Chapter 6

Regression 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)

6.1 Numerical Diagnostics

Diagnostics are used to check whether model assumptions are reasonable. Section 6.4 provides a graph for assessing model adequacy for very general regression models while the first three sections of this chapter focus on diagnostics for the multiple linear regression model with iid constant variance symmetric errors. Under this model,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$ where the errors are iid from a symmetric distribution with $\text{E}(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$.

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,

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{e},$$

$$\boldsymbol{X} = [X_1, X_2, ..., X_p] = [\mathbf{1}, \boldsymbol{U}],$$

$\mathbf{1}$ 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 kth column of $\boldsymbol{U}$ is the $n \times 1$ vector of the

jth 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

$$\overline{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}_i \tag{6.1}$$

and

$$\boldsymbol{C} = \text{Cov}(\boldsymbol{U}) = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{6.2}$$

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 $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{OLS}$. Recall that the vector of fitted values =

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H}$ is the *hat matrix.* Recall that the ith *residual* $r_i = Y_i - \widehat{Y}_i$. *Case* (or *leave one out* or *deletion*) diagnostics are computed by omitting the ith case from the OLS regression. Following Cook and Weisberg (1999a, p. 357), let

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{6.3}$$

denote the $n \times 1$ vector of fitted values for estimating $\boldsymbol{\beta}$ with OLS without the ith case. Denote the jth element of $\widehat{\boldsymbol{Y}}_{(i)}$ by $\widehat{Y}_{(i),j}$. It can be shown that the variance of the ith residual $\text{VAR}(r_i) = \sigma^2(1-h_i)$. The usual estimator of the error variance is

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-p}.$$

The (internally) *studentized residual*

$$\widehat{e}_i = \frac{r_i}{\widehat{\sigma}\sqrt{1-h_i}}$$

has zero mean and unit variance.

Definition 6.1. The ith *leverage* $h_i = \boldsymbol{H}_{ii}$ is the ith diagonal element of the hat matrix $\boldsymbol{H}$. The ith *squared (classical) Mahalanobis distance*

$$\text{MD}_i^2 = (\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T\boldsymbol{C}^{-1}(\boldsymbol{u}_i - \overline{\boldsymbol{u}}).$$

The ith *Cook's distance*

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{6.4}$$

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

Proposition 6.1. a) (Rousseeuw and Leroy 1987, p. 225)

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

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)

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

When the statistics CD_i, h_i and MD_i are large, case i may be an outlier or *influential* case. Examining a stem plot or 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 $\text{CD}_i > 0.5$ and that cases with $\text{CD}_i > 1$ 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 $\text{trace}(\boldsymbol{H}) = \sum_{i=1}^n h_i = p$. Rousseeuw and Leroy (1987, p. 220 and p. 224) suggest using $h_i > 2p/n$ and $\text{MD}_i^2 > \chi^2_{p-1,0.95}$ 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 Proposition 6.1c) 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 plots such as residuals versus fitted values and fitted values versus the response are probably the *most effective techniques* for detecting cases that effect the fitted values when the multiple linear regression model is a good approximation for the bulk of the data. In fact, these diagnostics may be useful for perhaps up to 90% of such data sets while residuals from robust regression and Mahalanobis distances from robust estimators of multivariate location and dispersion may be helpful for perhaps another 3% of such data sets.

6.2 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), by far the *most 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.

Similarly if there are two nontrivial predictors, say x_2 and x_3, make a three-dimensional (3D) plot with Y on the vertical axis, x_2 on the horizontal axis and x_3 on the out of page axis. Rotate the plot about the vertical axis, perhaps superimposing the OLS plane. As the plot is rotated, linear combinations of x_2 and x_3 appear on the horizontal axis. If the OLS plane $b_1 + b_2x_2 + b_3x_3$ fits the data well, then the plot of $b_2x_2 + b_3x_3$ versus Y should scatter about a straight line. See Cook and Weisberg (1999a, ch. 8).

In general there are more than two nontrivial predictors and in this setting two plots are **crucial for any multiple linear regression analysis,** regardless of the regression estimator (eg OLS, L_1 etc.). The first plot is a scatterplot of the fitted values $\widehat{Y}_i$ versus the residuals r_i, and the second plot is a scatterplot of the fitted values $\widehat{Y}_i$ versus the response Y_i.

Definition 6.2. 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 multiple linear regression model with iid constant variance symmetric errors *is not sustained.* In other words, if the variables in the residual plot show some type of dependency, eg increasing variance or a curved pattern, then the multiple linear regression model may be inadequate. The following proposition shows 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 multiple linear regression 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.*

Proposition 6.2. 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 $Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = r_i(\boldsymbol{b})$. QED

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 6.3. 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 FF plot. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$.

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 ith residual from the jth 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 ith observation. Similarly, let the ith fitted value from the jth fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

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

$$\leq \|\boldsymbol{x}_i\| \; (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{6.5}$$

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. Adding the response Y to the scatterplot matrix of fitted values can also be useful.

To illustrate the RR plot, we examined two moderately-sized data sets (in Chapter 1) with four *R/Splus* estimators: OLS, ALMS = the default version of `lmsreg`, ALTS = the default version of `ltsreg` and the MBA estimator described in Chapter 7. In the 2007 version of *R*, the last three estimators change with each call.

Example 6.1. Gladstone (1905-6) 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 (on p. 7) shows the RR plot. The five infants seem to be "good leverage points" in than 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 (on p. 8) 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 6.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We chose to predict *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index.* Observation 9 was deleted since it had missing values. Five individuals, numbers 62-66, 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 7.1 (on p. 246) shows the RR plot for *Splus-2000* implementations of `lmsreg` and `ltsreg`. Only the MBA estimator makes the absolute residuals large. Problem 6.1 shows how to create RR and FF plots.

Example 6.3. Figure 1.6 (on p. 16) is nearly identical to a response plot. Since the plotted points do not scatter about the identity line, the multiple linear regression model is not appropriate. Nevertheless,

$$Y_i \propto (\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{OLS})^3.$$

In Chapter 12 it will be shown that the response plot is useful for visualizing the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ in 1D regression models where

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}.$$

6.3 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 6.4. 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 7.1 on p. 7 and p. 246, respectively.

3. Make an FF plot. See Problem 6.1.

4. Display the residual plots from several different estimators. See Figure 1.2 on p. 8.

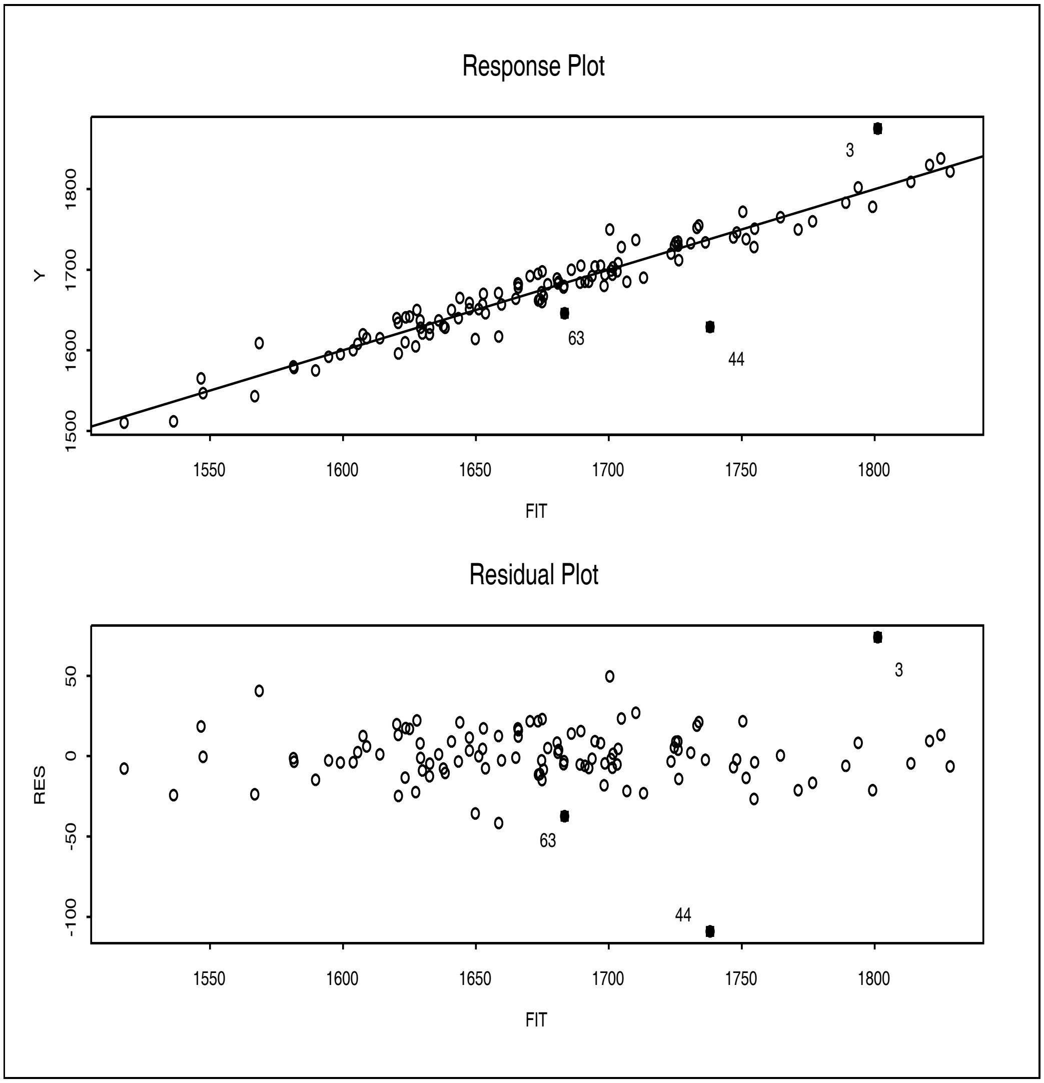

Figure 6.1: Residual and Response Plots for the Tremearne Data

5. Display the response plots from several different estimators. This can be done by adding Y to the FF plot.

6. Make a scatterplot matrix of several diagnostics such as leverages, Cook's distances and studentized residuals.

Example 6.4. 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 6.1 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. Two other cases have residuals near fifty.

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 1.4 on p. 12–13 should be used.

6.4 A Simple Plot for Model Assessment

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. Many important statistical models have the form

$$Y_i = m(x_{i1}, ..., x_{ip}) + e_i = m(\boldsymbol{x}_i^T) + e_i \equiv m_i + e_i \tag{6.6}$$

for $i = 1, ..., n$ where the zero mean error e_i is independent of $\boldsymbol{x}_i$. Additional assumptions on the errors are often made.

The above class of models is very rich. Many anova models, categorical models, nonlinear regression, nonparametric regression, semiparametric and time series models have this form. An additive error *single index model* uses

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{6.7}$$

The *multiple linear regression model* is an important special case. A *multi–index model* with additive error has the form

$$Y = m(\boldsymbol{\beta}_1^T \boldsymbol{x}, ..., \boldsymbol{\beta}_k^T \boldsymbol{x}) + e \tag{6.8}$$

where $k \geq 1$ is as small as possible. Another important special case of model (6.6) is the *response transformation model* where

$$Z_i \equiv t^{-1}(Y_i) = t^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}_i + e_i)$$

and thus

$$Y_i = t(Z_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i. \tag{6.9}$$

There are several important regression models that do not have additive errors including generalized linear models. If

$$Y = g(\boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{6.10}$$

then the regression has 1–dimensional structure while

$$Y = g(\boldsymbol{\beta}_1^T \boldsymbol{x}, ..., \boldsymbol{\beta}_k^T \boldsymbol{x}, e) \tag{6.11}$$

has k–dimensional structure if $k \geq 1$ is as small as possible. These models do not necessarily have additive errors although models (6.7) and (6.8) are important exceptions.

Definition 6.5 (Cook and Weisberg 1997, 1999a, ch. 17): A plot of $\boldsymbol{a}^T\boldsymbol{x}$ versus Y for various choices of $\boldsymbol{a}$ is called a *model checking plot*.

This plot is useful for model assessment and emphasizes the goodness of fit of the model. In particular, plot each predictor x_j versus Y, and also plot $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y if model (6.10) holds. Residual plots are also used for model assessment, but residual plots emphasize lack of fit.

The following notation is useful. Let $\hat{m}$ be an estimator of m. Let the ith predicted or fitted value $\hat{Y}_i = \hat{m}_i = \hat{m}(\boldsymbol{x}_i^T)$, and let the ith residual $r_i = Y_i - \hat{Y}_i$.

Definition 6.6. Then a *fit–response plot* or *FY plot* is a plot of $\hat{Y}$ versus Y.

Application 6.1. Use the FY plot to check the model for goodness of fit, outliers and influential cases.

To understand the information contained in the FY plot, first consider a plot of m_i versus Y_i. Ignoring the error in the model $Y_i = m_i + e_i$ gives $Y = m$ which is the equation of the *identity line* with unit slope and zero intercept. The vertical deviations from the identity line are $Y_i - m_i = e_i$. The reasoning for the FY plot is very similar. The line $Y = \hat{Y}$ is the identity line and the vertical deviations from the line are the residuals $Y_i - \hat{m}_i = Y_i - \hat{Y}_i = r_i$. Suppose that the model $Y_i = m_i + e_i$ is a good approximation to the data and that $\hat{m}$ is a good estimator of m. If the identity line is added to the plot as a visual aid, then the plotted points will scatter about the line and the variability of the residuals can be examined.

For a given data set, it will often be useful to generate the FY plot, residual plots, and model checking plots. An advantage of the FY plot is that if the model is not a good approximation to the data or if the estimator $\hat{m}$ is poor, then detecting deviations from the identity line is simple. Also, residual variability is easier to judge against a line than a curve. On the other hand, model checking plots may provide information about the form of the conditional mean function $E(Y|\boldsymbol{x}) = m(\boldsymbol{x}^T)$. See Chapter 12.

Many numerical diagnostics for detecting outliers and influential cases on the fit have been suggested, and often this research generalizes results from Cook (1977, 1986) to various models of form (6.6). Information from these diagnostics can be incorporated into the FY plot by highlighting cases that have large absolute values of the diagnostic.

The most important example is the multiple linear regression (MLR) model. For this model, the FY plot is the response plot. If the MLR model holds and the errors e_i are iid with zero mean and constant variance σ^2, then the plotted points should scatter about the identity line with no other pattern.

When the bulk of the data follows the MLR model, 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. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit a MLR estimator 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 bulk of the data but not the cluster, then the cluster points may be outliers.

To see why gaps are important, suppose that OLS was used to obtain $\hat{Y} = \hat{m}$. 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.

A commonly used diagnostic is Cook's distance CD_i. Assume that OLS is used to fit the model and to make the FY plot $\hat{Y}$ versus Y. Then 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 $\overline{Y}$.

Now suppose that the MLR model is incorrect. If OLS is used in the FY plot, and if $Y = g(\boldsymbol{\beta}^T \boldsymbol{x}, e)$, then the plot can be used to visualize g for many data sets (see Ch. 12). Hence the plotted points may be very far from linear. The plotted points in FY plots created from other MLR estimators may not be useful for visualizing g, but will also often be far from linear.

An advantage of the FY plot over numerical diagnostics is that while it depends strongly on the model m, defining diagnostics for different fitting methods can be difficult while the FY plot is simply a plot of $\hat{Y}$ versus Y. For the MLR model, the FY plot can be made from any good MLR estimator, including OLS, least absolute deviations and the *R/Splus* estimator `lmsreg`.

Example 6.2 (continued): Figure 6.2 shows the response plot and residual plot for the Buxton data. 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 response plot with $CD_i > \min(0.5, 2p/n)$ were highlighted. 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. Hence *masking* occurred for both Cook's distances and for OLS residuals, but not for OLS fitted values. Figure 7.1 on p. 246 shows that plots using `lmsreg` and `ltsreg` were similar, but MBA was effective.

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 11.3 and Chapter 12 is also effective for detecting outliers and other departures from the MLR model.)

Example 6.5. Hawkins, Bradu, and Kass (1984) present a well known artificial data set where the first 10 cases are outliers while cases 11-14 are good leverage points. Figure 6.3 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

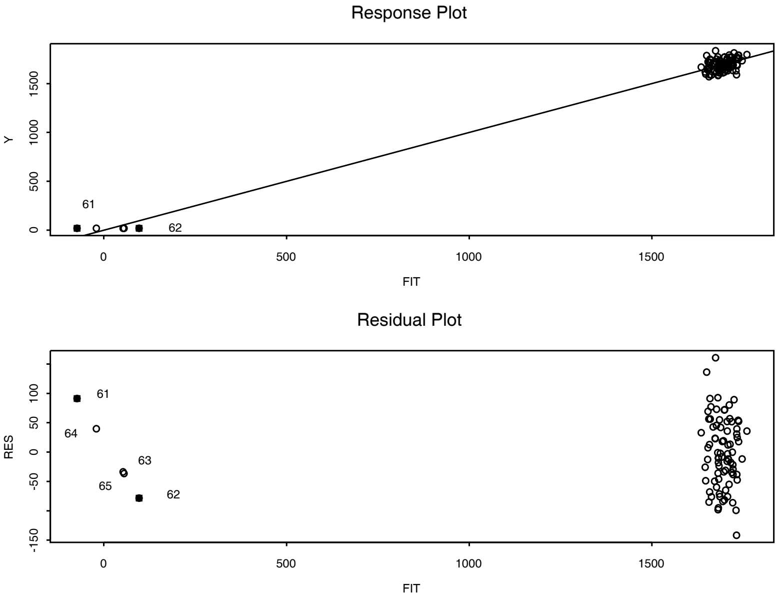

Figure 6.2: Plots for Buxton Data

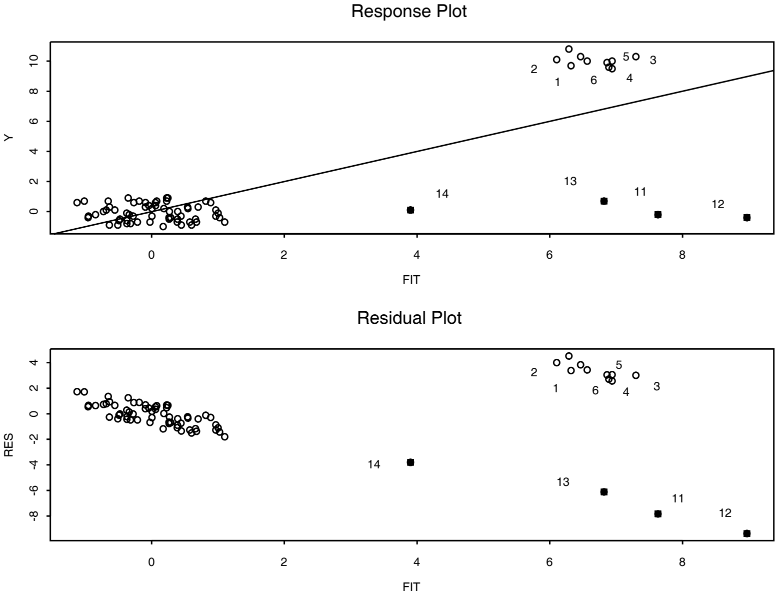

Figure 6.3: Plots for HBK Data

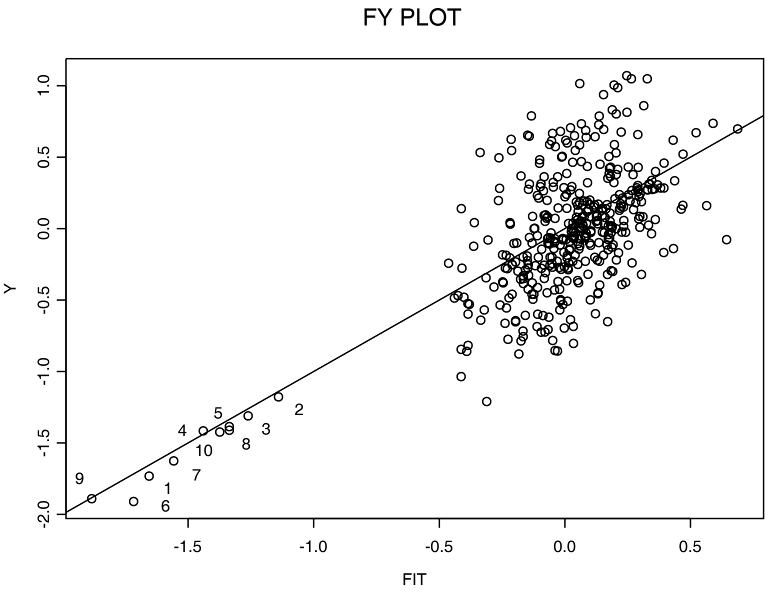

Figure 6.4: Projection Pursuit Regression, Artificial Data

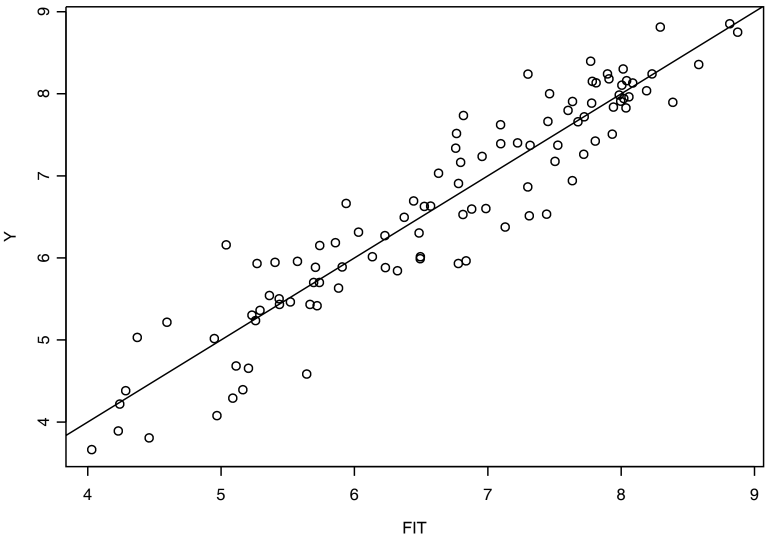

Figure 6.5: Fit–Response Plot for Log(Lynx) Data

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 11.3 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 response plot and residual plot is usually able to detect the outliers.

Example 6.6. MathSoft (1999a, p. 245-246) gives an FY plot for simulated data. In this example the simulated data set is modified by planting 10 outliers. Let x_1 and x_2 be iid uniform $U(-1, 1)$ random variables, and let $Y = x_1 x_2 + e$ where the errors e are iid $N(0, 0.04)$ random variables. The artificial data set uses 400 cases, but the first 10 cases used $Y \sim N(-1.5, 0.04)$, $x_1 \sim N(0.2, 0.04)$ and $x_2 \sim N(0.2, 0.04)$ where Y, x_1, and x_2 were independent. The model $Y = m(x_1, x_2) + e$ was fitted nonparametrically without using knowledge of the true regression relationship. The fitted values $\hat{m}$ were obtained from the *Splus* function `ppreg` for *projection pursuit regression* (Friedman and Stuetzle, 1981). The outliers are easily detected with the FY plot shown in Figure 6.4.

Example 6.7. The lynx data is a well known time series concerning the number Z_i of lynx trapped in a section of Northwest Canada from 1821 to 1934. There were $n = 114$ cases and MathSoft (1999b, p. 166-169) fits an AR(11) model $Y_i = \beta_0 + \beta_1 Y_{i-1} + \beta_2 Y_{i-2} + \cdots + \beta_{11} Y_{i-11} + e_i$ to the data where $Y_i = \log(Z_i)$ and $i = 12, 13, ..., 114$. The FY plot shown in Figure 6.5 suggests that the AR(11) model fits the data reasonably well. To compare different models or to find a better model, use an FF plot of Y and the fitted values from several competing time series models. See Problem 6.4.

6.5 Complements

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). More advanced works include Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985) and Chatterjee and Hadi (1988). Hoaglin and Welsh (1978) examines the hat matrix while Cook (1977) introduces Cook's distance.

Some other papers of interest include Barrett and Gray (1992), Gray (1985), Hadi and Simonoff (1993), Hettmansperger and Sheather (1992), Velilla (1998), and Velleman and Welsch (1981).

Hawkins and Olive (2002, p. 141, 158) suggest using the RR and FF plots. Typically RR and FF plots are used if there are several estimators for one fixed model, eg 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. The FFλ plot is an FF plot where the fitted values were obtained from competing power transformation models indexed by the power transformation parameter $\lambda \in \Lambda_c$. Variable selection uses both FF and RR plots.

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 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.

Some papers about the single index model include Aldrin, Bϕlviken, and Schweder (1993), Härdle, Hall, and Ichimura (1993), Naik and Tsai (2001), Simonoff and Tsai (2002), Stoker (1986) and Weisberg and Welsh (1994). Also see Olive (2004b). An interesting paper on the multi–index model is Hristache, Juditsky, Polzehl, and Spokoiny (2001).

The fact that the fit–response (FY) plot is 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. The FY plot is not limited to models of the form (6.6) since the plot can be made as long as fitted values $\hat{Y}$ can be obtained from the model. If $\hat{Y}_i \approx Y_i$ for $i = 1, ..., n$ then the plotted points will scatter about the identity line. Section 5.3 and Olive (2007) use the response plot to explain prediction intervals. Zheng and Agresti (2000) suggest using corr$(Y, \hat{Y})$ as a R^2 type measure.

6.6 Problems

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs** and the command *source("A:/robdata.txt")* **to download the data. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *MLRplot*, will display the code for the function. Use the `args` command, eg *args(MLRplot)*, to display the needed arguments for the function.

6.1*. 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 7) 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 7.1, but in about three plots the MBA estimator will also pass through the outliers.

If you use *R*, type the following command and include the plot in *Word.*

```
> library(MASS)
> rrplot2(buxx,buxy)
```

If you use *Splus*, type the following command and include the plot in *Word.*

```
> library(MASS)
> rrplot(buxx,buxy)
```

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.*

If you use *R*, type the following command and include the plot in *Word*.

```
> library(MASS)
> ffplot2(buxx,buxy)
```

If you use *Splus*, type the following command and include the plot in *Word*.

```
> ffplot(buxx,buxy)
```

6.2. a) If necessary, enter the two *source* commands above Problem 6.1. 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)
```

6.3. This problem makes the fit–response plot for the lynx data in Example 6.7.

a) Check that the lynx data is in *Splus* by typing the command *help(lynx)*. A window will appear if the data is available.

b) For *Splus*, enter the following *Splus* commands to produce the FY plot. Include the plot in *Word*. The command abline(0,1) adds the identity line.

```
> Y <- log(lynx)
> out <- ar.yw(Y)
> FIT <- Y - out$resid
> plot(FIT,Y)
> abline(0,1)
```

For *R*, enter the following *R* commands to produce the FY plot. Include the plot in *Word*. The command abline(0,1) adds the identity line.

```
> library(stats)
> data(lynx)
> Y <- log(lynx)
> out <- ar.yw(Y)
> Yts <- Y[12:114]
> FIT <- Yts - out$resid[12:114]
> plot(FIT,Yts)
> abline(0,1)
```

6.4*. Following Lin and Pourahmadi (1998), consider the lynx time series data and let the response $Y_t = \log_{10}(lynx)$. Moran (1953) suggested the autoregressive AR(2) model $\hat{Y}_t = 1.05 + 1.41Y_{t-1} - 0.77Y_{t-2}$. Tong (1977) suggested the AR(11) model $\hat{Y}_t = 1.13Y_{t-1} - .51Y_{t-2} + .23Y_{t-3} - 0.29Y_{t-4} + .14Y_{t-5} - 0.14Y_{t-6} + .08Y_{t-7} - .04Y_{t-8} + .13Y_{t-9} + 0.19Y_{t-10} - .31Y_{t-11}$. Brockwell and Davis (1991, p. 550) suggested the AR(12) model $\hat{Y}_t = 1.123 + 1.084Y_{t-1} - .477Y_{t-2} + .265Y_{t-3} - 0.218Y_{t-4} + .180Y_{t-9} - 0.224Y_{t-12}$. Tong (1983) suggested the following two self–exciting autoregressive models. The SETAR(2,7,2) model uses $\hat{Y}_t = .546 + 1.032Y_{t-1} - .173Y_{t-2} + .171Y_{t-3} - 0.431Y_{t-4} + .332Y_{t-5} - 0.284Y_{t-6} + .210Y_{t-7}$ if $Y_{t-2} \leq 3.116$ and $\hat{Y}_t = 2.632 + 1.492Y_{t-1} - 1.324Y_{t-2}$, otherwise. The SETAR(2,5,2) model uses $\hat{Y}_t = .768 + 1.064Y_{t-1} - .200Y_{t-2} + .164Y_{t-3} - 0.428Y_{t-4} + .181Y_{t-5}$ if $Y_{t-2} \leq 3.05$ and $\hat{Y}_t = 2.254 + 1.474Y_{t-1} - 1.202Y_{t-2}$, otherwise. The FF plot of the fitted values and the response can be used to compare the models. Type the *rpack* command `fflynx()` (in *R*, 1st type `library(stats)` and `data(lynx)`).

a) Include the resulting plot and correlation matrix in *Word*.

b) Which model seems to be best? Explain briefly.

c) Which two pairs of models gave very similar fitted values?

6.5. This problem may not work in R. Type *help(ppreg)* to make sure that *Splus* has the function `ppreg`. Then make the FY plot for Example 6.6 with the following commands. Include the plot in *Word*.

```
> set.seed(14)
> x1 <- runif(400,-1,1)
> x2 <- runif(400,-1,1)
> eps <- rnorm(400,0,.2)
> Y <- x1*x2 + eps
> x <- cbind(x1,x2)
> x[1:10,] <- rnorm(20,.2,.2)
> Y[1:10] <- rnorm(10,-1.5,.2)
> out <- ppreg(x,Y,2,3)
> FIT <- out$ypred
> plot(FIT,Y)
> abline(0,1)
```

Arc problems

Warning: The following problem uses data from the book's webpage. Save the data files on a disk. Get in *Arc* and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:).* Then click twice on the data set name.

Using material learned in Chapters 5–6, analyze the data sets described in Problems 6.6–6.16. Assume that the response variable $Y = t(Z)$ and that the predictor variable $X_2, ..., X_p$ are functions of remaining variables $W_2, ..., W_r$. Unless told otherwise, the full model $Y, X_1, X_2, ..., X_p$ (where $X_1 \equiv 1$) should use functions of every variable $W_2, ..., W_r$ (and often $p = r + 1$). (In practice, often some of the variables and some of the cases are deleted, but we will use all variables and cases, unless told otherwise, primarily so that the instructor has some hope of grading the problems in a reasonable amount of time.) See pages 176–180 for useful tips for building a full model. **Read the description of the data** provided by *Arc*. Once you have a good full model, perform forward selection and backward elimination. Find the model that minimizes $C_p(I)$ and find the smallest value of k such that $C_p(I) \leq 2k$. The minimum C_p model often has too many terms while the 2nd model sometimes has too few terms.

a) Give the output for your full model, including $Y = t(Z)$ and R^2. If it is not obvious from the output what your full model is, then write down the full model. Include a response plot for the full model. (This plot should be linear).

b) Give the output for your final submodel. If it is not obvious from the output what your submodel is, then write down the final submodel.

c) Give between 3 and 5 plots that justify that your multiple linear regression submodel is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

6.6. For the file *bodfat.lsp*, described in Example 1.4, use $Z = Y$ but do not use X_1 as a predictor in the full model. Do parts a), b) and c) above.

6.7*. For the file *boston2.lsp*, described in Examples 1.6, 12.6 and 12.7 use $Z = (y =)$ CRIM. Do parts a), b) and c) above Problem 6.6.

Note: $Y = \log(CRIM), X_4, X_8$, is an interesting submodel, but more predictors are probably needed.

6.8*. For the file *major.lsp*, described in Example 6.4, use $Z = Y$. Do parts a), b) and c) above Problem 6.6.

Note: there are 1 or more outliers that affect numerical methods of variable selection.

6.9. For the file *marry.lsp*, described below, use $Z = Y$. This data set comes from Hebbler (1847). The census takers were not always willing to count a woman's husband if he was not at home. Do not use the predictor X_2 in the full model. Do parts a), b) and c) above Problem 6.6.

6.10*. For the file *museum.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 6.6.

This data set consists of measurements taken on skulls at a museum and was extracted from tables in Schaaffhausen (1878). There are at least three groups of data: humans, chimpanzees and gorillas. The OLS fit obtained from the humans passes right through the chimpanzees. Since *Arc* numbers cases starting at 0, cases 47–59 are apes. These cases can be deleted by highlighting the cases with small values of Y in the scatterplot matrix and using the *case deletions* menu. (You may need to maximize the window containing the scatterplot matrix in order to see this menu.)

i) Try variable selection using all of the data.

ii) Try variable selection without the apes.

If all of the cases are used, perhaps only X_1, X_2 and X_3 should be used in the full model. Note that $\sqrt{Y}$ and X_2 have high correlation.

6.11*. For the file *pop.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 6.6.

This data set comes from Ashworth (1842). Try transforming all variables to logs. Then the added variable plots show two outliers. Delete these two cases. Notice the effect of these two outliers on the p–values for the coefficients and on numerical methods for variable selection.

Note: then $\log(Y)$ and $\log(X_2)$ make a good submodel.

6.12*. For the file *pov.lsp*, described below, use i) $Z = flife$ and ii) $Z = gnp2 = gnp + 2$. This dataset comes from Rouncefield (1995). Making *loc* into a factor may be a good idea. Use the commands *poverty>Make factors* and select the variable *loc*. For ii), try transforming to logs and deleting the 6 cases with $gnp2 = 0$. (These cases had missing values for *gnp*. The file *povc.lsp* has these cases deleted.) Try your final submodel on the data that includes the 6 cases with $gnp2 = 0$. Do parts a), b) and c) above Problem 6.6.

6.13*. For the file *skeleton.lsp*, described below, use $Z = y$.

This data set is also from Schaaffhausen (1878). At one time I heard or read a conversation between a criminal forensics expert with his date. It went roughly like "If you wound up dead and I found your femur, I could tell what your height was to within an inch." Two things immediately occurred to me. The first was "no way" and the second was that the man must not get many dates! The files *cyp.lsp* and *major.lsp* have measurements including *height*, but their $R^2 \approx 0.9$. The skeleton data set has at least four groups: stillborn babies, newborns and children, older humans and apes.

a) Take logs of each variable and fit the regression on log(Y) on log(X_1), ..., log(X_{13}). Make a residual plot and highlight the case with the with the smallest residual. From the *Case deletions* menu, select *Delete selection from data set.* Go to *Graph&Fit* and again fit the regression on log(Y) on log(X_1), ..., log(X_{13}) (you should only need to click on *OK*). The output should say that case 37 has been deleted. Include this output for the full model in *Word*.

b) Do part b) above Problem 6.6.

c) Do part c) above Problem 6.6.

6.14. Activate *big-mac.lsp* in *Arc.* Assume that a multiple linear regression model holds for $t(y)$ and some terms (functions of the predictors) where y is BigMac = hours of labor to buy Big Mac and fries. Using techniques you have learned in class find such a model. (Hint: Recall from Problem 5.27 that transforming all variables to logs and then using the model constant, log(service), log(TeachSal) and log(TeachTax) was ok but the residuals did not look good. Try adding a few terms from the minimal C_p model.)

a) Write down the full model that you use (eg a very poor full model is $\exp(BigMac) = \beta_1 + \beta_2 \exp(EngSal) + \beta_3 (TeachSal)^3 + e$) and include a response plot for the full model. (This plot should be linear). Give R^2 for the full model.

b) Write down your final model (eg a very poor final model is $\exp(BigMac) = \beta_1 + \beta_2 \exp(EngSal) + \beta_3 (TeachSal)^3 + e$).

c) Include the least squares output for your model and between 3 and 5 plots that justify that your multiple linear regression model is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

6.15. This is like Problem 6.14 with the BigMac data. Assume that a multiple linear regression model holds for $t(Y)$ and for some terms (usually powers or logs of the predictors). Using the techniques learned in class, find such a model. Give output for the full model, output for the final submodel and use several plots to justify your choices. These data sets, as well as the BigMac data set, come with *Arc*. See Cook and Weisberg (1999a). (**INSTRUCTOR: Allow 2 hours for each part.**)

```
            file          response Y
a)      allomet.lsp       BRAIN
b)      casuarin.lsp      W
c)      evaporat.lsp     Evap
d)      hald.lsp          Y
e)      haystack.lsp     Vol

f)      highway.lsp      rate
(from the menu Highway, select ‘‘Add a variate" and type
 sigsp1 = sigs + 1. Then you can transform sigsp1.)
g)      landrent.lsp      Y
h)      ozone.lsp        ozone
i)      paddle.lsp       Weight
j)      sniffer.lsp       Y
k)      water.lsp         Y
```

i) Write down the full model that you use and include the full model residual plot and response plot in *Word*. Give R^2 for the full model.

ii) Write down the final submodel that you use.

iii) Include the least squares output for your model and between 3 and 5 plots that justify that your multiple linear regression model is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

6.16*. a) Activate *buxton.lsp* (you need to download the file onto your disk *Floppy 3 1/2 A:*). From the "Graph&Fit" menu, select "Fit linear LS." Use *height* as the response variable and *bigonal breadth*, *cephalic index*, *head length* and *nasal height* as the predictors. Include the output in *Word.*

b) Make a response plot (L1:Fit-Values in H and height in V) and residual plot (L1:Fit-Values in H and L1:Residuals in V) and include both plots in *Word.*

c) In the residual plot use the mouse to move the curser just above and to the left of the outliers. Hold the leftmost mouse button down and move the mouse to the right and then down. This will make a box on the residual plot that contains the outliers. Go to the "Case deletions menu" and click on *Delete selection from data set.* From the "Graph&Fit" menu, select "Fit linear LS" and fit the same model as in a) (the model should already be entered, just click on "OK"). Include the output in *Word.*

d) Make a response plot (L2:Fit-Values in H and height in V) and residual plot (L2:Fit-Values in H and L2:Residuals in V) and include both plots in *Word.*

e) Explain why the outliers make the MLR relationship seem much stronger than it actually is. (Hint: look at R^2.)