1. Simple Linear Regression Model

The major purpose of regression analysis is to explore the dependence of one variable, say *Y*, on another variable, say *X*. This implies exploring the existence of a functional form between *X* and *Y*. In particular and to simplify the relationship, the functional form is assumed to take a linear form. Before we begin our survey of regression methods, let us introduce some of the uses for regression strategies:

1. Descriptions: As a researcher, you may wish to seek some sort of descriptive relationship between a set of measured variables. At this point, you are making the fewest assumptions in your search for such a relationship. This relationship may or may not help to justify a possible deterministic relationship at this time; however, you are at least establishing some sort of connection with the sample of data you are currently working with. *Example: A sociologist may be interested in establishing a relationship between the final occupational status of an individual and the educational level of that individual as well as their parents' educational level.*

2. Coefficient Estimation: When analysing the data, the researcher may have a theoretical, deterministic relationship in mind. Whether this is linear or nonlinear, the use of regression analysis can provide evidence for such a theory (but note that we never say that we can prove such a theory - we can only provide evidence for such a theory). Of particular interest will be the magnitudes and signs of the coefficients, which will yield insight into the research questions at hand. *Example: A botanist may be interested in estimating the coefficients for an established model used for relating a certain plant's weight with the amount of water it receives, the nutrients in the soil, and the amount of sunlight exposure.*

3. Prediction: A researcher may be primarily concerned with predicting some response variable at given levels of other input variables. These predictions may be crucial in planning, monitoring, altering, or evaluating a process or system. For prediction to be valid, various assumptions must be made and met in this case. Most notably, you must not extrapolate beyond the range of the data since the estimation is only valid within the domain of the sampled data. *Example: An estate agent has a 20- year history of the home selling prices for those properties that she sold throughout her career as well as the home's total square footage, the year it was built, and the assessed value. She will put a new home on the market and wants to be able to predict that home's selling given the values of the other variables provided that they do not extend outside the domain of her data.*

4. Control: Regression models may be used for monitoring and controlling systems such that the functional relationship continues over time. If it does not, then continual modification of the model must occur. *Example: A manufacturer of semiconductors continuously monitors the camber measurement on the substrate to be within certain limits. During this process, a variety of measurements in the system are recorded, such as lamination temperature, firing temperature, and lamination pressure. These inputs are always controlled within certain limits and if the camber measurement exceeds the designed limits, then the manufacturer must take corrective action.*

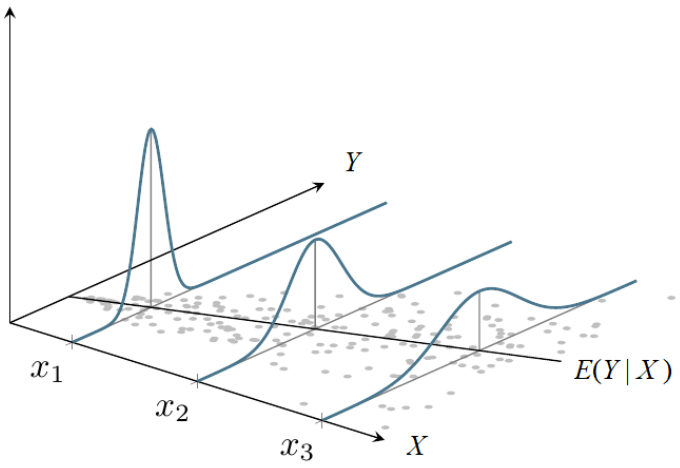
5. Variable Selection or Screening: A researcher may be faced with many independent variables and just one dependent variable. Since it is not feasible, nor necessarily informative, to model the dependent variable as a function of all of the independent variables, a search can be conducted to focus on only a subset of the independent variables that explain a significant amount of the variation in the dependent variable. Historical data may also be used to help in this decision process. *Example: A wine producer may be interested in assessing how the composition of his wine relates to sensory evaluations. A score for the wine's aroma is given by a judging panel and 25 elemental concentrations are recorded from that wine. It is then desired to see which elements explain a significant amount of the variation in the aroma scores.*

Population Regression Line

A regression model describes how the mean values of Y can relate to (be determined/predicted by) given values of X. Suppose $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ are the realisations of the random variable pairs (X, Y), then the population regression line is represented as:

$$E(Y \mid X = x_i) = \alpha + \beta x_i \qquad i = 1, ..., N$$

Geometrically, a population regression line is the locus of the conditional means of Y for a fixed value of X.



How about the relationship of individual value of Y given a value of X ?

At each value of X, ie $x_1, ..., x_N$, individual value of Y is around the mean and deviate from it by

$$\varepsilon_i = y_i - E(Y \mid X = x_i) \qquad \text{or} \qquad y_i = E(Y \mid X = x_i) + \varepsilon_i$$

Therefore, the overall simple linear regression model for individuals in the population from which the sample has been taken can be written as:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Where α and β are the unknown intercept and slope of the regression line, and ε, called the error (disturbance) term, is a random variable that stands for the approximation (simplification) of the relationship between X and Y.

Notes

- The variables X and Y are observable, whereas ε is not.
- In different contexts, the variable X is called the *right-hand-side variable/independent variable/regressor/predictor*, and the variable Y is similarly called the *left-hand-side variable/dependent variable/regressand/response*.

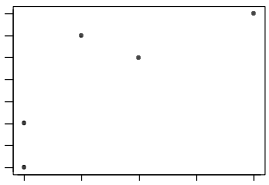
Assumptions on the Regression Model

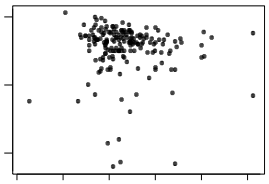
1. The observed values of *X* are fixed (independent of the error term).
2. The unobservable error terms have a mean zero, $E(\varepsilon) = 0$.
3. The error terms have the same constant (unknown) variance, $Var(\varepsilon) = \sigma^2$.
4. The error terms are uncorrelated, $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i . \varepsilon_j) = 0 \qquad \forall i \neq j$.

Sample Regression Line

Suppose that we take a random sample of size *n* (finite) from *X* and *Y*. That means pairs of observations as $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

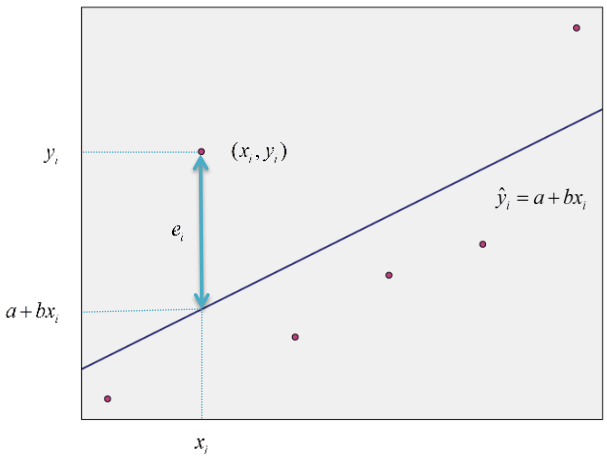
Having assumed a linear relationship between *X* and *Y*, the question is how to draw a line that fits the observations best. In other words, how to find good estimates of α and β. Naturally, if the number of observations is small, the visual inspection of the plot may lead to a reasonable fit. But a more formal approach is required for larger samples.





Method of Least Squares

The idea is to minimise the vertical distances between observations and the sample regression line. That is, the *least squares* (LS) estimates of the coefficients α and β are the values *a* and *b* for which the sum of squared residuals e_i is a minimum. The residuals are the differences between the observed (actual) values of *Y* and the fitted (predicted/estimated) values of *Y*, $e_i = y_i - \hat{y}_i$.



$$\min_{a,b} \; ESS = \sum_n e_i^2 = \sum_n (y_i - \hat{y}_i)^2 = \sum_n (y_i - a - bx_i)^2$$

The solution to the minimisation problem above is a set of *normal equations* which are solved to obtain the LS estimators of the population coefficients*.

$$b = \frac{\sum_{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{n}(x_i - \bar{x})^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$a = \bar{y} - b\bar{x}$$

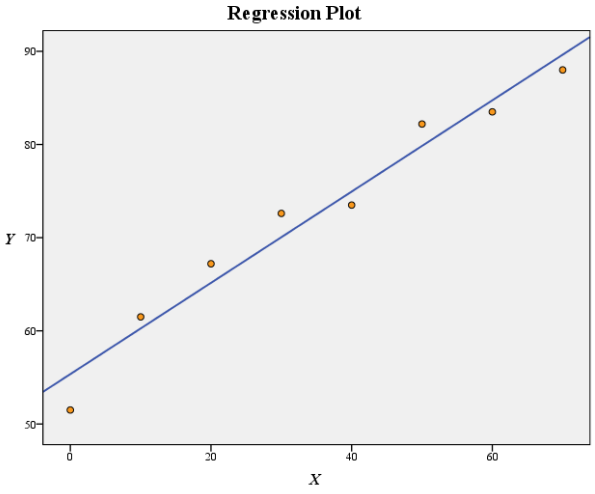
Notes

- It can be shown that $\sum e = 0$ and $\sum ex = 0$.
- The use of the LS method is justified by the *Gauss-Markov* theorem that states in the class of all linear and unbiased estimators, the LS estimators have the smallest variance.

Example 1.1

In an experiment, the weight of a given salt which dissolved in 100 grams of water was observed at 8 different temperatures. The results were as follows:

Temperature (*X*)	0	10	20	30	40	50	60	70
Weight (*Y*)	51.5	61.5	67.2	72.6	73.5	82.2	83.5	88.0



* The derivation of normal equations and LS estimators is shown in the appendix.

x	y	x^2	y^2	xy
0	51.5	0	2652.25	0
10	61.5	100	3782.25	615
20	67.2	400	4515.84	1344
30	72.6	900	5270.76	2178
40	73.5	1600	5402.25	2940
50	82.2	2500	6756.84	4110
60	83.5	3600	6972.25	5010
70	88.0	4900	7744.00	6160
280	**580**	**14000**	**43096.44**	**22357**

$$b = \frac{22357 - 8 \times 35 \times 72.5}{14000 - 8 \times 35^2} = \frac{2057}{4200} = .489762$$
$$a = 72.5 - .489762 \times 35 = 55.358333$$

$$\hat{y} = 55.36 + .49x$$

Notes

- In many cases the theory tells you which variable is the independent variable and which one is the dependent variable. In other cases, we arbitrarily decide how to set the variables. Notice that the formulae for *a* and *b* are not symmetrical in *X* and *Y*.
- The sample regression line always goes through the point $(\bar{x}, \bar{y})$.

Inference About β

So far, we have specified the population regression line as $Y = \alpha + \beta X + \varepsilon$ and a set of four assumptions that justifies the use of the LS estimation method. To study the sampling distribution of the LS estimators we further need to specify the distributional assumption for the error term.

If the assumptions 1-4 hold and if it is assumed that

$$\varepsilon \sim N(0, \sigma^2)$$

Then, an unbiased estimator for the error variance is $s^2 = \frac{\sum e_i^2}{n-2}$.

- An easy formula for calculating *ESS* is $\sum e^2 = \sum y^2 - a\sum y - b\sum xy$.

Moreover, the sampling distribution of b is:

$$b \sim N(\beta, \sigma_b^2)$$

An unbiased estimator for σ_b^2 is $s_b^2 = \frac{s^2}{\sum (x - \bar{x})^2}$. Hence, we can conclude that:

$$\frac{b - \beta}{s_b} \sim t_{(n-2)}.$$

Using this statistic, confidence intervals for β can be constructed and hypotheses on certain values for β can be tested.

Confidence Intervals & Hypothesis Tests

The sampling distribution of β suggests that a 95% CI for β is:

$$\Pr(-t_{.025} \le \frac{b - \beta}{s_b} \le t_{.025}) = \Pr(b - t_{.025}.s_b \le \beta \le b + t_{.025}.s_b) = .95$$

Example 1.2

From the previous example: $a = 55.358333, b = .489762, t_{.025}(\nu = 6) = 2.45$.

$$\sum e^2 = 43096.44 - 55.358333(580) - .489762(22357) = 38.997826$$
$$s^2 = \frac{38.997826}{8-2} = 6.4996$$
$$s_b^2 = \frac{6.4996}{4200} \qquad s_b = .03934$$

$$\Pr\left(.49 - 2.45(.03934) \le \beta \le .49 + 2.45(.03934)\right) =$$
$$\Pr\left(.39 \le \beta \le .59\right) = .95$$

Clearly, this CI can be used to test $H_0 : \beta = \beta_0$.

Alternatively, one can use the fact that under H_0:

$$t_0 = \frac{b - \beta_0}{s_b} \sim t(n-2)$$

Note

- An important value for β_0 is zero, ie testing the significance of β.

Example 1.3

From the previous example:

$$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$$

$$t_0 = \frac{.489762 - 0}{.03934} = 12.45$$

$$t_{.025}(6) = 2.45$$

$\therefore$ Reject H_0. X has a significant effect on Y.

Prediction

One of the most important applications of regression models is *prediction/forecast*. The idea is to use the regression line, based on a sample of size *n*, to find a corresponding value for the dependent variable for a specified value of the independent variable, x_{n+1}.

$$\hat{y}_{n+1} = a + bx_{n+1}$$

Note that since we are using the sample regression line, $\hat{y}_{n+1}$ is just an estimate (a prediction) of the true value of Y for a given value of X. Hence, we need to complement this with a confidence interval.

(i) If we are interested in the actual value of $\hat{y}_{n+1}$, then a 95% **PI** is:

$$\hat{y}_{n+1} \pm t_{.025,(n-2)}.s.\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

(ii) Alternatively, we may be interested in estimating (predicting) $E(\hat{y}_{n+1} | x_{n+1})$, that is the average value of the dependent variable when the independent variable is fixed at x_{n+1}. This time the **CI** is:

$$\hat{y}_{n+1} \pm t_{.025,(n-2)}.s.\sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Note

- For both intervals, the error is the smallest when $x_{n+1} = \bar{x}$.

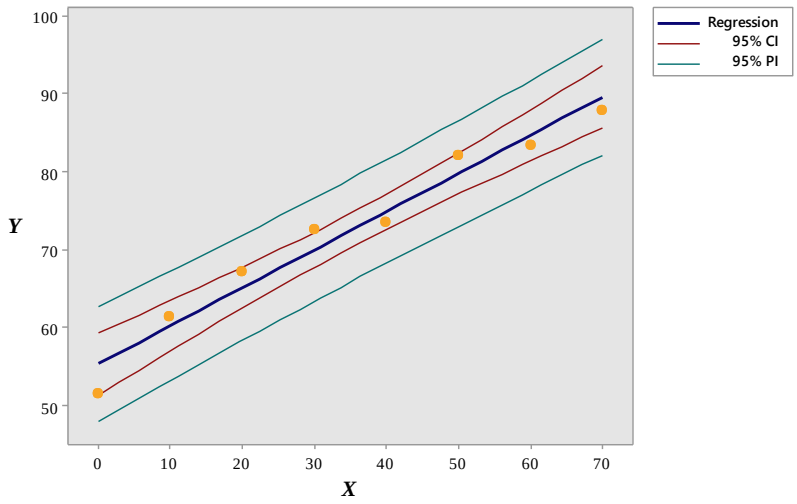
Example 1.4

For the previous example, suppose x is fixed at $x_{n+1} = 25^{\circ C}$, then

$$\hat{y} = 55.358 + .48976(25) = 67.602$$

The PI is $$67.602 \pm 2.45 \times 2.55\sqrt{1 + \frac{1}{8} + \frac{(25-35)^2}{4200}} = (60.914, 74.291).$$

The CI is $$67.602 \pm 2.45 \times 2.55\sqrt{\frac{1}{8} + \frac{(25-35)^2}{4200}} = (65.195, 70.010).$$



The Explanatory Power of Regression Model & Analysis of Variance

The regression analysis is an attempt to employ information on an independent variable, *X*, to *explain* the behaviour of a dependent variable, *Y*. The observations on *Y* show a certain amount of variation within the sample. The question is what proportion of this variation is explained by *X*.

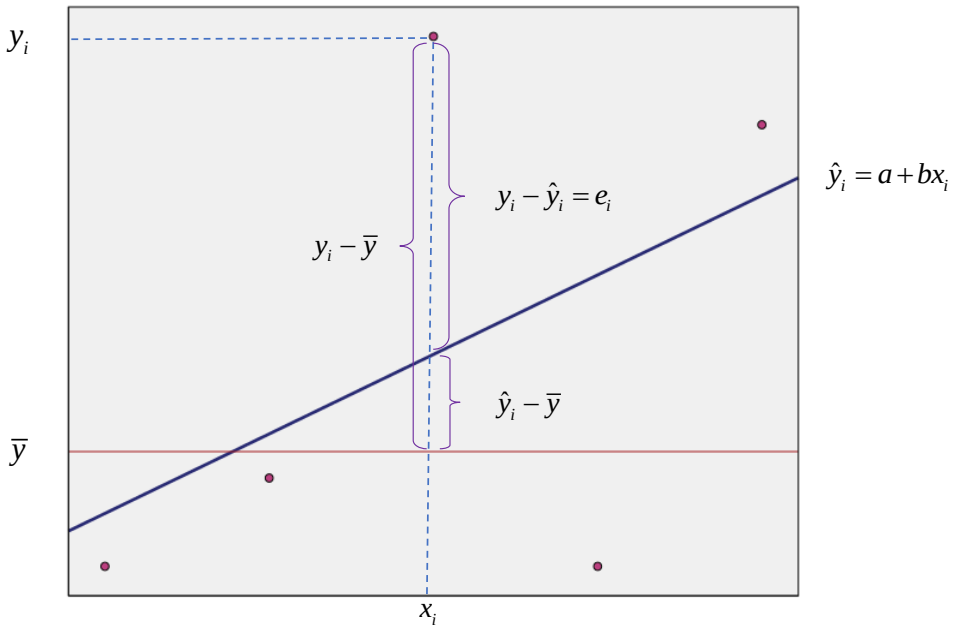
Recall that:

$$e = y - \hat{y} \qquad \text{or} \qquad y = \hat{y} + e$$

Note that the residual, *e*, represents the part of behaviour of *y* that cannot be explained by *x*.

Subtracting $\bar{y}$ from both sides yields:

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + e$$



Squaring both sides and summing over *n* observations:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum e^2$$

This equation has an important interpretation:

Total sample variation = Explained variation by the regression + Unexplained variation

Total Sum of Squares = Regression Sum of Squares + Error (Residual) Sum of Squares

TSS = RSS +ESS

These results show how a measure for the *goodness of fit* can be defined.

Coefficient of Determination, R^2

The coefficient of determination, R^2, is defined as the proportion of the total variation which is explained by the regression.

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

It follows from this definition that:

$$0 \le R^2 \le 1$$

Note
In a simple linear regression $R^2 = r^2$.

Analysis of Variance

The values defined in the previous section can conveniently be arranged in a table of ANOVA.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Ratio
	SS	df	MS	F_0
Regression	RSS	1	RSS/1	RMS/EMS
Error	ESS	*n*-2	ESS/(*n*-2)	
Total	TSS	*n*-1		

Example 1.5

Using the data from the previous example:

$$RSS = b^2 \sum (x - \bar{x})^2 = 1007.4$$
$$ESS = \sum e^2 = 39.0$$
$$TSS = \sum (y - \bar{y})^2 = \sum y^2 - n\bar{y}^2 = 1046.4$$

Source	SS	df	MS	F_0
Regression	1007.4	1	1007.4	154.99
Residual	39.0	6	6.5	
Total	1046.4	7		

F_0 is used to test $H_0 : \beta = 0$ against $H_0 : \beta \neq 0$:

$$F_0 = 154.99$$
$$F_{.05}(1, 6) = 5.99$$
$$\therefore \text{Reject } H_0.$$

Moreover, $R^2 = \frac{1007.4}{1046.4} = .963$, meaning that 96.3% of the variation in *y* is explained by the regression. A very good fit!

Notes

- The result of this *F* test is the same as the *t* test, since $F = t^2$, (in our example $12.45^2 \cong 154.99$).
- By definition *Residual MS* $= s^2$.
- From the value of R^2 we can conclude that $r = \sqrt{.963} = .981$ and *t* =12.45 can be used to test the significance of *r*. In other words, $H_0 : \rho = 0$ is equivalent to $H_0 : \beta = 0$.

Appendix

1. Proofs of Some Useful Relationships

$$\sum (x-\bar{x})(y-\bar{y}) = \sum (xy - \bar{x}y - \bar{y}x + \bar{x}\bar{y}) = \sum xy - \bar{x}\sum y - \bar{y}\sum x + n\bar{x}\bar{y} =$$

$$\sum xy - \frac{\sum x}{n}\sum y - \frac{\sum y}{n}\sum x + n\frac{\sum x}{n}\frac{\sum y}{n} = \sum xy - \frac{\sum x \sum y}{n} \quad (1)$$

$$= \sum xy - \frac{n\sum x \sum y}{n \times n} = \sum xy - n\bar{x}\bar{y} \quad (2)$$

Similar to (1) and (2) we also have:

$$\sum (x-\bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = \sum x^2 - n\bar{x}^2$$

$$\sum (y-\bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = \sum y^2 - n\bar{y}^2$$

2. Derivation of LS Estimators

$$\min_{a,b} : S = \sum e^2 = \sum (y-a-bx)^2$$

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum (y-a-bx) = 0 \\ \dfrac{\partial S}{\partial b} = -2\sum x(y-a-bx) = 0 \end{cases}$$

$$\begin{cases} \sum (y-a-bx) = 0 \quad \Rightarrow \quad \sum e = 0 \\ \sum x(y-a-bx) = 0 \quad \Rightarrow \quad \sum xe = 0 \end{cases}$$

$$\begin{cases} \sum y - na - b\sum x = 0 \quad (3) \\ \sum xy - a\sum x - b\sum x^2 = 0 \quad (4) \end{cases}$$

Equations (3) and (4) are called the ***normal equations***.

Divide (3) by n and solve for a:

$$\boxed{a = \bar{y} - b\bar{x}}$$

Substitute for a in (4):

$$\sum xy - (\bar{y} - b\bar{x})\sum x - b\sum x^2 = 0$$
$$\sum xy - \bar{y}\sum x + b\bar{x}\sum x - b\sum x^2 = 0$$

Factorise $-b$ and re-arrange:

$$\sum xy - \bar{y}\sum x = b(\sum x^2 - \bar{x}\sum x)$$

Set $\sum x = n\bar{x}$ and solve for b:

$$\boxed{b = \frac{\sum xy - n\overline{xy}}{\sum x^2 - n\bar{x}^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}}$$

3. Mean & Variance of LS Estimators

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum (x-\bar{x})y - \bar{y}\sum (x-\bar{x})}{\sum (x-x)^2} = \frac{\sum (x-\bar{x})y}{\sum (x-\bar{x})^2}$$

Let $\dfrac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = k_i$

Then $b = \sum k_i y_i$, this shows that b is a linear estimator in terms of y.

Note that:

$$\sum k = \sum \left(\frac{(x-\bar{x})}{\sum (x-\bar{x})^2} \right) = \frac{\sum (x-\bar{x})}{\sum (x-\bar{x})^2} = 0 \qquad (5)$$

$$\sum k^2 = \sum \left(\frac{(x-\bar{x})}{\sum (x-\bar{x})^2} \right)^2 = \frac{1}{\sum (x-\bar{x})^2} \qquad (6)$$

$$\sum kx = \sum \left(\frac{(x-\bar{x})x}{\sum (x-\bar{x})^2} \right) = 1 \qquad (7)$$

$\left(\text{Since } \sum (x-\bar{x})^2 = \sum (x-\bar{x})(x-\bar{x}) = \sum (x-\bar{x})x - \bar{x}\sum (x-\bar{x}) = \sum (x-\bar{x})x\right)$

Now:

$$b = \sum k_i y_i = \sum k_i(\alpha + \beta x_i + \varepsilon_i) = \alpha \sum k_i + \beta \sum k_i x_i + \sum k_i \varepsilon_i$$

Using (5) and (7)

$$b = \beta + \sum k_i \varepsilon_i \tag{8}$$

Taking expected value and noting that x is fixed:

$$E(b) = \beta + \sum x_i E(\varepsilon_i)$$

$$\boxed{\begin{array}{c} E(b) = \beta \\ b \text{ is an unbiased estimator of } \beta \end{array}}$$

From (8)

$$(b - \beta) = \sum k_i \varepsilon_i$$

Now:

$$\begin{aligned} Var(b) &= E\left(b - E(b)\right)^2 = E(b-\beta)^2 = E\left(\sum k_i \varepsilon_i\right)^2 \\ &= E(k_1\varepsilon_1 + k_2\varepsilon_2 + \ldots + k_n\varepsilon_n)^2 \\ &= E(k_1^2\varepsilon_1^2 + k_2^2\varepsilon_2^2 + \ldots + 2k_1k_2\varepsilon_1\varepsilon_2 + \ldots) \end{aligned}$$

Note that $E(\varepsilon_i^2) = \sigma^2$ and $E(\varepsilon_i \varepsilon_j) = 0, (i \neq j)$. Hence:

$$Var(b) = \sigma^2 \sum k_i^2$$

Using (6)

$$\boxed{Var(b) = \frac{\sigma^2}{\sum (x - \bar{x})^2}}$$

Similarly, one can show that:

$$\boxed{\begin{array}{l} E(a) = \alpha \\ Var(a) = \left(\dfrac{\sum x^2}{n \sum (x - \bar{x})^2} \right) \sigma^2 \\ Cov(a,b) = \left(\dfrac{-\bar{x}}{\sum (x - \bar{x})^2} \right) \sigma^2 = -\bar{x} Var(b) \end{array}}$$

Practical Week 1

The data used in **Example 1** is on Moodle as Salt.sav, a SPPS data file.

1. Retrieve the data to re-calculate different sections of the example. Make your own notes of the results which include scatter diagram together with regression line, linear regression analysis, including descriptive statistics and correlations.

2. Import the data into RStudio and repeat the above procedures in R.