

Essential Regression and Experimental Design for Chemists and Engineers



David D. Steppan

Joachim Werner

Robert P. Yeater

Copyright 1998

To our wives.

Thank you for all your support and tolerance of our programming addiction.

Preface

Essential Regression and Experimental Design for Chemists and Engineers was developed as an easy-to-use book with an accompanying software package which allows non-statisticians to analyze experimental designs and quantitative data using polynomial and multiple linear regression in a straightforward and understandable manner. From our experience as chemists and engineers, these two variations of regression analysis are the ones used most often to analyze data. They are the "essential" tools in data analysis. Recognizing the widespread use of Microsoft Office® software, we developed Essential Regression software as a MS Excel® Add-In (compiled Excel Macro). The user can work in the familiar and powerful data analysis environment of Excel and does not have to learn a new statistical software package. Other benefits from working directly in Excel are that it trivializes some of the most time consuming steps of regression analysis when compared to large conventional packages because the entire input and output of the regression lies within a standard spreadsheet workbook which eliminates the need to learn a new interface. They include:

1. setting up data input tables
2. creation, customization and printing of graphs
3. transfer of the regression analysis to other software packages (word processors, presentation software) for a final report
4. printing, saving, and recalling old results

The book and software are intuitive and guide the reader through the process of setting up a regression model and analyzing it. The software also contains an on-line help file which contains thumbnail descriptions of the significance of the output of the regression analysis and detailed instructions on how to use the software. This help file is no substitute for reading and understanding the book.

The book and software describe and implement all the tools needed for a complete linear regression analysis. Up to about 20 independent variables or regressors can be selected in a multiple regression, and second and third order models (including interactions) can easily

be set up using the built-in dialogs. In the Polynomial Regression module, up to ninth order polynomials can be constructed. There are limitations with respect to the number of data points. The accompanying software is best suited for small and intermediate data sets of 50 to several 100 data points. This is a size which most often occurs in "everyday problems" encountered by students and scientists. It was not developed to handle large data sets of several thousand and more data points used by, for example, sociologists or pharmacological researchers.

The following approach is repeated throughout the book. A theoretical discussion of a statistical technique is presented followed by chapters which explain the features of the software pertaining to the theory discussed before. The sequence in which the theory is introduced follows an order which is most likely employed by the user: introduction to regression and types of models, ANOVA, hypothesis testing, outlier analysis, and graphical evaluation including surface plots. At the end of the book, a tutorial is included with data sets (also included in the Excel spreadsheets which come with the software) which are analyzed to illustrate the utility of the software. All the analyses presented can readily be reproduced by the reader. The book starts with the usual discussion of coefficient of variation and ANOVA analysis. It contains a variety of sections on different statistical parameters and residual analyses useful for model adequacy checking. For example there are sections on stepwise regression ("auto fitting") techniques, the effect of response and factor transforming, and the detection of outlier, influence and leverage points. Although the treatment of linear regression is very complete, the book is not intended as a fundamental theoretical textbook of linear regression aimed at statisticians. It is intended to teach regression to non-statisticians by applying linear regression to real data sets.

Experimental design is covered as it relates directly to regression analysis. This restricts the design package to factors and responses that are continuous, quantitative variables. Screening designs including full and fractional (Resolution 3-5) 2 level factorial and Plackett-Burman designs are covered. Response surface modeling (RSM) designs

including face centered, circumscribed and inscribed central composite designs and Box-Behnken designs are included. The advantages and disadvantages of the various design types are covered. Advanced ideas such as aliasing, orthogonality, rotatability and sequential experimentation are explained.

The software accompanying *Essential Regression and Experimental Design for Chemists and Engineers* delivers all the tools necessary for a thorough, complete experimental design and linear regression analysis combined with easy handling and impressive output possibilities which rival the features of much more expensive and much less intuitive statistics packages.

Even as we go forward toward an electronic society, traditional publishing media (books) show no signs of being dethroned as the way to learn detailed technical concepts. However, books with illustrative examples and software that can be immediately applied do represent a vast improvement over a solely traditional approach. We believe that this "learning by doing" approach, along with a reasonably complete fundamental treatment represents an ideal way to learn new and useful technology. This is especially true for well-known and well defined concepts such as regression. We hope that you find *Essential Regression and Experimental Design for Chemists and Engineers* a good example of this new hybrid type of book.

Dave Steppan

Joachim Werner

Bob Yeater

Gibsonia, PA

Bethel Park, PA

Moundsville, WV

June 1998

Contents

1.	Regression Models, Variables, Coefficients.....	12
1.1	Theoretical Background	12
1.1.1	Introduction to Linear Regression	12
1.1.2	Transformation of Variables	13
1.1.3	Regression Model Equations	15
1.1.4	The Least-Squares Method.....	17
1.1.5	Confidence Limits for Regression Coefficients and Observations	23
1.1.6	Intercept-free Regression Models	25
1.2	Application: Regress Menu, Input Dialogs of Essential Regression	26
1.2.1	Overview	26
1.2.2	Regress Menu	27
1.2.3	Multiple and Polynomial Regression Input Dialog Boxes	30
2.	Tests for Significance of the Regression Model and Parameters	35
2.1	Theoretical Background	35
2.1.1	Introduction into Hypothesis Testing.....	35
2.1.2	Test for Significance of the Regression Model.....	36
2.1.3	Test of Significance on Individual Regression Coefficients.....	40
2.1.4	Test for Lack of Fit	42
2.2	Application: Multiple and Polynomial Regression Main Dialog (I): Model Term Selection, ANOVA, and Coefficients Table	43
2.2.1	Overview	43
2.2.2	Input Area of Main Dialog	45
2.2.3	Output Area of Main Dialog (I): ANOVA Table and Regression Coefficients Table.....	46
3.	Regression Diagnostics and Model Adequacy Checking.....	47
3.1	Theoretical Background	47
3.1.1	Overview	47
3.1.2	Coefficients of Multiple Determination for Intercept Models.....	47
3.1.3	Coefficients of Multiple Determination for No-Intercept Models	49

3.1.4 Residuals, Standardized Residuals and Outliers.....	50
3.1.5 R^2 for Prediction, Precision Index and Coefficient of Variation	55
3.1.6 Tests for Multicollinearity, Variance Inflation Factors.....	56
3.1.7 Autocorrelation.....	57
3.2 Application: Multiple and Polynomial Regression Main Dialog (II): Regression	
Summary, Residual Analysis, Outlier Analysis, and VIFs	60
3.2.1 Output Area of Main Dialog (II): Summary of Regression and VIFs.....	60
3.2.2 Outlier Button.....	62
3.2.3 Response Transformation in Essential Regression.....	62
. Graphs button.....	63
4. Model Optimization.....	64
4.1 Theoretical Background	64
4.1.1 The Problem of Finding the Best Regression Model.....	64
4.1.2 Performing All Possible Regressions and Criteria For Finding the Best Model	65
4.1.3 Stepwise Regression: Forward Selection of Variables	66
4.1.4 Stepwise Regression: Backward Elimination of Variables	67
4.1.5 Automatic Model Optimization	68
4.1.6 Transformation of the Response	69
4.2 Application: Multiple and Polynomial Regression Main Dialog (III): AutoRegress	
Area.....	69
4.2.1 Overview	69
4.2.2 Perform All Possible Regressions	70
4.2.3 Stepwise Regression in Essential Regression	70
5. Essential Regression Output	72
5.1 Graphical Evaluation of Residuals.....	72
5.2 Predicting Observations.....	76
5.3 Application: Essential Regression XLS Output Worksheet.....	77
5.3.1 Make XLS Button-Overview.....	77
5.3.2 ANOVA Table, Regression Coefficients Table, and Correlation Matrix	79
5.3.3 Tabular Output of Observations, Predictions, Residuals, and Outliers	80

5.3.4 Printed Output	82
5.3.5 Prediction of New Observations	82
5.3.6 Finding Input Variables for Given Output (Optimization Problem).....	84
5.3.7 Graphs: Scatter Plots, Confidence Limits, 3D- Plots, and Animations	86
5.3.8 Deleting or Duplicating an Output Sheet	93
5.3.9 Starting A New Regression from Output Sheet.....	94
6. Experimental Design	94
6.1 Introduction	94
6.2 Screening Designs	96
6.2.1 Two Level Full Factorial Designs	96
6.2.2 Two Level Fractional Factorial Designs.....	99
6.2.3 Using the EED Software for a Two Level Fractional Factorial Design	105
6.2.4 Plackett-Burman Designs	118
6.3 Orthogonality and Rotatability	123
6.4 Response Surface Modeling (RSM) Designs.....	125
6.4.1 Inscribed Central Composite Designs	126
6.4.2 Circumscribed Central Composite Designs	132
6.4.3 Face Centered Central Composite Designs	134
6.4.4 Box-Behnken Designs	135
6.5 Summary	137
7. Quick Guide and Tutorial	140
7.1 Important Reminder	140
7.2 Installation	141
7.3 Loading Essential Regression into MS Excel	141
7.4 Performing a Regression Analysis using the ER_Test Data	142
7.5 Unloading Essential Regression	154
7.6 Loading Essential Experimental Design into MS Excel	155
7.7 Creating a simple experimental design and analyzing it with Essential Experimental Design (EED).....	155
7.8 Unloading Essential Experimental Design	160

8.	Literature	161
9.	Index.....	163

Tables

Table 3-1: selected critical values of d as given by Durbin and Watson.....	60
Table 6-1: Full factorial experimental design for two factors	96
Table 6-2: Two level full factorial design with three factors	98
Table 6-3: Main design for a 2 level full factorial experiment for three factors	102
Table 6-4: Definition of design resolutions.....	104
Table 6-5: Output of Experiments sheet (I).....	108
Table 6-6: Output of Experiments sheet (II), design table	108
Table 6-7: Generators from the Aliasing sheet.....	109
Table 6-8: Defining Words from the Aliasing sheet	109
Table 6-9: Alias report output.....	111
Table 6-10: Simulated Design.....	113
Table 6-11: Main design for a Resolution 3 two level fractional factorial design.....	119
Table 6-12: Experiments sheet output for design in Table 6-11 (I)	119
Table 6-13: Experiments sheet output for design in Table 6-11 (II), design table.....	120
Table 6-14: Output for inscribed CCD design for two factors	126
Table 6-15: Inscribed CCD design for two factors with four centerpoints.....	128
Table 6-16: inscribed CCD design for three factors.....	129
Table 6-17: EED output for a circumscribed central composite design for 3 factors	133
Table 6-18: EED output for 3-factor-face centered CCD	135
Table 6-19: Box-Behnken design for 3 factors	136
Table 6-20: Number of factors and runs for each type of experimental design	138
Table 6-21: Possible and recommended number of centerpoints	138

Figures

Figure 1-1: Regress Menu	28
Figure 1-2: Multiple Regression Input Dialog	30
Figure 1-3: Polynomial Regression Input Dialog	31
Figure 2-1: ER Main Dialog with Model Term Selection, ANOVA, and Coefficients Table	44
Figure 3-1: Main Dialog, Output Summary and VIF table are highlighted	61
Figure 4-1: AutoRegress Area	69
Figure 4-2: Result of analysis of all possible y transformations	72
Figure 5-1: Normal Probability Plot of Rankits vs. Residuals.....	73
Figure 5-2: Plot of residuals vs. case with a possible trend (indicated by line).....	74
Figure 5-3: Patterns in residuals vs. predicted response plots and possible transformations to stabilize the variance	75
Figure 5-4: Typical plot of the predicted y and the Confidence Interval for the Mean Response at 95% significance level vs. the regressor, X.....	77
Figure 5-5: 'Worksheet Created' Message after pressing the <i>Make XLS</i> button.....	78
Figure 5-6: Buttons in XLS sheet.....	79
Figure 5-7: Print Selection Dialog.....	82
Figure 5-8: Predict Y Value Dialog.....	83
Figure 5-9: Solver Dialog after selecting the <i>Optimize</i> button	85
Figure 5-10: Solver Dialog after performing an optimization.....	86
Figure 5-11: Graph area of output sheet with 2D scatter plot of predicted vs. observed response Y including trend line and regression equation	87
Figure 5-12: 3D-graph area of output sheet with surface plot for 2-regressor models	89
Figure 5-13: 3D-graph area of output sheet with surface plot for 3-regressor model	92
Figure 5-14: "Movie" options dialog.....	93
Figure 6-1: DOE menu	105
Figure 6-2: EED Main Design Dialog	106
Figure 6-3: Factor Specification Dialog	107
Figure 6-4: Data Simulation Input Dialog	112

Figure 6-5: Input Model Coefficients Dialog.....	113
Figure 6-6: Multiple Regression Input Dialog of EED.....	114
Figure 6-7: Multiple Regression Main Dialog.....	116
Figure 6-8: Multiple Regression Main Dialog.....	117
Figure 6-9: Multiple Regression Main Dialog.....	121
Figure 6-10: EED Main Dialog.....	126
Figure 6-11: Graphical representation of inscribed CCD design for two factors.....	129
Figure 6-12: Graphical representation of inscribed CCD design for three factors	130
Figure 6-13: Graphical representation of Box-Behnken design.....	137

1. Regression Models, Variables, Coefficients

1.1 *Theoretical Background*

1.1.1 Introduction to Linear Regression

“Regression” is derived from the Latin word *regredi* meaning “to go back to”, or to “take refuge to” or “to resort to”. Actually, when we perform a “regression”, we are taking resort to approximating an observed, empirical variable (output, response) by an estimated one, based on a functional relationship between the estimated variable (we will call it y_{est}) and one or more regressor or input variables x_1, x_2, \dots, x_i . We often have to do this when we try to describe data sets, when parameters in known scientific equations have to be estimated, when we try to develop new models describing and even predicting a specific response, or when we try to control and optimize processes.

The value of the estimated variable y_{est} depends on the functional relationship with the regressors or input variables and therefore y_{est} is also called “dependent” variable. Ideally, the regressor variables do not depend on anything else than the will of the data analyst, who can chose their settings. Thus, they are called “independent” variables.

Developing this functional relationship we have to keep in mind that we cannot expect empirical data to be explained without any residual doubt. What we actually try to do is to “explain” the response with the set of the input variables as well as possible. This means, we have to account for the residual ambiguity the error contribution. Possible sources for error are random or measurement error, and the “lack-of-fit” error caused by the inaccuracies of our estimation function. Our ultimate goal in regression is to minimize this lack-of-fit error.

One can easily see that the functional relationship between y_{est} and the regressors can take many forms. The same is true for the definition of the estimation error and the way to

minimize it. These are the reasons for the numerous variants within the area of regression analysis.

Linear Regression simply means that the functional relationship between y_{est} and the regressors can be expressed by a linear equation or, in other words, a sum of terms including the error:

$$y_{est} = b_0 + b_1x_1 + b_2x_2 + \dots + b_ix_i + error \quad \text{Eq. 1-1}$$

For the case of only one regressor variable (x_1), equation (1-1) can be reduced to the familiar equation of a straight line, plus the error term, with b_0 being the intercept, and b_1 being the slope:

$$y_{est} = b_0 + b_1x_1 + error \quad \text{Eq. 1-2}$$

Equation (1-2) describes the case of “simple” Linear Regression, giving us the best fit line through data points in a x-y-plot.

1.1.2 Transformation of Variables

With more than one independent x variable, we perform *Multiple Linear Regression*, sometimes contracted to *Multiple Regression*, although the term “linear” is essential for defining the method. Let us use the letter j as an index for the independent variables running from 1 to i. In equation (1-1), b_0 is referred to as the *constant term*, meaning it gives the expected value for y with all x_j set to zero. The b_j are the regression coefficients for the respective x_j . Simply put, they describe to the magnitude of the effect of a unit change of the corresponding x_j given that the other regressors present are kept constant (the coefficient b_j for a given x_j without the other independent variables present could be different!). If x_1 changes from 1 to 2 units, and the other regressors are kept constant, y_{est} will change to $y_{est} + b_1$. To make the individual b_j independent of the scaling units of the

variables and thus comparable with respect to their magnitude, the independent variables and also the response can be transformed.

The transformation techniques commonly used consist of a *centering* of the variables which can be followed by a normalization step to transform the scale (also called *scaling of the variables*).

Centering can be done by simply subtracting the average over all data points of a given regressor variable x_j (or the response y) from the variable at the given data point. In addition to that, a division by the respective average transforms all the variables to the same scale. We will use the index k for the data points, with k running from 1 to n (the total number of data points). Following equation exemplifies these procedures for the independent variables, x_j :

$$z_{jk} = x_{jk} - \overline{x_k} \quad \text{Eq. 1-3}$$

or

$$z_{jk} = \frac{x_{jk} - \overline{x_k}}{\overline{x_k}} \quad \text{Eq. 1-4}$$

In the so-called *unit normal scaling*, for a given data point, the difference between a variable x_{jk} (y_k) and the average of the variable over all data points, $\overline{x_k}$ ($\overline{y_k}$), is divided by the sample standard deviation (s) of this variable. The scaled variables have a mean of zero and a standard deviation and variance of 1.

$$z_{jk} = \frac{x_{jk} - \overline{x_k}}{s_k} \quad \text{Eq. 1-5}$$

$$s_j^2 = \frac{\sum_{k=1}^n (x_{jk} - \overline{x_k})^2}{n-1} \quad \text{Eq. 1-6}$$

The *unit length scaling* uses the so-called corrected sum of squares instead of the standard deviation. The corrected sum of squares is simply the numerator of the expression for sample variance, i.e., the sum of the squared differences between individual variable and the average of the variable over all data points. This leads to variables with a mean of zero and a “unit length” of 1.

$$w_{jk} = \frac{x_{jk} - \overline{x_k}}{S_{jj}^2} \quad \text{Eq. 1-7}$$

$$S_{jj} = \sum_{k=1}^n (x_{jk} - \overline{x_k})^2 \quad \text{Eq. 1-8}$$

When applied not only to the independent variables, but also to the response, all these scaling techniques remove the constant term b_0 or the intercept from the model equation or, in other words, the estimate for the constant term b_0 becomes zero by definition. The new regression coefficients obtained after scaling are so-called *standardized* regression coefficients (sometimes called *betas*). Many statistical computer programs scale the model variables by default and report both betas and “raw” regression coefficients.

The main reason for using scaling techniques is to reduce the possibility of round-off errors in the calculations when using the raw variables, especially if these variables differ significantly in magnitude.

1.1.3 Regression Model Equations

It is important to realize that linear regression also includes model equations which contain “higher-order terms” (quadratic, cubic, etc.) derived from the independent variables. The functional relationship between y_{est} and the x_j has to be linear in the coefficients b_j , not necessarily linear in the x_j ! For example, the following equation (1-9) is also a linear regression model equation:

$$y_{est} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2 + error \quad \text{Eq. 1-9}$$

With

$$x_3 = x_1x_2$$

$$x_4 = x_1^2$$

$$x_5 = x_2^2$$

Equation (1-9) becomes the obviously linear equation (1-10):

$$y_{est} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + error \quad \text{Eq. 1-10}$$

A polynomial such as in equation (1-11) used to approximate our response variable also constitutes a linear regression model, since it is linear in the coefficients b_j :

$$y_{est} = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + error \quad \text{Eq. 1-11}$$

This special case is referred to as *Polynomial Regression Model*. It is just another variant of the general Linear Regression method.

Equation (1-9) represents what we refer to as a full quadratic model equation or full second order equation. It contains two linear terms or first order terms x_1 and x_2 , their squares or second order terms x_1^2 and x_2^2 , and the second order *interaction* x_1x_2 between the two linear terms. This is a very common model for responses depending on two regressors analyzed by Linear Regression. It can be modified to a second order model without the interaction, or to a model with linear terms and their interaction only. Second order or quadratic models, either complete or restricted, are frequently used for Linear Regression. Third order models (cubic terms or third order interactions, such as $x_1x_2x_3$, or $x_1^2x_2$) are less common. Generally, the higher the order of the model, the more likely the model equation simply “connects the dots” of the data points rather than fits a meaningful regression function through the data. This effect can easily be reproduced when using a software package such as MS Excel® to fit a polynomial curve through data points while consecutively increasing the order of the polynomial. Our goal in Linear Regression is to find the best of all approximated functions without just connecting the dots, i.e., without simply fitting the error.

Sometimes apparently nonlinear relationships can be transformed into model equations suitable for linear regression:

$$y = \frac{b_1 x_1}{e^{b_2 x_2}} \quad \text{Eq. 1-12}$$

becomes

$$\ln y = \ln b_1 + \ln x_1 - b_2 x_2 \quad \text{Eq. 1-13}$$

Equation (1-13) obviously can be used as a model equation for linear regression with one response ($\ln y$), a constant term ($\ln b_1$), and two regressor terms ($\ln x_1$, $b_2 x_2$). However, be aware that we have performed a nonlinear transformation. Therefore, the estimates of the coefficients may differ from a nonlinear regression using the original variables!

1.1.4 The Least-Squares Method

The method used to find the coefficients b_j of our general model equation (1-1) is called *least squares estimation*. This means that the error term we used in the model equations is defined as the difference between observed response variable y and estimated y_{est} for a given setting of the x_j at each data point. The total error must somehow be defined by summations over all data points or “cases”. Since we assume a random distribution of the individual errors with a mean of zero, a simple summation would ideally lead to zero. At least it leads to negative and positive differences canceling each other out. This can be avoided by squaring the errors for each data point and sum these squares. The desired optimum regression model then has to give us a minimum for this sum of squared errors, hence “least squares estimation method”.

A set of data consisting of n points can be considered a sample of the entire “population” of data points. A given point or “experiment” is defined by the settings of the i input factors or independent variables x_1, x_2, \dots, x_i of our model and the dependent variable or response y_k at that given experiment or “run”. Whereas the general equation (1-1) can be

considered the “population model equation”, our data set forms a system of n linear *sample equations*:

experiment, run

sample equations, Eq. 1-14a-c

1	$y_1 = b_{01} + b_1x_{11} + b_2x_{21} + \dots + b_ix_{i1} + error_1$
2	$y_2 = b_{02} + b_1x_{12} + b_2x_{22} + \dots + b_ix_{i2} + error_2$
n	$y_n = b_{0n} + b_1x_{1n} + b_2x_{2n} + \dots + b_ix_{in} + error_n$

Each of these equations can be rearranged to bring the error term on the left side. Then the square of the sums of all error terms can easily be defined:

experiment, run

sample equations for error, Eq. 1-15a-c

1	$error_1 = y_1 - b_{01} - b_1x_{11} - b_2x_{21} - \dots - b_ix_{i1}$
2	$error_2 = y_2 - b_{02} - b_1x_{12} - b_2x_{22} - \dots - b_ix_{i2}$
n	$error_n = y_n - b_{0n} - b_1x_{1n} - b_2x_{2n} - \dots - b_ix_{in}$

Again, using k as the running index for the n experiments or data points and j as the index for the i independent variables, the *Sum of Squared Errors (SSE)* is obtained by summing the squares of the right hand sides of the equations above:

$$SSE = \sum_{k=1}^n (y_k - b_0 - b_1x_{1k} - b_2x_{2k} - \dots - b_ix_{ik})^2$$

$$SSE = \sum_{k=1}^n (y_k - b_0 - \sum_{j=1}^i b_jx_{jk})^2$$

Eq. 1-16a+b

The coefficients b which meet the least squares criterion can be calculated by setting the partial derivatives of SSE with respect to each b_j (b_0 included) to zero. This is the well-known procedure for finding extrema of functions from which derivatives can be obtained.

Thus, the minimum for SSE is defined by:

$$\left. \frac{dS}{db_0} \right|_{b_0, b_1, \dots, b_i} = 0 \quad \text{and} \quad \left. \frac{dS}{db_j} \right|_{b_0, b_1, \dots, b_i} = 0 \quad \text{Eq. 1-17a+b}$$

which leads to

$$-2 \sum_{k=1}^n (y_k - b_0 - \sum_{j=1}^i b_j x_{jk}) = 0 \quad \text{Eq. 1-18}$$

and

$$-2 \sum_{k=1}^n (y_k - b_0 - \sum_{j=1}^i b_j x_{jk}) x_{jk} = 0 \quad \text{Eq. 1-19}$$

These relationships form a system of $(i+1) = p$ equations. Each equation can be rearranged with the y terms on the right hand side. We arrive at the so-called *least squares normal equations* (1-20,21). There are $i+1 = p$ of these, i for each of the coefficients b_j of the independent variables x_j (equations 1-21), and one more for the “constant” or “intercept” b_0 (equation 1-20). The total number of unknowns in our system of equations, p , is also called the number of *parameters* in the regression model equation.

$$nb_0 + b_1 \sum_{k=1}^n x_{1k} + b_2 \sum_{k=1}^n x_{2k} + \dots + b_i \sum_{k=1}^n x_{ik} = \sum_{k=1}^n y_k \quad \text{Eq. 1-20}$$

$$b_0 \sum_{k=1}^n x_{1k} + b_1 \sum_{k=1}^n x_{1k}^2 + b_2 \sum_{k=1}^n x_{1k} x_{2k} + \dots + b_i \sum_{k=1}^n x_{1k} x_{ik} = \sum_{k=1}^n x_{1k} y_k$$

.....

Eq. 1-21

$$b_0 \sum_{k=1}^n x_{ik} + b_1 \sum_{k=1}^n x_{ik} x_{1k} + b_2 \sum_{k=1}^n x_{2k} x_{1k} + \dots + b_i \sum_{k=1}^n x_{ik}^2 = \sum_{k=1}^n x_{ik} y_k$$

Basically, we are dealing with a two-dimensional problem here. Looking at our set of sample equations above, we have n equations with the $p=(i + 1)$ unknown parameters in each one from our model equation (dimensions $n \times p$). After transformation to the normal equations we obtain a system of p equations with p terms in each one ($p \times p$). Thus, it is not surprising that the most elegant and convenient way to solve the problem of finding the set of regression coefficients b which gives a minimum for SSE entails matrix algorithms. We are not going to go through this procedure in every gory detail. Other books do that, and they are written by real mathematicians (see recommendations in the Literature section of this book). Suffice it to say that the starting point of the calculations is the matrix notation (1-22) for the system of sample equations (bold small letters or words denote vectors, bold capital letters symbolize matrices!):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{error} \quad \text{Eq. 1-22}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1i} \\ 1 & x_{21} & x_{22} & \dots & x_{2i} \\ \dots & \dots & \dots & \dots & x_{3i} \\ 1 & x_{n1} & x_{n2} & \dots & x_{ni} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_i \end{bmatrix} + \begin{bmatrix} error_1 \\ error_2 \\ \dots \\ error_n \end{bmatrix} \quad \text{Eq. 1-23}$$

The n dependent variables or outputs become the $n \times 1$ vector \mathbf{y} , the p parameters (independent variables plus one constant or intercept) are represented by the product of the $n \times p$ matrix \mathbf{X} and the $n \times 1$ vector \mathbf{b} , and the n error terms of the n observations or runs form the $n \times 1$ vector \mathbf{error} .

The solution of the least squares problem can be obtained through a series of matrix transformation. We will give the final steps. The least squares criterion of $SSE=0$ leads to:

$$\mathbf{X'Xb} = \mathbf{X'y} \quad \text{Eq. 1-24}$$

Equation 1-24 is simply the matrix representation of the normal equations (1-20,21).

Finally, the vector of the estimated coefficients **b** is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{Eq. 1-25}$$

\mathbf{X}' denotes the transpose of \mathbf{X} , where the indices for the dimensions n and p are exchanged (rows become columns and vice versa). \mathbf{X}^{-1} is the **inverse matrix** of \mathbf{X} , which means the matrix product $\mathbf{X}\mathbf{X}^{-1}$ becomes the unitary matrix \mathbf{E} with diagonal elements of value 1 and non-diagonal values of zero.

Consequently, finding the regression coefficients b_j which meet the least squares criterion boils down to a series of matrix and vector transformations and multiplications. This is a task which is ideally suited for computers, and fortunately, we do not have to worry anymore about having to perform this tedious work manually. However, one caveat of equation (1-25) is the fact that the matrix product $(\mathbf{X}'\mathbf{X})^{-1}$, a quadratic matrix with the dimensions $p \times p$ (again, p = number of independent variables plus the constant term or number of **p**arameters), has to be calculated. This matrix, however, sometimes cannot be calculated if there is a high degree of collinearity between the columns of the matrix of the independent variables \mathbf{X} , i.e., if the regressors are not linearly independent. This can happen, for instance, when performing Polynomial Regression, where the regressor variables are different orders of one input and thus are strongly correlated. In computer programs, this can lead to error messages such as “division by zero”. This is one of the numerous problems caused by *multicollinearity* (see also Chapter 3.1.6).

Using matrix notation, the vector of the fitted or predicted responses, \mathbf{y}_{est} , can be calculated by

$$\mathbf{y}_{\text{est}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{Eq. 1-26}$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ has the dimensions $n \times n$ and is called the *hat matrix*. It plays an important role in regression analysis, especially regarding model adequacy checking (Chapters 3 and 5).

It is important to realize that the regression coefficients b_j calculated by the least-squares method are estimated parameters. According to the *Gauss-Markov-Theorem*, they are the linear unbiased estimators with the least error variance compared to all other unbiased estimators. This error variance can also be estimated and depends on the Sum of Squared Errors (SSE), introduced above, and the *degrees of freedom* of the regression model and of the residual error, respectively.

Assuming a data set with n data points, the total number of degrees of freedom in Linear Regression is $(n-1)$. The number of degrees of freedom occupied by the regression model is equal to the number of regression coefficients associated with regressors, i.e., the number of coefficients minus the intercept or constant term b_0 . In the equations above, we used i to denote this number. The number of degrees of freedom left for the error calculations is the total number minus the number occupied by the model, i.e., $(n-1)-i$. One can see that for the Linear Regression model with one constant term, there is another way for defining the error degrees of freedom: since the total number of terms in the model including the intercept is just $p = i + 1$, the error degrees of freedom (f_{error}) can be defined as:

$$f_{\text{error}} = (n - p) = \text{number of data points} - \text{number of model terms (including intercept)}$$

The Sum of Squared Errors (SSE) divided by the error degrees of freedom gives the so-called *Mean Squared Error (MSE)*:

$$MSE = \frac{SSE}{(n - p)} \quad \text{Eq. 1-27}$$

The Mean Squared Error (MSE) is equal to the unbiased estimator of the error variance, σ^2 . This is a model-dependent estimator of the error variance. Its value depends on the regression model used in the least-squares calculations. The square root of the model-dependent error variance MSE used as an estimator for the model-dependent “standard

deviation” and is called the *Standard Error* of the regression model. This should not be confused with the Standard Errors for the individual regression coefficients or observations as described in the following chapter.

1.1.5 Confidence Limits for Regression Coefficients and Observations

In linear regression, we assume that the error terms are uncorrelated random variables, they do not depend on each other (if they depend on each other, we call this *autocorrelation*, see Chapter 3.1.7.), and that they follow a normal distribution. Plotting the error distribution should ideally give a bell-shaped curve with a mean of zero and a standard deviation σ . It follows from the model equation (1-1) that the responses also have to be random variables. Due to the random error and lack of fit, there exists a probability distribution for a given y_k at each possible setting of the x_{kj} . Therefore, when reporting results of regression analyses, the estimates of the expected errors and confidence limits are essential. They determine the range where we can find the actual response with a certain probability. Actually, the expected value $y_{k(est)}$ is the mean of a distribution for a given setting or data point k .

The confidence limit or interval (CI) depends on the *confidence level* α or the probability, that the “actual” response can be found in the given confidence range. A t-distribution with the $(n-p)$ error degrees of freedom is used to estimate these confidence regions. In most cases, the confidence intervals are calculated at the 95% probability level. A higher probability leads to wider confidence ranges and vice versa.

Let us define the i regressors plus intercept for this setting as a row vector $\mathbf{x}_k' = [1, x_{1k}, x_{2k}, \dots, x_{ik}]$. The confidence interval CI at 95% probability level around the expected mean is defined by:

$$CI_{y_{k(est)}} = \pm t_{(95, n-p)} \sqrt{MSE \mathbf{x}_k' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_k} \quad \text{Eq. 1-28}$$

The expression under the square root is also referred to as *Standard Error for Mean Response*. Note that both Standard Error and CI depend on the location of the data point in x space!

If we want to predict new responses based on settings which do not occur in our data, we have to use a wider confidence range to reflect the increased uncertainty:

$$CI_{y_{new(est)}} = \pm t_{(95, n-p)} \sqrt{MSE(1 + \mathbf{x}_{new}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{new})} \quad \text{Eq. 1-29}$$

In this case, the square root term is also called *Standard Error for Prediction*. Again, the value for CI in equation (1-29) depends on the settings for the x_j of the new data point! These confidence ranges will be discussed again in Chapter 5 in connection with the prediction module of Essential Regression.

By the same token, we can define a *confidence limit* or *confidence range* around the estimated regression coefficients b_j which depends on the *confidence level* or the probability, that the “actual” regression parameter can be found in the given confidence range. At the 95% level, the equation for the confidence limit of a given b_j is:

$$CI_{b_j} = \pm t_{95, n-p} \sqrt{MSE[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}} \quad \text{Eq. 1-30}$$

The term in square brackets denotes the j th diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix we used above in the calculations of the least squares estimators (Equation 1-25). The square root term is also called the *Standard Error* of the regression coefficient b_j .

1.1.6 Intercept-free Regression Models

In the beginning of this introduction, we defined equation (1-1) as the basic model equation for our derivation of the least-squares method. This equation contains a coefficient b_0 symbolizing a constant term. This constant is also called *intercept* because, when using only one independent variable x , b_0 gives the value of y at $x = 0$, i.e., the point

where a graph of $y = b_0 + b_1 x$ intercepts the y-axis. However, we can also define a model equation with b_0 set to zero by definition. We arrive at what is referred to as a *no-intercept model* or *intercept-free model*. For the matrix equation (1-25) used to calculate the regression coefficients, this means that b_0 in the coefficient vector \mathbf{b} is zero by definition.

In situations where the dependent variable or the response can only be zero when all the independent variables are zero, intercept-free models appear appropriate. This occurs most often when analyzing physical or chemical relationships. Without this background information implying an intercept of zero, however, both intercept and no-intercept models have to be evaluated carefully. Sometimes, a scatter diagram of the data seems to indicate that the graph can be extended through the origin. However, if the available data are remote from the origin, such an extrapolation can lead to erroneous conclusions. If both an intercept and no-intercept model are possible, the Mean Squared Error (MSE) is a good basis for comparison. The smaller MSE indicates the better model.

1.2 *Application: Regress Menu, Input Dialogs of Essential Regression*

1.2.1 Overview

The contents of the previous chapter give the theoretical background for the following description of the Regress Menu and the Input Dialogs of Essential Regression (ER). Most of the topics covered theoretically in the previous chapter will be practically applied here and guide through the first steps of arriving at a regression model using ER. This concept of presenting the underlying theory first followed by the practical application within ER will be continued throughout this book. Admittedly, the resulting order of introduction of the different theoretical aspects of Linear Regression is sometimes in contrast to the didactic approach used in most textbooks on Linear Regression. However, by following the logical sequence of the dialogs and menus of ER, we intend to facilitate the use of ER, especially for first time users, and simultaneously give a theoretical background of Linear Regression.

We assume the user has an Excel worksheet open with the data to be evaluated in tabular form, i.e., the cases or data points are arranged in rows, the variables in columns. We also assume the top row contains header information for each column (variable names). The cursor should highlight the leftmost cell in the header row of the table. We refer to this cell as the *pivot cell*.

There are certain limitations to the variable and sheet names used. They should not contain hyphens, slashes, plus and minus signs or similar characters which could be misread for mathematical expressions. These characters could cause errors which would lead to a termination of the program.

After loading ER, an additional menu option, *Regress*, becomes available in the Main Menu of MS Excel. When selecting either the *Multiple Regression* or *Polynomial Regression* option, ER reads the header information of the data table from left to right and displays the column headers as possible variables (regressors and responses) in the corresponding list boxes of the *Multiple* or *Polynomial Regression Input Dialogs*. In these dialogs, the user can select the desired independent variables (factors, regressors), the dependent variable (response), and the type or the order of the regression model. For Multiple Regression, ER offers linear, quadratic, and cubic models, and second and third-order models with and without interaction. For Polynomial Regression, the user can choose from first to ninth-order polynomials in one variable. In addition, the input dialogs allow the user to transform the independent variables (regressors) either by centering or standardization (scaling). Furthermore, the user can choose between an *intercept* or *non-intercept model*. Also, the probability level of the confidence intervals for the regression coefficients can be adjusted here.

1.2.2 Regress Menu

This Menu is accessible in the MS Excel Main Menu once ER is loaded as an Add-in.



Figure 1-1: Regress Menu

Multiple Regression

This option opens the *Multiple Regression Input* dialog box. Essential Regression performs a Multiple Linear Regression based on the least squares method.

Polynomial Regression

This option opens the *Polynomial Regression Input* dialog box. Essential Regression performs a Polynomial Regression which uses a linear or higher order polynomial of *one* variable (predictor, regressor, independent variable, x) to describe the response (dependent variable, y).

Analyze Design and Simulate Data

These two menu options are used in connection with Essential Experimental Design (EED) described in detail in Chapter 6. To use the Analyze Design option, a worksheet created in EED has to be the starting point. However, the Simulate Data menu item can also be used in Essential Regression (ER) to create a simulated data set based on given

input variables and predefined regression coefficients (see Chapter 6 and the Quick Guide for more details).

Relink Buttons

This menu option relinks the buttons of the XLS output sheet created with Essential Regression to the Essential Regression Add-In. This is sometimes necessary if the worksheet buttons don't work despite the fact that the Add-In is loaded in memory. This will happen if ER is moved from the directory it was in when the worksheet was created. See Chapter 5 for more details about the XLS output sheet.

Duplicate Regression

Activating this menu option generates copies of the current XLS output worksheet (see Chapter 5) generated by Essential Regression. A XLS output sheet must be the active sheet.

Help

Opens the Essential Regression on-line help.

Unload

Removes the Regress Menu from the Excel Main Menu and unloads the Essential Regression Add-In file. If the Add-In is loaded, but the *Regress* menu is not visible, it can be reactivated by using the {Ctrl+M} key combination.

About

Gives information about the current version of Essential Regression and the system it is installed on.

1.2.3 Multiple and Polynomial Regression Input Dialog Boxes

The input dialog boxes appear when the user activates either the *Multiple Regression* or *Polynomial Regression* option in the *Regress* menu

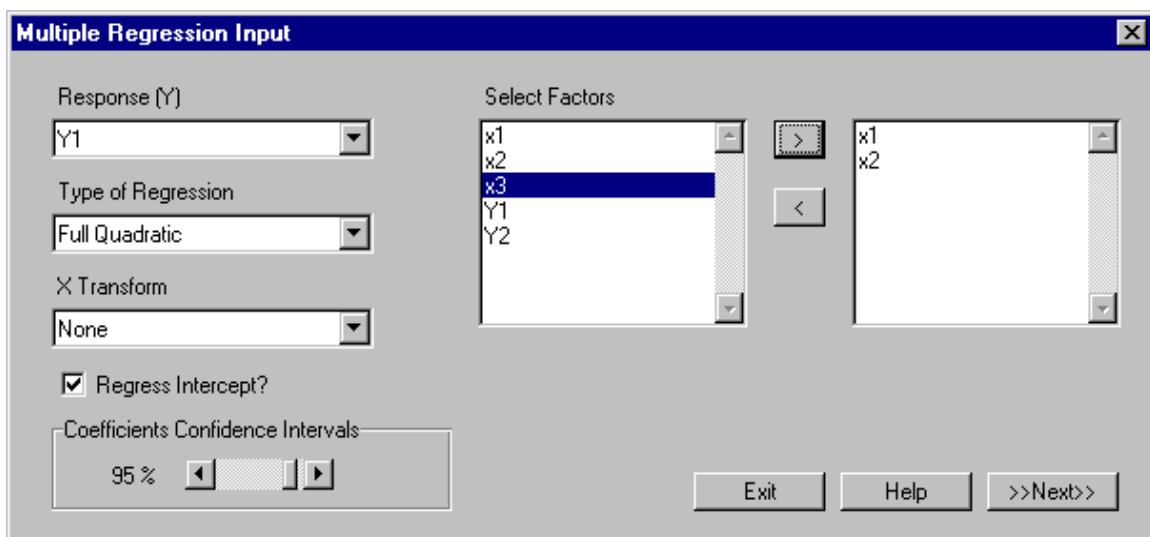


Figure 1-2: Multiple Regression Input Dialog

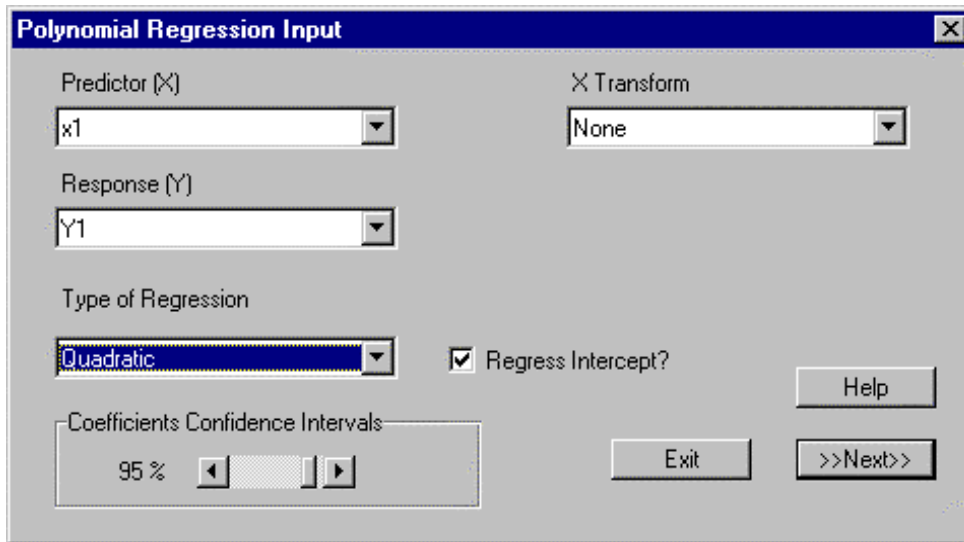


Figure 1-3: Polynomial Regression Input Dialog

Predictor(X)

(Polynomial Regression)

Choose one predictor or regressor variable x for a polynomial model.

Select Factors

(Multiple Regression)

Choose independent factors or variables (regressors) x_j for the regression model. Up to nine independent factors can be selected.

Type of Regression

(Polynomial Regression)

Specify the order of the regression model. ER allows the user to specify linear (1st order regression, "simple" linear regression with one predictor), quadratic, cubic etc., up to to 9th order polynomials of the regressor or independent variable x .

Type of Regression

(Multiple Regression)

Specify the order and type of the regression model. "Full Quadratic" and "Full Cubic" contain all higher order terms including interactions. Subsets of these models are "Interaction" (no quadratic terms, but linear-linear interactions), "Squared Interaction" (no cubic terms, but squared-linear interactions), "2nd order, no interaction" (only linear and quadratic terms), and "3rd order, no interaction" (linear, quadratic, and cubic terms without interactions).

Response(Y)

Choose the response or dependent variable Y

Regress Intercept?

Checked

A constant parameter which is independent of the settings for the x_j is used in the regression model. A so-called *intercept-model* is used as described in the previous chapter.

Unchecked

Specifies a *non-intercept regression model* or a “regression through the origin”. There is no constant parameter in the regression model. See remarks at the end of the previous chapter regarding the utility of non-intercept-models.

Centering or standardizing the response *together* with all the independent variables (see *X Transform* below) creates a non-intercept model by definition, and ER will produce a constant term of zero even when the *Regress Intercept* box is checked.

X Transform

This list box gives several options for transformations of the predictor (regressor) variable(s) x_j . Transformations of the x_j can be performed both in the Multiple and Polynomial Regression mode.

None

No transformation; regression model will be based on raw data.

Center All Terms

Centers both linear and higher order terms of the predictor (regressor) variable(s) x_j after calculating the higher order terms from the linear terms. Centering is done by dividing the difference between the maximum and the average of a given x_j by the average.

Standardize All Terms

Standardizes both linear and higher order terms of the predictor (regressor) variable(s) x_j after calculating the higher order terms from the linear terms. Standardization is done by dividing the difference between the maximum and the average of a given x_j by the standard deviation. This corresponds to the *unit normal scaling* method described in the previous chapter. The resulting transformed variables have a mean of zero and a standard deviation of 1.

Center Linear Terms

Centers only linear terms of the predictor (regressor) variable(s) x_j and calculates the higher order terms from the centered linear terms. Centering is done by dividing the difference between the maximum and the average of a given x_j by the average.

Standardize Linear Terms

Standardizes only linear terms of the predictor (regressor) variable(s) x_j and calculates the higher order terms from the standardized linear terms. Standardization is done by

dividing the difference between the maximum and the average of a given x_j by the standard deviation. This corresponds to the *unit normal scaling* method described in the previous chapter. The resulting transformed variables have a mean of zero and a standard deviation of 1.

It is important to emphasize that only the regressor or independent variables x_j are transformed after selecting one of the transformation options in the *X Transform* option! The response can be independently transformed using the *Y Trans* option which is part of the *Main Dialog* discussed in the next chapter. Centering or standardizing the response *together* with the all independent variables using the *All Terms* options creates a non-intercept model, and ER will produce a constant term of zero even when the *Regress Intercept* box is checked.

As discussed in the previous section of this chapter, centering or scaling the variables can be helpful when higher-order terms or polynomials are used in the regression model or if the variables differ significantly in magnitude. These conditions can lead to *ill-conditioning* of the matrix of the independent variables. This means that the matrix inversion used for the calculation of the regression coefficients can become inaccurate and significant error is introduced in the estimation of the coefficients.

Coefficients Confidence Intervals

Specifies the probability or significance level of the confidence intervals of regression coefficients and predicted responses. By default, it is set to a probability level of 95%. Increasing this number leads to wider confidence limits and vice versa. Other than that, it has no effect on the regression model.

Next

Opens the *Polynomial or Multiple Regression Main Dialog*

2. Tests for Significance of the Regression Model and Parameters

2.1 Theoretical Background

2.1.1 Introduction into Hypothesis Testing

In the previous chapter, we have shown that Linear Regression allows us to estimate a response variable depending on the values or settings of one or more independent variables. By applying the so-called least-squares technique, we can fit a model equation containing one or more independent variables by minimizing the residual error measured by the sum of squared deviations between the actual and the estimated responses. We do so by calculating estimates for the regression coefficients, i.e., the coefficients of the model variables including the intercept or constant term. However, this does not tell us if the calculated coefficients or the model equation actually have a statistical significance. In other words, does the linear relationship we defined when setting up the model equation have any meaning when compared to the error in the data? Does an individual regression coefficient for a given variable have any significance or could we drop it from the model without sacrificing the quality of the result? These questions are behind the *tests for significance of the regression model and the individual regression coefficients*.

In these situations, statisticians tend to define so-called *null hypotheses*. In order to test the significance of the model, they assume the worst case scenario by saying: “The null hypothesis is true if there is no linear relationship between any of the independent variables”. This is equivalent to the equations:

$$H_0: b_1 = b_2 = \dots b_i = 0 \quad \text{Eq. 2-1}$$

$$H_1: b_j \neq 0 \text{ for at least one } j \quad \text{Eq. 2-2}$$

with H_0 denoting the null hypothesis, H_1 being the rejection of the null hypothesis, and $b_1 \dots b_i$ representing the intercept and the regression coefficients of the i independent

variables in our model equation (1-1). *If H_0 is rejected, there is at least one independent variable significantly contributing to the linear model, and we can conclude that there exists a functional relationship between the response and at least one of the variables.*

Similarly, the hypotheses for the individual coefficients b_j can be defined:

$$\mathbf{H}_0: \mathbf{b}_j = 0 \quad \text{Eq. 2-3}$$

$$\mathbf{H}_1: \mathbf{b}_j \neq 0 \quad \text{Eq. 2-4}$$

If H_0 is rejected, the respective coefficient significantly contributes to the model. If H_0 cannot be rejected, the corresponding variable can be eliminated from the model equation.

2.1.2 Test for Significance of the Regression Model

The null hypothesis for the regression model (eq. 2-1a) is simply tested by comparing the effect or variability caused by the regression model to the overall error. This comparison is based on the so-called *Total Sum of Squares* (S_{yy}), the *Regression Sum of Squares* (SSR), and the *Sum of Squared Errors* or *Error Sum of Squares* (SSE).

In Linear Regression, we define the total variability in the n observations as the sum of the squared differences between a the responses y_i ($k=1 \dots n$) and the average of all responses, \bar{y} . This is also called the *Total Sum of Squares*, S_{yy} .

$$S_{yy} = \sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n y_k^2 - \frac{\left(\sum_{k=1}^n y_k \right)^2}{n} \quad \text{Eq. 2-5}$$

By the same token, the *Regression Sum of Squares*, SSR , which gives the variability in the response y explained by the model equation, is defined the sum of the squared differences between a the estimated responses $y_{i(\text{est})}$ ($k=1 \dots n$) and the average of all responses, \bar{y} :

$$SSR = \sum_{k=1}^n (y_{k(est)} - \bar{y})^2 = \sum_{k=1}^n y_{k(est)}^2 - \frac{\left(\sum_{k=1}^n y_k\right)^2}{n} \quad \text{Eq. 2-6}$$

We already defined the so-called *Error Sum of Squares*, SSE, in the previous chapter. It is the sum of the squared *residuals*, or, in other words, the sum of the squared differences between the observed responses, y_k , and the predicted or estimated responses based on the model, $y_{k(est)}$.

$$SSE = \sum_{k=1}^n (y_k - y_{k(est)})^2 \quad \text{Eq. 2-7}$$

The total variability in the observations is the sum of the variability or effect caused by the regression model, SSR and the error contribution. So, instead of using equation (2-6), the effect or variability caused by the regression model can be found by calculating the difference between the total variability or Total Sum of Squares and the Residual or Error Sum of Squares:

$$S_{yy} = SSR + SSE \quad \text{Eq. 2-8}$$

$$SSR = S_{yy} - SSE \quad \text{Eq. 2-9}$$

Associated with S_{yy} , the Total Sum of Squares, are $n-1$ degrees of freedom, with n being the number of data points in the regression. One degree of freedom has been “lost” or used up by the constraint that the sum of all the differences ($y_k - \bar{y}$) is zero.

The number of degrees of freedom for the model, associated with the Regression Sum of Squares, SSR, equals the number of coefficients b_j , without the constant term, b_0 . This is equal to the number i of independent variables, or model terms without the constant, if present.

The residual or error degrees of freedom are found by subtracting the degrees of freedom for the model from the degrees of freedom for the Total Sum of Squares. This is

equivalent to the difference between the number of data points, n , and the number of terms in the model including the constant, p .

$$\begin{aligned} f_{S_{yy}} &= n - 1 \\ f_{SSR} &= i \\ p &= i + 1 \\ f_{SSE} &= (n - 1) - i = n - (i + 1) = n - p \end{aligned} \quad \text{Eq. 2-10a-d}$$

The test for the significance of the regression model is performed as an *analysis-of-variance* procedure by calculating the ratio between the Regression Sum of Squares (SSR) and the Error Sum of Squares (SSE) and comparing the result to the F-statistic with the appropriate degrees of freedom at a given significance level.

$$F_0 = \frac{SSR / f_{SSR}}{SSE / f_{SSE}} = \frac{SSR / i}{SSE / (n - 1 - i)} = \frac{MSR}{MSE} \quad \text{Eq. 2-11}$$

From the previous chapter, we already know that the division of SSE by the error degrees of freedom gives the Mean Squared Error, MSE. By the same token, the term SSR/ i is called the *Regression Mean Square*, MSR.

The null hypothesis H_0 is rejected if F_0 is greater than the corresponding critical value F_{crit} of the F-distribution for a given significance level with i and $(n-1-i)$ degrees of freedom. In other words, for a significance level α , *the hypothesis that the regression model is not significant* can be rejected at the α -level if $F_0 > F_{\text{crit}} = F_{\alpha, i, n-1-i}$. Note that the significance level α stands for the probability that the null hypothesis is true, i.e., the model is not significant. Usually, significance levels α of 0.10, 0.05, and 0.01 are used to determine critical values F_{crit} , where *decreasing significance levels indicate a higher confidence for the model*. The values F_{crit} for the F distribution increase with decreasing significance level α and increasing degrees of freedom f_{SSR} for the regression model, and they decrease with increasing degrees of freedom f_{SSE} for the error contribution. For a given model, the larger the value of MSR/MSE, the lower the significance level α leading to critical values for F_{crit} which are smaller than F_0 , and the higher the confidence level for the *significance of the model*, i.e. a rejection of H_0 . On the other hand, increasing the number of model terms for a given data set, i.e., increasing f_{SSR} and decreasing f_{SSE} , can lead to a decrease of MSR

and an increase of MSE up to a point where the F_0 becomes smaller than F_{crit} and the model is no longer significant. If this occurs at significance levels α of higher than 0.1, the model is considered to be no longer significant.

An **analysis-of-variance-** or ANOVA-table such as Table 2-1 is commonly used to summarize the test for significance of the model which we just described. There are variations in the layout of this table. In computer programs, usually the significance level α is calculated and given in addition to the corresponding value of $F_0 = \text{MSR}/\text{MSE}$, so we do not have to look up the values for F_{crit} in a table anymore. For example, if the computed value for α is .076, then the model is significant at the 0.1 level, but not significant at the 0.05 level. Again, smaller values for the computed significance levels (also called *error probabilities*) indicate more significant, i.e., “better” models!

Table 2-1: ANOVA table for a model with i regressor variables and n observations.

<i>Source of Variation</i>	<i>Sum of Squares</i>	<i>Degrees of Freedom</i>	<i>Mean Square</i>	F_0	<i>Significance α or Error Probability P</i>
Regression Model	SSR	i	$\text{MSR} = \text{SSR}/i$	MSR/MSE	$= P(H_0: F_0 \leq F_{\text{crit}})$
Residual (Error)	SSE	$n-1-i$	MSE $= \text{SSE}/(n-1-i)$		
Total	S_{yy}	$n-1$			

2.1.3 Test of Significance on Individual Regression Coefficients

The significance test on the regression model tells us if at least one of the regression coefficients is different from zero. We have to perform another test to be able to assess the significance of the individual coefficients. This test forms the basis for model optimization by adding or deleting coefficients (see *Backward Elimination*, *Forward Selection*, and *Autofitting* in Chapter 4). A model with many coefficients is not necessarily

the best, and a model with only a few coefficients might improve dramatically by adding another, but we have to know which coefficient actually plays a significant role in the model.

The underlying null hypothesis was described above. A t-test statistic is used to test this hypothesis:

$$t_0 = \frac{b_j}{\sqrt{MSE * C_{jj}}} \quad \text{Eq. 2-12}$$

C_{jj} is the diagonal element pertaining to the coefficient b_j of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ which we introduced in the description of the least squares method in chapter 1 (Eq. 1-24). Note that the square root term equals the so-called standard error of the individual regression coefficient b_j .

Similar to the F-test used for checking the model significance, we compare the calculated t_0 to the critical t-value t_{crit} for a given significance level α and the error degrees of freedom, $n-1-i$. Note that there can be differences in the tables of the t-distribution given in the literature depending on the definition of α . In most tables, the t-distribution is given for the so-called *two-sided* or *two-tailed* significance level. In this case, the critical value we look for is $t_{\alpha, n-1-i}$. This means, the error probability or significance on each side of the two-tailed t-distribution is defined as $\alpha/2$. For a one-sided t-test, the fraction under the positive or negative tail of the distribution is defined as α . If the table lists the one-sided α levels, we have to look for $t_{\alpha/2, n-1-i}$. The built-in t-distribution of MS Excel® uses the first notation. For instance, $t_{0.05,1}$ can be calculated by entering the worksheet function “=TINV(.05,1)” and results in the value 12.706. Certain books, however, list one-sided significance levels, where $t_{0.05,1}$ is listed as 6.314, and $t_{0.025,1}$ gives 12.706.

If the calculated value for t_0 is larger than t_{crit} , we reject the null hypothesis at the given significance level. For instance, with $\alpha=0.05$, we would say that there is only a 5% error probability that the corresponding coefficient is not significant. Note that this significance

is based on the presence of all the other regressor variables in the model. It might change dramatically with a different set of regressor variables.

The results of the significance tests on the coefficients are usually listed in a table such as Table 2-2. In the P-value column, “*tin*v” denotes the probability or α -level for the calculated t_0 -value. CI stands for confidence interval. The expression for the confidence intervals of the coefficients b_j was already introduced in Chapter 1.1.5.

Table 2-2: Parameter Table for a model with i regressor variables (or $p = i+1$ parameters) and n observations.

<i>Variable</i>	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Statistic</i> (t_0)	<i>P-value or</i> <i>a for t_0</i>	<i>Lower</i> <i>CI</i>	<i>Upper</i> <i>CI</i>
Intercept	b_0					
X_j	b_j	$\sqrt{MSE[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$	$\frac{b_j}{\sqrt{MSE[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}}$	$= \text{tin}v(a, (n-1-i))$	$b_j - t_{a, n-p} \sqrt{MSE[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$	$b_j + t_{a, n-p} \sqrt{MSE[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$

2.1.4 Test for Lack of Fit

If replicate measurements are present, i.e., responses based on the same settings for the independent variables, a test can be performed which gives the significance of the replicate error in comparison to the model dependent error. In other words, the test splits the Residual or Error Sum of Squares, SSE, into a contribution from the *pure error*, which is based on the replicate measurements, and a fraction which is due to the *lack of fit* based on the model performance. Let us assume we have m data points based on different settings for the independent variables, and r_k replicates for a given observation y_k . The total number of data points is then:

$$n = \sum_{k=1}^m r_k$$

The so-called *Pure Error Sum of Squares*, *SSPE*, is obtained from the summation of the squared differences between the r replicate measurements and their average for each setting and then summing over all the m different settings.

$$SSPE = \sum_{k=1}^m \sum_{j=1}^r (y_{kj} - \bar{y}_k)^2 \quad \text{Eq. 2-13}$$

SSPE is associated with $(n-m)$ degrees of freedom. The *Sum of Squares for Lack of Fit* can be obtained by subtracting *SSPE* from *SSE*. It is associated with $m-2$ degrees of freedom. Similar to the F-test for significance of the model, the test statistic for lack of fit is given by

$$F_0 = \frac{SSLOF / (m-2)}{SSPE / (n-m)} = \frac{MSLOF}{MSPE} \quad \text{Eq. 2-14}$$

If F_0 is larger than the critical value F_{crit} for a given significance level α with $m-2$ and $n-m$ degrees of freedom, the lack of fit error is significant, i.e., there might be contributions in the regressor-response relationship not accounted for by the model. When performed on a linear (first order) model, this test indicates curvature if F_0 is significant.

2.2 *Application: Multiple and Polynomial Regression Main Dialog (I): Model Term Selection, ANOVA, and Coefficients Table*

2.2.1 Overview

In chapter 1.2, we described how to select Polynomial vs. Multiple Regression, how to pick input and response variables, how to define the order of the regression model (linear, quadratic etc.) and, finally, how to specify a model with or without intercept. After completing these steps and clicking Next in the Polynomial or Multiple Regression Input Dialogs, ER will bring up the Main Dialog (see Figure 2-1).

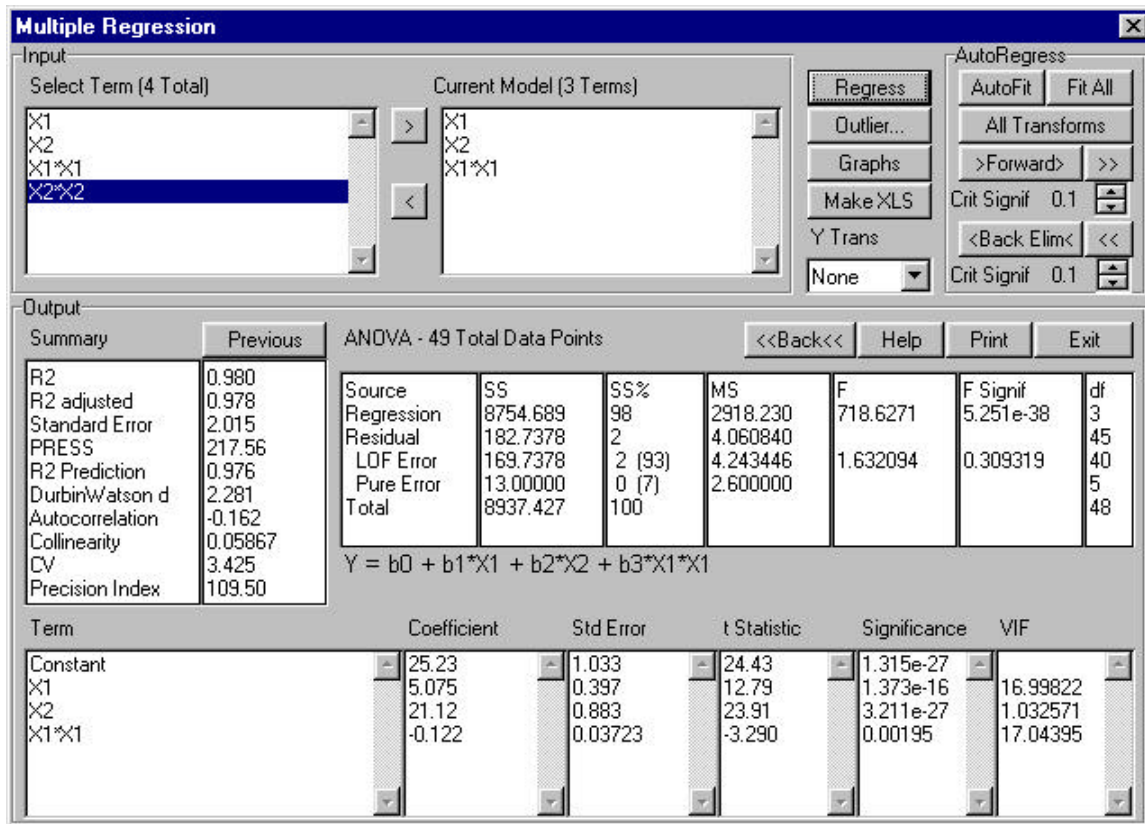


Figure 2-1: ER Main Dialog with Model Term Selection, ANOVA, and Coefficients Table

The Main Dialog is divided into several sections. In this chapter, we will describe the Input area at the top and the ANOVA and Regression Coefficient Tables, which are part of the Output Area in the bottom half of the dialog box. The remaining sections deal with “Model Adequacy Checking”, including Outlier Analysis, and automatic fitting by using forward selection and backward elimination. These sections will be discussed later in this book when we talk about how to determine how to arrive at the best possible regression model and how to assess its reliability. For now, let’s assume that we have picked our variables and just want to know what the model parameters look like and if the model is significant. Like in Chapter 1.2., we will go through the features of the Main Dialog in the chronological order most likely employed by the user.

2.2.2 Input Area of Main Dialog

Select Term Window

This window contains the terms which can be derived from the independent variables picked in the Input dialog (i.e., linear, higher order, and interaction terms, according to the order of regression selected by Type of Regression in the Input Dialog Box). By selecting terms and clicking the ">" button, terms can be selected for the regression model and will appear in the Initial Model window. By clicking the ">" button without selection, terms are transferred into the Initial Model window in order of their appearance in the Select Term window. Note that any subset of the model terms which are part of the full model can be selected.

Initial Model Window

This window contains all the terms which were selected from the Select Term window by using the ">" button. Terms can be eliminated from the model by selecting them in the Initial Model window and using the "<" button. Clicking this button without any terms selected removes the terms from the window in order of their appearance.

Regress Button

Starts the Linear Regression using the model terms in the Initial Model window.

2.2.3 Output Area of Main Dialog (I): ANOVA Table and Regression Coefficients Table

ANOVA

Output window for the *Analysis of Variance* table. This table gives the Regression Sum of Squares (SS) for the regression model (SSR), the Residual Sum of Squares for the error contribution (SSE), and the Total Sum of Squares for the overall variability. In addition, a percent contribution for the model and the error is calculated. The Mean Square for the regression model (MSR) and error (MSE, see Chapter 1.1 and 2.1 for a definition of these terms) are calculated by dividing the Sum of Squares by the respective degrees of freedom for each term. The MSE is used to calculate the F-statistic ($= \text{MSR}/\text{MSE}$). The resulting model significance is shown in the ANOVA table (F Signif). As described in the previous chapter 2.1., low values of F Signif indicate a high model significance. A value of .05 indicates a significant model at the 95% significance level. A lack-of-fit test (LOF test) is performed when replicates are present in the data set and the resulting Pure Error (SSPE) and Lack-of-Fit (SSLOF) contributions to the Regression Sum of Squares (SSR) as well as a F statistic for the significance of the LOF are shown in this case. Once a regression is performed, the current model equation is shown below the ANOVA table.

Regression Coefficients Window

This is the window at the bottom of the output area of the Main Dialog. It shows the terms (variables) used in the model and their regression coefficients with standard errors, t-statistic, and significance level for each coefficient. Another column shows the variance inflation factors (VIFs). We are going to talk about these in detail in Chapter 3. Each list box showing a specific parameter list can be scrolled independently. However, when a specific row is highlighted in a list box, the corresponding rows in the remaining list boxes will be highlighted automatically. This setup allows for more flexibility when dealing with parameter lists which are longer than the window area.

3. Regression Diagnostics and Model Adequacy Checking

3.1 Theoretical Background

3.1.1 Overview

The previous Chapter described how we can determine if a regression model and the individual regression coefficients are significant at the level we deem appropriate (90% or .1, 95% or .05 etc.). However, we still have to figure out if this model actually describes our data adequately. How good is the “fit” of the predicted data compared to the “real data”? Are there differences between fitted and experimental data which are larger or smaller than expected? How “good” are the “real data”, anyway?. Are there data points which we have to discard because they might be erroneous? Are there data points which might have an unusually strong influence on the regression results ? These questions are behind what is commonly called *Model Adequacy Checking*.

3.1.2 Coefficients of Multiple Determination for Intercept Models

Everybody who has ever performed a simple Linear Regression, maybe only as a straight line fit, knows that there are parameters called R (correlation coefficient) or R^2 which somehow describe the quality of the fit. Most people consider these parameters as most important in assessing the quality of a regression model. This chapter will show that we have to be very careful in relying exclusively on these parameters when evaluating a regression model.

R^2 , the *coefficient of determination* in Simple Linear Regression is called *coefficient of multiple determination* in Multiple Linear Regression. It is defined by the ratio of the Regression Sum of Squares (SSR) over the Total Sum of Squares (S_{yy}) or, which is equivalent, by one minus the ratio of the Error Sum of Squares (SSE) over the Total Sum of Squares (S_{yy}):

$$R^2 = \frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}} \quad \text{Eq. 3-1}$$

Eq. 3-1 explains why R^2 can only range between 0 and 1. One can think of R^2 as the fraction of total variability in the data (S_{yy}) explained by the regression model (SSR). Sometimes ($R^2 \cdot 100$) is called the percentage of total variability explained by the model. R^2 can also be described as an indicator of the proportion of variability around the average of the observed responses. However a R^2 value close to unity does not necessarily guarantee a good model. One has to keep in mind that adding a regressor variable always increases R^2 due to the increase in SSR. For instance, when performing a curve fit of scientific data using a polynomial model, it is always possible to increase the R^2 by adding higher order terms. Ultimately, this leads to a very complicated fitted curve which basically just connects the dots in our graph. This does not mean that this “model” has any real significance. Such a model with a very impressive R^2 close to 1 might perform very poorly in predicting new data. This is what is called an *overfitted model*.

The square root of R^2 is the *multiple correlation coefficient* between the response, y , and the regressor variables in the model. In linear curve fitting, the correlation coefficient R between y and x can range between -1 and +1 corresponding to a negative or positive slope of x versus y . In Multiple Linear Regression, R equals the correlation coefficient between the observed responses and the predicted responses and ranges from 0 to 1. To visualize the “quality of the fit” of a regression model, sometimes a plot of observed vs. predicted responses is used with a fitted straight line giving the correlation coefficient R or R^2 . Keep in mind, however, that this does not give any information about the adequacy and predictive power of a model.

By taking into account the degrees of freedom in the model, we can define a so-called *adjusted coefficient of determination* or R^2_{adjusted} . Whereas the ordinary R^2 always increases or at least stays constant when adding new model terms, R^2_{adjusted} can actually decrease, thus giving an indication if a new coefficient actually improves the model or might lead to *overfitting*:

$$R_{adj}^2 = 1 - \frac{MSE}{S_{yy} / (n - 1)} = 1 - \frac{SSE / (n - p)}{S_{yy} / (n - 1)} = 1 - \frac{(n - 1)}{(n - p)}(1 - R^2) \quad \text{Eq. 3-2}$$

As defined in Chapters 1 and 2, n denotes the number of data points, $p = k+1$ stands for the number of parameters in the model including the intercept (k = number of regressors). The difference $(n-p)$ decreases when a new regressor is added. This means that, for the new model, SSE has to decrease correspondingly for MSE to become smaller. Only in this case, R_{adj}^2 increases. When evaluating a regression model, R^2 and R_{adj}^2 should be compared. If they differ substantially, the model could be *overfitted*.

3.1.3 Coefficients of Multiple Determination for No-Intercept Models

Generally, R^2 should not be used to compare intercept and no-intercept models. For a model without an intercept, R^2 as defined in 3.1.2 describes the proportion of variability around the origin which can be explained by the regression model. This value can be larger in an intercept-free model than in a model with intercept even though the Mean Square for the Error (MSE) is smaller for the intercept model. Suffice it to say that the MSE (sometimes called RMS Error) is the more appropriate parameter for a comparison between intercept and no-intercept regression models.

3.1.4 Residuals, Standardized Residuals and Outliers

In Linear Regression, the difference between an observed response for a given data point, y_k , and the predicted response, $y_{k(est)}$, is called *residual*. We have already shown in the previous chapters that the sum of the squared errors, which in fact is the sum of the squared residuals, divided by the error degrees of freedom gives MSE, the Mean Square Error for the regression model. But the significance of the residuals does not only lie in this calculation. After calculating a model, a thorough analysis of the residuals is very important to evaluate the adequacy of the regression. The most commonly used methods in residual analysis are:

1. Normal Probability Plots of the Residuals.
2. Plots of the Residuals vs. the Predicted Responses.

3. Outlier Analysis using threshold or cut off values.

From Chapter 1, we know that one of the assumptions of Linear Regression is that the errors or residuals are normally distributed. This can be checked by plotting the residuals in a so-called *normal probability plot*. This can be done manually with normal probability paper by plotting the individual residuals $e_1 \dots e_k$, ranked in increasing order, against the cumulative probability $P_k = (k-1/2)/n$. In a computer program such as ER, the ranked residuals are plotted against the *expected normal value* or *rankit*, which is equal to the inverse of the normal cumulative distribution for a given cumulative probability P_k . In such a plot, the points should form a straight line if the residuals are perfectly normally distributed. In reality, the plot is usually slightly s-shaped, which can be tolerated if the deviation from linearity is not too bad. A pronounced s-shape, however, indicates a distribution with heavy “tails”, i.e. the residuals should be inspected for outliers.

In addition to inspecting the normal probability plots of residuals, it is helpful to plot the residuals versus the predicted responses. If the residuals are not correlated with the value of the predicted response, then this plot should look like a horizontal band on both sides of the expected average for the residuals, zero. If the pattern looks dramatically different, it indicates that the error variance is not constant and depends on the response. Usually, transformations in the regressors or the response are employed to correct this model inadequacy. The shape of the residual vs. predicted response plot can indicate which transformation of the response y could improve the model. For example, if the variance of the residuals increases proportionally with the estimated responses, the plot looks like a “funnel” becoming wider at higher values of the estimated response. In this case a transformation of y to the square root \sqrt{y} could improve the model.

When performing residual analysis, it is sometimes convenient to inspect the *standardized residuals* rather than the raw residuals. Standardized residuals are obtained by dividing the

residuals by their “standard deviation”, or the square root of MSE, also called the *standard error of the regression* (see also Chapter 1).

$$d_k = \frac{e_k}{\sqrt{MSE}} \quad \text{Eq. 3-3}$$

This scales the residuals in units of the standard error, which can be used to define threshold values for *outliers*, i.e., residuals which are so large that they indicate that either the model or the response for the respective data point is erroneous. A cut off value of 3 standard errors is commonly used to distinguish outliers among standardized residuals.

For smaller data sets, so-called *studentized residuals* are more appropriate for residual analysis. Studentized residuals are obtained by dividing the residuals by their exact standard error, rather than the averaged standard error as in Eq. 3-3. The k^{th} diagonal element of the *hat matrix* (see Chapter 1), h_{kk} , where i denotes the k^{th} data point, can be used to calculate the studentized residuals:

$$r_k = \frac{e_k}{\sqrt{MSE(1 - h_{kk})}} \quad \text{Eq. 3-4}$$

Since variances of residuals of remote data points tend to be smaller, the studentized residual of a data point which is outside the bulk of the data tends to become larger. Remote data points sometimes can affect the fit significantly, especially in small data sets. They become *influential points*. Hence, besides indicating outliers similar to the standardized residuals, studentized residuals help detect these influential points.

Influential points can generally be defined as cases which affect the model coefficients dramatically. Therefore, it is interesting to perform the regression without a given data point and determine, if the new model with $n-1$ cases is able to *predict* the withheld observation. This idea is the basis for the calculation of the *prediction error sum of squares* (PRESS) and the PRESS residuals. PRESS residuals ($e_{(k)}$), sometimes called *deleted residuals*, are defined by

$$e_{(k)} = \frac{e_k}{(1 - h_{kk})} \quad \text{Eq. 3-5}$$

Again, h_{kk} denotes the k^{th} diagonal element of the hat matrix. Residuals from data points with large values for h_{kk} will have large PRESS residuals and will be influential. If, for a data point, the difference between the raw residual and the press residual is large, the underlying model with this data point will exhibit a good fit, but the model without this point will predict this response poorly.

Based on the PRESS residuals defined above, another type of residual is sometimes used to detect outliers and influential points. It is called *R student* or *externally studentized residual*. It is calculated by scaling the residual according to the variance $S_{(k)}^2$ which is obtained when fitting the data without the respective data point (that is why it is called externally studentized):

$$S_{(k)}^2 = \frac{(n - p)MSE - e_k^2 / (1 - h_{kk})}{n - p - 1} \quad \text{Eq. 3-6}$$

$$t_k = \frac{e_k}{\sqrt{S_{(k)}^2 (1 - h_{kk})}} \quad \text{Eq. 3-7}$$

R student and studentized residuals will be equivalent if $S_{(k)}^2$ and MSE are similar in value. With influential points, however, these two variances will differ dramatically, and R student will become more sensitive in these cases.

Potentially influential points can be detected by inspecting the value of the respective hat matrix diagonal element, h_{kk} . This value depends on the location of the respective data point in the space defined by the regressor variables. A high value of h_{kk} indicates a potentially influential, remote location in x-space. A cutoff value of $2p/n$ (p = number of parameters in the model, n = number of data points) can be used to detect potentially influential or *leverage points*.

Another statistic measures the squared distance between the estimated response for a given data point based on all data points and the response obtained after deleting the respective case. It is called *Cook's Distance* and defined by

$$D_k = \frac{r_k^2}{p} \frac{h_{kk}}{(1 - h_{kk})} \quad \text{Eq. 3-8}$$

Since D_k contains the product of the squared studentized residual and the term $h_{kk}/(1-h_{kk})$, it is affected by the fit of the model and the distance of the data point from the rest of the data. Points for which $D_k > 1$ are usually considered influential.

The parameter DFBETAS can be used to determine the influence of a data point on the individual regression coefficients of the model. Consequently, as many DFBETAS as there are coefficients in the model have to be calculated for each case .

$$DFBETAS_{j,k} = \frac{r_{jk}}{\sqrt{\mathbf{r}_j' \mathbf{r}_j}} \frac{tk}{(1 - h_{kk})} \quad \text{Eq. 3-9}$$

The vector \mathbf{r}_j is the j^{th} row of the $p \times n$ matrix \mathbf{R} which is derived from the \mathbf{X} matrix of the regressors:

$$\mathbf{R} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \quad \text{Eq. 3-10}$$

As in Chapter 1, the index j denotes the coefficient, k stands for the k^{th} datapoint. For the constant term, $j=0$. A cutoff value of $2\sqrt{(n)}$ can be used to determine observations which are influential for a given coefficient.

Another statistic, called DFFITS, has been defined to detect the influence of an observation on the fitted or predicted response:

$$DFFITS_k = \left(\frac{h_{kk}}{(1 - h_{kk})} \right)^{1/2} t_k \quad \text{Eq. 3-11}$$

A commonly used cutoff value is $2\sqrt{(p/n)}$. Points with larger DFFITS have a considerable effect on the fitted values.

Finally, there is a term which is used to describe the influence of an observation on the precision of the observation. It is called COVRATIO and defined by:

$$COVRATIO_k = \left(\frac{S_{(k)}^2}{MSE} \right)^p \left(\frac{1}{1 - h_{kk}} \right) \quad \text{Eq. 3-12}$$

A high leverage data point will lead to a high value for COVRATIO, unless this point is an outlier. If $COVRATIO_k$ for an observation y_k is greater than 1, the data point will improve the precision of the model, if COVRATIO is smaller than 1, the inclusion of this data point led to a decrease of precision.

3.1.5 R^2 for Prediction, Precision Index and Coefficient of Variation

In the previous chapter, we defined the PRESS residual. The sum of the squared PRESS residuals is the *Prediction Error Sum of Squares* or PRESS:

$$PRESS = \sum_k e_{(k)}^2 = \sum_k \left(\frac{e_k}{1 - h_{kk}} \right)^2 \quad \text{Eq. 3-13}$$

PRESS is calculated from residuals which are based on a regression model with one data point removed. Thus, it can be used to calculate an approximate R^2 , which indicates the predictive power of the model. Analogous to R^2 , the R^2 for Prediction is defined by:

$$R_{prediction}^2 = 1 - \frac{PRESS}{S_{yy}} \quad \text{Eq. 3-14}$$

So, R^2 , adjusted R^2 , and R^2 for Prediction together are very convenient to get a quick impression of the overall fit of the model and the predictive power based on one data point removed. In a good model, these three parameters should not be too different from each other. However, for small data sets, it is very likely that every data point is influential. In these cases, a high value for R^2 for prediction cannot be expected. This reflects the fact that robust model equations are not very likely based on only a few data points. This does

not mean the calculated model is not adequate for the data, but it highlights that predictive models need a more extensive data base to predict with a reasonable statistical confidence.

The Pure Error Sum of Squares (SSPE) we described earlier in connection with the lack of fit test can be considered an estimate for the model independent error variance. This makes it possible to compute an expression for the potential predictive performance of the regression model by comparing the range of the fitted responses to their average standard error. In ER, we defined a *precision index* obtained by calculating the ratio between the range of fitted or predicted responses and the average standard error derived from SSPE:

$$\text{Precision Index} = \frac{y_{est,max} - y_{est,min}}{\sqrt{\frac{pSSPE}{n}}} \quad \text{Eq. 3-15}$$

The larger the precision index the more satisfactory the underlying model can be expected to perform when predicting new values. A low precision index close to 1 indicates that the predicted variability in the responses is only of the order of magnitude of the replicate measurement error.

Finally, the unexplained variability in the data, given by the standard error of regression or the square root of MSE can be compared to the average response. This ratio times 100 is called *coefficient of variation (C.V.)*.

$$C.V. = \frac{\sqrt{MSE}}{\bar{y}} * 100 \quad \text{Eq. 3-16}$$

Clearly, a small value for C.V. is obtained if the fit is good, i.e., MSE is small.

3.1.6 Tests for Multicollinearity, Variance Inflation Factors

We mentioned in Chapter 1 that, ideally, the independent variables in a regression model are orthogonal, i.e., there exists no linear relationship among them. In real life this is not easy to accomplish. Sometimes there are “hidden” relationships between regressor variables. In the case of polynomial models and higher order regression models they are

obvious. Linear or near linear relationships between regressor variables can cause a problem called *multicollinearity*.

Multicollinearity can have severe effects on the estimation of the least-squares regression coefficients. The estimates can become unstable due to an increase in the variances of the coefficients, and the model can become inadequate.

A simple test for multicollinearity is the inspection of the $\mathbf{X}'\mathbf{X}$ matrix of the regressors in correlation form. When applying unit length scaling to the regressors (see equation 1-7), the $\mathbf{X}'\mathbf{X}$ matrix shows main diagonals of 1 and off-diagonals r_{ij} equal to the correlation between the regressors x_i and x_j (one of the add-ins which come with MS Excel allows the user to create this correlation matrix of the regressors). If the regressors are linearly independent, the correlation between the regressors should be close to zero. The determinant of the correlation matrix can assume values from 0 to 1. If the value is 1, then the regressors are perfectly orthogonal, if the value is 0, there exists an exact linear relationship among them. In addition to the correlation matrix, the value of the determinant is given in ER as a “correlation parameter”.

The inverse of the $\mathbf{X}'\mathbf{X}$ matrix in correlation form, $[\mathbf{X}'\mathbf{X}]^{-1}$, offers another possibility to check for multicollinearity. The diagonal elements of this matrix give an indication for the combined effect of the dependencies among the regressors on the variance of the given regression coefficient. They are called *variance inflation factors (VIFs)*. Large VIFs (>10) indicate that the estimate for the respective coefficient could be severely affected by linear dependencies of the regressor.

3.1.7 Autocorrelation

In Linear Regression, we assume that the errors are uncorrelated with respect to the time sequence of the corresponding experiments or data points. Well-defined time intervals for experiments are used for so-called *time-series data*. Uncorrelated errors imply that the value of any error term has no effect on the value of the neighboring error terms when

arranged by their sequential order over time. A serial correlation of the errors is called *autocorrelation*. Autocorrelation affects the variance of the least-squares estimates and may lead to an underestimation of MSE and confidence intervals. In hypothesis testing, it could lead to erroneous results indicating a false significance of regressors.

Residual plots vs. time can be helpful in detecting autocorrelation among errors. If the errors increase or decrease steadily with time, so that we find clusters of residuals with the same sign, we speak of *positive autocorrelation*. *Negative autocorrelation*, on the other hand, leads to residuals alternating in sign too rapidly when plotted vs. time.

A more systematic approach to detecting autocorrelation is based on the assumption that the errors or residuals are correlated via a linear or first-order relationship such as Eq. 3-17 (t = index for time).

$$e_t = \rho e_{t-1} + a_t \quad \text{Eq. 3-17}$$

For uncorrelated errors, we expect that the parameter ρ equals zero. Positively autocorrelated errors should give a positive value for ρ and vice versa. An estimate of this *autocorrelation parameter* is simply the slope of the linear regression line through the residuals (errors) sorted in time order. It can be used to transform the original regressor and response variables in order to eliminate the effects of autocorrelation:

$$x_t' = x_t - \rho x_{t-1}$$

$$y_t' = y_t - \rho y_{t-1}$$

The *Durbin-Watson test* is most often used to determine if there exists positive autocorrelation via hypothesis testing. A test statistic is used to determine if ρ in Eq. 3-16 is zero or significantly larger than zero. The *Durbin-Watson parameter* is defined by

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad \text{Eq. 3-18}$$

.Similar to the F- and t-statistic, threshold or cut-off values for d depend on the degrees of freedom, i.e., the number of data points and number of model terms. For each data set, there exists two bounds for d (d_L = lower bound, d_U = upper bound). If d lies in between these bounds, the test is inconclusive. However, $d < d_L$, indicates autocorrelation, $d > d_U$ indicates no autocorrelation. Tables giving boundary values d depending on probability level and degrees of freedom are available. As a rule of thumb, values of 1.5 and 2.5 can be used as lower and upper cutoffs in many cases.

Table 3-1: selected critical values of d as given by Durbin and Watson

<i>n</i>	<i>1 regressor</i>		<i>2 regressors</i>		<i>3 regressors</i>	
	lower	upper	lower	upper	lower	upper
15	.08	1.36	.95	1.54	.821	.75
20	1.20	1.41	1.10	1.54	1.00	1.68
25	1.20	1.45	1.21	1.55	1.12	1.66
50	1.50	1.59	1.46	1.63	1.42	1.67
100	1.65	1.69	1.63	1.72	1.61	1.74

(one-sided probability level of .05 (95%), excluding the intercept) (J. Durbin and G. S. Watson, *Biometrika*, Vol. 38, 1951)

3.2 *Application: Multiple and Polynomial Regression Main Dialog (II): Regression Summary, Residual Analysis, Outlier Analysis, and VIFs*

3.2.1 Output Area of Main Dialog (II): Summary of Regression and VIFs

After performing a regression, the summary area contains a list of the parameters described in detail in the previous chapter (see also Fig. 2-4). In addition to the ANOVA table and the regression coefficients window, which give indications for the significance of the model and the parameters, this list allows the user to get a quick overview of the quality of the fit and the predictive power of the model. For a theoretical introduction of the parameters listed, see the previous chapter.

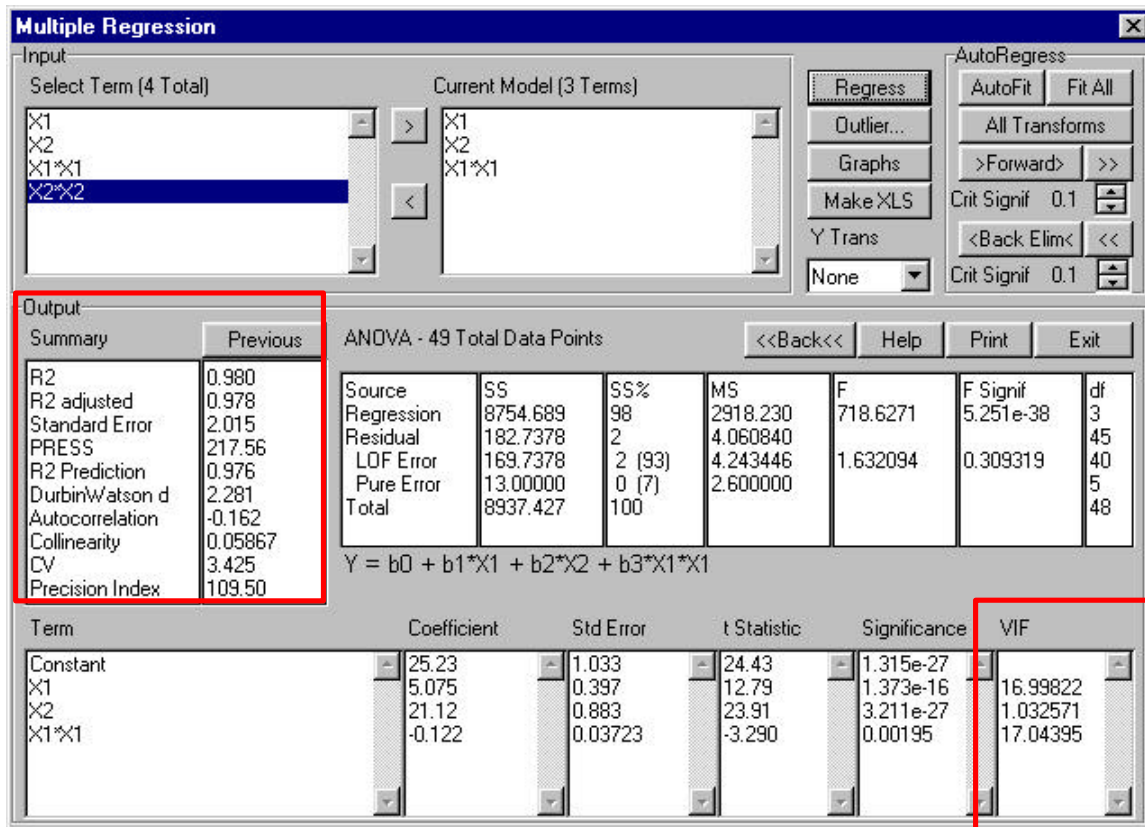


Figure 3-1: Main Dialog, Output Summary and VIF table are highlighted

Previous

This button opens a window which shows the summary results of the previous model, which is convenient when comparing a new model to the previous one and gives important clues when comparing results in a stepwise regression analysis.

VIFs

VIFs, which measure multicollinearity, can be found in the last column of the regression coefficients window at the bottom of the output area of the Main Dialog. As mentioned before, large VIFs (>10) indicate multicollinearity among the regressors.

3.2.2 Outlier Button

The outlier button opens a dialog box which, after performing a regression, gives a convenient overview of a residual analysis based on the regression model used. Listed are,

if present, standard residual outliers (absolute standard residual >3), potentially influential leverage points (diagonal element of the hat matrix, $h_{kk}, >2p/n$), and potentially influential observations or cases based on Cook's distance D ($D>1$). A detailed tabular and graphical residual and outlier analysis using all data points is possible after creating the XLS output sheet containing the complete Essential Regression analysis by using the *Make XLS* button (see Chapter 5.3).

3.2.3 Response Transformation in Essential Regression

As described in Chapter 3.1.4 about residuals, a response transformation can be useful to stabilize the variance of the model. Essential Regression supplies the user with a selection of possible response or y transformations. The *Y Transformation* drop-down list box contains the following options:

- *None* no transformation (default)
- *ln(y)* uses the natural logarithm (\ln) of y in the analysis
- *1/y* uses the reciprocal value of y
- *exp(y)* uses the expression e^y
- *sqrt(y)* uses the square root of y
- *center* centers the response (see Chapter 1.1.2)
- *standardize* standardizes the response (see Chapter 1.1.2)

When selected, the response will be transformed before performing the regression analysis. The original data in the spreadsheet, however, will remain untouched. Note, however, that logarithmic and square root transformations cannot be performed on negative numbers, and the exponential transformation cannot be used with very large numbers.

3.2.4 Graphs button

The *Graphs* button opens another dialog which shows a variety of scatter plots useful for graphical residual analysis. After performing a regression analysis, this complements the

tabular residual and outlier analysis performed when pressing the *Outlier* button. The following graphs for residual analysis and model adequacy checking are included:

- predicted vs. observed response
- raw residuals vs. predicted response
- standardized residuals vs. predicted response
- studentized residuals vs. predicted response
- expected normal value (rankit) vs. raw residuals
- expected normal value (rankit) vs. standardized residuals
- expected normal value (rankit) vs. studentized residuals
- response vs. individual regressors
- response vs. observation or case (to detect trends over time)
- raw residuals vs. observation or case

The dialog showing the graphs contains the following buttons:

<<*Graph n of m*>> buttons

Scrolls back and forth through the selection of graphs. The total number *m* of graphs varies with the number of terms (regressors) in the model

Add Trendline and *Remove Trendline* button

The user can add a regression line to the graphs to visualize possible trends in the plotted data.

Exit

Pressing this button takes us back to the Main Dialog (this button does not quit ER!)

The dialog shown after pressing the *Graphs* button is for visual inspection only. The options of Essential Regression which allow the user to generate editable, storable, and printable output are described in Chapter 5 of this book.

4. Model Optimization

4.1 *Theoretical Background*

4.1.1 The Problem of Finding the Best Regression Model

Quite often, we do not really know how many of a given pool of potential factors are really significant contributors to an effect or response. Also, there might be interactions and/or higher order effects we have to consider. Also, “real world” data can be inconsistent and saturated with outliers leading to misspecified, unrealistic regression models. Normally, when in doubt, we start out by including all the possible regressors into the model equation. We might end up with a regression model that contains all of our potential input candidates, but our nonstatistical intuition and a look at the model adequacy tests tell us more or less instantly that something is wrong. Usually, the model shows a very good fit as judged by the R^2 . However, the adjusted R^2 and the MSE are quite likely to cause concern. Also, some of the regression coefficients might be characterized by low significance (high P-values, see Chapter 2.1.3). What we need is a technique to select a reasonable subset of variables and/or their interactions in order to arrive at a model which is satisfactory. “Satisfactory” usually means that the “fit”, i.e., the R^2 can be lower than in the full model, but the significance of the remaining factors in the model, the adjusted R^2 , the R^2 for prediction and the MSE are lower in the “optimized” model. What we try to achieve is a trade-off between accuracy when reproducing the historical data and predictability or reliability when applying the model to new data.

Technically, when selecting a subset of potential regressors for an optimized model we introduce bias into our coefficient estimates. This is one of the main reasons why all methods for finding optimized regression models are somewhat controversial and, more importantly, they cannot guarantee us to yield the one best model. However, when keeping unnecessary variables in the model, we might actually end up with a higher

variance of the coefficient estimates or the predictions even though the estimates are unbiased. So, what we have to do is to find the model which reduces the variance more than increasing it due to the bias introduced by selecting a subset of variables.

4.1.2 Performing All Possible Regressions and Criteria For Finding the Best Model

The original set of statistical methods implemented in the *Analysis Toolpak* of Microsoft Excel® does not contain automated methods for optimizing regression model equations. Instead, the user has to “optimize” the model by more or less randomly picking variables and trying to find a better model by trial and error. Or, one has to calculate all possible regression models to pick out the best. This can be quite tedious when there is a plethora of potential regression variables, especially if one has to do this manually, one model at a time, in Excel®. But what are the criteria for the “best model”? One popular method uses the *Coefficient of Multiple Determination*, R^2 , to assess the “fit” of a model equation. However, usually, this value tends to get greater with the addition of more variables, irrespective of the significance of the variable added to the model. For a given number of variables, however, the R^2 can be used to determine the best among this subset of model equations. When comparing models with different numbers of variables, the adjusted R^2 (see Chapter 3.1.2) is more meaningful. This parameter can grow even if the number of variables decreases! In other words, the “best” model would be the one with the highest adjusted R^2 ! This is, by the way, equivalent to looking for the model with the lowest MSE (Mean Square Error, see Chapter 1.1.4). When comparing models with similar adjusted R^2 and MSE values, it can be helpful to take a look at the R^2 for Prediction (see Chapter 3.1.5).

In Essential Regression, the user can perform a quick scan of all possible regression models containing no more than 5 regressors to compare their R^2 and adjusted R^2 values. This automated feature is a welcome addition to the methods for variable selection described in the following chapters.

4.1.3 Stepwise Regression: Forward Selection of Variables

Sometimes, our intuition is a good starting point when selecting the more relevant variables from a pool of candidate regressors. However, there are standardized techniques available which use specified criteria for telling us if a new model is actually better than the previous one. In *Essential Regression*, we implemented Forward Selection, Backward Elimination, and the combination of both, which we called “Automatic Model Optimization” or “AutoFit”. We think this is one of the most welcome features of *Essential Regression* adding a substantial amount of convenience to the Multiple Regression analysis.

When applying the Forward Selection method, we start out with a model containing no regressors beside the intercept. New regressors are added to the model one at a time and the F-statistic introduced in Chapter 2.1.2 is used to decide if the additional regressor variable actually improves the model. In other words, the first regressor variable we pick is the one leading to the highest F-value when testing the significance of the regression model using only one regressor and the intercept (see Chapter 2.1.2). Stepwise Regression methods use a threshold value for the lowest possible F, usually called F_{in} (or the highest corresponding probability value, P_{in}) which determines if any regressor is deemed significant enough to start building a model. The next regressor which is added is the one which gives the highest partial F-statistic or, in other words, which shows the highest *partial correlation* with the response after accounting for the effects of the other variables already in the model. If the model already contains the variable x_1 , and x_2 is the new regressor, the partial F-statistic used to find the next variable can be expressed as

$$F = \frac{SSR(x_1, x_2, b_0) - SSR(x_1, b_0)}{MSE(x_1, x_2, b_0)} \quad \text{Eq. 4-1}$$

where SSR stands for the Regression Sum of Squares, MSE is the Mean Square Error, and b_0 denotes the intercept. If there is no regressor which would exceed the predefined F_{in} , the Forward Selection procedure stops.

4.1.4 Stepwise Regression: Backward Elimination of Variables

The reader will learn with only a mild surprise that Backward Elimination works in the opposite direction of Forward Selection. We start with a model possibly loaded with redundant regressor variables and try to strip it down to the really meaningful core. Actually, this is the approach which is more widely used because it allows the analyst to get an idea of the quality of the most comprehensive model before removing variables. In surface response modeling, quite often the models with quadratic terms and with or without interactions between linear terms are used as a starting point. The selection criteria for removing or eliminating a variable is the partial F-statistic, as in Forward Selection. However, now a F_{out} -value is defined indicating the threshold F-value below which a regressor can be eliminated.

4.1.5 Automatic Model Optimization

If a large number of potential regressors is in our “pool of candidates”, there is also a large number of possible regression models. Proceeding through the Stepwise Regression process in only one direction (forward or backward) does not necessarily give us the same answer. Regressors which are important at the point when we add them to the model might actually become insignificant when more regressors are added, and a variable removed from the model might have become much more significant at a later point had we left it in the model. There are bifurcations in the paths leading to an endpoint in the Stepwise Regression methods which might lead to better end results, but we can’t know unless we go back and choose a different route. Fortunately there is a method available which combines both the Forward and Backward techniques of Stepwise Regression. This

method is sometimes referred to as *Stepwise Regression*. In Essential Regression, we called it *Autofitting* because it allows us to step through the process of finding a subset model from a selection of regression models almost “automatically”. This is done by adding a new variable according to the Forward Selection method and by then reevaluating the variables already in the model using the partial F-statistic described previously. When necessary, one of the variables is then removed via Backward Elimination. Obviously, two threshold F-values (F_{in} and F_{out}) or the corresponding P-values have to be defined in order to follow this procedure. They do not necessarily have to be the same values. In fact, F_{in} is usually greater than F_{out} to make it more difficult to add another variable to the model.

Who ever had to go through the tedious phase of selecting a good model when many regressors are present will appreciate the advantages offered by an “automated” selection process. However, care has to be taken in making sure that the result is really meaningful. In other words, maintain caution when using the Stepwise Regression procedures and do not accept physically meaningless model equations just because they mathematically are the optimum of the *Autofitting* process.

4.1.6 Transformation of the Response

We mentioned the transformation of the response variable, y , in Chapters 3.1.2 and 3.1.4 in connection with residuals and outliers (Chapter 3.1.4). As described there, sometimes a pattern in the plot of the residuals vs. the response variable indicates that the error variance is not constant and depends on the response. Model inadequacies such as these can lead to inadequate models even after performing a thorough Stepwise Regression . Transformations of the response can be employed to find a new “starting point” for the Stepwise Regression.

4.2 Application: Multiple and Polynomial Regression Main Dialog (III): AutoRegress Area

4.2.1 Overview



Figure 4-1:
AutoRegress
Area

The *AutoRegress* area can be found in the upper right hand corner of the Multiple and Polynomial Regression Main Dialogs. We assume the user has selected potential regressor variables and a response in the Input Dialog and has continued to the Main Dialog. All possible variables are listed in the *Select Term* list box of the Input Area. In order to find an optimized model, the user can now apply Forward Selection, Backward Elimination, *Autofitting* and or Response Transformation. The subsequent paragraphs describe the functionality behind the buttons in the *AutoRegress* area.

4.2.2 Perform All Possible Regressions

Fit All button

After pressing this button, Essential Regression will ask the user to specify the maximum number of regressors to include. Depending on the number of model terms, the maximum number of models can be quite large (hundreds or more). The program will then calculate the R^2 and adjusted R^2 values of all possible model equations based on the variables listed in the *Select Term* list box. A new Excel worksheet will be generated with a sorted list of the models and parameters.

4.2.3 Stepwise Regression in Essential Regression

Critical Significance Spinners

This allows the user to predefine the threshold values of the critical significance for Forward Selection and Backward Elimination which correspond to the F_{in} and F_{out} values

described in the previous chapter. Remember, however, that the lower the P-values, the more significant the regressor, and lower critical significance values mean that it becomes more difficult to enter or delete a regressor variable to or from the model. By default, values for the critical significance of 0.1 for both the Forward and Backward step are preselected. However, these values do not have to be the same.

>Forward> button

By pressing this button, a Forward Selection step will be performed based on the predefined critical significance. The “>>” button to the right of this button performs a continuous Forward Selection procedure until there is no regressor left which falls below the critical significance value. To start the Forward Selection, no variable needs to be selected, i.e., listed in the *Current Model* list box.

<Backward Elimination< button

Performs a Backward Elimination step based on the predefined critical significance. The “<<” button to the right of this button performs a continuous Backward Elimination procedure until there is no regressor left which exceeds the critical significance value. To start the Backward Elimination, variables need to be selected, i.e., listed in the *Current Model* list box. Usually, one uses the full model to start with a Backward Elimination.

AutoFit Button

Starts the automated selection of the “best” model using repeated Forward and Backward Stepwise Regression until no further improvement can be detected. Note that the currently evaluated regressor variable is indicated in the Excel® status bar. When this procedure starts, all regressor variables are removed from the *Current Model* list box. If successful, a message “Autofit converged!” will indicate that the procedure has terminated.

If Essential Regression cannot find any variable to add or delete, a message box will come up indicating this result to the user. If either the stepwise or the continuous procedures are

successful, the new model will be displayed in the Summary, the ANOVA table, and the regression coefficients window in the Main Dialog.

All Transforms Button

This button allows the user to quickly perform a series of regression analyses based on the current variables in the model and using all of the response or y transformations given in the *Y Trans* drop-down list box (see also Chapter 3.1.4). This is useful to decide which y transformation could be a good starting point for a Stepwise Regression. Essential Regression will come up with a dialog window showing the R^2 values of all transformations and will indicate the best model. After confirming the dialog, the selected model will be displayed in the Main Dialog.

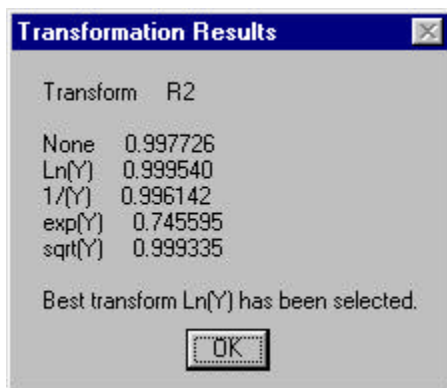


Figure 4-2: Result of analysis of all possible y transformations

5. Essential Regression Output

5.1 Graphical Evaluation of Residuals

In Chapter 3, we already have discussed the mathematical foundation of residuals and outliers and the benefits of both tabular and graphical residual analysis when assessing the quality of our regression model. However, since this part of the book deals especially with

the graphical and tabular output capabilities of ER, we are going to explain in more detail the most important aspects of using graphs in model adequacy checking. This section complements sections of Chapters 3.1.4 and 3.2.3, and the less-experienced user is well-advised to read those first.

The figure below shows a typical Normal Probability Plot of the Expected Normal Values (Rankit) vs. the Residuals. Plots like this are extremely important when trying to decide if the “error structure” behaves as expected, i.e., if the errors are distributed normally. If they are, the residuals will fall on a straight line. If the residual plot is pronouncedly S-shaped, with both ends turning away from the straight line, the error distribution is said to be “heavy-tailed”. In this case, an outlier analysis becomes important, and the tables of the parameters introduced in Chapter 3.1.4. (Residuals, Cook’s Distance, etc.) have to be studied for cases exceeding the threshold values.

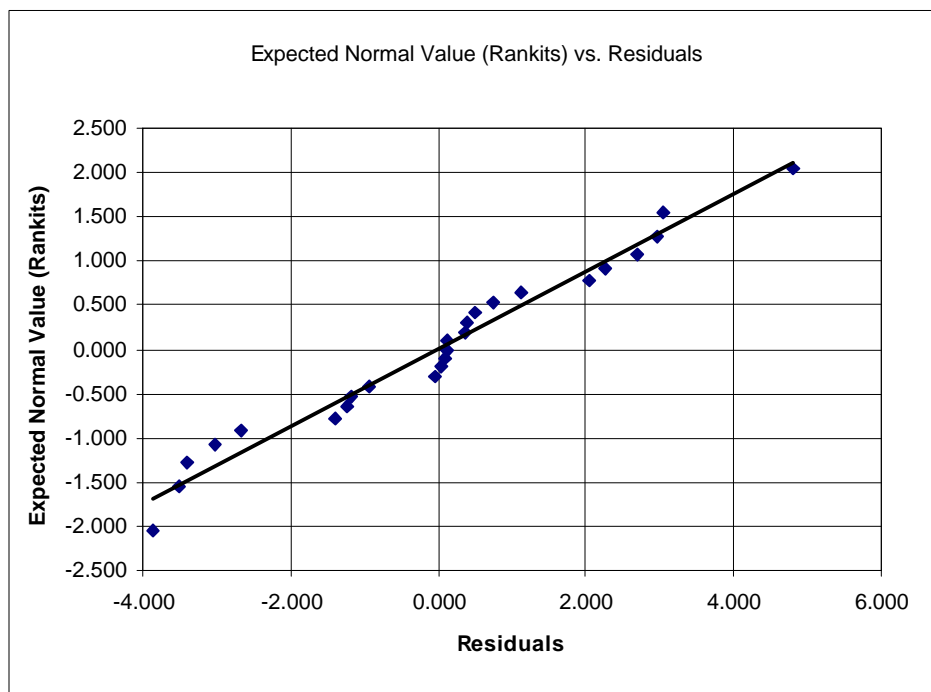


Figure 5-1: Normal Probability Plot of Rankits vs. Residuals

When trying to find outliers or explain unusual residuals, it can be useful to simply plot the residuals vs. the cases sorted by the case number. This is of importance when the cases (observations, experiments) are sorted by time, for instance, and can help find hidden

trends in the residuals or simply “bad” results. In the next Figure, for example, a possible trend in the residuals can be detected.

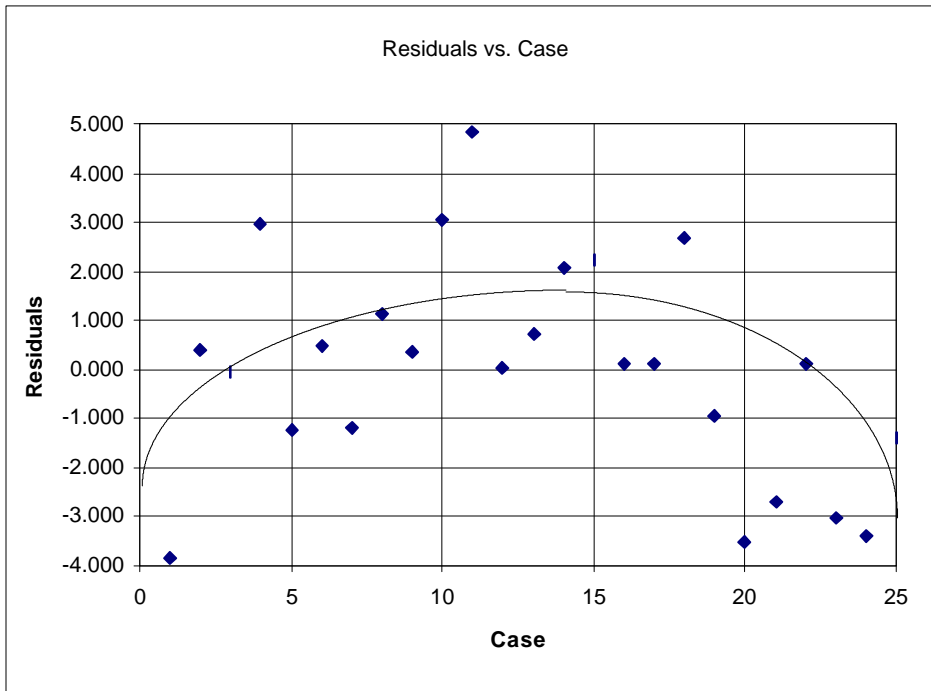


Figure 5-2: Plot of residuals vs. case with a possible trend (indicated by line).

As we mentioned in Chapter 3, the residuals or errors should not correlate with the response. When plotting the residuals vs. the expected or predicted response, they should form a band around 0 with a constant width, i.e., the variance should be stable with respect to the predicted response. If there is a different pattern in this plot, it can help us to find a transformation for the regressor variables or the response leading to a better model, i.e., a regression model with a stable residual variance.

The following plots of residuals vs. expected (predicted) response show a few of the patterns which can occur. The first two plots show typical “funnel” patterns indicating that the error variance increases or decreases with increasing y predicted. The double-bow pattern in the next graph can occur when the predicted y is a proportion between 0 and 1. The U-shape in the last graph indicates nonlinearity. In this case, other regressors or

higher-order-terms might have to be included in the model. The y transformation which might help stabilizing the variance is shown in each graph.

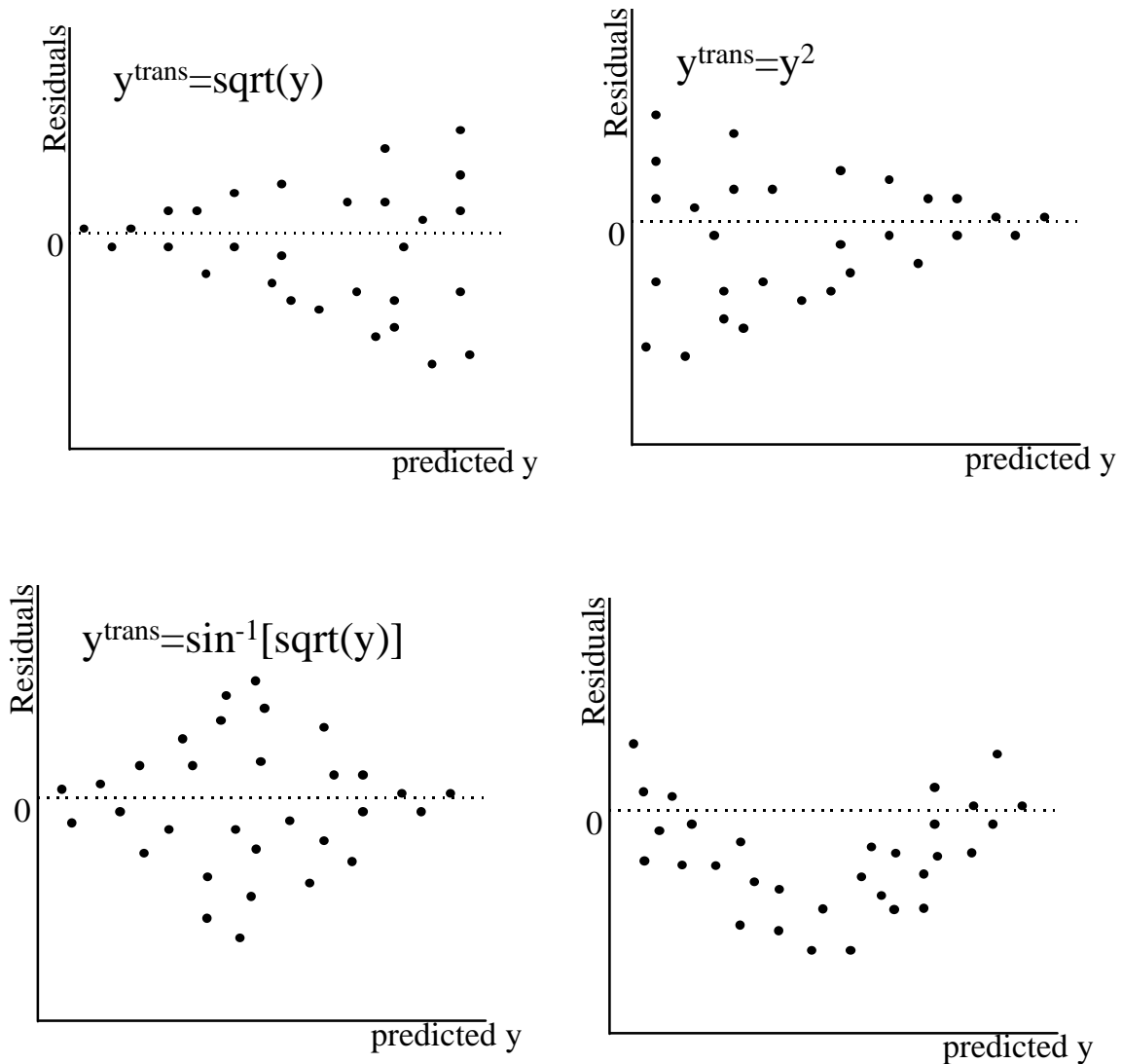


Figure 5-3: Patterns in residuals vs. predicted response plots and possible transformations to stabilize the variance

5.2 *Predicting Observations*

The reason for developing a regression model is not only to fit historical data, but also to be able to predict future observations. A good regression model should enable us to do so. However, in Chapter 1 we already introduced the concept of confidence intervals and uncertainty with respect to predictions. When calculating predicted responses, we must not forget that there is a confidence range associated with each prediction. We have to say, for example: “The predicted response will be y plus/minus the confidence range at the given probability level”.

The reader might recall that there are different equations used for the confidence limits for the mean response and the confidence limit for new observations (see Equations 1-28 and 1-29). The first is used when the **mean response** for a series of experiments or cases at a given data point within the range of the historical data has to be calculated. The second equation gives the confidence limit for a **single new observation** within or outside the range of historical data. This confidence range is wider to reflect the increased uncertainty associated with the prediction of a single response.

Also, both confidence limits vary with the location of the data point in x -space, i.e., the range of the regressor variables. The intervals have a minimum at the center of this range. The more the data point is located at the periphery of the range of the data used to generate the model, the wider the confidence limit, i.e., the more uncertain the prediction. One can see that clearly in Figure 5-4 for the case of Simple Linear Regression (only one regressor plus intercept). In cases where we exceed the range of original data, we actually perform an extrapolation. A model that fits well within the range of original data might perform badly when extrapolating! Unless there is sound physical evidence for the validity of our model outside the range of original data, every extrapolation outside the range of original data is inherently unreliable!

When only one or two variables are in the model, it is relatively easy to determine if a given prediction constitutes an extrapolation. However, for more complicated models we

can calculate the expression $\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$ which was used to determine the confidence limits (Equations 1-28 and 1-29). If this value for a given setting exceeds the range of the hat matrix diagonals (see Chapter 3) in our historical data set, we perform an extrapolation. This calculation is done in Essential Regression automatically every time we predict a new response.

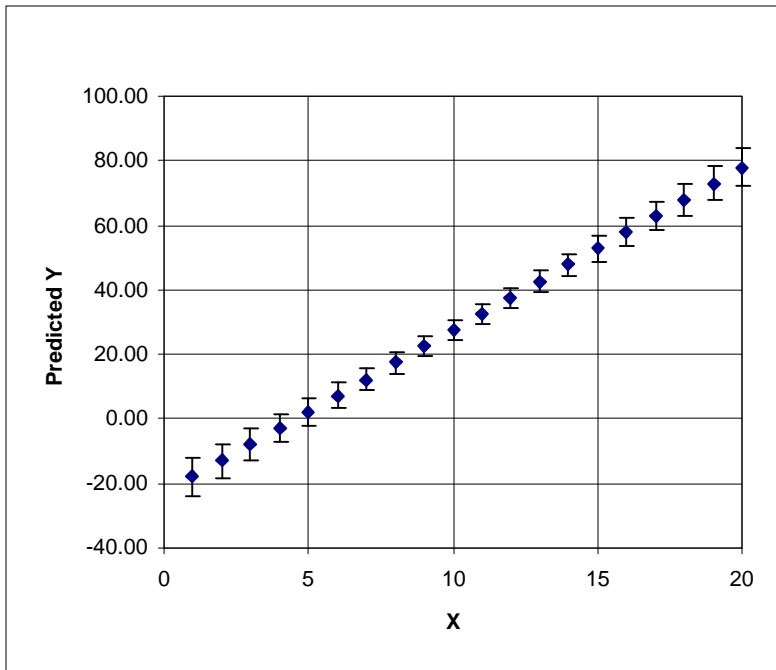


Figure 5-4: Typical plot of the predicted y and the Confidence Interval for the Mean Response at 95% significance level vs. the regressor, X.

5.3 *Application: Essential Regression XLS Output Worksheet*

5.3.1 Make XLS Button-Overview

So far, we have described the features of Essential Regression which allow the user to quickly get an impression of the quality of the regression model by producing dialog and message boxes, showing the parameters important for model adequacy checking in a summarizing fashion. All these features, however, produce only temporary results which change when a new model equation is chosen. To obtain a permanent and also more

detailed output after a regression analysis is complete, the user has to press the *Make XLS button* in the Main Dialog. This starts a procedure which creates another Excel®-worksheet (XLS-sheet) in the currently active workbook containing the data. This new output sheet contains all the information already discussed in connection with the main dialog and, in addition more detailed tables and extended graphical features including surface- and contour plots for models with two or more regressors. When the output sheet is generated, the following message appears:

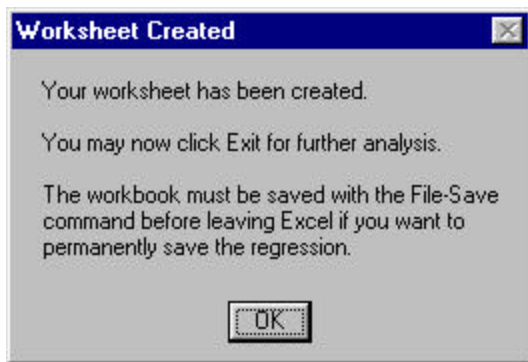


Figure 5-5: ‘Worksheet Created’ Message after pressing the *Make XLS* button

Note that the modified workbook still needs to be saved to make the changes permanent!

	A	B	
1			
2	Reregress		obs =
3	Delete		
4			
5	Predict		
6	Graph		
7	Data		R
8	Regression		R ²
9	Optimize		R ² adj
10	Confidence		Stand
11	Outlier		# Point
12	Print		PRES
13	R Matrix		R ² for F
14	Surfaces		Durbin
15			First O
16			Colline
17			Coeffic
18			
19			

Figure 5-6: Buttons in XLS sheet

By default, the new output worksheet generated by Essential Regression is named *Sheet name of data sheet_Rn*, with *n* counting the output sheets already generated from the data sheet.

The output sheet consists of several areas separated from each other and containing the information described in detail further below. A series of buttons in the top rows of the “A” column of the spreadsheet allow the user to jump to specified output areas and back. Keep in mind that all the areas are on this one spreadsheet. Every area contains a *Back* button which allows the user to jump back to the starting point on the spreadsheet.

This worksheet created after pressing *Make XLS* is a normal Excel spreadsheet! This means, the user can copy, edit and print any area on this sheet! However, moving around the different areas can obviously affect the buttons which take the user to these areas. For this reason, we recommend not moving the output areas.

5.3.2 ANOVA Table, Regression Coefficients Table, and Correlation Matrix

Located in Columns “C” through “H” next to the buttons, the user will find the regression model equation, and the already familiar summary and ANOVA tables for the regression model (see Chapter 1 for details).

Regression button

This button takes the user to the table of the regression coefficients and significance tests (see Chapter 1)

R Matrix button

This button allows the user to inspect the R or correlation matrix of the regressors. As explained in Chapter 3.1.6, this inspection can be useful to find out if regressors are linearly correlated with each other.

5.3.3 Tabular Output of Observations, Predictions, Residuals, and Outliers

Data button

By pressing this button, the user is taken to an area of the spreadsheet containing a detailed table of the regression input data, observed and predicted responses, and a plethora of additional parameters allowing for a thorough analysis of residuals, outliers, and influential observations. If applicable, the cutoff values for certain parameters are given above the respective columns. The table contains the following columns:

1. *Case*, case or observation number in the order of the raw data table

2. x_1, \dots , columns for the k regressor variables including interactions and higher order terms
3. *obs*, observed responses
4. *Predicted obs*, predicted observations
5. *Residuals*, raw residuals
6. *Standardized Residuals*
7. *Studentized Residuals*
8. *PRESS Residuals*
9. *R Student*
10. *DFFITS*
11. *Covariance Ratios*
12. *Std Error Prediction*, the Standard Error for the prediction of new observations
13. *Std Error Mean*, the Standard Error for the prediction of the mean response
14. *P% Confid Int Pred*, the Confidence Range for the prediction of new observations at the predefined significance level P%
15. *P% Confid Int Mean*, the Confidence Range for the prediction of the mean response at the predefined significance level P%
16. *+P % Confid Int Pred*, the upper Confidence Limit for the prediction of new observations at the predefined significance level P%
17. *-P % Confid Int Pred*, the lower Confidence Limit for the prediction of new observations at the predefined significance level P%
18. *+P % Confid Int Mean*, the upper Confidence Limit for the prediction of the mean response at the predefined significance level P%
19. *-P % Confid Int Mean*, the lower Confidence Limit for the prediction of the mean response at the predefined significance level P%
20. *Hat Diagonal*, the value of the hat matrix diagonal element for this observation
21. *Cook's Distance*
22. *Cumulative Probability* of the residual of the given observation
23. *Expected Normal Value (Rankits)* of the residual of the given observation
24. *dfbetas*, k columns of dfbeta values for each of the k regressor variables

A detailed discussion of these parameters and their significance is given in Chapter 3.

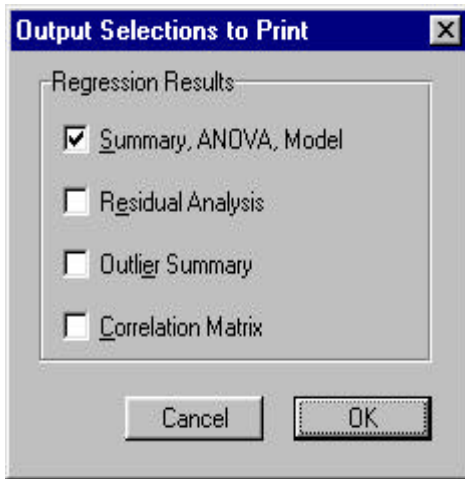
Outlier button

Shows a summary table of outliers or influential observations, similar to the *Outlier* button on the Main Dialog. If any of the observations exceeds one of the cutoff values for Standard Residual, Cook's Distance, and hat matrix diagonal, it is listed here (for an explanation of these terms, see Chapter 3).

5.3.4 Printed Output

Print button

The output areas discussed in the previous chapter can be printed after pressing this



button. A dialog window asks the user to specify the areas to be printed. After pressing “OK”, the *Print Preview* window of Excel will be displayed. Here, the user can make modifications to the page format, etc. After pressing “Print”, the areas shown in the preview will be printed. Note that printing is also possible by selecting an area directly on the output worksheet and printing using the standard Excel print functions!

Figure 5-7:Print Selection Dialog

5.3.5 Prediction of New Observations

Predict button

A new observation or the expected mean value for a given observation can be predicted using this button. A dialog box asks the user to specify the settings for the regressor variables.

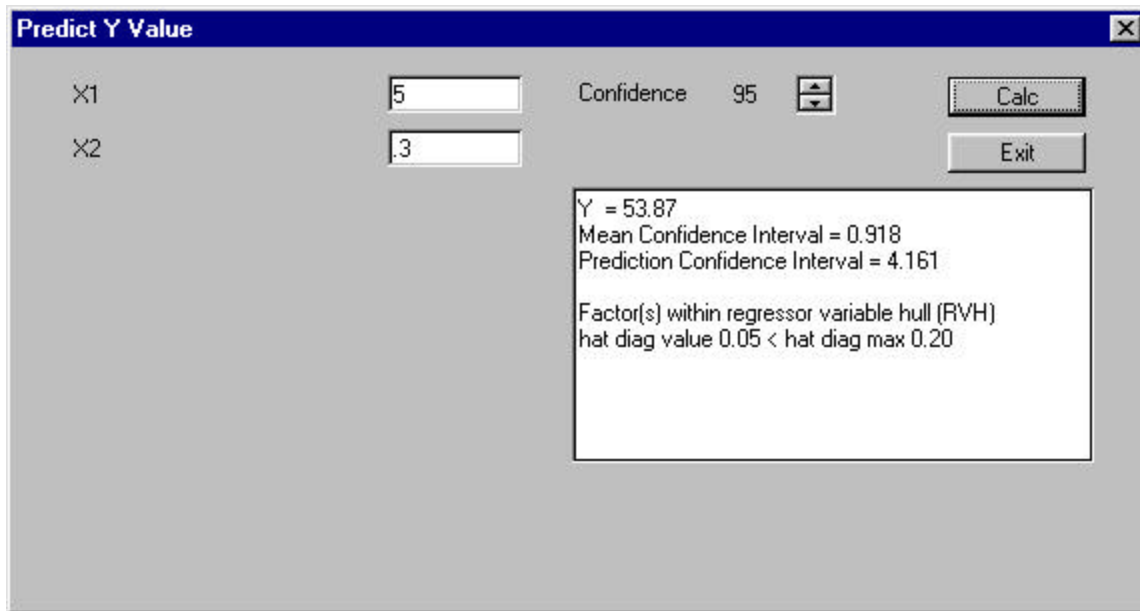


Figure 5-8: Predict Y Value Dialog

Also, a spinner control is provided to adjust the probability level of the confidence intervals. The higher the given number, the wider the confidence limits will be (to increase the probability that the predicted value lies in the confidence range, the wider the range must be!). The *Calc* button starts the calculation. The result is shown in the edit box. As described earlier, the *Mean Confidence Interval* will always be narrower than the *Prediction Confidence Interval*. If the chosen settings for the regressors constitute an extrapolation, the message at the bottom of the prediction output will say that the “Factor(s) are outside the regressor variable hull (RVH)”. The expression *hat diag* stands for diagonal value of the hat matrix. If this value, calculated from the given settings of the regressors, exceeds the maximum value of the original data, *hat diag max*, the current prediction extrapolates the data (see Chapter 5.2).

The *Exit* button closes this dialog.

5.3.6 Finding Input Variables for Given Output (Optimization Problem)

Optimization is here defined as the process of finding “optimized” settings of the regressors in the model in order to obtain a predefined output or response value.

Optimize button

After pressing this button a matrix is displayed showing the input variables (regressors, X-values), their min and max values, their average values (means) and the current settings of the regressors. By default, these current values are set equal to the average values. The corresponding response value is displayed below this matrix as “Y”.

In addition, the **Solver Add-In** of Excel is loaded and the Solver Dialog is displayed. If, after pressing this button, an error message is displayed indicating that “solver.xla” cannot be found, the Solver Add-in has to be added to the Add-In List in Excel by selecting the *Add-Ins* option in the Excel *Tools* menu!

The Solver Dialog displayed by Essential Regression already contains the appropriate criteria. (Note: For some unknown reason, in Microsoft Excel 97, this automatic procedure sometimes will not work, and the user must select the Solver Add-In from the menu and manually click o.k.!). The user only has to select the appropriate “Equal to:” option button and, if a specific value for the response, Y, is desired, enter this value. Note that, by selecting the Max or Min option, Solver will try to find the settings for the regressors which give the highest or lowest value for Y!

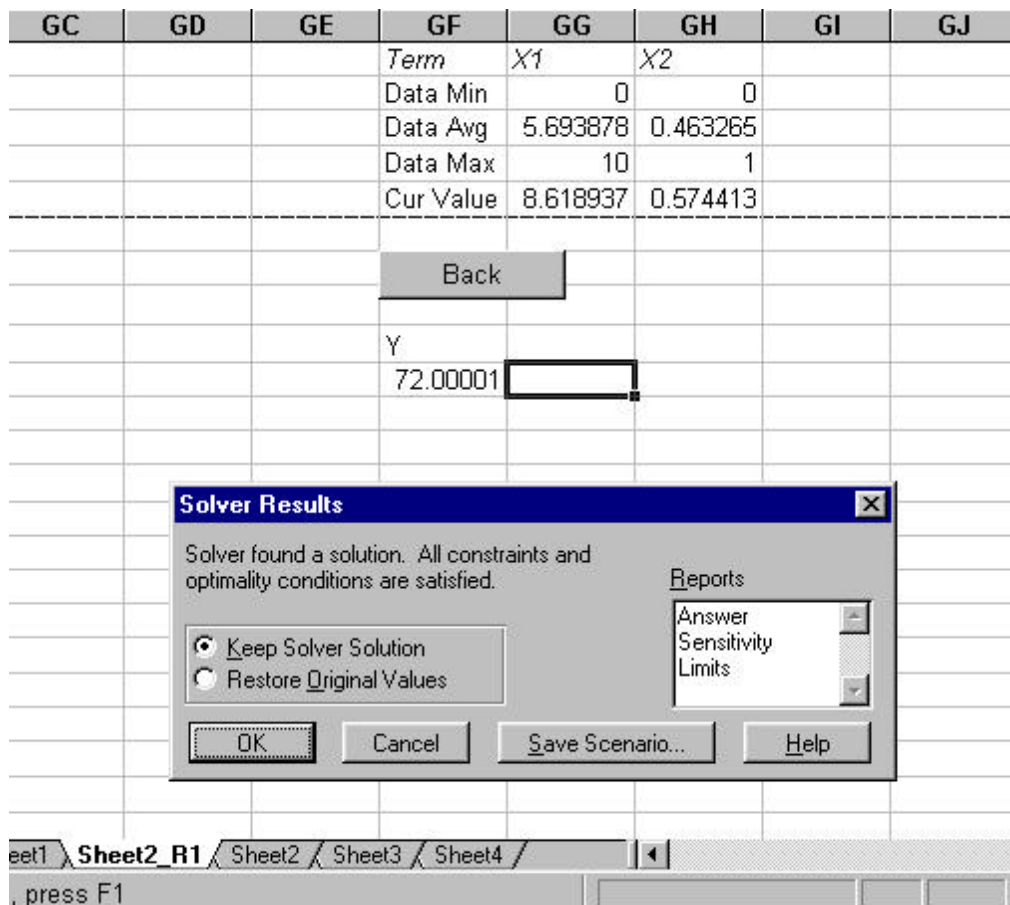


Figure 5-10: Solver Dialog after performing an optimization

The optimize area is on the same worksheet as the other output area. As usual, the *Back* button next to the regressor matrix takes the user back to the top left corner of the output worksheet.

5.3.7 Graphs: Scatter Plots, Confidence Limits, 3D- Plots, and Animations

Essential Regression contains a plethora of graphical output capabilities. Two-dimensional scatter plots are included to visualize the relationships between the columns of the data table accessible through the *Data* button as described further above. For regression models with two or more variables, spatial surface plots and their two-dimensional projections, also called contour plots, are available. All these graphs can be selected, edited, copied, and printed as standard Excel graphs.

Graph button

After clicking this button, the cursor jumps to the graph area. This area contains all the scatter plots which can be generated from the columns of the data table described in the section about the *Data* button. All the graphs can be viewed in a single embedded Excel chart. The y and x variables for the desired plot can be selected from the corresponding drop-down list boxes located above the vertical axis and below the right end of the horizontal axis. The graph is updated automatically according to the selection. It is important to remember that the graphs can be extensively formatted with the normal MS Excel graph editor. Double clicking on the graph will activate the graph editor.

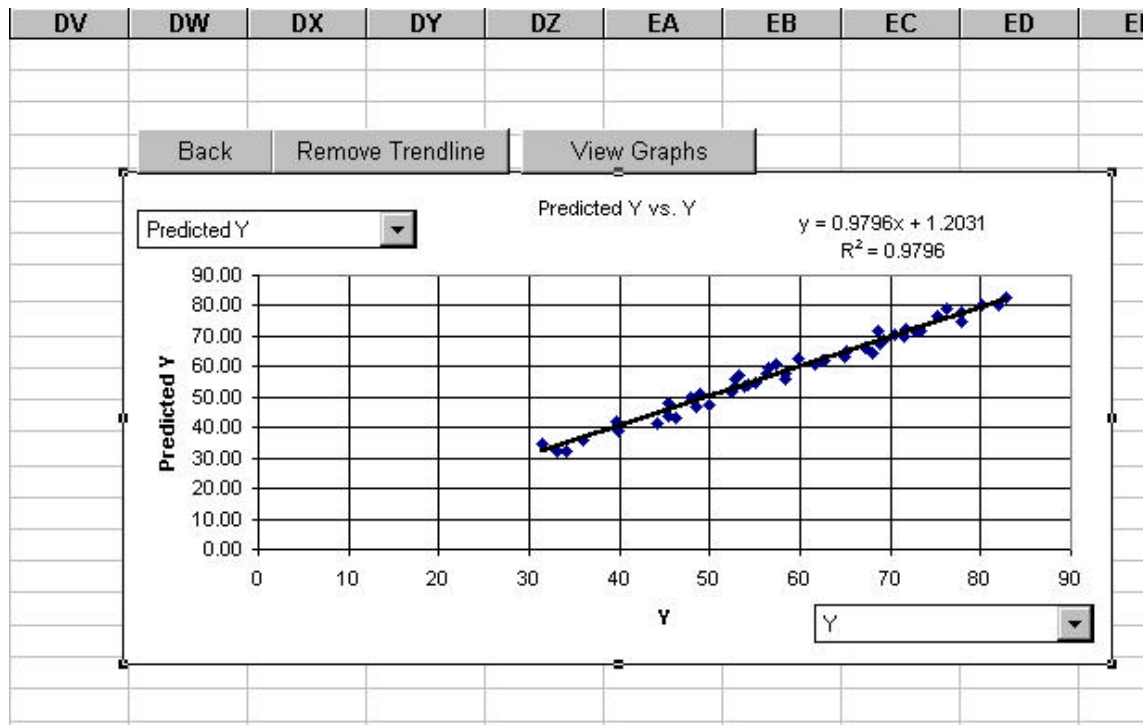


Figure 5-11: Graph area of output sheet with 2D scatter plot of predicted vs. observed response Y including trend line and regression equation

The graphs can be manipulated in several ways:

Add Trendline-Remove Trendline toggle button

Adds or removes linear trendline, R^2 -value, and regression equation for given selection of axis variables.

View Graphs-Graph m of n toggle button

Browses through a selection of standard plots to evaluate the regression model. Included are 8 standard plots similar to the selection we described in Chapter 3.2.4 (*Graphs* button on the Main Dialog) which are helpful for model adequacy checking:

1. Predicted Response vs. Observed Response
2. Raw Residuals vs. Predicted Response
3. Standardized Residuals vs. Predicted Response
4. Studentized Residuals vs. Predicted Response
5. Raw Residuals Normal Probability Plot
6. Standardized Residuals Normal Probability Plot
7. Studentized Residuals Normal Probability Plot
8. Raw Residuals vs. Case

In addition, there are plots of the observed response vs. the individual regressor variables.

Confidence button

Pressing this button takes us to a graph area showing all possible scatter plots of the predicted response vs. the columns of the data table described in the *Data* button section. In addition, the confidence ranges for the predefined probability level are shown in the graphs. A *Mean-Prediction* toggle button allows the user to switch between the (narrower) confidence range for the mean response and the (wider) confidence range for the prediction of new responses. The *Add Trendline-Remove Trendline* toggle button works as described in the previous paragraph.

Surfaces button

This button brings up the surface and contour plot dialog. In a dialog box, the user can choose two different regressor variables (independent variables) for the x_1 and x_2 axis of the plot. The third axis will be used for the response Y (dependent variable). After selecting the regressors (x_1 and x_2 variables for the plot), the 3D-graph area of the output worksheet with a surface plot of the response vs. the two selected regressors is displayed. In addition, a matrix of all the regressors in the model is displayed similar to the one in the optimization area described further above.

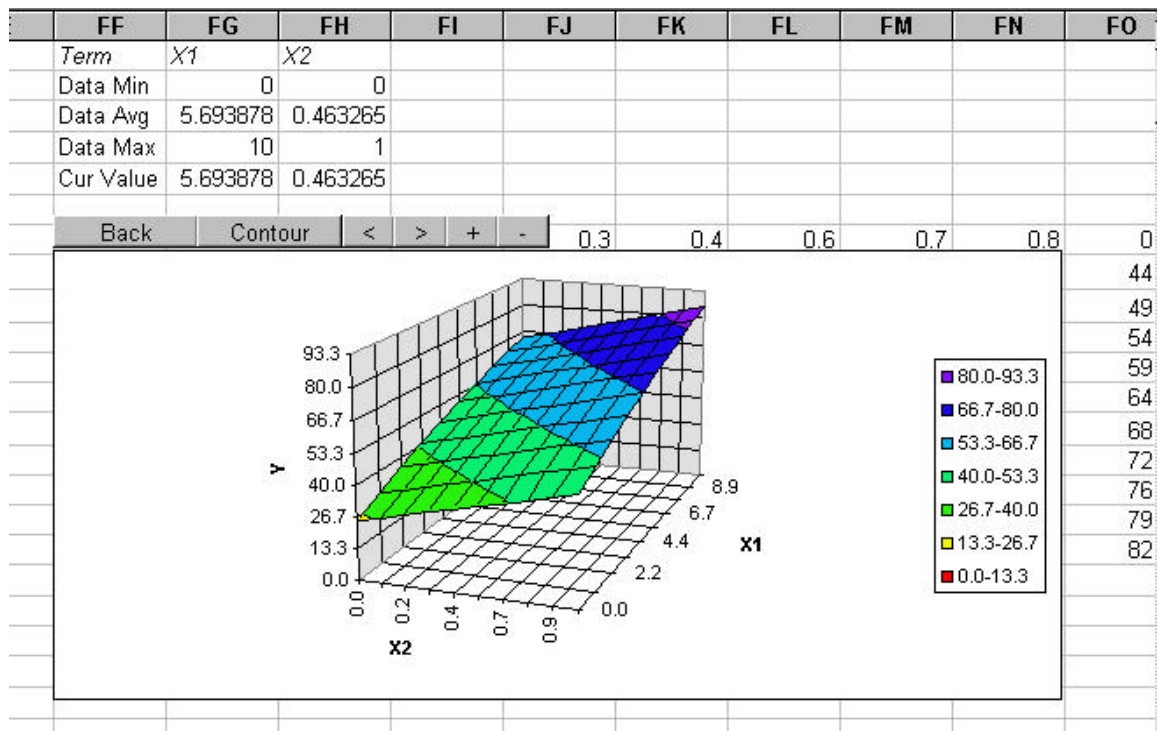


Figure 5-12: 3D-graph area of output sheet with surface plot for 2-regressor models

It is important to remember that the graphs can be extensively formatted with the normal MS Excel graph editor. Double clicking on the graph will activate the graph editor.

There are several possibilities to manipulate the plots:

3d-contour toggle button

This button to the right of the Back button allows the user to switch back and forth between a 3D (surface) and 2D (contour) display.

<, > buttons

With the arrow buttons, the user can rotate the plots to the left (<) or right (>).

+, - buttons

The plus (+) and minus (-) buttons allow the user to increase or decrease the number of levels (= colors) shown in the plots.

Above the surface/contour plot area, a table is displayed which gives minimum, maximum, average, and current values for the regressors (terms) of the regression model. By default, the current value is set to the average value for each regressor. Note that all the regressors in the model are given, not only the ones used in the surface/contour plot. This table offers the following possibilities for further manipulation of the graphs.

By changing the Minimum/Maximum values of the x_1 and x_2 variable (regressors used in the graph) in the regressor matrix, the *scale* of the corresponding axis (x_1 and/or x_2) can be adjusted. The graph will be updated immediately after the new values are entered in the table.

If more than two regressors are in the model, by default, the regressors which are not used as x_1 or x_2 variables in the surface/contour plot are set to their average values. To see the effects of changes in these additional variables, simply change the corresponding current values for the respective variable in the regressor matrix. The graph will be updated immediately after the change. As an additional feature in Essential Regression, this can be done in the form of an automated animation:

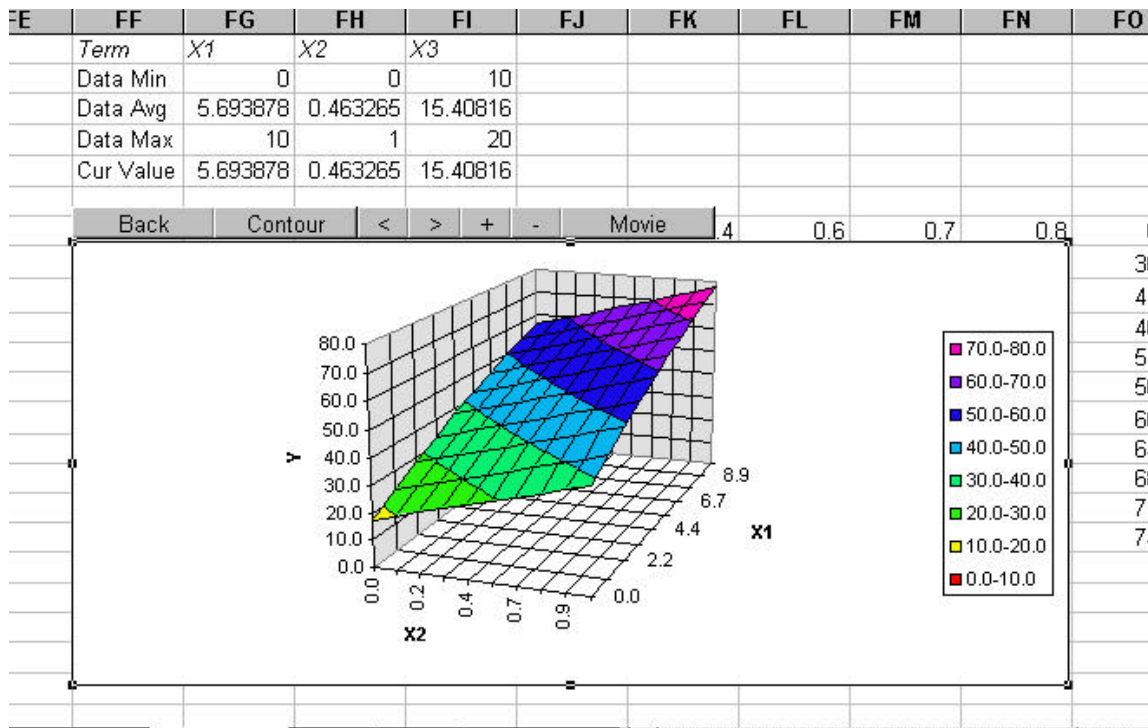


Figure 5-13: 3D-graph area of output sheet with surface plot for 3-regressor model

Movie button (for >2 regressors in the model)

If your model has more than 2 variables, you will find this button above the graph area. The “movie” feature allows you to incrementally change the value of one variable while plotting the response vs. two other variables. If you select to loop through these changes in the movie dialog window, the effect resembles an animation or movie with the surface moving up and down according to the value of the changed variable.

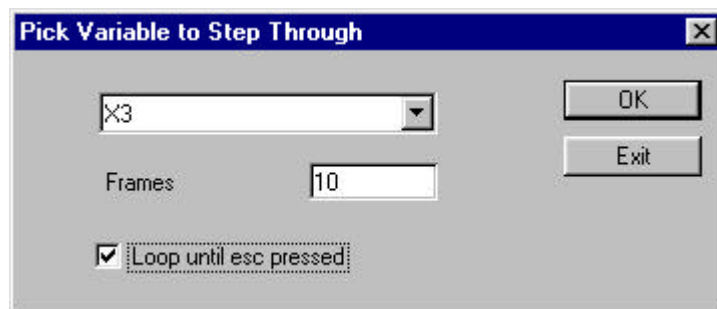


Figure 5-14: “Movie” options dialog

This effect becomes even more dramatic when preventing Excel from autoscaling the response-axis. The user can do that in the conventional way by selecting the graph editor in Excel and changing the settings for the scale of the respective axis.

Since the graphs produced by this procedure are standard Excel graphs, they can be copied to other spreadsheets or Windows applications like MS Powerpoint or Word. So, surface or contour plots with different settings of a regressor not depicted in the graph can be arranged on a page to visualize the effects of this regressor. This generates a quasi four-dimensional graphical representation of the regression model.

5.3.8 Deleting or Duplicating an Output Sheet

Delete button

By pressing this button, the Essential Regression output sheet will be removed from the current workbook. This will also keep track of the changes in the workbook pertaining to Essential Regression such as sheet numbering etc. It is recommended to follow this procedure rather than deleting the sheet by simply choosing *Delete Sheet* from the Excel *Edit* menu.

Duplicate Regression menu option in the Regress menu

Activating this menu option generates copies of the current XLS or output worksheet. It also duplicates the worksheet containing the original data for the regression. The XLS or output sheet must be the active sheet, otherwise an error message is returned!

5.3.9 Starting A New Regression from Output Sheet

Reregress Button

After examining the output sheet, the user might want to continue the analysis of the data set using the current model on this output sheet as a starting point. This is made easy by using the *Reregress* button. Pressing this button will start up the Essential Regression Main Dialog with the current model already selected! Obviously, this is a very convenient

feature compared to starting up Essential Regression from the *Regress* Menu and going through the model selection process again!

6. Experimental Design

6.1 Introduction

The whole area of experimental design is a very large field which has enjoyed a renewed industrial interest in the past two decades. A reasonably complete treatment of the topic would encompass an entire book. We cannot do this. What we will do is cover experimental design as it relates directly to regression analysis. Therefore, we will confine ourselves to covering some classic experimental designs whose analysis is a multiple regression. A prerequisite is that all the design factors are continuous **quantitative** variables. In contrast to **qualitative** variables, quantitative variables are easily measured and described by real numbers. Reactor temperature and reactor pressure are quantitative variables whereas catalyst type is a qualitative variable. A good experimental design methodology allow us to properly distribute our experiments within our factor space so that we can minimize the number of experiments required to develop a statistically sound relationship between factors and a response. The use of qualitative variables in the design and analysis of experiments is beyond the scope of this book.

In the usual jargon of experimental design, the variables which we are looking to make a correlation or regression with a measurable output are called **factors** and the output we are trying to predict is called the **response**. If one is trying to elucidate functional relationships between quantitative factors and a response, multiple regression is the tool required to accomplish this. Therefore, all of the methods and techniques covered up to this point in the book will be completely applicable to analyzing all of the designs presented here. This chapter will concentrate on explaining how one chooses an appropriate design for the problem he is trying to solve and the consequences and tradeoffs involved.

By the end of the chapter we hope to convince the reader that

- 1) when one chooses an experimental design he is also choosing a response regression model
- 2) for quantitative factors, multiple regression is a appropriate tool to analyze the design
- 3) smaller, sequential designs consisting of screening designs followed by response surface modeling design are preferable to single “megadesigns”

We will start by covering designs used for **screening**. These designs are used to determine if a factor is important or not. They are normally done to gain insight into which factors are important in a particular process. This is followed up by **response surface modeling** (RSM) where more details regression models are used to determine response behavior. In this chapter, as in all the others, the included software, Essential Experimental Design (EED), will develop all the experimental designs. Once again, Essential Regression (ER) will do the analysis. We would humbly submit that this modest experimental design and analysis software package can meet the experimental design needs of many of chemists and engineers.

6.2 Screening Designs

The goal of screening is to narrow the a long list of potentially important factors into those that are really important with a known amount of statistical confidence. How one would intuitively accomplish this is by running a given factor at two levels (a high level and a low level) and seeing if varying the level of this factor had any effect on the response. The simplest design for accomplishing this is the **Two level full factorial design**. In this case, a brute force approach is taken. Every factor is run with all the other factors at all their possible settings.

6.2.1 Two Level Full Factorial Designs

Consider the full factorial experimental design below in Table 6-1 for two factors generated by the Essential Experimental Design (EED) software. The low setting for a factor is given as -1 and the high setting 1. In this case six total runs are required (four runs for the main design and two **centerpoints**). Centerpoints are experiments added to the design whose settings are at the midpoint of every factor. The response regression model for this design is

$$Response = b_0 + b_1x + b_2x + b_3x_1x_2 \quad \text{Eq. 6-1}$$

Table 6-1: Full factorial experimental design for two factors

Run	Factor 1	Factor 2
1	-1	-1
2	-1	1
3	1	-1
4	1	1
5	0	0
6	0	0

We can see by doing this design we can estimate the linear effect of each factor (x_1, x_2) and an interaction term. These linear terms or linear effects are often referred to as **main effects**. Similarly, the EED output for a two level full factorial design with three factors is shown in Table 6-2.

Table 6-2: Two level full factorial design with three factors

Run	Factor 1	Factor 2	Factor 3
1	-1	-1	-1
2	-1	-1	1
3	-1	1	-1
4	-1	1	1
5	1	-1	-1
6	1	-1	1
7	1	1	-1
8	1	1	1
9	0	0	0
10	0	0	0

$$Response = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_2x_3 + b_6x_1x_3 + b_7x_1x_2x_3 \quad \text{Eq. 6-2}$$

It is becoming clear that the terms that can be estimated from a Two level full factorial design are main effects and all the possible interactions from 2 way up to n way where n is the number of factors. It is also clear that higher order terms such as x^2 can not be estimated with this design. This is obvious from the inspecting the main design. With only two levels of each factor it is not possible to estimate anything higher than a linear effect. However, centerpoints have been added to the basic design. They do add a third level of each factor to the design. A p^{th} order polynomial requires $p+1$ levels of each factor. From earlier discussions we know that repeated points in a multiple regression are necessary if one is to estimate Pure Error and the Lack-of-Fit (LOF) error. The LOF test provides a direct test of order of the regression. By looking at the significance of the LOF fit test we can see if **curvature** or a higher order term of the factors is present or not. We will show this directly in an example in the next section.

The addition of centerpoints does not effect the estimates of the coefficients except for the value of the constant. The number of centerpoints one chooses to do depends greatly on the preference of the experimenter and the difficulty of doing the experiments. If experimentation is relatively easy, we recommend four centerpoints for 2 level factorial screening designs. Essential Experimental Design defaults to 2, the minimum required to estimate LOF, but it allows a user selected number of up to five or as few as zero. We highly recommend centerpoints be used in all experimental designs. If our design contained any qualitative variables (like type of catalyst), centerpoints would not exist. This would greatly limit our analysis and interpretation of the experiment. We will not be discussing how to deal with these kinds of cases. In general, we like to use three or four centerpoints, if it is feasible, for full and fractional factorial screening designs.

From further inspection of the experimental designs it will be apparent that a computer program is not required to construct either the response regression model or the design matrix. One simply has to run all factors at all levels of all the other factors. The model is all possible linear combinations of the factors. The number of runs required for a 2 level full factorial design is 2^n where n is the number of factors plus the number of centerpoints. This causes the number of experiments to rise rapidly. For five factors 32 runs are required in the main design. However, we will discuss some more **efficient** designs which require fewer experiments to determine the same number of coefficients in the response regression model, especially as the number of factors to screen becomes larger.

6.2.2 Two Level Fractional Factorial Designs

While full factorial designs are very useful they tend to become very large beyond three factors. From our previous discussion it is apparent that much of the additional work with increasing factors is probably not worth it. For example, for five factors, 34 runs are required. For the vast majority of natural phenomena only 15 terms are of interest in the full response regression model. They are the five main effects and ten two way interaction terms. (In fact, there are so many terms beyond two way interactions as the number of

factors increases and they are important in so few cases we have deliberately dropped the higher order interactions out of the EED response regression model for two level full factorial designs. Otherwise, they become a major nuisance.) This means half the experiments in the design are used to estimate three, four and five way interactions. Typically these interactions are not significant and not of interest to the experimenter. Fortunately for the practicing experimenter, statisticians have solved the problem of how to properly **fractionate** a full factorial design to estimate main effects and two way interactions without having to do all the experiments required to estimate all higher order interaction terms.

Typically 2 level full factorial designs are fractionated by taking 2^k fractions. That is to say, one can take a half fraction, quarter fraction, eighth fraction and so on. Recalling that a 2 level full factorial has 2^n runs our fractional factorial design will have 2^{n-p} runs with the stipulation that $n > p$. The obvious questions that arise are 1) which experiments do we take? (we know how many to take) and 2) what effect does the fractionation have on our response regression model?

Consider the main design for a 2 level full factorial experiment for three factors (a,b, and c) shown below. We have multiplied out the values of the interaction terms and split the design into two half fractions based on the value of $a*b*c$. One could take a half fraction of the full factorial design based on the runs that $a*b*c = 1$ (**principle fraction**) or $a*b*c = -1$ (**complimentary fraction**). These two fractions have different shadings in the table. In the principle fraction $c = a*b$ and in the complimentary fraction $c = -a*b$. These equations are called the **generators** for the design. The word abc is called the **defining word** for the design. The **defining relation** is developed by setting the defining words equal to plus or minus one depending on the fraction of interest.

Table 6-3: Main design for a 2 level full factorial experiment for three factors

Run	a	b	c	a*b	b*c	a*c	a*b*c
1	-1	-1	1	1	-1	-1	1
2	-1	1	-1	-1	-1	1	1
3	1	-1	-1	-1	1	-1	1
4	1	1	1	1	1	1	1
5	-1	-1	-1	1	1	1	-1
6	-1	1	1	-1	1	-1	-1
7	1	-1	1	-1	-1	1	-1
8	1	1	-1	1	-1	-1	-1

While taking a half fraction of the full factorial design reduces the number of experiments by one half the effect on the response regression model is not immediately obvious.

Further examination of the design table yields some insight. The following columns have the same values in the principle fraction (a and b*c, b and a*c, c and a*b). Therefore, by taking the a half fraction of this design it is not possible to discern the difference of response dependent on a, b*c or a + b*c. The terms a and b*c are said to be **aliased** with each other. We have paid a penalty for fractionating the design. We are able to estimate fewer terms in the response regression model. Rather than eliminating single terms completely, terms become aliased together when fractionating. In this case

$$Response = b_0 + b_1(a+bc) + b_2(b+ac) + b_3(c+ab) \quad \text{Eq. 6-3}$$

The terms that are aliased together may be easily derived from the "multiplying" the factor of interest with the defining relation. In this case, $a*1 = a*abc$ which gives $a = a^2bc$.

Since the factor columns are either plus or minus one, the square of any value is unity.

Therefore, $a^2 = b^2 = c^2 = 1$ and the final aliases are $a = bc$, $b = ac$ and $c = ab$. Therefore, the following response regression models would yield the same result as 6-3 (to within a factor of 2 on the coefficients).

$$Response = b_0 + b_1(bc) + b_2(ac) + b_3(ab) \quad \text{Eq. 6-4}$$

$$Response = b_0 + b_1(a) + b_2(b) + b_3(c) \quad \text{Eq. 6-5}$$

In short, it is not possible to resolve all the terms in a full factorial regression model with a fractional factorial experiment. Two Level Fractional Factorial designs are classified in terms of their **resolution**. A design is of resolution r if no term of f factors is aliased with another term of less than $r-f$ factors. The previous half fraction design is a resolution 3 design. Each main factor (a, b, c where $f = 1$) is aliased with an interaction term (bc, ac, ab where $r-f = 2$).

- Resolution 2 - Main effects are aliased with other main effects.
- Resolution 3 - Main effects are not aliased with each other but with 2 way interactions. Two way interaction are aliased with main effects and maybe other 2 way interactions.
- Resolution 4 - Main effects are not aliased with 2 way interactions. Two way interactions are aliased with other two way interactions.
- Resolution 5 - Main effects are not aliased with either main effects or 2 way interactions. Two way interactions are not aliased with each other or main effects.

These definitions are summarized in the table below. Understanding resolution is necessary to choose an appropriate fractionated design. Resolution forms a basis for running the EED software. In the next section we will confirm that the design resolution is equal to the number of letters in the smallest defining word for the design. The resolution one chooses has a direct effect on the response regression model. By linear, we mean linear with respect to main effects.

Table 6-4: Definition of design resolutions

Resolution	<i>Aliases of</i> Main Effects	<i>Aliases of</i> Two Way Interactions	Response Regression Model
2	Main Effects		
3	Two Way	Main Effects, Two Way Interactions	Linear
4	None	Two Way Interactions	Linear
5	None	None	Linear + 2 Way Interactions

The above demonstration makes for a relatively simple case. As the number of factors rises, the amount of fractionation we can do and still realize a design of reasonable resolution rises quickly. In the next section we will use the EED software to create a Resolution 3 design for six factors.

With fractionated designs it does not automatically follow that there will be separate fractions for each resolution. In other words, for a particular number of factors a resolution 5 and resolution 3 fraction may only exist (i.e. no resolution 4 fraction exists). The reader need not worry about this level of detail when using the EED software. In this case, a resolution 4 design will not be available. Only a resolution 3 and resolution 5 designs could be chosen by the user.

6.2.3 Using the EED Software for a Two Level Fractional Factorial Design

The Essential Experimental Design (EED) software is launched by opening the EED22.xla file from within Microsoft Excel® Version 5 or 7/95. After seeing the main startup screen, experiments are launched by the new DOE menu to the immediate right of the Edit Menu.

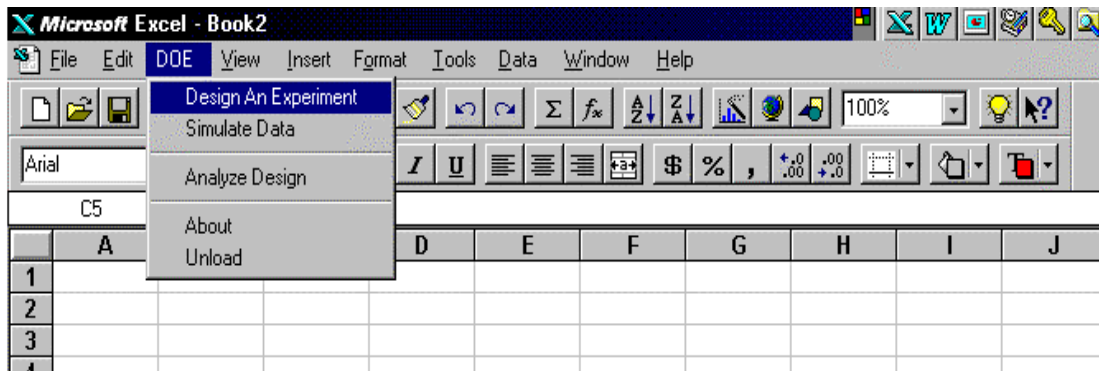


Figure 6-1: DOE menu

Selecting Design An Experiment menu item brings up the main design dialog.

Figure 6-2: EED Main Design Dialog

The Design an Experiment dialog box has several input sections. Basic input information common to any experimental design is required in the top Input section. The number of factors and responses are specified here. The number of centerpoints can also be specified (provided all factors are quantitative). The user can specify if the experimental runs should be randomized and if the aliasing structure should be determined.

The design choices are categorized into two main types, screening and response surface designs. The screening designs are separated by resolution. In fact, resolution is the users guide to selecting a screening design. One can see as the resolution increases, so does the number of runs required. We will begin by doing a resolution 3 screening design on six

factors. We can see that 10 total runs are required for this design. Eight runs for the main design and two recommended centerpoints. We can simultaneously see that the full factorial requires 64 runs while the resolution 3 design requires just 8 runs for the main design. This means that the resolution 3 design is an eighth fraction of the full factorial. Clicking on the Make Experiment button brings up another dialog box for specifying the factors.

The dialog box titled "Factor Definition" contains a table with four columns: Factor Name, Units, Low Value, and High Value. The table lists factors a through f. Factors b, c, d, e, and f have a Low Value of -1 and a High Value of 1. Factor a has a Low Value of -1 and a High Value of 1. Below the table, there are input fields for Factor Name, Units, Low Value, and High Value, with "a" entered in the Factor Name field. At the bottom, there are buttons for "< Back <" and "OK".

Factor Name	Units	Low Value	High Value
a		-1	1
b		-1	1
c		-1	1
d		-1	1
e		-1	1
f		-1	1

Below the table, the input fields are as follows:

Factor Name	Units	Low Value	High Value
a		-1	1

Buttons: < Back < , OK

Figure 6-3: Factor Specification Dialog

Here the user needs to further specify the factors in terms of their names, units (if desired), high and low values. Clicking OK yields the experimental design in the form of an Excel workbook. On the Experiments sheet we see the main design and a brief description of what the design is and the response regression model.

Table 6-5: Output of Experiments sheet (I)

Fractional Factorial, Resolution 3
6 Factors
2 Centerpoints
Linear Model with 7 terms
Response = $b_0 + b_1*a + b_2*b + b_3*c + b_4*d + b_5*e + b_6*f$

Table 6-6: Output of Experiments sheet (II), design table

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>Resp_1</i>	<i>Resp_2</i>
1	1	1	1	1	1	1		
2	-1	1	1	-1	-1	1		
3	1	-1	1	-1	1	-1		
4	-1	-1	1	1	-1	-1		
5	1	1	-1	1	-1	-1		
6	-1	1	-1	-1	1	-1		
7	1	-1	-1	-1	-1	1		
8	-1	-1	-1	1	1	1		
9	0	0	0	0	0	0		
10	0	0	0	0	0	0		

EED lists generators, defining words, and aliases for fractional factorial designs with less than 16 factors. On the Aliasing sheet we can see generators and defining words (at the bottom).

Table 6-7: Generators from the Aliasing sheet

Generators
d = ab
e = ac
f = bc

Table 6-8: Defining Words from the Aliasing sheet

Defining Words
abd
ace
bcf
cebd
cfad
bfae
def

We can see that the design was made from three generators. This makes sense since $2^3=8$ and this is an eighth fraction of a full factorial design. However, we have 7 defining words. Where did they come from? The first three come directly from the generators as in our previous example. The rest come from all possible linear combinations of the first three defining words. For example, the fourth defining word comes from multiplying the first two defining words together and remembering that a squared term is equal to unity. So $abd*ace = a^2bdce = bdce$ or $cebd$ and so on. Multiplying a given factor times all the defining words gives the aliases. For simplicity, the EED software drops four way and higher terms from the alias report.

Table 6-9: Alias report output

Factor	Aliases
a	bd, ce, cdf, bef
b	ad, cf, cde, aef
c	ae, bf, bde, adf
d	ab, ef, bce, acf
e	ac, df, bcd, abf
f	bc, de, acd, abe
ab	d, ef, bce, acf
ac	e, df, bcd, abf
ad	b, cf, cde, aef
ae	c, bf, bde, adf
af	cd, be, bdf, cef, abc, ade
bc	f, de, acd, abe
bd	a, ce, cdf, bef
be	cd, af, ade, abc, cef, bdf
bf	c, ae, adf, bde
cd	be, af, abc, ade, bdf, cef
ce	a, bd, bef, cdf
cf	b, ad, aef, cde
de	f, bc, abe, acd
df	e, ac, abf, bcd
ef	d, ab, acf, bce

Once again, we can see that the design resolution is equal to the number of letters in the smallest defining word. If we select the simulate data option from the DOE menu we can make some experimental data. In this first dialog we can choose which variables we want to be important and the form of the equation whose coefficients we will specify. In this case, we have selected a linear model.

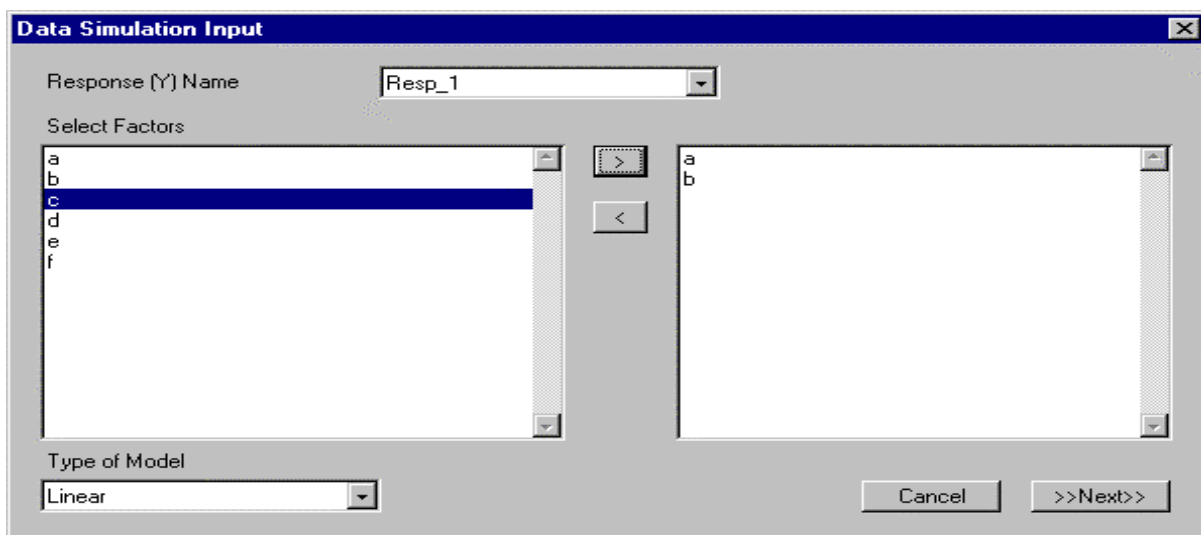


Figure 6-4: Data Simulation Input Dialog

Clicking next brings on the Input Model Coefficients dialog. Here, we can input coefficients for the model terms, enter a constant term and a noise standard deviation value. The noise standard deviation is the coefficient that random values pulled from a **standard normal distribution** (mean of zero and standard deviation of one) are multiplied by. In general, the higher the noise coefficient the noisier the simulated data will be. Some noise in the data is necessary to realistically model an experiment and to avoid singularities in the analysis, especially the LOF test.

Input Model Coefficients

Specify Coefficients

Possible Model Terms

a
b

Initial Coefficients

5
10

To change a coefficient click it and type in its new value below

a 5

Other Info

Constant 0

Noise Standard Deviation 2

Cancel

Make Data

Figure 6-5: Input Model Coefficients Dialog

Table 6-10: Simulated Design

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>Resp_1</i>	<i>Resp_2</i>
1	1	1	1	1	1	1	14.14	14.14
2	-1	1	1	-1	-1	1	3.45	3.45
3	1	-1	1	-1	1	-1	-3.41	-3.41
4	-1	-1	1	1	-1	-1	-16.49	-16.49
5	1	1	-1	1	-1	-1	15.09	15.09
6	-1	1	-1	-1	1	-1	1.13	1.13
7	1	-1	-1	-1	-1	1	-6.33	-6.33
8	-1	-1	-1	1	1	1	-17.12	-17.12
9	0	0	0	0	0	0	1.17	6.17
10	0	0	0	0	0	0	-1.10	3.90

For our second response we have copied the first response with one notable exception. Instead of using just the old centerpoints (Experiments 9 and 10), we have forced some curvature into the response by adding 5 to the old centerpoints. We will now illustrate how one can analyze the design. By comparing the LOF analysis with these two responses, one with no significant curvature (Response 1) and one with significant curvature (Response 2) we will show how curvature can be detected with replicated centerpoints in this two level fractional factorial design. We can now analyze the design by selecting Analyze Design from the DOE menu. This brings up an alternative startup procedure for Essential Regression analysis. The first dialog that will appear is shown below.

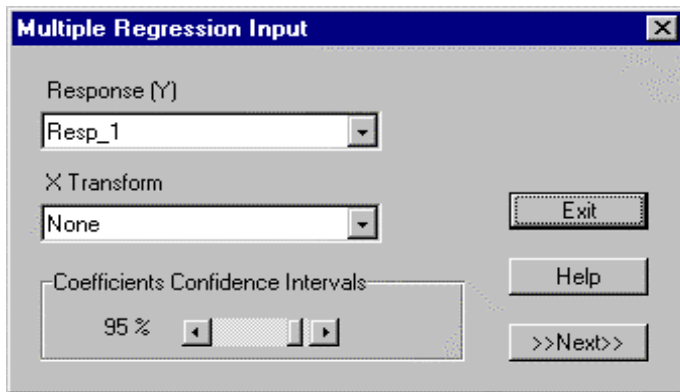


Figure 6-6: Multiple Regression Input Dialog of EED

The user is prompted for the confidence level of the confidence interval calculations, the response for which the regression will be done, and X or factor transformation to be specified. The notable difference between this startup and the usual Essential Regression startup is that the user can not arbitrarily pick regression model terms (i.e., full quadratic, cubic, etc.). The obvious reason for this is the experimental design chosen restricts the maximum order of the model terms. In this case we will be restricted to linear model terms. This becomes very clear after clicking Next. This brings up the main regression dialog. If one allows the program to AutoFit the data by clicking the Auto button the dialog box shown below will be the result.

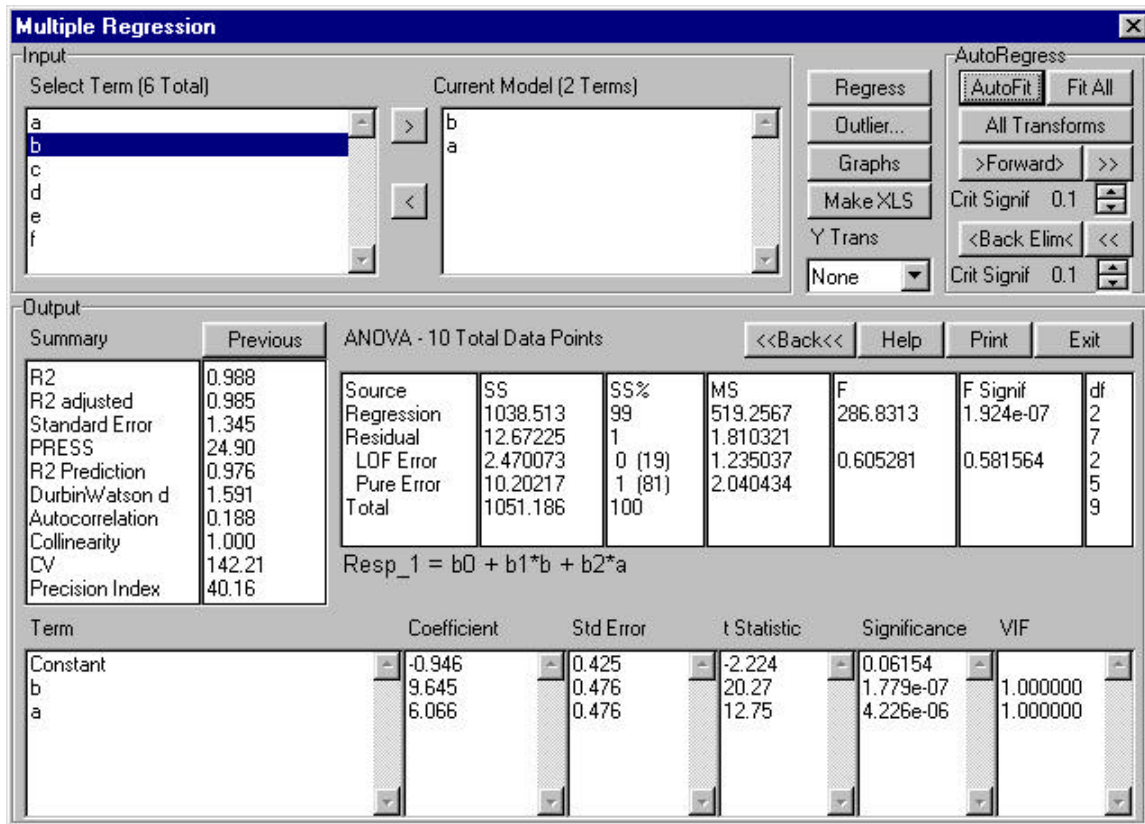


Figure 6-7: Multiple Regression Main Dialog

There are only linear terms available for inclusion into the regression. In this case the AutoFit feature performs well with the default settings. The coefficients for a and b are close to their "true" values of 5 and 10. We know that this particular response has no curvature since no higher order terms were used to create it. The LOF test indicates that LOF significance is high, or that the probability of getting this high a LOF value from random chance alone is low. This result may be interpreted that the correct functional dependence of a and b is linear and not quadratic or higher. The reader may try to improve on the model developed using the AutoFit routine by adding or removing terms from the model. In this case it is probably not possible to improve on the model. Consider the output below from doing the identical analysis on Response 2.

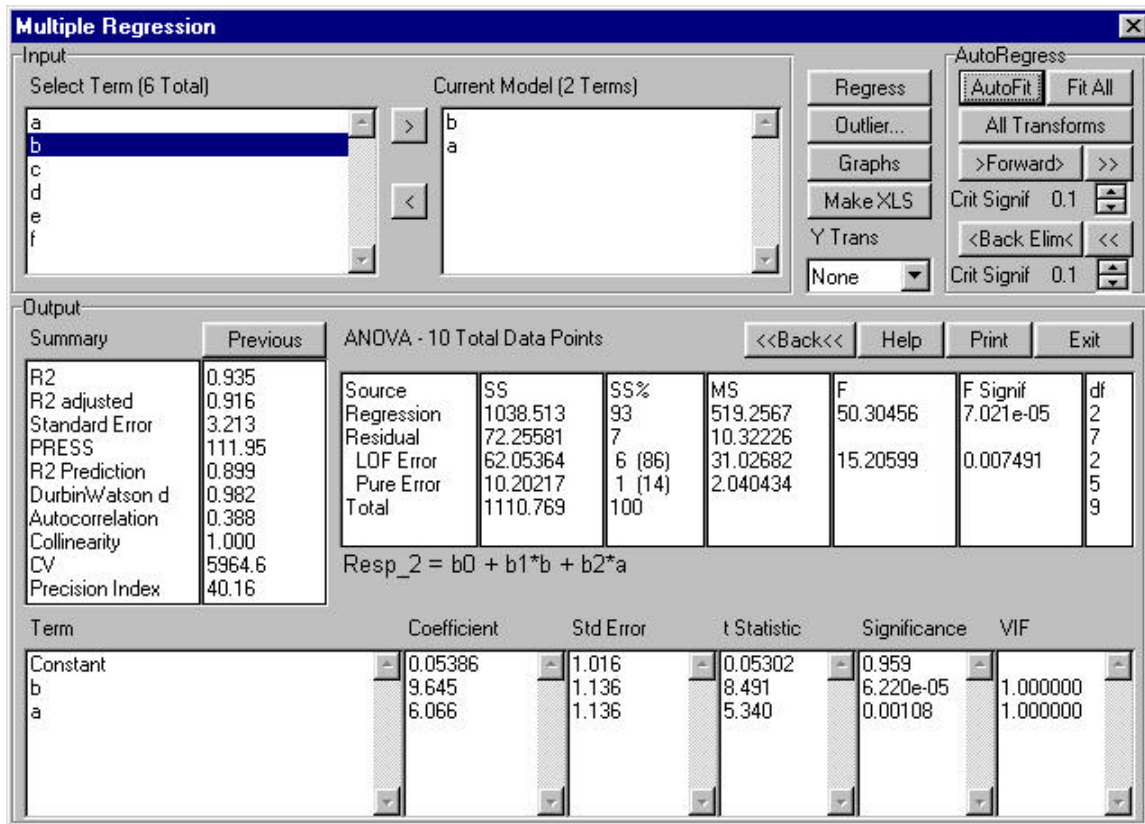


Figure 6-8: Multiple Regression Main Dialog

It is interesting to note that the coefficients of a and b are unchanged even though the centerpoints were significantly shifted to guarantee curvature in the response. The adjusted coefficient of variation is somewhat reduced but still very good. The LOF test reveals that there is very significant evidence of missing higher order terms or LOF. Note that the LOF test is generally more important to inspect if one feels that the important factors have been captured in the regression. That is to say, for good adjusted coefficients of variation with all the factors of interest. The LOF test may not indicate any missing higher order terms but this is not so important if major factors are missing.

At this point the reader is strongly encouraged to run the EED/ER software in simulation mode. Setup screening design and simulate data with varying levels of noise which may be thought of as experimental or measurement error. We think you will be surprised at what you find. The error does not have to get very high in order for unimportant terms to

appear to be important and other problems to arise. This is especially true when one screens higher numbers of factors in fewer and fewer runs (these designs are said to approach saturation, see next section).

6.2.4 Plackett-Burman Designs

In the table below we have determined how many runs are required in the main design for a Resolution 3 two level fractional factorial design. The runs shaded in gray are said to be **saturated**. This is because we know from linear algebra it is not possible to determine $n+1$ unknowns without $n+1$ independent equations. Since we need to determine n coefficients plus a constant (intercept), $n+1$ runs are always required at a minimum. The cases highlighted in gray are perfectly efficient or saturated. We are not doing any extra runs than are absolutely required. If one looks one row past a saturated row, the fractionated two level factorial designs are significantly less efficient. Plackett-Burman designs help fill the increasing void of inefficiency for 11, 19, 23 and 27 factors (12, 20, 24 and 28 run designs respectively) by providing designs that are saturated. This can result in a significant saving of effort. For example, for 11 factors the Plackett-Burman design is saturated and requires 12 runs whereas the Resolution 3 two level fractional factorial design requires 16 runs (33% more). However, nothing in life is free. The Plackett-Burman design has a price to pay. It has a very complicated alias structure. Main effects are not aliased with each other but all main effects are aliased with all two way interactions. This can make for a situation which makes the designs difficult to interpret properly.

Table 6-11: Main design for a Resolution 3 two level fractional factorial design

Number of Factors (<i>n</i>)	Number of Runs	<i>n</i> +1
2	4	3
3	4	4
4	8	5
5	8	6
6	8	7
7	8	8
8	16	9
9	16	10
10	16	11
11	16	12
12	16	13
13	16	14
14	16	15
15	16	16
16	32	17

Consider the experimental design below where we have generated data with the Simulate Data menu option with the expression $\text{Resp}_1 = 6 + 1 \cdot \text{Noise} + 10 \cdot a \cdot b$.

Table 6-12: Experiments sheet output for design in Table 6-11 (I)

Plackett-Burman Design, Resolution 3
8 Factors
2 Centerpoints
Linear Model with 9 terms
Response = $b_0 + b_1 \cdot a + b_2 \cdot b + b_3 \cdot c + b_4 \cdot d + b_5 \cdot e + b_6 \cdot f + b_7 \cdot g + b_8 \cdot h$

Table 6-13: Experiments sheet output for design in Table 6-11 (II), design table

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>Resp_1</i>
1	1	-1	1	-1	-1	-1	1	1	-3.9051
2	1	1	-1	1	-1	-1	-1	1	17.491
3	-1	1	1	-1	1	-1	-1	-1	-2.8844
4	1	-1	1	1	-1	1	-1	-1	-4.3799
5	1	1	-1	1	1	-1	1	-1	15.902
6	1	1	1	-1	1	1	-1	1	15.034
7	-1	1	1	1	-1	1	1	-1	-4.5596
8	-1	-1	1	1	1	-1	1	1	16.476
9	-1	-1	-1	1	1	1	-1	1	16.506
10	1	-1	-1	-1	1	1	1	-1	-6.0701
11	-1	1	-1	-1	-1	1	1	1	-3.0353
12	-1	-1	-1	-1	-1	-1	-1	-1	17.719
13	0	0	0	0	0	0	0	0	6.6746
14	0	0	0	0	0	0	0	0	4.96

A preliminary AutoFit with the default settings first yields no significant model terms. Increasing the forward step significance to 0.2 yields the model below. However, the coefficient of variation is poor. It is clear that the convoluted alias structure is making a number of linear terms seem important. This problem of an inability to detect interaction terms without concomitant linear terms is not unique to Plackett-Burman designs. It is common to all screening designs near saturation (the same problem would be encountered if we used a resolution 3 fractional two level factorial design). If we used a resolution 4 design, we would not find any model terms to be significant. The message that should be received is that if one is concerned that two factors may only be present as an interaction (e.g. a reaction rate) a higher resolution screening design is in order. For this particular case, actually a resolution 5 screening design will catch the $a*b$ interaction term right off from the start. However, this requires 64 runs and is not an attractive course of action.

Normally, nobody wants to run a screening experiment that far from saturation. At this point the reader may feel that these experimental design techniques are not so helpful.

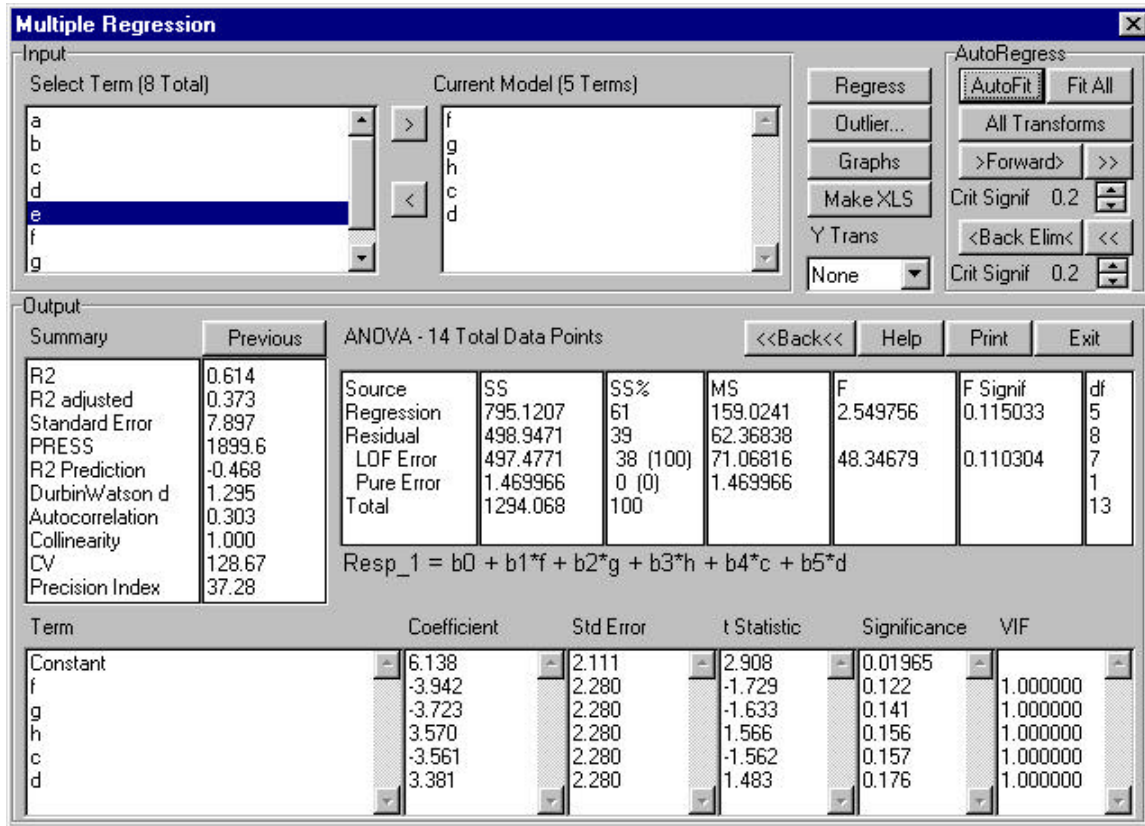


Figure 6-9: Multiple Regression Main Dialog

However, something we should point out at this juncture is that often, terms not in the response regression model for a design can be estimated. In truth, in a saturated design, there are enough degrees of freedom to simultaneously evaluate all coefficients in the response regression model. When one has trouble finding any significant terms in the response regression model, we can often investigate higher order and interaction terms not present in the response regression model. This benefit is a direct consequence of following good experimental design form by using replicated centerpoints. We may run out of degrees of freedom if we add too many terms but the software will let us know. In this case there are 28 two way interaction terms. Obviously we can not simultaneously

estimate all 28 as we have only 12 runs in the basic Plackett-Burman design. However, we may evaluate them one at a time with a stepwise regression routine.

How can we easily do this? By using Essential Regression to analyze the design directly. We can do this by opening ER22.xla. If the Regress menu is not visible strike Ctrl-m (m is for menu). Positioning the cursor at the top of the “a” or first factor column one should click the Regress Menu and choose Multiple Regression. In the first dialog, select an linear + interaction model, factors a-h and Resp_1 as the response. A single forward selection step will immediately show the extreme importance of the a*b interaction. However, the terms g*h, c*h and b*d will also show up as significant via forward selection. However, since they improve the overall fit so very marginally we might correctly assume they are of very minor importance.

If one performs the same screening experiment with a resolution 4 two level fractional factorial design (this requires 16 runs in the main design vs. 12 for the Plackett-Burman) no terms are significant. Since this design has the same response regression model as the previous Plackett-Burman design and a cleaner alias structure (linear terms are aliased with three way interactions only) we can be more confident that no linear terms are important. Going on and fitting an interactions model we find that the a*b interaction is very important on the first forward selection step. This is a bit of a fluke as we can ascertain from either the alias structure or another forward selection step. The regression crashes when evaluating a model with a*b and c*g, or a*b and d*h or a*b and e*h. This is because they are aliased together and cause a singularity in the main regression routine. In fact, if we run the regression four times with only one of these terms we get the same result. The clean alias structure gives one no basis for preferring one interaction term over the other. In contrast, the Plackett-Burman design clearly shows the a*b interaction to be more significant if one runs the regression with only one of the potentially important interaction terms. In general, estimating terms not in the response model equation is not desirable since the design is not orthogonal with respect to the new model. We will discuss this further in the next section.

Plackett-Burman designs are good resolution 3 designs. However, if a main effect shows up as important it is not possible to know which specific interaction terms it is aliased with (it is aliased with all of them). In this case study this was not an issue. A two level fractional factorial design does not have this problem. For this reason we usually recommend staying away from Plackett-Burman designs unless the cost penalty of experimentation is very high.

6.3 Orthogonality and Rotatability

Full and two level fractional factorial designs are said to be first order **orthogonal designs**. That means that if one is using these designs to fit a first order model of the following form

$$y_{est} = b_0 + b_1x_1 + b_2x_2 + \dots + b_ix_i + error \quad \text{Eq. 6-6}$$

for an n run design we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1i} \\ 1 & x_{21} & x_{22} & \dots & x_{2i} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{ni} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_i \end{bmatrix} + \begin{bmatrix} error_1 \\ error_2 \\ \dots \\ error_n \end{bmatrix} \quad \text{Eq. 6-7}$$

which can be expressed in matrix form as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{error} \quad \text{Eq. 6-8}$$

For a design to be orthogonal the matrix multiplication ($\mathbf{X'X}$) must yield the identity matrix. The advantage using an orthogonal design for determining the coefficients in 6-6 is that the variance of the coefficients is minimized (precision is maximized) over any other non-orthogonal design. It turns out that all two level full and fractional factorial designs with resolution greater than or equal to 3 with standardized factor coding (± 1 factor levels) are orthogonal for a linear model. Similarly, resolution 5 or greater two level fractional factorial designs are orthogonal for an interactions model. It should now be

clear to the reader that the EED software always gives the response regression model for which a chosen screening design is orthogonal.

While orthogonality leads to a minimization of coefficient variance, there is another variance that is important when considering experimental designs. It is the variance of the prediction of the response or **prediction variance**. Of course we would like to predict the response at any point in the factor design space with equal or uniform confidence intervals but this is not possible. With some reflection it may be more obvious that we can predict with more confidence in areas of the factor space where we have measured data in contrast to regions where we are extrapolating and interpolating. Since uniform prediction variance is not possible the next best condition is where the variance of the prediction is symmetric about the center of the factor space. This means contours of constant variance form concentric rings about the center of the factor space. Designs which meet this criteria of equal precision of prediction in all directions are called **rotatable**. All two level full and fractional factorial designs with resolution greater than or equal to three with standardized factor coding (± 1 factor levels) are rotatable.

Since all two level full and fractional factorial designs are both orthogonal and rotatable this makes them very sound designs from a theoretical viewpoint. Adding centerpoints to these designs as we recommend does not cause the designs to become non-orthogonal. Centerpoints do not affect the estimates of any of the coefficients. However, they do change the estimate of the constant and they do affect the variance of the prediction. However, these two level full and fractional factorial designs with added centerpoints remain rotatable. Clearly, we know the value of the response more precisely if we measure the centerpoint 10 times (by adding centerpoints to the design) compared to measuring no centerpoints. We remain strong advocates of using replicated centerpoints in all experimental designs.

6.4 Response Surface Modeling (RSM) Designs

In contrast to screening designs the objective of RSM designs is to identify the detailed dependence of different factors on a response. In this case, one is fairly certain that all factors are important and a full quadratic model is the response regression model. For example for two factors the response regression model is

$$\text{Response} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2 + \text{error} \quad \text{Eq. 6-9}$$

In all the RSM designs we will present there is no aliasing between the terms of the full quadratic response model. Aliasing with higher order terms may well be present. In order for one to properly access a quadratic term, a minimum of three levels of each factor is required. At this point the reader may think we are heading in the direction of describing three level full and fractional factorial designs. We are not. All the designs which we will describe and that are fully implemented in the EED software have the three level factorial designs solidly beat from a statistical viewpoint. They are more efficient (require fewer experiments) and have better predictive properties. They are rotatable or nearly rotatable whereas three level factorial designs are not.

The first class of designs we will cover are **Central Composite Designs** (CCD). We will discuss three flavors of this design (**inscribed, circumscribed and face-centered**). One other class of RSM designs we think merits attention is the Box-Behnken designs. They will follow the CCD designs.

6.4.1 Inscribed Central Composite Designs

Rather than starting with a long theoretical discussion let us start with using EED to make an inscribed CCD design for two factors.

Design an Experiment

Input Data

Number of Factors: 2 # of Centerpoints: 4 Number of Responses: 1

☒ All Factors are Quantitative ☒ Show Aliasing (if applicable) ☒ Randomize Worksheet

2 Level Screening

Many Factors

☐ Fractional Factorial Resolution 3 4 Runs

☐ Placket Burman 12 Runs

Higher Resolution

☐ Fractional Factorial Resolution 4 4 Runs

☐ Fractional Factorial Resolution 5 4 Runs

☐ Full Factorial 4 Runs

Response Surface Designs

Second Order Models

☒ Central Composite 8 Runs

☐ Box - Behnken Runs

Central Composite Options

☐ Circumscribed (Min & Max=Star Points)

☒ Inscribed (Star Points outside Min & Max)

☐ Face Centered

Current Design (12 Runs)

Central Composite design for 2 Factors

Model is quadratic

8 model runs and 4 centerpoints

Exit

Help

Make DOE

Figure 6-10: EED Main Dialog

Leaving the minimum and maximum levels for each factor at -1 and 1 and naming the first factor a and the second b gives the following output:

Table 6-14: Output for inscribed CCD design for two factors

Central Composite Design
2 Factors
4 Centerpoints
Quadratic Model with 6 terms
Response = $b_0 + b_1*a + b_2*b + b_3*a*a + b_4*b*b + b_5*a*b$

Table 6-15: Inscribed CCD design for two factors with four centerpoints

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>Resp_1</i>
1	-1	-1	
2	1	-1	
3	-1	1	
4	1	1	
5	-1.414	0	
6	1.414	0	
7	0	-1.414	
8	0	1.414	
9	0	0	
10	0	0	
11	0	0	
12	0	0	

The aliasing information is "Main effects and two way interactions are not aliased with each other but may be aliased with three way and higher interactions".

Looking at the response regression model we can see that quadratic terms are present in contrast to a screening design. The experiments look somewhat familiar. The first four runs are the same as a Two level full factorial design for two factors. The last four runs are replicated centerpoints. The only new feature for this case is the runs 5-8. They are called axial or star points and are illustrated in red in Figure 6-11.

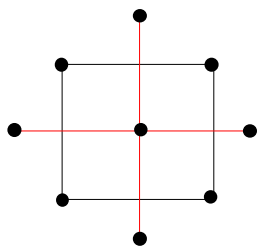


Figure 6-11: Graphical representation of inscribed CCD design for two factors

Let us look at using another inscribed CCD design for three factors:

Table 6-16: inscribed CCD design for three factors

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>Resp_</i> <i>l</i>
1	-1	-1	-1	
2	1	-1	-1	
3	-1	1	-1	
4	1	1	-1	
5	-1	-1	1	
6	1	-1	1	
7	-1	1	1	
8	1	1	1	
9	-1.682	0	0	
10	1.682	0	0	
11	0	-1.682	0	
12	0	1.682	0	
13	0	0	-1.682	
14	0	0	1.682	
15	0	0	0	
16	0	0	0	
17	0	0	0	
18	0	0	0	

Graphically the axial points may be represented as shown.

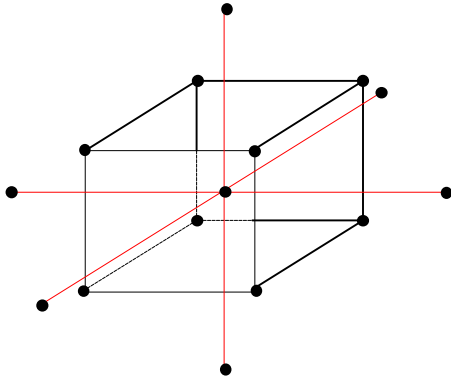


Figure 6-12: Graphical representation of inscribed CCD design for three factors

The pattern that emerges is the one which does describe the inscribed CCD design for n factors.

1. A main Two level full factorial design, or two level fractional factorial design of resolution 5 or higher. A total of F runs is required for this factorial portion.
2. Axial or start points along the factor axes beyond the minimum and maximum values of the main Two level full factorial design
3. Replicated centerpoints.

The reader may be attuned to a practical implication of the first point. It is often possible to use most or all of a 2 level screening design in a RSM design. Suppose we screen on 8 factors and find out three are important. Often we can use many of the runs from the fractionated screening 2 level factorial design. This reusing (or recycling to use a nineties term) is the ultimate elegance in the practice of sequential designed experiments, resulting in true reductions in the number of runs required. Intuitively one might expect savings, however the magnitude may be surprising. Suppose we have a process with 11 potentially

important factors of which four are important. If we wanted to estimate a full quadratic model for all eleven factors, we could go straight to an inscribed CCD. This would require a staggering 2075 runs (2048 for the full factorial, 22 axial runs and 5 centerpoints). Another approach would be to screen with a resolution 4 fractional two level factorial design which would require 35 runs (32 runs for the main design and 3 centerpoints). This would be followed by an inscribed CCD of 28 runs. This **sequential experimentation** has a maximum of 63 runs. The exact same information would be obtained at the end of the two different experiments. If the experimenter is willing to reuse centerpoints and other runs from the screening design this number can be reduced still more. We heartily endorse the concept of sequential experimentation as both a practical and theoretically sound approach to experimental design. It is another reason to use relatively clean two level fractional factorial screening designs with replicated centerpoints.

The reader may be wondering how the number of centerpoints are chosen as well as the placement of the axial points. It is obvious that the axial points are not a fixed distance from the origin from the two previous examples. It is important to remember in response surface modeling we usually are looking for some optimum or will try to predict values throughout the factor space. Since we know beforehand that the factors are important, the chances of a factor not appearing in the final response regression model is small. We will want to make predictions throughout the factor space. This places a premium on minimizing the variance of prediction and maximizing its symmetry (rotatability). The axial distance a is chosen to assure that the designs are rotatable. Calculating a from

$$a = \sqrt[4]{F} \quad \text{Eq. 6-10}$$

assures that the inscribed CCD design is rotatable. The number of added centerpoints has no impact on the rotatability of the design or the value of a . However, for reducing the prediction variance, especially in the center of the factor space, 3-5 centerpoints are

recommended. The EED software defaults to four centerpoints and allows the user to select a maximum of five or a minimum of three.

6.4.2 Circumscribed Central Composite Designs

However, there are cases when we do not want to have the maximum values we specify in the minimum and maximum values of the factors multiplied by a to get the axial points. Instead we wish the minimum and maximum values we specify to be the values of the star points. The minimum and maximum values for the factorial part of the design needs to be scaled to an appropriate level. Clearly, if we know that we have absolute upper limits on certain factors we might want to specify them as the axial points. Table 6-17 below shows a EED output for a circumscribed central composite design, which does this desired scaling, for 3 factors.

The circumscribed central composite design, is not really different than the inscribed central composite design, it just has its factor levels scaled such that the axial points are to the user specified minimum and maximum level. Therefore, it has the same number of total runs, same number of centerpoints and same axial distance value a is used. In order to see this one must divide $1/0.595 = 1.682$ as it was in the previous inscribed case for three factors.

Table 6-17: EED output for a circumscribed central composite design for 3 factors

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>Resp_1</i>
1	-0.595	-0.595	-0.595	
2	0.595	-0.595	-0.595	
3	-0.595	0.595	-0.595	
4	0.595	0.595	-0.595	
5	-0.595	-0.595	0.595	
6	0.595	-0.595	0.595	
7	-0.595	0.595	0.595	
8	0.595	0.595	0.595	
9	-1	0	0	
10	1	0	0	
11	0	-1	0	
12	0	1	0	
13	0	0	-1	
14	0	0	1	
15	0	0	0	
16	0	0	0	
17	0	0	0	
18	0	0	0	

One has to remember that if he ultimately thinks he will be performing a CCD, the axial points might present a problem. Otherwise, the factorial points from a screening design may not be reusable in the final CCM. Planning and forethought yield savings in effort and experimentation. If there are no problems going beyond the limits of the screening factorial experiment, then inscribed central composite designs really lend themselves to sequential design as described in the previous section. Otherwise, the circumscribed

designs described here might be necessary. Consequently, the screening design upper and lower limits may have to be narrowed or more experiments performed.

6.4.3 Face Centered Central Composite Designs

This is a very special case of the central composite where the axial distance a is the same as the minimum and maximum values of the factorial portion of the design. The axial distance value $a = 1$. For three factors, the axial points are on the centers of a cube and hence the name face centered CCD. In general, this is not a desirable thing to do as we have emphasized that the value of a determines whether or not the CCD is rotatable. In fact, the face centered central composite design is not rotatable. For three factors the EED output is shown in Table 6-18.

In this case the number of centerpoints is reduced to 2. This is all that is required to get a reasonably even variance of prediction throughout the design space. When is it appropriate to apply these designs? When the factors have clear boundaries that can not be exceeded. For instance, percent conversion of a raw material or Shore D hardness which have intrinsic limits of 0 and 100. Factors that have measurement scales that intrinsically have fixed upper and lower limits. In these cases, rotatability is not a major concern.

Table 6-18: EED output for 3-factor-face centered CCD

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>Resp_</i> <i>l</i>
1	-1	-1	-1	
2	1	-1	-1	
3	-1	1	-1	
4	1	1	-1	
	-1	-1	1	
6	1	-1	1	
7	-1	1	1	
8	1	1	1	
9	-1	0	0	
10	1	0	0	
11	0	-1	0	
12	0	1	0	
13	0	0	-1	
14	0	0	1	
15	0	0	0	
16	0	0	0	

6.4.4 Box-Behnken Designs

These are an unusual class of 3 level designs appropriate for fitting second order response models. They are rotatable or nearly rotatable depending on the number of factors.

Below is a Box-Behnken design for 3 factors.

Table 6-19: Box-Behnken design for 3 factors

<i>Exp #</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>Resp_</i> <i>l</i>
1	-1	-1	0	
2	1	-1	0	
3	-1	1	0	
4	1	1	0	
5	-1	0	-1	
6	1	0	-1	
7	-1	0	1	
8	1	0	1	
9	0	-1	-1	
10	0	1	-1	
11	0	-1	1	
12	0	1	1	
13	0	0	0	
14	0	0	0	
15	0	0	0	
16	0	0	0	

The design is shown graphically in Figure 6-13. The features which immediately jump out when inspecting these designs is that they are "corner free". No runs are done at the design corners. There are no experiments where at least one of the factors is not at its midpoint. In contrast, to the inscribed and circumscribed central composite designs, there are no star or axial points so each factor appears at only three (not five) levels. At first glance, these may not seem like potential candidates for sequential experimentation. However, closer inspection reveals that they are balanced blocks within the design which does make them candidates for sequential experimentation. In this case we can see, three

blocks of 2 factor 2 level factorial designs. In fact all of the Box-Behnken designs are balanced block designs. Since the factors not being studied in a particular block are always set at their midpoint, corners never appear. One should especially consider using these designs when one is not interested in predicting behavior in the corners of the design space.

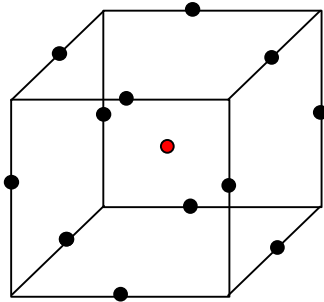


Figure 6-13: Graphical representation of Box-Behnken design

6.5 Summary

The EED software has limits on the number of factors for each type of experimental design. They are summarized in Table 6-20 below. Obviously for an arbitrary number of factors only certain designs are available. For more than 7 factors only screening designs are available. The software handles all of this automatically. Choices that are not supported cannot be selected as their option buttons become "grayed out". Obviously, the EED software was designed so that an experiment will not exceed 64 runs (without centerpoints). This is a hefty and realistic number.

The software will also help the user pick the design with the highest resolution possible for a given number of experiments. For example for 3 factors, resolution 4 and resolution 5 two level fractional factorial designs both require 8 runs. The software automatically selects the highest possible resolution design and makes the lower resolution unavailable.

In other words, the user is always guided into obtaining the maximum amount of information for the number of experiments performed.

Table 6-20: Number of factors and runs for each type of experimental design

Experimental Design	Number of Factors	Number of Runs
Full 2 Level Factorial	2-6	4-64
Fractional Factorial (Resolution 5)	2-8	4-64
Fractional Factorial (Resolution 4)	2-11	4-32
Fractional Factorial (Resolution 3)	2-31	4-32
Plackett-Burman	2-27	12-28
Central Composite Designs	2-6	8-44
Box-Behnken	3-7	12-56

Table 6-21: Possible and recommended number of centerpoints

Design	EED Allows	Number Recommended (Minimum/Preferred)
Full Factorial	0-5	2,3
Fractional Factorial (All Resolutions)	0-5	2,3
Plackett-Burman	0-5	2,3
Central Composite (Circumscribed and Inscribed)	3-4	4,4
Central Composite Face Centered	2-5	2,3
Box-Behnken	3-5	4,4

We would like to re-emphasize and summarize some of the important points made earlier. The first is that one should always use some repeated points in an experimental design to estimate lack-of-fit (LOF). Table 6-21 summarizes the possible and recommended number of centerpoints for the various designs.

Replicating the centerpoint in two-level fractional factorial designs has advantages because the LOF test can be interpreted from a curvature point of view. This was shown directly in section 6.2.1. Replicating other points in the design does not afford this judgment of the presence or absence of curvature in the response regression model.

It is important that the repeated centerpoints represent *replicate* samples and not merely *repeat measures*. A repeated measure is really just measuring the response of a given experiment more than once rather than completely starting over. Consider an experiment where a spectrometer is measuring some absorbance of a painted panel. Repeated measures would correspond to multiple measurements made on a single sprayed panel. A true replicate would start from a separate batch of raw materials, a separate substrate and another spraying of the substrate. This truly captures the variation in the entire coating process. Repeated measures in this case could be used to determine the variability in the spectrometer.

Another issue when doing an experiment is the possibility that there is some other bias floating around that could disturb the analysis of the design. Consider an experiment that is performed outside. If all the experiments that performs have a high level of factor A are done in the morning when the ambient temperature is cooler, we can not separate the effect of ambient temperature and factor A. These two factors are said to be *confounded*. If we feel a variable like ambient temperature is important it should be included as a design factor. But if we do not, we should try to minimize its potential effect on the analysis through *randomization* of the design. The main dialog in EED does have a randomize checkbox for this very reason. This allows there to be a completely random way of running the experiments. Therefore, it is unlikely that factor A will be confounded with

ambient temperature. Furthermore, even if ambient temperature were slightly significant, it would not interfere with the analysis of the experiment since its effect would be distributed across the design. It would contribute to the lack of fit error.

In closing we would want to point out some of the dangers of using an experimental design and blindly running an automatic fitting of the response. This often will lead to terms in the response which may not be important. We recommend that the user manually use sequential forward selection and backward elimination steps keeping an eye on the adjusted R^2 value. If for a given forward selection the adjusted R^2 goes up only very slightly and the term does not seem to be physically feasible we would tend to go without that term. One has to remember that, in general, one does not have a great excess of data in analyzing an experimental design so it is relatively easy for a term to seem important by chance alone. If one had a large amount of data (>50 points) then this precaution need not be made. Running the EED/ER package in simulation will demonstrate our point very strongly.

7. Quick Guide and Tutorial

7.1 *Important Reminder*

Essential Regression and Essential Experimental Design are compiled Microsoft Excel® Macros (Add-ins). In other words, Microsoft Excel is needed to run them. They were developed for Microsoft Excel Versions 5.0c and later. We recommend using Microsoft Excel 7.0 for Windows 95 or Excel 97 (Version 8.0). It has not been tested for versions of Excel beyond 97. We cannot guarantee it will work on newer versions. We will try and upgrade the software if necessary when later versions of Excel arrive.

7.2 **Installation**

Essential Experimental Design and Essential Regression come on a 3.5" disk. To install the software, run setup.exe on the disk.

*Insert the disk in your diskette drive. In Windows 3.x, if the disk drive is named drive A, start the File Manager and activate the file list for drive A. You can either double-click the setup.exe file in the file list (Windows 3.x) or select **File, Run**, and then type "setup.exe". In Windows 95, you can either double-click the "My computer" icon, then do the same with the "3 1/2 Floppy [A:]" icon and then double-click the "setup.exe" file, or you click on the **Start** button, select **Run** and type "a:\setup.exe".*

By default, setup.exe will install the program files to "C:\eregress". You can choose a different destination if you prefer. Setup will also install a program group "Essential Regression" in the start menu (Windows 95) or the Program Manager (Windows 3.x).

Note: Setup.exe will not install the data file **er_test.xls** which is also on the program diskette. Please copy this file manually from the diskette to the directory in which you installed Essential Regression (C:\eregress by default).

7.3 **Loading Essential Regression into MS Excel**

From within MS Excel

In Excel, with at least one empty workbook open, select the **File, Open** menu. Locate ER22.xla in c:\eregress (or the directory you installed the program into) and open it. This will start the Add-In and, after an introductory screen, add a new **Regress** menu to the Excel main menu bar between the **File** and **View** menus.

From outside MS Excel

In Windows 95, select **Start → Programs → Essential Regression → Essential Regression**. In Windows 3.x, double-click the Essential Regression Icon in the Essential Regression Program Group.

Like any other Excel file, Essential Regression (ER22.xla) can also be opened directly in the File Manager (Windows 3.x) or Explorer (Windows 95) .

MS Excel will start up (or simply become the active application if it was started up before), and, after the introductory screens, a new **Regress** menu will be added to the Excel main menu bar between the **File** and **View** menus.

7.4 Performing a Regression Analysis using the ER_Test Data

Note: Paragraphs in italics are meant only to point to additional features of Essential Regression. You are not supposed to execute them. However, if you do so, your screen could look different from what is given in the text and you should go back to the point before you “took the detour”.

Open the Excel workbook “er_test.xls” which you should find in the Essential Regression program directory (provided you copied it from the program diskette, see chapter “Installation”). On the “data” worksheet, this workbook contains a small data set. The regressor variables X1 and X2 and the response, Y, are arranged in columns, the observations are arranged in rows. Any data table to be analyzed with Essential Regression should be arranged like this. The “A” column contains the index number of each observation. In columns “B” and “C”, you’ll find the effects or x variables. Column “D” contains the response or Y variable.

Cell “B1” is highlighted in red. It is the leftmost cell in the header row of the range with useful data (not counting the index column). We call it the “pivot cell” of the data table. Please select this cell.

Note: It is important to select the pivot cell in a data table before launching Essential Regression!

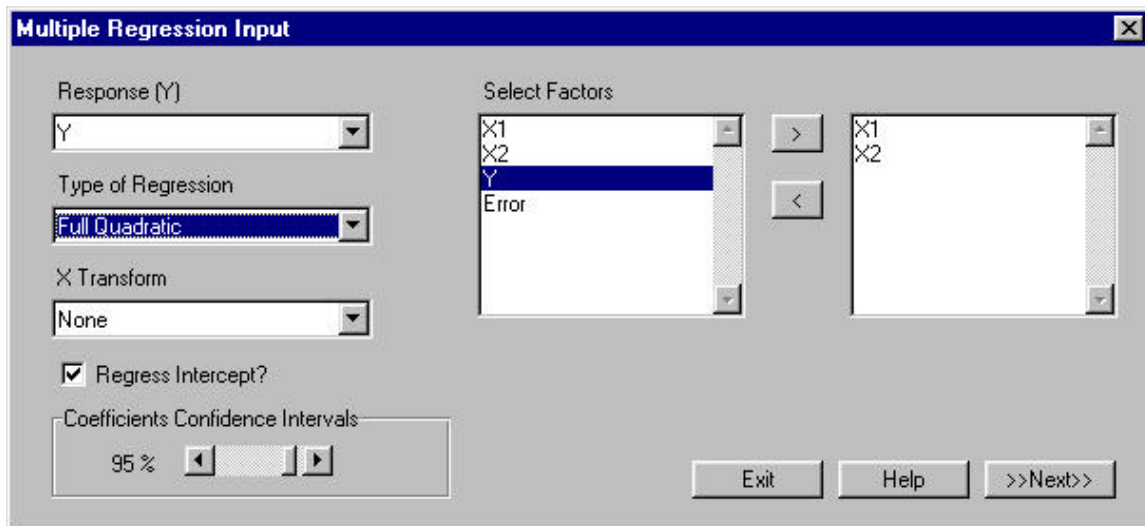
In the **Regress** menu, select **Multiple Regression**. This will activate the **Multiple Regression Input Dialog**.

To select the response variable, click the “down arrow” in the *Response (Y)* drop-down list box. The list box should show the variables. Select the “Y” variable as the response.

Now focus on the two “Select Factors” windows in the dialog box. To select factors or input variables, add “X1” and “X2” from the list in the left window to the right window by using the “>” button between the windows. Do not add the response or y variable to the right window. If this happens, you can remove it using the “<” button.

Go to the “Type of Regression” drop-down box and select “Full Quadratic” from the list.

Do not change the remaining options. The dialog box should now look like this:



Click “>>Next>>”. This opens the “Multiple Regression” Main Dialog.

In the upper left quadrant of the “Multiple Regression” Main Dialog you’ll find the “Select Term” window with a list of all possible terms in the model based on the “Full quadratic” model selected in the previous dialog: linear, squared, and interaction terms. Note that

Essential Regression creates this list for you automatically. Using the arrow buttons, model terms can be added or deleted from the model after selecting them in the corresponding window. The terms currently in the model are listed in the “Current Model” Window. Note that any subset of the regression model selected in the previous dialog under “Type of Regression” can be created.

Select “X1” in the “Select Term” window and click the “>” button. Repeat this with “X2”. This creates a linear model with these two terms. To perform the regression, click the “Regress” button to the right of the “Current Model” window. This executes the regression analysis and the dialog should now look like this:

The screenshot shows the 'Multiple Regression' dialog box with the 'Regress' button highlighted. The 'Current Model' contains two terms: X1 and X2. The 'Output' section displays the following data:

Summary		ANOVA - 49 Total Data Points							Equation	
R2	0.984	Source	SS	SS%	MS	F	F Signif	df	$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2$	
R2 adjusted	0.984	Regression	10116.00	98	5058.001	1431.968	3.751e-42	2		
Standard Error	1.879	Residual	162.4813	2	3.532203			46		
PRESS	183.65	LOF Error	105.3988	1 (65)	4.053801	1.420330	0.212282	26		
R2 Prediction	0.982	Pure Error	57.08250	1 (35)	2.854125			20		
DurbinWatson d	1.879	Total	10278.48	100				48		
Autocorrelation	0.03102									
Collinearity	0.975									
CV	3.221									
Precision Index	120.71									

Term	Coefficient	Std Error	t Statistic	Significance	VIF
Constant	26.79	0.653	41.00	7.226e-38	
X1	3.990	0.09319	42.81	1.039e-38	1.025300
X2	20.60	0.825	24.99	2.102e-28	1.025300

The “Multiple Regression” Main Dialog displays most of the results needed to evaluate a regression model instantly. In the “Output” area, the “Summary”, “ANOVA”, and regression coefficients or “Term” window show the parameters needed to assess the

quality of the selected model. For example, you can see that the coefficient of determination R^2 for the linear model is .984, the adjusted R^2 is .984, and the so-called R^2 for prediction, estimating the prediction accuracy of the model, is .982. In the ANOVA table, the F-value is high (1432), and the F-significance is very low ($3.75e-42$), indicating a highly significant regression model.

What if you want to evaluate other models based on the selected variables “X1” and “X2”? How does the full quadratic model compare to the linear model ?

Select the first term in the “Select term” window and click the “>” button repeatedly until all the 5 possible terms are in the model. Note that the “Output” area is cleared when doing that. Now click on “Regress” again. The dialog should look like this:

The screenshot shows the 'Multiple Regression' dialog box. In the 'Input' section, 'Select Term (5 Total)' contains X1, X2, X1*X1, X1*X2, and X2*X2. The 'Current Model (5 Terms)' list also contains these five terms. The 'Regress' button is highlighted. The 'Output' section shows the 'Summary' tab with the following data:

Summary	Previous
R2	0.989
R2 adjusted	0.987
Standard Error	1.642
PRESS	148.06
R2 Prediction	0.986
DurbinWatson d	2.036
Autocorrelation	-0.03534
Collinearity	0.000250
CV	2.814
Precision Index	86.75

The ANOVA table for 49 total data points is also displayed:

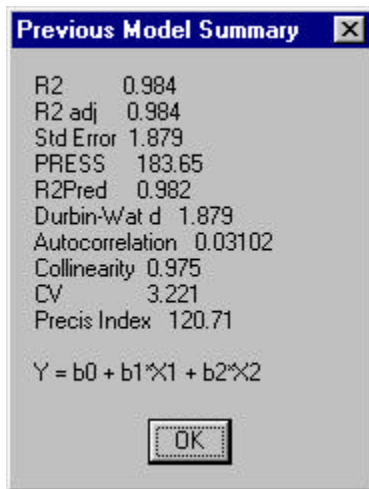
Source	SS	SS%	MS	F	F Signif	df
Regression	10162.57	99	2032.515	754.0187	1.064e-40	5
Residual	115.9098	1	2.695576			43
LOF Error	58.82726	1 (51)	2.557707	0.896144	0.602971	23
Pure Error	57.08250	1 (49)	2.854125			20
Total	10278.48	100				48

The regression equation is: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_1 + b_4X_1X_2 + b_5X_2X_2$

The coefficients table is as follows:

Term	Coefficient	Std Error	t Statistic	Significance	VIF
Constant	25.57	1.085	23.56	3.296e-26	
X1	5.304	0.357	14.87	1.476e-18	19.68504
X2	17.29	3.208	5.390	2.796e-06	20.33866
X1*X1	-0.145	0.03532	-4.101	0.000179	18.51996
X1*X2	0.07454	0.281	0.265	0.792	10.93036
X2*X2	2.747	3.188	0.861	0.394	21.06522

These are the results for the full quadratic model. You can see that all three R^2 parameters have improved. This can be checked easily by clicking the “Previous” button. The “Previous model summary” is displayed:



However, a look at the “ANOVA” and coefficients window shows that the F-value has decreased (from 1432 to 754), and, more obviously, some of the model terms have a low significance, i.e., the probability output for the t-statistic in the coefficients window shows numbers >0.1 (remember: the **smaller** the significance number in the table, the **more** significant the term).

Apparently, our model contains “unnecessary” terms. How can we find out fast what is the “best” model among the possible combinations of linear, quadratic, and interaction terms? In Essential Regression, we have the possibility to perform **forward and backward stepwise regression** based on a threshold significance which can be adjusted by the user.

You’ll find buttons for forward selection or backward elimination of model terms in the “AutoRegress” area in the upper right corner of the dialog. For example, using the full quadratic model with 5 terms, we could use the “<<Back Elim<<” button now to remove insignificant terms from the model in a stepwise fashion.

Another possibility is the use of the “Fit All” Button (can be used with no model terms selected in the Main Dialog) to get a list of all possible models (31) sorted by decreasing R^2 and R^2 adjusted. If you do that, you’ll get another worksheet with a list of all possible subsets of our 5 term quadratic model.

However, one of the exceptional features of Essential Regression is the “AutoFit”, i.e., the automatic selection of the “best” model using repeated forward and backward stepwise regression until no further improvement can be detected.

Using the dialog as shown above as a starting point, press the “AutoFit” button in the “AutoRegress” area (upper left corner). Note that the progress is indicated in the Excel status bar at the bottom of the screen. After a few seconds, you should get the following message:



Click “OK”, and the dialog should look like this:

Multiple Regression

Input
Select Term (5 Total)

Current Model (3 Terms)

Output
Summary

ANOVA - 49 Total Data Points

Y = b0 + b1*X1 + b2*X2 + b3*X1*X1

Term	Coefficient	Std Error	t Statistic	Significance	VIF
Constant	24.89	0.732	34.00	9.688e-34	
X1	5.340	0.341	15.67	7.642e-20	18.37279
X2	20.43	0.714	28.64	1.571e-30	1.028682
X1*X1	-0.141	0.03470	-4.078	0.000183	18.27476

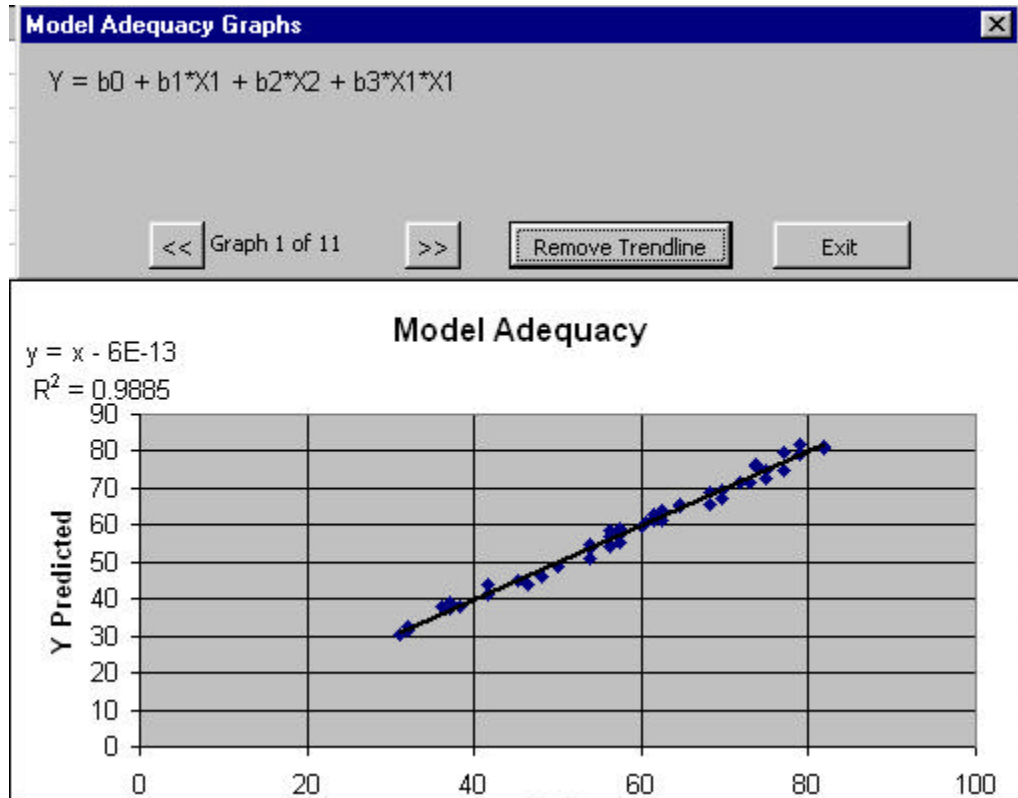
The selected model contains the terms “X1”, “X2”, “X1*X1”, and the constant term or intercept. Note that this model does not generally have higher R^2 terms than the full quadratic model (the R^2 for prediction is only slightly higher), but the F-value is higher (1284) (or, meaning the same, the “F-Significance” value is lower), indicating a more significant model. All the model terms are highly significant, indicated by the very low “Significance” values in the coefficients window.

If you execute the “Fit All” option described further above, you’ll see that the model the “AutoFit” came up with is actually not the best model available in terms of the R^2 -values. However, the 3 “better” models all have insignificant, i.e., redundant terms!

The “Multiple Regression” dialog allows to perform model adequacy checking. The “outlier” button produces a list showing outliers , leverage, and influential cases in our

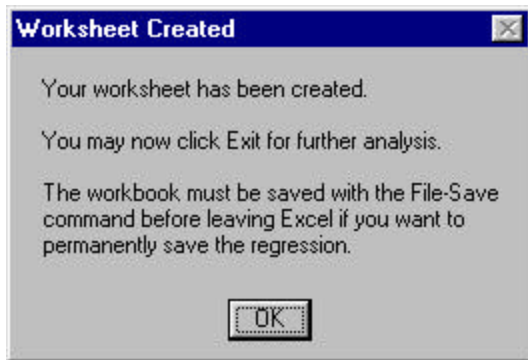
database. The “Graph” button opens another dialog which shows a variety of scatter plots useful for residual analysis.

For example, click the “Graph” button and then “Add Trendline” in the graph dialog. It should look like this:



This graph shows a plot of the y-values predicted by the model (“Y predicted”) vs. the observed “Y” values and the corresponding linear trend line. As you can see, a variety of plots is available which can be selected with the arrow buttons.

So far we only could see the results of the regression analysis in dialog boxes. Now, we will create a permanent Excel output worksheet. Exit the graph dialog shown above and press the “Make XLS” button in the main dialog. After a few seconds, the following message should appear. Click “OK” and then “Exit” in the main dialog.



After exiting the main dialog, the output sheet ("data_1") should be the active window. Note the buttons on the left hand side in the first column. By pressing these buttons, you can perform a series of useful actions:

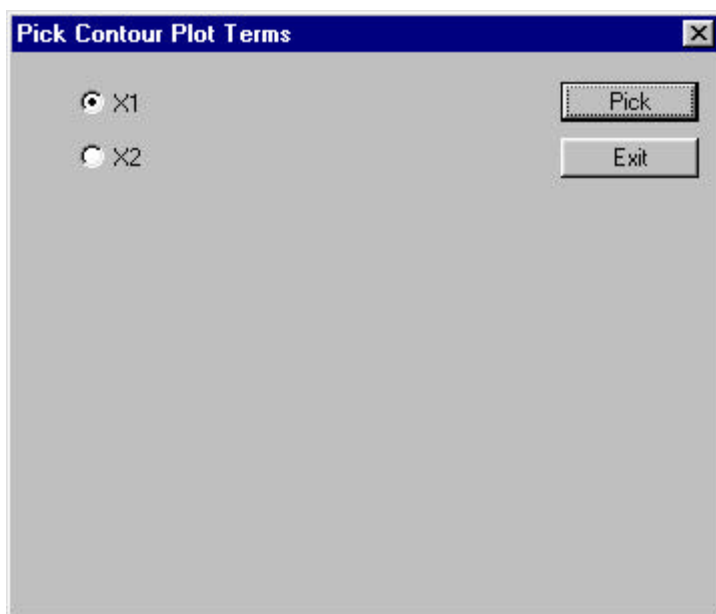
- Reregress the model (goes back to the Main Dialog),
- Delete the output sheet if needed,
- Predict new responses based on new data points,
- see scatter plots similar to the ones described above for residual analysis ("Graph"),
- evaluate a data table including residual analysis for each data point,
- go to a regression coefficients table like the one in the main dialog,
- "optimize", i.e., find a set of inputs which gives a specific output,
- check the confidence ranges for the regression in a scatter plot,
- view the outlier table,
- print selected output ranges from the sheet,
- look at the correlation matrix (R matrix).

Finally, the "surface" button allows you to see a 3D surface of your regression model equation, provided there is more than one variable in your model.

In our example, the equation we arrived at after using "AutoFit" contained "X1" and "X2" (as the squared term). On the output Excel sheet you just created, press the "Surfaces" button. In the next message box, click "OK":



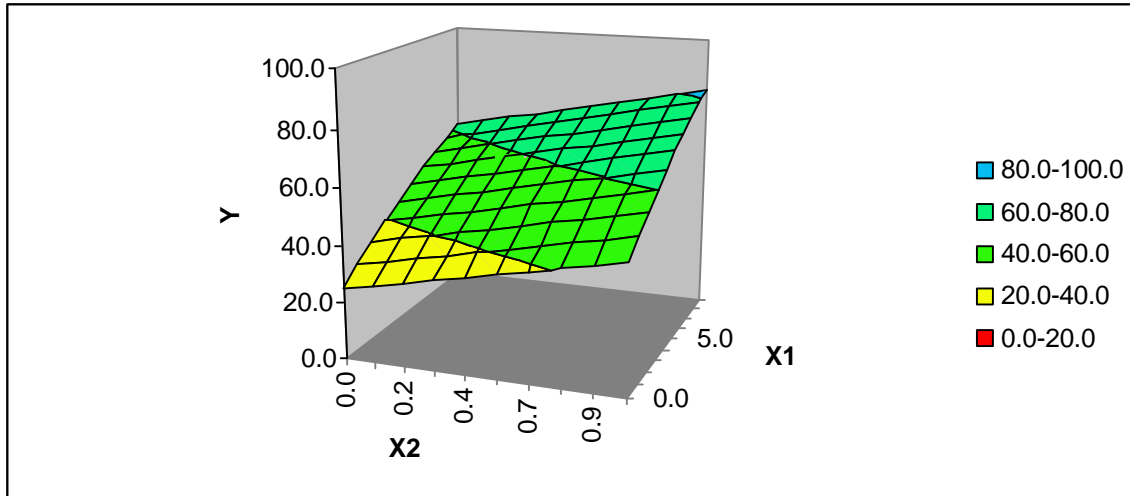
The next dialog shows a list of the available variables to plot. In our case, there is only one 3D plot possible: the reponse ("Y") vs. "X1" and "X2".



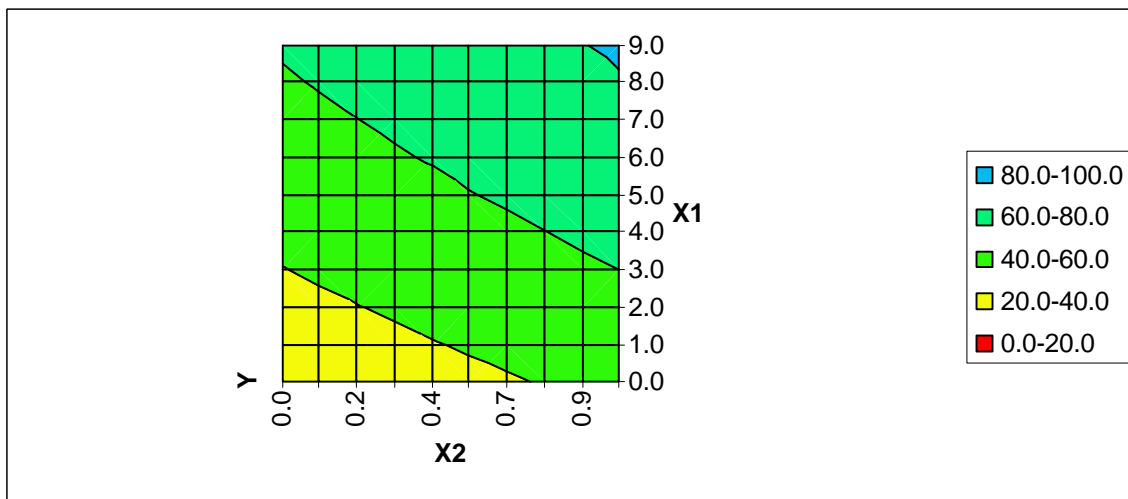
Select "X1" and click the "Pick" button. In the next message box, click "OK".



Select "X2", and click the "Pick" button again. Essential Regression creates the surface plot, and you should see the following graph on your worksheet:



If you click the “Contour” button above the graph, you get the 2D representation of the surface, a contour plot. The “Contour” button changes to a “3d” button. Pressing it brings back a surface plot.



You can rotate the graphs by using the “<” and “>” keys. Also, you can increase or decrease the number of levels by clicking “+” or “-“, respectively. In our example, we have used the “+” key a few times to bring out more colors.

If your model has more than 2 variables, you will find another button above the graph area with the caption “movie”. The “movie” feature allows you to incrementally change the value of one variable while plotting the response vs. two other variables. If you loop

through these changes, the effect resembles an animation or movie with the surface changing according to the value of the changed variable.

Pressing the “Back” button at any time takes you back to the starting point, i.e., the upper left corner of the worksheet.

Make sure that you save the Excel worksheet before closing Excel if you wish to keep the output. Basically, this sheet generated by Essential Regression (ER) is a standard Excel worksheet linked to ER through the added buttons.

This tutorial is intended to lead you through a relatively simple regression analysis while emphasizing the features of Essential Regression which allow for a quick assessment of the model. There are many more features explained in detail in the previous chapters.

7.5 *Unloading Essential Regression*

In Excel simply select the **Regress, Unload** menu option this will close Essential Regression and remove the **Regress** menu from Excel.

7.6 Loading Essential Experimental Design into MS Excel

From within MS Excel

In Excel, with at least one empty workbook open, select the **File, Open** menu. Locate EED22.xla in c:\eregress (or the directory you installed the program into) and open it. This will start the Add-in and, after an introductory screen, add a new **DOE** menu to the Excel main menu bar between the **File** and **View** menus.

From outside MS Excel

In Windows 95, select **Start, Programs→Essential Regression →Essential Experimental Design**. In Windows 3.x, double-click the Essential Experimental Design Icon in the Essential Regression Program Group.

Like any other Excel file, Essential Experimental Design (EED22.xla) can also be opened directly in the File Manager (Windows 3.x) or Explorer (Windows 95) .

MS Excel will start up (or simply become the active application if it was started up before), and, after the introductory screens, a new **DOE** menu will be added to the Excel main menu bar between the **File** and **View** menus.

7.7 Creating a simple experimental design and analyzing it with Essential Experimental Design (EED)

We assume EED is loaded and the **DOE** menu is visible. First, select the **Design An Experiment** option in the **DOE** menu. This brings up the Design an Experiment Dialog. We are going to create a circumscribed **central composite design** (CCD) with 3 factors and 4 center points to assess curvature and experimental error. Please make the appropriate selections. The dialog should look like this before you continue:

Design an Experiment

Input Data

Number of Factors: 3 # of Centerpoints: 4 Number of Responses: 1

☒ All Factors are Quantitative ☒ Show Aliasing (if applicable) ☒ Randomize Worksheet

2 Level Screening

Many Factors

☐ Fractional Factorial Res 3 4 Runs

☐ Plackett - Burman 12 Runs

Higher Resolution

☐ Fractional Factorial Res 4 8 Runs

☐ Fractional Factorial Res 5 8 Runs

☐ Full Factorial 8 Runs

Response Surface Designs

Second Order Models

☒ Central Composite 14 Runs

☐ Box - Behnken 12 Runs

Central Composite Options

☒ Circumscribed (Min & Max=Star Points)

☐ Inscribed (Star Points outside Min & Max)

☐ Face Centered

Current Design (18 Runs)

Central Composite design for 3 Factors

Model is quadratic

14 model runs and 4 centerpoints

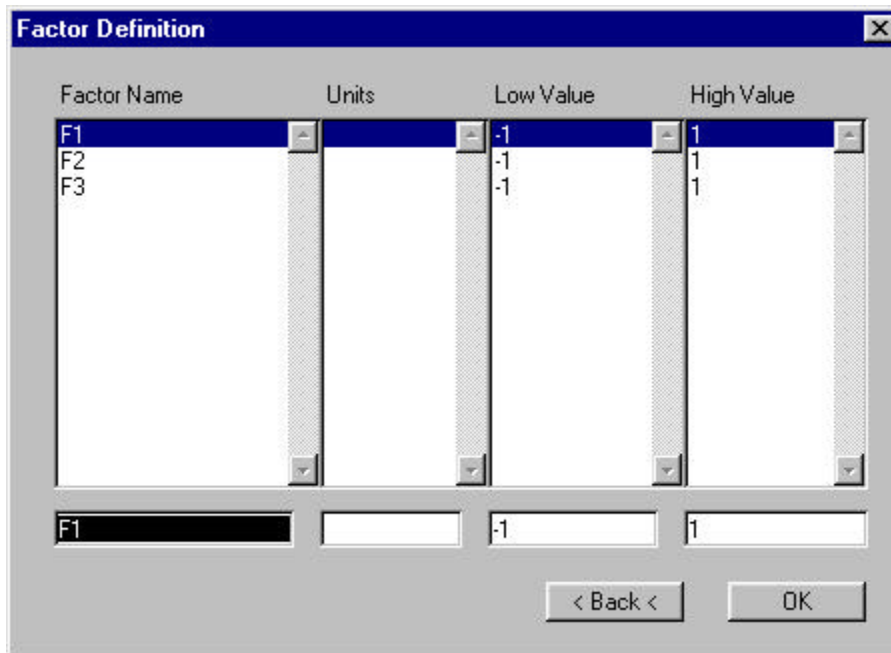
Exit

Help

Make DOE

In the colored section at the bottom, the dialog shows that our design has 18 runs or experiments (including the center points), and that the underlying model has quadratic terms.

Press the “Make DOE” button. EED creates the “Aliasing” worksheets giving information how certain effects are aliased with others, and the Factor Definition Dialog will be displayed:



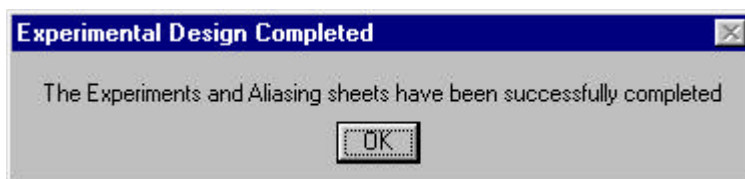
The Factor Definition dialog box is shown with a table of factors. The table has four columns: Factor Name, Units, Low Value, and High Value. The first row is highlighted in blue and contains F1, an empty unit field, -1, and 1. The second row contains F2, an empty unit field, -1, and 1. The third row contains F3, an empty unit field, -1, and 1. Below the table, there are input fields for Factor Name (F1), Units (empty), Low Value (-1), and High Value (1). At the bottom right, there are buttons for '< Back <' and 'OK'.

Factor Name	Units	Low Value	High Value
F1		-1	1
F2		-1	1
F3		-1	1

Factor Name: F1 Units: Low Value: -1 High Value: 1

< Back < OK

Here, you can set the lows and highs for the design factors. For our purposes, simply accept -1 and 1 as low and high settings for the design and continue with “OK”. EED will create the “Experiments” worksheet and the following confirmation message will appear. Simply press “OK” to continue:

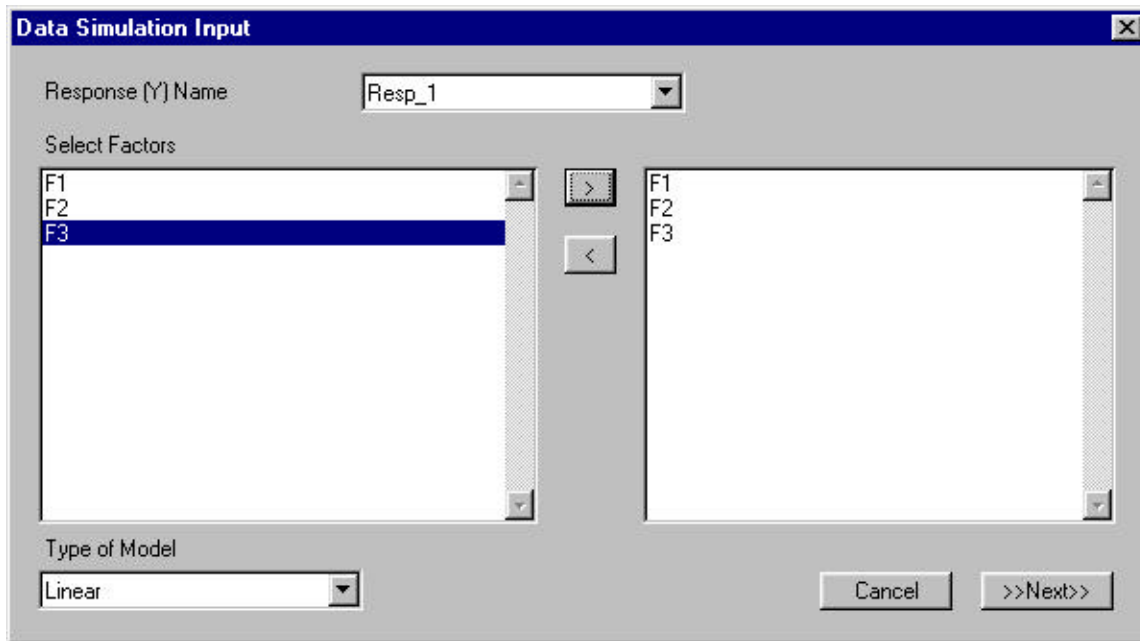


The Experimental Design Completed dialog box is shown with a message: 'The Experiments and Aliasing sheets have been successfully completed'. At the bottom center, there is an 'OK' button.

The Experiments and Aliasing sheets have been successfully completed

OK

In the “Experiments” worksheet, you’ll find information about your design and the underlying model. Let us pretend we would conduct the 18 experiments necessary to analyze the model. In EED, we can simulate this process. In the **DOE** menu, select **Simulate Data**. This will bring up the Data Simulation Input Dialog. Accept “Resp_1” as the response name and select the Factors F1, F2, and F3 as the model factors. Further, let’s assume we have a linear model (you can change the model type in the “Type of Model” list box at the bottom of the dialog). The dialog should now look like this:

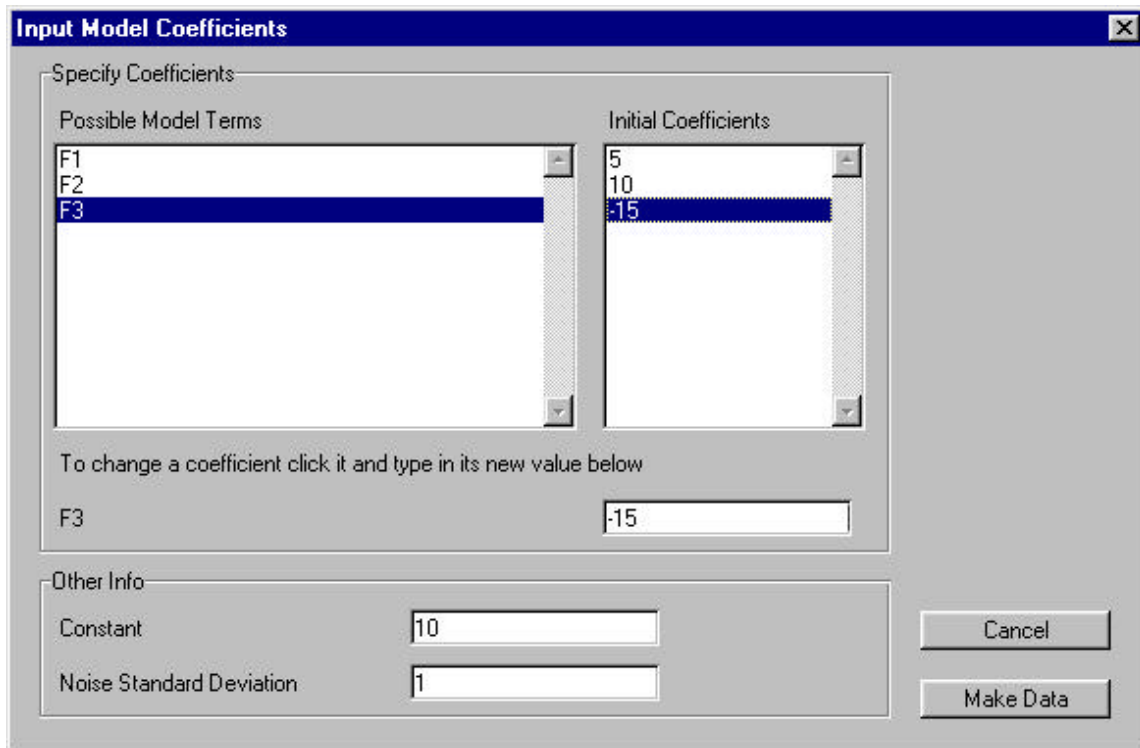


The dialog box is titled "Data Simulation Input". It contains the following elements:

- Response (Y) Name:** A text box containing "Resp_1".
- Select Factors:** Two list boxes with arrows between them. The left list box contains "F1", "F2", and "F3", with "F3" selected. The right list box contains "F1", "F2", and "F3".
- Type of Model:** A dropdown menu set to "Linear".
- Buttons:** "Cancel" and ">>Next>>" buttons at the bottom right.

Press the ">>Next>>" button. This will bring up the Input Model Coefficients Dialog. Select "F1" (factor 1) in the window listing the "Possible Model Terms". Then type the value for the coefficient for factor 1 into the edit box shown below the "Initial Coefficients" window. The cursor should be activated in this edit box by default so that, after selecting "F1", you should be able to type directly. Enter "5" as the value for the coefficient.

Repeat these steps for the factors F2 and F3 using "10" and "-15" as coefficients. After that, enter "10" as a value for the constant term in the model and leave the noise standard deviation at 1. The dialog should then look like this:



The dialog box is titled "Input Model Coefficients". It contains two main sections: "Specify Coefficients" and "Other Info".

Specify Coefficients:

- Possible Model Terms:** A list box containing F1, F2, and F3. F3 is currently selected.
- Initial Coefficients:** A list box containing 5, 10, and -15. -15 is currently selected.

Below these lists, a text instruction reads: "To change a coefficient click it and type in its new value below".

Other Info:

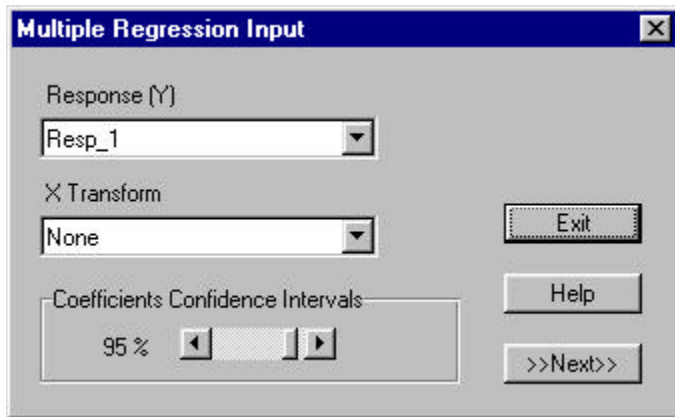
- Constant:** A text box containing the value 10.
- Noise Standard Deviation:** A text box containing the value 1.

At the bottom right, there are two buttons: "Cancel" and "Make Data".

This concludes the model definition. What we have done is to simulate a linear regression model as the basis for our experimental design. Press the “Make Data” button, and EED will calculate “responses” for each experiment on the “Experiments” worksheet. Note that the data table now contains data in the response column.

Let’s pretend we do not know the exact model equation which we just have used to calculate our data. The next step will be a multiple regression to come up with a model which describes our data best.

To perform this task, select **Analyze Design** in the **DOE** menu (the “Experiments” worksheet should be the active sheet when doing this). This will launch Essential Regression in “EED mode”, and a Multiple Regression Input Dialog different from the one shown in Chapter 7.4 will come up:



By default, this dialog selects “Resp_1” as the column of the data table containing the response. Accept the defaults and click “>>Next>>”. This will bring up the Multiple Regression Main Dialog (see Chapter 7.4). At this point, simply click the “AutoFit” button and have Essential Regression find the best model. The outcome depends on the data you simulated as described previously. The random error or noise term we introduced can lead to different results as far as the optimized model is concerned. However, the model you end up with should contain F1, F2, and F3 as highly significant factors and possibly another, higher order term with less significance.

You could click the “Fit all” button in the Multiple Regression Main Dialog and find out which model is the “best”, based on R^2 and R^2 adjusted. If you limit the number of factors to 4, this should not take unreasonably long.

Also, note that you are now in Essential Regression (ER). You can use all the features of ER including 3D- graphing. Since you have 3 variables, you can use the “movie” feature in the surface plot area of the output sheet (described in chapter 7.4).

7.8 Unloading Essential Experimental Design

In Excel simply select the **DOE, Unload** menu option this will close Essential Experimental Design and remove the **DOE** menu from Excel.

8. Literature

This book was meant as a supplement to the Essential Regression and Essential Experimental Design Add-Ins. We put in as much information about Linear Regression and DOE as we thought was reasonable to enable any user of this software to perform a meaningful analysis. We are aware that, by doing so, we had to cut corners here and there and sometimes even leave out topics which, in the eyes of a really serious reader (or a statistician), should have been discussed.

For people interested in the fundamentals and the mathematical details, we recommend studying some of the following publications. We, not being statisticians by trade, necessarily had to distill much of the information presented in these literature references into this book and, hopefully did not make too many mistakes in doing so. We think everybody applying statistics on a regular basis should peruse some of the books listed below:

Douglas C. Montgomery and Elizabeth A. Peck, “Introduction to Linear Regression Analysis”, 2nd Ed. 1992, John Wiley & Sons, Inc., New York, NY (ISBN 0-471-53387-4).

Raymond H. Myers, Douglas C. Montgomery, “Response Surface Methodology, 1995, John Wiley & Sons, Inc., New York, NY (ISBN 0-471-58100-3).

Douglas C. Montgomery, “Design and Analysis of Experiments”, 3rd Ed., 1991, John Wiley & Sons, New York, NY (ISBN 0-471-52000-4).

Lyman Ott, “An Introduction to Statistical Methods and Data Analysis”, 3rd Ed. 1988, PWS-Kent Publishing Co. Boston, MA (ISBN 0-534-91926-X).

Jay L. Devore, “Probability and Statistics for Engineers and the Sciences”, 3rd Ed. 1991, Brooks/Cole Publishing Co., Pacific Grove, CA (ISBN 0-534-14352-0)

For readers interested in the details of nonlinear regression analysis:

Douglas M. Bates, Donald G. Watts, *Nonlinear Regression Analysis and its Applications*,
John Wiley & Sons, Inc., New York, NY (ISBN 0-471-81643-4).

9. Index

A

adjusted R^2 , 52, 61, 62, 63, 66, 130, 135
alias, 102, 109, 111, 113
Analysis of Variance, 44
ANOVA table, 38, 44, 57, 68, 75, 135
Autocorrelation, 55

C

centering, 14, 26, 33
centerpoints, 91, 92, 93, 99, 100, 106, 108, 112,
115, 118, 120, 121, 122, 124, 127, 128, 129
central composite design (CCD), 144
Coefficient of Multiple Determination, 62
confounded, 129
Cook's distance, 59
correlation, 45, 46, 54, 55, 63, 75, 89, 140
COVRATIO, 52
curvature, 41, 92, 106, 107, 108, 129, 144

D

defining relation, 94, 95
defining word, 94, 96, 101, 102, 103
dependent variable, 17, 20, 25, 26, 27, 31, 85
DFBETAS, 51
DFFITS, 51, 76

E

EED, 28, 90, 91, 94, 96, 97, 98, 101, 102, 108, 115,
116, 122, 123, 124, 125, 127, 128, 129, 130, 144,
145, 146, 148
ER, 25, 26, 27, 28, 30, 31, 33, 41, 48, 53, 54, 60, 69,
90, 108, 130, 143, 149

Error Sum of Squares, 35, 36, 37, 40, 41, 45, 52, 53
Essential Experimental Design, 28, 90, 91, 93, 98,
130, 131, 144, 149, 150

F

face centered CCD, 124, 125
factors, 17, 26, 30, 44, 54, 61, 89, 90, 91, 92, 93, 94,
95, 96, 97, 99, 100, 101, 108, 109, 111, 113, 116,
117, 118, 119, 120, 121, 122, 123, 124, 125, 126,
127, 128, 129, 133, 144, 146, 147, 149
fitted value, 51
Forward Selection, 39, 63, 64, 65, 66, 67
fractionate, 94

H

hat matrix, 22, 49, 50, 59, 73, 76, 77, 79
hypothesis testing, 55

I

independent variable, 13, 14, 15, 17, 18, 19, 20, 21,
25, 26, 27, 30, 31, 33, 34, 35, 36, 40, 43, 53, 85
influential observation, 59, 75, 77
Input Dialog, 25, 26, 41, 43, 66, 133, 146, 148
inscribed CCD, 116, 117, 119, 120, 121
interaction, 16, 26, 31, 43, 91, 94, 96, 111, 112, 113,
114, 134, 136

L

lack of fit, 23, 40, 41, 53, 130
linear model, 35, 104, 106, 114, 134, 135, 146
Linear Regression, 13, 16, 22, 26, 27, 34, 35, 43, 45,
46, 47, 48, 54, 72, 150

M

Main Dialog, 33, 41, 42, 44, 58, 60, 66, 68, 74, 77, 84, 89, 133, 134, 137, 140, 149

main effects, 91, 92, 94, 96, 109

multicollinearity, 21, 54, 58

Multiple Regression Input Dialog, 41, 133, 148

N

normal equations, 19, 20, 21

O

orthogonal, 53, 54, 113, 114, 115

outliers, 48, 49, 50, 59, 61, 65, 68, 69, 75, 77, 138

P

polynomial model, 30, 46, 54

Polynomial Regression Input Dialog, 26

prediction, 24, 49, 52, 61, 72, 76, 79, 84, 115, 121, 122, 124, 135, 138

prediction variance, 115, 122

PRESS, 49, 50, 52, 76

principle fraction, 94, 95

probability, 23, 24, 27, 33, 37, 39, 40, 48, 56, 57, 63, 72, 79, 84, 107, 136

pure error, 40

Q

quadratic model, 16, 116, 121, 135, 136, 137, 138

qualitative, 89, 93

quantitative, 89, 90, 99

R

R^2 , 45, 46, 47, 52, 61, 62, 63, 66, 68, 84, 130, 135, 136, 137, 138, 149

R^2 adjusted, 137, 149

R^2 for prediction, 52, 61, 135, 138

R^2 adjusted, 46

randomization, 129

Regression analysis, 59, 63, 106

regression coefficient, 13, 15, 20, 21, 22, 23, 24, 25, 27, 28, 33, 34, 38, 39, 44, 45, 51, 54, 57, 58, 68, 75, 135, 140

regression model, 15, 16, 17, 19, 22, 23, 25, 26, 30, 31, 32, 33, 34, 35, 36, 37, 38, 41, 42, 43, 44, 45, 46, 47, 52, 53, 58, 61, 62, 63, 64, 68, 70, 72, 73, 75, 82, 84, 86, 88, 90, 91, 93, 94, 95, 96, 100, 106, 112, 113, 115, 116, 118, 121, 129, 134, 140, 148

regressor, 12, 13, 14, 17, 21, 27, 30, 32, 33, 38, 40, 41, 46, 47, 50, 54, 55, 57, 63, 64, 66, 67, 70, 72, 76, 77, 78, 79, 81, 82, 84, 85, 86, 88, 132

repeat measure, 129

replicate, 40, 41, 53, 129

Residual Sum of Squares, 44

residuals, 36, 47, 48, 49, 50, 52, 55, 59, 60, 65, 68, 69, 70, 75, 76

resolution, 96, 97, 99, 103, 111, 113, 114, 115, 120, 121, 127

Response Surface, 150

response surface modeling, 90, 121

response variable, 16, 17, 34, 41, 55, 65, 133

RMS, 47

rotatability, 121, 124

S

saturation, 109, 111

scaling, 13, 14, 15, 27, 32, 33, 50, 54, 122

scatter diagram, 25

scatter plot, 60, 82, 83, 84, 139, 140

screening, 90, 93, 99, 108, 111, 113, 115, 116, 118, 120, 123, 127

serial correlation, 55

Simple Linear Regression, 45, 72

SSE, 18, 19, 20, 22, 35, 36, 37, 38, 40, 41, 44, 45, 47

SSR, 35, 36, 37, 38, 44, 45, 46, 64

Standard Error, 23, 24, 40, 76

standard normal distribution, 104

standardized regression coefficient, 15

stepwise regression, 58, 113, 136, 137

Sum of Squares, 35, 36, 37, 38, 40, 41, 44, 45, 52,
53, 64

s_{yy} , 35, 36, 38, 45, 46

T

transformation, 14, 17, 20, 32, 33, 48, 59, 65, 68, 70,
71, 106

V

variables, 12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 25,
26, 28, 30, 31, 32, 33, 34, 35, 36, 38, 40, 41, 42,
43, 44, 46, 50, 53, 55, 61, 62, 63, 64, 65, 66, 67,
68, 70, 72, 76, 77, 78, 80, 82, 83, 84, 85, 86, 87,
89, 93, 103, 132, 133, 135, 141, 143, 149

variance, 14, 15, 22, 23, 37, 38, 44, 48, 50, 53, 54,
55, 59, 62, 65, 70, 71, 114, 115, 121, 122, 124

VIFs, 44, 54, 58