An-Najah National University

Faculty of Graduated Studies

Estimation of Multivariate Multiple Linear Regression Models and Applications

By

Jenan Nasha’t Sa’eed Kewan

Supervisor

Dr. Mohammad Ass’ad

Co-Supervisor

Dr. Ali Barakat

This Thesis is Submitted in Partial Fulfillment of Requirements for the Degree of Master of Mathematics, Faculty of Graduate Studies, An-National University, Nablus, Palestine.

2015

Estimation of Multivariate Multiple Linear Regression Models and Applications

By

Jenan Nasha't Sa'eed Kewan

This thesis was successfully defended on 12 /5 / 2015 and approved by:

Defense Committee Members **Signature**

- **Dr. Mohammad Ass'ad** **(Supervisor)** ………………….
- **Dr. Ali Barakat** **(Co- Supervisor)** …A. Barakat…
- **Dr. Mahmoud Almanassra** **(External Examiner)** ………………….
- **Dr. Jihad Abdallah** **(Internal Examiner)** ………………….

Dedication

To the most wonderful person in presence, who encourages me, supports me, and believes in me. The person who taught me that there is no such thing called impossible. To “my father”.

To the greatest heart in the world, the cause of happiness in my life, to “my mother”.

To my lovely brothers “Jehad” and “Raslan”.

To my cute sister “Razan”.

To my lovely uncle, brother, and friend “Emad”.

Acknowledgment

First of all I am grateful to The Almighty Allah for helping me to complete this thesis, Praise and thanks to Allah.

I would like to express my sincere gratitude and appreciation to my supervisor Dr. Mohammad Ass’ad for his helpful guidance, inspiring ideas, invaluable advice, and immense knowledge. I am grateful to him for devoting his time to help me in completing this thesis and teach me how to be different and distinctive. I am indebted to him for his support and encouragement. I would also like to express my sincere thanks to Dr. Ali Barakat, my co-supervisor, for his insight, invaluable advice, guidance, and assistance. Thanks and appreciations to Dr. Mohammad Ass’ad and Dr. Ali Barakat for their enthusiasm, patience, and motivation. I hope I have made them proud.

Thanks and appreciations to the members of committee discussion Dr. Mahmoud Almanassra and Dr. Jihad Abdallah for their efforts, time, and patience.

Special thanks go to Dr. Fayez Mahameed, Psychology Department, An-Najah National University, who helped me to get standard questionnaires and analyze them. Thanks to all people who helped me to complete my work on the case study; thanks to Dr. Sulaiman Kayed, Alquds Open University. Thanks to Dr. Saleh Affaneh, Arab American University. Thanks to my cousin Miss Nisreen Al-Haj, An-Najah National University. Special word of thanks also goes to my brothers, Jehad, Emad, and Raslan for their efforts in spreading out the questionnaires and get the software programs that I needed in my work. Special thanks to Dr. Mohammad Al-Sayed, Faculty of Engineering, An- Najah National University, for his help in Matlab software.

Finally, thanks to all my friends and for every person who supported me, helped me, and believed in me.

الإقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Estimation of Multivariate Multiple Linear Regression Models and Applications

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو جزء منها لم يقدم من قبل لنيل أي درجة أو بحث علمي أو بحثي لدى أي مؤسسة تعليمية او بحثية أخرى.

Declaration

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

Student's name: جنان نشأت سعيد كيوان **اسم الطالبة:**

Signature: **التوقيع:**

Date: 12/5/2015 **التاريخ:**

Table of contents

List of Figures

List of Tables

Estimation of Multivariate Multiple Linear Regression Models and Applications

By

Jenan Nasha’t Sa’eed Kewan

Supervisor

Dr. Mohammad Ass’ad

Co-Supervisor

Dr. Ali Barakat

Abstract

Regression is a statistical method which is widely used in various fields of science for defining the relationships between the variables in the form of an equation to estimate the parameters, the strength and direction of the relationships. This main objective of this thesis was to study the Multivariate Multiple Linear Regression Models which relate more than one dependent variable with more than one independent variable.

Four types of regression models were considered; Simple Regression, Multiple Regression, Multivariate Regression, and Multivariate Multiple Regression. The method of least squares was used in estimating the multivariate multiple linear regression models. Then multivariate analysis of variance (MANOVA) was used to test the usefulness of the estimated models. Several software programs were used to achieve this objective, such as Stata, Matlab, Minitab, SPSS, and SAS.

The multivariate multiple regression model was applied to simulated data and to real data. A case study was constructed from three universities in Palestine; An-Najah National University, Arab American University, and Alquds Open University. A sample of size 350 students from these universities was considered to study the relationship between two psychological variables (self concept and achievement motivation) and the cumulative average of the student, which were the response variables. Then we showed how these three responses were affected by some predictors such as tawjihi average, English score level exam, budget per day, absents per semester, the study program of the student, and the university he/she is studying in. The results showed, according to the p-values, that the study program has no effect on both self concept and achievement motivation of a student. When we test the effect of the study program on these two responses, the p-values were 0.5674, and 0.2227, respectively. On the other hand, the cumulative average was affected by the study program, the p-value was 0.0501. The cumulative average of the student was also affected by the university, it was with null p-value. But the student's self concept and achievement motivation were not affected by the university, the p-values were 0.2530, 0.4352, respectively. Also, data for each university was analyzed separately. The results for An-Najah National University showed that the twajihi average controlled the responses very well and the student's cumulative average was affected by the study program. The results of The Arab American University showed that the three responses were affected by tawjihi average and English score, but they were not affected by the study program. Finally, the results of Alquds Open University showed that the three responses were affected by just one factor which is the tawjihi average.

Chapter One

Introduction

1.1 Overview

In statistics, the term linear model is used in different ways according to the context. The most common occurrence is in connection with regression models. Linear models describe a continuous response variable as a function of one or more predictor variables. They can help understand and predict the behavior of complex systems or analyze experimental, financial, and biological data. The term linear model is often taken as synonymous with linear regression model.

The term “regression” was first coined by Sir Francis Galton, an accomplished 19th century scientist. He tried to describe a biological phenomenon [10]. The phenomenon was that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. Rather, the characteristics in the offspring regress towards a mediocre point (a point which has since been identified as the mean). By measuring the heights of hundreds of people, Galton observed that children’s heights tend to ‘revert’ to the average height of the population rather than diverting from it, (a phenomenon also known as regression toward the mean) [11]. GaIton originally used the word ‘reversion’ to describe this tendency and some years later used the word ‘regression’ instead. He was able to quantify regression to the mean, and estimate the size of the effect. Galton wrote that, “the average regression of the offspring is a constant fraction of their respective mid-parental deviations”. This means that the difference between a child and its parents for some characteristic is proportional to its parents' deviation from typical people in the population. If its parents are each two inches taller than the averages for men and women, on average, it will be shorter than its parents by some factor (which, today, we would call one minus the regression coefficient) times two inches. For height, Galton estimated this coefficient to be about 2/3: the height of an individual will measure around a midpoint that is two thirds of the parents’ deviation from the population average. This is incorrect, since a child receives its genetic makeup exclusively from its parents. There is no generation-skipping in genetic material: any genetic material from earlier ancestors than the parents must have passed through the parents. The phenomenon is better understood if we assume that the inherited trait (height) is controlled by a large number of recessive genes. Exceptionally tall individuals must be homozygous for increased height mutations on a large proportion of these loci. But the loci which carry these mutations are not necessarily shared between two tall individuals, and if these individuals mate, their offspring will be on average homozygous for "tall" mutations on fewer loci than either of their parents. In addition, height is not entirely genetically determined, but also subject to environmental influences during development, which make offspring of exceptional parents even more likely to be closer to the average than their parents.

For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context around the 20th century. George Udny Yule initially studied to be a civil engineer, but through the influence of the famous statistician, Karl Pearson, he turned his attention to theoretical and inferential statistics. They completed pioneering work developing multiple regression models [20] [27]. Yule wrote, speaking about the typical fit curve of y as a function of x over many data points: “It is a fact attested by statistical experience that these means do not lie chaotically all over the table, but range themselves more or less closely round a smooth curve, which we will name the curve of regression of x on y. So regression methods evolve from finding the curve of regression, which itself is the best fit for groups of observations after allowing some of the variation to be declared “unexplained” and left in a noise term. This is advance from mere fitting or solving where you might be trying to explain all of the observed variation in n-individuals using as many as n-variables [27].

The earliest form of regression was the method of least squares, which was published by Legendre in 1805 [19], and by Gauss in 1809 [12]. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821 [13], including a version of the Gauss–Markov theorem. In 1806, Legendre published new methods to determine the orbits of comets. His method involved three observations taken at equally spaced intervals and he assumed that the comet followed a parabolic path so that he ended up with more equations than there were unknowns [1]. However, Gauss published his version of the least squares method in 1809 and, while acknowledging that it appeared in Legendre's book, Gauss still claimed priority for himself. This greatly hurt Legendre, leading to one of the infamous priority disputes in the history of mathematics.

1.2 Regression Analysis

Regression analysis is a statistical process for estimating the relationships among variables. This includes estimating the parameters of the regression model, showing the strength and direction of the relationships, and assessing the estimated model. Regression analysis includes many techniques for modeling and analyzing several variables. When the focus is on the relationship between a dependent variable y (also called a response variable) and one or more independent variables x_i (called the predictors or the explanatory variables) [18]. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, and the other independent variables are held fixed. It is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. Regression analysis is applied in many areas of academic and applied sciences such as social sciences, medical researches, biology, meteorology, psychology, chemistry and economics.

Regression analysis tries to find answers to questions such as “Is there a relationship between dependent and independent (explanatory) variables? If there is, what is the power of this relationship? How good is this relationship? What kind of a relationship is there between the variables?” Regression analysis can be defined as the expression of the relationship between dependent and independent variables in the form of a mathematical function. Most studies assume a linear relationship between independent and dependent variables [8]. The parameters of the regression models usually unknown and can be estimated using different methods. One of the most commonly used prediction techniques is the method of least squares which will be used in this research. The correlation coefficient and the coefficient of determination will indicate the strength of the estimated relationships, and the sign of the correlation coefficient will be the indicator for the direction of this relationships. Regression analysis is an easily comprehensible method. Today, it has a wide area of usage and applications with the help of statistical package softwares such as SPSS, Minitab, Matlab, SAS, and Stata.

A regression model in which only one independent variable is used to predict the value of one dependent variable is called simple regression model. Whereas a regression model in which more than one independent variable is used to predict the value of one dependent variable is called multiple regression model. Also, a regression model in which only one independent variable is used to predict the value of many dependent variables is called multivariate regression model. Whereas a regression model in which more than one independent variable is used to predict the value of more than one dependent variable is called multivariate multiple regression model. Many people confuse between the multivariate regression model and the multivariate multiple regression model and consider them the same. Here, we separate between the two models and give an explanation about the multivariate regression model, derive its formulas, write the model in matrix form, and give examples. So, we consider four types of regression models; ***Simple Regression***, ***Multiple Regression***, ***Multivariate Regression***, and ***Multivariate Multiple Regression***.

The main objective of this research was to study the ***Multivariate Multiple Regression Model***. We constructed two main examples to estimate this model. The first one using simulation, data from different distributions has been generated on three psychological variables, four academic variables (standardized test scores) for 20 high school students. We were interested in how the set of psychological variables is related to the academic variables. In the second example, A case study was generated about university students in Palestine. Data for 350 university students has been collected on three dependent variables; two psychological variables (self concept, achievement motivation) and cumulative average of the student. We were interested in how the set of the dependent variables is related to the set of independent variables which were tawjihi average, English score level exam, budget, absents, and the study program of the student.

After estimating the regression model and expressing the relationship between dependent and independent variables in the form of a mathematical function, we have to find answers to the questions; How good is the relationship? And what is the power of this relationship? We gave answers for these questions through assessing the estimated models by testing the goodness of fit for the models. One of the most important techniques for assessing the regression model is the analysis of variance approach (ANOVA). We used ANOVA to test for the significance of the estimated regression model in simple and multiple regression in chapter 2, i-e ANOVA can be used when we have only one dependent variable. However, ANOVA can’t be used when we have more than one dependent variable; instead, we use multivariate analysis of variance (MANOVA). MANOVA has been applied to test the significance of the estimated multivariate multiple regression models in the simulated data and in the case study. The MANOVA output for the first case study showed that the predictor variables; tawjihi, English score, budget, and absent were good predictors for the response variables. Our decision was according to the p-values of the four predictors which were 0.0000, 0.0001, 0.0011, and 0.0010 respectively. Small p-values indicate us to reject the null hypothesis that a certain variable has no effect on the responses. The results of MANOVA will be explained in details in chapters three and four.

Chapter Tow

Estimating Linear Regression Models

Regression analysis is used to answer questions about how one variable depends on one or more other variables. For instance, does diet correlate with cholesterol level, and does this relationship depend on other factors, such as age, smoking status, and level of exercise?

Regression models can answer these questions. They describe the relationship between a dependent variable, which is diet in our example, and an independent variable or variables, which are cholesterol level, age, smoking status, and level of exercise.

2.1 Linear Regression Models

2.1.1 Definitions, Basic Concepts, and Examples

– A **regression model** is a mathematical equation that describes the relationship between two or more variables, sometimes we call it regression equation.

– And by **linear regression model** we mean a model that assumes a linear relationship between two or more variables.

– Types of linear regression models:

- **Simple Linear Regression:** When we consider the relationship between one dependent variable and one independent variable, we use Simple Linear Regression.

- **Multiple Linear Regression**: When we consider the relationship between one dependent variable and more than one independent variable, we use Multiple Regression.
- **Multivariate linear regression**: When we consider the relationship between more than one dependent variable and one independent variable, we use Multivariate Regression.
- **Multivariate Multiple Linear Regression:** When we consider the relationship between more than one dependent variable and more than one independent variable, we use Multivariate Regression. This is the model that we are most interested in and will study it in details later in chapter 3.

- ➢ **Correlation analysis** is a statistical approach used to measure the strength of the relationship among variables. The term correlation most often refers to the linear association between two quantities or variables, that is, the tendency for one variable to increase or decrease as the other increases or decreases, in a straight-line trend or relationship. Correlation and regression analysis are related in the sense that they both deal with relationships among variables.
- ➢ The **correlation coefficient** (also called the Pearson linear correlation coefficient) is a numerical index of the strength of relationship between two variables. Values of the correlation coefficient are always between -1 and +1. A correlation coefficient of +1 indicates that the two variables are perfectly related in a positive linear sense. A correlation coefficient of -1 indicates that two variables are perfectly

related in a negative linear sense. A correlation coefficient of 0 indicates that there is no linear relationship between the two variables.

- The **population correlation coefficient, ρ** (rho), measures the strength and direction of the linear association between the variables.
- The **sample correlation coefficient, r,** is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations. The closer r is to +1, the stronger the positive correlation is. The closer r is to -1, the stronger the negative correlation is. If |r| = 1 exactly, the two variables are perfectly correlated. A value of zero for r does not mean that there is no correlation.

Regression models are widely used for prediction. We can predict the value of a response variable from knowledge of the values of one or more explanatory variables. Below are some examples:

- We may be interested to know whether the sale price of home will rise next year.
- We may wish to examine whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, and price of cigarettes.
- A meteorologist may forecast it will rain tomorrow.
- An executive of an insurance company may predict there will be more road accidents and casualties next year.
- A researcher in education may claim that educational success depends on intelligence, economic and social class of a student.

2.1.2 Regression Equation

The simple regression equation takes the algebraic form for a straight line: $y = mx + b$, where m is the slope of the line, and b is the y-intercept. It is known from algebra that a line is identified by its slope (the angle of the line describing the change in y per unit x) and intercept (where the line crosses the y axis). So regression describes the relation between x and y with just such a line. We look for the equation or formula for the straight line that minimizes the total error.

We can use the regression equation to predict the value of the dependent variable at fixed values of the independent variable(s).

Generally, when we have k predictor variables the formula of regression equation takes the form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \;=\; x_i^T \beta + \varepsilon_i\,, \qquad i = 1,2,\ldots,n.$$

where T denotes the transpose, so that $x_i^T\beta$ is the inner product between vectors x_i and β.

Often these n equations are stacked together and written in vector form as:

$$Y = X\beta + \varepsilon,$$

where

- Y is the column vector of n observations of the response variable. The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables.

- X is called the design matrix consisting of column vectors of observations on the predictor variables.

Remark:

- Usually a constant is included as one of the predictor variables. For example we can take $x_{i1} = 1$ for $i = 1, ..., n$. The corresponding element of β is called the intercept. Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero. That is, the regression equation will take the form: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$, where β_0 represents the intercept parameter.
- Sometimes one of the predictor variables can be a non-linear function of another predictor, as in polynomial regression. The model remains linear as long as it is linear in the parameter vector β.

- β is the column vector of coefficients to be estimated. Statistical estimation and inference in linear regression focuses on β.
- ε is the error term, or noise. This variable captures all other factors which influence the dependent variable y_i other than the regressors x_i.

In fact, we don't know the exact population regression line (values of the regression coefficients), and the goal of linear regression methods is to find the "best" choices of the estimated values for the constants $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ to make the regression formula as "accurate" as possible. The regression line that we obtain from a sample provides an estimate of the population regression line. The estimated regression model:

$$E(\ y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \quad \text{(population)}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \qquad \text{(sample)}$$

where

$\hat{y}$: is the estimated y value.

$\hat{\beta}_i$: are the estimated values of the regression coefficients.

$\hat{\beta}_0$: is the estimate of the regression intercept.

The individual random error terms ε_i have a mean of zero.

- A **residual** (or the error term) is the difference between the observed response y and the predicted response $\hat{y}$, $\varepsilon_i = y_i - \hat{y}_i$.

2.1.3 Linear Regression Assumptions

The following are major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- Error values (ε) are statistically independent. This assumes that the errors of the response variables are uncorrelated with each other. Some methods are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.

- The probability distribution of the errors is normal with mean zero.
- The probability distribution of the errors has constant variance.
- The independent variables are measured with no error. The observed values of x are assumed to be a set of known constants. In other words, the predictor variables are assumed to be error-free.
- Linearity: The underlying relationship between the x variable and the y variable is linear. More generally, that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Because the predictor variables are treated as fixed values, linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently.
- The predictors are linearly independent, that is, it is not possible to express any predictor as a linear combination of the others [29].

Sometimes one of the predictors can be a non-linear function of another predictor or of the data, as in polynomial regression [21] and segmented regression [22]. The model remains linear as long as it is linear in the parameter vector β. For example, consider a situation where a small ball is being tossed up in the air and then we measure its heights of ascent h_i at various moments in time t_i. Physics tells us that, ignoring the drag, the relationship can be modeled as:

$$h_i = \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i ,$$

where β_1 determines the initial velocity of the ball, β_2 is proportional to the standard gravity, and ε_i is due to measurement errors. Linear regression can be used to estimate the values of β_1 and β_2 from the measured data. This model is non-linear in the time variable, but it is linear in the parameters β_1 and β_2. If we take regressors $x_i = (x_{i1}, x_{i2}) = (t_i, t_i^2)$, the model takes on the standard form:

$$h_i = x_i^T\beta + \varepsilon_i \quad [25]$$

Remark: In statistics, the independence and normality assumptions about the errors are called the Gauss–Markov conditions [23].

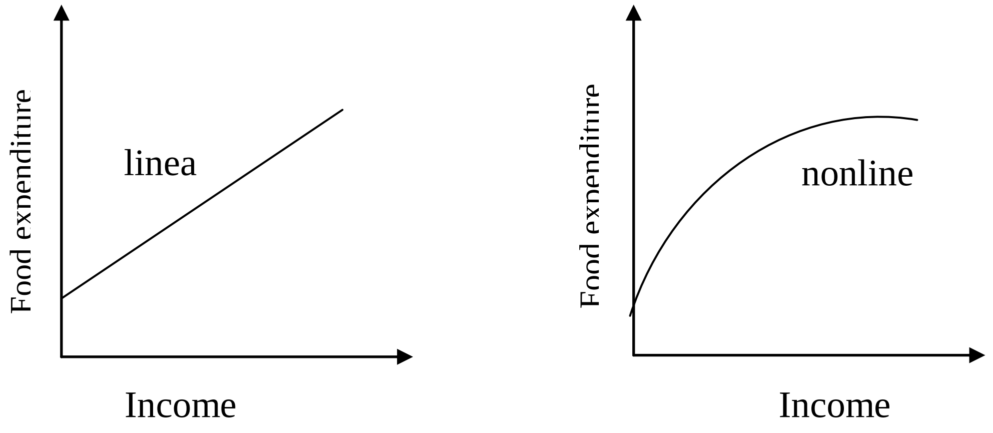
2.2 Nonlinear regression

The assumption of linearity in regression models requires that the relationship among the variables must be linear. That is a straight-line relationship between the variables (the response variable is a linear combination of the parameters and of the predictor variables). In linear regression we try to find the best straight line fitted to data, but, sometimes the true relationship that we want to model is curved. For example, if something is growing exponentially, which means growing at a steady rate, the relationship between x and y is curve. To fit something like this, we need non-linear regression, which is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables.

Nonlinear regression is a general technique to fit a curve through the data. It fits data to any equation that defines y as a function of x and one or more parameters. It finds the values of those parameters that generate the curve which comes closest to the data (minimizes the sum of the squares of the vertical distances between data points and curve).

Both linear and nonlinear regression find the estimated values of the parameters (slope and intercept for linear regression) that make the line (in linear regression) or the curve(in nonlinear regression) come as close as possible to the data.

The two diagrams in Figure 2.1 show a linear and a nonlinear relationship between the dependent variable food expenditure and the independent variable income. A linear relationship between income and food expenditure, shown in Figure 2.1 (a), indicates that as income increases, the food expenditure always increases at a constant rate. A nonlinear relationship between income and food expenditure, as shown in Figure 2.1 (b), shows that as income increases, the food expenditure increases. Although, after a point, the rate of increase in food expenditure is lower for every subsequent increase in income.



2.1(a) 2.1 (b)

Figure 2.1: Relationship between food expenditure and income

(a) Linear relationship (b) non linear relationship

Many relationships in biology and other fields of science do not follow a straight line. To analyze such data, you have two choices:

- Use nonlinear regression methods:
  Like the ordinary least squares (OLS) approach which gives the best-fit curve that minimizes the sum of squared residuals.
- Do mathematical transformations, to force the data into a linear relationship. Then use linear regression. Although these techniques are commonly used, they should be avoided. They are less accurate than nonlinear regression, and are not any easier [14].

For example, consider the nonlinear regression problem,

$$y = Ae^{bx} u$$

This is an exponential growth equation. b is the growth rate. u is a random error term. If we take the logarithm of both sides of that equation, we get:

$$\ln(y) = \ln(A) + bx + \ln(u),$$

this equation has logarithms in it, but they relate in a linear way. It is in the form $y = a + bx + \text{error}$, except that y, a, and the error are logarithms. This means that if we create a new variable, the base-e logarithm of y, written as ln(y), we can use the linear regression methods to fit

$$\ln(y) = \ln(A) + bx + \ln(u)$$

That is a way of fitting the curve

$y = Ae^{bx}$ to the data [3].

In general, there is no closed-form expression for the best-fitting parameters, as there is in linear regression. Usually numerical optimization algorithms are applied to determine the best-fitting parameters. Again, in contrast to linear regression, there may be many local minima of the function to be optimized and even the global minimum may produce a biased estimate. In practice, estimated values of the parameters are used in conjunction with the optimization algorithm to attempt to find the global minimum of a sum of squares [4], [9].

2.3 The Method of Least Squares

Least squares linear regression (also known as “ordinary least squares”, “OLS”, or often just “least squares”), is one of the most basic and most commonly used prediction techniques known to humankind, with applications in fields as diverse as statistics, finance, medicine, economics, and psychology.

The mathematical concept of least squares is the basis for several methods to fit certain types of curves and surfaces to data. Problems of fitting curves and surfaces have a history spanning several millennia. The basic idea of the method of least squares is easy to understand. It may seem unusual that when several people measure the same quantity, they usually do not obtain the same results. In fact, if the same person measures the same quantity several times, the results will vary. What then is the best estimate for the true measurement?

Why is least squares so popular?

Least squares is such an extraordinarily popular technique that often when people use the phrase “linear regression” they are in fact referring to “least squares regression”. Much of the use of least squares can be attributed to several factors:

- It is one of the earliest general prediction methods known to humankind.
- Its implementation on modern computers is efficient, so it can be very quickly applied even to problems with hundreds of features and tens of thousands of data points.
- It is easier to analyze mathematically than many other regression techniques.
- It is not too difficult for non-mathematicians to understand at a basic level.

- It is the optimal technique in a certain sense in certain special cases. In particular, if the system being studied truly is linear with additive independent normally distributed errors (i.e. with mean zero and constant variance), then the constants solved for by least squares are in fact the most likely coefficients to have been used to generate the data [28].

There is also the **Gauss-Markov theorem** which states that if the system we are modeling is linear with additive noise (error), and the random variables representing the errors made by our ordinary least squares model are uncorrelated from each other, and if the distributions of these random variables all have the same variance and a mean of zero, then the least squares method is the best unbiased linear estimator of the model coefficients, in that the coefficients it leads to have the smallest variance. So, the method of least squares gives a way to find the best estimate, assuming that the errors (i.e. the differences from the true value) are random and unbiased [23].

The goal of the least squares estimate is to choose the constants β_0, β_1, β_2, …, β_k so that our linear formula

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

is as accurate as possible. But what do we mean by “accurate”? By far the most common form of linear regression used is least squares regression, which provides us with a specific way of measuring “accuracy” and hence gives a rule for how precisely to choose our “best” constants β_0, β_1, β_2, …,

β_k once we are given a set of training data. The Least squares method says that we are to choose these constants so that for every example point in our training data we minimize the sum of the squared differences between the actual dependent variable (the exact value of y_i) and our predicted value for the dependent variable ($\hat{y}_i$). In other words, we want to select $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_k$ to minimize the sum of the values (observed y – predicted y)^2 for each training point, which is the same as minimizing the sum of the values

$$(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}))^2 = (y_i - \hat{y}_i)^2 = \varepsilon_i^2$$

To illustrate the idea of least squares in a better way, we will take the simplest form of a linear model. The relationship between two variables x and y. Let us go back to our example of the relationship between income (x) and food expenditure (y). Suppose we take a sample of seven households from a small city and collect information on their incomes and food expenditure (in hundreds of dollars) in a certain month. The data obtained was as given in table 2.1.

Table 2.1: Income and Food Expenditure of seven households

Income	Food expenditure
55	14
83	24
38	13
61	16
33	9
49	15
67	17

In this table, we have a pair of observations for each of the seven households. Each pair consists of one observation on income and a second on food expenditure. For example, the first household's income was $5500 and its food expenditure was $1400. If we plot all seven pairs of values, we obtain a scatterplot as shown in Figure 2.2.

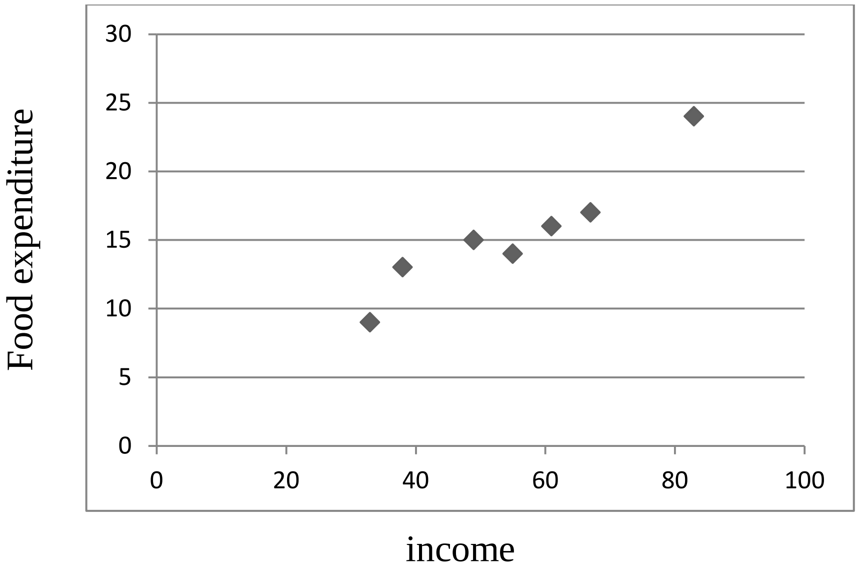
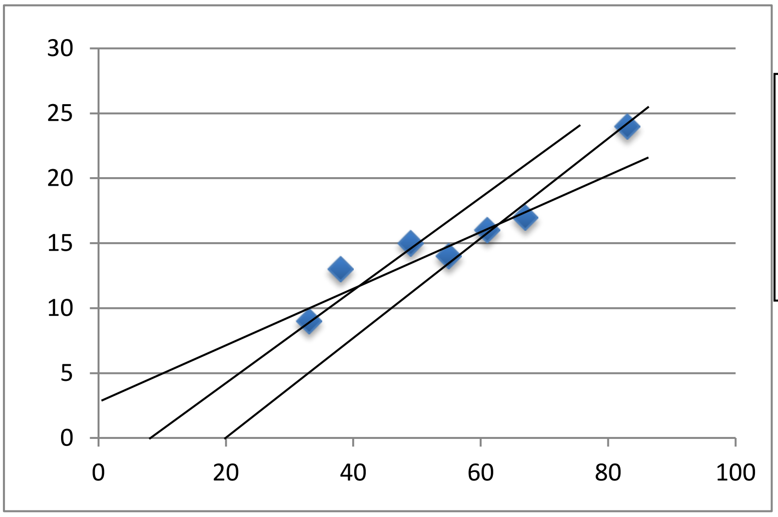


Figure 2.2: scatterplot of food expenditure and income

We would like to find a line that best describes the relationship between the variables. How do we determine which line is best?



Which line best describes the relationship between x and y?

Figure 2.3: Lines that may describe the relationship between food expenditure and income

The best line will be the one that is “closest” to the points on the scatterplot. In other words, the best line is the one that minimizes the total distance between itself and all the observed data points. We want to find the line that minimizes the vertical distance between itself and the observed points on the scatterplot. Here we have three different lines in figure 2.3 that may describe the relationship between x (income) and y (food expenditure). In Figure 2.4, ε is the vertical distance between the actual position of a household and the point on the regression line.

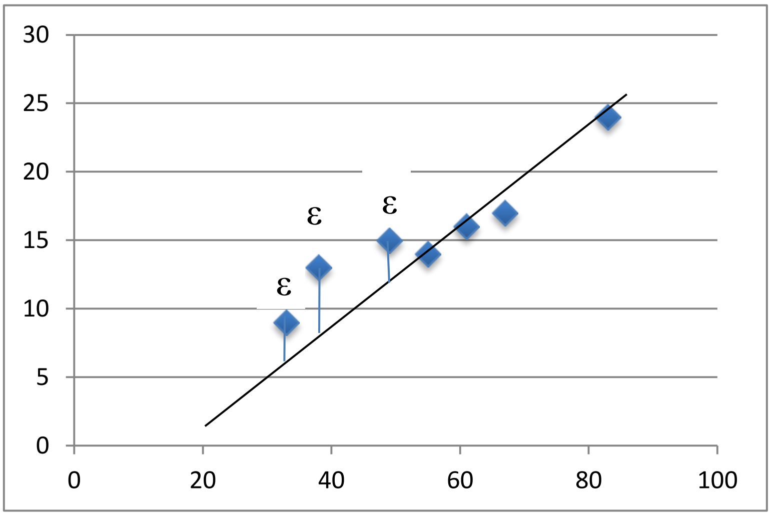


Figure 2.4: The best fit line for the relationship between food expenditure and income

The value of an error is positive if the point that gives the actual food expenditure is above the regression line and negative if it is below the regression line. The sum of these errors is always zero. In other words, the sum of the actual food expenditures for the seven households included in the sample will be the same as the sum of food expenditures predicted from the regression model. Thus,

$$\Sigma\varepsilon_i = \Sigma(y_i - \hat{y}_i) = 0$$

Hence, to find the line that best fits the scatter of points, we can’t minimize the sum of errors. Instead, we minimize **the sum of squares of the errors**, (denote it as **SSE**), which is obtained by adding the squares of errors. Thus,

$$SSE= \Sigma\varepsilon_i^2= \Sigma(y_i - \hat{y}_i)^2$$

Now we can define the **Least Square Method** to be the best line that minimizes the sum of squared vertical differences between the points and the line. That is, to find estimated values for the coefficients β_0, β_1, β_2, …, β_k, that give the minimum SSE [21].

Now, after this overview, we will show how to derive the least squares regression formulas to estimate various regression models. But note that the formulas that we will use are for estimating a sample regression line. Suppose we have access to a population data set. We can find the population regression line by using the same formulas with a little adaptation. We replace the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_k$ (which are for the sample) by other coefficients, say β_0, β_1, β_2, …, β_k (to be coefficients for the population), and the sample size n by N, the population size. Then the population regression equation is written as:

$$\mu_{y|x} = \beta_0 +\beta_1 x_{i1} + \beta_2 x_{i2} +\ldots+ \beta_k x_{ik} + \varepsilon_i$$

where $\mu_{y|x}$ is read as “the mean value of y for a given x”. When plotted on a graph, the points on this population regression line give the average value of y for the corresponding values of x. These average values of y are denoted by $\mu_{y|x}$.

2.4 Simple Linear Regression Model

We begin with the simplest situation, the simple linear regression model which involves only one dependent variable (y) and one independent variable(x) and states that the true mean of the dependent variable changes at a constant rate as the value of the independent variable increases or decreases. The equation of the line relating y to x is called the simple linear regression equation.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where,

y: the dependent (response) variable.

x: the independent (explanatory) variable.

β_0: the y-intercept, the value of y when x = 0.

β_1: the slope, the expected change in y relative to one unit increase in x.

ε: is the random error.

The estimate of the simple linear regression equation is given by substituting the least squares estimates into equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{y}$ is the expected value of y for a given value of x.

The simple linear regression model has two coefficients β_0 and β_1, which are to be estimated from the data, that is we want to find $\hat{\beta}_0$ and $\hat{\beta}_1$. If there was no random error in y_i, any two data points (x_i, y_i) could be used to solve explicitly for the values of the parameters. The random variation in y, however, causes each pair of observed data points to give different results. (All estimates would be identical only if the observed data fell exactly on the straight line). The method of least squares will be used to combine all the information to give one solution which is “best” by some criterion.

The least squares estimation procedure uses the criterion that the best solution must give the smallest possible sum of squared deviations of the criterion observed y_i from the estimates of their true means provided by the solution. The best fit regression line is the line that minimizes the sum of errors.

We want to minimize

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2,$$

This is a quadratic expression and it reaches its minimum value when its derivatives vanish. So, by taking the derivative of SSE with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and setting them to zero gives the following set of equations (called the normal equations):

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = \frac{\partial\left(\sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2\right)}{\partial \hat{\beta}_0}$$

$$= 2 \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \frac{\partial(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))}{\partial \hat{\beta}_0}$$

$$= -2 \sum_{i=1}^{n} [y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

We set this partial derivative equal to zero:

$$-2\ \Sigma[y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 \qquad \Sigma[y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$$

$$\Sigma y_i = n\ \hat{\beta}_0 + \hat{\beta}_1\ \Sigma x_i$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = \frac{\partial \left(\sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 \right)}{\partial \hat{\beta}_1}$$

$$= 2 \sum_{i=1}^{n} [y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \frac{\partial (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))}{\partial \hat{\beta}_1}$$

$$= -2\ \Sigma x_i [y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)].$$

Set the partial derivative equal to zero:

$$-2\ \Sigma x_i [y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 \qquad \Sigma x_i [y_i – (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$$

$$\Sigma x_i y_i - \ \hat{\beta}_0\ \Sigma x_i\ - \hat{\beta}_1 \Sigma x_i^2 = 0$$

$$\Sigma x_i y_i = \hat{\beta}_0\ \Sigma x_i + \hat{\beta}_1 \Sigma x_i^2$$

Now we must solve this system of two normal equations:

$$\Sigma y_i = n\ \hat{\beta}_0 + \hat{\beta}_1\ \Sigma x_i$$

$$\Sigma x_i y_i = \hat{\beta}_0\ \Sigma x_i + \hat{\beta}_1 \Sigma x_i^2$$

The system can be solved to get: $\hat{\beta}_1 = \frac{n\ \Sigma x_i\ y_i - \Sigma x_i\ y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$

and

$$\hat{\beta}_0 = \frac{\Sigma x_i}{n} + \frac{1}{n} . \frac{\Sigma x_i^2\ \Sigma x_i - \Sigma x_i\ \Sigma x_i\ y_i}{n\ \Sigma x_i^2 - (\Sigma x_i)^2} = \frac{\Sigma x_i}{n} - \hat{\beta}_1 \frac{\Sigma x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \qquad [21]$$

We can rewrite $\hat{\beta}_1$ given the following information about what are called "sums of squares",

$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{1}{n} \Sigma x_i \Sigma y_i = \Sigma x_i y_i - n \bar{x} \bar{y}$

$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{1}{n} (\Sigma x_i)^2 = \Sigma x_i^2 - n \bar{x}^2$

$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{1}{n} (\Sigma y_i)^2 = \Sigma y_i^2 - n \bar{y}^2$

Therefore, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

So, the line that minimizes the sum of squared errors has the following slope and y-intercept estimated parameters:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

To illustrate the method, let us consider the following example.

Example (2.1):

Find the "best" regression line for the data on income and food expenditure on the seven households given in table 2.1. Use income as independent variable and food expenditure as dependent variable.

Solution:

We want to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Table 2.2 shows the calculations required for the computation of $\hat{\beta}_0$ and $\hat{\beta}_1$. We denote the independent variable (income) by x and the dependent variable (food expenditure) by y, both in hundreds of dollars.

Table 2.2: Data on Income and Food Expenditure

x	y	xy	x^2
55	14	770	3025
83	24	1992	6889
38	13	494	1444
61	16	976	3721
33	9	297	1089
49	15	735	2401
67	17	1139	4489
Σx= 386	Σy= 108	Σxy= 6403	Σx^2= 23058

- First, we find Σx, Σy, Σxy, Σx^2 as shown in the table. And compute $\bar{x}$ and $\bar{y}$:

$$\bar{x} = \Sigma x/n = 386/7 = 55.1429$$

$$\bar{y} = \Sigma y/n= 108/7 = 15.4286$$

- Compute Sxy and Sxx :

$$S_{xy} = \Sigma x_i\, y_i - \frac{1}{n}\,\Sigma x_i\, \Sigma y_i = 6403 - \frac{1}{7}\,.(386).\,(108) = 447.57$$

$$S_{xx} = \Sigma x_i^2 \;\; - \frac{1}{n}\,(\Sigma x_i)^2 \;\; = 23058 - \frac{1}{7}\;\;.(386)^2 \qquad = 1772.85$$

- Compute $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\text{Sxy}}{\text{Sxx}} = \frac{447.57}{1772.85} = 0.2525$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\, \bar{x} = 15.4286 - (0.2525)(55.1429) = 1.505$$

Thus, our estimated regression model is:

$$\hat{y} = 1.505 + 0.2525\ x$$

Using this estimated regression model, we can find the predicted value of y for any fixed value of x (during the month in which the data has been collected). For instance, suppose we randomly select a household whose monthly income is \$6100, so that x = 61. The predicted value of food expenditure for this household is:

$$\hat{y} = 1.505 + 0.2525\,(61) = \$16.9075 \text{ hundred} = \$1690.75$$

In other words, based on our regression line, we predict that a household with a monthly income of \$6100 is expected to spend \$1690.75 per month on food. In our data on seven households, there is a one household whose income is \$6100. The actual food expenditure for that household is \$1600 (see Table 2.2). The difference between the actual and predicted values gives the error of prediction ε.

$$\varepsilon = y - \hat{y} = 16 - 16.9075 = \$ \ {-0.9075} \text{ hundred} = \$\text{-}90.75$$

- $\hat{\beta}_0$ = 1.505, is the expected value of y when x=0. That is, a household with no income is expected to spend \$150.5 per month on food.
- The value of $\hat{\beta}_1$ in the regression model gives the change in y due to increase of one unit in x. That is, for every one dollar increase in income, a household food expenditure is predicted to increase by \$0.2525.
- We can find the correlation coefficient r which can be computed from the formula :

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{447.57}{\sqrt{(1772.85)(125.7143)}} = 0.94805$$

Goodness of Fit of the model

Once we fit a model, a natural question comes to our mind, how good is the fit? Are the explanatory variables useful in prediction? To answer these questions we need to assess the regression model, either using t-test or using ANOVA for regression. In general, the purpose of analysis of variance (ANOVA) is to test for significant differences between means. But in regression models it consists of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance.

It is easy to show that the total variation of the response variable y can be decomposed into two parts: the residual variation of y (error sum of squares (SSE)) and the explained variation of y (regression sum of squares (SSR)).

Consider the total sum of squares:

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$, which we usually rewrite as:

$$SST = SSE + SSR$$

- SST stands for the “total sum of squares”, this is essentially the total variation in the data set. That is, the total variation of food expenditure.
- SSR stands for “sum of squares due to regression” - this is the squared variation around the mean of the estimated food expenditure. This is sometimes called the total variation explained by the regression.

- SSE stands for “sum of squares due to error” - this is simply the sum of the squared residuals, and it is the variation in the y variable that remains unexplained after taking into account the variable x.

For the simple regression case, these are computed as:

SST = $S_{yy} = \Sigma y_i^2 - n\,\bar{y}^2$

SSR = $\hat{\beta}_1 S_{xy}$

SSE = SST – SSR

Each sum of squares can be divided by an appropriate constant (degrees of freedom) to get the mean sum of squares due to regression MSR, and the mean sum of squares due to error MSE.

It is often useful to summarize the decomposition of the variation in y in terms of an analysis of variance (ANOVA). In such a case the total explained and unexplained variations in y are converted into variances by dividing by the appropriate degrees of freedom. This helps develop a formal procedure to test the goodness of fit by the regression line.

Initially we set the null hypothesis that the fit is not good. In other words, our hypothesis is that the overall regression is not significant in a sense that the explanatory variable is not able to explain the response variable in a satisfactory way.

Table 2.3: ANOVA table for simple regression

Source of Variation (Source)	**Sum of Squares (SS)**	**Degrees of Freedom (df)**	**Mean Square (MS)**	**F statistic**
Regression	SSR	1	MSR= $\frac{\text{SSR}}{1}$	F = $\frac{\text{MSR}}{\text{MSE}}$
Error	SSE	n-2	MSE= $\frac{\text{SSE}}{\text{n}-2}$	
Total	SST	n-1		

From the ANOVA table we can easily conclude the overall regression is significant at the 5% level of significance, i.e., the OLS regression line adequately fits the data.

If the calculated value of the statistic falls in the critical region, we reject the null hypothesis and conclude that the regression coefficient is significant. In other words, we say that the explanatory variable has significant effect on the response variable. The critical region (or the rejection region) is determined by the value of F-tabulated, $F_{\alpha,1,n-2}$.

If the value of the statistic falls outside the critical region, we do not reject the null hypothesis and conclude that the regression coefficient is not significant, i.e., the explanatory variable has no significant effect on the response variable.

Now, back to our example on income and food expenditure, if we were asked to assess the usefulness of the model at the significance level 0.05 (α = 0.05)

- First, set the null hupothesis (H_0) and the alternative hypothesis (H_1):

$$H_0: \beta_1 = 0 \qquad \text{versus} \qquad H_1: \beta_1 \neq 0$$

- Compute the quantities SST, SSR, and SSE:

SST = S_{yy} = Σy_i^2 - n □² = 1792 – $\frac{(\mathbf{108})^{\mathbf{2}}}{\mathbf{7}}$ = 125.71

SSR = $\hat{\beta}_1$ S_{xy} = 0.2525 . (447.57) = 113.01

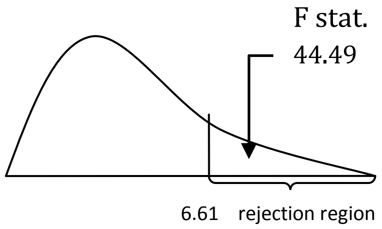
SSE = SST – SSR = 125.71 – 113.01 = 12.7

- Put the calculations in ANOVA table and compute the rested quantities:

Source	**SS**	**(df)**	**(MS)**	**F statistic**
Regression	113.01	1	113.01	44.49
Error	12.7	5	2.54	
Total	125.71	6		

The tabulate F value is:

$F_{\mathbf{0.05,1,5}}$ = 6.61



Since the F-statistic falls in the rejection region, we reject the null hypothesis. That is, the income is useful to explain the food expenditure in a satisfactory way.

- From the assumptions about the errors, that they have a constant variance σ^2, and since it is a population parameter, so we can't know for certain what its value is. Therefore, it is usually estimated by s^2 =

$\frac{\text{SSE}}{\text{n}-2}$, “MSE”, which is provided in the regression output under the name “standard error”.

We define the coefficient of determination r^2, which is the portion of the total variation in the dependent variable that is explained by variation in the independent variable (it is a measure of the explanatory power of the model). And it can be defined by the percentage of the response variable variation that is explained by a linear model,

$$r^2 = \frac{\text{SSR}}{\text{SST}} \ , 0 \le r^2 \le 1, \quad r^2 = \frac{113.01}{125.71} = 0.8989$$

In general, the higher the R-squared, the better the model fits your data.

2.5 Multiple Linear Regression Model

In this section we present one more complicated model, and develop the normal equations and solution to the normal equations for a more general linear model involving finite number of independent variables. We present multiple regression analysis in matrix notation. In this model we consider the relationship between one dependent variable and more than one independent variable.

The linear model for relating a dependent variable to k independent variables is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

The multiple linear regression can be thought of as an extension of simple linear regression, where there are k explanatory variables, or simple linear regression can be thought as a special case of multiple linear regression, where k = 1. Also here we use the least squares method to find the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize the sum of squares $\Sigma(y - \hat{y})^2$

The method is to write the following formulas,

- $S_{x1y} = \hat{\beta}_1 S_{x_1x_1} + \hat{\beta}_2 S_{x_1x_2} + \ldots + \hat{\beta}_k S_{x_1x_k}$
- $S_{x_2y} = \hat{\beta}_1 S_{x_1x_2} + \hat{\beta}_2 S_{x_2x_2} + \ldots + \hat{\beta}_k S_{x_2x_k}$

  .
  .
  .
- $S_{x_ky} = \hat{\beta}_1 S_{x_1x_k} + \hat{\beta}_2 S_{x_2x_k} + \ldots + \hat{\beta}_k S_{x_k x_k}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_{11} - \hat{\beta}_2 \bar{\bar{x}}_2 - \ldots - \hat{\beta}_k x\square_k$

These equations are called normal equations, so we can solve these equations to find the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Note that, because the normal equations are linear, and because there are as many equations as unknown regression coefficients (k+1), there is usually unique solution for the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ [6].

To illustrate how the method of least squares work in this model, we will consider a very simple example of multiple regression model in the case of just two explanatory variables, that is the formula of the model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, and through this example we will see how to compute the regression coefficients by doing similar steps like in simple regression in a developed manner.

Example (2.2):

A researcher is interested in predicting the average value of y, length of time that an individual can continue a physical exercise, on the basis of two predicting variables, x_1, the average number of cigarettes smoked per day, and x_2, the ratio of weight in kilograms to height in meters. The following data summary for 20 individuals:

$\Sigma y = 360$, $\Sigma x_1 = 200$, $\Sigma x_2 = 900$, $\Sigma y^2 = 7162$, $\Sigma x_1^2 = 3398$ $\Sigma x_2^2 = 41058$,

$\Sigma x_1 y = 2669$, $\Sigma x_2 y = 16034$, $\Sigma x_1 x_2 = 9298$.

Obtain the multiple regression equation.

Solution:

- $\bar{y} = 18$, $\bar{x}_1 = 10$, $\bar{x}_2 = 45$.
- Compute S_{x_1y}, S_{x_2y}, $S_{x_1x_1}$, $S_{x_1x_2}$, $S_{x_2x_2}$

$$S_{x_1y} = \Sigma x_1 y - n\,\bar{x}_1\bar{y} = 2669 - (20)(10)(18) = -931$$

$$S_{x_2y} = \Sigma x_2 y - n\,\bar{x}_2\bar{y} = 16034 - (20)(45)(18) = -166$$

$$S_{x_1x_1} = \Sigma x_1^2 - n\,\bar{x}_1^2 = 3398 - (20)(10)^2 = 1398$$

$$S_{x_1x_2} = \Sigma x_1 x_2 - n\,\bar{x}_1\bar{x}_2 = 9298 - (20)(10)(45) = 298$$

$$S_{x_2x_2} = \Sigma x_2^2 - n\,\bar{x}_2^2 = 41058 - (20)(45)^2 = 558$$

- Set the normal equations to find the estimated coefficients

$$S_{x_1y} = \hat{\beta}_1 S_{x_1x_1} + \hat{\beta}_2 S_{x_1x_2}$$

$$S_{x_2y} = \hat{\beta}_1 S_{x_1x_2} + \hat{\beta}_2 S_{x_2x_2}$$

$$-931 = 1398\,\hat{\beta}_1 + 298\,\hat{\beta}_2$$

$$-166 = 298\,\hat{\beta}_1 + 558\,\hat{\beta}_2$$

these are two equations in two unknowns, solving these equations we get:

$\hat{\beta}_1 = -0.68$ and $\hat{\beta}_2 = 0.066$

And $\hat{\beta}_0 = 18 - (-0.68)(10) - (0.066)(45) = 21.83$

The multiple regression equation is given by

$$\hat{y} = 21.83 - 0.68\,x_1 + 0.066\,x_2$$

Using this estimated regression model, we can find the predicted value of y for specific values of x_1 or x_2. For instance, suppose we randomly select a person who smokes seven cigarettes per day, so that $x_1= 7$, and that this person weighted 70 kg and is 1.7 meter tall, so that $x_2= 41.176$. The predicted value of time to continue a physical exercise for this person is:

$\hat{y} = 21.83 - (0.68)(7) + (0.066)(41.176) = 19.787$ unit of time (minute, say)

- The value of $\hat{\beta}_1$ indicates that for every one more cigarette smoked in a day, the length of time for a person to continue an exercise is expected to decrease by 0.68 minute. Keeping the ratio of weight to height fixed.

- The value of $\hat{\beta}_2$ indicates that if the ratio of weight to height for a person increases by one unit, the length of time for a person to continue an exercise is expected to increase by 0.066 minute. Keeping the number of cigarette smoked in a day fixed.

Goodness of Fit of the model

As in simple regression we use ANOVA for regression to test the goodness of the multiple regression models.

Our hypothesis is that the overall regression is not significant in a sense that the explanatory variables are not able to explain the response variable in a satisfactory way.

$SST = S_{yy} = \Sigma y_i^2 - n\,\bar{y}^2$

$SSR = \hat{\beta}_1 S_{x_1y} + \hat{\beta}_2 S_{x_2y}$

$SSE = SST – SSR$ [6]

Table 2.4: ANOVA table for multiple regression

Source of Variation (Source)	**Sum of Squares (SS)**	**Degrees of Freedom (df)**	**Mean Square (MS)**	**F statistic**
Regression	SSR	K	$MSR=\frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-k-1	$MSE=\frac{SSE}{n-k-1}$	
Total	SST	n-1		

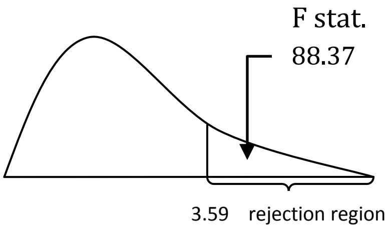
At 5% level of significance, the rejection region is determined by $F_{0.05,k,n-k-1}$.

we want to test

H_0:$\beta_1 = \beta_2 = 0$ versus H_1: not all β_i's equal to zero

Source	SS	Df	MS	F statistic
Regression	622.124	2	311.062	88.37
Error	59.876	17	3.52	
Total	682	19		

$F_{0.05,2,17} = 3.59$



Since the F-statistic falls in the rejection region, we reject the null hypothesis. That is, both cigarettes and ratio of weight to height are useful for predicting the length of time an individual need to continue an exercise in a satisfactory way.

- $r^2 = \frac{SSR}{SST} = \frac{622.124}{682} = 0.9122$, the value of r^2 is close to 1, this means that the explanatory power of the model is very good.
- The correlation coefficient $r = \sqrt{r^2} = 0.955$, this indicates that the variables can be considered very highly correlated.

Multiple Regression Model in Matrix Notation:

Here we begin by representing the linear model in matrix form. For the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \text{ for all } i= 1,2,\ldots,n$$

where the error terms assumed to have the following properties:

- $E(\varepsilon_i) = 0$.
- $Var\ (\varepsilon_i) = \sigma^2$ (constant).
- $Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$

The subscript i denotes the observational unit from which the observations on y and the k independent variables were taken. The second subscript designates the independent variable. The sample size is denoted with n, i = 1, . . . , n, and k denotes the number of independent variables. There are (k + 1) estimated coefficients $\hat{\beta}_j$, j = 0, . . . , k when the linear model includes the estimated intercept $\hat{\beta}_0$.

Four matrices are needed to express the linear model in matrix notation:

Y : the n×1 column vector of observations on the dependent variable y.

X: the n × (k+1) matrix consisting of a column of ones, which is labeled 1, followed by the k column vectors of the observations on the independent variables.

β: the (k+1) × 1 vector of parameters to be estimated.

ε: the n × 1 vector of random errors.

With these definitions, the linear model can be written as

$$Y = X\beta + \varepsilon$$

or :

$$\underset{(n \times 1)}{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{pmatrix}} = \underset{(n \times (k+1))}{\begin{pmatrix} 1 & x_{11} & x_{12} & . \; . \; . & x_{1k} \\ 1 & x_{21} & x_{22} & . \; . \; . & x_{2k} \\ 1 & x_{31} & x_{32} & . \; . \; . & x_{3k} \\ . & . & & & . \\ . & & & . \; . \; . \; . & . \\ 1 & x_{n1} & x_{n2} & . \; . \; . & x_{nk} \end{pmatrix}} \underset{((k+1) \times 1)}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix}} + \underset{(n \times 1)}{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_n \end{pmatrix}}$$

The assumptions become:

- $E(\varepsilon) = 0$
- $\mathrm{Cov}(\varepsilon) = E(\varepsilon\, \varepsilon^T) = \sigma^2 I$.

The elements of a particular row of X, say row i, are the coefficients on the corresponding parameters in β. Notice that β_0 has the constant multiplier 1 for all observations; hence, the column vector 1 is the first column of X.

The normal equations and their solution

For the least square estimation our main objective is to find a vector of parameters which minimizes the error sum of squares

$SSE = \Sigma\, \varepsilon_i^2 = \varepsilon^T \varepsilon$

In matrix notation, the normal equations are written as:

$$X^T X\, \hat{\beta} = X^T Y$$

The normal equations are always consistent and hence will always have a solution of the form

$$\hat{\beta} = (X^{T}X)^{-1} X^{T}Y$$

If $X^{T}X$ has an inverse, then the normal equations have a unique solution given by

$$\hat{\beta} = (X^{T}X)^{-1} X^{T}Y$$

here $\hat{Y} = X\hat{\beta}$ the predicted value of the response variable (in matrix form), and $\varepsilon = Y - \hat{Y}$ [21].

Example (2.3):

The data in table 2.5 relate grams plant dry weight ,y, to percent soil organic matter, x_1, and kilograms of supplemental soil nitrogen added per 1000 square meters, x_2. Obtain the multiple regression equation.

Table 2.5

y	x_1	x_2
78.5	7	2.6
74.3	1	2.9
104.3	11	5.6
87.6	11	3.1
95.9	7	5.2
109.2	11	5.5
102.7	3	7.1

Solution: We want to find the estimated coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for the regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \text{ using matrices}$$

- First, write the variables in matrix form

$$Y = \begin{pmatrix} 78.5 \\ 74.3 \\ 104.3 \\ 87.6 \\ 95.9 \\ 109.2 \\ 102.7 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 7 & 2.6 \\ 1 & 1 & 2.9 \\ 1 & 11 & 5.6 \\ 1 & 11 & 3.1 \\ 1 & 7 & 5.2 \\ 1 & 11 & 5.5 \\ 1 & 3 & 7.1 \end{pmatrix}$$

- Find the matrix X^TX

$$X^TX = \begin{pmatrix} 7 & 51 & 32 \\ 51 & 471 & 235 \\ 32 & 235 & 163.84 \end{pmatrix}$$

- Find the inverse of X^TX

$$(X^TX)^{-1} = \begin{pmatrix} 1.7996 & -0.0685 & -0.2532 \\ -0.0685 & 0.0101 & -0.0011 \\ -0.2532 & -0.0011 & 0.0571 \end{pmatrix}$$

- Find X^TY

$$X^TY = \begin{pmatrix} 652.5 \\ 4915.3 \\ 3103.7 \end{pmatrix}$$

$$\text{- Now, } \hat{\beta} = (X^TX)^{-1} X^TY = \begin{pmatrix} 1.7996 & -0.0685 & -0.2532 \\ -0.0685 & 0.0101 & -0.0011 \\ -0.2532 & -0.0011 & 0.0571 \end{pmatrix} \begin{pmatrix} 652.5 \\ 4915.3 \\ 3103.7 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 51.6 \\ 1.5 \\ 6.72 \end{pmatrix}$$

We get the multiple regression equation:

$$\hat{y} = 51.6 + 1.5\, x_1 + 6.72\, x_2$$

From the regression equation, for every additional one percent soil organic matter, the plant dry weight is expected to increase by approximately 1.5 gram, keeping the number of supplemental soil nitrogen fixed. Also, for every additional one kilogram supplemental soil nitrogen added per 1000 square meter, the dry weight is expected to increase by approximately 6.72 gram, keeping the percent soil organic matter fixed.

We can find the predicted value for each observation of the response variable by computing the matrix $\hat{Y} = X\hat{\beta}$, and the matrix of errors $\varepsilon = Y - \hat{Y}$, as shown,

$$\hat{Y} = \begin{pmatrix} 79.5320 \\ 72.5645 \\ 105.6914 \\ 88.8833 \\ 97.0125 \\ 105.0191 \\ 103.7970 \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} -1.0320 \\ 1.7355 \\ -1.3914 \\ -1.2833 \\ -1.1125 \\ 4.1809 \\ -1.0970 \end{pmatrix}$$

2.6 Estimating regression models using software

Fitting linear regression models is very important and widely used. When we deal with simple regression with just two variables, it was easy. Then we access to multiple regression models, and in chapter three we will introduce more complicated regression models. We studied slight examples in the case of two explanatory variables and one response spending a lot of time in doing calculations. So, what if we want to study the relationship between five independent variables and one response each of 30 observations? How much time will we need to fit the model?

Many statistical softwares have been developed in order to do statistical analysis to save both effort and time. These can be used in fitting regression models, like SPSS, Minitab, SAS, Stata and Matlab. In this section we show Matlab and Minitab, ans SAS output for the examples done by hand before and see the results. In fact, Matlab is not a statistical software but it can fit regression models even if you have as many variables (dependent and independent) using just few commands, and provides you with accurate results [5].

Example (2.1')

In this example we will solve example (2.1), the relationship between food expenditure and income using Matlab and SAS.

Matlab output:

```
Linear regression model:     y ~ 1 + x1

Estimated Coefficients:

                  Estimate      SE          tStat       pValue

                  ________    ________    _______     _________

   (Intercept)     1.5073      2.1742     0.69327      0.51902

       x1          0.25246     0.037883   6.6641       0.0011485

Number of observations: 7, Error degrees of freedom: 5

Root Mean Squared Error: 1.6

R-squared: 0.899,  Adjusted R-Squared 0.879

F-statistic vs. constant model: 44.4, p-value = 0.00115
```

Explanation for the outputs:

- The first column of "Estimates" give the estimated values of regression coefficients, the intercept "$\hat{\beta}_0$" and the coefficient of x_1

"$\hat{\beta}_1$". $\hat{\beta}_0$ was found to be 1.505 and here it is 1.5073. And $\hat{\beta}_1$ was found to be 0.2525, here it is 0.25246.

- The second and third columns of "SE" and "tstat" give the standard error and the test statistic, respectively. Which are used in making inferences and test hypothesis about the estimated regression coefficients.
- The fourth column give the p-value for each coefficient, which is used in decision making when doing test hypothesis about the estimated regression parameters separately.

  In statistics, **the p-value** is a statistic that is used for testing a statistical hypothesis. Before performing the test a threshold value is chosen, called the significance level of the test, traditionally 5%. The p-value is also defined as the probability, under the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed. The rule is that we reject the null hypothesis if the p-value is less than or equal the significance level and not to reject it if the p-value is greater than the significant level.
- The Root Mean Squared Error (RMSE): which is the square root of the MSE, that is $\sqrt{\frac{\mathrm{SSE}}{\mathrm{df}}}$. RMSE represents the sample standard deviation of the differences between predicted values and observed values. That is, RMSE is the standard deviation of the variation of observations around the regression line.
- R-squared is the measure of the explanatory power of the model (or it is a statistical measure of how close the data are to the fitted

regression line), $r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}$, here it is 0.899 very close to the value computed by hand, it was 0.8989.

- Adjusted R-Squared: sometimes, introducing extra variables can lead to spurious results and can interfere with the proper estimation of slopes for the important variables. So, in order to penalize an excess of variables, we consider the adjusted R^2, which is computed by $r^2 = \frac{\mathrm{SSR}/\mathrm{k}}{\mathrm{SST}/\mathrm{n}-1}$. The adjusted r^2 thus divides numerator and denominator by their degrees of freedom.
- Finally, F-statistic: it is the statistic computed by dividing MSR by MSE (F= $\frac{\mathrm{MSR}}{\mathrm{MSE}}$). Rejecting or not rejecting the null hypothesis depends on the value of the F-statistic. Here it is 44.4 very close to the value computed before (44.49). The p-value = 0.00115 less than 0.05 so again, the null hypothesis is rejected which is the same decision we made when we test the goodness of the model before that the income was useful in predicting food expenditure.

SAS output:

The SAS System

Model: MODEL1

Dependent Variable: y

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	112.99285	112.99285	44.41	0.0011
Error	5	12.72143	2.54429		
Corrected Total	6	125.71429			

Root MSE	1.59508	**R-Square**	0.8988
Dependent Mean	15.42857	**Adj R-Sq**	0.8786
Coeff Var	10.33850		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	1.50733	2.17424	0.69	0.5190
X	1	0.25246	0.03788	6.66	0.0011

Example (2.3')

Now, we will solve example 2.3, the relationship between grams plant dry weight ,y, percent soil organic matter, x_1, and kilograms of supplemental soil nitrogen added per 1000 square meters, x_2. Based on the table 2.5 given before. We will use Minitab and SAS to find the estimated multiple regression equation.

Minitab output:

Regression Analysis: Y versus X1, X2

The regression equation is

Y = 51.6 + 1.50 X_1 + 6.72 X_2

Predictor	Coef	SE Coef	T	P
Constant	51.570	3.523	14.64	0.000
X1	1.4974	0.2636	5.68	0.005
X2	6.7233	0.6274	10.72	0.000

S = 2.62587 R-Sq = 97.4% R-Sq(adj) = 96.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1053.83	526.91	76.42	0.001
Residual Error	4	27.58	6.90		
Total	6	1081.41			

Notice that the p-value is 0.001 which is less than 0.05, i.e. the plant dry weight depends on the percent soil organic matter and kilograms of supplemental soil nitrogen added per 1000 square meters. And they are highly correlated. SAS out put:

The SAS System

Model: MODEL1

Dependent Variable: y

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	1053.82777	526.91388	76.42	0.0007
Error	4	27.58081	6.89520		
Corrected Total	6	1081.40857			

Root MSE	2.62587	**R-Square**	0.9745
Dependent Mean	93.21429	**Adj R-Sq**	0.9617
Coeff Var	2.81703		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	51.56970	3.52258	14.64	0.0001
x1	1	1.49741	0.26360	5.68	0.0047
x2	1	6.72326	0.62735	10.72	0.0004

Note that SAS gives more accurate results than Minitab.

Chapter Three

Multivariate Multiple Linear Regression Model

In chapter two we studied two types of regression models in which one response variable is affected by one predictor variable or a set of predictor variables. These two regression models were called simple and multiple regression models. Sometimes the model or the relationship among variables may be more complicated. In such cases, more than one response variable can be affected by one predictor or by the same set of predictors. For example, if we want to study the relationship between eating habits and playing sports and how they affect some other variables, such as blood pressure, cholesterol, and weight. In this example, we may let eating habits and playing sports be the explanatory variables (independent variables) and let blood pressure, cholesterol, and weight be the responses (dependent variables). The regression model used to relate these variables in such cases is called multivariate multiple linear regression model. Before we introduce the multivariate multiple regression model we should take a look at the multivariate regression model.

3.1 Multivariate Linear Regression Model

In this section we present multivariate regression model, in which we consider the relationship between more than one dependent variable and one independent variable. This model is similar to the multiple regression model in solving the normal equations and estimate the regression parameters. These parameters are easily estimated using matrix form.

Suppose that the number of response variables is m, so we have n observations for each y_i, i = 1,2,…,m. The general formula for the multivariate regression model is given by:

$$y_i = \beta_{0i} + \beta_{1i}\,x_1 + \varepsilon_i$$

$$\hat{y}_i = \hat{\beta}_{0i} + \hat{\beta}_{1i}\,x_1\ , \quad i = 1,2,\dots, m$$

There are two parameters for each response to be estimated when the linear model includes the intercept β_0.

Four matrices are needed to express the linear model in matrix notation:

Y : the n×m matrix of observations on the dependent variable y .

X: the n × 2 matrix consisting of a column of ones, which is labeled 1, followed by the column vector of the observations on the independent variable.

β: the 2 × m matrix of parameters to be estimated.

ε: the n × m matrix of random errors.

$$\underset{(n\ x\ m)}{\begin{pmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ y_{31} & y_{32} & \dots & y_{3m} \\ . & . & \dots & . \\ . & . & \dots & . \\ y_{n1} & y_{n2} & \dots & y_{nm} \end{pmatrix}} = \underset{(n\ x\ 2)}{\begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \\ . & . \\ . & . \\ 1 & x_{n1} \end{pmatrix}} \underset{(2\ x\ m)}{\begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \end{pmatrix}} + \varepsilon_{\ (nxm)}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \varepsilon_{31} & \varepsilon_{32} & \cdots & \varepsilon_{3m} \\ . & \cdots & & . \\ . & . & \cdots & . \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{pmatrix}_{(nxm)}$$

If X^TX has an inverse, then the normal equations have a unique solution given by

$$\hat{\beta} =(X^TX)^{-1} X^TY$$

Example (3.1)

Given the following data in table 3.1, find the multivariate regression equations.

Table 3.1

x_1	y_1	y_2
0	1	-1
1	4	-1
2	3	2
3	8	3
4	9	2

Solution:

- The regression equations takes the form:

$$\hat{y}_1= \hat{\beta}_{01} + \hat{\beta}_{11}x_1$$

$\hat{y}_2 = \hat{\beta}_{02} + \hat{\beta}_{12}\, x_1$

- Write the data in matrix form

$$Y = \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \qquad X^T X = \begin{pmatrix} 5 & 10 \\ 10 & 30 \end{pmatrix}$$

$$(X^T X)^{-1} = \begin{pmatrix} 0.6 & -0.2 \\ -0.2 & 0.1 \end{pmatrix}, \quad \text{and} \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\beta} = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$$

Thus, the estimated multivariate linear regression equations:

$\hat{y}_1 = 1 + 2\, x_1$, and $\hat{y}_2 = -1 + x_1$

The following is Stata output for the estimated multivariate regression model of example 3.1 by just using the command

```
. mvreg y1 y2 = c.x

Equation        Obs   Parms        RMSE          "R-sq"          F            P
------------------------------------------------------------------------------------------------
y1               5       2        1.414214        0.8696         20          0.0208
y2               5       2        1.154701        0.7143         7.5         0.0714
------------------------------------------------------------------------------------------------
```

```
|   Coef.       Std. Err.       t         P>|t|            [95% Conf. Interval]
-------------+------------------------------------------------------------------------------------
y1   |
x1 |     2        .4472136       4.47       0.021          .5767667       3.423233
_cons |     1        1.095445       0.91       0.429          -2.486195      4.486195
-------------+-------------------------------------------------------------------------------------
y2   |
x1 |      1        .3651484       2.74       0.071          -.1620651      2.162065
_cons |    -1        .8944272      -1.12       0.345          -3.846467      1.846467
------------------------------------------------------------------------------------------------------
```

The estimated regression equations are given by:

$\hat{y}_1 = 1 + 2x_1$

$\hat{y}_2 = -1 + x_1$ which are exactly the equations found before.

According to the p- values, we can say that y_1 is affected by x_1, but y_2 is not affected by x_1.

3.2 Multivariate Multiple Linear Regression Model

In multivariate multiple linear regression model we consider the relationship between more than one dependent variable and more than one independent variable. That is, we extend the regression model to the situation where we have measured m responses $y_1, y_2, \ldots, y_m$ and the same set of k predictors $x_1, x_2, \ldots, x_k$ is used in a sample of size n, then each response variable follows its own regression model:

$$y_1 = \beta_{01} + \beta_{11} x_1 + \beta_{21} x_2 + \ldots + \beta_{k1} x_k + \varepsilon_1$$

$$y_2 = \beta_{02} + \beta_{12} x_1 + \beta_{22} x_2 + \ldots + \beta_{k2} x_k + \varepsilon_2$$

.

.

$$y_m = \beta_{0m} + \beta_{1m} x_1 + \beta_{2m} x_2 + \ldots + \beta_{km} x_k + \varepsilon_m$$

The error term $\varepsilon^T = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m]$ has $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \Sigma$. Thus the error terms associated with different responses may be correlated [17].

β_{ij}, i= 1,2,…,k and j= 1,2,…,m, is the estimated regression coefficient of the jth response in the effect of the ith predictor. β_{0j} is the intercept parameter of the jth response.

To represent the model in matrix form, we need four types of matrices:

Y: the n × m matrix consisting of m column vectors of the observations on each of the dependent variables;

X: the n × (k+1) matrix consisting of a column of ones followed by the k column vectors of the observations on the independent variables;

β: the (k+1)×m matrix consisting of column vectors of parameters to be estimated.

ε: the n × m matrix consisting of column vectors of random errors.

The linear model can be written as:

$$Y_{(n\times m)} = X_{(n\times (k+1))}\beta_{((k+1)\times m)} + \varepsilon_{(n\times m)}$$

with $E(\varepsilon_j) = 0$ and $\text{cov}(\varepsilon_j, \varepsilon_k) = \sigma_{jk}.I$; $j,k = 1,2,\ldots,m$

Note also that the m observed responses on the j^{th} trial have covariance matrix $\Sigma = \{ \sigma_{jk} \}$ but observations from different trials are uncorrelated. Here β and σ_{jk} are unknown parameters.

The matrix of responses,

$$Y_{n\times m} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ . & . & & . \\ . & & \cdots & . \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix} = \left[y_{(1)} \quad y_{(2)} \quad \cdots \quad y_{(m)} \right]$$

where $y_{(j)}$ is the column vector of n measurements of the j^{th} variable, j= 1,2,...,m. That is,

$$y_{(j)} = \left[y_{ij} \right] \quad \text{for } i= 1,2,\ldots,n$$

The design matrix X:

$$X_{(n\times(k+1))} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ . & . & & & . \\ . & & . & \cdots & . \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

Note that the rows of X correspond to observations, the columns to independent variables.

$$\beta_{((k+1) \times m)} = \begin{pmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \ldots & \beta_{1m} \\ . & . & & . \\ . & & \ldots & . \\ \beta_{k1} & \beta_{k2} & \ldots & \beta_{km} \end{pmatrix} = \begin{pmatrix} \beta_{(1)} & \beta_{(2)} & \ldots & \beta_{(m)} \end{pmatrix}$$

where $\beta_{(j)}$ are the (k+1) regression coefficients in the model for the j^{th} variable, j= 1,2,…,m. That is

$$\beta_{(j)} = \begin{pmatrix} \beta_{ij} \end{pmatrix} \quad \text{for } i= 0,1,2,\ldots,k.$$

$$\varepsilon_{(n \times m)} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \ldots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \ldots & \varepsilon_{2m} \\ . & . & & . \\ . & & \ldots & . \\ \varepsilon_{n1} & \varepsilon_{n2} & \ldots & \varepsilon_{nm} \end{pmatrix} = \begin{pmatrix} \varepsilon_{(1)} & \varepsilon_{(2)} & \ldots & \varepsilon_{(m)} \end{pmatrix}$$

where each $\varepsilon_{(j)}$ vector represents the residuals for each of the m response variables. That is,

$$\varepsilon_{(j)} = \begin{pmatrix} \varepsilon_{ij} \end{pmatrix}$$

Also the m observed responses on the j^{th} trial have covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2m} \\ . & . & & . \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_{mm} \end{pmatrix}$$

Because the covariance of the i^{th} random variable with itself is simply the random variable's variance, so, each element on the principal diagonal of the covariance matrix is the variance of one of the random variables. Also, because the covariance of the i^{th} random variable with the j^{th} one is the same as the covariance of the j^{th} random variable with the i^{th} one, every covariance matrix is symmetric. In addition, every covariance matrix is positive semi-definite. That is, for every non-zero column vector z of n real numbers, $z^T\Sigma\, z \geq 0$ [2].

The strategy in the least squares is the same as in the simple and multiple linear regression models. First, we calculate the sum of squared residuals and, second, find a set of estimators that minimize the sum

$$SSE= \Sigma\varepsilon_i{}^2 = \varepsilon^T\varepsilon$$

By solving the normal equation

$$X^TX\,\hat{\beta} = X^TY$$

we get the solution in the form

$$\hat{\beta} = (X^TX)^{-1} X^TY \qquad [17]$$

Using the least squares estimator for β, we can obtain predicted values:

$$\hat{Y} = X\hat{\beta}$$

Note that $\hat{\beta}$ is unbiased estimator for β, i.e. $E(\hat{\beta}) = \beta$

$$E(\hat{\beta}) = E((X^TX)^{-1} X^TY) = (X^TX)^{-1} X^T E(X\, \beta + \varepsilon) = (X^TX)^{-1} X^TX\, \beta = I.\, \beta = \beta$$

One might ask what the advantage is of doing all m regressions at once rather than doing m separate ones. The main reason is to gather strength from having several variables. For example, suppose one has an analysis of variance comparing drugs on a number of health-related variables. It may be that no single variable shows significant differences between drugs, but the variables together show strong differences. Using the overall model can also help deal with multiple comparisons, e.g., when one has many variables, there is a good chance at least one shows significance even when there is nothing going on.

To illustrate the estimation of multivariate multiple regression using the method of least squares, let us consider the following example.

Example 3.2

Suppose we had the following six sample observations, as shown in Table 3.2, on two independent variables (palatability and texture) and two dependent variables (purchase intent and overall quality) for some product.

Table 3.2 Data on six sample observation

Palatability (x_1)	Texture (x_2)	Overall quality (y_1)	Purchase intent (y_2)
65	71	63	67
72	77	70	70
77	73	72	70
68	78	75	72
81	76	89	88
73	87	76	77

Use the data to estimate the multivariate multiple linear regression model.

Solution:

- We have two dependent variables and two independent variables, so we had to find the estimated coefficients $\hat{\beta}_{01}$, $\hat{\beta}_{02}$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, $\hat{\beta}_{21}$ and $\hat{\beta}_{22}$ for the regression model:

  $\hat{y}_1 = \hat{\beta}_{01} + \hat{\beta}_{11}\, x_1 + \hat{\beta}_{21}\, x_2$

  $\hat{y}_2 = \hat{\beta}_{02} + \hat{\beta}_{12}\, x_1 + \hat{\beta}_{22}\, x_2$

- First, we have to write the data in matrix form:

$$Y_{(6\times2)} = \begin{pmatrix} 63 & 67 \\ 70 & 70 \\ 72 & 70 \\ 75 & 72 \\ 89 & 88 \\ 76 & 77 \end{pmatrix}, \quad X_{(6\times3)} = \begin{pmatrix} 1 & 65 & 71 \\ 1 & 72 & 77 \\ 1 & 77 & 73 \\ 1 & 68 & 78 \\ 1 & 81 & 76 \\ 1 & 73 & 87 \end{pmatrix}$$

- Now, we find the (3x3) X^TX matrix: (using matlab)

$$X^TX = \begin{pmatrix} 6 & 436 & 462 \\ 436 & 31852 & 33591 \\ 462 & 33591 & 35728 \end{pmatrix}$$

- Find the inverse of the matrix X^TX

$$(X^TX)^{-1} = \begin{pmatrix} 62.5606 & -0.3783 & -0.4533 \\ -0.3783 & 0.0060 & -0.0007 \\ -0.4533 & -0.0007 & 0.0066 \end{pmatrix}$$

- The estimated regression coefficients is given by:

  $\hat{\beta} = (X^TX)^{-1} X^TY$

$$\hat{\beta}_{(3\times2)} = \begin{pmatrix} -37.5012 & -21.4323 \\ 1.1346 & 0.9409 \\ 0.3795 & 0.3514 \end{pmatrix}$$

- So now, we get the multivariate multiple regression equations:

$$\hat{y}_1 = -37.5012 + 1.1346\, x_1 + 0.3795\, x_2$$

and $$\hat{y}_2 = -21.4323 + 0.9409\, x_1 + 0.3514\, x_2$$

Using this estimated regression model, we can find the predicted values of $\hat{y}_1$ and $\hat{y}_2$ for specific values of x_1 and x_2. For example, suppose we randomly select a product which has a palatability 68 and a texture of 78, that is, $x_1 = 68$ and $x_2 = 78$. Then the predicted value of the overall quality for this product is:

$\hat{y}_1 = -37.5012 + 1.1346\,(68) + 0.3795\,(78) = 69.2526$

and the predicted value of the purchase intent is:

$\hat{y}_2 = -21.4323 + 0.9409\,(68) + 0.3514\,(78) = 69.9581$

- We can write the matrix of estimated values

$$\hat{Y} = X\hat{\beta} = \begin{pmatrix} 63.1912 & 64.6779 \\ 73.4103 & 73.3727 \\ 77.5652 & 76.6714 \\ 69.2514 & 69.9607 \\ 83.2420 & 81.4892 \\ 78.3399 & 77.8281 \end{pmatrix}$$

- Also, we can write the residuals matrix:

$$\varepsilon = Y - \hat{Y} = \begin{pmatrix} -0.1912 & 2.3221 \\ -3.4103 & -3.3727 \\ -5.5652 & -6.6714 \\ 5.7486 & 2.0393 \\ 5.7580 & 6.5108 \\ -2.3399 & -0.8281 \end{pmatrix}$$

Note that each column sums in the residuals matrix is approximately zero, which is agree with the theory. That is, the sum of errors is always equal to zero, and in the form of matrices, the sum of each column is zero.

- We can obtain the covariance matrix for the two observed responses by using Matlab

$$\Sigma = \begin{pmatrix} 74.1667 & 63.0000 \\ 63.0000 & 58.0000 \end{pmatrix}$$

The elements on the diagonal of the covariance matrix 74.1667, and 58 are the variances of y_1(overall quality) and y_2 (purchase intent), respectively. And the off diagonal value (63) is the covariance between y_1 and y_2.

3.3 Assessing Multivariate Multiple Models using ***MANOVA***

After fitting the multivariate multiple model we have to show how good the model is? In simple and multiple regression models we use ANOVA to test the goodness of the model. Here we will use MANOVA for assessing the multivariate multiple regression models.

The MANOVA or multivariate analysis of variance is a way to test the hypothesis that one or more independent variables, or factors, have an effect on a set of two or more dependent variables. We do MANOVA instead of a series of one-at-a-time ANOVAs.

Why MANOVA?

- Supposedly to reduce the experiment-wise level of Type I error (rejecting the null hypothesis when it is in fact true) . The so-called overall test or omnibus test protects against this inflated error probability only when the null hypothesis is true. If you follow up a significant multivariate test with a bunch of ANOVAs on the individual variables without adjusting the error rates for the individual tests, there's no "protection". (probability of any type I errors increases with number of variables)

- Another reason to do MANOVA: none of the individual ANOVAs may produce a significant main effect on the response variables, but in combination they might, which suggests that the variables are more meaningful taken together than considered separately.
- MANOVA takes into account the intercorrelations among the responses [15].

MANOVA is a generalization of ANOVA allowing multiple dependent variables. Where sums of squares appear in univariate analysis of variance, in multivariate analysis of variance certain positive-definite matrices appear. The diagonal entries are the same kinds of sums of squares that appear in univariate ANOVA. The off-diagonal entries are corresponding sums of products.

The multiplication X^TX generates a $(k+1)\times(k+1)$ matrix where the diagonal elements are sums of squares of each of the independent variables and the off-diagonal elements are the sums of product between independent variables. The general form of X^TX is

$$\begin{pmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} & \ldots & \Sigma x_{ik} \\ \Sigma x_{i1} & \Sigma x_{i1}^{2} & \Sigma x_{i1}x_{i2} & \ldots & \Sigma x_{i1}x_{ik} \\ \Sigma x_{i2} & \Sigma x_{i1}x_{i2} & \Sigma x_{i2}^{2} & \ldots & \Sigma x_{i2}x_{ik} \\ . & . & . & & . \\ . & . & . & & . \\ \Sigma x_{ik} & \Sigma x_{i1}x_{ik} & \Sigma x_{i2}x_{ik} & \ldots & \Sigma x_{ik}^{2} \end{pmatrix}$$

summation in all cases is over i =1 to n, the n observations of the data.

The elements of the matrix product X^TY are the sums of products between each independent variable and the dependent variables:

$$X^TY = \begin{pmatrix} \Sigma\, y_{i1} & \Sigma y_{i2} & \ldots & \Sigma y_{im} \\ \Sigma x_{i1}y_{i1} & \Sigma x_{i1}y_{i2} & \ldots & \Sigma\, x_{i1}y_{im} \\ . & . & & . \\ . & . & & . \\ \Sigma x_{ik}y_{i1} & \Sigma x_{ik}y_{i2} & \ldots & \Sigma x_{ik}y_{im} \end{pmatrix}$$

The first row is the sum of products between the vector of ones (the first column of X) and Y.

The null hypothesis in MANOVA is that the predictor variables (or sometimes we test for a particular predictor variables) do not influence the response variables. That is, all the coefficients $\hat{\beta}_j$, j = 1,2,…,m as column vectors are equal to zero.

In most of the statistical programs used to calculate MANOVAs there are four multivariate measures used for assessing the multivariate multiple regression model: Wilks' lambda, Pillai's trace, Lawley–Hotelling trace, and Roy's largest root [7].

Popular computer-package programs routinely calculate four multivariate test statistics such as Matlab, Minitab, Stata, and SAS. To understand these multivariate test statistics and how they are computed, we need to find two matrices, E, the m×m error sum of squares and cross product matrix

$$E = n\hat{\Sigma}$$

$$\text{where } \hat{\Sigma} = \frac{1}{n}\,\varepsilon^{T}\,\varepsilon = \frac{1}{n}\,(Y - X\hat{\beta})^{T}\,(Y - X\hat{\beta})$$

The second matrix is H, the m×m hypothesis sum of squares and cross product matrix

$$H = n\,(\hat{\Sigma}_1 - \hat{\Sigma})$$

where $\hat{\Sigma}_1$ varies according to the null hypothesis [17].

MANOVA is based on these two matrices or on the eigenvalues (λ_i) of the product matrix (HE^{-1}).

The first statistic is Wilks' lamda, it is computed using the formula:

$$\text{WiIks' lambda} = \prod_{i=1}^{m} \frac{1}{1+\lambda i} = \frac{|E|}{|H+E|}$$

The second statistic is Pillai's trace (the trace of an n×n-matrix is defined to be the sum of the elements on the main diagonal),

$$\text{PilIai's trace} = \sum_{i=1}^{m} \frac{\lambda i}{1+\lambda i} = \text{trace}\ [\ H(H + E)^{-1}\ ]$$

The third test statistic is Hotelling-Lawley's trace and it is given by the formula:

$$\text{Hotelling-Lawley trace} = \sum_{i=1}^{m} \lambda i = \text{trace}\ [\ HE^{-1}\ ]$$

The fourth statistic is Roy's largest root.

$$\text{Roy's greatest root} = \max_{i}\ (\lambda i)$$

Now, we will use the data in example 3.2 to compute the four multivariate test statistics to test the null hypothesis that the two predictor variables do not influence the responses. In other words, the responses do not depend on any one of the predictor variables x_1, x_2.

H_0: $\beta_1 = \beta_2 = 0$ versus H_1: at least one of β_j's not equal to zero

First, we need to find the matrices E and H

$$E= \varepsilon^T \varepsilon = (Y- X\hat{\beta})^T (Y- X\hat{\beta})$$

Under the null hypothesis, the design matrix X will be reduced to (6x1) matrix, the column vector of ones, call it X_{new}. So, we find the new matrix of estimated parameters $\hat{\beta}_{new}$ under H_0 using the formula

$$\hat{\beta}_{new} = (X_{new}{}^T X_{new})^{-1} X_{new}{}^T Y$$

$$\text{Then , } \hat{\Sigma}_1 = \frac{1}{n} (Y- X_{new}\, \hat{\beta}_{new}\ )^T (Y- X_{new}\, \hat{\beta}_{new}\ )$$

$$H= n\ (\hat{\Sigma}_1 - \hat{\Sigma})$$

$$H = \begin{pmatrix} 256.5203 & 215.6649 \\ 215.6649 & 181.4906 \end{pmatrix}, \quad E = \begin{pmatrix} 114.3130 & 99.3351 \\ 99.3351 & 108.5094 \end{pmatrix}$$

Now, we find the matrix HE^{-1}

$$HE^{-1} = \begin{pmatrix} 2.5277 & -0.3265 \\ 2.1183 & -0.2666 \end{pmatrix}$$

The eigenvalues of the matrix HE^{-1} are:

$\lambda_1 = 2.2533$, and $\lambda_2 = 0.0078$

So, the multivariate test statistics are:

WiIks'lambda = $\prod_{i=1}^{2} \frac{1}{1+\lambda_i} = \frac{|E|}{|H+E|} = 0.3050$

PilIai's trace = $\sum_{i=1}^{2} \frac{\lambda_i}{1+\lambda_i}$ = trace [$H(H+E)^{-1}$]= 0.7004

Hotelling-Lawley trace = $\sum_{i=1}^{2} \lambda_i$= trace [HE^{-1}] = 2.2611

And Roy's greatest root = $\max_i (\lambda_i) = 2.2533$

In the same way we can test that the responses do not depend on the variable x_1 that is:

H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$

Under the null hypothesis, the regression model become $\hat{Y} = X_2\hat{\beta}_2$, that is the second column of the design matrix will be deleted. The error sum of squares and cross product matrix E is the same one which was computed before, but the hypothesis sum of squares and cross product matrix will be

$$H= n (\hat{\Sigma}_1 - \hat{\Sigma})$$

$$\text{where } \hat{\Sigma}_1= \frac{1}{n} (Y- X_2 \hat{\beta}_2)^T (Y- X_2 \hat{\beta}_2)$$

$$\text{and} \quad \hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T Y$$

$$H = \begin{pmatrix} 214.9619 & 178.2623 \\ 178.2623 & 147.8282 \end{pmatrix}, \quad HE^{-1} = \begin{pmatrix} 2.2147 & -0.3846 \\ 1.8366 & -0.3189 \end{pmatrix}$$

The eigenvalues of the matrix HE^{-1} are:

$\lambda_1 = 1.8957$, and $\lambda_2 = 0.0000$

The multivariate test statistics are:

Wilks'lambda = $\prod_{i=1}^{2} \frac{1}{1+\lambda_i} = \frac{|E|}{|H+E|} = 0.3453$

PilIai's trace = $\sum_{i=1}^{2} \frac{\lambda_i}{1+\lambda_i}$ = trace [$H(H + E)^{-1}$]= 0.6547

Hotelling-Lawley trace = $\sum_{i=1}^{2} \lambda_i$= trace [HE^{-1}] = 1.8957

And Roy's greatest root = $\max_i (\lambda_i) = 1.8957$

Finally, under the null hypothesis that the responses do not depend on the variable x_2, the regression model become $\hat{Y}= X_1\hat{\beta}_1$, that is the third column of the design matrix will be deleted. The error sum of squares and cross product matrix E is the same one which was computed before, but the hypothesis sum of squares and cross product matrix will be

$$H= n\,(\hat{\Sigma}_1 - \hat{\Sigma})$$

$$\text{where}\quad \hat{\Sigma}_1= \frac{1}{n}\,(Y- X_1\,\hat{\beta}_1)^T\,(Y- X_1\,\hat{\beta}_1)$$

$$\text{and}\qquad \hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$$

$$H = \begin{pmatrix} 21.8720 & 20.2554 \\ 20.2554 & 18.7583 \end{pmatrix}, \qquad HE^{-1} = \begin{pmatrix} 0.1424 & 0.0563 \\ 0.1319 & 0.0521 \end{pmatrix}$$

The eigenvalues of the matrix HE^{-1} are:

$\lambda_1 = 0.1945$, and $\lambda_2 = 0.0000$

The multivariate test statistics are:

WiIks'lambda = $\prod_{i=1}^{2} \frac{1}{1+\lambda_i} = \frac{|E|}{|H+E|} = 0.8371$

PilIai's trace = $\sum_{i=1}^{2} \frac{\lambda_i}{1+\lambda_i}$ = trace [$H(H + E)^{-1}$]= 0.1629

Hotelling-Lawley trace = $\sum_{i=1}^{2} \lambda_i$= trace [HE^{-1}] = 0.1945

And Roy's greatest root = $\max_i\ (\lambda_i) = 0.1945$

The following is MANOVA output for the same example (3.2) using Stata

```
                    Number of obs =      6

              W = Wilks' lambda          L = Lawley-Hotelling trace

              P = Pillai's trace          R = Roy's largest root

      Source |  Statistic       df      F(df1,    df2) =     F           Prob>F

   -----------+------------------------------------------------------------------------

      Model | W  0.3050        2       4.0       4.0        0.81         0.5781 e

            | P   0.7004               4.0       6.0        0.81         0.5626 a
```

```
           | L    2.2611            4.0     2.0      0.57      0.7184 a
           | R   2.2533             2.0     3.0      3.38      0.1704 u
           |-------------------------------------------------------------------------
  Residual |                  3
-----------+--------------------------------------------------------------------------
        x1 | W   0.3453      1        2.0     2.0      1.90      0.3453 e
           | P    0.6547              2.0     2.0      1.90      0.3453 e
           | L    1.8957              2.0     2.0      1.90      0.3453 e
           | R   1.8957               2.0     2.0      1.90      0.3453 e
           |-------------------------------------------------------------------------
        x2 | W   0.8371      1        2.0     2.0      0.19      0.8371 e
           | P   0.1629               2.0     2.0      0.19      0.8371 e
           | L    0.1945              2.0     2.0      0.19      0.8371 e
           | R   0.1945               2.0     2.0      0.19      0.8371 e
           |-------------------------------------------------------------------------
  Residual |                  3
-----------+--------------------------------------------------------------------------
     Total |                  5
--------------------------------------------------------------------------------------
               e = exact, a = approximate, u = upper bound on F
```

Explanation for MANOVA output:

- Source : this indicates the predictor variable in question. In our model, we are looking for palatability, and texture (x_1, x_2, respectively).
- Statistic : this is the test statistic for the given source listed in the prior column and the multivariate statistic indicated with the letter

(W, P, L or R). For each independent variable, there are four multivariate test statistics calculated.

- df : this is the number degrees of freedom. Here, we have 2 predictors and our dataset has 6 observations, so we have 2 degrees of freedom for the hypothesis, 3 residual degrees of freedom, and 5 total degrees of freedom.
- F(df1, df2), F : the first two columns (df1 and df2) list the degrees of freedom used in determining the F statistics. The third column lists the F statistic for the given source and multivariate test.
- Prob > F : this is the p-value associated with the F statistic of a given effect and test statistic. The null hypothesis that a given predictor has no effect on either of the outcomes is evaluated with regard to this p-value. For a given alpha level, if the p-value is less than alpha (usually 0.05), the null hypothesis is rejected. If not, then we fail to reject the null hypothesis.
- e = exact, a = approximate, u = upper bound on F : this indicates how the F statistic was calculated (whether it was an exact calculation, an approximation, or an upper bound) for each of the multivariate tests.

The null hypothesis that we want to test in MANOVA is

H_0: $\beta_{(1)} = \beta_{(2)} = \ldots = \beta_{(m)} = 0$

versus the alternative hypothesis H_1 that at least one of the $\beta_{(j)}$ is different from zero.

where $\beta_{(j)}$ is the vector of regression coefficient for the j^{th} dependent variable for j= 1,2,…,m.

i-e H_0: $\beta_{(1)} = \beta_{(2)} = 0$ vs H_1: at least one of the $\beta_{(j)}$ is different from zero.

In our example, based on the last column of MANOVA output, we do not reject the null hypothesis because the p-values given are greater than the significance level (0.05). That is, palatability and texture have no effect on the overall quality and the purchase intent for a product.

3.4 Generating Data Using Simulation

Simulation is the imitation of the operation of a real-world process or system over time. It is used with scientific modeling of natural systems or human systems. Simulation is also used when the real system cannot be engaged, because it may not be accessible. Or it may be dangerous or unacceptable to engage. Or it is being designed but not yet built. Or it may simply not exist [24].

A computer simulation is an attempt to model a real-life or hypothetical situation on a computer so that it can be studied to see how the system works. By changing variables in the simulation, predictions may be made about the behavior of the system. It is a tool to virtually investigate the behavior of the system under study. Computer simulation has become a useful part of modeling many natural systems in physics, chemistry, biology, and human systems in economics and social.

Traditionally, the formal modeling of systems has been via a mathematical model, which attempts to find analytical solutions enabling the prediction of the behavior of the system from a set of parameters and initial conditions. There are many different types of computer simulation. The common feature they all share is the attempt to generate a sample of representative scenarios for a model in which a complete enumeration of all possible states would be prohibitive or impossible. Several software packages exist for running computer-based simulation such as Monte Carlo simulations, stochastic simulations, and multiscale simulations that make the modeling almost effortless [31].

The term simulation is used in different ways by different people. As we used here, simulation is defined as the process of creating a model of an existing or proposed system in order to identify and understand those factors which control the system and to predict (forecast) the future behavior of the system. Almost any system which can be quantitatively described using equations or rules can be simulated.

Simulation is a powerful and important tool because it provides a way in which alternative designs, plans and policies can be evaluated without having to experiment on a real system, which may be prohibitively costly, time-consuming, or simply impractical to do. That is, it allows you to ask "What if?" questions about a system without having to experiment on the actual system itself.

In this section, we will use simulation to generate data from different distributions such as normal distribution, logistic distribution and exponential distribution. Then the data will be used to fit multivariate multiple regression models. We are interested in generating data from various distributions using Matlab. Then, we will analyze the data and compare the results.

3.4.1 Simulation Using Normal Distribution

A hypothetical dataset of 20 high school students was drawn from a normal distribution using Matlab. The data was on three psychological variables (locus of control, self concept, and motivation) and four academic variables (standardized test scores in reading, writing, science, and art). We were interested in how the set of psychological variables is related to the academic variables.

The symbols :

y_1: Locus of control

y_2: self concept

y_3: motivation

x_1: read

x_2: write

x_3: science

x_4: art

(The simulated data is given in appendix I)

Stata software was used to get the estimated coefficients, the test statistic for each estimated parameter, standard errors, confidence intervals, the p-values and MANOVA output.

```
   Equation      Obs  Parms       RMSE      "R-sq"          F                 P

 -------------------------------------------------------------------------------------------

 locus_of_c~l     20      5         1.123921      0.1953       0.9100091         0.4831

 self_concept      20     5          0.7086149    0.3057       1.651137          0.2134

  motivation       20     5          0.7560022    0.3623      2.130425          0.1273

 -------------------------------------------------------------------------------------------

              |     Coef.         Std. Err.         t         P>|t|         [95% Conf. Interval]

-------------+------------------------------------------------------------------------------
locus_of_control

           x1 |   .0874673      .0680478       1.29      0.218        -.0575731      .2325077

           x2 |   .033673        .0710768       0.47      0.642        -.1178236      .1851696

          x3 |  -.0611048      .059823       -1.02      0.323        -.1886145      .0664048

          x4 |   .0459898      .1041279      0.44       0.665        -.1759537     .2679333

      _cons |   -8.032441     10.02414      -0.80       0.435         -29.3984      13.33352
```

```
-------------+------------------------------------------------------------------------------------

self_concept

        x1 |   -.0884473    .0429031     -2.06     0.057       -.1798931     .0029985

      x2 |   -.006916      .0448128     -0.15     0.879       -.1024323     .0886003

        x3 |     .0659879     .0377175       1.75     0.101        -.014405       .1463809

      x4 |     .0291023      .0656511       0.44     0.664        -.1108296     .1690343

   _cons |     .7964386      6.320071       0.13     0.901        -12.67447      14.26735

 ------------+------------------------------------------------------------------------------------

motivation

      x1 |     -.0325539     .0457722     -0.71     0.488        -.130115       .0650071

      x2 |     -.0766052     .0478096     -1.60     0.130        -.178509       .0252986

       x3 |     -.0706611       .0402398      -1.76      0.099         -.1564301     .015108

      x4 |      .0263419      .0700414       0.38     0.712        -.1229478     .1756316

  _cons |      11.86579       6.742714       1.76     0.099        -2.505961      26.23755

---------------------------------------------------------------------------------------------------
```

From the results above, we can write the estimated multivariate multiple regression model as following:

$\hat{y}_1 = -8.032441 + 0.0874673\ x_1 + 0.033673\ x_2 - 0.0611048\ x_3 + 0.0459898\ x_4$

$\hat{y}_2$ = 0.7964386 - 0.0884473 x_1 – 0.006916 x_2+ 0.0659879 x_3+ 0.0291023 x_4

$\hat{y}_3$ =11.86579 - 0.0325539 x_1 - 0.0766052 x_2 - 0.0706611 x_3 +0.0263419 x_4

From this system of three equations we can see that:

- As the test score of read increases by one grade, the student's locus of control is expected to increase by 0.0875, keeping the other variables fixed.
- When the test score of read increases by one grade, the student's self concept is expected to decrease by 0.0884, keeping the other variables fixed.
- The student's motivation is expected to decrease by 0.0766 as the test score of write increases by one grade, keeping the other variables fixed

MANOVA Output : In Stata, MANOVA output includes four multivariate test statistics for each predictor variable. The four tests are listed above the output table. For each of the four test statistics, an F statistic and associated p-value are also displayed.

```
              W = Wilks' lambda      L = Lawley-Hotelling trace
                          P = Pillai's trace         R = Roy's largest root

     Source |     Statistic      df      F(df1,    df2) =      F        Prob>F

  -----------+-------------------------------------------------------------------------

     Model | W   0.3630      4      12.0      34.7      1.35      0.2369 a

           | P   0.7938             12.0      45.0      1.35      0.2259 a
           | L   1.3391             12.0      35.0      1.30      0.2611 a
           | R   0.9259              4.0      15.0      3.47      0.0338 u
           |-------------------------------------------------------------------------

   Residual |                 15

  -----------+-------------------------------------------------------------------------

        x1 | W   0.7421      1       3.0      13.0     1.51       0.2595 e

           | P   0.2579              3.0      13.0    1.51       0.2595 e
           | L   0.3475              3.0      13.0    1.51       0.2595 e
           | R   0.3475              3.0      13.0    1.51       0.2595 e
           |---------------------------------------------------------------------------
        x2 | W   0.8535      1       3.0      13.0     0.74       0.5450 e
           | P   0.1465              3.0      13.0     0.74       0.5450 e
           | L   0.1716              3.0      13.0     0.74       0.5450 e
           | R   0.1716              3.0      13.0     0.74       0.5450 e
           |---------------------------------------------------------------------------
        x3 | W   0.6226      1       3.0      13.0     2.63       0.0945 e
           | P   0.3774              3.0      13.0     2.63       0.0945 e
           | L   0.6062              3.0      13.0     2.63       0.0945 e
           | R  0.6062               3.0      13.0     2.63       0.0945 e
           |---------------------------------------------------------------------------
        x4 | W   0.9524      1       3.0      13.0     0.22       0.8833 e
           | P   0.0476              3.0      13.0     0.22       0.8833 e
           | L   0.0499              3.0      13.0     0.22       0.8833 e
           | R  0.0499               3.0      13.0     0.22       0.8833 e
           |---------------------------------------------------------------------------

  Residual |                 15

 -----------+---------------------------------------------------------------------------

     Total |                 19

 ----------------------------------------------------------------------------------------

                    e = exact, a = approximate, u = upper bound on F
```

We can test the null hypothesis that the coefficients for the variable x_1 (read) are equal to zero in all three equations.

(1) [locus_of_control]x1= 0

(2) [self_concept]x1 =0

(3) [motivatiion] x1 = 0

F(3, 15) = 1.74 Prob > F = 0.2022

Because the p-value is greater than the significance level 0.05, we don't reject the null hypothesis. The coefficients for the variable x_1 (read) are equal to zero in all three equations. The same test can be done for the other explanatories (write, science and art).

For the variable write,

(1) [locus_of_control]x2= 0

(2)[self_concept]x2 =0

(3)[motivatiion] x2 = 0

F(3, 15) = 0.86 Prob > F = 0.4842

Again the p-value is greater than the significance level 0.05, so we don't reject the null hypothesis. The coefficients for the variable x_2 (write) are equal to zero in all three equations.

Test for science

(1) [locus_of_control]x3 = 0

(2) [self_concept]x3 = 0

(3) [motivation] x3 = 0

F(3, 15) = 3.03 Prob > F = 0.0621

The result is to not reject the null hypothesis. That is, the coefficients for the variable x_3 (science) are equal to zero in all three equations.

Finally, for the variable art:

(1) [locus_of_control]x4 = 0

(2) [self_concept]x4 = 0

(3) [motivation]x4 = 0

F(3, 15) = 0.25 Prob > F = 0.8603

The p-value indicates that we don’t reject the null hypothesis. That is, the coefficients for the variable x_4 (art) are equal to zero in all three equations.

3.4.2 Simulation Using Logistic Distribution

In probability theory and statistics, the Logistic distribution is a continuous probability density function that is symmetric and uni-modal. It resembles the normal distribution in shape but has heavier tails. In practical applications, the two distributions cannot be distinguished from one another.

The probability density function (p.d.f.) of the logistic distribution is given by:

$$f(x) = \frac{1}{\sigma} \cdot \frac{e^{-\frac{x-\mu}{\sigma}}}{[1+e^{-\frac{x-\mu}{\sigma}}]^2} \text{ , } -\infty < x < \infty \text{ ; } -\infty < \mu < \infty \text{ ; } \sigma > 0$$

$$\text{mean} = \mu \quad \text{and} \quad \text{variance} = \frac{\pi^2\sigma^2}{3}$$

Compared with the Normal distribution, $N(\mu,\sigma^2)$, the variance of the Logistic is different from the variance of the normal only by the scaling value of $\frac{\pi^2}{3}$ [26].

One of the most common applications is in logistic regression, which is used for modeling categorical dependent variables (e.g., yes-no choices or a choice of more than two possibilities), much as standard linear regression is used for modeling continuous variables [16].

Suppose we have a hypothetical data set of 20 high school students drawn from Logistic distribution. The data is on the three psychological variables (locus of control, self concept and motivation) and four academic variables (standardized test scores in reading, writing, science and art).

(The simulated data is given in appendix II).

```
Equation        Obs  Parms    RMSE      "R-sq"         F            P
--------------------------------------------------------------------------------------------------
locus_of_c~l     20     5       2.182174     0.3271     1.823242       0.1769
self_concept    20     5       1.557655     0.4208     2.724891       0.0691
motivation       20     5       1.573219     0.5215     4.086314      0.0195
--------------------------------------------------------------------------------------------------
             |    Coef.     Std. Err.       t       P>|t|       [95% Conf. Interval]
-------------+------------------------------------------------------------------------------------
locus_of_control
```

```
          x1 |  -.0516        .0923795    -0.56    0.585    -.2485023    .1453022
          x2 |  -.1094598   .1006543    -1.09    0.294    -.3239992    .1050797
          x3 |   .2879711   .1184109     2.43    0.028     .0355843    .540358
          x4 |  -.063793     .0894625    -0.71    0.487    -.2544779    .1268918
       _cons | -3.195589    11.87159    -0.27    0.791    -28.49928     22.1081
-------------+------------------------------------------------------------------
self_concept
          x1 |   .0170904    .0659413     0.26    0.799    -.1234602     .157641
          x2 |  -.059127     .0718479    -0.82    0.423    -.2122671    .0940132
          x3 |   .2136971    .0845228     2.53    0.023     .0335412    .3938531
          x4 |  -.0184996    .0638591    -0.29    0.776    -.1546122    .1176129
       _cons | -11.30352    8.474044    -1.33    0.202    -29.36551    6.758481
-------------+------------------------------------------------------------------
motivation
          x1 |  -.0643461    .0666002    -0.97    0.349    -.2063011    .0776088
          x2 |   .1165073    .0725658     1.61    0.129    -.038163     .2711776
          x3 |   .0886091    .0853673     1.04    0.316    -.093347     .2705652
          x4 |   .1926915    .0644972     2.99    0.009     .0552189    .3301641
       _cons |  -23.46869   8.558716    -2.74    0.015    -41.71117    5.226223
--------------------------------------------------------------------------------
```

The estimated multivariate multiple regression model is:

$\hat{y}_1 = -3.195589 - 0.0516\, x_1 - 0.1094598\, x_2 + 0.2879711x_3 - 0.063793\, x_4$

$\hat{y}_2 = -11.30352 + 0.0170904\, x_1 - 0.059127\, x_2 + 0.2136971x_3 - 0.0184996\, x_4$

$\hat{y}_3 = -23.46869 -0.0643461\, x_1 + 0.1165073\, x_2 + 0.0886091\, x_3 +0.1926915\, x_4$

Output for MANOVA,

```
                    Number of obs =      20

              W = Wilks' lambda         L = Lawley-Hotelling trace
              P = Pillai's trace        R = Roy's largest root


  Source |   Statistic        df     F(df1,    df2) =      F        Prob>F
 -----------+-------------------------------------------------------------------------
   Model | W   0.2039        4      12.0      34.7         2.38      0.0229 a
         | P   1.1071               12.0      45.0         2.19      0.0287 a
         | L    2.4180              12.0      35.0         2.35      0.0244 a
         | R   1.4086                4.0      15.0         5.28      0.0074 u
         |-------------------------------------------------------------------------
Residual |                     15
 -----------+-------------------------------------------------------------------------
      x1 | W   0.9211        1       3.0      13.0         0.37      0.7752 e
         | P    0.0789                 3.0      13.0         0.37      0.7752 e
         | L    0.0856                 3.0      13.0         0.37      0.7752 e
         | R   0.0856                 3.0      13.0         0.37      0.7752 e
         |-------------------------------------------------------------------------
      x2 | W   0.7514       1       3.0       13.0         1.43       0.2781 e
         | P   0.2486                3.0      13.0         1.43       0.2781 e
         | L   0.3308                3.0      13.0         1.43       0.2781 e
         | R   0.3308                3.0      13.0         1.43      0.2781 e
         |-------------------------------------------------------------------------
      x3 | W   0.5091     1       3.0       13.0         4.18      0.0282 e
         | P   0.4909                3.0      13.0         4.18      0.0282 e
```

```
           | L   0.9642           3.0     13.0        4.18     0.0282 e
           | R   0.9642           3.0     13.0        4.18    0.0282 e
           |-----------------------------------------------------------------
        x4 | W  0.6032     1      3.0     13.0        2.85     0.0784 e
           | P   0.3968           3.0     13.0        2.85     0.0784 e
           | L   0.6577           3.  0   13.0        2.85     0.0784 e
           | R   0.6577           3.0     13.0        2.85     0.0784 e
           |-----------------------------------------------------------------
  Residual |               15
  ---------+-----------------------------------------------------------------
     Total |               19
  ---------------------------------------------------------------------------
              e = exact, a = approximate, u = upper bound on F
```

We can test the null hypothesis that the coefficients for the variable x_1 (read) are equal to zero in all three equations.

(1)[locus_of_control]x1= 0

(2)[self_concept]x1 =0

(3)[motivatiion] x1 = 0

F(3, 15) = 0.43   Prob > F = 0.7357

Because the p-value is greater than the significance level 0.05, we don’t reject the null hypothesis. That is, the coefficients for the variable x_1 (read) are equal to zero in all three equations.

For the variable write,

(1) [locus_of_control]x2= 0

(2)[self_concept]x2 =0

(3)[motivatiion] x2 = 0

F(3, 15) = 1.65 Prob > F = 0.2193

The p-value is greater than the significance level 0.05, so we don’t reject the null hypothesis. That is, the coefficients for the variable x_2 (write) are equal to zero in all three equations.

Test for science

(1)[locus_of_control]x3 = 0

(2)[self_concept]x3 = 0

(3)[motivation] x3 = 0

F(3, 15) = 4.82 Prob > F = 0.0152

The p-value is less than the significance level 0.05, so we reject the null hypothesis. That is, the coefficients for the variable x_3 (science), taken for all three outcomes together, are statistically significant.

Finally, for the variable art:

(1)[locus_of_control]x4 = 0

(2)[self_concept]x4 = 0

(3)[motivation]x4 = 0

```
F( 3, 15)  =  3.29          Prob > F = 0.0499
```

The p-value indicates that we reject the null hypothesis. That is, the coefficients for the variable x_4 (art) taken for all three outcomes together are statistically significant.

3.4.3 Simulation Using Exponential Distribution

Suppose we have a hypothetical dataset of 20 high school students drawn from Exponential distribution. The data is on the three psychological variables (locus of control, self concept and motivation) and four academic variables (standardized test scores in reading, writing, science and art).

(The simulated data is given in appendix III).

```
Equation        Obs  Parms    RMSE       "R-sq"         F                P
---------------------------------------------------------------------------------------------------
locus_of_c~l    20     5       .7524345     0.3562      2.074457       0.1351
self_concept    20     5       .7979952     0.1790     .8174282        0.5337
motivation      20     5       .9459599     0.1845     .8486324        0.5162
---------------------------------------------------------------------------------------------------
 |      Coef.          Std. Err.        t          P>|t|        [95% Conf. Interval]
----------+----------------------------------------------------------------------------------------
locus_of_control
      x1 |  -.0002182      .0038171     -0.06      0.955     -.0083543       .0079178
      x2 |  -.0033952      .0039473    -0.86       0.403    -.0118087        .0050183
      x3 |  -.003846       .0022943    -1.68       0.114    -.0087362        .0010442
      x4 |  -.0072875      .0045457    -1.60       0.130    -.0169764        .0024013
  _cons |  1.839243       .3930391      4.68       0.000       1.0015          2.676986
```

```
----------+-------------------------------------------------------------------------------------
self_concept
       x1 |  -.0015016      .0040483    -0.37      0.716     -.0101303      .0071271
       x2 |   .0019333      .0041863     0.46      0.651     -.0069896      .0108563
       x3 |  -.0012859      .0024332    -0.53      0.605      -.0064722     .0039004
       x4 |  -.0083373      .0048209    -1.73      0.104      -.0186128     .0019382
  _cons | 1.475811        .4168381     3.54      0.003       .5873419      2.364281
----------+-------------------------------------------------------------------------------------
motivation
       x1 |  -.0044473      .0047989    -0.93      0.369       -.014676      .0057813
       x2 |   .0087889      .0049626     1.77      0.097       -.0017886    .0193663
       x3 |   .0007853      .0028844     0.27      0.789       -.0053626    .0069333
       x4 |   -.004499       .0057148     -0.79      0.443       -.0166798    .0076818
  _cons |    1.04685       .4941284      2.12      0.051       -.00636        2.10006
-------------------------------------------------------------------------------------------------
```

The estimated multivariate multiple regression model is:

$$\hat{y}_1 = 1.839243 - 0.0002182\, x_1 - 0.0033952\, x_2 - 0.003846 x_3 - 0.0072875\, x_4$$

$$\hat{y}_2 = 1.475811 - 0.0015016\, x_1 + 0.0019333\, x_2 - 0.0012859 x_3 - 0.0083373\, x_4$$

$$\hat{y}_3 = 1.04685 - 0.0044473\, x_1 + 0.0087889\, x_2 + 0.0007853\, x_3 - 0.004499 x_4$$

MANOVA output:

```
                      Number of obs =     20

                W = Wilks' lambda        L = Lawley-Hotelling trace

                 P = Pillai's trace           R = Roy's largest root

   Source |    Statistic         df      F(df1,    df2) =     F          Prob>F
```

```
-----------+-----------------------------------------------------------------------
     Model | W   0.4027       4      12.0     34.7      1.19      0.3308 a
           | P    0.6961             12.0     45.0      1.13      0.3589 a
           | L    1.2402             12.0      35.0      1.21      0.3177 a
           | R   1.0020               4.0      15.0      3.76      0.0261 u
           |-----------------------------------------------------------------------
  Residual |                   15
-----------+------------------------------------------------------------------------
        x1 | W   0.9448      1        3.0      13.0       0.25      0.8576 e
           | P   0.0552                3.0      13.0       0.25      0.8576 e
           | L   0.0585                3.0      13.0       0.25      0.8576 e
           | R   0.0585               3.0      13.0       0.25       0.8576 e
           |-----------------------------------------------------------------------
        x2 | W   0.7822      1        3.0      13.0      1.21       0.3463 e
           | P   0.2178                3.0      13.0      1.21       0.3463 e
           | L   0.2784                3.0      13.0      1.21       0.3463 e
           | R   0.2784               3.0      13.0      1.21       0.3463 e
           |-----------------------------------------------------------------------
        x3 | W  0.7825       1       3.0       13.0     1.20       0.3470 e
           | P   0.2175               3.0       13.0     1.20       0.3470 e
           | L   0.2779               3.0       13.0     1.20       0.3470 e
           | R  0.2779               3.0        13.0     1.20      0.3470 e
           |-----------------------------------------------------------------------
        x4 | W  0.6747        1       3.0      13.0      2.09      0.1512 e
           | P   0.3253                3.0      13.0      2.09     0.1512 e
           | L   0.4821                3.0      13.0      2.09      0.1512 e
```

```
                                          93
          | R   0.4821           3.0     13.0    2.09      0.1512 e

          |-----------------------------------------------------------------------

 Residual |                 15

 -----------+-------------------------------------------------------------------------

    Total |                 19

 -------------------------------------------------------------------------------------
```

We can test the null hypothesis that the coefficients for the variable x_1 (read) are equal to zero in all three equations.

(1)[locus_of_control]x1= 0

(2)[self_concept]x1 =0

(3)[motivatiion] x1 = 0

F(3, 15) = 0.29   Prob > F = 0.8303

Because the p-value is greater than the significance level 0.05, we don’t reject the null hypothesis. That is, the coefficients for the variable x_1 (read) are equal to zero in all three equations.

For the variable write,

(1) [locus_of_control]x2= 0

(2)[self_concept]x2 =0

(3)[motivatiion] x2 = 0

F(3, 15) = 1.39   Prob > F = 0.2837

The p-value is greater than the significance level 0.05, so we don’t reject the null hypothesis. That is, the coefficients for the variable x_2 (write) are equal to zero in all three equations.

Test for science

(1)[locus_of_control]x3 = 0

(2)[self_concept]x3 = 0

(3)[motivation] x3 = 0

F(3, 15) = 1.39 Prob > F = 0.2844

Here we do not reject the null hypothesis because the p-value is greater than the significance level 0.05. That is, the coefficients for the variable x_3 (science) are equal to zero in all three equations.

Finally, for the variable art:

(1)[locus_of_control]x4 = 0

(2)[self_concept]x4 = 0

(3)[motivation]x4 = 0

F(3, 15) = 2.41 Prob > F = 0.1075

The p-value indicates that we do not have to reject the null hypothesis. That is, the coefficients for the variable x_4 (art) taken for all three outcomes together, are not statistically significant.

After fitting the three regression models of data from the three distributions, normal, logistic, and exponential, the results showed that the regression model which was obtained from the logistic distribution was the best. The decision was according to the p-values as we seen from the hypothesis tests. Also, the explanatory power of the model which was obtained from the logistic distribution was the highest.

Chapter Four

Application

In this chapter real data set will be analyzed to study the relationships among variables. Data was collected on some variables such as tawjihi average, English score level exam, daily budget and number of absents per semester, and the objective of the case study was to show how these variables affect or control some other variables such as self concept, achievement motivation and cumulative average. Data for these variables was collected using standardized questionnaires on 350 university students from three universities: An-Najah National University, Arab American University and Alquds Open University.

The data was analyzed using Stata 13 software. Stata's capabilities include data management, statistical analysis, graphics, simulations, regression analysis and custom programming. The name Stata is a syllabic abbreviation of the words statistics and data [30].

Before analyzing the data we have to test the normality of the responses. Here the SPSS output for test of normality followed by normality plots.

Table 4.1 Test of normality for the response variables

Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
self_concept	.069	350	.000	.966	350	.000
motivation	.031	350	.200*	.996	350	.494
cumavg4	.045	350	.083	.985	350	.001

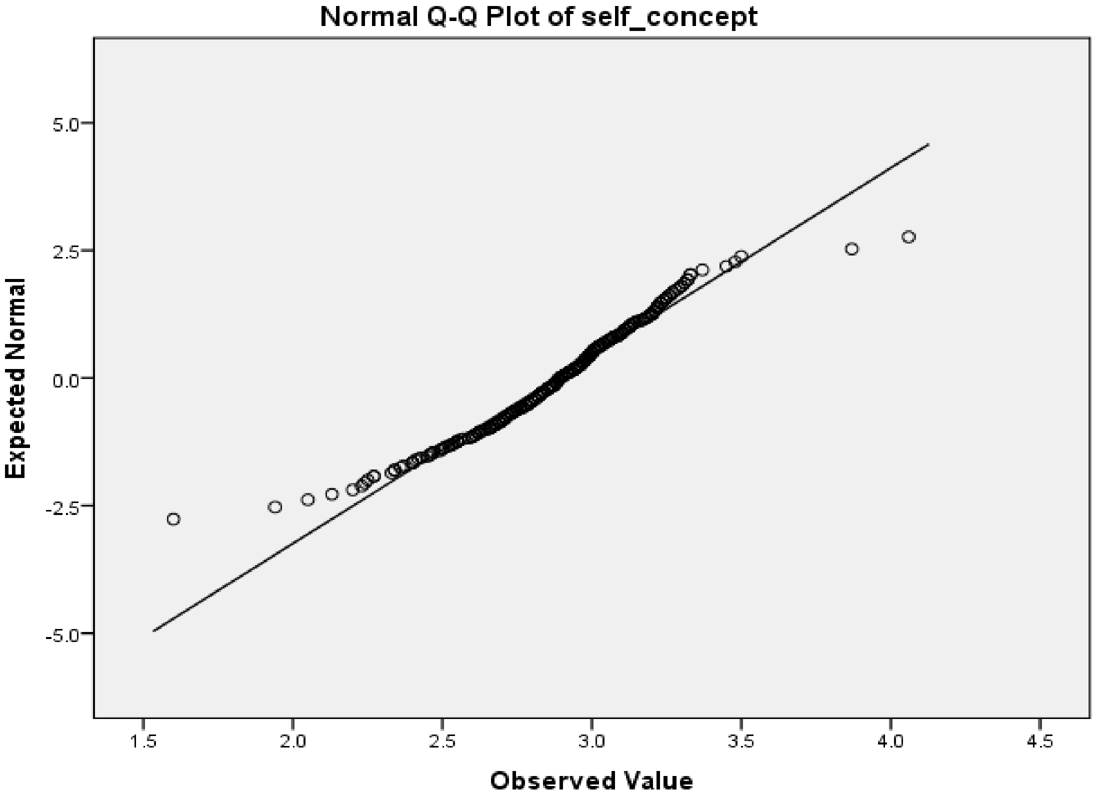


Figure 4.1: Normal Q-Q plot of self_concept

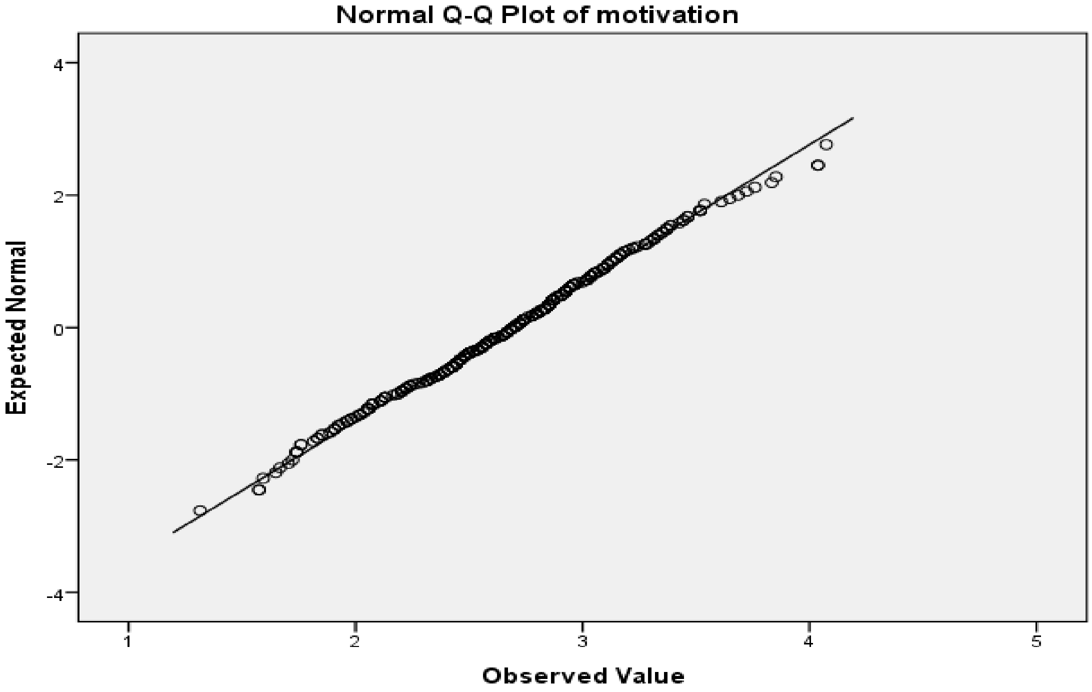


Figure 4.2: Normal Q-Q plot of motivation

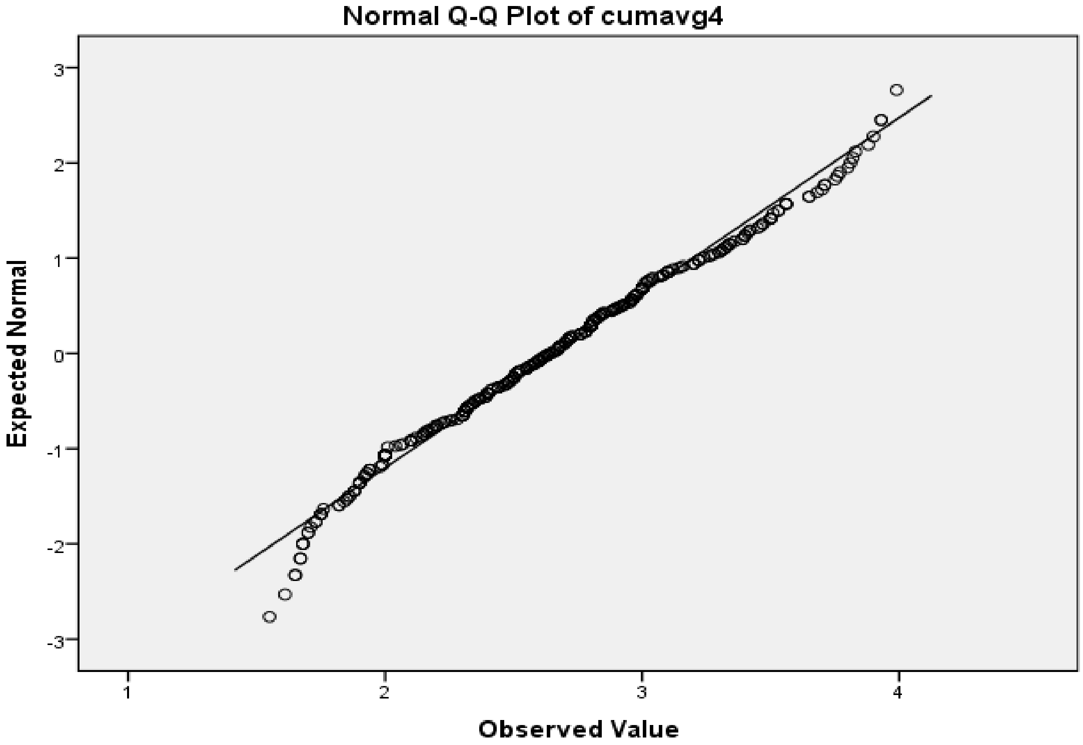


Figure 4.3: Normal Q-Q plot of cumulative average

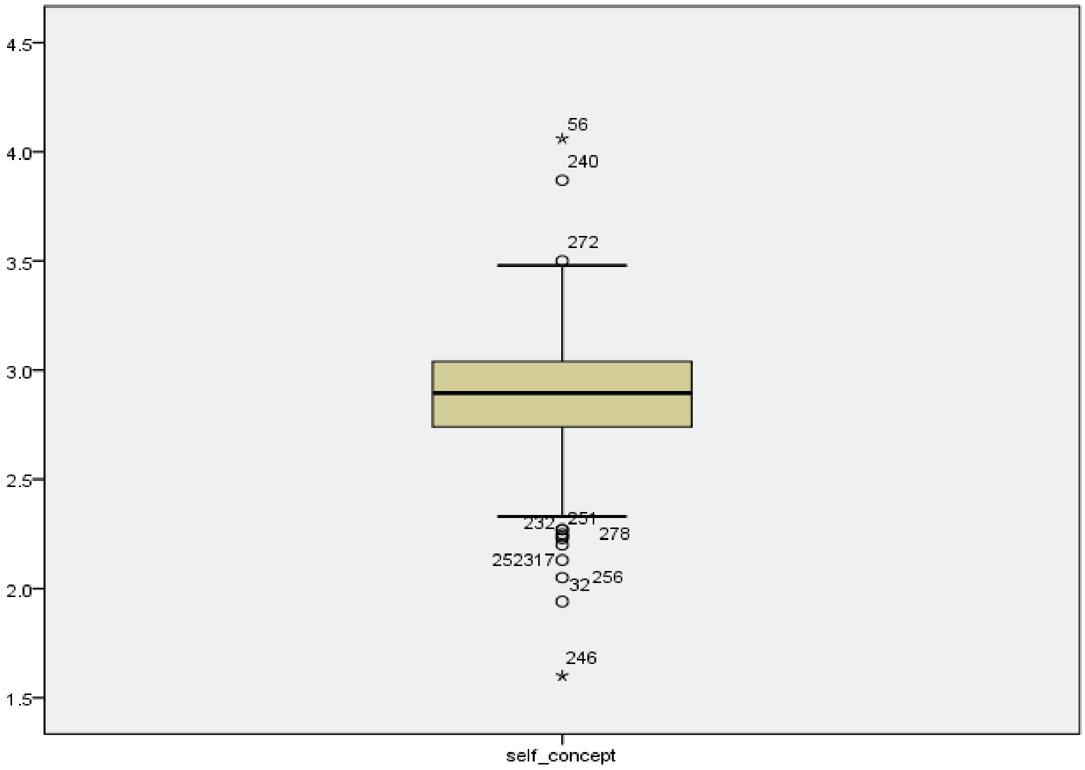


Figure 4.4: Box plot of self_concept

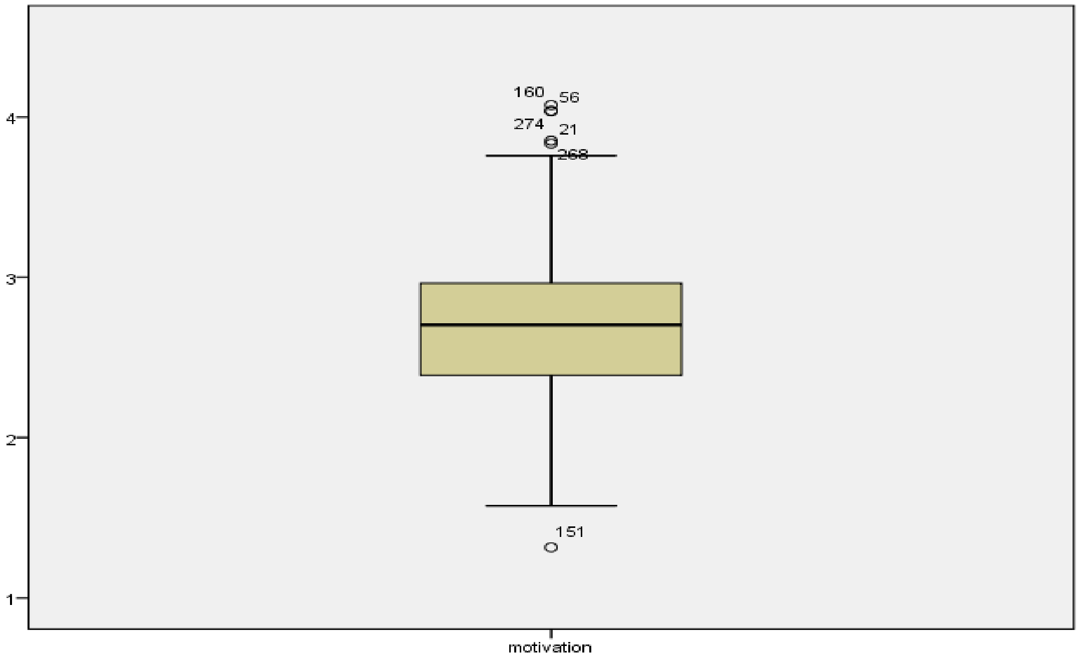


Figure 4.5: Box plot of motivation

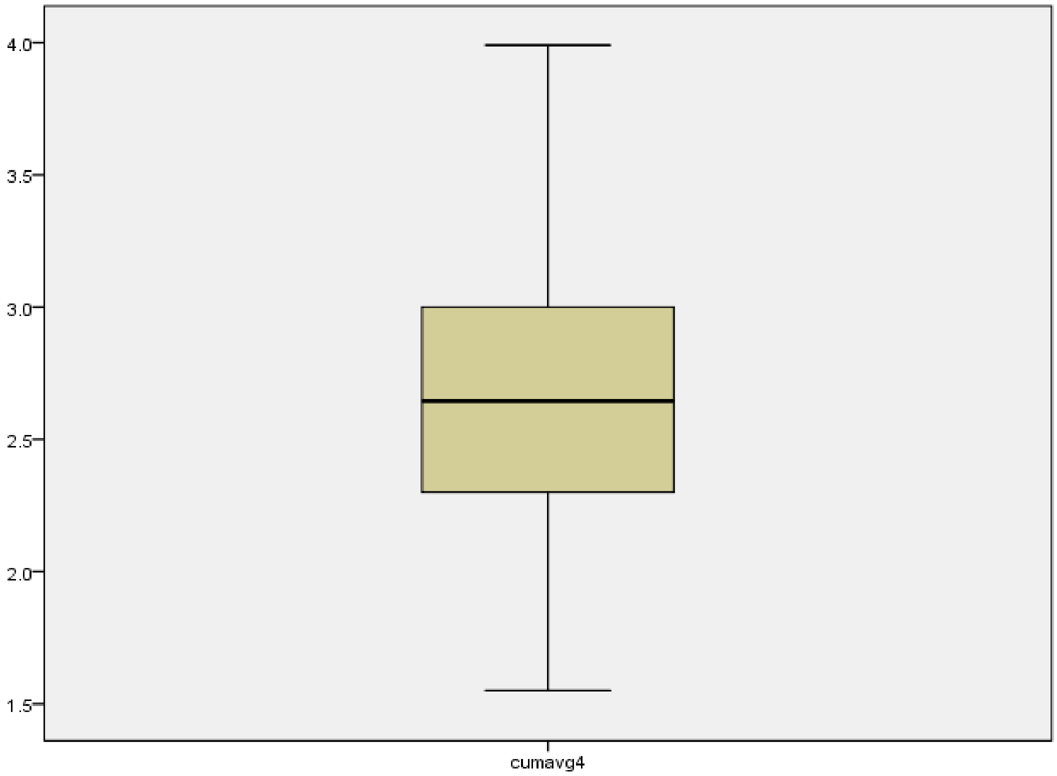


Figure 4.6: Box plot of cumulative average

4.1 Case Study 1: Analysis of Real Data Set with Continuous predictor variables

In this section, a sample of the 350 students were considered regardless the university or college that the student belongs to. As said before, seven variables were considered; three responses (self concept, achievement motivation, and cumulative average), and four predictors (tawjihi average, English score level exam, daily budge,t and number of absents per semester).

The following is a summary for the data

```
. summarize  self_concept  motivation cumavrg  tawjihi  english budget absents
      Variable |     Obs         Mean           Std. Dev.          Min            Max
    -----------+-------------------------------------------------------------------------------
  self_concept|      350         2.881286        0.271814          1.6            4.06
     motivation|     350        2.6774           0.4787392        1.31           4.07
       cumavrg|      350         2.656343       0.5458484         1.55           3.99
    ----------+-------------------------------------------------------------------------------
       tawjihi |     350        81.774           10.74564           55.6           99.5
       english |     350         61.86           16.52749           23             98
       budget |      350         32.94571         20.5946            5             200
      absents |      350         4.614286         6.282036            0              35
   ---------------------------------------------------------------------------------------------
```

In Stata, two commands are needed to get a multivariate multiple regression, manova and mvreg. The manova command will indicate if all of the equations, taken together, are statistically significant. And the mvreg command gives the coefficients, standard errors, etc., for each of the predictors in each part of the model. When running manova command, it is necessary to use the c. in front of the predictors to identify them as continuous variables, because, by default, the manova command assumes all predictor variables are categorical.

```
. manova self_concept motivation cumavrg= c.tawjihi c.english c.budget c.absents

                        Number of obs =     350

                W = Wilks' lambda             L = Lawley-Hotelling trace
                P = Pillai's trace            R = Roy's largest root


     Source |   Statistic        df     F(df1,    df2) =        F        Prob>F
  -----------+--------------------------------------------------------------------
      Model |  W  0.6736         4        12.0     907.8       12.19      0.0000 a
            |  P   0.3323                12.0    1035.0      10.74      0.0000 a
            |  L   0.4759                12.0    1025.0      13.55      0.0000 a
            |  R  0.4571                  4.0     345.0      39.43      0.0000 u
            |--------------------------------------------------------------------
   Residual |                   345
  -----------+--------------------------------------------------------------------
    tawjihi |  W  0.8244          1        3.0      343.0      24.36      0.0000 e
            |  P   0.1756                 3.0      343.0      24.36      0.0000 e
            |  L   0.2130                 3.0      343.0      24.36      0.0000 e
            |  R  0.2130                  3.0      343.0      24.36      0.0000 e
            |--------------------------------------------------------------------
    english |  W  0.9369          1        3.0      343.0       7.70      0.0001 e
            |  P  0.0631                  3.0      343.0       7.70      0.0001 e
```

```
             |  L  0.0674              3.0      343.0      7.70      0.0001 e
             |  R  0.0674              3.0      343.0      7.70      0.0001 e
             |-----------------------------------------------------------------
      budget | W  0.9543       1       3.0      343.0      5.47      0.0011 e
             | P   0.0457              3.0      343.0      5.47      0.0011 e
             | L   0.0478              3.0      343.0      5.47      0.0011 e
             | R  0.0478               3.0      343.0      5.47      0.0011 e
             |-----------------------------------------------------------------
     absents | W  0.9535       1       3.0      343.0      5.58      0.0010 e
             | P  0.0465               3.0      343.0      5.58      0.0010 e
             | L  0.0488               3.0      343.0      5.58      0.0010 e
             | R   0.0488              3.0      343.0      5.58      0.0010 e
             |-----------------------------------------------------------------
    Residual |                345
-------------+-----------------------------------------------------------------
       Total |                349
-------------------------------------------------------------------------------
                 e = exact, a = approximate, u = upper bound on F
```

The test for the overall model, shown in the section labeled Model (under source), indicates that the model is statistically significant, regardless of the type of multivariate criteria that is used (i.e. all of the p-values are less than 0.05). Below the overall model tests are the multivariate tests for each of the predictor variables. And we see from the last column of p-values that each of predictors is statistically significant.

We use mvreg to obtain estimates of the coefficients in our model.

```
. mvreg
  Equation        Obs   Parms      RMSE        "R-sq"         F              P
-----------------------------------------------------------------------------------------------
self_concept      350      5        .2669567     0.0465      4.203963        0.0025
motivation        350      5        .4793062     0.0091      0.7936825       0.5299
cumavrg           350      5        .467025      0.2763      32.93712        0.0000
-----------------------------------------------------------------------------------------------
             |    Coef.       Std. Err.       t          P>|t|          [95% Conf. Interval]
 ------------+---------------------------------------------------------------------------------
self_concept
    tawjihi |  .0042879      .0013811    3.10       0.002         .0015715      .0070043
    english |  .0001206      .0008938     0.13      0.893        -.0016373     .0018785
     budget |  -.0014926     .0006999    -2.13      0.034        -.0028693     -.000116
    absents |  .0001949      .0022901     0.09      0.932        -.0043094      .0046992
      _cons |  2.571464      .1190967    21.59      0.000         2.337217      2.805711
------------+-----------------------------------------------------------------------------------
motivation
   tawjihi |    .00152        .0024797      0.61      0.540       -.0033572       .0063972
   english |   -.0028165    .0016047      -1.76      0.080       -.0059728       .0003398
    budget |   -.000028      .0012567      -0.02      0.982       -.0024997      .0024438
  absents |    .0003267    .0041117        0.08      0.937       -.0077605     .0084138
     _cons |    2.726749    .2138316      12.75      0.000        2.306172      3.147327
 ------------+-----------------------------------------------------------------------------------
cumavrg
    tawjihi |   .0184153     .0024161       7.62      0.000         .0136631       .0231675
    english |   .0068497     .0015636      4.38      0.000         .0037743       .0099251
```

```
     budget |  -.0038057    .0012245     -3.11    0.002      -.0062141    -.0013973

    absents |  -.0163848    .0040064     -4.09    0.000      -.0242648    -.0085048

      _cons |   .9277133    .2083526      4.45    0.000       .5179121     1.337514

-----------------------------------------------------------------------------------------------
```

The first table gives the number of observations, number of parameters, RMSE, R-squared, F-ratio, and p-value for each of the three models.

- According to the column labeled P, each of the self concept and cumulative average models are statistically significant. But the model of motivation is not.
- In the column labeled by R-sq, we see that the four predictor variables explain approximately 5%, 0.9%, and 28% of the variance in the outcome variables self concept, motivation, and cumulative average, respectively.
- The second table contains the coefficients, their standard errors, test statistics (t), p-values, and 95% confidence interval, for each predictor variable in the model. The coefficients are interpreted in the same way that coefficients of simple or multiple regression are interpreted. For example, looking at the top of the table, a one unit increase in tawjihi average is associated with 0.00428 unit increase in the predicted value of self concept for a student.

The test command can be applied to test different hypothesis.

For the first test, the null hypothesis is that the coefficients for the variable tawjihi are equal to zero in all three equations.

```
. test tawjihi
```

```
 ( 1)  [self_concept]tawjihi = 0

 ( 2)  [motivation]tawjihi = 0

 ( 3)  [cumavrg]tawjihi = 0

       F ( 3, 345 ) = 24.50,            Prob > F =    0.0000
```

The result of this test is to reject the null hypothesis that the coefficients for tawjihi across the three equations are simultaneously equal to zero, in other words, the coefficients for tawjihi, taken for all three outcomes together, are statistically significant.

The same test can be done for the other three predictors, English, budget, and absents.

```
. test english

 ( 1)  [self_concept]english = 0

 ( 2)  [motivation]english = 0

 ( 3)  [cumavrg]english = 0

        F ( 3, 345 ) = 7.75,            Prob > F =    0.0001

. test budget

 ( 1)  [self_concept]budget = 0

 ( 2)  [motivation]budget = 0

 ( 3)  [cumavrg]budget = 0
```

```
          F ( 3, 345 ) = 5.50,        Prob > F =    0.0011

 . test absents

 ( 1)  [self_concept]absents = 0

 ( 2)  [motivation]absents = 0

 ( 3)  [cumavrg]absents = 0

       F ( 3, 345 ) = 5.61,           Prob > F =    0.0009
```

The results of these tests indicate that we must reject the null hypothesis. That is, the coefficients for English, taken for all three outcomes together, are statistically significant. And the coefficients for budget, taken for all three outcomes together, are statistically significant. And the coefficients for absents, taken for all three outcomes together, are statistically significant.

We can also test the null hypothesis that the coefficient for the variable English in the equation with cumulative average as the outcome is equal to the coefficient for English in the equation with motivation as the outcome. Another way of stating this null hypothesis is that, the effect of English score on the cumulative average is equal to the effect of English score on motivation.

```
 . test[cumavrg]english= [motivation]english

 ( 1)  - [motivation]english + [cumavrg]english = 0
```

```
F ( 1, 345) = 17.64,        Prob > F =    0.0000
```

The result of this test indicates that the difference between the coefficients of English with cumulative average and motivation as the outcome is significantly different from 0. In other words, the coefficients are significantly different.

Now, we will test the null hypothesis that the coefficient for the variable tawjihi in the equation with cumulative average as the outcome is equal to the coefficient for tawjihi in the equation with self concept as the outcome.

```
. test[cumavrg]tawjihi= [self_concept]tawjihi

( 1)  - [self_concept]tawjihi + [cumavrg]tawjihi = 0

       F ( 1, 345) = 23.76,            Prob > F =    0.0000
```

The result indicates that the difference between the coefficients of tawjihi with cumulative average and self concept as the outcome is significantly different from 0.

4.2 Case Study 2: Analysis of Real Data Set with Categorical Predictor Variable (program)

In this section, the same sample of 350 observations would be considered but, we would be interested in the collage or the program that the student belongs to regardless at which university he is studying. Our interest is on three programs (or specializations); economics, engineering, and science. We have three response variables (self concept, motivation, and cumulative average), and five predictors (tawjihi average, English score level exam, daily budget, number of absents per semester, and the program that the student belongs to).

```
  . summarize self_concept motivation cumavrg tawjihi english budget absents

      Variable |    Obs          Mean            Std. Dev.         Min          Max

  ----------------+-------------------------------------------------------------------------------

self_concept|    350          2.881286        0.271814          1.6           4.06

   motivation |    350         2.6774           0.4787392         1.31          4.07

     cumavrg |   350        2.656343        0.5458484         1.55         3.99

        tawjihi |   350          81.774           10.74564          55.6          99.5

       english |    350          61.86             16.52749          23            98

-----------------+-------------------------------------------------------------------------------

       budget |    350          32.94571        20.5946              5            200

     absents |    350         4.614286         6.282036            0            35
```

```
. tabulate prog

            prog |      Freq.       Percent             Cum.

     ------------+------------------------------------------------

    economic |      144           41.14               41.14

 engineering |      108           30.86                72.00

       science |      98             28.00             100.00

      -----------+------------------------------------------------

         Total |       350           100.00
```

```
. manova  self_concept motivation cumavrg= c.tawjihi c.english c.budget c.absents prog

                              Number of obs =     350

                       W = Wilks' lambda            L = Lawley-Hotelling trace

                        P = Pillai's trace               R = Roy's largest root
```

```
   Source |  Statistic       df      F(df1,    df2) =       F         Prob>F

--------------+---------------------------------------------------------------------------------

    Model | W 0.6556       6     18.0       965.0        8.63        0.0000 a

          | P 0.3550             18.0      1029.0        7.67        0.0000 a

          | L  0.5092            18.0      1019.0        9.61        0.0000 a

          | R  0.4762             6.0       343.0       27.22        0.0000 u

          |-------------------------------------------------------------------------------

 Residual |                343

--------------+---------------------------------------------------------------------------------

   tawjihi | W  0.8308     1       3.0        341.0      23.14        0.0000 e

          | P  0.1692             3.0       341.0      23.14        0.0000 e

          | L 0.2036              3.0       341.0      23.14        0.0000 e

          | R 0.2036              3.0       341.0      23.14        0.0000 e

          |-------------------------------------------------------------------------------

   english | W 0.9376      1       3.0        341.0       7.56        0.0001 e

           | P 0.0624             3.0       341.0       7.56        0.0001 e

           | L 0.0665             3.0       341.0       7.56        0.0001 e

           | R   0.0665           3.0       341.0       7.56        0.0001 e

           |------------------------------------------------------------------------------

    budget | W 0.9544      1      3.0       341.0       5.43          0.0012 e

           | P 0.0456             3.0      341.0       5.43          0.0012 e

           | L  0.0478            3.0       341.0      5.43          0.0012 e

           | R 0.0478             3.0      341.0       5.43          0.0012 e

           |------------------------------------------------------------------------------

   absents |W 0.9610        1     3.0       341.0       4.61          0.0035 e

           | P 0.0390             3.0       341.0      4.61          0.0035 e
```

```
             | L 0.0405            3.0      341.0      4.61       0.0035 e

             | R 0.0405            3.0      341.0      4.61       0.0035 e

             |-----------------------------------------------------------------------

        prog |W 0.9733      2      6.0      682.0      1.55       0.1604 e

             |P 0.0268             6.0      684.0      1.55       0.1604 a

             | L 0.0273            6.0      680.0      1.55       0.1604 a

             |R 0.0225             3.0      342.0      2.57       0.0543 u

             |-----------------------------------------------------------------------

    Residual |              343

  -----------+------------------------------------------------------------------------

       Total |              349

      ---------------------------------------------------------------------------------

                     e = exact , a = approximate  ,  u = upper bound on F

. mvreg

 Equation        Obs   Parms     RMSE          "R-sq"          F              P

--------------------------------------------------------------------------------------

self_concept     350      7        0.2672918      0.0496      2.984838     0.0074

motivation        350      7        0.478601        0.0178      1.033593     0.4031

 cumavrg          350     7        0.4643145       0.2889      23.2218      0.0000

----------------------------------------------------------------------------------------

            |     Coef.          Std. Err.       t         P>|t|        [95% Conf. Interval]

------------+---------------------------------------------------------------------------

self_concept

   tawjihi |  .0038579       .001539      2.51     0.013       .0008308     .0068851

 english |   .0001651       .000898      0.18     0.854      -.0016012    .0019314

 budget |  -.0015163       .0007073   -2.14     0.033      -.0029075   -.0001251
 absents|  -.00017          .0023185   -0.07      0.942      -.0047303    .0043903

     prog |

        2  |   .0269295       .0387385    0.70      0.487      -.0492654    .1031244

        3  |   -.0139693      .0368504   -0.38      0.705     -.0864506    .0585119

  _cons |  2.601938        .125458     20.74     0.000      2.355174    2.848702

----------+--------------------------------------------------------------------------------------

motivation

 tawjihi |  -.0005721       .0027557    -0.21      0.836      -.0059924    .0048482

english |  -.0028728       .001608      -1.79      0.075      -.0060355    .0002899

budget |   .0001402       .0012665      0.11      0.912      -.0023508    .0026312

absents| -.0004423       .0041514     -0.11      0.915     -.0086077    .0077232

prog

       2  |    .1204565       .0693634      1.74      0.083      -.0159747    .2568876

       3  |    .0521768       .0659827      0.79      0.430      -.077605      .1819585

_cons |   2.847534        .2246396     12.68    0.000       2.40569      3.289379

---------+---------------------------------------------------------------------------------------

cumavrg

 tawjihi |   .0203167      .0026735       7.60      0.000       .0150582    .0255752

english |  .0066932       .00156            4.29      0.000      .0036249    .0097615

budget | -.0037428       .0012286      -3.05      0.002     -.0061594   -.0013261

absents| -.0149192      .0040275      -3.70      0.000     -.02284       -.0069975

    prog |

        2  |  -.1175353     .0672929       -1.75       0.082     -.2498939   .0148234

        3  |   .0440806     .0640131        0.69        0.492    -.0818271    .1699883

  _cons |   .7969961    .217934            3.66       0.000      .3683408   1.225651

--------------------------------------------------------------------------------------------------
```

Notes from tables:

- A one unit increase in tawjihi average is associated with a 0.0203167 increase in the predicted value of cumulative average.
- If the budget increases by one unit, the cumulative average is expected to decrease by 0.0037428 .
- As the absents increases by one unit, the predicted value of motivation will decrease by 0.0004423.

For the first test, the null hypothesis is that the coefficients for the variable tawjihi are equal to zero in all three equations.

```
. test tawjihi

 ( 1)  [self_concept]tawjihi = 0

 ( 2)  [motivation]tawjihi = 0

 ( 3)  [cumavrg]tawjihi = 0

           F ( 3, 343) = 23.28,        Prob > F =    0.0000
```

The result of this test is to reject the null hypothesis. That is, the coefficients for tawjihi, taken for all three outcomes together, are statistically significant.

The same test is done for the variables English, budget, and absents

```
. test english

 ( 1)  [self_concept]english = 0
 ( 2)  [motivation]english = 0

 ( 3)  [cumavrg]english = 0

          F ( 3, 343) = 7.60,          Prob > F =    0.0001,

. test budget

 ( 1)  [self_concept]budget = 0

 ( 2)  [motivation]budget = 0

 ( 3)  [cumavrg]budget = 0

             F ( 3, 343) = 5.46,          Prob > F =    0.0011

. test absents

 ( 1)  [self_concept]absents = 0

 ( 2)  [motivation]absents = 0

 ( 3)  [cumavrg]absents = 0

           F ( 3, 343) = 4.64,         Prob > F =    0.0034
```

Second, we can test the null hypothesis that the coefficients for the variable tawjihi in the equation of cumulative average as the outcome is equal to the coefficient for tawjihi in the equation of motivation as the outcome.

```
. test[cumavrg]tawjihi= [motivation]tawjihi

 ( 1)  - [motivation]tawjihi + [cumavrg]tawjihi = 0
```

```
      F ( 1, 343) = 28.27,        Prob > F =   0.0000
```

The result of this test indicates that the difference between the coefficients for tawjihi with cumulative average and motivation as the outcome is significantly different from 0, in other words, the coefficients are significantly different.

```
. test[self_concept]budget= [motivation]budget

 ( 1)  [self_concept]budget - [motivation]budget = 0

      F ( 1, 343) = 2.04,         Prob > F =   0.1545
```

Here the decision is not to reject the null hypothesis. That is, the coefficients of budget are not significantly different in the two equations. In other words, that the effect of budget on the self concept is equal to the effect of budget on motivation.

```
. test[cumavrg]absents= [motivation]absents

 ( 1)  - [motivation]absents + [cumavrg]absents = 0

     F ( 1, 343) = 5.98,          Prob > F =   0.0149
```

The test indicates to reject the null hypothesis. That is, the effect of absents on the motivation is different from the effect of absents on cumulative average.

We can test the null hypothesis that the coefficients for prog=2 and prog=3 are simultaneously equal to zero in the equation of cumulative average. When used to test the coefficients for dummy variables that form a single categorical predictor, this type of test is sometimes called an overall test for the effect of the categorical predictor .

```
 . test[cumavrg]2.prog [cumavrg]3.prog

( 1)  [cumavrg]2.prog = 0

 ( 2)  [cumavrg]3.prog = 0

       F ( 2, 343) = 3.02,            Prob > F =    0.0501
```

The result indicates that the two coefficients together are not significantly different from 0. In other words, the overall effect of program on cumulative average is not statistically significant.

```
. test[self_concept]2.prog [self_concept]3.prog

( 1)  [self_concept]2.prog = 0

 ( 2)  [self_concept]3.prog = 0

       F ( 2, 343) = 0.57,            Prob > F =    0.5674

. test[motivation]2.prog [motivation]3.prog

( 1)  [motivation]2.prog = 0

 ( 2)  [motivation]3.prog = 0

            F ( 2, 343) = 1.51,       Prob > F =    0.2227
```

Achievement motivation and self concept are not affected by the study program of the student.

In general, the estimated multivariate multiple regression model was good. We saw that the self concept, achievement motivation, and the cumulative average for a university student is affected by the four predictor variables (tawjihi average, English score, budget, and absents). Moreover, the study program has no effect on student's self concept and his achievement motivation, the p-values were 0.5674, and 0.2227, respectively. The student's cumulative average is affected by the study program. The p-value was 0.0501.

4.3 Case Study 3: Analysis of Real Data Set with Categorical Predictor Variable (university)

In this section, the same sample of 350 observations would be considered. But, we would be interested in the university that the student is studying in regardless the study program. Our interest is on three universities; An-Najah National University, Arab American University, and Alquds Open University. We have three response variables (self concept, motivation, and cumulative average), and five predictors (tawjihi average, English score level exam, daily budget, number of absents per semester, and the university that the student studying in).

```
. summarize self_concept motivation cumavrg tawjihi english budget absents

    Variable |   Obs        Mean          Std. Dev.        Min        Max

   ------------+---------------------------------------------------------------------------

self_concept |   350      2.881286       .271814          1.6         4.06

  motivation |   350       2.6774         .4787392        1.31        4.07

     cumavrg |   350      2.656343      .5458484        1.55        3.99

      tawjihi |    350       81.774        10.74564        55.6        99.5

      english |    350       61.86          16.52749         23          98

   ------------+---------------------------------------------------------------------------

      budget |   350      32.94571        20.5946           5          200

     absents |   350       4.614286       6.282036         0           35

. tabulate univ

         univ |    Freq.       Percent          Cum.

    -----------+--------------------------------------------------

       alquds |     91           26.00            26.00

arbamerican |   111           31.71            57.71

        najah |   148           42.29          100.00

    -----------+--------------------------------------------------

        Total |    350          100.00

. manova self_concept motivation cumavrg= c.tawjihi c.english c.budget c.absents univ

                                    Number of obs =     350

                      W = Wilks' lambda            L = Lawley-Hotelling trac e

                      P = Pillai's trace              R = Roy's largest root

    Source |   Statistic        df          F(df1,    df2) =            F           Prob>F

  -----------+-------------------------------------------------------------------------------------

    Model | W   0.5777        6          18.0       965.0          11.48          0.0000 a
```

```
           | P    0.4491            18.0    1029.0      10.07      0.0000 a
           | L    0.6848            18.0    1019.0      12.92      0.0000 a
           | R   0.6111              6.0     343.0      34.93      0.0000 u
           |------------------------------------------------------------------------------------
  Residual |               343
-----------+------------------------------------------------------------------------------------
   tawjihi | W   0.7595      1        3.0     341.0      35.99      0.0000 e
           | P   0.2405               3.0     341.0      35.99      0.0000 e
           | L   0.3166               3.0     341.0      35.99      0.0000 e
           | R   0.3166               3.0     341.0      35.99      0.0000 e
           |------------------------------------------------------------------------------------
   english | W  0.9618       1        3.0     341.0       4.51      0.0040 e
           | P  0.0382                3.0     341.0       4.51      0.0040 e
           | L  0.0397                3.0     341.0       4.51      0.0040 e
           | R 0.0397                 3.0     341.0       4.51      0.0040 e
           |---------------------------------------------------------------------------------
    budget | W 0.9590        1        3.0     341.0       4.86      0.0025 e
           | P  0.0410                3.0     341.0       4.86      0.0025 e
           | L  0.0427                3.0     341.0       4.86      0.0025 e
           | R 0.0427                 3.0     341.0       4.86      0.0025 e
           |---------------------------------------------------------------------------------
   absents | W 0.9767        1        3.0     341.0       2.71      0.0448 e
           | P  0.0233                3.0     341.0       2.71      0.0448 e
           | L   0.0239               3.0     341.0       2.71      0.0448 e
           | R   0.0239               3.0     341.0       2.71      0.0448 e
           |--------------------------------------------------------------------------------
```

```
      univ | W 0.8577      2      6.0     682.0        9.06     0.0000 e

           | P  0.1424            6.0     684.0        8.74     0.0000 a

           | L   0.1657           6.0     680.0        9.39     0.0000 a

           | R  0.1645            3.0     342.0       18.75     0.0000 u

           |------------------------------------------------------------------------------

  Residual |                343

   -----------+------------------------------------------------------------------------------

     Total |                349

       ----------------------------------------------------------------------------------

. mvreg

Equation         Obs    Parms    RMSE        "R-sq"           F                P

---------------------------------------------------------------------------------------------

self_concept     350      7      .2666633      0.0541       3.268718        0.0039

motivation        350      7      .4795371       0.0139        .8065809        0.5653

cumavrg          350      7     .4383563       0.3662      33.02444          0.0000

---------------------------------------------------------------------------------------------

           |       Coef.        Std. Err.         t          P>|t|          [95% Conf. Interval]

------------+-------------------------------------------------------------------------------

self_concept

    tawjihi |   .0023985      .0017954       1.34       0.182       -.0011329       .0059299

    english |   .0005523     .0009302        0.59       0.553       -.0012773       .0023819

     budget |  -.0015061     .0007052       -2.14       0.033       -.0028932      -.0001191

   absents |  -.0008609     .0023968       -0.36      0.720      -.0055753        .0038534

       univ |

         2  |  .0501958      .0414407        1.21       0.227       -.031314          .1317056

         3  |  .0797626      .048241          1.65       0.099       -.0151228        .1746479
    _cons |  2.654934     .1302952      20.38    0.000       2.398655      2.911212
-------------+----------------------------------------------------------------------------------
motivation
     tawjihi |  .003846       .0032287       1.19      0.234       -.0025044      .0101965
    english | -.0033581     .0016728      -2.01      0.045       -.0066483     -.000068
    budget |  .0000665     .0012681       0.05      0.958       -.0024278      .0025607
   absents |   .0014272     .0043102       0.33      0.741        -.0070506     .009905
        univ |
           2  |  -.0892385     .0745222      -1.20      0.232         -.2358165     .0573395
           3  |  -.1000668     .0867511      -1.15      0.250         -.2706979     .0705643
      _cons |  2.632467     .2343082      11.24      0.000         2.171605      3.093329
-------------+----------------------------------------------------------------------------------
cumavrg
     tawjihi | .0298896       .0029514      10.13      0.000          .0240845      .0356948
   english |  .004178        .0015291        2.73      0.007         .0011704       .0071856
    budget | -.003339        .0011592       -2.88      0.004        -.0056191      -.001059
   absents | -.0109577      .0039401       -2.78      0.006        -.0187074      -.0032079
        univ |
           2  | -.4404709      .0681225       -6.47      0.000        -.5744614      -.3064805
           3  | -.4936445      .0793012       -6.22      0.000        -.6496225      -.3376666
      _cons |  .4627001      .2141867        2.16      0.031         .0414154       .8839848
--------------------------------------------------------------------------------------------------

. test tawjihi

 ( 1)  [self_concept]tawjihi = 0

 ( 2)  [motivation]tawjihi = 0
 ( 3)  [cumavrg]tawjihi = 0

              F ( 3, 343) = 36.20,        Prob > F =    0.0000

. test english

 ( 1)  [self_concept]english = 0

 ( 2)  [motivation]english = 0

 ( 3)  [cumavrg]english = 0

           F ( 3, 343) = 4.54,         Prob > F =    0.0039

. test budget

 ( 1)  [self_concept]budget = 0

 ( 2)  [motivation]budget = 0

 ( 3)  [cumavrg]budget = 0

      F ( 3, 343) = 4.88,              Prob > F =    0.0024

. test absents

 ( 1)  [self_concept]absents = 0

 ( 2)  [motivation]absents = 0

 ( 3)  [cumavrg]absents = 0

           F ( 3, 343) = 2.73,         Prob > F =    0.0438
```

The four tests indicate that the coefficients for the tawjihi, english, budget, and absents, taken for all three outcomes together, are statistically significant.

Now, we want to test the overall effect for the university on the three responses, cumulative average, self concept, and achievement motivation.

```
test[cumavrg]2.univ [cumavrg]3.univ

 ( 1)  [cumavrg]2.univ = 0

 ( 2)  [cumavrg]3.univ = 0

          F ( 2, 343) = 24.30,       Prob > F =    0.0000
```

The null hypothesis is rejected. That is, the cumulative average is affected by the university that the student belong to.

```
. test[motivation]2.univ [motivation]3.univ

 ( 1)  [motivation]2.univ = 0

 ( 2)  [motivation]3.univ = 0

        F ( 2, 343) = 0.83,          Prob > F =    0.4352

. test[self_concept]2.univ [self_concept]3.univ

 ( 1)  [self_concept]2.univ = 0

 ( 2)  [self_concept]3.univ = 0

          F ( 2, 343) = 1.38,        Prob > F =    0.2530
```

Achievement motivation and self concept were not affected by the university that the student is studying in.

In this model, all the predictors we used were useful in predicting the responses. We saw that the self concept, achievement motivation, and the cumulative average for a university student is determined by the tawjihi average, English score, budget, and absents per semester. The variable university affects the cumulative average for the student, but is has no effect on his achievement motivation or his self concept.

4.4 Case Study 4: Data Analysis of An-Najah National University

In this section we are interested in studying the effect of tawjihi average, English score level exam, budget per day, number of absents per semester, and the program that the student belongs to on the responses self concept, achievement motivation, and cumulative average of a student. Data for 148 students from An- Najah National University has been collected and analyzed.

```
. summarize self_concept motivation cumavrg tawjihi english budget absents

    Variable |   Obs        Mean        Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------------------
self_concept |   148     2.928378      .241934         1.94       4.06
  motivation |   148     2.67973       .4470279        1.57       4.04
     cumavrg |   148     2.687297      .5534095        1.65       3.82
     tawjihi |   148     88.3223       7.087277        69.1       99.5
     english |   148     60.74324      18.35631         23         98
-------------+-----------------------------------------------------------------------------
```

```
      budget |   148        30.48649        15.09088          5           100

     absents |   148        6.722973        6.853988          0            30

. tabulate prog

        prog |      Freq.       Percent         Cum.

   ------------+---------------------------------------------

    economic |      32            21.62          21.62

 engineering |      62            41.89          63.51

     science |      54            36.49         100.00

      ----------+---------------------------------------------

       Total |     148           100.00

. manova self_concept motivation cumavrg= c.tawjihi c.english c.budget c.absents prog

                       W = Wilks' lambda          L = Lawley-Hotelling trace

                       P = Pillai's trace             R = Roy's largest root

  Source |   Statistic            df           F(df1,    df2) =            F            Prob>F

 -----------+-------------------------------------------------------------------------------------

   Model | W    0.6625          6          18.0         393.6           3.43          0.0000 a

         | P    0.3544                     18.0         423.0           3.15          0.0000 a

         | L    0.4840                     18.0         413.0           3.70          0.0000 a

         | R    0.4269                      6.0         141.0          10.03          0.0000 u

         |-----------------------------------------------------------------------------------

Residual |                      141

   ----------+-----------------------------------------------------------------------------------

   tawjihi | W    0.7683          1          3.0          139.0          13.97          0.0000 e

           | P    0.2317                     3.0          139.0          13.97          0.0000 e

           | L    0.3015                     3.0          139.0          13.97          0.0000 e

           | R    0.3015                     3.0          139.0          13.97          0.0000 e
           |-----------------------------------------------------------------------------
   english | W   0.9857      1     3.0     139.0     0.67     0.5717 e
           | P   0.0143            3.0     139.0     0.67     0.5717 e
           | L   0.0145            3.0     139.0     0.67     0.5717 e
           | R   0.0145            3.0     139.0     0.67     0.5717 e
           |-----------------------------------------------------------------------------
    budget | W   0.9750      1     3.0     139.0     1.19     0.3173 e
           | P   0.0250            3.0     139.0     1.19     0.3173 e
           | L   0.0256            3.0     139.0     1.19     0.3173 e
           | R   0.0256            3.0     139.0     1.19     0.3173 e
           |-----------------------------------------------------------------------------
   absents | W   0.9768      1     3.0     139.0     1.10     0.3514 e
           | P   0.0232            3.0     139.0     1.10     0.3514 e
           | L   0.0237            3.0     139.0     1.10     0.3514 e
           | R   0.0237            3.0     139.0     1.10     0.3514 e
           |-----------------------------------------------------------------------------
      prog | W   0.9111      2     6.0     278.0     2.21     0.0425 e
           | P   0.0905            6.0     280.0     2.21     0.0421 a
           | L   0.0958            6.0     276.0     2.20     0.0430 a
           | R   0.0706            3.0     140.0     3.29     0.0225 u
           |-----------------------------------------------------------------------------
  Residual |                 141
-----------+-----------------------------------------------------------------------------
     Total |                 147
      -----------------------------------------------------------------------------------
              e = exact, a = approximate, u = upper bound on F
```

```
. mvreg

Equation        Obs    Parms       RMSE        "R-sq"            F                P

-------------------------------------------------------------------------------------------------

self_concept     148      7         .2430859       0.0317        .7683536         0.5960

motivation        148      7         .4478248        0.0374        .912881          0.4876

cumavrg          148      7         .4751208        0.2930       9.739242         0.0000

---------------------------------------------------------------------------------------------------

             |    Coef.            Std. Err.          t           P>|t|            [95% Conf. Interval]

 -------------+--------------------------------------------------------------------------------------

self_concept

    tawjihi |   .0050275       .0035389       1.42        0.158       -.0019686         .0120235

   english |   .0000777      .0011186        0.07       0.945       -.0021338         .0022892

   budget |  -.0004821      .001355         -0.36       0.723       -.0031609         .0021966

  absents |   .0010951      .0030215       0.36       0.718      -.0048781         .0070684

      prog |

          2  |   .0334322       .0589287         0.57         0.571       -.0830658        .1499303

          3  |   .0272275       .0556512         0.49         0.625       -.0827911        .1372461

    _cons |   2.463017       .3205536          7.68         0.000       1.829305         3.09673

-------------+--------------------------------------------------------------------------------------

motivation

    tawjihi |   .0012794      .0065195           0.20         0.845        -.0116092        .0141679

   english |  -.0023383      .0020608          -1.13         0.258        -.0064124        .0017358

   budget |   .0027088       .0024963          1.09         0.280        -.0022261        .0076438

  absents |  -.0005061       .0055663         -0.09          0.928        -.0115104        .0104981

      prog |

         2  |   .1679357       .1085614          1.55          0.124         -.0466827        .3825541
```

```
          3  |   .1796755    .1025234     1.75    0.082    -.0230062    .3823573

       _cons |   2.493681    .5905397     4.22    0.000     1.326224    3.661137
-------------+----------------------------------------------------------------
cumavrg
     tawjihi |   .0442983    .0069168     6.40    0.000     .0306242    .0579724

     english |   .0017713    .0021864     0.81    0.419    -.0025511    .0060937

      budget |   -.003879    .0026484    -1.46    0.145    -.0091148    .0013568

     absents |  -.0102067    .0059056    -1.73    0.086    -.0218817    .0014682

        prog |

          2  |  -.1888269    .1151785    -1.64    0.103    -.4165268    .0388731

          3  |   .1500583    .1087725     1.38    0.170    -.0649773     .365094

       _cons |  -1.121599    .6265346    -1.79    0.076    -2.360215    .1170174
------------------------------------------------------------------------------

. test tawjihi

 ( 1)  [self_concept]tawjihi = 0

 ( 2)  [motivation]tawjihi = 0

 ( 3)  [cumavrg]tawjihi = 0

       F ( 3, 141) =  14.17,        Prob > F =    0.0000

. test english

 ( 1)  [self_concept]english = 0

 ( 2)  [motivation]english = 0

 ( 3)  [cumavrg]english = 0
```

```
    F ( 3, 141) = 0.68,            Prob > F =    0.5658

. test budget

 ( 1)  [self_concept]budget = 0

 ( 2)  [motivation]budget = 0

 ( 3)  [cumavrg]budget = 0

          F ( 3, 141) = 1.20,         Prob > F =    0.3108

. test absents

 ( 1)  [self_concept]absents = 0

 ( 2)  [motivation]absents = 0

 ( 3)  [cumavrg]absents = 0

        F ( 3, 141) = 1.12,         Prob > F =    0.3448
```

Note that the three responses are controlled by only one explanatory variable which is tawjihi average, and they are not affected by english score or budget or absents.

Now, we want to test the overall effect of the study program on the three responses.

```
  . test[cumavrg]2.prog [cumavrg]3.prog

 ( 1)  [cumavrg]2.prog = 0

 ( 2)  [cumavrg]3.prog = 0
```

```
         F ( 2, 141) = 4.79,            Prob > F =    0.0098

. test[self_concept]2.prog[self_concept]3.prog

 ( 1)  [self_concept]2.prog = 0

 ( 2)  [self_concept]3.prog = 0

            F ( 2, 141) = 0.19,           Prob > F =    0.8301

. test[motivation]2.prog[motivation]3.prog

 ( 1)  [motivation]2.prog = 0

 ( 2)  [motivation]3.prog = 0

      F ( 2, 141) = 1.81,               Prob > F =    0.1675
```

Note that the cumulative average for a student from An-Najah university depends on the student’s program (economics, engineering, or science). But his self concept and achievement motivation is not related with the program.

4.5 Case Study 5: Data Analysis of Arab American University

In this section we will do similar analysis as in section 4.4, but here we collect data from Arab American University. The same set of dependent and independent variables will be used. Data for 111 students from Arab American University has been collected and analyzed.

```
. summarize self_concept motivation cumavrg tawjihi english budget absents

    Variable |     Obs          Mean        Std. Dev.       Min        Max

  ---------------+------------------------------------------------------------------------

self_concept |    111       2.874234        .2357331        2.23        3.37

  motivation |    111       2.655045        .4615663       1.31        4.07

     cumavrg |   111        2.531622      .6612475        1.55        3.99

       tawjihi |   111      80.79009       9.563747         60           99

       english |   111     61.47748         15.24279        34           98

  ---------------+------------------------------------------------------------------------

        budget |  111      36.88288         22.85637          5          160

       absents |  111      3.567568       5.941566          0           35

. tabulate prog

          prog |     Freq.       Percent          Cum.

    ------------+-----------------------------------------------

   economic |      55           49.55           49.55

engineering |      33            29.73          79.28

     science |      23          20.72         100.00

    -----------+-----------------------------------------------

        Total |     111          100.00

. manova self_concept motivation cumavrg= c.tawjihi c.english c.budget c.absents prog

                          W = Wilks' lambda      L = Lawley-Hotelling trace

                          P = Pillai's trace     R = Roy's largest root

  Source |    Statistic         df         F(df1,    df2) =          F           Prob>F

 -----------+------------------------------------------------------------------------------------------

    Model | W   0.3656       6         18.0        289.0          6.86        0.0000 a

             | P   0.6824                  18.0       312.0           5.10        0.0000 a
```

```
           | L   1.6059            18.0     302.0        8.98      0.0000 a
           | R   1.5233             6.0     104.0       26.40      0.0000 u
           |---------------------------------------------------------------------------
  Residual |                104
-----------+---------------------------------------------------------------------------
   tawjihi | W   0.7638      1       3.0     102.0       10.52      0.0000 e
           | P   0.2362              3.0     102.0       10.52      0.0000 e
           | L   0.3093              3.0     102.0       10.52      0.0000 e
           | R   0.3093              3.0     102.0       10.52      0.0000 e
           |---------------------------------------------------------------------------
   english | W   0.8933      1       3.0     102.0        4.06      0.0090 e
           | P   0.1067              3.0     102.0        4.06      0.0090 e
           | L   0.1195              3.0     102.0        4.06      0.0090 e
           | R   0.1195              3.0     102.0        4.06      0.0090 e
           |---------------------------------------------------------------------------
    budget | W   0.9600      1       3.0     102.0        1.42      0.2426 e
           | P   0.0400              3.0     102.0        1.42      0.2426 e
           | L   0.0416              3.0     102.0        1.42      0.2426 e
           | R   0.0416              3.0     102.0        1.42      0.2426 e
           |---------------------------------------------------------------------------
   absents | W   0.9632      1       3.0     102.0        1.30      0.2784 e
           | P   0.0368              3.0     102.0        1.30      0.2784 e
           | L   0.0383              3.0     102.0        1.30      0.2784 e
           | R   0.0383              3.0     102.0        1.30      0.2784 e
           |---------------------------------------------------------------------------
      prog | W   0.9569      2       6.0     204.0        0.76      0.6047 e
```

```
              | P   0.0435              6.0        206.0      0.76      0.5990 a

              | L   0.0445              6.0        202.0      0.75      0.6104 a

              | R   0.0258              3.0        103.0      0.89      0.4508 u

              |---------------------------------------------------------------------------------

     Residual |                 104

   -----------+---------------------------------------------------------------------------------

        Total |                 110

         ---------------------------------------------------------------------------------

. mvreg

Equation         Obs   Parms      RMSE          "R-sq"          F              P

------------------------------------------------------------------------------------------

self_concept     111      7         .2324416       0.0808       1.522901        0.1779

motivation        111      7         .4665309       0.0341        .6118812       0.7203

cumavrg           111      7         .4492732      0.5636        22.38111        0.0000

------------------------------------------------------------------------------------------

            |      Coef.        Std. Err.         t          P>|t|          [95% Conf. Interval]

-----------+---------------------------------------------------------------------------------

self_concept

    tawjihi |   .0012131      .0034369       0.35       0.725          -.0056025      .0080286

    english |   .0020039      .0020537       0.98       0.331          -.0020687      .0060766

    budget |  -.0009854      .0010534      -0.94       0.352          -.0030744      .0011036

   absents |  -.0073005     .0039917      -1.83       0.070          -.0152162      .0006153

        prog |

          2  |   -.041195      .0592177       -0.70       0.488          -.1586258      .0762359

          3  |   -.105489      .0692771        -1.52      0.131          -.242868          .03189

     _cons |   2.74953       .2242891       12.26      0.000           2.304756       3.194303
```

```
------------+-------------------------------------------------------------------------------------
motivation
    tawjihi |   .0028553     .0068982       0.41      0.680        -.0108241     .0165347
    english |  -.004791      .004122       -1.16      0.248        -.0129652     .0033831
     budget |  -.0007213     .0021144      -0.34      0.734        -.0049141     .0034716
    absents |  -.0037589     .0080118      -0.47      0.640        -.0196465     .0121287
       prog |
          2  |  -.0028096    .1188551      -0.02      0.981        -.2385038     .2328846
          3  |  -.1162751    .1390452      -0.84      0.405        -.3920071     .1594568
      _cons |   2.783846     .4501682       6.18      0.000         1.891146     3.676547
------------+-------------------------------------------------------------------------------------
cumavrg
    tawjihi |   .0362646     .006643        5.46      0.000         .0230911     .049438
    english |   .0102503     .0039696       2.58      0.011         .0023785     .0181221
     budget |  -.0032896     .0020361      -1.62      0.109        -.0073273     .0007482
    absents |  -.0026304     .0077154      -0.34      0.734        -.0179303     .0126695
       prog |
          2  |  -.1143693    .1144585      -1.00      0.320        -.3413448     .1126062
          3  |   .0672728    .1339017       0.50      0.616        -.1982593     .332805
      _cons |  -.8775826     .4335158      -2.02      0.045        -1.737261    -.0179045
--------------------------------------------------------------------------------------------------

. test tawjihi

 ( 1)  [self_concept]tawjihi = 0

 ( 2)  [motivation]tawjihi = 0

 ( 3)  [cumavrg]tawjihi = 0
```

```
         F ( 3, 104) = 10.72,         Prob > F =    0.0000

. test english

 ( 1)  [self_concept]english = 0

 ( 2)  [motivation]english = 0

 ( 3)  [cumavrg]english = 0

              F ( 3, 104) = 4.14,       Prob > F =    0.0081

. test budget

 ( 1)  [self_concept]budget = 0

 ( 2)  [motivation]budget = 0

 ( 3)  [cumavrg]budget = 0

              F ( 3, 104) = 1.44,         Prob > F =    0.2345

. test absents

 ( 1)  [self_concept]absents = 0

 ( 2)  [motivation]absents = 0

 ( 3)  [cumavrg]absents = 0

        F ( 3, 104) = 1.33,            Prob > F =    0.2700
```

Note that the three responses are controlled by two explanatory variables which are tawjihi average, and English score. And they are not affected by budget or absents.

```
  . test[cumavrg]2.prog [cumavrg]3.prog

   ( 1)  [cumavrg]2.prog = 0

   ( 2)  [cumavrg]3.prog = 0

         F ( 2, 104) = 1.04,            Prob > F =    0.3575

. test[self_concept]2.prog [self_concept]3.prog

 ( 1)  [self_concept]2.prog = 0

 ( 2)  [self_concept]3.prog = 0

         F ( 2, 104) = 1.16             Prob > F =    0.3175

. test[motivation]2.prog [motivation]3.prog

 ( 1)  [motivation]2.prog = 0

 ( 2)  [motivation]3.prog = 0

          F ( 2, 104) = 0.42,           Prob > F =    0.6578
```

These tests show that the self concept, the achievement motivation, and the cumulative average for a student in Arab American University is not affected by his program or specialization.

```
. test[motivation]tawjihi= [cumavrg]tawjihi
```

```
 ( 1)  [motivation]tawjihi - [cumavrg]tawjihi = 0

       F ( 1, 104) = 11.87,            Prob > F =    0.0008
```

This test indicates that the tawjihi average doesn’t influence achievement motivation in the same way that it influences cumulative average.

4.6 Case Study 6: Data Analysis of Alquds Open University

In this section we will analyze data which has been collected from Alquds Open University. The same set of dependent and independent variables will be used. Our sample is of 91 students.

```
. summarize self_concept motivation cumavrg tawjihi english budget absents

    Variable |    Obs          Mean            Std. Dev.          Min          Max
-------------+--------------------------------------------------------------------
self_concept |     91        2.813297         .3385231            1.6          3.87
  motivation |     91        2.700879         .5484415            1.57         4.04
     cumavrg |     91        2.758132         .2991763            1.65         3.46
     tawjihi |     91        72.32418          9.588029           55.6         95.7
     english |     91        64.14286         14.74492            36           92
-------------+--------------------------------------------------------------------
      budget |     91        32.14286         24.58707            10           200
     absents |     91        2.461538          4.450487            0           28
----------------------------------------------------------------------------------

. tabulate prog

    program |     Freq.     Percent        Cum.
------------+-----------------------------------
   economic |        57       62.64        62.64
```

```
engineering |      13       14.29      76.92

    science |      21       23.08     100.00

------------+-----------------------------------

      Total |      91      100.00

. manova self_concept motivation cumavrg= c.tawjihi c.english c.budget c.absents prog

                           Number of obs =      91

                      W = Wilks' lambda         L = Lawley-Hotelling trace

                      P = Pillai's trace        R = Roy's largest root

     Source |    Statistic     df      F(df1,    df2) =       F       Prob>F

-----------+---------------------------------------------------------------------

      Model | W   0.5561       6        18.0      232.4       2.98     0.0001 a

            | P   0.4955                18.0      252.0       2.77     0.0002 a

            | L   0.7082                18.0      242.0       3.17     0.0000 a

            | R   0.5618                 6.0       84.0       7.86     0.0000 u

            |---------------------------------------------------------------------

   Residual |                  84

-----------+---------------------------------------------------------------------

    tawjihi | W  0.6969        1          3.0       82.0      11.89     0.0000 e

            | P   0.3031                 3.0       82.0      11.89     0.0000 e

            | L   0.4350                 3.0       82.0      11.89     0.0000 e

            | R   0.4350                 3.0       82.0      11.89     0.0000 e

            |---------------------------------------------------------------------

    english | W  0.9548        1          3.0       82.0       1.29     0.2822 e

            | P   0.0452                 3.0       82.0       1.29     0.2822 e

            | L   0.0473                 3.0       82.0       1.29     0.2822 e

            | R   0.0473                 3.0       82.0       1.29     0.2822 e
```

```
           |-----------------------------------------------------------------------------------
    budget | W   0.9383      1      3.0      82.0      1.80     0.1540 e
           | P   0.0617             3.0      82.0      1.80     0.1540 e
           | L   0.0658             3.0      82.0      1.80     0.1540 e
           | R   0.0658             3.0      82.0      1.80     0.1540 e
           |-----------------------------------------------------------------------------------
   absents | W   0.9761      1      3.0      82.0      0.67     0.5742 e
           | P   0.0239             3.0      82.0      0.67     0.5742 e
           | L   0.0244             3.0      82.0      0.67     0.5742 e
           | R   0.0244             3.0      82.0      0.67     0.5742 e
           |-----------------------------------------------------------------------------------
      prog | W   0.9475      2      6.0     164.0      0.75     0.6124 e
           | P   0.0528             6.0     166.0      0.75     0.6101 a
           | L   0.0551             6.0     162.0      0.74     0.6149 a
           | R   0.0490             3.0      83.0      1.36     0.2620 u
           |-----------------------------------------------------------------------------------
  Residual |                 84
-----------+-----------------------------------------------------------------------------------
     Total |                 90
-----------------------------------------------------------------------------------------------
           e = exact, a = approximate, u = upper bound on F

. mvreg

Equation          Obs  Parms     RMSE        "R-sq"          F            P
-----------------------------------------------------------------------------------------------

self_concept       91     7     .3393834     0.0619     .9240481      0.4820

motivation         91     7     .5409383     0.0920     1.419006      0.2170
```

```
cumavrg            91     7        .2524411       0.3355       7.068105        0.0000

-----------------------------------------------------------------------------------------------

             |     Coef.       Std. Err.        t          P>|t|        [95% Conf. Interval]

-------------+---------------------------------------------------------------------------------

self_concept

    tawjihi |   .0012982      .003942       0.33        0.743        -.006541       .0091374

    english |   .0006519       .002601       0.25        0.803        -.0045205      .0058243

     budget |  -.0024512      .0014657    -1.67        0.098        -.0053659      .0004636

    absents |   .0031719      .0082124     0.39        0.700        -.0131593      .0195031

        prog |

           2  |   .1308551       .1050235     1.25        0.216        -.0779957      .3397058

           3  |  -.0060587       .0890248    -0.07        0.946        -.1830943      .1709769

      _cons |   2.731275        .2888951     9.45        0.000         2.156775      3.305774

-------------+---------------------------------------------------------------------------------

motivation

    tawjihi |    .0088431        .0062832     1.41         0.163        -.0036517      .0213378

   english |  -.0054313        .0041457    -1.31         0.194        -.0136755      .0028129

   budget |   -.000708         .0023362    -0.30         0.763        -.0053538      .0039378

 absents |   .0177188         .0130896     1.35         0.179        -.0083113      .0437488

      prog |

         2  |    .3296563        .1673954     1.97         0.052        -.0032279     .6625405

         3  |    .0942573        .1418953     0.66         0.508        -.1879171     .3764317

    _cons |   2.319986         .4604657     5.04         0.000         1.404299      3.235672

-------------+---------------------------------------------------------------------------------

cumavrg

    tawjihi |   .0164738        .0029322      5.62          0.000          .0106429      .0223048
```

```
  english |   .0021727      .0019347      1.12         0.265          -.0016746      .00602

  budget |  -.0014283      .0010902     -1.31         0.194          -.0035964      .0007397

absents |   .0007444       .0061086      0.12         0.903          -.0114031     .0128919

     prog |

         2   |  -.048863       .0781189     -0.63        0.533           -.2042109     .106485

           3  |  -.0378827      .0662187     -0.57       0.569          -.1695658      .0938004

     _cons |  1.487113       .2148867       6.92       0.000           1.059787      1.914439
------------------------------------------------------------------------------------------------------

. test tawjihi

 ( 1)  [self_concept]tawjihi = 0

 ( 2)  [motivation]tawjihi = 0

 ( 3)  [cumavrg]tawjihi = 0

       F ( 3, 84) = 12.18,              Prob > F =    0.0000

. test english

 ( 1)  [self_concept]english = 0

 ( 2)  [motivation]english = 0

 ( 3)  [cumavrg]english = 0

           F ( 3, 84) = 1.33,             Prob > F =    0.2717

. test budget

 ( 1)  [self_concept]budget = 0
 ( 2)  [motivation]budget = 0

 ( 3)  [cumavrg]budget = 0

       F ( 3, 84) = 1.84,            Prob > F =    0.1458

. test absents

 ( 1)  [self_concept]absents = 0

 ( 2)  [motivation]absents = 0

 ( 3)  [cumavrg]absents = 0

        F ( 3, 84) = 0.68,           Prob > F =    0.5642
```

Note that the three responses are controlled by one explanatory variable which is tawjihi average. And they are not affected by English score or budget or absents.

```
  . test[cumavrg]2.prog [cumavrg]3.prog

 ( 1)  [cumavrg]2.prog = 0

 ( 2)  [cumavrg]3.prog = 0

           F ( 2, 84) = 0.29,        Prob > F =    0.7457

. test[self_concept]2.prog [self_concept]3.prog

 ( 1)  [self_concept]2.prog = 0

 ( 2)  [self_concept]3.prog = 0
```

```
      F ( 2, 84) = 0.84              Prob > F =    0.4360

. test[motivation]2.prog [motivation]3.prog

 ( 1)  [motivation]2.prog = 0

 ( 2)  [motivation]3.prog = 0

        F ( 2, 84) = 1.97,           Prob > F =    0.1463
```

Note that the cumulative average, self concept, and achievement motivation for a student from Alquds Open University do not depend on student’s program (economics, engineering, or science).

Conclusion

This thesis addresses the multivariate multiple linear regression model, in which a set of dependent variables are controlled or affected by a set of independent variables. The method of least squares has been used for estimating the multivariate multiple regression model. This method is one of the most commonly used prediction techniques. The main objective of this method is to find the m vectors of parameters which minimize the error sum of squares

$$SSE = \Sigma\varepsilon_i^2 = \Sigma\varepsilon^T\varepsilon$$

where m represents the number of dependent variables. After estimating the regression model, multivariate analysis of variance (MANOVA) was used to test the usefulness of our estimated model.

The multivariate multiple regression model was applied to simulated data from different distributions, such as normal, logistic, and exponential distribution. Also, a case study was constructed to study the effect of some independent variables (tawjihi average, English score level exam, budget, and absents) on three response variables (self concept, achievement motivation, and the cumulative average) for university students from three collages or specializations; economics, engineering, and science. Data were collected from three universities in Palestine: An-Najah National University, Arab American University and Alquds University. The case study was divided into six sub-cases:

Case 1: We studied the relationship among the explanatory and response variables regardless the study program of the student and the university he is studying in. And we found that the four predictors, tawjihi average, English score, budget, and absents affect the responses well.

Case 2: We studied the relationship among the explanatory and response variables but here we were interested in the study program regardless the university. The results showed that the program affects the cumulative average but not too much, and it has no effect on self concept or achievement motivation for a student. The other four predictors had more effect and control the responses very well.

Case 3: We studied the relationship among the explanatory and response variables but here we were interested in the university that the student is studying in regardless the study program. The results showed that the university affect the cumulative average, but has no effect on his self concept or achievement motivation for students.

Case 4: We studied the relationship among the explanatory and response variables for a sample from An-Najah National University students who study economics, engineering, and science. The results showed that only the tawjihi average controls the responses, the other predictors had no effect. Also, the cumulative average for a student from An-Najah University depends on the study program. But his/her self concept and achievement motivation was not related with the study program.

Case 5: We studied the relationship among the explanatory and response variables for a sample from Arab American University students who study economics, engineering, and science. The results showed that the three responses were controlled by two predictors which were tawjihi average, and English score. And they were not affected by budget or absents. We found, also, that self concept, achievement motivation, and cumulative average for a student from The Arab American University were not affected by the study program.

Case 6: We studied the relationship among the explanatory and response variables for a sample from Alquds University students who study economics, engineering, and science. The results showed that just the tawjihi average affect the response variables. And the English score, budget, and absents had no effect on the three responses. Also, we found that the cumulative average, self concept, and achievement motivation for a student from Alquds University do not depend on the study program.

References

[1] Agarwal P. Ravi, Sen K. Syamal**. Creators of Mathematical and Computational Sciences,** 218-219.

[2] Alvin C. Rencher, Schaalje G. Bruce, **Linear Models In Statistics**, Department of Statistics, Brigham Young University, Provo, Utah. second edition.

[3] Baker L. Samuel, **Non-Linear Regression** © 2006-2008.

[4] Bates, D. M. & D. G. Watts, **Nonlinear Regression Analysis and Its Applications**. New York: Wiley 1988.

[5] Brown H. Scott, **Multiple Linear Regression Analysis**: A Matrix Approach with MATLAB, Auburn University Montgomery. Alabama Journal of Mathematics, Spring/Fall 2009.

[6] Byrkit R. Donald, **Statistics Today**, **A Comprehensive Introduction**, university of West Florida, January 1987.

[7] Carey, Gregory. **"Multivariate Analysis of Variance (MANOVA): I. Theory"**. Retrieved March 22, 2011.

[8] Fox J., **Applied Regression Analysis**, Linear Models and Related Methods. (1997).

[9] Fox J., **Nonlinear Regression and Nonlinear Least Squares, Appendix to An R and S-PLUS Companion to Applied Regression**. January 2002.

[10] Galton F., (1886). **Regression towards Mediocrity in Hereditary Stature,** Journal of the Anthropological Institute, 15, 246-263.

[11] Galton F., **"Kinship and Correlation (reprinted 1989)".** Statistical Science (Institute of Mathematical Statistics), 1989, 4 (2): 80–86.

[12] Gauss C.F., **Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum**, 1809.

[13] Gauss C.F., **Theoria combinationis observationum erroribus minimis obnoxiae**, 1821-1823.

[14] Gordon K. Smyth**, Nonlinear regression**, Encyclopedia of Environmetrics (ISBN 0471 899976), 2002, Volume 3, pp 1405–1411.

[15] Hair F. Joseph, Anderson E. Rolph, Tatham L. Ronald, Black C. William, **Multivariate Data Analysis**, fifth edition, chapter 6, pp. 326- 352.

[16] Hosmer W. David, Jr.‹Stanley Lemeshow‹Rodney X. Sturdivant, **Applied Logistic Regression**, third edition.

[17] Johnson A. Richard, Wichern W. Dean, “**Applied Multivariate Statistical Analysis**”, sixth editon. 360- 398.

[18] Kutner M. H., Nachtsheim C. J., and Neter J. (2004), **"Applied Linear Regression Models",** 4th edition.

[19] Legendre A.M. , **Nouvelles méthodes pour la détermination des orbites des comètes**, Firmin Didot, Paris, 1805.

[20] Pearson, Karl; Yule, G.U.; Blanchard, Norman; Lee,Alice (1903). **"The Law of Ancestral Heredity"**. Biometrika (Biometrika Trust) 2 (2): 211–236.

[21] Rawlings O. John, Pantula G. Sastry, Dickey A. David, **Applied Regression Analysis**: A Research Tool, second edition.

[22] Ritzema H. P., **Drainage Principles and Applications**, second revised edition 1994, Publ. 16, chapter 6, pp. 175-224.

[23] Sen A. & Srivastava M., **Regression Analysis: Theory, Methods, and Applications.**

[24] Sokolowski, J.A., Banks, C.M.. **Principles of Modeling and Simulation**. Hoboken, NJ: Wiley, 2009, p. 6. ISBN 978-0-470-28943-3.

[25] Stigler, Stephen M. (November 1974). , **"Gergonne's 1815 paper on the design and analysis of polynomial regression experiments"**. Historia Mathematica 1 (4): 431–439.

[26] Stockute R., Veaux A., Johnson P., **Logistic Distribution.**

[27] Yule, G. Udny (1897). **"On the Theory of Correlation".** Journal of the Royal Statistical Society (Blackwell Publishing) 60 (4): 812–54.

websites:

[28] http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/

[29] http://en.wikipedia.org/wiki/Linear_regression

[30] http://www.stata.com

[31] http://plato.stanford.edu/entries/simulations-science/#TypComSim

Appendix I

Simulated data from normal distribution

We generate the variables y_1,y_2,y_3,x_1,x_2,x_3,and x_4 for a sample of 20 observation using the following command

```
Variable= random('Normal',mean,st.deviation,20,1)
x=[x1 x2 x3 x4]
y= [y1 y2 y3]
```

```
x=                                               y=

83.6957   71.0379   78.0973   66.9688      -0.4140    0.4136    0.4759

81.0677   70.3079   60.5619   67.7441      -0.4383   -0.5771    1.4122

82.5666   68.0983   67.9601   72.2206       2.0034    0.1440    0.0226

81.7019   76.1529   64.7135   67.7729       0.9510   -1.6387   -0.0479

74.7411   68.6233   67.4555   66.4195      -0.4320   -0.7601    1.7013

78.3344   75.4409   71.2714   68.3627       0.6489   -0.8188   -0.5097

84.8988   78.1483   70.5522   69.3565      -0.3601    0.5197   -0.0029

79.8257   74.9911   67.1571   71.6373       0.7059   -0.0142    0.9199

82.3297   75.3724   73.1080   66.1895       1.4158   -1.1555    0.1498

75.9740   73.4874   72.4896   71.2693      -1.6045   -0.0095    1.4049

80.2581   69.0693   72.5895   70.312        1.0289   -0.6898    1.0341

82.4012   74.8247   68.2975   75.2975       1.4580   -0.6667    0.2916

74.5539   78.8433   74.1943   69.1473       0.0475    0.8641   -0.7777

81.3904   81.9530   72.6428   69.4621       1.7463    0.1134    0.5667

79.2726   73.2792   80.0351   65.3146       0.1554    0.3984   -1.3826

76.2419   68.4907   74.2538   64.4589      -1.2371    0.8840    0.2445
```

```
79.8499   75.6654   74.6277   71.2420       -2.1935    0.1803    0.8084

72.4148   76.5051   70.2119   69.0020       -0.3334    0.5509    0.2130

71.4881   74.0922   64.8465   72.0150        0.7135    0.6830    0.8797

75.2923   70.4044   68.5151   66.4294        0.3174    1.1706    2.0389
```

Appendix II

Simulated data from logistic distribution

We generate the variables y_1,y_2,y_3,x_1,x_2,x_3,and x_4 for a sample of 20 observation using the following command

```
Variable= random('logistic',mean,st.deviation,20,1)
x=[x1 x2 x3 x4]
y= [y1 y2 y3]
```

```
x=                                         |    y=

84.3162   75.7953   74.1845   69.7516      |  4.4466    2.3400   -0.1762

61.7533   74.7636   65.2750   63.1192      |  0.1603   -2.1534   -2.1788

68.0895   83.3824   74.0797   75.1078      |  0.8805    1.0750    5.3748

82.7951   80.5192   83.9875   61.2380      |  7.5838    1.0267   -0.6987

81.6797   79.0670   77.4963   66.4971      | -0.9059    0.2487   -0.8600

80.4180   63.3330   60.5580   65.3315      | -0.3453   -1.4882   -2.7158

83.9726   64.8272   67.8099   70.4946      | -0.1409    0.3939   -0.8557

83.5283   65.6730   67.8574   60.8181      |  1.1745   -0.8476   -3.0240

85.0948   80.5041   73.1079   68.4665      |  1.5042   -1.8650    0.0217

76.3791   86.2247   71.5876   61.4220      | -2.1948   -1.3093    1.1605

83.2479   78.0836   75.2844   61.7295      | -1.5292    2.1422    0.5368

80.9106   67.4693   67.8308   75.1666      | -0.5769   -2.5646   -2.3150

78.3201   78.8321   64.6039   66.4489      | -2.8115   -1.1391   -2.4307

69.1053   66.6515   60.5799   71.6808      |  0.0876   -2.8681    1.2497

85.0668   66.9344   74.8771   83.1993      | -0.6818   -0.2342    2.2556

77.3037   77.3139   64.5952   68.9131      | -1.5460   -4.3079    0.1353
```

```
81.7534   72.1458   68.1817   73.2897       -1.3313    2.1664   -2.0994

84.2100   77.5434   70.8314   74.5691        2.2558   -1.4073    1.5562

71.4206   79.3759   65.1431   68.9157        0.7327   -2.2732   -0.6718

72.3210   76.3435   72.3352   72.5732       -0.1263   -0.8125   -0.8762
```

Appendix III

Simulated data from exponential distribution

We generate the variables y_1,y_2,y_3,x_1,x_2,x_3,and x_4 for a sample of 20 observation using the following command

```
Variable= random('exponential',mean,20,1)
x=[x1 x2 x3 x4]
y= [y1 y2 y3]
```

```
x =

1.7840     6.5873    24.4613    51.6007

180.4220   114.8770    40.7676    95.1524

37.2888    58.6935    19.5216    69.7697

61.7514    20.2162     0.3170    28.8991

82.1781   226.7630     2.6805   124.1980

93.2705     3.8618    43.7754    89.4123

19.3243    18.8731     2.5759   113.2860

0.3357    40.7349   151.0180   114.4070

117.5260   118.5570   207.7020    78.2800

17.2896    48.8081    83.2706     8.9215

114.1470    46.0655    38.1078    52.6877

0.5369     0.4042    44.3142    63.4491

15.4224    10.9779     7.2813   120.3350

60.0241     2.6825    43.0617     2.2098

22.1388    27.1055    58.7562    62.8476

48.7512    63.5303    42.7882    11.8318
```

```
y =

2.4080    1.2329    2.0113

1.1365    0.1621    0.3983

0.6706    0.0925    0.5604

2.8034    0.4474    1.7733

0.3206    1.3650    1.9129

0.5860    2.4229    0.7422

0.6361    0.1764    0.0964

0.1864    0.5366    0.5939

0.1523    0.0533    3.4131

0.2370    2.7964    2.9213

1.1462    0.5368    0.2168

0.7936    1.2549    0.7955

0.2847    0.1891    0.9606

2.2085    1.6556    0.2362

2.2096    0.8152    1.0098

1.3098    0.9329    0.6305
```

```
14.8378    4.7062   23.7366   33.9923        0.6450    0.1905    0.3402

72.1977   51.4531  286.5700   68.3580        0.0277    0.3903    0.1376

182.9780  102.3340   15.5395    9.172        0.3419    1.5721    1.1126

36.8168   64.7926  136.3840   16.9582        1.1652    1.1454    0.4306
```

جامعة النجاح الوطنية

كلية الدراسات العليا

تقدير نماذج الانحدار الخطي متعدد المتغيرات المتعدد وتطبيقات

إعداد

جنان نشأت سعيد كيوان

إشراف

د. محمد نجيب أسعد

د. علي بركات

قدمت هذه الأطروحة استكمالاً لمتطلبات الحصول على درجة الماجستير في الرياضيات بكلية الدراسات العليا في جامعة النجاح الوطنية في نابلس، فلسطين

2015

تقدير نماذج الانحدار الخطي متعدد المتغيرات المتعدد وتطبيقات

إعداد

جنان نشأت سعيد كيوان

إشراف

د. محمد نجيب أسعد

د. علي بركات

الملخص

الانحدار هو أحد أهم الأساليب الاحصائية وأوسعها استخداماً في مختلف العلوم من حيث تحديد العلاقة بين المتغيرات على هيئة معادلة لتقدير معلماتها وقوة واتجاه هذه العلاقة. وقد قامت الباحثة في هذه الرسالة بدراسة أحد أهم نماذج الانحدار وهو الانحدار الخطي متعدد المتغيرات المتعدد، والذي يدرس العلاقة بين عدد من المتغيرات المعتمدة ومدى تاثرها واعتمادها على عدد من المتغيرات المستقلة. حيث قامت الباحثة باستخدام طريقة المربعات الصغرى لتقدير معلمات نموذج الانحدار الخطي متعدد المتغيرات المتعدد واستخدمت تحليل التباين المتعدد (MANOVA) من أجل تقييم النموذج المقدّر واختبار الفرضيات حول المعلمات المقدرة من خلال استخدام عدد من البرامج للحصول على أفضل النتائج مثل: Stata, SPSS, Minitab, Matlab, SAS.

كما وقامت الباحثة باستخدام المحاكاة من خلال برنامج الماتلاب لتوليد بيانات من عدة توزيعات احتمالية واستخدمت البيانات لبناء وتقدير نماذج الانحدار متعدد المتغيرات المتعدد. وأخيراً، قامت الباحثة بتطبيق نموذج الانحدار الخطي متعدد المتغيرات المتعدد على مثال واقعي، حيث قامت بدراسة عينة من طلاب الجامعات في فلسطين شملت كل من: جامعة النجاح الوطنية، الجامعة العربية الأمريكية، وجامعة القدس المفتوحة– فرع نابلس. ودرست العلاقة بين المعدل التراكمي للطالب وبعض العوامل النفسية مثل مفهوم الذات ودافعية الانجاز لدى الطالب، ومدى تأثر هذه العوامل بعدد من المتغيرات المستقلة مثل: معدل التوجيهي، علامة امتحان قبول اللغة الانجليزية عند الالتحاق بالجامعة، المصروف اليومي، الغيابات في الفصل الدراسي،

والكلية الملتحق بها الطالب. وتمت المقارنة بين الجامعات، كما قامت بدراسة العلاقة بين هذه المتغيرات في كل جامعة على حدة. وأظهرت النتائج أن العوامل النفسية لم تتأثر في الغالب بأي من المتغيرات، بينما المعدل التراكمي تأثر بالجامعة بشكل واضح كما تأثر بالكلية الملتحق بها الطالب. وأن العوامل المعتمدة وتحديداً المعدل التراكمي لطلاب جامعة النجاح تأثرت بشكل واضح بمعدل التوجيهي وبالكلية التي يدرس بها الطالب. بينما طلاب الجامعة العربية الأمريكية فقد كان لمعدل التوجيهي وعلامة امتحان اللغة الانجليزية أثر كبير على كل من المعدل التراكمي ومفهوم الذات ودافعية الانجاز لدى الطالب ولم تتأثر هذه العوامل بالكلية التي يدرس بها. أما طلاب جامعة القدس المفتوحة، فإن العامل الوحيد الذي أثر على كل متغير من المتغيرات الثلاثة كان معدل التوجيهي، ولم يكن لأي عامل آخر أي تأثير على المعدل التراكمي أو مفهوم الذات أو دافعية الانجاز.