Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 48 (2012) 9202-9205

ISSN: 2229-712X

Assessing the fit of regression models using residuals in a multiplpe regression diagnotics

Osemeke Reuben F[1] and Desmond Ekokotu[2]
[1]Department of Mathematics, College of Education, Agbor.
[2]Department of Mathematics, Delta State, Nigeria.

ARTICLE INFO

Article history:
Received: 6 May 2012;
Received in revised form:
15 June 2012;
Accepted: 29 June 2012;

Keywords
Error validation,
Stepwise regression,
Dynamic scatter plots,
Regression diagnostics.

ABSTRACT

A dynamic graphical display and regression diagnostics is proposed for examining the validating assumption of the error term in a multiple regression modeling. Residual plots were used to detect the regression assumption of homocesdasticity tendency. Independence of the error term was diagnosed through Durbin Watson Test statistic or scatter plots of residuals against time order. Letter value displays were used to detect approximate normality through mid-summaries values. The error term validation were characterized with the residuals following an even distribution of scatter plots along a horizontal line through zero point, high (r^2), minimal standard error of estimates for the predictors, the Cp statistic close to P+1 which means a small biased, the P value less than 0.05(level of significant), and lack of bivariate high correlations among the predictors

Introduction

In multiple regression modeling, the behavioral pattern of the set of independent variables (X_i, i = 1, 2……P) offers an opportunity for validity of numerical and graphical diagnostics for judging the adequacy of a regression model. See Chatterjee and Price (1991) or Fox (1991).The examination and plotting of residuals to detect adequacies in a fitted multiple regression model is a useful and recommended practice (e.g., Daniel and Wood 1971, Weisberg 1980). Residuals (Y_i-Y^) are deviations from regression line and are generated from a fitted model of the linear trend Y^i = $b_o+b_1X_1+……. +b_nX_n$. Residuals are used to check model accuracy of regression assumption and other diagnostics checks. Before, we begin to evaluate the residuals to check for model accuracy and validate regression assumption, we must first of all, determine the amount of collinearity among the set of independent variable (X_i, i = 1, 2……P) which is evaluated through the use of variance inflationary factor (VIF) or through correlation matrix, notifying bivariate variable with high correlation (generally those of 0.90 and above) which is the first indication of substantial collinearity. Unfortunately, there are no well defined critical value for which is needed to have a large VIF. Some authors such as Chatterjee and Price (1991) suggest ten (10) as being large enough to indicate a problem. David Levine, Mark Berenson and David Stephan (1998) suggest five (5) as a rule of thumb. This implies that any set of independent variables that is greater than 5 should be deleted from the model but less than 5 should be retain for further analysis. Subsequently, a stepwise regression analysis has to be carried out to ascertain the best subsets approach to model building. The best independent variables are selected through Cp≤P+1. The residuals are generated from the fitted model and used to evaluate the aptness of the fitted model and check for goodness of fit tests.

The regression assumption shows that the forecast, confidence intervals and economic insights yielded by a regression model may be (at best) efficient and seriously unbiased. Cook and Weisberg (1982) proposed both the graphical method and the regression diagnostics for an effective assessment of residuals. In general, the scatter plots are the starting points for checking the model assumption of the regression analysis

Steps Involved in Model Building

Step 1: Choose a set of predictors to be considered for inclusion in the regression model

Step 2: Fit a full regression model that includes all the predictors to be considered so that the (VIF) for each predictor can be determined

Step 3: Determine whether any predictors have a VIF > 5

There are three possible results that can occur

(a) None of the predictors have a VIF > 5.If this is the case ,proceed to step 5

(b) One of the predictor has a VIF > 5.If this is the case, eliminate that predictor and proceed to step 5

(c) More than one of the predictors has a VIF > 5.If this is the case, eliminate the covariates that has the highest VIF and go back to step 2

Step 4: Perform a best-subsets regression with the remaining predictors to obtain the best models (in terms of Cp) for a given number of predictors

Step 5: List all models that have Cp statistic ≤ (P+1)

Step 6: Among those models listed in step 5, choose a best model for prediction

Step 7: Perform a complete analysis of the model chosen including residual analysis in determining regression assumption

Model Building

Example 1: Developing a model building process that considers standby hours(y) based on influence of total staff present(x1), remote hours(x2), dubner hours(x3), and total labor hours(x4). We begin our analysis by first measuring the amount of collinearity that exists among the set of predictors

Tele:
E-mail addresses: reubitosmek@yahoo.com

From Table 1, we observe that none of the predictors have VIF > 5. We accept all the predictors and perform a best subset regression to obtain the best models. The VIF values are relatively small, ranging from 2.0 for total staff hours to a low value of 1.2 for remote hours. There is little evidence of collinearity among the set of predictors

The Stepwise Regression Approach to Model Building

We now continue our analysis of these data by attempting to determine the best subset of all covariates that yield an adequate and appropriate model without having to use the complete model. We begin by describing a widely used search procedure called stepwise regression, which attempts to find the best regression model without examining all possible regression. Once a best model has been found, residuals are used to evaluate the fitness of the model. An important feature of this stepwise regression is that a predictor that has entered into the model at an early stage may subsequently be removed once other predictors are considered. That is, in stepwise regression, variables are either added or deleted from the regression model at each step of the model building process. The stepwise procedure terminates with the selection of a best fitting model when no additional variables can be added to or deleted from the last model fitted

Here we look for values of P+1 and Cp where Cp ≤ (P+1)

Mallows (1973) suggested that the best criterion often used in the evaluating of competing models is based on the Cp statistic. In choosing a model, we look for models with a Cp≤P+1 which means a small biased. The Cp statistic is defined as

Cp = $(1-R^2p)\ (n-T)/1-R^2_T-[1-2(p+1)]$ fig 1

Where

P = Number of independent variables included in the regression model

T = Total number of parameters (including the intercept) to be estimated in the full regression model

R^2p = Coefficient of multiple determination for a regression model that has P independent variables

$R^2_{T\,=}$ Coefficient of multiple determination for a full regression model that contains all T estimated parameters

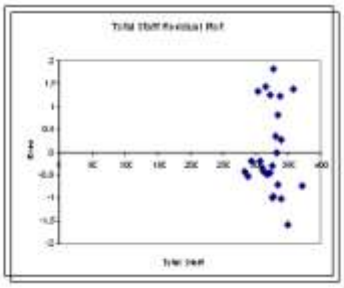
Using the above equation in fig 1 to compute Cp statistic for the Total Staff Present and Remote hours, we have n= 26, P = 2, T = 4+1= 5, $R^2p = 0.490$, $R^2_T = 0.623$

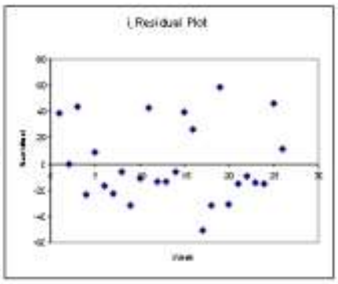
Cp = (1-0.49)(26-5)/1-0.623-[26-2(2+1)]

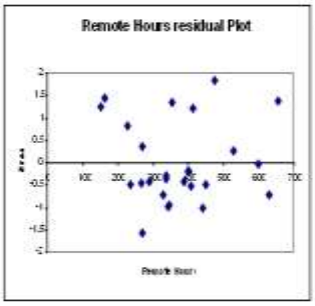
Cp = 8.42

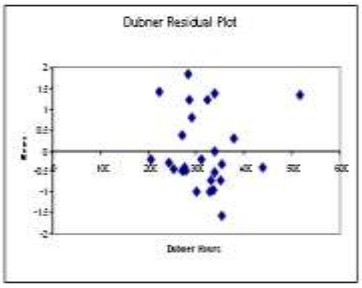
From Table 2, we observed that only the model with all four (4) independent variables considered contain a Cp≤P+1 which means a small biased and are choosing for prediction. For other illustration and comments on interpretation, see Gorman and Toman (1996), Mallows (1973) or Daniel and Wood (1980).Now that the explanatory variables to be included in the model have been selected, a residual analysis should be undertaken to evaluate the aptness of the fitted model

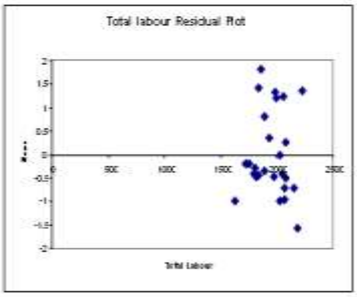
5: Dynamic scatter plots to show error validation

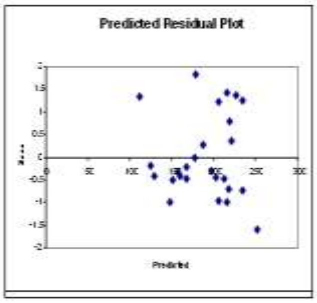












The scatter plots of standardized residuals versus total staff, the remote hours, the dubner hours , total labour hours and the i residual plots, all reveal no apparent pattern or relationship between the residuals. The residuals appear to be evenly spread above and below zero (o) for different values of Xi. The residuals are constant across all range of predictors (homoscesdasticity).This shows that the coefficient estimates are unbiased, the standard error of the estimate are correct and the statistical inference is valid . The i residuals plots shows independence of the error term.

Assessing Normality Assumption using Letter Value Displays

The basis for this is a numerical summary display called the letter value display. Letter values are similar to percentiles of the data and are defined by their depth. The Median, the hinges, the eights, and the sixteenths are the start of the sequence of the letter values. They are defined as follows

Depth of the median:	d(M) = (n+1)/2.
Depth of hinges:	d(H) = ([d(M)] + 1)/2
Depth of eights:	d(E) = ([d (H)] +1)/2
Depth of sixteenths:	d(D) = ([d (E)] +1)/2

The remaining depths are found by continuing the pattern. They are labeled C, B, A, Z, Y, X.

To find the letter values, first order the data (the standardized residuals). The lower hinge is the observation at a distance d(H) from smallest observations, the upper hinge is the observation at a distance d(H) from the largest observations. Similarly, the lower and upper eights are the observations at a depth d(E) and so on. The midpoint for a given depth is the average of the upper and lower letter values at the depth. The spread is (upper –lower).

Normality of this letter values are ensure when the midsummaries values for the residuals are the same across all values,

From table 3, we notice that there is an upward trend in the mid-point values, but the trend are approximate the same and this suggests that the errors are normally distributed

Testing for independence of the error term

Our Durbin Watson test statistic = 2.220. Finding critical values of Durbin Watson statistic with α = 0.05 level of significance

Thus, for our data, with four predictors,(P=4) and 26 observations (n=26), lower critical value(dl) = 1.06 and upper critical value(du) = 1.76. Because Durbin Watson (D) = 2.220> 1.06 and 1.76, we conclude that there is no evidence of autocorrelation among the residuals. The i residuals seems independence and the fitted model Y = -0.330.83+1.2456X1i-0.1184X2i-0.2917x3i+0.1305X4i seems appropriate because of the presence of independence among the residuals.

Regression Statistics

The simplest and most obvious means of identifying collinearity is an examination of the correlation matrix for the independent variables. The presence of high correlations (generally those of 0.90 and above) is the first indication of substantial collinearity. From Table 5, it is very obvious that the pairwise correlations are very small which shows that there is absent of bivariate collinearity among the set of predictors

The multiple r = 0.789.R^2 is 0.623.Adjusted R^2 is 0.551 and standard error for each predictor is very minimal. The coefficient of multiple determination computed as 0.623 means that 62.3% of the variation in standby hours(y) can be explained by the variation in total staff present, remote hours, dubner hours and total labor hours. The sample fitted model Y = -0.330.83+1.2456X1i-0.1184X2i-0.2917x3i+0.1305X4i is good for prediction. The P value in Table 6 for each covariate is less than 0.05 which is an indication that the predictors are statistically significant; hence, the model is good for prediction

Conclusion

The potential validation of regression assumption is examined through a thorough examination of diagnostics regression measures, dynamic scatter plots display, $Cp \leq P+1$ which means a small biased and test statistic in a multiple regression analysis. Using these techniques, the researcher is able to have an insight in the validation of regression assumption. The value of employing these techniques are well documented in books by Belsley, Kuh and Welsch (1980), Gunst and Mason (1980), Cook and Weisberg (1982) and Montgomery and Peak (1982).

In all these, analysis of residuals after the validation were characterized with even distribution of scatter plots, improve in r^2, significance and statistical relationship among the independent variables, minimal standard error estimate and independence of the error term through residual plots against time or through Durbin Watson Test. Approximate normality were identified through midsummaries values in letter value display.

We therefore regard residual analysis as an indispensable tool of regression analysis. It facilitates the job of the analysis

Recommendation

In dealing with the use of residuals in detecting the validations of the regression assumption, the writer suggests that checking these assumptions carry significant benefits for the researchers, making sure an analysis meets the associated assumptions helps avoid Type 1 and Type п errors. Attending to issues as adequacy due to homocesdasticity, normality term, independence of the error term often boost effect sizes, usually a desirable outcome. The writer suggests other methods of residual validations. The probability plots, histogram stem and leaf display, box ands whisker plot and Goodness of fit test can be use to validate error term of normality. The non-parametric test can be used to detect error validations. Other diagnostics measures are F statistics, T test, P value and confidence interval.

The writer also suggest that apart from using stepwising to select the best regression model, variance inflation factor and correlation matrix in choosing the best regression models, other methods that can be used to diagnose predictors that are statistically significant are tolerance value, factorial exploration, the residual mean square, the multiple correlation coefficient, the adjusted multiple correlation coefficient, the press, turkey's rule and freehand method. The writer suggests also that areas of residual violations of regression assumption and transformation should be look into

References

Belsley, D.A. Kuh, E, and Welsch, R.E. (1980), "Regression Diagnostics", New York: John Wiley.

Belsley, D.A. (1991), "Conditioning Diagnostics", New York: John Wiley

Cook, D.R.D., and Weisberg. S. (1982), "Residual and Influence in Regression", New York: Chapman and Hall.

Chatterjee and Price. (1991) or Fox. (1991), "Useful and recommended practice", (e.g., Daniel and Wood 1971, Weisberg 1980).

Daniel and Wood, F.S. (1980), "Fitting Equations to Data": New York: Wiley,

Daniel, C.W, and F. S, "Wood. (1980), "Fitting Equations to Data", 2d Ed": New York: Wiley.

Draper, N.R, and Smith, H. (1996), "Applied regression Analysis": Wiley and Sons, Inc 430 pp.

Gorman, J.W. and Torman, R.J (1966), "Selection of variables for fitting equations to data": Technometrices, 8, 27-5

Gunt, R.F, and Mason, R.L. (1980), "Regression Analysis and its Application": New York: Marcel Dekker

Mallows, C. L. (1973), "Some comments on Cp", Technometrics, 15,661-675

Montgomery, D.C., and Peck, E.A. (1982), "Introduction to linear Regression Analysis", New York: John Wiley.

Chatterjee, S. and Price, B. (1991), Regression Diagnostics: New York: John Wiley

Fox, J.(1991),Regression Diagnostics, Newbury Park, CA :Sage

Gujarati, D.N.(1988),Basic Econometrices, New York : McGraw-Hill

Kutner, Nachtsheim, Neter, Applied Linear Regression Models, 4th edition, McGraw-Hill, Irwin, 2004

Marquardt, D.W. (1970)" Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Nonlinear estimation", Techno metrics 12/2) 591,605-07

O" Brien, Robert M.2007. " A Caution Regarding Rules of Thumb for Variance Inflation Factors" Quantity and Quality

Table 1:

Regression Models	Collinearity Statistics(VIF)
Total staff present model for (X_1) and other X	1.707
Remote hours model for (X_2) and all other X	1.233
Dubner hours model for (X_3) and other X	1.459
Total Labour Hours. Model (X_4) and all other X	1.999

Table 2: Best Subsets Approach to Model Building

Models	$C_{p\ \text{statistic}}$	P+1	R^2	Adj.R^2	Std Error	Consider This Model for prediction
X_1	13.32152	2	0.367	0.34	38.62	No
X_1X_2	8.419	3	0.490	0.445	35.4	No
$X_1X_2X_3$	7.8418	4	0.536	0.473	34.50	No
$X_1X_2X_3X_4$	5	5	0.623	0.551	31.84	Yes
$X_1X_2X_4$	9.345	4	0.509	0.442	35.49	No
X_1X_3	10.65	3	0.45	0.402	36.74	No
$X_1X_3X_4$	7.712	4	0.538	0.475	34.44	No
X_1X_4	14.798	3	0.375	0.321	39.16	No
X_2	33.208	2	0.009	-0.032	48.28	No
X_2X_3	32.31	3	0.061	-0.02	48.01	No
$X_2X_3X_4$	12.14	4	0.459	0.385	37.26	No
X_2X_4	23.25	3	0.224	0.1560	43.65	No
X_3	30.39	2	0.059	0.021	47.03	No
X_3X_4	11.82	3	0.43	0.38	37.45	No
X_4	24.2	2	0.171	0.14	44.16	No

Table 3

N = 26	Lower	Upper	Mid	Spread
M = 13.5	-0.3600	-0.3600	-0.3600	0.000
H = 7.25	-0.58	1.085	0.2525	1.665
E = 4.125	-0.925	1.19	0.1325	2.115
D = 2.57	-1.135	1.255	0.1	2.39
1	-1.33	1.60	0.135	2.93

Table 4

$\alpha = 0.05$								
	P = 1		P = 2		P = 3		P = 4	
N	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76

Table 5: Pairwise Correlation Matrix: An alternative method of Table 1

	X_1	X_2	X_3	X_4
X_1	1.000	.390	0.036	.571
X_2	.390	1.000	0.022	.241
X_3	0.036	0.022	1.000	.395
X_4	.571	.246	0.395	1.000

Table 6

	Coefficients	Std Error	P value
Intercept	-330.83	110.895	0.007
Total Staff	1.2456	0.412	0.006
Remote	-0.1184	0.054	0.04
Dubner	-0.297	0.118	0.0199
Total Labour	0.1305	0.059	0.039