LINEAR REGRESSION[1]

The linear regression model $\hat{\mathbf{y}}_i = \mathbf{F}(\mathbf{x}_i)\mathbf{k} + \boldsymbol{\varepsilon}_i$; i=1,2,…,N (3.1)

where **F**(**x**$_i$) is an *m×p dimensional* matrix which depends only on **x**$_i$ and it is independent of the parameters.

If matrix **F** is the independent variables ➔ *linear regression* model:

$$\hat{\mathbf{y}}_i = \mathbf{X}_i\,\mathbf{k} + \boldsymbol{\varepsilon}_i \quad ; \quad i=1,2,\ldots,N \qquad (2.5)$$

(i) The *simple linear regression model*

$$\hat{y}_i = k_1 x_i + k_2 + \varepsilon_i \qquad (3.3a)$$

or in matrix notation

$$\hat{y}_i = [x_i , 1]\begin{bmatrix} k_1 \\ k_2 \end{bmatrix} + \varepsilon_i \qquad (3.3b)$$

(ii) The *multiple linear regression model*

$$\hat{y}_i = k_1 x_{1i} + k_2 x_{2i} + \ldots + k_{p-1}x_{p-1,i} + k_p + \varepsilon_i (3.4a)$$

or in matrix notation

$$\hat{y}_i = [x_{1i} , x_{2i}, \ldots, x_{p-1,i} , 1]\begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_p \end{bmatrix} + \varepsilon_i \qquad (3.4b)$$

or more compactly as

$$\hat{y}_i = \mathbf{x}_i^T \mathbf{k} + \varepsilon_i \qquad (3.4c)$$

where $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{p-1,i} , 1]^T$ is the augmented *p-dimensional* vector of independent variables ($p=n+1$).

[1] Englezos, P. and N. Kalogerakis, "*Applied Parameter Estimation for Chemical Engineers*", Marcel-Dekker, New York, 2001

(iii) The *multiresponse linear regression model* with *m* response variables, *(m×n)* independent variables and *p (=n+1)* parameters,

$$\begin{aligned}
\hat{y}_{1i} &= k_1 x_{11i} + k_2 x_{12i} + \ldots + k_{p-1} x_{1,p-1,i} + k_p + \varepsilon_{1i} \\
\hat{y}_{2i} &= k_1 x_{21i} + k_2 x_{22i} + \ldots + k_{p-1} x_{2,p-1,i} + k_p + \varepsilon_{2i} \\
&\vdots \\
\hat{y}_{mi} &= k_1 x_{m1i} + k_2 x_{m2i} + \ldots + k_{p-1} x_{m,p-1,i} + k_p + \varepsilon_{mi}
\end{aligned} \tag{3.5a}$$

or in matrix notation

$$\begin{bmatrix} \hat{y}_{1i} \\ \hat{y}_{2i} \\ \vdots \\ \hat{y}_{mi} \end{bmatrix} = \begin{bmatrix} x_{11i} & x_{12i} & \cdots & x_{1,p-1,i} & 1 \\ x_{21i} & x_{22i} & \cdots & x_{2,p-1,i} & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ x_{m1i} & x_{m2i} & \cdots & x_{m,p-1,i} & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_p \end{bmatrix} + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{mi} \end{bmatrix} \tag{3.5b}$$

or more compactly as

$$\hat{\mathbf{y}}_i = \mathbf{X}_i \mathbf{k} + \boldsymbol{\varepsilon}_i \tag{3.5c}$$

where the matrix $\mathbf{X}_i$ is defined as

$$\mathbf{X}_i = \begin{bmatrix} x_{11i} & x_{12i} & \cdots & x_{1,p-1,i} & 1 \\ x_{21i} & x_{22i} & \cdots & x_{2,p-1,i} & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ x_{m1i} & x_{m2i} & \cdots & x_{m,p-1,i} & 1 \end{bmatrix} \tag{3.6}$$

It should be noted that in linear regression books **X** is often defined for the simple or multiple linear regression model and it contains *all* the measurements. In our case, index i explicitly denotes the ith measurement and matrix $\mathbf{X}_i$ represents the values of the independent variables from the ith experiment.

THE LINEAR LEAST SQUARES OBJECTIVE FUNCTION

Given N measurements of the response variables (output vector), the parameters are obtained by minimizing the *Linear Least Squares* (LS) objective function

$$S_{LS}(\mathbf{k}) = \sum_{i=1}^{N} [\hat{\mathbf{y}}_i - \mathbf{X}_i\mathbf{k}]^T \mathbf{Q}_i [\hat{\mathbf{y}}_i - \mathbf{X}_i\mathbf{k}] \qquad (3.8)$$

where $\mathbf{Q}_i$ is an $m \times m$ weighting matrix. Depending on f $\mathbf{Q}_i$, we have the following cases:

Simple Linear Least Squares. In this case we use $\mathbf{Q}_i = \mathbf{I}$ in Equation 3.8. This choice of $\mathbf{Q}_i$ yields ML estimates of the parameters if the error terms in each response variable and for each experiment (ε_{ij}, i=1,…N; j=1,…,m) are all identically and independently distributed (i.i.d) normally with zero mean and variance, σ_e^2. Namely, $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$ and $COV(\boldsymbol{\varepsilon}_i) = \sigma_e^2 \mathbf{I}$ where $\mathbf{I}$ is the $m \times m$ identity matrix.

Weighted Least Squares (WLS) Estimation. In this case the weighting matrix is kept the same for all experiments, $\mathbf{Q}_i = \mathbf{Q}$ for all i=1,…,N in Equation 3.8. This choice of $\mathbf{Q}_i$ yields ML estimates of the parameters if the error terms in each response variable and for each experiment (ε_{ij}, i=1,…N; j=1,…,m) are independently distributed normally with zero mean and constant variance. Namely, the variance of a particular response variable is constant from experiment to experiment; however, different response variables have different variances, i.e.,

$$COV(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} \sigma_{e_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{e_2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{em}^2 \end{bmatrix} \;;\; i=1,2,\ldots,N \qquad (3.11)$$

which can be written as

$$COV(\boldsymbol{\varepsilon}_i) = \sigma^2\, diag(v_1, v_2, \ldots, v_m) \;;\; i=1,2,\ldots,N \qquad (3.12)$$

where σ^2 is an unknown scaling factor and v_1, v_2,…,v_m are known constants. ML estimates are obtained if the constant weighting matrix $\mathbf{Q}$ have been chosen as

$$\mathbf{Q} = diag(v_1^{-1}, v_2^{-1}, \ldots, v_m^{-1}) \;;\; i=1,2,\ldots,N \qquad (3.13)$$

Generalized Least Squares (GLS) Estimation. In this case we minimize a weighted SSE with non-constant weights. The weighting matrices differ from experiment to experiment. ML estimates of the parameters are obtained if we choose

$$\mathbf{Q}_i = [COV(\boldsymbol{\varepsilon}_i)]^{-1} \;;\; i=1,2,\ldots,N \qquad (3.14)$$

LINEAR LEAST SQUARES ESTIMATION

The computation of the parameter estimates is accomplished by minimizing the *least squares* (LS) objective function given by Equation 3.8 and using the stationary criterion

$$\frac{\partial S_{LS}(\mathbf{k})}{\partial \mathbf{k}} = \mathbf{0} \qquad (3.15)$$

yields a linear equation of the form.

$$\mathbf{A}\,\mathbf{k} = \mathbf{b} \qquad (3.16)$$

where the *(p×p) dimensional* matrix **A** is given by.

$$\mathbf{A} = \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}_i \mathbf{X}_i \qquad (3.17a)$$

and the *p-dimensional* vector **b** is given by.

$$\mathbf{b} = \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}_i \hat{\mathbf{y}}_i \qquad (3.17b)$$

Solution of the above linear equation yields the least squares estimates of the parameter vector, **k***,

$$\mathbf{k}^* = \left[\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}_i \hat{\mathbf{y}}_i\right] \qquad (3.18)$$

For the *single response* linear regression model (*m*=1), Eqns (3.17a) and (3.17b) reduce to

$$\mathbf{A} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T Q_i \qquad (3.19a)$$

and

$$\mathbf{b} = \sum_{i=1}^{N} \mathbf{x}_i \hat{y}_i Q_i \qquad (3.19b)$$

where Q_i is a scalar weighting factor and $\mathbf{x}_i$ is the augmented *p-dimensional* vector of independent variables $[x_{1i}\ ,\ x_{2i},\ \ldots,\ x_{p-1,i}\ ,\ 1]^T$. The optimal parameter estimates are obtained from

$$\mathbf{k}^* = \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T Q_i\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{x}_i \hat{y}_i Q_i\right] \qquad (3.20)$$

In practice, the solution of Equation 3.16 for the estimation of the parameters is not done by computing the inverse of matrix **A**. Instead, one may perform first an eigenvalue decomposition of the real symmetric matrix **A** which provides significant additional information about potential ill-conditioning of the parameter estimation problem.

STATISTICAL INFERENCES

Inference on the Parameters

The least squares estimator has several desirable properties. Namely, the parameter estimates are normally distributed, unbiased (i.e., $E(\mathbf{k}^*)=\mathbf{k}$) and their covariance matrix is given by

$$COV(\mathbf{k}^*) = \sigma_\varepsilon^2 \mathbf{A}^{-1} \qquad (3.30)$$

where matrix **A** is given by Equation 3.17a or 3.19a. An estimate, $\hat{\sigma}_\varepsilon^2$□ of the variance σ_ε^2 is given by

$$\hat{\sigma}_\varepsilon^2 = \frac{S_{LS}(\mathbf{k}^*)}{(\text{d.f.})} \qquad (3.31)$$

where (d.f.)=(N-*mp*) are the *degrees of freedom*, namely the total number of measurements minus the number of unknown parameters. Note that *m* is the number of response variables and *p* the number of parameters,

The corresponding *(1-α)100% marginal confidence interval* for each parameter, k_i, i=1,2,…,p, is

$$k_i^* - t_{\alpha/2}^{\nu}\hat{\sigma}_{k_i} \le \ k_i \ \le k_i^* + t_{\alpha/2}^{\nu}\hat{\sigma}_{k_i} \qquad (3.33)$$

where $t_{\alpha/2}^{\nu}$ is obtained from the tables of Student's T-distribution with ν=N*m*-□*p* degrees of freedom.

The standard error of parameter k_i, $\hat{\sigma}_{k_i}$, is obtained as the square root of the corresponding diagonal element of the inverse of matrix **A** multiplied by $\hat{\sigma}_\varepsilon$, i.e.,

$$\hat{\sigma}_{k_i} = \hat{\sigma}_\varepsilon \sqrt{\left\{\mathbf{A}^{-1}\right\}_{ii}} \qquad (3.34)$$

Practically, for ν≥30 we can use the approximation $t_{\alpha/2}^{\nu} \approx z_{\alpha/2}$ where $z_{\alpha/2}$ is obtained from the tables of the standard normal distribution. That is why when the degrees of freedom are high, the 95% confidence intervals are simply taken as twice the standard error (recall that $z_{0.025}$=1.96 and $t_{0.025}^{30}$=2.042).

Inference on the Expected Response Variables

The *predicted mean response* of the linear regression model at $\mathbf{x}_0$ is $\mathbf{y}_0 = \mathbf{X}_0\mathbf{k}^*$. {for the standard multiresponse linear regression model where $\mathbf{F}(\mathbf{x}_0) \equiv \mathbf{X}_0$,}
Although the error term $\boldsymbol{\varepsilon}_0$ is not included, there is some uncertainty in the predicted mean response due to the uncertainty in $\mathbf{k}^*$. The covariance matrix of the predicted mean response is given by

$$COV(\mathbf{y}_0) = \mathrm{x}_0^T COV(\mathbf{k}^*)\mathrm{x}_0 \qquad (3.35b)$$

The *(1-a)100% confidence interval* of y_{i0} (i=1,…,*m*), the ith element of the response vector $\mathbf{y}_0$ at $\mathbf{x}_0$ is given below

$$y_{i0} - t^{\nu}_{\alpha/2}\hat{\sigma}_{y_{i0}} \le \mu_{y_{i0}} \le y_{i0} + t^{\nu}_{\alpha/2}\hat{\sigma}_{y_{i0}} \qquad (3.36)$$

The standard error of y_{i0}, $\hat{\sigma}_{y_{i0}}$, is the square root of the ith diagonal element of *COV*($\mathbf{y}_0$), namely,

$$\hat{\sigma}_{y_{i0}} = \hat{\sigma}_{\varepsilon}\sqrt{\left\{\mathbf{X}_0^T\mathbf{A}^{-1}\mathbf{X}_0\right\}_{ii}} \qquad (3.37b)$$

For the *single* response y_0 in the case of simple or multiple linear regression (i.e., *m*=1), the *(1-α)100% confidence interval* of y_0 is,

$$y_0 - t^{\nu}_{\alpha/2}\hat{\sigma}_{y_0} \le \mu_{y_0} \le y_0 + t^{\nu}_{\alpha/2}\hat{\sigma}_{y_0} \qquad (3.38a)$$

or equivalently

$$\mathbf{x}_0^T\mathbf{k}^* - t^{\nu}_{\alpha/2}\hat{\sigma}_{y_0} \le \mu_{y_0} \le \mathbf{x}_0^T\mathbf{k}^* + t^{\nu}_{\alpha/2}\hat{\sigma}_{y_0} \qquad (3.38b)$$

where $t^{\nu}_{\alpha/2}$ is obtained from the tables of Student's T-distribution with ν=(N-p) degrees of freedom and $\hat{\sigma}_{y_0}$ is the *standard error of prediction* at $\mathbf{x}_0$. This quantity usually appears in the standard output of many regression computer packages. It is computed by

$$\hat{\sigma}_{y_0} = \hat{\sigma}_{\varepsilon}\sqrt{\mathbf{x}_0^T\mathbf{A}^{-1}\mathbf{x}_0} \qquad (3.39)$$

In all the above cases we presented confidence intervals for the *mean* expected response rather than a *future observation (future measurement)* of the response variable, $\hat{\mathbf{y}}_0$.

In this case, besides the uncertainty in the estimated parameters, we must include the uncertainty due to the measurement error ($\boldsymbol{\varepsilon}_0$).

The corresponding *(1-α)100% confidence interval* for the multi-response linear model is

$$y_{i0} - t^{\nu}_{\alpha/2}\hat{\sigma}_{\hat{y}_{i0}} \le \hat{y}_{i0} \le y_{i0} + t^{\nu}_{\alpha/2}\hat{\sigma}_{\hat{y}_{i0}} \quad ; i=1,\ldots,m \qquad (3.40)$$

where the corresponding standard error of $\hat{y}_{i0}$ is given by

$$\hat{\sigma}_{\hat{y}_{i0}} = \hat{\sigma}_{\varepsilon}\sqrt{1 + \left\{\mathbf{X}_0^T\mathbf{A}^{-1}\mathbf{X}_0\right\}_{ii}} \qquad (3.41)$$

For the case of a single response model (i.e., *m*=1), the (*1-α*)*100% confidence interval* of $\hat{y}_0$ is,

$$\mathbf{x}_0^T\mathbf{k}^* - t^{\nu}_{\alpha/2}\hat{\sigma}_{\hat{y}_0} \le \hat{y}_0 \le \mathbf{x}_0^T\mathbf{k}^* + t^{\nu}_{\alpha/2}\hat{\sigma}_{\hat{y}_0} \qquad (3.42)$$

where the corresponding standard error of $\hat{y}_0$ is given by

$$\hat{\sigma}_{\hat{y}_0} = \hat{\sigma}_{\varepsilon}\sqrt{1 + \mathbf{x}_0^T\mathbf{A}^{-1}\mathbf{x}_0} \qquad (3.43)$$

SEE the example next to appreciate the difference between equations 3.39 and 3.43

EXAMPLE[2]

Consider a set of data for 10 cars (y=miles per gallon; x-weight in tones)

Car No	1	2	3	4	5	6	7	8	9	10
Miles per gallon (y)	17.9	16.5	16.4	16.8	18.8	15.5	17.5	16.4	15.9	18.3
Weight in tones (x)	1.35	1.90	1.70	1.80	1.30	2.05	1.60	1.80	1.85	1.40

Estimated Model equation: y=23.75-4.03 x

Suppose that we are interested in all cars weighing 1.7 tons. Then the estimated average mileage is 23.75-4.03 (1.7)=16.899 miles per gallon.

CONFIDENCE?

We can be 90 % confident (α=0.1) that the average gas mileage for cars weighing 1.7 tons lies between 16.689 and 17.109 per gallon.

We can be 90 % confident that the gas mileage for any individual automobile weighing 1.7 tons lies between 16.209 and 17.589 miles per gallon.

This shows that the prediction interval used to predict the gas mileage for a single auto is wider than that used to predict the average mileage for a group of automobiles.

[2] Milton and Arnold, *Introduction to Probabilities and Statistics*, 2nd edition, Wiley, 1990.

similar. The difference is that the former entails the term

$$\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \qquad (\mu_{Y/x_0})$$

whereas the corresponding term in the latter is a little larger, namely

$$\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \qquad (Y|x_0)$$

This is to be expected since we should be able to estimate an average response more precisely than we can predict an individual observation. Graphically, the confidence band on $\mu_{Y|x}$ will be contained in the corresponding prediction band for $Y|x$. This idea is illustrated in Fig. 11.7.

$\mu_{Y/x_0} = b_0 + b_1 X$

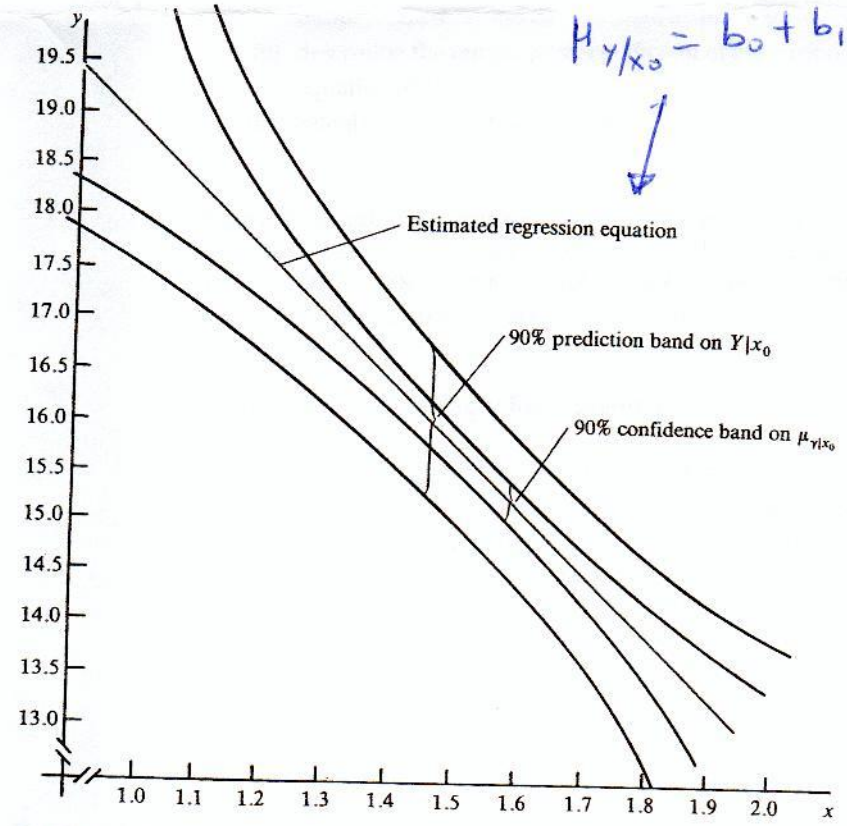

FIGURE 11.7
Relative positions of 90% confidence bands on $\mu_{Y|x}$ and $Y|x$.

SOLUTION OF MULTIPLE LINEAR REGRESSION PROBLEMS

Problems that can be described by a multiple linear regression model (i.e., they have a single response variable, m=1) can be readily solved by using Microsoft Excel^{TM}

Procedure for using Microsoft Excel^{TM} for Windows

Step 1. First the data are entered in columns. The single dependent variable is designated by y whereas the independent ones by x_1, x_2, x_3 etc.

Step 2. Select cells below the entered data to form a rectangle [5 × p] where p is the number of parameters sought after; e.g. in the equation $y=k_1x_1+k_2x_2+k_3$ you are looking for k_1, k_2 and k_3. Therefore p would be equal to 3. Note: Excel casts the p-parameter model in the following form: $y=m_1x_1+m_2x_2+\ldots+m_{p-1}x_{p-1}+b$.

Step 3. Now that you have selected an area [5 × p] on the spreadsheet, go to the f_x (Paste Function button) and click.

Step 4. Click on Statistical on the left scroll menu and click on LINEST on the right scroll menu; then hit OK. A box will now appear asking for the following

Known Y`s
Known X`s
Const
Stats

Step 5. Click in the text box for known values for the singe response variable y; then go to the Excel sheet and highlight the y values.

Step 6 Repeat Step 5 for the known values for the independent variables x_1, x_2, etc. by clicking on the box containing these values. This the program lets you highlight the area that encloses all the x values (x_1, x_2, x_3, etc….).

Step 7. Set the logical value *Const=true* if you wish to calculate a y value.

Step 8. Set the logical value *Stats=true* if you wish the program to return additional regression statistics.

Step 9. Now that you have entered all the data you do not hit the OK button but instead press *Control-Shift-Enter*. This command allows all the elements in the array to be displayed. If you hit OK you will only see one element in the array.

Once the above steps have been followed, the program returns the following information on the worksheet displayed in a [5 × p] table where p is the number of parameters

1st row: parameter values

	m_{p-1}	m_{p-2}	…	m_2	m_1	b
or	k_{p-1}	k_{p-2}	…	k_2	k_1	k_p

2nd row: Standard errors for the estimated parameter values

se (k_{p-1}) se(k_{p-2}) … se(k_2) se(k_1) se(k_p)

3rd row: Coefficient of determination and standard error of the y value

R^2 sev

4th row: F statistic and the number of degrees of freedom (d.f.)

5th row: Information about the regression

ssreg (regression sum of squares) ssresid (residual sum of squares)

Example

In Table 3.2 a set of data that relate the pH with the charge on wood fibers are provided. We seek to establish a correlation of the form given by Equation 3.47 by fitting the equation to the charge (Q) versus pH data given in Table 3.2.

$$Q=C_1+C_2(pH)+C_3(pH)^2+C_4(pH)^3 \qquad (3.47)$$

Solution

Equation 3.47 is written in the following form

$$y=k_1x_1+k_2x_2+k_3x_3+k_4 \qquad (3.48a)$$

where $y=Q$, $k_1=C_2$, $k_2=C_3$, $k_3=C_4$, $k_4=C_1$, $x_1=pH$, $x_2=(pH)^2$ and $x_3=(pH)^3$.

Table 3.2 Charge on Wood Fibers

pH	Charge (Q)	Calculated Charge (Q)
2.8535	19.0	16.9
3.2003	32.6	34.5
3.6347	52.8	54.0
4.0910	71.4	71.6
4.5283	86.2	86.3
5.0390	99.6	100.7
5.6107	115.4	114.0
6.3183	130.7	127.2
7.0748	138.4	138.4
7.7353	144.1	147.0
8.2385	151.6	153.3
8.8961	159.9	162.1
9.5342	172.2	171.9
10.0733	183.9	181.9
10.4700	193.1	190.5
10.9921	200.7	203.9

Excel casts the model in the following form

$$y = m_1x_1 + m_2x_2 + m_3x_3 + b \qquad (3.48b$$

where y=Q, $m_1=k_1= C_2$, $m_2=k_2=C_3$, $m_3=k_3=C_4$, x_1=pH $x_2=(pH)^2$ and $x_3=(pH)^3$.

The program returns the following results

m_3	m_2	m_1	b
k_3	k_2	k_1	k_4
C_4	C_3	C_2	C_1
0.540448	-12.793	113.4188	-215.213
0.047329	0.9852	6.380755	12.61603
0.998721	2.27446	#N/A	#N/A
3122.612	12	#N/A	#N/A
48461.41	62.07804	#N/A	#N/A

In Table 3.2 the calculated charge values by the model are also shown.

NOTE: R^2=0.998721, F=3122.612, 12=degrees of freedom (16-4=12)