5 Estimation of Linear Model Parameters Using Least Squares

This chapter deals with methods to estimate parameters of linear parametric models using ordinary least squares (OLS). The univariate case is first reviewed along with equations for the uncertainty in the model estimates as well as in the model predictions. Several goodness-of-fit indices to evaluate the model fit are also discussed, and the assumptions inherent in OLS are highlighted. Next, multiple linear models are treated, and several notions specific to correlated regressors are presented. The insights which residual analysis provides are discussed, and different types of remedial actions to improper model residuals are addressed. Other types of linear models such as splines and models with indicator variables are discussed. Finally, a real-world case study analysis which was meant to verify whether actual field tests supported the claim that a refrigerant additive improved chiller thermal performance is discussed.

5.1 Introduction

The analysis of observational data or data obtained from designed experiments often requires the identification of a statistical model or relationship which captures the underlying structure of the system from which the sample data was drawn. A model is a relation between the variation of one variable (called the *dependent or response variable*) against that of other variables (called *independent or regressor variables*). If observations (or data) are taken of both response and regressor variables under various sets of conditions, one can build a mathematical model from this information which can then be used as a predictive tool under different sets of conditions. How to analyze the relationships among variables and determine a (if not "the") optimal relation, falls under the realm of *regression model building or regression analysis.*

Models, as stated in Sect. 1.1, can be of different forms, with mathematical models being of sole concern in this book. These can divided into:

(i) *parameteric models* which can be a single function (or a set of functions) capturing the variation of the response variable in terms of the regressors. The intent is to identify both the model function and determine the values of the parameters of the model along with some indication of their uncertainty; and

(ii) *nonparametric models* where the relationship between response and regressors is such that a mathematical model in the conventional sense is inadequate. Nonparameteric models are treated in Sect. 9.3 in the framework of time series models and in Sect. 11.3.2 when dealing with artificial neural network models.

The parameters appearing in parametric models can be estimated in a number of ways, of which ordinary least squares (OLS) is the most common and historically the oldest. Other estimation techniques are described in Chap. 10. There is a direct link between how the model parameters are estimated and the underlying joint probability distributions of the variables, which is discussed below and in Chap. 10. In this chapter, only *models linear in the parameters* are addressed which need not necessarily be linear models (see Sect. 1.2.4 for relevant discussion). However, often, the former are loosely referred to as linear parametric models.

5.2 Regression Analysis

5.2.1 Objective of Regression Analysis

The objectives of regression analysis are: (i) to identify the "best" model among several candidates in case the physics of the system does not provide an unique mechanistic relationship, and (ii) to determine the "best" values of the model parameters; with "best" being based on some criterion yet to be defined. Desirable properties of estimators, which are viewed as random variables, have been described in Sect. 4.7.2, and most of these concepts apply to parameter estimation of regression models as well.

T. Agami Reddy, *Applied Data Analysis and Modeling for Energy Engineers and Scientists,*
DOI 10.1007/978-1-4419-9613-8_5, © Springer Science+Business Media, LLC 2011

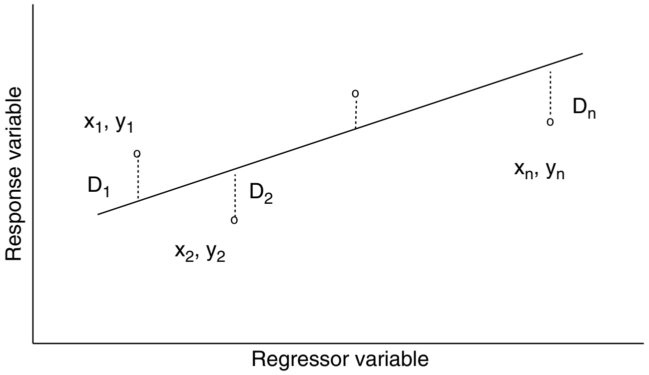


Fig. 5.1 Ordinary Least Squares Regression (OLS) is based on finding the model parameters which minimize the squared sum of the vertical deviations or residuals or $\left(\sum_{i=1}^{n} D_i^2\right)^{1/2}$

5.2.2 Ordinary Least Squares

Once a set of data is available, what is the best model which can be fit to the data. Consider the (x, y) set of n data points shown in Fig. 5.1. The criterion for "best fit" should be objective, intuitively reasonable and relatively easy to implement mathematically. One would like to minimize the deviations of the points from the prospective regression line. The method most often used is the *method of least squares* where, as the name implies, the "best fit" line is interpreted as one which minimizes the sum of the squares of the residuals. Since it is based on minimizing the squared deviations, it is also referred to as the Method of Moments Estimation (MME). The most common and widely used sub-class of least squares is the ordinary least squares (OLS) where, as shown in Fig. 5.1, squared sum of the vertical differences between the line and the observation points are minimized, i.e., $\min(D_1^2 + D_2^2 + \cdots + D_n^2)$. Another criterion for determining the best fit line could be to minimize the sum of the absolute deviations, i.e., $\min(|D_1| + |D_2| + \cdots |D_n|)$. However, the mathematics to deal with absolute quantities becomes cumbersome and restrictive, and that is why historically, the method of least squares was proposed and developed. Inferential statistics plays an important part in regression model building because identification of the system structure via a regression line from sample data has an obvious parallel to inferring population mean from sample data (discussed in Sect. 4.2.1). The intent of regression is to capture or "explain" via a model the variation in y for different x values. Taking a simple mean value of y (see Fig. 5.2a) leaves a lot of the variation in y unexplained. Once a model is fit, however, the unexplained variation is much reduced as the regression line accounts for some of the variation that is due to x (see Fig. 5.2b, c). Further, the assumption of normally distributed variables, often made in inferential theory, is also presumed for the distribution of the population of y values at each x value (see Fig. 5.3). Here, one notes that when slices of data are made at different values of x, the individual y distributions are *close to normal with equal variance*.

5.3 Simple OLS Regression

5.3.1 Traditional Simple Linear Regression

Let us consider a simple linear model with two parameters, a and b, given by:

$$y = a + b \cdot x \tag{5.1}$$

The parameter 'a' denotes the model intercept, i.e., the value of y at x=0, while the parameter 'b' is the slope of the straight line represented by the simple model (see Fig. 5.4). The objective of the regression analysis is to determine the numerical values of the parameters a and b which result in the model given by Eq. 5.1 able to *best* explain the variation of y about its mean $\bar{y}$ as the numerical value of the regressor variable x changes. Note that the slope parameter b explains the *variation* in y due to that in x. It does not necessarily follow that this parameter accounts for more of the observed *absolute* magnitude in y than does the intercept parameter term a. For any y value, the total deviation can be partitioned into two pieces: explained and unexplained (recall the ANOVA approach presented in Sect. 4.3 which is based on the same conceptual approach). Mathematically,

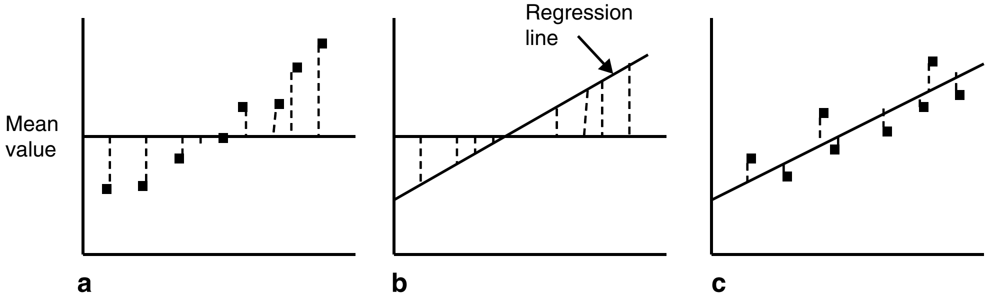


Fig. 5.2 Conceptual illustration of how regression explains or reduces unexplained variation in the response variable. It is important to note that the variation in the response variable is taken in reference to its mean value. **a** total variation (before regression), **b** explained variation (due to regression), **c** residual variation (after regression)

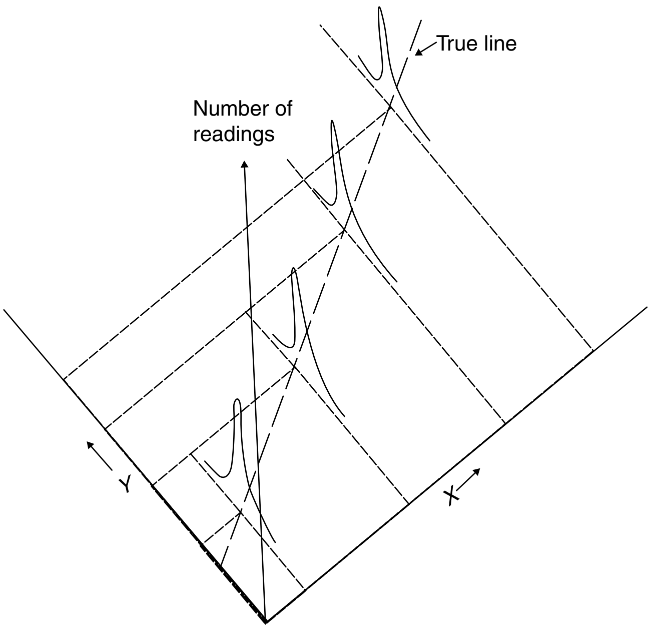


Fig. 5.3 Illustration of normally distributed errors with equal variances at different discrete slices of the regressor variable values. Normally distributed errors is one of the basic assumptions in OLS regression analysis. (From Schenck 1969 by permission of McGraw-Hill)

$$\begin{aligned}\sum_{i=1}^{n}(\mathrm{y_i}-\bar{\mathrm{y}})^2 &= \sum_{i=1}^{n}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^2+\sum_{i=1}^{n}(\hat{\mathrm{y}}_\mathrm{i}-\bar{\mathrm{y}})^2 \\ \text{or}\quad \mathrm{SST} &= \qquad \mathrm{SSE} \qquad + \qquad \mathrm{SSR}\end{aligned} \tag{5.2}$$

where

$\mathrm{y_i}$ is the individual response at observation i,

$\bar{\mathrm{y}}$ the mean value of $\mathrm{y_i}$ of the n observations,

$\hat{\mathrm{y}}_i$ the value of y estimated from the regression model for observation i,

SST = total sum of squares,

SSE = error sum of squares or sum of the residuals which reflects the variation about the regression line (similar to Eq. 4.20 when dealing with ANOVA type of problems), and

SSR = regression sum of squares which reflects the amount of variation in y explained by the model (similar to treatment sum of squares of Eq. 4.19).

These quantities are conceptually illustrated in Fig. 5.4. The sum of squares minimization implies that one wishes to minimize SSE, i.e.

$$\sum_{i=1}^{n}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^2=\sum_{i=1}^{n}(\mathrm{y_i}-\mathrm{a}-\mathrm{bx_i})^2=\sum_{i=1}^{n}\varepsilon_i^2 \tag{5.3}$$

where ε is called the model residuals or error.

From basic calculus, the model residuals are minimized when:

$$\frac{\partial\sum\varepsilon^2}{\partial \mathrm{a}}=0 \quad\text{and}\quad \frac{\partial\sum\varepsilon^2}{\partial \mathrm{b}}=0$$

The above two equations lead to the following equations (called the *normal equations*):

$$\begin{aligned}\mathrm{na}+\mathrm{b}\sum\mathrm{x} &= \sum\mathrm{y}\quad\text{and}\\ \mathrm{a}\sum\mathrm{x}+\mathrm{b}\sum\mathrm{x}^2 &= \sum\mathrm{xy}\end{aligned} \tag{5.4}$$

where n is the number of observations. This leads to the following expressions of the most "probable" OLS values of a and b:

$$\mathrm{b}=\frac{\mathrm{n}\sum\mathrm{x_iy_i}-\left(\sum\mathrm{x_i}\right)\left(\sum\mathrm{y_i}\right)}{\mathrm{n}\sum\mathrm{x_i^2}-\left(\sum\mathrm{x_i}\right)^2}=\frac{S_{xy}}{S_{xx}} \tag{5.5a}$$

$$\mathrm{a}=\frac{\left(\sum\mathrm{y_i}\right)\left(\sum\mathrm{x_i^2}\right)-\left(\sum\mathrm{x_iy_i}\right)\left(\sum\mathrm{x_i}\right)}{\mathrm{n}\sum\mathrm{x_i^2}-\left(\sum\mathrm{x_i}\right)^2}=\bar{y}-b\cdot\bar{x} \tag{5.5b}$$

where

$$\begin{aligned}S_{xy}&=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\quad\text{and}\\ S_{xx}&=\sum_{i=1}^{n}(x_i-\bar{x})^2\end{aligned} \tag{5.6}$$

Fig. 5.4 The value of regression in reducing unexplained variation in the response variable as illustrated by using a single observed point. The total variation from the mean of the response variable is partitioned into two portions: one that is explained by the regression model and the other which is the unexplained deviation, also referred to as model residual

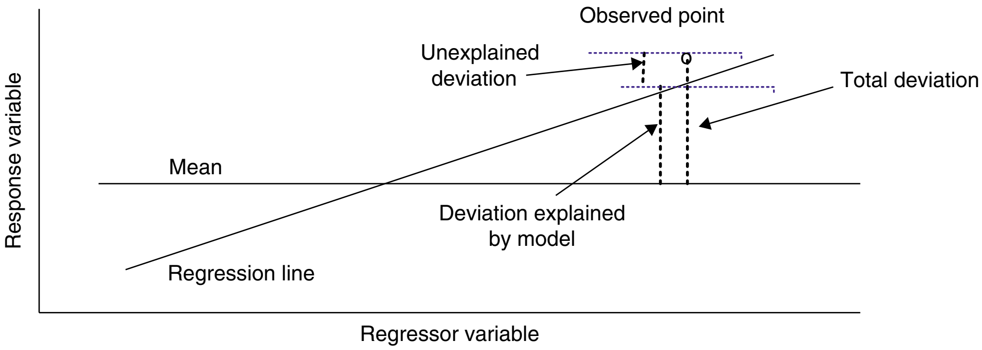


Table 5.1 Data table for Example 5.3.1

Solids reduction x (%)	Chemical oxygen demand, y (%)	Solids reduction x (%)	Chemical oxygen demand, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

How the least squares regression model reduces the unexplained variation in the response variable is conceptually illustrated in Fig. 5.4.

Example 5.3.1:[1] *Water pollution model between solids reduction and chemical oxygen demand*
In an effort to determine a regression model between tannery waste (expressed as solids reduction) and water pollution (expressed as chemical oxygen demand), sample data (33 observation sets) shown in Table 5.1 were collected. Estimate the parameters of a linear model.

The regression line is estimated by first calculating the following quantities:

$$\sum_i^{33} x_i = 1104, \quad \sum_i^{33} y_i = 1124,$$
$$\sum_i^{33} x_i \cdot y_i = 41,355, \quad \sum_i^{33} x_i^2 = 41,086$$

Subsequently Eqs. 5.5a and b are used to compute:

$$b = \frac{(33)(41,355) - (1104)(1124)}{(33)(41,086) - (1104)^2} = 0.9036$$

$$a = \frac{1124 - (0.903643)(1104)}{33} = 3.8296$$

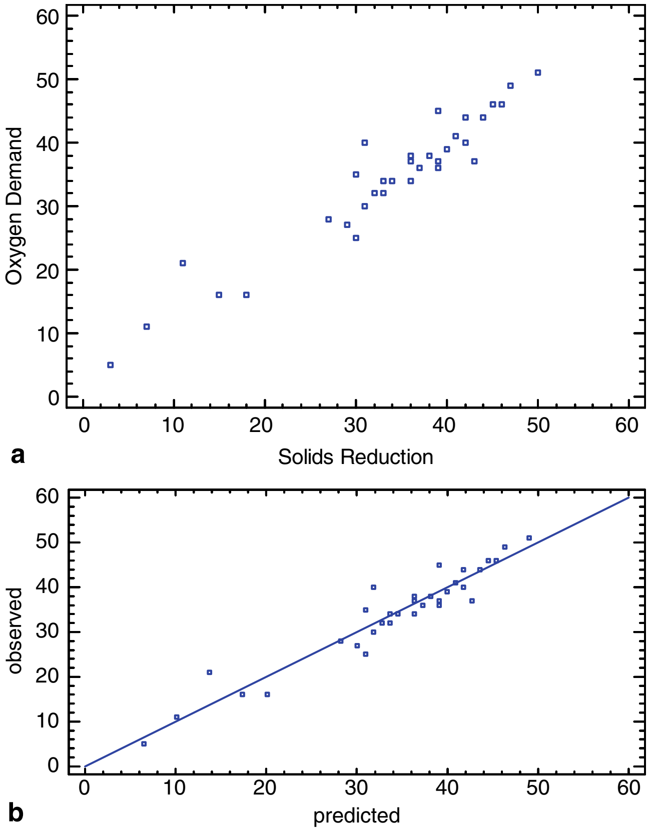


Fig. 5.5 **a** Scatter plot of data **b** Plot of observed versus OLS model predicted values of the y variable

Thus, the estimated regression line is:

$$\hat{y} = 3.8296 + 0.9036 \cdot x$$

The above data is plotted as a scatter plot in Fig. 5.5a. How well the regression model performs compared to the measurements is conveniently assessed from the observed vs predicted plot such as Fig. 5.5b. Tighter scatter of the data points around the regression line indicates more accurate model fit.

The regression line can be used for prediction purposes. The value of y at, say, x = 50 is simply:

$$\hat{y} = 3.8296 + (0.9036)(50) = 49 \qquad ■$$

5.3.2 Model Evaluation

(a) The most widely used measure of model adequacy or goodness-of-fit is the *coefficient of determination* R^2 where $0 \leq R^2 \leq 1$:

$$R^2 = \frac{\text{explained variation of y}}{\text{total variation of y}} = \frac{SSR}{SST} \qquad (5.7a)$$

[1] From Walpole et al. (1998) by © permission of Pearson Education.

For a perfect fit $R^2=1$, while $R^2=0$ indicates that either the model is useless or that no relationship exists. For a univariate linear model, R^2 is identical to the square of the Pearson correlation coefficient r (see Sect. 3.4.2). R^2 is a misleading statistic if models with different number of regressor variables are to be compared. The reason for this is that R^2 does not account for the number of degrees of freedom, it cannot but increase as additional variables are included in the model even if these variables have very little explicative power.

(b) A more desirable goodness-of-fit measure is *the corrected or adjusted* $\bar{R}^2$, computed as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{5.7b}$$

where n is the total number of observation sets, and k is the number of model parameters (for a simple linear model, k=2).

Since $\bar{R}^2$ concerns itself with variances and not variation, this eliminates the incentive to include additional variables in a model which have little or no explicative power. Thus, $\bar{R}^2$ is the right measure to use during identification of a parsimonious[2] model when multiple regressors are in contention. However, it should not be used to decide whether an intercept is to be added or not. For the intercept model, $\bar{R}^2$ is the proportion of variability measured by the sum of squares *about the mean* which is explained by the regression. Hence, for example, $\bar{R}^2 = 0.92$ would imply that 92% of the variation in the dependent variable about its mean value is explained by the model.

(c) Another widely used estimate of the magnitude of the absolute error of the model is the *root mean square error* (RMSE), defined as follows:

$$\mathrm{RMSE} = \left(\frac{\mathrm{SSE}}{\mathrm{n-k}}\right)^{1/2} \tag{5.8a}$$

where SSE is the sum of square error defined as

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (\mathrm{y_i - a - b \cdot x_i})^2. \tag{5.8b}$$

The RMSE is an absolute measure and its range is $0 \le \mathrm{RMSE} \le \infty$. Its units are the same as those of the y variable. It is also referred to as "*standard error of the estimate*".

A normalized measure is often more appropriate: the *coefficient of variation* of the RMSE (or CVRMSE or simply CV), defined as:

$$CV = \frac{RMSE}{\bar{\mathrm{y}}} \tag{5.8c}$$

Hence, a CV value of say 12% implies that the root mean value of the unexplained variation in the dependent variable y is 12% of the mean value of y.

Note that the CV defined thus is based on absolute errors. Hence, it tends to place less emphasis on deviations between model predictions and observations which occur at lower numerical values of y than at the high end. Consequently, the measure may inadequately represent the goodness of fit of the model over the entire range of variation under certain circumstances. An alternative definition of CV based on *relative mean deviations* is:

$$CV^* = \left\{ \frac{1}{(n-k)} \sum_{i=1}^{n} \left[\frac{(y_i - \hat{y}_i)}{y_i} \right]^2 \right\}^{1/2} \tag{5.8d}$$

If CV and CV* indices differ appreciably for a particular model, this would suggest that the model may be inadequate at the extreme range of variation of the response variable. Specifically, if CV*>CV, this would indicate that the model deviates more at the lower range, and vice versa.

(d) The *mean bias error* (MBE) is defined as the mean difference between the actual data values and model predicted values:

$$\mathrm{MBE} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{y_i - \hat{y}_i})}{\mathrm{n-k}} \tag{5.9a}$$

Note that when a model is identified by OLS, the model MBE of the original set of regressor variables used to identify the model should be zero (to within round-off errors of the computer). Only when, say, the model identified from a first set of observations is used to predict the value of the response variable under a second set of conditions will MBE be different than zero. Under such circumstances, the MBE is also called the mean *simulation or prediction error.* A normalized MBE (or NMBE) is often used, and is defined as the MBE given by Eq. 5.9a divided by the mean value $\bar{\mathrm{y}}$:

$$\mathrm{NMBE} = \frac{MBE}{\bar{y}} \tag{5.9b}$$

Competing models can be evaluated based on the CV and the NMBE values; i.e., those that have low CV and NMBE values. Under certain circumstances, one model may be preferable to another in terms of one index but not the other. The analysts is then perplexed as to which index to pick as the primary one. In such cases, the specific intent of how the model is going to be subsequently applied should be considered which may suggest the model selection criterion.

While fitting regression models, there is the possibility of "overfitting", i.e., the model fits part of the noise in the data along with the system behavior. In such cases, the model is likely to have poor predictive ability which often the analyst is unaware of. A statistical index is defined later (Eq. 5.42) which can be used to screen against this possibility. A better

[2] Parsimony in the context of regression model building is a term denoting the most succinct model, i.e., one without any statistically superfluous regressors.

way to minimize this effect is to randomly partition the data set into two (say, in proportion of 80/20), use the 80% portion of the data to develop the model, calculate the *internal predictive indices* CV and NMBE (following Eqs. 5.8c and 5.9b), use the 20% portion of the data and predict the y values using the already identified model, and finally calculate the *external or simulation* indices CV and NMBE. The competing models can then be compared, and a selection made, based on both the internal and external predictive indices. The simulation indices will generally be poorer than the internal predictive indices; larger discrepancies are suggestive of greater over-fitting, and vice versa. This method of model evaluation which can avoid model over-fitting is referred to as *holdout sample cross-validation* or simply cross-validation. Note, however, that though the same equations are used to compute the CV and NMBE indices, the degrees of freedom (df) are different. While df=n-k for computing the internal predictive errors where n is the number of observations used for model building, df=m for computing the external indices where m is the number of observations in the cross-validation set.

(e) The *mean absolute deviation* (MAD) is defined as the mean *absolute* difference between the actual data values and model predicted values:

$$\mathrm{MAD} = \frac{\sum\limits_{i=1}^{n} |\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}|}{\mathrm{n} - \mathrm{k}} \tag{5.10}$$

Example 5.3.2: Using the data from Example 5.3.1 repeat the exercise using your spreadsheet program. Calculate, R^2, RMSE and CV values.

From Eq. 5.2, SSE=323.3 and SSR=3713.88. From this SST=SSE+SSR=4037.2.

Then from Eq. 5.7a, R^2=92.0%, while from Eq. 5.8a, RMSE=3.2295, from which CV=0.095=9.5%. ■

5.3.3 Inferences on Regression Coefficients and Model Significance

Even after the overall regression model is found, one must guard against the fact that there may not be a significant relationship between the response and the regressor variables, in which case the entire identification process becomes suspect. The *F-statistic,* which tests for significance of the overall regression model, is defined as:

$$\begin{aligned} \mathrm{F} &= \frac{\text{variance explained by the regression}}{\text{variance not explained by the regression}} \\ &= \frac{SSR}{SSE} \cdot \frac{\mathrm{n} - \mathrm{k}}{\mathrm{k} - 1} \end{aligned} \tag{5.11}$$

Thus, the smaller the value of F, the poorer the regression model. It will be noted that the F-statistic is directly related to R^2 as follows:

$$\mathrm{F} = \frac{\mathrm{R}^2}{(1\text{–}\mathrm{R}^2)} \cdot \frac{\mathrm{n} - \mathrm{k}}{\mathrm{k} - 1} \tag{5.12}$$

Hence, the F-statistic can alternatively be viewed as being a *measure to test the R^2 significance itself.* In the case of univariate regression, the F-test is really the same as a t-test for the significance of the slope coefficient. In the general case, the F-test allows one to test the joint hypothesis of whether *all* coefficients of the regressor variables are equal to zero or not.

Example 5.3.3: Calculate the F-statistic for the model identified in Example 5.3.1. What can you conclude about the significance of the fitted model? From Eq. 5.11,

$$F = \left(\frac{3713.88}{323.3}\right) \cdot \left(\frac{33 - 2}{2 - 1}\right) = 356$$

which clearly indicates that the overall regression fit is significant. The reader can verify that Eq. 5.12 also yields an identical value of F. ■

Note that the values of coefficients a and b based on the given sample of n observations are only estimates of the true model parameters α and β. If the experiment is repeated over and over again, the estimates of a and b are likely to vary from one set of experimental observations to another. OLS estimation assumes that the model residual ε is a random variable with zero mean. Further, it is assumed that the residuals ε_i at specific values of x are randomly distributed, which is akin to saying that the distributions shown in Fig. 5.3 at specific values of x are normal and have equal variance.

After getting an overall picture of the regression model, it is useful to study the significance of each individual regressor on the overall statistical fit in the presence of all other regressors. The *student t-statistic* is widely used for this purpose and is applied to each regression parameter:

For the slope parameter:

$$\mathrm{t} = \frac{\mathrm{b} - 1}{s_b} \tag{5.13a}$$

where the estimated standard deviation of parameter "b" is $\mathrm{s_b} = RMSE/\sqrt{S_{xx}}$.

For the intercept parameter:

$$\mathrm{t} = \frac{\mathrm{a} - 0}{s_a} \tag{5.13b}$$

where the estimated standard deviation of parameter "a" is

$$\mathrm{s_a} = RMSE.\left(\frac{\sum\limits_{i}^{n} {x_i}^2}{n.S_{xx}}\right)^{1/2}$$

where b and a are the estimated slope and intercept coefficients, β and α the hypothesized true values, and RMSE is given by Eq. 5.8a. Estimated standard deviations of the coefficients b and a, given by Eqs. 5.13a and b, are usually referred to as *standard errors of the coefficients.* Basically, the t-test as applied to regression model building is a formal statistical test to determine how significantly different an individual coefficient is from zero in the presence of the remaining coefficients. Stated simply, it enables an answer to the following question: would the fit become poorer if the regressor variable in question is not used in the model at all?

The confidence intervals, assuming the model residuals to be normally distributed, are given by:

For the slope:

$$b - \frac{t_{\alpha/2} \cdot RMSE}{\sqrt{S_{xx}}} < \beta < b + \frac{t_{\alpha/2} \cdot RMSE}{\sqrt{S_{xx}}} \tag{5.14a}$$

For the intercept:

$$\begin{aligned} a - \frac{t_{\alpha/2} \cdot RMSE \cdot \sqrt{\sum_i^n x_i{}^2}}{\sqrt{n \cdot S_{xx}}} < \alpha < \\ a + \frac{t_{\alpha/2} \cdot RMSE \cdot \sqrt{\sum_i^n x_i{}^2}}{\sqrt{n \cdot S_{xx}}} \end{aligned} \tag{5.14b}$$

where $t_{\alpha/2}$ is the value of the t distribution with df=(n−2) and S_{xx} is defined by Eq. 5.6.

Example 5.3.4: In Example 5.3.1, the estimated value of b=0.9036. Test the hypothesis that β =1.0 as against the alternative that <1.0.

H_0: β=1.0
H_1: β<1.0

From Eq. 5.6a, S_{xx}=4152.1. Using Eq. 5.13a, with *RMSE*=3.2295

$$t = \frac{0.9036 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92$$

with n−2=31 degrees of freedom.

From Table A.4, the one-sided critical t-value for 95% CL=1.697. Since the computed t-value is greater than the critical value, one can reject the null hypothesis and conclude that there is strong evidence to support β<1 at the 95% confidence level. ■

Example 5.3.5: In Example 5.3.1, the estimated value of a=3.8296. Test the hypothesis that α =0 as against the alternative that $\alpha \neq 0$ at the 95% confidence level.

H_0: α=0
H_1: $\alpha \neq 0$

Using Eq. 5.13b,

$$t = \frac{3.8296 - 0}{3.2295/\sqrt{41,086/(33)(4152.18)}} = 2.17$$

with n−2=31 degrees of freedom.

Again, one can reject the null hypothesis, and conclude that $\alpha \neq 0$ at 95% CL. ■

Example 5.3.6: Find the 95% confidence interval for the slope term of the linear model identified in Example 5.3.1.

Assuming a two-tailed test, $t_{0.05/2}$=2.045 for 31 degrees of freedom. Therefore, the 95% confidence interval for β given by Eq. 5.14a is:

$$0.9036 - \frac{(2.045)(3.2295)}{(4152.18)^{1/2}} < \beta < 0.9036 + \frac{(2.045)(3.2295)}{(4152.18)^{1/2}}$$
$$0.8011 < \beta < 1.0061$$ ■

Example 5.3.7: Find the 95% confidence interval for the intercept term of the linear model identified in Example 5.3.1.

Again, assuming a two-tailed test, and using Eq. 5.14b, the 95% confidence interval for α is:

$$\begin{aligned} 3.8296 - \frac{(2.045)(3.2295)\sqrt{41{,}086}}{[(33)(4152.18)]^{1/2}} < \alpha < \\ 3.8296 + \frac{(2.045)(3.2295)\sqrt{41{,}086}}{[(33)(4152.18)]^{1/2}} \\ 0.2131 < \alpha < 7.4461 \end{aligned}$$ ■

5.3.4 Model Prediction Uncertainty

A regression equation can be used to predict future values of y *provided* the x value is within the domain of the original data from which the model was identified. One differentiates between the two types of predictions (similar to the confidence limits of the mean treated in Sect. 4.2.1.b):

(a) **mean response** or *standard error of regression* where one would like to predict the mean value of y for a large number of repeated x_0 values. The mean value is directly deduced from the regression equation while the variance is:

$$\sigma^2(\hat{y}_0) = MSE \cdot \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \tag{5.15}$$

Note that the first term within the brackets, namely (MSE/n) is the standard error of the mean (see Eq. 4.2) while the other term is a result of the standard error of the slope coefficient. The latter has the effect of widening the uncertainty bands at either end of the range of variation of x.

(b) **individual or specific response** or *standard error of prediction* where one would like to predict the specific value of y for a specific value x_0. This error is larger than the error in the mean response by an amount equal to the RMSE. Thus,

$$\sigma^2(\hat{y}_0) = MSE \cdot \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \qquad (5.16)$$

Finally, the 95% CL for the individual response at level x_0 is:

$$y_0 = \hat{y}_0 \pm t_{0.05/2} \cdot \sigma(\hat{y}_0) \qquad (5.17)$$

where $t_{0.05/2}$ is the value of the t-student distribution at a significance level of 0.05 for a two-tailed error distribution. It is obvious that the *prediction intervals* for individual responses are wider than those of the mean response called *confidence levels* (see Fig. 5.6). Note that Eqs. 5.16 and 5.17 strictly apply when the errors are normally distributed.

Some texts state that the data set should be at least five to eight times larger than the number of model parameters to be identified. In case of *short data sets,* OLS may not yield robust estimates of model uncertainty and resampling methods are advocated (see Sect. 10.6.2).

Example 5.3.8: Calculate the 95% confidence limits (CL) for predicting the mean response for x = 20.

First, the regression model is used to calculate $\hat{y}_0$ at $x_0 = 20$:

$$\hat{y}_0 = 3.8296 + (0.9036)(20) = 21.9025$$

Using Eq. 5.15,

$$\sigma(\hat{y}_0) = (3.2295)\left[\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}\right]^{1/2} = 0.87793$$

Further, from Table A.4, $t_{0.05/2} = 2.04$ for d.f. = 33–2 = 31. Using Eq. 5.15 yields the confidence interval for the mean response

$$21.9025 - (2.04)(0.87793) < \mu(\hat{y}_{20}) < 21.9025 + (2.04)(0.87793)$$

or,

$$20.112 < \mu(\hat{y}_{20}) < 23.693 \text{ at } 95\% \text{ CL.}$$ ■

Example 5.3.9: Calculate the 95% prediction limits (PL) for predicting the individual response for x = 20.

Using Eq. 5.16,

$$\sigma(\hat{y}_0) = (3.2295)\left[1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}\right]^{1/2} = 3.3467$$

Further, $t_{0.05/2} = 2.04$. Using Eq. 5.17 yields

$$21.9025 - (2.04)(3.3467) < \hat{y}_{20} < 21.9025 + (2.04)(3.3467)$$

or

$$15.075 < \hat{y}_{20} < 28.730.$$ ■

5.4 Multiple OLS Regression

Regression models can be classified as:

(i) *single variate or multivariate,* depending on whether only one or several regressor variables are being considered;

(ii) *single equation or multi-equation* depending on whether only one or several response variables are being considered; and

(iii) *linear or non-linear,* depending on whether the model is linear or non-linear in its function. Note that the distinction is with respect to the parameters (and not its variables). Thus, a regression equation such as $y = a + b \cdot x + c \cdot x^2$ is said to be linear in its parameters {a, b, c} though it is non-linear in the regressor variable

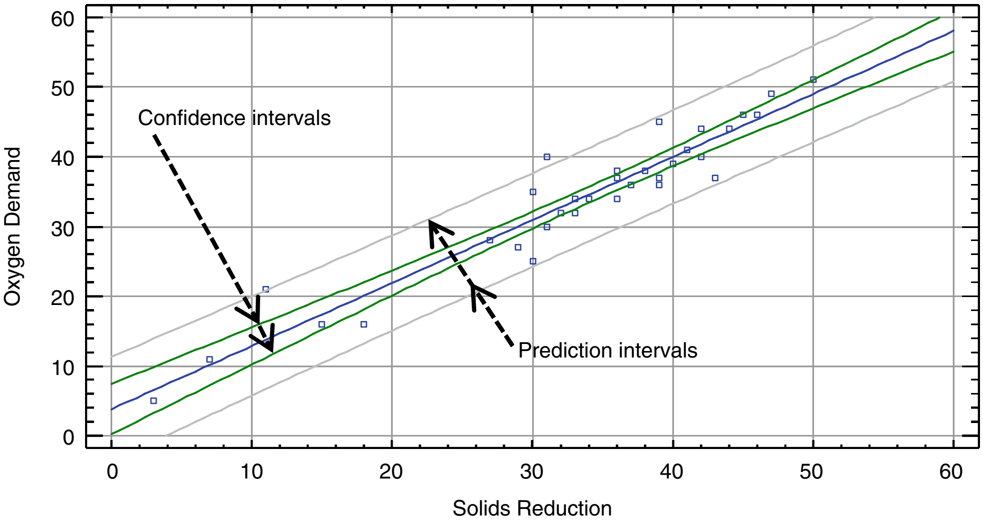


Fig. 5.6 95% confidence intervals and 95% prediction intervals about the regression line

x (see Sect. 1.2.3 for a discussion on classification of mathematical models).

Certain simple single variate equation models are shown in Fig. 5.7. Frame (a) depicts simple linear models (one with a positive slope and another with a negative slope), while (b) and (c) are higher order polynomial models which, though non-linear in the function, are models linear in their parameters. The other figures depict non-linear models. Because of the relative ease in linear model building, data analysts often formulate a linear model even if the relationship of the data is not strictly linear. If a function such as that shown in frame (d) is globally non-linear, and if the domain of the experiment is limited say to the right knee of the curve (bounded by points c and d), then a linear function in this region could be postulated. Models tend to be preferentially framed as linear ones largely due to the simplicity in the subsequent analysis and the prevalence of solution methods based on matrix algebra.

5.4.1 Higher Order Linear Models: Polynomial, Multivariate

When more than one regressor variable is known to influence the response variable, a multivariate model will explain more of the variation and provide better predictions than a single

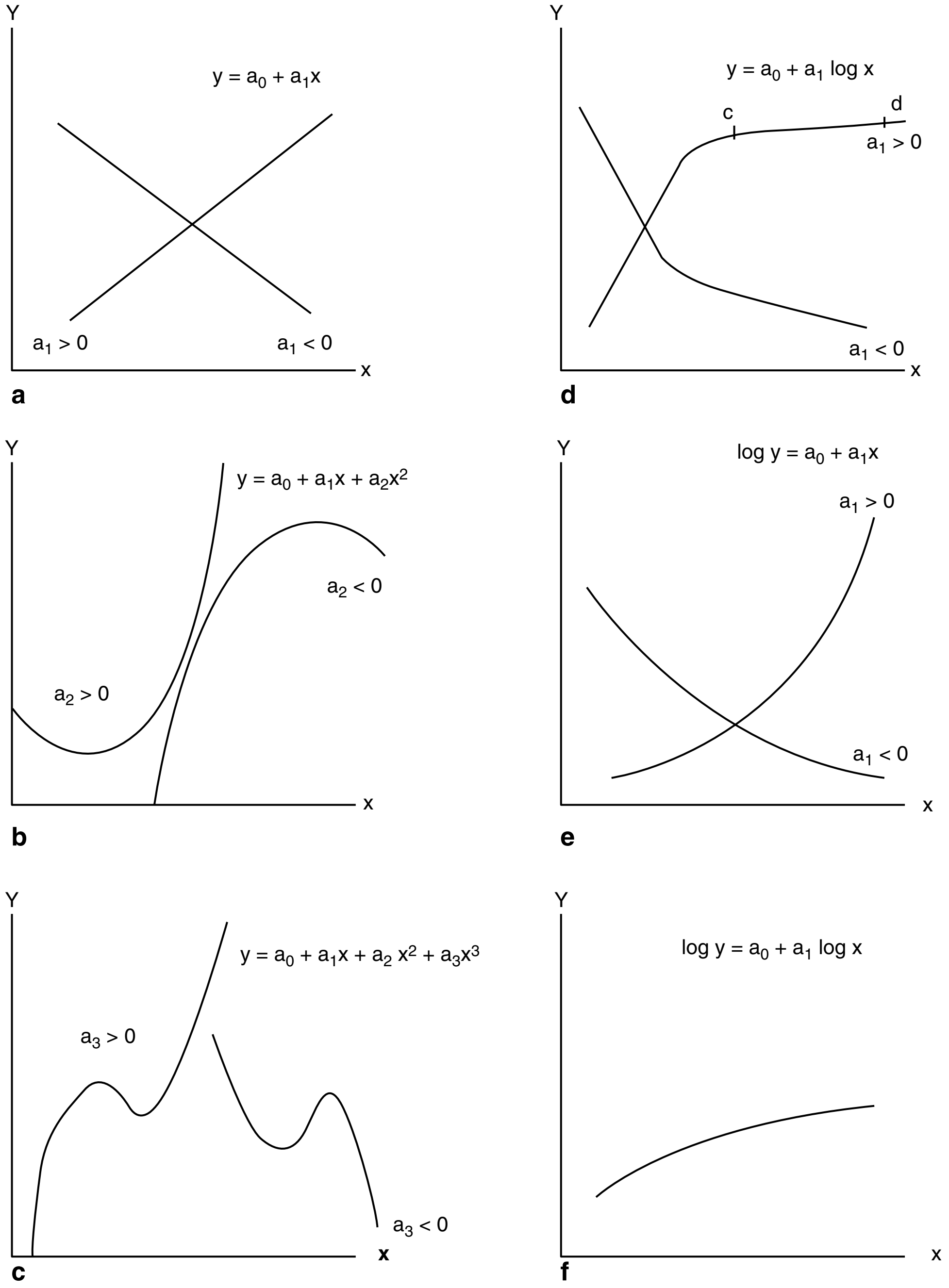


Fig. 5.7 General shape of regression curves. (From Shannon 1975 by © permission of Pearson Education)

variate model. The parameters of such a model need to be identified using multiple regression techniques. This section will discuss certain important issues regarding multivariate, single-equation models linear in the parameters. For now, the treatment is limited to regressors which are *uncorrelated or independent.* Consider a data set of n readings that include k regressor variables. The corresponding form, called the *additive multiple linear regression* model, is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (5.18a)$$

where ε is the error or unexplained variation in y. Note the lack of any interaction terms, and hence the term "additive". The simple interpretation of the model parameters is that β_i measures the unit influence of x_i on y (i.e., denotes the slope $\frac{dy}{dx_i}$). Note that this is strictly true only when the variables are really independent or uncorrelated, which, often, they are not.

The same model formulation is equally valid for a k-th degree polynomial regression model which is a special case of Eq. 5.18a with $x_1=x$, $x_2=x^2$ …

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon \qquad (5.19)$$

Let x_{ij} denote the i^{th} observation of parameter j. Then Eq. 5.18a can be re-written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad (5.18b)$$

Often, it is most convenient to consider the "normal" transformation where the regressor variables are expressed as a difference from the mean (the reason why this form is important will be discussed in Sect. 6.3 while dealing with experimental design methods). Specifically, Eq. 5.18a transforms into

$$\begin{aligned} y =& \beta_0{}' + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) \\ &+ \cdots + \beta_k(x_k - \bar{x}_k) + \varepsilon \end{aligned} \qquad (5.18c)$$

An important special case is the quadratic regression model when k=2. The straight line is now replaced by parabolic curves depending on the value of β (i.e., either positive or negative). Multivariate model development utilizes some of the same techniques as discussed in the single variable case. The first step is to identify all variables that can influence the response as predictor variables. It is the analyst's responsibility to identify these potential predictor variables based on his or her knowledge of the physical system. It is then possible to plot the response against all possible predictor variables in an effort to identify any obvious trends. The greatest single disadvantage to this approach is the sheer labor involved when the number of possible predictor variables is high.

A situation that arises in multivariate regression is the concept of variable synergy, or commonly called *interaction between variables* (this is a consideration in other problems; for example, when dealing with design of experiments). This occurs when two or more variables interact and impact system response to a degree greater than when the variables operate independently. In such a case, the *first-order linear model with two interacting regressor variables* takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2 + \varepsilon \qquad (5.20)$$

How the interaction parameter affects the shape of the family of curves is illustrated in Fig. 5.8. The origin of this model function is easy to derive. The lines for different values of regressor x_1 are essentially parallel, and so the slope terms for both models are equal. Let the model with the first regressor be: $y = a' + bx_1$, while the intercept be given by: $a' = f(x_2) = a + cx_2$. Combining both equations results in: $y = a + bx_1 + cx_2$. This corresponds to Fig. 5.8a. For the interaction case, both the slope and the intercept terms are function of x_2. Hence, representing $a' = a + bx_1$ and $b' = c + dx_1$, then:

$$y = a + bx_1 + (c + dx_1)x_2 = a + bx_1 + cx_2 + dx_1 x_2$$

which is identical in structure to Eq. 5.20.

Simple linear functions have been assumed above. It is straightforward to derive expressions for higher order models by analogy. For example, the *second-order (or quadratic) model without interacting variables* is:

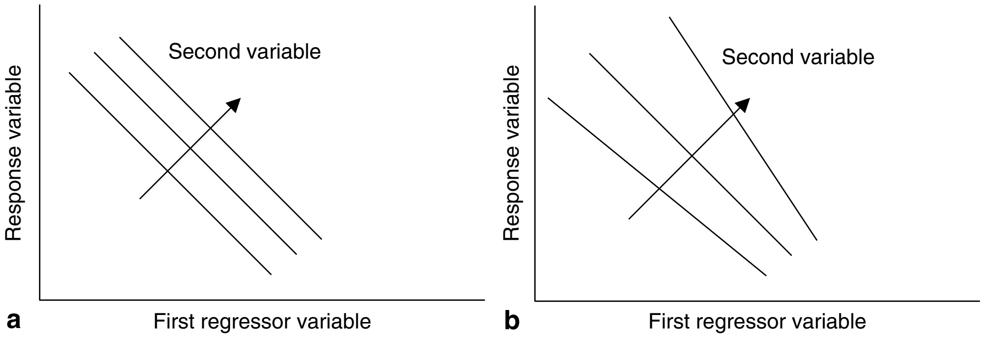


Fig. 5.8 Plots illustrating the effect of interaction among the regressor variables. **a** Non-interacting. **b** Interacting

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon \quad (5.21)$$

For a second order model *with interacting terms,* the corresponding expression can be derived as follows:

Consider the linear polynomial model with one regressor:

$$y = b_0 + b_1 x_1 + b_2 {x_1}^2 \quad (5.22)$$

If the parameters $\{b_0, b_1, b_2\}$ can themselves be expressed as second-order polynomials of another regressor x_2, the full model which has nine regression parameters is:

$$\begin{aligned} y = {} & b_{00} + b_{10}x_1 + b_{01}x_2 + b_{11}x_1x_2 \\ & + \mathrm{b}_{20}x_1^2 + b_{02}x_2^2 + b_{21}x_1^2x_2 \\ & + b_{12}x_1x_2^2 + b_{22}x_1^2x_2^2. \end{aligned} \quad (5.23)$$

The most general additive model, which imposes little structure to the relationship is given by:

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k) + \varepsilon \quad (5.24)$$

where the form of $f_i(x_i)$ are unspecified.

Note that synergistic behavior can result in two or more variables working together to "overpower" another variable's prediction capability. As a result, it is necessary to always check the importance (the relative value of either the t- or F-values) of each individual predictor variable while performing multivariate regression. Those variables with t- or F-values that are insignificant should be omitted from the model and the remaining predictors used to estimate the model parameters. The stepwise regression method described in Sect. 5.7.4 is based on this approach.

5.4.2 Matrix Formulation

When dealing with multiple regression, it is advantageous to resort to matrix algebra because of the compactness and ease of manipulation it offers without loss in clarity. Though the solution is conveniently provided by a computer, a basic understanding of matrix formulation is nonetheless useful. In matrix notation (with **y'** denoting the transpose of **y**), the linear model given by Eq. 5.18 can be expressed as follows (with the matrix dimension shown in subscripted brackets):

$$\mathrm{Y}_{(\mathrm{n},1)} = \mathrm{X}_{(\mathrm{n},\mathrm{p})}\beta_{(\mathrm{p},1)} + \varepsilon_{(\mathrm{n},1)} \quad (5.25)$$

where p is the number of parameters in the model = k + 1 (for a linear model), n is the number of observations
and

$$\mathrm{Y}' = [y_1\ y_2 \ldots y_n], \quad \beta' = [\beta_0\ \beta_1 \ldots \beta_\mathrm{k}], \quad \varepsilon' = [\varepsilon_1\ \varepsilon_2 \ldots \varepsilon_\mathrm{n}] \quad (5.26\mathrm{a})$$

and

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}. \quad (5.26\mathrm{b})$$

The descriptive measures applicable for a single variable can be extended to multivariables of order p (= k + 1), and written in compact matrix notation.

5.4.3 OLS Parameter Identification

The approach involving minimization of SSE for the univariate case (Sect. 5.3.1) can be generalized to multivariate linear regression. Here, the parameter set β is to be identified such that the sum of squares function L is minimized:

$$\mathrm{L} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_i^2 = \varepsilon'\varepsilon = (\mathrm{Y} - \mathrm{X}\beta)'(\mathrm{Y} - \mathrm{X}\beta) \quad (5.27)$$

or,

$$\frac{\partial \mathrm{L}}{\partial \beta} = -2\mathrm{X}'\mathrm{Y} + 2\mathrm{X}'\mathrm{X}\beta = 0 \quad (5.28)$$

which leads to the system of normal equations

$$\mathrm{X}'\mathrm{X}b = \mathrm{X}'\mathrm{Y}. \quad (5.29)$$

From here,

$$b = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y} \quad (5.30)$$

provided matrix X is not singular and where b is the least square estimator matrix of β.

Note that X'X is called the *variance-covariance matrix* of the estimated regression coefficients. It is a symmetrical matrix with the main diagonal elements being the sum of squares of the elements in the columns of X (i.e., the variances) and the off-diagonal elements being the sum of the cross-products (i.e., the covariances). Specifically,

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \mathrm{n} & \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}1} & \cdots & \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{ik}} \\ \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}1} & \sum_{i=1}^{n} \mathrm{x}_{\mathrm{i}1}^2 & \cdots & \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}1} \cdot \mathrm{x}_{\mathrm{ik}} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{ik}} & \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{ik}} \cdot \mathrm{x}_{\mathrm{i}1} & \cdots & \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{ik}}^2 \end{bmatrix}. \quad (5.31)$$

Under OLS regression, **b** is an unbiased estimator of β with the variance-covariance matrix var(b) given by:

$$\operatorname{var}(b) = \sigma^2 (X'X)^{-1} \tag{5.32}$$

where σ^2 is the mean square error of the model error terms

$$= (\text{sum of square errors})/(n - p) \tag{5.33}$$

An unbiased estimator of σ^2 is s^2, where

$$s^2 = \frac{\varepsilon'\varepsilon}{n-p} = \frac{y'y - b'x'y}{n-p} = \frac{SSE}{n-p} \tag{5.34}$$

For predictions within the range of variation of the original data, the mean and individual response values are normally distributed with the variance given by the following:

(a) For the *mean* response at a specific set of x_0 values, called the *confidence level,* under OLS

$$\operatorname{var}(\hat{y}_0) = s^2 \left[X_0 (X'X)^{-1} X_0'\right] \tag{5.35}$$

(b) The variance of an *individual* prediction, called the *prediction level,* is

$$\operatorname{var}(\hat{y}_0) = s^2 \left[\mathbf{1} + X_0 (X'X)^{-1} X_0'\right] \tag{5.36}$$

where **1** is a column vector of unity.

Confidence limits at a significance level α are:

$$y_0 \pm t(n - k, \alpha/2) \cdot \operatorname{var}^{1/2}(\hat{y}_0) \tag{5.37}$$

Example 5.4.1: *Part load performance of fans (and pumps)* Part-load performance curves do not follow the idealized fan laws due to various irreversible losses. For example, decreasing the flow rate by half of the rated flow does not result in a 1/8th decrease in its rated power consumption. Hence, actual tests are performed for such equipment under different levels of loading. The performance tests of the flow rate and the power consumed are then normalized by the rated or 100% load conditions called part load ratio (PLR) and fractional full-load power (FFLP) respectively. Polynomial models can then be fit between these two quantities. Data assembled in Table 5.2 were obtained from laboratory tests on a variable speed drive (VSD) control which is a very energy efficient control option.

(a) What is the matrix X in this case if a second order polynomial model is to be identified of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$?

Table 5.2 Data table for Example 5.4.1

PLR	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
FFLP	0.05	0.11	0.19	0.28	0.39	0.51	0.68	0.84	1.00

(b) Using the data given in the table, identify the model and report relevant statistics on both parameters and overall model fit.

(c) Compute the confidence bands and the prediction bands at 0.05 significance level for the response at values of PLR = 0.2 and 1.00 (i.e., the extreme points).

Solution

(a) The independent variable matrix X given by Eq. 5.26b is:

$$X = \begin{bmatrix} 1 & 0.2 & 0.05 \\ 1 & 0.3 & 0.11 \\ 1 & 0.4 & 0.19 \\ 1 & 0.5 & 0.28 \\ 1 & 0.6 & 0.39 \\ 1 & 0.7 & 0.51 \\ 1 & 0.8 & 0.68 \\ 1 & 0.9 & 0.84 \\ 1 & 1 & 1 \end{bmatrix}$$

(b) The results of the regression are shown below:

Parameter	Estimate	Standard error	t-statistic	P-value
CONSTANT	−0.0204762	−0.0173104	−1.18288	0.2816
PLR	0.179221	0.0643413	2.78547	0.0318
PLR^2	0.850649	0.0526868	16.1454	0.0000

Analysis of Variance

Source	Sum of squares	Df	Mean square	F-ratio	P-value
Model	0.886287	2	0.443144	5183.10	0.0000
Residual	0.000512987	6	0.0000854978		
Total (Corr.)	0.8868	8			

Goodness-of-fit $R^2 = 99.9\%$, Adjusted $R^2 = 99.9\%$, RMSE = 0.009246

Mean absolute error (MAD) = 0.00584. The equation of the fitted model is (with appropriate rounding)

$$\text{FFLP} = -0.0205 + 0.1792^*\text{PLR} + 0.8506^*\text{PLR}^2$$

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between FFLP and PLR at the 95% confidence level. However, the p-value of the constant term is large, and a model without an intercept term is probably more appropriate; thus, such an analysis ought to be performed, and its results evaluated. The values shown are those provided by the software package. There are too many significant decimals, and so the analyst should round these off appropriately while reporting the results (as shown above).

(c) The 95% confidence and the prediction intervals are shown in Fig. 5.9. Because the fit is excellent, these are very

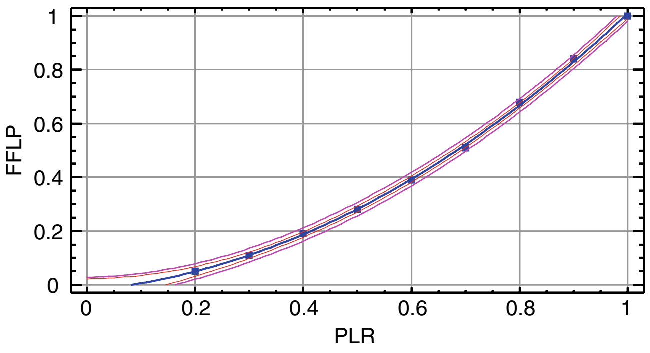


Fig. 5.9 Plot of fitted model with 95% CL and 95% PL bands

narrow and close to each other. The predicted values as well as the 95% CL and PL for the two data points are given in the table below. Note that the uncertainty range is relatively much larger at the lower value than at the higher range.

	Predicted	95% Prediction Limits		95% Confidence Limits	
x	y	Lower	Upper	Lower	Upper
0.2	0.0493939	0.0202378	0.0785501	0.0310045	0.0677834
1.0	1.00939	0.980238	1.03855	0.991005	1.02778

■

Example 5.4.2: Table 5.3 gives the solubility of oxygen in water in (mg/L) at 1 atm pressure for different temperatures and different chloride concentrations in (mg/L).

(a) Plot the data and formulate two different models to be evaluated

(b) Evaluate both models and identify the better one. Give justification for your choice

(c) Report pertinent statistics for model parameters as well as overall model fit

(a) The above data is plotted in Fig. 5.10a. One notes that the series of plots are slightly non-linear but parallel suggesting a higher order model without interaction terms. Hence, first order and second order polynomial models without interaction are logical models to investigate.

Table 5.3 Solubility of oxygen in water (mg/L) with temperature and chloride concentration

Temperature (°C)	Chloride concentration in water (mg/L)			
	0	5,000	10,000	15,000
0	14.62	13.73	12.89	12.10
5	12.77	12.02	11.32	10.66
10	11.29	10.66	10.06	9.49
15	10.08	9.54	9.03	8.54
20	9.09	8.62	8.17	7.75
25	8.26	7.85	7.46	7.08
30	7.56	7.19	6.85	6.51

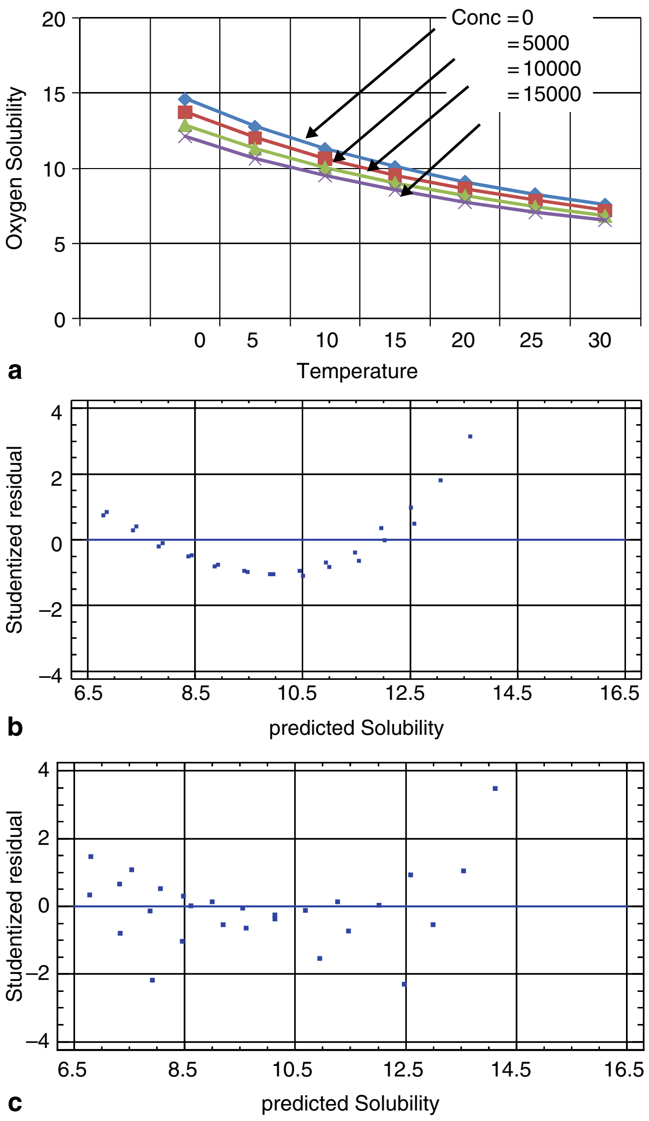


Fig. 5.10 **a** Plot of data. **b** Residual pattern for the first order model. **c** Residual pattern for the second order model

(b1) Analysis results of the first order model without interaction term:

$R^2 = 96.83\%$, Adjusted $R^2 = 96.57\%$, RMSE $= 0.41318$

Parameter	Estimate	Standard error	t-statistic	P-value
CONSTANT	13.6111	0.175471	77.5686	0.0000
Chloride Concentration	−0.000109857	0.000013968	−7.86489	0.0000
Temperature	−0.206786	0.00780837	−26.4826	0.0000

Analysis of Variance

Source	Sum of squares	Df	Mean square	F-ratio	P-value
Model	130.289	2	65.1445	381.59	0.0000
Residual	4.26795	25	0.170718		
Total (Corr.)	134.557	27			

The equation of the fitted model is:

Solubility = 13.6111 − 0.000109857 * Chloride Concentration − 0.206786 * Temperature

The model has excellent R^2 with all coefficients being statistically significant, but the model residuals are very ill-behaved since a distinct pattern can be seen (Fig. 5.10b). This issue of how model residuals can provide diagnostic insights into model building will be explored in detail in Sect. 5.6.

(b2) Analysis results for the second order model without interaction term:

The OLS regression results in $R^2 = 99.26\%$, Adjusted $R^2 = 99.13\%$, RMSE = 0.20864, Mean absolute error = 0.14367. This model is distinctly better with higher R^2 and lower RMSE. Except for one term (the square of the concentration), all parameters are statistically significant. The residual pattern is less distinct, but the residuals are still patterned (Fig. 5.10c). It would be advisable to investigate other functional forms, probably non-linear or based on some mechanistic insights.

Parameter	Estimate	Standard error	t-statistic	P-value
CONSTANT	14.1183	0.112448	125.554	0.0000
Temperature	−0.325	0.0142164	−22.8609	0.0000
Chloride concentration	−0.000118643	0.0000246866	−4.80596	0.0001
Temperature^2	0.00394048	0.000455289	8.65489	0.0000
Chloride concentration^2	5.85714E-10	1.57717E-9	0.371371	0.7138

Analysis of Variance

Source	Sum of squares	Df	Mean square	F-ratio	P-value
Model	133.556	4	33.3889	767.02	0.0000
Residual	1.0012	23	0.0435305		
Total (Corr.)	134.557	27			

■

5.4.4 Partial Correlation Coefficients

The simple correlation coefficient between two variables has already been introduced previously (Sect. 3.4.2). Consider the multivariate linear regression (MLR) model given by Eq. 5.18. If the regressors are uncorrelated, then the simple correlation coefficients provide a direct indication of the influence of the individual regressors on the response variable. Since regressors are often "somewhat" correlated, the concept of the simple correlation coefficient can be modified to handle such interactions. This leads to the concept of *partial correlation coefficients*. Assume a MLR model with only two regressors: x_1 and x_2. The procedure to compute the partial correlation coefficient r_{yx_1} between y and x_1 will make the concept clear:

Step 1: Regress y vs x_2 so as to identify a prediction model for $\hat{y}$

Step 2: Regress x_1 vs x_2 so as to identify a prediction model for $\hat{x}_1$

Step 3: Compute new variables (in essence, the model residuals): $y^* = y - \hat{y}$ and $x_1^* = x_1 - \hat{x}_1$

Step 4: The partial correlation r_{yx_1} between y and x_1 is the simple correlation coefficient between y^* and $x_1{}^*$

Note that the above procedure allows the linear influence of x_2 to be removed from both y and x_1, thereby enabling the partial correlation coefficient to describe only the effect of x_2 on y which is not accounted for by the other variables in the model. This concept plays a major role in the process of stepwise model identification described in Sect. 5.7.4.

5.4.5 Beta Coefficients and Elasticity

Beta coefficients β^* are occasionally used to make statements about the relative importance of the regressor variables in a multiple regression model (Pindyck and Rubinfeld 1981). These coefficients are the parameters of a linear regression model with each variable normalized by subtracting its mean and dividing by its standard deviation:

$$\frac{y-\bar{y}}{\sigma_y} = \beta_1^* \frac{x_1-\bar{x}_1}{\sigma_{x1}} + \beta_2^* \frac{x_2-\bar{x}_2}{\sigma_{x2}} + \cdots \varepsilon \tag{5.38}$$

or

$$y^* = \beta_1^* x_1{}^* + \beta_2^* x_2{}^* + \cdots \varepsilon$$

The β^* matrix can be directly deduced from the original slope parameter "b" of the un-normalized MLR model as:

$$\beta^* = b \cdot \frac{\sigma_x}{\sigma_y} \tag{5.39}$$

For example, the beta coefficient $\beta^* = 0.7$ can be interpreted to mean that one standard deviation in the regressor variable leads to a 0.7 standard deviation in the dependent variable. For a two-variable model, β^* is the simple correlation between the two variables. The rescaling associated with the normalized regression makes it possible to compare the individual values of β^* directly, i.e., the relative importance of the different regressors can be directly evaluated against each other, provided the regressors are uncorrelated with each other. A variable with a high β^* coefficient should account for more of the variance in the response variable (variance is not to be confused with contribution). The square of the β^* weights are indicative of the relative effects of the respective variables on the variation of the response variable.

The beta coefficients indicate or represent the marginal effect of the standardized regressors on the standardized response variable. Often, one is interested in deducing the effect of a fractional (or percentage) change of a regressor j on the dependent variable. This is provided by the *elasticity*

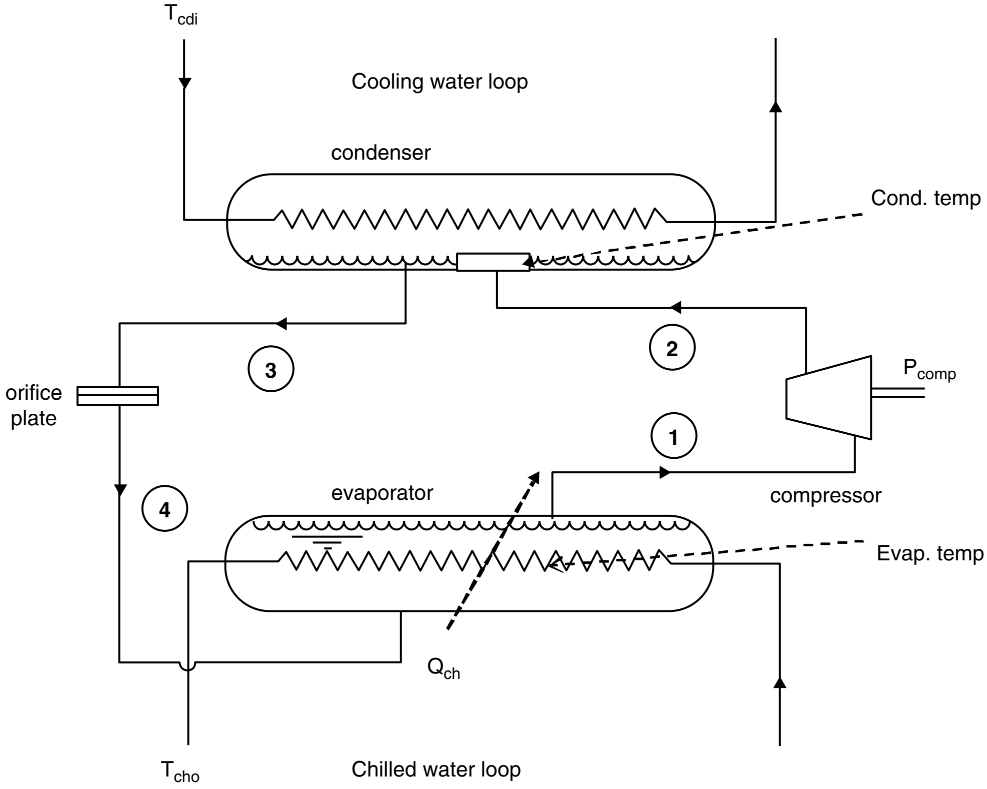


Fig. 5.11 Sketch of a flooded-type centrifugal chiller with two water loops showing the various regressors often used to develop the performance model for COP

of y with respect to say x_j which is usually evaluated at their mean values as:

$$E_j = b_j \cdot \frac{\bar{x}_j}{\bar{y}} \approx \frac{\partial y}{\bar{y}} \Big/ \frac{\partial x_j}{\bar{x}_j} \tag{5.40}$$

Elasticities can take both positive or negative values. Large values of elasticity imply that the regressor variable is very responsive to changes in the regressor variables. For non-linear functions, elasticities can also be calculated at the point of interest rather than at the mean point. The interpretation of elasticities is straightforward. If $E_j = 1.5$, this implies that a 1% increase in the mean of the regressor variable will result in a 1.5% increase in y.

Example 5.4.3: *Beta coefficients for ascertaining importance of driving variables for chiller thermal performance*

The thermal performance of a centrifugal chiller is characterized by the Coefficient of Performance (COP) which is the dimensionless ratio of the cooling thermal capacity (Q_{ch}) and the compressor electric power (P_{comp}) in consistent units. A commonly used performance model for the COP is one with three regressors, namely the cooling load Q_{ch}, the condenser inlet temperature T_{cdi} and chiller leaving temperature T_{cho} (see Fig. 5.11). The condenser and evaporator temperatures shown are those of the refrigerant as it changes phase.

A data set of 107 performance points from an actual chiller was obtained whose summary statistics are shown in the table below. An OLS regression yielded a model with $R^2 = 90.1\%$ whose slope coefficients b_j are also shown in Table 5.4 along with the beta coefficients and the elasticity computed from Eqs. 5.39 and 5.40 respectively. One would conclude looking at the elasticity values that T_{cdi} has the most influence on COP followed by Q_{ch}, while that of T_{cho} is very small. A 1% increase in Q_{ch} increases COP by 0.431% while a 1% increase in T_{cdi} would decrease COP by 0.603%. The beta coefficients, on the other hand, take into account the range of variation of the variables. For example, the load variable Q_{ch} can change from 20 to 100% while T_{cdi} usually changes only by 15°C or so. Thus, beta coefficients express the change in the COP of 0.839 in terms of one standard deviation change in Q_{ch} (i.e., a load change of 88.1 kW) while a comparable one standard deviation change in T_{cdi} (of 4.28°C) would result in a *decrease* of 0.496 in COP. ■

Table 5.4 Associated statistics of the four variables, results of the OLS regression and beta coefficients

	Response	Regressors		
	COP	Q_{ch} (kW)	T_{cdi} (°C)	T_{cho} (°C)
Mean	3.66	205.8	23.66	7.37
St. dev	0.806	88.09	4.283	2.298
Min	2.37	86	16.01	3.98
Max	4.98	361.4	29.95	10.94
Slope coeff. b		0.0077	−0.0933	0.0354
beta_coeff. (Eq. 5.39)		0.839	−0.496	0.101
Elasticity (Eq. 5.40)		0.431	−0.603	0.071

5.5 Assumptions and Sources of Error During OLS Parameter Estimation

5.5.1 Assumptions

The ordinary least square (OLS) regression method:

(i) enables simple or multiple linear regression models to be identified from data, which can then be used for future prediction of the response variable along with its uncertainty bands, and

(ii) allows statistical statements to be made about the estimated model parameters.

No statistical assumptions are used to obtain the OLS estimators for the model coefficients. When nothing is known regarding measurement errors, OLS is often the best choice for estimating the parameters. However, in order to make statistical statements about these estimators and the model predictions, it is necessary to acquire information regarding the measurement errors. Ideally, one would like the error terms ε_i to be normally distributed, without serial correlation, with mean zero and constant variance. The implications of each of these four assumptions, as well as a few additional ones, will be briefly addressed below since some of these violations may lead to biased coefficient estimates as well as distorted estimates of the standard errors, confidence intervals, and statistical tests.

(a) *Errors should have zero mean:* If this is not true, the OLS estimator of the intercept will be biased. The impact of this assumption not being correct is generally viewed as the least critical among the various assumptions. Mathematically, this implies that expected value $E(\varepsilon_i)=0$.

(b) *Errors should be normally distributed:* If this is not true, statistical tests and confidence intervals are incorrect for small samples though the OLS coefficient estimates are unbiased. Figure 5.3 which illustrates this behavior has already been discussed. This problem can be avoided by having large samples, and verifying that the model is properly specified.

(c) *Errors should have constant variance:* This violation of the basic OLS assumption results in increasing the standard errors of the estimates and widening the model prediction confidence intervals (though the OLS estimates themselves are unbiased). In this sense, there is a loss in statistical power. This condition is expressed mathematically as, var $(y_i)=\sigma^2$. This issue is discussed further in Sect. 5.6.3.

(d) *Errors should not be serially correlated:* This violation is equivalent to have less *independent* data, and also results in a loss in statistical power with the same consequences as (c) above. Serial correlations may occur due to the manner in which the experiment is carried out. Extraneous factors, i.e., factors beyond our control (such as the weather, for example) may leave little or no choice as to how the experiments are executed. An example of a reversible experiment is the classic pipe-friction experiment where the flow through a pipe is varied so as to cover both laminar and turbulent flows, and the associated friction drops are observed. Gradually increasing the flow one way (or decreasing it the other way) may introduce biases in the data which will subsequently also bias the model parameter estimates. In other circumstances, certain experiments are irreversible. For example, the loading on a steel sample to produce a stress-strain plot has to be performed by gradually increasing the loading till the sample breaks, one cannot proceed in the other direction. Usually the biases brought about by the test sequence are small, and this may not be crucial. In mathematical terms, this condition, for a first order case, can be written as $E(\varepsilon_i.\varepsilon_{i+1})=0$. This assumption, which is said to be hardest to verify, is further discussed in Sect. 5.6.4.

(e) *Errors should be uncorrelated with the regressors:* The consequences of this violation result in OLS coefficient estimates being biased and the predicted OLS confidence intervals understated, i.e., narrower. This violation is a very important one, and is often due to "mis-specification error" or underfitting. Omission of influential regressor variables and improper model formulation (assuming a linear relationship when it is not) are likely causes. This issue is discussed at more length in Sect. 10.4.1.

(f) *Regressors should not have any measurement error:* Violation of this assumption in some (or all) regressors will result in biased OLS coefficient estimates for those (or all) regressors. The model can be used for prediction but the confidence limits will be understated. Strictly speaking, this assumption is hardly ever satisfied since there is always some measurement error. However, in most engineering studies, measurement errors in the regressors are not large compared to the random errors in the response, and so this violation may not have important consequences. As a rough rule of thumb, this violation becomes important when the errors in x reach about a fifth of the random errors in y, and when multi-collinearity is present. If the errors in x are known, there are procedures which allow unbiased coefficient estimates to be determined (see Sect. 10.4.2). Mathematically, this condition is expressed as var $(x_i)=0$.

(g) *Regressor variables should be independent of each other:* This violation applies to models identified by multiple regression when the regressor variables are correlated among each other (called multicollinearity). This is true even if the model provides an excellent fit to the data. Estimated regression coefficients, though unbiased, will tend to be unstable (their values tend to change greatly when a data point is dropped or added), and the OLS standard errors and the prediction intervals will be understated. Multicollinearity is likely to be problem only when one (or more) of the correlation coefficients among the regressors exceeds 0.85 or so. Sect. 10.3 deals with this issue at more length.

5.5.2 Sources of Errors During Regression

Perhaps the most crucial issue during parameter identification is the type of measurement inaccuracy present. This has a direct influence on the estimation method to be used. Though statistical theory has more or less neatly classified this behavior into a finite number of groups, the data analyst is often stymied by data which does not fit into any one category. Remedial action advocated does not seem to entirely remove the adverse data conditioning. A certain amount of experience is required to surmount this type of adversity, which, further, is circumstance-specific. As discussed earlier, there can be two types of errors:

(a) **measurement error.** The following sub-cases can be identified depending on whether the error occurs:

(i) in the dependent variable, in which case the model form is:

$$\mathrm{y_i} + \delta_\mathrm{i} = \beta_0 + \beta_1 \mathrm{x_i} \tag{5.41a}$$

(ii) in the regressor variable, in which case the model form is:

$$\mathrm{y_i} = \beta_0 + \beta_1(\mathrm{x_i} + \gamma_i) \tag{5.41b}$$

(iii) in both dependent and regressor variables:

$$\mathrm{y_i} + \delta_\mathrm{i} = \beta_0 + \beta_1(\mathrm{x_i} + \gamma_i) \tag{5.41c}$$

Further, the errors δ and γ (which will be jointly represented by ε) can have an additive error, in which case, $\varepsilon_\mathrm{i} \neq \mathrm{f(y_i,\ x_i)}$, or a multiplicative error: $\varepsilon_\mathrm{i} = \mathrm{f(y_i,\ x_i)}$, or worst still, a combination of both. Section 10.4.1 discusses this issue further.

(b) **model misspecification error.** How this would affect the model residuals ε_i is difficult to predict, and is extremely circumstance-specific. Misspecification could be due to several factors, for example, one or more important variables have been left out of the model, or the functional form of the model is incorrect. Even if the physics of the phenomenon or of the system is well understood and can be cast in mathematical terms, identifiability constraints may require that a simplified or macroscopic model be used for parameter identification rather than the detailed model (see Sect. 10.2). This is likely to introduce both bias and random noise in the parameter estimation process except when model R^2 is very high ($\mathrm{R}^2 > 0.9$). This issue is further discussed in Sect. 5.6. Formal statistical procedures do not explicitly treat this case but limit themselves to type (a) errors and more specifically to case (i) assuming purely additive or multiplicative errors. The implicit assumptions in OLS and their implications, if violated, are described below.

5.6 Model Residual Analysis[3]

5.6.1 Detection of Ill-Conditioned Model Residual Behavior

The availability of statistical software has resulted in routine and easy application of OLS to multiple linear models. However, there are several underlying assumptions that affect the individual parameter estimates of the model as well as the overall model itself. Once a model has been identified, the general tendency of the analyst is to hasten and use the model for whatever purpose intended. However, it is extremely important (and this phase in often overlooked) that an assessment of the model be done to determine whether the OLS assumptions are met, otherwise the model is likely to be deficient or misspecified, and yield misleading results. In the last few decades, there has been much progress made on how to screen model residual behavior so as to provide diagnostics insight into model deficiency or misspecification.

A few idealized plots illustrate some basic patterns of improper residual behavior which are addressed in more detail in the later sections of this chapter. Figure 5.12 illustrates the effect of omitting an important dependence which suggests that an additional variable is to be introduced in the model

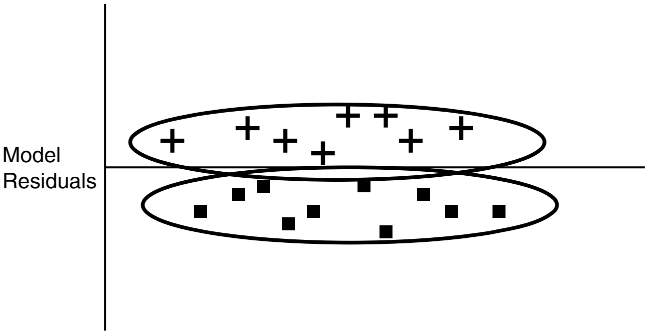


Fig. 5.12 The residuals can be separated into two distinct groups (shown as *crosses* and *dots*) which suggest that the response variable is related to another regressor not considered in the regression model. This residual pattern can be overcome by reformulating the model by including this additional variable. One example of such a time-based event system change is shown in Fig. 9.15 of Chap. 9.

[3] Herschel: "… almost all of the greatest discoveries in astronomy have resulted from the consideration of what … (was) termed residual phenomena".

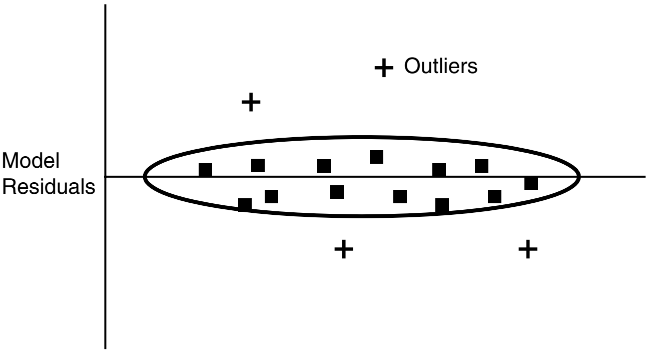


Fig. 5.13 Outliers indicated by crosses suggest that data should be checked and/or robust regression used instead of OLS

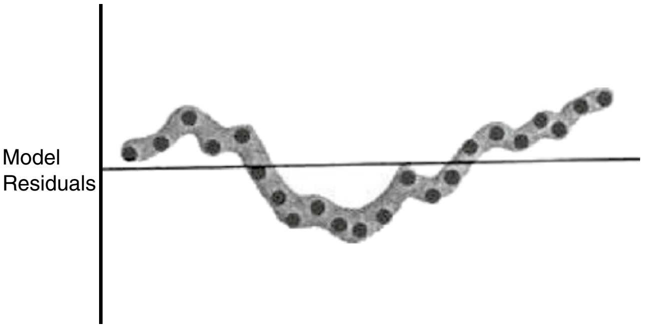


Fig. 5.16 Serial correlation is indicated by a pattern in the residuals when plotted in the sequence the data was collected, i.e., when plotted against time even though time may not be a regressor in the model

which distinguishes between the two groups. The presence of outliers and the need for more robust regression schemes which are immune to such outliers is illustrated in Fig. 5.13. The presence of non-constant variance (or heteroscedasticity) in the residuals is a very common violation and one of several possible manifestations is shown in Fig. 5.14. This particular residual behavior is likely to be remedied by using a log transform of the response variable instead of the variable itself. Another approach is to use weighted least squares estimation procedures described later in this chapter. Though non-constant variance is easy to detect visually, its cause is difficult to identify. Figure 5.15 illustrates a typical behavior which arises when a linear function is used to model a quadratic variation. The proper corrective action will increase the predictive accuracy of the model (RMSE will be lower), result in the estimated parameters being more efficient (i.e., lower standard errors), and most importantly, allow more sound and realistic interpretation of the model prediction uncertainty bounds.

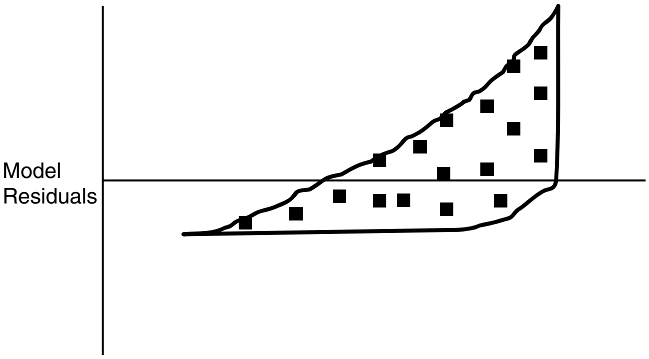


Fig. 5.14 Residuals with bow shape and increased variability (i.e., error increases as the response variable y increases) indicate that a log transformation of y is required

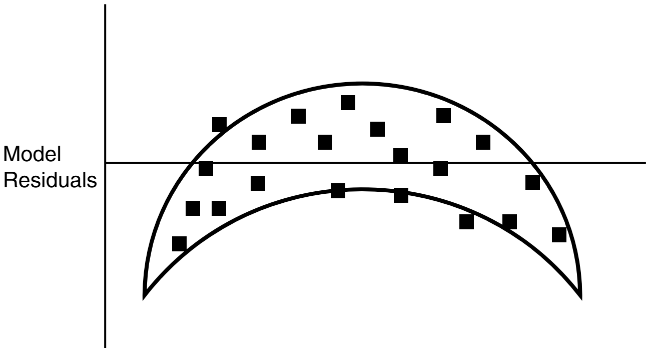


Fig. 5.15 Bow-shaped residuals suggest that a non-linear model, i.e. a model with a square term in the regressor variable to be evaluated

Figure 5.16 illustrates the occurrence of serial correlations in time series data which arises when the error terms are not independent. Such patterned residuals occur commonly during model development and provide useful insights into model deficiency. Serial correlation (or autocorrelation) has special pertinence to time series data (or data ordered in time) collected from in-situ performance of mechanical and thermal systems and equipment. Autocorrelation is present if adjacent model residuals, i.e., residuals show a trend or a pattern of clusters above or below the zero value that can be discerned visually. Such correlations can either suggest that additional variables have been left out of the model (model-misspecification error), or could be due to the nature of the process itself (called pure or "pseudo" autocorrelation). The latter is due to the fact that equipment loading over a day would follow an overall cyclic curve (as against random jumps from say full load to half load) consistent with the diurnal cycle and the way the system is operated. In such cases, positive residuals would tend to be followed by positive residuals, and vice versa. Time series data and models are treated further in Chap. 9.

Problems associated with *model underfitting and overfitting* are usually the result of a failure to identify the non-random pattern in time series data. Underfitting does not capture enough of the variation in the response variable which the corresponding set of regressor variables can possibly explain. For example, all four models fit to their respective sets of data as shown in Fig. 5.17, have identical R^2 values and t-statistics but are distinctly different in how they capture the data variation. Only plot (a) can be described by a linear model. The data in (b) needs to be fitted by a higher order

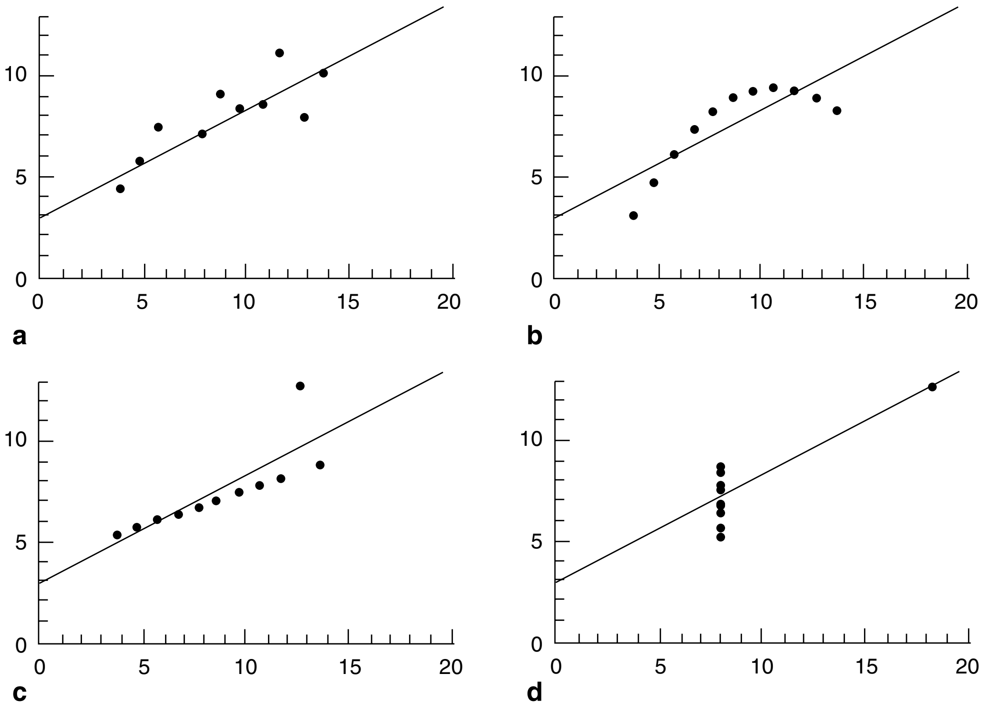


Fig. 5.17 Plot of the data (x, y) with the fitted lines for four data sets. The models have identical R^2 and t-statistics but only the first model is a realistic model. (From Chatterjee and Price 1991 by permission of John Wiley and Sons)

model, while one data point in (c) and (d) distorts the entire model. Blind model fitting (i.e., relying only on model statistics) is, thus, inadvisable.

Overfitting implies capturing randomness in the model, i.e., attempting to fit the noise in the data. A rather extreme example is when one attempts to fit a model with six parameters to six data points which have some inherent experimental error. The model has zero degrees of freedom and the set of six equations can be solved without error (i.e., RMSE = 0). This is clearly unphysical because the model parameters have also "explained" the random noise in the observations in a deterministic manner.

Both underfitting and overfitting can be detected by performing certain statistical tests on the residuals. The most commonly used test for white noise (i.e., uncorrelated residuals) involving model residuals is the **Durbin-Watson (DW)** statistic defined by:

$$\mathrm{DW} = \sum_{\mathrm{i}=2}^{\mathrm{n}} \left(\varepsilon_{\mathrm{i}} - \varepsilon_{\mathrm{i}-1}\right)^2 \Big/ \sum_{\mathrm{i}=1}^{n} \varepsilon_{\mathrm{i}}{}^2 \qquad (5.42)$$

where ε_{i} is the residual at time interval i, defined as $\varepsilon_{\mathrm{i}} = \mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{i}}$.

If there is no serial or autocorrelation present, the expected value of DW is 2. If the model underfits, DW would be less than 2 while it would be greater than 2 for an overfitted model, the limiting range being 0–4. Tables are available for approximate significance tests with different numbers of regressor variables and number of data points. Table A.13 assembles lower and upper critical values of DW statistics to test autocorrelation. For example, if n = 20, and the model has three variables (p = 3), the null hypothesis that the correlation coefficient is equal to zero can be rejected at the 0.05 significance level if its value is either below 1.00 or above 1.68. Note that the critical values in the table are one-sided, i.e., apply to one tailed distributions.

It is important to note that the DW statistic is only sensitive to correlated errors in adjacent observations, i.e., when only first-order autocorrelation is present. For example, if the time series has seasonal patterns, then higher autocorrelations may be present which the DW statistic will be unable to detect. More advanced concepts and modeling are discussed in Sect. 9.5 while treating stochastic time series data.

5.6.2 Leverage and Influence Data Points

Most of the aspects discussed above relate to identifying general patterns in the residuals of the entire data set. Another issue is the ability to identify subsets of data that have an unusual or disproportionate influence on the estimated model in terms of parameter estimation. Being able to flag such influential subsets of individual points allows one to investigate their validity, or to glean insights for better experimental design since they may contain the most interesting system behavioral information. Note that such points are not necessarily "bad" data points which should be omitted, but should be viewed as being "distinctive" observations in the overall data set. Scatter plots reveal such outliers easily for single regressor situations, but are inappropriate for multivariate cases. Hence, several statistical measures have been proposed to deal with multivariate situations, the influence and leverage indices being widely used (Belsley et al. 1980; Cook and Weisberg 1982).

The *leverage* of a point quantifies the extent to which that point is "isolated" in the x-space, i.e., its distinctiveness in terms of the regressor variables. It has a large impact on the numerical values of the model parameters being estimated. Consider the following matrix (called the hat matrix):

$$H = X(X'X)^{-1}X' \tag{5.43}$$

If one has a data set with two regressors, the order of the H matrix would be (3×3), i.e, equal to the number of parameters in the model (constant plus the two regressor coefficients). The diagonal element p_{ii} can be related to the distance between x_i and $\overline{x}$, and is defined as the *leverage* of the ith data point. Since the diagonal elements have values between 0 and 1, their average value is equal to (p/n) where n is the number of observation sets. Points with $p_{ii}>3$ (p/n) are regarded as points with high leverage (sometimes the threshold is taken as 2 (p/n).

Large residuals are traditionally used to highlight suspect data points or data points unduly affecting the regression model. Instead of looking at residuals ε_i, it is more meaningful to study a normalized or scaled value, namely the standardized residuals or R-student residuals, where

$$\text{R-student} = \frac{\varepsilon_i}{\text{RMSE} \cdot [1-p_{ii}]^{1/2}} \tag{5.44}$$

Points with |R-student|>3 can be said to be influence points which corresponds to a significance level of 0.01. Sometimes a less conservative value of 2 is used corresponding to the 0.05 significance level, with the underlying assumption that residuals or errors are Gaussian.

A data point is said to be *influential* if its deletion, singly or in combination with a relatively few others, cause statistically significant changes in the fitted model coefficients. See Sect. 3.5.3 for a discussion based on graphical considerations of this concept. There are several measures used to describe influence, a common one is DFITS:

$$\text{DFITS}_i = \frac{\varepsilon_i (p_{ii})^{1/2}}{s_i (1-p_{ii})^{1/2}} \tag{5.45}$$

where ε_i is the residual error of observation i, and s_i is the standard deviation of the residuals without considering the ith residual. Points with $DFITS \geq 2[p/(n-p)]^{1/2}$ are flagged as influential points.

Both the R-student statistic and the DFITS indices are often used to detect influence points. In summary, just because a point has high leverage does not make it influential. It is advisable to identify points with high leverage, and, then, examine them to determine whether they are influential as well.

Influential observations can impact the final regression model in different ways (Hair et al. 1998). For example, in Fig. 5.18a, the model residuals are not significant and the two influential observations shown as filled dots reinforce the general pattern in the model and lower the standard error of the parameters and of the model prediction. Thus, the two points would be considered to be leverage points which are beneficial to our model building. Influential points which adversely impact model building are illustrated in Fig. 5.18b and c. In the former, the two influential points almost totally account for the observed relationship but would not have been identified as outlier points. In Fig. 5.18c, the two influential points have totally altered the model identified, and the actual data points would have shown up as points with large residuals which the analyst would probably have identified as spurious.

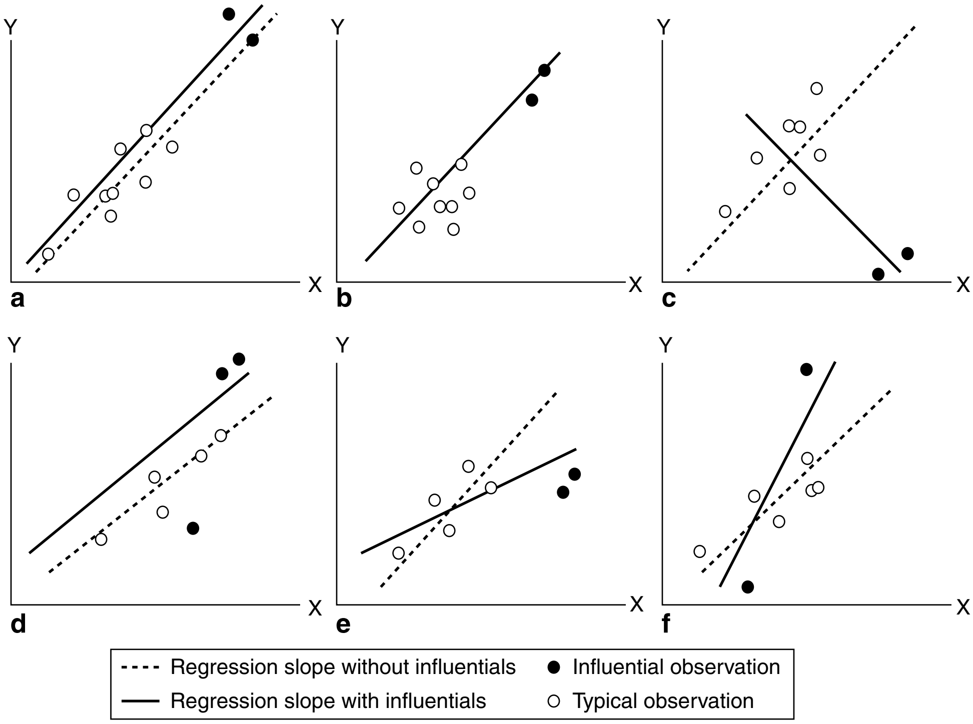


Fig. 5.18a–f Common patterns of influential observations. (From Hair et al. 1998 by © permission of Pearson Education)

The next frame (d) illustrates the instance when an influential point changes the intercept of the model but leaves the slope unaltered. The two final frames, Fig. 5.18e and f, illustrate two, hard to identify and rectify, cases when two influential points reinforce each other in altering both the slope and the intercept of the model though their relative positions are very much different. Note that data points that satisfy both these statistical criteria, i.e., are both influential and have high leverage, are the ones worthy of closer scrutiny. Most statistical programs have the ability to flag such points, and hence performing this analysis is fairly straightforward.

Thus, in conclusion, individual data points can be outliers, leverage or influential points. Outliers are relatively simple to detect and to interpret using the R-student statistic. Leverage of a point is a measure of how unusual the point lies in the x-space. An influence point is one which has an important affect on the regression model when that particular point were to be removed from the data set. Influential points are the ones which need particular attention since they provide insights about the robustness of the fit. In any case, all three measures (leverage p_{ii}, DFITS and R-student) provide indications as to the role played by different observations towards the overall model fit. Ultimately, the decision of deciding whether to retain or reject such points is somewhat based on judgment.

Example 5.6.1: *Example highlighting different characteristic of outliers or residuals versus influence points.*
Consider the following made-up data (Table 5.5) where x ranges from 1 to 10, and the model is $y=10+1.5*x$ to which random normal noise $\varepsilon=[0, \sigma=1]$ has been added to give y (second column). The last observation has been intentionally corrupted to a value of 50 as shown.

How well a linear model fits the data is depicted in Fig. 5.19. The table of unusual residuals shown below lists all observations which have Studentized residuals greater than 2.0 in absolute value. Note that observation 10 is flagged as an unusual residual (not surprising since this was intentionally corrupted) and no observation has been identified as influential despite it being very much of an outlier (the studentized value is very large—recall that a value of 3.0 would indicate a 99% CL). Thus, the error in one point seems to be overwhelmed by the well-behaved nature of the other nine points. This example serves to highlight the different characteristic of outliers versus influence points. ■

Row	x	y	Predicted		Studentized residual
			y	Residual	
10	10.0	50.0	*37.2572*	12.743	*11.43*

Table 5.5 Data table for Example 5.6.1

x	y[0,1]	y1
1	11.69977	11.69977
2	12.72232	12.72232
3	16.24426	16.24426
4	19.27647	19.27647
5	21.19835	21.19835
6	23.73313	23.73313
7	21.81641	21.81641
8	25.76582	25.76582
9	29.09502	29.09502
10	28.9133	*50*

5.6.3 Remedies for Non-uniform Model Residuals

Non-uniform model residuals or heteroscedasticity can be due to: (i) the nature of the process investigated, (ii) noise in the data, or (iii) the method of data collection from samples which are known to have different variances. Three possible generic remedies for non-constant variance are to (Chatterjee and Price 1991):

(a) **introduce additional variables into the model and collect new data:** The physics of the problem along with model residual behavior can shed light into whether certain key variables, left out in the original fit, need to be introduced or not. This aspect is further discussed in Sect. 5.6.5;

(b) **transform the dependent variable:** This is appropriate when the errors in measuring the dependent variable may follow a probability distribution whose variance is a function of the mean of the distribution. In such cases, the model residuals are likely to exhibit heterosce-

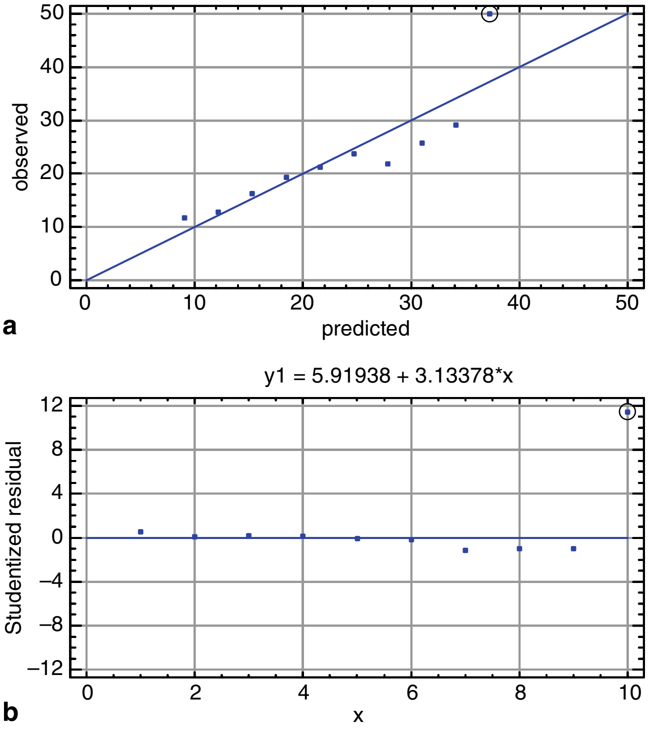


Fig. 5.19 **a** Observed vs predicted plot. **b** Residual plot versus regressor

Table 5.6. Transformations in dependent variable y likely to stabilize non-uniform model variance

	Variance of y in terms of its mean μ	Transformation
Poisson	μ	$y^{1/2}$
Binomial	$\mu(1-\mu)/n$	$\sin^{-1}(y)^{1/2}$

dasticity which can be removed by using exponential, Poisson or Binomial transformations. For example, a variable which is distributed Binomially with parameters "n and p" has mean (n.p.) and variance [n.p.(1−p)] (Sect. 2.4.2). For a Poisson variable, the mean and variance are equal. The transformations shown in Table 5.6 will stabilize variance, and the distribution of the transformed variable will be closer to the normal distribution.

The logarithmic transformation is also widely used in certain cases to transform a non-linear model into a linear one (see Sect. 9.5.1). When the variables have a large standard deviation compared to the mean, working with the data on a log scale often has the effect of dampening variability and reducing asymmetry. This is often an effective means of removing heteroscedascity as well. However, this approach is valid only when the magnitude of the residuals increase (or decrease) with that of one of the variables.

Example 5.6.2: *Example of variable transformation to remedy improper residual behavior*

The following example serves to illustrate the use of variable transformation. Table 5.7 shows data from 27 departments in a university with y as the number of faculty and staff and x the number of students.

A simple linear regression yields a model with R-squared=77.6% and a RMSE=21.7293. However, the residuals reveal an unacceptable behavior with a strong funnel behavior (see Fig. 5.20a).

Table 5.7 Data table for Example 5.6.2

	x	y		x	y
1	294	30	15	615	100
2	247	32	16	999	109
3	267	37	17	1,022	114
4	358	44	18	1,015	117
5	423	47	19	700	106
6	311	49	20	850	128
7	450	56	21	980	130
8	534	62	22	1,025	160
9	438	68	23	1,021	97
10	697	78	24	1,200	180
11	688	80	25	1,250	112
12	630	84	26	1,500	210
13	709	88	27	1,650	135
14	627	97			

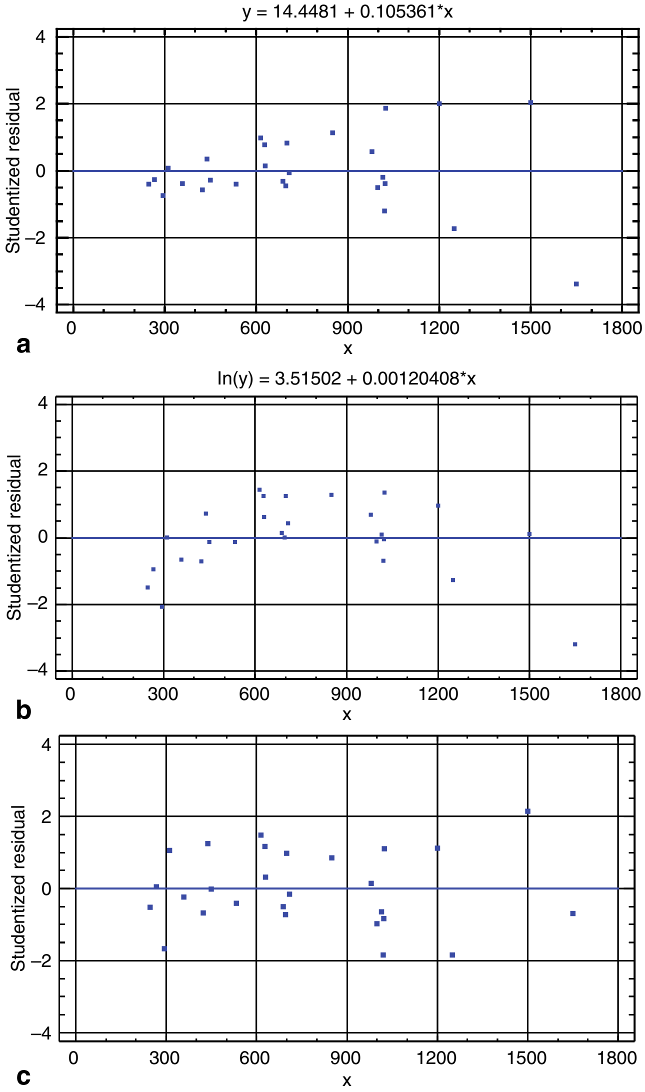


Fig. 5.20 **a** Residual plot of linear model. **b** Residual plot of log transformed linear model. **c** Residual plot of log transformed linear model

Instead of a linear model in y, a linear model in ln(y) is investigated. In this case, the model R-squared=76.1% and RMSE=0.252396. However, these statistics should NOT be compared directly since the y variable is no longer the same (in one case, it is "y"; in the other "ln y").

Let us not look into this aspect, but rather study the residual behavior. Notice that a linear model does reduce some of the improper residual variance but the inverted u shape behavior is indicative of model mis-specification (see Fig. 5.20b).

Finally, using a quadratic model along with the ln transformation results in a model:

$$\ln(y) = 2.8516 + 0.00311267{}^*x - 0.00000110226{}^*x^2$$

The residuals shown in Fig. 5.20c are now quite well behaved as a result of such a transformation. ■

(c) **perform weighted least squares.** This approach is more flexible and several variants exist (Chatterjee and Price 1991). As described earlier, OLS model residual behavior can exhibit non-uniform variance (called heteroscedasticity) even if the model is structurally complete, i.e., the model is not mis-specified. This violates one of the standard OLS assumptions. In a multiple regression model, detection of heteroscedasticity may not be very straight-forward since only one or two variables may be the culprits. Examination of the residuals versus each variable in turn along with intuition and understanding of the physical phenomenon being modeled can be of great help. Otherwise, the OLS estimates will lack precision, and the estimated standard errors of the model parameters will be wider. If this phenomenon occurs, the model identification should be redone with explicit recognition of this fact.

During OLS, the sum of the model residuals of all points are minimized with no regard to the values of the individual points or to points from different domains of the range of variability of the regressors. The basic concept of weighted least squares (WLS) is to simply assign different weights to different points according to a certain scheme. Thus, the general formulation of WLS is that the following function should be minimized:

$$\text{WLS function} = \sum \mathrm{w_i}\left(\mathrm{y_i} - \beta_0 - \beta_1 \mathrm{x_{1i}} \cdots - \beta_p \mathrm{x_{pi}}\right)^2 \tag{5.46}$$

where w_i are the weights of individual points. These are formulated differently depending on the weighting scheme selected which, in turn, depends on prior knowledge about the process generating the data.

(c-i) Errors Are Proportional to x Resulting in Funnel-Shaped Residuals Consider the simple model $y=\alpha+\beta x+\varepsilon$ whose residuals ε have a standard deviation which increases as the regressor variable (resulting in the funnel-like shape in Fig. 5.21). Assuming a weighting scheme such as $\mathrm{var}(\varepsilon_i) = k^2 x_i^2$, transforms the model into:

$$\frac{\mathrm{y}}{\mathrm{x}} = \frac{\alpha}{\mathrm{x}} + \beta + \frac{\varepsilon}{x} \quad \text{or } \mathrm{y}' = \alpha \mathrm{x}' + \beta + \varepsilon' \tag{5.47}$$

Note that the variance of ε' is constant and equals k^2. If the assumption about the weighting scheme is correct, the transformed model will be homoscedastic, and the model parameters α and β will be efficiently estimated by OLS (i.e., the standard errors of the estimates will be optimal).

The above transformation is only valid when the model residuals behave as shown in Fig. 5.21. If residuals behave differently, then different transformations or weighting schemes will have to be explored. Whether a particular transformation is adequate or not can only be gauged by the behavior of the variance of the residuals. Note that the analyst has to perform two separate regressions: one an OLS regression in order to determine the residual amounts of the individual data points, and then a WLS regression for final parameter identification. This is often referred to as *two-stage estimation.*

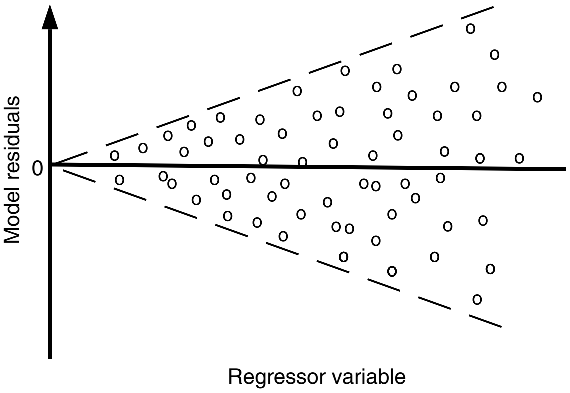


Fig. 5.21 Type of heteroscedastic model residual behavior which arises when errors are proportional to the magnitude of the x variable

(c-ii) Replicated Measurements with Different Variance It could happen, especially with models involving one regressor variable only and when the data is obtained in the framework of a designed experimental study (as against observational or non-experimental data), that one obtains replicated measurements on the response variable corresponding to a set of fixed values of the explanatory variables. For example, consider the case when the regressor variable x takes several discrete values. If the physics of the phenomenon cannot provide any theoretical basis on how to select a particular weighty scheme, then this has to be determined experimentally from studying the data. If there is an increasing pattern in the heteroscedascity present in the data, this could be modeled either by a logarithmic transform (as illustrated in Example 5.6.2) or a suitable variable transformation. Here, another more versatile approach which can be applied to any pattern of the residuals is illustrated. Each observed residual ε_{ij} (where the index for discrete x values is i, and the number of observations at each discrete x value is j=1, 2, … n_i) is made up of two parts, i.e., $\varepsilon_{ij} = (\mathrm{y_{ij}} - \bar{\mathrm{y}}_\mathrm{i}) + (\bar{\mathrm{y}}_\mathrm{i} - \hat{\mathrm{y}}_\mathrm{ij})$. The first part is referred to as pure error while the second part measures lack of fit. An assessment of heteroscedasticity is based on pure error. Thus, the WLS weight may be estimated as $\mathrm{w_i} = 1/s_i^2$ where the mean square error is:

$$\mathrm{s_i^2} = \frac{\sum (\mathrm{y_{ij}} - \bar{\mathrm{y}}_\mathrm{i})^2}{(n_i - 1)} \tag{5.48}$$

Alternatively a model can be fit to the mean values of x and the s_i^2 values in order to smoothen out the weighting function, and this function used instead. Thus, this approach would also qualify as a two-stage estimation process. The following example illustrates this approach.

Table 5.8 Measured data, OLS residuals deduced from Eq. 5.49a and the weights calculated from Eq. 5.49b

x	y	Residual ε_i	w_i
1.15	0.99	0.26329	0.9882
1.90	0.98	−0.59826	1.7083
3.00	2.60	−0.2272	6.1489
3.00	2.67	−0.1572	6.1489
3.00	2.66	−0.1672	6.1489
3.00	2.78	−0.0472	6.1489
3.00	2.80	−0.0272	6.1489
5.34	5.92	0.435964	15.2439
5.38	5.35	−0.17945	13.6185
5.40	4.33	−1.22216	12.9092
5.40	4.89	-0.66216	12.9092
5.45	5.21	−0.39893	11.3767
7.70	7.68	−0.48358	0.9318
7.80	9.81	1.53288	0.8768
7.81	6.52	−1.76847	0.8716
7.85	9.71	1.37611	0.8512
7.87	9.82	1.463402	0.8413
7.91	9.81	1.407986	0.8219
7.94	8.50	0.063924	0.8078
9.03	9.47	−0.20366	0.4694
9.07	11.45	1.730922	0.4614
9.11	12.14	2.375506	0.4535
9.14	11.50	1.701444	0.4477
9.16	10.65	0.828736	0.4440
9.37	10.64	0.580302	0.4070
10.17	9.78	−1.18802	0.3015
10.18	12.39	1.410628	0.3004
10.22	11.03	0.005212	0.2963
10.22	8.00	−3.02479	0.2963
10.22	11.90	0.875212	0.2963
10.18	8.68	−2.29937	0.3004
10.50	7.25	−4.0927	0.2696
10.23	13.46	2.423858	0.2953
10.03	10.19	−0.61906	0.3167
10.23	9.93	−1.10614	0.2953

Example 5.6.3:[4] *Example of weighted regression for replicate measurements*

Consider the data given in Table 5.8 of replicate measurements of y taken at different values of x (which vary slightly).

A scatter plot of this data and the simple OLS linear model are shown in Fig. 5.22a. The regressed model is:

$$y = -0.578954 + 1.1354^{*}x \quad \text{with} \quad R^2 = 0.841 \quad \text{and} \quad \text{RMSE} = 1.4566 \tag{5.49a}$$

Note that the intercept term in the model is not statistically significant (p-value = 0.4 for the t-statistic), while the overall model fit given by the F-ratio is significant. The model residuals of a simple OLS fit are shown in Fig. 5.22b.

Coefficients

Parameter	Least squares estimate	Standard error	t-statistic	P-value
Intercept	−0.578954	0.679186	−0.852423	0.4001
Slope	1.1354	0.086218	13.169	0.0000

Analysis of Variance

Source	Sum of squares	Df	Mean square	F-ratio	P-value
Model	367.948	1	367.948	173.42	0.0000
Residual	70.0157	33	2.12169		
Total (Corr.)	437.964	34			

The residuals of a simple linear OLS model shown in Fig. 5.22b reveal, as expected, marked heteroscadascity. Hence, the OLS model is bound to lead to misleading uncertainty bands even if the model predictions themselves are not biased. The model residuals from the above model are also shown in the table. Subsequently, the mean and the mean square error s_i^2 are calculated following Eq. 5.48 to yield the following table:

$\hat{x}$	s_i^2
3	0.0072
5.39	0.373
7.84	1.6482
9.15	0.8802
10.22	4.1152

Then, because of the pattern exhibited, a second order polynomial OLS model is regressed to this data (see Fig. 5.22c):

$$s_i^2 = 1.887 - 0.8727.\bar{x} + 0.9967.\bar{x}^2 \text{ with } R^2 = 0.743 \tag{5.49b}$$

The regression weights w_i can thus be deduced by using individual values of x_i instead of $\bar{x}$ in the above equation. The values of the weights are also shown in the data table. Finally, a weighted regression is performed following Eq. 5.46 (most statistical packages have this capability) resulting in:

$$y = -0.942228 + 1.16252^{*}x \quad \text{with} \quad R^2 = 0.896 \quad \text{and} \quad \text{RMSE} = 1.2725. \tag{5.49c}$$

[4] From Draper and Smith (1981) by permission of John Wiley and Sons.

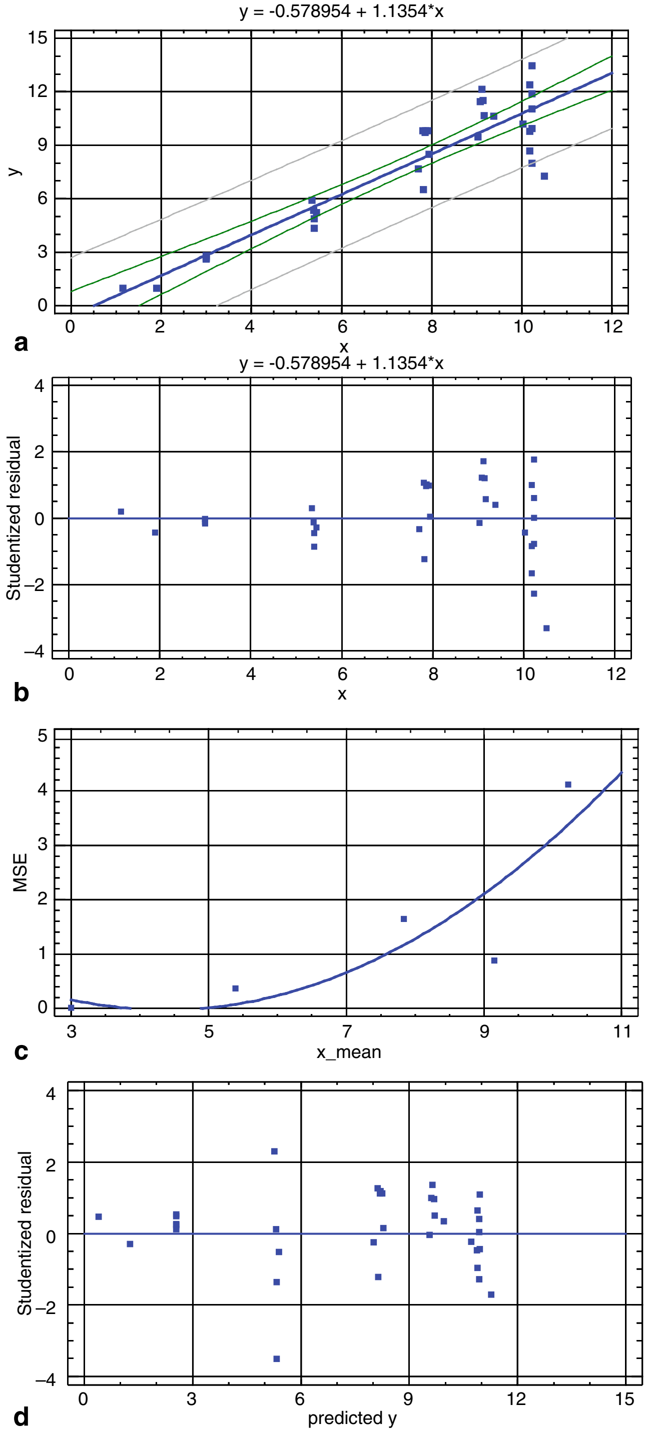


Fig. 5.22 **a** Data set and OLS regression line of observations with non-constant variance and replicated observations in x. **b** Residuals of a simple linear OLS model fit (Eq. 5.49a). **c** Residuals of a second order polynomial OLS fit to the mean x and mean square error (MSE) of the replicate values (Eq. 5.49b). **d** Residuals of the weighted regression model (Eq. 5.49c)

The residual plots are shown as Fig. 5.22d. Though the goodness of fit is only slightly better than the OLS model, the real advantage is that this model will have better prediction accuracy and realistic prediction errors. ■

(c-iii) Non-patterned Variance in the Residuals A third type of non-constant residual variance is one when no pattern is discerned with respect to the regressors which can be discrete or vary continuously. In this case, a practical approach is to look at a plot of the model residuals against the response variable, divide the range in the response variable into as many regions as seem to have different variances, and calculate the standard deviation of the residuals for each of these regions. In that sense, the general approach parallels the one adopted in case (c-ii) when dealing with replicated values with non-constant variance; however, now, no model such as 5.49b is needed. The general approach would involve the following steps:

- First, fit an OLS model to the data;
- Next, discretize the domain of the regressor variables into a finite number of groups and determine ε_i^2 from which the weights w_i for each of these groups can be deduced;
- Finally, perform a WLS regression in order to estimate the efficient model parameters.

Though this two-stage estimation approach is conceptually easy and appealing for simple models, it may become rather complex for multivariate models, and moreover, there is no guarantee that heteroscedasticity will be removed entirely.

5.6.4 Serially Correlated Residuals

Another manifestation of improper residual behavior is serial correlation (discussed in Sect. 5.6.1). As stated earlier, one should distinguish between the two different types of autocorrelation, namely pure autocorrelation and model-misspecification, though often it is difficult to discern between them. The latter is usually addressed using the weight matrix approach (Pindyck and Rubinfeld 1981) which is fairly formal and general, but somewhat demanding. Pure autocorrelation relates to the case of "pseudo" patterned residual behavior which arises because the regressor variables have strong serial correlation. This serial correlation behavior is subsequently transferred over to the model, and thence to its residuals, even when the regression model functional form is close to "perfect". The remedial approach to be adopted is to transform the original data set prior to regression itself. There are several techniques of doing so, and the widely-used Cochrane-Orcutt (CO) procedure is described. It involves the use of generalized differencing to alter the linear model into one in which the errors are independent. The *two stage first-order CO procedure* involves:

(i) fitting an OLS model to the original variables;
(ii) computing the first-order serial correlation coefficient ρ of the model residuals;
(iii) transforming the original variables y and x into a new set of pseudo-variables:

$$y_t{}^* = y_t - \rho \cdot y_{t-1} \quad \text{and} \quad x_t{}^* = x_t - \rho \cdot x_{t-1} \tag{5.50}$$

(iv) OLS regressing of the pseudo variables y* and x* to re-estimate the parameters of the model;

(v) Finally, obtaining the fitted regression model in the original variables by a back transformation of the pseudo regression coefficients:

$$b_0 = b_0{}^*/(1-\rho) \quad \text{and} \quad b_1 = b_1{}^* \tag{5.51}$$

Though two estimation steps are involved, the entire process is simple to implement. This approach, when originally proposed, advocated that this process be continued till the residuals become random (say, based on the Durbin-Watson test). However, the current recommendation is that alternative estimation methods should be attempted if one iteration proves inadequate. This approach can be used during parameter estimation of MLR models provided only one of the regressor variables is the cause of the pseudo-correlation. Also, a more sophisticated version of the CO method has been suggested by Hildreth and Lu (Chatterjee and Price 1991) involving only one estimation process where the optimal value of ρ is determined along with the parameters. This, however, requires non-linear estimation methods.

Example 5.6.4: *Using the Cochrane-Orcutt procedure to remove first-order autocorrelation*

Consider the case when observed pre-retrofit data of energy consumption in a commercial building support a linear regression model as follows:

$$E_i = a_o + a_1 T_i \tag{5.52}$$

where

T = daily average outdoor dry-bulb temperature,

E_i = daily total energy use predicted by the model,

i = subscript representing a particular day, and,

a_o and a_1 are the least-square regression coefficients

How the above transformation yields a regression model different from OLS estimation is illustrated in Fig. 5.23 with year-long daily cooling energy use from a large institutional building in central Texas. The first-order autocorrelation coefficients of cooling energy and average daily temperature were both equal to 0.92, while that of the OLS residuals was 0.60. The Durbin-Watson statistic for the OLS residuals (i.e. untransformed data) was DW=3 indicating strong residual autocorrelation, while that of the CO transform was 1.89 indicating little or no autocorrelation. Note that the CO transform is inadequate in cases of model mis-specification and/or seasonal operational changes. ■

5.6.5 Dealing with Misspecified Models

An important source of error during model identification is *model misspecification* error. This is unrelated to measurement error, and arises when the functional form of the model is not appropriate. This can occur due to:

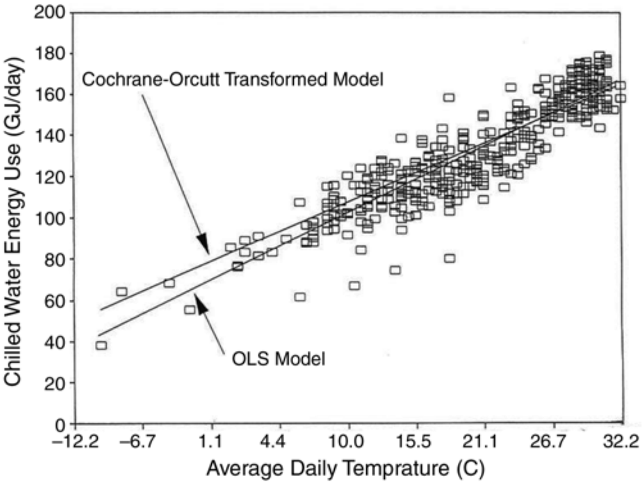


Fig. 5.23 How serial correlation in the residuals affects model identification (Example 5.6.4)

(i) *inclusion of irrelevant variables:* This does not bias the estimation of the intercept and slope parameters, but generally reduces the efficiency of the slope parameters, i.e., their standard errors will be larger. This source of error can be eliminated by, say, step-wise regression or simple tests such as t-tests;

(ii) *exclusion of an important variable:* This case will result in the slope parameters being both biased and inconsistent.

(iii) *assumption of a linear model:* This arises when a linear model is erroneously assumed, and

(iv) *incorrect model order:* This corresponds to the case when one assumes a lower or higher model than what the data warrants.

The latter three sources of errors are very likely to manifest themselves in improper residual behavior (the residuals will show sequential or non-constant variance behavior). The residual analysis may not identify the exact cause, and several attempts at model reformulations may be required to overcome this problem. Even if the physics of the phenomenon or of the system is well understood and can be cast in mathematical terms, experimental or identifiability constraints may require that a simplified or macroscopic model be used for parameter identification rather than the detailed model. This could cause model misspecification, especially so if the model is poor.

Example 5.6.5: *Example to illustrate how inclusion of additional regressors can remedy improper model residual behavior*

Energy use in commercial buildings accounts for about 18% of the total energy use in the United States and consequently, it is a prime area of energy conservation efforts. For this purpose, the development of baseline models, i.e., models of energy use for a specific end-use before energy conservation measures are implemented, is an important modeling activity for monitoring and verification studies.

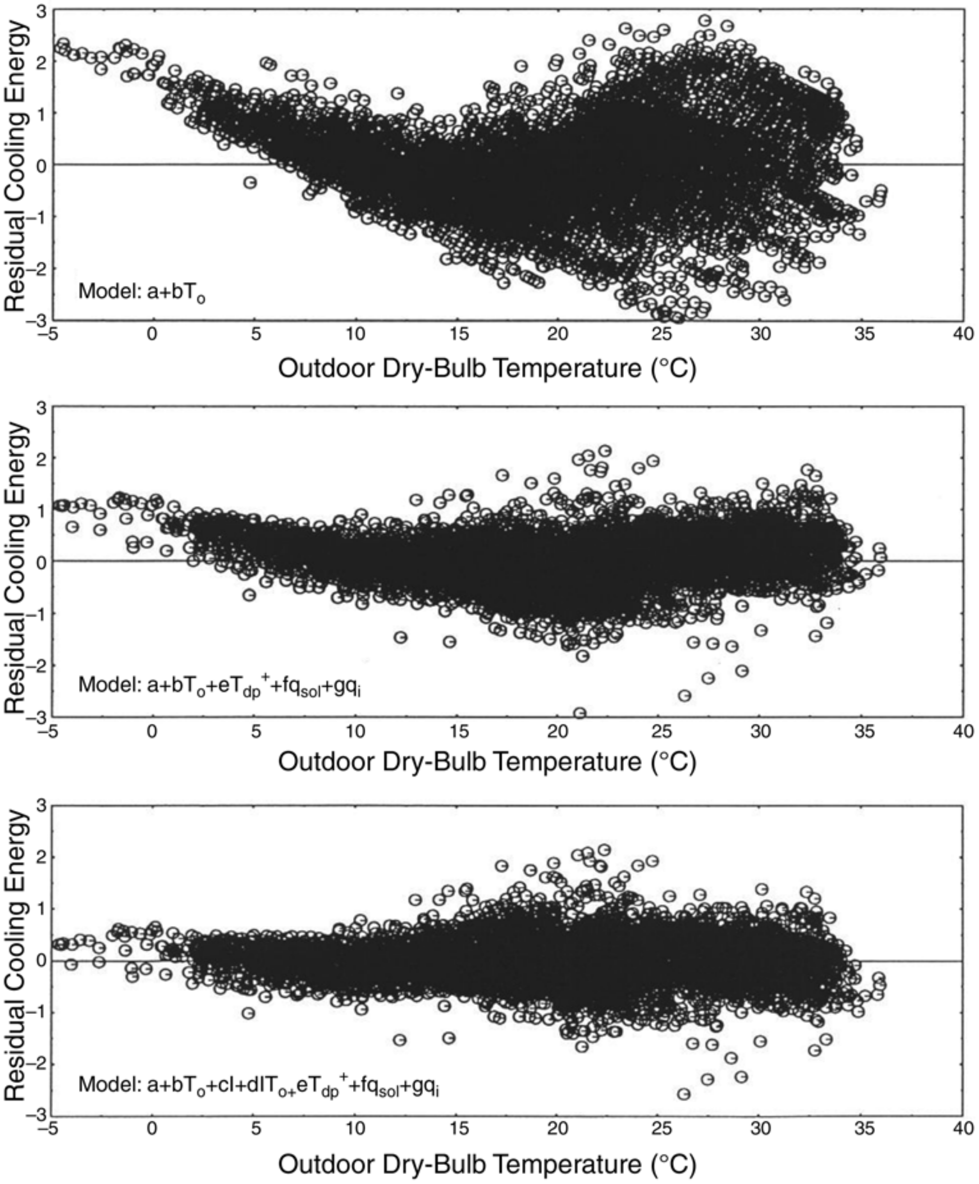


Fig. 5.24 Improvement in residual behavior for a model of hourly energy use of a variable air volume HVAC system in a commercial building as influential regressors are incrementally added to the model. (From Katipamula et al. 1998)

Let us illustrate the effect of improper selection of regressor variables or model misspecification for modeling measured thermal cooling energy use of a large commercial building operating 24 hours a day under a variable air volume HVAC system (Katipamula et al. 1998). Figure 5.24 illustrates the residual pattern when hourly energy use is modeled with only the outdoor dry-bulb temperature (T_o). The residual pattern is blatantly poor exhibiting both non-constant variance as well as systematic bias in the low range of the x-variable. Once the outdoor dew point temperature (T_{dp}^{+})[5], the global horizontal solar radiation (q_{sol}) and the internal building heat loads q_i (such as lights and equipment) are introduced in the model, the residual behavior improves significantly but the lower tail is still present. Finally, when additional terms involving indicator variables I to both intercept (T_o) are introduced, (described in Sect. 5.7.2), an acceptable residual behavior is achieved. ■

5.7 Other OLS Parameter Estimation Methods

5.7.1 Zero-Intercept Models

Sometimes the physics of the system dictates that the regression line pass through the origin. For the linear case, the model assumes the form:

$$y = \beta_1 x + \varepsilon \tag{5.53}$$

[5] Actually the outdoor humidity impacts energy use only when the dew point temperature exceeds a certain threshold which many studies have identified to be about 55°F (this is related to how the HVAC is controlled in response to human comfort). This type of conditional variable is indicated as a + superscript.

The interpretation of R^2 under such a case is not the same as for the model with an intercept, and this statistic cannot be used to compare the two types of models directly. Recall that the R^2 value designated the percentage variation of the response variable *about its mean* explained by that of the regressor variable. For the no-intercept case, the R^2 value explains the percentage variation of the response variable *about the origin* explained by that of the regressor variable. Thus, when comparing both models, one should decide on which is the better model based on their RMSE values.

5.7.2 Indicator Variables for Local Piecewise Models—Spline Fits

Spline functions are an important class of functions, described in numerical analysis textbooks in the framework of interpolation, which allow distinct functions to be used over different ranges while maintaining continuity in the function. They are extremely flexible functions in that they allow a wide range of locally different behavior to be captured within one elegant functional framework. Thus, a globally nonlinear function can be decomposed into simpler local patterns. Two cases arise.

(a) The simpler case is one where it is known which points lie on which trend, i.e., when the physics of the system is such that the location of the structural break or "hinge point" x_c of the regressor is known. The simplest type is the piece-wise *linear* spline (as shown in Fig. 5.25), with higher order polynomial splines up to the third degree being also used often to capture non-linear trends. The objective here is to formulate a linear model and identify its parameters which best describe data points in Fig. 5.25. One cannot simply divide the data into two, and fit each region with a separate linear model since the constraint that the model be continuous at the hinge point would be violated. A model of the following form would be acceptable:

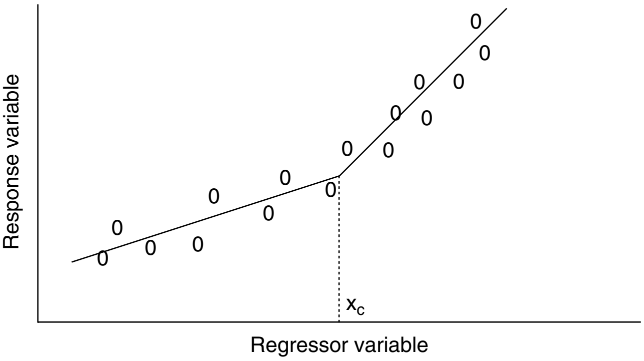


Fig. 5.25 Piece-wise linear model or first-order spline fit with hinge point at x_c. Such models are referred to as *change point models* in building energy modeling terminology

$$y = \beta_0 + \beta_1 x + \beta_2 (x - x_c) I \qquad (5.54a)$$

where the indicator variable

$$I = \begin{cases} 1 & \text{if } x > x_c \\ 0 & \text{otherwise} \end{cases} \qquad (5.54b)$$

Hence, for the region $x \le x_c$, , the model is:

$$y = \beta_0 + \beta_1 x \qquad (5.55)$$

and for the region $x > x_c$ $y = (\beta_0 - \beta_2 x_c) + (\beta_1 + \beta_2)x$. Thus, the slope of the model is β_1 before the break and $(\beta_1 + \beta_2)$ afterwards. The intercept term changes as well from β_0 before the break to $(\beta_0 - \beta_2 x_c)$ after the break. The logical extensions to linear spline models with two structural breaks or to higher order splines involving quadratic and cubic terms are fairly straightforward.

(b) The second case arises when the change point is not known. A simple approach is to look at the data, identify a "ball-park" range for the change point, perform numerous regression fits with the data set divided according to each possible value of the change point in this ball-park range, and pick that value which yields the best overall R-square or RMSE. Alternatively, the more accurate but more complex approach is to cast the problem as a nonlinear estimation method with the change point variable as one of the parameters.

Example 5.7.1: *Change point models for building utility bill analysis*

The theoretical basis of modeling monthly energy use in buildings is discussed in several papers (for example, Reddy et al. 1997). The interest in this particular time scale is obvious—such information is easily obtained from utility bills which are usually on a monthly time scale. The models suitable for this application are similar to linear spline models, and are referred to as **change point models** by building energy analysts. A simple example is shown below to illustrate the above equations. Electricity utility bills of a residence in Houston, TX have been normalized by the number of days in the month and assembled in Table 5.9 along with the corresponding month and monthly mean outdoor temperature values for Houston (the first three columns of the table). The intent is to use Eq. 5.54 to model this behavior.

The scatter plot and the trend lines drawn in Fig. 5.26 suggest that the change point is in the range 17–19°C. Let us perform the calculation assuming a value of 17°C. Defining an indicator variable:

$$I = \begin{cases} 1 & \text{if } x > 17^\circ\text{C} \\ 0 & \text{otherwise} \end{cases}$$

Table 5.9 Measured monthly energy use data and calculation step for deducing the change point independent variable assuming a base value of 17°C

Month	Mean outdoor temperature (°C)	Monthly mean daily electric use (kWh/m^2/day)	x (°C)	$(x-17°C)I$ (°C)
Jan	11	0.1669	11	0
Feb	13	0.1866	13	0
Mar	16	0.1988	16	0
Apr	21	0.2575	21	4
May	24	0.3152	24	7
Jun	27	0.3518	27	10
Jul	29	0.3898	29	12
Aug	29	0.3872	29	12
Sept	26	0.3315	26	9
Oct	22	0.2789	22	5
Nov	16	0.2051	16	0
Dec	13	0.1790	13	0

Based on this assumption, the last two columns of the table have been generated to correspond to the two regressor variables in Eq. 5.54. A linear multiple regression yields:

$$y = 0.1046 + 0.005904x + 0.00905(x - 17)I$$
$$\text{with} \quad R^2 = 0.996 \quad \text{and} \quad \text{RMSE} = 0.0055$$

with all three parameters being significant. The reader can repeat this analysis assuming a different value for the change point (say $x_c = 18°C$) in order to study the sensitivity of the model to the choice of the change point value. Though only three parameters are determined by regression, this is an example of a four parameter (or 4-P) model in building science terminology. The fourth parameter is the change point x_c which also needs to be determined. Software programs have been developed to determine the optimal value of x_c (i.e., that which results in minimum RMSE of different possible choices of x_c) following a numerical search process akin to the one described in this example. ■

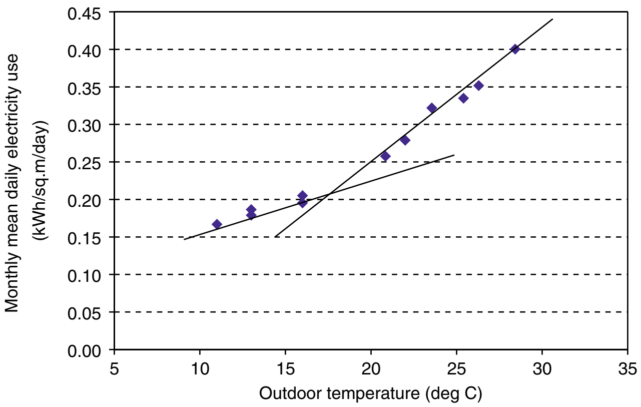


Fig. 5.26 Piece-wise linear regression lines for building electric use with outdoor temperature. The change point is the point of intersection of the two lines. The combined model is called a change point model, which, in this case, is a four parameter model given by Eq. 5.54

5.7.3 Indicator Variables for Categorical Regressor Models

The use of indicator (also called dummy) variables has been illustrated in the previous section when dealing with spline models. They are also used in cases when shifts in either the intercept or the slope are to be modeled with the *condition of continuity now being relaxed.* The majority of variables encountered in mechanistic models are quantitative, i.e., the variables are measured on a numerical scale. Some examples are temperature, pressure, distance, energy use and age. Occasionally, the analyst comes across models involving qualitative variables, i.e., regressor data that belong in one of two (or more) possible categories. One would like to evaluate whether differences in intercept and slope between categories are significant enough to warrant two separate models or not. This concept is illustrated by the following example.

Whether the annual energy use of a regular commercial buildings is markedly higher than that of another certified as being energy efficient is to be determined. Data from several buildings which fall in each group is gathered to ascertain whether the presumption is supported by the actual data. Factors which affect the *normalized energy use* (variable y) of both experimental groups are conditioned floor area (variable x_1) and outdoor temperature (variable x_2). Suppose that a linear relationship can be assumed with the same intercept for both groups. One approach would be to separate the data into two groups: one for regular buildings and one for efficient buildings, and develop regression models for each group separately. Subsequently, one could perform a t-test to determine whether the slope terms of the two models are significantly different or not. However, the assumption of constant intercept term for both models may be erroneous, and this may confound the analysis. A better approach is to use the entire data and adopt a modeling approach involving indicator variables.

Let model 1 be for regular buildings:

$$y = a + b_1 x_1 + c_1 x_2$$

and, model 2 be for energy efficient buildings:

$$y = a + b_2\, x_1 + c_2\, x_2 \tag{5.56}$$

The complete model (or model 3) would be formulated as:

$$y = a + b_1\, x_1 + c_1\, x_2 + b_2(I \cdot x_1) + c_2(I \cdot x_2) \tag{5.57}$$

where I is an indicator variable such that

$$I = \begin{cases} 1 & \text{for energy efficient buildings} \\ 0 & \text{for regular buildings} \end{cases}$$

Note that a basic assumption in formulating this model is that the intercept is unaffected by the building group. Formally, one would like to test the null hypothesis H: $b_2=c_2=0$. The hypothesis is tested by constructing an F statistic for the comparison of the two models. Note that model 3 is referred to as the *full model* (FM) or as the pooled model. Model 1, when the null hypothesis holds, is the *reduced model* (RM). The idea is to compare the goodness-of-fit of the FM and that of the RM. If the RM provides as good a fit as the FM, then the null hypothesis is valid. Let SSE(FM) and SSE(RM) be the corresponding model sum of square error or squared model residuals. Then, the following F-test statistic is defined:

$$F = \frac{[\mathrm{SSE(RM)} - \mathrm{SSE(FM)}]/(\mathrm{k\text{-}m})}{\mathrm{SSE(FM)}/(\mathrm{n\text{-}k})} \tag{5.58}$$

where n is the number of data sets, k is the number of parameters of the FM, and m the number of parameters of the RM. If the observed F value is larger than the tabulated value of F with (n-k) and (k-m) degrees of freedom at the pre-specified significance level (provided by Table A.6), the RM is unsatisfactory and the full model has to be retained. As a cautionary note, this test is strictly valid only if the OLS assumptions for the model residuals hold.

Example 5.7.2: *Combined modeling of energy use in regular and energy efficient buildings*

Consider the data assembled in Table 5.10. Let us designate the regular buildings by group (A) and the energy efficient buildings by group (B), with the problem simplified by assuming both types of buildings to be located in the same geographic location. Hence, the model has *only one regressor variable* involving floor area. The complete model with the indicator variable term given by Eq. 5.57 is used to verify whether group B buildings consume less energy than group A buildings.

The full model (FM) given by Eq. 5.57 reduces to the following form since only one regressor is involved:

Table 5.10 Data table for Example 5.7.2

Energy use (y)	Floor area (x_1)	Bldg type	Energy use (y)	Floor area (x_1)	Bldg type
45.44	225	A	32.13	224	B
42.03	200	A	35.47	251	B
50.1	250	A	33.49	232	B
48.75	245	A	32.29	216	B
47.92	235	A	33.5	224	B
47.79	237	A	31.23	212	B
52.26	265	A	37.52	248	B
50.52	259	A	37.13	260	B
45.58	221	A	34.7	243	B
44.78	218	A	33.92	238	B

$y=a+b_1x_1+b_2I\cdot x_2$ where the variable I is an indicator variable such that it is 0 for group A and 1 for group B. The null hypothesis is that H_0: $b_2=0$. The reduced model (RM) is $y=a+b_{.}x_1$.

The estimated model is $y=14.2762+0.14115\ x_1-13.2802\ (I\cdot x_2)$. The analysis of variance shows that the SSR(FM) = 7.7943 and SSR(RM) = 889.245. The F statistic in this case is:

$$F = \frac{(889.245 - 7.7943)/1}{7.7943/(20-3)} = 1922.5$$

One can thus safely reject the null hypothesis, and state with confidence that buildings built as energy-efficient ones consume energy which is statistically lower than those which are not.

It is also possible to extend the analysis and test whether both slope and intercept are affected by the type of building. The FM in this case is $y=a_1+b_1\ x_1+c(I)+d(I\cdot x_1)$ where I is an indicator variable which is, say 0 for Building A and 1 for Building B. The null hypothesis in this case is that $c=d=0$. This is left for the interested reader to solve. ■

5.7.4 Assuring Model Parsimony—Stepwise Regression

Perhaps the major problem with multivariate regression is that the "independent" variables are not really independent but collinear to some extent (how to deal with collinear regressors by transformation is discussed in Sect. 9.3). In multivariate regression, a thumb rule is that the number of variables should be less than four times the number of observations (Chatfield 1995). Hence, with n = 12, the number of variables should be at most 3 or less. Moreover, some authors go so far as stating that multivariate regression models with more than 4–5 variables are suspect. There is, thus, a big benefit in identifying models that are parsimonious. The more straightforward approach is to use the simpler (but formal) methods to identify/construct the "best" model linear in the parameters if the comprehensive set of all feasible/possible regressors of the model is known (Draper and Smith 1981; Chatterjee and Price 1991):

(a) *All possible regression models*: This method involves: (i) constructing models of different basic forms (single variate with various degrees of polynomials and multi-variate), (ii) estimating parameters that correspond to all possible predictor variable combinations, and (iii) then selecting one considered most desirable based on some criterion. While this approach is thorough, the computational effort involved may be significant. For example, with p possible parameters, the number of model combinations would be p^2. However, this may be moot if the statistical analysis program being used contains such a capability. The only real drawback is that blind curve fitting may suggest a model with no physical justification which in certain applications may have undesirable consequences. Further, it is advised that the cross-validation scheme should be used to avoid overfitting (see Sect. 5.3.2-d).

In any case, one needs a statistical criterion to determine, if not the "best[6]" model, then, at least a subset of desirable models from which one can be chosen based on the physics of the problem. One could use the adjusted R-square given by Eq. 5.7b which includes the number of model parameters. Another criterion for model selection is the *Mallows C_p statistic* which gives a normalized estimate of the total expected estimation error for all observations in the data set and takes account of both bias and variance:

$$C_p = \frac{SSE}{\sigma^2} + (2p - n) \tag{5.59}$$

where SSE is the sum of square errors (see Eq. 5.2), σ^2 is the variance of the residuals with the full set of variables, n is the number of data points, and p is the number of parameters in the specific model. It can be shown that the expected value of C_p is p when there is no bias in the fitted equation containing p terms. Thus "good" or desirable model possibilities are those whose C_p values are close to the corresponding number of parameters of the model.

Another automatic selection approach to handling models with large number of possible parameters is the *iterative approach* which comes in three variants.

(b-1) Backward Elimination Method: One begins with selecting an initial model that includes the full set of possible predictor variables from the candidate pool, and then successively dropping one variable at a time on the basis of their contribution to the reduction of SSE. The OLS method is used to estimate all model parameters along with t-values for each model parameter. If all model parameters are statistically significant, the model building process stops. If some model parameters are not significant, the model parameter of least significance (lowest t-value) is omitted from the regression equation, and the reduced model is refit. This process continues until all parameters that remain in the model are statistically significant.

(b-2) Forward Selection Method: One begins with an equation containing no regressors (i.e., a constant model). The model is then augmented by including the regressor variable with the highest simple correlation with the response variable. If this regression coefficient is significantly different from zero, it is retained, and the search for a second variable is made. This process of adding regressors one-by-one is terminated when the last variable entering the equation has an insignificant regression coefficient or when all the variables are included in the model. Clearly, this approach involves fitting many more models than in the backward elimination method.

(b-3) Stepwise Regression Method: This is one of the more powerful model building approaches and combines both the above procedures. Stepwise regression begins by computing correlation coefficients between the response and each predictor variable. The variable most highly correlated with the response is then allowed to "enter the regression equation". The parameter for the single-variable regression equation is then estimated along with a measure of the goodness of fit. The next most highly correlated predictor variable is identified, given the current variable already in the regression equation. This variable is then allowed to enter the equation and the parameters re-estimated along with the goodness of fit. Following each parameter estimation, t-values for each parameter are calculated and compared to t-critical to determine whether all parameters are still statistically significant. Any parameter that is not statistically significant is removed from the regression equation. This process continues until no more variables "enter" or "leave" the regression equation. In general, it is best to select the model that yields a reasonably high "goodness of fit" for the fewest parameters in the model (referred to as *model parsimony*). The final decision on model selection requires the judgment of the model builder, and on mechanistic insights into the problem. Again, one has to guard against the danger of overfitting by performing a cross-validation check.

When a black-box model is used containing several regressors, step-wise regression would improve the robustness of the model by reducing the number of regressors in the model, and thus hopefully reduce the adverse effects of multicollinearity between the remaining regressors. Many packages use the F-test indicative of the overall model instead of the t-test on individual parameters to perform the step-wise regression. A value of F=4 is often chosen. It is suggested that step-wise regression not be used in case the regressors are highly correlated since it may result in non-robust models. However, the backward procedure is said to better handle such situations than the forward selection procedure.

A note of caution is warranted in using stepwise regression for engineering models based on mechanistic considerations. In certain cases, stepwise regression may omit a regressor which ought to be influential when using a particular data set, while the regressor is picked up when another data set is used. This may be a dilemma when the model is to be used for subsequent predictions. In such cases, discretion based on physical considerations should trump purely statistical model building.

[6] Actually, there is no "best" model since random variables are involved. A better term would be "most plausible" and should include mechanistic considerations, if appropriate.

Example 5.7.3:[7] *Proper model identification with multivariate regression models*

An example of multivariate regression is the development of model equations to characterize the performance of refrigeration compressors. It is possible to regress compressor manufacturer's tabular data of compressor performance using the following simple bi-quadratic formulation (see Fig. 5.11 for nomenclature).

$$y = C_0 + C_1 \cdot T_{cho} + C_2 \cdot T_{cdi} + C_3 \cdot T_{cho}^2 + C_4 \cdot T_{cdi}^2 + C_5 \cdot T_{cho} \cdot T_{cdi} \quad (5.60)$$

where y represents either the compressor power (P_{comp}) or the cooling capacity (Q_{ch}).

OLS is then used to develop estimates of the six model parameters, C_0–C_5, based on the compressor manufacturer's data. The biquadratic model was used to estimate the parameters for compressor cooling capacity (in Tons) for a screw compressor. The model and its corresponding parameter estimates are given below. Although the overall curve fit for the data was excellent (R^2=99.96%), the t-values of two parameter estimates (C_2 and C_4) are clearly insignificant.

A second stage regression is done omitting these regressors resulting in the following model and coefficient t-values shown in Table 5.11.

$$y = C_0 + C_1 \cdot T_{cho} + C_3 \cdot T_{cho}^2 + C_5 \cdot T_{cho} \cdot T_{cdi}$$

All of the parameters in the simplified model are significant and the overall model fit remains excellent: R^2=99.5%. ■

5.8 Case Study Example: Effect of Refrigerant Additive on Chiller Performance[8]

The objective of this analysis was to verify the claim of a company which had developed a refrigerant additive to improve chiller COP. The performance of a chiller before (called pre-retrofit period) and after (called post-retrofit period) addition of this additive was monitored for several months to determine whether the additive results in an improvement in chiller performance, and if so, by how much. The same four variables described in Example 5.4.3, namely two temperatures (T_{cho} and T_{cdi}), the chiller thermal cooling load (Q_{ch}) and the electrical power consumed (P_{comp}) were measured in intervals of 15 min. Note that the chiller COP can be deduced from the last two variables. Altogether, there were 4,607 and 5,078 data points for the pre-and post periods respectively.

Table 5.11 Results of the first and second stage model building

With all parameters			With significant parameters only	
Coefficient	Value	t-value	Value	t-value
C_0	152.50	6.27	114.80	73.91
C_1	3.71	36.14	3.91	11.17
C_2	−0.335	−0.62	–	–
C_3	0.0279	52.35	0.027	14.82
C_4	−0.000940	−0.32	–	–
C_5	−0.00683	−6.13	−0.00892	−2.34

Step 1: Perform Exploratory Data Analysis At the onset, an exploratory data analysis should be performed to determine the spread of the variables, and their occurrence frequencies during the pre- and post-periods, i.e., before and after addition of the refrigerant additive. Further, it is important to ascertain whether the operating conditions during both periods are similar or not. The eight frames in Fig. 5.27 summarize the spread and frequency of the important variables. It is noted that though the spreads in the operating conditions are similar, the frequencies are different during both periods especially in the condenser temperatures and the chiller load variables. Figure 5.28 suggests that $COP_{post}>COP_{pre}$. An ANOVA test with results shown in Table 5.12 and Fig. 5.29 also indicates that the mean of post-retrofit power use is statistically different at 95% confidence level as compared to the pre-retrofit power.

t Test to Compare Means

Null hypothesis: mean (COP_{post}) = mean (COP_{pre})

Alternative hypothesis: mean (COP_{post}) ≠ mean (COP_{pre}) assuming equal variances:

$$t = 38.8828, \ \text{p-value} = 0.0$$

The null hypothesis is rejected at α=0.05.

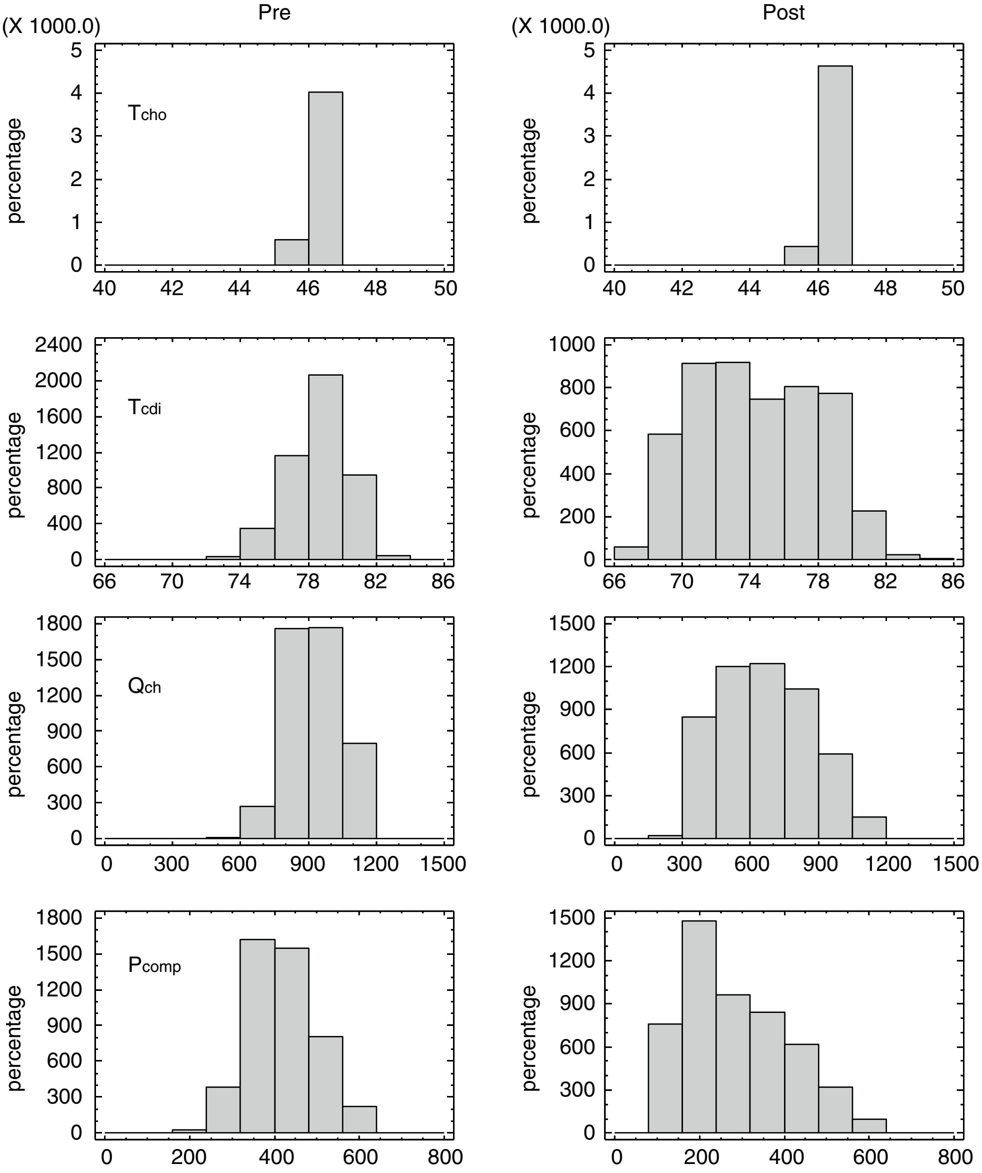
Of particular interest is the confidence interval for the difference between the means, which extends from 0.678 to 0.750. Since the interval does not contain the value 0.0, there is a statistically significant difference between the means of the two samples at the 95.0% confidence level. However, it would be incorrect to infer that $COP_{post}>COP_{pre}$ *since the operating conditions are different, and thus one should not use this approach to draw any conclusions.* Hence, a regression model based approach is warranted.

Step 2: Use the Entire Pre-retrofit Data to Identify a Model The GN chiller models (Gordon and Ng 2000) are described in Pr. 5.13. The monitored data is first used to compute the variables of the model given by the regression model Eqs. 5.70 and 5.71. Then, a linear regression is performed which is given below along with standard errors of the coefficients shown within parenthesis:

[7] From ASHRAE (2005) © American Society of Heating, Refrigerating and Air-conditioning Engineers, Inc., www.ashrae.org.

[8] The monitored data was provided by Ken Gillespie for which we are grateful.

Fig. 5.27 Histograms depicting the range of variation and frequency of the four important variables before and after the retrofit (pre=4,607 data points, post=5,078 data points). The condenser water temperature and the chiller load show much larger variability during the post period



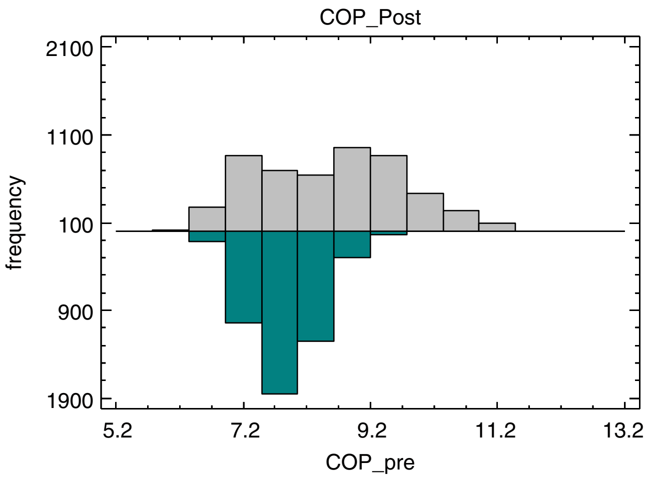


Fig. 5.28 Histogram plots of Coefficient of Performance (COP) of chiller before and after retrofit. Clearly, there are several instances when $COP_{post} > COP_{pre}$ but that could be due to operating conditions. Hence, a regression modeling approach is clearly warranted

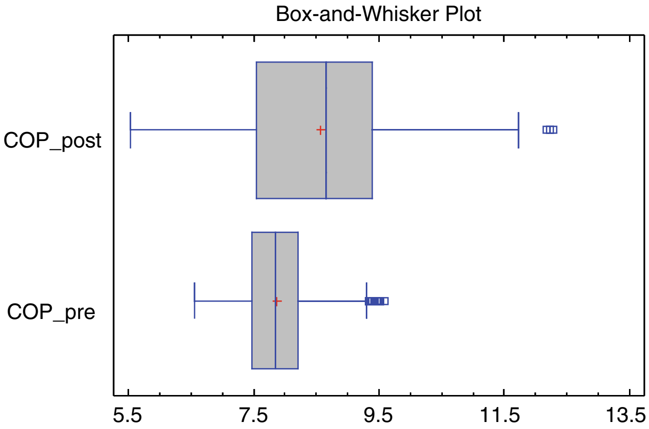


Fig. 5.29 ANOVA test results in the form of box-and-whisker plots for chiller COP before and after addition of refrigerant additive

Table 5.12 Results of the ANOVA Test of comparison of means at significance level of 0.05

95.0% confidence interval for mean of COP_{post}:	8.573 ±0.03142 = [8.542, 8.605]
95.0% confidence interval for mean of COP_{pre}:	7.859 ±0.01512 = [7.844, 7.874]
95.0% confidence interval for the difference between the means assuming equal variances:	0.714 ±0.03599 = [0.678, 0.750]

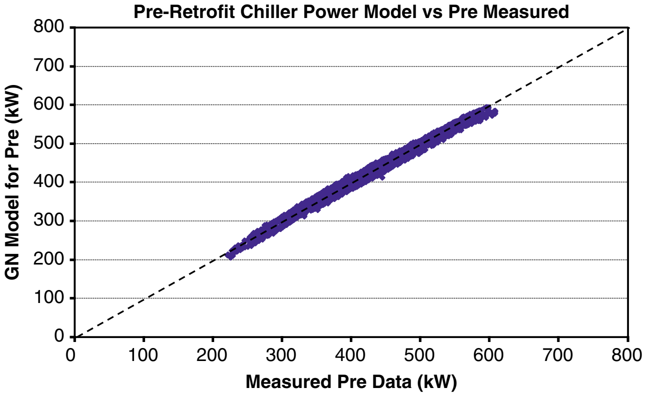


Fig. 5.30 X–Y plot of chiller power during pre-retrofit period. The overall fit is excellent (RMSE=9.36 kW and CV=2.24%), and except for a few data points, the data seems well behaved. Total number of data points=4,607

$$y = \underset{(0.00163)}{-0.00187} \cdot x_1 \underset{(15.925)}{+261.2885} \cdot x_2 \underset{(0.000111)}{+0.022461} \cdot x_3$$
$$\text{with adjusted } R^2 = 0.998 \qquad (5.61)$$

This model is then re-transformed into a model for power using Eq. 5.76, and the error statistics using the pre-retrofit data are found to be: RMSE=9.36 kW and CV=2.24%. Figure 5.30 shows the x–y plot from which one can visually evaluate the goodness of fit of the model. Note that the mean power use=418.7 kW while the mean model residuals=0.017 kW (negligibly close to zero, as it should be. This step validates the fact that the spreadsheet cells have been coded correctly with the right formulas).

Step 3: Calculate Savings in Electrical Power The above chiller model representative of thermal performance of the chiller without refrigerant additive is used to estimate savings by first predicting power use for each 15 min interval using the two operating temperatures and the load corresponding to the 5,078 post-retrofit data points. Subsequently, savings in chiller power are deduced for each of the 5,078 data points:

$$\begin{aligned} Power\ savings &= Model\text{-}predicted\ pre\text{-}retrofit\ use \\ &\quad - measured\ post\text{-}retrofit\ use \end{aligned} \qquad (5.62)$$

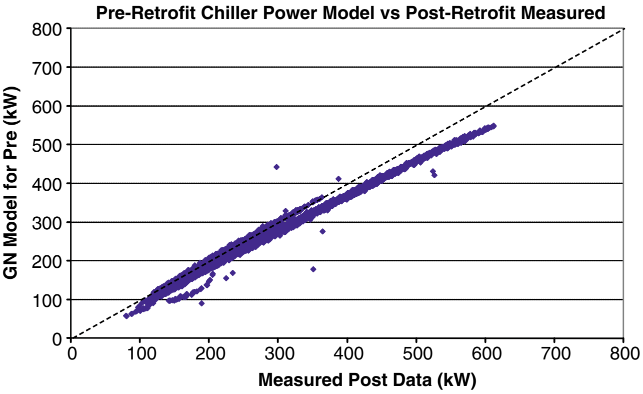


Fig. 5.31 Difference in X–Y plots of chiller power indicating that post-retrofit values are higher than those during pre-retrofit period (mean increase=21 kW or 7.88%). One can clearly distinguish two operating patterns in the data suggesting some intrinsic behavioral change in chiller operation. Entire data set for the post-period consisting of 5,078 observations has been used in this analysis

It is found that mean power savings=−21.0 kW (i.e., **an increase in power use**) or a decrease of 7.88% in the measured mean power use of 287.5 kW. Figure 5.31 visually illustrates the extent to which power use during the post-retrofit period has increased as compared to the pre-retrofit model. Overlooking the few outliers, one notes that there are two patterns: a larger number of data points indicating that post-retrofit electricity power use was much higher and a smaller set when the difference is little to nil. The reason for the onset of two distinct patterns in operation is worthy of a subsequent investigation.

Step 4: Calculate Uncertainty in Savings and Draw Conclusions The uncertainty arises from two sources: prediction model and power measurement errors. The latter are usually small, about 0.1% of the reading, which in this particular case is less than 1 kW. Hence, this contribution can be neglected during an initial investigation such as this one. The model uncertainty is given by:

$$\begin{aligned} &\text{absolute uncertainty in power use savings or reduction} \\ &= (\text{t\_value x RMSE}) \end{aligned} \qquad (5.63)$$

The t-value at 90% confidence level=1.65 and RMSE of model (for pre-retrofit period)=9.36 kW.

Hence the calculated increase in power due to refrigerant additive = −21.0 kW ± 15.44 kW at 90% CL. Thus, one would conclude that the refrigerant additive is actually penalizing chiller performance by 7.88% since electric power use is increased.

Note: The entire analysis was redone by cleaning the post-retrofit data so as to remove the dual sets of data (see Fig. 5.31). Even then, the same conclusion was reached.

Table 5.13 Data table for Problem 5.1

Temperature t (°C)	0	10	20	30	40	50	60	70	80	90	100
Specific volume v (m^3/kg)	206.3	106.4	57.84	32.93	19.55	12.05	7.679	5.046	3.409	2.361	1.673
Sat. vapor enthalpy kJ/kg	2501.6	2519.9	2538.2	2556.4	2574.4	2592.2	2609.7	2626.9	2643.8	2660.1	2676

Problems

Pr. 5.1 Table 5.13 lists various properties of saturated water in the temperature range 0–100°C.

(a) Investigate first order and second-order polynomials that fit saturated vapor enthalpy to temperature in °C. Identify the better model by looking at R^2, RMSE and CV values for both models. Predict the value of saturated vapor enthalpy at 30°C along with 95% confidence intervals and 95% prediction intervals.

(b) Repeat the above analysis for specific volume but investigate third-order polynomial fits as well. Predict the value of specific volume at 30°C along with 95% confidence intervals and 95% prediction intervals.

Pr. 5.2 Tensile tests on a steel specimen yielded the results shown in Table 5.14.

(a) Assuming the regression of y on x to be linear, estimate the parameters of the regression line and determine the 95% confidence limits for $x=4.5$

(b) Now regress x on y, and estimate the parameters of the regression line. For the same value of y predicted in (a) above, determine the value of x. Compare this value with the value of 4.5 assumed in (a). If different, discuss why.

(c) Compare the R^2 and CV values of both models.

(d) Plot the residuals of both models

(e) Of the two models, which is preferable for OLS estimation.

Pr. 5.3 The yield of a chemical process was measured at three temperatures (in °C), each with two concentrations of a particular reactant, as recorded in Table 5.15.

(a) Use OLS to find the best values of the coefficients a, b, and c in the equation: $y=a+bt+cx$.

(b) Calculate the R^2, RMSE, and CV of the overall model as well as the SE of the parameters

(c) Using the β coefficient concept described in Sect. 5.4.5, determine the relative importance of the two independent variables on the yield.

Table 5.14 Data table for Problem 5.2

Tensile force x	1	2	3	4	5	6
Elongation y	15	35	41	63	77	84

Table 5.15 Data table for Problem 5.3

Temperature, t	40	40	50	50	60	60
Concentration, x	0.2	0.4	0.2	0.4	0.2	0.4
Yield y	38	42	41	46	46	49

Pr. 5.4 *Cost of electric power generation versus load factor and cost of coal*

The cost to an electric utility of producing power (C_{Ele}) in mills per kilowatt-hr ($\$10^{-3}$/kWh) is a function of the load factor (LF) in % and the cost of coal (C_{coal}) in cents per million Btu. Relevant data is assembled in Table 5.16.

(a) Investigate different models (first order and second order with and without interaction terms) and identify the best model for predicting C_{Ele} vs LF and C_{Coal}. Use stepwise regression if appropriate. (Hint: plot the data and look for trends first).

(b) Perform residual analysis

(c) Calculate the R^2, RMSE, and CV of the overall model as well as the SE of the parameters

Pr. 5.5 *Modeling of cooling tower performance*

Manufacturers of cooling towers often present catalog data showing outlet-water temperature T_{co} as a function of ambient air wet-bulb temperature (T_{wb}) and range (which is the difference between inlet and outlet water temperatures). Table 5.17 assembles data for a specific cooling tower. Identify an appropriate model (investigate first order and second order polynomial models for T_{co}) by looking at R^2, RMSE and CV values, the individual t-values of the parameters as well as the behavior of the overall model residuals.

Pr. 5.6 *Steady-state performance testing of solar thermal flat plate collector*

Solar thermal collectors are devices which convert the radiant energy from the sun into useful thermal energy that goes to heating, say, water for domestic or for industrial applications. Because of low collector time constants, heat capacity effects are usually small compared to the hourly time step

Table 5.16 Data table for Problem 5.4

LF	85	80	70	74	67	87	78	73	72	69	82	89
C_{Coal}	15	17	27	23	20	29	25	14	26	29	24	23
C_{Ele}	4.1	4.5	5.6	5.1	5.0	5.2	5.3	4.3	5.8	5.7	4.9	4.8

Table 5.17 Data table for Problem 5.5

	T_{wb} (°C)				
Range (°C)	20	21.5	23	23.5	26
10	25.89	26.65	27.49	27.78	29.38
13	26.40	27.11	27.90	28.18	29.75
16	26.99	27.64	28.38	28.66	30.18
19	27.65	28.24	28.94	29.20	30.69
22	28.38	28.92	29.58	29.83	31.28

used to drive the model. The steady-state useful energy q_C delivered by a solar flat-plate collector of surface area A_C is given by the Hottel-Whillier-Bliss equation (Reddy 1987):

$$q_c = A_c F_R \Big[I_T \eta_n - U_L (T_{Ci} - T_a) \Big]^+ \tag{5.64}$$

where F_R is called the heat removal factor and is a measure of the solar collector performance as a heat exchanger (since it can be interpreted as the ratio of actual heat transfer to the maximum possible heat transfer); η_n is the optical efficiency or the product of the transmittance and absorptance of the cover and absorber of the collector at normal solar incidence; U_L is the overall heat loss coefficient of the collector which is dependent on collector design only, I_T is the radiation intensity on the plane of the collector, T_{ci} is the temperature of the fluid entering the collector, and T_a is the ambient temperature. The $^+$ sign denotes that only positive values are to be used, which physically implies that the collector should not be operated if q_C is negative i.e., when the collector loses more heat than it can collect (which can happen under low radiation and high T_{ci} conditions).

Steady-state collector testing is the best manner for a manufacturer to rate his product. From an overall heat balance on the collector fluid and from Eq. 5.64, the expressions for the instantaneous collector efficiency η_c under normal solar incidence are:

$$\eta_C \equiv \frac{q_C}{A_C I_T} = \frac{(mc_p)_C (T_{Co} - T_{Ci})}{A_C I_T} \\ = \left[F_R \eta_n - F_R U_L \left(\frac{T_{Ci} - T_a}{I_T} \right) \right] \tag{5.65}$$

where m_c is the total fluid flow rate through the collectors, c_{pc} is the specific heat of the fluid flowing through the collector, and T_{ci} and T_{co} are the inlet and exit temperatures of the fluid to the collector. Thus, measurements (of course done as per the standard protocol, ASHRAE 1978) of I_T, T_{ci} and T_{co} are done under a pre-specified and controlled value of fluid flow rate. The test data are plotted as η_c against reduced temperature $[(T_{Ci} - T_a)/I_T]$ as shown in Fig. 5.32. A linear fit is made to these data points by regression, from which the values of $F_R\ \eta_n$ and $F_R\ U_L$ are easily deduced.

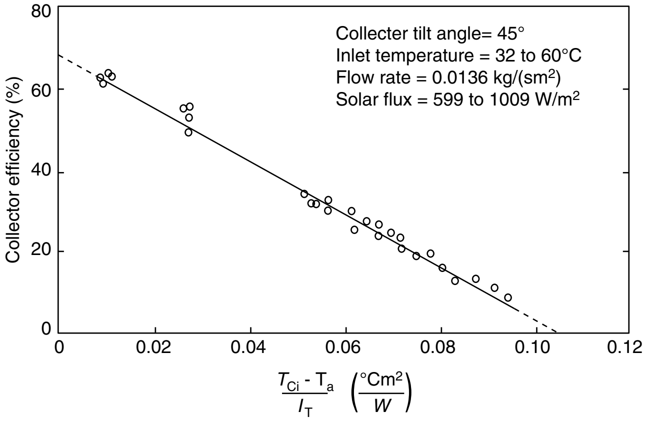


Fig. 5.32 Test data points of thermal efficiency of a double glazed flat-plate liquid collector with reduced temperature. The regression line of the model given by Eq. 5.65 is also shown. (From ASHRAE (1978) © American Society of Heating, Refrigerating and Air-conditioning Engineers, Inc., www.ashrae.org)

If the same collector is testing during different days, slightly different numerical values are obtained for the two parameters $F_R\eta_n$ and F_RU_L which are often, but not always, within the uncertainty bands of the estimates. Model misspecification (i.e., the model is not perfect, for example, it is known that the collector heat losses are not strictly linear) is partly the cause of such variability. This is somewhat disconcerting to a manufacturer since this introduces ambiguity as to which values of the parameters to present in his product specification sheet.

The data points of Fig. 5.32 are assembled in Table 5.18. Assume that water is the working fluid.

(a) Perform OLS regression using Eq. 5.65 and identify the two parameters $F_R\eta_n$ and F_RU_L along with their standard errors. Plot the model residuals, and study their behavior.
(b) Draw a straight line visually through the data points and determine the x-axis and y-axis intercepts. Estimate the $F_R\eta_n$ and F_RU_L parameters and compare them with those determined from (a).
(c) Calculate the R^2, RMSE and CV values of the model
(d) Calculate the F-statistic to test for overall model significance of the model
(e) Perform t-tests on the individual model parameters
(f) Use the model to predict collector efficiency when $I_T = 800$ W/m², $T_{ci} = 35$°C and $T_a = 10$°C

Table 5.18 Data table for Problem 5.6

x	y (%)	x	y (%)	x	y (%)	x	y (%)
0.009	64	0.051	30	0.064	27	0.077	20
0.011	65	0.052	30	0.065	26	0.080	16
0.025	56	0.053	31	0.065	24	0.083	14
0.025	56	0.056	29	0.069	24	0.086	14
0.025	52.5	0.056	29	0.071	23	0.091	12
0.025	49	0.061	29	0.071	21	0.094	10
0.050	35	0.062	25	0.075	20		

(g) Determine the 95% CL intervals for the mean and individual responses for (f) above.

(h) The steady-state model of the solar thermal collector assumes the heat loss term given by $[UA(T_{ci}-T_a]$ to be linear with the temperature difference between collector inlet temperature and the ambient temperature. One wishes to investigate whether the model improves if the loss term is to include an additional second order term:
   (i) Derive the resulting expression for collector efficiency analogous to Eq. 5.65?
   (Hint: start with the fundamental heat balance equation—Eq. 5.64)
   (ii) Does the data justify the use of such a model?

Pr. 5.7[9] *Dimensionless model for fans or pumps*
The performance of a fan or pump is characterized in terms of the head or the pressure rise across the device and the flow rate for a given shaft power. The use of dimensionless variables simplifies and generalizes the model. Dimensional analysis (consistent with fan affinity laws for changes in speed, diameter and air density) suggests that the performance of a centrifugal fan can be expressed as a function of two dimensionless groups representing flow coefficient and pressure head respectively:

$$\Psi = \frac{SP}{D^2\omega^2\rho} \quad \text{and} \quad \Phi = \frac{Q}{D^3\omega} \tag{5.66}$$

where SP is the static pressure, Pa; D the diameter of wheel, m; ω the rotative speed, rad/s; ρ the density, kg/m^3 and Q the volume flow rate of air, m^3/s.

For a fan operating at constant density, it should be possible to plot one curve Ψ vs Φ that represents the performance at all speeds. The performance of a certain 0.3 m diameter fan is shown in Table 5.19.

Table 5.19 Data table for Problem 5.7

Rotation ω (Rad/s)	Flow rate Q (m^3/s)	Static pressure SP (Pa)	Rotation ω (Rad/s)	Flow rate Q (m^3/s)	Static pressure SP (Pa)
157	1.42	861	94	0.94	304
157	1.89	861	94	1.27	299
157	2.36	796	94	1.89	219
157	2.83	694	94	2.22	134
157	3.02	635	94	2.36	100
157	3.30	525	63	0.80	134
126	1.42	548	63	1.04	122
126	1.79	530	63	1.42	70
126	2.17	473	63	1.51	55
126	2.36	428			
126	2.60	351			
126	3.30	114			

(a) First, plot the data and formulate two or three promising functions.

(b) Identify the best function by looking at the R^2, RMSE and CV values and also at the residuals.

Assume density of air at STP conditions to be 1.204 kg/m^3

Pr. 5.8 Consider the data used in Example 5.6.3 meant to illustrate the use of weighted regression for replicate measurements with non-constant variance. For the same data set, identify a model using the logarithmic transform approach similar to that shown in Example 5.6.2

Pr. 5.9 *Spline models for solar radiation*
This problem involves using splines for functions with abrupt hinge points. Several studies have proposed correlations to predict different components of solar radiation from more routinely measured components. One such correlation relates the fraction of hourly diffuse solar radiation on a horizontal radiation (I_d) and the global radiation on a horizontal surface (I) to a quantity known as the hourly atmospheric clearness index ($k_T = I/I_0$) where I_0 is the extraterrestrial hourly radiation on a horizontal surface at the same latitude and time and day of the year (Reddy 1987). The latter is an astronomical quantity and can be predicted almost exactly. Data has been gathered (Table 5.20) from which a correlation between $(I_d/I) = f(k_T)$ needs to be identified.

(a) Plot the data and visually determine likely locations of hinge points. (Hint: there should be two points, one at either extreme).

(b) Previous studies have suggested the following three functional forms: a constant model for the lower range, a second order for the middle range, and a constant model for the higher range. Evaluate with the data provided whether this functional form still holds, and report pertinent models and relevant goodness-of-fit indices.

Table 5.20 Data table for Problem 5.9

k_T	(I_d/I)	k_T	(I_d/I)
0.1	0.991	0.5	0.658
0.15	0.987	0.55	0.55
0.2	0.982	0.6	0.439
0.25	0.978	0.65	0.333
0.3	0.947	0.7	0.244
0.35	0.903	0.75	0.183
0.4	0.839	0.8	0.164
0.45	0.756	0.85	0.166
		0.9	0.165

[9] From Stoecker (1989) by permission of McGraw-Hill.

Table 5.21 Data table for Problem 5.10

Balance point temp. (°C)	25	20	15	10	5	0	−5
VBDD (°C-Days)	4,750	3,900	2,000	1,100	500	100	0

Pr. 5.10 *Modeling variable base degree-days with balance point temperature at a specific location*

Degree-day methods provide a simple means of determining annual energy use in envelope-dominated buildings operated constantly and with simple HVAC systems which can be characterized by a constant efficiency. Such simple single-measure methods capture the severity of the climate in a particular location. The variable base degree day (VBDD) is conceptually similar to the simple degree-day method but is an improvement since it is based on the actual balance point of the house instead of the outdated default value of 65°F or 18.3°C (ASHRAE 2009). Table 5.21 assembles the VBDD values for New York City, NY from actual climatic data over several years at this location.

Identify a suitable regression curve for VBDD versus balance point temperature for this location and report all pertinent statistics.

Pr. 5.11 *Change point models of utility bills in variable occupancy buildings*

Example 5.7.1 illustrated the use of linear spline models to model monthly energy use in a commercial building versus outdoor dry-bulb temperature. Such models are useful for several purposes, one of which is for energy conservation. For example, the energy manager may wish to track the extent to which energy use has been increasing over the years, or the effect of a recently implemented energy conservation measure (such as a new chiller). For such purposes, one would like to correct, or normalize, for any changes in weather since an abnormally hot summer could obscure the beneficial effects of a more efficient chiller. Hence, factors which change over the months or the years need to be considered explicitly in the model. Two common normalization factors include changes to the conditioned floor area (for example, an extension to an existing wing), or changes in the number of students in a school. A model regressing monthly utility energy use against outdoor temperature is appropriate for buildings with constant occupancy (such as residences) or even offices. However, buildings such as schools are practically closed during summer, and hence, the occupancy rate needs to be included as the second regressor. The functional form of the model, in such cases, is a multi-variate change point model given by:

$$
\begin{aligned}
y = {} & \beta_{0,un} + \beta_0 f_{oc} + \beta_{1,un}x + \beta_1 f_{oc}x \\
& + \beta_{2,un}(x - x_c)I + \beta_2 f_{oc}(x - x_c)I
\end{aligned}
\tag{5.67}
$$

where x and y are the monthly mean outdoor temperature (T_o) and the electricity use per square foot of the school (E) respectively, and $f_{oc} = N_{oc}/N_{total}$ represents the fraction of days in the month when the school is in session (N_{oc}) to the total number of days in that particular month (N_{total}). The factor f_{oc} can be determined from the school calendar. Clearly, the unoccupied fraction $f_{un} = 1 - f_{oc}$.

The term I represents an indicator variable whose numerical value is given by Eq. 5.54b. Note that the change point temperatures for occupied and unoccupied periods are assumed to be identical since the monthly data does not allow this separation to be identified.

Consider the monthly data assembled (shown in Table 5.22).

(a) Plot the data and look for change points in the data. Note that the model given by Eq. 5.67 has 7 parameters of which x_c (the change point temperature) is the one which makes the estimation non-linear. By inspection of the scatter plot, you will assume a reasonable value for this variable, and proceed to perform a linear regression as illustrated in Example 5.7.1. The search for the best value of x_c (one with minimum RMSE) would require several OLS regressions assuming different values of the change point temperature.

Table 5.22 Data table for Example 5.11

Year	Month	E (W/ft²)	T_o (°F)	f_{oc}	Year	Month	E (W/ft²)	T_o (°F)	f_{oc}
94	Aug	1.006	78.233	0.41	95	Aug	1.351	81.766	0.39
94	Sep	1.123	73.686	0.68	95	Sep	1.337	76.341	0.71
94	Oct	0.987	66.784	0.67	95	Oct	0.987	65.805	0.68
94	Nov	0.962	61.037	0.65	95	Nov	0.938	56.714	0.66
94	Dec	0.751	52.475	0.42	95	Dec	0.751	52.839	0.41
95	Jan	0.921	49.373	0.65	96	Jan	0.921	49.270	0.65
95	Feb	0.947	53.764	0.68	96	Feb	0.947	55.873	0.66
95	Mar	0.876	59.197	0.58	96	Mar	0.873	55.200	0.57
95	Apr	0.918	65.711	0.66	96	Apr	0.993	66.221	0.65
95	May	1.123	73.891	0.65	96	May	1.427	78.719	0.64
95	Jun	0.539	77.840	0	96	Jun	0.567	78.382	0.1
95	Jul	0.869	81.742	0	96	Jul	1.005	82.992	0.2

(b) Identify the parsimonious model, and estimate the appropriate parameters of the model. Note that of the six parameters appearing in Eq. 5.67, some of the parameters may be statistically insignificant, and appropriate care should be exercised in this regard. Report appropriate model and parameter statistics.

(c) Perform a residual analysis and discuss results.

Pr. 5.12 *Determining energy savings from monitoring and verification (M&V) projects*

A crucial element in any energy conservation program is the ability to verify savings from measured energy use data—this is referred to as monitoring and verification (M&V). Energy service companies (ESCOs) are required, in most cases, to perform this as part of their services. Figure 5.33 depicts how energy savings are estimated. A common M&V protocol involves measuring the monthly total energy use at the facility for whole year before the retrofit (this is the baseline period or the pre-retrofit period) and a whole year after the retrofit (called the post-retrofit period). The time taken for implementing the energy saving measures (called the "construction period") is neglected in this simple example. One first identifies a baseline regression model of energy use against ambient dry-bulb temperature T_o during the pre-retrofit period $E_{pre}=f(T_o)$. This model is then used to predict energy use during each month of the post-retrofit period by using the corresponding ambient temperature values. The difference between model predicted and measured monthly energy use is the energy savings during that month.

$$\begin{aligned} \textit{Energy savings} &= \textit{Model-predicted pre-retrofit use} \\ &\quad - \textit{measured post-retrofit use} \end{aligned} \tag{5.68}$$

The determination of the annual savings resulting from the energy retrofit and its uncertainty are finally determined. It is very important that the uncertainty associated with the savings estimates be determined as well for meaningful conclusions to be reached regarding the impact of the retrofit on energy use.

You are given monthly data of outdoor dry bulb temperature (T_o) and area-normalized whole building electricity use WB_e) for two years (Table 5.23). The first year is the pre-retrofit period before a new energy management and control system (EMCS) for the building is installed, and the second is the post-retrofit period. Construction period, i.e., the period it takes to implement the conservation measures is taken to be negligible.

(a) Plot time series and x–y plots and see whether you can visually distinguish the change in energy use as a result of installing the EMCS (similar to Fig. 5.33);

(b) Evaluate at least two different models (with one of them being a model with indicator variables) for the pre-retrofit period, and select the better model;

Table 5.23 Data table for Problem 5.12

Pre-retrofit period			Post-retrofit period		
Month	T_o (°F)	WB_e (W/ft²)	Month	T_o (°F)	WB_e (W/ft²)
1994-Jul	84.04	3.289	1995-Jul	83.63	2.362
Aug	81.26	2.827	Aug	83.69	2.732
Sep	77.98	2.675	Sep	80.99	2.695
Oct	71.94	1.908	Oct	72.04	1.524
Nov	66.80	1.514	Nov	62.75	1.109
Dec	58.68	1.073	Dec	57.81	0.937
1995-Jan	56.57	1.237	1996-Jan	54.32	1.015
Feb	60.35	1.253	Feb	59.53	1.119
Mar	62.70	1.318	Mar	58.70	1.016
Apr	69.29	1.584	Apr	68.28	1.364
May	77.14	2.474	May	78.12	2.208
Jun	80.54	2.356	Jun	80.91	2.070

(c) Use this baseline model to determine month-by-month energy use during the post-retrofit period representative of energy use had not the conservation measure been implemented;

(d) Determine the month-by-month as well as the annual energy savings (this is the "model-predicted pre-retrofit energy use" of Eq. 5.68);

(e) The ESCO which suggested and implemented the ECM claims a savings of 15%. You have been retained by the building owner as an independent M&V consultant to verify this claim. Prepare a short report describing your analysis methodology, results and conclusions. (Note: you should also calculate the 90% uncertainty in the savings estimated assuming zero measurement uncertainty. Only the cumulative annual savings and their uncertainty are required, not month-by-month values).

Pr. 5.13[10] *Grey-box and black-box models of centrifugal chiller using field data*

You are asked to evaluate two types of models: physical or gray-box models versus polynomial or black-box models. A brief overview of these is provided below.

(a) Gray-Box Models The *Universal Thermodynamic Model* proposed by Gordon and Ng (2000) is to be used. The GN model is a simple, analytical, universal model for chiller performance based on first principles of thermodynamics and linearized heat losses. The model predicts the dependent chiller COP (defined as the ratio of chiller (or evaporator) thermal cooling capacity Q_{ch} by the electrical power P_{comp} consumed by the chiller (or compressor) with specially chosen independent (and easily measurable) parameters such as the fluid (water or air) inlet temperature from the condenser

[10] Data for this problem is given in Appendix B.

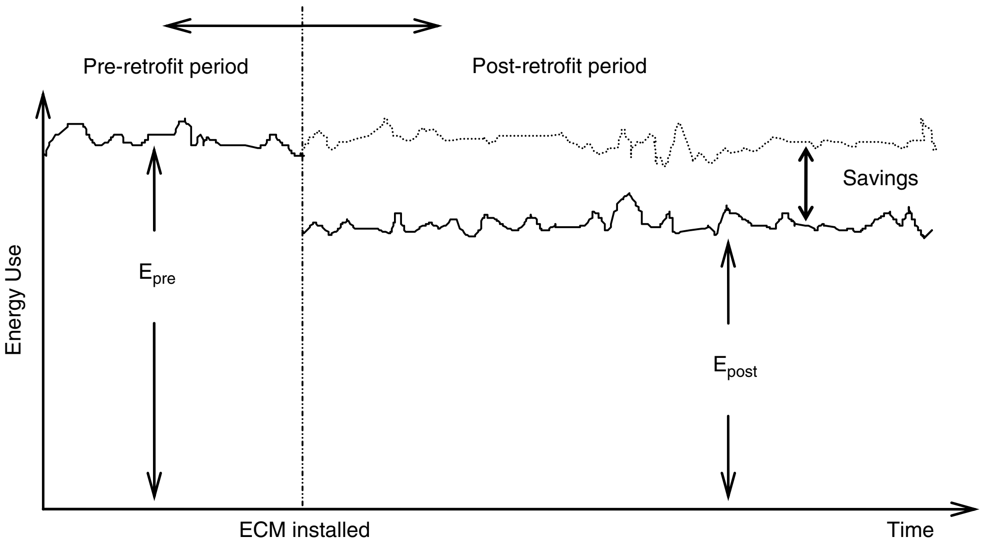


Fig. 5.33 Schematic representation of energy use prior to and after installing energy conservation measures (ECM) and of the resulting energy savings

T_{cdi}, fluid temperature leaving the evaporator (or the chilled water return temperature from the building) T_{cho}, and the thermal cooling capacity of the evaporator (similar to the figure for Example 5.4.3). The GN model is a three-parameter model which, for parameter identification, takes the following form:

$$\left(\frac{1}{COP}+1\right)\frac{T_{cho}}{T_{cdi}}-1 = a_1\frac{T_{cho}}{Q_{ch}}+a_2\frac{(T_{cdi}-T_{cho})}{T_{cdi}Q_{ch}}+a_3\frac{(1/COP+1)Q_{ch}}{T_{cdi}} \tag{5.69}$$

where the **temperatures are in absolute units,** and the parameters of the model have physical meaning in terms of irreversibilities:

$a_1=\Delta s$, the total internal entropy production rate in the chiller due to internal irreversibilities,

$a_2=Q_{leak}$, the rate of heat losses (or gains) from (or in to) the chiller,

$a_3=R=\frac{1}{(mCE)_{cond}}+\frac{1-E_{evap}}{(mCE)_{evap}}$ i.e., the total heat exchanger thermal resistance which represents the irreversibility due to finite-rate heat exchanger, and m is the mass flow rate, C the specific heat of water, and E is the heat exchanger effectiveness.

The model applies both to unitary and large chillers operating under steady state conditions. Evaluations by several researchers have shown this model to be very accurate for a large number of chiller types and sizes. If one introduces:

$$x_1=\frac{T_{cho}}{Q_{ch}}, x_2=\frac{(T_{cdi}-T_{cho})}{T_{cdi}Q_{ch}}, x_3=\frac{(1/COP+1)Q_{ch}}{T_{cdi}}$$
$$\text{and}\quad y=\left(\frac{1}{COP}+1\right)\frac{T_{cho}}{T_{cdi}}-1 \tag{5.70}$$

Eq. 5.69 assumes the following linear form:

$$y=a_1x_1+a_2x_2+a_3x_3 \tag{5.71}$$

Although most commercial chillers are designed and installed to operate at constant coolant flow rates, *variable condenser water flow operation* (as well as evaporator flow rate) is being increasingly used to improve overall cooling plant efficiency especially at low loads. In order to accurately correlate chiller model performance under variable condenser flow, an analytic model as follows was developed:

$$\frac{T_{cho}(1+1/COP)}{T_{cdi}}-1-\frac{1}{(V\rho C)_{cond}}\frac{(1/COP+1)Q_{ch}}{T_{cdi}} = c_1\frac{T_{cho}}{Q_{ch}}+c_2\left(\frac{T_{cdi}-T_{cho}}{Q_{ch}T_{cdi}}\right)+c_3\frac{Q_{ch}(1+1/COP)}{T_{cdi}} \tag{5.72}$$

If one introduces

$$x_1=\frac{T_{cho}}{Q_{ch}},\quad x_2=\frac{T_{cdi}-T_{cho}}{Q_{ch}T_{cdi}},\quad x_3=\frac{(1/COP+1)Q_{ch}}{T_{cdi}}$$

and

$$y=\frac{T_{cho}(1/COP+1)}{T_{cdi}}-1-\frac{1}{(V\rho C)_{cond}}\frac{(1/COP+1)Q_{ch}}{T_{cdi}} \tag{5.73}$$

where V, ρ and c are the volumetric flow rate, the density and specific heat of the condenser water.

For the variable condenser flow rate, Eq. 5.72 becomes

$$y=c_1x_1+c_2x_2+c_3x_3 \tag{5.74}$$

(b) Black-Box Models Whereas the structure of a gray box model, like the GN model, is determined from the underlying physics, the black box model is characterized as having no (or sparse) information about the physical problem incorporated in the model structure. The model is regarded as a black box and describes an empirical relationship between input and output variables. The commercially available DOE-2 building energy simulation model (DOE-2 1993) relies on the same parameters as those for the physical model, but uses a second order linear polynomial model instead. This "standard" empirical model (also called a *multivariate polynomial linear model or MLR*) has 10 coefficients which need to be identified from monitored data:

$$\begin{aligned} COP &= b_0 + b_1 T_{cdi} + b_2 T_{cho} \\ &\quad + b_3 Q_{ch} + b_4 {T_{cdi}}^2 + b_5 T_{cho}^2 + b_6 Q_{ch}^2 \\ &\quad + b_7 T_{cdi} T_{cho} + b_8 T_{cdi} Q_{ch} + b_9 T_{cho} Q_{ch} \end{aligned} \quad (5.75)$$

These coefficients, unlike the three coefficients appearing in the GN model, have no physical meaning and their magnitude cannot be interpreted in physical terms. Collinearity in regressors and ill-behaved residual behavior are also problematic issues. Usually one needs to retain in the model only those parameters which are statistically significant, and this is best done by step-wise regression.

Table B.3 in Appendix B assembles data consisting of 52 sets of observations from a 387 ton centrifugal chiller with variable condenser flow data. A sample hold-out cross-validation scheme will be used to guard against over-fitting. Though this is a severe type of split, **use the first 36 data points as training data and the rest (shown in italics) as testing data.**

(a) You will use the three models described above (Eqs. 5.71, 5.74 and 5.75) to identify suitable regression models. Study residual behavior as well as collinearity issues between regressors. Identify the best forms of the GN and the MLR model formulations.

(b) Evaluate which of these models is superior in terms of their external prediction accuracy The GN and MLR models have different y-values and so you cannot use the statistics provided by the regression package directly. You need to perform subsequent calculations in a spreadsheet using the power as the basis of comparing model accuracy and reporting internal and external prediction accuracies. For the MLR model, this is easily deduced from the model predicted COP values. For the GN model with constant flow, rearranging terms of Eq. 5.71 yields the following expression for the chiller electric power P_{ch}:

$$P_{comp} = \frac{Q_{ch}(T_{cdi} - T_{cho}) + a_1 T_{cdi} T_{cho} + a_2(T_{cdi} - T_{cho}) + a_3 {Q^2}_{ch}}{T_{cho} - a_3 Q_{ch}} \quad (5.76)$$

(c) Report all pertinent steps performed in your analysis and present your results succinctly.

Helpful tips:

(i) Convert temperatures into degrees Celsius, Q_{ch} into kW and volumetric flow rate V into L/s for unit consistency (work in SI units)

(ii) For the GN model, all temperatures should be in absolute units

(iii) Degrees of freedom (d.f.) have to be estimated correctly in order to compute RMSE and CV. For internal prediction, d.f. = n − p where n is the number of data points and p the number of model parameters. For external prediction accuracy, d.f. = m where m is the number of data points.

Pr. 5.14[11] Effect of tube cleaning in reducing chiller fouling

A widespread problem with liquid-cooled chillers is condenser fouling which increases heat transfer resistance in the condenser and results in reduced chiller COP. A common remedy is to periodically (every year or so) brush-clean the insides of the condenser tubes. Some practitioners question the efficacy of this process though this is widely adopted in the chiller service industry. In an effort to clarify this ambiguity, an actual large chiller (with refrigerant R11) was monitored during normal operation for 3 days before (9/11-9/13-2000) and 3 days after (1/17-1/19-2001) tube cleaning was done. Table B.4 (in Appendix B) assembles the entire data set of 72 observations for each period. This chiller is similar to the figure for Example 5.4.3.

Analyze, using the GN model described in Pr. 5.13, the two data sets, and determine the extent to which the COP of the chiller has improved as a result of this action. Prepare a report describing your analysis methodology, your analysis results, the uncertainty in your results, your conclusions, and any suggestions for future analysis work.

References

ASHRAE, 1978, Standard 93-77: Methods of Testing to Determine the Thermal Performance of Solar Collectors, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA.

ASHRAE, 2005. *Guideline 2-2005: Engineering Analysis of Experimental Data*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA.

ASHRAE, 2009. *Fundamentals Handbook*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA.

Belsley, D.A., E. Kuh, and R.E Welsch, 1980, *Regression Diagnostics*, John Wiley & Sons, New York.

Chatfield, C., 1995. *Problem Solving: A Statistician's Guide*, 2nd Ed., Chapman and Hall, London, U.K.

Chatterjee, S. and B. Price, 1991. *Regression Analysis by Example*, 2nd Edition, John Wiley & Sons, New York.

[11] Data for this problem is given in Appendix B.

Cook, R.D. and S. Weisberg, 1982. *Residuals and Influence in Regression*, Chapman and Hall, New York.

DOE-2, 1993. Building Energy simulation software developed by Lawrence Berkeley National Laboratory with funding from U.S. Department of Energy, http://simulationresearch.lbl.gov/

Draper, N.R. and H. Smith, 1981. *Applied Regression Analysis*, 2nd Ed., John Wiley and Sons, New York.

Gordon, J.M. and K.C. Ng, 2000. *Cool Thermodynamics*, Cambridge International Science Publishing, Cambridge, UK

Hair, J.F., R.E. Anderson, R.L. Tatham, and W.C. Black, 1998. *Multivariate Data Analysis*, 5th Ed., Prentice Hall, Upper Saddle River, NJ,

Katipamula, S., T. A. Reddy and D. E. Claridge, 1998. "Multivariate regression modeling", *ASME Journal of Solar Energy Engineering,* vol. 120, p.177, August.

Pindyck, R.S. and D.L. Rubinfeld, 1981. *Econometric Models and Economic Forecasts*, 2nd Edition, McGraw-Hill, New York, NY.

Reddy, T.A., 1987. *The Design and Sizing of Active Solar Thermal Systems*, Oxford University Press, Clarendon Press, U.K., September.

Reddy, T.A., N.F. Saman, D.E. Claridge, J.S. Haberl, W.D. Turner W.D., and A.T. Chalifoux, 1997. Baselining methodology for facility-level monthly energy use- part 1: Theoretical aspects, *ASHRAE Transactions*, v.103 (2), American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA.

Schenck, H., 1969. *Theories of Engineering Experimentation*, Second Edition, McGraw-Hill, New York.

Shannon, R.E., 1975. *System Simulation: The Art and Science*, Prentice-Hall, Englewood Cliffs, NJ.

Stoecker, W.F., 1989. *Design of Thermal Systems*, 3rd Edition, McGraw-Hill, New York.

Walpole, R.E., R.H. Myers, and S.L. Myers, 1998. *Probability and Statistics for Engineers and Scientists*, 6th Ed., Prentice Hall, Upper Saddle River, NJ