OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS

Correlation and Regression Analysis

TEXTBOOK

ORGANISATION OF ISLAMIC COOPERATION

STATISTICAL ECONOMIC AND SOCIAL RESEARCH
AND TRAINING CENTRE FOR ISLAMIC COUNTRIES

OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS

Correlation and Regression Analysis

TEXTBOOK

{{Dr. Mohamed Ahmed Zaid}}

ORGANISATION OF ISLAMIC COOPERATION

STATISTICAL ECONOMIC AND SOCIAL RESEARCH
AND TRAINING CENTRE FOR ISLAMIC COUNTRIES

Kudüs Cad. No: 9, Diplomatik Site, 06450 Oran, Ankara – Turkey

Telephone +90 – 312 – 468 6172

Internet www.sesric.org

E-mail statistics@sesric.org

ISBN: xxx-xxx-xxxx-xx-x

Cover design by Publication Department, SESRIC.

For additional information, contact Statistics Department, SESRIC.

CONTENTS

ACRONYMS

r	Pearson Coefficient of Correlation
τ	Kendall's Tau Coefficient of Correlation
ρ	Spearman Coefficient of Correlation
R^2	Coefficient of Determination
α	Significance Level
P_value	Calculated Significance value (probability value)

SPSS	Statistical Package for Social Science OR Statistical Product for Solutions Services
CAPMAS	Central Agency of Public Mobilization and Statistics (Statistic office of Egypt)

ACKNOWLEDGEMENT

Prepared jointly by the Central Agency of Public Mobilization and Statistics (CAPMAS) in Cairo, Egypt and the Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC) under the OIC Accreditation and Certification Programme for Official Statisticians (OIC-CPOS) supported by Islamic Development Bank Group (IDB), this textbook on Correlation and Regression Analysis covers a variety topics of how to investigate the strength , direction and effect of a relationship between variables by collecting measurements and using appropriate statistical analysis. Also this textbook intends to practice data of labor force survey year 2015, second quarter (April, May, June), in Egypt by identifying how to apply correlation and regression statistical data analysis techniques to investigate the variables affecting phenomenon of employment and unemployment.

UNIT 1
INTRODUCTION

1.1. Preface

The goal of statistical data analysis is to understand a complex, real-world phenomenon from partial and uncertain observations. It is important to make the distinction between the mathematical theory underlying statistical data analysis, and the decisions made after conducting an analysis. Where there is a subjective part in the way statistical analysis yields actual human decisions. Understanding the risk and the uncertainty behind statistical results is critical in the decision-making process.

In this textbook, we will study the relation and association between phenomena through the correlation and regression statistical data analysis, covering in particular how to make appropriate decisions throughout applying statistical data analysis.

In regards to technical cooperation and capacity building, this textbook intends to practice data of labor force survey year 2015, second quarter (April, May, June), in Egypt by identifying how to apply correlation and regression statistical data analysis techniques to investigate the variables affecting phenomenon of employment and unemployment.

There are many terms that need introduction before we get started with the recipes. These notions allow us to classify statistical techniques within multiple axes.

Prediction consists of learning from data, and predicting the outcomes of a random process based on a limited number of observations, the term "**predictor**" can be misleading if it is interpreted as the ability to predict even beyond the limits of the data. Also, the term "**explanatory variable**" might give an impression of a causal effect in a situation in which inferences should be limited to identifying associations. The terms "**independent**" and "**dependent**" variable are less subject to these interpretations as they do not strongly imply cause and effect

Observations are independent realizations of the same random process; each observation is made of one or several variables. Mainly variables are either numbers, or elements belonging to a finite set "finite number of values". The first step in an analysis is to understand what your observations and variables are.

Study is **univariate** if you have one variable. It is **Bivariate** if there are two variables and **multivariate** if at least two variables. Univariate methods are typically simpler. That being said, univariate methods may be used on multivariate data, using one dimension at a time. Although interactions between variables cannot be explored in that case, it is often an interesting first approach.

1.2. What Are correlation and regression

Correlation quantifies the degree and direction to which two variables are related. Correlation does not fit a line through the data points. But simply is computing a correlation coefficient that tells how much one variable tends to change when the other one does. When r is 0.0, there is no relationship. When r is positive, there is a trend that one variable goes up as the other one goes up. When r is negative, there is a trend that one variable goes up as the other one goes down.

With correlation, it doesn't have to think about cause and effect. It doesn't matter which of the two variables is call dependent and which is call independent, if the two variables swapped the degree of correlation coefficient will be the same.

The sign (+, -) of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association, e.g. A correlation of r = - 0.8 suggests a strong, negative association (reverse trend) between two variables, whereas a correlation of r = 0.4 suggest a weak, positive association. A correlation close to zero suggests no linear association between two continuous variables.

Linear regression finds the best line that predicts dependent variable from independent variable. The decision of which variable calls dependent and which calls independent is an important matter in regression, as it'll get a different best-fit line if you swap the two. The line that best predicts independent variable from dependent variable is not the same as the line that predicts dependent variable from independent variable in spite of both those lines have the same value for R^2. Linear regression quantifies goodness of fit with R^2, if the same data put into correlation matrix the square of r degree from correlation will equal R^2 degree from regression. The sign (+, -) of the regression coefficient indicates the direction of the effect of independent variable(s) into dependent variable, where the degree of the regression coefficient indicates the effect of the each independent variable into dependent variable.

1.3. Assumptions of parametric and non parametric Statistics

Parametric statistics are the most common type of inferential statistics, which are calculated with the purpose of generalizing the findings of a sample to the population it represents. Parametric tests make assumptions about the parameters of a population, whereas nonparametric tests do not include such assumptions or include fewer. For instance, parametric tests assume that the sample has been randomly selected from the population it represents and that the distribution of data in the population has a known underlying distribution. The most common distribution assumption is that the distribution is normal. Other distributions include the binomial distribution (logistic regression) and the Poisson distribution (Poisson regression), and non-parametric tests are sometimes called "distribution-free" tests. Additionally, parametric statistics require that the data are measured using an interval or ratio scale, whereas nonparametric statistics use data that are measured with a nominal or ordinal scale. There are three types of commonly used nonparametric correlation coefficients (Spearman R, Kendall Tau, and Gamma coefficients), where parametric correlation coefficients (Pearson)

It’s commonly thought that the need to choose between a parametric and nonparametric test occurs when your data fail to meet an assumption of the parametric test. This can be the case when you have both a small sample size and non normal data. The decision often depends on whether the mean or median more accurately represents the center of your data’s distribution.

- If the mean accurately represents the center of your distribution and your sample size is large enough, consider a parametric test because they are more powerful.
- If the median better represents the center of your distribution, consider the nonparametric test even when you have a large sample.

In general, parametric methods make more assumptions than non-parametric methods. If those extra assumptions are correct, parametric methods can produce more accurate and precise estimates. They are said to have more statistical power. However, if assumptions are incorrect, parametric methods can be very misleading for that reason they are often not considered robust. On the other hand, parametric formulae are often simpler to write down and faster to compute. In some cases, but not all, their simplicity makes up for their non-robustness, especially if care is taken to examine diagnostic statistics.

1.4. Test of Significance level

In linguistic, "significant" means important, while in Statistics "significant" means probably true (not due to chance). A research finding may be true without being important. When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.

Significance levels show you how likely a pattern in your data is due to chance. The most common level, used to mean something is good enough to be believed, is "0.95". This means that the finding has a 95% chance of being true which also means that the finding has a confidence degree 95% of being true. No statistical package will show you "95%" or ".95" to indicate this level. Instead it will show you ".05," meaning that the finding has a five percent (.05) chance of not being true "error", which is the converse of a 95% chance of being true. To find the significance level, subtract the number shown from one. For example, a value of ".01" means that there is a confidence degree 99% (1-.01=.99) chance of it being true.

In other words the significance level α "alpha level" for a given hypothesis test is a value for which a P-value "calculated value" less than or equal to α is considered statistically significant. Typical value levels for α are 0.1, 0.05, and 0.01. These value levels correspond to the probability of observing such an extreme value by chance. For example, if the P-value is 0.0082, so the probability of observing such a value by chance is less that 0.01, and the result is significant at the 0.01 level.

UNIT 2
Correlation Analysis

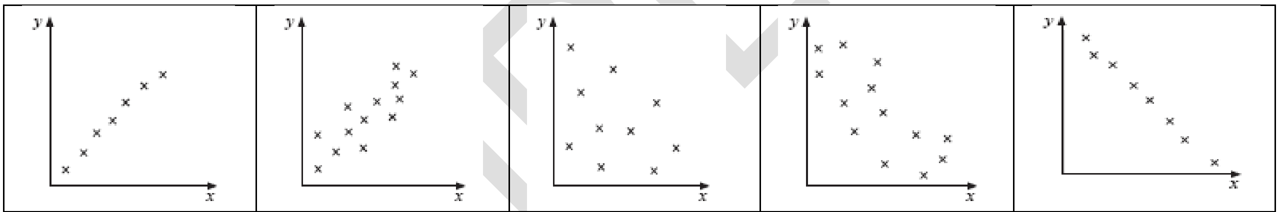
2.1. Definition

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other. However, correlation does not imply causation. There may be an unknown factor that influences both variables similarly.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

- Correlation is **Positive** or direct when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases, and so called inverse or contrary correlation.



a	b	c	d	e

If the points plotted were all on a straight line we would have perfect correlation, but it could be positive or negative as shown in the diagrams above,

a. Strong positive correlation between x and y. The points lie close to a straight line with y increasing as x increases.
b. Weak, positive correlation between x and y. The trend shown is that y increases as x increases but the points are not close to a straight line
c. No correlation between x and y; the points are distributed randomly on the graph.
d. Weak, negative correlation between x and y. The trend shown is that y decreases as x increases but the points do not lie close to a straight line
e. Strong, negative correlation. The points lie close to a straight line, with y decreasing as x increases

Correlation can have a value:

3. 1 is a perfect positive correlation
4. 0 is no correlation (the values don't seem linked at all)
5. -1 is a perfect negative correlation

The value shows how good the correlation is (not how steep the line is), and if it is positive or negative. Usually, in statistics, there are three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

2.2. Assumption of Correlation

Employing of correlation rely on some underlying assumptions. The variables are assumed to be independent, assume that they have been randomly selected from the population; the two variables are normal distribution; association of data is homoscedastic (homogeneous), homoscedastic data have the same standard deviation in different groups where data are heteroscedastic have different standard deviations in different groups and assumes that the relationship between the two variables is linear. The correlation coefficient is not satisfactory and difficult to interpret the associations between the variables in case if data have outliers.

An inspection of a scatterplot can give an impression of whether two variables are related and the direction of their relationship. But it alone is not sufficient to determine whether there is an association between two variables. The relationship depicted in the scatterplot needs to be described qualitatively. Descriptive statistics that express the degree of relation between two variables are called correlation coefficients. A commonly employed correlation coefficient are Pearson correlation, Kendall rank correlation and Spearman correlation.

Correlation used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied. The results of the analysis, however, need to be interpreted with care, particularly when looking for a causal relationship.

2.3. Bivariate Correlation

Bivariate correlation is a measure of the relationship between the two variables; it measures the strength and direction of their relationship, the strength can range from absolute value 1 to 0. The stronger the relationship, the closer the value is to 1. Direction of The relationship can be positive (direct) or negative (inverse or contrary); correlation generally describes the effect that two or more phenomena occur together and therefore they are linked For example, the positive relationship of .71 can represent positive correlation between the statistics degrees and the science degrees. The student who has high degree in statistics has also high degree in science and vice versa.

The Pearson correlation coefficient is given by the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where $\bar{x}$ is the mean of variable x values, and $\bar{y}$ is the mean of variable y values.

Example – Correlation of statistics and science tests

A study is conducted involving 10 students to investigate the association between statistics and science tests. The question arises here; is there a relationship between the degrees gained by the 10 students in statistics and science tests?

Table (2.1) Student degree in Statistic and science

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Notes: the marks out of 30

Suppose that (x) denotes for statistics degrees and (y) for science degree

Calculating the mean $(\bar{x}\ ,\ \bar{y})$;

$$\bar{x} = \frac{\sum x}{n} = \frac{173}{10} = 17.3\ ,\ \ \bar{y} = \frac{\sum y}{n} = \frac{200}{10} = 20$$

Where the mean of statistics degrees $\bar{x}$ = 17.3 and the mean of science degrees $\bar{y}$ = 20

Table (2.2) Calculating the equation parameters

Statistics	Science					
x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
20	20	2.7	7.29	0	0	0
23	25	5.7	32.49	5	25	28
8	11	-9.3	86.49	-9	81	83
29	24	11.7	136.89	4	16	46
14	23	-3.3	10.89	3	9	-9.9
12	16	-5.3	28.09	-4	16	21.2
11	12	-6.3	39.69	-8	64	50.4
21	21	3.7	13.69	1	1	3.7
17	22	-0.3	0.09	2	4	-0.6
18	26	0.7	0.49	6	36	4.2
173	200	0	356.1	0	252	228

$$\sum(x-\bar{x})^2 = 356.1\ ,\ \sum(y-\bar{y})^2 = 252\ ,$$
$$\sum(x-\bar{x})(y-\bar{y}) = 228$$

Calculating the Pearson correlation coefficient;

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{228}{\sqrt{356.1}\sqrt{252}}$$
$$= \frac{228}{(18.8706)(15.8745)} = \frac{228}{299.5614} = 0.761$$

Other solution

Also; the Pearson correlation coefficient is given by the following equation:

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

Table (2.3) Calculating the equation parameters

x	y	xy	x^2	y^2
20	20	400	400	400
23	25	575	529	625
8	11	88	64	121
29	24	696	841	576
14	23	322	196	529
12	16	192	144	256
11	12	132	121	144
21	21	441	441	441
17	22	374	289	484
18	26	468	324	676
173	200	3688	3349	4252

Required calculation
$\sum x = 173$, $\sum y = 200$ $\sum xy = 3688$ $\sum x^2 = 3349$ $\sum y^2 = 4252$

Calculating the Pearson correlation coefficient by substitute in the aforementioned equation;

$$\boldsymbol{r} = \frac{\mathbf{3688} - \frac{(\mathbf{173})(\mathbf{200})}{\mathbf{10}}}{\sqrt{\left(\mathbf{3349} - \frac{(\mathbf{173})^2}{\boldsymbol{n}\mathbf{10}}\right)\left(\mathbf{4252} - \frac{(\mathbf{200})^2}{\mathbf{10}\boldsymbol{n}}\right)}} = \frac{\mathbf{228}}{\sqrt{(\mathbf{356.1})(\mathbf{252})}} = \frac{\mathbf{228}}{\mathbf{299.5614}} = \mathbf{0.761}$$

Pearson Correlation coefficient r **=** 0.761 exactly the same output of the first equation.

The calculation shows a strong positive correlation (0.761) between the student's statistics and science degrees. This means that as degrees of statistics increases the degrees of science increase also. Generally the student who has a high degree in statistics has high degree in science and vice versa.

2.4. Partial Correlation

The Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables. Correlations are measures of linear association. Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association**.**

Partial correlation is the correlation between two variables after removing the effect of one or more additional variables. Suppose we want to find the correlation between y and x controlling by W. This is called the partial correlation and its symbol is $r_{YX.W}$. This command is specifically for the case of one additional variable. In this case, the partial correlation can be computed based on standard correlations between the three variables as follows:

$$r_{YX.W} = \frac{r_{XY} - r_{XW} r_{YW}}{\sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)}}$$

$r_{YX.W}$ Stands for the correlation between x and y controlling by W.

As with the standard correlation coefficient, a value of +1 indicates a perfect positive linear relationship, a value of -1 indicates a perfect negative linear relationship, and a value of 0 indicates no linear relationship. *For more information see unit 4 of this book.*

2.5. Correlation Coefficients Pearson, Kendall and Spearman

Correlation is a Bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. Usually, in statistics, we measure three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

Pearson r correlation: Pearson correlation is widely used in statistics to measure the degree of the relationship between linear related variables. For example, in the stock market, if we want to measure how two commodities are related to each other, Pearson correlation is used to measure the degree of relationship between the two commodities. The following formula is used to calculate the Pearson correlation coefficient r: See Example

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Kendall's Tau rank correlation: Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables. If we consider two samples, x and y, where each sample size is n, we know that the total number of pairings with x y is n (n-1)/2.

The following formula is used to calculate the value of Kendall rank correlation:

$$\tau = \frac{n_c \; - \; n_d}{\frac{1}{2}\, n(n-1)}$$

Where:

τ = Kendall rank correlation coefficient

n_c = number of concordant (Ordered in the same way).

n_d = Number of discordant (Ordered differently).

Kendall’s Tau Basic Concepts

Definition 1: Let $x_1, \ldots, x_n$ be a sample for random variable x and let $y_1, \ldots, y_n$ be a sample for random variable y of the same size n. There are $C(n, 2)$ possible ways of selecting distinct pairs (x_i, y_i) and (x_j, y_j). For any such assignment of pairs, define each pair as concordant, discordant or neither as follows:

- Concordant © if ($x_i > x_j$ and $y_i > y_j$) or ($x_i < x_j$ and $y_i < y_j$)
- Discordant (D) if ($x_i > x_j$ and $y_i < y_j$) or ($x_i < x_j$ and $y_i > y_j$)
- Neither if $x_i = x_j$ or $y_i = y_j$ (i.e. ties are not counted).

Observation: To facilitate the calculation of $C – D$ it is best to first put all the x data elements in ascending order. If x and y are perfectly positively correlated, then all the values of y would be in ascending order too, and so if there are no ties then $C = C\,(n, 2)$ and $\tau = 1$.

Otherwise, there will be some inversions. For each i, count the number of $j > i$ for which $x_j < x_i$. This sum is D. If x and y are perfectly negatively correlated, then all the values of y would be in descending order, and so if there are no ties then $D = C\,(n, 2)$ and $\tau = -1$.

An example of calculating Kendall's Tau correlation

To calculate a ***Kendall's Tau correlation*** coefficient on same data without any ties we use the following data:

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Table (2.4) Set rank to the data

data				Arranged Rank	
statistics (degree)	science (degree)	Rank (statistics)	Rank (science)	Rank (science)	Rank (statistics)
20	20	4	7	1	5
23	25	2	2	2	2
8	11	10	10	3	1
29	24	1	3	4	7
14	23	7	4	5	6
12	16	8	8	6	3
11	12	9	9	7	4
21	21	3	6	8	8
17	22	6	5	9	9
18	26	5	1	10	10

Continued Table (2.4) Calculating the Number of Concordant C and Discordant (D)

D	C											
		1	--									
1		2	D	--								
2		3	D	D	--							
	3	4	C	C	C	--						
1	3	5	D	C	C	C	--					
3	2	6	D	D	C	C	D	--				
3	3	7	C	D	D	C	C	D	--			
	7	8	C	C	C	C	C	C	C	--		
	8	9	C	C	C	C	C	C	C	C	--	
	9	10	C	C	C	C	C	C	C	C	C	--
			1	2	3	4	5	6	7	8	9	10
10	35	Total of (D) and (C)										

Then substitute into the main equation

$$\tau = \frac{n_c \; - \; n_d}{\frac{1}{2}\, n(n-1)} \qquad\qquad \tau = \frac{35 - \; 10}{\frac{1}{2} * 10(10-1)}$$

$$\tau = \frac{25}{45} = 0.556$$

Kendall's Tau coefficient τ = 0.556; this indicates a moderate positive relationship between the ranks individuals obtained in the statistics and science exam. This means the higher you ranked in statistics, the higher you ranked in science also, and vice versa.

Calculating Kendall's Tau manually can be very tedious without a computer and is rarely done without a computer. Large dataset make it almost impossible to do manually by hand. . For more information see unit4 in this book

Spearman rank correlation: Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

The following formula is used to calculate the Spearman rank correlation coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where:

ρ = Spearman rank correlation coefficient
di= the difference between the ranks of corresponding values Xi and Yi
n= number of value in each data set.

The Spearman correlation coefficient, ρ, can take values from +1 to -1. A ρ of +1 indicates a perfect association of ranks, a ρ of zero indicates no association between ranks and a ρ of -1 indicates a perfect negative association of ranks. The closer ρ to zero, the weaker the association between the ranks.

An example of calculating Spearman's correlation

To calculate a Spearman rank-order correlation coefficient on data without any ties use the following data:

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Table (2.5) Calculating the Parameters of Spearman rank Equation:

statistics (degree)	science (degree)	Rank (statistics)	Rank (science)	| d |	d^2
20	20	4	7	3	9
23	25	2	2	0	0
8	11	10	10	0	0
29	24	1	3	2	4
14	23	7	4	3	9
12	16	8	8	0	0
11	12	9	9	0	0
21	21	3	6	3	9
17	22	6	5	1	1
18	26	5	1	4	16

Where d = absolute difference between ranks and d^2 = difference squared.
Then calculate the following:

$$\sum d_i^2 = 9+0+0+4+9+0+0+9+1+16 = 48$$

Then substitute into the main equation as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad ; \qquad \rho = 1 - \frac{6*48}{10(10^2-1)}$$

$$\rho = 1 - \frac{288}{990} \qquad ; \qquad \rho = 1 - 0.2909$$

$$\rho = 0.71$$

Hence, we have a ρ= 0.71; this indicates a strong positive relationship between the ranks individuals obtained in the statistics and science exam. This means the higher you ranked in statistics, the higher you ranked in science also, and vice versa.

So; the **Pearson r correlation** coefficient = 0.761 and ***Spearman's correlation*** = 0.71 for the same data which means that correlation coefficients for both techniques are approximately equal. For more information see unit4 in this book

2.6 Exercises

Study is conducted involving 14 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Table (2.6) Gestational age and their Weight at birth

Infant No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gestational age	34.7	36	29.3	40.1	35.7	42.4	40.3	37.3	40.9	38.3	38.5	41.4	39.7	39.7
Birth Weight	1895	2030	1440	2835	3090	3827	3260	2690	3285	2920	3430	3657	3685	3345

Applying the proper method; Estimate the association between gestational age and infant birth weight.

(Guide values $r = 0.882$, $\rho = 0.779$, $\tau = 0.641$)

UNIT 3
Regression Analysis

3.1. Definition

Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences which involves identifying and evaluating the relationship between a dependent variable and one or more independent variables, which are also called predictor or explanatory variables. It is particularly useful for assess and adjusting for confounding. Model of the relationship is hypothesized and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

Linear regression explores relationships that can be readily described by straight lines or their generalization to many dimensions. A surprisingly large number of problems can be solved by linear regression, and even more by means of transformation of the original variables that result in linear relationships among the transformed variables.

When there is a single continuous dependent variable and a single independent variable, the analysis is called a **simple linear regression analysis**. This analysis assumes that there is a linear association between the two variables. **Multiple regression** is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

Independent variables are characteristics that can be measured directly; these variables are also called predictor or explanatory variables used to predict or to explain the behavior of the dependent variable.

Dependent variable is a characteristic whose value depends on the values of independent variables.

Reliability and Validity:

- Does the model make intuitive sense? Is the model easy to understand and interpret?
- Are all coefficients statistically significant? (p-values less than .05)
- Are the signs associated with the coefficients as expected?
- Does the model predict values that are reasonably close to the actual values?
- Is the model sufficiently sound? (High R-square, low standard error, etc.)

3.2. Objectives of Regression Analysis

Regression analysis used to explain variability in dependent variable by means of one or more of independent or control variables and to analyze relationships among variables to answer; the question of how much dependent variable changes with changes in each of the independent's variables, and to forecast or predict the value of dependent variable based on the values of the independent's variables.

The primary objective of regression is to develop a linear relationship between a response variable and explanatory variables for the purposes of prediction, assumes that a functional linear relationship exists, and alternative approaches (functional regression) are superior.

3.3. Assumption of Regression Analysis

The regression model is based on the following assumptions.

- The relationship between independent variable and dependent is linear.
- The expected value of the error term is zero
- The variance of the error term is constant for all the values of the independent variable, the assumption of homoscedasticity.
- There is no autocorrelation.
- The independent variable is uncorrelated with the error term.
- The error term is normally distributed.
- On an average difference between the observed value (yi) and the predicted value (ˆyi) is zero.
- On an average the estimated values of errors and values of independent variables are not related to each other.
- The squared differences between the observed value and the predicted value are similar.
- There is some variation in independent variable. If there are more than one variable in the equation, then two variables should not be perfectly correlated.

Intercept or Constant

- Intercept is the point at which the regression intercepts y-axis.
- Intercept provides a measure about the mean of dependent variable when slope(s) are zero.
- If slope(s) are not zero then intercept is equal to the mean of dependent variable minus slope $\times$ mean of independent variable.

Slope

- Change is dependent variable as we change independent variable.
- Zero Slope means that independent variable does not have any influence on dependent variable.
- For a linear model, slope is not equal to elasticity. That is because; elasticity is percent change in dependent variable, as a result one percent change in independent variable.

3.4. Simple Regression Model

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable $\mathbf{y}$, based on the value of an independent variable $\boldsymbol{x}$.

- One variable, denoted ($\boldsymbol{x}$), is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted ($\boldsymbol{y}$), is regarded as the **response**, **outcome**, or **dependent** variable.

Mathematically, the regression model is represented by the following equation:

$$\mathbf{y} = \beta_0 \;\pm\; \beta_1\, \boldsymbol{x}_1 \;\pm\; \varepsilon_1$$

Where

- $\boldsymbol{x}$ independent variable.
- $\mathbf{y}$ dependent variable.
- β_1 The Slope of the regression line
- $\boldsymbol{\beta_0}$ The intercept point of the regression line and the y axis.
- $\boldsymbol{n}$ Number of cases or individuals.
- $\sum \boldsymbol{x}\mathbf{y}$ Sum of the product of dependent and independent variables.
- $\sum \boldsymbol{x}$ = Sum of independent variable.
- $\sum \mathbf{y}$ = Sum of dependent variable.
- $\sum \boldsymbol{x^2}$ = Sum of square independent variable.

$$\boldsymbol{\beta_1} = \frac{\boldsymbol{n} \sum \boldsymbol{x}\mathbf{y} - \sum \boldsymbol{x} \sum \mathbf{y}}{\boldsymbol{n} \sum \boldsymbol{x^2} - (\sum \boldsymbol{x})^2}$$

$$\boldsymbol{\beta_0} = \bar{\mathbf{y}} - \boldsymbol{\beta_1}\bar{\boldsymbol{x}}$$

Example – linear Regression of patient's age and their blood pressure

A study is conducted involving 10 patients to investigate the relationship and effects of patient's age and their blood pressure.

Table (3.1) calculating the linear regression of patient's age and blood pressure

Obs	Age x	BP y	xy	x^2
1	35	112	3920	1225
2	40	128	5120	1600
3	38	130	4940	1444
4	44	138	6072	1936
5	67	158	10586	4489
6	64	162	10368	4096
7	59	140	8260	3481
8	69	175	12075	4761
9	25	125	3125	625
10	50	142	7100	2500
Total	491	1410	71566	26157

Required calculation
$\sum x = 491$
$\sum y = 1410$
$\sum xy = 71566$
$\sum x^2 = 26157$

Calculating the mean $(\bar{x}\ ,\ \bar{y})$ ***;***

$$\bar{x} = \frac{\sum x}{n} = \frac{491}{10} = 49.1\ ,\ \bar{y} = \frac{\sum y}{n} = \frac{1410}{10} = 141$$

Calculating the regression coefficient;

$$\beta_1 = \frac{n\sum x\mathrm{y} - \sum x \sum \mathrm{y}}{n\sum x^2 - (\sum x)^2} \qquad \beta_1 = \frac{10 * 71566 - 491 * 1410}{10 * 26157 - (491)^2}$$

$$\beta_1 = \frac{715660 - 692310}{261570 - 241081} \qquad \beta_1 = \frac{23350}{20489} = 1.140$$

$$\beta_0 = \bar{\mathrm{y}} - \beta_1 \bar{x} \qquad \beta_0 = 141 - 1.140 * 49.1$$

$$\beta_0 = 141 - 55.974 \qquad \beta_0 = 85.026$$

Then substitute the regression coefficient into the regression model

$$\boldsymbol{Estimated\ blood\ pressure}\ (\hat{\mathrm{Y}}) = 85.026 + 1.140\ age$$

Interpretation of the equation;

Constant (intercept) value β_0 = 85.026 indicates that blood pressure at age zero.

Regression coefficient $\beta_1 =$ 1.140 indicates that as age increase by one year the blood pressure increase by 1.140

Table (3.2) Applying the value of age to the regression Model to calculate the estimated blood pressure ($\hat{Y}$) coefficient of determination (R^2) as follows:

Obs	Age	BP							
	x	y	$\hat{Y}$	$\hat{Y} - \bar{y}$	$(\hat{Y} - \bar{y})^2$	$(\gamma - \hat{Y})$	$(\gamma - \hat{Y})^2$	$(\gamma - \bar{\gamma})$	$(\gamma - \bar{\gamma})^2$
1	35	112	124.926	-16.074	258.373	-12.926	167.081	-29	841
2	40	128	130.626	-10.374	107.620	-2.626	6.896	-13	169
3	38	130	128.346	-12.654	160.124	1.654	2.736	-11	121
4	44	138	135.186	-5.814	33.803	2.814	7.919	-3	9
5	67	158	161.406	20.406	416.405	-3.406	11.601	17	289
6	64	162	157.986	16.986	288.524	4.014	16.112	21	441
7	59	140	152.286	11.286	127.374	-12.286	150.946	-1	1
8	69	175	163.686	22.686	514.655	11.314	128.007	34	1156
9	25	125	113.526	-27.474	754.821	11.474	131.653	-16	256
10	50	142	142.026	1.026	1.053	-0.026	0.001	1	1
Total	491	1410	1410	0.000	2662.750	0.000	622.950	0	3284

Table (3.3) Equation of ANOVA table for simple linear regression;

Source of Variation	*Sums of Squares*	*Df*	*Mean Square*	*F*
Regression	$\Sigma(\hat{Y} - \bar{Y})^2$	*1*	$SS_{reg} / 1$	MS_{reg} / MS_{res}
Residual	$\Sigma(Y - \hat{Y})^2$	$N - 2$	$SS_{res} / (N - 2)$	
Total	$\Sigma(Y - \bar{Y})^2$	$N - 1$		

Continued Table (3.3) Calculating the ANOVA table values for simple linear regression;

Source of Variation	*Sum of Squares*	*Df*	*Mean Square*	*F*
Regression	*2662.75*	*1*	*2662.75 / 1 =2662.75*	*2662.75 / 77.86875= 34.195*
Residual	*622.95*	*8*	*622.95 / 8 =77.86875*	
Total	*3284*	*9*		

Calculating the coefficient of determination (R^2)

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{Regression\ Sum\ of\ Square\ (SSR)}{Total\ Sum\ of\ Square\ (SST)}$$

Then substitute the values from ANOVA table

$$R^2 = \frac{2662.75}{3284} = 0.810$$

We can say that 81% of the variation in the blood pressure rate is explained by age.

3.5. Multiple Regressions Model

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a dependent variable (target or criterion variable) based on the value of two or more independent variables (predictor or explanatory variables). Multiple regression allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time and lecture attendance "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

Mathematically, the multiple regression model is represented by the following equation:

$$Y = \beta_0 \pm \beta_i X_i \dots\dots\dots\dots \pm \beta_n X_n \pm u$$

Where:

- X_i *to* X_n Represent independent variables.
- **Y** Dependent variable.
- β_1 The regression coefficient of variable x_1
- β_2 The regression coefficient of variable x_2
- β_0 The intercept point of the regression line and the y axis.

By using method of deviation

- $\bar{y}$ The mean of dependent variable values.
- $\overline{X_1}$ The mean of X_1 independent variable values.
- $\sum y = \sum(Y - \bar{Y})$
- $\sum x_1 = \sum(X_1 - \overline{X_1})$

- $\overline{X_2}$ The mean of X_2 independent variable values.
- $\sum x_1 y = \sum(x_1 * y)$
- $\sum x_2 y = \sum(x_2 * y)$
- $\sum x_1 x_2 = \sum(x_1 * x_2)$
- $\left(\sum x_2^2\right) =$ ***Sum of square of*** x_2
- $\sum x_2 = \sum(X_2 - \overline{X_2})$
- $\left(\sum x_1^2\right) =$ ***Sum of square of*** x_1
- $\left(\sum x_2^2\right) =$ ***Sum of square of*** x_2
- $\left(\sum x_1^2\right) =$ ***Sum of square of*** x_1

$$\beta_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2\, y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \qquad \beta_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1\, y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2}$$

Example – Multiple Regression of students exam performance, revision time and lecture attendance

A study is conducted involving 10 students to investigate the relationship and affects of revision time and lecture attendance on exam performance.

Table (3.4) Students exam performance, revision time and lecture attendance

Obs	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Y	**40**	**44**	**46**	**48**	**52**	**58**	**60**	**68**	**74**	**80**
X_1	**6**	**10**	**12**	**14**	**16**	**18**	**22**	**24**	**26**	**32**
X_2	**4**	**4**	**5**	**7**	**9**	**12**	**14**	**20**	**21**	**24**

Stands for

(Y) Exam performance

(X_1) Revision time

(X_2) Lecture attendance.

Table (3.5) Calculating the coefficient of regression

				y	x_1	x_2	x_1* **y**	x_2* **y**	x_1* x_2
Obs	**Y**	X_1	X_2	**Y-$\overline{Y}$**	X_1-$\overline{X_1}$	X_2-$\overline{X_2}$			
1	**40**	**6**	**4**	**-17**	**-12**	**-8**	**204**	**136**	**96**
2	**44**	**10**	**4**	**-13**	**-8**	**-8**	**104**	**104**	**64**
3	**46**	**12**	**5**	**-11**	**-6**	**-7**	**66**	**77**	**42**
4	**48**	**14**	**7**	**-9**	**-4**	**-5**	**36**	**45**	**20**
5	**52**	**16**	**9**	**-5**	**-2**	**-3**	**10**	**15**	**6**
6	**58**	**18**	**12**	**1**	**0**	**0**	**0**	**0**	**0**
7	**60**	**22**	**14**	**3**	**4**	**2**	**12**	**6**	**8**
8	**68**	**24**	**20**	**11**	**6**	**8**	**66**	**88**	**48**
9	**74**	**26**	**21**	**17**	**8**	**9**	**136**	**153**	**72**
10	**80**	**32**	**24**	**23**	**14**	**12**	**322**	**276**	**168**
	570	**180**	**120**				**956**	**900**	**524**

Continued Table (3.5) Calculating the coefficient of regression

	x_1^2	x_2^2	$\widehat{Y}$	$\widehat{y}^2$	u	u^2	y^2
Obs				$(\widehat{Y}-\overline{Y})^2$	$(y-\widehat{y})$	$(y-\widehat{y})^2$	$(Y-\overline{Y})^2$
1	**144**	**64**	**40.32**	**278.22**	**-0.320**	**0.1024**	**289**
2	**64**	**64**	**42.92**	**198.25**	**1.080**	**1.1664**	**169**
3	**36**	**49**	**45.33**	**136.19**	**0.670**	**0.4489**	**121**
4	**16**	**25**	**48.85**	**66.42**	**-0.850**	**0.7225**	**81**
5	**4**	**9**	**52.37**	**21.44**	**-0.370**	**0.1369**	**25**
6	**0**	**0**	**57**	**0.00**	**1.000**	**1**	**1**
7	**16**	**4**	**61.82**	**23.23**	**-1.820**	**3.3124**	**9**
8	**36**	**64**	**69.78**	**163.33**	**-1.780**	**3.1684**	**121**
9	**64**	**81**	**72.19**	**230.74**	**1.810**	**3.2761**	**289**
10	**196**	**144**	**79.42**	**502.66**	**0.580**	**0.3364**	**529**
	576	**504**		**1620.47**	**0.000**	**13.6704**	**1634**

$$\overline{X_1} = \frac{\sum X_1}{n} \;;\; \overline{X_2} = \frac{\sum X_2}{n} \;;\; \overline{Y} = \frac{\sum Y}{n}$$

$$\overline{X_1} = \frac{\sum 180}{10} = 18 \;;\; \overline{X_2} = \frac{\sum 120}{10} = 12 \;;\; \overline{Y} = \frac{\sum 570}{10} = 57$$

$$\beta_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2\, y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \qquad \beta_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1\, y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_1 = \frac{(956)(504) - (900)(524)}{(576)(504) - (524)^2} = 0.65 \qquad \beta_2 = \frac{(900)(576) - (956)(524)}{(576)(504) - (524)^2} = 1.11$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} \qquad \beta_0 = 57 - (0.65 * 18) - (1.11 * 12) = 31.98$$

Regression Model;

$$\hat{y}_1 = 31.98 + 0.65\, X_{1i} + 1.11\, X_{2i} \pm u$$

Table (3.6) Equation of the ANOVA table for simple linear regression;

Source of Variation	***Sums of Squares***	***Df***	***Mean Square***	***F***
Regression	$(\widehat{Y}-\overline{Y})^2$	***K- 1***	SS_{reg} / K-1	MS_{reg} / MS_{res}
Residual	$(y-\widehat{y})^2$	***N – K***	SS_{res} / (N – K)	
Total	$(Y-\overline{Y})^2$	***N – 1***		

Continued Table (3.6) Calculating the ANOVA table for simple linear regression;

Source of Variation	***Sum of Squares***	***Df***	***Mean Square***	***F***
Regression	***1620.33***	***2***	***1620.47 / 2 =810.235***	***810.235 / 1.953= 414.87***
Residual	***13.67***	***7***	***13.6704 / 7 =1.953***	
Total	***1634***	***9***		

Coefficient of Determination (R^2)

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{SSR}{SST} = 1 - \frac{\sum u_i^2}{\sum y_i^2} = 1 - \frac{SSe}{SST} \quad R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{SSR}{SST} = \frac{1620.47}{1634} = 0.9917 = 99.17\,\%$$

$$R^2 = 1 - \frac{\sum u_i^2}{\sum y_i^2} = 1 - \frac{SSe}{SST} = 1 - \frac{13.67}{1634} = 1 - 0.008365 = 0.9917 = 99.17\,\%$$

We can say that 99.17% of the variation in the exam performance variable is explained by revision time and lecture attendance variables.

Adjusted R^2

The adjusted R Square value is adjusted for the number of variables included in the regression equation. This is used to estimate the expected shrinkage in R Square that would not generalize to the population because our solution is over-fitted to the data set by including too many independent variables. If the adjusted R Square value is much lower than the R Square value, it is an indication that our regression equation may be over-fitted to the sample, and of limited generalize ability.

$$AdjR^2 = 1 - \frac{n-1}{n-k} * (1 - R^2) = 1 - \frac{9}{7} * (1 - 0.9917) = 0.989 = 98.9\%$$

For the mentions example, R Square = **0.9917** and the Adjusted R Square = **0.989**. These values are very close, anticipating minimal shrinkage based on this indicator.

3.6. Exercises

A study examined the heat generated during the hardening of Portland cements, which was assumed to be a function of the chemical composition, the following variables were measured, where

x1 : amount of tricalcium aluminate x2 : amount of tricalcium silicate
x3 : amount of tetracalcium alumino ferrite x4 : amount of dicalcium silicate
Y : heat evolved in calories per gram of cement.

Data

No	Y	X_1	X_2	X_3	X_4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Investigate the relationship and affects of tricalcium aluminate, tricalcium silicate, tetracalcium alumino ferrite and dicalcium silicate on heat evolved in calories per gram of cement.

UNIT 4
Applied Example
Using Statistical Package

4.1. Preface

In regards to technical cooperation and capacity building, this textbook intends to practice data of labor force survey year 2015, second quarter (April, May, June), in Egypt by identifying how to apply correlation and regression statistical data analysis techniques to investigate the variables affecting phenomenon of employment and unemployment.

In the previous two unit this textbook deliberately to illustrate the equation or formula of correlation and regression to demonstrate the components of each equation enabling the students to understand the meaning of correlation and regression and its mathematics calculation to be able to express its meaning and how to perform the functions of them. But calculating statistics measurements and indices manually can be very tedious without a computer and is rarely done without a computer. Large dataset make it almost impossible to do manually by hand.

So; in this section intends to apply an example of correlation and regression analysis using statistical package (SPSS) to practice data of labor force survey year 2015, second quarter, in Egypt to investigate the variables affecting phenomenon of employment and unemployment.

Egypt statistical office ***CAPMAS*** is considered under presidential decree no. 2915 of 1964 the official source for data and statistical information collection, preparation, processing, dissemination and giving official nature of the statistical figures in A.R.E. Also the responsible for Implementation of statistics and data collection of various kinds, specializations, levels and performs many of the general censuses and economic surveys. One of the key aims of CAPMAS is to complete unified and comprehensive statistical work to keep up with all developments in various aspects of life and unifying standards, concepts and definitions of statistical terms, development of comprehensive information system as a tool for planning and development in all fields.

Labor Force Sample survey was conducted for the first time in Egypt on November 1957, and continued periodically, until it is finally settled as a quarterly issued since 2007. Starting from January 2008 has been development the methodology that used to collect data to be representative of the study reality during the research period, by dividing the sample for each governorate into (5) parts and fulfillment each part separately by periodicity at the middle of the month during three months (in the middle and end of the month). It measures the manpower and the civilian labor force and provides data on employment and unemployment adding to their characteristics such as geographical, gender and age distribution.

4.1.1 Manpower:

The manpower includes the whole population excluding (population out of manpower):

i. Children less than 6 years old
ii. Persons of 65 years or more who don't work or have no desire to work and don't seek for work
iii. Totally disabled persons.

4.1.2 Labor Force Definition:

All the individuals which their ages range are from 15 years old (the minimum age of employment according to the Egyptian labor law) to 65 years old (the retirement age) whether they are actually taking part by their physical or mental efforts in an activity related to the production of commodities and services. Starting from year 2008 data is for population 15 years old and more.

4.1.3 Employed Definition:

All the individuals which their ages range are from 15 years old and more who are work in any field related to the Economic Sector part time (Minimum One hour) during the short period of the survey (One week) either in and out the establishments.

4.1.4 Unemployed Definition:

All the individuals whom their ages range are from (15- 64 years old) who's had the ability to work, want it and search for work but don't find.

4.1.5 Objectives of Egypt labor force survey:

- Measuring the size of the Egyptian civilian labor force and its characteristics.
- Measuring the level of employment in different geographical areas in state.
- Monitor the Geographic distribution of employed population according to gender, age, educational status, employment status, occupation, economic activity, sector, stability in work, and work hours.
- Measuring unemployment level in various geographical areas in state.
- Monitor the Geographical distribution of unemployed person according to gender, age, educational status, duration of unemployment, type of unemployment (previously worked, he never work), occupation and economic activity for who ever worked before.

4.1.6 Survey implementation:

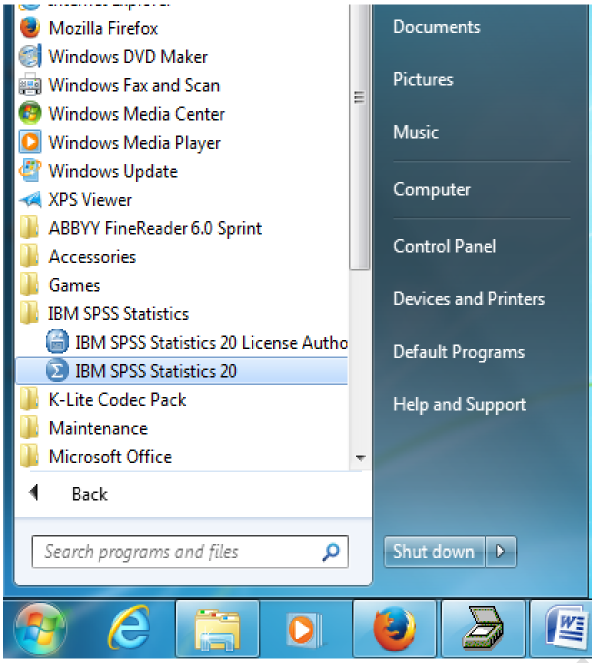
Labor force survey is a household survey conducted periodically quarterly to take in consideration the effect of seasonality on employment and unemployment. An annual aggregated bulletin published yearly in addition to this quarterly bulletin.

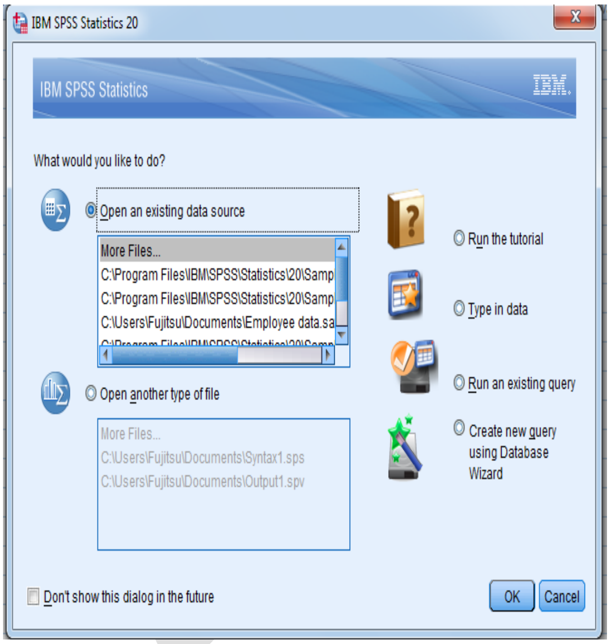
4.1.7 Sample Design:

Sample of labor force survey is a two stage stratified cluster sample and self weighted to extend practical. Sample size designed for each quarter is 22896 households with a total 91584 households per year, allocated over all governorates (urban. rural) in proportion to the size of urban and rural residents in each governorate. For more information visit: www.capmas.gov.eg

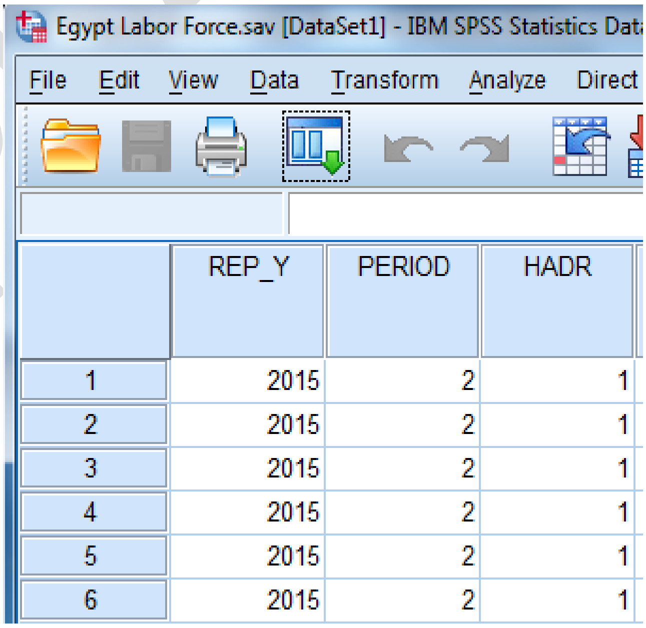
To Start SPSS:

From the Windows start menu choose: Programs >> IBM SPSS Statistics >> IBM SPSS Statistics 20; as follows:





Open Data File name "Egypt Labor Force 2015 Second quarter"



Total Number of Sample (Manpower) survey 2015 second quarter before weight is 56495 individuals where labor force is 27482 individuals whom their ages are 15 + years old and 29014 out of labor force.

4.2. Bivariate Correlation

In This part of the textbook will apply Bivariate Correlations analysis to measure the relationship strength and direction of the labor force survey interviewers characteristics like Education status, age, Marital Status, Gender, Residence and Labor force using SPSS Statistics computer Package, the procedure computes Pearson's correlation coefficient, Spearman's rho, and Kendall's tau-b with their significance levels. Two variables can be perfectly related, but if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association. So; a non-parametric distribution free test Spearman's rho, and Kendall's tau-b correlation is applied.

To run a correlations analysis, from the menus choose: Analyze >>Correlate >> Bivariate …..

- Select Labor force, Education status, age, Marital Status, gender and Residence as analysis variables.
- For Test of Significance. Select two-tailed.
- For correlation coefficient Select Pearson , Kendall's tau-b and Spearman
- Click Ok.

The Bivariate Correlations procedure computes the pairwise associations for a set of variables in larger set, cases are included in the computation when the two variables have no missing values, irrespective of the values of the other variables in the set. The classification of variables as follows:

1. Labor force dummy variable; employed =1 , unemployed = 0.
2. Education status continues (scale) variable.
3. Age continues (scale) variable.
4. Marital status, Dummy variable; (married, step of married, divorced and widowed) =1, (less than age and never married) = 0.
5. Gender, dummy variable; female =1 male = 0.
6. Residence, dummy variable rural = 1, urban = 0.

Table (4-1) Bivariate correlation matrix for mentioned variables.

Correlations						
		education status	age	marital status	gender	Residence
Pearson correlation						
Labor force	Coefficient	.244**	.260**	.316**	.251**	-.043-**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	27482	27482	27482	27482	27482
Kendall's tau_b correlation						
Labor force	Coefficient	.165**	.139**	.184**	.395**	-.037-**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	27482	27482	27482	27482	27482
Spearman's rho correlation						
Labor force	Coefficient	.227**	.206**	.239**	.307**	-.038-**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	27482	27482	27482	27482	27482

****. Correlation is significant at the 0.01 level (2-tailed).**

***. Correlation is significant at the 0.05 level (2-tailed).**

The Pearson correlation coefficient, Kendall's tau_b correlation coefficient and Spearman's rho correlation coefficient have slightly different in values of coefficient and the relationships between variables are statistically significant at 0.01 level. There is a significant fairly weak or moderate positive correlation between Labor force and (Education status, Age, Marital status and Gender) means that as education increase by one education level or age increase by one year or individual marital status and female gender, the opportunity to have job is fairly weak or moderate. But there is no significantly difference in job opportunity for individuals who live in rural or urban area. All of which are due to political instability and complicated security situation that hurt the economics, tourism and foreign investment causes low growth rates, soaring unemployment, widening budget deficits and dwindling foreign reserves.

4.3. Partial Correlation

The Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between Labor force, education status and gender while controlling for the effects of marital status, age and residence.

To obtain partial correlations: Analyze >>Correlate >> Partial

- Select Labor force, Education status and Gender as the variables.
- Select Marital Status, Age and Residence as the control variable**.**
- Click Options
- Click (check) Zero-order correlations and then click Continue.
- In the main Partial Correlations dialog, click Ok to run the procedure.

Table (4-2) Partial Correlation Matrix

Correlations					
Control Variables			Labor force	Education status	Gender
Marital Status & Age & Residence	Labor force	Correlation	1.000	.128	.224
		Significance (2-tailed)	.	.000	.000
		df	0	27482	27482
	Education status	Correlation	.128	1.000	-.155-
		Significance (2-tailed)	.000	.	.000
		df	27482	0	27482
	Gender	Correlation	.224	-.155-	1.000
		Significance (2-tailed)	.000	.000	.
		df	27482	27482	0

Table (4-1) shows the zero-order correlations which mean correlations of all variables without any control variables and table (4-2) shows the partial correlation controlling variables which are Marital Status, age and Residence controlling for the relationship of education status and gender variables with labor force.

The zero-order correlation and partial correlation between labor force and education status is slightly difference but indeed, both fairly weak or moderate respectively as (0.0.244) and(0.128) and statistically significant ($p < 0.001$), and the same case of correlation between labor force and female gender which is respectively equals (0.251) and (0.224). On interpretation of this finding removing the effects of controlling variables reduces the correlation between the other variables to almost half in case of labor force and education status. Where slight effect in case of labor force and gender.

4.4. Linear Regression Model

In this section Linear Regression analysis will be apply between labor force as a dependent variable and Age, Gender, Residence, Education status and Marital as an independent variables that best predict the value of the dependent variable. To carry out multiple regression using SPSS Statistics, as well as interpret and report the results from this test. However, the different assumptions need to be understood in order to have a valid result. See

In the Linear Regression dialog box, select a numeric dependent variable. Select one or more numeric independent variables.

To Obtain a Linear Regression Analysis From the menus chooses: Analyze >> Regression >> Linear…

- Select labor force as the dependent variable.
- Select Age, Gender, Residence, Education status and Marital status as independent variables.
- Click **OK** in the Linear Regression dialog box.

<u>Out put</u>

Table (4.3) Model Summary

Model	R	R square	Adjusted R square	Std. Error of the Estimate
1	.414[a]	.172	.172	.656

a. Predictors: (Constant), age, gender, Residence, education status, marital status

Table (4.3) shows the coefficient of determination (R^2)' as a whole, the regression does an extreme weak job of modeling Labor force. Where the coefficient of determination (R^2) is nearly 17.2% of the variation in labor force is explained by the model.

Table (4.4) ANOVA[a]

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	7803.895	5	1560.779	3632.378	.000[b]
Residual	37670.931	27476	.430		
Total	45474.826	27481			

a. Dependent Variable: Labor force

The ANOVA table (4.4) reports a significant F statistic, indicating that using the model is better than guessing the mean.

The coefficient table (4.5) shows that model fit looks positive, the first section of the coefficients table shows that there are too many predictors in the model. There are several significant coefficients, indicating that these variables contribute much to the model. Where residence is non-significant coefficients indicating that residence do not contribute much to the model.

Table (4.5) Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.555	.011		48.751	.000
Education status	.036	.001	.175	50.730	.000
Gender	.348	.005	.241	75.947	.000
Marital status	.080	.003	.192	27.509	.000
Residence	-.002-	.005	-.001-	-.337-	.736
Age	.001	.000	.029	4.359	.000

a. Dependent Variable: Labor force

To determine the relative importance of the significant predictors, look at the standardized coefficients (Beta). Even though Age has a small coefficient compared to Gender, Marital Status and Education status actually contributes more to the model because it has a larger absolute standardized coefficient, where gender is highly important than married status and education status.

4.5. Stepwise Analysis Methods

Stepwise regression is an approach to selecting a subset of effects for a regression model. It is a step-by-step iterative construction of a regression model that involves automatic selection of independent variables; it interactively explores which predictors seem to provide a good fit. It improves a model's prediction performance by reducing the variance caused by estimating unnecessary terms. The Stepwise platform also enables you to explore all possible models and to conduct model averaging.

To Obtain Stepwise Regression Analysis From the menus chooses: Analyze >> Regression >> Linear…

In the dialogue box

- Select labor force as the dependent variable.
- Select Age, Gender, Residence, Education status and Marital Status as independent variables.
- In method Tab change "Enter" to "Stepwise"
- Click OK in the Linear Regression dialogue box.

The output viewer appears and shows the following tables:

The Model Summary table (4.6) presents details of the overall correlation and R^2 and Adjusted R^2 values for each step along with the amount of R^2 Change. In the first step, as can be seen from the footnote beneath of the Model Summary table, marital status was entered into the model. The R^2 with that predictor in the model was .100. On the second step, positive affect was added to the model. The R^2 with both predictors (marital status, gender) in the model was .147; thus, we gained .047 in the value of R^2 (.147 – .100 = .047), and this is reflected in the R^2 Change for that step. At the end of the fourth step, R^2 value has reached .172. Note that this value is identical to the R^2 value we obtained under the standard "Enter" method.

Table (4.6) Models Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.316[a]	.100	.100	.683
2	.383[b]	.147	.147	.665
3	.414[c]	.171	.171	.656
4	.414[d]	.172	.172	.656
a. Predictors: (Constant), marital status				
b. Predictors: (Constant), marital status, gender				
c. Predictors: (Constant), marital status, gender, education status				
d. Predictors: (Constant), marital status, gender, education status, age				

ANOVA table (4.7) displays the results of the analysis, there are 27481 (N-1) total degrees of freedom. With four predictors, the Regression effect has 4 degrees of freedom. Table shows four F-tests, one for each step of the procedure, all steps had overall Significant results (P-value = .000). The Regression effect is statistically significant indicating that prediction of the dependent variable is accomplished better than can be done by chance.

Table (4.7) ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4544.074	1	4544.074	9733.555	.000[b]
	Residual	40930.752	27480	.467		
	Total	45474.826	27481			
2	Regression	6663.288	2	3331.644	7526.076	.000[c]
	Residual	38811.538	27479	.443		
	Total	45474.826	27481			
3	Regression	7795.536	3	2598.512	6046.275	.000[d]
	Residual	37679.290	27478	.430		
	Total	45474.826	27481			
4	Regression	7803.846	4	1950.962	4540.490	.000[e]
	Residual	37670.980	27477	.430		
	Total	45474.826	27481			
a. Dependent Variable: Labor force						
b. Predictors: (Constant), marital status						
c. Predictors: (Constant), marital status, gender						
d. Predictors: (Constant), marital status, gender, education status						
e. Predictors: (Constant), marital status, gender, education status, age						

There should also be a Coefficients table (4.8), showing the linear regression equation coefficients for the various model variables. The "B" values are the coefficients for each variable, that is, they are the value which the variable's data should be multiplied by in the final linear equation we might use to predict long term Labor force with. The "Constant" is the intercept equivalent in the equation. The figures should be Significance at 0.05 or below to have confidence degree 95 %, this means that finding has 95% chance of being true.

Table (4.8) Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.134	.004		281.457	.000
	marital status	.131	.001	.316	98.659	.000
2	(Constant)	.693	.007		92.387	.000
	marital status	.121	.001	.291	92.702	.000
	gender	.313	.005	.217	69.190	.000
3	(Constant)	.561	.008		71.730	.000
	marital status	.091	.001	.219	64.341	.000
	gender	.344	.004	.239	76.470	.000
	education status	.035	.001	.174	51.328	.000
4	(Constant)	.552	.008		68.573	.000
	marital status	.080	.003	.192	27.565	.000
	gender	.348	.005	.241	76.011	.000
	education status	.036	.001	.176	51.507	.000
	age	.001	.000	.029	4.398	.000
a. Dependent Variable: Labor force						

This Coefficients table gives Beta coefficients so that you can construct the regression equation. Notice that the betas change, depending on which predictors are included in the model. These are the weights that you want, for the model number four that includes marital status, gender, education status and age the predictors are ranked according to strength of their effectiveness considering the Beta values from the highest to lowest effect as (gender .241, marital status .192, education status .176, age .029).

There should also be an Excluded Variables table (4.9) showing the variables removed from each model.

Table (4.9) Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	education status	.139[b]	40.131	.000	.134	.838
	gender	.217[b]	69.190	.000	.228	.987
	Residence	-.028-[b]	-8.596-	.000	-.029-	.998
	age	-.084-[b]	-12.427-	.000	-.042-	.222
2	education status	.174[c]	51.328	.000	.171	.823
	Residence	-.027-[c]	-8.709-	.000	-.029-	.998
	age	-.008-[c]	-1.180-	.238	-.004-	.216
3	Residence	-.002-[d]	-.672-	.502	-.002-	.972
	age	.029[d]	4.398	.000	.015	.214
4	Residence	-.001-[e]	-.337-	.736	-.001-	.967

a. Dependent Variable: Labor Force

b. Predictors in the Model: (Constant), marital status

c. Predictors in the Model: (Constant), marital status, gender

d. Predictors in the Model: (Constant), marital status, gender, education status

e. Predictors in the Model: (Constant), marital status, gender, education status, age

The last table (4.9) Variables Excluded from the Equation, just lists the variables that weren’t included in the model at each step, where residence are removed from four model because of its null effect to the dependent variable Labor force.

4.6. Exercises

Practice one of the official statistics or surveys conducted by the country statistics office, to examine the relationship between the survey parameters and study the factors affected the phenomena or the survey problem.

GLOSSARY

Exploratory methods
Allow us to get a preliminary look at a dataset through basic statistical aggregates and interactive visualization.

Statistical inference
Statistical inference consists of getting information about an unknown process through partial and uncertain observations. In particular, estimation entails obtaining approximate quantities for the mathematical variables describing this process.

Decision theory
Allows us to make decisions about an unknown process from random observations, with a controlled risk.

Correlation Coefficient
A measure of the degree to which variation of one variable is related to variation in one or more other variables. The most commonly used correlation coefficient indicates the degree to which variation in one variable is described by a straight line relation with another variable.

Positive Correlation
A relationship between two variables in which both variables move in the same direction. A positive correlation exists when as one variable increases, the other variable also increases and vice versa. In statistics, a perfect positive correlation is represented by the value +1, 0 where 0 indicates no correlation and +1 indicates a perfect positive correlation.

Negative Correlation
An inverse or contrary relationship between two variables such that they move in opposite directions. In an inverse correlation with variables A and B, as A increases, B would decrease, and vice versa. In statistical terminology, an inverse correlation is denoted by the correlation coefficient r having a value between -1 and 0, with r = -1 indicating perfect inverse correlation.

Pearson r correlation
Pearson r correlation is widely used in statistics to measure the degree of the relationship between linear related variables.

Kendall rank correlation
Kendall rank correlation is a non-parametric test provides a distribution free test of independence and a measure of the strength of dependence between two variables. Kendall's rank correlation is satisfactory used if it is difficult to interpret the independence between the two variables when the null hypothesis of independence between them is rejected, by reflecting the strength of the dependence between the variables being compared.

Spearman rank correlation
Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. Spearman's rank correlation is satisfactory for testing a null hypothesis of independence between two variables but it is difficult to interpret when the null hypothesis is rejected. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the

distribution of the data, and is the appropriate correlation analysis when the variables are measured on an ordinal scale.

Homoscedastic
Homoscedastic data have the same standard deviation in different groups where data are

Heteroscedastic
Heteroscedastic have different standard deviations in different groups and assumes that the relationship between the two variables is linear.

Simple Regression
Regression involving variables one of which may be regarded as dependent and one which may be regarded as independent, If we calculate a regression of, say, weight on height, that regression is called a simple regression**.**

Independent variables
Are characteristics that can be measured directly; these variables are also called predictor or explanatory variables used to predict or to explain the behavior of the dependent variable.

Dependent variable
Is a characteristic whose value depends on the values of independent variables.

Coefficient of Determination R2
The coefficient of determination (R^2) is a measure of the proportion of variance of a predicted outcome. With a value of 0 to 1, the coefficient of determination is calculated as the square of the correlation coefficient (r) between the sample and predicted data. The coefficient of determination shows how well a regression model fits the data. Its value represents the percentage of variation that can be explained by the regression equation. A value of 1 means every point on the regression line fits the data; a value of 0.5 means only half of the variation is explained by the regression. The coefficient of determination is also commonly used to show how accurately a regression model can predict future outcomes.

READING LIST

Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Psychology Press. ISBN 0-8058-2223-2.

Foster, Dean P., & George, Edward I. (1994). The Risk Inflation Criterion for Multiple Regression. Annals of Statistics, 22(4). 1947–1975. doi:10.1214/aos/1176325766

Hazewinkel, Michiel, ed. (2001), "Correlation (in statistics)", Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4

Harrell, F. E. (2001) "Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis," Springer-Verlag, New York.

Oestreicher, J. & D. R. (February 26, 2015). Plague of Equals: A science thriller of international disease, politics *and drug discovery*. California: Omega Cat Press. p. 408. ISBN 978-0963175540.

REFERENCES

Aldrich, John (1995). "Correlations Genuine and Spurious in Pearson and Yule". Statistical Science 10 (4): 364–376. doi:10.1214/ss/1177009870. JSTOR 2246135.

Anscombe, Francis J. (1973). "Graphs in statistical analysis". The American Statistician 27: 17–21. doi:10.2307/2682899. JSTOR 2682899.

Dowdy, S. and Wearden, S. (1983). "Statistics for Research", Wiley. ISBN 0-471-08602-9 pp 230.

Francis, DP; Coats AJ; Gibson D (1999). "How high can a correlation coefficient be?". Int J Cardiol 69 (2): 185–199. doi:10.1016/S0167-5273(99)00028-5.

Kendall, M. G. (1955) "Rank Correlation Methods", Charles Griffin & Co.

Lopez-Paz D. and Hennig P. and Schölkopf B. (2013). "The Randomized Dependence Coefficient", "Conference on Neural Information Processing Systems" Reprint

Mahdavi Damghani, Babak (2012). "The Misleading Value of Measured Correlation". Wilmott 2012 (1): 64–73. doi:10.1002/wilm.10167.

Mahdavi Damghani B. (2013). "The Non-Misleading Value of Inferred Correlation: An Introduction to the Cointelation Model". Wilmott Magazine. doi:10.1002/wilm.10252.

Nikolić, D; Muresan, RC; Feng, W; Singer, W (2012). "Scaled correlation analysis: a better way to compute a cross-correlogram". European Journal of Neuroscience: 1–21. doi:10.1111/j.1460-9568.2011.07987.x.

Rodgers, J. L.; Nicewander, W. A. (1988). "Thirteen ways to look at the correlation coefficient". The American Statistician 42 (1): 59–66. doi:10.1080/00031305.1988.10475524. JSTOR 2685263.

Székely, G. J. Rizzo; Bakirov, N. K. (2007). "Measuring and testing independence by correlation of distances". Annals of Statistics 35 (6): 2769–2794. doi:10.1214/009053607000000505.

Székely, G. J.; Rizzo, M. L. (2009). "Brownian distance covariance". Annals of Applied Statistics 3 (4): 1233–1303. doi:10.1214/09-AOAS312.

Thorndike, Robert Ladd (1947). Research problems and techniques (Report No. 3). Washington DC: US Govt. print. off.

Yule, G.U and Kendall, M.G. (1950), "An Introduction to the Theory of Statistics", 14th Edition (5th Impression 1968). Charles Griffin & Co. pp 258–270