Multiple-Linear, Polynomial and Nonlinear Regression

Basic Concepts (1)

Let us assume that there is a set of *N* data points of a dependent variable y_i versus x_{1i}, x_{2i}, … x_{ni}, where $\mathbf{x}_1$, $\mathbf{x}_2$,… $\mathbf{x}_n$ are n independent (explanatory) variables. A particular model to be fitted to the data is of the form

$$y_i = g(x_{1i}, x_{2i} \ldots x_{ni}, \beta_0, \beta_1 \ldots \beta_m) \qquad (1)$$

where $\beta_0, \beta_1 \cdots \beta_m$ are *m*+1 parameters of the model. The least-squares error approach is most often used to find the parameters of Equation (1).

The statistical assumption behind the least-squares error method for parameter estimation is that the measured value of the dependent variable has a deterministic and a stochastic part. The stochastic part is often denoted as an error, ε_i .

$$y_i = g(x_{1i}, x_{2i} \ldots x_{ni}, \beta_0, \beta_1 \ldots \beta_m) \pm \varepsilon_i \qquad (1a)$$

It is further assumed that the origin of ε_i is measurement error, which is randomly distributed.

Multiple-Linear, Polynomial and Nonlinear Regression

Basic Concepts (2)

An infinite number of measurements would be required to obtain the true values of the parameters $\beta_0, \beta_1 \cdots \beta_m$ .Because a sample always contains a finite number of measurements, the calculated parameters are always approximations for the true values. They are denoted with a circumflex. Thus, $\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_m$ are the calculated values of the parameters and $\hat{y}_i$ is the calculated estimate for the dependent variable .

In the least-squares error approach, the estimates $\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_m$ are found so that they minimize the following function:

$$F = \sum_{i=1}^{N} \left[y_i - g(x_{1i}, x_{2i} \ldots x_{ni}, \beta_0, \beta_1 \ldots \beta_m) \right]^2 \qquad (2)$$

where *F* is the sum of squares of the errors.The particular mathematical technique of finding the set of the parameter values that minimizes the function *F* depends on the form of the function $g(\mathbf{x}_i, \boldsymbol{\beta})$. If the parameters appear in linear expressions in the function *g* (in multiple-linear and polynomial regressions, for example), the minimization can be carried out by solving a system of linear equations (the normal equations). Often models where the parameters appear in nonlinear expressions can be transformed to linear models by transformation of variables.

Graphic information for checking the quality of the fit .

An assessment of the quality of the fit of a particular model and comparison between different models is based on graphic and numeric information.

The measured (y_i) and the calculated ($\hat{y}_i$) values of the dependent variable can be plotted versus x_i (if there is a single independent variable) or versus i, the point number (if there are several independent variables). The distance between the experimental values and the calculated curve can serve as an indication for the quality of the fit. These distances are amplified using the "residual plot". In this plot the model error is plotted (usually versus y_i) , where

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad (3)$$

A random distribution of the residuals around zero indicates that the model represents correctly the set of data. A definite trend or pattern in the residual plot may indicate either lack of fit of the model or that the assumed error distribution for the data (random error distribution in **y**) is not correct.

Numeric information for checking the quality of the fit (1).

The most frequently used numeric indicator of the quality of the fit is the standard error of the estimate which represents the sample variance, and given by

$$s^2 = \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{N - (m+1)} \qquad (4)$$

Thus the sample variance is the sum of squares of errors divided by the degrees of freedom (where the number of parameters, m+1, is subtracted from the number of data points, N) and is a measure for the variability of the actual y_i values from the predicted $\hat{y}_i$ values. Smaller variance means a better fir of the model to the data. It should be emphasized that when the sample variance is used for comparison of different models, the same independent variable (transformed or non-transformed) should be used in Equation (4) for all the models. The variance is an un-scaled variable which can take any value from zero to infinity. Consequently the variance alone cannot be used for judging the goodness of fit between the data and a respective model.

Numeric information for checking the quality of the fit (2).

The linear correlation coefficient (R^2) is often used to judge the quality of the fit between the regression model and the experimental data. The correlation coefficient represents the ratio between the sum of squares about the mean due to regression to the total sum of squares, and is obtained by

$$R^2 = \frac{\sum_{i=1}^{N} \left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2} \qquad (5)$$

where $\overline{y}$ is the sample mean of the dependent variable. The value of *R* is bounded: $0 \le R \le 1$. If *R* is close to 1 there is a strong correlation between the variables, whereas a value close to zero indicates a weak or no correlation.

Confidence intervals on the parameter values are very useful indicators of the fit between the model and the data. The discussion concerning the confidence intervals is postponed after discussing the solution techniques of the normal equation

Simple Linear Regression (1)

The simplest example of least squares approximation is fitting a straight line to a set of paired measurements (or observations), (x_1, y_1), (x_2, y_2)… (x_N, y_N).

$$\hat{y}_i = \beta_0 + \beta_1 x_i \qquad (7)$$

Introducing this model into Equation (2) yields

$$F = \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 \qquad (8)$$

The criterion for optimality requires that $\frac{\partial F}{\partial \beta_0} = 0$ and $\frac{\partial F}{\partial \beta_1} = 0$ Thus,

$$\begin{aligned} \frac{\partial F}{\partial \beta_0} &= \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right](-2) = 0 \\ \frac{\partial F}{\partial \beta_1} &= \sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right](-2x_i) = 0 \end{aligned} \qquad (9)$$

Simple Linear Regression (2)

After rearrangement

$$\begin{aligned} \beta_0 N + \beta_1 \sum x_i &= \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i \end{aligned} \qquad (10)$$

Using Cramer's rule to solve for β_0 and β_1 yields

$$\begin{aligned} \beta_0 &= \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2} \\ \beta_1 &= \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2} \end{aligned} \qquad (11)$$

Example 1. Fitting a Straight Line to Thermal Conductivity Data

Thermal conductivity of low-pressure gases can be fairly well correlated, over small temperature ranges, with a linear equation (straight line). A linear equation should be fitted to the thermal conductivity data of air shown in Table 1 and the appropriateness of the linear model should be assessed.

Table 1. Thermal conductivity of Air[a]

No.	Temperature	Thermal Conductivity*10^6
	F °	cal/s•cm•°C
1	-40	50.09
2	-20	52.15
3	0	54.22
4	20	56.24
5	40	58.31
6	60	60.34
7	80	62.2
8	100	64.22
9	120	66.04

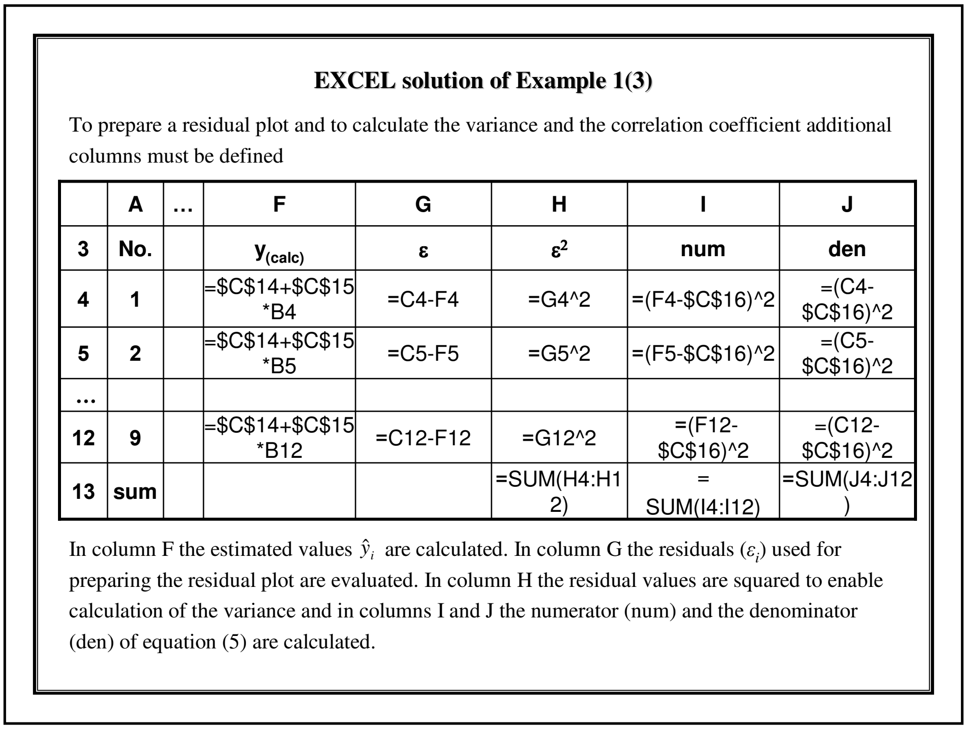
EXCEL solution of Example 1 (1).

	A	B	C	D	E
3	**No.**	**x**	**y**	**x²**	**xy**
4	**1**	-40	50.09	=B4^2	=B4*C4
5	**2**	-20	52.15	=B5^2	=B5*C5
	...				
12	**9**	120	66.04	=B12^2	=B12*C12
13	**sum**	=SUM(B4:B12)	=SUM(C4:C12)	=SUM(D4:D12)	=SUM(E4:E12)
14	**β_0**		=(C13*D13-B13*E13)/(A12*D13-B13^2)		
15	**β_1**		=(A12*E13-B13*C13)/(A12*D13-B13^2)		

The x_i values are stored in column B, the y_i values are stored in column C, the x_i^2 values are calculated in column D and the x_iy_i values are calculated in column E. The respective sums are calculated in row 13. In cells C14 and C15 the various terms are introduced into Equation (11) in order to calculate β_0 and β_1. The numerical results obtained are shown below.

EXCEL solution of Example 1 (2)

	A	B	C	D	E
3	**No.**	**x**	**y**	**x²**	**xy**
4	**1**	-40	50.09	1600	-2003.6
5	**2**	-20	52.15	400	-1043
	...				
12	**9**	120	66.04	14400	7924.8
13	**sum**	360	523.81	38400	23354
14	**β_0**		54.1988		
15	**β_1**		0.1001		

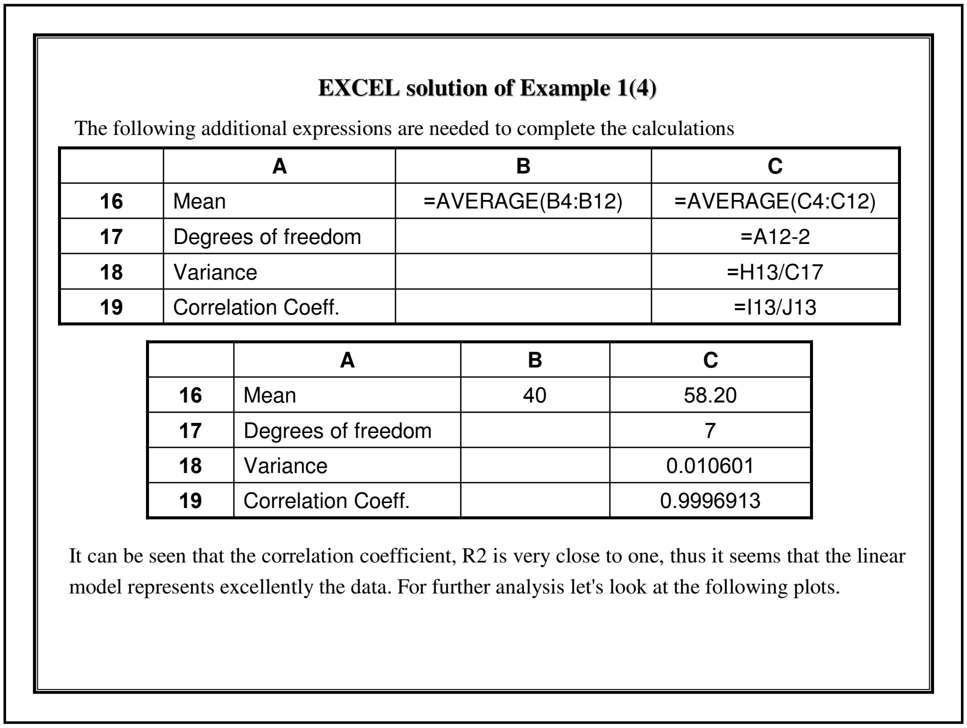


EXCEL solution of Example 1(3)

To prepare a residual plot and to calculate the variance and the correlation coefficient additional columns must be defined

	A	…	F	G	H	I	J
3	**No.**		**$y_{(calc)}$**	**ε**	**$ε^2$**	**num**	**den**
4	**1**		=$C$14+$C$15*B4	=C4-F4	=G4^2	=(F4-$C$16)^2	=(C4-$C$16)^2
5	**2**		=$C$14+$C$15*B5	=C5-F5	=G5^2	=(F5-$C$16)^2	=(C5-$C$16)^2
…							
12	**9**		=$C$14+$C$15*B12	=C12-F12	=G12^2	=(F12-$C$16)^2	=(C12-$C$16)^2
13	**sum**				=SUM(H4:H12)	= SUM(I4:I12)	=SUM(J4:J12)

In column F the estimated values $\hat{y}_i$ are calculated. In column G the residuals (ε_i) used for preparing the residual plot are evaluated. In column H the residual values are squared to enable calculation of the variance and in columns I and J the numerator (num) and the denominator (den) of equation (5) are calculated.



EXCEL solution of Example 1(4)

The following additional expressions are needed to complete the calculations

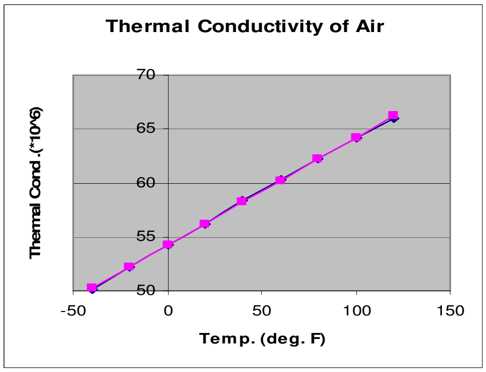
	A	B	C
16	Mean	=AVERAGE(B4:B12)	=AVERAGE(C4:C12)
17	Degrees of freedom		=A12-2
18	Variance		=H13/C17
19	Correlation Coeff.		=I13/J13

	A	B	C
16	Mean	40	58.20
17	Degrees of freedom		7
18	Variance		0.010601
19	Correlation Coeff.		0.9996913

It can be seen that the correlation coefficient, R2 is very close to one, thus it seems that the linear model represents excellently the data. For further analysis let's look at the following plots.

EXCEL solution of Example 1

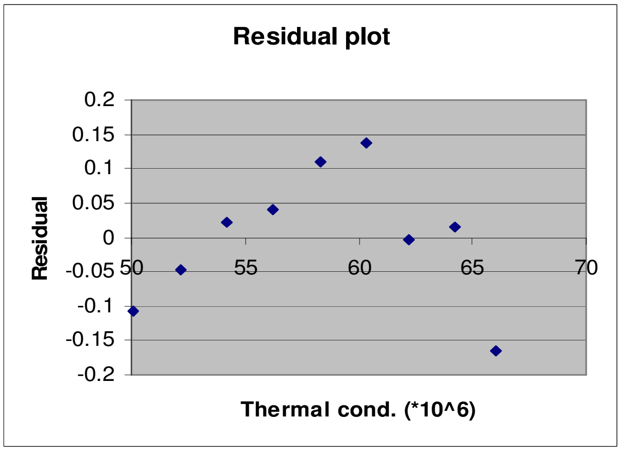
Measured and Calculated Values of Thermal Conductivity



This plot also indicates very good fit. The calculated and experimental points are actually indiscernible in this plot. But in the following residual plot the errors are not randomly distributed around zero indicating that the model can probably be further improved.

EXCEL solution of Example 1

Residual Plot



Multiple Linear Regression (1)

Let us develop the equations to be solved for the case of two independent variables where a linear model is to be fitted into a set of measurements (or observations), (x_{11}, x_{21}, y_1), (x_{12}, x_{22}, y_2)… (x$_{1N}$, x$_{2N}$, y_N). The linear model is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \qquad (12)$$

Introducing this model into Equation (2) yields

$$F = \sum_{i=1}^{N} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) \right]^2 \qquad (13)$$

The criterion for optimality requires that $\frac{\partial F}{\partial \beta_0} = 0$, $\frac{\partial F}{\partial \beta_1} = 0$ and $\frac{\partial F}{\partial \beta_2} = 0$. Thus,

$$\begin{aligned}
\frac{\partial F}{\partial \beta_0} &= \sum_{i=1}^{N} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) \right](-2) = 0 \\
\frac{\partial F}{\partial \beta_1} &= \sum_{i=1}^{N} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) \right](-2x_{1i}) = 0 \qquad (14) \\
\frac{\partial F}{\partial \beta_2} &= \sum_{i=1}^{N} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) \right](-2x_{2i}) = 0
\end{aligned}$$

Multiple Linear Regression (2)

After rearrangement and bringing into matrix-vector form we get

$$\begin{bmatrix} N & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{bmatrix} \qquad (15)$$

For the general case of m independent (explanatory) variables the matrix form of the normal equations can be more easily obtained by defining the following matrices

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1N} & \cdots & x_{mN} \end{bmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \qquad (16)$$

The normal equation is defined

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y} \qquad (17)$$

Multiple Linear Regression (3)

Confidence Intervals

This is a system of linear equations is solved for the *m*+1 coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ … $\hat{\beta}_m$.

If there is no free parameter in the model (thus $\beta_0 = 0$) then the first column of the numbers 1 (one) should be removed from the matrix **X** and the first element, β_0 should be removed from the vector **β**. It should be emphasized that in such case the number of parameters in the model is *m* (instead of m+1, when there is a free parameter).

Solution of the system of equations (17) provides estimates for the parameter values. The uncertainty in these approximate parameter values can be estimates using the definition of the confidence intervals

$$\hat{\beta}_i - ts\sqrt{a_{ii}} \leq \beta_i \leq \hat{\beta}_i + ts\sqrt{a_{ii}} \quad i = 0,1\ldots m \qquad (18)$$

where *t* is the statistical *t*-distribution value corresponding to the degrees of freedom and the % confidence selected, $s = \sqrt{s^2}$ the standard deviation (square root of the variance) and a_{ii} is the i^{th} diagonal element of the $\mathbf{X}^T\mathbf{X}$ matrix. The 95% confidence intervals are used the most often.

Confidence Intervals (2)

Confidence intervals are very useful indicators of the fit between the model and the data. A better model fit and more precise data lead to narrow confidence intervals, while a poor model fit and/or imprecise data cause wide confidence intervals. Furthermore, confidence intervals which are larger (in absolute value) than the respective parameter values often indicate that the model contains superfluous parameters/ explanatory variables.

Table 2. t-values corresponding to 95%confidence and ν degrees of freedom

ν	t-value	ν	t-value	ν	t-value
1	12.7062	16	2.1199	31	2.0395
2	4.3027	17	2.1098	32	2.0369
:					
14	2.1448	29	2.0452	44	2.0154
15	2.1315	30	2.0423	45	2.0141

Example 2. Heat Evolved During Hardening Of Portland Cement

Woods *et al*(1932) investigated the integral heat of hardening of cement as a function of composition. The independent variables represent weight percent of the clinker compounds: x_1-tricalcium aluminate ($3CaO \cdot Al_2O_3$), x_2-tricalcium silicate ($3CaO \cdot SiO_2$), x_3-tetracalcium alumino-ferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$), and x_4- β-dicalcium silicate ($3CaO \cdot SiO_2$). The dependent variable, y is the total heat evolved (in calories per gram cement) in a 180-day period.

No.	x1	x2	x3	x4	y
1	7	26	6	60	78.7
2	1	29	15	52	74.3
:					
12	11	66	9	12	113.3
13	10	68	8	12	109.4

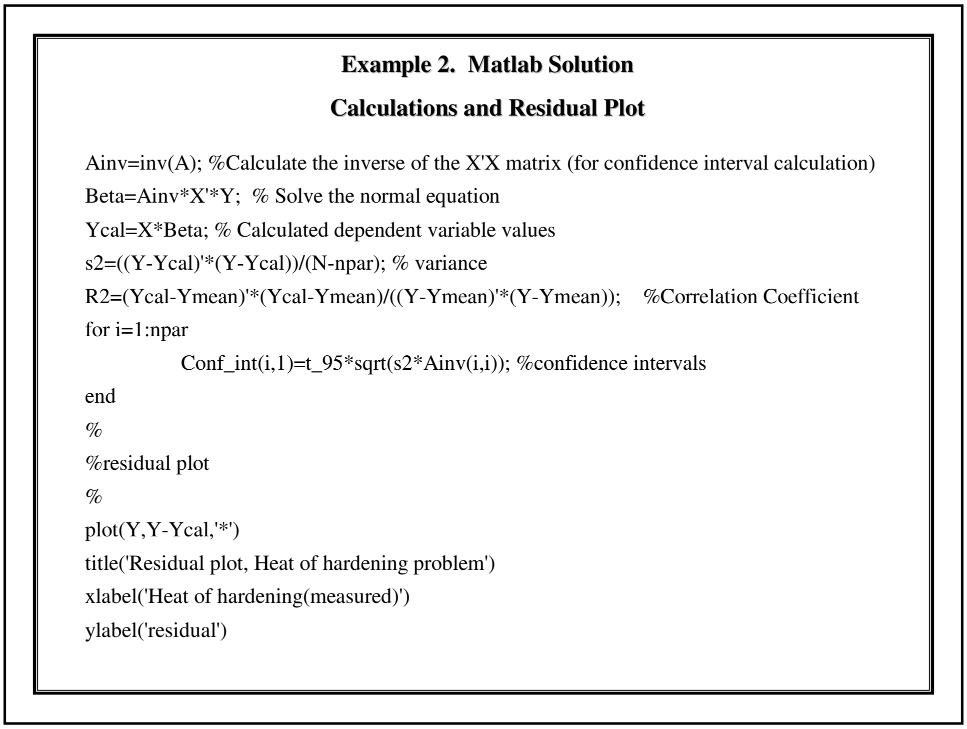
Calculate the coefficients of a linear model representation of y as function of x_1, x_2, x_3, and x_4, calculate the variance and the correlation coefficient R^2 and the confidence intervals. Prepare a residual plot.

Consider the cases when the model includes and does not include a free parameter.

Example 2. Matlab Solution

Data Input and Normal Matrix

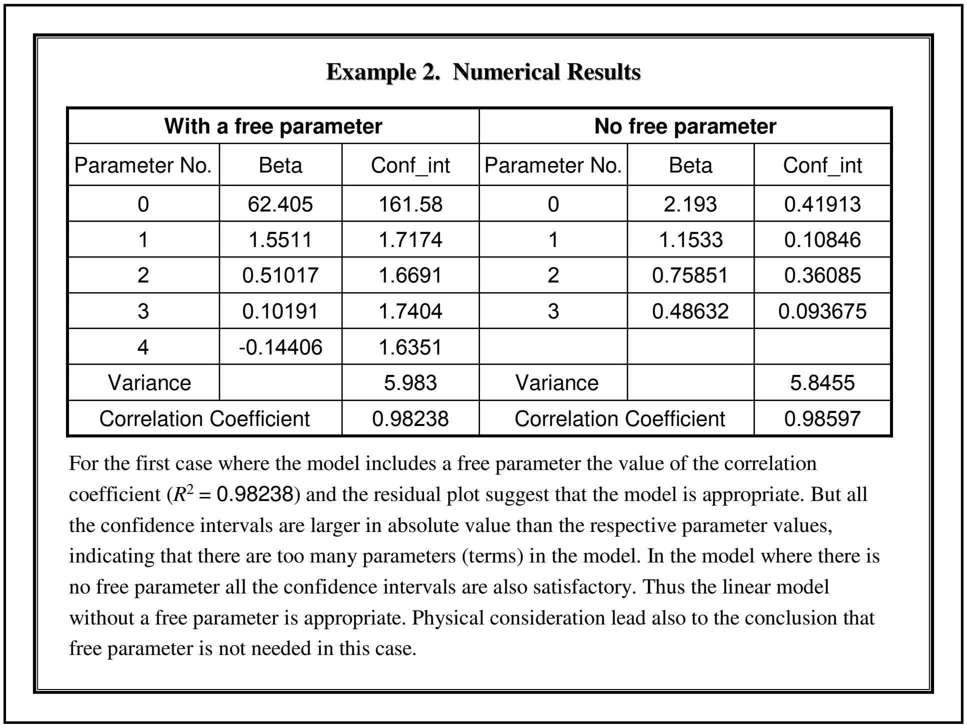
```
% filename heat_hardening.m
clear, clc, format short g, format compact
X=[7 1 11 11 7 11 3 1 2 21 1 11 10
26 29 56 31 52 55 71 31 54 47 40 66 68
6 15 8 8 6 9 17 22 18 4 23 9 8
60 52 20 47 33 22 6 44 22 26 34 12 12]';
Y=[78.5 74.3 104.3 87.6 95.9 109.2 102.7 72.5 93.1 115.9 83.8 113.3 109.4]';
Ymean=mean(Y);
N=13;     % No. of data points
npar=5;   % No. of parameters - with a free parameter
t_95=2.306;  % t-value for 95%confidence interval - with a free parameter
%npar=4;   % No. of parameters - no free parameter
%t_95=2.2622; % t-value for 95%confidence interval - no free parameter
e=ones(N,1);
X=[e X];      % Add column of ones to the X matrix (with free parameter)
A=X'*X;
```



Example 2. Matlab Solution

Calculations and Residual Plot

```
Ainv=inv(A); %Calculate the inverse of the X'X matrix (for confidence interval calculation)
Beta=Ainv*X'*Y;  % Solve the normal equation
Ycal=X*Beta; % Calculated dependent variable values
s2=((Y-Ycal)'*(Y-Ycal))/(N-npar); % variance
R2=(Ycal-Ymean)'*(Ycal-Ymean)/((Y-Ymean)'*(Y-Ymean));    %Correlation Coefficient
for i=1:npar
        Conf_int(i,1)=t_95*sqrt(s2*Ainv(i,i)); %confidence intervals
end
%
%residual plot
%
plot(Y,Y-Ycal,'*')
title('Residual plot, Heat of hardening problem')
xlabel('Heat of hardening(measured)')
ylabel('residual')
```

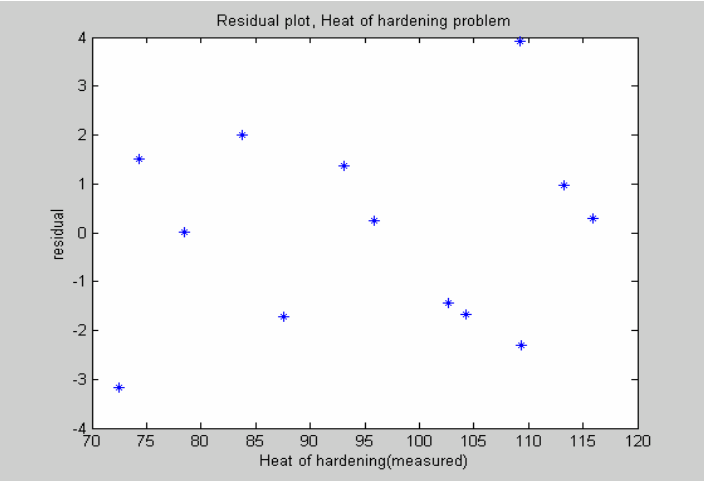


Example 2. Numerical Results

With a free parameter			**No free parameter**		
Parameter No.	Beta	Conf_int	Parameter No.	Beta	Conf_int
0	62.405	161.58	0	2.193	0.41913
1	1.5511	1.7174	1	1.1533	0.10846
2	0.51017	1.6691	2	0.75851	0.36085
3	0.10191	1.7404	3	0.48632	0.093675
4	-0.14406	1.6351			
Variance		5.983	Variance		5.8455
Correlation Coefficient		0.98238	Correlation Coefficient		0.98597

For the first case where the model includes a free parameter the value of the correlation coefficient ($R^2 = 0.98238$) and the residual plot suggest that the model is appropriate. But all the confidence intervals are larger in absolute value than the respective parameter values, indicating that there are too many parameters (terms) in the model. In the model where there is no free parameter all the confidence intervals are also satisfactory. Thus the linear model without a free parameter is appropriate. Physical consideration lead also to the conclusion that free parameter is not needed in this case.

Example 2. Residual Plot



Generalized Multiple Linear Regression .

General models of the form

$$g(\hat{y})_i = \hat{\beta}_0 + \hat{\beta}_1 f_1(x_{1i}, x_{2i} \ldots) + \hat{\beta}_2 f_2(x_{1i}, x_{2i} \ldots) \ldots \hat{\beta}_m f_m(x_{1i}, x_{2i} \ldots) \quad (19)$$

Can be brought into the form, which is appropriate for multiple linear regression

$$\hat{y}'_i = \hat{\beta}_0 + \hat{\beta}_1 x'_{1i} + \hat{\beta}_2 x'_{2i} \ldots + \hat{\beta}_m x'_{mi} \quad (20)$$

by transforming the variables $\hat{y}' = g(\hat{y})$, $x'_1 = f_1(x_1, x_2 \ldots)$, $x'_2 = f_2(x_1, x_2 \ldots)$ …etc.

The normal (16 and 17) can be solved for the coefficients β_0, β_1,… β_m, after the values of y' are introduced into the vector **Y** and x'_j introduced into the matrix **X**.

In polynomial regression the transformations: $x'_1 = x$, $x'_2 = x^2$ and $x'_m = x^m$ used.

In Riedel's equation $\log(P) = \hat{\beta}_0 + \hat{\beta}_1 (1/T) + \hat{\beta}_2 \log(T) + \hat{\beta}_3 T^2$; for vapor pressure correlation the transformations; $y' = \log(P)$, $x'_1 = 1/T$, $x'_2 = \log(T)$ and $x'_3 = T^2$ should be used.

Example 3. Fitting a 2nd Order Polynomial Thermal Conductivity Data

A 2nd order polynomial should be fitted to the thermal conductivity data of air shown in Table 1 to which a straight line was fitted in Example 1 .

The same Matlab program that was used for solution of Example 2 can be used only the containts of the **X** matrix the **Y** vector, *N*, *npar* and *t*_95 should be changed.

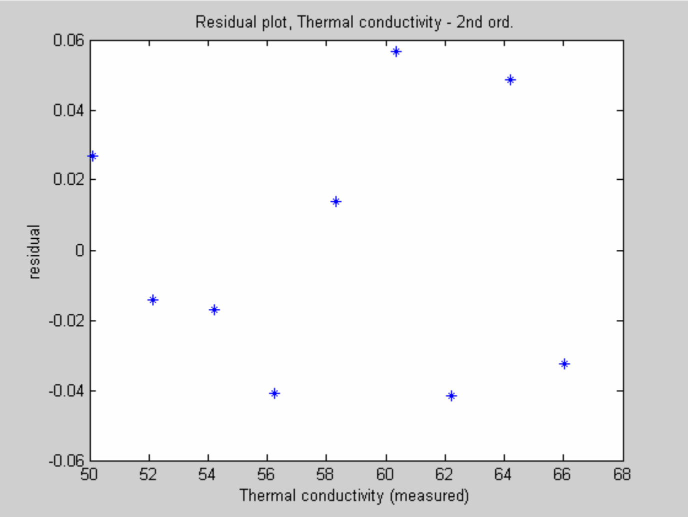
```
T=[-40 -20 0 20 40 60 80 100 120]';

Y=[50.09 52.15 54.22 56.24 58.31 60.34 62.2 64.22 66.04]';

N=9;     % No. of data points

for i=1:N

   X(i,:)=[T(i) T(i)^2];

end

npar=3;   % No. of parameters - with a free parameter

t_95=2.4469;  % t-value for 95%confidence interval - with a free parameter
```

Example 3. MATLAB results

2nd order polynomial			**Straight line**		
Parameter No.	Beta	Conf_int	Parameter No.	Beta	Conf_int
0	54.237	0.047897	0	54.1988	0.10265
1	0.10291	0.0014	1	0.1001	0.001572
2	-3.57E-05	1.52E-05			
Variance		0.001908	Variance		0.010601
Correlation Coefficient		0.99995	Correlation Coefficient		0.9996913

From the solution it can be seen that the residual plot of the 2nd order polynomial representation is randomly distributed the variance and the confidence intervals for the polynomial representation are smaller than for the straight line and R2 is closer to one. Thus, the polynomial represents the data correctly.

Example 3. Residual Plot



Physical Properties Correlation

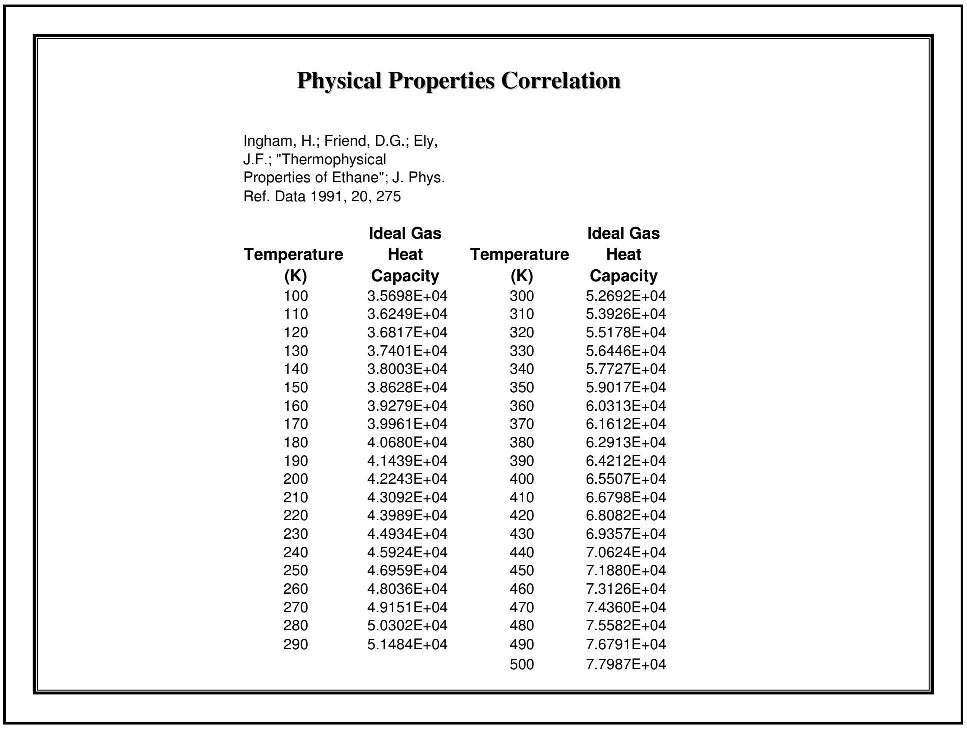
Determine appropriate correlations for heat capacity, vapor pressure, and liquid viscosity of ethane. The data files are given and also the data are available in Appendix F. Compare those correlations with the expressions suggested by the Design Institute for Physical Properties, DIPPR[2].

(a) Compare third-degree and fifth-degree polynomials for the correlation of the heat capacity data (Table A of Appendix F) using both POLYMATH and Excel by examining the respective variances, confidence intervals, and residual plots.

(b) Use Excel to compare the fifth-degree polynomial for the correlation of the heat capacity data (Table B of Appendix F) with the two DIPPR recommended correlations for the appropriate temperature intervals.

(c) Utilize multiple linear regression in Excel to fit the Wagner equation to the vapor pressure of ethane data found in Table C of Appendix F. Comment on the applicability of the Wagner equation for correlating these data. Compare the correlation obtained by the Wagner equation with that of the Riedel equation recommended by DIPPR.

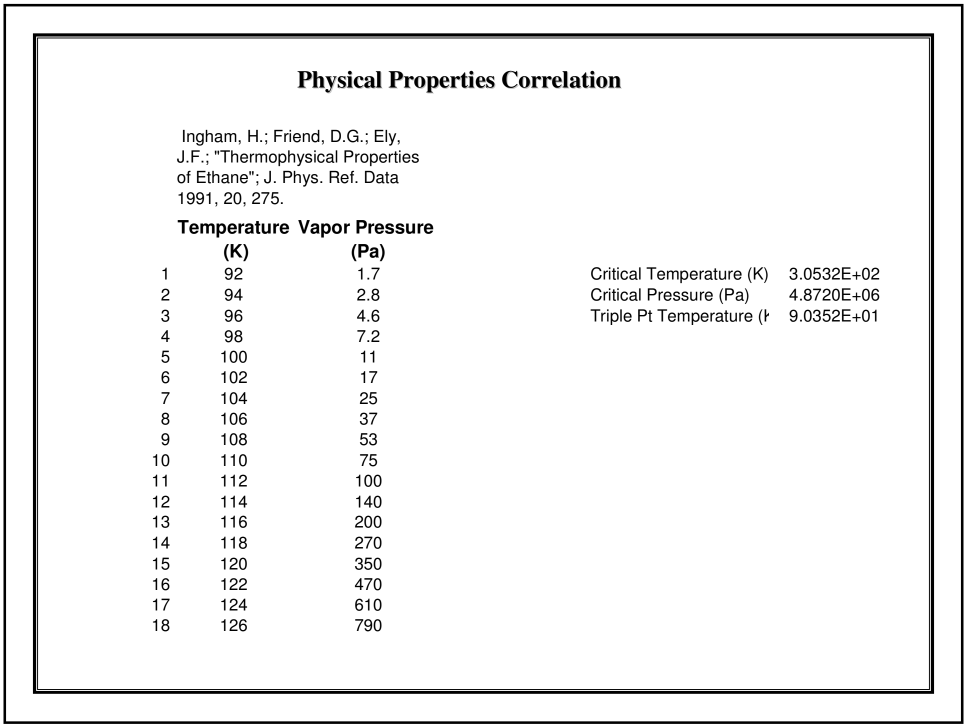
(d) Use nonlinear regression to fit the Antoine equation to the liquid viscosity data of ethane data found in Table D of Appendix F. Initial estimates of the nonlinear regression parameters should be obtained by linear regression. Verify nonlinear regression results in both POLYMATH and Excel. Compare the correlation obtained by the Antoine equation with that of the Riedel equation recommended by DIPPR.



Physical Properties Correlation

Ingham, H.; Friend, D.G.; Ely, J.F.; "Thermophysical Properties of Ethane"; J. Phys. Ref. Data 1991, 20, 275

Temperature (K)	Ideal Gas Heat Capacity	Temperature (K)	Ideal Gas Heat Capacity
100	3.5698E+04	300	5.2692E+04
110	3.6249E+04	310	5.3926E+04
120	3.6817E+04	320	5.5178E+04
130	3.7401E+04	330	5.6446E+04
140	3.8003E+04	340	5.7727E+04
150	3.8628E+04	350	5.9017E+04
160	3.9279E+04	360	6.0313E+04
170	3.9961E+04	370	6.1612E+04
180	4.0680E+04	380	6.2913E+04
190	4.1439E+04	390	6.4212E+04
200	4.2243E+04	400	6.5507E+04
210	4.3092E+04	410	6.6798E+04
220	4.3989E+04	420	6.8082E+04
230	4.4934E+04	430	6.9357E+04
240	4.5924E+04	440	7.0624E+04
250	4.6959E+04	450	7.1880E+04
260	4.8036E+04	460	7.3126E+04
270	4.9151E+04	470	7.4360E+04
280	5.0302E+04	480	7.5582E+04
290	5.1484E+04	490	7.6791E+04
		500	7.7987E+04



Physical Properties Correlation

Ingham, H.; Friend, D.G.; Ely, J.F.; "Thermophysical Properties of Ethane"; J. Phys. Ref. Data 1991, 20, 275.

	Temperature (K)	Vapor Pressure (Pa)
1	92	1.7
2	94	2.8
3	96	4.6
4	98	7.2
5	100	11
6	102	17
7	104	25
8	106	37
9	108	53
10	110	75
11	112	100
12	114	140
13	116	200
14	118	270
15	120	350
16	122	470
17	124	610
18	126	790

Critical Temperature (K)	3.0532E+02
Critical Pressure (Pa)	4.8720E+06
Triple Pt Temperature (K	9.0352E+01

Nonlinear Regression

If one or more of the parameters of the model are included in nonlinear expressions the estimation of the parameters cannot be carried out by solving a system of linear equations. The most general approach to solving a nonlinear regression problem is by using optimization programs to minimize F (Equation 2) numerically while changing the parameters β_0, $\beta_1 \ldots \beta_m$. For individual cases, more specific simpler techniques can be used.

Example 4. Fitting Parameters to the Antoine equation

The following table presents data of vapor pressure versus temperature for benzene. Correlate the data using the Antoine equation.

No.	Temperature, *T*	Pressure, *P*
	°C	Mm Hg
1	-36.7	1
2	-19.6	5
:		
9	60.6	400
10	80.1	760

Example 4. Solution (1)

The Antoine equation is a widely used vapor pressure correlation that utilizes the parameters A, B, and C. It can be expressed by

$$\log(P) = A + \frac{B}{T + C} \qquad (22)$$

Defining and introducing Equation (22) into Equation (2) gives

$$F = \sum_{i=1}^{n} \left[y_i - \left(A + \frac{B}{T_i + C} \right) \right]^2 \qquad (23)$$

Example 4. Solution (2)

Differentiating *F* with respect to A, B, and C and equating to zero, yield

$$\frac{\partial F}{\partial A} = \sum_{i=1}^{n}\left[y_i - \left(A + \frac{B}{T_i + C}\right)\right](-2) = 0$$
$$\frac{\partial F}{\partial B} = \sum_{i=1}^{n}\left[y_i - \left(A + \frac{B}{T_i + C}\right)\right](\frac{-2}{T_i + C}) = 0 \qquad (24)$$
$$\frac{\partial F}{\partial C} = \sum_{i=1}^{n}\left[y_i - \left(A + \frac{B}{T_i + C}\right)\right]\left[\frac{2}{(T_i + C)^2}\right] = 0$$

This is a system of three nonlinear algebraic equations with three unknowns. The MATLAB function to be used when solving this problem as a system of three equations follows.

MATLAB function for solving Example 4 as a system of three nonlinear equations

```
filename fun_antoine
%Solving for Antoine equation parameters as a system of equations
function f=fun_antoine(x)
A=x(1); B=x(2); C=x(3);
T=[-36.7 -19.6 -11.5 -2.6 7.6 15.4 26.1 42.2 60.6 80.1]';
P=[1 5 10 20 40 60 100 200 400 760]';
Y=log10(P);
 f1=0; f2=0; f3=0;
 for i=1:10
    f1=f1+(-Y(i)+A+B/(T(i)+C));
    f2=f2-(Y(i)-A-B/(T(i)+C))/(T(i)+C);
    f3=f3+(Y(i)-A-B/(T(i)+C))*B/(T(i)+C)^2;
 end
 f(1,1) = f1 ; f(2,1) = f2 ; f(3,1) = f3 ;
```

MATLAB function for solving Example 4

This function can be used with a main program which apply, for example, the multi-dimensional Newton-Raphson method (see the main program of Example 3 in the previous chapter of systems of nonlinear algebraic equations). Starting from the initial estimate A = 6, B = -700, and C = 150 the program converges to the solution: A = 5.7673, B = -677.09, and C = 153.89.

It should be mentioned that when solving this problem as a system of nonlinear equations, good initial estimates for the parameters should be provided otherwise most solution methods will diverge. An alternative approach which does not require close initial estimates involves conversion of the problem to a single nonlinear equation. For a specified value of C, the first two equations of the system (24) are linear and can be solved for A and B. Then the third equation of this system is used for calculating $f(C)$. The MATLAB function for carrying out this calculation follows.

MATLAB function for solving Example 4 as a single nonlinear equation

```
%filename fun_antoine2
%Solving for Antoine equation parameters as a single nonlinear equation
function fC=fun_antoine(C)
T=[-36.7 -19.6 -11.5 -2.6 7.6 15.4 26.1 42.2 60.6 80.1]';
P=[1 5 10 20 40 60 100 200 400 760]';
Y=log10(P);
x(:,1)=[1./(T+C)];
den=10*(x'*x)-sum(x)^2;
A=(sum(Y)*(x'*x)-sum(x)*(x'*Y))/den;
B=(10*(x'*Y)-sum(x)*sum(Y))/den;
 fC=0;
 for i=1:10
   fC=fC+(Y(i)-A-B/(T(i)+C))*B/(T(i)+C)^2;
 end
```

This function can be used in conjunction with the main programs for solving single nonlinear equations that are provided in the first chapter. Using the bisection method for example (main program of Example 3 in Chapter 1) starting from C_{01} = 50 and C_{02} = 350, it converges to the solution C = 153.93 in 12 iterations.