Chapter 5

Multiple Linear Regression

In the multiple linear regression model,

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

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

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

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

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

Often the first column of $\boldsymbol{X}$ is $X_1 \equiv \boldsymbol{x}^1 = \mathbf{1}$, the $n \times 1$ vector of ones. The ith case $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. If the e_i are iid with zero mean and variance σ^2, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

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

Most regression methods attempt to find an estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals where the ith residual

$r_i(\boldsymbol{b}) = r_i = Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = Y_i - \hat{Y}_i$. The order statistics for the absolute residuals are denoted by

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

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

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

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

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

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

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

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

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

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

Definition 5.5. The *iid constant variance symmetric error model* uses the assumption that the errors $e_1, ..., e_n$ are iid with a pdf that is symmetric about zero and $\text{VAR}(e_1) = \sigma^2 < \infty$.

In the location model, $\hat{\boldsymbol{\beta}}_{OLS} = \overline{Y}$, $\hat{\boldsymbol{\beta}}_{L_1} = \text{MED}(n)$ and the Chebyshev estimator is the *midrange* $\hat{\boldsymbol{\beta}}_{L_\infty} = (Y_{(1)} + Y_{(n)})/2$. These estimators are simple

to compute, but computation in the multiple linear regression case requires a computer. Most statistical software packages have OLS routines, and the L_1 and Chebyshev fits can be efficiently computed using linear programming. The L_1 fit can also be found by examining all

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

subsets of size p where $n! = n(n-1)(n-2)\cdots 1$ and $0! = 1$. The Chebyshev fit to a sample of size $n > p$ is also a Chebyshev fit to some subsample of size $h = p+1$. Thus the Chebyshev fit can be found by examining all $C(n, p+1)$ subsets of size $p+1$. These two combinatorial facts will be very useful for the design of high breakdown regression algorithms described in Chapters 7 and 8.

5.1 A Graphical Method for Response Transformations

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

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

$$t_{\lambda_o}(Y_i) \equiv Y_i^{(\lambda_o)} = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{5.6}$$

If λ_o was known, then $Z_i = Y_i^{(\lambda_o)}$ 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 and symmetric about 0. A frequently used family of transformations is given in the following definition.

Definition 5.6. Assume that **all** of the values of the response variable Y_i are **positive**. Then the *power transformation family*

$$t_\lambda(Y_i) \equiv Y_i^{(\lambda)} = \frac{Y_i^\lambda - 1}{\lambda} \tag{5.7}$$

for $\lambda \neq 0$ and $Y_i^{(0)} = \log(Y_i)$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as $\Lambda_c = \{0, \pm 1/4, \pm 1/3, \pm 1/2, \pm 2/3, \pm 1\}$. This family is a special case of the response transformations considered by Tukey (1957).

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_c, then sometimes $\hat{\lambda}_n$ will converge (eg ae) to $\lambda^* \in \Lambda_c$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable.

This section follows Cook and Olive (2001) closely and proposes a graphical method for assessing response transformations under model (5.6). The appeal of the proposed method rests with its simplicity and its ability to show the transformation against the background of the data. The method uses the two plots defined below.

Definition 5.7. An FFλ plot is a scatterplot matrix of the fitted values $\hat{Y}^{(\lambda_j)}$ for $j = 1, ..., 5$ where $\lambda_1 = -1$, $\lambda_2 = -0.5$, $\lambda_3 = 0$, $\lambda_4 = 0.5$ and $\lambda_5 = 1$. These fitted values are obtained by OLS regression of $Y^{(\lambda_i)}$ on the predictors. For $\lambda_5 = 1$, we will usually replace $\hat{Y}^{(1)}$ by $\hat{Y}$ and $Y^{(1)}$ by Y.

Definition 5.8. For a given value of $\lambda \in \Lambda_c$, a *transformation plot* is a plot of $\hat{Y}$ versus $Y^{(\lambda)}$. Since $Y^{(1)} = Y - 1$, we will typically replace $Y^{(1)}$ by Y in the transformation plot.

Remark 5.1. Our convention is that a plot of W versus Z means that W is on the horizontal axis and Z is on the vertical axis. We may add fitted OLS lines to the transformation plot as visual aids.

Application 5.1. Assume that model (5.6) is a useful approximation of the data for some $\lambda_o \in \Lambda_c$. Also assume that each subplot in the FFλ plot is strongly linear. To estimate $\lambda \in \Lambda_c$ graphically, make a transformation plot for each $\lambda \in \Lambda_c$. If the transformation plot is linear for $\tilde{\lambda}$, then $\hat{\lambda}_o = \tilde{\lambda}$. (If more than one transformation plot is linear, contact subject matter experts

and use the simplest or most reasonable transformation.)

By "strongly linear" we mean that a line from simple linear regression would fit the plotted points very well, with a correlation greater than 0.95. We introduce this procedure with the following example.

Example 5.1: 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 Y 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. This remark applies also to many of the diagnostic methods for response transformations in the literature. For example, the influence diagnostics studied by Cook and Wang (1983) and others are largely numerical.

To use the graphical method, we first check the assumption on the FFλ plot. Figure 5.1 shows the FFλ plot meets the assumption. The smallest sample correlation among the pairs in the scatterplot matrix is about 0.9995. Shown in Figure 5.2 are transformation plots of $\hat{Y}$ versus $Y^{(\lambda)}$ for four values of λ. 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 5.2a to form along a linear scatter in Figure 5.2c. 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.

The next example illustrates that the transformation plots can show characteristics of data that might influence the choice of a transformation by the usual Box–Cox procedure.

Example 5.2: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The response is *muscle mass* M in grams, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log W$ of the *shell width* W,

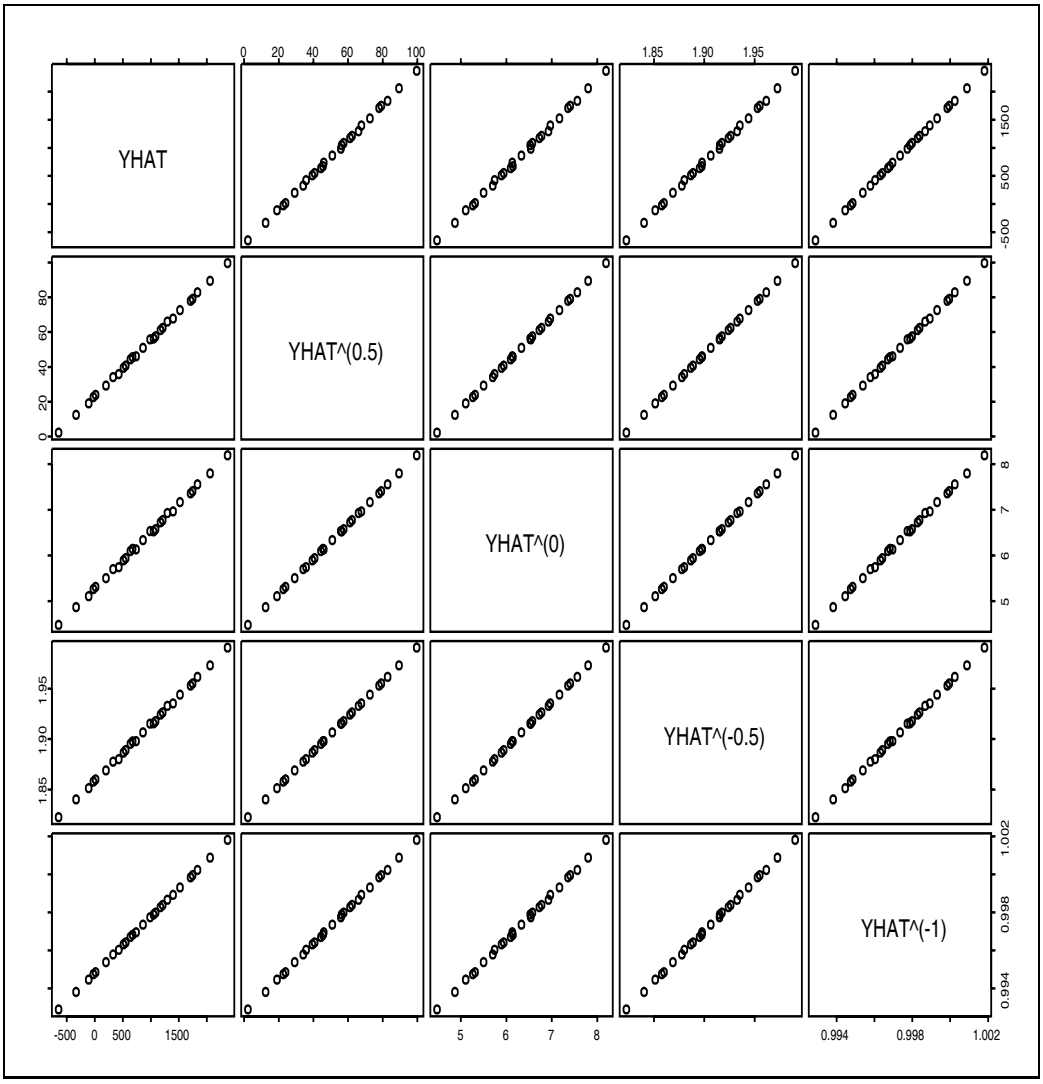

Figure 5.1: FFλ Plot for the Textile Data

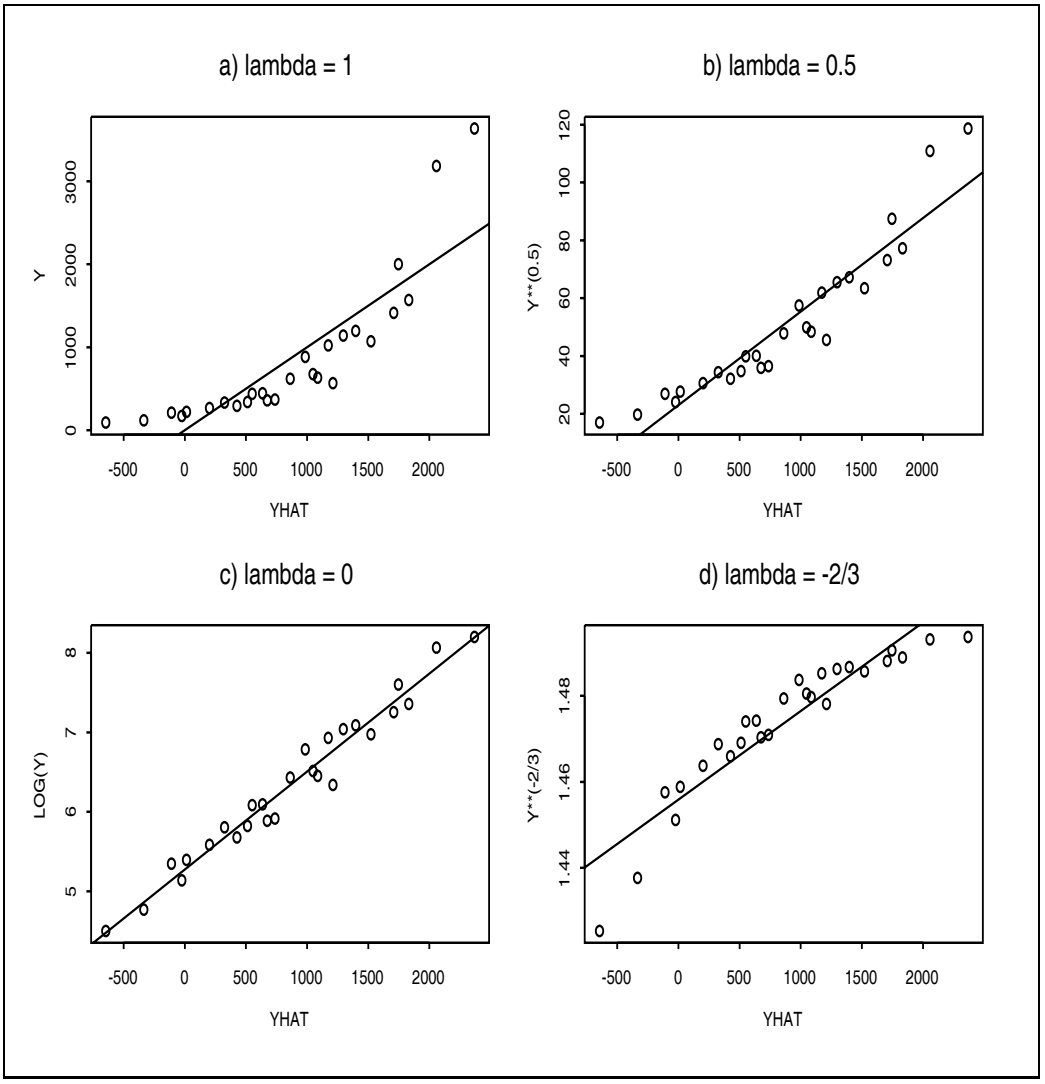

Figure 5.2: Four Transformation Plots for the Textile Data

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.

The FFλ plot (not shown, but very similar to Figure 5.1) exhibits strong linear relations, the correlations ranging from 0.9716 to 0.9999. Shown in Figure 5.3 are transformation plots of $Y^{(\lambda)}$ versus $\hat{Y}$ 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 OLS line in Figure 5.3c, 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 plots of $\hat{Y}$ versus $Y^{(\lambda)}$ indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

The essential point of this example is that observations that influence the choice of power transformation are often easily identified in a transformation plot of $\hat{Y}$ versus $Y^{(\lambda)}$ when the FFλ subplots are strongly linear.

The easily verified assumption that there is strong linearity in the FFλ plot is needed since if $\lambda_o \in [-1, 1]$, then

$$\hat{Y}^{(\lambda)} \approx c_\lambda + d_\lambda \hat{Y}^{(\lambda_o)} \tag{5.8}$$

for all $\lambda \in [-1, 1]$. Consequently, for any value of $\lambda \in [-1, 1]$, $\hat{Y}^{(\lambda)}$ is essentially a linear function of the fitted values $\hat{Y}^{(\lambda_o)}$ for the true λ_o, although we do not know λ_o itself. However, to estimate λ_o graphically, we could select any fixed value $\lambda^* \in [-1, 1]$ and then plot $\hat{Y}^{(\lambda^*)}$ versus $Y^{(\lambda)}$ for several values of λ and find the $\lambda \in \Lambda_c$ for which the plot is linear with constant variance. This simple graphical procedure will then work because a plot of $\hat{Y}^{(\lambda^*)}$ versus $Y^{(\lambda)}$ is equivalent to a plot of $c_{\lambda^*} + d_{\lambda^*}\hat{Y}^{(\lambda_o)}$ versus $Y^{(\lambda)}$ by Equation (5.8). Since the plot of $\hat{Y}^{(1)}$ versus $Y^{(\lambda)}$ differs from a plot of $\hat{Y}$ versus $Y^{(\lambda)}$ by a

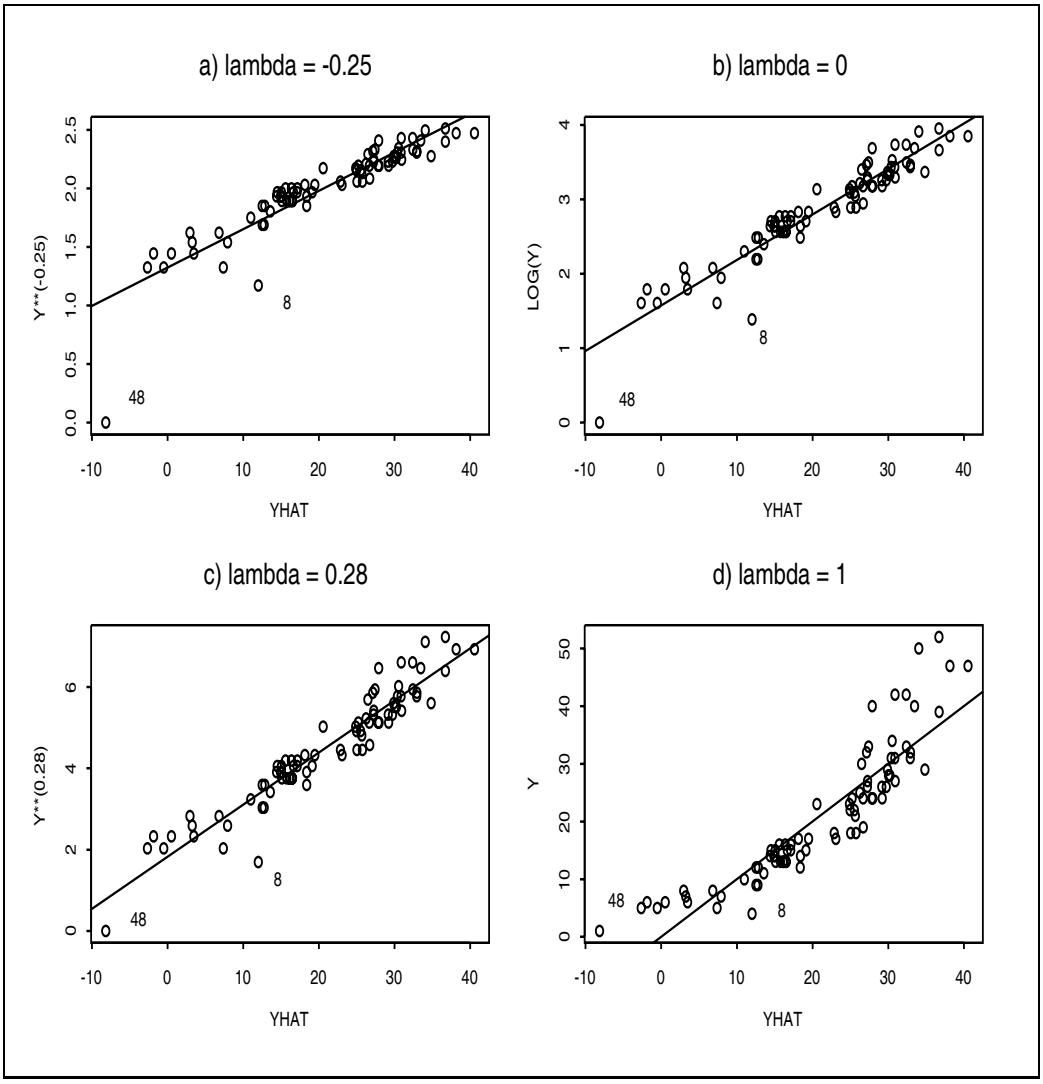

Figure 5.3: Transformation Plots for the Mussel Data

constant shift, we take $\lambda^* = 1$, and use $\hat{Y}$ instead of $\hat{Y}^{(1)}$. By using a single set of fitted values $\hat{Y}$ on the horizontal axis, influential points or outliers that might be masked in plots of $\hat{Y}^{(\lambda)}$ versus $Y^{(\lambda)}$ for $\lambda \in \Lambda_c$ will show up unless they conform on *all* scales.

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 $\lambda \in \Lambda_c$. Example 5.2 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.

There are at least two interesting facts about the strength of the linearity in the FFλ plot. First, the FFλ correlations are frequently all quite high for many data sets when no strong linearities are present among the predictors. Let $\boldsymbol{x} = (x_1, \boldsymbol{w}^T)^T$ where $x_1 \equiv 1$ and let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\eta}^T)^T$. Then $\boldsymbol{w}$ corresponds to the nontrivial predictors. If the conditional predictor expectation $E(\boldsymbol{w}|\boldsymbol{w}^T\boldsymbol{\eta})$ is linear or if $\boldsymbol{w}$ follows an elliptically contoured distribution with second moments, then for *any* λ (not necessarily confined to a selected Λ), the *population* fitted values $\hat{Y}_{\rm pop}^{(\lambda)}$ are of the form

$$\hat{Y}_{\rm pop}^{(\lambda)} = \alpha_\lambda + \tau_\lambda \boldsymbol{w}^T\boldsymbol{\eta} \tag{5.9}$$

so that any one set of population fitted values is an exact linear function of any other set provided the τ_λ's are nonzero. See Cook and Olive (2001). This result indicates that sample FFλ plots will be linear when $E(\boldsymbol{w}|\boldsymbol{w}^T\boldsymbol{\eta})$ is linear, although Equation (5.9) does not by itself guarantee high correlations. However, the strength of the relationship (5.8) can be checked easily by inspecting the FFλ plot.

Secondly, if the FFλ subplots are not strongly linear, and 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**. 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

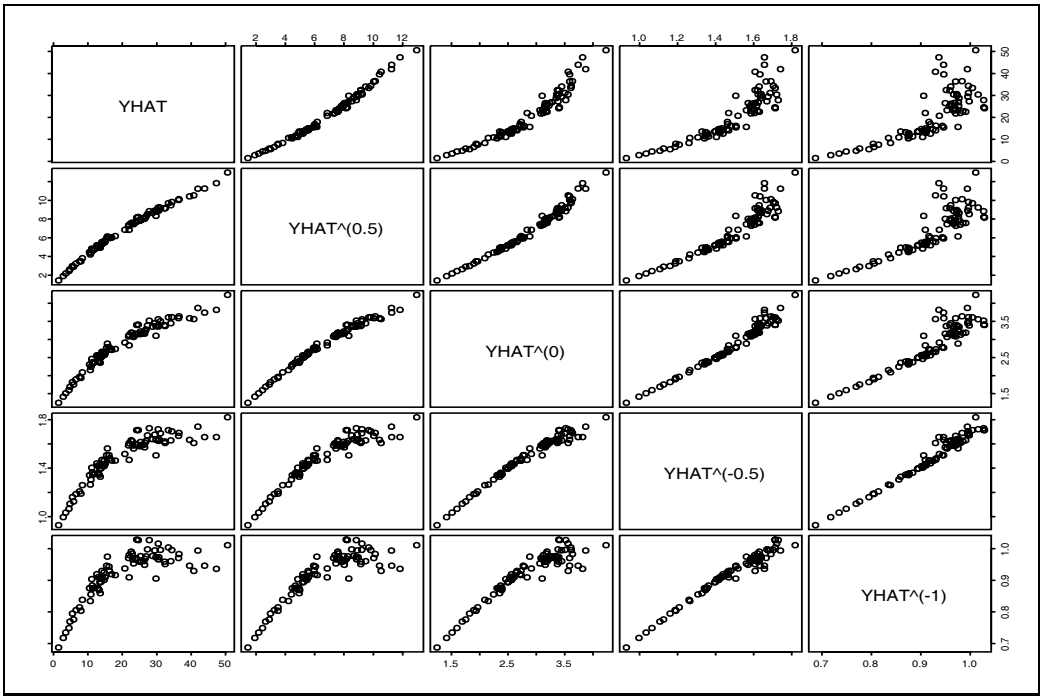

Figure 5.4: FFλ Plot for Mussel Data with Original Predictors

for estimating a power transformation of the response in regression, but the likelihood comes from the assumed multivariate normal distribution of $\boldsymbol{w}^{\boldsymbol{\lambda}}$.) More will be said about predictor transformations in Sections 5.3 and 12.3.

Example 5.3: 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. The FFλ plot for this regression is shown in Figure 5.4. The sample correlations in the plots range between 0.76 and 0.991 and there is notable curvature in some of the plots. Figure 5.5 shows the scatterplot matrix of the predictors L, H, W and S. Again nonlinearity is present. Figure 5.6 shows that taking the log transformations of W and S results in a linear scatterplot matrix for the new set of predictors L, H, $\log W$, and $\log S$. Then the search for the response transformation can be done as in Example 5.2.

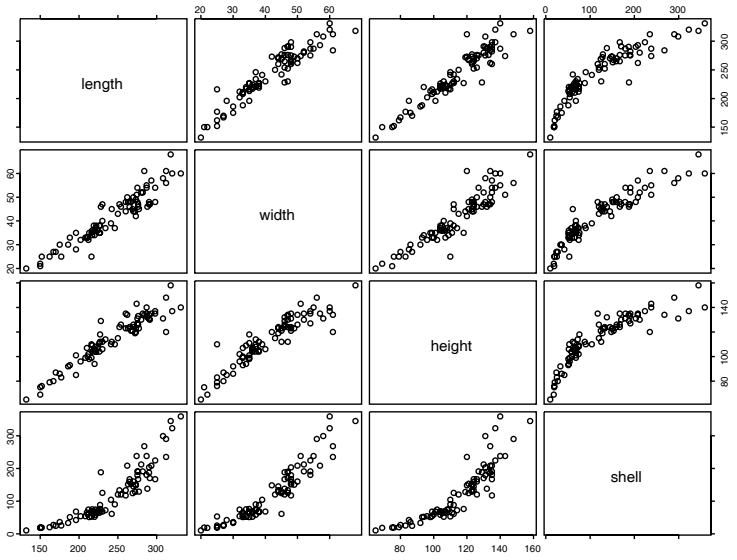

Figure 5.5: Scatterplot Matrix for Original Mussel Data Predictors

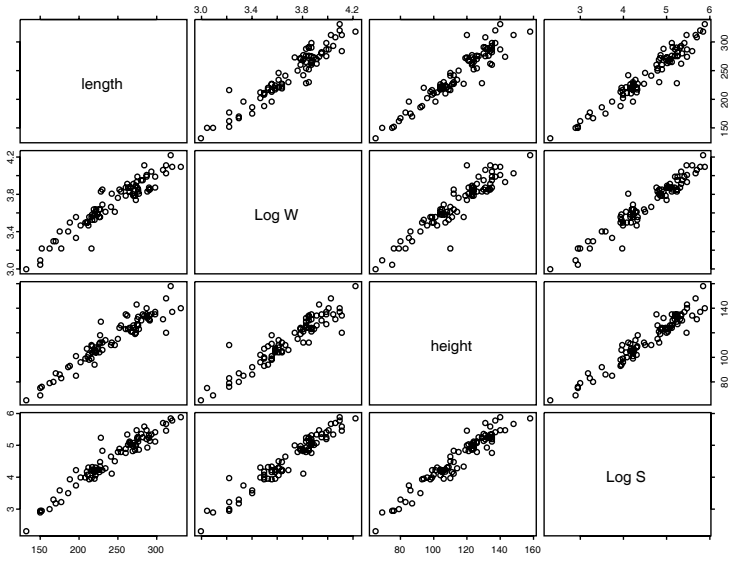

Figure 5.6: Scatterplot Matrix for Transformed Mussel Data Predictors

5.2 Assessing 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. This section follows Olive and Hawkins (2005) closely. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_E^T\boldsymbol{x}_E + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + e \qquad (5.10)$$

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

$$Y = \boldsymbol{\beta}_I^T\boldsymbol{x}_I + \boldsymbol{\beta}_O^T\boldsymbol{x}_O + e. \qquad (5.11)$$

Definition 5.9. The model $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{\beta}_I^T\boldsymbol{x}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The *sufficient predictor* (SP) is the linear combination of the predictor variables used in the model. Hence the full model is $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ and the submodel is $SP = \boldsymbol{\beta}_I^T\boldsymbol{x}_I$.

Notice that the full model is a submodel. The estimated sufficient predictor (ESP) is $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ 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 (5.10) holds. Then

$$SP = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}_S^T\boldsymbol{x}_S = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T\boldsymbol{x}_{I/S} + \boldsymbol{0}^T\boldsymbol{x}_O = \boldsymbol{\beta}_I^T\boldsymbol{x}_I \qquad (5.12)$$

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 $\text{corr}(\boldsymbol{\beta}^T\boldsymbol{x}_i, \boldsymbol{\beta}_I^T\boldsymbol{x}_{I,i}) = 1.0$ for the population model if $S \subseteq I$.

This section proposes a graphical method for evaluating candidate submodels. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i$ 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 $r_{I,i} = Y_i - \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ and $\hat{Y}_{I,i} = \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ where $i = 1, ..., n$. Two important summary statistics for a multiple linear regression model are R^2, the proportion of the variability of Y explained by the nontrivial predictors in the model, and the estimate $\hat{\sigma}$ of the error standard deviation σ.

Definition 5.10. 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.

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 (5.10) holds and that a good estimator 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 5.2. 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. For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the F_I 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. Also recall that

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

where MSE is the error mean square for the full model. Notice that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$. Remark 5.3 below suggests that for subsets I with k terms, submodels with $C_p(I) \leq 2k$ are especially interesting.

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 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 5.1. 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 $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R_I^2$ and intercept $a = \overline{Y}(1 - R_I^2)$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and R_I^2 is the coefficient of multiple determination from the candidate model.
FF 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 $b = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.
v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
RR Plot: vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\text{corr}(r, r_I)]^2$ and

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

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

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

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 $[\text{corr}(z, w)]^2$.

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

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

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),

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

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

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

Remark 5.2. Note that for large n, $C_p(I) < k$ or $F_I < 1$ will force corr(ESP,ESP(I)) to be high. If the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ are not the OLS estimators, the plots will be similar to the OLS plots if the correlation of the fitted values from OLS and the alternative estimators is high (≥ 0.95).

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 5.3. 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 5.1 vi) implies that if $C_p(I) \leq k$ then corr(r, r_I) and 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 corr(r, r_I). Proposition 5.1 vi) implies that if

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

then 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, use the $C_p = k$ line for large values of k, but also consider models close to or under the $C_p = 2k$ line when $k \leq p/2$.

Example 5.4. 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

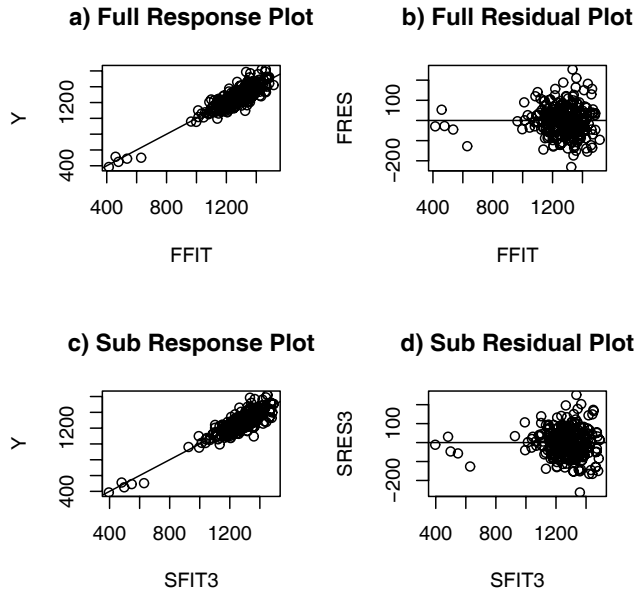

Figure 5.7: Gladstone data: comparison of the full model and the submodel.

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 5.7 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 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

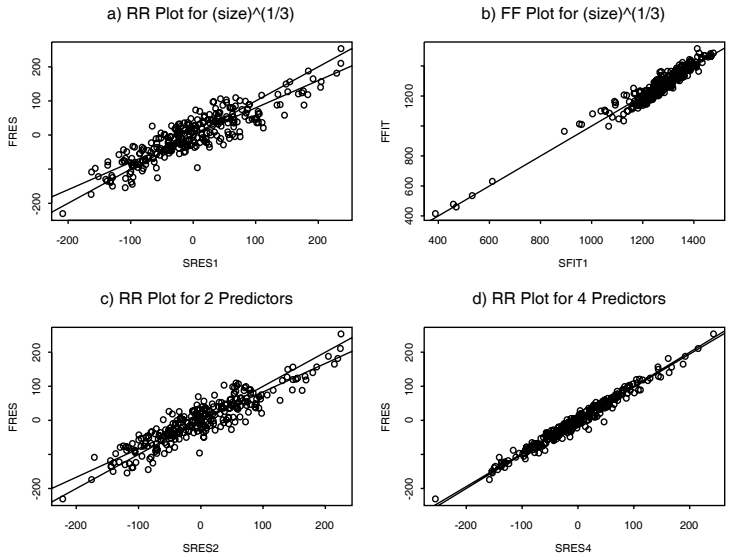

Figure 5.8: Gladstone data: submodels added $(size)^{1/3}$, sex, age and finally $breadth$.

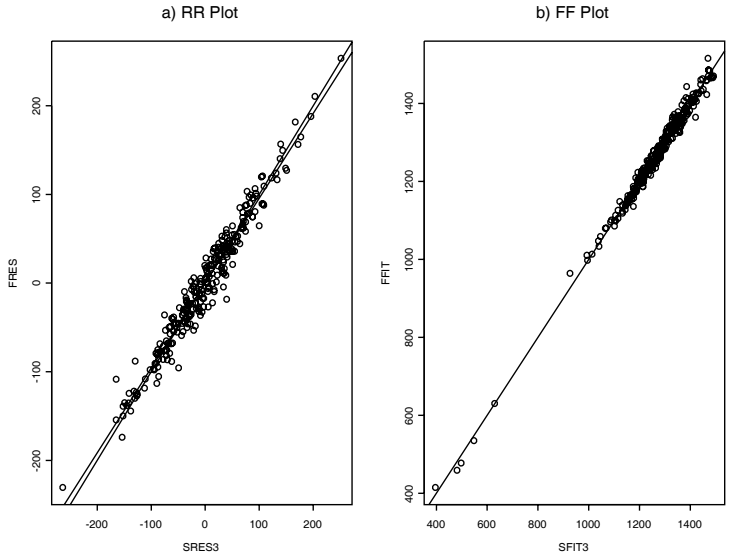

Figure 5.9: Gladstone data with Predictors $(size)^{1/3}$, sex, and age

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 5.8d 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 5.8b) was very high, but in the RR plot the OLS line did not coincide with the identity line (Figure 5.8a). Next *sex* was added to I, but again the OLS and identity lines did not coincide in the RR plot (Figure 5.8c). Hence *age* was added to I. Figure 5.9a 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 5.9b 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 5.5. 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 FF and RR plots are shown in Figure 5.10. The identity line and OLS lines were added to the plots as visual aids. The FF plot shows one outlier, the third case, that is clearly separated

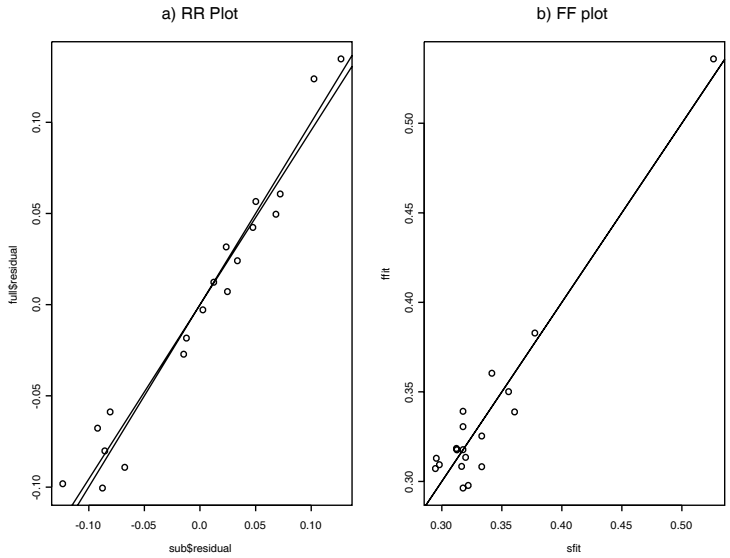

Figure 5.10: FF and RR Plots for Rat Data

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 5.11 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. For the MLR model, standard case diagnostics

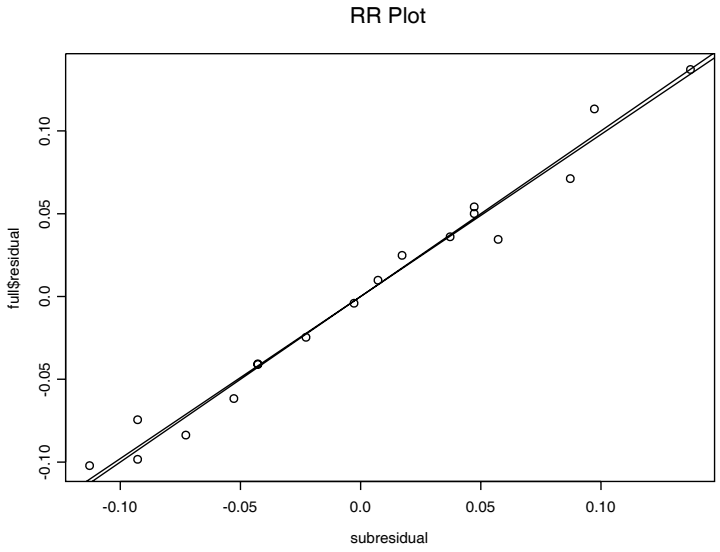

Figure 5.11: RR Plot With Outlier Deleted, Submodel Uses No Predictors

are the full model residuals r_i and the Cook's distances

$$\mathrm{CD}_i = \frac{r_i^2}{p\hat{\sigma}^2(1-h_i)}\frac{h_i}{(1-h_i)}, \tag{5.13}$$

where h_i is the leverage and $\hat{\sigma}^2$ is the usual estimate of the error variance. (See Chapter 6 for more details about these quantities.)

Definition 5.11. The *RC plot* is a plot of the residuals r_i versus the Cook's distances CD_i.

Though two-dimensional, the RC plot shows cases' residuals, leverage, and influence together. Notice that cases with the same leverage define a parabola in the RC plot. In an ideal setting with no outliers or undue case leverage, the plotted points should have an evenly-populated parabolic shape. This leads to a graphical approach of making the RC plot, temporarily deleting cases that depart from the parabolic shape, refitting the model and regenerating the plot to see whether it now conforms to the desired shape.

The cases deleted in this approach have atypical leverage and/or deviation. Such cases often have substantial impact on numerical variable selection methods, and the subsets identified when they are excluded may be

very different from those using the full data set, a situation that should cause concern.

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 both prediction and description.

A thorough subset selection analysis will use the RC plots in conjunction with the more standard numeric-based algorithms. This suggests running the numerical variable selection procedure on the entire data set and on the "cleaned data" set with the influential cases deleted, keeping track of interesting models from both data sets. For a candidate submodel I, let $C_p(I, c)$ denote the value of the C_p statistic for the cleaned data. The following two examples help illustrate the procedure.

Example 5.6. Ashworth (1842) presents a data set of 99 communities in Great Britain. The response variable Y = *log(population in 1841)* and the predictors are x_1, x_2, x_3 and a constant where x_1 is log(*property value in pounds in 1692*), x_2 is log(*property value in pounds in 1841*), and x_3 is the log(*percent rate of increase in value*). The initial RC plot, shown in Figure 5.12a, is far from the ideal of an evenly-populated parabolic band. Cases 14 and 55 have extremely large Cook's distances, along with the largest residuals. After deleting these cases and refitting OLS, Figure 5.12b shows that the RC plot is much closer to the ideal parabolic shape. If case 16 had a residual closer to zero, then it would be a very high leverage case and would also be deleted.

Table 5.1 shows the summary statistics of the fits of all subsets using all cases, and following the removal of cases 14 and 55. The two sets of results are substantially different. On the cleaned data the submodel using just x_2 is the unique clear choice, with $C_p(I, c) = 0.7$. On the full data set however, none of the subsets is adequate. Thus cases 14 and 55 are responsible for all indications that predictors x_1 and x_3 have any useful information about Y. This is somewhat remarkable in that these two cases have perfectly ordinary values for all three variables.

Example 5.4 (continued). Now we will apply the RC plot to the Gladstone data using Y = *brain weight*, x_1 = *age*, x_2 = *height*, x_3 = *head height*, x_4 = *head length*, x_5 = *head breadth*, x_6 = *head circumference*, x_7 = *cephalic index*, x_8 = *sex*, and $x_9 = (size)^{1/3}$. All submodels contain a constant.

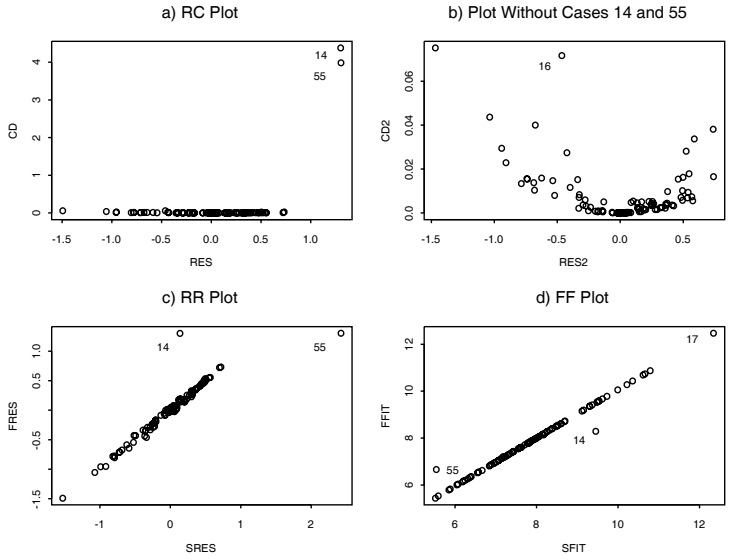

Figure 5.12: Plots for the Ashworth Population Data

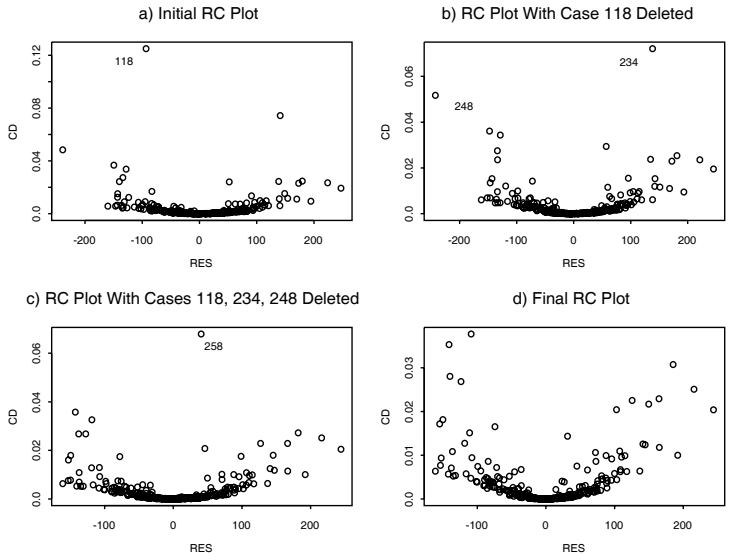

Figure 5.13: RC Plots for the Gladstone Brain Data

Table 5.1: Exploration of Subsets – Ashworth Data

		All cases		2 removed	
Subset I	k	SSE	$C_p(I)$	SSE	$C_p(I,c)$
x_1	2	93.41	336	91.62	406
x_2	2	23.34	12.7	17.18	0.7
x_3	2	105.78	393	95.17	426
x_1, x_2	3	23.32	14.6	17.17	2.6
x_1, x_3	3	23.57	15.7	17.07	2.1
x_2, x_3	3	22.81	12.2	17.17	2.6
All	4	20.59	4.0	17.05	4.0

Table 5.2: Some Subsets – Gladstone Brain Data

		All cases		Cleaned data	
Subset I	k	SSE $\times 10^3$	$C_p(I)$	SSE$\times 10^3$	$C_p(I,c)$
x_1, x_9	3	1486	12.6	1352	10.8
x_8, x_9	3	1655	43.5	1516	42.8
x_1, x_8, x_9	4	1442	6.3	1298	2.3
x_1, x_5, x_9	4	1463	10.1	1331	8.7
x_1, x_5, x_8, x_9	5	1420	4.4	1282	1.2
All	10	1397	10.0	1276	10.0

Table 5.2 shows the summary statistics of the more interesting subset regressions. The smallest C_p value came from the subset x_1, x_5, x_8, x_9, and in this regression x_5 has a t value of -2.0. Deleting a single predictor from an adequate regression changes the C_p by approximately $t^2 - 2$, where t stands for that predictor's Student's t in the regression – as illustrated by the increase in C_p from 4.4 to 6.3 following deletion of x_5. Analysts must choose between the larger regression with its smaller C_p but a predictor that does not pass the conventional screens for statistical significance, and the smaller, more parsimonious, regression using only apparently statistically significant predictors, but (as assessed by C_p) possibly less accurate predictive ability.

Figure 5.13 shows a sequence of RC plots used to identify cases 118, 234, 248 and 258 as atypical, ending up with an RC plot that is a reasonably

Table 5.3: 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	

evenly-populated parabolic band. Using the C_p criterion on the cleaned data suggests the same final submodel I found earlier – that using a constant, $x_1 = age$, $x_8 = sex$ and $x_9 = size^{1/3}$.

The five cases (230, 254, 255, 256 and 257) corresponding to the five infants were well separated from the bulk of the data and have higher leverage than average, and so good exploratory practice would be to remove them also to see the effect on the model fitting. The right columns of Table 5.2 reflect making these 9 deletions. As in the full data set, the subset x_1, x_5, x_8, x_9 gives the smallest C_p, but x_5 is of only modest statistical significance and might reasonably be deleted to get a more parsimonious regression. What is striking after comparing the left and right columns of Table 5.2 is that, as was the case with the Ashworth data set, the adequate C_p values for the cleaned data set seem substantially smaller than their full-sample counterparts: 1.2 versus 4.4, and 2.3 versus 6.3. Since these C_p for the same p are dimensionless and comparable, this suggests that the 9 cases removed are primarily responsible for any additional explanatory ability in the 6 unused predictors.

Multiple linear regression data sets with cases that influence numerical variable selection methods are common. Table 5.3 shows results for seven interesting data sets. The first two rows correspond to the Ashworth data in Example 5.6, the next 2 rows correspond to the Gladstone Data in Example 5.4, 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 as files `pop.lsp, cbrain.lsp, cyp.lsp` and `major.lsp`. 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 5.5. Rows 11 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 5.3. 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 5.4. For the rat data, the final submodel is the one given in Example 5.5: none of the 3 nontrivial predictors was used.

Table 5.3 and simulations suggest that if the subset I has k predictors, then using the $C_p(I) \leq 2k$ 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.

5.3 Asymptotically Optimal Prediction Intervals

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

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

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

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

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

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

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

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

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

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

Definition 5.12. Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$

for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

Rule of thumb 5.1. Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$.

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

Definition 5.13. Consider the iid error MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$. Then **regression function** is the hyperplane

$$E(Y) \equiv E(Y|\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{5.14}$$

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

$$\hat{Y}_f = x_{f,1}\hat{\beta}_1 + \cdots + x_{f,p}\hat{\boldsymbol{\beta}}_p = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}. \tag{5.15}$$

The **point estimator** of $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$ given $\boldsymbol{x} = \boldsymbol{x}_f$ is also $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$. Assume that the MLR model contains a constant β_1 so that $x_1 \equiv 1$. The large sample 100 $(1 - \alpha)\%$ confidence interval (CI) for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{1-\alpha/2,n-p}se(\hat{Y}_f) \tag{5.16}$$

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

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

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

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

Theorem 5.2: Huber (1981, p. 157-160). Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

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

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

Definition 5.14. A large sample $100(1-\alpha)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\alpha$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable Y_f is independent of $Y_1, ..., Y_n$. Then the 100 $(1-\alpha)\%$ PI for Y_f is

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

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

$$se(pred) = \sqrt{MSE\ (1+h_f)}.$$

The interpretation of a 100 $(1-\alpha)$% PI for a random variable Y_f is similar to that of a CI. Collect data, then form the PI, and repeat for a total of k times where k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for m of the PIs follows a binomial$(k, \rho = 1 - \alpha)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number L, say. Secondly, the CI for $E(Y_f|\boldsymbol{x}_f)$ given in Definition 5.13 tends to work well for the iid error MLR model if the sample size is large while the PI in Definition 5.14 is made under the assumption that the e_i are iid $N(0, \sigma^2)$ and may not perform well if the normality assumption is violated.

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

Assume that the iid error MLR model is valid so that e is from some distribution with 0 mean and variance σ^2. Olive (2007) shows that if $1 - \delta$ is the asymptotic coverage of the classical nominal $(1-\alpha)100$% PI (5.17), then

$$1 - \delta = P(-\sigma z_{1-\alpha/2} < e < \sigma z_{1-\alpha/2}) \geq 1 - \frac{1}{z^2_{1-\alpha/2}} \tag{5.18}$$

where the inequality follows from Chebyshev's inequality. Hence the asymptotic coverage of the nominal 95% PI is at least 73.9%. The 95% PI (5.17) was often quite accurate in that the asymptotic coverage was close to 95% for a wide variety of error distributions. The 99% and 90% PIs did not perform as well.

Let ξ_α be the α percentile of the error e, ie, $P(e \leq \xi_\alpha) = \alpha$. Let $\hat{\xi}_\alpha$ be the sample α percentile of the residuals. Then the results from Theorem

5.2 suggest that the residuals r_i estimate the errors e_i, and that the sample percentiles of the residuals $\hat{\xi}_\alpha$ estimate ξ_α. For many error distributions,

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

This result suggests that

$$\sqrt{\frac{n}{n-p}} r_i \approx e_i.$$

Using

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}, \tag{5.19}$$

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

$$(\hat{Y}_f + a_n\hat{\xi}_{\alpha/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\alpha/2}). \tag{5.20}$$

This PI is very similar to the classical PI except that $\hat{\xi}_\alpha$ is used instead of σz_α to estimate the error percentiles ξ_α. The large sample coverage $1-\delta$ of this nominal $100(1-\alpha)\%$ PI is asymptotically correct: $1-\delta = 1-\alpha$.

Example 5.8. For the Buxton (1920) data suppose that the response Y = *height* and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five outliers were deleted leaving 82 cases. Figure 5.14 shows a response plot of the fitted values versus the response Y with the identity line added as a visual aid. The plot suggests that the model is good since the plotted points scatter about the identity line in an evenly populated band although the relationship is rather weak since the correlation of the plotted points is not very high. The triangles represent the upper and lower limits of the semiparametric 95% PI (5.20). Notice that 79 (or 96%) of the Y_i fell within their corresponding PI while 3 Y_i did not. A plot using the classical PI (5.17) would be very similar for this data.

When many 95% PIs are made for a single data set, the coverage tends to be higher or lower than the nominal level, depending on whether the difference of the estimated upper and lower percentiles for Y_f is too high or too small. For the classical PI, the coverage will tend to be higher than 95% if se(pred) is too large (MSE $> \sigma^2$), otherwise lower (MSE $< \sigma^2$).

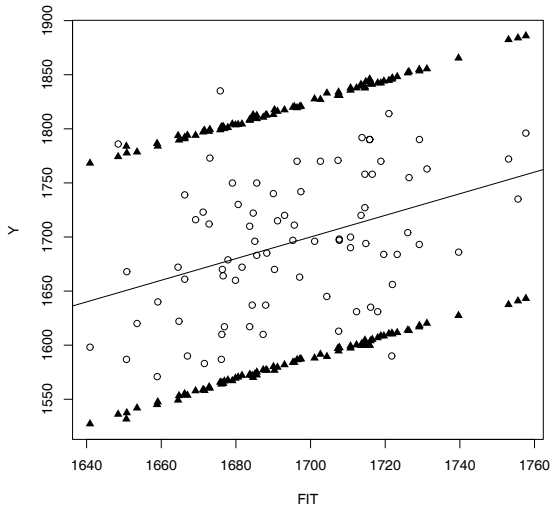

Figure 5.14: 95% PI Limits for Buxton Data

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

Given output showing $\hat{\boldsymbol{\beta}}_i$ and given $\boldsymbol{x}_f$, $se(pred)$ and $se(\hat{Y}_f)$, Example 5.9 shows how to find $\hat{Y}_f$, a CI for $E(Y_f|\boldsymbol{x}_f)$ and a PI for Y_f. Below Figure 5.14 is shown typical output in symbols.

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

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

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

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

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

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

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

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

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

$$\hat{Y}_f \pm \sqrt{\frac{n}{n-p}} \max(|\hat{\xi}_{\alpha/2}|, |\hat{\xi}_{1-\alpha/2}|) \sqrt{(1 + h_f)} \tag{5.21}$$

which has asymptotic coverage

$$1 - \delta = P[-\max(|\xi_{\alpha/2}|, |\xi_{1-\alpha/2}|) < e < \max(|\xi_{\alpha/2}|, |\xi_{1-\alpha/2}|)]. \tag{5.22}$$

Notice that $1-\alpha \leq 1-\delta \leq 1-\alpha/2$ and $1-\delta = 1-\alpha$ if the error distribution is symmetric.

In the simulations described below, $\hat{\xi}_\alpha$ will be the sample percentile for the PIs (5.20) and (5.21). A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\alpha) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\hat{\xi}_{\alpha_1}, \hat{\xi}_{1-\alpha_2})$ correspond to the interval with the smallest distance. Then the 100 $(1-\alpha)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n \hat{\xi}_{\alpha_1}, \hat{Y}_f + b_n \hat{\xi}_{1-\alpha_2}). \tag{5.23}$$

In the simulations, we used $a_n = b_n$ where a_n is given by (5.19).

Table 5.4: N(0,1) Errors

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

Table 5.5: t_3 Errors

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

Table 5.6: Exponential(1) −1 Errors

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

A small simulation study compares the PI lengths and coverages for sample sizes $n = 50, 100$ and 1000 for several error distributions. The value $n = \infty$ gives the asymptotic coverages and lengths. The MLR model with $E(Y_i) = 1 + x_{i2} + \cdots + x_{i8}$ was used. The vectors $(x_2, ..., x_8)^T$ were iid $N_7(\mathbf{0}, \boldsymbol{I}_7)$. The error distributions were N(0,1), t_3, and exponential(1) −1. Also, a small sensitivity study to examine the effects of changing $(1 + 15/n)$ to $(1+k/n)$ on the 99% PIs (5.20) and (5.23) was performed. For $n = 50$ and k between 10 and 20, the coverage increased by roughly 0.001 as k increased by 1.

The simulation compared coverages and lengths of the classical (5.17), semiparametric (5.20), asymptotically conservative (5.21) and asymptotically optimal (5.23) PIs. The latter 3 intervals are asymptotically optimal for symmetric unimodal error distributions in that they have the shortest asymptotic length that gives the desired asymptotic coverage. The semiparametric PI gives the correct asymptotic coverage if the unimodal errors are not symmetric while the PI (5.21) gives higher coverage (is conservative). The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where Y_f fell within the nominal $100(1-\alpha)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1-\delta_n)$ distribution where $1 - \delta_n$ converges to the asymptotic coverage $(1 - \delta)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0014$, 0.0031 and

0.0042 for $p = 0.01, 0.05$ and 0.1, respectively. Hence an observed coverage $\hat{p} \in (.986, .994)$ for 99%, $\hat{p} \in (.941, .959)$ for 95% and $\hat{p} \in (.887, .913)$ for 90% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Tables 5.4–5.6 show the results of the simulations for the 3 error distributions. The letters *c, s, a* and *o* refer to intervals (5.17), (5.20), (5.21) and (5.23) respectively. For the normal errors, the coverages were about right and the semiparametric interval tended to be rather long for $n = 50$ and 100. The classical PI asymptotic coverage $1 - \delta$ tended to be fairly close to the nominal coverage $1 - \alpha$ for all 3 distributions and $\alpha = 0.01, 0.05$, and 0.1.

5.4 A Review of MLR

The **simple linear regression** (SLR) model is $Y_i = \beta_1 + \beta_2 X_i + e_i$ where the e_i are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The parameters β_1, β_2 and σ^2 are **unknown constants** that need to be estimated. (If the X_i are random variables, then the model is conditional on the X_i's. Hence the X_i's are still treated as constants.)

The normal SLR model adds the assumption that the e_i are iid $\text{N}(0, \sigma^2)$. That is, the error distribution is normal with zero mean and constant variance σ^2.

The response variable Y is the variable that you want to predict while the predictor (or independent or explanatory) variable X is the variable used to predict the response.

A **scatterplot** is a plot of W versus Z with W on the horizontal axis and Z on the vertical axis and **is used to display the conditional distribution** of Z given W. For SLR the scatterplot of X versus Y is often used.

For SLR, $E(Y_i) = \beta_1 + \beta_2 X_i$ and the line $E(Y) = \beta_1 + \beta_2 X$ is the regression function. $\text{VAR}(Y_i) = \sigma^2$.

For SLR, the **least squares estimators** $\hat{\beta}_1$ and $\hat{\beta}_2$ minimize the least squares criterion $Q(\eta_1, \eta_2) = \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)^2$. For a fixed η_1 and η_2, Q is the sum of the squared vertical deviations from the line $Y = \eta_1 + \eta_2 X$.

The least squares (OLS) line is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

and $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

By the **chain rule,**

$$\frac{\partial Q}{\partial \eta_1} = -2 \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)$$

and

$$\frac{d^2 Q}{d\eta_1^2} = 2n.$$

Similarly,

$$\frac{\partial Q}{\partial \eta_2} = -2 \sum_{i=1}^{n} X_i (Y_i - \eta_1 - \eta_2 X_i)$$

and

$$\frac{d^2 Q}{d\eta_1^2} = 2 \sum_{i=1}^{n} X_i^2.$$

The OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the **normal equations**:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \text{ and}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2.$$

For SLR, $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ is called the ith fitted value (or predicted value) for observation Y_i while the ith **residual** is $r_i = Y_i - \hat{Y}_i$.

The error (residual) sum of squares $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2$.

For SLR, the mean square error MSE $= SSE/(n-2)$ is an unbiased estimator of the error variance σ^2.

Properties of the OLS line:

i) the residuals sum to zero: $\sum_{i=1}^{n} r_i = 0$.

ii) $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$.

iii) The independent variable and residuals are uncorrelated:

$$\sum_{i=1}^{n} X_i r_i = 0.$$

iv) The fitted values and residuals are uncorrelated: $\sum_{i=1}^{n} \hat{Y}_i r_i = 0$.

v) The least squares line passes through the point $(\overline{X}, \overline{Y})$.

Knowing how to use output from statistical software packages is important. Shown below is an output only using symbols and an actual *Arc* output.

Coefficient Estimates where the Response = Y

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

```
R Squared:                R^2
Sigma hat:                sqrt{MSE}
Number of cases:            n
Degrees of freedom:        n-2

Summary Analysis of Variance Table
Source         df     SS    MS    F           p-value
Regression      1     SSR   MSR  Fo=MSR/MSE   p-value for beta_2
Residual       n-2    SSE   MSE
-----------------------------------------------------------------
Response      = brnweight
Terms         = (size)
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   305.945         35.1814         8.696     0.0000
size       0.271373        0.00986642     27.505     0.0000

R Squared:                0.74058
Sigma hat:                83.9447
Number of cases:             267
Degrees of freedom:          265

Summary Analysis of Variance Table
Source         df       SS            MS           F     p-value
Regression      1  5330898.      5330898.     756.51    0.0000
Residual      265  1867377.       7046.71
```

Let the $p \times 1$ vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and let the $p \times 1$ vector $\boldsymbol{x}_i = (1, X_{i,2}, ..., X_{i,p})^T$. Notice that $X_{i,1} \equiv 1$ for $i = 1, ..., n$. Then the **multiple linear regression** (MLR) model is

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

for $i = 1, ..., n$ where the e_i are iid with $E(e_i) = 0$ and VAR$(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The parameters β_1, β_2, ..., β_p and σ^2 are **unknown constants** that need to be estimated.

In matrix notation, these n equations become

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

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,2} & X_{1,3} & \dots & X_{1,p} \\ 1 & X_{2,2} & X_{2,3} & \dots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,2} & X_{n,3} & \dots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

The first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, the $n \times 1$ vector of ones. The ith case $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. If the e_i are iid with zero mean and variance σ^2, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2. (If the X_i are random variables, then the model is conditional on the X_i's. Hence the X_i's are still treated as constants.)

The normal MLR model adds the assumption that the e_i are iid N$(0, \sigma^2)$. That is, the error distribution in normal with zero mean and constant variance σ^2. Simple linear regression is a special case with $p = 2$.

The response variable Y is the variable that you want to predict while the predictor (or independent or explanatory) variables $X_1, X_2, ..., X_p$ are the variables used to predict the response. Since $X_1 \equiv 1$, sometimes $X_2, ..., X_p$ are called the predictor variables.

For MLR, $E(Y_i) = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and the hyperplane $E(Y) = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \boldsymbol{x}^T \boldsymbol{\beta}$ is the regression function. VAR$(Y_i) = \sigma^2$.

The **least squares estimators** $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ minimize the least squares criterion $Q(\boldsymbol{\eta}) = \sum_{i=1}^{n}(Y_i - \eta_1 - \eta_2 X_{i,2} - \cdots - \eta_p X_{i,p})^2 = \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$. For a fixed $\boldsymbol{\eta}$, Q is the sum of the squared vertical deviations from the hyperplane $H = \eta_1 + \eta_2 X_2 + \cdots + \eta_p X_p$.

The least squares estimator $\hat{\boldsymbol{\beta}}$ satisfies the MLR normal equations

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

and the least squares estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

The vector of *predicted* or *fitted values* is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ where the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The ith entry of $\hat{\boldsymbol{Y}}$ is the ith fitted value (or predicted value) $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{i,2} + \cdots + \hat{\beta}_p X_{i,p} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ for observation Y_i while the ith **residual** is $r_i = Y_i - \hat{Y}_i$. The vector of residuals is $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.

The (residual) error sum of squares $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2$. For MLR, the MSE $= SSE/(n-p)$ is an unbiased estimator of the error variance σ^2.

After obtaining the least squares equation from computer output, **predict** Y for a given $\boldsymbol{x} = (1, X_2, ..., X_p)^T$: $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$.

Know the meaning of the least squares multiple linear regression output. Shown on the next page is an output only using symbols and an actual *Arc* output.

The 100 $(1-\alpha)$ % CI for β_k is $\hat{\beta}_k \pm t_{1-\alpha/2,n-p}\ se(\hat{\beta}_k)$. If $\nu = n - p > 30$, use the N(0,1) cutoff $z_{1-\alpha/2}$. The corresponding 4 step t–test of hypotheses has the following steps, and makes sense if there is no interaction.
i) State the hypotheses Ho: $\beta_k = 0$ Ha: $\beta_k \neq 0$.
ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find the p–value from output or use the t–table: p–value =

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

Use the normal table or $\nu = \infty$ in the t–table if the degrees of freedom $\nu = n - p > 30$.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

Response = Y
Coefficient Estimates

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

```
R Squared:                R^2
Sigma hat:                sqrt{MSE}
Number of cases:            n
Degrees of freedom:         n-p

Summary Analysis of Variance Table
```

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

```
Response      = brnweight
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   99.8495         171.619         0.582     0.5612
size       0.220942        0.0357902       6.173     0.0000
sex        22.5491         11.2372         2.007     0.0458
breadth   -1.24638         1.51386        -0.823     0.4111
circum     1.02552         0.471868        2.173     0.0307

R Squared:                0.749755
Sigma hat:                82.9175
Number of cases:             267
Degrees of freedom:          262

Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      4  5396942.      1349235.     196.24    0.0000
Residual      262  1801333.       6875.32
```

Recall that Ho is rejected if the p–value $< \alpha$. As a benchmark for this textbook, use $\alpha = 0.05$ if α is not given. If Ho is rejected, then conclude that X_k is needed in the MLR model for Y given that the other $p-2$ nontrivial predictors are in the model. If you fail to reject Ho, then conclude that X_k is not needed in the MLR model for Y given that the other $p-2$ nontrivial predictors are in the model. Note that X_k could be a very useful individual predictor, but may not be needed if other predictors are added to the model. It is better to use the output to get the test statistic and p–value than to use formulas and the t–table, but exams may not give the relevant output.

Be able to perform the 4 step ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\beta_2 = \cdots = \beta_p = 0$ Ha: not Ho
ii) Find the test statistic $Fo = MSR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

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

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is a MLR relationship between Y and the predictors X_2, ..., X_p. If you fail to reject Ho, conclude that there is not a MLR relationship between Y and the predictors X_2, ..., X_p.

Be able to find i) the point estimator $\hat{Y}_f = \boldsymbol{x}_f^T \boldsymbol{Y}$ of Y_f given $\boldsymbol{x} = \boldsymbol{x}_f = (1, X_{f,2}, ..., X_{f,p})^T$ and
ii) the 100 $(1-\alpha)$% CI for $\mathrm{E}(Y_f) = \boldsymbol{x}_f^T \boldsymbol{\beta} = E(\hat{Y}_f)$. This interval is $\hat{Y}_f \pm t_{1-\alpha/2,n-p} se(\hat{Y}_f)$. Generally $se(\hat{Y}_f)$ will come from output.

Suppose you want to predict a new observation Y_f where Y_f is independent of $Y_1, ..., Y_n$. Be able to find
i) the point estimator $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$ and the
ii) the 100 $(1-\alpha)$% prediction interval (PI) for Y_f. This interval is $\hat{Y}_f \pm t_{1-\alpha/2,n-p} se(pred)$. Generally $se(pred)$ will come from output. Note that Y_f is a random variable not a parameter.

Full model

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

Reduced model

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

```
Summary Analysis of Variance Table for the Full Model
Source          df       SS            MS           F     p-value
Regression       6   260467.       43411.1      87.41     0.0000
Residual        69   34267.4       496.629

Summary Analysis of Variance Table for the Reduced Model
Source          df       SS            MS           F     p-value
Regression       2   94110.5       47055.3      17.12     0.0000
Residual        73   200623.       2748.27
```

Know how to perform the 4 step **change in SS F test**. Shown is an actual *Arc* output and an output only using symbols. Note that both the full and reduced models must be fit in order to perform the change in SS F test. Without loss of generality, assume that the X_i corresponding to the β_i for $i \geq q$ are the terms to be dropped. Then the **full** MLR model is $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i$ while the **reduced model** is $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_q X_{i,q} + e_i$. Then the change in SS F test has the following 4 steps:
i) Ho: the reduced model is good Ha: use the full model
ii) $F_R =$

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

iii) p–value = $P(F_{df_R - df_F, df_F} > F_R)$. (Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$).
iv) Reject Ho if the p–value $< \alpha$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Given two of SSTO $= \sum_{i=1}^n (Y_i - \overline{Y})^2$, SSE $= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n r_i^2$, and SSR $= \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, find the other sum of squares using the formula SSTO = SSE + SSR.

Be able to find $R^2 = SSR/SSTO =$ (sample correlation of Y_i and $\hat{Y}_i)^2$.

Know i) that the covariance matrix of a random vector $\boldsymbol{Y}$ is $\text{Cov}(\boldsymbol{Y}) = E[(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$.
ii) $E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y})$.
iii) $\text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

Given the least squares model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, be able to show that
i) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and
ii) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

A matrix $\boldsymbol{A}$ is idempotent if $\boldsymbol{AA} = \boldsymbol{A}$.

An **added variable plot** (also called a partial regression plot) is used to give information about the test $Ho : \beta_i = 0$. The points in the plot cluster about a line with slope $= \hat{\beta}_i$. If there is a strong trend then X_i is needed in the MLR for Y given that the other predictors $X_2, ..., X_{i-1}, X_{i+1}, ..., X_p$ are in the model. If there is almost no trend, then X_i may not be needed in the MLR for Y given that the other predictors $X_2, ..., X_{i-1}, X_{i+1}, ..., X_p$ are in the model.

The **response plot** of $\hat{Y}_i$ versus Y is used to check whether the MLR model is appropriate. If the MLR model is appropriate, then the plotted points should cluster about the identity line. The squared correlation $[\text{corr}(Y_i, \hat{Y}_i)]^2 = R^2$. Hence the clustering is tight if $R^2 \approx 1$. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. Know how to decide whether the MLR model is appropriate by looking at a response plot.

The **residual plot** of $\hat{Y}_i$ versus r_i is used to detect departures from the MLR model. If the model is good, then the plot should be ellipsoidal with no trend and should be centered about the horizontal axis. Outliers and patterns such as curvature or a fan shaped plot are bad. Be able to tell a good residual plot from a bad residual plot.

Know that for any MLR, the above two plots should be made.

Other residual plots are also useful. Plot $\boldsymbol{X}_{i,j}$ versus r_i for each nontrivial predictor variable $X_j \equiv \boldsymbol{x}^j$ in the model and for any potential predictors X_j not in the model. Let $r_{[t]}$ be the residual where $[t]$ is the time order of the trial. Hence [1] was the 1st and $[n]$ was the last trial. Plot the time order t versus $r_{[t]}$ if the time order is known. Again, trends and outliers suggest that the model could be improved. A box shaped plot with no trend suggests that the MLR model is good.

The **FF plot** of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ and the **RR plot** of $r_{I,i}$ versus r_i can be used to check whether a candidate submodel I is good. 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 in the RR plot (the OLS line is the identity line in the FF plot). 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 to predict Y. The two plots are also useful for examining the reduced model in the change in SS F test. Note that if the candidate model seems to be good, the usual MLR checks should still be made. In particular, the response plot and residual plot (of $\hat{Y}_{I,i}$ versus $r_{I,i}$) need to be made for the submodel.

The plot of the residuals $Y_i - \overline{Y}$ versus r_i is useful for the Anova F test of $Ho : \beta_2 = \cdots = \beta_p = 0$ versus Ha: not Ho. If Ho is true, then the plotted points in this special case of the RR plot should cluster tightly about the identity line.

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. It is often useful to transform predictors if strong nonlinearities are apparent in the scatterplot matrix.

For the graphical method for choosing a **response transformation**, the **FF**λ plot should have very high correlations. Then the transformation plots can be used. Choose a transformation such that the **transformation plot** is linear. Given several transformation plots, you should be able to find the transformation corresponding to the linear plot.

There are several guidelines for **choosing power transformations**. First, 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**

$$-1,\ -2/3,\ -0.5,\ -1/3,\ -0.25,\ 0,\ 0.25,\ 1/3,\ 0.5,\ 2/3,\ \text{and}\ \ 1.$$

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger. See Cook and Weisberg (1999a, p. 86).

For example, in the plot of *shell* versus *height* in Figure 5.5, 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.

Next, 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.

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 the cube root transformation $w^{1/3}$ may be useful. Consider the ladder of powers. No transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal transformation.

Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if Y = weight and X_1 = volume = $X_2 * X_3 * X_4$, then Y versus $X_1^{1/3}$ and $\log(Y)$ versus $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ 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.

There are also several guidelines for **building a MLR model**. Suppose that variable Z is of interest and variables $W_2, ..., W_r$ have been collected along with Z. Make a scatterplot matrix of $W_2, ..., W_r$ 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 nontrivial predictor variables $X_2, ..., X_p$ 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 OLS output. The statistic R^2 gives the proportion of the variance of Y explained by the predictors and is of some importance.

Variable selection is closely related to the change in SS F test. You are seeking a subset I of the variables to keep in the model. The submodel I

will always contain a constant and will have $k-1$ nontrivial predictors where $1 \leq k \leq p$. Know how to find candidate submodels from output.

Forward selection starts with a constant $= W_1 = X_1$. Step 1) $k = 2$: compute C_p for all models containing the constant and a single predictor X_i. Keep the predictor $W_2 = X_j$, say, that corresponds to the model with the smallest value of C_p.
Step 2) $k = 3$: Fit all models with $k = 3$ that contain W_1 and W_2. Keep the predictor W_3 that minimizes C_p. ...
Step j) $k = j+1$: Fit all models with $k = j+1$ that contains $W_1, W_2, ..., W_j$. Keep the predictor W_{j+1} that minimizes C_p. ...
Step $p-1$): Fit the full model.

Backward elimination: All models contain a constant $= U_1 = X_1$. Step 1) $k = p$: 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 2) $k = p-1$: fit each model with $p-1$ 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 3) $k = p-2$: fit each model with $p-2$ predictors and a constant. Delete the predictor U_{p-1} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-2}$. ...
Step j) $k = p-j+1$: fit each model with $p-j+1$ predictors and a constant. Delete the predictor U_{p-j+2} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-j+1}$. ...
Step $p-1$) $k = 2$. 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.

Rule of thumb for variable selection (assuming that the cost of each predictor is the same): find the submodel I_m with the minimum C_p. If I_m uses k_m predictors, do not use any submodel that has more than k_m predictors. Since the minimum C_p submodel **often has too many predictors**, also look at the submodel I_o with the smallest value of k, say k_o, such that $C_p \leq 2k$ and $k_o \leq k_m$. This submodel **may have too few predictors**. So look at the predictors in I_m but not in I_o and see if they can be deleted or not. (If $I_m = I_o$, then it is a good candidate for the best submodel.)

Assume that the full model has p predictors including a constant and that

the submodel I has k predictors including a constant. Then we would like properties i) – xi) below to hold. Often we can not find a submodel where i) – xi) all hold simultaneously. Given that i) holds, ii) to xi) are listed in decreasing order of importance with ii) – v) much more important than vi) – xi).

i) Want $k \leq p < n/5$.
ii) The response plot and residual plots from both the full model and the submodel should be good. The corresponding plots should look similar.
iii) Want k small but $C_p(I) \leq 2k$.
iv) Want $\mathrm{corr}(\hat{Y}, \hat{Y}_I) \geq 0.95$.
v) Want the change in SS F test using I as the reduced model to have p-value ≥ 0.01. (So use $\alpha = 0.01$ for the change in SS F test applied to models chosen from variable selection. Recall that there is very little evidence for rejecting Ho if p-value ≥ 0.05, and only moderate evidence if $0.01 \leq$ p-value < 0.05.)

vi) Want $R^2_I > 0.9R^2$ and $R^2_I > R^2 - 0.07$.
vii) Want MSE(I) to be smaller than or not much larger than the MSE from the full model.
viii) Want hardly any predictors with p-value ≥ 0.05.
xi) Want only a few predictors to have $0.01 <$ p-value < 0.05.

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. See Chapter 6.

5.5 Complements

Chapters 2–4 of Olive (2007d) covers MLR in much more detail.

Algorithms for OLS are described in Datta (1995), Dongarra, Moler, Bunch and Stewart (1979), and Golub and Van Loan (1989). Algorithms for L_1 are described in Adcock and Meade (1997), Barrodale and Roberts (1974), Bloomfield and Steiger (1980), Dodge (1997), Koenker (1997), Koenker and d'Orey (1987), Portnoy (1997), and Portnoy and Koenker (1997). See Harter (1974a,b, 1975a,b,c, 1976) for a historical account of linear regression. Draper (2000) provides a bibliography of more recent references.

Early papers on transformations include Bartlett (1947) and Tukey (1957). In a classic paper, Box and Cox (1964) developed numerical methods for es-

timating λ_o in the family of power transformations. It is well known that the Box–Cox normal likelihood method for estimating λ_o can be sensitive to remote or outlying observations. Cook and Wang (1983) suggested diagnostics for detecting cases that influence the estimator, as did Tsai and Wu (1992), Atkinson (1986), and Hinkley and Wang (1988). Yeo and Johnson (2000) provide a family of transformations that does not require the variables to be positive.

According to Tierney (1990, p. 297), one of the earliest uses of dynamic graphics was to examine the effect of power transformations. In particular, a method suggested by Fowlkes (1969) varies λ until the normal probability plot is straight. McCulloch (1993) also gave a graphical method for finding response transformations. A similar method would plot $Y^{(\lambda)}$ vs $\hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{x}$ for $\lambda \in \Lambda$. See Example 1.5. Cook and Weisberg (1982, section 2.4) surveys several transformation methods, and Cook and Weisberg (1994) described how to use an inverse response plot of fitted values versus Y to visualize the needed transformation.

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). Some useful ideas for variable selection when outliers are present are given by Burman and Nolan (1995), Ronchetti and Staudte (1994), and Sommer and Huggins (1996).

In the variable selection problem, the FF and RR plots can be highly informative for 1D regression models as well as the MLR model. Results from Li and Duan (1989) suggest that the FF and RR plots will be useful for variable selection in models where Y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^T \boldsymbol{x}$ (eg GLMs), provided that no strong nonlinearities are present in the predictors (eg if $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ and the nontrivial predictors $\boldsymbol{w}$ are iid from an elliptically contoured distribution). See Section 12.4.

Chapters 11 and 13 of Cook and Weisberg (1999a) give excellent discussions of variable selection and response transformations, respectively. They also discuss the effect of deleting terms from the full model on the mean and variance functions. It is possible that the full model mean function $E(Y|\boldsymbol{x})$ is linear while the submodel mean function $E(Y|\boldsymbol{x}_I)$ is nonlinear.

Several authors have used the FF plot to compare models. For example, Collett (1999, p. 141) plots the fitted values from a logistic regression model versus the fitted values from a complementary log–log model to demonstrate that the two models are producing nearly identical estimates.

Section 5.3 followed Olive (2007) closely. See Di Bucchianico, Einmahl, and Mushkudiani (2001) for related intervals for the location model and Preston (2000) for related intervals for MLR. For a review of prediction intervals, see Patel (1989). Cai, Tian, Solomon and Wei (2008) show that the Olive intervals are not optimal for symmetric bimodal distributions. For theory about the shorth, see Grübel (1988). Some references for PIs based on robust regression estimators are given by Giummolè and Ventura (2006).

5.6 Problems

Problems with an asterisk * are especially important.

5.1. Suppose that the regression model is $Y_i = 7 + \beta X_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta) = \sum_{i=1}^{n} (Y_i - 7 - \eta X_i)^2$.

a) What is $E(Y_i)$?

b) Find the least squares estimator $\hat{\beta}$ of β by setting the first derivative $\frac{d}{d\eta} Q(\eta)$ equal to zero.

c) Show that your $\hat{\beta}$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta^2} Q(\eta) > 0$ for all values of η.

5.2. The location model is $Y_i = \mu + e_i$ for $i = 1, ..., n$ where the e_i are iid with mean $E(e_i) = 0$ and constant variance $\text{VAR}(e_i) = \sigma^2$. The least squares estimator $\hat{\mu}$ of μ minimizes the least squares criterion $Q(\eta) = \sum_{i=1}^{n} (Y_i - \eta)^2$. To find the least squares estimator, perform the following steps.

a) Find the derivative $\frac{d}{d\eta} Q$, set the derivative equal to zero and solve for

η. Call the solution $\hat{\mu}$.

b) To show that the solution was indeed the global minimizer of Q, show that $\frac{d^2}{d\eta^2}Q > 0$ for all real η. (Then the solution $\hat{\mu}$ is a local min and Q is convex, so $\hat{\mu}$ is the global min.)

5.3. The normal error model for simple linear regression through the origin is

$$Y_i = \beta X_i + e_i$$

for $i = 1, ..., n$ where $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables.

a) Show that the least squares estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

b) Find $E(\hat{\beta})$.

c) Find VAR($\hat{\beta}$).

(Hint: Note that $\hat{\beta} = \sum_{i=1}^{n} k_i Y_i$ where the k_i depend on the X_i which are treated as constants.)

```
Output for Problem 5.4
Full Model Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      6   265784.       44297.4     172.14    0.0000
Residual       67   17240.9       257.327

Reduced Model Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      1   264621.       264621.     1035.26    0.0000
Residual       72   18403.8       255.608
```

5.4. Assume that the response variable Y is *height*, and the explanatory variables are $X_2 =$ *sternal height*, $X_3 =$ *cephalic index*, $X_4 =$ *finger to ground*, $X_5 =$ *head length*, $X_6 =$ *nasal height*, $X_7 =$ *bigonal breadth*. Suppose that the full model uses all 6 predictors plus a constant ($= X_1$) while the reduced

model uses the constant and *sternal height*. Test whether the reduced model can be used instead of the full model using the above output. The data set had 74 cases.

```
Output for Problem 5.5
Full Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  9   16771.7   1863.52   1479148.9     0.0000
Residual  235   0.29607   0.0012599

Reduced Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  2   16771.7   8385.85   6734072.0     0.0000
Residual  242  0.301359   0.0012453

Coefficient Estimates, Response = y, Terms  = (x2 x2^2)
Label      Estimate     Std. Error    t-value    p-value
Constant   958.470      5.88584       162.843     0.0000
x2        -1335.39     11.1656       -119.599     0.0000
x2^2       421.881      5.29434        79.685     0.0000
```

5.5. The above output comes from the Johnson (1996) STATLIB data set *bodyfat* after several outliers are deleted. It is believed that $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$ where Y is the person's bodyfat and X_2 is the person's density. Measurements on 245 people were taken and are represented by the output above. In addition to X_2 and X_2^2, 7 additional measurements $X_4, ..., X_{10}$ were taken. Both the full and reduced models contain a constant $X_1 \equiv 1$.

a) Predict Y if $X_2 = 1.04$. (Use the reduced model $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$.)

b) Test whether the reduced model can be used instead of the full model.

5.6. Suppose that the regression model is $Y_i = 10 + 2X_{i2} + \beta_3 X_{i3} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta_3) = \sum_{i=1}^{n} (Y_i - 10 - 2X_{i2} - \eta_3 X_{i3})^2$. Find the least squares es-

timator $\hat{\beta}_3$ of β_3 by setting the first derivative $\frac{d}{d\eta_3}Q(\eta_3)$ equal to zero. Show that your $\hat{\beta}_3$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta_3^2}Q(\eta_3) > 0$ for all values of η_3.

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

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

```
Output for Problem 5.9
Label      Estimate   Std. Error    t-value    p-value
Constant  -5.07459    1.85124        -2.741     0.0076
log[H]     1.12399    0.498937         2.253     0.0270
log[S]     0.573167   0.116455         4.922     0.0000

R Squared: 0.895655 Sigma hat: 0.223658 Number of cases: 82
(log[H] log[S]) (4 5)
Prediction = 2.2872, s(pred) = 0.467664,
Estimated population mean value  = 2.2872, s = 0.410715
```

5.9. The output above was produced from the file *mussels.lsp* in *Arc*. Let Y = log(M) where M is the muscle mass of a mussel. Let $X_1 \equiv 1$, $X_2 =$ log(H) where H is the height of the shell, and let $X_3 =$ log(S) where S is the shell mass. Suppose that it is desired to predict Y_f if $\log(H) = 4$ and $\log(S) = 5$, so that $\boldsymbol{x}_f' = (1, 4, 5)$. Assume that $se(\hat{Y}_f) = 0.410715$ and that se(pred) = 0.467664.

a) If $\boldsymbol{x}_f' = (1, 4, 5)$ find a 99% confidence interval for $E(Y_f)$.

b) If $\boldsymbol{x}_f' = (1, 4, 5)$ find a 99% prediction interval for Y_f.

5.10*. a) Show $C_p(I) \leq k$ iff $F_I \leq 1$.

b) Show $C_p(I) \leq 2k$ iff $F_I \leq p/(p-k)$.

```
Output for Problem 5.11 Coefficient Estimates Response = height
Label              Estimate  Std. Error    t-value    p-value
Constant           227.351   65.1732         3.488     0.0008
sternal height     0.955973   0.0515390     18.549     0.0000
finger to ground   0.197429   0.0889004      2.221     0.0295

R Squared: 0.879324    Sigma hat:  22.0731

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

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

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

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

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

5.12. Suppose that it is desired to predict the weight of the brain (in grams) from the cephalic index measurement. The output below uses data from 267 people.

```
predictor   coef          Std. Error    t-value    p-value
Constant   865.001        274.252         3.154     0.0018
cephalic   5.05961        3.48212         1.453     0.1474
```

Do a 4 step test for $\beta_2 \neq 0$.

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

5.14. Suppose that the 95% confidence interval for β_2 is $(-17.457, 15.832)$. Suppose only a constant and X_2 are in the MLR model. Is X_2 a useful linear predictor for Y? If your answer is no, could X_2 be a useful predictor for Y? Explain.

5.15*. a) For $\lambda \neq 0$, expand $f(\lambda) = y^\lambda$ in a Taylor series about $\lambda = 1$. (Treat y as a constant.)

b) Let

$$g(\lambda) = y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda}.$$

Assuming that

$$y\ [\log(y)]^k \approx a_k + b_k y,$$

show that

$$g(\lambda) \approx \frac{[\sum_{k=o}^{\infty}(a_k + b_k y)\ \ \frac{(\lambda-1)^k}{k!}] - 1}{\lambda}$$

$$= [(\frac{1}{\lambda}\sum_{k=o}^{\infty} a_k \frac{(\lambda-1)^k}{k!}) - \frac{1}{\lambda}] + (\frac{1}{\lambda}\sum_{k=o}^{\infty} b_k \frac{(\lambda-1)^k}{k!})y$$

$$= a_\lambda + b_\lambda y.$$

c) Often only terms $k = 0, 1$, and 2 are kept. Show that this 2nd order expansion is

$$\frac{y^\lambda - 1}{\lambda} \approx \left[\frac{(\lambda-1)a_1 + \frac{(\lambda-1)^2}{2}a_2 - \ 1}{\lambda}\right] + \left[\frac{1 + b_1(\lambda-1) + b_2\frac{(\lambda-1)^2}{2}}{\lambda}\right] y.$$

```
Output for problem 5.16.
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
```

5.16. From the output from backward elimination given above, what are two good candidate models for predicting Y? (When listing terms, DON'T FORGET THE CONSTANT!)

Output for Problem 5.17.

	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_I^2	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 change in F test	1.0	0.902	0.622	0.004

5.17. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The forward 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 5.18.

	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_I^2	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 change in F test	1.0	0.550	0.272	0.015

5.18*. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The forward response plot and residual plot for the full model L1 was good. Model L2 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 5.19.

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

5.19. 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:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *Tplt*, will display the code for the function. Use the `args` command, eg *args(Tplt)*, to display the needed arguments for the function.

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

b) Download the *R/Splus* function `ffL` that makes a FFλ plot.

c) 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 programs `ffL` and `Tplt` given in a) and b). Type **ls()** to see if the programs were downloaded correctly.

To make an $FF\lambda$ plot, type the following command.

```
ffL(nx,y)
```

Include the $FF\lambda$ 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".

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

```
Tplt(nx,y)
```

The first plot will be for $\lambda = -1$. Move the curser 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. When you get a plot that clusters about the OLS line which is included in each

plot, 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.

e) 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*.

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

5.21. a) Download the *R/Splus* functions `piplot` and `pisim`.

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

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

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

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

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the northeast corner and click. Problems 5.22–5.27 use data sets that come with *Arc* (Cook and Weisberg 1999a).

5.22*. 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.

5.23. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp*.

The response variable Y is the mussel muscle mass M, and the explanatory variables are $X_2 = S =$ shell mass, $X_3 = H =$ shell height, $X_4 = L =$ shell length and $X_5 = W =$ shell width.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *S*, *H*, *L*, *W* in the "Terms/Predictors" box, *M* in the "Response" box and click on *OK*.

a) To get a response plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *M* in the V–box. Copy the plot into *Word.*

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *L1:Residuals* in the V–box. Copy the plot into *Word.*

d) Based on the residual plot, what MLR assumption seems to be violated?

e) Include the regression output in *Word*.

f) Ignoring the fact that an important MLR assumption seems to have been violated, do any of predictors seem to be needed given that the other predictors are in the model? CONTINUED

g) Ignoring the fact that an important MLR assumption seems to have been violated, perform the ANOVA F test.

5.24*. 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?

5.25*. 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 an FFλ plot 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.

Next use the menu commands "Graph&Fit>Scatterplot-matrix of" to create a dialog window. Select L5:Fit-Values, L4:Fit-Values, L3:Fit-Values, L2 :Fit-Values, and L1:Fit-Values. Then click on "OK." Include the resulting $FF\lambda$ plot in *Word*.

b) Use the menu commands "Graph&Fit>Plot of" to create a dialog window. Double click on L5:Fit-Values and double click on Cycles^{-1}, Cycles$^{-0.5}$, log[Cycles], Cycles$^{0.5}$ or Cycles until the resulting plot in linear. Include the plot of $\widehat{Y}$ versus $Y^{(\lambda)}$ that is linear in *Word*. Use the OLS fit as a visual aid. What response transformation do you end up using?

5.26. 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 $Y = \log Vol$, $X_2 = \log D$ and $X_3 = \log Ht$. 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 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 two lines. Include the plot in *Word*.

e) Next put the residuals from the submodel on the V axis and $\log(Ht)$ on the H axis. 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. 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? **Remark on the information given by each of the 6 plots**. (Some of the plots will suggest that log(Ht) is needed while others will suggest that log(Ht) is not needed.)

5.27*. a) In this problem we want to build a MLR model to predict $Y = g(BigMac)$ for some power transformation g. 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,

use the "Transformations" menu and select "Transform to logs". Include the resulting scatterplot 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 $C_p \leq 2k$ 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*.

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.

5.28*. (Scatterplot in *Arc*.) Activate the *cbrain.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description.

a) Make a plot of *age* versus brain weight *brnweight*. The commands "Graph&Fit > Plot of" will bring down a menu. Put *age* in the **H** box and *brnweight* in the **V** box. Put *sex* in the **Mark by** box. Click *OK*. Make the **lowess bar** on the plot read .1. Open *Word*.

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

b) For a given age, which gender tends to have larger brains?

c) At what age does the brain weight appear to be decreasing?

5.29. (SLR in *Arc*.) Activate *cbrain.lsp*. Brain weight and the cube root of size should be linearly related. To add the cube root of size to the data set, use the menu commands "cbrain > Transform." From the window, select *size* and enter 1/3 in the **p:** box. Then click *OK*. Get some output with commands "Graph&Fit > Fit linear LS." In the dialog window, put *brnweight* in **Response,** and $(size)^{1/3}$ in **terms**.

a) Cut and paste the output (from *Coefficient Estimates* to *Sigma hat*) into *Word*. Write down the least squares equation $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$.

b) If $(size)^{1/3} = 15$, what is the estimated brnweight?

c) Make a plot of the fitted values versus the residuals. Use the commands "Graph&Fit > Plot of" and put "L1:Fit-values" in **H** and "L1:Residuals" in **V**. Put sex in the **Mark by** box. Put the plot into *Word*. Does the plot look ellipsoidal with zero mean?

d) Make a plot of the fitted values versus y = brnweight. Use the commands "Graph&Fit > Plot of" and put "L1:Fit-values in **H** and *brnweight* in **V**. Put *sex* in **Mark by.** Put the plot into *Word*. Does the plot look linear?

5.30*. 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 model was chosen by forward selection?

c) What model was chosen by backward elimination?

d) Which model do you prefer?

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 two lines. 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.

5.31. a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*. This data set consists of various measurements taken on men from Cyprus around 1920. Let the response Y = *height* and X = *cephalic index* = 100(head breadth)/(head length). Use *Arc* to get the least squares output and include the relevant output in *Word*.

b) Intuitively, the cephalic index should not be a good predictor for a person's height. Perform a 4 step test of hypotheses with Ho: $\beta_2 = 0$.

5.32. a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*.

The response variable Y is *height*, and the explanatory variables are a constant, X_2 = *sternal height* (probably height at shoulder) and X_3 = *finger*

to ground.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *sternal height* and *finger to ground* in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*.

Include the output in *Word*. Your output should certainly include the lines from "Response = height" to the ANOVA table.

b) Predict Y if $X_2 = 1400$ and $X_3 = 650$.

c) Perform a 4 step ANOVA F test of the hypotheses with
Ho: $\beta_2 = \beta_3 = 0$.

d) Find a 99% CI for β_2.

e) Find a 99% CI for β_3.

f) Perform a 4 step test for $\beta_2 = 0$.

g) Perform a 4 step test for $\beta_3 = 0$.

h) What happens to the conclusion in g) if $\alpha = 0.01$?

i) The *Arc* menu "L1" should have been created for the regression. Use the menu commands "L1>Prediction" to open a dialog window. Enter 1400 650 in the box and click on *OK*. Include the resulting output in *Word*.

j) Let $X_{f,2} = 1400$ and $X_{f,3} = 650$ and use the output from i) to find a 95% CI for $E(Y_f)$. Use the last line of the output, that is, se = $S(\hat{Y}_f)$.

k) Use the output from i) to find a 95% PI for Y_f. Now se(pred) = s(pred).

l) Make a residual plot of the fitted values vs the residuals and make the response plot of the fitted values versus Y. Include both plots in *Word*.

m) Do the plots suggest that the MLR model is appropriate? Explain.

5.33. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*.

The response variable Y is *height*, and the explanatory variables are X_2 = *sternal height* (probably height at shoulder) and X_3 = *finger to ground*.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *sternal height* and *finger to ground* in the "Terms/Predictors" box,

height in the "Response" box and click on *OK*.

a) To get a response plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *height* in the V–box. Copy the plot into *Word*.

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *L1:Residuals* in the V–box. Copy the plot into *Word*.

d) Based on the residual plot, does a linear model seem reasonable?

5.34. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*.

The response variable Y is *height*, and the explanatory variables are X_2 = *sternal height*, X_3 = *finger to ground*, X_4 = *bigonal breadth* X_5 = *cephalic index* X_6 = *head length* and X_7 = *nasal height*. Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter the 6 predictors (in order: X_2 1st and X_7 last) in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*. This gives the *full model*. For the *reduced model*, only use predictors 2 and 3.

a) Include the ANOVA tables for the full and reduced models in *Word*.

b) Use the menu commands "Graph&Fit>Plot of..." to get a dialog window. Place *L2:Fit-Values* in the H–box and *L1:Fit-Values* in the V–box. Place the resulting plot in *Word*.

c) Use the menu commands "Graph&Fit>Plot of..." to get a dialog window. Place *L2:Residuals* in the H–box and *L1:Residuals* in the V–box. Place the resulting plot in *Word*.

d) Both plots should cluster tightly about the identity line if the reduced model is about as good as the full model. Is the reduced model good?

e) Perform the 4 step change in SS F test (of Ho: the reduced model is good) using the 2 ANOVA tables from part (a). The test statistic is given in Section 5.4.

5.35. 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.