CHAPTER 1: Basic Concepts of Regression Analysis

Prof. Alan Wan

Table of contents

Introduction

- *Regression analysis* is a statistical technique used to describe relationships among variables.
- The simplest case to examine is one in which a variable Y, referred to as the *dependent* or *target* variable, may be related to one variable X, called an *independent* or *explanatory* variable, or simply a *regressor*.

Introduction

- *Regression analysis* is a statistical technique used to describe relationships among variables.
- The simplest case to examine is one in which a variable Y, referred to as the *dependent* or *target* variable, may be related to one variable X, called an *independent* or *explanatory* variable, or simply a *regressor*.
- If the relationship between Y and X is believed to be linear, then the equation for a line may be appropriate:

$$Y = \beta_1 + \beta_2 X,$$

where β_1 is an intercept term and β_2 is a slope coefficient.

Introduction

- *Regression analysis* is a statistical technique used to describe relationships among variables.
- The simplest case to examine is one in which a variable Y, referred to as the *dependent* or *target* variable, may be related to one variable X, called an *independent* or *explanatory* variable, or simply a *regressor*.
- If the relationship between Y and X is believed to be linear, then the equation for a line may be appropriate:

$$Y = \beta_1 + \beta_2 X,$$

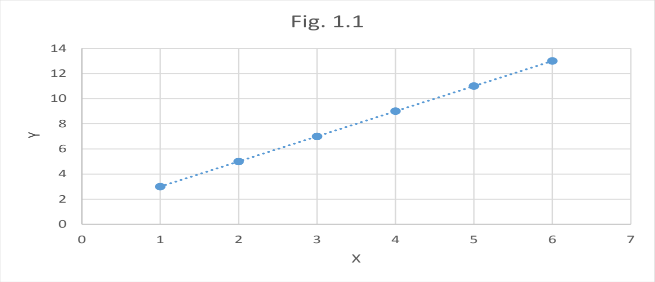
  where β_1 is an intercept term and β_2 is a slope coefficient.
- In simplest terms, the purpose of regression is to try to find the *best fit* line or equation that expresses the relationship between Y and X.

Introduction

- Consider the following data points

X	1	2	3	4	5	6
Y	3	5	7	9	11	13

- A graph of the (x, y) pairs would appear as



Introduction

- Regression analysis is not needed to obtain the equation that describes Y and X because it is readily seen that $Y = 1 + 2X$.
- This is an *exact* or *deterministic* relationship.

Introduction

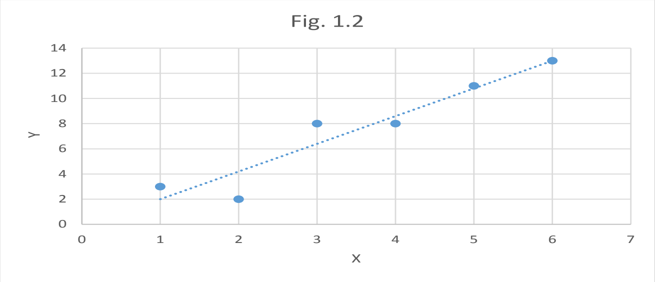
- Regression analysis is not needed to obtain the equation that describes Y and X because it is readily seen that $Y = 1 + 2X$.
- This is an *exact* or *deterministic* relationship.
- Deterministic relationships are sometimes (although very rarely) encountered in business environments. For example, in accounting:

$$\begin{aligned} \text{assets} &= \text{liabilities} + \text{owner equity} \\ \text{total costs} &= \text{fixed costs} + \text{variable costs} \end{aligned}$$

- In business and other social science disciplines, deterministic relationships are the exception rather than the norm.

Introduction

- Data encountered in a business environment are more likely to appear like the data points in this graph, where Y and X largely obey an approximately linear relationship, but it is not an exact relationship:



Introduction

- Still, it may be useful to describe the relationship in equation form, expressing Y as X alone - the equation can be used for forecasting and policy analysis, allowing for the existence of errors (since the relationship is not exact).
- So how to fit a line to describe the "broadly linear" relationship between Y and X when the (x, y) pairs do not all lie on a straight line?

Approaches to Line Fitting

- Consider the pairs (x_i, y_i). Let $\hat{y}_i$ be the "predicted" value of y_i associated with x_i if the fitted line is used. Define $e_i = y_i - \hat{y}_i$ as the *residual* representing the "error" involved.

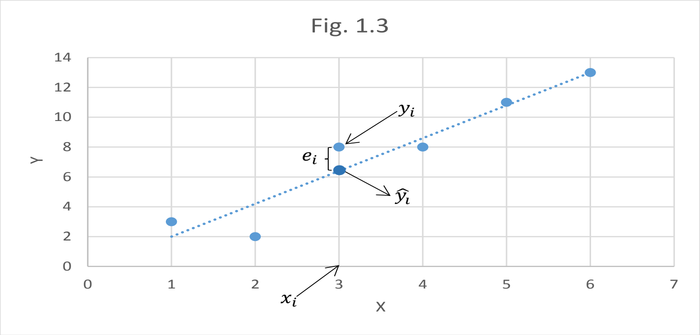
Approaches to Line Fitting

- Consider the pairs (x_i, y_i). Let $\hat{y}_i$ be the "predicted" value of y_i associated with x_i if the fitted line is used. Define $e_i = y_i - \hat{y}_i$ as the *residual* representing the "error" involved.
- If over- and under-predictions of the same magnitude are considered to be equally undesirable, then the object would be to fit a line to make the absolute error as small as possible, but noting that the sample contains n observations and given the relationship is inexact, it would not be possible to minimise all e_i's simultaneously.

Approaches to Line Fitting

- Consider the pairs (x_i, y_i). Let $\hat{y}_i$ be the "predicted" value of y_i associated with x_i if the fitted line is used. Define $e_i = y_i - \hat{y}_i$ as the *residual* representing the "error" involved.
- If over- and under-predictions of the same magnitude are considered to be equally undesirable, then the object would be to fit a line to make the absolute error as small as possible, but noting that the sample contains n observations and given the relationship is inexact, it would not be possible to minimise all e_i's simultaneously.
- Thus, our criterion must be based on some aggregate measures.

Approaches to Line Fitting



Approaches to Line Fitting

Several approaches may be considered:

- Eye-balling

Approaches to Line Fitting

Several approaches may be considered:

- Eye-balling
- Minimise the sum of the errors, i.e., $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)$

Approaches to Line Fitting

Several approaches may be considered:

- Eye-balling
- Minimise the sum of the errors, i.e., $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)$
- Minimise the sum of the absolute errors, $\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |(y_i - \hat{y}_i)|$.
  Although use of this criterion is gaining popularity, it is not the one most commonly used because it involves the application of linear programming. As well, the solution may not be unique.

The Least Squares Approach

- By far, the most common approach to estimating a regression equation is the *least squares* approach.

The Least Squares Approach

- By far, the most common approach to estimating a regression equation is the *least squares* approach.
- This approach leads to a fitted line that minimises the sum of the squared errors, i.e.,

$$
\begin{aligned}
\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2 .
\end{aligned}
$$

The Least Squares Approach

- To find the values of b_1 and b_2 that lead to the minimum,

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) = 0 \qquad (1)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2} = -2 \sum_{i=1}^{n} x_i (y_i - b_1 - b_2 x_i) = 0 \qquad (2)$$

- Equations (1) and (2) are known as *normal equations*.

The Least Squares Approach

- Solving the two normal equations leads to

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_1 &= \bar{y} - b_2\bar{x} \end{aligned}$$

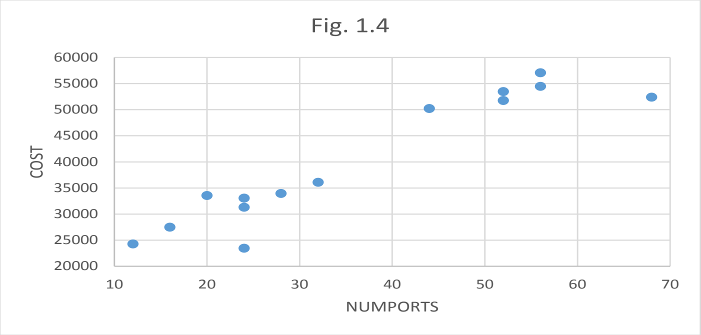
or

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ b_1 &= \bar{y} - b_2\bar{x} \end{aligned}$$

The Least Squares Approach

- **Example 1.1** The cost of adding a new communication node at a location not currently included in the network is of concern to a major manufacturing company. To try to predict the price of new communication nodes, data were observed on a sample of 14 existing nodes. The installation cost (Y=COST) and the number of ports available for access (X=NUMPORTS) in each existing node were available information.
- A scatter plot of the data is shown overleaf.

Approaches to Line Fitting



The Least Squares Approach

- We find

$$\sum_{i=1}^{n} x_i y_i = 23107792, \qquad \bar{x} = 36.2857,$$

$$\bar{y} = 40185.5, \quad \sum_{i=1}^{n} x_i^2 = 22576$$

- Using our least squares formulae,

$$\begin{aligned} b_2 &= \frac{23107792 - (14)(36.2857)(40185.5)}{22576 - (14)(36.2857)^2} \\ &= 650.169 \\ b_1 &= 40185.5 - (650.169)36.2857 = 16593.65 \end{aligned}$$

- The results obtained from EXCEL are shown overleaf.

The Least Squares Approach

Output 1.1: SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941928423
R Square	0.887229154
Adjusted R Square	0.877831584
Standard Error	4306.914458
Observations	14

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1751268376	1751268376	94.41048161	4.88209E-07
Residual	12	222594145.8	18549512.15		
Total	13	1973862522			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	16593.64717	2687.049999	6.17541437	4.75984E-05	10739.06816	22448.22618
X Variable 1	650.1691724	66.91388832	9.716505628	4.88209E-07	504.3763341	795.9620108

The Least Squares Approach

- So, the estimated equation relating the price of the new communication nodes to the number of access ports to the included at the node is estimated to be

$$\widehat{COST} = 16593.65 + 650.169 NUMPORTS$$

- Thus, the estimated installation cost for a node with 40 access ports is

$$42600.41 = 16593.65 + 650.169(40)$$

The Least Squares Approach

- **Example 1.2** A real estate appraiser uses the square footage of houses to derive individual appraisal values on each house. The sales values and size of 100 houses are available. The least squares results by EXCEL are shown in the output overleaf. The results show that the estimated equation is determined as

$$\widehat{VALUE} = -50034.607 + 72.82 SIZE$$

- If size were the only factor thought to be of importance in determining value, this equation could be used as a basis for appraisal. But obviously, other factors need to be considered. Developing an equation that includes more than one explanatory variable leads to the multiple regression approach.

The Least Squares Approach

Output 1.2: SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.815274956
R Square	0.664673254
Adjusted R Square	0.661251553
Standard Error	27270.25391
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.44459E+11	1.44459E+11	194.252262	5.49151E-25
Residual	98	72879341312	743666748.1		
Total	99	2.17338E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	- 50034.6065	7422.677496	-6.740776032	1.0951E-09	-64764.6684	-35304.544
X Variable 1	72.8203802	5.22480275	13.93744102	5.4915E-25	62.45192918	83.1888312

The Least Squares Approach

- The least squares approach has several advantages:
  - The objective function $\sum_{i=1}^{n} e_i^2$ is strictly convex, meaning that the solution is always unique.

The Least Squares Approach

- The least squares approach has several advantages:
  - The objective function $\sum_{i=1}^{n} e_i^2$ is strictly convex, meaning that the solution is always unique.
  - The solution is easy to compute.

The Least Squares Approach

- The least squares approach has several advantages:
  - The objective function $\sum_{i=1}^{n} e_i^2$ is strictly convex, meaning that the solution is always unique.
  - The solution is easy to compute.
  - The least squares approach leads to a set of well-established inference procedures (to be discussed in Chapter 2).

The Least Squares Approach

- The least squares approach has several advantages:
  - The objective function $\sum_{i=1}^{n} e_i^2$ is strictly convex, meaning that the solution is always unique.
  - The solution is easy to compute.
  - The least squares approach leads to a set of well-established inference procedures (to be discussed in Chapter 2).
  - The solution has an optimal mathematical property (to be discussed in Chapter 2).

The Least Squares Approach

- There are also disadvantages associated with the least squares approach:
  - The solution is sensitive to "outliers".

The Least Squares Approach

- There are also disadvantages associated with the least squares approach:
  - The solution is sensitive to "outliers".
  - The objective function $\sum_{i=1}^{n} e_i^2$ is symmetric, meaning that over- and under-estimation of the same magnitude will be equally penalised. In some situations, one of the two errors may be considered as more serious than the other.

The Least Squares Approach

- There are also disadvantages associated with the least squares approach:
  - The solution is sensitive to "outliers".
  - The objective function $\sum_{i=1}^{n} e_i^2$ is symmetric, meaning that over- and under-estimation of the same magnitude will be equally penalised. In some situations, one of the two errors may be considered as more serious than the other.
- That said, the advantages of least squares outweigh its disadvantages in most situations encountered in practice.

Statistical Model

- Thus far, the regression can be viewed as a descriptive statistic.
- However, the power of regression is not restricted to its use as a descriptive statistic for a particular sample, but more in its ability to draw inferences or generalisations about the entire population of values for the variables X and Y.

Statistical Model

- Thus far, the regression can be viewed as a descriptive statistic.
- However, the power of regression is not restricted to its use as a descriptive statistic for a particular sample, but more in its ability to draw inferences or generalisations about the entire population of values for the variables X and Y.
- To draw inference about a population from a sample, we must make some assumptions about how X and Y are related in the population. These assumptions will be spelt out in details in Chapter 2. Most of these assumptions describe an "ideal" situation. Later, some of these assumptions will be relaxed and we demonstrate modifications to the basic least squares approach that provide a method that is still suitable.

Statistical Model

- The basic assumption is the population regression line, written as

$$\mu_{Y|X} = \beta_1 + \beta_2 X,$$

where $\mu_{Y|X}$ is the conditional mean of Y given X.
- To explain, consider Example 1.1. Suppose X=NUMPORTS and Y=COST. Consider all possible communication nodes with 30 access ports. $\mu_{Y|X=30}$ is the average value of all communication nodes with 30 access ports.
- Suppose the computation of $\mu_{Y|X}$ can be done for a number of X values. The equation $\mu_{Y|X} = \beta_1 + \beta_2 X$ assumes that all of the conditional means lie on a straight line.

Statistical Model

- However, even for a given number of ports (say, X=30), the installation costs (Y's) are not all equal to the average $\mu_{Y|X=30}$; the actual costs are distributed around the point $\mu_{Y|X=30}$. The same holds for Y's associated with other X values.
- In other words, the actual Y's are distributed around the population regression line. Because of this variation, it is convenient to rewrite the equation representing an individual's response as

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

where $\epsilon_i = y_i - \mu_{Y|X=x_i}$ is called the disturbance.

Statistical Model

- To allow statistical inference from a sample to a population, assumptions are usually made about the disturbances, for example, $E(\epsilon_i) = 0$, the variance of each ϵ_i is equal to a common constant, the ϵ_i's across different i's are uncorrelated, etc. These will be spelt out in Chapter 2.

Statistical Model

- To allow statistical inference from a sample to a population, assumptions are usually made about the disturbances, for example, $E(\epsilon_i) = 0$, the variance of each ϵ_i is equal to a common constant, the ϵ_i's across different i's are uncorrelated, etc. These will be spelt out in Chapter 2.
- b_1 and b_2 are "estimators" of β_1 and β_2 respectively; the actual numerical values of b_1 and b_2 based on a given sample of observations are "estimates" of β_1 and β_2 respectively. b_1 and b_2 are not the same as β_1 and β_2.

Statistical Model

- Because b_1 and b_2 are derived based on the principle of least squares, they are called least squares estimators. Specifically, to distinguish them from other more complex form of least squares estimators, we call b_1 and b_2 the *ordinary least squares* (O.L.S.) estimators.

Statistical Model

- Because b_1 and b_2 are derived based on the principle of least squares, they are called least squares estimators. Specifically, to distinguish them from other more complex form of least squares estimators, we call b_1 and b_2 the *ordinary least squares* (O.L.S.) estimators.
- The more complex least squares estimators arise, for example, when the variance of ϵ_i varies across i's.

Statistical Model

- Because b_1 and b_2 are derived based on the principle of least squares, they are called least squares estimators. Specifically, to distinguish them from other more complex form of least squares estimators, we call b_1 and b_2 the *ordinary least squares* (O.L.S.) estimators.
- The more complex least squares estimators arise, for example, when the variance of ϵ_i varies across i's.
- It is instructive to recognise that ϵ_i and e_i are intrinsically different; ϵ_i is associated with the true regression model, while e_i arises from the estimation process.

Multiple Linear Regression

- In practice, more often than not, Y is determined by more than one factor. For example, In Example 1.2, size is rarely the only factor of importance in determining housing prices. Obviously, other factors need to be considered. A regression that contains more than one explanatory variable is called a multiple regression model.
- **Example 1.3** Observations are available for twenty five households on their annual total expenditure on non-durable goods and services (Y), annual disposable income (X_2), and stocks of liquid assets they hold (X_3). The figures are in thousands of dollars. The regression model is therefore:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i; \quad i = 1, \cdots, 25.$$

Multiple Linear Regression

- Let b_1, b_2 and b_3 be the estimators of β_1, β_2 and β_3 respectively.
- Differentiating $\sum_{i=1}^{n} e_i^2$ with respect to b_1, b_2 and b_3 yields the following normal equations:

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_1 - b_2x_{i2} - b_3x_{i3}) = 0 \tag{3}$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2} = -2\sum_{i=1}^{n}x_{i2}(y_i - b_1 - b_2x_{i2} - b_3x_{i3}) = 0 \tag{4}$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_3} = -2\sum_{i=1}^{n}x_{i3}(y_i - b_1 - b_2x_{i2} - b_3x_{i3}) = 0 \tag{5}$$

Multiple Linear Regression

- Equations (3), (4) and (5) can be solved for b_1, b_2 and b_3, but the solutions in terms of ordinary algebra are messy, and their algebraic complexity increases as k increases.
- To work with the general linear model, it simplifies matters if we make use of the linear algebra notations.

The linear model in matrix form

- When there are k coefficients and $k-1$ explanatory variables, the complete set of n observations can be written in full as

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_k x_{nk} + \epsilon_n \end{aligned}$$

where x_{ij} denotes the i^{th} observation of the j^{th} explanatory variable.

The linear model in matrix form

- In matrix algebra notations, these equations can be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{22} & x_{23} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The linear model in matrix form

- Thus, Y is a $n \times 1$ vector containing all of the observations on the dependent variable, X is $n \times k$ matrix containing all the observations on the explanatory variables (including the constant term), β is a $k \times 1$ vector of unknown coefficients we wish to estimate, and ϵ is $n \times 1$ a vector of disturbances.
- In compact notations, our model therefore becomes

$$Y = X\beta + \epsilon$$

The linear model in matrix form

Let

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \text{ be the O.L.S. estimator of } \beta.$$

Define

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = y - Xb$$

Note that $\sum_{i=1}^{n} e_i^2 = e'e$.

The linear model in matrix form

Thus,

$$
\begin{aligned}
e'e &= (Y - Xb)'(Y - Xb) \\
&= Y'Y - b'X'Y - Y'Xb + b'X'Xb \\
&= Y'Y - 2b'X'Y + b'X'Xb
\end{aligned}
$$

Note that if a and b are $k \times 1$, and A is $k \times k$ and symmetric, then

- $\frac{\partial a'b}{\partial b} = \frac{\partial b'a}{\partial b} = a.$
- $\frac{\partial b'Ab}{\partial b} = 2Ab$

The linear model in matrix form

Applying these results,

$$\frac{\partial e'e}{\partial b} = -2X'Y + 2X'Xb = 0,$$

leading to

$$\begin{aligned} X'Xb &= X'Y \quad \text{or} \\ b &= (X'X)^{-1}X'Y, \end{aligned} \tag{6}$$

provided that $X'X$ is non-singular.

The linear model in matrix form

The O.L.S. residual vector is $e = Y - Xb$. Note that

$$\begin{aligned} X'e &= X'(Y - Xb) \\ &= X'Y - (X'X)(X'X)^{-1}X'Y \\ &= 0, \end{aligned}$$

implying that the residual vector is uncorrelated with each explanatory variable.

The linear model in matrix form

For the special case of a simple linear regression model,

$$X'X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \text{ and } X'Y = \begin{bmatrix} \sum_{i=1} y_i \\ \sum_{i=1} x_i y_i \end{bmatrix}$$

giving

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

or

$$nb_1 + b_2 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_1 \sum_{i=1}^n x_i + b_2 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

as in equations (1) and (2).

The linear model in matrix form

Refer to Example 1.3,

$$(X'X)^{-1}X'y = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i3} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & \sum_{i=1}^{n} x_{i2}x_{i3} \\ \sum_{i=1}^{n} x_{i3} & \sum_{i=1}^{n} x_{i2}x_{i3} & \sum_{i=1}^{n} x_{i3}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \sum_{i=1}^{n} x_{i3}y_i \end{bmatrix}$$

giving

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 25 & 4080.1 & 14379.1 \\ 4080.1 & 832146.8 & 2981925 \\ 14379.1 & 2981925 & 11737267 \end{bmatrix}^{-1} \begin{bmatrix} 4082.34 \\ 801322.7 \\ 2994883 \end{bmatrix}$$

The linear model in matrix form

Or

$$\begin{aligned}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} &= \begin{bmatrix} 0.202454971 & -0.001159287 & 0.000046500 \\ -0.001159287 & 0.000020048 & -0.000003673 \\ 0.000046500 & -0.000003673 & 0.000000961 \end{bmatrix} \begin{bmatrix} 4082.34 \\ 801322.7 \\ 2994883 \end{bmatrix} \\ &= \begin{bmatrix} 36.789 \\ 0.332 \\ 0.125 \end{bmatrix}\end{aligned}$$

These estimates concur with the results produced by EXCEL shown overleaf.

The linear model in matrix form

Output 1.3: SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.891730934
R Square	0.795184059
Adjusted R Square	0.776564428
Standard Error	38.43646324
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	126186.6552	63093.33	42.70676	2.66073E-08
Residual	22	32501.95754	1477.362		
Total	24	158688.6128			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	36.79008153	17.29448828	2.127272	0.044851	0.923508064	72.65665499
X Variable 1	0.33183046	0.172101064	1.928114	0.066847	-0.025085301	0.688746221
X Variable 2	0.125785793	0.037687964	3.337559	0.002984	0.04762574	0.203945847

The linear model in matrix form

Hence the estimated regression equation is:

$$\hat{y}_i = 36.79 + 0.3318x_{i2} + 0.1258x_{i3}$$

A family with annual disposable income of \$50,000 and liquid assets worth \$100,000 is predicted to spend

$$\begin{aligned} \hat{y}_i &= 36.79 + 0.3318(50) + 0.1258(100) \\ &= 65.96 \end{aligned}$$

thousand dollars on non-durable goods and services in a year.