Design of Experiments: A gentle introduction

FITTING REGRESSION MODELS

Fabio Stella

University of Milano-Bicocca

fabio.stella@unimib.it

LECTURE LEARNING OBJECTIVES

1) Understand fitting linear regression models using the method of least squares.
2) Understand basic regression model inference, including tests for significance of regression, tests on individual model parameters and confidence intervals on the parameters.
3) Know how to use the model to estimate the mean response at a specific point and construct the associated confidence interval.
4) Know how to use the model to predict a new response at a specific point and construct the associated prediction interval.
5) Know the difference between a confidence interval on the mean response and a prediction interval on a new response and when each interval is appropriate.
6) Know how to use basic regression model diagnostics, such as residual plots, the PRESS statistic, and measures of leverage and influence.
7) Know how to test for lack of fit of a linear regression model.

Fitting Regression Models: Introduction

In many problems two or more variables are related, and it is of interest to model and explore this relationship.

For example, in a **chemical process** the **yield of product** is related to the **operating temperature**. The chemical engineer may want to **build a model relating yield to temperature** and then use the model **for prediction**, **process optimization**, or **process control**.

In general, suppose that there is a single **dependent variable or response** y that depends on k **independent or regressor variables**, for example, $x_1, x_2, \ldots x_k$.

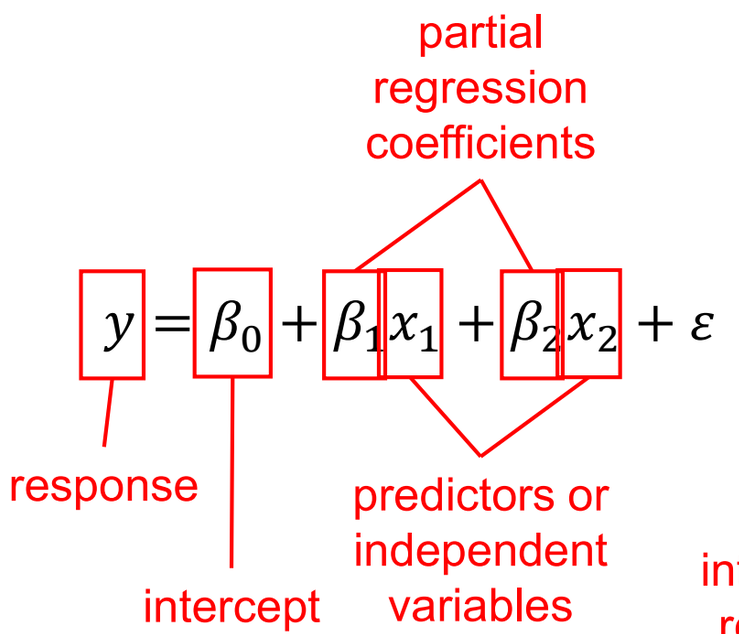
The relationship between these variables is characterized by a mathematical model called a **regression model**.

The regression model is **fit to a set of sample data**.

In some instances, the experimenter knows the exact form of the true functional relationship between y and $x_1, x_2, \ldots x_k$, in some cases not, thus it must approximate it.

We will focus on fitting linear regression models. To illustrate, suppose that we wish to develop an empirical model relating the viscosity of a polymer to the temperature and the catalyst feed rate.

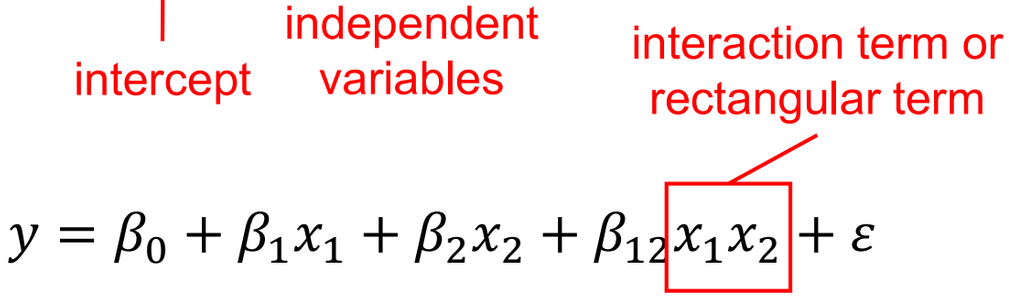
A model that might describe this relationship is

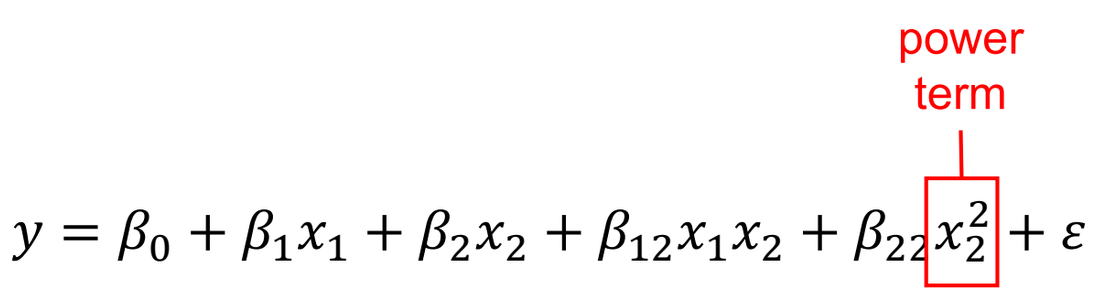


y , viscosity
x_1, temperature
x_2, catalyst feed rate

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple Linear Regression model





The method of **least squares** is typically used to estimate the regression coefficients in a multiple linear regression model.

Suppose we are given a dataset where a response variable y and k regressor variables $x_1, x_2, \ldots, x_k$ are measured on n samples.

Data for Multiple Linear Regression

y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

We assume that the error term ε in the model is such that the following conditions hold

- $\mathrm{E}(\varepsilon) = 0$
- $\mathrm{V}(\varepsilon) = \sigma^2$
- $\{\varepsilon_i\}$ are uncorrelated random variables.

$$\varepsilon_i \sim NID(0, \sigma^2)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,2, \ldots, n$$

The **least squares function** is

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

Data for Multiple Linear Regression

y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

The solution to the normal equations will be the least squares **estimators of the regression coefficients**

$$\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_k$$

The method of least squares chooses the β_k's in so that the sum of the squares of the errors, ε_i, is minimized.

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,2, \ldots, n$$

Fitting Regression Models: Tests on Individual Regression Coefficients and Groups of Coefficients

Viscosity Data for Example 10.1 (Viscosity in Centistokes @ 100°C)

Observation	Temperature (x_1, °C)	Catalyst Feed Rate (x_2, lb/h)	Viscosity
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

Regression Analysis

The regression equation is

Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36 R-Sq = 92.7% R-Sq (adj) = 91.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

y, viscosity

x_1, temperature

x_2, catalyst feed rate

In multiple linear regression problems, certain **tests of hypotheses about the model parameters** are helpful in **measuring the usefulness of the model**.

We describe several important hypothesis-testing procedures, which require that the **errors ε_i** in the model be **normally and independently distributed with mean zero and variance σ^2**.

$$\varepsilon_i \sim NID(0, \sigma^2) \longrightarrow y_i \sim NID\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2\right)$$

We now go into details of the following two hypothesis tests

- Test for Significance of Regression
- Tests on Individual Regression Coefficients and Groups of Coefficients

The test for significance of regression is a test to determine whether a linear relationship exists between the response variable y and a subset of the regressor variables $x_1, x_2, \ldots, x_k$.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \beta_j \neq 0, for\ at\ least\ one\ j$$

Rejection of H_0 implies that at least one of the regressor variables $x_1, x_2, \ldots, x_k$ contributes significantly to the model

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

The test procedure involves an analysis of variance partitioning of the **total sum of squares** into a **sum of squares due to the model** (or to regression) and a **sum of squares due to residual** (or error), say

$$SS_T = SS_R + SS_E = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

If the null hypothesis is true

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_k$$

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{n-k-1}$$

SS_R and SS_E are independent

The test procedure is to compute

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

If

$$F_0 > F_{\alpha,k,n-k-1} \longrightarrow \text{Reject } H_0$$

Otherwise if

$$F_0 \leq F_{\alpha,k,n-k-1} \longrightarrow \text{Not Reject } H_0$$

Alternatively, we could use the p-value approach to hypothesis testing and reject H_0 if the p-value for the statistic F_0 is less than α.

If the null hypothesis is true

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_k$$

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{n-k-1}$$

SS_R and SS_E are independent

The test procedure is to compute

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Fitting Regression Models: Test for Significance of Regression

Viscosity Data for Example 10.1 (Viscosity in Centistokes @ 100°C)

Observation	Temperature (x_1, °C)	Catalyst Feed Rate (x_2, lb/h)	Viscosity
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

Regression Analysis

The regression equation is
Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36 R-Sq = 92.7% R-Sq (adj) = 91.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

p-value

Measures the amount of reduction in the variability of y obtained by using the regressor variables x_1 and x_2 in the model.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

The p-value for F_0 is very small, so we would conclude that at least one of the two variables—temperature (x_1) and feed rate (x_2)—has a nonzero regression coefficient.

However, a large value of R^2 does not necessarily imply that the regression model is a good one.

Adding a variable to the model will always increase R^2, regardless of whether the additional variable is statistically significant or not.

Thus, it is possible for models that have large values of R^2 to yield poor predictions of new observations or estimates of the mean response.

Therefore, we prefer to use an adjusted statistic

$$R^2_{adj} = 1 - \frac{\frac{SS_E}{(n-p)}}{\frac{SS_T}{(n-1)}} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

Measures the amount of reduction in the variability of y obtained by using the regressor variables x_1 and x_2 in the model.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Regression Analysis

The regression equation is
Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36 R-Sq = 92.7% R-Sq (adj) = 91.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

R^2 vs R^2_{adj} not a dramatic difference.

In general, R^2_{adj} will not always increase as variables are added to the model.
In fact, if unnecessary terms are added, the value of R^2_{adj} will often decrease.

However, a large value of R^2 does not necessarily imply that the regression model is a good one.

Adding a variable to the model will always increase R^2, regardless of whether the additional variable is statistically significant or not.

Thus, it is possible for models that have large values of R^2 to yield poor predictions of new observations or estimates of the mean response.

Therefore, we prefer to use an adjusted statistic

$$R^2_{adj} = 1 - \frac{\frac{SS_E}{(n-p)}}{\frac{SS_T}{(n-1)}} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

Measures the amount of reduction in the variability of y obtained by using the regressor variables x_1 and x_2 in the model.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Regression Analysis

The regression equation is
Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36 R-Sq = 92.7% R-Sq (adj) = 91.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

When R^2 and R^2_{adj} differ dramatically, there is a good chance that nonsignificant terms have been included in the model.

We are frequently interested in testing hypotheses on the individual regression coefficients.

Such tests would be useful in **determining the value of each regressor variable** in the regression model.

For example, the model might be more effective with the inclusion of additional variables or perhaps with the deletion of one or more of the variables already in the model.

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease.

$$SS_T = SS_R + SS_E = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model.

Furthermore, adding an unimportant variable to the model can actually increase the mean square error, thereby decreasing the usefulness of the model.

The hypotheses for testing the significance of any individual regression coefficient, for example β_j

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

$t_0 \leq t_{\alpha/2,n-k-1}$ → Not Reject H_0 → x_j can be deleted

Regression Analysis

The regression equation is
Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36 R-Sq = 92.7% R-Sq (adj) = 91.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

A procedure can also be used to investigate the contribution of a *subset* of the regressor variables to the model

It is often necessary to construct confidence interval estimates for the regression coefficients $\{\beta_j\}$ and for other quantities of interest from the regression model.

The development of a procedure for obtaining these confidence intervals requires that we assume the errors $\{\varepsilon_i\}$ to be normally and independently distributed with mean zero and variance σ^2.

$$\hat{\beta}_{j0} - t_{\alpha/2,n-k-1}\, se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_{j0} + t_{\alpha/2,n-k-1}\, se(\hat{\beta}_j)$$

We will construct a 95 percent confidence interval for the parameter β_1 in Example 10.1. Now $\hat{\beta}_1 = 7.62129$, and because $\hat{\sigma}^2 = 267.604$ and $C_{11} = 1.429184 \times 10^{-3}$, we find that

$$\hat{\beta}_1 - t_{0.025,13}\sqrt{\hat{\sigma}^2 C_{11}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,13}\sqrt{\hat{\sigma}^2 C_{11}}$$

$$7.62129 - 2.16\sqrt{(267.604)(1.429184 \times 10^{-3})} \le \beta_1 \le 7.62129 + 2.16\sqrt{(267.604)(1.429184 \times 10^{-3})}$$

$$7.62129 - 2.16(0.6184) \le \beta_1 \le 7.62129 + 2.16(0.6184)$$

and the 95 percent confidence interval on β_1 is

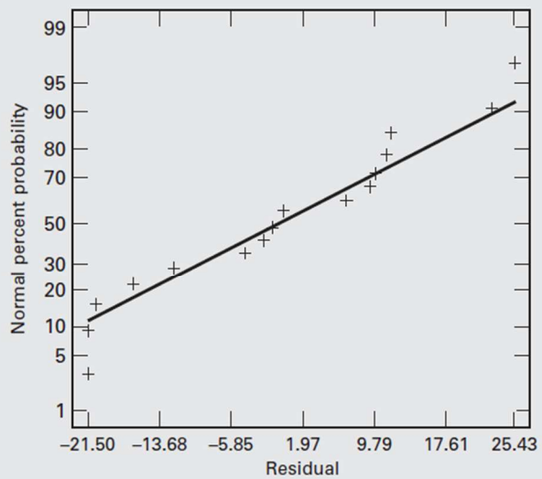
$$6.2855 \le \beta_1 \le 8.9570$$

Model adequacy checking is an important part of the data analysis procedure.

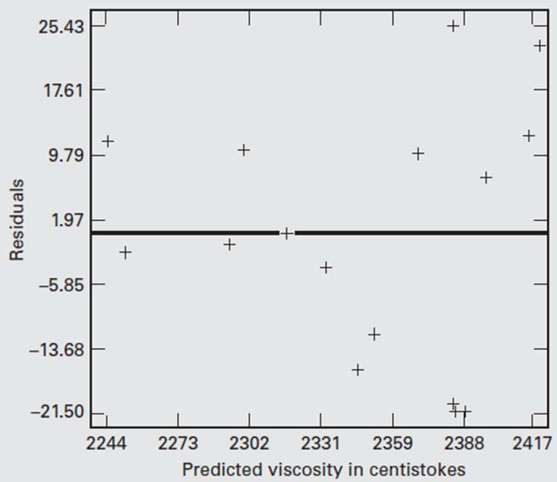
In general, it is always necessary to

1. examine the fitted model to ensure that it provides an adequate approximation to the true system and
2. verify that none of the least squares regression assumptions are violated.

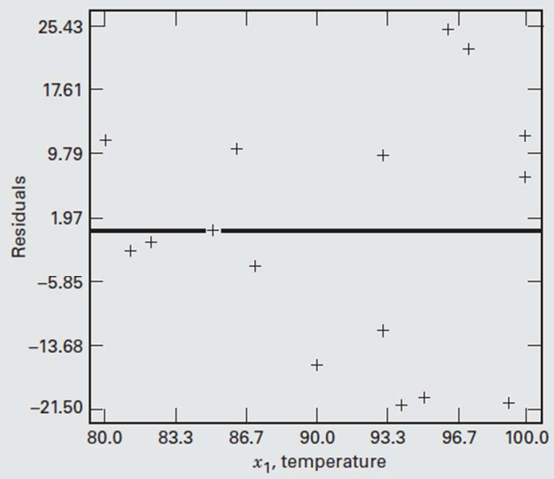
The regression model will probably give poor or misleading results unless it is an adequate fit. A first diagnostic is given by **residual plots**.



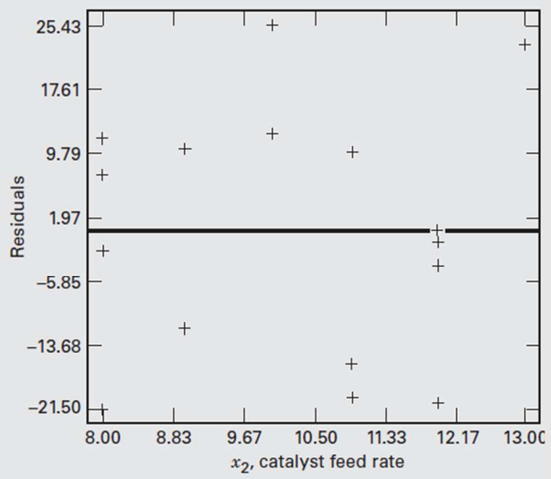
■ FIGURE 10.1 Normal probability plot of residuals, Example 10.1



■ FIGURE 10.2 Plot of residuals versus predicted viscosity, Example 10.1



■ FIGURE 10.3 Plot of residuals versus x_1 (temperature), Example 10.1



■ FIGURE 10.4 Plot of residuals versus x_2 (feed rate), Example 10.1

In addition to residual plots, other **model diagnostics** are frequently useful in regression.

Standardized and Studentized Residuals. Many model builders prefer to work with scaled residuals in contrast to the ordinary least squares residuals. These scaled residuals often convey more information than do the ordinary residuals.

- **Standardized Residuals**. $d_i = \dfrac{e_i}{\hat{\sigma}} \qquad i = 1, 2, \ldots, n \qquad \hat{\sigma} = \sqrt{MS_E}$

  Most of the standardized residuals should lie in the interval $-3 \leq d_i \leq 3$, and any observation with a standardized residual outside of this interval is potentially unusual with respect to its observed response.

  These outliers should be carefully examined because they may represent something as simple as a data-recording error or something of more serious concern, such as a region of the regressor variable space where the fitted model is a poor approximation to the true response surface.

- **Studentized Residuals**. $r_i = \dfrac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \qquad i = 1, 2, \ldots, n \qquad \hat{\sigma}^2 = MS_E$

  The studentized residuals have constant variance $V(r_i) = 1$ regardless of the location of $\mathbf{x}_i$ when the form of the model is correct.

  In many situations the variance of the residuals stabilizes, particularly for large data sets, thus studentized reisuals become extremly close to standardized residuals.

- **PRESS Residuals.** $$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \qquad i = 1, 2, \ldots, n$$

  We fit the regression model to the remaining $n-1$ observations and use this equation to predict the withheld observation y_i.

  Denoting this predicted value y_i, we may find the prediction error for point i as $e(i) = y_i - \hat{y}_i$. The prediction error is often called the i^{th} PRESS residual.

  This procedure is repeated for each observation $i = 1, \ldots, n$, producing a set of n PRESS residuals $e(1), e(2), \ldots, e(n)$.

  Then the PRESS statistic is defined as the sum of squares of the n PRESS residuals as in

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

  Thus, PRESS uses each possible subset of $n-1$ observations as an estimation data set, and every observation in turn is used to form a prediction data set.

$$R^2_{\text{Prediction}} = 1 - \frac{\text{PRESS}}{SS_T}$$

- **Leverage Points.**

The disposition of points in x space is important in determining model properties.

In particular, remote observations potentially have disproportionate leverage on the parameter estimates, predicted values, and the usual summary statistics.

Letting $p = rank(\mathbf{X})$, when $h_{ii} > \frac{2p}{n}$ the observation x_i is a high-leverage point.

Observation i	y_i	Predicted Value $\hat{y}_i$	Residual e_i	h_{ii}	Studentized Residual	D_i	R-Student
1	2256	2244.5	11.5	0.350	0.87	0.137	0.87
2	2340	2352.1	−12.1	0.102	−0.78	0.023	−0.77
3	2426	2414.1	11.9	0.177	0.80	0.046	0.79
4	2293	2294.0	−1.0	0.251	−0.07	0.001	−0.07
5	2330	2346.4	−16.4	0.077	−1.05	0.030	−1.05
6	2368	2389.3	−21.3	0.265	−1.52	0.277	−1.61
7	2250	2252.1	−2.1	0.319	−0.15	0.004	−0.15
8	2409	2383.6	25.4	0.098	1.64	0.097	1.76
9	2364	2385.5	−21.5	0.142	−1.42	0.111	−1.48
10	2379	2369.3	9.7	0.080	0.62	0.011	0.60
11	2440	2416.9	23.1	0.278	1.66	0.354	1.80
12	2364	2384.5	−20.5	0.096	−1.32	0.062	−1.36
13	2404	2396.9	17.1	0.289	0.52	0.036	0.50
14	2317	2316.9	0.1	0.185	0.01	0.000	<0.01
15	2309	2298.8	10.2	0.134	0.67	0.023	0.66
16	2328	2332.1	−4.1	0.156	−0.28	0.005	−0.27

$$\frac{2p}{n} = \frac{2(3)}{16} = 0.375$$

no leverage points in these data

- **Influence on regression coefficients**.

  Cook (1977, 1979) has suggested using a measure of the squared distance D_i between the least squares estimate based on all n points $\hat{\beta}$ and the estimate obtained by deleting the i point, say $\hat{\beta}(i)$.

  That is, we usually consider observations for which $D_i > 1$ to be influential.

Observation i	y_i	Predicted Value $\hat{y}_i$	Residual e_i	h_{ii}	Studentized Residual	D_i	R-Student
1	2256	2244.5	11.5	0.350	0.87	0.137	0.87
2	2340	2352.1	−12.1	0.102	−0.78	0.023	−0.77
3	2426	2414.1	11.9	0.177	0.80	0.046	0.79
4	2293	2294.0	−1.0	0.251	−0.07	0.001	−0.07
5	2330	2346.4	−16.4	0.077	−1.05	0.030	−1.05
6	2368	2389.3	−21.3	0.265	−1.52	0.277	−1.61
7	2250	2252.1	−2.1	0.319	−0.15	0.004	−0.15
8	2409	2383.6	25.4	0.098	1.64	0.097	1.76
9	2364	2385.5	−21.5	0.142	−1.42	0.111	−1.48
10	2379	2369.3	9.7	0.080	0.62	0.011	0.60
11	2440	2416.9	23.1	0.278	1.66	0.354	1.80
12	2364	2384.5	−20.5	0.096	−1.32	0.062	−1.36
13	2404	2396.9	17.1	0.289	0.52	0.036	0.50
14	2317	2316.9	0.1	0.185	0.01	0.000	<0.01
15	2309	2298.8	10.2	0.134	0.67	0.023	0.66
16	2328	2332.1	−4.1	0.156	−0.28	0.005	−0.27

there is no strong evidence of influential observations in these data

Adding center points to a 2^k factorial design allows the experimenter to obtain an estimate of **pure experimental error**.

This allows the partitioning of the residual sum of squares SS_E into two components;

$$SS_E = SS_{PE} + SS_{LOF}$$

SS_{LOF}: sum of squares due to lack of fit

SS_{PE}: sum of squares due to pure error

We may give a general development of this partitioning in the context of a regression model.

Suppose that we have n_i observations on the response at the i^{th} level of the regressors $\mathbf{x}_i$, $i = 1, \ldots, m$.

Let y_{ij} denote the j^{th} observation on the response at $\mathbf{x}_i$, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$.

There are $n = \sum_{i=1}^{m} n_i$ total observations.

We may write the ij^{th} residual as

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

where $\bar{y}_i$ is the average of the n_i observations at $\mathbf{x}_i$.

If the fitted values $\hat{y}_i$ are close to the corresponding average responses $\bar{y}_i$, then there is a strong indication that the regression function is linear.

If the $\hat{y}_i$ deviate greatly from the $\bar{y}_i$, then it is likely that the regression function is not linear.

sum of squares due to pure error

$$SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

sum of squares due to lack of fit

$$SS_{LOF} = \sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2$$

Squaring and summing over i and j

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

$$F_0 = \frac{SS_{LOF}/(m-p)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} > F_{\alpha,m-p,n-m}$$

regression function is not linear

sum of squares due to pure error

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

sum of squares due to lack of fit

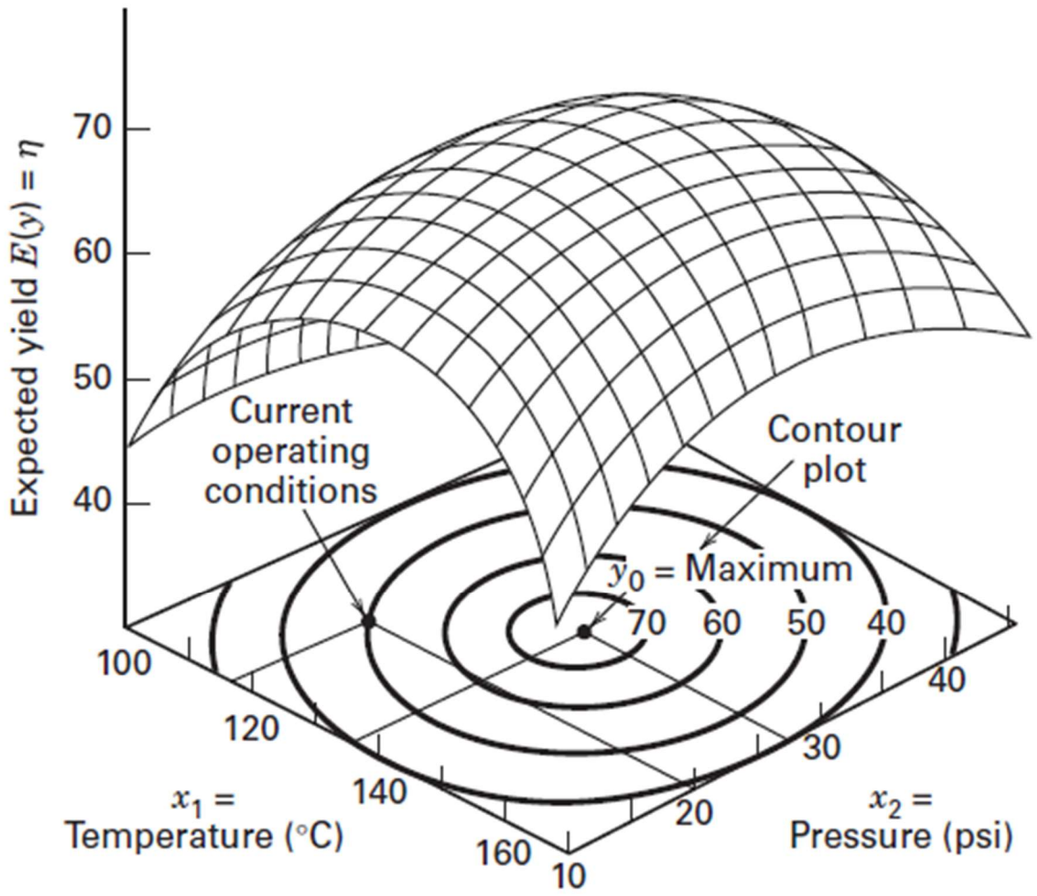
$$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

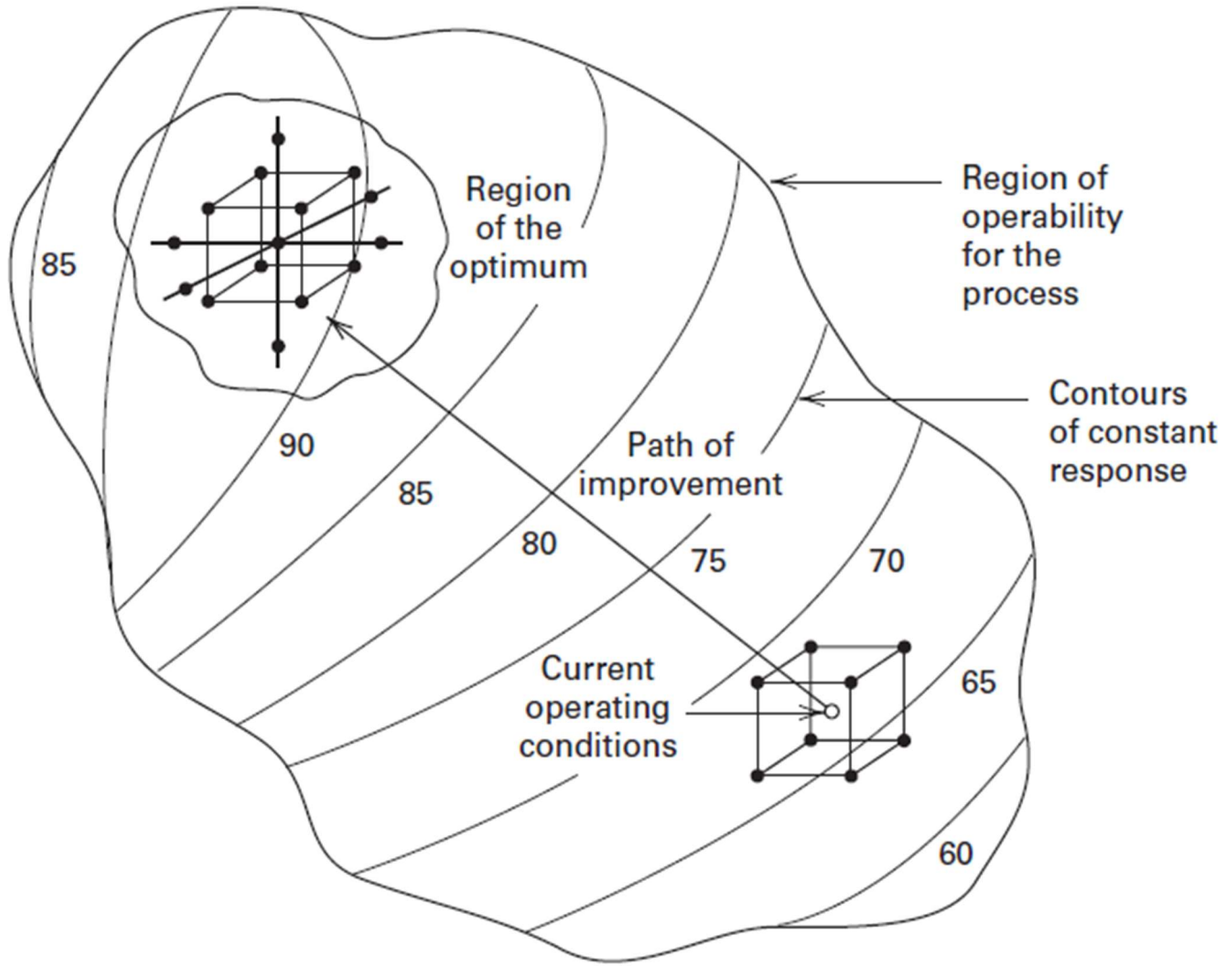
$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

Squaring and summing over i and j

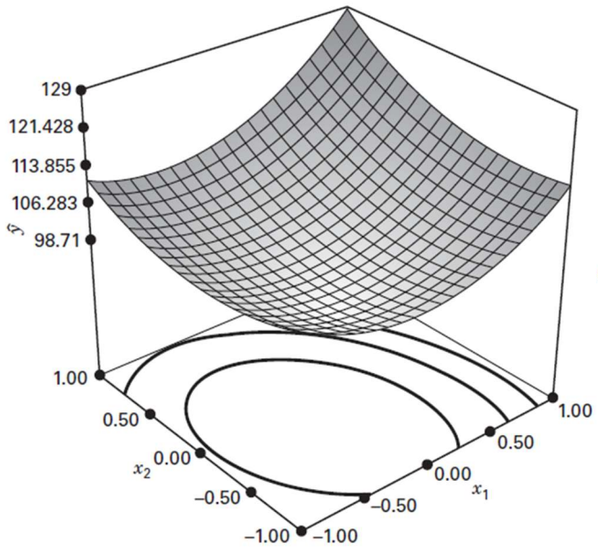
$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

Pure quadratic					
Curvature	1.51	1	1.51	0.093	0.7802
Pure error	48.75	3	16.25		
Cor total	5781.20	19			
Model	5535.81	5	1107.16	59.02	<0.000
A	*1870.56*	*1*	*1870.56*	*99.71*	*<0.000*
C	*390.06*	*1*	*390.06*	*20.79*	*0.0005*
D	*855.56*	*1*	*855.56*	*45.61*	*<0.000*
AC	*1314.06*	*1*	*1314.06*	*70.05*	*<0.000*
AD	*1105.56*	*1*	*1105.56*	*58.93*	*<0.000*
Pure quadratic					
curvature	1.51	1	1.51	0.081	0.7809
Residual	243.87	13	18.76		
Lack of fit	*195.12*	*10*	*19.51*	*1.20*	*0.4942*
Pure error	*48.75*	*3*	*16.25*		
Cor total	5781.20	19			

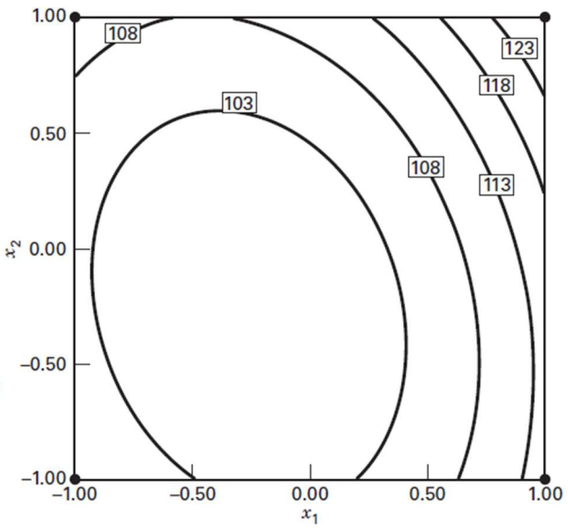

Expected yield $E(y) = \eta$
70
60
50
40
Current operating conditions
Contour plot
y_0 = Maximum
70 60 50 40
100
120
140
160
x_1 = Temperature (°C)
40
30
20
10
x_2 = Pressure (psi)


85
Region of the optimum
Region of operability for the process
Contours of constant response
90
85
80
75
70
Path of improvement
Current operating conditions
65
60

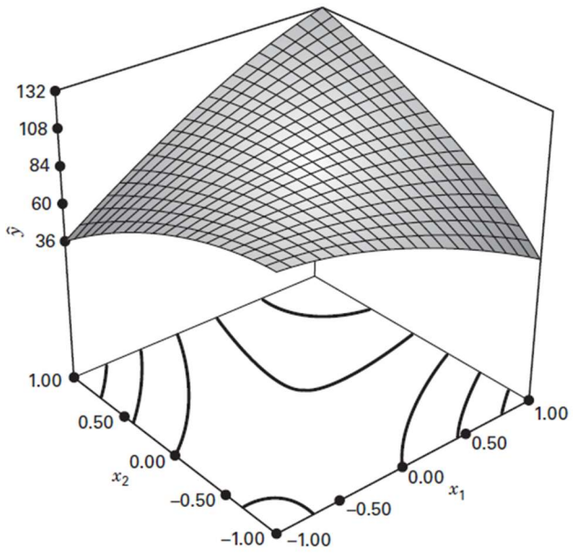
Fitting Regression Models: Response surface



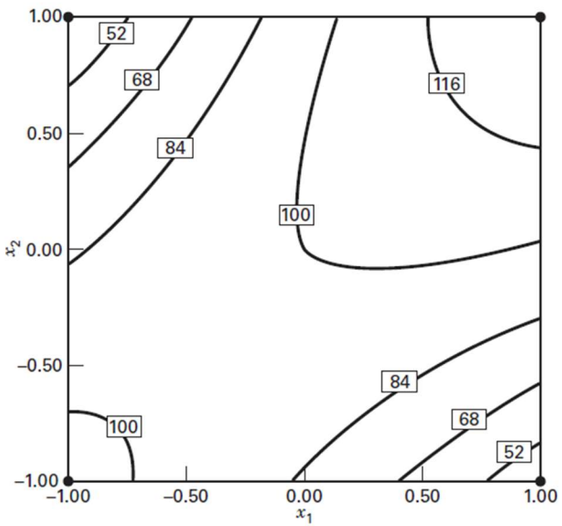
(a) Response surface



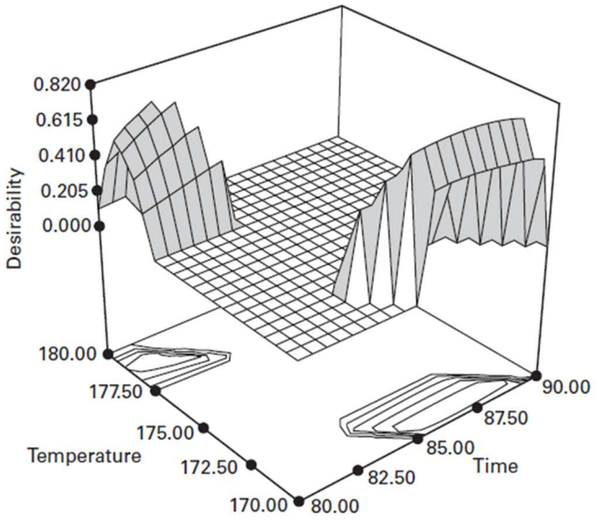
(b) Contour plot



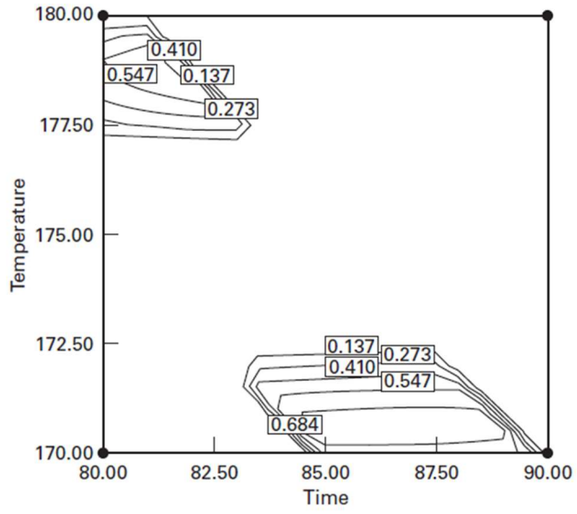
(a) Response surface



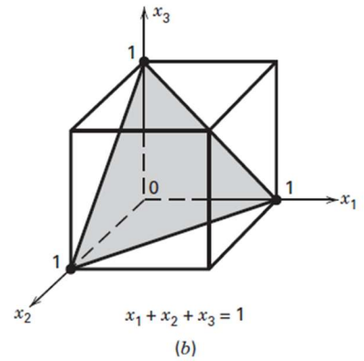
(b) Contour plot



(a) Response surface

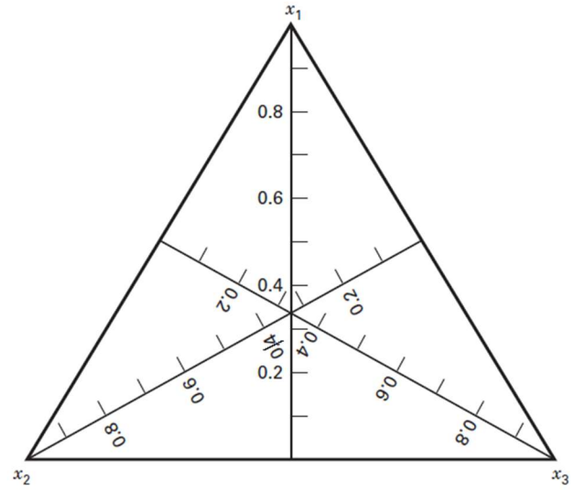


(b) Contour plot



(b)

■ **FIGURE 11.39**
Constrained factor space for mixtures with (a) $p = 2$ components and (b) $p = 3$ components



■ **FIGURE 11.40** **Trilinear coordinate system**

Designed experiments can also be successfully applied to computer simulation models of physical systems.

In such applications, the data from the experimental design is used to build a model of the system being modeled by the computer simulation—a metamodel—and optimization is carried out on the metamodel.

The assumption is that if the computer simulation model is a faithful representation of the real system, then optimization of the model will result in adequate determination of the optimum conditions for the real system.

Generally, there are two types of simulation models,

- **Stochastic**; the output responses are random variables. Often standard experimental design techniques can be applied to the output from a stochastic simulation model, although a number of specialized techniques have been developed. Sometimes polynomials of higher order than the usual quadratic response surface model are used.
- **Deterministic**; the output responses are not random variables. Deterministic simulation models are often used by engineers and scientists as computer-based design tools.

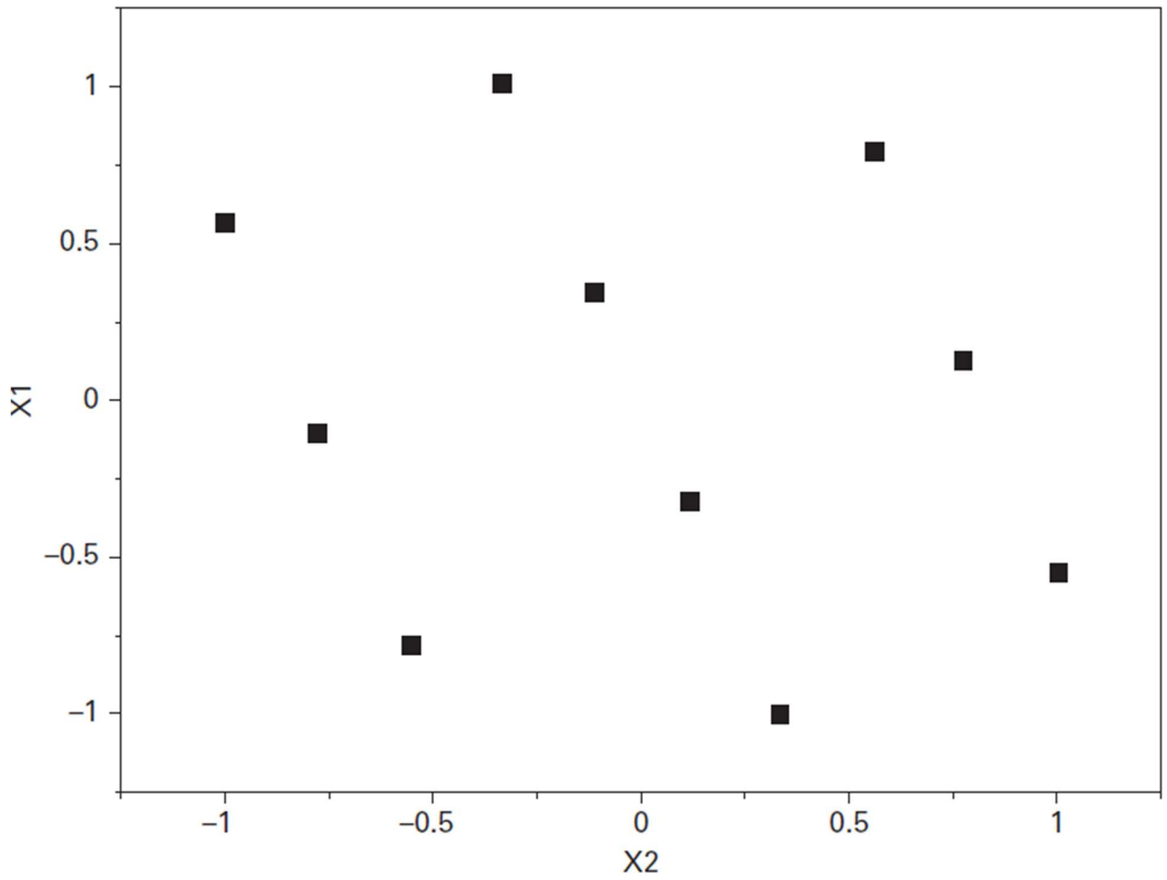
In recent years, various types of **space-filling designs** have been suggested for computer experiments.

Space-filling designs are often thought to be **particularly appropriate for deterministic computer models** because in general they **spread** the **design points out nearly evenly or uniformly** (in some sense) throughout the **region of experimentation**.

This is a **desirable** feature if the experimenter **doesn't know the form of the model that is required**, and believes that interesting phenomena are likely to be found in different regions of the experimental space.

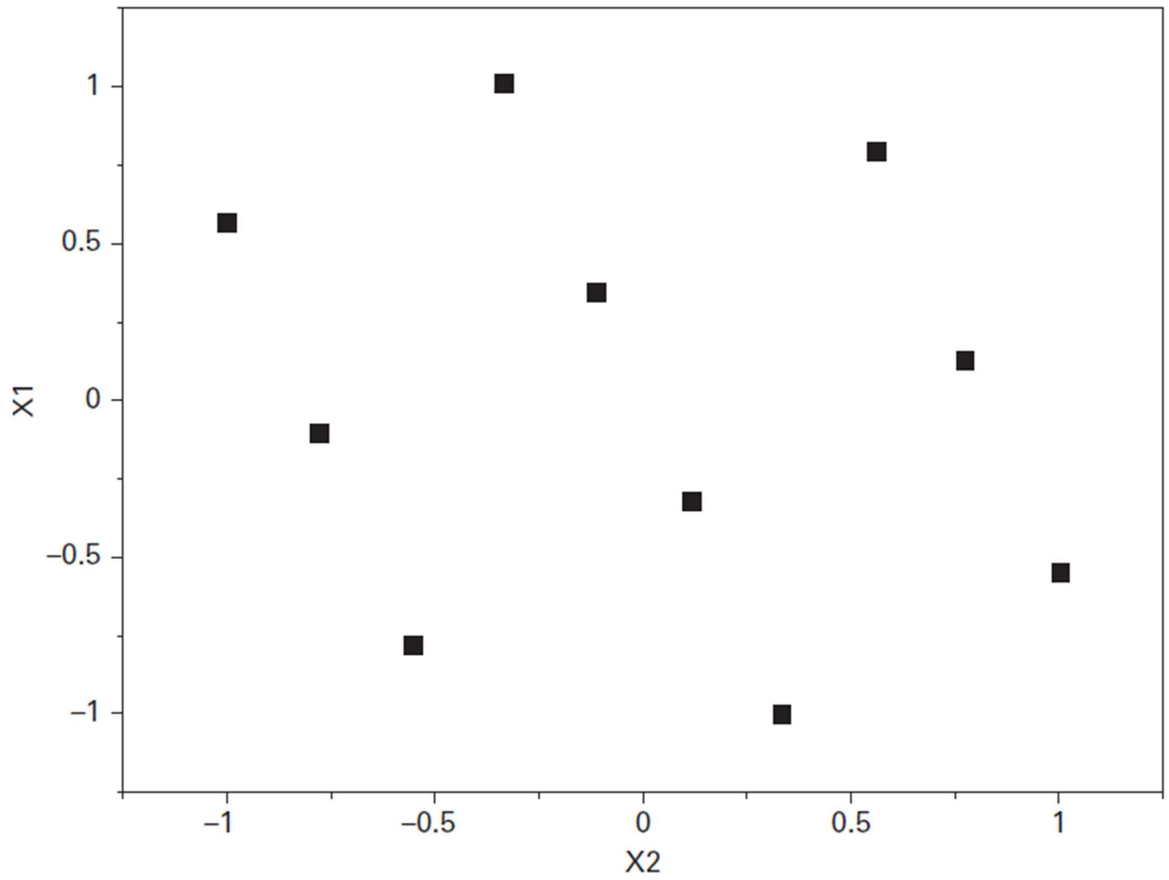
Furthermore, most space-filling designs do **not contain any replicate runs**.

For a **deterministic computer model** this is **desirable**, because **a single run** of the computer model **at a design point provides all of the information about the response at that point**.

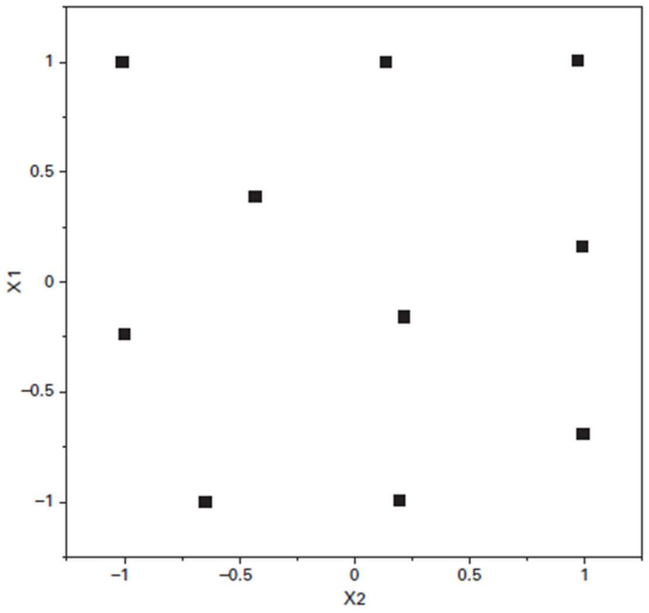


■ FIGURE 11.34 A 10-run Latin hypercube design

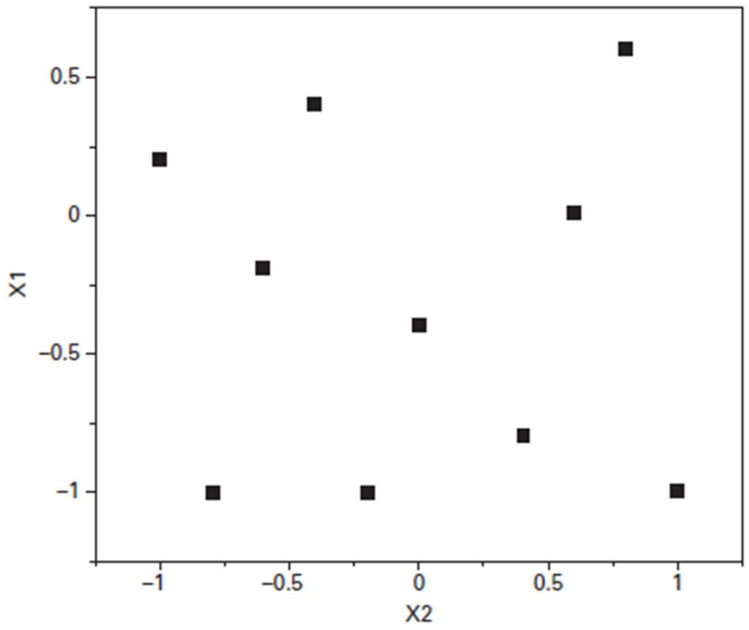
A **Latin hypercube** in n runs for k factors in an $n \times k$ matrix where each column is a random permutation of the levels $1, 2, \ldots, n$.



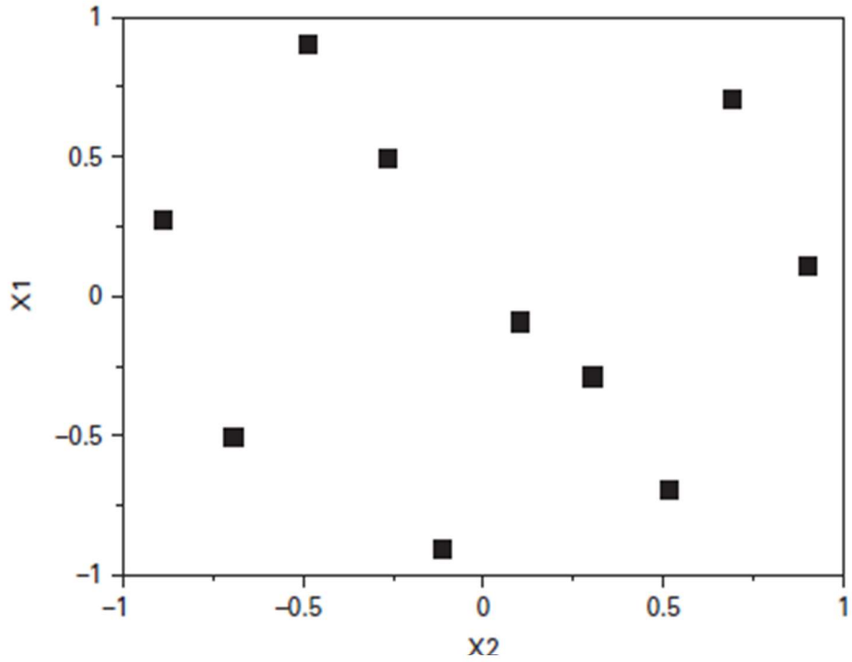
■ FIGURE 11.34 A 10-run Latin hypercube design



■ FIGURE 11.35 A 10-run sphere-packing design



■ FIGURE 11.37 A 10-run maximum entropy design



■ FIGURE 11.36 A 10-run uniform design