REVIEW ARTICLE

Introduction to Multivariate Regression Analysis

Alexopoulos EC

Department of Public Health, Medical School, University of Patras, Rio Patras, Greece

Key words: multivariate regression analysis

Corresponding author: Evangelos Alexopoulos, Department of Public Health, Medical School, University of Patras, 26500 Rio Patras, Greece, e-mail: ecalexop@med.uoa.gr

Statistics are used in medicine for data description and inference. Inferential statistics are used to answer questions about the data, to test hypotheses (formulating the alternative or null hypotheses), to generate a measure of effect, typically a ratio of rates or risks, to describe associations (correlations) or to model relationships (regression) within the data and, in many other functions. Usually point estimates are the measures of associations or of the magnitude of effects. Confounding, measurement errors, selection bias and random errors make unlikely the point estimates to equal the true ones. In the estimation process, the random error is not avoidable. One way to account for is to compute p-values for a range of possible parameter values (including the null). The range of values, for which the *p-value* exceeds a specified alpha level (typically 0.05) is called confidence interval. An interval estimation procedure will, in 95% of repetitions (identical studies in all respects except for random error), produce limits that contain the true parameters. It is argued that the question if the pair of limits produced from a study contains the true parameter could not be answered by the ordinary (frequentist) theory of confidence intervals[1]. Frequentist approaches derive estimates by using probabilities of data (either p-values or likelihoods) as measures of compatibility between data and hypotheses, or as measures of the relative support that data provide hypotheses. Another approach, the Bayesian, uses data to improve existing (prior) estimates in light of new data. Proper use of any approach requires careful interpretation of statistics[1,2].

The goal in any data analysis is to extract from raw information the accurate estimation. One of the most important and common question concerning if there is statistical relationship between a response variable (Y) and explanatory variables (Xi). An option to answer this question is to employ regression analysis in order to *model* its relationship. There are various types of regression analysis. The type of the regression model depends on the type of the distribution of Y; if it is continuous and approximately normal we use linear regression model; if dichotomous we use logistic regression; if Poisson or multinomial we use log-linear analysis; if time-to-event data in the presence of censored cases (survival-type) we use Cox regression as a method for modeling. By modeling we try to predict the outcome (Y) based on values of a set of predictor variables (Xi). These methods allow us to assess the impact of multiple variables (covariates and factors) in the same model[3,4].

In this article we focus in linear regression. Linear regression is the procedure that estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable which should be quantitative. Logistic regression is similar to a linear regression but is suited to models where the dependent variable is dichotomous. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model.

Linear equation

In most statistical packages, a curve estimation procedure produces curve estimation regression statistics and related plots for many different models (linear, logarithmic, inverse, quadratic, cubic, power, S-curve, logistic, exponential etc.). It is essential to plot the data in order to determine which model to use for each depedent variable. If the variables appear to be related linearly, a simple linear regression model can be used but in the case that the variables are not linearly related, data transformation might help. If the transformation does not help then a more complicated model may be needed. It is strongly advised to view early a scatterplot of your data; if the plot resembles a mathematical function you recognize, fit the data to that type of model. For example, if the data resemble an exponential function, an exponential model is to be used. Alternatively, if it is not obvious which model best fits the data, an option is to try several models and select among them. It is strongly recommended to screen the data graphically (e.g. by a scatterplot) in order to determine how the independent and dependent variables are related (linearly, exponentially etc.)[4-6].

The most appropriate model could be a straight line, a higher degree polynomial, a logarithmic or exponential. The strategies to find an appropriate model include the forward method in which we start by assuming the very simple model i.e. a straight line ($Y = a + bX$ or $Y = b_0 + b_1X$). Then we find the best estimate of the assumed model. If this model does not fit the data satisfactory, then we assume a more complicated model e.g. a

2nd degree polynomial ($Y=a+bX+cX^2$) and so on. In a backward method we assume a complicated model e.g. a high degree polynomial, we fit the model and we try to simplify it. We might also use a model suggested by theory or experience. Often a straight line relationship fits the data satisfactory and this is the case of simple linear regression. The simplest case of linear regression analysis is that with one predictor variable[6,7].

Linear regression equation

The purpose of regression is to predict Y on the basis of X or to describe how Y depends on X (regression line or curve)

$$X_1, X_2, \ldots, X_k \Rightarrow Y$$

The Xi ($X_1, X_2, \ldots, X_k$) is defined as “predictor”, “explanatory” or “independent” variable, while Y is defined as “dependent”, “response” or “outcome” variable.

Assuming a linear relation in population, mean of Y for given X equals α+βX i.e. the “population regression line”.

If Y = a + bX is the estimated line, then the fitted $\hat{Y}i = a + bXi$ is called the fitted (or predicted) value, and $Yi - \hat{Y}i$ is called the residual.

The estimated regression line is determined in such way that Σ (residuals)² to be the minimal i.e. the standard deviation of the residuals to be minimized (residuals are on average zero). This is called the “least squares” method. In the equation

$$\mathbf{Yi = a + bXi}$$

b is the slope (the average increase of outcome per unit increase of predictor)

a is the intercept (often has no direct practical meaning)

A more detailed (higher precision of the estimates a and b) regression equation line can also be written as

Yi = a + bXi + σ_{res} where **σ_{res}** = residual standard deviation = sd

Further inference about regression line could be made by the estimation of confidence interval (95%CI for the slope b). The calculation is based on the standard error of b:

$$\text{se(b)} = \frac{S_{res}}{\sqrt{S_{xx}}} = \frac{S_{res}}{\sqrt{\Sigma x_i^2 - (\Sigma x)^2 / n}}$$

so, 95% CI for β is b ± t0.975*se(b) [t-distr. with df = n-2]

and the test for H0: β=0, is t = b / se(b) [p-value derived from t-distr. with df = n-2].

If the p value lies above 0.05 then the null hypothesis is not rejected which means that a straight line model in X does not help predicting Y. There is the possibility that the straight line model holds (slope = 0) or there is a curved relation with zero linear component. On the other hand, if the null hypothesis is rejected either the straight line model holds or in a curved relationship the straight line model helps, but is not the best model. Of course there is the possibility for a type II or type I error in the first and second option, respectively. The standard deviation of residual (σ_{res}) is estimated by

$$\sigma_{res} = \sqrt{\frac{\sum (residuals)^2}{n-2}} \text{ or } \sqrt{\frac{\sum (Y_i - Y_{fit})^2}{n-2}}$$

The standard deviation of residual (σ_{res}) characterizes the variability around the regression line i.e. the smaller the σ_{res}, the better the fit. It has a number of degrees of freedom. This is the number to divide by in order to have an unbiased estimate of the variance. In this case df = n-2, because two parameters, α and β, are estimated[7].

Multiple linear regression analysis

As an example in a sample of 50 individuals we measured: Y = toluene personal exposure concentration (a widespread aromatic hydrocarbon); X1 = hours spent outdoors; X2 = wind speed (m/sec); X3 = toluene home levels. Y is the continuous response variable (“dependent”) while X1, X2, …, Xp as the predictor variables (“independent”) [7]. Usually the questions of interest are how to predict Y on the basis of the X′s and what is the “independent” influence of wind speed, i.e. corrected for home levels and other related variables? These questions can in principle be answered by multiple linear regression analysis.

In the multiple linear regression model, Y has normal distribution with mean

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta\rho X\rho + \sigma(Y),\ \ sd(Y) = \sigma\ (independent\ of\ X's)}$$

The model parameters $\beta_0 + \beta_1 + \ldots + \beta\rho$ and σ must be estimated from data.

β_0 = intercept

$\beta_{1\ \ldots}\ \beta\rho$ = regression coefficients

$\sigma = \sigma_{res}$ = residual standard deviation

Interpretation of regression coefficients

In the equation $\mathbf{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta\rho X\rho}$

β_1 equals the mean increase in Y per unit increase in Xi , while other Xi’s are kept fixed. In other words βi is influence of Xi corrected (adjusted) for the other X’s. The estimation method follows the least squares criterion.

If $b_0, b_1, \ldots, b\rho$ are the estimates of $\beta_0, \beta_1, \ldots, \beta\rho$ then the “fitted” value of Y is

$$Yfit = b_0 + b_1 X_1 + \ldots + b\rho X\rho$$

The b0, b1, … , bρ are computed such that $\sum (Y\text{-}Y_{fit})^2$ to be minimal. Since Y – Yfit is called the residual; one can also say that the sum of squared residuals is minimized.

In our example, the statistical packages give the following estimates or regression coefficients (bi) and standard errors (se) for toluene personal exposure levels.

Predictor Xi	Bi	se (bi)
Time spent outdoors (hours)	0.582	0.191
Home levels (μg/m³)	0.554	0.053
Wind speed (m/sec)	-54.15	18.24

Then the regression equation for toluene personal exposure levels would be:

Tpers = 0.582 time outdoors + 0.554 Thome + (-54.15) wind speed

The estimated coefficient for time spent outdoors (0.582) means that the estimated mean increase in toluene personal levels is 0.582 $\mu g/m^3$ if time spent outdoors increases 1 hour, while home levels and wind speed remain constant. More precisely one could say that individuals differing one hour in the time that spent outdoors, but having the same values on the other predictors, will have a mean difference in toluene xposure levels equal to 0.582 $\mu g/m^3$ [8].

Be aware that this interpretation does not imply any causal relation.

Confidence interval (CI) and test for regression coefficients

95% CI for βi is given by bi ± t0.975*se(bi) for df= n-1-p (df: degrees of freedom)

In our example that means that the 95% CI for the coefficient of time spent outdoors is 95%CI: - 0.19 to 0.49

The test for H_0 (βi = 0) is $t = \frac{b_i}{se(b_i)}$ (t-distr. with df = n–1– p)

As in example if we test the H_0: β humidity = 0 and find P = 0.40, which is not significant, we assumed that the association between between toluene personal exposure and humidity could be explained by the correlation between humididty and wind speed[8].

In order to estimate the standard deviation of the residual (Y – Yfit), i.e. the estimated standard deviation of a given set of variable values in a population sample, we have to estimate σ

$$\sigma = \text{Sres} = \sqrt{\sum \frac{(\text{residual})^2}{n-p-1}}$$

The number of degrees of freedom is df = n – (p + 1), since p + 1 parameters are estimated.

The ANOVA table gives the total variability in Y which can be partitioned in a part due to regression and a part due to residual variation:

$$\sum(Y-\bar{Y})^2 = \sum(Y_{fit}-\bar{Y})^2 = \sum(Y-Y_{fit})^2$$

total sum of quares	=	sum of squares due to regression	+	residuals sum of squares
SStotal	=	SSreg	+	SSres

With degrees of freedom (n – 1) = p + (n – p – 1)

In statistical packages the ANOVA table in which the partition is given usually has the following format [6]:

Source	SS	Df	MS	F	P	R²
Regression						
Residual						
Total						

SS: "sums of squares"; df: Degrees of freedom; MS: "mean squares" (SS/dfs); F: F statistics (see below)

As a measure of the strength of the linear relation one can use R. R is called the multiple correlation coefficient between Y, predictors (X1, … Xp) and Yfit and R square is the proportion of total variation explained by regression (R^2=SSreg / SStot).

Test on overall or reduced model

Model: Y= $\beta_0 + \beta_1 X1 + \ldots + \beta\rho X\rho$ + residual

In our example Tpers = $\beta_0 + \beta_1$ time outdoors + β_2 Thome +β_3 wind speed + residual

The null hypothesis (H_0) is that there is no regression overall i.e. $\beta_1 = \beta_2 = \ldots + \beta\rho = 0$

The test is based on the proportion of the SS explained by the regression relative to the residual SS. The test statistic (F= MSreg / MSres) has F-distribution with df1 = p and df2 = n – p – 1 (F- distribution table). In our example F= 5.49 (P<0.01)

If now we want to test the hypothesis Ho: $\beta_1 = \beta_2 = \beta_5 = 0$ (k = 3)

In general k of p regression coefficients are set to zero under H0. The model that is valid if H_0=0 is true is called the "reduced model". The Idea is to compare the explained variability of the model at hand with that of the reduced model.

The test statistic (F):

$$F = \frac{\left(SS_{reg}(\text{full mod el}) - SS_{reg}(\text{reduced mod el})\right)/k}{MS_{res}(\text{full mod el})}$$

follows a F-distribution with df_1 = k and df_2 = n – p – 1.

If one or two variables are left out and we calculate SS reg (the statistical package does) and we find that the test statistic for F lies between 0.05 < P < 0.10, that means that there is some evidence, although not strong, that these variables together, independently of the others, contribute to the prediction of the outcome.

Assumptions

If a linear model is used, the following assumptions should be met. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and the independent variables should be linear, and all observations should be independent. So the assumptions are: independence; linearity; normality; homoscedasticity. In other words the residuals of a good model should be normally and randomly distributed i.e. the unknown σ does not depend on X ("homoscedasticity")[2,4,6,9].

Checking for violations of model assumptions

To check model assumptions we used **residual analysis**. There are several kinds of residuals most commonly used are the standardized residuals (ZRESID) and the studentized residuals (SRESID) [6]. If the model is correct, the residuals should have a normal distribution with mean zero and constant sd (i.e. not depending on X). In order to check this we can plot residuals against X. If the variation alters with increasing X, then there is violation of homoscedasticity. We can also use the Durbin-Watson test for serial correlation of the residuals and casewise diagnostics for the cases meeting the selection criterion (outliers above n standard deviations). The residuals are (zero mean) independent, normally distributed with constant standard deviation (homogeneity of variances)[4,6].

To discover deviations form linearity and homogeneity of variables we can plot residuals against each predictor or against predicted values. Alternatively by using the PARTIAL plot we can assess linearity of a predictor variable. The partial plot for a predictor X_1 is a plot of residuals of Y regressed on other X's and against residuals of Xi regressed on other X's. The plot should be linear. To check the normality of residuals we can use an histogram (with normal curve) or a normal probability plot[6,7].

The goodness-of-fit of the model is assessed by studying the behavior of the residuals, looking for "special observations / individuals" like outliers, observations with high "leverage" and influential points. Observations deserving extra attention are outliers i.e. observations with unusually large residual; high leverage points: unusual x - pattern, i.e. outliers in predictor space; influential points: individuals with high influence on estimate or standard error of one or more β's. An observation could be all three. It is recommended to inspect individuals with large residual, for outliers; to use distances for high leverage points i.e. measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the regression model. For influential points use influence statistics i.e. the change in the regression coefficients (DfBeta(s)) and predicted values (DfFit) that results from the exclusion of a particular case. Overall measure for influence on all β's jointly is "Cook's distance" (COOK). Analogously for standard errors overall measure is COVRATIO[6].

Deviations from model assumptions

We can use some tips to correct some deviation from model assumptions. In case of curvilinearity in one or more plots we could add quadratic term(s). In case of non homogeneity of residual sd, we can try some transformation: **log Y** if Sres is proportional to predicted Y; **square root of Y** if Y distribution is Poisson-like; **1/Y** if $Sres^2$ is proportional to predicted Y; $\mathbf{Y^2}$ if $Sres^2$ decreases with Y. If linearity and homogeneity hold then non-normality does not matter if the sample size is big enough ($n \geq 50$-100). If linearity but not homogeneity hold then estimates of β's are correct, but not the standard errors. They can be corrected by computing the "robust" se's (sandwich, Huber's estimate)[4,6,9].

Selection methods for Linear Regression modeling

There are various selection methods for linear regression modeling in order to specify how independent variables are entered into the analysis. By using different methods, a variety of regression models from the same set of variables could be constructed. Forward variable selection enters the variables in the block one at a time based on entry criteria. Backward variable elimination enters all of the variables in the block in a single step and then removes them one at a time based on removal criteria. Stepwise variable entry and removal examines the variables in the block at each step for entry or removal. All variables must pass the tolerance criterion to be entered in the equation, regardless of the entry method specified. A variable is not entered if it would cause the tolerance of another variable already in the model to drop below the tolerance criterion[6]. In a model fitting the variables entered and removed from the model and various goodness-of-fit statistics are displayed such as R2, R squared change, standard error of the estimate, and an analysis-of-variance table.

Relative issues

Binary logistic regression models can be fitted using either the logistic regression procedure or the multinomial logistic regression procedure. An important theoretical distinction is that the logistic regression procedure produces all statistics and tests using data at the individual cases while the multinomial logistic regression procedure internally aggregates cases to form subpopulations with identical covariate patterns for the predictors based on these subpopulations. If all predictors are categorical or any continuous predictors take on only a limited number of values the mutinomial procedure is preferred. As previously mentioned, use the Scatterplot procedure to screen data for multicollinearity. As with other forms of regression, multicollinearity among the predictors can lead to biased estimates and inflated standard errors. If all of your predictor variables are categorical, you can also use the loglinear procedure.

In order to explore **correlation** between variables, Pearson or Spearman correlation for a pair of variables r (Xi, Xj) is commonly used. For each pair of variables (Xi, Xj) Pearson's correlation coefficient (r) can be computed. Pearson's r (Xi; Xj) is a measure of linear association between two (ideally normally distributed) variables. R^2 is the proportion of total variation of the one explained by the other ($R^2 = b * Sx/Sy$), identical with regression. Each correlation coefficient gives measure for association between two variables without taking other variables into account. But there are several useful correlation concepts involving more variables. The **partial correlation coefficient** between Xi and Xj, adjusted for other X`s e.g. r (X1; X2 / X3). The partial correlation coefficient can be viewed as an adjustment of the simple correlation taking into account the effect of a control variable: r(X ; Y / Z) i.e. correlation between X and Y controlled for Z. The **multiple correlation coefficient** between one X and several other X`s e.g. r (X1 ; X2 , X3 , X4) is a measure of association between one variable and several other variables r (Y ; X1, X2, …, Xk). The multiple correlation coefficient between Y and X1, X2,…, Xk is defined as the simple Pearson correlation coefficient r (Y ; Yfit) between Y and its fitted value in the regression model: Y = β0 + β1X1+ βkXk + residual. The square of r (Y; X1, …, Xk) is interpreted as the proportion of variability in Y that can be explained by X1, …, Xk. The null hypothesis [H_0: ρ (Y : X1, …, Xk) = 0] is tested with the F-test for overall regression as it is in the multivariate regression model (see above)[6,7]. The **multiple-partial correlation coefficient** between one X and several other X`s adjusted for some other X`s e.g. r (X1 ; X2 , X3 , X4 / X5 , X6). The multiple partial correlation coefficient equal the relative increase in % explained variability in Y by adding X1,…, Xk to a model already containing Z1, …, Zρ as predictors[6,7].

Other interesting cases of multiple linear regression analysis include: **the comparison of two group means**. If for example we wish to answer the question if mean HEIGHT differs between men and women?

In the simple linear regression model:

$$\text{HEIGHT} = \beta o + \beta_1 \text{ SEX with SEX} = 1 \text{ for women and SEX} = 2 \text{ for men}$$

Testing β1 = 0 is equivalent with testing

$\text{HEIGHT}_{\text{MEN}} = \text{HEIGHT}_{\text{WOMEN}}$ by means of Student's t-test

The linear regression model assumes a normal distribution of HEIGHT in both groups, with equal σ. This is exactly the model of the **two-sample t-test**. In the case of comparison of several group means, we wish to answer the question if mean HEIGHT differ between different SES classes?

SES: 1 (low); 2 (middle) and 3 (high) (socioeconomic status)

We can use the following linear regression model:

$$\text{HEIGHT} = \beta o + \beta_1 X_1 + \beta_2 X_2 \text{ with } X_1 = 1 \text{ if SES is low and } X_1 = 0 \text{ otherwise and } X_2 = 1 \text{ if SES is middle and } X_2 = 0 \text{ otherwise}$$

Then β_1 and β_2 are interpreted as:

β_1 = difference in mean HEIGHT between low and high class

β_2 = difference in mean HEIGHT between middle and high class

Testing $\beta_1 = \beta_2 = 0$ is equivalent with the "**one-way ANalysis Of VAriance F-test**". The statistical model in both cases is in fact the same[4,6,7,9].

Analysis of covariance (ANCOVA)

If we wish to compare a continuous variable Y (e.g. HEIGHT) between groups (e.g. men and women) corrected (adjusted or controlled) for one or more covariables X (confounders) (e.g. X = age or weight) then the question is formulated: Are means of HEIGHT of men and women different, if men and women of equal weight are compared? Be aware that this question is different from that if there is a difference between the means of HEIGHT for men and women? And the answers can be quite different! The difference between men and women could be opposite, larger or smaller than the crude if corrected. In order to estimate the corrected difference the following multiple regression model is used:

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \text{residual}$$

where Y: response variable (for example HEIGHT); Z: grouping variable (for example Z = 0 for men and Z = 1 for women); X: covariable (confounder) (for example weight).

So, for men the regression line is $y = \beta_0 + \beta_2 X$ and for women is $y = (\beta_0 + \beta_1) + \beta_2 X$.

This model assumes that regression lines are parallel. Therefore β_1 is the vertical difference, and can be interpreted as the: for X corrected difference between the mean response Y of the groups. If the regression lines are not parallel, then difference in mean Y depends on value of X. This is called "**interaction**" or "**effect modification**".

A more complicated model, in which interaction is admitted, is:

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 Z{*}X + \text{residual}$$

regression line men: $y = \beta_0 + \beta_2 X$

regression line women: $y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X$

The hypothesis of the absence of "effect modification" is tested by H_0: $\beta_3 = 0$

As an example, we are interested to answer what is - the corrected for body weight - difference in HEIGHT between men and women in a population sample?

We check the model with interaction:

$$\text{HEIGHT} = \beta_0 + \beta_1 \text{ SEX} + \beta_2 \text{ WEIGHT} + \beta_3 \text{ SEX} * \text{WEIGHT} + \text{residual}$$

By testing β_3=0, a p-value much larger than 0.05 was calculated. We assume therefore that there is no interaction i.e. regression lines are parallel. Further **Analysis of Covariance for ≥ 3 groups** could be used if we ask the difference in mean HEIGHT between people with different level of education (primary, medium, high), corrected for body weight. In a model where the three lines may be not parallel we have to check for interaction (effect modification)[7]. Testing the hypothesis that coefficient of interactions terms equal 0, it is reasonable to assume a model without interaction. Testing the hypothesis H_0: $\beta_1 = \beta_2 = 0$, i.e. no differences between education level when corrected for weight, gives the result of fitting the model, for which the P-values for Z_1 and Z_2 depend on your choice of the reference group. The purposes of ANCOVA are to correct for confounding and increase of precision of an estimated difference.

In a **general ANCOVA model** as:

$$\mathbf{Y = \beta_0 + \beta_1 Z_1 + \ldots\ \beta_{k-1}\ Z_{k-1} + \beta_k\ X_1 + \ldots + \beta_{k+p-1}\ X_p + res}$$

where Y the response variable; k groups (dummy variables $Z_1, Z_2, \ldots, Z_{k-1}$) and $X_1, \ldots, X_p$ confounders

there is a straightforward extension to arbitrary number of groups and covariables.

Coding categorical predictors in regression

One always has to figure out which way of coding categorical factors is used, in order to be able to interpret the parameter estimates. In "**reference cell**" coding, one of the categories plays the role of the reference category ("reference cell"), while the other categories are indicated by dummy variables. The β`s corresponding to the dummies that are interpreted as the difference of corresponding category with the reference category. In "**difference with overall mean**" coding in the model of the previous example: [$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots +$ residual], the β_0 is interpreted as the overall mean of the three levels of education while β_1 and β_2 are interpreted as the deviation of mean of primary and medium from overall mean, respectively. The deviation of the mean of high level from overall mean is given by ($- \beta_1 - \beta_2$). In "**cell means**" coding in the previous model (without intercept): [$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 \ldots +$ residual], β_1 is the mean of primary, β_2 the middle and β_3 of the high level education[6,7,9].

Conclusions

It is apparent to anyone who reads the medical literature today that some knowledge of biostatistics and epidemiology is a necessity. The goal in any data analysis is to extract from raw information the accurate estimation. But before any testing or estimation, a careful data editing, is essential to review for errors, followed by data summarization. One of the most important and common question is if there is statistical relationship between a response variable (Y) and explanatory variables (Xi). An option to answer this question is to employ regression analysis. There are various types of regression analysis. All these methods allow us to assess the impact of multiple variables on the response variable.

References

1. Rothman KJ, Greenland S. Modern Epidemiology, 2nd ed. Lippincot- Raven 1998.
2. Altman DG. Practical Statistics for Medical Research. Chapman & Hall/CRC, 1991.
3. Rosner BA. Fundamentals of Biostatistics, 4th ed. Duxbury, 1995.
4. Draper NR, Smith H. Applied Regression Analysis. Wiley Series in Probability and Statistics, 1998.
5. Munro BH. Statistical Methods for Health Care Research, 5th ed. Lippincott Williams & Wilkins, 2005.
6. SPSS 15.0 Command Syntax Reference 2006, SPSS Inc., Chicago Ill.
7. Stijnen T, Mulder PGH. Classical methods for data analyses. NIHES program, Rotterdam, 1999.
8. Alexopoulos EC, Chatzis C, Linos A. An analysis of factors that influence personal exposure to toluene and xylene in residents of Athens, Greece. BMC Public Health. 2006; 6: 50.
9. Shedecor GW, Cochran WG. Staistical Methods, 8nd ed. Iowa State Univ Press, 1989.