{MAD}(w_i) = \text{median}(|w_i - \text{median}(\text{w}_\text{i})|)$.

**Example 5.11 (continued):** Figure 5.9 shows the response plot and residual plot for the Buxton data. Notice that the OLS fit passes through the outliers, but the response plot is resistant to $Y$–outliers since $Y$ is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$, only two of the outliers had large Cook's distance and only one case had a large $WCD_i$. Hence *masking* occurred for the Cook's distances, the $WCD_i$ and for the OLS residuals, but not for the OLS fitted values. Figure 6.1 shows that plots using `lmsreg` and `ltsreg` were similar, but MBA was effective. Figure 5.9 was made with the following R commands.

```
source("G:/rpack.txt"); source("G:/robdata.txt")
mlrplot4(buxx,buxy) #right click Stop twice
```

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. (Using the trimmed views of Section 6.1 is also effective for detecting outliers and other departures from the MLR model.)

**Example 5.13.** Hawkins et al. (1984) present a well known artificial data set where the first 10 cases are outliers while cases 11-14 are good leverage points. Figure 5.10 shows the residual and response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the response plot. Since the good cases 11-14 have the largest Cook's distances and absolute OLS residuals, *swamping* has occurred. (Masking has also occurred since the outliers have small Cook's distances, and some of the outliers have smaller OLS residuals than clean cases.) To determine whether both clusters are outliers or if one cluster consists of good leverage points, cases in both clusters could be given weight zero and the resulting response plot created. (Alternatively, response plots based on the `tvreg` estimator of Section 6.1 could be made where the cases with weight one are highlighted. For high levels of trimming, the identity line often passes through the good leverage points.)

The above example is typical of many "benchmark" outlier data sets for MLR. In these data sets traditional OLS diagnostics such as Cook's distance and the residuals often fail to detect the outliers, but the combination of the

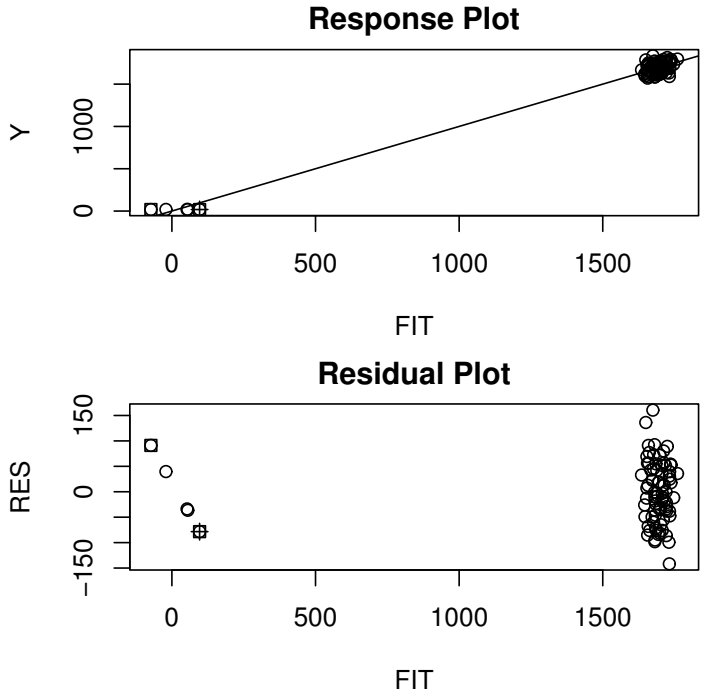


**Fig. 5.9** Plots for Buxton Data

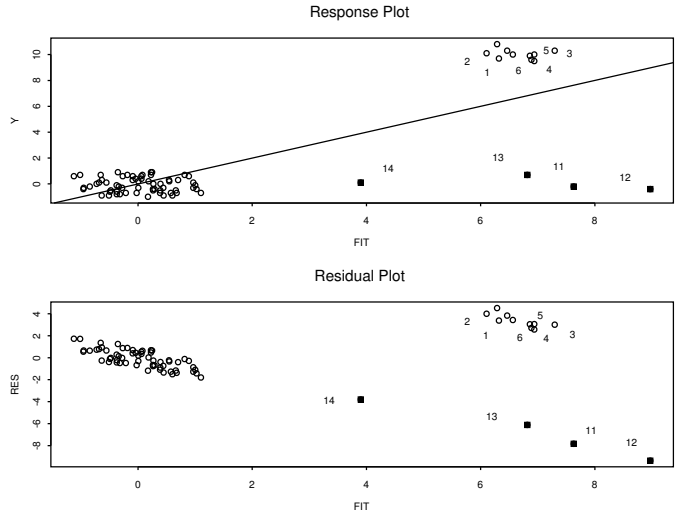


**Fig. 5.10** Plots for HBK Data

response plot and residual plot is usually able to detect the outliers. The $CD_i$ and $WCD_i$ are the most effective when there is a single cluster about the identity line as in Example 5.12. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail.

**Example 5.14.** Wood (1973) provides data where the octane number is predicted from 3 feed compositions and the log of a combination of process conditions. The OLS response and residual plots in Figure 5.11 suggest that the model is linear but the constant variance assumption may not be reasonable. There appear to be three groups of data. For this data, none of the cases had large $CD_i$ or $WCD_i$. Tremendous profit can be gained by raising the octane number by one point, and the two cases with the largest fitted values $\hat{Y} \approx 97$ were of the greatest interest.

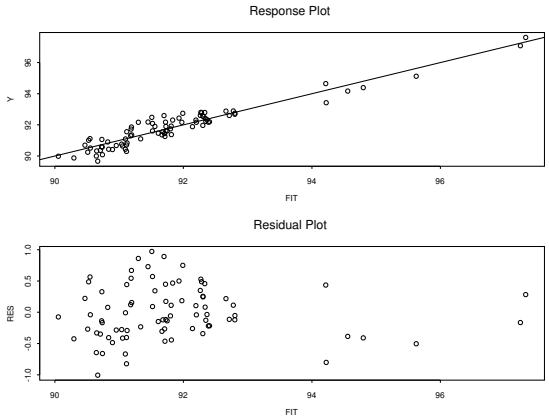


**Fig. 5.11** Octane Data

## 5.8 MLR Breakdown and Equivariance

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data, and are discussed below. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{5.32}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{5.33}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{5.34}$$

If the design matrix $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the vector of dependent variables $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set.

**Definition 5.29. Regression Equivariance:** Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \tag{5.35}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{r}(\boldsymbol{W}, \boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that the residuals are invariant under this type of transformation, and note that if $\boldsymbol{u} = -\widehat{\boldsymbol{\beta}}$, then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$. Also see Problem 5.6.

**Definition 5.30. Scale Equivariance:** Let $c$ be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{5.36}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = c\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Scale equivariance implies that if the $Y_i$'s are stretched, then the fits and the residuals should be stretched by the same factor.

**Definition 5.31. Affine Equivariance:** Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{5.37}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}}$, and $\boldsymbol{r}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that both the predicted values and the residuals are invariant under an affine transformation of the predictor variables.

**Definition 5.32. Permutation Invariance:** Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript $T$ denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{5.38}$$

Hence if $\boldsymbol{W} = \boldsymbol{PX}$ and $\boldsymbol{Z} = \boldsymbol{PY}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{PX}, \boldsymbol{PY}) = \boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

**Remark 5.3.** OLS has the above invariance properties, but most Statistical Learning alternatives such as lasso and ridge regression do not have all four properties. Hence Remark 7.11 is used to fit the data with $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Then obtain $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$.

The remainder of this section gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the $i$th row corresponds to the $i$th case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with $i$th row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any $d_n$ of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma \equiv \gamma_n = d_n/n$, and the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of $\gamma_n$ needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

**Definition 5.33.** Let $1 \leq d_n \leq n$. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the *breakdown value* of $T$ is

$$B(T, \boldsymbol{W}) = \min\left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$.

**Definition 5.34.** *High breakdown* regression estimators have $\gamma_n \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in *general position*: any $p$ clean cases give a unique estimate of $\boldsymbol{\beta}$. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED$(|r_i|)$ can be made arbitrarily large. The result implies that if the breakdown value $\leq 0.5$, breakdown can be computed using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$. Similarly $\hat{\boldsymbol{\beta}}$ is high breakdown if the median squared residual or the $c_n$th largest absolute residual $|r_i|_{(c_n)}$ or squared residual $r^2_{(c_n)}$ stay bounded under high contamination where $c_n \approx n/2$. Note that $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \leq M$ for some constant $M$ that depends on $T$ and $\boldsymbol{W}$ but not on the outliers if the number of outliers $d_n$ is less than the smallest number of outliers needed to cause breakdown.

**Theorem 5.11.** If the breakdown value $\leq 0.5$, computing the breakdown value using the median absolute residual $\text{MED}(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 5.33.

**Proof.** Consider any contaminated data set $\boldsymbol{W}_d^n$ with $i$th row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \leq M$ for some constant $M$ if $d < d_n$, then the median absolute residual $\text{MED}(|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|)$ is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d_n < n/2$.

If the median absolute residual is bounded by $M$ when $d < d_n$, then $\|\hat{\boldsymbol{\beta}}\|$ is bounded provided fewer than half of the cases line on the hyperplane (and so have absolute residual of 0), as shown next. Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the $i$th case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence $\text{MED}(|r_i|) = \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d_n/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where $q$ is the number of "good" cases that lie on a hyperplane of lower dimension than $p$. In the literature it is usually assumed that the original data are in *general position*: $q = p - 1$. □

Suppose that the clean data are in general position and that the number of outliers is less than the number needed to make the median absolute residual and $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down by adding a large positive or negative constant to the $Y$ values of the outliers, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and $\text{MED}(|r_i|)$ stay bounded where the bounds depend on the clean data $\boldsymbol{W}$ but not on the outliers even if the number of outliers is nearly as large as $n/2$. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the $Y_i$'s are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not $\infty$. If both $\boldsymbol{x}$ and $Y$ can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but may still have small residuals. For example, move the outliers along the regression hyperplane formed by the clean cases.

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p - 1$ "clean" cases and $d_n$ "contaminated" cases. Hence $d_n + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if $T$ possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then $T$ is neither regression nor affine equivariant. The breakdown value of $T$ is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the $Y_i$'s are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to $\infty$. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to $\infty$ if they can be constructed so that the estimator is no longer defined, e.g. so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The examples following some results on norms may help illustrate these points.

**Definition 5.35.** Let $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \mathbb{R}^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \ \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and for all scalars $a$, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$.

**Definition 5.36.** Let $\boldsymbol{G}$ be an $n \times p$ matrix. Then $\|\boldsymbol{G}\|$ is a *matrix norm* if
mn1) $\|\boldsymbol{G}\| \geq 0$ for every $n \times p$ matrix $\boldsymbol{G}$ with equality iff $\boldsymbol{G}$ is the zero matrix,
mn2) $\|a\boldsymbol{G}\| = |a| \ \|\boldsymbol{G}\|$ for all scalars $a$, and
mn3) $\|\boldsymbol{G} + \boldsymbol{H}\| \leq \|\boldsymbol{G}\| + \|\boldsymbol{H}\|$ for all $n \times p$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$.

**Example 5.15.** The *q-norm* of a vector $\boldsymbol{y}$ is $\|\boldsymbol{y}\|_q = (|y_1|^q + \cdots + |y_n|^q)^{1/q}$. In particular, $\|\boldsymbol{y}\|_1 = |y_1| + \cdots + |y_n|$, the *Euclidean norm* $\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$, and $\|\boldsymbol{y}\|_\infty = \max_i \ |y_i|$. Given a matrix $\boldsymbol{G}$ and a vector norm $\|\boldsymbol{y}\|_q$ the *q-norm* or *subordinate matrix norm* of matrix $\boldsymbol{G}$ is $\|\boldsymbol{G}\|_q = \max_{\boldsymbol{y} \neq \boldsymbol{0}} \dfrac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}$. It can be shown that the *maximum column sum norm* $\|\boldsymbol{G}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} |g_{ij}|$, the *maximum row sum norm* $\|\boldsymbol{G}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{p} |g_{ij}|$, and the *spectral norm* $\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}$. The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})}.$$

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm, $\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \ \|\boldsymbol{y}\|_q$. Let $J = J_m = \{m_1, ..., m_p\}$ denote the $p$ cases in the $m$th elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{e}_J\|. \quad (5.39)$$

The following results (Golub and Van Loan 1989, pp. 57, 80) on the Euclidean norm are useful. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J = (x_{mi,j})$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (5.40)$$

$$\max_{i,j} |x_{mi,j}| \leq \|\boldsymbol{X}_J\| \leq p \ \max_{i,j} |x_{mi,j}|, \text{ and} \quad (5.41)$$

$$\frac{1}{p \ \max_{i,j} |x_{mi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \tag{5.42}$$

*From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.*

**Example 5.16.** Suppose the response values $Y$ are near 0. Consider the fit from an elemental set: $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ and examine Equations (5.40), (5.41), and (5.42). Now $\|\boldsymbol{b}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{Y}_J\|$, and *since $x$-outliers make $\|\boldsymbol{X}_J\|$ large, $x$-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ towards zero not towards $\infty$.* The $x$-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

**Example 5.17.** Without loss of generality, assume that the clean $Y$'s are contained in an interval $[a, f]$ for some $a$ and $f$. Assume that the regression model contains an intercept $\beta_1$. Then there exists an estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}}_M\| \leq \max(|a|, |f|)$ if $d_n < n/2$.

**Proof.** Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\hat{\boldsymbol{\beta}}_M\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\hat{\boldsymbol{\beta}}_M$ is equal to the median absolute deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, i = 1, ..., n) \leq f - a$ if $d_n < \lfloor (n+1)/2 \rfloor$. □

Note that $\hat{\boldsymbol{\beta}}_M$ is a poor high breakdown estimator of $\boldsymbol{\beta}$ and $\hat{Y}_i(\hat{\boldsymbol{\beta}}_M)$ tracks the $Y_i$ very poorly. If the data are in general position, a high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, pp. 29, 206) conjectured that high breakdown regression estimators can not be computed cheaply, and that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following theorem shows that these two conjectures are false.

**Theorem 5.12.** If the clean data are in general position and the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ can be found by computing OLS on the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w \ \text{MAD}(Y_1, ..., Y_n)]$ where $w \geq 1$ (so at least half of the cases are used).

**Proof.** Note that $\hat{\boldsymbol{\beta}}_w$ is obtained by computing OLS on the set $J$ of the $n_j$ cases which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm w\text{MAD}(n)]$$

where $w \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$ which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual $r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (w \text{ MAD}(n))^2$ if the $i$th case is in $J$. Hence the weighted LS fit $\hat{\boldsymbol{\beta}}_w$ is the OLS fit to the cases in $J$ and has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_w) \leq n_j (w \text{ MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_w)|, ..., |r_n(\hat{\boldsymbol{\beta}}_w)|) \leq \sqrt{n_j}\ w \text{ MAD}(n) < \sqrt{n}\ w \text{ MAD}(n) < \infty.$$

Thus the estimator $\hat{\boldsymbol{\beta}}_w$ has a median absolute residual bounded by $\sqrt{n}\ w \text{ MAD}(Y_1, ..., Y_n)$. Hence $\hat{\boldsymbol{\beta}}_w$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for constant $c = 0$, $\hat{\boldsymbol{\beta}}_w = \mathbf{0}$, and for $c \neq 0$ the set of cases used remains the same under scale transformations and OLS is scale equivariant. □

Note that if $w$ is huge and $\text{MAD}(n) \neq 0$, then the high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust. Even if $w = 1$, the HB estimator $\hat{\boldsymbol{\beta}}_w$ only resists large $Y$ outliers.

## 5.9 MLR Concentration Algorithms

Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trial fit to be used in the resistant estimator.

**Definition 5.37.** Suppose $c = c_n \approx n/2$. The LMS($c$) criterion is

$$Q_{LMS}(\boldsymbol{b}) = r_{(c)}^2(\boldsymbol{b}) \tag{5.43}$$

where $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$ are the ordered squared residuals, and the LTS($c$) criterion is

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r_{(i)}^2(\boldsymbol{b}). \tag{5.44}$$

The LTA($c$) criterion is

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c} |r(\boldsymbol{b})|_{(i)} \tag{5.45}$$

where $|r(\boldsymbol{b})|_{(i)}$ is the $i$th ordered absolute residual.

Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. Also see Hawkins and Olive (1999ab). These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion. LMS, LTA, and LTS have $O(n^p)$ or $O(n^{p+1})$ complexity. See Bernholt (2005), Hawkins and Olive (1999b), Klouda (2015), and Mount et al. (2014). Estimators with $O(n^4)$ or higher complexity take too long to compute. LTS and LTA are $\sqrt{n}$ consistent while LMS has the lower $n^{1/3}$ rate. See Kim and Pollard (1990), Čížek (2006, 2008), and Mašiček (2004). If $c = n$, the LTS and LTA criteria are the OLS and $L_1$ criteria. See Olive (2008, 2017b: ch. 14) for more on these estimators.

Concentration algorithms are widely used since impractical brand name estimators, such as LMS, LTA, and LTS, take too long to compute. The FLTS concentration algorithm, defined in Definition 5.40, use $K$ starts and attractors. The letter "F" is used since a fixed number of $K$ starts, such as $K = 500$, is used. A *start* is an initial estimator of $\boldsymbol{\beta}$, and an *attractor* is an estimator of $\boldsymbol{\beta}$ obtained by refining the start. For example, let the start be an estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$. Find the half set of $c_n$ cases with the smallest squared residuals $r_i^2$ where $r_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{b}$. Compute OLS on this set. This process could be iterated for $k$ concentration steps, producing an attractor.

**Definition 5.38.** For multiple linear regression, an *elemental set* is a set of $p$ cases.

Some notation is needed for algorithms that use many elemental sets. Let

$$J \equiv J_m = \{m_1, ..., m_p\}$$

denote the set of indices for the $m$th elemental set. Since there are $n$ cases, $m_1, ..., m_p$ are $p$ distinct integers between 1 and $n$. For example, if $n = 7$ and $p = 3$, the first elemental set may use cases $J_1 = \{1, 7, 4\}$, and the second elemental set may use cases $J_2 = \{5, 3, 6\}$. The data for the $m$th elemental set is $(\boldsymbol{Y}_{J_m}, \boldsymbol{X}_{J_m})$ where $\boldsymbol{Y}_{J_m} = (Y_{m1}, ..., Y_{mp})^T$ is a $p \times 1$ vector, and the $p \times p$ matrix

$$\boldsymbol{X}_{J_m} = \begin{bmatrix} \boldsymbol{x}_{m1}^T \\ \boldsymbol{x}_{m2}^T \\ \vdots \\ \boldsymbol{x}_{mp}^T \end{bmatrix} = \begin{bmatrix} x_{m1,1} & x_{m1,2} & \dots & x_{m1,p} \\ x_{m2,1} & x_{m2,2} & \dots & x_{m2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{mp,1} & x_{mp,2} & \dots & x_{mp,p} \end{bmatrix}.$$

Then the elemental fit is a hyperplane that passes through the $p$ cases of the elemental set. For $p = 2$, the hyperplane is a line.

**Definition 5.39.** The *elemental fit* from the $i$th elemental set $J_i$ is the OLS estimator $\hat{\boldsymbol{\beta}}_{J_i} = (\boldsymbol{X}_{J_i}^T\boldsymbol{X}_{J_i})^{-1}\boldsymbol{X}_{J_i}^T\boldsymbol{Y}_{J_i} = \boldsymbol{X}_{J_i}^{-1}\boldsymbol{Y}_{J_i}$ applied to the cases corresponding to the elemental set provided that the inverse of $\boldsymbol{X}_{J_i}$ exists.

**Definition 5.40.** A *start* is an initial trial fit and an *attractor* is the final fit generated by the algorithm from the start. Let $\boldsymbol{b}_{0,j}$ be the $j$th start and compute all $n$ residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. Let $\lfloor n/2 \rfloor \leq c_n \leq \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$. i) For an *FLTS concentration algorithm*, at the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for $k$ steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. The result of the iteration $\boldsymbol{b}_{k,j}$ is called the $j$th attractor where $j = 1, ..., K$. The final FLTS concentration algorithm estimator uses the attractor that minimizes the LTS criterion.

ii) For an *FLTA concentration algorithm*, at the next iteration, the $L_1$ estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest absolute residuals $|r_i(\boldsymbol{b}_{0,j})|$. This iteration can be continued for $k$ steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$ where $\boldsymbol{b}_{k,j}$ is the $j$th attractor and $j = 1, ..., K$. The final FLTA concentration algorithm estimator uses the attractor that minimizes the LTA criterion.

iii) The FLMS concentration algorithm uses the $L_\infty$ estimator and the LMS criterion.

Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$ concentration steps, i.e., the attractors are the starts.

**Definition 5.41.** The *elemental basic resampling algorithm* uses $K$ elemental starts that are equal to the attractors (hence $k = 0$). Compute the attractors $\boldsymbol{b}_{0,1}, ..., \boldsymbol{b}_{0,K}$, and the elemental basic resampling estimator uses the attractor that minimizes the (e.g. LMS, LTA, or LTS) criterion.

The elemental concentration and elemental resampling algorithms use $K$ elemental fits where $K$ is a fixed number that does not depend on the sample size $n$, e.g. $K = 500$. Note that an estimator can not be consistent for $\boldsymbol{\theta}$ unless the number of randomly selected cases goes to $\infty$, except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If $K = K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

**Theorem 5.13:** a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

**Proof:** a) Note that you can not get a consistent estimator by using $Kh$ randomly selected cases since the number of cases $Kh$ needs to go to $\infty$ for consistency except in degenerate situations.

b) Contaminating all $Kh$ cases in the $K$ elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. □

**Remark 5.4.** The number of randomly selected elemental sets needs to go to $\infty$ as $n \to \infty$ to get a consistent estimator. The $L_1$ estimator and

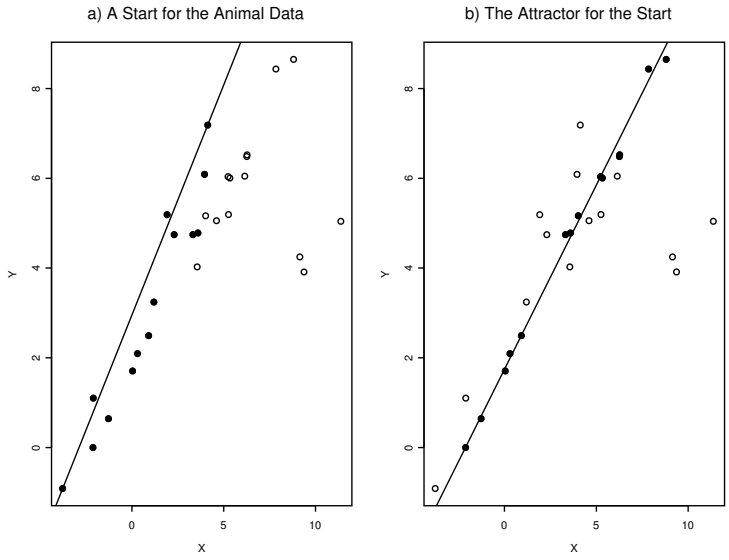


**Fig. 5.12** The Highlighted Points are More Concentrated about the Attractor

the sample median (when $n$ is odd) are consistent and both estimators are determined by an elemental set, but all $n$ cases are used to choose those elemental sets.

**Remark 5.5.** Theorem 5.13 shows that shows that the elemental basic resampling PROGRESS estimators of Rousseeuw (1984) and Rousseeuw and Leroy (1987) are zero breakdown and inconsistent. Yohai's two stage estimators, such as MM, need initial consistent high breakdown estimators such as LMS, but were implemented with the inconsistent zero breakdown elemental estimators such as `lmsreg`. See Hawkins and Olive (2002, p. 157). You can get consistent estimators if $K = K_n \to \infty$. If the concentration algorithm is iterated to convergence, it is not known whether the resulting estimator is consistent or not. The Hubert et al. (2008) claim that LTS can be computed efficiently by FLTS = Fast-LTS is false. See similar results below Theorem 3.15 for multivariate location and dispersion.

**Example 5.18.** As an illustration of the FLTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57). The response $Y$ is the *log brain weight* and the predictor $x$ is the *log body weight* for 25 mammals and 3 dinosaurs (outliers with the highest body weight). Suppose that the first elemental start uses cases 20 and 14, corresponding to mouse and man. Then the start $\boldsymbol{b}_{s,1} = \boldsymbol{b}_{0,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest absolute residuals $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{0,1}) = 12.101$. Figure 5.12a

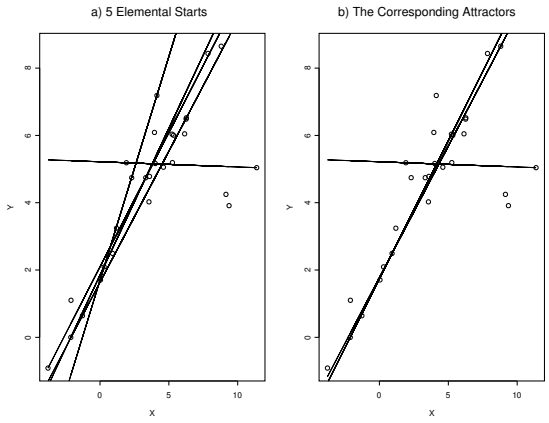


**Fig. 5.13** Starts and Attractors for the Animal Data

shows the scatterplot of $x$ and $y$. The start is also shown and the 14 cases corresponding to the smallest absolute residuals are highlighted. The $L_1$ fit to these $c$ highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990$. The iteration consists of finding the cases corresponding to the $c$ smallest absolute residuals, obtaining the corresponding $L_1$ fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA($c$) criterion evaluated at the attractor is $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172$. Figure 5.12b shows the attractor and that the $c$ highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 5.12a. Figure 5.13a shows 5 randomly selected starts while Figure 5.13b shows the corresponding attractors. Notice that the elemental starts have more variability than the attractors, but if the start passes through an outlier, so does the attractor.

**Remark 5.6.** Consider drawing $K$ elemental sets $J_1, ..., J_K$ with replacement to use as starts. For multivariate location and dispersion, use the attractor with the smallest MCD criterion to get the final estimator. For multiple linear regression, use the attractor with the smallest LMS, LTA, or LTS criterion to get the final estimator. For $500 \leq K \leq 3000$ and $p$ not much larger than 5, the elemental set algorithm is very good for detecting certain "outlier configurations," including i) a mixture of two regression hyperplanes that cross in the center of the data cloud for MLR (not an outlier configuration since outliers are far from the bulk of the data) and ii) a cluster of outliers that can often be placed close enough to the bulk of the data so that an MB, RFCH, or RMVN DD plot can not detect the outliers. However, the outlier resistance of elemental algorithms that use $K$ elemental sets decreases rapidly

as $p$ increases. All practical estimators have outlier configurations where they perform poorly. If $p$ is small, elemental algorithms tend to have trouble when there is a weak regression relationship for the bulk of the data and a cluster of outliers that are not good leverage points (do not fall near the hyperplane followed by the bulk of the data). The Buxton (1920) data set is an example.

Suppose the MLR data set has $n$ cases where $d$ are outliers and $n - d$ are "clean" (not outliers). The the outlier proportion $\gamma = d/n$. Suppose that $K$ elemental sets are chosen with replacement and that it is desired to find $K$ such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the $K$ elemental sets is clean) $\approx 1 - [1 - (1 - \gamma)^p]^K$ by independence. Hence $\alpha \approx [1 - (1 - \gamma)^p]^K$ or

$$K \approx \frac{\log(\alpha)}{\log([1 - (1 - \gamma)^p])} \approx \frac{\log(\alpha)}{-(1 - \gamma)^p} \tag{5.46}$$

using the approximation $\log(1 - x) \approx -x$ for small $x$. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \dfrac{3}{(1 - \gamma)^p}$. Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn. If the start passes through an outlier, so does the attractor. For concentration algorithms for multivariate location and dispersion, if the start passes through a cluster of outliers, sometimes the attractor would be clean. See Figures 3.9–3.15.

**Table 5.4** Largest $p$ for a 95% Chance of a Clean Subsample.

| | | | | $K$ | | | | |
|---|---|---|---|---|---|---|---|---|
| $\gamma$ | 500 | 3000 | 10000 | $10^5$ | $10^6$ | $10^7$ | $10^8$ | $10^9$ |
| 0.01 | 509 | 687 | 807 | 1036 | 1265 | 1494 | 1723 | 1952 |
| 0.05 | 99 | 134 | 158 | 203 | 247 | 292 | 337 | 382 |
| 0.10 | 48 | 65 | 76 | 98 | 120 | 142 | 164 | 186 |
| 0.15 | 31 | 42 | 49 | 64 | 78 | 92 | 106 | 120 |
| 0.20 | 22 | 30 | 36 | 46 | 56 | 67 | 77 | 87 |
| 0.25 | 17 | 24 | 28 | 36 | 44 | 52 | 60 | 68 |
| 0.30 | 14 | 19 | 22 | 29 | 35 | 42 | 48 | 55 |
| 0.35 | 11 | 16 | 18 | 24 | 29 | 34 | 40 | 45 |
| 0.40 | 10 | 13 | 15 | 20 | 24 | 29 | 33 | 38 |
| 0.45 | 8 | 11 | 13 | 17 | 21 | 25 | 28 | 32 |
| 0.50 | 7 | 9 | 11 | 15 | 18 | 21 | 24 | 28 |

Notice that the number of subsets $K$ needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors

$p$ but is free of $n$. Hawkins and Olive (2002) showed that if $K$ is fixed and free of $n$, then the resulting elemental or concentration algorithm (that uses $k$ concentration steps), is inconsistent and zero breakdown. See Theorem 5.13. Nevertheless, many practical estimators tend to use a value of $K$ that is free of both $n$ and $p$ (e.g. $K = 500$ or $K = 3000$). Such algorithms include ALMS = FLMS = `lmsreg` and ALTS = FLTS = `ltsreg`. The "A" denotes that an algorithm was used. The "F" means that a fixed number of trial fits ($K$ elemental fits) was used and the criterion (LMS or LTS) was used to select the trial fit used in the final estimator.

To examine the outlier resistance of such inconsistent zero breakdown estimators, fix both $K$ and the contamination proportion $\gamma$ and then find the largest number of predictors $p$ that can be in the model such that the probability of finding at least one clean elemental set is high. Given $K$ and $\gamma$, $P$(at least one of $K$ subsamples is clean) $= 0.95 \approx$
$1 - [1 - (1 - \gamma)^p]^K$. Thus the largest value of $p$ satisfies $\dfrac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{5.47}$$

if the sample size $n$ is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 5.4 shows the largest value of $p$ such that there is a 95% chance that at least one of $K$ subsamples is clean using the approximation given by Equation (5.47). Hence if $p = 28$, even with one billion subsamples, there is a 5% chance that none of the subsamples will be clean if the contamination proportion $\gamma = 0.5$. Since clean elemental fits have great variability, an algorithm needs to produce many clean fits in order for the best fit to be good. When contamination is present, all $K$ elemental sets could contain outliers. Hence basic resampling and concentration algorithms that only use $K$ elemental starts are doomed to fail if $\gamma$ and $p$ are large.

**Theorem 5.14.** Let $h = p$ be the number of randomly selected cases in an elemental set, and let $\gamma_o$ be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If $n$ is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{5.48}$$

**Proof.** As in Remark 3.5, if the contamination proportion $\gamma$ is fixed, then the probability of obtaining at least one clean subset of size $h$ with high probability (say $1 - \alpha = 0.8$) is given by $0.8 = 1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts $K$ and solve this equation for $\gamma$. □

The value of $\gamma_o$ depends on $c \geq n/2$ and $h$. To maximize $\gamma_o$, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided

every clean start is able to at least partially separate inliers (clean cases) from outliers. However, if $h = p = 50$, this proportion drops to 11%.

**Theorem 5.15.** If the clean data are in general position and if a high breakdown start is added to an FLTA, FLTS, or FLMS concentration algorithm, then the resulting estimator is HB.

**Proof.** Concentration reduces (or does not increase) the corresponding HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. □

For example, consider the LTS($c_n$) criterion. Suppose the ordered squared residuals from the high breakdown $m$th start $\boldsymbol{b}_{0m}$ are obtained. If the data are in general position, then $Q_{LTS}(\boldsymbol{b}_{0m})$ is bounded even if the number of outliers $d_n$ is nearly as large as $n/2$. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the $c_n$ smallest squared residuals $r^2_{(i)}(\boldsymbol{b}_{0m})$ for $i = 1, ..., c_n$. Denote these cases by $i_1, ..., i_{c_n}$. Then $Q_{LTS}(\boldsymbol{b}_{1m}) =$

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m}) = Q_{LTS}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce or at least do not increase the LTS criterion. If $c_n = (n+1)/2$ for $n$ odd and $c_n = 1+n/2$ for $n$ even, then the LTS criterion is bounded iff the median squared residual is bounded.

Theorem 5.15 can be used to show that the following two estimators are high breakdown. The estimator $\hat{\boldsymbol{\beta}}_B$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `hbreg` estimator of Definition 6.15.

**Definition 5.42.** Make an OLS fit to the $c_n \approx n/2$ cases whose $Y$ values are closest to the MED$(Y_1, ..., Y_n) \equiv$ MED$(n)$ and use this fit as the start for concentration. Define $\hat{\boldsymbol{\beta}}_B$ to be the attractor after $k$ concentration steps. Define $\boldsymbol{b}_{k,B} = 0.9999\hat{\boldsymbol{\beta}}_B$.

**Theorem 5.16.** If the clean data are in general position, then $\hat{\boldsymbol{\beta}}_B$ and $\boldsymbol{b}_{k,B}$ are high breakdown regression estimators.

**Proof.** The start can be taken to be $\hat{\boldsymbol{\beta}}_w$ with $w = 1$ from Theorem 5.12. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 5.15. Multiplying a HB estimator by a positive constant does not change the breakdown value, so $\boldsymbol{b}_{k,B}$ is HB. □

The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to a consistent estimator on a large class of iid zero mean symmetric error distributions, although the outlier resistance of the HB

estimator is poor. The following result may not hold if $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ estimates $\boldsymbol{\beta}_{LMS}$ where $\boldsymbol{\beta}_C \neq \boldsymbol{\beta}_{LMS}$. Then $\boldsymbol{b}_{k,B}$ could have a smaller median squared residual than $\hat{\boldsymbol{\beta}}_C$ even if there are no outliers. The two parameter vectors could differ because the constant term is different if the error distribution is not symmetric. For a large class of symmetric error distributions, $\boldsymbol{\beta}_{LMS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_C \equiv \boldsymbol{\beta}$, then the ratio $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}))/\text{MED}(r_i^2(\boldsymbol{\beta})) \to 1$ as $n \to \infty$ for any consistent estimator of $\boldsymbol{\beta}$. The estimator below has two attractors, $\hat{\boldsymbol{\beta}}_C$ and $\boldsymbol{b}_{k,B}$, and the probability that the final estimator $\hat{\boldsymbol{\beta}}_D$ is equal to $\hat{\boldsymbol{\beta}}_C$ goes to one under the strong assumption that the error distribution is such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ are consistent estimators of $\boldsymbol{\beta}$.

**Theorem 5.17.** Assume the clean data are in general position, and that the LMS estimator is a consistent estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_C$ be any practical consistent estimator of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_D = \hat{\boldsymbol{\beta}}_C$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \leq \text{MED}(r_i^2(\boldsymbol{b}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_D = \boldsymbol{b}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_D$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

**Proof.** The estimator is HB since the median squared residual of $\hat{\boldsymbol{\beta}}_D$ is no larger than that of the HB estimator $\boldsymbol{b}_{k,B}$. Since $\hat{\boldsymbol{\beta}}_C$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error $e^2$. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}_C$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_D$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

## 5.10 Complements

Following Cook and Weisberg (1999a, p. 396), a *residual plot* is a plot of a function of the predictors versus the residuals $r$, while a *model checking plot* is a plot of a function of the predictors versus the response. Researchers need to know what are the most important residual and model checking plots. For the *1D regression model* of Definition 1.1, the most important model checking plot is the *response plot* of $\hat{h}(\boldsymbol{x})$ versus $Y$, and the most important residual plot is the plot of $\hat{h}(\boldsymbol{x})$ versus $r$. If $p = 1$ so there is a single predictor $x$, then $h(x) = \hat{h}(x) = x$ and the response plot is widely used. For $p > 2$ the response plot is more important than any residual plot, but is not yet widely used.

Application 5.1 was suggested by Olive (2004b). An advantage of this graphical method is that it works for linear models: that is, for multiple linear regression and for many experimental design models. Notice that if the plotted points in the transformation plot follow the identity line, then the plot is also a response plot. The method is also easily performed for MLR methods other than least squares. Plotting the residual plots can also be useful,

but they do not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55). Response, residual, and transformation plots also very useful for outlier detection for linear models.

Cook and Olive (2001) also suggest a graphical method for selecting and assessing response transformations for linear models where the "transformation plot" of $\hat{Z}_i$ versus $W_i$ is made for each of the seven values of $\lambda \in \Lambda_L$.

In a classic paper, Box and Cox (1964) developed numerical methods for estimating $\lambda_o$ in the family of power transformations. This method also works for many experimental design models. It is well known that the Box–Cox normal likelihood method for estimating $\lambda_o$ can be sensitive to remote or outlying observations. Also see Tukey (1957). Yeo and Johnson (2000) provide a family of transformations that does not require the variables to be positive.

Section 5.4 followed Olive (2007) closely. See Di Bucchianico, Einmahl, and Mushkudiani (2001) for related intervals for the location model and Preston (2000) for related intervals for MLR. For a review of prediction intervals, see Patel (1989). Cai, Tian, Solomon, and Wei (2008) show that the Olive (2007) intervals are not optimal for symmetric bimodal distributions. Some references for PIs based on robust regression estimators are given by Giummolè and Ventura (2006). Chapter 7 gives PIs for after variable selection.

Excellent introductions to OLS diagnostics include Fox (1991) and Cook and Weisberg (1999a, p. 161-163, 183-184, section 10.5, section 10.6, ch. 14, ch. 15, ch. 17, ch. 18, and section 19.3). Hoaglin and Welsh (1978) examines the hat matrix while Cook (1977) introduces Cook's distance. Some other papers of interest include Hettmansperger and Sheather (1992), Velilla (1998), and Velleman and Welsch (1981).

Olive (2005) suggests using residual, response, RR, and FF plots to detect outliers while Hawkins and Olive (2002, p. 141, 158) suggest using the RR and FF plots. The four plots are best for $n > 5p$. Typically RR and FF plots are used if there are several estimators for one fixed model, e.g. OLS versus $L_1$ or frequentist versus Bayesian for multiple linear regression, or if there are several competing models. An advantage of the FF plot is that the response $Y$ can be added to the plot. FF and RR plots are useful for variable selection. Park, Kim, and Kim (2012) show response plots are competitive with the best robust regression methods for outlier detection on some outlier data sets that have appeared in the literature.

Rousseeuw and van Zomeren (1990) suggest that Mahalanobis distances based on "robust estimators" of location and dispersion can be more useful than the distances based on the sample mean and covariance matrix. They show that a plot of robust Mahalanobis distances $\mathrm{RD}_i$ versus residuals from "robust regression" can be useful.

Several authors have suggested using the response plot to visualize the coefficient of determination $R^2$ in multiple linear regression. See for example

Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280). Anderson-Sprecher (1994) provides an excellent discussion about $R^2$.

The fact that response plots are extremely useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. Certainly in any multiple linear regression analysis, the response plot and the residual plot of $\hat{Y}$ versus $r$ should always be made. Section 5.4 and Olive (2007) use the response plot to explain prediction intervals.

For more on the behavior of fits from randomly selected elemental sets, see Hawkins and Olive (2002), Olive (2008), and Olive and Hawkins (2007a).

## 5.11 Problems

**Problems with an asterisk * are especially important.**

**5.1.** Show that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is idempotent, that is, show that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}^2 = \boldsymbol{H}$.

**5.2.** Show that $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is idempotent, that is, show that $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

```
Output for Problem 5.3 Coefficient Estimates Response = height
Label              Estimate  Std. Error    t-value    p-value
Constant           227.351   65.1732         3.488     0.0008
sternal height     0.955973   0.0515390     18.549     0.0000
finger to ground   0.197429   0.0889004      2.221     0.0295

R Squared: 0.879324    Sigma hat:  22.0731

Summary Analysis of Variance Table
Source         df       SS        MS           F    p-value
Regression      2   259167.    129583.     265.96    0.0000
Residual       73   35567.2    487.222
```

**5.3.** The output above is from the multiple linear regression of the response $Y = \textit{height}$ on the two nontrivial predictors *sternal height* = height at shoulder and *finger to ground* = distance from the tip of a person's middle finger to the ground.

a) Consider the plot with $Y_i$ on the vertical axis and the least squares fitted values $\hat{Y}_i$ on the horizontal axis. Sketch how this plot should look if the multiple linear regression model is appropriate.

b) Sketch how the residual plot should look if the residuals $r_i$ are on the vertical axis and the fitted values $\hat{Y}_i$ are on the horizontal axis.

c) From the output, are *sternal height* and *finger to ground* useful for predicting *height*? (Perform the ANOVA F test.)

**5.4.** Suppose that the scatterplot of $X$ versus $Y$ is strongly curved rather than ellipsoidal. Should you use simple linear regression to predict $Y$ from $X$? Explain.

**5.5.** Suppose that the 95% confidence interval for $\beta_2$ is $[-17.457, 15.832]$. Suppose only a constant and $X_2$ are in the MLR model. Is $X_2$ a useful linear predictor for $Y$? If your answer is no, could $X_2$ be a useful predictor for $Y$? Explain.

**5.6.** Assume that the model has a constant $\beta_1$ so that the first column of $\boldsymbol{X}$ is $\mathbf{1}$. Show that if the regression estimator is regression equivariant, then adding $\mathbf{1}$ to $\boldsymbol{Y}$ changes $\hat{\beta}_1$ but does not change the slopes $\hat{\beta}_2, ..., \hat{\beta}_p$.

**5.7.** By the OLS CLT, under mild regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V})$. If $\boldsymbol{A}$ is a constant $k \times p$ matrix with rank $k$, what is the limiting distribution of $\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\boldsymbol{\beta})$?

**Problems using R.** Some $R$ code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *tplot*, will display the code for the function. Use the `args` command, e.g. *args(tplot)*, to display the needed arguments for the function.

**5.8*.** a) Download the $R$ function `tplot` that makes the transformation plots for $\lambda \in \Lambda_L$.

b) Use the following $R$ command to make a $100 \times 3$ matrix. The columns of this matrix are the three nontrivial predictor variables.

```
nx <- matrix(rnorm(300),nrow=100,ncol=3)
```

Use the following command to make the response variable Y.

```
y <- exp( 4 + nx%*%c(1,1,1) + 0.5*rnorm(100) )
```

This command means the MLR model $\log(Y) = 4 + X_2 + X_3 + X_4 + e$ will hold where $e \sim N(0, 0.25)$.

To find the response transformation, you need the program `tplot` given in a). Type **ls()** to see if the programs were downloaded correctly.

c) To make the transformation plots type the following command.

```
tplot(nx,y)
```

The first plot will be for $\lambda = -1$. Move the cursor to the plot and hold the **rightmost mouse key** down (and in *R*, highlight **stop**) to go to the next plot. Repeat these *mouse* operations to look at all of the plots. The identity line is included in each plot. When you get a plot where the plotted points cluster about the identity line with no other pattern, include this transformation plot in *Word* by pressing the **Ctrl** and **c** keys simultaneously. This will copy the graph. Then in *Word* use the menu commands "File>Paste". You should get the log transformation.

d) Type the following commands.

```
out <- lsfit(nx,log(y))
ls.print(out)
```

Use the mouse to highlight the created output and include the output in *Word*.

e) Write down the least squares equation for $\widehat{\log(Y)}$ using the output in d).

**5.9.** a) Download the *R* functions `piplot` and `pisim`.

b) The command `pisim(n=100, type = 1)` will produce the mean length of the classical, semiparametric, conservative and asymptotically optimal PIs when the errors are normal, as well as the coverage proportions. Give the simulated lengths and coverages.

c) Repeat b) using the command `pisim(n=100, type = 3)`. Now the errors are EXP(1) - 1.

d) Download `robdata.txt` and type the command
`piplot(cbrainx,cbrainy)`. This command gives the semiparametric PI limits for the Gladstone data. Include the plot in *Word*.

e) The infants are in the lower left corner of the plot. Do the PIs seem to be better for the infants or the bulk of the data? Explain briefly.

**5.10*.** a) After entering the two *source* commands above, enter the following command.

```
> MLRplot(buxx,buxy)
```

Click the rightmost mouse button (and in *R* click on *Stop*). The response plot should appear. Again, click the rightmost mouse button (and in *R* click on *Stop*). The residual plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

b) The response variable is *height*, but 5 cases were recorded with heights about 0.75 inches tall. The highlighted squares in the two plots correspond to cases with large Cook's distances. With respect to the Cook's distances, what is happening, swamping or masking?

c) *RR plots:* One feature of the MBA estimator (see Chapter 6) is that it depends on the sample of 7 centers drawn and changes each time the function is called. In ten runs, about seven plots will look like Figure 6.1, but in about three plots the MBA estimator will also pass through the outliers. Make the RR plot by pasting the commands for this problem into *R*, and include the plot in *Word*.

d) *FF plots: the plots in the top row will cluster about the identity line if the MLR model is good or if the fit passes through the outliers.* Make the FF plot by pasting the commands for this problem into *R*, and include the plot in *Word*.

**5.11.** a) If necessary, enter the two *source* commands above Problem 5.7. The `diagplot` function makes a scatterplot matrix of various OLS diagnostics.

b) Enter the following command and include the resulting plot in *Word*.

```
> diagplot(buxx,buxy)
```

**5.12.** This problem fits OLS to $n$ inliers and $k$ outliers. The inliers follow the model $Y = x + e$ (the mean function is the identity line) while the outliers are a near point mass with $(x, y) \approx (20, -20)$. Copy and paste the commands for this problem into *R*. Then copy and paste the four plots into *Word*.

The first three plots a), b), and c) use 1 outlier and $n = 10, 100$, and 1000. The OLS line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$ is added to each plot. When $n = 10$, the OLS line is tilted away from the identity line. There is still some tilt for $n = 100$ but little tilt for $n = 1000$. Plot d) uses 40 outliers but 10000 inliers, and the OLS line is close to the identity line. (The outlier resistance occurs since OLS minimizes $\sum r_i^2$. If the OLS line goes through the outliers, then the inliers are fit badly. If there are enough inliers, then fitting the inliers well and the outliers poorly leads to a lower OLS criterion than fitting the outliers well. One outlier can tilt OLS arbitrarily badly, but the one outlier needs to be very far from the bulk of the data if the number of inliers is large. A small percentage of outliers, e.g. 1%, can tilt OLS even if the outliers are not very fall from the bulk of the data.)