http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (5): 884-891, 2006
ISSN 1812-5638

A Software Tool for Regression Analysis and its Assumptions

Sona Mardikyan and Osman N. Darcan
Department of Management Information Systems, Bogaziçi University, Istanbul, Turkey

Abstract: Nowadays, among the forecasting methods, the most important one is the regression analysis. In this method, the aim is to estimate the population regression model as much as accurate by taking as basis the sample regression function. Its results are valid under certain assumptions and the violations of these assumptions cause the invalidity of some properties of the estimators. In this study, a new object-oriented program concentrated only on the regression analysis and its assumptions has been developed using Java, to carry out this analysis more easily and in a shorter time. In this program, regression model selection, regression and correlation analysis with Least Square method, one test for every assumption and solution methods has been presented. All the results of the analysis are illustrated by using a multiple regression example.

Key words: Regression analysis, autocorrelation, hetreoscedasticity, normality, multicollinearity

INTRODUCTION

Regression analysis is one of the most widely used statistical technique, in all fields of the business administration, economics, social sciences, engineering, physical, chemical and biological sciences. It is a method to evaluate the relationship between one or more independent variables (explanatory variables) and one dependent variable (explained) variable, so that the latter can be predicted from the others. The linear regression model is based on a set of assumptions that not only describe the model but also imply appropriate estimation and inference procedures. Diagnostic methods are used to examine the appropriateness of assumptions and to locate unusual characteristics of the data that may influence conclusions. Therefore, the determination and the solution procedures for the assumptions' validity are very important and discussed in most of text-books in econometrics (Greene, 2000; Gujarati, 2003) and in regression analysis (Montgomery *et al*., 2001; Kleinbaum *et al*., 1998; Ryan, 1997; Miles and Shevlin, 2001).

Today, regression analysis exists in many statistical programs such as SPSS, SAS, STATISTICA, STATA, MINITAB, EVIEWS. In almost all of these programs, parameters of the regression equation, standard errors, correlation coefficients and related tests are presented similarly under regression analysis title. However, methods for investigating assumptions' validity and solution analysis are under different titles or they can only be handled by Macros as in SPSS. For this reason, in order to be able to apply directly the regression analysis together with its assumptions, a new program has been developed in Java programming language.

REGRESSION ANALYSIS AND ITS ASSUMPTIONS

The general form of the multiple linear regression models with (p-1) independent variables is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ...+ \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad I = 1,...,n \tag{1}$$

where Y is the dependent variable, $X_1, X_2,\ldots, X_{p-1}$ are the independent variables, ε is the random residual (disturbance) term, $\beta_0, \beta_1,\ldots, \beta_{p-1}$ are the regression coefficients and n is the sample size. In matrix terms, formula (1) becomes (Neter *et al*., 1983):

$$Y = X\beta+\varepsilon \tag{2}$$

where Y is a nx1 vector of observations, β is px1 vector of regression parameters, X is nxp matrix of constants, ϵ is a nx1 vector of independent normal random variables.

The fundamental assumptions of the above model are as follows (Greene, 2000):

No autocorrelation: Given any two X values, the correlation between any two ε_i and ε_j $(i\neq j)$ is zero.

Corresponding Author: Sona Mardikyan, Bogaziçi Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, Hisar Kampus, 34342 Bebek, Istanbul Tel: +90 212 359 69 42 GSM: +90 532 204 78 87

Homoscedasticity: Given the value X, the variance of ε_i is a positive constant number σ^2.

Normality: Given the value of X, the ε_i has a normal distribution around the $E(Y_i)$. Thus, the conditional mean value of ε_i is zero.

No multicollinearity: No exact linear relationships between the independent variables.

The most common method of estimating the parameters of the linear regression model is the Least Squares method where the vector of estimated regression coefficients is denoted as b that can be calculated using the following formula:

$$b = (X'X)^{-1}XY. \quad (3)$$

Testing and solution procedures of the assumptions: No autocorrelation: The autocorrelation can be detected by graphical methods in which the residuals or standardized residuals are simply plotted against time or dependent variable. There also exist some powerful quantitative tests such as the Durbin-Watson test (Durbin and Watson, 1950, 1951, 1971), Number of Runs test (Geary, 1970), Farebrother test (Farebrother, 1980), Von-Neumann Ratio test.

Among these tests the most frequently used one is the Durbin-Watson test in which the distribution of the test statistic d depends on X matrix and it is also closely related to ρ which is the autocorrelation coefficient. Durbin and Watson showed that d lies between two critical values d_L and d_U and they depend only on n and p. As stated in Gujarati (2003), the decision procedure is as follows (Table 1).

If the autocorrelation coefficient ρ is known, to handle the autocorrelation the most popular method is the method of Generalized Differences. In this method, firstly the generalized difference equation is obtained by using the transformed variables in the regression model. Then, the Least Squares method is applied to estimate the parameters of this model. If ρ is not known, the same method can be applied by estimating ρ based on Durbin-Watson d statistic or directly from residuals.

No heteroscedasticity: Contrary to homoscedasticity, if the conditional (upon given X_i) variance of ε_i increases as X_i increases, we say that there is heteroscedasticity. To test this assumption, in addition to the graphical methods, there are some known formal tests. These are Spearman's Rank Correlation test, White test (White, 1980), Breush-Pagan-Godfrey test (Breush and Pagan, 1979; Godfrey, 1987), Koenker-Bassett test (Koenker and Bassett, 1982).

Table 1: Decision rules for Durbin-Watson test

d Value ranges	Decision
$0 < d < d_L$	Reject H_0 (positive autocorrelation)
$d_L \le d \le d_U$	No decision
$d_U < d < 4-d_U$	Do not reject H_0 (no autocorrelation)
$4-d_U \le d \le 4-d_L$	No decision
$4-d_L \le d \le 4$	Reject H_0 (negative autocorrelation)

In the Spearman's Rank Correlation test, the absolute value of the residuals and the independent variable X_i (assuming that the variance of the error terms are positively related to this variable) values are ranked according to an ascending or descending order. Spearman's rank correlation coefficient r_s is calculated based on the differences in the ranks assigned to these variables. Assuming that the population rank correlation coefficient $\rho_s = 0$, the significance of the r_s can be tested by the t-test with df = n-2.

To handle the heteroscedasticity, the most popular method is the Weighted Least Squares (WLS) in which the estimators are computed by the following formula (Montgomery *et al.*, 2001):

$$b_w = (X'WX)^{-1} X'WY \quad (4)$$

where W is a diagonal matrix with weights w_1, w_2,...,w_n on diagonals. In this method, the smaller weights are assigned to the observations with large variances than observations with small variances. By selecting w_i as $1/\sigma_i^2$, it is shown that, the WLS estimators minimize the weighted residual sum of squares (Myers, 1990). If the σ_i^2 is not known, then taking into consideration the nature of the variance, certain assumptions can be made to determine the weights. The followings are the most powerful assumptions;

$$\sigma_i^2 = \sigma^2 X_i^2; \quad \sigma_i^2 = \sigma^2 X_i; \quad \sigma_i^2 = \sigma 2[E(Y)]^2 \quad (5)$$

Normality: To test the normality, several tests are discussed in the literature. In graphical methods, histogram of the residuals and the normal probability plot can be used to analyze the normality assumption of the model. In the formal methods, the most frequently used tests are Chi-Square Goodness of Fit test, Jarque-Bera test (Jarque and Bera, 1987), Kolmogorov-Smirnov test.

In the Jarque-Bera test, the test statistic JB value is computed based on the skewness and the kurtosis of the residulas. Under the null hypothesis that the residuals are normally distributed, the test statistic follows a chi-square distribution with df = 2.

The most popular ways to find a model in which the normality assumption is satisfied, are to increase the sample size, to transform variables, to apply WLS or Robust Regression (Huber, 2004; Rousseeuw and Leroy, 2003).

No multicollinearity: There are some multicollinearity diagnostics (Myers, 1990) which are not certain but can aid to detect severe multicollinearity.

- Some or all the simple correlation coefficients among the independent variables are very high.
- High determination coefficient is obtained but t-tests show that none or very few of the regression coefficients are statistically different than zero.
- Variance Inflation Factors (VIF) is a measure of collinearity and is defined as

$$VIF_i = \frac{1}{1 - R_i^2} \quad (6)$$

where R_i^2 is the determination coefficient in the regression of independent variable X_i on the remaining independent variables of the model. In general, if any VIF exceeds 10, corresponding variable is said to be highly collinear.

To handle multicollinearity, the following rules can be used depending on the nature of the data: using prior information, omitting highly collinear data, obtaining new data or transforming data. Another approach is the respecification of the model by employing different subsets of the independent variables. The most popular approach for the model selection is the Stepwise Regression method in which only a small number of subsets are evaluated by either adding or deleting independent variables one at a time. The criterion for adding or deleting an independent variable can be stated equivalently in terms of error sum of squares reduction or F statistic (Neter *et al*., 1983). The procedure terminates when no additional independent variables can enter on the basis of F_{IN} and no independent variables in the model can be eliminated on the basis of F_{OUT}.

In the estimation procedures to combat multicollinearity, the Ridge Regression proposed by Hoerl and Kennard (1970a, 1970b) is the most powerful method. In this method, the estimators are obtained by introducing into the least squares normal equations a biasing constant c = 0 and can be calculated using the following formula:

$$b_R = (X'X + cI)^{-1}X'Y \quad (7)$$

This constant reflects the amount of bias in the estimators. The difficulty of this model is to determine the optimum c value. A commonly used method is to list simultaneously the values of the Ridge Regression estimators calculated for different values of c, usually between 0 and 1. Then by examining the list, the smallest value of c for which the regression coefficients become consistent, is chosen.

A PROGRAM FOR REGRESSION ANALYSIS

General Overview of the Program: In our program, the user is faced with a window including four main menus as shown in Fig.1. Under these menus, the file operations, model selection process, regression and correlation analysis, assumptions validity tests and solution methods are presented.

The user can create a new file to enter the data, open an existing file, display the data in tabular form, edit, save and print the file by using the menu items under the "File" menu. A variety of models can be produced by using the "Model" menu. Correlation and regression analysis can be applied to these models by the "Regression" menu. As to the "Assumptions" menu, there are four subitems, each related to the main assumptions of the regression analysis. When the user selects a subitem, a new submenu appears next to it. Each menu involves one popular test to detect the validity of the related assumption and the solution procedure(s) that can be used in case of invalidity. For multicollinearity, VIF values are calculated as a measure of diagnostic, Stepwise and Ridge Regression methods can be applied to the model as solution procedures. The most popular test Durbin-Watson and method of Generalized Differences are subitems of autocorrelation menu. Spearman Rank Correlation test is provided to detect the homoscedasticity and in case of invalidity Weighted Least Squares method can be applied to the model. The reason of selecting Spearman Rank Correlation test among others is that it can determine not only the existence of the heteroscedasticity but also the independent variable(s) that may have a relation with the residuals of the model. The last subitem is devoted to normality assumption and includes the Jarque-Bera test.

In our program, all the outcomes of these analyses are written on the main screen one after the other as they are selected. For the sake of illustration, we choose an

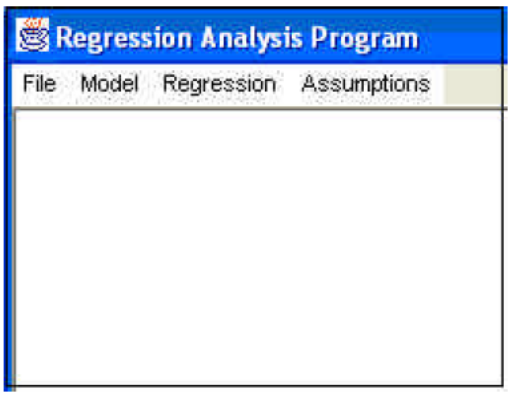


Fig. 1: Main window of the program

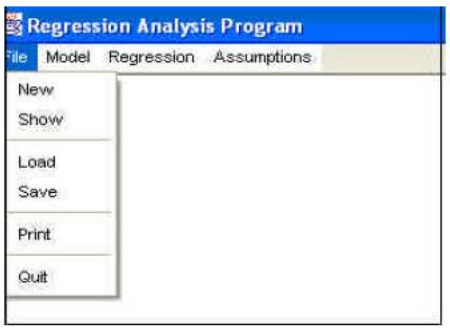


Fig. 2: "File" menu window

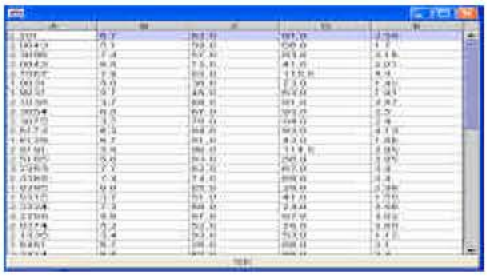
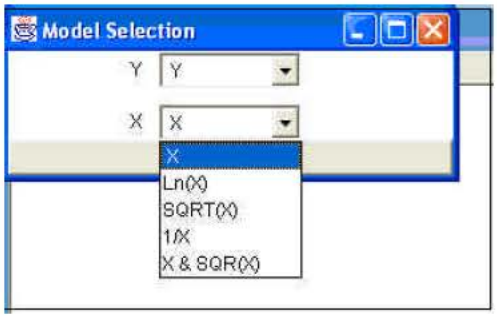


Fig 3: "Show" menu window

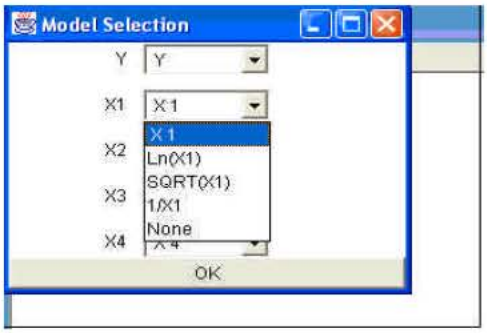
example of a multiple regression data that consist of 54 observations of four potential independent variables and a dependent variable (Neter *et al.* 1983). All the results of this example are verified by the widely used statistical program, SPSS.

File menu: It is designed to perform file operations like in other Windows based programs (Fig. 2). "New" subitem provides to create a new file and enter the data in it, whereas the "Load" provides to open an existing file. It is possible to view and change the data by selecting the "Show" option at any time of the program. The window shown in Fig. 3 is obtained by applying "Show" option to our example after entering it by the "New" option. The first column includes dependent variable whereas the others include the independent variables in the model. The file is saved whenever "Save" option is selected. "Print" option is available to take printouts and "Quit" option to terminate the program.

Model menu: The user is asked through a dialogue window to build up the model with the data entered by using "New" or "Load" options of "File" menu. Through "Model Selection" window, the variables can be used directly as they are entered or various transformations can also be applied to them. In the simple regression, two variables, X and Y will appear in the list. For Y the



(a) Simple regression



(b) Multiple regression

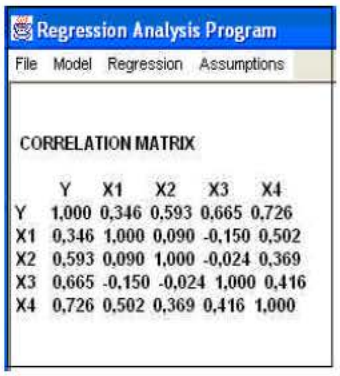
Fig. 4: Windows used for model selection

following transformations are available: Ln(Y), SQRT(Y) and 1/Y. For the independent variable, available options such as Ln(X), SQRT(X), 1/X and X and SQR(X) are shown in Fig. 4-a. The last item (X and SQR(X)) can be selected when a quadratic form of the model is considered.

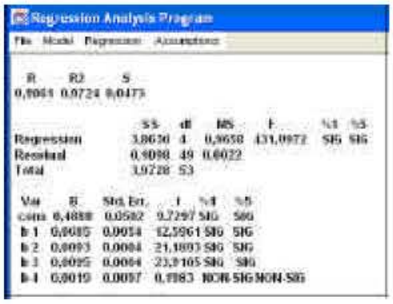
In the multiple regression, the dependent and independent variables are automatically listed in "Model Selection" window. For the dependent variable, the same options as in the simple model are valid and for independent variables, the last option is replaced by the "None" option that enables the user to exclude the independent variable from of the model. Hence, multiple regression models can be specified with some of the entered variables. As a result, the researcher is able to form linear models or models that have been transformed to linear form by applying various transformations to their variables. Deciding suitable model for data by comparing the results of the analysis is left to the researcher. For this purpose, returning to "Model Selection" menu at any point of the program, changing the variable set as described above then applying analysis from the beginning is made possible.

The window shown in Fig. 4-b will appear when the "Model Selection" menu is selected for our example. No transformations are applied to the variables; that is the first menu item is directly selected to construct the model for the further analysis.

Regression menu: After the model has been developed as described in the previous section, the user can apply the regression analysis through this menu in which there are "Correlations" and "Regression Equation" subitems. When the user selects the "Correlations" option, the simple correlation coefficients between the variables of the model are computed and shown on the main screen. The second option "Regression Equation" is used to apply the regression analysis to the model. As in other statistical programs, to form the regression equation, first the Least Squares estimators are computed. Following that, the multiple correlation and determination coefficients, the standard error of the model and the ANOVA table are given. The last part of the report includes the values, the standard errors and t-test results of the estimators. Here, the part that is different from other programs is that test results are interpreted for 1% and 5% significance levels.



(a) "Correlations" results



(b) "Regression equation" results

Fig. 5: Windows of "Regression" menu

In our examples, when the "Correlations" and "Regression Equation" options are selected the program generates the following outcomes as shown in Fig. 5. All the figures following this figure illustrate the analysis results of this example.

According to these results, it is clear that, the independent variables can describe the variability of dependent variable at 0.986 ratio in the population. The F test of the model and t-tests of b_1, b_2 and b_3 coefficients are found to be significant, but t-test result of b_4 coefficient is insignificant for both levels.

Assumptions menu: As mentioned before, in this part of the program, the main purpose is to test the four main assumptions with one chosen method. In case of violations, the user can obtain new regression models easily by one of the proper methods whose menu item becomes automatically enabled by the program.

Multicollinearity submenu: There are three subitems under this menu as shown in Fig. 6. The VIF values of all independent variables of the chosen model are calculated and added to the report by selecting the first option of this menu. By taking these values into consideration with

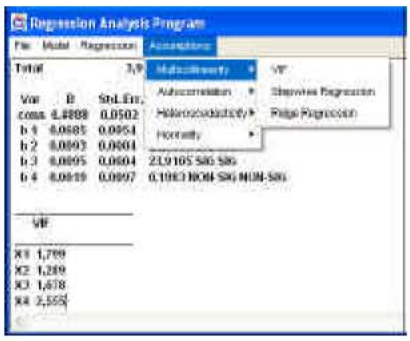


Fig. 6: VIF values

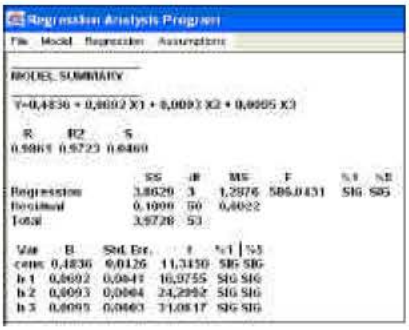


Fig. 7: Stepwise Regression results

regression analysis results, it can be investigated that which of the independent variables have a linear relation between others at an important level.

Considering our example, VIF values are computed as shown in Fig. 6.

In this model, even though the obtained VIF values are not greater than 10, some of the simple correlation coefficients between the independent variables are high. In this case, in order to choose the variables which have significant effect on the dependent variable but have no relation with other independent variables, the Stepwise Regression method can be applied. The second item of this submenu provides this method to the user. The results shown in Fig. 7 are obtained when the Stepwise method is applied to this example. According to the result of this method, the efficient set of independent variables includes only X_1, X_2 and X_3.

A useful method to include all the variables in the model is the Ridge Regression method which can be applied by the third item of this submenu. If this item is selected, the user is faced with a new window in which various values of the biasing constant c and the corresponding biased estimators are shown. The biasing constant varies from 0 to 0.99 by increasing 0.01 at each step. The user selects the suitable c value for which estimators become consistent and enters it into the textbox that corresponds to the title of "Enter c".

The Ridge Regression parameters are obtained as shown in Fig. 8, when this method is applied to our example. According to these results, the user can specify the best value of c as about 0.25. When this value is entered to the required box and "Find Model" button is pressed, the program will build up the Ridge Regression model as shown in Fig. 9.

Fig. 8: Window for determining "c" value in the Ridge Regression

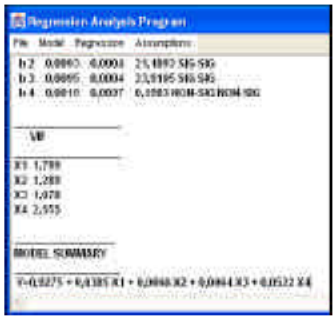


Fig. 9: Window of the Ridge Regression results

Autocorrelation submenu: It provides to apply the Durbin-Watson test and the method of the Generalized Differences which can be used in case of autocorrelation.

Before implementing the Durbin-Watson test, first, regression analysis should be applied to the chosen model by using "Regression Equation" or "Stepwise Regression" options. Similar to other statistical programs, the Durbin-Watson test statistic is calculated and added to the main screen when this option is selected. Additionally, in our program, the existence of the autocorrelation is determined by interpreting this value for both significance levels based on the rules given in Table 1. This feature that does not exist in the other examined programs facilitates the analysis of the researchers. If the autocorrelation is detected for at least one of the significance levels or a decision could not be made, then the second option is automatically activated by the program. As mentioned before, the Generalized Differences method estimates the parameters of the model by the Least Squares method after the transformations are applied to the variables. Thus, the new parameters are calculated and the generalized difference equation is added to the main screen automatically by the program, when this option is selected. In order to determine whether the autocorrelation exists in the newly constructed model, the Durbin-Watson test is applied automatically and the outcomes are interpreted in respect to the same criteria. Hence, the researcher both implements the Generalized Differences method to the model by only one option and checks whether or not the new model involves autocorrelation. We believe that this is furthermore an important advantage which does not exist in the other programs.

To test our example for autocorrelation, after applying the Stepwise Regression method, the Durbin-Watson test results are obtained as shown in Fig. 10.

According to these results, the model does not have the autocorrelation for both significance levels. As mentioned before, the method of the Generalized Differences will not become active (Fig. 10) and therefore can not be applied.

Heteroscedasticity submenu: The user can apply the Spearman Rank Correlation test to the model by selecting the first option of this submenu. Through this option, the Spearman Correlation coefficients between each independent variable and the residual are computed and their significance is automatically interpreted. When all

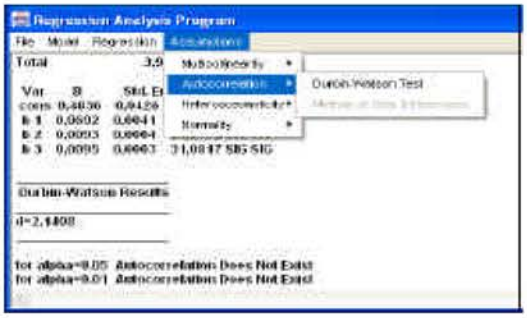


Fig. 10: Window of the Durbin-Watson test results

Spearman Correlation coefficients are insignificant, the second option of this title, "Weighted Least Squares" does not become active because homoscedasticity is valid. This feature will also provide considerable time savings to researchers.

In case of heteroscedasticity, a new dialog window is displayed to list all the independent variables which have significant Spearman Correlation coefficient for at least one significance level. The user is asked to choose one of the independent variable from the list, so that WLS method can be applied by using various weights calculated from this independent variable. In our program, 9 different weight values are considered and calculated using the following formula:

$$1/X^{\gamma}$$

where, γ varies from-2 to 2 incremented by 0.5 to obtain each weight. Like in other programs, it is required to compute the maximum likelihood function of the models obtained by the different weights, in order to decide the best estimate.

Considering our example, the Spearman Rank Correlation test results are obtained as in Fig. 11. As seen from here, only the correlation between X_1 and residuals is found to be significant. In this case, the option of "Weighted Least Squares" has been activated automatically and if it is chosen, in the opened window, only X_1 is listed as shown in the same figure. Figure 12 shows the results of the WLS method. Here, the weight of X_1 that maximizes the Maximum Likelihood function is the $1/X_1^{-2}$. Therefore, in the model summary the WLS equation is obtained based on this weight. Moreover, the correlation and determination coefficients and the ANOVA table of this model are added to the main screen. Similar to the other parts of the program the significance of the parameters are interpreted in the last part of the results.

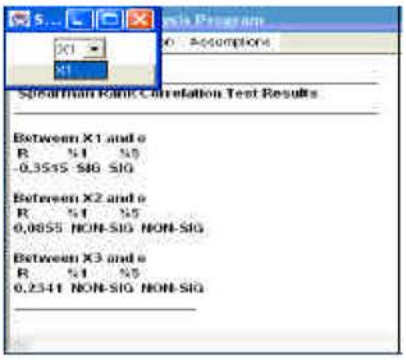


Fig. 11: Window of the Spearman Rank Correlation test results

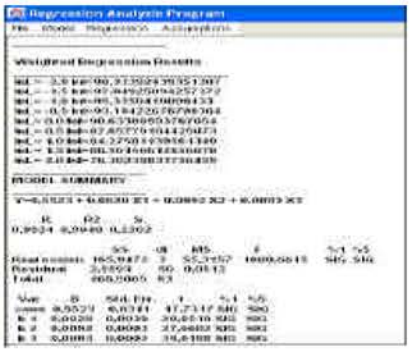


Fig. 12: Window of the WLS method results

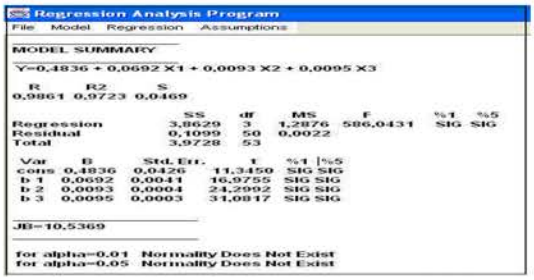


Fig. 13: Window of the Jarque-Bera test results

The method can be applied separately by selecting different independent variables from the list provided by the program. The determination of the best model is left to the researcher at this point.

Normality submenu: The last subitem for investigating the normality assumption is the Jarque-Bera test. The program computes and interprets the test statistic of this method for the model estimated by "Regression Equation" or "Stepwise Regression" options.

The results are obtained as shown in Fig. 13 after applying the Stepwise Regression method to our example.

In the current status of the program, an algorithmic method to handle the violation of this assumption is not included in the program. But, as mentioned before the WLS method can be used to overcome this violation and therefore estimate a suitable model.

CONCLUSIONS

In our program, all the required tests and methods for the regression analysis are combined under a single menu and applicable easily without needing any additional computations or operations. This property enables the researchers to concentrate on the outcomes faster. Another property of the program is that, the transformation functions that are commonly used to apply Least Squares to models which can't be explained linearly are collected in the lists which are opened below the variables. Thus, for each variable, the different functions can be selected easily and furthermore for the multiple regression models, the independent variables may be excluded from the model by the same way. The tests for the assumptions' validity and the solution methods can be applied consecutively on the different models obtained by the model variations in a fast way and therefore the appropriate model for the data can be chosen easily. The other significant superiority of the program is that, in addition to the given outcomes by the other programs, it makes comments on the test results and orientates the researcher for his further analysis at the following steps. For this purpose, within all tests, the test statistics are interpreted by comparing them to the related table values automatically for the widely used 1% and 5% significance levels.

As a result, we believe that the program facilitates the model construction and the work to be done for testing the assumptions' validity in the regression analysis. Furthermore, it provides a platform independent environment to the user and other tests and solution methods for assumptions' validity can also be added easily.

REFERENCES

Breush, T. and A. Pagan, 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47: 1287-1294.

Durbin, J. and G. Watson, 1950. Testing for serial correlation in least squares regression-I. Biometrica, 37: 409-428.

Durbin, J. and G. Watson, 1951. Testing for serial correlation in least squares regression-II. Biometrica, 38: 159-178.

Durbin, J. and G. Watson, 1971. Testing for serial correlation in least squares regression-II. Biometrica, 58: 1-42.

Farebrother, R.W., 1980. The Durbin-watson test for serial correlation when there is no intercept in the regression. Econometrica, 48: 1553-1563.

Geary, R.C., 1970. Relative efficiency of count of sign changes for assessing residual autocorrelation in least regression. Biometrica, 57: 123-127.

Godfrey, L., 1987. Testing for multiplicative heteroscedasticity. J. Econometrics, 8: 227-236.

Greene, W.H., 2000. Econometric Analysis. 4th Edn., Macmillan Publishing Company, New York.

Gujarati, D.N., 2003. Basic Econometrics. 4th Edn., McGraw-Hill, New York.

Hoerl, A.E. and R.W. Kennard, 1970a. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12: 55-67.

Hoerl, A.E. and R.W. Kennard, 1970b. Ridge regression: Applications to nonorthogonal problems. Technometrics, 12: 69-82.

Huber, P.J., 2004. Robust Statistics. Wiley-Interscience, New York.

Jarque, C.M. and A.K. Bera, 1987. A test for normality of observations and regression residuals. Intl. Stat. Rev., 55: 163-172.

Kleinbaum, D.G., L. L. Kupper, K.E. Muller and A. Nizam, 1998. Applied Regression Analysis and Other Multivariable Methods. 3rd Edn., Duxbury Press, Washington.

Koenker, R. and G. Bassett, 1982. Robust tests for heteroscedasticity based on regression quantiles. Econometrica, 50: 43-61.

Miles, J. and M. Shevlin, 2001. Applying regression and correlation. SAGE Publications, London.

Montgomery, D., E. Peck and G.G. Vining, 2001. Introduction to Linear Regression Analysis, 3rd Edn., Jon Wiley and Sons, New York.

Myers, R.H., 1990. Classical and Modern Regression with Applications, 2nd Edn., Duxbury, Washington.

Neter, J., W. Wasserman and M.H. Kutner, 1983. Applied Linear Regression Models. Richard. D. Irwin, Inc., Illinois.

Rousseeuw, P.J. and A.M. Leroy, 2003. Robust Regression and Outlier Detection. Wiley-Interscience, New York.

Ryan, T.P., 1997. Modern Regression Methods. John Wiley and Sons, Inc., New York,

White, H., 1980. A Heteroscedasticity- Consistance covariance matrix estimator and a direct test of heteroscedasticity. Econometrica, 48: 817-838.