3.0 What is regression?

Regression analysis gives information on the relationship between a response variable and one or more independent variables to the extent that information is contained in the data. The goal of regression analysis is to express the response variable as a function of the predictor variables.

Once regression analysis relationship is obtained, it can be used to predict values of the response variable, identify variables that most affect response, or verify hypothesized casual models of the response.

3.1 Linear regression

Linear regression is the most popular regression model. In this model we wish to predict response to *n* data points (x_1,y_1), (x_2,y_2), ..., (x_n, y_n) data by a regression model given by.

$$y = a_0 + a_1 x$$

Where a_0 and a_1 are the constants of the regression model.

A measure of goodness of fit, that is, how $a_0 + a_1 x$ predicts the response variable *y* is the magnitude of the residual, ε_i at each of the *n* data points.

$$\varepsilon_i = y_i - (a_0 + a_1 x_i)$$

Ideally, if all the residuals ε_i are zero, one may have found an equation in which all the points lie on the model. Thus, ***minimization*** of the residual is an objective of obtaining regression coefficients.

The most popular method to minimize the residual is the least squares method, where the estimates of the constants of the models are chosen such that the sum of the squared residuals is minimized, that is minimize $\sum_{i=1}^{n} \varepsilon_i^2$.

Let us use the least squares criterion where we minimize

$$S_r = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

S_r is called the sum of the square of the residuals.

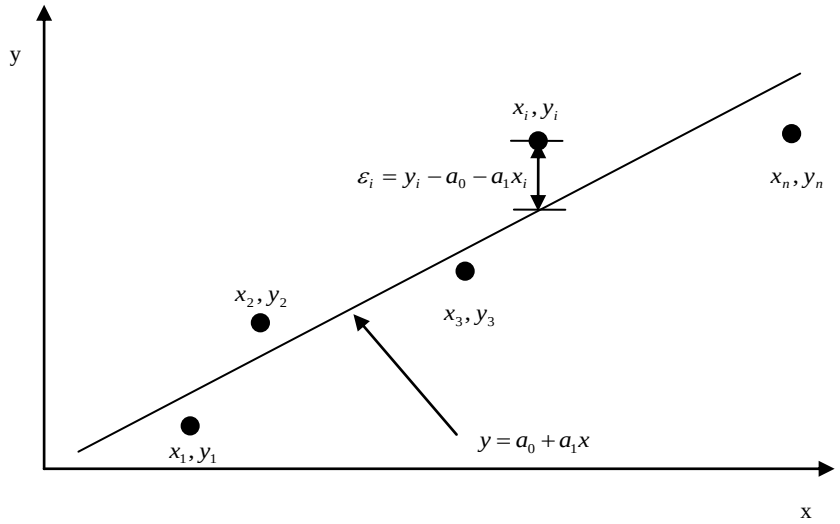


Figure 3.1 Linear regression of y vs. x data showing residuals at a typical point, x_i.

To find a_0 and a_1, we minimize S_r with respect to a_0 and a_1:

$$\frac{\partial S_r}{\partial a_0} = 2\sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i\right)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = 2\sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i\right)(-x_i) = 0$$

Giving

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = 0$$

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = 0$$

Noting that $\sum_{i=1}^{n} a_0 = a_0 + a_0 + \ldots + a_0 = n a_0$

$$n a_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (3.1)$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (3.2)$$

Solving the above equations gives:

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Or from equation (3.2)

$$a_o = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{a_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - a_1 \bar{x}$$

Example 3.1

The following *y* vs. *x* data is given

x	1	7	13	19	25
y	1	49	169	361	625

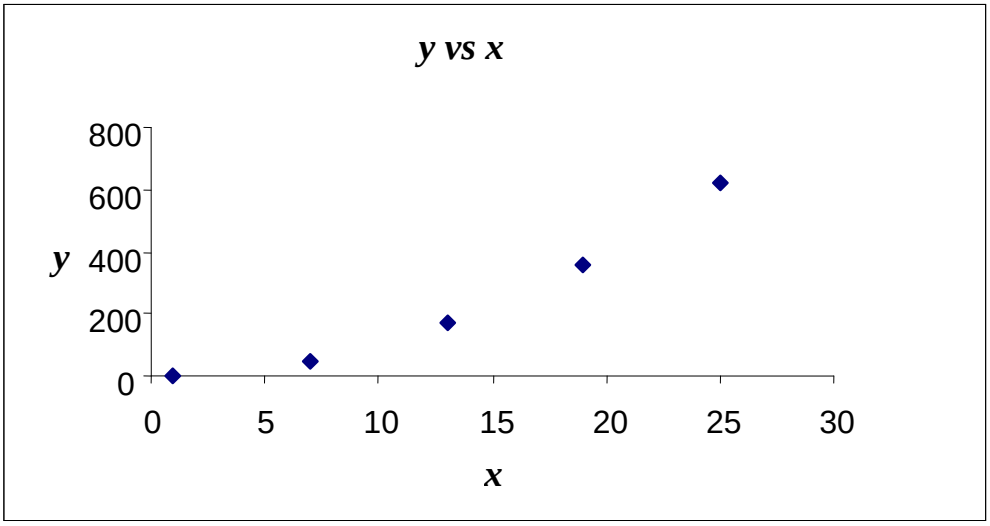


Figure 3.1 Data points of the y vs x data

Although $y = x^2$ is an exact fit to the data, a scientist thinks that $y = a_0 + a_1 x$ can explain the data. Find constants of the model, a_0, and a_1,

Solution

First find the constants of the assumed model

$$y = a_0 + a_1 x$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$n = 5$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{5} x_i y_i = 1 \times 1 + 7 \times 49 + 13 \times 169 + 19 \times 361 + 25 \times 625 = 25025$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{5} x_i^2 = 1^2 + 7^2 + 13^2 + 19^2 + 25^2 = 1205$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{5} y_i = 1 + 49 + 169 + 361 + 625 = 1205$$

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{5} x_i = 1+7+13+19+25 = 65$$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$a_1 = \frac{5(25025) - (65)(1205)}{5(1205) - (65)^2} = 26$$

$$a_0 = \bar{y} - a_1 \bar{x} = \frac{1205}{5} - 26\frac{65}{5} = (241) - 26(13) = -97$$

This gives

$$y = a_0 + a_1 x$$

$$y = -97 + 26x$$

Example 3.2

The following table gives the value of density of saturated water for various temperatures of saturated stream.

Temp^oC (= T)	100	150	200	250	300
Density kg/m^3 (= D)	958	917	865	799	712

a) Use curve fitting to fit the results to a first-order polynomial $D = A + BT$.

b) Find the densities when the temperatures are 130^oC and 275^oC respectively.

Solution:

a_0 and a_1 can be computed by constructing the following table:

T_i	D_i	T_i^2	$T_i D_i$
100	958	10000	95800
150	917	22500	137550
200	865	40000	173000
250	799	62500	199750
300	712	90000	213600
∑1000	4251	225000	819700

$$a_1 = \frac{5\times 819700 - 1000\times 4251}{5\times 225000 - (1000)^2} = -1.22$$

$$a_0 = \frac{4251}{5} - a_1\frac{1000}{5} = 1094.2$$

D=1094.2-1.22×T

To compare the predicted values to the experimental values:

T_i	D_i	D_i(estimated) D=1094.2-1.22×T
100	958	972.2
150	917	911.2
200	865	850.2
250	799	789.2
300	712	728.2

D(130)= 1094.2-1.22×130=935.6

D(175)= 1094.2-1.22×175=880.7

3.2 Polynomial Models

Given *N* data points $(x_1, y_1), (x_2, y_2)$. . , (x_N, y_N) use least squares method to regress the data to an n^{th} order polynomial.

In the development, we use *n* as the degree of the polynomial and *N* as the number of data pairs (x_i, y_i). We will always have $N > n+1$ in the following.

Assume the functional relationship for fitting

$$Y(x) = a_0 + a_1 x + a_2 x^2 + \mathrm{L}\ + a_n x^n$$

with errors defined by

$$e_i = y_i - Y(x_i) = y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - \mathrm{L}\ - a_n x_i^{\,n},$$

in which $i = 1, 2, 3, \ldots, N$.

We minimize the sum of error squares,

$$S = \sum_{i=1}^{N} e_i^{\,2} = \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - \mathrm{L}\ - a_n x_i^{\,n})^2 .$$

At the minimum, all the first partial derivatives with respect to a_i's vanish. We have

$$\frac{\partial S}{\partial a_0} = 0 = 2\sum_{i=1}^{N} (y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - \mathrm{L}\ - a_n x_i^{\,n})(-1),$$

$$\frac{\partial S}{\partial a_1} = 0 = 2\sum_{i=1}^{N} (y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - \mathrm{L}\ - a_n x_i^{\,n})(-x_i),$$

$$\frac{\partial S}{\partial a_2} = 0 = 2\sum_{i=1}^{N} (y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - \mathrm{L}\ - a_n x_i^{\,n})(-x_i^{\,2}),$$

M

$$\frac{\partial S}{\partial a_n} = 0 = 2\sum_{i=1}^{N} (y_i - a_0 - a_1 x_i - a_2 x_i^{\,2} - \mathrm{L}\ - a_n x_i^{\,n})(-x_i^{\,n}),$$

Rearrange them to get

$$a_0 N + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^{\,2} + \mathrm{L}\ + a_n \sum_{i=1}^{N} x_i^{\,n} = \sum_{i=1}^{N} y_i ,$$

$$a_0\sum_{i=1}^{N}x_i + a_1\sum_{i=1}^{N}x_i^{\,2} + a_2\sum_{i=1}^{N}x_i^{\,3} + \mathrm{L} + a_n\sum_{i=1}^{N}x_i^{\,n+1} = \sum_{i=1}^{N}x_i y_i \,,$$

$$a_0\sum_{i=1}^{N}x_i^{\,2} + a_1\sum_{i=1}^{N}x_i^{\,3} + a_2\sum_{i=1}^{N}x_i^{\,4} + \mathrm{L} + a_n\sum_{i=1}^{N}x_i^{\,n+2} = \sum_{i=1}^{N}x_i^{\,2} y_i \,,$$

M

$$a_0\sum_{i=1}^{N}x_i^{\,n} + a_1\sum_{i=1}^{N}x_i^{\,n+1} + a_2\sum_{i=1}^{N}x_i^{\,n+2} + \mathrm{L} + a_n\sum_{i=1}^{N}x_i^{\,2n} = \sum_{i=1}^{N}x_i^{\,n} y_i \,,$$

or, in matrix form,

$$\begin{bmatrix} N & \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_i^{\,2} & \mathrm{L} & \sum_{i=1}^{N}x_i^{\,n} \\ \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_i^{\,2} & \sum_{i=1}^{N}x_i^{\,3} & \mathrm{L} & \sum_{i=1}^{N}x_i^{\,n+1} \\ \sum_{i=1}^{N}x_i^{\,2} & \sum_{i=1}^{N}x_i^{\,3} & \sum_{i=1}^{N}x_i^{\,4} & \mathrm{L} & \sum_{i=1}^{N}x_i^{\,n+2} \\ \mathrm{M} & & & & \\ \sum_{i=1}^{N}x_i^{\,n} & \sum_{i=1}^{N}x_i^{\,n+1} & \sum_{i=1}^{N}x_i^{\,n+2} & \mathrm{L} & \sum_{i=1}^{N}x_i^{\,2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \mathrm{M} \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N}y_i \\ \sum_{i=1}^{N}x_i y_i \\ \sum_{i=1}^{N}x_i^{\,2} y_i \\ \mathrm{M} \\ \sum_{i=1}^{N}x_i^{\,n} y_i \end{bmatrix}. \qquad (3.3)$$

Equations (3.3) represent a linear system. However, this system is usually ill-conditioned and round-off errors can distort the solution of a_i's. Up to degree-3 or 4, the problem is not too great. It is very infrequent to use a degree higher than 4.

Example 3.3

Rotameter calibration data (flow rate versus Rotameter reading) are as follows:

Rotameter Reading R	10	30	50	70	90
Flow rate V(L/min)	20	52.1	84.6	118.3	151

a) Using curve fitting to fit the calibration data to second order polynomial.

b) Calculate the flowrate (V) at rotameter reading R=73.

Solution:

a) 2nd order polynomial

$$S_r = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - a_o - a_1 x_i - a_2 x_i)^2$$

$$\frac{dS_r}{da_o} = 2\sum_{i=1}^{n}(y_i - a_o - a_1 x_i - a_2 x_i^2)\times(-1) = 0$$

$$\frac{dS_r}{da_1} = 2\sum_{i=1}^{n}(y_i - a_o - a_1 x_i - a_2 x_i^2)\times(-x_i) = 0 \qquad (1)$$

$$\frac{dS_r}{da_2} = 2\sum_{i=1}^{n}(y_i - a_o - a_1 x_i - a_2 x_i^2)\times(-x_i^2) = 0$$

Re arranging above equations

$$a_o n + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$

$$a_o \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \qquad (2)$$

$$a_o \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

Making required table

	R	V	R^2	R^3	R^4	RV	R^2y
	10	20	100	1000	10000	200	2000
	30	52.1	900	27000	810000	1563	46890
	50	84.6	2500	125000	6250000	4230	211500
	70	118.3	4900	343000	2401000	8281	579670
	90	151	8100	729000	6561000	13590	1223100
∑	250	426	16500	1225000	9669000	27864	2063160

By substitution in equation 2

$$5a_o + 250a_1 + 16500a_2 = 426$$

$$250a_o + 16500a_1 + 1225000a_2 = 27864$$

$$16500a_o + 1225000a_1 + 96690000a_2 = 2063160$$

Solving above equation simultaneously gives;

ao =3.8786 , a1 =1.5981 , a2 = 4.2857×10^{-4}

then

$$V = 3.8786 + 1.5981 \times R + 4.2857 \times 10^{-4} \times R^2$$

B)

$$V(73) = 3.8786 + 1.5981 \times 73 + 4.2857 \times 10^{-4} \times 73^2 = 122.83$$

3.3 Nonlinear Data

Whenever data from experimental tests are not linear, we need to fit to them some function other than a first-degree polynomial. Popular forms that are tried are the power form

$y = ax^b$

or the exponential form

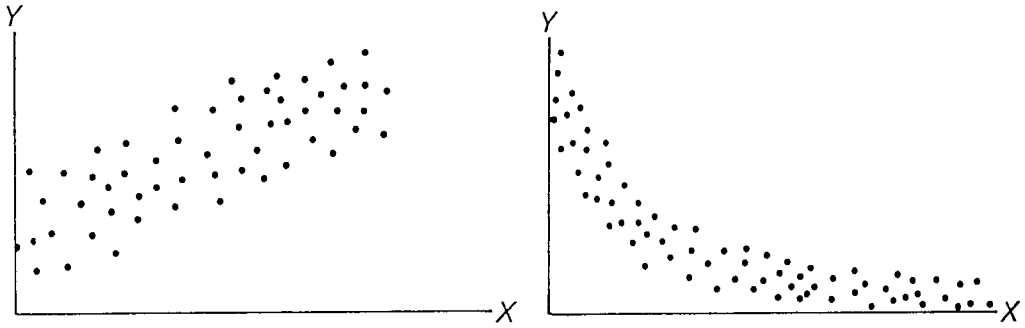
$y = ae^{bx}$.

Since such nonlinear equations are much more difficult to solve than linear equations, they are usually linearized by taking logarithms before determining the parameters:

$\ln y = \ln a + b \ln x$,

or

$\ln y = \ln a + bx$.



linear non-linear

Figure 3.2 Linear vs non-linear data

In cases when such linearization of the function is not desirable, or when no method of linearization can be discovered, graphical methods are frequently used; one merely plots the experimental values and sketches in a curve that seems to fit well.

Example 3.4

The progress of a homogeneous chemical reaction is followed and it is desired to evaluate the rate constant and the order of the reaction. The rate law expression for the reaction is known to follow the power function form $-r = kC^n$

Use the data provided in the table to obtain n and k.

C_A(gmol/l)	4	2.25	1.45	1.0	0.65	0.25	0.006
$-r_A$(gmol/l · s)	0.398	0.298	0.238	0.198	0.158	0.098	0.048

Solution

Taking the natural log of both sides of Equation, we obtain

$\ln(-r) = \ln(k) + n\ln(C)$

Let

$$z = \ln(-r)$$

$$w = \ln(C)$$

$$a_0 = \ln(k) \quad \text{implying that} \quad k = e^{a_0}$$

$$a_1 = n$$

We get

$$z = a_0 + a_1 w$$

This is a linear relation between z and w, where

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} w_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2}$$

$$a_0 = \left(\frac{\sum_{i=1}^{n} z_i}{n}\right) - a_1 \left(\frac{\sum_{i=1}^{n} w_i}{n}\right)$$

Table: Kinetics rate law using power function

i	C	$-r$	w	z	$w \times z$	w^2
1	4	0.398	1.3863	-0.92130	-1.2772	1.9218
2	2.25	0.298	0.8109	-1.2107	-0.9818	0.65761
3	1.45	0.238	0.3716	-1.4355	-0.5334	0.13806
4	1	0.198	0.0000	-1.6195	0.0000	0.00000
5	0.65	0.158	-0.4308	-1.8452	0.7949	0.18557
6	0.25	0.098	-1.3863	-2.3228	3.2201	1.9218
7	0.006	0.048	-5.1160	-3.0366	15.535	26.173
$\sum_{i=1}^{7}$			-4.3643	-12.391	16.758	30.998

$$n = 7$$

$$\sum_{i=1}^{7} w_i = -4.3643$$

$$\sum_{i=1}^{7} z_i = -12.391$$

$$\sum_{i=1}^{7} w_i z_i = 16.758$$

$$\sum_{i=1}^{7} w_i^2 = 30.998$$

From above equations

$$a_1 = \frac{7\times(16.758)-(-4.3643)\times(-12.391)}{7\times(30.998)-(-4.3643)^2}$$

$$= 0.31943$$

$$a_0 = \frac{-12.391}{7} - (.31943)\frac{-4.3643}{7}$$

$$= -1.5711$$

Then

$$k = e^{-1.5711}$$

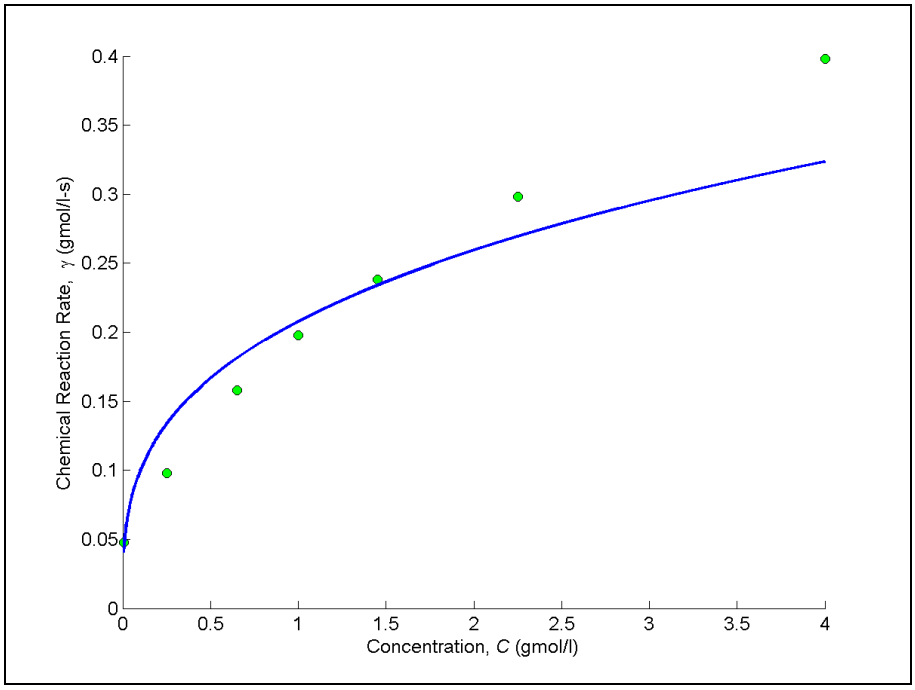
$$= 0.20782$$

$$n = a_1$$

$$= 0.31941$$

Finally, the model of progress of that chemical reaction is

$$-r = 0.20782 \times C^{0.31941}$$



Example 3.5

It is suspected from theoretical considerations that the rate of water flow from a firehouse is proportional to some power of the nozzle pressure. Assume pressure data is more accurate. You are transforming the data.

Flow rate, F (gallons/min)	96	129	135	145	168	235
Pressure, p (psi)	11	17	20	25	40	55

What is the exponent b of the nozzle pressure in the regression model $F = ap^b$

Solution

The linearization of the above data is done as follows.

$$F = ap^b$$

$$\ln(F) = \ln(a) + b\ln(p)$$

$$z = a_0 + bx$$

Where

$$z = \ln(F)$$

$$x = \ln(p)$$

$$a_0 = \ln(a)$$

Implying

$$a = e^{a_0}$$

There is a linear relationship between z and x.

Linear regression constants are given by

$$b = \frac{n\sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i z_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Since

$$n = 6$$

$$\sum_{i=1}^{6} x_i z_i = \ln(11)\times\ln(96) + \ln(17)\times\ln(129) + \ln(20)\times\ln(135) + \ln(25)\times\ln(145) + \ln(40)\times\ln(168) + \ln(55)\times\ln(235) = 96.208$$

$$\sum_{i=1}^{6} x_i = \ln(11) + \ln(17) + \ln(20) + \ln(25) + \ln(40) + \ln(55) = 19.142$$

$$\sum_{i=1}^{6} z_i = \ln(96) + \ln(129) + \ln(135) + \ln(145) + \ln(168) + \ln(235) = 29.890$$

$$\sum_{i=1}^{6} x_i^2 = (\ln(11))^2 + (\ln(17))^2 + (\ln(20))^2 + (\ln(25))^2 + (\ln(40))^2 + (\ln(55))^2 = 62.779$$

then

$$b = \frac{6\times 96.208 - 19.142\times 29.890}{6\times 62.779 - 19.142^2}$$

$$= \frac{577.25 - 572.15}{376.67 - 366.41}$$

$$= 0.49721$$

Example 3.6

The following data have been obtained for the decomposition of benzene diazonium chloride to chlorobenzene:

T (K)	313	319	323	328	333
k (s^{-1})	0.0043	0.0103	0.018	0.0355	0.0717

From this data, determine the pre-exponential factor *A* and activation energy *E*, assuming that the rate constant follows an Arrhenius form.

$$k = A\exp\left(\frac{-E}{RT}\right)$$

Solution:

$$\ln k = \ln A - \frac{E}{RT}$$

$$y = \ln k$$

$$x = 1/T$$

$$a_o = \ln A$$

$$a_1 = \frac{-E}{R}$$

We get

$$y = a_o + a_1 x$$

	T (K)	k (s^{-1})	x=1/T	y=ln k	x^2	xy
	313	0.0043	0.00319	-5.44914	1.02073e-05	-0.01741
	319	0.0103	0.00313	-4.57561	9.82695e-06	-0.01434
	323	0.018	0.00310	-4.01738	9.58506e-06	-0.01244
	328	0.0355	0.00305	-3.33822	9.29506e-06	-0.01018
	333	0.0717	0.00300	-2.63526	9.01803e-06	-0.00791
$\sum$			0.01548	-20.0156	4.79324e-05	-0.06228

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = -14612$$

$$a_o = \bar{y} - a_1 \bar{x} = 41.2272$$

$$a_o = \ln A = 41.2272 \Rightarrow$$

$$A = \exp(40.2272) = 8.0303 \times 10^{17}$$

$$a_1 = -E/R \Rightarrow$$

$$E = -a_1 \times R = -(-14612) \times 8.314 = 121480$$

A Matlab program for solving example 3.6 is listed in Table 3.1.

Table (3.1) Matlab code and results for solution example (3.6)

Matlab Code	**T=[313,319,323,328,333];** **K=[0.0043,0.0103,0.018,0.0355,0.0717];** **x=1./T;** **y=log(K);** **Poly=polyfit(x,y,1);** **E=-Poly(1)*8.314** **Ao=exp(Poly(2))**
Results	**E =** **1.2148e+05** **Ao =** **8.0303e+17**

The comparison between experimental and predicted k values is shown in below figure: