COST ESTIMATION & INTERPRETATION.

3.1 Introduction:

In this topic, you will learn how to determine and evaluate the cost functions or models using statistical tools. Cost estimation is particularly critical for **construction industry. Large construction projects** are often obtained on the basis of competitive bids. The contractors that bid on these projects must have accurate cost estimation methods to **win their share of the bids and to be profitable.**

3.2 Definitions of terms:

Cost estimation is the process of developing a well-defined relationship between a cost object (dependent/response variable) and its cost driver(s) [independent variable(s)] for purpose of predicting/ estimating the cost [Blocher et al 2005**].**

Alternatively, a cost function is a mathematical description of how a cost change with changes in the level of an activity (cost driver) relating to that cost [Horngren et al 2006].

3.3 Role of cost estimation in strategic management:

Cost estimation facilitates strategic management in two ways:

1. It helps **predict future costs** using either previously identified activity based cost drivers, volume based cost drivers, structural based cost drivers or executional cost drivers.
2. It helps **identify the key cost drivers for a cost object** (i.e. maintenance cost can be affected by machine running hours, number of operators e.t.c.) and which of these cost drivers are more **useful** in predicting costs.

3.4 Using Cost Estimation To Predict Future Costs:

Applications of cost estimation include:

(a) To facilitate strategic positioning analysis:

(b) To facilitate value chain analysis:

(c) To facilitate target costing and life cycle costing:

3.5 Cost estimation for different types of cost drivers:

- Recall from introduction to managerial accounting in second year that, the cost estimation methods can be used in any of the four types of cost drivers namely: **activity, volume , structural and executional based drivers.**
- The relationships between **costs and activity based or volume based cost drivers** often are best fit by the **linear cost estimation methods** since these relationships are at least approximately linear within the relevant range of the firm's operations.
- **Structural cost drivers** involve long term plans and decisions that have a strategic impact on the firm. These decisions **include *manufacturing experience, scale of product, product or production technology, and product or production complexity***.
- ***Technology and complexity*** issues often lead management to use activity based costing and linear estimation methods.
- In contrast, **experience and scale often require nonlinear methods.** As a cost driver, experience **represents the reduction in unit cost due to learning**. The effect on total cost due to experience is nonlinear: that is, costs decrease with increased manufacturing experience. Similarly, the relationship between the structural cost driver, scale and total cost is nonlinear.

- **Scale i**s the term used to describe the ***manufacturing of similar products that differ in size***- for example, pipe valves of different capacity. A common effect of scale is that total manufacturing cost increases more rapidly than the increase in the size of the product i.e. the manufacture of 22-inch industrial valve requires more than twice the cost of an 11-inch valve. The relationship between manufacturing cost and the valve size can be predicted by a mathematical estimation model called **power law** that is used in industrial engineering.

3.6 Types Of Cost Functions or models:

Strictly speaking, there are two types of cost functions namely below:

1. Linear cost functions i.e.

Y = a + bX + e. ……. A single cost driver

Or

$y=a+b_1x_1+b_2x_2+------b_nx_n+e$	…multiple cost drivers

2. Nonlinear cost function – learning curves

A) Cumulative average time model ($\overline{Y_x}$):

$\overline{Y_x} = a x^b$

Where **($\overline{Y_x}$) = cumulative average time required to produce the first x-units.**

a= time taken to produce the first unit.

X= the number of units under consideration.

$$b = \frac{\log r}{\log 2}$$

Where r = learning curve improvement rate.

B) Cumulative total time model (Y_x)

$$y_x = (\bar{y})(x)$$

Or

$$Y_x = a\,x^{(b+1)}$$

3.7 Six steps of cost estimation:

At this point we are going to learn six steps followed in estimating the cost function or model. These are:

1. Select/identify/define the cost object (dependent/response variable)
2. Select or identify or determine the cost diver(s) or (explanatory or predictor or determinant variable).
3. Collect data on both the cost object and the cost driver(s)
4. Plot the data on a scattered diagram
5. Estimate the cost function i.e. Y= a + bX + e.
6. Evaluate the reliability of the estimated cost function

Step one: Identify (select) the cost object –(Y variable)

- The cost object is also referred to as either **dependent or response or criterion variable**.

- It represents the cost to be estimated and its choice depends on the purpose of the cost function. Examples include maintenance, labour, product costs e.t.c.

Step two: Select or identify or determine the cost driver(s) or - (x variable)

- The cost driver is also referred to as either **explanatory or predictor or determinant variable or the activity base.**
- Cost driver(s) represent(s) a factor(s) that causes variations in the cost object i.e. predicts the dependent variable (costs).
- Cost drivers are variables which explain changes in the cost object. They are causal factors used in the estimation of the cost. When a cost is an indirect cost, the independent variable is called **a cost allocation base**.
- The cost driver or activity base is a measure of whatever causes the incurrence of variable cost. The most common activity bases include**: direct labour hours, machine hours, units produced or sold, number of miles driven by salespersons, the number of letters typed by a secretary, the number of occupied beds in a hospital etc**.
- Identifying the cost driver(s) is the most important step in developing the cost estimate. A number of relevant drivers might exist, and some might not be immediately obvious.
- The cost object and the cost driver(s) must have an **economic plausibility and a logical relationship i.e. a cause and effect relationship and be measurable.**
- **Economic plausibility** means that the relationship (describing how changes in the cost driver lead changes in the cost being considered) is based on **a physical relationship, a contract, or**

knowledge of operation and makes economic sense to the manager and the management accountant.

- This means that all the individual items of costs included in the dependent variable should have the same cost driver or a pair of cost drivers.

Step three: Collect data on both the cost object and the cost driver(s)

- Let us note that this is usually the most difficult step in cost analysis.
- The data collected must be ***consistent and accurate*.**
- **Consistent data** means that each period of data is calculated on the same accounting basis and all transactions are properly recorded in the period in which they occurred.
- The **accuracy** of the data depends on the nature of the source. Sometimes, data developed within the firm are very reliable, as a result of management policies and procedures to ensure accuracy. In addition, accuracy varies among external sources of data, including governmental sources, trade and industries publications, universities e.t.c. Internal sources include company records, interviews with managers, special studies like pilot studies etc.
- The choice of cost driver(s) require(s) trade-offs between the relevance of the driver(s) and the consistency and accuracy of the data.
- One must obtain a sufficient number of past observations (data) so as to derive acceptable cost functions. Consider the example below:

Example 1:

Let's us assume that Mombasa port ltd is a company in charge of off-loading cargo from the ships docking at the port. The management is trying to develop the cost model for predicting future costs of

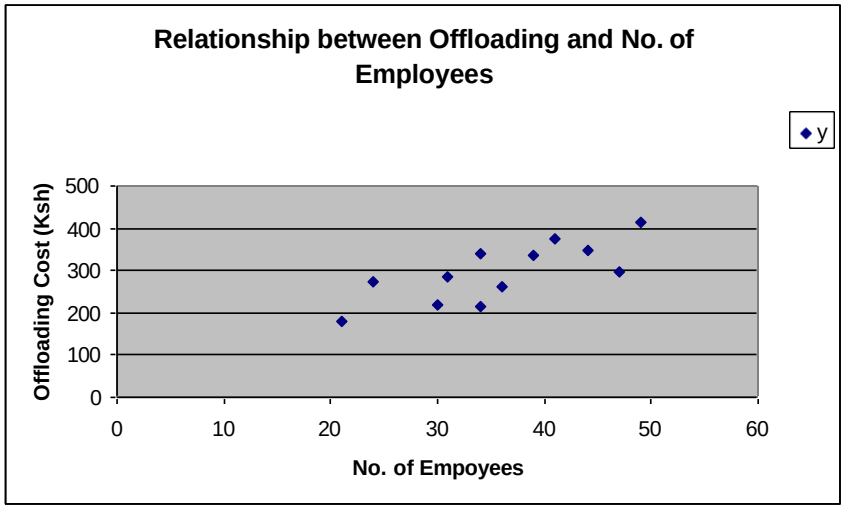
labour. The management believes that off-loading costs is a function of the number of employees working at the port. Data extracted from their records for the year **20x14** is as presented below:

Table 1: Number of employees & off-loading costs

Month	Number of employees	Off-loading costs
	[000s]	Kshs.[000s]
Jan	34	340
Feb	44	346
Mar	31	287
April	36	262
May	30	220
June	49	416
July	39	337
Aug	21	180
Sep	41	376
Oct	47	295
Nov	34	215
Dec	24	275

Step four: Plotting the data on a scattered diagram

- In this step we graph the data on a scattered diagram so as to identify unusual patterns. If we consider the data on example 1 above, the scattered diagram is presented below.
  - **Chart 1 : A scattered diagram**



- From the scatter diagram, **one can then make an intelligent guess of the most likely form of the model (function) e.g. linear, cubic, quadratic e.t.c.**

The main uses of a scatter-graph are:

1. It helps us identify the general nature of the observations or the general relationship between cost driver and cost. It provides a visual indication of the relationship between the cost object and the cost driver like a linear, cubic, quadratic relationship etc.
2. It helps us identify Outliers or extreme (abnormal) observations (observations outside the general pattern). They are unusual data points that strongly influence the model. These outliers should be investigated to ascertain whether they should be included or excluded from the analysis.
3. Last but not least, it helps us identify the relevant range i.e. band or range of activity within which the relationship between the cost object and the cost driver(s) is valid i.e. between 21 and 49 employees.

Step five: Estimate the cost function

- After exploring a variety of cost relationships, we should then select a method that can best be used in predicting the dependent variable.
- Therefore, at this step the analyst will be required to select and employ an estimation method that can accurately estimate the cost. The analysis can either be qualitative or quantitative in nature depending on the type of cost. Examples of these methods include;
    a. Engineering methods (Work measurements).
    b. Accounts analysis (Accounts Classifications or accounts inspection) method.

c. Scattered-graph method.
d. Two point (High-Low) method.
e. Regression analysis (least square) method.
f. Simulation analysis method.

- These methods we have mentioned differ in their ability to provide superior accuracy in cost estimation relative to the cost of the expertise and resources required. The management accountant should choose a method with the best precision/cost trade-off of the cost estimation objectives.
- These methods are not mutually exclusive and different methods may be used for different cost categories.

Step six: Evaluate the reliability of the cost function

- Our focus on this step is to evaluate the cost driver. This is a critical final step that considers the potential for error when the estimate is prepared. It involves considering the ***completeness and appropriateness of cost driver(s) selected in step 2, the consistence and accuracy of data selected in step 3, the study of graphs in step 4 and the precision of the method selected in step 5.***
- A common method of assessing the accuracy of an estimation method is to compare the estimates to the actual results over time i.e. a firm that predict overhead costs each year, over a 10 year period will result in 10 estimation errors to be evaluated. These errors can be evaluated using the mean absolute percentage error **(MAPE).**

- The **MAPE** is computed by taking the absolute value of each error, and then averaging these errors and converting the results to a percentage of the actual values of overheads.
- The main aim is to evaluate the cost driver(s) of the estimated cost function.

There are three (3) main tests for evaluating the cost function namely;

(a) The logical relationship test

(b) The goodness of fit tests

(c) Specification tests

<u>The logical relationship test:</u>

- The main question is that ''do you expect the cost driver **(X)** to cause a change in the cost object**(Y)?**'' i.e.' is there a cause and effect relationship, does the relationship make sense?'
- Without a cause-and-effect relationship, managers will be unable to estimate or predict costs and, therefore, will have difficulties managing those costs.

A cause and effect relationship might arise as a result of either:

1. **A physical relationship between the level of activity and costs**: i.e. when units of production is used as an activity measure that affects direct materials. Producing more units requires more direct materials, which results in higher total direct material costs. **Or**
2. **A contractual arrangement:** i.e. cost of materials or direct labour required or specified in the contract. **Or**

3. **Knowledge of operations:** i.e. when a number of parts is used as the activity measure of ordering costs. A product with many parts will incur higher ordering costs than a product with few parts.

NB: Managers must be careful not to interpret a high correlation or connection, in the relationship between two variables to mean that either variable causes the other.

- For example, higher production generally results in higher material costs and higher labour costs. Material and labour costs are highly correlated, but neither causes the other.
- Only a cause-and-effect relationship- and not merely correlation- establishes an economically plausible relationship between the level of an activity and its costs.

The goodness of fit tests

- The main question is 'how well does the estimated function (model) fit the actual data?
- The main aim is to test the strength or significance of the model.

There are two forms of tests for the goodness of fit namely:

1. **Tests on the overall cost model (function).**
2. **Tests of individual coefficients (cost drivers)**

Tests on the overall (whole) cost model (function).

This is done by using the following tests:

1. Coefficient of determination **(R –squared or r^2).**
2. Standard error of estimate **(Se).**
3. The **F**-test.

Tests of individual coefficients (cost drivers)

There are two categories of tests namely:

1. Testing the slope or coefficient(s) of the cost driver(s)
2. Testing the Y-intercept (a)-(TFC)

Testing the slope or coefficient(s) of the cost driver(s) i.e. bi

This is accomplished by using the following tests:

1. Coefficient of correlation **(r)**
2. Standard error of the slope **(Sb)**
3. **Z or T**-test.

- The main question is '**is the slope coefficient(s)-bi- statistically significant (that is, different from zero)?**

Testing the Y-intercept (TFC)

This is accomplished by computing the standard error of the Y-intercept **(Sa)**

Note that detailed discussion on the goodness of fit will be covered under regression analysis.

Specification tests (specification analysis)

- **Specification analysis** is the testing of the assumptions of the cost function.
- These are tests we use to test or evaluate the validity of the regression assumptions.

The assumptions of a linear relationship are:

(a) The relationship between the cost object (y) and the cost driver(s) (X) is linear within the relevant range.

(b) The cost driver or independent variable X is assumed to be known and is used to predict the cost object or dependent variable Y

(c) The errors (e) or the residuals given by $\sum(Y-\hat{Y})^2$ are assumed to:

1. Normally distributed
2. Have an expected value (mean) of Zero (0).
3. Have a constant variance- This is referred to as homoscedasticity. If not constant we have heteroscedasticity
4. Error are Independent of each other i.e. they are not serially correlated or there is no autocorrelation.

(a) <u>Linearity within the relevant range</u>

Where there is only one independent variable, the easiest way to check linearity is to study the data plotted on a scattered diagram and then make an intelligent guess.

(b) <u>Normality of residuals:</u>

- The error terms **(e**) are normally distributed around the cost function. This assumption is necessary in making inferences about **a, b,** $\hat{y}$**.**
- Use a histogram to test normal distribution of error terms (a chart of frequency against error terms).
- Also one can study the scattered diagram to test for normality of residuals.
- The error terms have a mean of zero i.e. expected value of errors E (e) =0.

(c) <u>Constant variance of residuals (error) terms or homoscedasticity.</u>

- **Residual or error or disturbance term (e) is** the vertical deviation of the observed value **Y** from the cost line estimate $\hat{y}$;

Where **e= Y-**$\hat{y}$. **Y** represents the observed or actual value while $\hat{y}$ is the estimated value of the function line.

- The assumption of constant variance implies that the residual terms are unaffected by the level of the cost driver. It also implies that there is a uniform scatter, or dispersion, of the data points about the estimated cost function.
- Violation of this assumption is called **heteroscedasticity (i.e. non constant error terms).**
- **Heteroscedasticity** does not affect the accuracy of the function estimates **a** and **b.** It does, however, reduce the reliability of the estimates of the standard errors and thus affect the precision with which inferences about the population parameters can be drawn from the function estimates.
- One can study the scattered diagram to test for constant variance of residuals.

(d) <u>Independence of residuals:</u>

- It is assumed that the residual term of one observation is not related to the residual term for any other observation.
- The problem of **serial correlation (also called autocorrelation)** in the residual arises when there is a systematic pattern in the sequence of residuals (errors) such that the residual in observation **n** conveys information about the residuals in observations **n+1, n+ 2 etc.**
- Non serial correlation means that an error term in observation one (period one) has no relationship with another observation.
- Autocorrelation is used in situation where we only have one cost driver (independent variable) while **multicollinearity** is used where we have more than one cost driver.

- **A scattered diagram** is used to identify autocorrelation i.e. errors move in the same direction and stick to each other around the estimated cost function.
- Serial correlation does not affect the accuracy of the function estimates of **a** and **b.** It does, however, affect the estimates of the standard errors of the coefficients, which in turn affect the precision with which inferences about the population parameters can be drawn from the function estimates.
- **Durbin-Watson** statistic is also another measure of serial correlation **(autocorrelation**) in the estimated residuals i.e. samples of 10 to 20 observations, a Durbin-Watson statistic in the range of **1.10 to 2.90** range indicates that the residuals are independent. Thus, an assumption of independence in the estimated residuals is reasonable for this cost model.
- Where we have more than one cost driver, the major problem is multicollinearity. We test **for multicollinearity** by using the**'t'-statistic i.e. a** cost driver with **a high 'T'-value should be used to develop the model.**
- **Multicollinearity** exists where two or more independent variables (cost drivers) are highly correlated with each other.

3.8 Regression analysis or least squire method

- This's a statistical method for obtaining the unique cost estimating equation that best fits a set of data points(observations) i.e. it identifies an estimated relationship between a dependent variable (cost object) and one or more independent variables [cost driver(s)].
- Where only one cost driver (one independent variable) is used, it called **a simple regression** analysis.

- Where two or more independent variables are used, it is called **multiple regression analysis.**
- It is based on the assumption that the sum of squares of the vertical deviations from the line established is the least possible.
- It fits the data by **minimizing the sum of squares** of the estimating **errors (e)** hence the reason why it is referred to as the **least square method. That is**

Minimize $\sum_{t=1}^{n}\left(y-\hat{y}\right)^{2}$

- Each error (e) is the vertical deviation (distance) measured from the regression line ($\hat{Y}$ **)** to the one of the actual data point (Y) i.e.

$$e = y - \hat{y}$$

The theoretical (general) linear model is given as:

$$y=a+b_1x_1+b_2x_2+------b_nx_n+e$$

Where Y = dependent variable (cost object)

Xi =ith predictor variable (cost driver) - (i = 1, 2 --- n).

bi=ith regression coefficient or actual change in Y for each unit change in Xi.

e=error term (residual value or disturbance term).

- The Error (residual) term (e) lumps together all the variations in Y which are not explained (not attributable to) by the predictor variable**(s).**
- An estimated cost model has no error term (e).

3.9 Evaluating Simple regression model

A simple regression equation in the form of $\hat{y} = a + bx$ is estimated by solving the following simultaneous equations:

$$\sum y = na + b\sum x \quad \text{..... (i)}$$

$$\sum xy = a\sum x + b\sum x^2 \quad \text{..... (ii)}$$

Then solve for (a) and (b) using matrices**, elimination or substitution methods.**

OR

The coefficients can be calculated as:

$$b = \frac{n\sum xy - \sum x.\sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a = \frac{\sum y\sum x^2 - \sum x\sum xy}{n\sum x^2 - (\sum x)^2}$$

Or: using the formulae below, try to compute for (a)

$$a = \frac{\sum y - b\sum x}{n} \quad \text{or } a = \left(\bar{y}\right) - b\left(\bar{x}\right)$$

Where:

$$\left(\bar{y}\right) = \frac{\sum y}{n} \quad \text{And} \quad \left(\bar{x}\right) = \frac{\sum x}{n}$$

Mombasa ltd illustration:

Periods	No. of employees (000)	Offloading Costs (000)			
	x	y	xy	x2	y2
Jan	34	340	11,560	1,156	115,600
Feb	44	346	15,224	1,936	119,716
Mar	31	287	8,897	961	82,369
April	36	262	9,432	1,296	68,644
May	30	220	6,600	900	48,400
June	49	416	20,384	2,401	173,056

July	39	337	13,143	1,521	113,569
Aug	21	180	3,780	441	32,400
Sep	41	376	15,416	1,681	141,376
Oct	47	295	13,865	2,209	87,025
Nov	34	215	7,310	1,156	46,225
Dec	24	275	6,600	576	75,625
N=12	$\sum x$ **=430**	$\sum y$ **=3,549**	$\sum xy$ **=132,211**	$\sum x^2$ **=16,234**	$\sum y^2$ **=1,104,005**

$$b = \frac{12*132,211 - 430*3,549}{12*16,234 - (430)^2} = \mathbf{6.10}$$

$$a = \frac{3,549*16,234 - 430*132,211}{12*16,234 - (430)^2} = \mathbf{77.08}$$

Hence a= 77.08 * 1000= Sh.77, 080

Therefore the estimated cost function or predicting function will be:

$\hat{y} = Sh77,080 + Sh.6.10x$

Suppose 10,000 employees will be used in the next year, what will be the offloading cost?

$\hat{y} = Sh77,080 + Sh.6.10(10,000) = \text{Sh } 138,080$

NB: Application of regression analysis usually involves twelve (12) or more observations (data points).

Evaluating Or Testing The Estimated Regression Cost Function

- At this point we test for the **precision and reliabilit**y of the estimated cost function.
- **Precision** refers to the accuracy of the estimates from the regression

- **Reliability** indicates whether the regression reflects an actual relationship among the variables, that is, is it likely to continue to predict accurately?

We can recall that t**here are three (3) main tests for evaluating the cost function namely;**

(a) The logical relationship test

(b) The goodness of fit tests

(c) Specification tests

3.9.1 Specification tests and logical relationship

NB: For specification tests and logical relationship refer to our earlier discussion in step six under section 3.7.

3.9.2 Testing/Evaluating The Goodness Of Fit

At this point we shall perform tests on the overall cost function and the coefficients**.**

3.9.2.1 Testing the whole model (entire cost function):

Here we shall compute coefficient of determination, standard error of estimate and F statistic.

A) Coefficient Of Determination (R-Squared or r^2)

- It measures the degree to which changes in the dependent variable (cost object) can be predicted or explained by changes in the independent variable(s) or cost driver(s).
- The number lies between **zero and one or as a percentage-**and often describes the explanatory power of the regression i.e. it measures the amount of variations in **Y** explained by **x**.
- A more reliable regression is one that has an R-squared **close to one (1)** i.e. as a rule of thumb R-squared should be **0.8** and **above**.

- A perfect linear relationship between cost (y) and cost driver (x) could result in R-squared being equal to **one** or **100%**.
- If viewed graphically, regression with a high R-squared show the data points lying near the regression line while that with low R- squared, the data points are scattered about.
- If the regression line estimated was to fit the actual observations perfectly, then all the observed points could lie on the regression line estimated.
- Reliability therefore, is based on the size of deviations of actual observations from the estimated values of the regression line established i.e. $Y = \hat{Y} + e$

Where Y= is actual observed point

$\hat{y}$ = are explained deviations

e = are unexplained deviations.

- This means that total deviations from the observed point **(Y)** can be given by the **explained deviations** ($\hat{y}$**) plus unexplained deviations (e).**
- Explained deviations (Variations) are measured by the coefficient of determination **(R-squared).**

$$r^2 = \frac{\sum\left(\hat{y} - \bar{y}\right)^2}{\sum\left(y - \bar{y}\right)^2} = \frac{Explained\ \text{variations}\,(SSR)}{Total\ \text{variations}\,(SST)}$$

Where $(\hat{y})$ = estimated cost value from the regression function.

Y= actual observation point.

$\left(\bar{y}\right)$ = the mean of y.

$$SST = \sum\left(y - \bar{y}\right)^2 \quad \textbf{Or}$$

$$\textbf{SST} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 1{,}104{,}005 - \frac{(3{,}549)^2}{12} = 54{,}388.25$$

$$SSR = \frac{\left[\sum xy - \frac{\sum x \sum y}{n}\right]^2}{\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]} \quad \text{Or} \quad SSR = \sum\left(\hat{Y} - \bar{Y}\right)^2$$

NB: consider the expression below.

$$\sum\left[y - \left(\bar{y}\right)\right]^2 = \sum\left[\left(\hat{y}\right) - \left(\bar{y}\right)\right]^2 + \sum\left[\left(y\right) - \left(\hat{y}\right)\right]^2$$

SST = SSR + SSE

Where SST=Total sum of squares or total variation

SSR= sum of square due to regression or variations due to regression.

SSE = Error sum of squares or unexplained variations.

$$SSR = \frac{\left[\sum xy - \frac{\sum x \sum y}{n}\right]^2}{\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]} \rightarrow SSR = \frac{\left[132{,}211 - \frac{430 * 3549}{12}\right]^2}{\left[16{,}234 - \frac{430^2}{12}\right]} = \textbf{30,746.6479}$$

NB: For a good regression model, SSR should be large as compared to SSE. In an ideal case SSE should be equal to zero.

SSE=SST-SSR.

Hence SSE= 54,388.25-30,746.6479= 23,641.6021

$$r^2 = 1 - \frac{SSE}{SST} \quad \textbf{Or} \quad r^2 = 1 - \frac{\underline{unexplained\ variations}}{total\ variations}$$

Remember:

$$SSE = \left(Y - \hat{Y}\right)^2$$

For computation purpose:

$$r^2 = \frac{\left(n\sum xy - \sum x \sum y\right)^2}{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}$$

For Mombasa ltd:

$$r^2 = \frac{\left[12*132{,}211 - 430*3{,}549\right]^2}{\left[12*16{,}234 - (430)^2\right]\left[12*1{,}104{,}005 - (3{,}549)^2\right]} = 0.5653 \text{ or } 56.53\%$$

- This means that about **56.53%** of the variations in off-loading cost are explained by variations in the number of employees whereas **43.47%** is explained by other independent variables (cost drivers) and the error term.
- This means that, the analyst might have left a major independent variable e.g. the number of hours worked. Therefore, more studies or research should be carried out.

B) Standard Error Of Estimate (Se)

- The coefficient of determination gives us an indication of the reliability of the estimated total cost based on the regression

equation but it does not provide an indication of the absolute size of the probable deviations from the true line. This information can be obtained by calculating the standard error of estimate **(Se).**

- The calculation of **Se** is necessary because the least square line is calculated from a sample data implying that other samples could probably result in different lines or different estimates.
- Obtaining the least square estimation of all the possible observation that might occur could result in the **'True least square line'.** In reality however, this is not possible and hence the need for **Se.**
- **The question that should be considered is 'how close is the sample estimate of the least square line to the true least square line?**
- **Se** enables one to estimate the confidence interval.
- **Se** is a measure of variability around the regression line and therefore it's similar to standard deviation under normal distribution.

$$Se = \sqrt{\frac{\sum\left(y - \hat{y}\right)^2}{n-k-1}} \quad \Rightarrow \quad Se = \sqrt{\frac{SSE}{n-k-1}}$$

Where **n-k-1** = degree of freedom.

n =number of observations

K = number of independent variables or cost drivers

1 = a constant.

For computation:

$$Se = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-k-1}}$$

$$Se = \sqrt{\frac{1{,}104{,}005 - 77.08 * 3{,}549 - 6.10 * 132{,}211}{12 - 1 - 1}}$$ **= 48.95**

The sample size, n, is reduced by 2 because 2 variables 'a' & 'b' in the regression equation had to be estimated from the sample observations.

Question:

Assume that 37,000 employees in the 13th month (i.e. Jan 20X15) could be used; estimate the off-loading costs for the 13th month and establish a 95% confidence interval on this cost.

Interval is given by:

$$Y = \hat{y} \pm t_c . Se$$

Where $\hat{y}$ = estimated value by regression line.

t_c = table value of 't' statistics.

Se = standard error of estimate.

$$t_c = t_{n-k-1,\alpha/2} = t_{12-1-1,0.025} = t_{10,0.025} = 2.2281$$

- **Degree of freedom (d.f) or n-k-1= 12-1-1= 10**
- **Area of significance (rejection area) or (α) is 2.5% or 0.025 =two tailed test**

$\hat{y}$ = 77.08 + 6.10*37 = Sh.302.78.

Interval will be:

[302.78- 2.2281(48.95) ≤ Y ≤ 302.78+ 2.2281(48.95)]

[Sh.193.715 ≤ Y ≤ sh.411.845]*1000

[Sh.193,715 ≤ Y ≤ sh.411,845] → the analyst is 95% confident that off-loading cost will lie within this interval.

NB : Where the number of observations is more than thirty (30) use **'Z'** statistics instead of 't' statistics.

C) F-Test Statistics

- We can test the significance of the regression results by using the F-statistics. The F-statistics is a ratio which compares the explained sum of squares and the unexplained sum of squares.

The Steps followed in the F- Test are:

1. State the hypothesis
2. State the significant level
3. State the test statistics
4. State the decision rule
5. Computation of F statistics
6. Conclusion

Step 1: State the hypothesis

- In this step we test the assumption or hypothesis of the slopes (**bi) or coefficients** or (independent variables) i.e. **HO → b_1, b2------ b_n = 0**. This expression implies that all parameters are equal to zero. That is; none of the predictor variable has any effect on the dependent variable (cost object) and thus the model is useless or the model is insignificant.
- **HA: bi ≠ 0 i.e.** at least one predictor variable has a significant effect (explanatory power) on the dependent variable(y) of the model or the model is significant.

Hypothesis to be tested:

Ho: b= 0 i.e. number of employees has no significant explanatory power on off- loading cost.

- **HA: b≠ 0 i.e.** number of employees has a significant explanatory power on off- loading cost.

Step 2: State the significant level

$\propto$ = 5%. Remember the confidence interval is at 95%

Step 3 State the test statistics

Computed F is given by:

$$F = \frac{SSR/k}{SSE/n-k-1} \quad \textbf{OR F=} \frac{r^2/k}{(1-r^2)/(n-k-1)}$$

For Mombasa ltd assuming 95% interval :

Table value of F [$f_{k, n-k-1, \propto}$] = 4.965

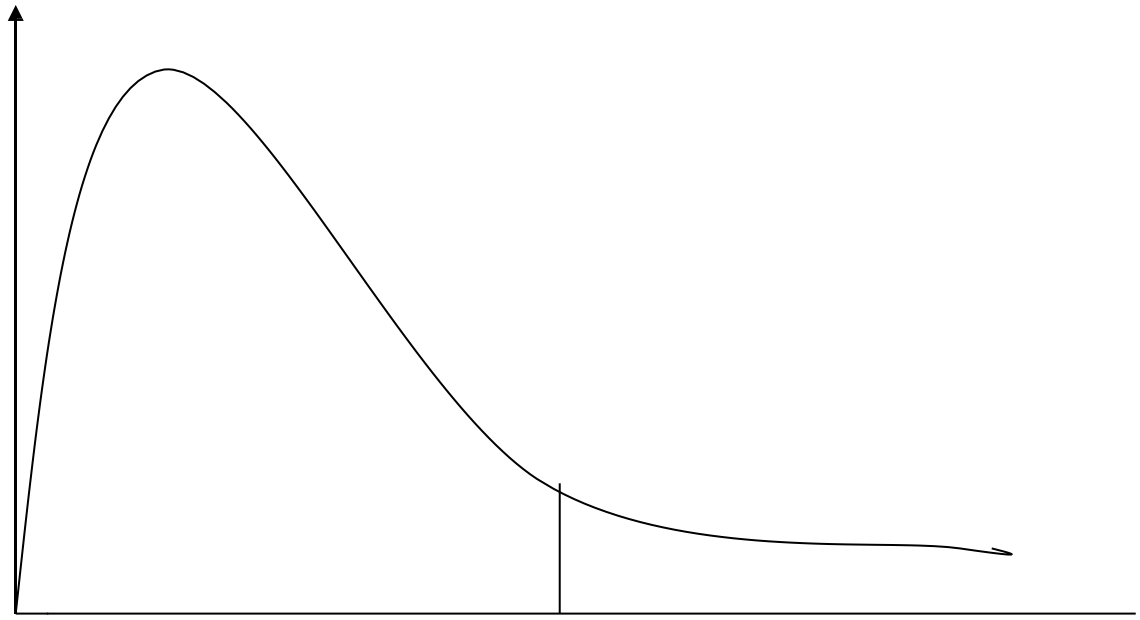
Where k=degree of freedom for the numerator i.e. 1

n-k-1= degree of freedom for the denominator i.e. 12-1-1=10

n-k-1 =10.

$\propto$ =significance area i.e. 5% or 0.05.

$f_{1, 10, 0.05}$ **= 4.965.**



4.965

k = 1

n-k-1 = 10

$\propto$ = 0.05

F_c = 4.965

Step 4 State the decision rule

RULE: If the computed value of [F] is greater than f_c (table value or $f_{k, n-k-1, \propto}$) then reject the null hypothesis and conclude that x has a significant explanatory power on dependent variable (y).

Step 5 Computation of F statistics.

$$F = \frac{SSR/k}{SSE/n-k-1} \quad \textbf{OR} \quad \textbf{F=} \frac{r^2/k}{(1-r^2)/(n-k-1)}$$

We substitute the figure as follows

$$F = \frac{30{,}746.6479/1}{23{,}641.6021/(12-1-1)}$$ **=13.0053 OR F=** $$\frac{0.5653/1}{1-0.5653/12-1-1}$$ **= 13.0044**

Step 6 conclusion

Since the **computed value (F=13.0053) is greater** than the **table value (f_c =4.96), reject the null hypothesis** and conclude that the cost driver (independent variable or the number of employees) has a significant explanatory power over variations in the cost object (dependent variable or off-loading cost**).**

NB: For a simple regression, testing the whole model is enough- doesn't require the testing of the slopes (individual regression coefficients). However, for a multiple regression analysis, testing the slopes (**bi**) is necessary.

3.9.2.2 Testing The Slope Or Coefficient(S) Of The Cost Driver(S) I.E. Bi

To test the slope, we shall compute the following:

1. Correlation coefficient (r)
2. Standard error of the slope (Sb)
3. Z or t statistics.

A) Coefficient of Correlation (r)

- It measures the degree (strength) and direction of association (relationship) between two variables.
- Used to determine the strength or weakness of the association between the response variable and the predictor variable (s). Unlike regression analysis, in correlation, no variable is considered dependent or independent.

- If the degree of association is very close, then it is possible to plot all the observations on a straight line and coefficient of correlation will almost be equal to one (1).

Coefficient of correlation (r) ranges between; $-1 \leq r \leq +1$

If **r = -1:** The two variables are said to be perfectly negatively correlated. This means that they move in opposite (different) directions.
If r = +1: The two variables are said to be perfectly positively correlated implying that they move in the same direction.

There are many measures of correlation, but the most commonly used is the **Pearsonia or product moment coefficient of correlation** as shown below:

$$r = \sqrt{r^2} \qquad \rightarrow r = \sqrt{0.565} = 0.752$$

Also, the coefficient of correlation can be computed as:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left\{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]\right\}}}$$

<u>NB:</u> The correlation coefficient between y (cost object) and x (cost driver) should be very high for the independent variable to influence variations in the dependent variable. As a rule of thumb correlation coefficient of 0.8 and above between the dependent variable(y) and the independent variable (x) is acceptable. However, the correlation coefficient between any two independent variables should be very low; otherwise the problem of multicollinearity could occur.

Alternatively **r** is given by:

$$r = \frac{\text{cov}(x, y)}{(\sigma x)(\sigma y)}$$

Or

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum(x - \bar{x})^2\right]\left[\sum(y - \bar{y})^2\right]}}$$

B) Standard error of the slope (Sb)

- The reliability of the regression coefficient is important because analyst focuses on the rate of variability rather than the absolute level of prediction.
- The **Sb** can be used to construct a confidence interval within which the True value of the unit variable cost lies.

$$S_b = \frac{S_e}{\sqrt{(\sum x^2) - \frac{(\sum x)^2}{n}}} \qquad = \frac{48.95}{\sqrt{16{,}234 - 430^2/12}} = 1.70$$

Interval for the true value of unit variable cost (B)

$$B = b \pm t_c . Sb$$

- Establishing a 95% confidence interval: **6.10-2.2281(1.7) ≤ B≤ 6.10+2.2281(1.7).**
- **[sh.2.312 ≤ B ≤ sh.9.89] → 95% confidence interval.**

C) Z or T statistics

These techniques are used to test the hypothesis that:

HO: b= 0. No significant relationship exists between the number of employees and off-loading cost**.**

HA: b≠ 0. A significant relationship exists between the number of employees and off- loading cost**. Or** Xi has a significant explanatory power in the model.

Test level of significance ($\propto$) at 5%

Use Z-statistics where the number of observations exceed thirty (**≥** 30) and 't'-statistics if the number of observations is less than thirty (**≤**30). Hence for Mombasa ltd we use 't' statistics.

Rule: If the computed value (**T)** is greater than the table value (t_c) then reject the null hypothesis (Ho)

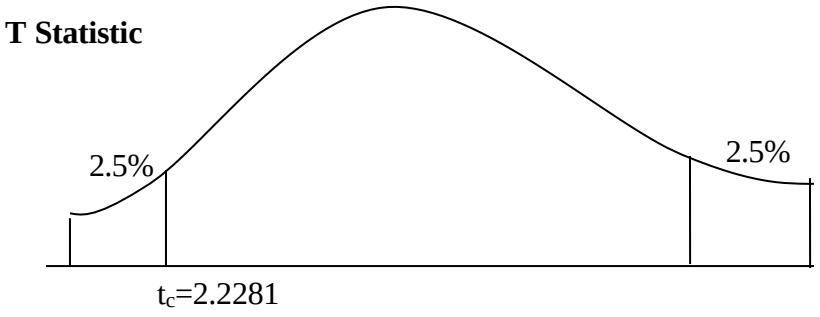
Computed value of t-statistics (T) is given by:

Computed $T = \frac{b}{S_b}$ $T = \frac{6.10}{1.70}$ **=3.5882**

Table value of t-statistics:

$t_{n-k-1,\propto} \rightarrow t_{12-1-1,\,0.025}$ **= 2.2281**

Area of significance (rejection area) is 2.5% on both sides i.e. see the diagram below:



Since the computed value (T=3.5882) is grater than the table value (t_c=2.2281); then reject the null hypothesis and conclude that a significant relationship exists between the number of employees (cost driver) and off-loading cost (dependent variable).

NB:

$$Z = \frac{X - \overline{X}}{\sigma}$$

Evaluating the Y intercept (a)

To evaluate the y intercept we compute Standard error of the Y-intercept (Sa). This is then used to construct a confidence interval.

Standard error of the Y-intercept (Sa) is given by::

$$Sa = Se.\sqrt{\frac{\sum x^2}{n\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]}}$$

Sa = $48.95 * \sqrt{\frac{16{,}234}{12\left[16{,}234 - 430^2/12\right]}}$ **= 62.6579.=62,657.9**

Determine the confidence interval at 95%

3.10 Strengths of regression analysis

1. It is more accurate as it gives an objective statistical estimate of the cost object by minimizing the sum squares of the deviations.
2. Provides greater mathematical precision as it considers all observations
3. Provides useful measures of determining and assessing the accuracy of the estimated cost function i.e. confidence intervals.

Limitations

1. Time consuming since it requires a lot of calculations.
2. It's a bit complex to apply especially to a layman.
3. If no care is taken in determining the data, it will greatly be influenced by outliers.

A dummy variable:

This is a **variable that is used to represent the presence or absence of a condition**. For example, a dummy variable can be used to **indicate seasonality** i.e. when estimating production cost and if production is always high in the month of March, a dummy variable with a value of **1** for March and **0** for the other months could be used (Blocher;2005).

3.11 MULTIPLE REGRESSION ANALYSIS

The least square regression equation discussed above (simple regression) was based on the assumption that total cost was determined by only one activity-based variable only (a single cost driver). However, other variables are likely to influence total cost. It means that the equation for the simple regression can be expanded to include more than one independent variable.

We can recall that the theoretical (general) linear model was given as:

$$y=a+b_1x_1+b_2x_2+------b_nx_n+e$$

Where Y = dependent variable (cost object)

a = non-variable cost item (fixed cost)

Xi =i^{th} predictor variable (cost driver) - (i = 1, 2 --- n).

bi=i^{th} regression coefficient or actual change in Y for each unit change in Xi.

e=error term (residual value or disturbance term).

- The Error (residual) term (e) lumps together all the variations in Y which are not explained (not attributable to) by the predictor variable**(s).**
- An estimated cost model has no error term (e).

Therefore, the estimated multiple regression will be given as follows:

$$\hat{Y} = a + b_1X_1 + b_2 X_2 + b_3 X_3 + \ldots\ldots\ldots + b_n X_n$$

For example labour costs can be affected by such drivers as labour hours, weight handled (material costs), machine hours, etc. These may have an effect on labour costs. The equation for estimating labour cost will be given as:

$$\hat{Y} = a + b_1X_1 + b_2 X_2 + b_3 X_3 + \ell$$

Where a, b_1 b_2, and b_3 are coefficients. ℓ = is a disturbance term that includes other factors that have an impact on labour (unexplained variations in labour cost normally measured by SSE) and X_1 = labour hours, X_2 =weight handled, X_3 = machine hours and Y = total cost (Lucey 2003)

<u>Two Independent Variables</u>

For two independent variables, the function will be of the form:

$$\hat{Y} = a + b_1X_1 + b_2 X_2$$

The normal equations for a regression equation with two independent variables are given below:

$$\sum y = an + b_1\sum x_1 + b_2\sum x_2$$

$$\sum x_1 y = a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1 x_2$$

$$\sum x_2 Y = a\sum x_2 + b_1\sum x_1 x_2 + b_2\sum x_2^2$$

(Drury 2004; Lucey 2003)

The manual calculation of multiple regression coefficients is tedious (laborious) but most computer systems have statistical packages which can be calculate the values of the individual coefficients, their standard errors, overall value of Y, confidence intervals for the regression line and so on. Tedious arithmetical calculations are covered in quantitative techniques. For computation purposes the student should have a better grasp of matrices which is extensively covered under quantitative techniques.

Example

A firm has discovered that its total overheads are dependent on labour hours, machine hours and unit produced. Analysis has produced the following multiple regression formula:

$\hat{Y}$ = £ 25,000 + 7.3X_1 + 4.8X_2 + 3.1 X_3

Where y = total overheads

X_1 =labour hours

X_2 = machine hours

X_3 = units produced

What are the predicted overheads in a period when there were 16,500 labour hours, 7,300 machine hours and 13,400 units produced?

Solution

$\hat{Y}$ = £ 25,000 + 7.3(16,500) + 4.8(7,300) + 3.1(13,400)

$\hat{Y}$ = £ 222,030 (Lucey 2003)

MULTI-COLLINEARITY

Multiple regression analysis is based on the assumption that the independent variables are not correlated with each other. When the independent variables are highly correlated with each other then it is very difficult and sometimes impossible to isolate or separate the effects of each of these variables on the dependent variable. This occurs when there is a simultaneous movement of two or more independent variables in the same direction and at approximately the same time. This condition is called **multicollinearity.**

We can use the **correlation matrix** to determine whether two independent variables are highly correlated. ***If a correlation value of more than 0.7 exists between two independent variables, then the problem of multi-collinearity is bound to occur.*** Alternatively if the correlation coefficient between the two variables is greater than the multiple correlation coefficients, then multi-collinearity problem will occur. To remove the problem of multi-collinearity, we drop one of the correlated variables. You can drop any of the variables. Therefore, there must be independence among the independent variables (cost drivers) (Drury 2004)

3.12 learning activities

QUESTION ONE: (Assignment- Required For Marking)

Given the following data

Maintenance cost	Y		33	21	40	38	46
Distance covered	Km		87	69	69	81	97
Load	Kg	5	11	4	9	7	10

Required:

a) Develop the estimating equation best describing these data
b) If distance covered was 83 and had a load of 7 kilograms, what maintenance would be expected
c) Evaluate the reliability of the estimated cost function

QUESTION TWO (Practice question)

The manager of Transporters ltd, a company that provides transportation services, is preparing a cost budget for the period starting from 30th June, 2015 to 31st July 2016. The following information is available from the records of transporters ltd.

Month	**Distance covered in KM**	**Transport costs**
July	345	Khs. 1,350

August	225	1,230
September	105	630
October	75	1,710
November	285	1,110
December	225	870
January	165	750
February	405	1,710
March	285	1,230
April	165	570
May	255	1,350
June	235	750
July	195	930

Required:

a) Estimate the cost function in the form of Y= a + bx under:

i) Regression analysis method. [7 **marks]**

ii) Evaluate the reliability of the estimated regression model in I above (Hint step six) **[18 marks]**

QUESTION THREE (Practice question)

The managing director of Protection security services ltd, a company that provides security services, is preparing a cost budget for the period starting from 30th June, 2014 to 31st July 2015. The following information is available from the records of Protect security ltd.

Month	**Security costs**	**Number of security guards**
	Ksh 000s	000s
July	400	190
August	360	110

September	160	30
October	520	20
November	320	150
December	240	110
January	200	70
February	560	230
March	360	150
April	140	70
May	400	130
June	200	50
July	260	90

Required:

a) Estimate the cost function Y =a + bx under the least square Method. **[6marks]**

b) Calculate the coefficient of determination (r^2). Comment on the answer **[3 marks]**

c) Compute the standard error of estimate (se) and establish a 98% confidence interval if 80,000 guards will be used in the month of August 2013. Assume the table value of t_c is 2.718 **[4 marks]**

d) Conduct an F test at 98% assuming the table value of $\mathbf{f}_c$ is 9.65. Clearly indicate the hypothesis and comment on the answer **[2Marks]**

e) Determine a 98% confidence interval for the true value of the slope **[4 marks]**

f) Explain how you can evaluate the reliability of an estimated cost model **[6marks]**