ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Determinants and Risk Factors of Adult Onset Diabetes and Hypertension in the Gaza Strip; Application of the Multivariate Regression Model

*** Mahmoud K. Okasha[1], and Hussien A. Abu Halima[1]**

1. Department of Applied Statistics, Al-Azhar University - Gaza, Palestine.

Manuscript Info

Manuscript History:

Received: 15 November 2015
Final Accepted: 26 December 2015
Published Online: January 2016

Key words:
Multivariate Methods,
M-estimate,
Maximum Likelihood Estimator,
MANOVA,
Normality Tests.

****Corresponding Author***

Mahmoud K. Okasha.

Abstract

Diabetes and hypertension are increasing rapidly around the globe, including in Palestine. These are highly related diseases, so it is useful to study them simultaneously. This paper aimed to discuss the application of a multivariate linear regression model to identify the risk factors that have the greatest effects on the severity of adult onset diabetes and hypertension on Palestinian patients. It studied the effects of several medical characteristics of Palestinian patients in Gaza City, as independent variables, on three measurements indicating patients' severity of diabetes and hypertension, as dependent variables. A real medical dataset of a random sample of 518 patients and 14 variables obtained from health clinics in Gaza City was used. The parameters of the model were estimated, and the multivariate analysis of variance was computed. For the sake of comparison, robust nonparametric methods based on M-estimators were also used, and similar results were obtained. The study concluded that the significant risk factors that had the greatest effects on the severity of diabetes and hypertension, according to both methods, were patients' age, weight, cholesterol/HDL ratio, and triglycerides.

Introduction:-

Okasha (2014) discussed the application of logistic regression analyses to identify the risk factors for the prevalence of hypertension in Palestine. He concluded that significant risk factors for hypertension include age, sex, smoking status, peptic ulcer disease, arthritis rheumatism, high cholesterol, hyperthyroidism (nonmalignant), fasting blood sugar, microalbuminurea, and locality type. This paper investigates the application of the multivariate linear regression model to identify risk factors for the severity of adult onset diabetes and hypertension simultaneously.

The multivariate linear model is a generalization of the univariate linear regression model and accommodates two or more response variables. This model allows for the generalization of univariate estimation results to a multivariate model with very few complications. In contrast, in such a situation, measures of association and hypothesis testing become far more complicated to derive and discuss. Johnson and Wichern (1992) defined multivariate linear regression as modeling the relationship between m responses and a single set of predictor variables, where each response variable is assumed to follow its own linear regression model.

There are several texts that treat multivariate linear models and multivariate analyses of variance (MANOVAs) extensively. The theory is presented in Rao (1973). Fox et al. (2007) described the implementation of multivariate regression methods in R, as well as their extension (e.g., from two to three dimensions and by scaling hypothesis ellipses and ellipsoids relative to error in a natural manner). The methods are incorporated in the plots package for R. Al-Marshadi (2010) considered the analysis of a multivariate regression experiment that is used frequently in a variety of research applications. He used a simulation study to compare five model selection criteria in terms of their ability to identify the right multivariate regression model with the right covariance structure and the right multivariate model structure. Liyun (2012) presented a new method to estimate the multivariate linear heteroscedastic regression model based on multivariate local polynomial estimation with a non-parametric technique. Olive (2013-a) showed that a robust multivariate regression estimator is asymptotically equivalent to the classical multivariate linear regression estimator because the probability that the robust estimator is equal to the classical estimator goes to one as the sample size n→ ∞ for a large class of iid zero mean error distributions. A good introduction to the MANOVA approach to repeated measures may be found in the work of O'Brien and Kaiser (1985), and many statistical tests and graphical methods are available to check the multivariate normality assumption. Burdenski (2000) reviewed several statistical and practical approaches, including the Q-Q plot, the box plot, the stem-and-leaf plot, and the Shapiro-Wilk and Kolmogorov-Smirnov tests to evaluate univariate normality; the contour and perspective plots to assess bivariate normality; and the chisquare Q-Q plot to check multivariate normality.

The basic assumptions of the multivariate regression model are normality and the homogenous variances of the residuals, conditional on the predictor's common covariance structure across observations and independent observations. Kakizawa (2009) examined the problem of testing the assumptions of the multivariate linear regression model. Four multivariate tests—Wilk's lambda, Pillai's trace, Hotelling-Lawley's trace and Roy's largest root—were examined. Testing the assumption of multivariate normality is an important issue, and many approaches, including graphical methods, are available to check it.

Research Issue:-

Hypertension, or high blood pressure, and diabetes are prevalent and important risk causes of premature mortality in the Gaza Strip. The primary goal of using multivariate linear regression in the analysis of this paper's dataset is to identify the most important risk factors affecting diabetes and blood pressure in Palestinian patients in the Gaza Strip. The theoretical aspects of the multivariate linear regression model and its assumptions are studied in detail in this paper. All methods are applied to the analysis of a real dataset of patients of hypertension and diabetes in the Gaza Strip.

Data Description:-

The data used in this study were obtained from the principal clinics of both the United Nations (UNRWA) and the government, and they comprise patients of adult onset diabetes and hypertension in the Gaza Strip from 2012 to 2015. The data include information on the prevalence of obesity, diabetes, and other cardiovascular risk factors among the patients. Of the sample patients, 427 (82.4%) were diagnosed in UNRWA clinics, whereas 91 (17.6%) were diagnosed in governmental clinics. Table 1 lists all of the variables and their names, labels, and measurement units.

Table 1: Names of the variables, their labels, and their measurement units.

Name of variable	Labels	Units
Stab.glu	Stabilized glucose	mg/dl
BP.S	First systolic blood pressure	mg/dl
BP.D	First diastolic blood pressure	mg/dl
Chol	Total cholesterol	mg/dl
HDL	High density lipoprotein	mg/dl
LDL	Low density lipoprotein	mg/dl
Ratio	Cholesterol/HDL ratio	mg/dl
Age	Age of patient	years
Gender	Gender of patient	
Weight	Weight of patient	Kg
Creatinine	Blood creatinine	mg/dl
BMI	Body mass index	Kg/ m^2
TG	Triglycerides	mg/dl
Urea	UREA	mg/dl

The data consist of 14 variables for 518 patients, 278 (53.7%) of whom were females. Among the variables, there are three indicators of the existence of diabetes and hypertension (stabilized glucose, median systolic blood pressure and median diastolic blood pressure); these are classified as dependent variables. There are also eleven variables that are thought to be risk factors for diabetes and hypertension diseases; these are classified as independent variables.

The Multivariate Linear Regression Model:-

A multivariate linear regression model is, ultimately, a combination of several univariate linear regression models. According to Jonson and Wichern (2007), a multivariate linear regression model is defined as the relationship between q response variables $Y_1, Y_2, \dots, Y_q$ and a single set of p-predictor variables $X_1, X_2, \dots, X_p$. Thus, we have a vector of dependent variables Y measured in relation to each set of independent variables X. Thus, we can write a regression model for each response variable as follows:

$$Y_{n\times q} = X_{n\times(p+1)}\ \beta_{(p+1)\times q} + \varepsilon_{n\times q} \qquad (1)$$

where $Y = [\ Y_1\ ,\ Y_2\ ,\ \dots\ ,\ Y_q\ ]^T$ is the dependent variable and X is the independent variable, which takes the form $X = [\ X_1\ , X_2\ ,\ \dots\ ,\ X_p\ ]^T$. We also have an error term, which can be written as $\varepsilon = [\ \varepsilon_1\ ,\ \varepsilon_2\ ,\ \dots\ ,\ \varepsilon_q\ ]^T$. The errors terms ($\varepsilon$) are assumed to have expectations and variances equal to $E(\varepsilon) = E\left[\ \varepsilon_1\ , \varepsilon_2\ , \dots, \varepsilon_q\ \right]^T = 0$ and $\text{Var}\left(\varepsilon\right) = \Sigma_{q\times q}$, respectively. The i^{th} observation can then be written as follows:

$$Y_{i\ j} = \beta_{j\,0} + \beta_{j1}X_{i\,1} + \beta_{j\,2}X_{i\,2} + \dots + \beta_{jp}X_{i\ p} + \varepsilon_{i\ j}\ ,\ \ i = 1,2,...,p;\ j = 1,2,...,q \qquad (2)$$

The multivariate linear regression model requires the following assumptions:

- The residuals $\varepsilon_{n\times q}$ should follow the multivariate normal distribution.
- The residuals should be independent.
- The residuals should have equal variances across observations, conditional on predictors (homogeneity assumption). (Latra et al., 2010)

Estimation of the parameters of the multivariate regression model:-

Consider the multivariate linear regression model $Y_j = X\ \beta_j + \varepsilon_j$ with the fitted value $\hat{Y} = X\ \hat{\beta}_{(\ p+1\ )\times q}$, where $\hat{\beta}$ is the estimate of β. For any choice of parameters $\hat{\beta} = \left(\ b_1\ , b_2\ , \dots, b_q\right)$, the resulting matrix of error terms is $\varepsilon_j = Y_j - X\ \beta_j$. Johnson and Wichern (2007) determined the least squares estimator $\hat{\beta}_.$ to be $\hat{\beta}_j = \left(X^TX\right)^{-1}X^TY_j$

The least squares estimator $\hat{\beta} = \left[\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_q\right]$ of the multivariate regression model with $E\left(\varepsilon_i\right) = X\beta$, $\text{Cov}\left(\varepsilon_i, \varepsilon_j\right) = \sigma_{ij}I$ for $i, j = 1,2,...,q$ with full $rank\ (X) = p+1 < n$, has many properties. The first and most important property is that the least squares estimator $\hat{\beta}$ is unbiased; that is, $E\left(\hat{\beta}\right) = \beta$ (Johnson & Wichern, 2002). The second property of the least squares estimator $\hat{\beta}$ is (Sawyer, 2010):

$$\text{Cov}\left(\ \hat{\beta}\ \right) = \left(X^TX\right)^{-1} \otimes \Sigma$$

There are two other useful properties regarding the estimated vectors of residuals $\hat{\varepsilon}_j$ (Johnson & Wichern, 2007). These are:

1) $E\left(\ \hat{\varepsilon}_j\ \right) = 0$ (3)

2) $$E\left[\ \hat{\varepsilon}_{j}^{T}\ \hat{\varepsilon}_{k}\ \right] = \left(\ n\ - p - 1\right)\sigma_{jk} \qquad (4)$$

For the multivariate linear regression model in (1), assuming that ε ~ $N_p(0,\Sigma)$, if $\operatorname{rank}(x) = p+1$ and $n \geq (p+1)+q$, then both the least squares and the maximum likelihood estimators of β follow the multivariate normal distribution, with:

$$E\left(\hat{\beta}\right) = \beta \quad , \quad \operatorname{cov}\left(\hat{\beta}_i\ , \hat{\beta}_k\right) = \sigma_{ik}\left(X^{T}X\right)^{-1} \qquad (5)$$

Furthermore, the maximum likelihood estimator of β is independent of the maximum likelihood estimator of the positive definite matrix $\sum$.

Estimation of μ and ∑:-

Let $x_1, x_2, \ldots, x_n$ be a random sample, where x follows the multivariate normal distribution with mean μ and covariance matrix $\sum$; that is, $x \approx N_p(\mu,\Sigma)$. To obtain the maximum likelihood estimates of parameters μ and Σ, given the sample data X, the multivariate normal likelihood function is given by:

$$L\left(\mu,\Sigma|X\ \right) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left[-\frac{1}{2}\left[\sum_{i=1}^{n}(x_i-\mu)\Sigma^{-1}(x_i-\mu)\right]\right]$$

$$= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} exp\left[-\frac{1}{2}\operatorname{tr}\left\{\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\bar{x})(\ x_i-\bar{x})' + n(\ \bar{x}-\mu)\ \Sigma^{-1}(\bar{x}-\mu)\right]\right\}\right]$$

Rencher and Chrestensen (2012) showed that the maximum likelihood estimators of μ and ∑ are given by, respectively:

$$\hat{\mu} = \bar{x} \quad \text{and} \qquad (6)$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)' = \frac{(n-1)}{n}S \qquad (7)$$

where $S = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)'$ is the unbiased estimator of $\sum$ (Kim & Timm, 1991) .

Analogous to the univariate case, the sampling distribution of the sample covariance can be summarized as follows (Johnson & Wichern, 2007):

1) The sample mean $\bar{x}$ has the normal distribution $\bar{x} \approx N_p\left(\mu, \frac{1}{n}\Sigma\right)$

2) $$(n-1)S = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)' \approx W_{n-1,\Sigma} \qquad (\ 8)$$

3) $Z_i = X_i - \bar{X} : N_p(0,\Sigma)$ and $(n-1)S = \sum_{i=1}^{n} Z_i Z_i'$ are distributed as a Wishart distribution with n-1 degree of freedom $W_{n-1}\left[(n-1)S\ |\Sigma\right]$. (9)

The probability density function of the Wishart distribution with n degrees of freedom is

$$f\left(M|\Sigma\right)=\frac{|M|^{(n-p-2)/2}\,exp\left[\dfrac{-tr\left(M\Sigma^{-1}\right)}{2}\right]}{2^{p(n-1)/2}\,\pi^{p(n-1)/4}\,|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma\left(\dfrac{1}{n}(n-i)\right)} \qquad (10)$$

where M is a positive definite matrix and $\Gamma(.)$ is the gamma function. Marden (2013) showed that the Wishart distribution possesses the following properties:

1) If $M_1 \approx W_{m1}(n_1,\Sigma)$, $M_2 \approx W_{m2}(n_2,\Sigma)$ and M_1 and M_2 are independent, then $M_1+M_2 \approx W_{m1+m2}(n_1+n_2,\Sigma)$.
2) If $A \approx W_m\left(A,\Sigma\right)$, then $CAC' \approx W_m\left(CAC', C\Sigma\Sigma'\right)$

The multivariate implications of the law of large numbers include that $\bar{X}$ converges in probability to μ as n increases without bound; that is, $P\left[-\varepsilon<\bar{X}-\mu<\varepsilon\right]$ approaches 1 as $n\to\infty$, and $P\left[-\varepsilon<S-\Sigma<\varepsilon\right]$ approaches 1 as $n\to\infty$. These results can be used to support the multivariate central limit theorem, which implies that $\bar{X}\approx N_p\left(\mu,\frac{1}{n}\Sigma\right)$ for n large relative to p. This result indicates that $n\left(\bar{X}-\mu\right)'S'\left(\bar{X}-\mu\right)$ is approximately $\chi_p{}^2$ for $n-p$ large (Rencher & Chrestensen, 2013).

Robust multivariate linear regression:-

A robust multivariate regression is based on multivariate MM-estimates. Hence, the large sample nonparametric prediction region and the large sample Willks test, Pillais test and Hotelling Lawley test using the robust estimator are asymptotically equivalent to their analogs using the classical estimator for a large class of error distribution (Olive, 2013-b).

Predictions from multivariate linear regression:-

For a given set of predictors $x_0=\left[\ 1,x_{01},....,x_{0p}\ \right]$, we can simultaneously estimate the mean responses $x_0\beta$ for all p response variables as $x_0\hat{\beta}=\hat{Y}_0=\left[Y_{01},Y_{02},...,Y_{0q}\right]$. Now, to find the mean of $x_0\hat{\beta}$ for a fixed value x_0 , we have $E\left(x_0\,\hat{\beta}\right)=x_0\,E\left(\ \hat{\beta}\ \right)=x_0\,\beta$. The estimation errors $x_0\,\hat{\beta}_i-x_0\,\beta_i$ and $x_0\,\hat{\beta}_j-x_0\,\beta_j$ for the i^{th} and j^{th}

response variables have a covariance given by:

$$\begin{aligned}Cov\left(\ x_0\,\hat{\beta}_i-x_0\,\beta_j\ \right)&=E\left[\ \left(x_0\,\hat{\beta}_i-E\left(x_0\,\hat{\beta}_i\right)\right)\left(x_0\,\beta_j-E\left(x_0\,\hat{\beta}_j\right)\right)\ \right]\\&=x_0\left[\ x_0\left(X^TX\right)^{-1}x_0{}^T\ \right]\sigma_{ij}\end{aligned} \qquad (11)$$

We can then compute a Hotelling T^2 statistic as:

$$T^2=\left(\frac{x_0\,\hat{\beta}-x_0\beta}{\sqrt{x_0\left(X^TX\right)^{-1}x_0{}^T}}\right)^T\left(\frac{n}{n-p-1}\hat{\Sigma}\right)^{-1}\left(\frac{x_0\,\hat{\beta}-x_0\,\beta}{\sqrt{x_0\left(X^TX\right)^{-1}x_0{}^T}}\right) \qquad (12)$$

Then, $100(1-\alpha)\%$ confidence ellipsoids for $x_0\beta$ are given by all $x_0\beta$ that satisfy:

$$\left(x_0\hat{\beta}-x_0\beta\right)^T\left(\frac{n}{n-p-1}\hat{\Sigma}\right)^{-1}\left(x_0\hat{\beta}-x_0\beta\right)\le x_0\left(X^TX\right)^{-1}x_0{}^T\left[\left(\frac{q(n-p-1)}{n-p-q}\right)F_{q,n-p-q}(\alpha)\right]$$

The simultaneous $100(1-\alpha)\%$ confidence intervals for the means for each response $E(Y_i)=x_0\ \beta_{(i)}$, as in Johson and Wichern (2007), are:

$$x_0\hat{\beta}_{(i)}\pm\sqrt{\left(\frac{q(n-p-1)}{n-p-q}\right)F_{q,n-p-q}(\alpha)}\sqrt{x_0\left(X^TX\right)^{-1}x_0{}^T\left(\frac{n}{n-p-1}\hat{\sigma}_{(ii)}\right)},i=1,2,...,q \tag{13}$$

where $\hat{\beta}_i$ is the i^{th} column of $\hat{\beta}$ and σ_{ii} is the i^{th} diagonal element of $\hat{\Sigma}$.

Generalized hypotheses tests:-

Consider the multivariate linear regression model and define the hypothesis that the responses do not depend on the predictor variables $X_{q+1}\ ,\ X_{q+2}\ ,...\ ,\ X_r$ as:

$$\begin{aligned}&H_0:\beta_{(m)}=0\\&H_1:\beta_{(m)}\ne 0\quad\text{for at least one }\beta_{(m)}\end{aligned} \tag{14}$$

where $\beta_{(m)}=\begin{bmatrix}\beta_{p-m}\\ \vdots\\ \beta_p\end{bmatrix}$. If we partition X in a similar manner, we have:

$$X_{p+1-m}=\begin{bmatrix}X_{(0)}\\ ....\\ X_{(m)}\end{bmatrix}\text{, and }\beta_{(p+1)\times q}=\begin{bmatrix}\beta_{(0)}\\ ....\\ \beta_{(m)}\end{bmatrix} \tag{15}$$

Then, Maiti (2014) wrote the general model as:

$$\begin{aligned}E(Y)&=E(X\beta+\varepsilon)=E(X\beta)=X\beta\\&=X_{(0)n\times(p+1-m)}\beta_{(0)(p+1-m)\times q}+X_{(m)n\times m}\beta_{(m)m\times q}\end{aligned} \tag{16}$$

There are several test statistics, including the likelihood ratio test, that have been proposed for testing a linear hypothesis $H_0:LB=0$ versus $H_a:LB\ne 0$, where L is a full rank $r\times p$ matrix. Some useful results from Su and Cook (2012) and Kakizawa (2009, as described in Olive, 2013-a) show that the Hotelling-Lawley test statistic is an extension of the partial F-test statistic. We can create the hypothesis sum of squares and cross-product as:

$$\Re=\hat{B}^T\ L^T\left[L\left(X^TX\right)^{-1}L^T\right]^{-1}L\ \hat{B} \tag{17}$$

If we define the matrix of residuals as $\hat{E}=Y-X\ \hat{B}$ and again refer to the residual sum of squares and cross-product matrix as:

$$\begin{aligned} W_e &= \hat{E}^T\,\hat{E} = \left(Z-\hat{Z}\right)^T\left(Z-\hat{Z}\right) = Z^T\,Z - Z^T X\,\hat{B} \\ &= Z^T\left[\;I_n - X\left(X^T X\right)^{-1} X^T\;\right] Z \end{aligned} \qquad (18)$$

then $\hat{\sum}_{\varepsilon} = \dfrac{W_e}{(n-p)}$.

The multivariate tests of hypotheses are based on $s = \min(q,m)$ nonzero latent roots $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m$ of the matrix $\Re$ relative to the matrix W_e. That is, the value of λ for which $det\left(\Re - \lambda W_e\right) = 0$ is the same as the ordinary latent roots of $W_e^{-1}\Re$ for which $det\left(\Re\, W_e^{-1} - \lambda\, I_m\right) = 0$. Using this property, several statistics have been developed. We describe some commonly used test statistics below.

- Roy‘s maximum root statistic, which can be defined as $\lambda_{max}\,(L) = \dfrac{\lambda_1}{1+\lambda_1}$.
- The Wilk lambda Λ test statistic, which can be defined as $\Lambda = \left|\,\left(\Re + W_e\right)^{-1} W_e\,\right| = \left|\,W_e^{\,-1}\Re + I\,\right|^{-1} = \prod_{i=1}^{m} \dfrac{1}{\left(1+\lambda_j\right)}$ and can be approximated to the F distribution.
- Pillai‘s trace statistic, which can be defined as $V(L) = tr\left[\left(\Re + W_e\right)^{-1}\Re\right] = \sum_{i=1}^{m} \dfrac{\lambda_i}{1+\lambda_i}$.
- The Hotelling-Lawley trace statistic, which can be defined as $U(L) = tr\left(W_e^{-1}\Re\right) = \sum_{i=1}^{m} \lambda_i$.

Each of these statistics is an alternative to Wilk’s lambda and performs well, particularly for large sample sizes. In a large sample size, Wilk’s lambda, Roy‘s greatest root and the Hotelling-Lawley trace test are nearly equivalent (Fox et al., 2007). Typically, some of the functions of one of the four above statistics are used to get p-values. The estimated p-values output often gives the p-values for all four test statistics.

Wilk’s lambda, as a multivariate probability distribution, is difficult to apply in multivariate linear regression models as it is. To transfer its distribution to a one-dimensional case, two types of likelihood ratio tests based on Wilk’s lambda have been built. These are the Bartlett's test and the Rao‘s test statistics, which are described below.

1) Bartlett's test:-

We can create the Bartlett's test statistic as follows:

$$R = -M\;\ell n\,[\Lambda] \qquad (19)$$

where M is a multiplier that equals $M = (n-r-1) - \dfrac{(m-r+q+1)}{2}$, n is the total number of observations, p is the number of x variables, q is the number of y variables, and m is the number of dropped variables from the independent side.

Then, the statistic in equation (12), $R = -M\,\ell n(\Lambda) \approx \chi^2{}_{m\,(r-q)}$. H_0 is rejected if $R = -M\,\ell n\,(\Lambda) \ge \chi^2{}_{m(r-q)}\,(\alpha)$.

2) Rao‘s test:-

Rao‘s test statistic for the above hypotheses can be written as follows:

$$\frac{1-\Lambda^{1/Q}}{\Lambda^{1/Q}}\left[\frac{MQ+1-{}^{mq}\!/_{2}}{mq}\right], \qquad \text{where } Q=\left[\frac{q^2m^2-4}{q^2+m^2-5}\right]^{1/2} \tag{20}$$

If we let $qm = 2$, $Q = 0$ and $Q = 1$ (unity), then the statistic becomes:

$$\frac{1-\Lambda^{1/Q}}{\Lambda^{1/Q}}\left[\frac{MQ+1-{}^{mq}\!/_{2}}{mq}\right] \approx F_{mq,\, MQ+1-{}^{mq}\!/_{2}} \ ,$$

and the statistics approximate the F distribution, with $(mq,\ MQ+1-{}^{mq}\!/_{2})$ degrees of freedom. Then, by setting:

$$D=\frac{1-\Lambda^{1/Q}}{\Lambda^{1/Q}}\left[\frac{MQ+1-{}^{mq}\!/_{2}}{mq}\right], \tag{21}$$

we reject H_0 when $D \geq F(\alpha)_{mq,\, MQ+1-{}^{mq}\!/_{2}}$.

Testing for Multivariate Normality:-

There are many tests for testing multivariate normality. These include the Shapiro-Wilk test, the generalized Shapiro-Wilk test by Villasenor-Alva; the Gonzalez-Estrada test; the generalized Shapiro-Francia test; the energy test; and Mardia's, Royston's, and Henze-Zirkler's multivariate normality tests. We now describe some of these tests, which are later applied to our dataset.

1) Mardia's Test:-

Mardia's multivariate normality test is based on multivariate extensions of skewness $\left(\hat{\beta}_{1,p}\right)$ and kurtosis $\left(\hat{\beta}_{2,p}\right)$, as follows:

$$\hat{\beta}_{1,p}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}{g_{ij}}^3 \qquad \text{and} \qquad \hat{\beta}_{2,p}=\frac{1}{n}\sum_{i=1}^{n}{g_{ii}}^2 \tag{22}$$

where $g_{ij}=\left(x_i-\bar{x}\right)' {S_n}^{-1}\left(x_j-\bar{x}\right)$, $S_n=\frac{1}{n}\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)'\right]$, and $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

2) Henze-Zirkler's Test:-

Henze-Zirkler’s test is based on a non-negative functional distance that measures the distance between two distribution functions. If data follow a multivariate normal distribution, the test statistic is approximately log-normally distributed, and it calculates the mean, variance, and smoothness parameters. Then, the mean and the variance are log-normalized, and the p-value is estimated (Henze & Zirkler, 1990). The test statistic of Henze-Zirkler’s multivariate normality test takes the following form:

$$HZ=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\frac{\beta^2}{2}D_{ij}}-2\left(1+\beta^2\right)^{-\frac{p}{2}}\sum_{i=1}^{n}e^{-\frac{\beta^2}{2(1+\beta^2)}D_i}+n\left(1+\beta^2\right)^{-\frac{p}{2}} \tag{23}$$

where P is the number of variables,

$D_{ij}=\left(x_i-x_j\right)' S^{-1}\left(x_i-x_j\right)$, and $D_i=\left(x_i-\bar{x}\right)' S^{-1}\left(x_i-\bar{x}\right)$

3) Royston's Test:-

Royston's test uses the Shapiro-Wilk/Shapiro-Francia statistic to test multivariate normality. If the kurtosis of the data is greater than three, then it uses the Shapiro-Francia test for leptokurtic distributions; otherwise, it uses the Shapiro-Wilk test for platykurtic distributions (Mecklin & Mundfrom, 2005). The polynomial coefficients are provided by Royston (1992) for different sample sizes.

$$\begin{aligned} \beta &= a_{0\beta} + a_{1\beta}x + a_{2\beta}x^2 + \ldots + a_{d\beta}x^d \\ \mu &= a_{0\mu} + a_{1\mu}x + a_{2\mu}x^2 + \ldots + a_{d\mu}x^d \\ log\left(\sigma\right) &= a_{0\sigma} + a_{1\sigma}x + a_{2\sigma}x^2 + \ldots + a_{d\sigma}x^d \end{aligned} \tag{24}$$

Royston's test statistic for multivariate normality then takes the following form:

$$H = \frac{e\sum_{j=1}^{p}\psi_j}{p} \quad \approx \quad \chi_e{}^2 \tag{25}$$

where e is the equivalent degrees of freedom and $\Phi(.)$ is the cumulative distribution function for standard normal distribution, such that:

$$e = \frac{p}{\left[1+(p-1)\,\bar{c}\,\right]} \tag{26}$$

$$\psi_j = \left\{\Phi^{-1}\left[\Phi\left(-Z_j\right)/2\right]\right\}^2 , \qquad j = 1, 2, \ldots, p$$

4) The energy statistic test:-

The energy test is based partly on the Euclidean distance between observations. The energy test uses the following statistic:

$$\hat{E}_{n,d} = n\left(\frac{2}{n}\sum_{j=1}^{n}E\left\|y_j - Z\right\| - 2\,\frac{\Gamma\left(\left(d+1\right)/2\right)}{\Gamma\left(d/2\right)} - \frac{1}{n^2}\sum_{j,k=1}^{n}E\left\|y_j - y_k\right\|\right) \quad \text{with}$$

$$E\left\|a - Z\right\| = \sqrt{2}\,\frac{\Gamma\left(\left(d+1\right)/2\right)}{\Gamma\left(d/2\right)} + \sqrt{\frac{2}{\pi}}\sum_{k=0}^{inf}\frac{\left(-1\right)^k \left\|a\right\|^{2k+2}}{k_! \, 2^k \left(2k+1\right)\left(2k+2\right)}\,\frac{\Gamma\left(\left(d+1\right)/2\right)\Gamma\left(k+1.5\right)}{\Gamma\left(\left(d/2\right)+k+1\right)}$$

The energy statistic test (Szkely & Rizzo, 2005) may be computed using R software through the energy package (Joenssen & Vogel, 2015; Rizzo & Szkely, 2008).

5) Normal probability plots for the multivariate regression model:-

Sometimes, different multivariate normal tests may come up with different results. In such cases, examining multivariate normal plots and hypothesis tests together can be quite useful for reaching a more reliable decision.

Certain notations are needed to discuss the diagnostic checks of the multivariate linear regression model plots, which are referred to as the DD plots. Letting the $p \times 1$ column vector T be a multivariate location estimator and the $p \times p$ symmetric positive definite matrix C be a dispersion estimator, then the i^{th} squared sample Mahalanobis distance is a scalar.

$$D_i{}^2 = D_i{}^2\left(T, C\right) = \left(x_i - T\right)^T C^{-1}\left(x_i - T\right) \quad \text{for each observation } x_i \,. \tag{27}$$

Notice that the Euclidean distance of x_i from the estimate of center T is $D_i\left(T , I_p \right)$. The notation MD is used to denote the classical Mahalanobis distances $MD_i = D_i\left(\bar{x} , S \right)$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $S = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)^T$, and RD denotes the distance $RD_i = D_i\left(T, C\right)$.

Using the robust RMVN estimator $\left(T, C\right)$ described in (Zhang et al., 2012), the classical estimator $\left(\bar{X}, S\right)$ is a consistent estimator of the population mean and covariance matrix $\left(\mu_x , \Sigma_x \right)$. Furthermore, the RMVN estimator $\left(T , C\right)$ is a consistent estimator of $\left(\mu_x , c\Sigma_x \right)$ for a large class of elliptically contoured distributions, where $c > 0$ depends on the distribution and c = 1 for the multivariate normal distribution.

The Q-Q plot, where “Q” stands for quantile, is a widely used graphical approach to evaluate the agreement between two probability distributions. Each axis refers to the quantiles of probability distributions to be compared, where one of the axes indicates the theoretical quantiles and the other indicates the observed quantiles.

For the multivariate linear model, any hypothesis test may be calculated from an analog of the univariate F, where p×p matrices H and E play the roles of the univariate sums of squares SS_H and SS_E . However, in the multivariate case, the variation against the null hypothesis may be large in one or more dimensions relative to the error variation n (E). The plots are implemented in two and three dimensions in the **heplots** package in the R software (Fox et al., 2013).

Application of the Multivariate Regression Model to Adult Diabetic and Hypertensive Patients' Data in the Gaza Strip:-

In this paper, we analyze a real dataset by applying the multivariate linear regression model to the dataset of adult onset diabetes and hypertension patient data in the Gaza Strip. The main purpose of this analysis is to study the important risk factors affecting the existence of diabetes and hypertension in a sample of patients.

The data consist of 14 variables for 518 patients. The dependent variables are indicators of the existence of diabetes and hypertension diseases; these include stabilized glucose (Stab.glu), median of three measurements of systolic blood pressure (BP.S), and median of three measurements of diastolic blood pressure (BP.D). The summary statistics of these variables are given in table (2).

A preliminary investigation of the data and the initial multivariate regression model for the original data indicated that neither the model's residuals nor the variables of the study were found to be normally distributed, except for the variable Age. For this, we applied the Box-Cox transformation (Bozdogan & Ramirez, 1997) to transform the underlying variables to normality. Accordingly, we added the prefix "T." to the name of each transformed variable.

Table 2: Descriptive statistics of dependent variables (Stab.glu, BP.S, and BP.D)

Variables	Mean	St. Dev.	Min.	Max.	Shapiro-test	
					Statistic (W)	P-value
Actual Data						
Stab.glu	136.34	65.38	68	381	0.782	0.000
BP.S	131.7	13.86	100	190	0.942	0.000
BP.D	86.42	9.95	61	110	0.977	0.000
Transformed Data						
T.Stab.glu	136.4	65.38	-71.47	344.15	0.999	1.000
T.BP.S	131.67	13.86	92.19	175.78	0.996	0.218
T.BP.D	86.42	9.95	54.75	116.03	0.996	0.195

The independent variables are thought to be risk factors for diabetes and hypertension diseases. Among them, gender is a categorical variable, such that there are 278 (53.7%) of females and 240 (46.3%) are males. All other independent variables are numeric. Table 3 shows that all of the independent variables became normally distributed following the transformation, whereas the p-values of Shapiro-Wilk test are all greater than the 0.05 significance level.

Table 3: Descriptive statistics of the quantitative independent variables

Variables	Mean	St. Dev.	Min.	Max.	Shapiro-Wilk test	
					Statistic	P-value
Actual Data						
Chol	192.4	40.91	105	311	0.990	0.0008
HDL	46.87	10.93	25.0	85.0	0.988	0.0004
LDL	118.8	35.89	41	258.0	0.988	0.0003
Ratio	4.3	1.3	2.11	10.72	0.907	<2.2e-16
TG	168.2	85.67	37.0	565	0.889	<2.2e-16
Urea	30.42	8.38	13.0	58.0	0.970	8.3e-09
Age	54.39	12.36	18.0	85.0	0.996	0.142
Weight	88.75	15.38	52.0	134.0	0.983	7.6e-06
BMI	32.26	5.84	18.36	49.1	0.986	5.8e-05
Transformed Data						
T.Chol	13.79	1.47	10.25	17.64	0.996	0.189
T.HDL	6.8	0.8	5	9.22	0.994	0.052
T.LDL	10.77	1.66	6.4	16.06	0.998	0.788
T.Ratio	0.497	0.07	0.31	0.69	0.996	0.136
T.TG	5.01	0.48	3.61	6.34	0.998	0.641
T.Urea	3.38	0.27	2.57	4.06	0.996	0.138
T.Weight	4.47	0.17	3.95	4.9	0.996	0.211
T.BMI	3.46	0.18	2.91	3.89	0.997	0.386
Creatinine	0.812	0.22	0.4	1.8	0.88	<2.2e-16

Multivariate regression model:-

The multivariate regression model that we estimate here is $Y = X\beta + \varepsilon$, as shown in Eq. (1) above, where Y represents the dependent variables, as follows:

$$Y = \begin{bmatrix} T.BP.S_1 & T.BP.D_1 & Stab.glu_1 \\ T.BP.S_2 & T.BP.D_2 & Stab.glu_2 \\ \vdots & \vdots & \vdots \\ T.BP.S_{518} & T.BP.D_{518} & Stab.glu_{518} \end{bmatrix} \tag{28}$$

The X matrix represents the independent variables: Gender, Age, T.Weight, T.BMI, T.Urea, T.Ratio, T.Chol, T.LDL, T.HDL, T.TG and Creatinine. It can be presented as follows:

$$X = \begin{bmatrix} 1 & Gender_1 & Age_1 & \cdots & Creatinine_1 \\ 1 & Gender_2 & Age_2 & \cdots & Creatinine_2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & Gender_{518} & Age_{518} & \cdots & Creatinine_{518} \end{bmatrix} \tag{29}$$

β represents the model coefficients, as follows:

$$\beta = \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{32} \\ \vdots & \vdots & \vdots \\ \beta_{111} & \beta_{211} & \beta_{311} \end{bmatrix} \tag{30}$$

and ε represents the residuals of the model, as follows:

$$\varepsilon = \begin{bmatrix} \varepsilon_{1T.BP.S} & \varepsilon_{1T.BP.D} & \varepsilon_{1Stab\,.glu} \\ \varepsilon_{2T.BP.S} & \varepsilon_{2T.BP.D} & \varepsilon_{2Stab\,.glu} \\ \vdots & \vdots & \vdots \\ \varepsilon_{518T.BP.S} & \varepsilon_{518T.BP.D} & \varepsilon_{518Stab\,.glu} \end{bmatrix}. \tag{31}$$

After estimating the model parameters, we can express the estimated full multivariate model, as shown in Eq. (1), where the estimated model's coefficients are as given in table (4).

Table 4: Estimates of the coefficients β of the independent variables X of the full multivariate regression model $Y = X\beta + \varepsilon$.

Variables	T.Stab.glu	T.BP.D	T.BP.S
(Intercept)	56.9072442	11.54046748	35.73184441
Gender	-2.5139733	0.52300914	0.40696655
Age	0.2809305	-0.05263272	0.28290046
T.Weight	29.8851910	6.48772314	7.24262435
T.TG	17.1850803	0.64073990	-0.14635055
T.Ratio	-4.3171957	51.34699122	63.14328683
T.BMI	-46.4010300	4.17283167	5.82961295
T.Urea	-3.8644056	0.14003085	1.15095296
T.Chol	1.3187449	3.10749514	2.65840877
T.LDL	1.5412040	-0.54814380	-0.06814503
T.HDL	-1.6212302	-4.79793599	-6.33509476

Testing the significance of the full model:-

To test the significance of the above full model, all tests described in section 9 were applied. Table 5 shows the results of applying the Pillai, Roy, Hotelling-Lawley, and Wilk lambda tests to test the significance of the full multivariate linear regression model. The results of all tests indicated that the estimated regression model was significant at the 0.05 level.

Table 5: Results of the Pillai, Roy, Hotelling-Lawley, and Wilk lambda tests for testing the significance of the multivariate regression model.

Test	Sources	Df	Statistics	Approx. F	Num Df	Den Df	Pr(>F)
Pillai	Model	11	0.24558	4.101	33	1518	0.000
	Residual	506					
Roy	Model	11	0.16587	7.630	11	506	0.000
	Residual	506					
Hotelling-Lawley	Model	11	0.27624	4.208	33	1508	0.000
	Residual	506					
Wilk	Model	11	0.77088	4.1571	33	1485.6	0.000
	Residual	506					

Testing the significance of the model's coefficients and the final multivariate regression model:-

Further investigation of the full model required testing the significance of the coefficients of each independent variable in the model. The results of the Pillai, Roy, Hotelling-Lawley, and Wilk lambda tests indicated that the variables Age, T.Weight, T.Ratio, and T.TG were statistically significant at 0.05 level. However, coefficients of all other variables were not statistically significant; these variables could, thus, be omitted from the full model. The results of testing the significance of the coefficients of the independent variables after omitting the insignificant variables are listed in table 6. These results indicate that all four variables—Age, T.weight, T.Ratio, and T.TG—are statistically significant at the 0.05 level.

Table 6 : Results of the Pillai, Roy, Hotelling, and Wilk tests for testing only the significant variables in the multivariate regression model at the 0.05 level

Variables	**Pillai**			**Roy**			**Hotelling-Lawley**			**Wilk**		
	Stat	**Approx. F**	**Pr(>F)**	**Stat**	**Approx. F**	**Pr(>F)**	**Stat**	**Approx. F**	**Pr(>F)**	**Stat**	**Approx. F**	**Pr(>F)**
Intercept	0.99	21501	0.000	126.2	21501	0.000	126.23	21501	0.000	0.01	21501.4	0.000
Age	0.13	25.10	0.000	0.15	25.10	0.000	0.15	25.10	0.000	0.87	25.1	0.000
T.Weight	0.04	7.00	0.000	0.04	7.00	0.000	0.04	7.00	0.000	0.96	7	0.000
T.Ratio	0.03	4.80	0.003	0.03	4.80	0.003	0.03	4.80	0.003	0.97	4.8	0.003
T.TG	0.02	3.10	0.027	0.02	3.10	0.027	0.02	3.10	0.027	0.98	3.1	0.027

Figure 2 displays the hypothesis-error (HE) plots to show the significance and effect size of the independent variables on the dependent variables in the final reduced model. The plot indicates a high significance and a high effect size of the independent variables on the dependent variables.

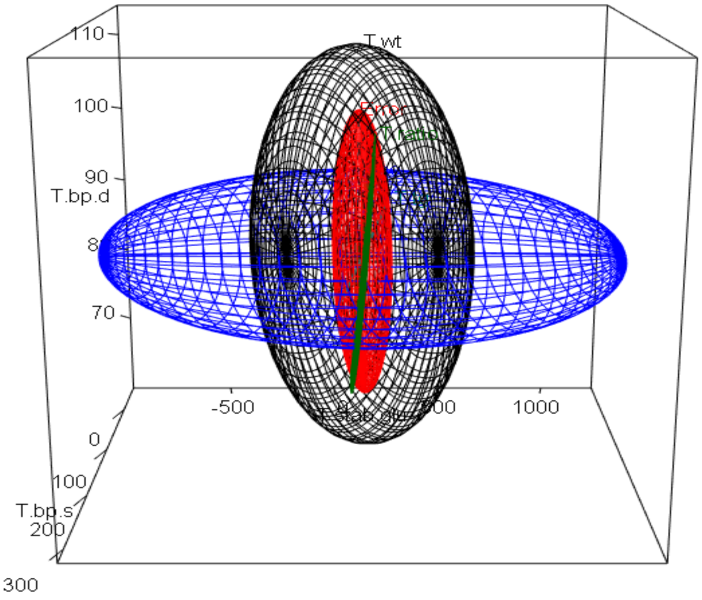

Figure (2): HE plot for the multivariate regression model.

Table 7 displays the estimates for the multivariate regression model's coefficients for each statistically significant independent variable's effect on the dependent variables at the 0.05 significance level. Noting that the positive influence indicates a similar pattern of change in both variables (and vise versa), we see in table 7 that the variable Age has positive influences on both the T.stab.glu and the T.BP.S variables and a weak negative influence on the T.BP.D variable. The T.Weight variable has a strong positive influence on both the T.BP.S and the T.BP.D variables and a negative influence on the T.stab.glu variable. The T.Ratio variable has a strong negative influence on all dependent variables. The variable T.TG has a strong positive influence on the T.stab.glu variable and a weak positive influence on the T.BP.D variable, while its influence on the T.BP.S variable is negative and weak.

Table 7: Estimates of the coefficients of the independent variables of the reduced multivariate regression model.

Variables	T.BP.S	T.BP.D	T.Stab.glu
(Intercept)	69.79	46.77	109.41
Age	0.30	-0.05	0.24
T.Weight	12.30	10.10	-9.25
T.Ratio	-16.40	-16.07	-51.44
T.TG	-0.22	1.04	16.11

The multivariate regression model for this dataset, as in Eq. (1), is thus estimated as $Y = X\hat{\beta} + \varepsilon$, where Y and X are as defined in Eqs. 28 and 29, respectively, and $\hat{\beta}$ is the estimated coefficients, which may be obtained from table 7, as follows:

$$\hat{\beta} = \begin{bmatrix} 69.79 & 46.77 & 109.41 \\ 0.30 & -0.05 & 0.24 \\ 12.30 & 10.10 & -9.25 \\ -16.40 & -16.07 & -51.44 \\ -0.22 & 1.04 & 16.11 \end{bmatrix}$$

To conduct a diagnostic check of the model's residuals, we tested the normality of the residuals using the tests described in section (10), the scatter diagram of the residual, and the Pearson correlation coefficient to check the multicollinearity of the model. Table 8 shows the results of testing for the multivariate normality of the residuals of the final model using different tests. The results of table 8 indicate that the p-value is greater than the 0.05 significance level for all tests. Therefore, the null hypothesis cannot be rejected, and we can conclude that the residuals of the final multivariate regression model follow the multivariate normal distribution. These results, together with the histogram and the q-q plot in Figure 3, indicate that the model's residuals are independent and multivariate normally distributed with a constant variance. This indicates that the fitted model is good for predicting stabilized glucose, median systolic blood pressure and median diastolic blood pressure levels in Palestinian patients in the Gaza Strip.

Table 8: Results of testing the multivariate normality of the model residual.

Test	Stat.	P-value
Mardia		
Chi-Skewness	8.547	0.576
Z.Kurtosis	0.477	0.633
Henze-Zirkler	0.856	0.403
Royston	0.063	0.996
Shapiro-Wilk	0.996	0.188
Generalized Shapiro-Wilk by Villasenor-Alva and Gonzalez-Estrada	0.997	0.628
Generalized Shapiro-Francia	0.995	0.073
Energy	0.963	0.159

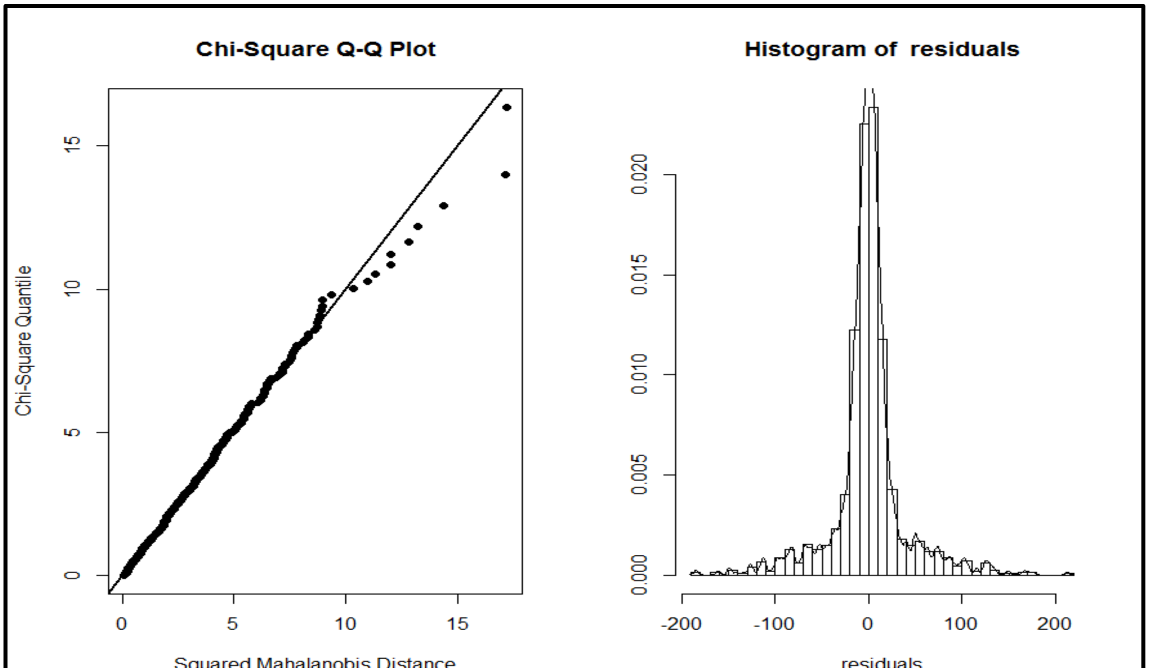

Figure 3: QQ plot and histogram of the residuals of the multivariate regression model.

Robust multivariate regression model based on multivariate MM estimates:-

Now, for the purpose of comparison, we estimate the multivariate regression model for the variables of this study using robust methods based on multivariate MM estimates. The bootstrap estimates of the regression coefficients were applied based on the Fast and Robust Bootstrap 999 repeated, and the p-values were computed based on the bias corrected and accelerated (BCA) method for testing the significance of the coefficients.

Firstly, the full multivariate linear model with all independent variables was estimated; the results indicate that only the independent variables T.TG and Age have significant effects on T.Stab.glu and T.BP.S, respectively, at the 0.05 level. There is no variable with a significant effect on T.BP.D. The final model that we estimated using the ordinary least squares multivariate regression method with four independent variables (Age, T.Weight, T.Ratio,and T.TG) was estimated again using robust methods. The results in table 9 indicate that all four variables have significant effects on the three-dimensional dependent variable at the 0.05 significance level.

Table 9: Results of the robust method based on the MM estimates of the multivariate model regression on the independent variables of Age, T.Weight, T.Ratio and T.TG.

Variables	**T.Stab.glu**		**T.BP.S**		**T.BP.D**	
	Coefficient	**p-value**	**Coefficient**	**p-value**	**Coefficient**	**p-value**
(intercept)	114.67	0.1910	**66.537**	**0.0000**	**42.903**	**0.0000**
Age	0.21	0.4070	**0.307**	**0.0000**	-0.053	0.1458
T.Weight	-8.46	0.5840	**13.246**	**0.0000**	**11.38**	**0.0000**
T.Ratio	-64.52	0.1930	-15.843	0.0595	**-14.275**	**0.0438**
T.TG	**16**	**0.0150**	-0.582	0.7239	0.551	0.5555

Note: Bold indicates a significant coefficient.

A comparison between the estimates of the coefficients of the robust final model and those of the ordinary least squares multivariate regression final model indicates that, though the signs of the coefficients are identical for both methods, the values of the coefficients are slightly different. This again indicates that the model is quite good.

Conclusions and Recommendations:-

From the discussion in the previous sections, we can conclude that the multivariate regression model is a good method for predicting the determinants of adult onset diabetes and hypertension diabetes (including stabilized glucose, median systolic blood pressure, and median diastolic blood pressure levels) among Palestinian patients in the Gaza Strip. The coefficients of the model were estimated, and their significances were tested. The fitted model proved appropriate for our data set. Most importantly, the fitted model managed to identify the most important determinants and risk factors having the greatest effects on adult onset diabetes and hypertension levels among patients in the Gaza Strip. Other findings that may be concluded from the previous sections are as follows:

1. The variables that have significant effects on adult onset diabetes and hypertension, according to the multivariate linear regression model at the 0.05 level, are Age, Weight, Ratio, and TG. The effect of Age was negative on PBD and positive on Stab.glu and PBS; the effect of patient weight was negative on Stab.glu and positive on PBS and PBD; the effect of Ratio was negative on all dependent variables; and the effect of TG was negative on PBD and positive on Stab.glu and PBS.
2. The robust method for estimating the multivariate regression model produced the same subset of independent variables with significant effects as the ordinary least squares method, although the estimates for the coefficients had slightly different values. This may be because the number of observations was very large.
3. There is still a high need to develop more information on the multivariate linear regression model, particularly with respect to methods of selecting the best subset of independent variables and conducting further hypothesis testing on the multivariate linear regression models' parameters.
4. We recommend conducting further research on the multicollinearity problem among both dependent and independent variables in the multivariate linear regression models.

References:-

AL-Marshadi, A. H. (2010). Comparison of model selection criteria for multivariate regression model with mixed model. *Journal of Applied Sciences Research, 6*(2), 107–120.

Bozdogan, H. & Ramirez, D. (1997). Testing for model fit: Assessing Box-Cox transformations of data to near normality. *Computational Statistics Quarterly, 3*, 203–213.

Burdenski, T. (2000). Evaluating univariate, bivariate, and multivariate normality using graphical and statistical procedures. *Multiple Linear Regression Viewpoints, 26*(2), 15–156 .

Fox, J., Friendly, M. & Monette, G. (2013). *Heplots: Visualizing tests in multivariate linear models. R package version 1.0-11*. Retrieved from http://CRAN.R-project.org/package=heplots/

Fox, J., Friendly, M. & Monette, G. (2007). Visual hypothesis tests in multivariate linear models: The heplots package for R. *DSC 2007: Directions in Statistical Computing*. Canada: McMaster University. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/heplots-dsc-paper.pdf

Henze, N. & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics—Theory and Methods, 19*(10), 3595–3617.

Joenssen, D. W. & Vogel, J. (2015). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, *84*(5), 1055- 1077.

Johnson, R. A. & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Johnson, R. A. & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Johnson, R. A. & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Kakizawa, Y. (2009). Third-order power comparisons for a class of tests for multivariate linear hypothesis under general distributions. *Journal of Multivariate Analysis, 100*, 473–496. Retrieved from http://dx.doi.org/10.1016/j.jmva.2008.06.002/

Kim, K. & Timm, N. (1991). *Univariate and multivariate general linear models: theory and applications with SAS* (2nd ed.). PA: University of Pittsburgh.

Latra, N. L., Linuwih, S., Purhadi & Suhartono. (2010). Estimation for multivariate linear mixed models. *International Journal of Basic & Applied Sciences*, *10*(6), 48-53.

Liyun, S. (2012). *Local polynomial estimation of heteroscedasticity in a multivariate linear regression model and its applications in economics*. Chongqing, China: School of Mathematics and Statistics, Chongqing University of Technology.

Maiti, J. (2014). *Applied multivariate statistical modeling*. Department of Management, IIT Kharagpur. DOER - Directory of Open Educational Resources, National Programme on Technology Enhanced Learning (NPTEL).

Marden, J. I. (2013). *Multivariate statistics*. *Old school.* University of Illinois at Urbana-Champaign. Retrieved from http://istics.net/pdfs/multivariate.pdf.

Mecklin, C. M. & Mundfrom D. J. (2005). A Monte Carlo comparison of the type I and type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation, 75*(2), 93–107.

O’Brien, R. G. & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs. In, *An Extensive Primer, Psychological Bulletin, 97*(2), 316-33.

Okasha, M. K. (2014). Prevalence Rates of Hypertension and Related Risk Factors in Palestine. *International Journal of Statistical Sciences (IJSS)*, 13, 55-72.

Olive, D. J. (2013-a). Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. *International Journal of Statistics and Probability*, *2*, 90–100. Retrieved from http://dx.doi.org/10.5539/ijsp.v2n1p90/

Olive, D. J. (2013-b). Plots for generalized additive models. *Communications in Statistics: Theory and Methods*, *42*, 2610–2628.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.

Rencher A. C. & Chrestensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Willy Series in Probability and Statistics, John Wiley & Sons.

Rencher, A. C. & Christensen, W. F. (2013). Methods of multivariate analysis. *International Statistical Review, 81*(2), 328–329.

Rizzo, M. L. & Szekely, G. L. (2008). *Energy: E-statistics (energy statistics). R package version 1.10*. Retrieved from http://CRAN.R-project.org/package=energy.

Royston, J. P. (1992). Approximating the Shapiro-Wilk W test for non-normality. *Statistics and Computing, 2*(3), 117–119.

Sawyer, S. (2010). *Multivariate linear models*. Technical Report, Washington University.

Su, Z. & Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika, 99*, 687–702. Retrieved from http://dx.doi.org/10.1093/biomet/ass024

Szkely, G. J. & Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis, 93*, 58–80.

Zhang, J., Olive, D.J., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability, 1*, 119–136.