Chapter 3

Building an MLR Model

Building a multiple linear regression (MLR) model from data is one of the most challenging regression problems. The "final full model" will have response variable $Y = t(Z)$, a constant x_1 and predictor variables $x_2 = t_2(w_2, ..., w_r), ..., x_p = t_p(w_2, ..., w_r)$ where the initial data consists of $Z, w_2, ..., w_r$. Choosing $t, t_2, ..., t_p$ so that the final full model is a useful MLR approximation to the data can be difficult.

Model building is an *iterative process*. Given the problem and data but no model, the model building process can often be aided by graphs that help visualize the relationships between the different variables in the data. Then a statistical model can be proposed. This model can be fit and inference performed. Then *diagnostics* from the fit can be used to check the assumptions of the model. If the assumptions are not met, then an alternative model can be selected. The fit from the new model is obtained, and the cycle is repeated. This chapter provides some tools for building a good full model.

Warning: Researchers often have a single data set and tend to expect statistics to provide far more information from the single data set than is reasonable. MLR is an extremely useful tool, but MLR is at its best when the final full model is known before collecting and examining the data. But it is very common for researchers to build their final full model by using the iterative process until the final model "fits the data well." Researchers should not expect that all or even many of their research questions can be answered from such a full model. If the final MLR full model is built from a single data set in order to fit that data set well, then typically inference from that model **will not be valid**. The model may be useful for describing

the data, but may perform very poorly for prediction of a future response. The model may suggest that some predictors are much more important than others, but a model that is chosen prior to collecting and examining the data is generally much more useful for prediction and inference. **A single data set is a great place to start an analysis, but can be a terrible way to end the analysis.**

Often a final full model is built after collecting and examining the data. This procedure is called "data snooping," and such models can not be expected to be reliable. If possible, spend about 1/8 of the budget to collect data and build an initial MLR model. Spend another 1/8 of the budget to collect more data to check the initial MLR model. If changes are necessary, continue this process until no changes from the previous step are needed, resulting in a tentative MLR model. Then spend between 3/4 and 1/2 of the budget to collect data assuming that the tentative model will be useful.

After obtaining a final full model, researchers will typically find a final submodel after performing variable selection. Even if the final full model was selected before collecting data, the final submodel, obtained after performing variable selection, may not be useful for inference.

Rule of thumb 3.1. If the MLR model is built using the variable selection methods from Section 3.4, then the final submodel can be used for description but will often not be useful for inference and prediction.

3.1 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. Power transformations are particularly effective, and the techniques of this section are often useful for general regression problems, not just for multiple linear regression. 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

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{3.1}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder", eg from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, eg 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 ± 2 and ± 3 can always be added.

Definition 3.1. 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 response.

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 software *Arc* gives two numbers, the minimum and maximum of the variable, along with the name of the variable. The software *R/Splus* labels the values of each variable in two places, see Example 3.2 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.

Rule of thumb 3.2. 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 λ *smaller*.
To spread *large* values of a variable, make λ *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^λ 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 labelling. Certainly if $Y|x = 9 \sim N(0, 1)$, then $Y|\sqrt{x} = 3 \sim N(0, 1)$. To see that Rule of thumb 3.2a 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 β_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 λ_i smaller. To spread large values of the variable, make λ_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 variables need spreading. Figures 1.8 and 10.4 b), 11.1 b) and 15.11 a) have this shape.

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 3.1. Examine Figure 3.1. 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

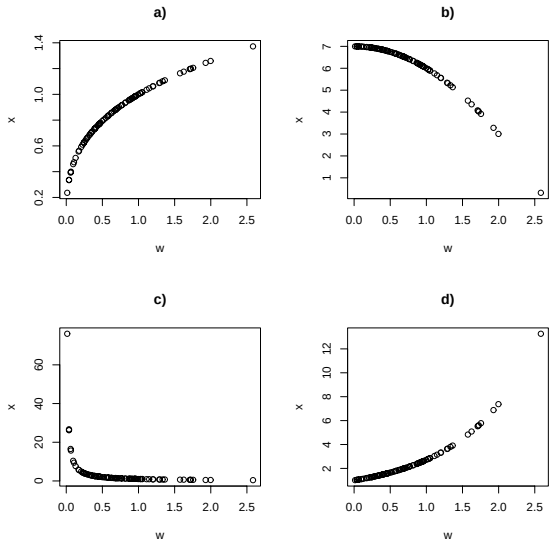

Figure 3.1: Plots to Illustrate the Bulging and Ladder Rules

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 small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 3.1a, 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 3.1b, 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 3.1c, 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 3.1d, small values of x need spreading. Notice that the plotted points bulge down towards large values of the horizontal variable.

Example 3.2: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand.

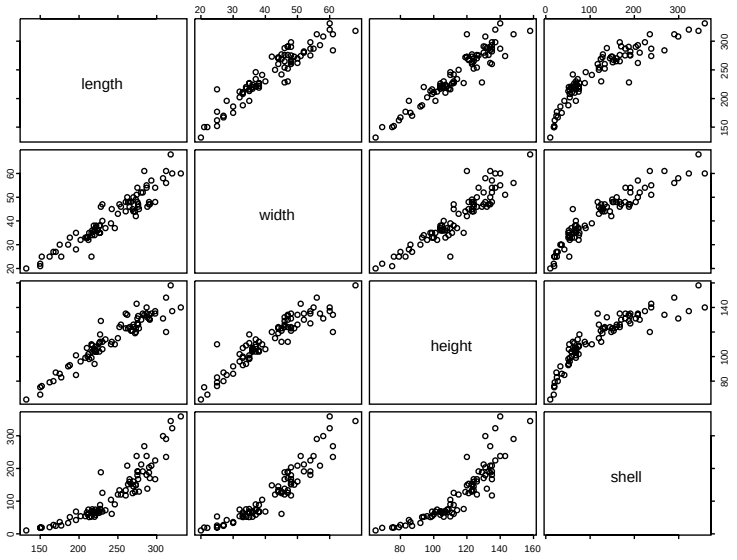

Figure 3.2: Scatterplot Matrix for Original Mussel Data Predictors

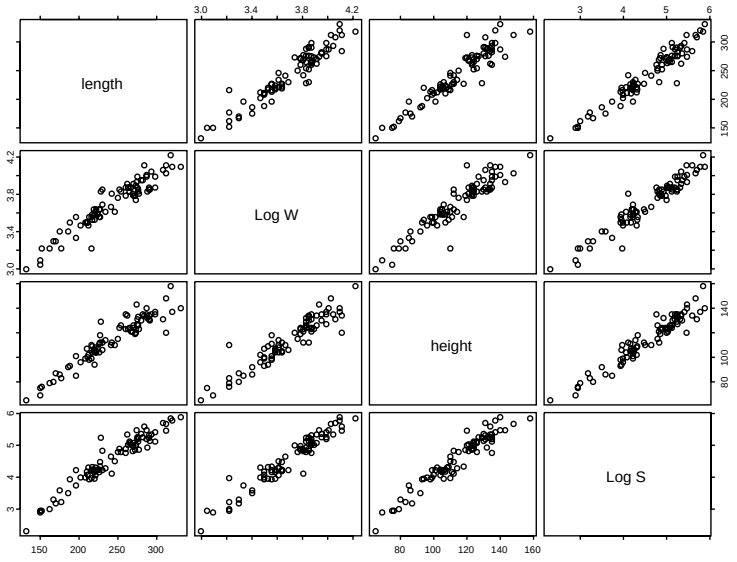

Figure 3.3: Scatterplot Matrix for Transformed Mussel Data Predictors

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 3.2 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 (made with R), 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 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.

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 λ from 1 and we tried $\log(W)$. Figure 3.3 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 3.2. Notice that the plot of W versus L and the plot of $\log(W)$ versus L both appear linear.

The plot of *shell* versus *height* in Figure 3.2 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.

3.2 Graphical Methods 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 λ_o,

such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i. \tag{3.2}$$

If λ_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 λ_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 3.2. 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 3.3. Assume that **all** of the values of the response variable Y_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{3.3}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as Λ_L. This family is a special case of the response transformations considered by Tukey (1957).

A graphical method for response transformations computes the "fitted values" $\hat{W}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_\lambda$ from the multiple linear regression model using $W_i = t_\lambda(Z_i)$ as the "response." Then a "response plot" of the $\hat{W}$ versus W is made for each of the seven values of $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the iid error MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$.

By adding the "response" Z to the scatterplot matrix, the methods of the previous section can also be used to suggest good values of λ, and it is usually a good idea to use predictor transformations to remove nonlinearities from the predictors before selecting a response transformation. 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. Recall our convention 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 3.2 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 *Arc* data set `evaporat.lsp`, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

Definition 3.4. 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 Λ_L, then sometimes $\hat{\lambda}_n$ will converge (eg 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 Λ_L. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and ± 3. Powers from numerical methods can also be added.

Application 3.1. 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 λ^*, 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 λ_o by adding $\hat{\lambda}$ to Λ_L.)

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 λ in decreasing order of importance are $1, 0, 1/2, -1$ and $1/3$. So the log transformation

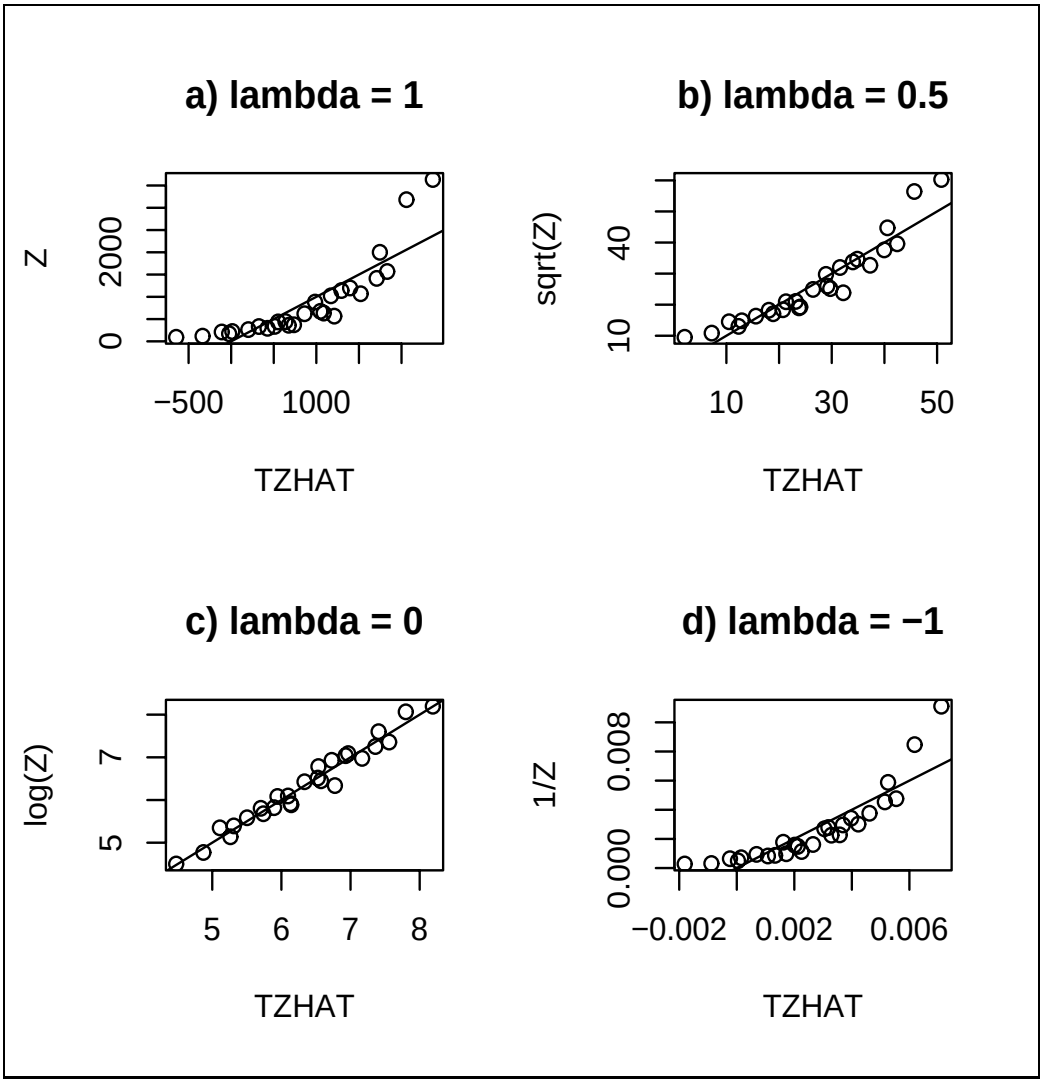

Figure 3.4: Four Transformation Plots for the Textile Data

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. In the following example, 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 3.3: 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 predictors *length, amplitude* and *load.* Using the normal profile log likelihood for λ_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 3.4 are transformation plots of $\hat{Z}$ versus Z^λ for four values of λ 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 λ causes all points along the curvilinear scatter in Figure 3.4a to form along a linear scatter in Figure 3.4c. Dynamic plotting using λ 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 3.4a shows that a response transformation is needed since the plotted points follow a nonlinear curve while Figure 3.4c 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 3.4a 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

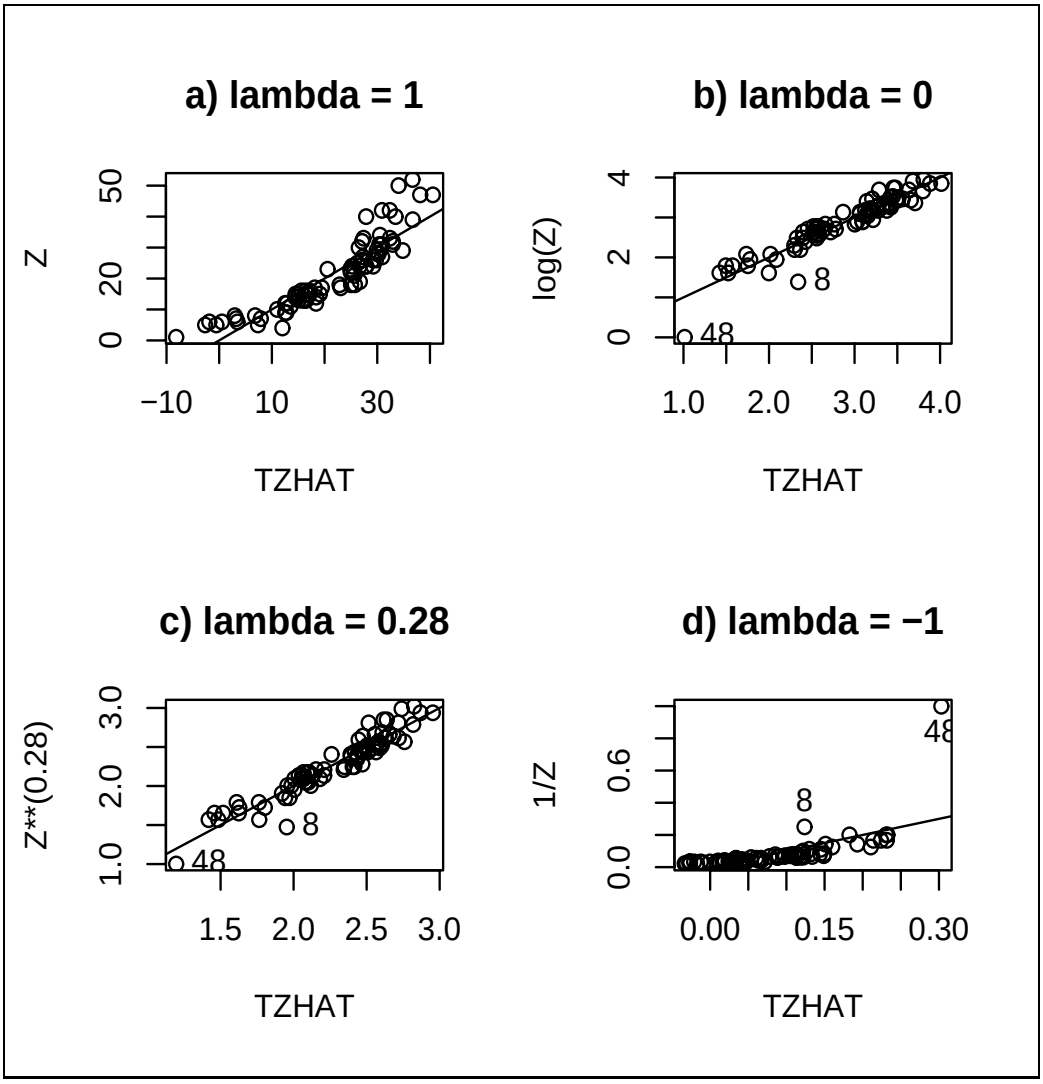

Figure 3.5: Transformation Plots for the Mussel Data

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 3.4: Mussel Data. Consider the mussel data of Example 3.2 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 3.5 are transformation plots for four values of λ. 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 3.5c, 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 λ_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 3.4 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 λ_o.

Example 3.5: 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 3.2 shows the scatterplot matrix of the predictors L, H, W and S. Again nonlinearity is present. Figure

3.3 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 3.4.

3.3 Main Effects, Interactions and Indicators

Section 1.7 explains interactions, factors and indicator variables in an abstract setting when $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x}^T\boldsymbol{\beta}$ is the sufficient predictor (SP). MLR is such a model. The interpretations given Section 1.7 in terms of the SP can be given in terms of $E(Y|\boldsymbol{x})$ for MLR since $E(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} = SP$ for MLR.

Definition 3.5. Suppose that the explanatory variables have the form $x_2, ..., x_k$, $x_{jj} = x_j^2$, $x_{ij} = x_i x_j$, $x_{234} = x_2 x_3 x_4$, et cetera. Then the variables $x_2, ..., x_k$ are *main effects*. A product of two or more different main effects is an *interaction*. A variable such as x_2^2 or x_7^3 is a *power*. An $x_2 x_3$ interaction will sometimes also be denoted as $x_2 : x_3$ or $x_2 * x_3$.

Definition 3.6. A *factor* W is a qualitative random variable. Suppose W has c categories $a_1, ..., a_c$. Then the factor is incorporated into the MLR model by using $c - 1$ indicator variables $x_{Wi} = 1$ if $W = a_i$ and $x_{Wi} = 0$ otherwise, where one of the levels a_i is omitted, eg, use $i = 1, ..., c - 1$. Each indicator variable has 1 degree of freedom. Hence the degrees of freedom of the $c - 1$ indicator variables associated with the factor is $c - 1$.

Rule of thumb 3.3. Suppose that the MLR model contains at least one power or interaction. Then the corresponding main effects that make up the powers and interactions should also be in the MLR model.

Rule of thumb 3.3 suggests that if x_3^2 and $x_2 x_7 x_9$ are in the MLR model, then x_2, x_3, x_7 and x_9 should also be in the MLR model. A quick way to check whether a term like x_3^2 is needed in the model is to fit the main effects models and then make a scatterplot matrix of the predictors and the residuals, where the residuals are on the top row. Then the top row shows plots of x_k versus r, and if a plot is parabolic, then x_k^2 should be added to the model. Potential predictors w_j could also be added to the scatterplot matrix. If the plot of w_j versus r shows a positive or negative linear trend add w_j to the model. If the plot is quadratic, add w_j and w_j^2 to the model. This technique is for quantitative variables x_k and w_j.

The simplest interaction to interpret is the interaction between a quantitative variable x_2 and a qualitative variable x_3 with 2 levels. Suppose that $x_3 = 1$ for level a_2 and $x_3 = 0$ for level a_1. Then a first order model with interaction is $SP = E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3$. This model yields two unrelated lines in the conditional expectation depending on the value of x_3: $E(Y|\boldsymbol{x}) = \beta_1 + \beta_3 + (\beta_2 + \beta_4)x_2$ if $x_3 = 1$ and $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ if $x_3 = 0$. If $\beta_4 = 0$, then there are two parallel lines: $E(Y|\boldsymbol{x}) = \beta_1 + \beta_3 + \beta_2 x_2$ if $x_2 = 1$ and $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ if $x_3 = 0$. If $\beta_3 = \beta_4 = 0$, then the two lines are coincident: $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ for both values of x_3. If $\beta_3 = 0$, then the two lines have the same intercept: $E(Y|\boldsymbol{x}) = \beta_1 + (\beta_2 + \beta_4)x_2$ if $x_3 = 1$ and $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ if $x_3 = 0$.

Notice that $\beta_4 = 0$ corresponds to no interaction. The estimated slopes of the two lines will not be exactly identical, so the two estimated lines will not be parallel even if there is no interaction. If the two estimated lines have similar slopes and do not cross, there is evidence of no interaction, while crossing lines is evidence of interaction provided that the two lines are not nearly coincident. Two lines with very different slopes also suggests interaction. In general, as factors have more levels and interactions have more terms, eg $x_2 x_3 x_4 x_5$, the interpretation of the model rapidly becomes very complex.

Example 3.6. Two varieties of cement that replace sand with coal waste products were compared to a standard cement mix. The response Y was the compressive strength of the cement measured after 7, 28, 60, 90 or 180 days of *curing time* $= x_2$. This cement was intended for sidewalks and barriers but not for construction. The data is likely from small batches of cement prepared in the lab, and is likely correlated; however, MLR can be used for exploratory and descriptive purposes. Actually using the different cement mixtures in the field (eg as sidewalks), would be very expensive. The factor *mixture* had 3 levels, 2 for the standard cement and 0 and 1 for the coal based cements. A plot of x_2 versus Y (not shown but see Problem 3.15), resembled the left half of a quadratic $Y = -c(x_2 - 180)^2$. Hence x_2 and x_2^2 were added to the model.

Figure 3.6 shows the response plot and residual plots from this model. The standard cement mix uses the symbol + while the coal based mixes use an inverted triangle and square. OLS lines based on each mix are added as visual aids. The lines from the two coal based mixes do not intersect, suggesting that there may not be an interaction between these two mixes.

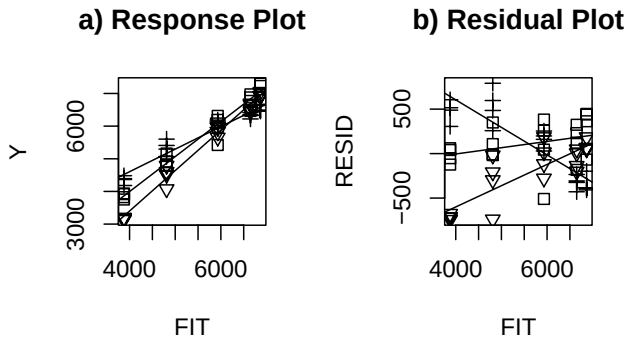

Figure 3.6: Plots to Illustrate Interaction for the Cement Data

There is an interaction between the standard mix and the two coal mixes since these lines do intersect. All three types of cement become stronger with time, but the standard mix has the greatest strength at early curing times while the coal based cements become stronger than the standard mix at the later times. Notice that the interaction is more apparent in the residual plot. Problem 3.15 adds a factor Fx_3 based on mix as well as the $x_2 * Fx_3$ and $x_2^2 * Fx_3$ interactions. The resulting model is an improvement, but there is still some curvature in the residual plot, and one case is not fit very well.

3.4 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e \qquad (3.4)$$

where e is an error, Y is the response variable, $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is a $k_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - k_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of k terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + \boldsymbol{x}_O^T \boldsymbol{\beta}_O + e. \tag{3.5}$$

Definition 3.7. The model $Y = \boldsymbol{x}^T \boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The **full model is always a submodel**. The *sufficient predictor* (SP) is the linear combination of the predictor variables used in the model. Hence the full model has $SP = \boldsymbol{x}^T \boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$.

The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ and the following remarks suggest that *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$. Suppose that S is a subset of I and that model (3.4) holds. Then

$$SP = \boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T \boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T \boldsymbol{0} = \boldsymbol{x}_I^T \boldsymbol{\beta}_I \tag{3.6}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$.

All too often, variable selection is performed and then the researcher tries to use the final submodel for inference as if the model was selected before gathering data. At the other extreme, it could be suggested that variable selection should not be done because inferences after variable selection are not valid. Neither of these two extremes is useful.

Ideally the model is known before collecting the data. After the data is collected, the MLR assumptions are checked and then the model is used for inference. Alternatively, a preliminary study can be used to collect data. Then the predictors and response can be transformed until a full model is built that seems to be a useful MLR approximation of the data. Then variable selection can be performed, suggesting a final model. Then this final model is the known model used before collecting data for the main part of the study.

In practice, the researcher often has one data set, builds the full model and performs variable selection to obtain a final submodel. In other words, an extreme amount of data snooping was used to build the final model. A major problem with the final MLR model (chosen after variable selection or data snooping) is that it is not valid for inference in that the p-values for the OLS t-tests and ANOVA F test are likely to be too small, while the p-value for the partial F test that uses the final model as the reduced model is likely to be too high. Similarly, the actual coverage of the nominal $100(1-\delta)\%$ prediction intervals tends to be too small and unknown (eg the nominal 95% PIs may only contain 83% of the future responses Y_f). Thus the model is likely to fit the data set from which it was built much better than future observations. Call the data set from which the MLR model was built the "training data," consisting of cases $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Then the future predictions tend to be poor in that $|Y_f - \hat{Y}_f|$ tends to be larger on average than $|Y_i - \hat{Y}_i|$. To summarize, a final MLR model selected after variable selection can be useful for description and exploratory analysis: the tests and intervals can be used for exploratory purposes, but are not valid for inference.

Generally the research paper should state that the model was built with one data set, and is useful for description and exploratory purposes, but should not be used for inference. The research paper should only suggest that the model is useful for inference if the model has been shown to be useful **on data collected after the model was built**. For example, if the researcher can collect new data and show that the model produces valid inferences (eg 97 out of 100 95% prediction intervals contained the future response Y_f), then the researcher can perhaps claim to have found a model that is useful for inference.

Other problems exist even if the full MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ is good. Let $I \subset \{1, ..., p\}$ and let $\boldsymbol{x}_I$ be the final vector of predictors. If $\boldsymbol{x}_I$ is missing important predictors contained in the full model, sometimes called *underfitting*, then the final model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ may be a very poor approximation to the data, in particular the full model may be linear while the final model may be nonlinear. Similarly the full model may satisfy $V(e_i) = \sigma^2$ while the constant variance assumption is violated by the submodel: $V(e_i) = \sigma_i^2$. These two problems are less severe if the joint distribution of $(Y, \boldsymbol{x}^T)^T$ is multivariate normal, since then $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ satisfies the constant variance MLR model regardless of the subset I used.

In spite of these problems, if the researcher has a single data set with many predictors, then usually variable selection must be done. Let $p-1$ be the number of nontrivial predictors and assume that the model also contains a constant. Also assume that $n > 10p$. If the MLR model found after variable selection has good response and residual plots, then the model may be very useful for descriptive and exploratory purposes.

Simpler models are easier to explain and use than more complicated models, and there are several other important reasons to perform variable selection. First, an MLR model with unnecessary predictors has a mean square error for prediction that is too large. Let $\boldsymbol{x}_S$ contain the necessary predictors, let $\boldsymbol{x}$ be the full model, and let $\boldsymbol{x}_I$ be a submodel. If (3.4) holds and $S \subseteq I$, then $E(Y|\boldsymbol{x}_I) = \boldsymbol{x}_I^T\boldsymbol{\beta}_I = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}^T\boldsymbol{\beta}$. Hence OLS applied to Y and $\boldsymbol{x}_I$ yields an unbiased estimator $\hat{\boldsymbol{\beta}}_I$ of $\boldsymbol{\beta}_I$. If (3.4) holds, $S \subseteq I$, $\boldsymbol{\beta}_S$ is a $k \times 1$ vector and $\boldsymbol{\beta}_I$ is a $j \times 1$ vector with $j > k$, then it is shown in Chapter 13 that

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 k}{n} = \frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Si}). \tag{3.7}$$

In particular, the full model has $j = p$. Hence having unnecessary predictors decreases the precision for prediction. Fitting unnecessary predictors is sometimes called *fitting noise* or *overfitting*. As an extreme case, suppose that the full model contains $p = n$ predictors, including a constant, so that the hat matrix $\boldsymbol{H} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. Then $\hat{Y} = Y$ so that $\mathrm{VAR}(\hat{Y}|\boldsymbol{x}) = \mathrm{VAR}(Y)$.

Secondly, often researchers are interested in examining the effects of certain predictors on the response. Recall that $\hat{\boldsymbol{\beta}}_i$ measures the effect of x_i given that all of the other predictors $x_1, ..., x_{i-1}, x_{i+1}, ..., x_p$ are in the model. If some of the predictors are highly correlated, then these predictors may not be needed in the MLR model given that the other predictors are in the model. Hence it will not be possible to examine the effects of these predictors on the response unless the MLR model is changed.

Thirdly, there may be an extremely expensive predictor x_p that researchers would like to omit. If x_p is not needed in the MLR model given that $x_1, ..., x_{p-1}$ are in the model, then x_p can be removed from the model, saving money.

A major assumption before performing variable selection is that the full model is good. A factor with c levels can be incorporated into the full

model by creating $c-1$ indicator variables. Sometimes the categories can be combined into fewer categories. For example, if the factor is race with levels white, black and other, new levels white and nonwhite may be useful for some data sets. Two rules of thumb are useful for building a full model. Notice that Rule of thumb 3.4 uses data snooping. Hence the full model and the submodels chosen after variable selection can be used for description and exploratory analysis, but should not be used for inference.

Rule of thumb 3.4. Remove strong nonlinearities from the predictors by making scatterplot matrices of the predictors and the response. If necessary, transform the predictors and the response using methods from Sections 3.1 and 3.2. Do not transform indicator variables. Each scatterplot matrix should contain the response entered as the last variable. Do not use more than 10 variables per scatterplot matrix. Hence if there are 90 predictor variables, make 10 scatterplot matrices. The first will contain $x_1, ..., x_9, Y$ and the last will contain $x_{81}, ..., x_{90}, Y$.

Often a variable x_i does not need to be transformed if the transformation does not increase the linearity of the plot of x_i versus Y. If the plot of x_i versus x_j is nonlinear for some x_j, try to transform one or both of x_i and x_j in order to remove the nonlinearity, but be careful that the transformation do not cause a nonlinearity to appear in the plots of x_i and x_j versus Y.

Rule of thumb 3.5. Let $x_{w1}, ..., x_{w,c-1}$ correspond to the indicator variables of a factor W. Either include all of the indicator variables in the model or exclude all of the indicator variables from the model. If the model contains powers or interactions, also include all main effects in the model (see Section 3.3).

Next we suggest methods for finding a good submodel. We make the simplifying assumptions that the full model is good, that all predictors have the same cost, that each submodel contains a constant and that there is no theory requiring that a particular predictor must be in the model. Also assume that $n \geq 5p$ and that the response and residual plots of the full model are good. Rule of thumb 3.5 should be used for the full model and for all submodels.

The basic idea is to obtain fitted values from the full model and the candidate submodel. If the candidate model is good, then the plotted points in a plot of the submodel fitted values versus the full model fitted values

should follow the identity line. In addition, a similar plot should be made using the residuals.

A problem with this idea is how to select the candidate submodel from the nearly 2^p potential submodels. One possibility would be to try to order the predictors in importance, say $x_1, ..., x_p$. Then let the kth model contain the predictors $x_1, x_2, ..., x_k$ for $k = 1, ..., p$. If the predicted values from the submodel are highly correlated with the predicted values from the full model, then the submodel is "good." All subsets selection, forward selection and backward elimination can be used (see Section 1.6), but criteria to separate good submodels from bad are needed.

Two important summaries for submodel I are $R^2(I)$, the proportion of the variability of Y explained by the nontrivial predictors in the model, and $MSE(I) = \hat{\sigma}_I^2$, the estimated error variance. See Definitions 2.15 and 2.16. Suppose that model I contains k predictors, including a constant. Since adding predictors does not decrease R^2, the adjusted $R_A^2(I)$ is often used, where

$$R_A^2(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

See Seber and Lee (2003, p. 400-401). Hence the model with the maximum $R_A^2(I)$ is also the model with the minimum $MSE(I)$.

For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p - k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k) - (n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}[\frac{SSE(I)}{SSE} - 1]$$

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. An extremely important criterion for variable selection is the C_p criterion.

Definition 3.7.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

From Section 1.6, recall that all subsets selection, forward selection and backward elimination produce one or more submodels of interest for $k =$

$2, ..., p$ where the submodel contains k predictors including a constant. The following proposition helps explain why C_p is a useful criterion and suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Olive and Hawkins (2005) show that this interpretation of C_p can be generalized to 1D regression models such as generalized linear models. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T \hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T \hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

Proposition 3.1. Suppose that a numerical variable selection method suggests several submodels with k predictors, including a constant, where $2 \leq k \leq p$.

a) The model I that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\mathrm{corr(r, r_I)} \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

c) As $\mathrm{corr}(r, r_I) \to 1$,

$$\mathrm{corr}(\boldsymbol{x}^{\mathrm{T}} \hat{\boldsymbol{\beta}}, \boldsymbol{x}_{\mathrm{I}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) = \mathrm{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Proof. These results are a corollary of Proposition 3.2 below. QED

Remark 3.1. Consider the model I_i that deletes the predictor x_i. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using Definition 3.7 and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor x_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases. The literature suggests using the $C_p(I) \leq k$ screen, but this screen tends to overfit: too many unimportant predictors are included in the model.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (ie, say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

Six graphs will be used to compare the full model and the candidate submodel. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$.

Definition 3.8. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the $C_p(I)$ criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. We will use the FF plot, RR plot, the response plots from the full and submodel, and the residual plots (of the fitted values versus the residuals) from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (3.4) holds and that a good estimator (such as OLS) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

For these plots to be useful, it is crucial to verify that a multiple linear regression (MLR) model is appropriate for the full model. **Both the response plot and the residual plot for the full model need to be used to check this assumption**. The plotted points in the response plot should cluster about the *identity line* (that passes through the origin with unit slope) while the plotted points in the residual plot should cluster about the horizontal axis (the line $r = 0$). Any nonlinear patterns or outliers in either plot suggests that an MLR relationship does not hold. Similarly, before accepting the candidate model, use the response plot and the residual plot from the candidate model to verify that an MLR relationship holds for

the response Y and the predictors $\boldsymbol{x}_I$. If the submodel is good, then the residual and response plots of the submodel should be nearly identical to the corresponding plots of the full model. Assume that all submodels contain a constant.

Application 3.2. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should nearly coincide near the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively.

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following proposition shows that the plotted points in the FF, RR and response plots will cluster about the identity line. Notice that the proposition is a property of OLS and holds even if the data does not follow an MLR model. Let $\text{corr}(x, y)$ denote the correlation between x and y.

Proposition 3.2. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.
Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.
ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.
FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity

line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.
iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\mathrm{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.
RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\mathrm{corr}(r, r_I)]^2$ and

$$\mathrm{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\mathrm{corr}(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\mathrm{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\mathrm{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i} \hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H}\boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\mathrm{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T\boldsymbol{r}_I/\boldsymbol{r}^T\boldsymbol{r}$. Since $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H}_I) = \boldsymbol{I}-\boldsymbol{H}$, the numerator $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{r}^T\boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I)+n-2k}} = \sqrt{\frac{n-p}{(p-k)F_I+n-p}}. \quad QED$$

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

Remark 3.2. Daniel and Wood (1980, p. 85) suggest using Mallows' graphical method for screening subsets by plotting k versus $C_p(I)$ for models close to or under the $C_p = k$ line. Proposition 3.2 vi) implies that if $C_p(I) \leq k$ or $F_I < 1$, then $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models I that satisfy the $C_p(I) \leq k$ screen will contain the true model S with high probability when n is large. This result does not guarantee that the true model S will satisfy the screen, hence overfit is likely (see Shao 1993). Let d be a lower bound on $\text{corr}(r, r_I)$. Proposition 3.2 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d_n \equiv \sqrt{1 - \frac{p}{n}}.$$

To reduce the chance of overfitting, consider models I with $C_p(I) \leq \min(2k, p)$. Models under both the $C_p = k$ line and the $C_p = 2k$ line are of interest.

Rule of thumb 3.6. a) After using a numerical method such as forward selection or backward elimination, let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted (recall that if the $c - 1$ indicator variables corresponding to a factor are deleted, then the factor has $c - 1$ degrees of freedom) and the jump in C_p is large, greater than 4, say.

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best subset found by the numerical procedure.

Rule of thumb 3.7. Assume that the full model has good response and residual plots and that $n > 5p$. Let subset I have k predictors, including a

constant. Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let I_{min} be the minimum C_p model and let I_I be the model with the fewest predictors satisfying $C_p(I_I) \leq C_p(I_{min}) + 1$. Do not use more predictors than model I_I to avoid overfitting. Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = corr($\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}$) ≥ 0.95.
iii) The plotted points in the FF plot (= EE plot for MLR) cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ (recall that $R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.

The following description of forward selection and backward elimination modifies the description of Section 1.6 slightly. Criterion such as AIC, $MSE(I)$ or $R^2_A(I)$ are sometimes used instead of C_p. For forward selection, the numerical method may add the predictor not yet in the model that has the smallest pvalue for the t test. For backward elimination, the numerical method may delete the variable in the model (that is not a constant) that has the largest pvalue for the t test.

Forward selection Step 1) $k = 1$: Start with a constant $w_1 = x_1$. Step 2) $k = 2$: Compute C_p for all models with $k = 2$ containing a constant and a single predictor x_i. Keep the predictor $w_2 = x_j$, say, that minimizes C_p.
Step 3) $k = 3$: Fit all models with $k = 3$ that contain w_1 and w_2. Keep the predictor w_3 that minimizes C_p. ...
Step j) $k = j$: Fit all models with $k = j$ that contains $w_1, w_2, ..., w_{j-1}$. Keep the predictor w_j that minimizes C_p. ...
Step p): Fit the full model.

Backward elimination: All models contain a constant $= u_1$. Step 0) $k = p$: Start with the full model that contains $x_1, ..., x_p$. We will also say that the full model contains $u_1, ..., u_p$ where $u_1 = x_1$ but u_i need not equal x_i for $i > 1$.
Step 1) $k = p - 1$: Fit each model with $k = p - 1$ predictors including a constant. Delete the predictor u_p, say, that corresponds to the model with the smallest C_p. Keep $u_1, ..., u_{p-1}$.
Step 2) $k = p-2$: Fit each model with $p-2$ predictors including a constant. Delete the predictor u_{p-1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-2}$.
...
Step j) $k = p - j$: fit each model with $p - j$ predictors including a constant. Delete the predictor u_{p-j+1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-j}$. ...
Step $p - 2$) $k = 2$. The current model contains u_1, u_2 and u_3. Fit the model u_1, u_2 and the model u_1, u_3. Assume that model u_1, u_2 minimizes C_p. Then delete u_3 and keep u_1 and u_2.

Heuristically, backward elimination tries to delete the variable that will increase C_p the least. An increase in C_p greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may use some other criterion: eg, delete the variable such that the submodel I with j predictors has a) the smallest $C_p(I)$ or b) the biggest p–value in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the model with $j + 1$ terms from the previous step (using the j predictors in I and the variable x^*_{j+1}) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease C_p the most. A decrease in C_p less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may use some other criterion, eg, add the variable such that the submodel I with j nontrivial predictors has a) the smallest $C_p(I)$ or b) the smallest p–value in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, et cetera. Recall that $ESP(I) = \hat{Y}_I$. Make a scatterplot matrix of the ESPs for M1, M2, M3, M4, M5 and Y. Good candidates should have estimated sufficient predictors that are highly correlated with the full model

ESP (the correlation should be at least 0.9 and preferably greater than 0.95). Similarly, make a scatterplot matrix of the residuals for M1, M2, M3, M4 and M5.

To summarize, the final submodel should have few predictors, few variables with large OLS t test p–values (0.01 to 0.05 is borderline), good response and residual plots and an FF plot (= EE plot) that clusters tightly about the identity line. If a factor has $c-1$ indicator variables, either keep all $c-1$ indicator variables or delete all $c-1$ indicator variables, do not delete some of the indicator variables.

Example 3.7. The pollution data of McDonald and Schwing (1973) can be obtained from STATLIB or the text's website. The response $Y = mort$ is the mortality rate and most of the independent variables were related to pollution. A scatterplot matrix of the first 9 predictors and Y was made and then a scatterplot matrix of the remaining predictors with Y. The log rule suggested making the log transformation with 4 of the variables. The summary output is shown on the following page. The response and residual plots were good. Notice that $p = 16$ and $n = 60 < 5p$. Also many p-values are too high.

```
Response      = MORT
Label       Estimate        Std. Error    t-value    p-value
Constant    1881.11         442.628         4.250     0.0001
DENS        0.00296328      0.00396521      0.747     0.4588
EDUC       -19.6669         10.7005        -1.838     0.0728
log[HC]    -31.0112         15.5615        -1.993     0.0525
HOUS       -0.401066        1.64372        -0.244     0.8084
HUMID      -0.445403        1.06762        -0.417     0.6786
JANT       -3.58522         1.05355        -3.403     0.0014
JULT       -3.84292         2.12079        -1.812     0.0768
log[NONW]   27.2397         10.1340         2.688     0.0101
log[NOX]    57.3041         15.4764         3.703     0.0006
OVR65      -15.9444         8.08160        -1.973     0.0548
POOR        3.41434         2.74753         1.243     0.2206
POPN       -131.823         69.1908        -1.905     0.0633
PREC        3.67138         0.778135        4.718     0.0000
log[SO]    -10.2973         7.38198        -1.395     0.1700
WWDRK       0.882540        1.50954         0.585     0.5618
```

```
R Squared:  0.787346    Sigma hat:  33.2178
Number of cases:  60 Degrees of freedom:   44

Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression     15   179757.       11983.8      10.86    0.0000
Residual       44   48550.5       1103.42
```

Shown below this paragraph is some output from forward selection. The minimum C_p model had $C_p = 7.353$ with 7 predictors. Deleting JANT from this model increased C_p to 17.763, suggesting that JANT is an important predictor. Notice that $C_p > 2k = 12$ for the model that deletes JANT.

```
Base terms: (log[NONW] EDUC log[SO] PREC)
                 df   RSS           |   k    C_I
Add: log[NOX]    54   72563.9       |   6   17.763
Add: JANT        54    72622.       |   6   17.815
Add: HOUS        54   74884.8       |   6   19.866
Add: POPN        54   75350.2       |   6   20.288
Add: log[HC]     54   75373.4       |   6   20.309
Add: JULT        54   75405.8       |   6   20.338
Add: OVR65       54   75692.2       |   6   20.598
Add: HUMID       54   75747.4       |   6   20.648
Add: DENS        54   75872.1       |   6   20.761
Add: POOR        54   75938.4       |   6   20.821
Add: WWDRK       54   75971.8       |   6   20.851

Base terms: (log[NONW] EDUC log[SO] PREC log[NOX])
                df   RSS           |   k    C_I
Add: JANT       53    58871.       |   7    7.353
Add: log[HC]    53   69233.3       |   7   16.744
Add: HOUS       53   70774.1       |   7   18.141
Add: POPN       53   71424.7       |   7   18.730
Add: POOR       53   72049.4       |   7   19.296
Add: OVR65      53   72337.1       |   7   19.557
Add: JULT       53   72348.6       |   7   19.568
Add: WWDRK      53   72483.1       |   7   19.690
```

```
Add: DENS       53   72494.9       |   7   19.700
Add: HUMID      53   72563.9       |   7   19.763
```

Output for backward elimination is shown below, and the minimum C_p model had $C_p = 6.284$ with 6 predictors. Deleting EDUC increased C_p to $10.800 > 2k = 10$. Since C_p increased by more than 4, EDUC is probably important.

```
Current terms: (EDUC JANT log[NONW] log[NOX] OVR65 PREC)
                     df   RSS           |   k    C_I
Delete: OVR65        54   59897.9       |   6    6.284
Delete: EDUC         54   66809.3       |   6   12.547
Delete: log[NONW]    54   73178.1       |   6   18.319
Delete: JANT         54   76417.1       |   6   21.255
Delete: PREC         54   83958.1       |   6   28.089
Delete: log[NOX]     54   86823.1       |   6   30.685

Current terms: (EDUC JANT log[NONW] log[NOX] PREC)
                     df   RSS           |   k    C_I
Delete: EDUC         55   67088.1       |   5   10.800
Delete: JANT         55   76467.4       |   5   19.300
Delete: PREC         55   87206.7       |   5   29.033
Delete: log[NOX]     55   88489.6       |   5   30.196
Delete: log[NONW]    55   95327.5       |   5   36.393
```

Taking the minimum C_p model from backward elimination gives the output shown below. The response and residual plots were OK although the correlation in the RR and FF plots was not real high. The R^2 in the submodel decreased from about 0.79 to 0.74 while $\hat{\sigma} = \sqrt{MSE}$ was 33.22 for the full model and 33.31 for the submodel. Removing nonlinearities from the predictors by using two scatterplots and the log rule, and then using backward elimination and forward selection, seems to be very effective for finding the important predictors for this data set. See Problem 3.17 in order to reproduce this example with the essential plots.

```
Response      = MORT
Label       Estimate        Std. Error    t-value    p-value
Constant    943.934         82.2254        11.480     0.0000
```

```
EDUC       -15.7263         6.17683        -2.546     0.0138
JANT       -1.86899         0.483572       -3.865     0.0003
log[NONW]   33.5514         5.93658         5.652     0.0000
log[NOX]    21.7931         4.29248         5.077     0.0000
PREC        2.92801         0.590107        4.962     0.0000

R Squared: 0.737644    Sigma hat:  33.305
Number of cases:  60  Degrees of freedom: 54

Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      5   168410.       33681.9      30.37    0.0000
Residual       54   59897.9       1109.22
```

Example 3.8. The FF and RR plots can be used as a diagnostic for whether a given numerical method is including too many variables. Gladstone (1905-1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index*. The *sex* (coded as 0 for females and 1 for males) of each subject was also included. The variable *cause* was coded as 1 if the cause of death was acute, 3 if the cause of death was chronic, and coded as 2 otherwise. A variable *ageclass* was coded as 0 if the age was under 20, 1 if the age was between 20 and 45, and as 3 if the age was over 45. Head *size*, the product of the *head length*, *head breadth*, and *head height*, is a volume measurement, hence $(size)^{1/3}$ was also used as a predictor with the same physical dimensions as the other lengths. Thus there are 11 nontrivial predictors and one response, and all models will also contain a constant. Nine cases were deleted because of missing values, leaving 267 cases.

Figure 3.7 shows the response plots and residual plots for the full model and the final submodel that used a constant, $size^{1/3}$, *age* and *sex*. The five cases separated from the bulk of the data in each of the four plots correspond to five infants. These may be outliers, but the visual separation reflects the small number of infants and toddlers in the data. A purely numerical variable selection procedure would miss this interesting feature of the data. We will first perform variable selection with the entire data set, and then examine the effect of deleting the five cases. Using forward selection and the C_p statistic

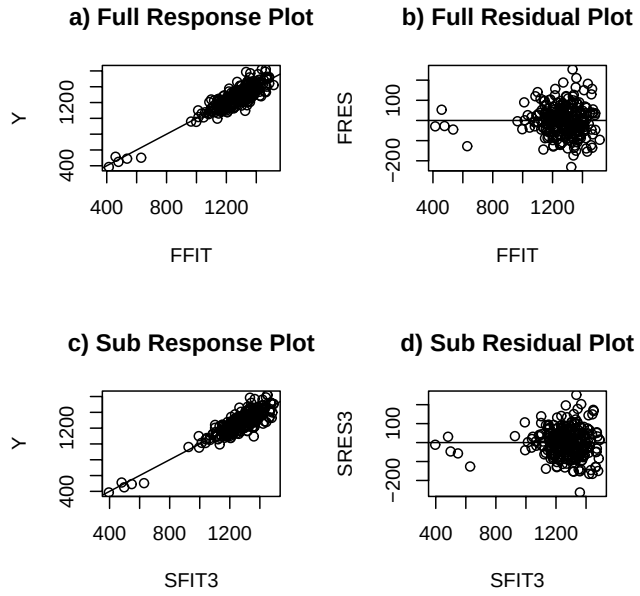

Figure 3.7: Gladstone data: comparison of the full model and the submodel.

on the Gladstone data suggests the subset I_5 containing a constant, $(size)^{1/3}$, *age*, *sex*, *breadth*, and *cause* with $C_p(I_5) = 3.199$. The p–values for breadth and cause were 0.03 and 0.04, respectively. The subset I_4 that deletes *cause* has $C_p(I_4) = 5.374$ and the p–value for *breadth* was 0.05. Figure 3.8d shows the RR plot for the subset I_4. Note that the correlation of the plotted points is very high and that the OLS and identity lines nearly coincide.

A scatterplot matrix of the predictors and response suggests that $(size)^{1/3}$ might be the best single predictor. First we regressed $Y =$ *brain weight* on the eleven predictors described above (plus a constant) and obtained the residuals r_i and fitted values $\hat{Y}_i$. Next, we regressed Y on the subset I containing $(size)^{1/3}$ and a constant and obtained the residuals $r_{I,i}$ and the fitted values $\hat{y}_{I,i}$. Then the RR plot of $r_{I,i}$ versus r_i, and the FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ were constructed.

For this model, the correlation in the FF plot (Figure 3.8b) was very high, but in the RR plot the OLS line did not coincide with the identity line (Figure 3.8a). Next *sex* was added to I, but again the OLS and identity lines did not coincide in the RR plot (Figure 3.8c). Hence *age* was added to I. Figure 3.9a

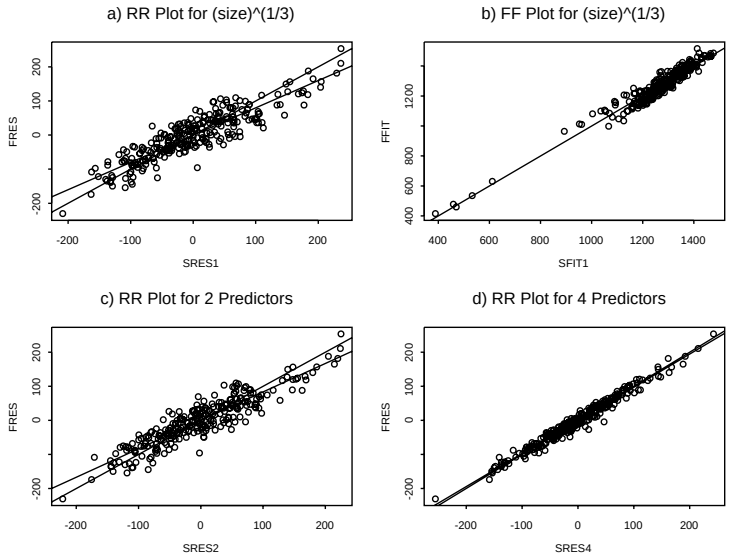

Figure 3.8: Gladstone data: submodels added $(size)^{1/3}$, sex, age and finally $breadth$.

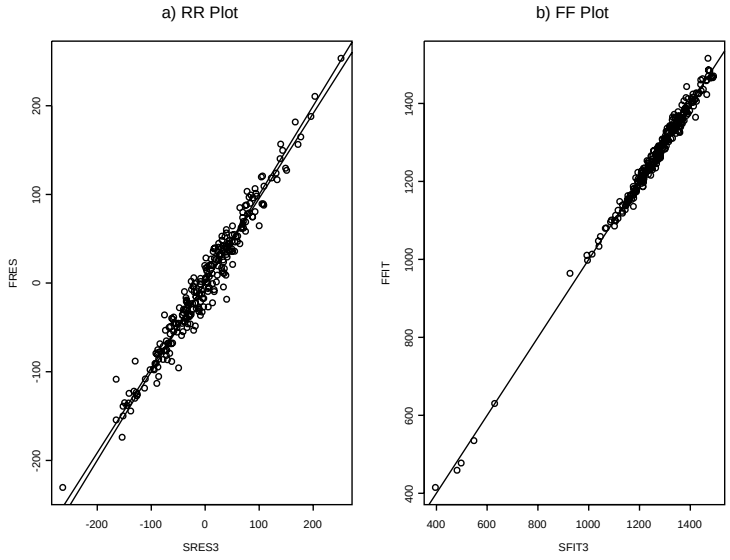

Figure 3.9: Gladstone data with Predictors $(size)^{1/3}$, sex, and age

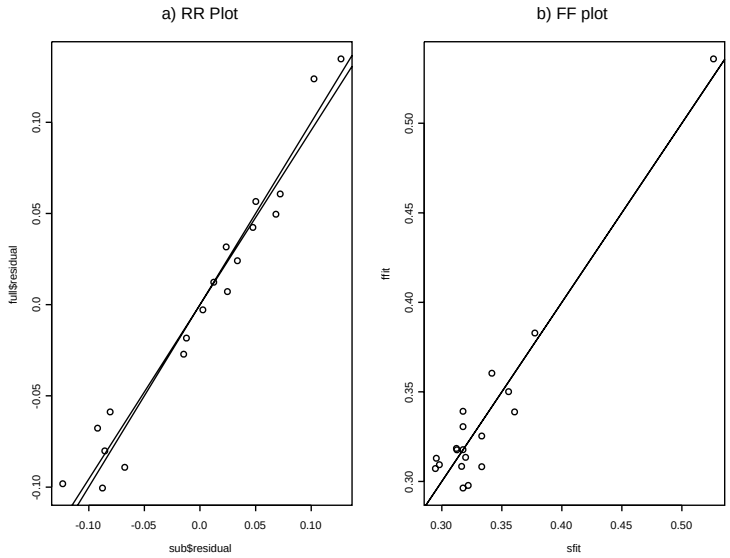

Figure 3.10: RR and FF Plots for Rat Data

shows the RR plot with the OLS and identity lines added. These two lines now nearly coincide, suggesting that a constant plus $(size)^{1/3}$, *sex*, and *age* contains the relevant predictor information. This subset has $C_p(I) = 7.372$, $R_I^2 = 0.80$, and $\hat{\sigma}_I = 74.05$. The full model which used 11 predictors and a constant has $R^2 = 0.81$ and $\hat{\sigma} = 73.58$. Since the C_p criterion suggests adding *breadth* and *cause*, the C_p criterion may be leading to an overfit.

Figure 3.9b shows the FF plot. The five cases in the southwest corner correspond to five infants. Deleting them leads to almost the same conclusions, although the full model now has $R^2 = 0.66$ and $\hat{\sigma} = 73.48$ while the submodel has $R_I^2 = 0.64$ and $\hat{\sigma}_I = 73.89$.

Example 3.9. Cook and Weisberg (1999a, p. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The data set is included as the file *rat.lsp* in the *Arc* software and can be obtained from the website (www.stat.umn.edu/arc/). The response Y is the fraction of the drug recovered from the rat's liver. The three predictors are the *body weight* of the rat, the *dose* of the drug, and the *liver weight*. The experimenter expected the response to be independent of the predictors, and 19 cases were used. However, the C_p criterion suggests

using the model with a constant, *dose* and *body weight*, both of whose coefficients were statistically significant. The RR and FF plots are shown in Figure 3.10. The identity line was added to both plots and the OLS line was added to the RR plot. The FF plot shows one outlier, the third case, that is clearly separated from the rest of the data.

We deleted this case and again searched for submodels. The C_p statistic is less than one for all three simple linear regression models, and the RR and FF plots look the same for *all* submodels containing a constant. Figure 2.2 shows the RR plot where the residuals from the full model are plotted against $Y - \overline{Y}$, the residuals from the model using no nontrivial predictors. This plot suggests that the response Y is independent of the nontrivial predictors.

The point of this example is that a subset of outlying cases can cause numeric second-moment criteria such as C_p to find structure that does not exist. The FF and RR plots can sometimes detect these outlying cases, allowing the experimenter to run the analysis without the influential cases. The example also illustrates that global numeric criteria can suggest a model with one or more nontrivial terms when in fact the response is independent of the predictors.

Numerical variable selection methods for MLR are very sensitive to "influential cases" such as outliers. Olive and Hawkins (2005) show that a plot of the residuals versus Cook's distances (see Section 3.5) can be used to detect influential cases. Such cases can also often be detected from response, residual, RR and FF plots.

Warning: deleting influential cases and outliers will often lead to better plots and summary statistics, but the cleaned data may no longer represent the actual population. In particular, the resulting model may be very poor for prediction.

Multiple linear regression data sets with cases that influence numerical variable selection methods are common. Table 3.1 shows results for seven interesting data sets. The first two rows correspond to the Ashworth (1842) data, the next 2 rows correspond to the Gladstone Data in Example 3.8, and the next 2 rows correspond to the Gladstone data with the 5 infants deleted. Rows 7 and 8 are for the Buxton (1920) data while rows 9 and 10 are for the Tremearne (1911) data. These data sets are available from the book's website. Results from the final two data sets are given in the last 4 rows. The last 2 rows correspond to the rat data described in Example 3.9. Rows 11

Table 3.1: Summaries for Seven Data Sets

influential cases file, response	submodel I transformed predictors	p, $C_p(I)$, $C_p(I,c)$
14, 55	$\log(x_2)$	4, 12.665, 0.679
pop, log(y)	$\log(x_1)$, $\log(x_2)$, $\log(x_3)$	
118, 234, 248, 258	$(size)^{1/3}$, age, sex	10, 6.337, 3.044
cbrain,brnweight	$(size)^{1/3}$	
118, 234, 248, 258	$(size)^{1/3}$, age, sex	10, 5.603, 2.271
cbrain-5,brnweight	$(size)^{1/3}$	
11, 16, 56	sternal height	7, 4.456, 2.151
cyp,height	none	
3, 44	x_2, x_5	6, 0.793, 7.501
major,height	none	
11, 53, 56, 166	log(LBM), log(Wt), sex	12, −1.701, 0.463
ais,%Bfat	log(Ferr), log(LBM), log(Wt), $\sqrt{Ht}$	
3	no predictors	4, 6.580, −1.700
rat,y	none	

and 12 correspond to the *Ais* data that comes with *Arc* (Cook and Weisberg, 1999a).

The full model used p predictors, including a constant. The final submodel I also included a constant, and the nontrivial predictors are listed in the second column of Table 3.1. For a candidate submodel I, let $C_p(I, c)$ denote the value of the C_p statistic for the *clean data* that omits influential cases and outliers. The third column lists p, $C_p(I)$ and $C_p(I, c)$ while the first column gives the set of influential cases. Two rows are presented for each data set. The second row gives the response variable and any predictor transformations. For example, for the Gladstone data $p = 10$ since there were 9 nontrivial predictors plus a constant. Only the predictor *size* was transformed, and the final submodel is the one given in Example 3.8. For the rat data, the final submodel is the one given in Example 3.9: none of the 3 nontrivial predictors was used.

Table 3.1 and simulations suggest that if the subset I has k predictors, then using the $C_p(I) \leq \min(2k, p)$ screen is better than using the conventional

$C_p(I) \leq k$ screen. The major and ais data sets show that deleting the influential cases may increase the C_p statistic. Thus interesting models from the entire data set and from the clean data set should be examined.

Example 3.10. Conjugated linoleic acid (CLA), occurs in beef and dairy products and appears to have many human health benefits. Joanne Numrich provided four data sets where the response was the amount of CLA (or related compounds) and the explanatory variables were feed components from the cattle diet. The data was to be used for descriptive and exploratory purposes. Several data sets had outliers with unusually high levels of CLA. These outliers were due to one researcher and may be the most promising cases in the data set. However, to describe the bulk of the data with OLS MLR, the outliers were omitted. In one of the data sets there are 33 cases and 25 predictors, including a constant. Regressing Y on all of the predictors gave $R^2 = .84$ and an ANOVA F test p-value of 0.223, suggesting that none of the predictors are useful. From Proposition 2.5, an $R^2 > (p-1)/(n-1) = .75$ is not very surprising. Remarks above Theorem 2.7 help explain why R^2 can be high with a high ANOVA F test p-value.

Of course just fitting the data to the collected variables is a poor way to proceed. Only variables $x_1, x_2, x_5, x_6, x_{20}$ and x_{21} took on more than a few values. Taking $\log(Y)$ and using variables x_2, x_9, x_{23}, and x_{24} seemed to result in an adequate model, although the number of distinct fitted values was rather small. See Problem 3.18 for more details.

3.5 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)

Diagnostics are used to check whether model assumptions are reasonable. This section focuses 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 $\mathrm{E}(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$. The zero mean and symmetry assumptions are often not very important.

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 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

$$\overline{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i \tag{3.8}$$

and

$$\boldsymbol{C} = \mathrm{Cov}(\boldsymbol{U}) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{3.9}$$

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

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{3.10}$$

denote the $n \times 1$ vector of fitted values from 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 3.9. 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*

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{3.11}$$

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

Proposition 3.3. 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 $\mathrm{CD}_i > 0.5$ 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 $\mathrm{trace}(\boldsymbol{H}) = \sum_{i=1}^{n} h_i = p$. It can be shown that $0 \leq h_i \leq 1$. 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 Proposition 3.3c) 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 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.

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 $b_1 + b_2 x_2 + b_3 x_3$ fits the data well, then the plot of $b_2 x_2 + b_3 x_3$ 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 the residual plot of the fitted values $\widehat{Y}_i$ versus the residuals r_i, and the second plot is the response plot of the fitted values $\widehat{Y}_i$ versus the response Y_i.

Recalling Definitions 2.11 and 2.12, residual and response plots are plots of $w_i = \boldsymbol{x}_i^T \boldsymbol{\eta}$ versus r_i and Y_i, respectively, where $\boldsymbol{\eta}$ is a known $p \times 1$ vector. The most commonly used residual and response plots takes $\boldsymbol{\eta} = \widehat{\boldsymbol{\beta}}$. 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. Proposition 2.1 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 multiple linear regression model holds. Recall that residual plots *magnify departures* from the model while the response plot emphasizes *how well the model fits the data.*

When the bulk of the data follows the MLR model, the following *rules of thumb* are useful for finding influential cases and outliers form the response and residual plots. 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, recall that the coefficient of determination

R^2 is equal to the squared correlation $(\text{corr}(Y, \hat{Y}))^2$. 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.

Information from numerical diagnostics can be incorporated into the response plot by highlighting cases that have large absolute values of the diagnostic. For example, the Cook's distance CD_i for the ith case tends to be large if $\hat{Y}_i$ is far from the sample mean $\overline{Y}$ and if the corresponding absolute residual $|r_i|$ is not small. If $\hat{Y}_i$ 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 $\overline{Y}$. Thus cases with large Cook's distances can often be found by examining the response and residual plots.

Example 3.11. 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 2.1 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.

3.6 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 3.10. *Outliers* are cases that lie far from the bulk of the data. Hence Y outliers are cases that have unusually large vertical distances from the MLR fit to the bulk of the data while $\boldsymbol{x}$ outliers are cases with predictors $\boldsymbol{x}$ that lie far from the bulk of the $\boldsymbol{x}_i$. 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 residual and response plots are very useful for detecting outliers. If there is a cluster of cases with outlying Ys, the identity line will often pass through the outliers. If there are two clusters with similar Ys, then the two plots may fail to show the clusters. Then using methods to detect $\boldsymbol{x}$ outliers may be useful.

Let the q continuous predictors in the MLR model be collected into vectors $\boldsymbol{u}_i$ for $i = 1, ..., n$. Let the $n \times q$ matrix $\boldsymbol{W}$ have n rows $\boldsymbol{u}_1^T, ..., \boldsymbol{u}_n^T$. Let the $q \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $q \times q$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a covariance estimator. Often $q = p - 1$ and only the constant is omitted from $\boldsymbol{x}_i$ to create $\boldsymbol{u}_i$.

Definition 3.11. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{u}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{u}_i - T(\boldsymbol{W})) \quad (3.12)$$

for each point $\boldsymbol{u}_i$. Notice that D_i^2 is a random variable (scalar valued).

The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(\boldsymbol{W}) = \overline{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i,$$

and

$$\boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - T(\boldsymbol{W}))(\boldsymbol{u}_i - T(\boldsymbol{W}))^T$$

and will be denoted by MD_i. When $T(\boldsymbol{W})$ and $\boldsymbol{C}(\boldsymbol{W})$ are robust estimators, $D_i = \sqrt{D_i^2}$ will sometimes be denoted by RD_i. We suggest using the Olive (2009) FCH estimator as the robust estimator. The sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value of the sample z-score $|z_i| = |(Y_i - \overline{Y})/\hat{\sigma}|$. Also notice that the Euclidean distance of $\boldsymbol{u}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_q)$ where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. Plot the MD_i versus the RD_i to detect outlying $\boldsymbol{u}$.

Definition 3.12: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

Olive (2002) shows that the plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the predictor distribution is multivariate normal (MVN), and will follow a line with zero intercept but non–unit slope if the distribution is elliptically contoured with nonsingular covariance matrix but not MVN. (Such distributions have linear scatterplot matrices. See Chapter 14.) Hence if the plotted points in the DD plot follow some line through the origin, then there is some evidence that outliers and strong nonlinearities have been removed from the predictors.

Example 3.12. 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.

Figure 3.11 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

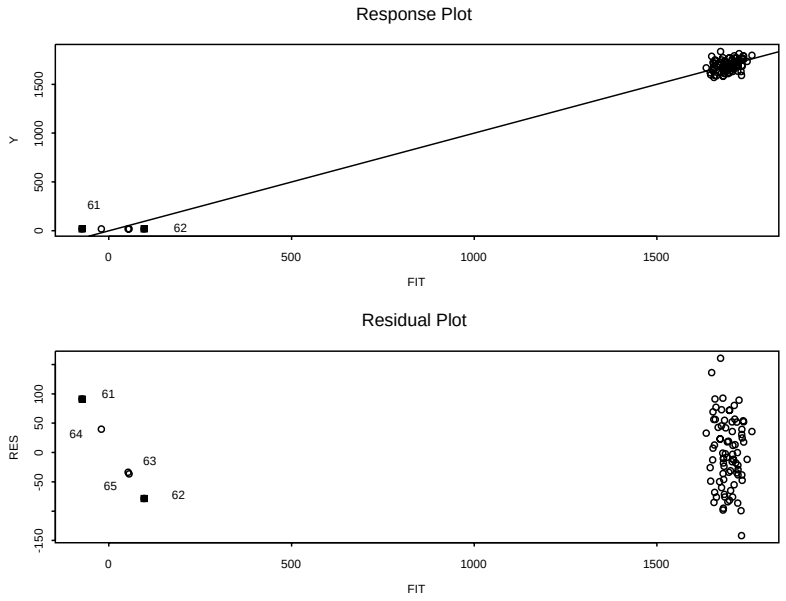

Figure 3.11: Residual and Response Plots for Buxton Data

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 3.12a shows the DD plot made from the four predictors *head length, nasal height, bigonal breadth,* and *cephalic index*. The five massive outliers correspond to head lengths that were recorded to be around 5 feet. Figure 3.12b is the DD plot computed after deleting these points and suggests that the predictor distribution is now much closer to a multivariate normal distribution.

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. 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.

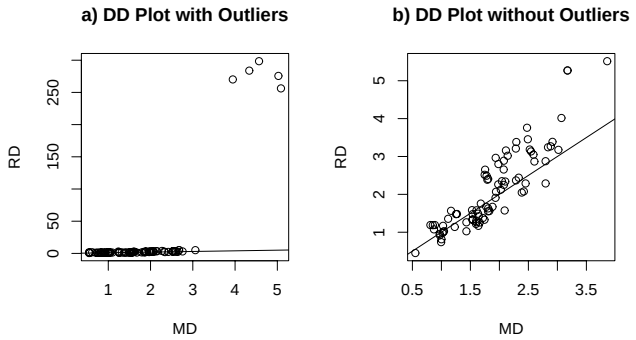

Figure 3.12: DD Plots for Buxton Data

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. The identity line and $r = 0$ line may pass right through a cluster of outliers, but the cluster of outliers can often be detected because there is a large gap between the cluster and the bulk of the data, as in Figure 3.11.

2. Make a DD plot of the predictors that take on many values (the continuous predictors).

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

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. A third, much more effective method, is to use the response and residual plots.

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 an additional rough guideline.

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.

3.7 Summary

1) Suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$, $x_1, x_2 > 0$ and that the plotted points follow a nonlinear one to one function. Consider the **ladder of powers** $-1, -0.5, -1/3, 0, 1/3, 0.5,$ and 1. The **ladder rule** says to spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger.

2) Suppose w is positive. The **log rule** says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$.

3) There are several guidelines for choosing power transformations. First, see the rule 1) and 2) above. Suppose that all values of the variable w to be transformed are positive. The log rule often works wonders on the data.

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. The **unit rule** says that if X_i and y have the same units, then use the same transformation of X_i and y. The **cube root rule** says that if w is a volume measurement, then cube root transformation $w^{1/3}$ may be useful. Consider the ladder of powers given in point 1). No transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation. Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if y = weight, X_1 = volume = $X_2 * X_3 * X_4$, then y vs. $X_1^{1/3}$ or $\log(y)$ vs. $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ may both work. Also if y 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.

4) To find a **response transformation**, make the transformation plots and choose a transformation such that the **transformation plot** is linear.

5) A factor (with c levels $a_1, ..., a_c$) is incorporated into the MLR model by using $c - 1$ indicator variables $x_{Wi} = 1$ if $W = a_i$ and $x_{Wi} = 0$ otherwise, where one of the levels a_i is omitted, eg, use $i = 1, ..., c - 1$.

6) For **variable selection**, the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The **full model is always a submodel**. The full model has $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$.

7) Make scatterplot matrices of the predictors and the response. Then **remove strong nonlinearities from the predictors using power transformations**. The log rule is very useful.

8) Either include all of the indicator variables for a factor in the model or exclude all of them. If the model contains powers or interactions, also include all main effects in the model.

9) After selecting a submodel I, make the response and residual plots for the full model and the submodel. Make the RR plot of $r_{I,i}$ versus r_i and the FF plot of $\hat{Y}_{I,i}$ versus Y_i. The submodel is good if the plotted points in the FF and RR plots cluster tightly about the identity line. In the RR plot, the OLS line and identity line can be added to the plot as visual aids. It should be difficult to see that the OLS and identity lines intersect at the origin, so the two lines should nearly coincide at the origin. If the FF plot looks good but the RR plot does not, the submodel may be good if the main goal of the

analysis is for prediction.

10) **Forward selection** Step 1) $k = 1$: Start with a constant $w_1 = x_1$.
Step 2) $k = 2$: Compute C_p for all models with $k = 2$ containing a constant and a single predictor x_i. Keep the predictor $w_2 = x_j$, say, that minimizes C_p.
Step 3) $k = 3$: Fit all models with $k = 3$ that contain w_1 and w_2. Keep the predictor w_3 that minimizes C_p. ...
Step j) $k = j$: Fit all models with $k = j$ that contains $w_1, w_2, ..., w_{j-1}$. Keep the predictor w_j that minimizes C_p. ...
Step p): Fit the full model.

Backward elimination: All models contain a constant $= u_1$. Step 0) $k = p$: Start with the full model that contains $x_1, ..., x_p$. We will also say that the full model contains $u_1, ..., u_p$ where $u_1 = x_1$ but u_i need not equal x_i for $i > 1$.
Step 1) $k = p - 1$: Fit each model with $k = p - 1$ predictors including a constant. Delete the predictor u_p, say, that corresponds to the model with the smallest C_p. Keep $u_1, ..., u_{p-1}$.
Step 2) $k = p-2$: Fit each model with $p-2$ predictors including a constant. Delete the predictor u_{p-1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-2}$.
...
Step j) $k = p - j$: fit each model with $p - j$ predictors including a constant. Delete the predictor u_{p-j+1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-j}$. ...
Step $p - 2$) $k = 2$. The current model contains u_1, u_2 and u_3. Fit the model u_1, u_2 and the model u_1, u_3. Assume that model u_1, u_2 minimizes C_p. Then delete u_3 and keep u_1 and u_2.

11) Let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined. Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked.

12) There are several guidelines for building a MLR model. Suppose that variable Z is of interest and variables $W_1, ..., W_r$ have been collected along

with Z. Make a scatterplot matrix of $W_1, ..., W_r$ and Z. (If r is large, several matrices may need to be made. Each one should include Z.) Remove or correct any gross outliers. It is often a good idea to transform the W_i to **remove any strong nonlinearities from the predictors**. Eventually you will find a response variable $Y = t_Z(Z)$ and predictor variable $X_1, ..., X_{p-1}$ for the **full model**. Interactions such as $X_k = W_i W_j$ and powers such as $X_k = W_i^2$ may be of interest. Indicator variables are often used in interactions but do not transform an indicator variable. The response plot for the full model should be linear and the residual plot should be ellipsoidal with zero trend. Find the LS output. The statistic R^2 gives the proportion of the variance of Y explained by the predictors and is of great importance. Use backwards elimination and forward selection with the $C_p(I)$ statistic to suggest candidate models I. As a rule of thumb, (assuming that the sample size n is much larger than the pool of predictors, eg $n > 5p$), make sure that $R_I^2 > 0.9R^2$ or $R_I^2 > R^2 - 0.07$. Often want the number of predictors k in the submodel to be small. We will almost always include a constant in the submodel. If the submodel seems to be good, make the response plot and residual plot for the submodel. They should be linear and ellipsoidal with zero trend, respectively. From the output, see if any terms can be eliminated (are there any predictors X_i such that the p–value for Ho:$\beta_i = 0 > 0.01$?)

13) Assume that the full model has good response and residual plots and than $n > 5p$. Let subset I have k predictors, including a constant. The following rules of thumb may be useful, but may not all hold simultaneously. Let I_{min} be the minimum C_p model and let I_I be the model with the fewest predictors satisfying $C_p(I_I) \leq C_p(I_{min}) + 1$. Do not use more predictors than model I_I to avoid overfitting. Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = corr($\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}$) ≥ 0.95.
iii) The plotted points in the FF plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ (recall that $R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.

ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.

14) Always check that the full model is good. If the candidate model seems to be good, the usual MLR checks should still be made. In particular, the response plot and residual plot need to be made for the submodel.

15) **Influence** is roughly (leverage)(discrepancy). The leverages h_i are the diagonal elements of the hat matrix $\boldsymbol{H}$ and measure how far $\boldsymbol{x}_i$ is from the sample mean of the predictors. Cook's distance is widely used, but the response plot and residual plot are the most effective tools for detecting outliers and influential cases.

3.8 Complements

With one data set, OLS is a great place to start but a bad place to end. If $n = 5kp$ where $k > 2$, it may be useful to take a random sample of n/k cases to build the MLR model. Then check the model on the full data set.

Predictor Transformations

One of the most useful techniques in regression is to remove gross nonlinearities in the predictors by using predictor transformations. The log rule is very useful for transforming highly skewed predictors. The linearizing of the predictor relationships could be done by using marginal power transformations or by transforming the joint distribution of the predictors towards an elliptically contoured distribution. The linearization might also be done by using simultaneous power transformations $\boldsymbol{\lambda} = (\lambda_2, \ldots, \lambda_p)^T$ of the predictors so that the vector $\boldsymbol{w}^{\boldsymbol{\lambda}} = (x_2^{(\lambda_2)}, ..., x_p^{(\lambda_p)})^T$ of transformed predictors is approximately multivariate normal. A method for doing this was developed by Velilla (1993). (The basic idea is the same as that underlying the likelihood approach of Box and Cox for estimating a power transformation of the response in regression, but the likelihood comes from the assumed multivariate normal distribution of $\boldsymbol{w}^{\boldsymbol{\lambda}}$.) The Cook and Nachtsheim (1994) procedure can cause the distribution to be closer to elliptical symmetry. Marginal Box-Cox transformations also seem to be effective. Power transformations can also be selected with slider bars in *Arc*. More will be said about predictor transformations in Section 15.3.

Suppose that it is thought that the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ could be improved by transforming x_j. Let $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{u}^T\boldsymbol{\eta} + \beta_j x_j$ where $\boldsymbol{u}^T\boldsymbol{\eta} = x_1\beta_1 +$

$\cdots + x_{j-1}\beta_{j-1} + x_{j+1}\beta_{j+1} + \cdots + x_p\beta_p$. Let $\tau(x_j)$ denote the unknown transformation.

Definition 3.13. Consider the OLS residuals $r_i(j) = Y_i - \boldsymbol{u}_i^T\hat{\boldsymbol{\eta}}$ obtained from the OLS regression of Y on $\boldsymbol{u}$. A *partial residual plot* or *component plus residual plot* or *ceres plot with linear augmentation* is a plot of the $r_i(j)$ versus x_j and is used to visualize τ.

Cook (1993) shows that partial residual plots are useful for visualizing τ provided that the plots of x_i versus x_j are linear. More general ceres plots, in particular ceres plots with smooth augmentation, can be used to visualize τ if $Y = \boldsymbol{u}^T\boldsymbol{\eta} + \tau(x_j) + e$ but the linearity condition fails.

The assumption that all values of x_1 and x_2 are positive for power transformation can be removed by using the modified power transformations of Yeo and Johnson (2000).

Response Transformations

Application 3.1 was suggested by Olive (2004b) and Olive and Hawkins (2009a). 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.

A variant of the method would plot the residual plot or both the response and the residual plot for each of the seven values of λ. Residual plots are also useful, but they no not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55).

Cook and Olive (2001) also suggest a graphical method for selecting and assessing response transformations under model (3.2). Cook and Weisberg (1994) show that a plot of Z versus $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ (swap the axis on the transformation plot for $\lambda = 1$) can be used to visualize t if $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$, suggesting that t^{-1} can be visualized in a plot of $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ versus Z.

If there is nonlinearity present in the scatterplot matrix of the nontrivial predictors, then **transforming the predictors to remove the nonlinearity will often be a useful procedure**. More will be said about response transformations for experimental designs in Section 5.3.

There has been considerable discussion on whether the response transfor-

mation parameter λ should be selected with maximum likelihood (see Bickel and Doksum 1981), or selected by maximum likelihood and then rounded to a meaningful value on a coarse grid Λ_L (see Box and Cox 1982 and Hinkley and Runger 1984). Suppose that no strong nonlinearities are present among the predictors $\boldsymbol{x}$ and that if predictor transformations were used, then the transformations were chosen without examining the response. Also assume that

$$Y = t_{\lambda_o}(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e.$$

Suppose that a transformation $t_{\hat{\lambda}}$ is chosen without examining the response. Results in Li and Duan (1989), Chen and Li (1998) and Chang and Olive (2009) suggest that if $\boldsymbol{x}$ has an approximate multivariate normal distribution, then the OLS ANOVA F, partial F and Wald t tests will have the correct level asymptotically, even if $\hat{\lambda} \neq \lambda_o$.

Now assume that the response is used to choose $\hat{\lambda}$. For example assume that the numerical Box Cox method is used. Then $\hat{\lambda}$ is likely to be variable unless the sample size is quite large, and considerable bias can be introduced, as observed by Bickel and Doksum (1981). Now assume that $\hat{\lambda}$ is chosen with the graphical method (and assume that ties are broken by using theory or by using the following list in decreasing order of importance $1, 0, 1/2, -1$ and $1/3$ so that the log transformation is chosen over the cube root transformation if both look equally good). Then $\hat{\lambda}$ will often rapidly converge in probability to a value $\lambda^* \in \Lambda_L$. Hence for moderate sample sizes, it may be reasonable to assume that the OLS tests have approximately the correct level. Let $W = t_{\hat{\lambda}}(Z)$ and perform the OLS regression of W on $\boldsymbol{x}$. If the response and residual plots suggest that the MLR model is appropriate, then the response transformation from the graphical method will be useful for description and exploratory purposes, and may be useful for prediction and inference.

The MLR assumptions always need to be checked after making a response transformation. Since the graphical method uses a response plot to choose the transformation, the graphical method should be much more reliable than a numerical method. Transformation plots should be made if a numerical method is used, but numerical methods are not needed to use the graphical method.

Variable Selection and Multicollinearity

The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous. Three important papers are

Jones (1946), Mallows (1973), and Furnival and Wilson (1974). Chatterjee and Hadi (1988, p. 43-47) give a nice account on the effects of overfitting on the least squares estimates. Also see Claeskins and Hjort (2003), Hjort and Claeskins (2003) and Efron, Hastie, Johnstone and Tibshirani (2004). Texts include Burnham and Anderson (2002), Claeskens and Hjort (2008) and Linhart and Zucchini (1986).

Cook and Weisberg (1999, p. 264-265) give a good discussion of the effect of deleting predictors on linearity and the constant variance assumption. Walls and Weeks (1969) note that adding predictors increases the variance of a predicted response. Also R^2 gets large. See Freedman (1983).

Discussion of biases introduced by variable selection and data snooping include Hurvich and Tsai (1990), Selvin and Stuart (1966) and Hjort and Claeskins (2003). This theory assumes that the full model is known before collecting the data, but in practice the full model is often built after collecting the data. Freedman (2005, p. 192–195) gives an interesting discussion on model building and variable selection.

Olive and Hawkins (2005) discuss influential cases in variable selection, as do Léger and Altman (1993).

The interpretation of Mallows C_p given in Proposition 3.2 is due to Olive and Hawkins (2005) and can be generalized to other 1D regression models. Other interpretations of the C_p statistic specific to MLR can be given. See Gilmour (1996). The C_p statistic is due to Jones (1946). Also see Kenard (1971).

The $AIC(I)$ statistic is often used instead of $C_p(I)$. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Find the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$. Then I_I is the initial submodel to examine, and often $I_I = I_{min}$. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$.

When there are strong linear relationships among the predictors, *multicollinearity* is present. Let R_k^2 be the coefficient of multiple determination when x_k is regressed on the remaining predictor variables, including a constant. The variance inflation factor is VIF(k) $= 1/(1 - R_k^2)$. Both R_k^2 and VIF(k) are large when multicollinearity is present. Following Cook and Weisberg (1999, p. 274), if s_k is the sample standard deviation of x_k, than the

standard error of $\hat{\beta}_k$ is

$$se(\hat{\beta}_k) = \frac{\sqrt{MSE}}{s_k\sqrt{n-1}}\frac{1}{1-R_k^2} = \frac{\sqrt{MSE}}{s_k\sqrt{n-1}}\sqrt{VIF(k)}.$$

Hence β_k becomes more difficult to estimate when multicollinearity is present. Variable selection is a useful way to reduce multicollinearity, and alternatives such as ridge regression are discussed in Gunst and Mason (1980). Belsley (1984) shows that centering the data before diagnosing the data for multicollinearity is not necessarily a good idea.

We note that the pollution data of Example 3.7 has been heavily analyzed in the ridge regression literature, but this data was easily handled by the log rule combined with variable selection. The pollution data can be obtained from this text's website, or from the STATLIB website:
(http://lib.stat.cmu.edu/).

The `leaps` function in *Splus* and `Proc Rsquare` in *SAS* can be used to perform all subsets variable selection with the C_p criterion. The `step` function in *R/Splus* can be used for forward selection and backward elimination.

Diagnostics

Excellent introductions to OLS diagnostics include Fox (1991) and Cook and Weisberg (1999, 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. Also see Velleman and Welsch (1981). Cook and Weisberg (1997, 1999 ch. 17) call a plot that emphasizes model agreement a *model checking plot*.

Outliers

Olive (2009) is an authoritative introduction to outlier detection. Some useful properties of the DD plot are given in Olive (2002). Theory for the FCH estimators is given in Olive (2009, ch. 10) and Olive and Hawkins (2009b).

3.9 Problems

Problems with an asterisk * are especially important.

```
Output for problem 3.1.
Current terms: (finger to ground nasal height sternal height)
                            df   RSS           |   k    C_I
Delete: nasal height        73   35567.2       |   3    1.617
Delete: finger to ground    73   36878.8       |   3    4.258
Delete: sternal height      73   186259.       |   3  305.047
```

3.1. From the output from backward elimination given on the previous page, what terms should be used in the MLR model to predict Y? (You can tell that the nontrivial variables are finger to ground, nasal height and sternal height from the "delete lines." DON'T FORGET THE CONSTANT!)

Output for Problem 3.2.

	L1	L2	L3	L4
# of predictors	10	6	4	3
# with $0.01 \leq$ p-value ≤ 0.05	0	0	0	0
# with p-value > 0.05	6	2	0	0
$R^2(I)$	0.774	0.768	0.747	0.615
$\text{corr}(\hat{Y}, \hat{Y}_I)$	1.0	0.996	0.982	0.891
$C_p(I)$	10.0	3.00	2.43	22.037
$\sqrt{MSE}$	63.430	61.064	62.261	75.921
p-value for partial F test	1.0	0.902	0.622	0.004

3.2. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The response plot and residual plot for the full model L1 was good. Model L3 was the minimum C_p model found. Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Output for Problem 3.3.

	L1	L2	L3	L4
# of predictors	10	5	4	3
# with $0.01 \leq$ p-value ≤ 0.05	0	1	0	0
# with p-value > 0.05	8	0	0	0
$R^2(I)$	0.655	0.650	0.648	0.630
$\text{corr}(\hat{Y}, \hat{Y}_I)$	1.0	0.996	0.992	0.981
$C_p(I)$	10.0	4.00	5.60	13.81
$\sqrt{MSE}$	73.548	73.521	73.894	75.187
p-value for partial F test	1.0	0.550	0.272	0.015

3.3. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The response plot and residual plot for the full model L1 was good. Model L2 was the minimum C_p model found.

a) Which model is I_I, the initial submodel to look at?

b) What other model or models, if any, should be examined?

Output for Problem 3.4.

k	CP	ADJUSTED R SQUARE	99 cases 2 outliers R SQUARE	RESID SS	MODEL VARIABLES
1	760.7	0.0000	0.0000	185.928	INTERCEPT ONLY
2	12.7	0.8732	0.8745	23.3381	B
2	335.9	0.4924	0.4976	93.4059	A
2	393.0	0.4252	0.4311	105.779	C
3	12.2	0.8748	0.8773	22.8088	B C
3	14.6	0.8720	0.8746	23.3179	A B
3	15.7	0.8706	0.8732	23.5677	A C
4	4.0	0.8857	0.8892	20.5927	A B C

k	CP	ADJUSTED R SQUARE	97 cases after deleting the 2 outliers R SQUARE	RESID SS	MODEL VARIABLES
1	903.5	0.0000	0.0000	183.102	INTERCEPT ONLY
2	0.7	0.9052	0.9062	17.1785	B
2	406.6	0.4944	0.4996	91.6174	A
2	426.0	0.4748	0.4802	95.1708	C
3	2.1	0.9048	0.9068	17.0741	A C
3	2.6	0.9043	0.9063	17.1654	B C
3	2.6	0.9042	0.9062	17.1678	A B
4	4.0	0.9039	0.9069	17.0539	A B C

3.4. The output above is from software that does all subsets variable selection. The data is from Ashworth (1842). The predictors were A = log(1692 property value), B = log(1841 property value) and C = log(percent increase in value) while the response variable is Y = log(1841 population).

a) The top output corresponds to data with 2 small outliers. From this output, what is the best model? Explain briefly.

b) The bottom output corresponds to the data with the 2 outliers removed. From this output, what is the best model? Explain briefly.

Problems using R/Splus.

Warning: Use the command *source("A:/regpack.txt")* **to download the programs. See Preface or Section 17.1.** Typing the name of the

`regpack` function, eg *tplot*, will display the code for the function. Use the `args` command, eg *args(tplot)*, to display the needed arguments for the function.

3.5*. You may also copy and paste R commands for this problem from (www.math.siu.edu/olive/reghw.txt).

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

b) Use the following *R/Splus* command to make a 100×3 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).

3.6. Download *cbrainx* and *cbrainy* from (www.math.siu.edu/olive/regdata.txt) into *R*. Either use the source command on regdata.txt if it is saved on a disk, or copy and paste the two files into *R*. Copy and paste the *R* commands for this problem from (www.math.siu.edu/olive/reghw.txt).

The data is the brain weight data from Gladstone (1905-6). The response Y is *brain weight* while the predictors are *age, breadth, cephalic, circum, headht, height, len, sex* and a constant. The *step* function can be used to perform forward selection and backward elimination in *R*.

a) Copy and paste the commands for this problem into *R*. The commands fit the full model, display the LS output and perform backward elimination using the AIC criterion. Copy and paste the output for backward elimination into *Word* (one page of output).

```
zx <- cbrainx[,c(1,3,5,6,7,8,9,10)]
zbrain <- as.data.frame(cbind(cbrainy,zx))
zfull <- lm(cbrainy~.,data=zbrain)
summary(zfull)
back <- step(zfull)
```

b) Want low AIC and as few predictors as possible. Backward elimination starts with the full model then deletes one nontrivial predictor at a time. The term <None> corresponds to the current model that does not eliminate any terms. The terms listed above <None> correspond to models that have smaller AIC than the current model. *R* stops when eliminating terms makes the AIC higher than the current model. Which terms, including a constant, were in this minimum AIC model?

c) Copy and paste the commands for this problem into *R*. The commands fit the null model that only contains a constant. Forward selection starts at the null model (corresponding to lower) and considers 8 nontrivial predictors (given by upper).

Copy and paste the output for forward selection into *Word* (two pages of output).

```
zint <- lm(cbrainy~1,data=zbrain)
forw <- step(zint,scope=list(lower=~1,
upper=~age+breadth+cephalic+circum+headht+height+len+sex),
direction="forward")
```

d) Forward selection in R starts with the null model and then adds a predictor *circum* to the model. Forward selection in R allows you to consider models with fewer predictors than the minimum AIC model (unlike backward elimination). Which terms, including a constant, were in the minimum AIC model?

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the northeast corner and click. Problems 3.7–3.11 use data sets that come with *Arc* (Cook and Weisberg 1999a).

3.7*. a) In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *big-mac.lsp*. Next use the menu commands "Graph&Fit> Plot of" to obtain a dialog window. Double click on *TeachSal* and then double click on *BigMac*. Then click on *OK*. These commands make a plot of $x = TeachSal$ = primary teacher salary in thousands of dollars versus $y = BigMac$ = minutes of labor needed to buy a Big Mac and fries. Include the plot in *Word*.

Consider transforming y with a (modified) power transformation

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

b) Should simple linear regression be used to predict y from x? Explain.

c) In the plot, $\lambda = 1$. Which transformation will increase the linearity of the plot, $\log(y)$ or $y^{(2)}$? Explain.

3.8*. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp*. Use the commands "Graph&Fit>Scatterplot Matrix of." In the dialog window select H, L, W, S and M (so select M last). Click on "OK" and include the scatterplot matrix in *Word*. The response M is the edible part of the mussel while the 4 predictors are shell measurements.

Are any of the marginal predictor relationships nonlinear? Is $E(M|H)$ linear or nonlinear?

3.9*. The file *wool.lsp* has data from a 3^3 experiment on the behavior of worsted yarn under cycles of repeated loadings. The response Y is the number of cycles to failure and the three predictors are the length, amplitude and load. Make five transformation plots by using the following commands.

From the menu "Wool" select "transform" and double click on *Cycles*. Select "modified power" and use $p = -1, -0.5$, 0 and 0.5. Use the menu commands "Graph&Fit>Fit linear LS" to obtain a dialog window. Next fit LS five times. Use *Amp*, *Len* and *Load* as the predictors for all 5 regressions, but use Cycles^{-1}, Cycles$^{-0.5}$, log[Cycles], Cycles$^{0.5}$ and Cycles as the response.

Use the menu commands "Graph&Fit>Plot of" to create a dialog window. Double click on L5:Fit-Values and double click on Cycles, double click on L4:Fit-Values and double click on Cycles$^{0.5}$, double click on L3:Fit-Values and double click on log[Cycles], double click on L2:Fit-Values and double click on Cycles$^{-0.5}$, double click on L1:Fit-Values and double click on Cycles^{-1}.

a) You may stop when the resulting plot in linear. Let Z = Cycles. Include the plot of $\widehat{Y}$ versus $Y = Z^{(\lambda)}$ that is linear in *Word*. Move the OLS slider bar to 1. What response transformation do you end up using?

b) Use the menu commands "Graph&Fit>Plot of" and put L5:Fit-Values in the H box and L3:Fit-Values in the V box. Is the plot linear?

3.10. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *bcherry.lsp*. The menu *Trees* will appear. Use the menu commands "Trees>Transform" and a dialog window will appear. Select terms *Vol*, *D*, and *Ht*. Then select the *log* transformation. The terms *log Vol*, *log D* and *log H* should be added to the data set. If a tree is shaped like a cylinder or a cone, then $Vol \propto D^2Ht$ and taking logs results in a linear model.

a) Fit the full model with $Y = \log Vol$, $X_1 = \log D$ and $X_2 = \log Ht$. Add the output that has the LS coefficients to *Word*.

b) Fitting the full model will result in the menu *L1*. Use the commands "L1>AVP–All 2D." This will create a plot with a slider bar at the bottom that says *log[D]*. This is the added variable plot for $\log(D)$. To make an added variable plot for $\log(Ht)$, click on the slider bar. Add the OLS line to the AV plot for $\log(Ht)$ by moving the *OLS slider bar* to 1, and add the zero line

by clicking on the "Zero line box". Include the resulting plot in *Word.*

c) Fit the reduced model that drops log(Ht). Make an RR plot with the residuals from the full model on the V axis and the residuals from the submodel on the H axis. Add the LS line and the identity line as visual aids. (Click on the *Options* menu to the left of the plot and type "y=x" in the resulting dialog window to add the identity line.) Include the plot in *Word*.

d) Similarly make an FF plot using the fitted values from the two models. Add the OLS line which is the identity line. Include the plot in *Word*.

e) Next put the residuals from the submodel on the V axis and $\log(Ht)$ on the H axis. Move the *OLS slider bar* to 1, and include this residual plot in *Word*.

f) Next put the residuals from the submodel on the V axis and the fitted values from the submodel on the H axis. Include this residual plot in *Word.*

g) Next put log(Vol) on the V axis and the fitted values from the submodel on the H axis. Move the *OLS slider bar* to 1, and include this response plot in *Word.*

h) Does $\log(Ht)$ seem to be an important term? If the only goal is to predict volume, will much information be lost if $\log(Ht)$ is omitted? **Beside each of the 6 plots, remark on the information given by the plot**. (Some of the plots will suggest that log(Ht) is needed while others will suggest that log(Ht) is not needed.)

3.11*. a) In this problem we want to build a MLR model to predict $Y = t(BigMac)$ where t is some power transformation. In *Arc* enter the menu commands "File>Load>Data>Arcg" and open the file *big-mac.lsp*. Make a scatterplot matrix of the variate valued variables and include the plot in *Word*.

b) The log rule makes sense for the BigMac data. From the scatterplot matrix, use the "Transformations" menu and select "Transform to logs". Include the resulting scatterplot matrix in *Word*.

c) From the "Mac" menu, select "Transform". Then select all 10 variables and click on the "Log transformations" button. Then click on "OK". From the "Graph&Fit" menu, select "Fit linear LS." Use log[BigMac] as the

response and the other 9 "log variables" as the Terms. This model is the full model. Include the output in *Word*.

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

e) Using the "L1" menu, select "Examine submodels" and try forward selection and backward elimination. Using the $C_p \leq \min(2k, p)$ rule suggests that the submodel using log[service], log[TeachSal] and log[TeachTax] may be good. From the "Graph&Fit" menu, select "Fit linear LS", fit the submodel and include the output in *Word*.

f) Make a response plot (L2:Fit-Values in H and log(BigMac) in V) and residual plot (L2:Fit-Values in H and L2:Residuals in V) for the submodel and include the plots in *Word*.

g) Make an RR plot (L2:Residuals in H and L1:Residuals in V) and FF plot (L2:Fit-Values in H and L1:Fit-Values in V) for the submodel and include the plots in *Word*. Move the OLS slider bar to 1 in each plot to add the identity line. For the RR plot, click on the *Options menu* then type $y = x$ in the long horizontal box near the bottom of the window and click on OK to add the identity line.

h) Do the plots and output suggest that the submodel is good? Explain.

Warning: The following problems 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.

3.12*. The following data set has 5 babies that are "good leverage points:" they look like outliers but should not be deleted because they follow the same model as the bulk of the data.

a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cbrain.lsp*. Select *transform* from the *cbrain* menu, and add $size^{1/3}$ using the power transformation option ($p = 1/3$). From *Graph&Fit,* select *Fit linear LS*. Let the response be *brnweight* and as terms include everything but *size* and *Obs*. Hence your model will include $size^{1/3}$. This regression will add *L1* to the menu bar. From this menu, select *Examine submodels.* Choose *forward selection.* You should get models including $k =$

2 to 12 terms including the constant. Find the model with the smallest $C_p(I) = C_I$ statistic and include all models with the same k as that model in *Word.* That is, if $k = 2$ produced the smallest C_I, then put the block with $k = 2$ into *Word.* Next go to the *L1* menu, choose *Examine submodels* and choose *Backward Elimination.* Find the model with the smallest C_I and include all of the models with the same value of k in *Word.*

b) What was the minimum C_p model was chosen by forward selection?

c) What was the minimum C_p model was chosen by backward elimination?

d) Which minimum C_p model do you prefer? Explain.

e) Give an explanation for why the two models are different.

f) Pick a submodel and include the regression output in *Word.*

g) For your submodel in f), make an RR plot with the residuals from the full model on the V axis and the residuals from the submodel on the H axis. Add the OLS line and the identity line y=x as visual aids. Include the RR plot in *Word.*

h) Similarly make an FF plot using the fitted values from the two models. Add the OLS line which is the identity line. Include the FF plot in *Word.*

i) Using the submodel, include the response plot (of $\hat{Y}$ versus Y) and residual plot (of $\hat{Y}$ versus the residuals) in *Word.*

j) Using results from f)-i), explain why your submodel is a good model.

3.13. Activate the *cyp.lsp* data set. Choosing no more than 3 nonconstant terms, try to predict *height* with multiple linear regression. Include a plot with the fitted values on the horizontal axis and height on the vertical axis. Is your model linear? Also include a plot with the fitted values on the horizontal axis and the residuals on the vertical axis. Does the residual plot suggest that the linear model may be inappropriate? (There may be outliers in the plot. These could be due to typos or because the error distribution has heavier tails than the normal distribution.) State which model you use.

3.14. Activate the insulation data, contributed by Elizabeth Spector, with the commands "File>Load>3 1/2 Floppy (A:)>insulation.lsp."

The data description should appear in the "Listener" window.

Then go to the "Graph&Fit" menu and choose "Plot of ..." and select "time" for the "H box" "y" for the "V box" and "type" for the "Mark by box". Then click on "OK" and a window with a plot should open.

a) The OLS popdown menu is the triangle below OLS. Select "Fit by marks–general" and then use the cursor to mover the small black box to 2 on the OLS slider bar. Then copy and paste the plot to *Word*. This command fits least squares quadratic functions to the data from each of the 5 types of insulation.

b) If there is no interaction, then the 5 curves will be roughly parallel and will not cross. The curves will cross if there is interaction. Is there interaction?

c) The top curve corresponds to no insulation and the temperature rapidly rose and then rapidly cooled off. Corn pith corresponds to curve 2. Is corn pith comparable to the more standard types of insulation 3–5?

3.15. Activate the *cement.lsp* data, contributed by Alyass Hossin. Act as if 20 different samples were used to collect this data. If 5 measurements on 4 different samples were used, then experimental design with repeated measures or longitudinal data analysis may be a better way to analyze this data.

a) From *Graph&Fit* select *Plot of*, place $x1$ in H, y in V and $x2$ in the *Mark by* box. From the OLS menu, select *Fit by marks–general* and move the slider bar to 2. Include the plot in *Word*.

b) A quadratic seems to be a pretty good MLR model. From the *cement* menu, select Transform, select x1, and place a 2 in the p box. This should add $x1^2$ to the data set. From *Graph&Fit* select *Fit linear LS*, select x1 and $x1^2$ as the terms and y as the response. Include the output in *Word*.

c) Make the response plot. Again from the OLS menu, select *Fit by marks–general* and move the slider bar to 1. Include the plot in *Word*. This plot suggests that there is an interaction: the CM cement is stronger for low curing times and weaker for higher curing times. The plot suggests that there may not be an interaction between the two new types of cement.

d) Place the residual plot in *Word*. (Again from the OLS menu, select *Fit by marks–general* and move the slider bar to 1.) The residual plot is slightly fan shaped.

e) From the *cement* menu, select *Make factors* and select x2. From the

cement menu, select *Make interactions* and select x1 and (F)x2. Repeat, selecting $x1^2$ and (F)x2. From *Graph&Fit* select *Fit linear LS*, select x1, $x1^2$, (F)x2, x1*(F)x2 and $x1^2$*(F)x2 as the terms and y as the response. Include the output in *Word*.

f) Include the response plot and residual plot in *Word*.

g) Next delete the standard cement in order to compare the two coal based cements. From *Graph&Fit* select *Scatterplot–matrix of*, then select x1, x2 and y. Hold down the leftmost mouse button and highlight the x2 = 2 cases. Then from the *Case deletions* menu, select *Delete selection from data set*. From *Graph&Fit* select *Fit linear LS*, select x1, $x1^2$, x2 as the terms and y as the response. Include the output in *Word*. The output suggests that the MA brand is about 320 psi less strong than the ME brand. (May need to add x2*x1 and x2*$x1^2$ interactions.)

h) Include the response plot and residual plot in *Word*. The residual plot is not particularly good.

3.16. This problem gives a slightly simpler model than Problem 3.15 by using the indicator variable x3 = 1 if standard cement (if x2 = 2) and x3 = 0 otherwise (if x2 is 0 or 1). Activate the *cement.lsp* data.

a) From the *cement* menu, select Transform, select x1, and place a 2 in the p box. This should add $x1^2$ to the data set. From the *cement* menu, select *Make interactions* and select x1 and x3.

b) From *Graph&Fit* select *Fit linear LS*, select x1, $x1^2$, x3 and x1*x3 as the terms and y as the response. Include the output in *Word*.

c) Make the response and residual plots. When making these plots, place x2 in the *Mark by* box. Include the plots in *Word*. Does the model seem ok?

3.17*. Get the McDonald and Schwing (1973) data *pollution.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk. Activate the *pollution.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > pollution.lsp." Scroll up the screen to read the data description. Often simply using the log rule on the predictors with $\max(x)/\min(x) > 10$ works wonders.

a) Make a scatterplot matrix of the first nine predictor variables and the response *Mort*. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select DENS, EDUC, HC, HOUS, HUMID,

JANT, JULT, NONW, NOX and MORT. Then click on *OK*.

A scatterplot matrix with slider bars will appear. Move the slider bars for NOX, NONW and HC to 0, providing the log transformation. In *Arc*, the diagonals have the min and max of each variable, and these were the three predictor variables satisfying the log rule. Open *Word.*

In *Arc*, use the menu commands "Edit > Copy." In *Word*, use the menu commands "Edit > Paste." This should copy the scatterplot matrix into the *Word* document. Print the graph.

b) Make a scatterplot matrix of the last six predictor variables and the response *Mort*. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select OVR65, POOR, POPN, PREC, SO, WWDRK and MORT. Then click on *OK*. Move the slider bar of SO to 0 and copy the plot into *Word*. Print the plot as described in a).

c) Click on the *pollution* menu and select *Transform*. Click on the *log transformations* button and select HC, NONW, NOX and SO. Click on *OK*.

Then fit the full model with the menu commands "Graph&Fit > Fit linear LS". Select MORT for the response. For the terms, select DENS, EDUC, log[HC], HOUS, HUMID, JANT, JULT, log[NONW], log[NOX], OVR65, POOR, POPN, PREC, log[SO] and WWDRK. Click on *OK*.

This model is the full model. To make the response plot use the menu commands "Graph&Fit >Plot of". Select MORT for the V-box and L1:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 1 to add the identity line. Copy the plot into *Word.*

To make the residual plot use the menu commands "Graph&Fit >Plot of". Select L1:Residuals for the V-box and L1:Fit-Values for the H-box. Click on *OK*. Copy the plot into *Word.* Print the two plots.

d) Using the "L1" menu, select "Examine submodels" and try forward selection. Using the "L1" menu, select "Examine submodels" and try backward elimination. You should get a lot of output including that shown in Example 3.7.

Fit the submodel with the menu commands "Graph&Fit > Fit linear LS". Select MORT for the response. For the terms, select EDUC, JANT, log[NONW], log[NOX], and PREC. Click on *OK*.

This model is the submodel suggested by backward elimination. To make the response plot use the menu commands "Graph&Fit >Plot of". Select MORT for the V-box and L2:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 1 to add the identity line.

Copy the plot into *Word.*

To make the residual plot use the menu commands "Graph&Fit >Plot of". Select L2:Residuals for the V-box and L2:Fit-Values for the H-box. Click on *OK*. Copy the plot into *Word.* Print the two plots.

e) To make an RR plot use the menu commands "Graph&Fit >Plot of". Select L1:Residuals for the V-box and L2:Residuals for the H-box. Click on *OK*. Move the OLS slider bar to one. On the window for the plot, click on *Options*. A window will appear. Type $y = x$ and click on *OK* to add the identity line. Copy the plot into *Word.* Print the plot.

f) To make an FF plot use the menu commands "Graph&Fit >Plot of". Select L1:Fit-Values for the V-box and L2:Fit-Values for the H-box. Click on *OK*. Move the OLS slider bar to one and click on *OK* to add the identity line. Copy the plot into *Word.*

g) Using the response and residual plots from the full model and submodel along with the RR and FF plots, does the submodel seem ok?

3.18. Get the Joanne Numrich data *c12.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk. Activate the *c12.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > c12.lsp." Scroll up the screen to read the data description. This data set is described in Example 3.10.

a) A bad model uses Y_1 and all 24 nontrivial predictors. There are many indicator variables. Click on the *CLA* menu and select *Transform.* Click on the *log transformations* button and select y_1. Click on *OK*.

b) Use the menu commands "Graph&Fit > Fit linear LS". Select log[y1] for the response. For the terms, select x1, x2, x8, x9, x10, x11, x18, x20, x23 and x24. Click on *OK*.

This model will be used as the full model. To make the response plot use the menu commands "Graph&Fit >Plot of". Select log[y1] for the V-box and L1:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 1 to add the identity line. Copy the plot into *Word.*

To make the residual plot use the menu commands "Graph&Fit >Plot of". Select L1:Residuals for the V-box and L1:Fit-Values for the H-box. Click on *OK*. Copy the plot into *Word.* Print the two plots.

c) As in Problem 3.17, use forward selection, backward elimination and plots to find a good submodel.

Using material learned in Chapters 2–3, analyze the data sets described in **Problems 3.19–3.29**. 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 $p = r$). (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.)

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 I_{min} that minimizes $C_p(I)$, find the model I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$ (it is possible that $I_I = I_{min}$), and find the smallest value of k such that $C_p(I) \leq \min(p, 2k)$. Model I_I often has too many terms while the 2nd model often 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). Also include a residual plot.

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.

3.19. For the file *bodfat.lsp*, described in Problem 2.2, use $Z = Y = bodyfat$ but do not use $X_1 = density$ as a predictor in the full model. You may use the remaining 13 nontrivial predictor variables. Do parts a), b) and c) above.

3.20*. For the file *boston2.lsp*, described in Examples 15.6 and 15.7 use $Z = (y =)$ CRIM. Do parts a), b) and c) above Problem 3.19.

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

3.21*. For the file *major.lsp*, described in Example 2.3, use $Z = Y$. Do parts a), b) and c) above Problem 3.19.

Note: there are 1 or more outliers that affect numerical methods of vari-

able selection.

3.22. 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 3.19.

3.23*. For the file *museum.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 3.19.

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.

3.24*. For the file *pop.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 3.19.

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.

3.25*. 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 3.19.

3.26*. 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 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 3.19.

c) Do part c) above Problem 3.19.

3.27. 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 3.11 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.

3.28. This is like Problem 3.27 with the BigMac data. Assume that a multiple linear regression model holds for $Y = t(Z)$ 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" Z
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.

3.29*. 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 cursor 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.)

Variable Selection in SAS

3.30. Copy and paste the *SAS* program for this problem from (www.math.siu.edu/olive/reghw.txt) into the *SAS* editor. Then perform the menu commands "Run>Submit" to obtain about 15 pages of output. Do not print out the output.

The key *SAS* code is shown below.

```
proc reg data=fitness;
    model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
    output out =a p = pred r = resid;
      model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
            / selection=forward;
      model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
            / selection=backward;
      model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
            / selection=cp best = 10;

proc rsquare cp data = fitness;
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;

proc plot data = a;
    plot resid*(pred);
    plot Oxygen*pred;

proc reg data=fitness;
    model Oxygen=Age  RunTime RunPulse MaxPulse;
    output out =sub p = pred r = resid;

proc plot data = sub;
    plot resid*(pred);
    plot Oxygen*pred;
run;
```

The data is from SAS Institute (1985, p. 695-704, 717-718). Aerobic fitness is being measured by the ability to consume oxygen. The response $Y = Oxygen$ (uptake rate) is expensive to measure, and it is hoped that the OLS $\hat{Y}$ can be used instead. The variables are *Age* in years, *Weight* in kg, *RunTime* = time in minutes to run 1.5 miles, *RunPulse* = heart rate

when Y is measured, *RestPulse* = heart rate while running and *MaxPulse* = maximum heart rate recorded while running.

The *selection* commands do forward selection, backward elimination and all subset selection where the best ten models with the lowest C_p are recorded. The proc rsquare command also does all subsets regression with the C_p criterion.

The plots give the response and residual plots for the full model and the submodel that used *Age, RunTime, RunPulse, MaxPulse* and a constant.

a) Was the above plot for the minimum C_p model?

b) Do the plots suggest that the submodel was good?

Variable Selection in Minitab

3.31. Get the data set *prof.mtb* as described in Problem 2.15. The data is described in McKenzie and Goldman (1999, p. ED-22-ED-23). Assign the response variable to be *instrucr* (the instructor rating from course evaluations) and the predictors to be *interest* in the course, *manner* of the instructor, and *course* = rating of the course.

a) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands "Stat>Regression>Regression" to get the regression window. Put *instrucr* in the **Response** and *interest*, *manner* and *course* in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

b) To get a response plot, use the commands "Graph>Plot," (double click) place *instrucr* in the **Y** box, and *Fits1* in the **X** box. Then click on **OK**. Print the plot by clicking on the graph and then clicking on the printer icon.

c) To make a residual plot, use the menu commands "Graph>Plot" to get a window. Place "Resi1" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

d) To perform all subsets regression, use the menu commands "Stat>Regression>Best Subsets" to get the regression window. Put *instrucr* in the **Response** and *interest*, *manner* and *course* in the **Free predictors** boxes. Which submodel is good?