

5. Regression Diagnostics

In the preceding chapters the broad principles of multiple linear regression analysis have been described. The main features of the computer output have been presented, and what specific features to look for have been identified. The following is a summary of the main aspects of multiple regression analysis:

1. Multiple regression analysis builds on univariate and bivariate analysis. Before venturing into multivariate analysis it is best to explore the data and the distribution of the response and explanatory variables. This is then followed by descriptive statistics and bivariate analysis in which the relationships between the response and each individual explanatory variable are further examined. Once this information has been compiled, the relationship between several explanatory variables and the response variable may be further explored. Multiple regression will provide the independent contribution of each explanatory variable to the prediction of the outcome while controlling for the influence of the other explanatory variables.
2. Multiple linear regression is a direct generalisation of the simple linear regression. The main difference is that in multiple regression the value of the dependent variable depends on several independent variables instead of one.
3. In addition to describing the relationship between the independent and dependent variables, multiple regression can be used to judge the relative importance of different independent variables. This is done by comparing their relative t-ratios.
4. Categorical variables may be included in the regression equation by giving them Yes/No values like 1/0. This is an advantage because one can look at possible variations in outcome in the presence or absence of a categorical variable. If need be a new categorical variable may be created to represent a numerical variable. For example a variable like age can be split into two categories like young/old. Such variables are often referred to as dummy variables. The ability to combine continuous and dummy variables is one of the main strengths of multiple regression analysis in dealing with biomedical data.

As with all statistical procedures multiple linear regression analysis rests on **basic assumptions** about the population from where the data have been derived. The results of the analysis are only reliable when these assumptions are satisfied. These assumptions are:

- The relationship between the explanatory variables and the outcome variable is linear. In other words, each increase by one unit in an explanatory variable is associated with a fixed increase in the outcome variable.
- The regression equation describes the mean value of the dependent variable for a set of independent variables.
- The individual data points of Y (the response variable) for each of the explanatory variables are normally distributed about the line of means (regression line).
- The variance of the data points about the line of means is the same for each explanatory variable.
- The explanatory variables are independent of each other i.e. knowing the value of one or more of the independent variables does not tell us anything about the others. This is often not the case in real life because many variables are correlated. Such correlations can lead to mistaken conclusions.

That these assumptions are not violated can be checked by the following means:

1. For a continuous variable X the easiest way of checking for a linear relationship with Y is by means of a scatter plot of Y against X . hence a scatter plot matrix comprising all candidate variables (including Y) is a prudent way to commence.
2. Normality of distribution of Y data points can be checked by plotting a histogram of the residuals. (See below).
3. Ways of checking for independence of explanatory variables from one another (i.e. lack of collinearity) has been discussed in Chapter 4. Another way of testing for multi collinearity is by the Durbin-Watson statistic and Variance Inflation Factor, both of which are discussed below.

In doing regression analysis the important issue is which explanatory variables to include in the model, and how to interpret the results. This is usually clear from the research question being asked. Even then it is often necessary to fit several alternative regression models to see which one turns out to be the best for the purpose. Interpretation requires care. For example one may choose a model with first one explanatory variable, then two, and later three as follows:

$$\begin{array}{l} \alpha + \beta_1 X_1 \\ \text{Or} \quad \alpha + \beta_1 X_1 + \beta_2 X_2 \\ \text{Or} \quad \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \end{array}$$

In the first model β_1 represents the total regression coefficient of Y on X_1 . In the second model β_1 is partial regression coefficient of Y on X_1 taking account of X_2 . In the third model β_1 is partial regression coefficient of Y on X_1 taking account of both X_2 and X_3 .

Thus the meaning of a regression coefficient depends on which explanatory variables have been included. It is also important to specify the units of the response and the explanatory variables. As we have seen the relative importance of independent variables can be judged by comparing the 't' ratios of the β coefficients and not by their numerical values. This is because the value of the β coefficient depends on the units of the explanatory variable concerned. Each explanatory variable is measured in its own respective units. For example, in the data set about 'Crime rates in U.S.' X_1 is measured in numbers of males aged 14 – 24 per 1000 total state population; X_3 is measured as mean numbers of years of schooling; and X_4 is measured in U.S. dollars. Comparison between the coefficients of variables measured in such disparate units is not possible, unless we can have some form of standardisation. The t-ratios achieve this, since $\beta \div \text{standard deviation of } \beta = \text{t-ratio}$. The larger the t-ratio the more important is the effect of the variable concerned.

The regression equation is often referred to as 'the model' because it helps one to understand a situation by identifying the main sources of influence on the response variable. In that case it is best to use just enough explanatory variables to capture the main features of the data set. This is the **principle of parsimony**. If two models explain the data about equally well, choose the simpler one. The methods of selecting the best model have been demonstrated earlier in Chapter 4.

Clinical interpretation of regression analysis.

Multiple linear regression analysis enables the researcher to model a data set for predicting an outcome variable from one or more explanatory variables. The model is represented by the regression equation, which is strictly valid only within the range of data observed. Trying to

predict the outcome outside the range of the data can be seriously misleading, and is not advised.

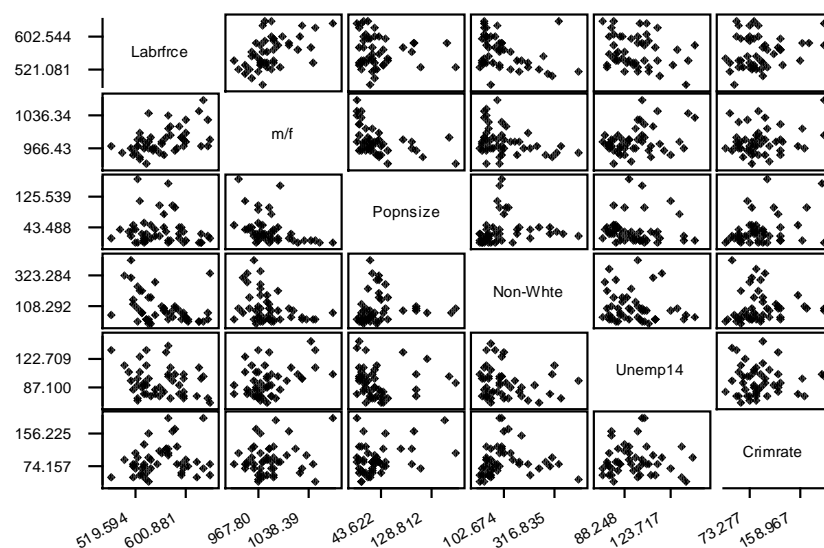
When several explanatory variables are responsible for an outcome, each subject provides a unique combination of these variables. A new subject on whom a clinician wishes to predict the outcome has also his own unique combination of variables. The measurements may or may not fall within the range of the original study. It is for this reason that prediction on the basis of a given model can only be a guide and is not infallible.

Multicollinearity

A problem, which commonly arises in all research and particularly in biomedical research, is that of multicollinearity. This was illustrated in Chapter 4. Multicollinearity means that some of the explanatory variables are not independent but are correlated. When multicollinearity is present it muddies the water and the regression coefficients become imprecise. It becomes difficult to assign the change in the dependent variable precisely to one or the other of the explanatory variables. The precision with which the parameters of each independent variable can be estimated is thereby reduced. It is therefore advisable that when the model has several explanatory variables the regression analysis should commence with first estimating the correlation coefficients between all the variables to be included in the model. In experimental research the investigator determines the explanatory variables, and the chances of multicollinearity are thereby much diminished. In nature, on the other hand, a number of variables are related. Hence in observational studies the chances of multicollinearity are ever present. As we saw with the 'Crime in U.S.' data, one can check for multicollinearity by means of the correlation matrix. In such a matrix when the correlation coefficient between two explanatory variables is above 0.8 one needs to be aware of possible collinearity. If the correlation coefficient is above 0.95 the problem is really serious.

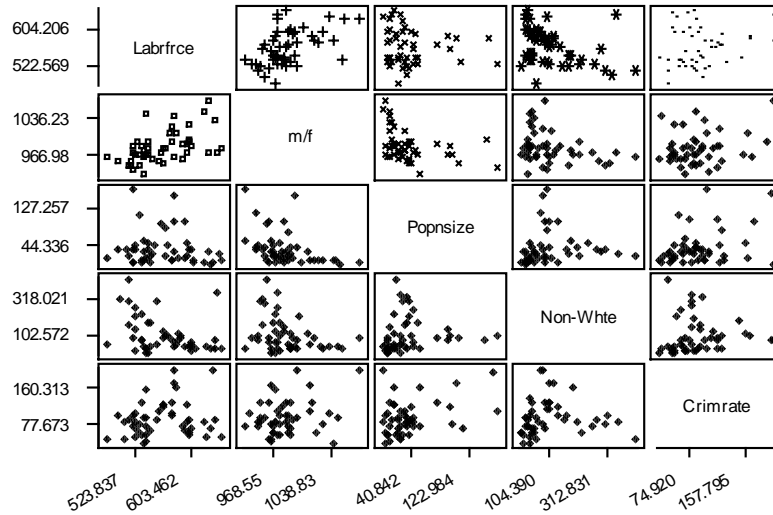
Correlations convey limited information. They may conceal problems like curvilinearity, outliers or clustering of data points. Scatterplots reveal more detail and can be scanned almost as quickly as correlation matrices. Most modern statistical software can provide scatterplot matrices. An example of scatterplot matrices for the data set 'Crime in U.S.' is shown below.

Scatterplot matrix of X6 to X10 against Crime rate



As we have already seen, another use of a scatter plot matrix is to identify what explanatory variables to choose for the first round of regression analysis. Those explanatory variables which appear to correlate well with Y, the outcome variable are to be the more likely candidates for inclusion.

Scatterplot matrix for variables X11 to X13 against Crime rate



A diagnostic approach to check for multicollinearity after performing regression analysis is to display the Variance Inflation factor (**VIF**). This is shown below for the ‘crime in U.S.’ data set using the best set of explanatory variables.

The regression equation is

$$\text{Crimrate} = -619 + 1.13 \text{ MalePop} + 1.82 \text{ YrsSchl} + 1.05 \text{ expend60} + 0.828 \text{ Unemp35} + 0.160 \text{ Incm} + 0.824 \text{ Pvrty}$$

Predictor	Coef	StDev	T	P	VIF
Constant	-618.5	108.2	-5.71	0.000	
MalePop	1.1252	0.3509	3.21	0.003	2.1
YrsSchl	1.8179	0.4803	3.79	0.001	3.1
expend60	1.0507	0.1752	6.00	0.000	2.9
Unemp35	0.8282	0.4274	1.94	0.060	1.4
Incm	0.15956	0.09390	1.70	0.097	8.7
Pvrty	0.8236	0.1815	4.54	0.000	5.6

S = 20.83 R-Sq = 74.8% R-Sq(adj) = 71.0%

VIF is a measure of how much the variance of an estimated regression coefficient increases if the explanatory variables are correlated. The length of the confidence interval for the parameter estimates is increased by the square root of the respective **VIF** as compared to uncorrelated variables. The higher the value of **VIF** the greater is the degree of collinearity. Some authors suggest that if the **VIF** is >10 there is strong evidence that collinearity is affecting the regression coefficients and consequently they are poorly estimated. For comparison we next examine the **VIF** in the full model of the same data set where we know that two explanatory variables are correlated.

Another check for collinearity is the **Durbin-Watson** statistic. Normally its value should lie between 0 and 4. A value close to 2 suggests no correlation; one close to 0 negative

correlation, and a value close to 4 positive correlation. Both VIF and the Durbin-Watson Statistic are illustrated in the regression analysis of the full model of the data set 'Crime in U.S.' in which as we know there is close correlation between two variables.

The regression equation is

$$\begin{aligned} \text{Crimrate} = & -692 + 1.04 \text{ MalePop} - 8.3 \text{ Sth-Rest} + 1.80 \text{ YrsSchl} + 1.61 \text{ expend60} \\ & - 0.67 \text{ Expend59} - 0.041 \text{ Labrfrce} + 0.165 \text{ m/f} - 0.041 \text{ Popnsize} \\ & + 0.0072 \text{ Non-White} - 0.602 \text{ Unempl4} + 1.79 \text{ Unemp35} + 0.137 \text{ Incm} \\ & + 0.793 \text{ Pvrty} \end{aligned}$$

Predictor	Coef	StDev	T	P	VIF
Constant	-691.8	155.9	-4.44	0.000	
MalePop	1.0398	0.4227	2.46	0.019	2.7
Sth-Rest	-8.31	14.91	-0.56	0.581	4.9
YrsSchl	1.8016	0.6497	2.77	0.009	5.0
expend60	1.608	1.059	1.52	0.138	94.6
Expend59	-0.667	1.149	-0.58	0.565	98.6
Labrfrce	-0.0410	0.1535	-0.27	0.791	3.7
m/f	0.1648	0.2099	0.78	0.438	3.7
Popnsize	-0.0413	0.1295	-0.32	0.752	2.3
Non-White	0.00717	0.06387	0.11	0.911	4.1
Unempl4	-0.6017	0.4372	-1.38	0.178	5.9
Unemp35	1.7923	0.8561	2.09	0.044	5.0
Incm	0.1374	0.1058	1.30	0.203	10.0
Pvrty	0.7929	0.2351	3.37	0.002	8.4

Durbin-Watson statistic = 1.48

Regression Diagnostics

Diagnostic procedures are intended to check how well the assumptions of multiple linear regression are satisfied. Infringement of these assumptions cast doubt on the validity of the conclusions drawn on the basis of the results. A number of checks and tests help us to ensure that analysis has proceeded within the bounds of the basic assumptions. The checks fall into two groups viz. (i). Those for checking the pattern of the residuals by means of residual plots, and (ii). Those for individual data points.

Residual plots are the best single check for violation of assumptions, such as:

- (i). Variance not being constant across the explanatory variables.
- (ii). Fitted relationships being non-linear.
- (iii). Random variation not having a normal distribution.

Residual, as we saw earlier, is the difference between the calculated mean value of Y (this is also the fitted value as determined by the regression line) and the actual observed value of Y for a given value of the explanatory variable. Thus the residuals tell us how well or otherwise the model fits the data. One problem with using residuals is that their values depend on the scale and units used. Since the residuals are in units of the dependent variable Y there are no cut-off points for defining what is a "large" residual. The problem is overcome by using standardised residuals. They are calculated by residual \div standard error of the residual. The standard error of each residual is different, and using standardised residuals helps one to get round the problem. For example, a standardised residual of 1 means that the residual is 1 standard deviation away from the regression line. A standardised residual of 2 means that the residual is 2 standard deviations away from the regression line. A standardised residual of 0 means that the point falls on the regression line. Most computer software can provide both the residual and standardised residual. We would expect about two-thirds of the standardised residuals to have values below 1, and almost all standardised residuals to have values below 2. Observations with standardised residuals exceeding 3 require close consideration as potential outliers.

Plotting residuals (resids) on the Y axis against fitted values (fits) on the X axis is a useful diagnostic procedure. If the model is appropriate for the data the plot should show an even scatter. Any discernible pattern in the plot means that the regression equation does not describe the data correctly, since pattern forms when the residuals are unevenly distributed about the regression line. Outliers may also get identified in such a plot.

In addition one also needs scatter plots with standardised residuals on the vertical axis and each predictor variable, by turn, on the horizontal axis. These should show the same amount of variation in the standardised residuals for all the predictors.

The use of residual plots for regression diagnostics is illustrated below employing the six best subset of explanatory variables from the data set ‘Crime in the U.S.’

The regression equation is

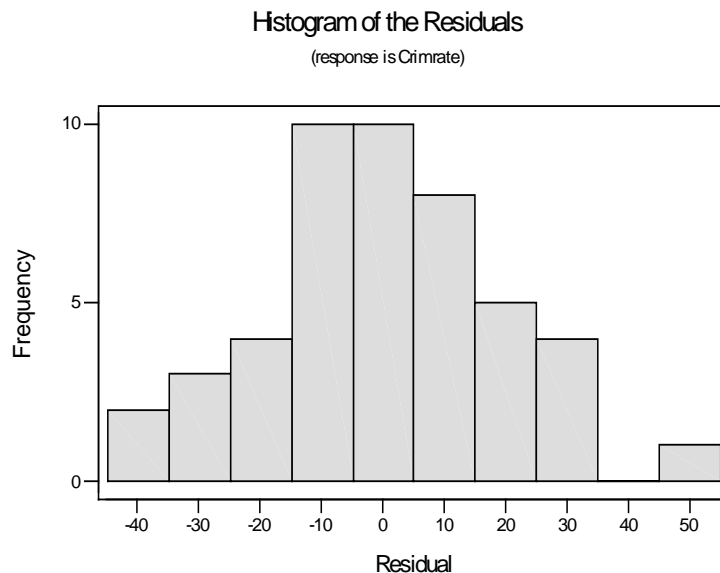
$$\text{Crimrate} = - 619 + 1.13 \text{ MalePop} + 1.82 \text{ YrsSchl} + 1.05 \text{ expend60} + 0.828 \text{ Unemp35} + 0.160 \text{ Incm} + 0.824 \text{ Pvrty}$$

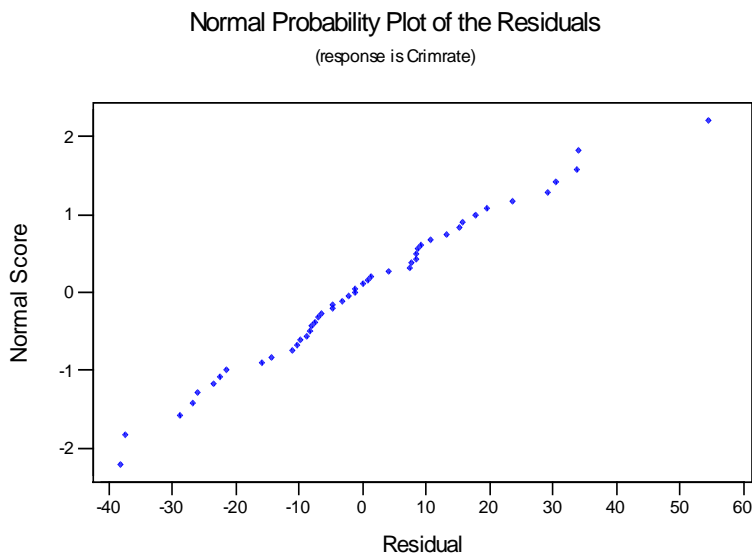
Predictor	Coef	StDev	T	P	VIF
Constant	-618.5	108.2	-5.71	0.000	
MalePop	1.1252	0.3509	3.21	0.003	2.1
YrsSchl	1.8179	0.4803	3.79	0.001	3.1
expend60	1.0507	0.1752	6.00	0.000	2.9
Unemp35	0.8282	0.4274	1.94	0.060	1.4
Incm	0.15956	0.09390	1.70	0.097	8.7
Pvrty	0.8236	0.1815	4.54	0.000	5.6

S = 20.83 R-Sq = 74.8% R-Sq(adj) = 71.0%

It is usual to examine four plots.

1. A histogram of the residuals will help to check the assumption about normal distribution.



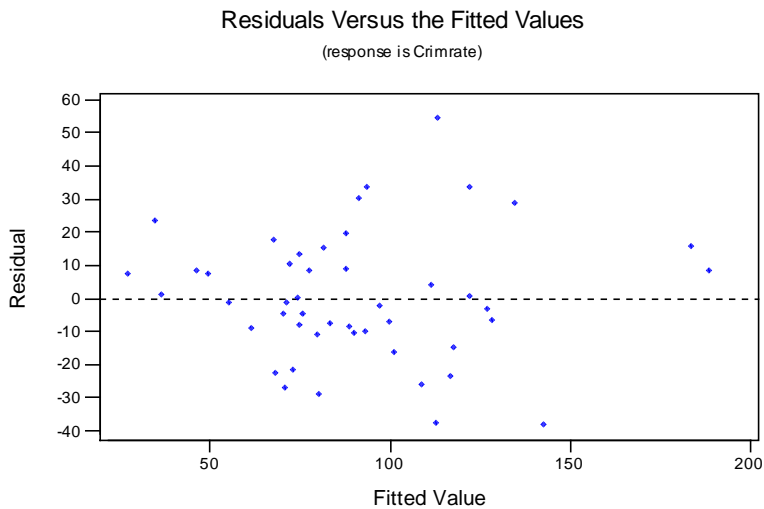


2. A normal Probability Plot of the residuals serves the same purpose by showing a straight line for normal distribution.

3. A plot of residuals against fitted values can show whether the fit is uniformly good or different for lower or higher values of Y.

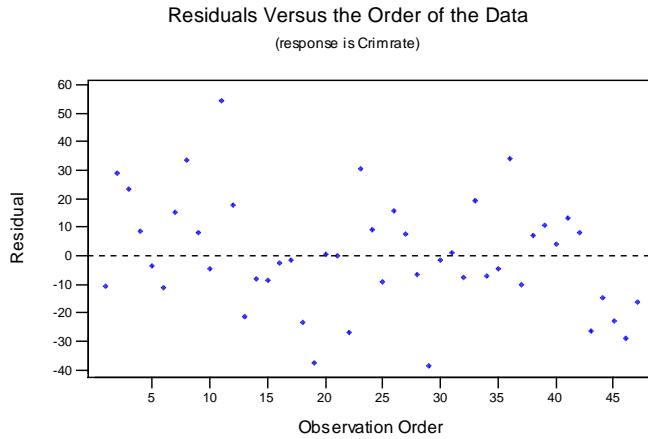
Where residuals get more scattered for larger 'Fitted Values' one may either fit a non-linear regression line, or transform the data. The most common transformation is log transformation.

Many relationships that have a curve in them respond well to log transformation, because a

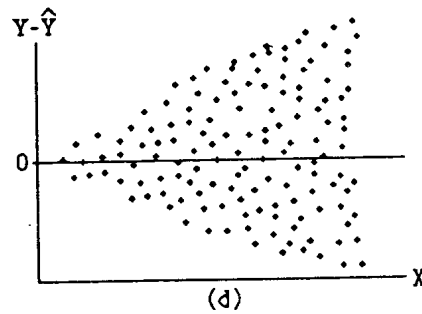
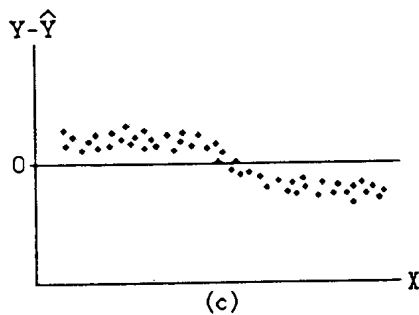
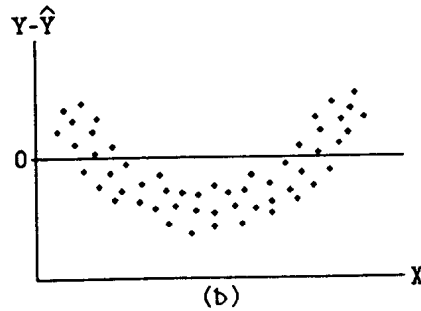
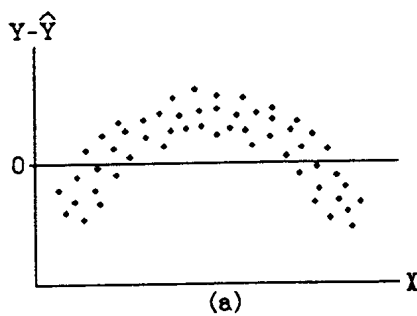


relationship like $Y = aX^n$ converts to a linear relationship like $\log Y = \log a + n \log X$

4. A plot of residuals versus fitted values by the order in which the data were entered helps to identify abnormal data points.



Patterns that suggest violation of assumptions are shown below for comparison.



Patterns (a) and (b) are seen when the relationship between X and Y is non-linear. In subtracting the fitted value of Y from Y (i.e. $Y - \hat{Y}$) a residual plot removes the linear tendency making other trends more obvious.

Pattern (c) suggests auto-correlation. Adjacent observations tend to have residuals of the same sign. The Durbin-Watson test is specifically designed to identify such a situation. Pattern (d) suggests that the variance in Y values is non-uniform (heterogeneous). The spread of Y values

is far greater for larger values of X than for small ones. This calls for transformation of data. As we have seen a log transformation stabilises the variance.

Checking individual observations

Leverage. Data points which are a long distance away from the rest of the data, can exercise undue influence on the regression line. A long distance away means an extreme value (either too low or too high compared to the rest). A point with a large residual is called an outlier. Such data points are of interest because they have an influence on the parameter estimates. Leverage is a way of checking on extreme values. Data points with high leverage have the *potential* of moving the regression line up or down as the case may be. Recall that the regression line represents the regression equation in a graphic form, and is represented by the β coefficients. High leverage points make our estimation of β coefficients inaccurate. Any conclusions drawn about which explanatory variables are related to the response variable could be misleading. Similarly any predictions made on the basis of the regression model could be wrong.

If p is the number of parameters in the model including the constant, then data points with a leverage value of $(3 \times p) \div n$ or 0.99 (whichever the smallest) are flagged with an X. Those data points with leverage value of $>5p/n$ are flagged with XX. These data points merit checking. The leverages for the data points in the above analysis are given below:

HI (leverages)					
0.099150	0.078397	0.210201	0.213855	0.138676	0.196886
0.109479	0.244874	0.098683	0.092566	0.156575	0.076160
0.136680	0.100084	0.124931	0.125860	0.159732	0.108542
0.112711	0.216581	0.098475	0.194544	0.138138	0.114861
0.160648	0.215233	0.212229	0.146591	0.307372	0.157140

In the current analysis 7 parameters have been measured including the constant; $(3 \times 7) \div 47$ gives an upper limit of 0.447 for leverage. This limit is not exceeded by any of the values given in the matrix above.

Dividing the number of parameters (i.e. $p + 1$) by N the number of subjects gives the mean value of h_i . As a rule of thumb take any $h_i >$ twice the mean (i.e. $> 2(p+1)/n$) as large. So for a data set with 3 parameters and 100 subjects the mean value of h_i is $2(3+1)/100 = 0.08$. And twice that is 0.16. Any value of $h_i > 0.16$ for this data set would be considered large.

Standardized residuals. It is calculated by dividing the residual by its standard error. A standardized residual greater than 2 requires close scrutiny since it indicates that an observation is unusual in the Y value.

Cook's Distance. If leverage gives us a warning about data points that have the *potential* of influencing the regression line then Cook's Distance indicates how much *actual influence* each case has on the slope of the regression line. Cook's Distance is a measure of the distance between coefficients calculated with and without the particular data points. It takes into account both leverage and residuals. Cook's D can be interpreted as a measure of how different the regression coefficients (including the intercept) would be if the particular observation is left out of the analysis altogether. Cook's D is thus a way of identifying data points that actually do exert too big an influence. Large values for Cook's Distance signify unusual observations. Values >1 require careful checking; those >4 are potentially serious outliers. Values of Cook's Distance in the analysis of the current data set are listed next.

COOK					
0.004386	0.025562	0.061248	0.008511	0.000682	0.012425
0.010511	0.160351	0.002788	0.000831	0.215658	0.009127
0.027902	0.002656	0.003815	0.000270	0.000110	0.024886
0.066366	0.000051	0.000000	0.071148	0.056398	0.004070
0.005943	0.028183	0.006282	0.002909	0.309624	0.000126
0.000086	0.002334	0.013923	0.001153	0.001790	0.156716
0.014432	0.003046	0.003569	0.000367	0.018654	0.004796
0.048927	0.009686	0.083666	0.062473	0.012167	

None of the coefficients is >1.

Once the diagnostic techniques have identified outliers or potentially influential points what should one do?

There are two possible explanations for the discrepancy:

1. Something is wrong with the particular data point, which is making it not fit the overall pattern emerging from the rest of the data.
2. The model is not correctly specified and does not describe the data, as it should.

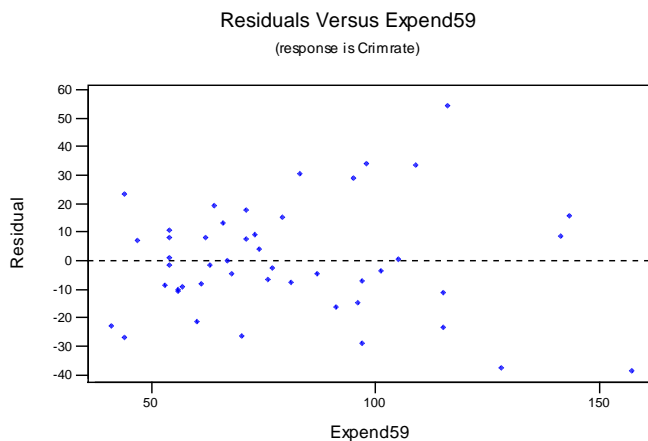
First step is to check the data for errors in data entry. If none is found, one should go back to the clinical and laboratory notes to check on the accuracy of the data, and make sure that there are no mistakes in transcribing the data.

Once it has been ascertained that there are no errors with regard to the data, the regression model needs checking. There are two common problems:

1. Some important variable may not have been included in the model.
2. Non-linear relationships or interactions between variables may not have been taken into account.

A plot of the residuals against potentially important explanatory variables can help to decide whether any merits inclusion in the model. If the variable in question has a large influence the plot should show a pattern. If it has no influence the points on the plot would be randomly distributed. Moreover, the researcher’s knowledge of the situation can provide a clue about the exclusion of other variables.

In selecting our data subset the variable expend 59 was left out because of collinearity with expend 60. This was correct. But just to illustrate a point the plot of residuals against expend 59 is given below:



A fan-shaped pattern is obvious indicating that expend 59 could be considered as a potential candidate for inclusion amongst explanatory variables. However, with the sub-set we have the residual plots as well as values for leverage and Cook's Distance are acceptable. It would be appropriate not to include expend 59, and to consider the sub-set of explanatory variables as satisfactory.

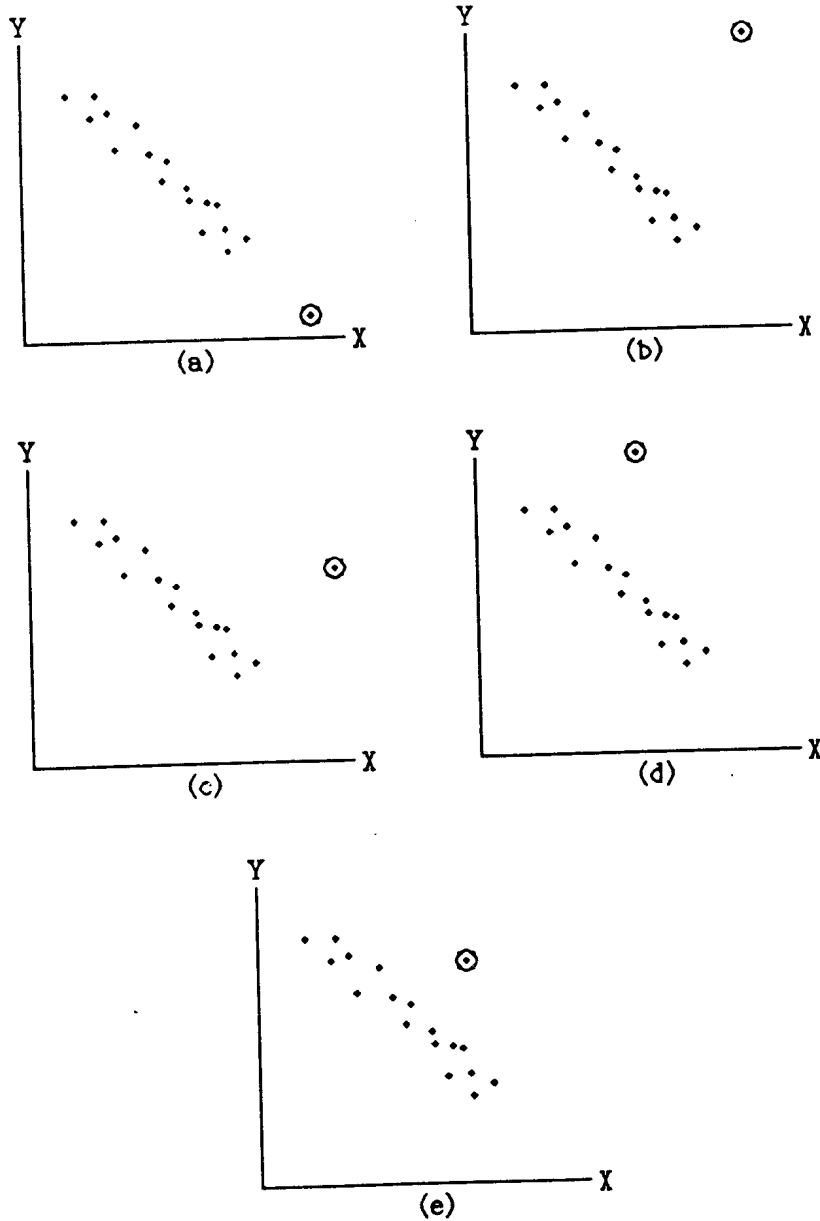
In this particular study excluding the variable 'Expend 59' did not matter much. But when two variables have been shown to be collinear and both merit inclusion, it is often useful to carry out the regression analysis using the difference between them as one X variable and the sum or mean as another.

There are often situations where these diagnostics do not give much help. In such cases one continues with the analysis strengthened in one's view that the diagnostic procedures have increased one's confidence in the results.

In summary, the diagnostics considered viz. Standardised residuals; plots of residuals; leverages (H_i) and Cook's Distance (Cook's D) statistics are ways of identifying points that do not fit with the regression model. These data points can then be checked and double-checked for errors.

Once the flagged outliers have been thoroughly checked the obvious next step is to ask whether the high value is due to biological diversity. In that case the outlier in question may turn out to be an exciting finding of the data and may lead to further research.

The main patterns of outliers are illustrated below:



- (a). Outlier is extreme in both X and Y co-ordinates, but not in pattern. Its removal is unlikely to alter the regression line.
- (b). Outlier is extreme in both X and Y, and also in the overall pattern. Its inclusion will strongly influence the regression line.
- (c). Outlier is extreme for X, but nearly average for Y.
- (d). Outlier is extreme in Y, not in X.
- (e). Outlier is extreme in pattern, but not in X or Y.