Contents

5

Regression - Selecting A Model

5.1 Automatic Variable Selection Procedures

In studies of business and economics, there are generally several variables that may serve as useful predictors of the response. In searching for a suitable representation of the data, there are a large number of models that are based on linear combinations of explanatory variables and virtually an infinite number of models that are based on nonlinear combinations. To search among the models based on linear combinations, several automatic procedures are available to select variables to be included in the model. These automatic procedures are easy to use, and will suggest one or more models that you can explore in further detail.

To illustrate how large the potential number of linear models is, suppose that there are only four variables, x_1, x_2, x_3 and x_4, under consideration for fitting a model to y. Without any consideration of multiplication or other nonlinear combinations of independent variables, how many possible models are there? The answer is 16. These are:

TABLE 5.1a Sixteen Possible Models

E $y = \beta_0$			1 model with no independent variables
E $y = \beta_0 + \beta_1 x_i$,	$i =$	$1, 2, 3, 4$	4 models with one independent variable
E $y = \beta_0 + \beta_1 x_i + \beta_2 x_j$,	$(i, j) =$	$(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)$	6 models with two independent variables
E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_j + \beta_3 x_k$,	$(i, j, k) =$	$(1,2,3), (1,2,4), (1,3,4), (2,3,4)$	4 models with three independent variables
E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$			1 model with all independent variables

Now suppose there were only three independent variables under consideration. Use the same logic to verify that there are eight possible models. How many linear models will there be if there are ten independent variables? The answer is 1,024, which is quite a few. In general, the answer is 2^k, where k is the number of independent variables. For example, 23 is 8, 24 is 16, and so on.

In any case, for a moderately large number of independent variables, there are many potential models that are based on linear combinations of independent variables. We would like a procedure to search quickly through a number of these potential models to give use more time to think about some of the more interesting aspects of the problem. One procedure for bringing explanatory variables into the model is *stepwise regression*. This procedure employs a series of t-tests to check the

"significance" of explanatory variables entered into, or deleted from, the model. The following is a description of the basic algorithm.

Stepwise Regression Algorithm

Suppose that the analyst has identified one variable as the response, y, and potential explanatory variables, x_1, x_2, ..., x_k.

(i) Consider all possible regressions using one explanatory variable. For each of the k regressions, compute $t(b_1)$, the t-ratio for the slope. Choose that variable with the largest t-ratio. If the t-ratio does not exceed a prespecified t-value (such as two), then do not choose any variables and halt the procedure.

(ii) Add a variable to the model from the previous step. The variable to enter is the one that makes the largest significant contribution. To determine the size of contribution, use the absolute value of the variable's t-ratio. To enter, the t-ratio must exceed a specified t-value in absolute value.

(iii) Delete a variable to the model from the previous step. The variable to be removed is the one that makes the smallest contribution. To determine the size of contribution, use the absolute value of the variable's t-ratio. To be removed, the t-ratio must be less than a specified t-value in absolute value.

(iv) Repeat steps #2 and #3 until all possible additions and deletions are performed.

When implementing this routine, some statistical software packages use an F-test in lieu of t-tests. Recall, when only one variable is being considered, that (t-ratio)^2 = F-ratio and thus these procedures are equivalent.

This algorithm is useful in that it quickly searches through a number of candidate models. However, there are several drawbacks:

- The procedure "snoops" through a large number of models and may fit the data "too well."
- There is no guarantee that the selected model is the best. The algorithm does not consider models that are based on nonlinear combinations of explanatory variables. It also ignores the presence of outliers and high leverage points.
- In addition, the algorithm does not even search all 2^k possible linear regressions.
- The algorithm uses one criterion, a t-ratio, and does not consider other criteria such as s, R^2, R, and so on.
- There is a sequence of significance tests involved. Thus, the significance level that determines the t-value is not meaningful.
- By considering each variable separately, the algorithm does not take into account the joint effect of independent variables.
- Purely automatic procedures may not take into account an investigator's special knowledge.

Because the procedure is not optimal, there are some simpler variants that are available. An advantage of these variants is that they are easier to explain. These include:

- Forward selection. Add one variable at a time without trying to delete variables.
- Backwards selection. Start with the full model and delete one variable at a time without trying to add variables.

Many of the criticisms of the basic stepwise regression algorithm can be addressed with modern computing software that is now widely available. We now consider each drawback, in reverse order. To respond to drawback number seven, many statistical software routines have options for forcing variables into a model equation. In this way, if other evidence indicates that one or more variables should be included in the model, then the investigator can force the inclusion of these variables.

For drawback number six, in the subsection on *suppressor variables* in Section 5.3, we will provide examples of variables that do not have important individual effects but are important when considered jointly. These combinations of variables may not be detected with the basic algorithm but will be detected with the backwards selection algorithm. Because the backwards procedure starts with all variables, it will detect, and retain, variables that are jointly important.

Drawback number five is really a suggestion about the way to use stepwise regression. Bendel and Afifi (1977) suggest using a cut-off smaller than you ordinarily might. For example, in lieu of using a t-value = 2, corresponding approximately to a 5% significance level, consider using a t-value ≈ 1.645, corresponding approximately to a 10% significance level. In this way, there is less chance of screening out variables that may be important. A lower bound, but still a good choice for exploratory work, is a cut-off as small as t-value = 1. This choice is motivated by an algebraic result: when a variable enters a model, s will decrease if the t-ratio exceeds one in absolute value.

To address drawbacks number three and four, we now introduce the *best regressions* routine. Best regressions is a useful algorithm that is now widely available in statistical software packages. The best regression algorithm searches over all possible combinations of explanatory variables, unlike stepwise regression, that adds and deletes one variable at a time. For example, suppose that there are four possible explanatory variables, x_1, x_2, x_3 and x_4, and the user would like to know what is the best two variable model. The best regression algorithm searches over all six models of the form E $y = \beta_0 + \beta_1 x_i + \beta_2 x_j$. Typically, a best regression routine recommends one or two models for each p coefficient model, where p is a number that is user specified. Because it has specified the number of coefficients to enter the model, it does not matter which of the criteria we use: R^2, R, or s.

The best regression algorithm performs its search by a clever use of the algebraic fact that, when a variable is added to the model, the error sum of squares does not increase. Because of this fact, certain combinations of variables included in the model need not be computed. An important drawback of the best regressions algorithm is that it still takes a considerable amount of time when the number of variables considered in the model is large.

Users of regression do not always appreciate the depth of drawback number one, *data-snooping*. Data-snooping occurs when the analyst fits a great number of models to a data set. We will address the problem of data-snooping in the subsection on model validation in Section 5.5. Here, we illustrate the effect of data-snooping in stepwise regression.

Illustration 5.1: Data-Snooping in Stepwise Regression

The idea of this illustration is due to Rencher and Pun (1980). Consider $n = 100$ observations of y and fifty explanatory variables, x_1, x_2, ..., x_{50}. The data we consider here were simulated using independent standard normal random variates. Because the variables were simulated independently, we are working under the null hypothesis of no relation between the response and the explanatory variables, that is, H_0: $\beta_1 = \beta_2 = ... = \beta_{50} = 0$. Indeed, when the model with all fifty explanatory variables was fit, it turns out that $s = 1.142$, $R^2 = 46.2\%$ and F-ratio = (Regression MS) / (Error MS) = 0.84. Using an F-distribution with $df_1 = 50$ and $df_2 = 49$, the 95th percentile is 1.604. In fact, 0.84 is the 27th percentile of this distribution, indicating that the p-value is 0.73. Thus, as expected, the data are in congruence with H0.

Next, a stepwise regression with t-value = 2 was performed. Two variables were retained by this procedure, yielding a model with $s = 1.05$, $R^2 = 9.5\%$ and F-ratio = 5.09. For an F-distribution with $df_1 = 2$ and $df_2 = 97$, the 95th percentile is F-value = 3.09. This indicates that the two variables are significant predictors of y. At first glance, this result is surprising. The data were generated so that y is unrelated to the independent variables. However, because F-ratio ¿ F-value, the F-test indicates that two independent variables are significantly related to y. The reason is that stepwise regression has performed many hypothesis tests on the data. For example, in Step 1, fifty tests were performed to find significant variables. Recall that a 5% level means that we expect to make roughly one mistake in 20. Thus, with fifty tests, we expect to find 50(0.05) = 2.5 "significant" variables, even under the null hypothesis of no relationship between y and the explanatory variables.

To continue, a stepwise regression with t-value = 1.645 was performed. Six variables were retained by this procedure, yielding a model with $s = 0.99$, $R^2 = 22.9\%$and F-ratio = 4.61. As before, an F-test indicates a significant relationship between the response and these six explanatory variables.

To summarize, using simulation we constructed a data set so that the explanatory variables have no relationship with the response. However, when using stepwise regression to examine the data, we "found" seemingly significant relationships between the response and certain subsets of the explanatory variables. This example illustrates a general caveat in model selection: when explanatory variables are selected using the data, t-ratios and F-ratios will be too large, thus overstating the importance of variables in the model.

Stepwise regression and best regressions are examples of *automatic variable selection procedures*. In your modeling work, you will find these procedures to be very useful because they can quickly search through several candidate models. However, these procedures do ignore nonlinear alternatives as well as the effect of outliers and high leverage points. The main point of the procedures is to mechanize certain routine tasks. This automatic selection approach can be extended and indeed, there are a number of so-called "expert systems" available in the market. For example, algorithms are available that "automatically" handle unusual points such as outliers and high leverage points. A model suggested by automatic variable selection procedures should be subject to the same careful diagnostic checking procedures as a model arrived at by any other means.

5.2 Residual Analysis

Recall the role of a residual in the linear regression model. A residual is a response minus the corresponding fitted value under the model. Because the model summarizes the linear effect of several independent variables, we may think of a residual as a response controlled for values of the independent variables. If the model is an adequate representation of the data, then residuals should closely approximate random errors. Random errors are used to represent the natural variation in the model; they represent the result of an unpredictable mechanism. Thus, to the extent that residuals resemble random errors, there should be no discernible patterns in the residuals. Patterns in the residuals indicate the presence of additional information that we hope to incorporate into the model. A lack of patterns in the residuals indicates that the model seems to account for the most important relationships in the data.

There are at least four types of patterns that can be uncovered through the *analysis of residuals*. In this section, we discuss the first two; residuals that are unusual and those that are related to other independent variables. We then introduce the third type, residuals that display a heteroscedastic pattern, in Section 5.6. In our study of longitudinal data models that begins in Chapter 9, we will introduced the fourth type, residuals that display patterns through time.

When examining residuals, it is usually easier to work with a *standardized residual*, a residual that has been rescaled to be dimensionless. As described in our first look at residual analysis in Section 3.5, we generally work with standardized residuals because we achieve some kind of carry-over of experience from one data set to another and may thus focus on relationships of interest. By using standardized residuals, we can train ourselves to look at a variety of residual plots and immediately recognize an unusual point when working in standard units.

There are a number of ways of defining a standardized residual. Using $\hat{e}_i = y_i - \hat{y}_i$ as the ith residual, here are three most commonly used ones:

$$\text{(a) } \frac{\hat{e}_i}{s}, \quad \text{(b) } \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}, \tag{5.1}$$

and

$$\text{(c) } \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

Here, h_{ii} is the ith leverage. It is calculated based on values of the explanatory variables and is defined in Section 5.3 below. Recall that s is the square root of the mean square error (Error MS). Similarly, define $s_{(i)}$ to be the square root of the Error MS when running a regression after having deleted the ith observation.

Now, the first choice of definition of standardized residuals in (a) is simple and is easy to explain. An easy calculation shows that the sample standard deviation of the residuals is approximately s and, indeed,s is often referred to as the residual standard deviation. Thus, it seems reasonable to standardize residuals by dividing by s.

The second choice is presented in (b) and, although more complex, is more precise.

Using theory from mathematical statistics, it turns out that the variance of the ith residual is exactly

$$Var(\hat{e}_i) = \sigma^2(1 - h_{ii}).$$

Note that this variance is smaller than the variance of the true error term, Var $(e_i) = \sigma^2$. Now, we can replace σ by its estimate, s. Then, this result leads to using the quantity $s(1 - hii)^{1/2}$ as an estimated standard deviation, or standard error, for i. Thus, we define the standard error of $\hat{e}_i$ to be

$$se(\hat{e}_i) = s\sqrt{1 - h_{ii}}.$$

Following the conventions introduced in Section 2.6, in this text we use $\hat{e}_i/se(\hat{e}_i)$ to be our *standardized residual*.

The third choice is a modification of the second and is sometimes termed a *studentized residual*. As noted in Section 3.5 and discussed further below, one important use of standardized residuals is to identify unusually large responses. Now, suppose that the ith response is unusually large and that this is measured through its residual. This unusually large residual will also cause the value of s to be large. Because the large effect appears in both the numerator and denominator, the standardized residual may not detect this unusual response. However, this large response will not inflate $s_{(i)}$ because it is constructed after having deleted the ith observation. Thus, when using studentized residuals we get a better measure of observations that have unusually large residuals. By omitting this observation from the estimate of σ, the size of the observation affects only the numerator i and not the denominator $s_{(i)}$.

Another advantage of the third choice is that the studentized residuals can be shown to be realizations from a t-distribution with $n - (k + 1)$ degrees of freedom, assuming the errors are drawn from a normal population. This knowledge of the precise distribution allows us to assess the degree of model fit, particularly in small samples. It is this relationship with the "Student's" t-distribution that suggests the name "studentized" residuals.

The Role of Residuals

One important role of residual analysis is to identify outliers, observations where the residual is unusually large. A good rule of thumb that is used by many statistical packages is that an observation is an outlier if the standardized residual exceeds two in absolute value. To the extent that the distribution of standardized residuals mimics the standard normal curve, we expect about only one in 20 observations, or 95%, to exceed two in absolute value and very few observations to exceed three.

Outliers provide a signal that an observation should be investigated to understand special causes associated with this point. An outlier is an observation that seems unusual with respect to the rest of the data set. It is often the case that the reason for this unusualness may be uncovered after additional investigation. Indeed, this may be the primary purpose of the regression analysis of a data set.

Consider a simple example of so-called *performance analysis*. Suppose we have available a sample of n salespeople and are trying to understand each person's second-year sales based on their first-year sales. To a certain extent, we expect that higher

first-year sales are associated with higher second-year sales. High sales may be due to a salesperson's natural ability, ambition, good territory, and so on. First-year sales may be thought of as a proxy variable that summarizes these factors. We expect variation in sales performance both cross-sectionally and across years. What is interesting is when one salesperson performs unusually well (or poorly) in the second year compared to their first-year performance. Residuals provide a formal mechanism for evaluating second-year sales after controlling for the effects of first-year sales.

Outliers are points that are not typical when compared to other observations in the data set. When summarizing the entire data set using regression techniques, there are a number of options available for handling outliers.

Options for Handling Outliers

- Include the observation in the usual summary statistics but comment on its effects. An outlier may be large but not so large as to skew the results of the entire analysis. If no special causes for this unusual observation can be determined, then this observation may reflect the variability of the data.
- Delete this observation from the data set. The observation may be determined to be unrepresentative of the population for which the sample is being used for inference. If this is the case, then there may be little information contained in the observation that can be used to make general statements about the population. This possibility means that we would omit the observation from the regression summary statistics and discuss it in our report as a separate case.
- Flag the observation with an indicator variable. If one or several special causes have been identified to explain an outlier, then these causes could be introduced into the modeling procedure formally by introducing a variable to indicate the presence (or absence) of these causes. This approach is similar to point deletion but allows the outlier to be formally included in the model formulation so that, if additional observations arise that are affected by the same special causes, then they can be handled on an automatic basis.

Another important role of residuals analysis is to help identify additional explanatory variables that may be used to improve the formulation of the model. If we have specified the model correctly, then residuals should resemble random errors and contain no discernible patterns. Thus, when comparing residuals to explanatory variables, we do not expect any relationships. If we do detect a relationship, then this suggests the need to control for this additional variable. This can be accomplished by introducing the additional variable into the regression model.

Relationships between residuals and explanatory variables can be quickly established using correlation statistics. However, if an explanatory variable is already included in the regression model, then the correlation between the residuals and an explanatory variable will be zero, by a result from matrix algebra. It is a good idea to reinforce this correlation with a scatter plot. Not only will a scatter plot of residuals versus explanatory variables reinforce graphically the correlation statistic, it will also serve to detect potential nonlinear relationships. For example, the quadratic

relationship illustrated in Section 3.5 could only be detected using a scatter plot, not a correlation statistic.

If you detect a relationship between the residuals from a preliminary model fit and an additional explanatory variable, then introducing this additional variable will not always improve your model specification. The reason is that the additional variable may be linearly related to the variables that are already in the model. If you would like a guarantee that adding an additional variable will improve your model, then construct an added variable plot, as described in Section 4.4.

To summarize, after a preliminary model fit, you should:

- Calculate summary statistics and display the distribution of (standardized) residuals to identify outliers.
- Calculate the correlation between the (standardized) residuals and additional explanatory variables to search for linear relationships.
- Create scatter plots between the (standardized) residuals and additional explanatory variables to search for nonlinear relationships.

Application: Stock Liquidity

An investor's decision to purchase a stock is generally made with a number of criteria in mind. First, the investor is usually looking for a high expected return. Second, because investors pay a premium for safety of their investment, they expect to earn higher returns for investing in stocks that are riskier. Thus, a second criterion is the riskiness of a stock which can be measured through the variability of the returns. Third, many investors are concerned with how long they are committing their capital to the purchase of a security. Many income stocks, such as utilities, regularly return portions of capital investments in the form of dividends. Other stocks, particularly growth stocks, return nothing until the sale of the security. Here, the philosophy is that the investor is better off reaping long-term profits that are hoped for in growth stocks than the immediate returns afforded by income stocks. Thus, the average length of investment in a security is another criterion. Fourth, investors are concerned with the ability to sell the stock at any time convenient to the investor. We refer to this fourth criterion as the *liquidity* of the stock. The more liquid is the stock, the easier it is to sell. To measure the liquidity, in this study we use the number of shares traded on an exchange over a specified period of time (called the VOLUME). We are interested in studying the relationship between the volume and other financial characteristics of a stock.

We begin this study with 126 companies whose options were traded on December 3, 1984. The stock data were obtained from Francis Emory Fitch, Inc. for the period from December 3, 1984 to February 28, 1985. For the trading activity variables, we examine the three months total trading volume (VOLUME), the three months total number of transactions (NTRAN), and the average time between transactions (AVGT). For the firm size variables, we use the opening stock price on January 2, 1985 (PRICE), the number of outstanding shares on December 31, 1984 (SHARE), and the market equity value (VALUE) obtained by taking the product of PRICE and SHARE. Finally, for the financial leverage, we examine the debt-to-equity ratio (DEB_EQ) obtained from the Compustat Industrial Tape and the Moody's manual. The data in SHARE are obtained from the Center for Research in Security Prices (CRSP) monthly tape.

After examining some preliminary summary statistics of the data, three companies were deleted because they either had an unusually large volume or high price. They are Teledyne and Capital Cities Communication, whose prices were more than four times the average price of the remaining companies, and American Telephone and Telegraph, whose total volume was more than seven times than the average total volume of the remaining companies. Based on additional investigation, the details of which are not presented here, these companies were deleted because they seemed to represent special circumstances that we would not wish to model. Table 5.1 summarizes the descriptive statistics of the stock liquidity variables based on the remaining 123 companies. For example, from Table 5.1 we see that the average time between transactions is about five minutes and this time ranges from a minimum of less than a minute to a maximum of about 20 minutes.

TABLE 5.1 Summary Statistics of the Stock Liquidity Variables

	Mean	Median	Standard deviation	Minimum	Maximum
VOLUME	13.423	11.556	10.632	0.658	64.572
AVGT	5.441	4.284	3.853	0.590	20.772
NTRAN	6436	5071	5310	999	36420
PRICE	38.80	34.37	21.37	9.12	122.37
SHARE	94.7	53.8	115.1	6.7	783.1
VALUE	4.116	2.065	8.157	0.115	75.437
DEB_EQ	2.697	1.105	6.509	0.185	53.628

Legend:
VOLUME: Total trading volume for the entire three months in million shares.
AVGT: Average transaction time interval measured in minutes.
NTRAN: Total number of transactions for the three months.
PRICE: Stock price at the opening on January 2, 1985 in U.S. dollars.
SHARE: Number of shares outstanding on December 31, 1984 in million shares.
VALUE: Market value in billion dollars (PRICE × SHARE).
DEB_EQ: Debt-to-equity ratio at the end of 1984.

Source: Francis Emory Fitch, Inc., Standard & Poor's Compustat, and University of Chicago's Center for Research on Security Prices.

Table 5.2 reports the correlation coefficients and Figure 5.1 provides the corresponding scatterplot matrix. If you have a background in finance, you will find it interesting to note that the financial leverage, measured by DEB_EQ, does not seem to be related to the other variables. From the scatterplot and correlation matrix, we see a strong relationship between VOLUME and the size of the firm as measured by SHARE and VALUE. Further, the three trading activity variables, VOLUME, AVGT and NTRAN, are all highly related to one another.

TABLE 5.2 Correlation Matrix of the Stock Liquidity Variables

	AVGT	NTRAN	PRICE	SHARE	VALUE	DEB_EQ
NTRAN	−0.668					
PRICE	−0.128	0.190				
SHARE	−0.429	0.817	0.177			
VALUE	−0.318	0.760	0.457	0.829		
DEB_EQ	0.094	−0.092	−0.038	−0.077	−0.077	
VOLUME	−0.674	0.913	0.168	0.773	0.702	−0.052

Figure 5.1 shows that the variable AVGT is inversely related to VOLUME and NTRAN is inversely related to AVGT. In fact, it turned out the correlation between

the average time between transactions and the reciprocal of the number of transactions was 99.98%! This is not so surprising when one thinks about how AVGT might be calculated. For example, on the New York Stock Exchange, the market is open from 10:00 A.M. to 4:00 P.M. For each stock on a particular day, the average time between transactions times the number of transactions is nearly equal to 360 minutes (= 6 hours). Thus, except for rounding errors because transactions are only recorded to the nearest minute, there is a perfect linear relationship between AVGT and the reciprocal of NTRAN.

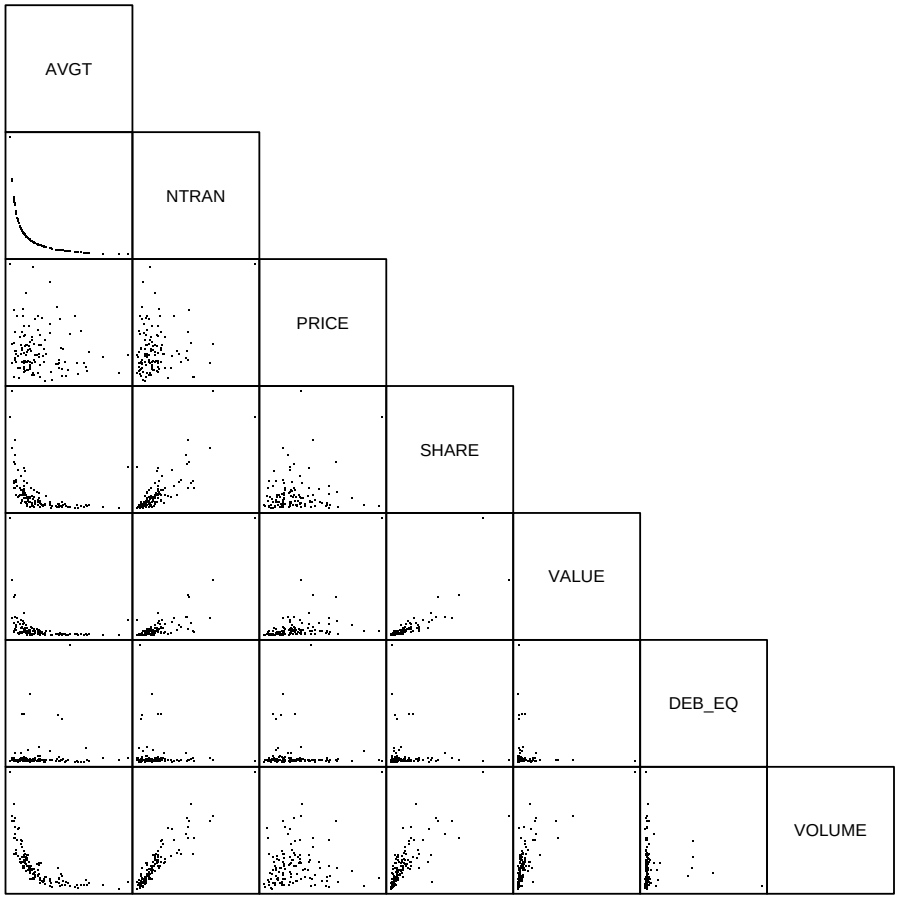

Fig. 5.1. Scatterplot matrix for stock liquidity variables The number of transactions variable (NTRAN) appears to be strongly related to the VOLUME of shares traded, and inversely related to AVGT.

To begin to understand the liquidity measure VOLUME, we first fit a regression model using NTRAN as an explanatory variable. The fitted regression model is:

VOLUME	=	1.65	+0.00183 NTRAN
std errors		(0.0018)	(0.000074)

with $R^2 = 83.4\%$ and $s = 4.35$. Note that the t-ratio for the slope associated with

NTRAN is $t(b_1) = b_1/se(b_1) = 0.00183/0.000074 = 24.7$, indicating that b_1 is about 24.7 standard errors above zero. Residuals were computed using this estimated model. To see if the residuals are related to the other explanatory variables, below is a table of correlations.

TABLE 5.2a First Table of Correlation

	AVGT	PRICE	SHARE	VALUE	DEB_EQ
RESID	-0.155	-0.017	0.055	0.007	0.078

Table 5.2a correlations between residuals and several explanatory variables. The residuals were created from a regression of VOLUME on NTRAN.

The correlation between the residual and AVGT and the scatter plot (not given here) indicates that there may be some information in the variable AVGT in the residual. Thus, it seems sensible to use AVGT directly in the regression model. Remember that we are interpreting the residual as the value of VOLUME having controlled for the effect of NTRAN.

We next fit a regression model using NTRAN and AVGT as an explanatory variables. The fitted regression model is:

VOLUME	=	4.41	-0.322 AVGT	+0.00167 NTRAN
std errors		(1.30)	(0.135)	(0.000098)

with $R^2 = 84.2\%$ and $s = 4.26$. Based on the t-ratio for AVGT, $t(b_1) = (-0.322)/0.135 = -2.39$, it seems as if AVGT is a useful explanatory variable in the model. Note also that s has decreased, indicating that R has increased.

TABLE 5.2b Second Table of Correlation

	PRICE	SHARE	VALUE	DEB_EQ
RESID	-0.015	0.096	0.071	0.089

The table (Table 5.2b) of correlations between the model residuals and other potential explanatory variables indicates that there does not seem to be much additional information in the explanatory variables. This is reaffirmed by the corresponding table of scatter plots in Figure 5.2. The histograms in Figure 5.2 suggest that although the distribution of the residuals is fairly symmetric, the distribution of each explanatory variable is skewed. Because of this, transformations of the explanatory variables were explored. Unfortunately, this line of thought provided no real improvements and thus the details are not provided here.

Thus, the firm size variables, although strongly related to volume, are not important determinants when other trading activity variables are entered into the model as explanatory variables. In Section 5.4 we will fit a model of VOLUME without using other trading activity variables as explanatory variables.

5.3 Leverage

The examination of unusual observations can be decomposed into the analysis of residuals and of high leverage points. In Section 3.5, we saw that a high leverage point is an observation containing an unusual set of explanatory variables. Such a point may turn out to be *influential* because it has a disproportionate effect on the

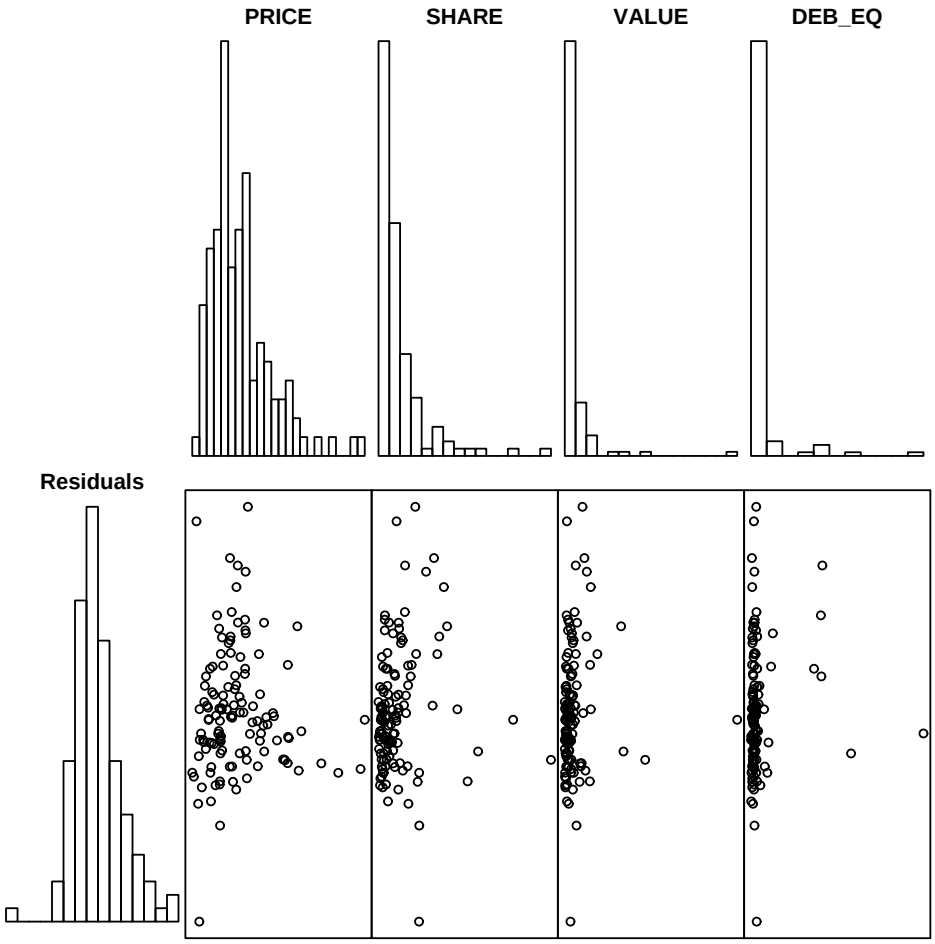

Fig. 5.2. Scatterplot matrix of the residuals from the regression of VOLUME on NTRAN and AVGT on the vertical axis and the remaining predictor variables on the horizontal axes.

overall regression fit. To illustrate this, an example from Section 3.5 shows how one observation in twenty can reduce the R^2 from 90% to 10% (point C).

The reason for this disproportionate effect is that regression slope estimates can be shown to be weighted averages of slopes, where the weights are determined by (squared Euclidean) distances between explanatory variables. This result is surprising because the regression estimates are defined as quantities that minimize the sum of squared deviations. To illustrate, recall in the case of regression using one variable that the least squares slope estimate is $b_1 = rs_y/s_x$. Using algebra, it can be checked that an alternative expression is

$$b_1 = \frac{\sum_{i=1}^{n} weight_i \; slope_i}{\sum_{i=1}^{n} weight_i}$$

Here, we have $weight_i = (x_i - \bar{x})^2$ and $slope_i = (y_i - \bar{y})/(x_i - \bar{x})$. That is, the ith slope is the slope between (x_i, y_i) and $(\bar{x}, \bar{y})$.

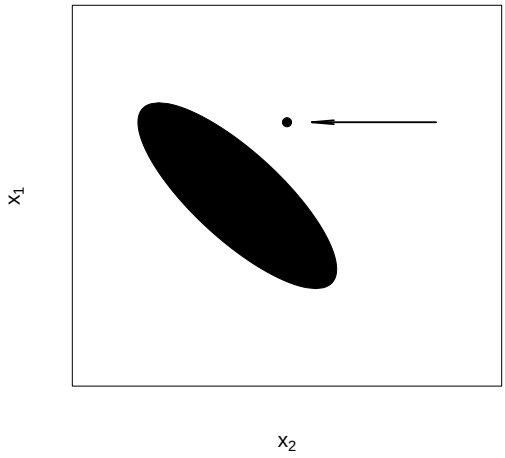

Fig. 5.3. The ellipsoid represents most of the data. The arrow marks an unusual point.

When there are more than two explanatory variables, it is difficult to determine graphically whether a point in unusual. With only one explanatory variable, it is easy to determine what is "unusual" merely by examining the histogram of the explanatory variable. With more than one variable, determining an unusual point is not as straightforward. Consider the fictitious data set represented in Figure 5.3. The point marked in the upper right hand corner is unusual. However, it is not unusual when examining the histogram of either x_1 or x_2. It is only unusual when the explanatory variables are considered jointly. For two explanatory variables, this is apparent when examining the data graphically. Because it is difficult to examine graphically data having more than two explanatory variables, we resort to the following numerical procedure.

Using matrix algebra, it can be shown that the fitted values can be expressed as a linear combination of responses. Thus, we have

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + ... + h_{ii}y_i + ... + h_{in}y_n.$$

The values h_{ij} are calculated using only the values of the explanatory variables. If you are interested, then the details of these calculations are in Section 4.7. From this expression, we see that the larger is h_{ii}, the larger is the effect that the ith response ($\hat{y}_i$) has on the corresponding fitted value (i). Thus, we call h_{ii} to be the *leverage* for the ith observation. Because the values h_{ii} are calculated based on the explanatory variables, the values of the response variable do not affect the calculation of leverages.

Large leverage values indicate that an observation may exhibit a disproportionate effect on the fit, essentially because it is distant from the other observations (when looking at explanatory variables). How large is large? Some guidelines are available from matrix algebra, where we have that

$$\frac{1}{n} \leq h_{ii} \leq 1$$

and

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{k+1}{n}.$$

Thus, each leverage is bounded by n^{-1} and one and the average leverage equals the number of regression coefficients divided by the number of observations. From these and related arguments, we use a widely adopted convention and declare an observation to be a *high leverage point* if the leverage exceeds three times the average, that is, if $h_{ii} > 3(k+1)/n$.

Having identified high leverage points, as with outliers it is important for the analyst to search for special causes that may have produced these unusual points. To illustrate, in Section 3.6 we identified the 1987 market crash as the reason behind the high leverage point. Further, high leverage points are often due to clerical errors in coding the data, which may or may not be easy to rectify. In general, the options for dealing with high leverage points are similar to those available for dealing with outliers.

Options for Handling Outliers.

(i) Include the observation in the summary statistics but comment on its effect. For example, an observation may barely exceed a cut-off and its effect may not be important in the overall analysis.

(ii) Delete the observation from the data set. Again, the basic rationale for this action is that the observation is deemed not representative of some larger population. An intermediate course of action between (1) and (2) is to present the analysis both with and without the high leverage point. In this way the impact of the point is fully demonstrated and the reader of your analysis may decide which option is more appropriate.

(iii) Choose another variable to represent the information. In some instances, another explanatory variables will be available to serve as a replacement. For example, in our Chapter 4 example on apartment rents, we used the indicator of two bedrooms variable to replace the square footage variable. Both variables provide information about the size of an apartment. Although an apartment may be unusually large causing it to be a high leverage point, it will only have one or two bedrooms in the sample we examined.

(iv) Use a nonlinear transformation of an explanatory variable, as described in Section 4.5. To illustrate, with our Stock Liquidity example in Section 5.1, we can transform the debt-to-equity DEB_EQ continuous variable into a variable that indicates the presence of "high" debt-to-equity. For example, we might code DE_IND = 1 if DEB_EQ > 5 and DE_IND = 0 if DEB_EQ ≤ 5. With this recoding, we still retain information on the financial leverage of a company without allowing the large values of DEB_EQ drive the regression fit.

In addition, some statisticians use "robust" estimation methodologies as an alternative to least squares estimation. The basic idea of these techniques is to reduce the effect of any particular observation. These techniques are useful in reducing the effect of both outliers and high leverage points. This tactic may be viewed as intermediate between one extreme procedure, ignoring the effect of unusual points, and another extreme procedure, giving unusual points full credibility by deleting them from the data set. The word *robust* is meant to suggest that these estimation methodologies are "healthy" even when attacked by an occasional bad observation (a germ). We have seen that this is not true for least squares estimates.

Cook's Distance

To quantify how unusual a point is, a measure that considers both the response and explanatory variables is *Cook's Distance*. This distance, D_i, is defined as

$$\begin{aligned} D_i &= \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1)s^2} \\ &= \left(\frac{\hat{e}_i}{se(\hat{e}_i)}\right)^2 \frac{h_{ii}}{(k+1)(1-h_{ii})}. \end{aligned} \tag{5.2}$$

The first expression provides a definition. Here, $\hat{y}_{j(i)}$is the prediction of the jth observation, computed leaving the ith observation out of the regression fit. To measure the impact of the ith observation, we compare the fitted values with and without the ith observation. Each difference is then squared and summed over all observations to summarize the impact.

After rescaling by $(k+1)s^2$, the second equation provides another interpretation of the distance D_i. The first part, $(\hat{e}_i/se(\hat{e}_i))^2$, is the square of the ith standardized residual. The second part, $h_{ii}/((k+1)(1-h_{ii}))$, is attributable solely to the leverage. Thus, the distance D_i is composed of a measure for outliers times a measure for leverage. In this way, Cook's distance accounts for both the response and explanatory variables.

To get an idea of the expected size of D_i for a point that is not unusual, recall that we expect the standardized residuals to be about one and the leverage h_{ii} to be about $(k+1)/n$. Thus, we anticipate that D_i should be about $1/n$. Another rule of thumb is to compare D_i to an F-distribution with $df_1 = k+1$ and $df_2 = n-(k+1)$ degrees of freedom. Values of D_i that are large compared to this distribution merit attention.

To illustrate, we return to our outlier Illustration 3.2 in Section 3.5. In this example, we considered 19 "good," or base, points plus each of the three types of unusual points, labelled A, B and C. Table 5.3 summarizes the calculations.

TABLE 5.3 Measures of unusual points for Example 3.1

Data	Standardized residual $\hat{e}_i/se(\hat{e}_i)$	Leverage h_{ii}	Cook's distance D_i
19 Base Points plus A	4.00	.067	.577
19 Base Points plus B	.77	.550	.363
19 Base Points plus C	-4.01	.550	9.832

As noted in Section 3.5, from the standardized residual column we see that both points A and C are outliers. To judge the size of the leverages, because there are $n = 20$ points, the leverages are bounded by 0.05 and 1.00 with the average leverage being $\bar{h} = 2/20 = 0.10$. Using 3 as a cut-off, both points B and C are high leverage points. Note that their values are the same. This is because, from Figure 3.12, the values of the explanatory variables are the same and only the response variable has been changed. The column for Cook's distance captures both types of unusual behavior. Because the typical value of D_i is $1/n$ or 0.05, Cook's distance provides one statistic to alert us to the fact that each point is unusual in one respect or another. In particular, point C has a very large D_i, reflecting the fact that it is both an outlier and a high leverage point. The 95th percentile of an F-distribution with $df_1 = 2$ and $df_2 = 18$ is 3.555. The fact that point C has a value of D_i that well exceeds this cut-off indicates the substantial influence of this point.

5.4 Collinearity

Collinearity, or *multicollinearity*, occurs when one explanatory variable is, or nearly is, a linear combination of the other explanatory variables. Intuitively, it is useful to think of the independent variables as being highly correlated with one another as an indication of collinearity. With collinear data, the explanatory variables may

provide little additional information over and above the information provided by the other explanatory variables. The issues are: Is collinearity important? If so, how does it affect our model fit and how do we detect it? To address the first question, consider a somewhat pathological example.

Illustration 5.2: Perfectly Correlated Independent Variables

Joe Finance was asked to fit the model E $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ to a data set. His resulting fitted model was $\hat{y} = -87 + x_1 + 18x_2$. The data set under consideration is:

i	1	2	3	4
y_i	23	83	63	103
x_{i1}	2	8	6	10
x_{i2}	6	9	8	10

Joe checked the fit for each observation. Joe was very happy because he fit the data perfectly! For example, for the third observation the fitted value is $\hat{y}_3 = -87 + 6 + 18(8) = 63$, which is equal to the third response, y_3. Because the response equals the fitted value, the residual is zero. You may check that this is true of each observation and thus the R^2 turned out to be 100%.

However, Jane Actuary came along and fit the model $\hat{y} = -7 + 9x_1 + 2x_2$. Jane performed the same careful checks that Joe did and also got a perfect fit. Who is right?

The answer is both and neither one. There are, in fact, an infinite number of fits. This is due to the perfect relationship $x_2 = 5 + x_1/2$ between the two explanatory variables.

This example serves to illustrate some important facts about collinearity.

Collinearity Facts

- The fact that there is high correlation (among independent variables) neither precludes us from getting good fits nor from making predictions of new observations. Note that in the above example we got perfect fits.
- Estimates of error variances and, therefore, tests of model adequacy, are still reliable.
- In cases of serious collinearity, standard errors of individual regression coefficients are larger than cases where, other things equal, serious collinearity does not exist. With large standard errors, individual regression coefficients may not be meaningful. Further, because a large standard error means that the corresponding t-ratio is small, it is difficult to detect the importance of a variable.

There are several useful devices for detecting collinearity. A matrix of correlation coefficients of explanatory variables is simple to create and is easy to interpret. This matrix quickly captures linear relationships between pairs of variables. A scatterplot matrix serves to provide a visual reinforcement of the summary statistics in the correlation matrix. Note that, for collinearity, we are only interested in detecting linear trends, so nonlinear relationships between variables are not an issue here. For example, we have seen that it is sometimes useful to retain both an explanatory

variable (x) and its square (x^2), despite the fact that there is a perfect (nonlinear) relationship between the two.

Variance Inflation Factors

Correlation and scatterplot matrices capture only relationships between pairs of variables. To capture more complex relationships among several variables, we use the concept of a *variance inflation factor (VIF)*. To define a *VIF*, suppose that the set of explanatory variables is labelled $x_1, x_2, ..., x_k$. Now, run the regression using x_j as the "response" and the other x's $(x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_k)$ as the explanatory variables. Denote the coefficient of determination from this regression by R_j^2. We interpret R_j^2 as the square of the multiple correlation coefficient between x_j and linear combinations of the other x's. From this coefficient of determination, we define the variance inflation factor

$$VIF_j = \frac{1}{1 - R_j^2} \quad \text{for} \quad j = 1, 2, ..., k.$$

A larger R_j^2 results in a larger VIF_j; this means greater collinearity between x_j and the other x's. Now, R alone is enough to capture the linear relationship of interest. However, we use VIF_j in lieu of R_j^2 as our measure for collinearity because of the algebraic relationship:

$$se(b_j) = s\frac{\sqrt{VIF_j}}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} = s\frac{\sqrt{VIF_j}}{s_{x_j}\sqrt{n-1}} \tag{5.3}$$

Here, $se(b_j)$ and s are standard errors and residual standard deviation from a full regression fit of y on $x_1, ..., x_k$. Further, s_{x_j} is the sample standard deviation of the jth variable x_j.

Thus, a larger VIF_j results in a larger standard error associated with the jth slope, b_j. (If you studied Section 4.7, then you will recall that $se(b_j)$ is s times the $(j+1)$st diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. The idea is that when collinearity occurs, the matrix $\mathbf{X}'\mathbf{X}$ has properties similar to the number zero. When we attempt to calculate the inverse of $\mathbf{X}'\mathbf{X}$, this is analogous to dividing by zero for scalar numbers.) As a rule of thumb, when VIF_j exceeds 10 (which is equivalent to $R_j^2 > 90\%$), we say that severe collinearity exists. This may signal is a need for action.

Illustration 5.3: Stock Liquidity Example - Continued

As an example, consider a regression of VOLUME on PRICE, SHARE and VALUE. Unlike the explanatory variables considered in Section 5.2, these three explanatory variables are not measures of trading activity. From a regression fit, we have $R^2 = 61\%$ and $s = 6.72$. From Table 5.3, we saw that $s_y = 10.6$, so $s = 6.72$ represents a considerable drop in our estimate of the variability. Statistics associated with the regression coefficients are in Table 5.4.

TABLE 5.4 Statistics from a Regression of VOLUME on PRICE, SHARE and VALUE

x_j	s_{x_j}	b_j	$se(b_j)$	$t(b_j)$	VIF_j
$PRICE$	21.37	−0.022	0.035	−0.63	1.5
$SHARE$	115.1	0.054	0.010	5.19	3.8
$VALUE$	8.157	0.313	0.162	1.94	4.7

You may check that the relationship in equation (6.3) is valid for each of the explanatory variables in Table 5.4. Because each VIF statistic is less than ten, there is little reason to suspect severe collinearity. This is interesting because you may recall that there is a perfect relationship between PRICE, SHARE and VALUE in that we defined the market value to be VALUE = PRICE × SHARE. However, the relationship is multiplicative, and hence is nonlinear. Because the variables are not linearly related, it is valid to enter all three into the regression model.

Still, we must check that nonlinear relationships are not approximately linear over the sampling region. Even though the relationship is theoretically nonlinear, if it is close to linear for our available sample, then problems of collinearity might arise. Figure 5.4 illustrates this situation.

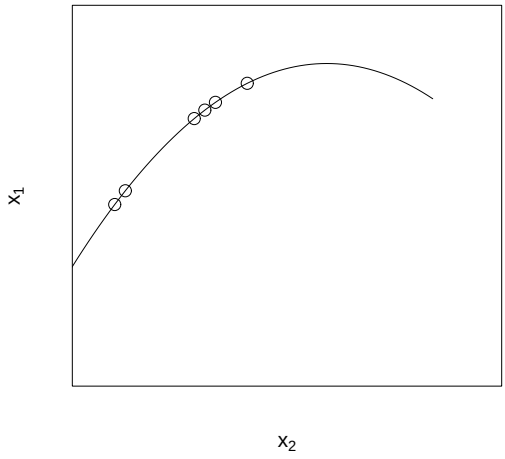

Fig. 5.4. The relationship between x_1 and x_2 is nonlinear. However, over the region sampled, the variables have close to a linear relationship.

What can we do in the presence of collinearity? One option is to center each variable, by subtracting its average and dividing by its standard deviation. For example, create a new variable $x^*_{ij} = (x_{ij} - \bar{x}_j)/s_{x_j}$. Occasionally, one variable appears as millions of units and another variable appears as fractions of units. Compared to the first mentioned variable, the second mentioned variable is close to a constant column of zeroes, at least if one uses single-precision (eight significant digits) arithmetic. If this is true, then the second variable looks very much like a linear shift of the constant column of ones corresponding to the intercept. This is a problem even using double-precision arithmetic because, with the least squares operations, we are implicitly squaring numbers that can make these columns appear even more similar.

This problem is simply a computational one and is easy to rectify. Simply recode the variables so that the units are of similar order of magnitude. Some data analysts

automatically center all variables to avoid these problems. This is a legitimate approach because regression techniques search for linear relationships; scale and location shifts do not affect linear relationships.

Another option is to simply not explicitly account for collinearity in the analysis but to discuss some of its implications when interpreting the results of the regression analysis. This approach is probably the most commonly adopted one. It is a fact of life that, when dealing with business and economic data, collinearity does tend to exist among variables. Because the data tends to be observational in lieu of experimental in nature, there is little that the analyst can do to avoid this situation.

When severe collinearity exists, often the only option is to remove one or more variables from the regression equation. In the best-case situation, an auxiliary variable that provides similar information and that eases the collinearity problem, is available to replace a variable. Similar to our discussion of high leverage points, a transformed version of the explanatory variable may also be a useful substitute. In some situations, such an ideal replacement is not available and we are forced to remove one or more variables. Deciding which variables to remove is a difficult choice. Sometimes automatic variables selection techniques, described in Section 5.1, can help determine an overall suitable model choice. When deciding among variables, often the choice will be dictated by the investigator's judgement as to which is the most relevant set of variables.

Collinearity and Leverage

Measures of collinearity and leverage share common characteristics, and yet are designed to capture different aspects of a data set. Both are useful for data and model criticism; they are applied after a preliminary model fit with the objective of improving model specification. Further, both are calculated using only the explanatory variables; values of the responses do not enter into either calculation.

Our measure of collinearity, the variance inflation factor, is designed to help us with model criticism. It is a measure calculated for each explanatory variable, designed to explain the relationship with other explanatory variables.

The leverage statistic is designed to help us with data criticism. It is a measure calculated for each observation to help us explain how unusual an observation is with respect to other observations.

Collinearity may be masked, or induced, by high leverage points, as pointed out by Mason and Gunst (1985) and Hadi (1988). Figures 5.5 and 5.6 provide illustrations of each case. These simple examples underscore an important point; data criticism and model criticism are not separate exercises.

The examples in Figures 5.5 and 5.6 also help us to see one way in which high leverage points may affect standard errors of regression coefficients. Recall, in Section 5.2, we saw that high leverage points may affect the model fitted values. In Figures 5.5 and 5.6, we see that high leverage points affect collinearity. Thus, from equation (5.3), we have that high leverage points can also affect our standard errors of regression coefficients.

Suppressor Variables

As we have seen, severe collinearity can seriously inflate standard errors of regression coefficients, other things equal. Because we rely on these standard errors for judging the usefulness of explanatory variables, our model selection procedures and inferences

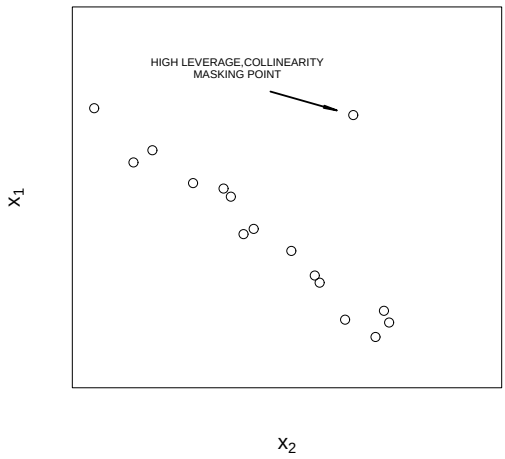

Fig. 5.5. With the exception of the marked point, x_1 and x_2 are highly linearly related.

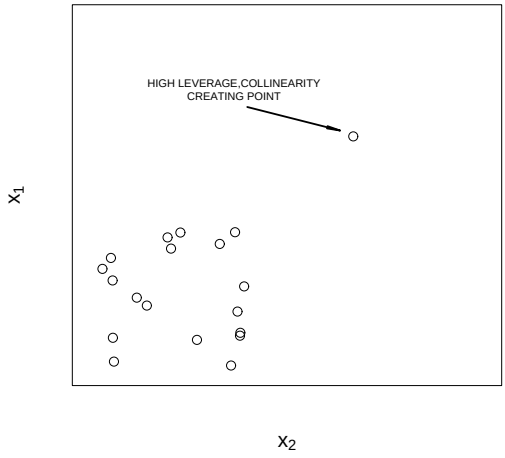

Fig. 5.6. The highly linear relationship between x_1 and x_2 is primarily due to the marked point.

may be deficient in the presence of severe collinearity. Despite these drawbacks, mild collinearity in a data set should not be viewed as a deficiency of the data set; it is simply an attribute of the available explanatory variables.

Even if one explanatory variable is nearly a linear combination of the others, that does not necessarily mean that the information that it provides is redundant. To illustrate, we now consider a *suppressor variable*, an explanatory variable that increases the importance of other explanatory variables when included in the model. Figure 5.7 shows a scatterplot matrix of a hypothetical data set of fifty observations. This data set contains a response and two explanatory variables. Table 5.5 is the corresponding matrix of correlation coefficients. Here, we see that the two explanatory variables are highly correlated. Now recall, for regression with one independent variable, that the correlation coefficient squared is the coefficient of determination. Thus, using Table 5.5, for a regression of y on x_1, the coefficient of determination is

$(.188)^2 = 3.5\%$. Similarly, for a regression of y on x_2, the coefficient of determination is $(-0.022)^2 = 0.04\%$. However, for a regression of y on x_1 and x_2, the coefficient of determination turns out to be a surprisingly high 80.7%. The interpretation is that individually, both x_1 and x_2 have little impact on y. However, when taken jointly, the two explanatory variables have a significant effect on y. Although Table 5.5 shows that x_1 and x_2 are strongly linearly related, this relationship does not mean that x_1 and x_2 provide the same information. In fact, in this example the two variables complement one another.

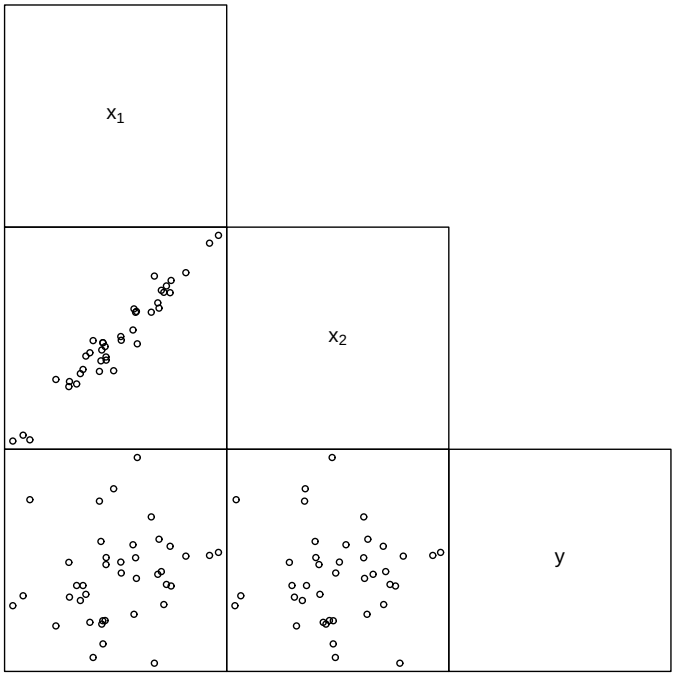

Fig. 5.7. Scatterplot matrix of a response and two explanatory variable for the suppressor variable example.

TABLE 5.5 Correlation Matrix
for the Supressor Example Corresponding to Figure 5.7

	x_1	x_2
x_2	0.972	
y	0.188	−0.022

5.5 Selection Criteria

There are several criteria available for selecting models. We introduced most of the basic criteria in Chapter 4. These criteria include the coefficient of determination (R^2), an adjusted version (R_a^2), the size of the typical error (s), and the t-ratio of each slope coefficient. Here, we discuss an additional quantity, the C_p statistic. According to the iterative procedure outlined in Section 4.0, we use these selection criteria in conjunction with diagnostic checks to arrive at candidate models.

C_p Statistic

Like the automatic variable selection procedures described in Section 5.1, the C_p statistic is used in the beginning stages of developing a regression model. To define this statistic, assume that we have available k explanatory variables $x_1, ..., x_k$ and we first run a regression to get s^2_{full} as the mean square error. Now, suppose that we are considering using only $p-1$ explanatory variables so that there are p regression coefficients. With these $p-1$ explanatory variables, we run a regression to get the error sum of squares $(Error\ SS)_p$. Thus, we are in the position to define

$$C_p = \frac{(Error\ SS)_p}{s^2_{full}} - (n - 2p)$$

The choice of p may vary from 1 to $k+1$. For example, in the case where $p = k+1$, all of the variables are included. In this case, we have

$$\begin{aligned} C_{k+1} &= \frac{(Error\ SS)_{k+1}}{s^2_{full}} - (n - 2(k+1)) \\ &= (n-(k+1))\frac{(Error\ MS)_{k+1}}{s^2_{full}} - (n - 2(k+1)) \\ &= (n-(k+1)) - (n - 2(k+1)) = k+1, \end{aligned}$$

because $(Error\ MS)_{k+1} = s^2_{full}$.

In general, if the model with p regression coefficients is correct, then we expect C_p to be close to p. The idea is that s^2_{full} should be close to σ^2 and, if the model is correct, then $(Error\ MS)_p$ should also be close to σ^2. Thus,

$$\begin{aligned} C_p &= (n-p)\frac{(Error\ MS)_p}{s^2_{full}} - (n - 2p) \\ &\approx (n-p)\frac{\sigma^2}{\sigma^2} - (n - 2p) = p. \end{aligned}$$

As a selection criterion, we choose the model with a "small" C_p coefficient, where small is taken to be relative to p. In general, models with smaller values of C_p are more desirable.

The C_p statistic measures the candidate model's mean square error relative to a full model mean square error. In general, we prefer models with a small C_p coefficient such that $C_p \approx p$. It may be, however, that the full model is poorly specified and that the resulting mean square error is inflated. In such cases, the value of C_p can be negative. This is not to say that the model with the smallest C_p is poor; it merely states that the full model is poorly specified.

Model Validation

Model validation is the process of confirming that our proposed model is appropriate, especially in light of the purposes of the investigation. Recall the iterative model formulation selection process described in Section 4.0. Using this iterative procedure, we examine the basic selection criteria as well as additional diagnostic checks

described in Sections 3.6, 5.1 and 5.2 to arrive at the final stage of model selection. An important criticism of this iterative process is that it is guilty of data-snooping, that is, fitting a great number of models to a single set of data. As we saw in Illustration 5.1 on data-snooping in stepwise regression, by looking at a large number of models we may actually overfit the data and understate the natural variation in our representation. Another drawback of the iterative fitting process is that we are implicitly using a sequence of tests of hypotheses to formulate our ideas about the candidate model. By doing a number of tests of hypothesis on a single data set, we may be working at a much higher significance level than we nominally prescribe.

We can respond to these criticisms by using a technique called *out-of-sample validation*. The idea is to have available two sets of data, one for model development and one for model validation. We initially develop one, or several, models on a first data set. The models developed from the first set of data are called our candidate models. Then, ideally, the relative performance of the *candidate models* could be measured on a second set of data. In this way, the data used to validate the model is unaffected by the procedures used to formulate the model.

Unfortunately, rarely will two sets of data be available to the investigator. However, we can implement the out-of-sample validation process by *splitting the data set* into two subsamples. We call these the *model development* and *validation subsamples*, respectively. To see how the data-splitting process works in the linear regression context, consider the following procedure.

Out-of-sample Validation Procedure

(i) Begin with a sample size of n and divide it into two subsamples, called the model development and validation subsamples. Let n_1 and n_2 denote the size of each subsample. In cross-sectional regression, do this split using a random sampling mechanism. Use the notation $i = 1, ..., n_1$ to represent observations from the model development subsample and $i = n_1 + 1, ..., n_1 + n_2 = n$ for the observations from the validation subsample. (In longitudinal data, we will use the first n_1 data points to predict the subsequent n_2 observations.) Figure 5.8 illustrates this procedure.

(ii) Using the model development subsample, fit a candidate model to the data set $i = 1, ..., n_1$.

(iii) Using the model created in Step 2 and the explanatory variables from the validation subsample, "predict" the dependent variables in the validation subsample, $\hat{y}_i$, where $i = n_1 + 1, ..., n_1 + n_2$. (To get these predictions, you may need to transform the dependent variables back to the original scale. This is discussed further in Section 5.6.)

(iv) Compute the *sum of squared prediction errors*

$$SSPE = \sum_{i=n_1+1}^{n_1+n_2} (y_i - \hat{y}_i)^2 \tag{5.4}$$

Repeat Steps 2 through 4 for each candidate model. Choose the model with the smallest *SSPE*.

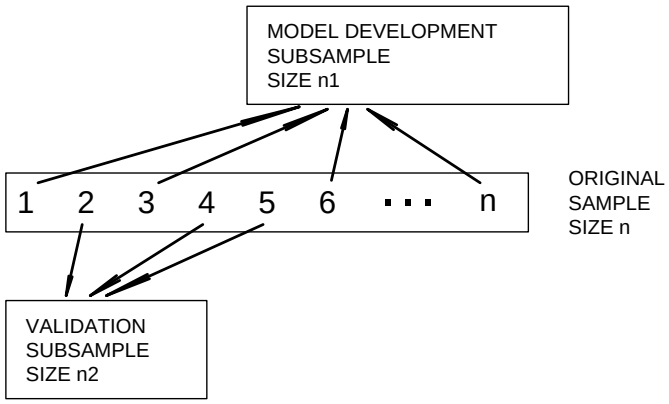

Fig. 5.8. For model validation, a data set of size n is randomly split into two subsamples.

There are a number of criticisms of the *SSPE*. First, it is clear that it takes a considerable amount of time and effort to calculate this statistic for each of several candidate models. However, as with many statistical techniques, this is merely a matter of having specialized statistical software available to perform the steps described above. Second, because the statistic itself is based on a random subset of the sample, its value will vary from analyst to analyst. This objection could be overcome by using the first n_1 observations from the sample. In most applications this is not done in case there is a lurking relationship in the order of the observations. Third, and perhaps most important, is the fact that the choice of the relative subset sizes, n_1 and n_2, is not clear. Various researchers recommend different proportions for the allocation. Snee (1977) suggests that data-splitting not be done unless the sample size is moderately large, specifically, $n \geq 2(k+1)+20$. The guidelines of Picard and Berk (1990) show that the greater the number of parameters to be estimated, the greater the proportion of observations needed for the model development subsample. As a rule of thumb, for data sets with 100 or fewer observations, use about 25-35% of the sample for out-of-sample validation. For data sets with 500 or more observations, use 50% of the sample for out-of-sample validation.

Because of these criticisms, several variants of the basic out-of-sample validation process are used by analysts. Although there is no theoretically best procedure, it is widely agreed that model validation is an important part of confirming the usefulness of a model.

PRESS Statistic

For small sample sizes, an attractive out-of-sample validation statistic is PRESS, the *Predicted Residual Sum of Squares*. To define the statistic, consider the following procedure where we suppose that a candidate model is available.

1. From the full sample, omit the ith point and use the remaining $n-1$ observations to compute regression coefficients.

2. Use the regression coefficients computed in step one and the explanatory variables for the ith observation to compute the predicted response, $\hat{y}_{(i)}$. This part of

the procedure is similar to steps one through three for calculating the *SSPE* statistic described above with $n_1 = n - 1$ and $n_2 = 1$.

3. Now, repeat this procedure for $i = 1, ..., n$, and define

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2. \tag{5.5}$$

As with *SSPE*, this statistic is calculated for each of several competing models. Under this criterion, we choose the model with the smallest *PRESS*.

At first glance, the statistic seems very computationally intensive in that it requires n regression fits to evaluate it. However, matrix algebra can be used to establish that

$$y_i - \hat{y}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}}. \tag{5.6}$$

Here, $\hat{e}_i$ and h_{ii} represent the ith residual and leverage from the regression fit using the complete data set. This yields

$$PRESS = \sum_{i=1}^{n} (\frac{\hat{e}_i}{1 - h_{ii}})^2, \tag{5.7}$$

which is a much easier computational formula. Thus, the *PRESS* statistic is less computationally intensive that *SSPE*.

Another important advantage of this statistic, when compared to SSPE, is that we do not need to make an arbitrary choice as to our relative subset sizes split. Indeed, because we are performing an "out-of-sample" validation for each observation, it can be argued that this procedure is more efficient, an especially important consideration when the sample size is small (say, less than 50 observations).

Because the model is re-fit for each point deleted, *PRESS* does not enjoy the appearance of independence between the estimation and prediction aspects, unlike *SSPE*. Further, out-of-sample validation is a general principle that is useful in a number of circumstances, including cross-sectional regression and time series. Although computationally attractive, the sample re-use principle that the *PRESS* statistic is based on is not as well understood for model selection purposes.

5.6 Handling Heteroscedasticity - Transformations

When fitting regression models to data, an important assumption is that the variability is common among all observations. This assumption of common variability is called *homoscedasticity* which stands for "same scatter." Indeed, the least squares procedure assumes that the expected variability of each observation is constant and gives the same weight to each observation when minimizing the sum of squared deviations. When the scatter varies by observation, the data are said to be *heteroscedastic*. Figure 5.9 is a plot of regression data with one explanatory variable where the scatter seems to increase as the explanatory variable increases.

To detect heteroscedasticity, a good idea is to perform a preliminary regression

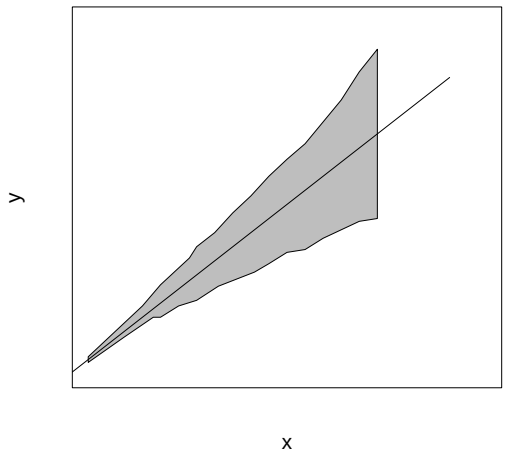

Fig. 5.9. The shaded area represents the data. The line is the true regression line.

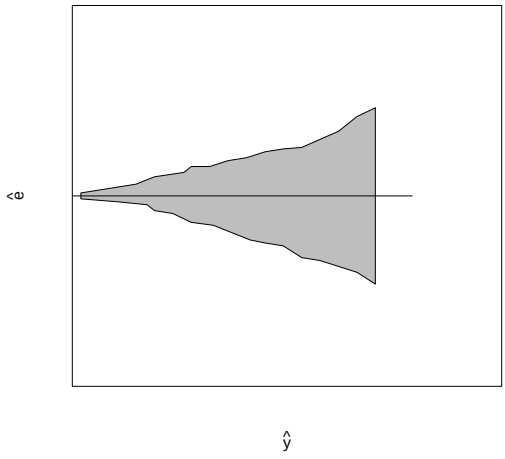

Fig. 5.10. Residuals plotted versus the fitted values for the data in Figure 5.9.

fit of the data and plot the residuals versus the fitted values. Figure 5.10 is an example of this plot. The preliminary regression fit removes many of the major patterns in the data and leaves the eye free to concentrate on other patterns that may influence the fit. We plot residuals versus fitted values because the fitted values are an approximation of the expected value of the response and, in many situations, the variability grows with the expected response.

Fortunately, when we perform regression fits on heteroscedastic data, it has been established that our resulting estimates are still unbiased. However, because the variability of the response differs from observation to observation, there is no common measure of variability, such as σ^2. Thus, many of our tests of hypotheses and confidence and prediction intervals are no longer valid. Further, our estimates are not as efficient as they could be, a particularly important point when the sample size is small.

When minimizing the sum of squared errors using heteroscedastic data, the expected variability of some observations is smaller than others. Intuitively, it seems reasonable that the smaller the variability of the response, the more reliable that response and the greater weight that it should receive in the minimization procedure. We will introduce a technique, called *weighted least squares*, in Chapter 8 that accounts for this "variable variability."

A simpler device that we pursue in this section is to simply transform the response variable. Even though this device is not always available, it has proven effective for a surprisingly large number of data sets. Many of the transforms that are used in practice can be expressed as part of the *Box-Cox family of transforms*. Within this family of transforms, in lieu of using the response y, we use a *transformed*, or *rescaled* version,$y^* = y^\lambda$. That is, the new response equals the old response raised to a specified power. Here, the power λ (lambda, a greek "el") is a number that is user specified. Typical values of λ that are used in practice are $\lambda = 1, 1/2, 0$ or -1. When we use $\lambda = 0$, we mean $y^* = ln(y)$, that is, the natural logarithmic transform. More formally, the Box-Cox family can be expressed as

$$y^* = \frac{y^\lambda - 1}{\lambda}$$

Because regression estimates are not affected by location and scale shifts, in practice we do not need to subtract one nor divide by λ when rescaling the response. (The advantage of the above expression is that, if we let λ approach 0, then y^* approaches $ln(y)$, from some straightforward calculus arguments.)

Transformation of the response may help to stabilize the variance and to achieve a symmetric distribution of responses, and thus of errors. To see the effects on the distribution of responses, Figure 5.11 displays the distribution of a set of responses and several transformations. The data were created by simulating 250 observations from the sum of five squared standard normal variates. As we can see from the histogram in the upper right hand of Figure 5.11, this distribution is skewed to the right. Histograms of the data transformed using the square root, logarithmic and negative reciprocal transformations are in the left hand column of Figure 5.11. We see that the square root and logarithmic transformation have served to symmetrize the distribution although the negative reciprocal transformation has produced a distribution that is skewed to the left. The three scatter plots in Figure 5.11, between y and each of $y^{1/2}, ln(y)$ and $-1/y$, indicate that each of these transformations is highly nonlinear.

Transforming the response variable has an effect on the distribution of the response and the expected variability of the response. If the only reason for investigating a transformation is to handle a potential nonlinear relation between a response and an explanatory variable, then it is simpler to transform the explanatory variable. We introduced this topic in Section 4.5.

Values of the transformation parameter $\lambda < 1$ serve to "shrink" spread out data. There are several special cases when certain choices of λ work well. For example, when the response is a type of *count data*, then the square root transform is suggested. Here, count data refers to the number of some entity, for example, the number of people in a family.

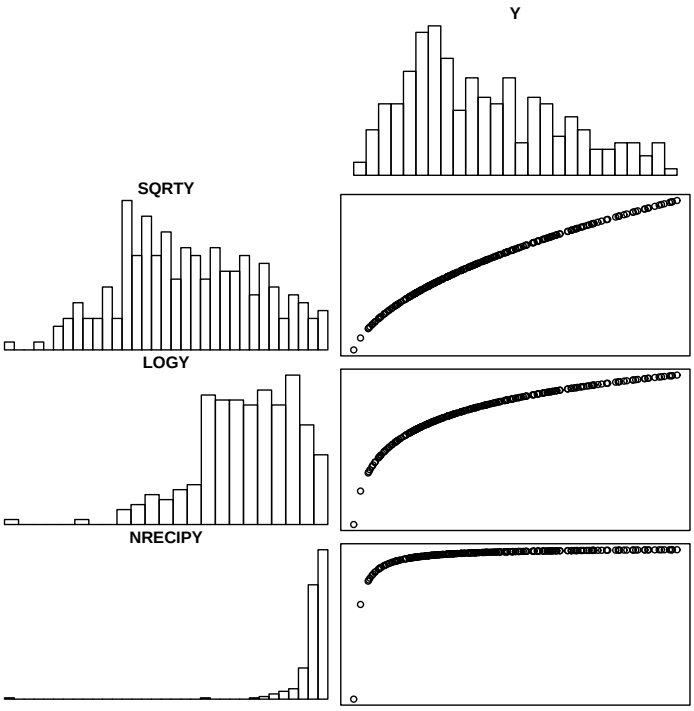

Fig. 5.11. Effects of the square root, logarithmic and negative reciprocal transformations on the distribution of a response.

Another example in which the use of logarithmic transformations arise naturally are in connection with *multiplicative models*. Here the general set-up is that response may be modeled as the expected response *times* the natural error, that is, $y = (\mathrm{E}y)e$. For example, consider modeling the response M_{od}, the number of migrants from a specified state of origin o to a specified destination state d. The so-called "gravity model" is (see, for example, Frees, 1992),

$$M_{od} = c\frac{P_o P_d}{D_{od}^c}(\frac{I_d}{I_o})^b(\frac{E_d}{E_o})^f e_{od}$$

for migration from the oth to dth state. Here, P is state population, I is state income, E is state (un)employment, D is distance between population centroids, a, b, c and f are parameters to be estimated, and e_{od} is the multiplicative error term. This model can be easily converted to the linear model via the logarithmic transform.

When using the natural logarithm (base e) transformation, there is a useful interpretation of the regression coefficients. Recall the idea that a partial slope is interpreted as the expected change in the response per unit change in the explanatory variables, holding other explanatory variables fixed. Thus, if the response is in natural logarithmic units, then a per unit change can be interpreted as a proportional (or, when multiplied by 100, percentage) change in the original units of the response. See the discussion towards the end of Section 5.7 for an illustration of this point.

5.7 Case Study: NFL Players' Compensation

In this section we report on a study of compensation of players in the National Football League (NFL). Our goal is to understand the relationship between a player's salary and various personal characteristics that may influence salary. These relationships could be useful in predicting or determining salaries of future players. This study may also be useful in salary arbitration matters. For example, based on certain characteristics of a player, the model could be used to determine an expected salary. The difference between an actual salary and that expected under the model could represent an important deviation of a player from his peers. The reason for this deviation would be the subject of arbitration. Of course, our usual caveats for interpreting regression models. In particular, the regression model establishes the size of the deviation, not the reasons why the deviation occurs.

Data Sources and Characteristics

Data for the players' salaries were provided by the *NFL Players' Association*. Of the 1,570 salary figures available at the beginning of the 1990 season, a random sample of 200 players was drawn. The salaries represent the response variable in this study. Also available from the Players' Association was the POSITION of the player, the round in which the player was DRAFTed and the years of experience of a player (YRS EXP). Additional personal characteristics of the players were collected from the *1990 Media Guide*. These characteristics included the number of regular season games PLAYED in the previous year and the number of regular season games STARTED in the previous year. Only one characteristic about the team that the player belongs to was investigated. This was the size of the city in which the team is domiciled (CITY POP). These data were collected from *The Statistical Abstract of the United States, 1990*. The conjecture here was that larger cities may have larger salaries to offset the higher cost of living compared to smaller metropolitan areas.

After a preliminary analysis of the data, it became clear that rookies (players with YRS EXP = 0) followed a different compensation market than veterans. Thus, 35 rookie players were excluded from the analysis. Further, two veteran players were unusual with respect to other players in the data set. These were Warren Moon, veteran quarterback of the Houston Oilers, and Bo Jackson, part-time player for the Los Angeles Raiders (and professional baseball's Kansas City Royals). The circumstances of each player are unusual and it was decided not to try to accommodate these unusual circumstances with the model. Thus, these players were also deleted, leaving us with a data set of size 163. Unfortunately, the media guide did not include information for three teams. For this reason, the data regarding games PLAYED and STARTED was missing for 26 veteran players.

An examination of plots and summary statistics reveals several interesting aspects of the data. To begin, Table 5.6 is a correlation matrix and Figure 5.12 displays a series of scatter plots of the response versus each of the explanatory variables. We see that games started and years of experience have a strong influence on salaries. The variable draft has a strong negative effect, indicating that the lower is the round in which a player was selected into the league (DRAFTed), the higher is the salary. A closer examination of the scatter plot matrix in Figure 5.12 reveals that the reciprocal of draft (1/DRAFT) may provide a better fit. The motivation for this is that there is a large difference between the first and second rounds in a draft when

compared to, say, differences between the ninth and tenth rounds. This is reflected in the reciprocal of draft but not the draft variable itself.

TABLE 5.6 Correlation Matrix

	DRAFT	YRS EXP	PLAYED	STARTED	CITYPOP
YRS EXP	-0.041				
PLAYED	-0.105	0.383			
STARTED	-0.284	0.410	0.516		
CITYPOP	-0.128	-0.015	0.186	0.156	
SALARY	-0.412	0.457	0.301	0.559	0.044

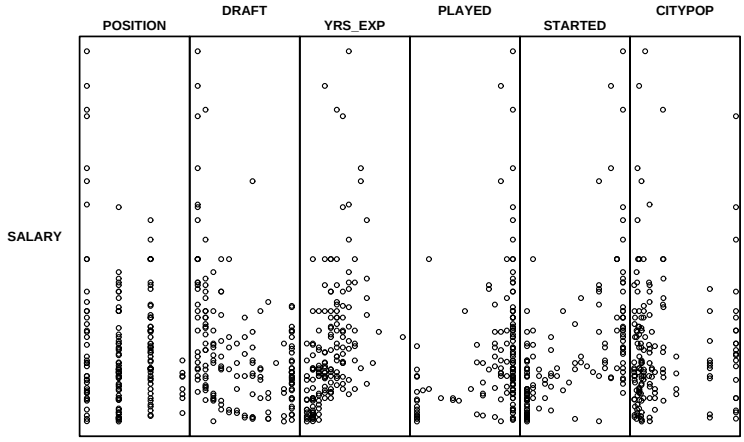

Fig. 5.12. Scatter plot of SALARY versus several explanatory variables. *Source: NFL Players' Association, 1990 Media Guide, and the 1990 Statistical Abstract of the U.S.*

The categorical variable POSITION was split into four indicator variables, one for offensive back (POSITION=1), defensive back (POSITION=2), lineman (POSITION=3) and kicker/punter (POSITION=4). It turned out that only the indicator for offensive back (OB) was important in the subsequent analysis. That is, once we control for other personal characteristics of a player, it did not seem to matter if the player was a defensive back, lineman or kicker/punter.

Finally, Figures 5.13 and 5.14 are histograms of the response variable SALARY. In Figure 5.13 we see that salaries are skewed to the right. Note that this is a real effect and not merely the result of one or two players earning large salaries. (Recall that we have already deleted two players with large salaries.) To "bring in" players with large salaries, we intend to argue that it is more appropriate to model the *logarithm* of the salaries as the response variable than the salary itself. Figure 5.14 is the histogram of the salaries but now the horizontal axis is on a logarithmic (base ten) scale. Using this scale, players with large salaries are still large but not as dramatically large as on the original scale.

Preliminary Regression Model

In anticipation of the model validation step, 18 observations were randomly selected. These observations were held out to be used in subsequent model validation. Thus, in the preliminary model development stages, you will find $163 - 18 = 145$ observations.

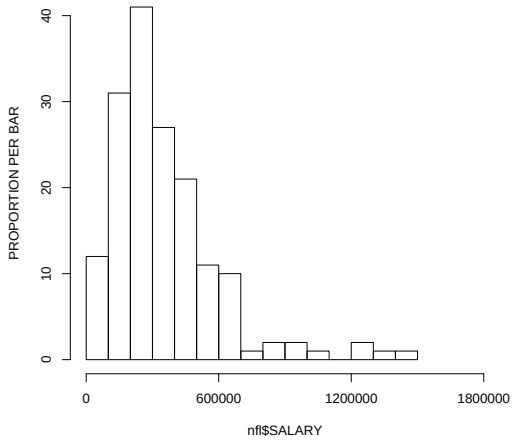

Fig. 5.13. Histogram of the salaries of 163 veteran NFL players.

One way of quickly arriving at a candidate model is to use the automatic variable selection procedures described in Section 5.1. Exhibits 5.1 and 5.2 are examples of the results based on forward and backwards stepwise regression routines. For these routines, the criterion for adding and deleting a variable was to declare the variable unimportant if the variable's t-ratio is less than two in absolute value. A comforting fact is that both algorithms wind up with the same recommended model.

EXHIBIT 5.1 Minitab output of the Forward Stepwise Regression Routine.

FORWARD STEPWISE REGRESSION OF SLARY
ON 7 PREDICTORS, WITH N = 121
N (CASES WITH ISSING OBS.)
= 24 N (ALL CASES) = 145

STEP	1	2	3	4
CONTANT	209372	45145	42902	9602
1/DRAFT	418194	399961	328178	302858
T-RATIO	7.18	7.71	5.89	5.62
YRS EXP		33465	25866	26596
T-RATIO		5.68	4.14	4.44
STARTED			8867	9701
T-RATIO			2.98	3.39
OB				125901
T-RATIO				3.36
S	215257	191570	185473	177836
R-SQ	30.25	45.22	49.09	53.59

EXHIBIT 5.2 Minitab Output of the Backward Stepwise Regression Routine.

BACKWARD STEPWISE REGRESSION OF SALARY
ON 7 PREDICTORS, WITH N = 121
N (CASES WITH MISSING OBS.)
= 24 N (ALL CASES) = 145

STEP	1	2	3	4
CONSTANT	-34307	-29475	-40280	9620
DRAFT	4750	4508	4629	
T-RATIO	0.77	0.74	0.77	
YRS EXP	27012	27010	26687	26596
T-RATIO	4.37	4.39	4.44	4.44
PLAYED	-1123	-1018		
T-RATIO	-0.29	-0.27		
STARTED	9880	9967	9627	9701
T-RATIO	3.11	3.17	3.35	3.39
CITYPOP	0.0008			
T-RATIO	0.24			
1/DRAFT	355802	352610	355126	302858
T-RATIO	3.99	4.01	4.08	5.61
OB	132132	131697	131966	125901
T-RATIO	3.41	3.41	3.44	3.36
S	179618	178875	178152	177836
R-SQ	53.88	53.86	53.83	53.59

Based on these procedures and further examination of the data, a preliminary fit of the data was made using SALARY as the response variable and 1/DRAFT, YRS EXP, STARTED and OB as explanatory variables. The model fit well. The adjusted coefficient of determination was $R_a^2 = 52\%$ and the size of the typical error was $s \approx 178,000$, a reduction from $s_y \approx 250,000$. Figure 5.15 is a plot of the standardized residuals versus the fitted values from this regression fit. It seems that there is a tendency to make large mistakes for individuals who have a high expected salary under the model. We call this tendency a heteroscedastic error. One way to compensate for this violation of model assumptions is to consider a transformed version of the response variable, logged salaries.

Before continuing, it should be noted that the choice of *base* that one uses in the logarithm does not affect the essential arguments here. The choice of base ten has especially desirable interpretations when one deals with monetary figures; Figure 5.14 is an example of this. Demographers and other scientists often prefer base two because it is the natural scale to see how often entities, such as populations, double over time. In this text, we use the natural logarithm that uses the base $e \approx 2.7182818$.

From the above discussion, the next step is to explore models using the logarithm of salary (LNSALARY) as the response. To illustrate further automatic variable selection procedures, in Exhibit 5.3 is the output of a best subsets regression proce-

dure. Using the criteria R^2, R_a^2, s and C_p, it seems as if the best model is the one using the same four explanatory variables, 1/DRAFT, YRS EXP, STARTED and OB, as above. This model is fit to the data. Figure 5.16 is a plot of standardized residuals of this model versus the fitted values. Based on this plot, no patterns are apparent. Further diagnostic checks were also made on the data. Although not included here, a plot of residuals versus years of experience indicated the potential need to include a quadratic term in years of experience. When this variable was included in the regression model, the t-ratio associated with this variable was -3.43, indicating a great deal of significance. Thus, this variable was also considered when validating our candidate models.

EXHIBIT 5.3 Minitab Output of the Best Subsets Regression Routine.

BEST SUBSETS REGRESSION of LNSALARY
121 cases used 24 cases contain missng values.

						YRS	PLAY	START	CITY	1/DRAFT	
		Adj.			DRAFT	EXP	ED	ED	OP		OB
Vars	R-sq	R-sq	C-p	s							
1	36.4	35.8	72.6	0.52227				X			
1	29.3	28.7	93.6	0.55044		X					
2	54.7	53.9	20.0	0.44260		X				X	
2	47.2	46.3	42.2	0.47757		X		X			
3	59.9	58.9	6.5	0.41819		X		X		X	
3	56.1	55.0	17.8	0.43746		X				X	X
4	61.8	60.5	2.7	0.40967		X		X		X	X
4	60.1	58.7	8.0	0.41908		X		X	X	X	
5	62.0	60.4	4.2	0.41043		X		X	X	X	X
5	61.8	60.2	4.7	0.41134	X	X		X		X	X

Insert Figure 5.15 here
Insert Figure 5.16 here

Model Validation

In this subsection, we consider three candidate models. The first uses SALARY as the response variable and uses 1/DRAFT, YRS EXP, STARTED and OB as explanatory variables. The second model has the same explanatory variables but uses LNSALARY as the response. The third candidate model also uses LNSALARY as the response and adds years of experience squared (EXP SQR) to the list of explanatory variables. These three models are compared based on their ability to predict the sample that was held out for validation purposes. This subset was randomly selected from the original sample. Thus, to the extent that observations are independent, any mischief that we may have inadvertently gotten into by overanalyzing the data should not affect the results of this independent validation. Of course, the drawback is that we are allowing our model choice to be driven by a small percentage of the data. One can always increase this percentage of the data, but then there are fewer observations upon which we can base our opinion as to what viable candidate models might be.

For the subset of 18 observations originally held out, it turns out that some of the media guide data were unavailable for two of the observations. Thus, only 16

observations, about 10% of the data set, were available for validation. The three fitted models are:

$$\begin{aligned}\text{Model 1} \quad : \widehat{\text{SALARY}} &= 9,6202 + 26,596 \text{ YRS EXP} + 9,701 \text{ STARTED} \\ &\quad +302,858 \text{ 1/DRAFT} + 125,901 \text{ OB}\end{aligned}$$

$$\begin{aligned}\text{Model 2} \quad : \widehat{\text{LNSALARY}} &= 11.6 + 0.0907 \text{ YRS EXP} + 0.0275\text{STARTED} \\ &\quad +0.721 \text{ 1/DRAFT} + 0.210 \text{ OB}\end{aligned}$$

$$\begin{aligned}\text{Model 3} \quad : \widehat{\text{LNSALARY}} &= 11.3 + 0.214 \text{ YRS EXP} + 0.0227 \text{ STARTED} \\ &\quad +0.717 \text{ 1/DRAFT} + 0.216 \text{ OB} - 0.00932 \text{ EXP SQR.}\end{aligned}$$

Insert Figure 5.17 here

For the values of the held-out explanatory variables, predictions were made for each model. For example, Figure 5.17 is a scatter plot of actual salaries of the 16 held-out players versus their predicted salaries based on the first model. To compare the models, predictions from Models 2 and 3 were exponentiated so that the predictions, as well as actual salaries, would be in dollars. For each model, the deviation between actual salary and predicted salary was calculated, squared and then summed over all 16 individuals. Table 5.7 presents the result of this procedure in the form of the sum of squared prediction errors *(SSPE)* statistic, which was first described in Section 5.5.

TABLE 5.7 Sum of Squared Prediction Errors for Three Competing Models

Model	SSPE(in Millions of Dollars Squared)
Model 1	0.2795
Model 2	0.1692
Model 3	0.2209

The SSPE statistics are not meaningful in and of themselves but they do allow for a meaningful comparison between models using different scales for the responses. Based on this work, it seems that Model 2 is the preferred choice.

Transformed Responses: PRESS and Regression Coefficient Interpretation

To complete the model validation process, the PRESS statistic was calculated for each model using the full data set. Because the responses are in different units, they must be converted to the same scale. That is, using equation (5.8) directly yields a *PRESS* statistic in $(\text{dollars})^2$ units for Model 1 and in $(\text{logged dollars})^2$ for Models 2 and 3. This conversion also needs to be done for the *SSPE* statistic. The conversion procedure for *SSPE* is similar to the one for the *PRESS* and is easier. For example, to convert the Model 2 *PRESS* statistic to $(\text{dollars})^2$ units, we use the following steps:

1. Run the regression model using LNSALARY as the response. Obtain residuals LNRESID and leverages HI.

2. Using equation (5.7), compute each omit i fitted value, $\text{LNSALARY}_{(i)} = \text{LNSALARY}_i - \text{LNRESID}_i/(1-\text{HI}_i)$.

3. Express each omit i fitted value in dollars through exponentiation, that is, define $\text{SALARY}_{(i)} = exp(\text{LNSALARY}_{(i)})$.

4. Summarize the omit i fitted values through the *PRESS* statistic using

$$\text{PRESS} = \sum_{i=1}^{n} (\text{SALARY}_I - \text{SALARY}_{(i)})^2.$$

The result of these calculations are in Table 5.8. In this table, we provide values of the *PRESS* statistic for each of the three models, for each type of unit. The *PRESS* statistic is presented in both (dollars)^2 and (logged dollars)^2 units because different results can be obtained depending on the type of unit. However, Table 5.8 shows that the third model is the best in terms of *PRESS*, regardless of the units of measurement.

TABLE 5.8 Predicted Residual Sum of Squares for Three Competing Models

Model	*PRESS*(in Millions of dollars squared)	*PRESS*(in logged dollars squared)
Model 1	4.3475	31.633
Model 2	4.3373	23.041
Model 3	3.7823	21.430

From our model validation processes, Model 1 was the poorest candidate model. This model displayed the highest *SSPE* and *PRESS* statistics. This presents a strong argument for using the logarithm of salaries as the response. The choice between Models 2 and 3 is less clear; Model 2 outperforms Model 3 based on the SSPE although the reverse is true for the PRESS criterion. You might elect to choose Model 2 on the *principle of parsimony*; choosing the simplest model possible to represent the real world. Here is the result of the Model 2 estimation procedure using the full sample of n=163 observations.

$$\widehat{\text{LNSALARY}} = 11.6 + 0.0861 \text{ YRS EXP} + 0.0332 \text{ STARTED}$$

$$\textit{std errors} \quad (0.0782) \qquad (0.01269) \qquad\qquad (0.005925)$$

$$+0.642 \text{ 1/DRAFT} + 0.170 \text{ OB.}$$
$$(0.1135) \qquad\qquad (0.08074)$$

$$s = 0.3996 \qquad R^2 = 62.4\% \qquad R_a^2 = 61.2\%$$

This model could then be used for predicting a player's salary for arbitration purposes. The idea here is that if various characteristics about a player are known, in particular, the years of experience, number of games started in the prior year, the round selected in his first year (DRAFT) and position played, then an estimate of the player's salary can be made. This estimate is not perfect but it is interpreted as

a weighted average of the many other players in the league having similar characteristics. In this way a player can be compared to his peers. We also have a measure of the natural variation in the data. That is, every player represents a special circumstance. We allow for individuality in the regression model and yet, at the same time, attempt to measure the impact of individuality through the concept of natural variation.

We have a special interpretation for the estimated coefficients when using a logarithmic response. When using the natural logarithm transformation, a per unit change can be interpreted as a proportional (or, when multiplied by 100, percentage) change in the original units of the response. Thus, we can interpret the coefficient associated with YRS EXP as meaning that salary increases by 8.61% per unit change in years of experience. The idea here is that, because the logarithm of salary increases by .0861, then the increase in salary is (SALARY) 1.0861. That is, suppose that YRS EXP increases by one unit, causing salary to go from $\text{SALARY}_{\text{old}}$ to $\text{SALARY}_{\text{new}}$. Then, the unit increase means that the log salary increase is .0861, or,

$$increase = .0861 = \ln(\text{SALARY}_{\text{new}}) - \ln(\text{SALARY}_{\text{old}}) = \ln(\frac{\text{SALARY}_{\text{new}}}{\text{SALARY}_{\text{old}}}).$$

$$\text{Thus, } \frac{\text{SALARY}_{\text{new}}}{\text{SALARY}_{\text{old}}} = e^{.0861} \approx 1.0861.$$

To sumarize, in business and economic studies we generally use the natural logarithmic transformation. This is due to (i) the interpretation of the regression coefficients as proportional changes and (ii) the natural logarithmic transformation is a member of the Box-Cox family of transforms.

5.8 Summary

Chapter 6 plays a pivotal role in our development of multiple linear regression models in Chapters 4 through 7. Chapter 4 introduced the model, with estimation and basic statistical inference ideas. Chapter 5 extended the linear regression model to include categorical independent variables. These two chapters are part of the "science" of statistics; ideas that are grounded on firm results in mathematical statistics and that are subject to relatively little disagreement among researchers. Chapter 6 provides tools and guidelines for selecting a model to represent a data set. This chapter is part of the "art" of statistics; different analysts, regardless of their diligence and thoughtfulness, will use different models to represent the same data set.

In this chapter, we discussed some of the common difficulties encountered when fitting regression models to data and provided practical suggestions for dealing with these difficulties. Unusual observations cause difficulties because regression models use weighted averages to estimate coefficients. In Sections 5.2 and 5.3, we explored two types of unusual observations, outliers and high leverage points. We think of these techniques as directed towards "data criticism" because the statistics vary at the observation level. In contrast, the techniques introduced in Sections 5.4 through 5.6 used statistics at the variable level. Specifically, the extent to which one independent variable duplicates the information contained in other variables was discussed

in Section 5.4. Criteria for selecting a regression model, and procedures for arriving at candidate models were discussed in Sections 5.1, 5.4, and 5.5. Heteroscedasticity, and how to handle this model difficulty using a rescaling, or transformation, of the response variables was addressed in Section 5.6. The case study in Section 5.7 provides an example to illustrate how to deal with the situation when several of these difficulties occur in a single data set.

Selecting a model is an inexact aspect of statistics. As such, you must carefully defend your choice of model selection. Model selection determines much of what we can say about a data set. Thus, when selecting a model, we are essentially reasoning with data. The arguments for, and against, choosing a specific model should be carefully spelled out. Chapter 7 summarizes the many lines of arguments that we can use when interpreting the results of a regression study. We will see, although the model selection formally drives the interpretation of results, that potential interpretations of results also influence our choice of a model.

Key Words, Phrases and Symbols - Chapter 5

After reading this chapter, you should be able to define each of the following important terms, phrases and symbols in your own words. If not, go to the page indicated and review the definition.

Insert Vocabulary here