Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 5 (10): 1792-1796, 2005
ISSN 1812-5654

Importance of Diagnostics in Multiple Regression Analysis

E. Eyduran,T. Özdemir and E. Alarslan
Department of Animal Science, Biometry Genetics Unit, Faculty of Agriculture,
University of Yüzüncü Yil, 65080 Van, Turkey

Abstract: The aim of this study was to obtain some valuable information from different diagnostics in Multiple Regression Analysis (MRA). Sample data set was composed of live weights at different periods (birth weight (X_1), live weights in 30th (X_2), 45th(X_3), 60th (X_4) and 75th (Y) days) of 18 Hamdani breed single-male lambs born in early March of 2001. According to results of MRA, although all independent variables including in model explained approximately 92% of variation in dependent variable, Y, the effect of only independent variable X_4 on dependent variable Y was significant ($p<0.01$). With respect to residual analysis, it could be said that the assumptions of normal distribution and homogeneity of error terms in MRA were provided. As the value of Durbin-Watson statistics equaled to 2.31, there was not a sequent correlation among error terms, that is, the assumption that error terms independent from each other was ensured. Considered the leverage and influence diagnostics calculating for observations of sample data set, only two observations (2nd and 16th observations) of all observations-both outliers and potential effective (influence) observations- should be carefully examined. It could be concluded that diagnostics would be an important statistics for researchers because they could give an idea about whether the basic assumptions would be provided for reliability of MRA, data set and goodness of fit.

Key words: Diagnostics, outliers, influence observation, Durbin-Watson

INTRODUCTION

Multiple Regression Analysis (MRA) is commonly used in all science fields. As being in other analysis techniques, MRA should be provided with some assumptions for reliable estimation of parameters: expected value of residual terms should be zero; residual terms should have a normal distribution; residual terms should be independent from each other; observation number should be more than parameter number; there should not be multicollinearity between or among independent variables[1,2].

The aim of MRA is to find the best set of the independent variables which can explain dependent variable on condition that the assumptions are provided[1,3]. Diagnostics are analysis techniques that given an idea about determining levels of unfavorable cases such as lack of model and heterogeneity of variances which can be encountered in data set[1,3,4].

This paper dealt with some problems by using diagnostics mentioned below. Therefore, the aim of this study was to obtain some valuable information from different diagnostics in Multiple Regression Analysis (MRA).

MATERIALS AND METHODS

Materials of this study were composed of 18 male-single lambs randomly selected from Hamdani lambs raised in Van province of Turkey. Data of body weights at different periods (birth weight, body weights at 45th, 60th and 75th days) of the lambs were recorded. The data set was analyzed by using SAS program[5].

MRA is used to explain effects of independent variables on dependent variables. Model of MRA can be written as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_1$$
$$I = 1, 2....n \qquad (1)$$

Where, Y, dependent variable; $X_1, X_2 ... X_k$ are independent variables, β_0, β_1, $\beta_2 ... \beta_k$ (regression coefficients (slopes) and ϵ_i random error.

Equation 1 can be rewritten as $Y=X\beta + \varepsilon$ in matrix notation where X, design matrix; β, coefficients vector of regression coefficients and ϵ, vector of random error. Regression coefficients can be estimated by Ordinary Least Square (OLS) Method. The method is based on

Corresponding Author: Ecevit Eyduran, Department of Animal Science, Biometry Genetics Unit, Faculty of Agriculture, Yüzüncü Yil University 65080, Van, Turkey Tel: +904322251024/2648

minimizing $\sum_{i=1}^{n} e_i^2 = Y - \hat{Y}$, difference between observed Y values with predicted $\hat{Y}_i$ values. $\hat{\beta} = (X'X)^{-1}(X'Y)$ is solved by using OLS then β_0, β_1, β_2,....β_k were calculated[2].

Diagnostics: Regression diagnostics are statistics used for detecting problems which are encountered in model or data set[1,3]. Let's examine Diagnostics by turns.

Leverage points diagnostics: The diagnostics composed of residual Analysis, standardized residuals, studentized residuals and Hat matrix.

Residual analysis: Residual, difference between observed Y values with predicted $\hat{Y}_i$ values, denotes by e_i. The term can be obtained by Eq. 2. The assumptions that variance and expected values of error terms in MRA should be fixed, which is denoted by var (e) = $\sigma^2 I$ and E (e) = 0 [2,4,6].

$$e_i = Y_i - \hat{Y}_i \quad (2)$$

Standardized residuals: The diagnostic, which is denoted by r_i, the ratio of each residual to standard deviation of all residuals[1-4, 6], can be written as follows:

$$r_i = \frac{e_i}{\sqrt{s^2(1-h_{ii})}} = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad (i = 1,2,\ldots n) \quad (3)$$

Where, e_i is residual, s term is the root of means squares of error and diagonal elements of hat matrix, h_{ii}.

Studentized residuals: Each residual is standardized with standard deviation which is calculated after it is released out of calculation[1-4,6]. After the ith observation is removed from data set, variance for ith residual is denoted by $s^2_{(i)}$, estimating from the rest of data set. $s^2_{(i)}$ can be calculated below:

$$S_i^2 = \frac{(n-p-1)s^2 - e_i^2 / (1-h_{ii})}{n-p-2} \quad (4)$$

Thus residual value converted to Student's t, denotes by r_i^* and can be written follow as:

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \quad (i = 1, 2...n) \quad (5)$$

In application[1], Eq. 5 can be expressed as Eq. 6:

$$r_i^* = \sqrt{\frac{n-p-2}{n-p-1-r_i^2}} \quad (i= 1, 2...n) \quad (6)$$

The i^{th} observation is an outlier if $|r_i^*| > 1.96$ or $|r_i^*| > 2$ The e_i, r_i and r_i^* values are based on studies related to effectiveness of model estimation[3,4]. In most observations, these three values can have similar results. It was reported that studentized residuals can be used as appropriate criterion in point of size of residuals[1,2].

Hat matrix: Consider a matrix H;

$$H = X(X^T X)^{-1} X^T \quad (7)$$

Where, X is data matrix containing independent variables. First column of matrix X is only 1's corresponding to intercept and matrix X^T is transpose of matrix X. The Eq. 7 is called as Hat Matrix whose diagonal elements are denoted by h_{ii}.

The h_{ii} value is an indicator of the leverage of data point concerning i^{th} observation from space centre of X variables (X_1, X_2... X_n). In other words, the value, which ranges from 0 to 1[2,4] as well as lies between 1/n and 1/r according to other author where n, is the number of observations and r is the number of i^{th} observation and/or is shown whether i^{th} observation will be an outlier in a space of X variables[6].

The critic value or cut-off value for the statistics is 2 p′/n where the number of parameters or independent variables and regression constant (intercept =1), respectively, is denoted by p' and p. For instance, let's we have 3 independent variables in a model. The number of parameters estimated equals to p′ = 3+1 = 4. Observations whose h_{ii} values are larger than 2 p′/n values can be expressed as outliers in place of X variables[1-4].

Durbin-Watson: The statistics whose optimum value ranges from 2 to 4 is used in determining sequent correlation among residuals[1-4,6,7] and is calculated as:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \quad (8)$$

Calculated value of Durbin-Watson statistics is compared with the table value(s) containing the cut-off values for the statistics[4].

Influence statistics: The influence diagnostics comprised of Cook Distance, Differences between the fits (DFFITS), Differences between the betas (DFBETAS) and Covariance ratio (COVRATIO).

Cook's distance: Cook's distance is shown the combined effects of i^{th} observation on all regression coefficients. Observations whose values are larger than the cut off value for Cook's distance, 4/n, can

be expressed as influential observations and said to be observations $\hat{\beta}$ effective on. The statistics can be calculated by Eq. 9:

$$D_i = \frac{\left(\hat{\beta}_i - \hat{\beta}\right)' (X'X) \left(\hat{\beta}_i - \hat{\beta}\right)}{p's^2} \quad (9)$$

Where $\hat{\beta}_i$, is calculated in the event of deletion of i^{th} observation and the other $\hat{\beta}$ is normally calculated[1-4].

Differences between the fits (DFFITS): The statistics is given the changes of predicted $\hat{Y}_i$ is given when ith observation is ignored and its expression can be written as follows:

$$(DFFITSi)^2 = \frac{\left(\hat{\beta}_i - \hat{\beta}\right)' (X'X) \left(\hat{\beta}_i - \hat{\beta}\right)}{\hat{\sigma}^2_{(i)}} \quad (10)$$

The cut value for the statistics is $2\sqrt{p'/n}$ and if DFFITS value of ith observation is larger than the cut value, it can be said to be effective of the observation on $\hat{Y}_i$[1,3,4].

As there is a close association between the statistics and Cook's distance, results of both statistics are similar[1-4,6].

Differences between the betas (DFBETAS): The statistics is based on measured influence of i^{th} observation on each regression coefficient and obtained from standardized differences between $\hat{\beta}_J$ and $\hat{\beta}_{J(i)}$.

$$(DFBETAS)_{J(i)} = \frac{\left(\hat{\beta}_J - \hat{\beta}_{J(i)}\right)}{s_{(i)}\sqrt{C_{ii}}} \quad (11)$$

Where, $\hat{\beta}_{J(i)}$ is obtained from ignored to i^{th} observation.

The cut off value for DFBETAS is $2\sqrt{n}$ and if DFBETAS value of i^{th} observation is larger than it, it can be said to be effective of i^{th} observation on j. regression coefficient[1-4,6].

Covariance ratio (COVRATIO): The statistics is the ratio of determinant of variance-covariance matrix calculated when ith observation is omitted to determinant of variance-covariance matrix calculated when all observations are considered.

The ratio, if closes to 1, influences of i^{th} observation on regression coefficients is small. If the ratio is larger than 1, its influence is larger compared to approximate ratio of 1.

The cut off values for COVRATIO are expressed as $COVRATIO_i \geq 1+3\ p'/n$ or $COVRATIO_i \leq 1-3\ p'/n$[1-3,6].

RESULTS AND DISCUSSION

Descriptive statistics of live weights at different periods of Hamdani breed 18 male-single randomly selected lambs born in early March of 2001 are presented in Table 1.

As examining in Table 2, correlations between different pairs of independent variables were more significant and much higher which showed an evidence for multicolinearity[1-4].

As shown in Table 3, the ratio of model explanation was 0.9186% in case of all being independent variables in model. In case of reliability of model, with coefficient of determination is much higher, assumptions (homogeneity of variance, expected value of error is zero) should be provided[1-4]. Because of context of assumptions and reasons mentioned, it is inevitable that the diagnostics should be taken into account for MRA. The effect of only 60th live weight as independent variable on 75th live weight was significant. Besides, with respect to result of stepwise elimination method that the most ideal set of independent variables was determined; the effect of only 60th live weight on 75th live weight was significant.

The statistics related to residuals analysis such as e_i, r_i and r^*_i, are used for determining problems which are encountered in data set and model[1,3,4,8].

As examining Table 4, Observation 2 and 16 are outliers with respect to the statistics. It is obviously seen that only two observations of all observations are exceeding the cut off values with ±2.

Although these two observations had unfavorable effects on the assumption mentioned above, it was not correct to remove them from data set[1,3,4].

With respect to h_{ii} statistics, only the 2nd observation can potentially affect the regression analysis in point of X value. As examined Cook's D and DFFITS, it is said that only observations 2nd and 16th on the results related to all regression coefficients can be effective. The Cook's and DFFITS had similar results which were in consistent with those reported by other authors[3,4].

Table 1: Descriptive statistics of live weights at different periods

Variable	N	Mean	SD	SE	Min.	Max.	CV (%)
Birth weight	18	2.46	0.33	0.078	1.90	3.00	13.45
Live weight in 30th day	18	5.13	1.22	0.288	3.00	6.98	23.82
Live weight in 45th day	18	6.49	1.61	0.379	3.70	9.78	24.76
Live weight in 60 th day	18	8.16	1.52	0.358	4.80	10.40	18.60
Live weight in 75 th day	18	10.38	1.85	0.436	6.74	12.80	17.81

Table 2: Correlation between all pair of variables

	X_1	X_2	X_3	X_4
X_2	0.151			
X_3	0.337	0.922**		
X_4	0.343	0.848**	0.906**	
Y	0.449	0.841**	0.910**	0.938**

* : $p < 0.05$, **: $p < 0.01$

Table 3: Results of regression analysis related to estimation of parameters

Variable	DF	Estimation of parameters	SE	value	p-values
Intercept	1	-0.229	1.368	-0.167	0.8699
Birth Weight	1	0.899	0.527	1.706	0.1117
Live Weight in 30th day	1	0.215	0.346	0.622	0.5449
Live Weight in 45th day	1	0.214	0.317	0.674	0.5123
Live Weight in 60th day	1	0.724	0.231	3.131	0.0080**

Model R^2 value : 0.9186 Model (%CV) : 5.81

Table 4: Results of residuals analysis concerning each observation

Observation	Yi	$\hat{Y}i$	e_i (Residual)	r_i	r_i^*
1	9.5600	9.9202	-0.3602	-0.704	-0.6896
2	11.8400	12.7509	-0.9109	-2.549*	-3.4640*
3	12.8000	12.2645	0.5355	1.156	1.1721
4	9.0200	8.8113	0.2087	0.399	0.3855
5	10.2600	10.2529	0.00715	0.013	0.0123
6	12.6400	12.4179	0.2221	0.420	0.4065
7	12.5600	12.2105	0.3495	0.657	0.6422
8	11.7400	11.7926	-0.0526	-0.102	-0.0982
9	8.4800	8.7608	-0.2808	-0.534	-0.5188
10	11.6600	11.7189	-0.0589	-0.110	-0.1053
11	11.6200	11.8292	-0.2092	-0.394	-0.3811
12	11.1600	10.9984	0.1616	0.297	0.2859
13	9.8600	9.7708	0.0892	0.162	0.1555
14	6.7400	6.5681	0.1719	0.365	0.3525
15	10.6600	10.4246	0.2354	0.445	0.4304
16	7.1600	8.6083	-1.4483	-3.252*	-7.2407*
17	8.3000	7.7471	0.5529	1.044	1.0476
18	10.8000	10.0132	0.7868	1.468	1.5443

Table 5: The cut off formulas and their values of influence statistic

Influence statistics	The cut off formulas of influence statistics	The cut off value
h_{ii}	$2p'/n$	0.555
Cook's D	$4/n$	0.222
DFFITS	$2\sqrt{\frac{P'}{n}}$	0.248
DFBETAS	$2/\sqrt{n}$	0.471
COVRATIO	$1\pm 3p'/n$	<0.1666 or >1.833

Table 6: Values of potential effective observation in point of influence statistics

					DFBETAS				
Gözlem	Cook's D	h_{ii}	Covratio	DFFITS	A	X1	X2	X3	X4
1	0.039	0.2806	1.7078	-0.4307	0.0238	-0.0878	-0.3599	0.2025	0.1464
2	2.406	0.6493	0.1330	-4.7130	-0.7725	0.6091	1.9862	-4.0001	2.5353
3	0.186	0.4100	1.4710	0.9771	0.1257	-0.6261	-0.1822	-0.0780	0.6236
4	0.010	0.2472	1.8638	0.2209	0.1558	-0.1303	0.0431	-0.0006	-0.0553
5	0.000	0.1485	1.7523	0.0052	0.0021	-0.0029	0.0012	-0.0010	0.0008
6	0.011	0.2328	1.8163	0.2239	-0.1051	0.0676	0.1436	-0.0598	-0.0320
7	0.025	0.2228	1.6214	0.3438	0.0592	-0.2052	-0.1160	0.0994	0.1119
8	0.001	0.2723	2.0424	-0.0600	0.0282	-0.0183	-0.0491	0.0369	-0.0041
9	0.018	0.2403	1.7581	-0.2918	-0.0412	-0.1098	-0.1407	0.0226	0.2026
10	0.001	0.2060	1.8706	-0.0536	0.0329	-0.0305	0.0075	0.0063	-0.0189
11	0.009	0.2266	1.8168	-0.2063	0.1543	-0.1648	-0.0508	0.0503	-0.0133
12	0.004	0.1840	1.7676	0.1358	0.0212	-0.0315	-0.1065	0.0689	0.0370
13	0.001	0.1649	1.7689	0.0691	0.0278	-0.0153	-0.0539	0.0369	0.0039
14	0.017	0.3903	2.3246	0.2820	0.2162	-0.0624	-0.0525	0.0875	-0.1521
15	0.012	0.2293	1.7934	0.2348	-0.1197	0.1712	-0.0014	-0.0540	0.0380
16	1.768	0.4552	0.0006	-6.6188	-1.0086	1.5116	1.1138	3.0988	-4.9146
17	0.065	0.2289	1.2494	0.5708	0.4216	-0.1443	-0.2339	0.2182	-0.2034
18	0.115	0.2108	0.7639	0.7982	-0.3707	0.5676	0.0818	-0.2940	0.1533

The cut off formulas and their values concerning the statistics are presented in Table 5. Based on the cut off values of Table 5, the values of potential effective observations in point of influence statistics are given in Table 6.

As to COVRATIO statistics, six observations (2, 4, 8, 10, 14 and 16) on fitted or predicted values were $\hat{Y}_i$ potential effective.

According to DFBETAS statistics, 2nd and 6th observations which were potentially effective influenced on intercept and all regression coefficients.

Durbin-Watson value for the data set was 2.31 which means that auto-correlation among residuals was not exist.

As a result, if points of observations with large leverage (outlier) are influential or potential, the observations (observation 2 and 16) should be carefully examined by researcher[4].

CONCLUSIONS

The aim of MRA is to determine the best set of independent variables most efficiently explaining variation of dependent variable, which is based on realizing the assumptions of MRA mentioned in introduction section. Diagnostics are given an idea about whether the basic assumptions will be provided or whether results of MRA will be reliable.

The most important results from this study can be summarized as;

Plot of residuals e_i versus fitted values gives an idea about whether assumptions of normal distribution and homogeneity of error terms will be supplied. In other words, the value of each residual should be in the interval of ±2 for ensuring the assumptions. Otherwise, ideal transformation as to scatter form of residuals e_i versus fitted values should be performed to dependent variable Y.

Being serial correlation among residuals means that residuals is not independent from each other. To provide this, the optimum cut off value of Durbin Watson statistics should be 2 to 4.

Consequently, it could be suggested that it would be useful to employ diagnostics in addition to MRA to make it more reliable due to discussed reasons above.

REFERENCES

1. Yazici, A.C., 1998. Analysis of diagnostics in multiple regression. M.SC Thesis, Ankara University, Institute of Natural and Applied Science, Ankara.
2. Johnson, R.A. and D.W. Wichern, 2002. Applied Multivariate Statistical Analysis. Prentice-Hall, Inc. Upper Saddle River, NJ 07458, pp: 354-383.
3. Yazici, A.C. and F. Gürbüz, 2002. Analysis of diagnostics in multiple regression. The 3rd National Animal Science Congress. Ankara University, Agriculture Faculty, Ankara, Turkey, pp: 361-370.
4. Chatterjee, P. and B. Price, 1991. Regression Analysis By Example. 2nd Ed., John Wiley and Sons, Inc., pp: 59-172.
5. SAS, 1998. SAS institute Inc. Cary,NC, USA.
6. http://www.data.princeton
7. Ergün, M.,1995. Statictics Applications with Computers for Scientific Researchs: SPSS for Windows, pp: 124-163.
8. Tabacnick, B.G. and L. S. Fidell, 2001. Using Multivariate Statistics. Allyn and Bacon, pp:111-218.