REVIEW OF RESEARCH

IMPACT FACTOR : 5.7631(UIF)   UGC APPROVED JOURNAL NO. 48514   ISSN: 2249-894X

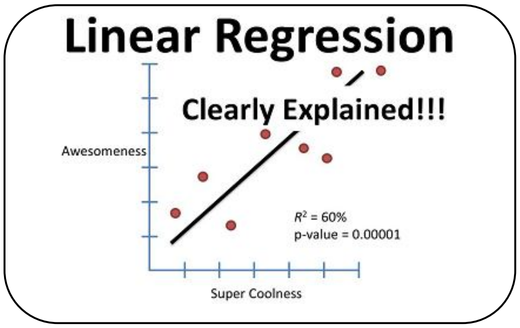
VOLUME - 8 | ISSUE - 5 | FEBRUARY - 2019

SINGLE OUTLIER TEST USING LINEAR REGRESSION MODEL ON THE UPPER BOUNDS OF TEST STATISTICS

Dr. Ravindra S. Acharya
Professor in Mathematics,
Vishwakarma Institute of Information Technology, Kondhava, Pune.

ABSTRACT :

An astonishing number of test statistics have been found to test the presence of an outsider in multiple linear regression models. Accurate critical values of these test statistics are not available and are usually obtained by first-order Bonferoni upper bound or large-scale simulation,used to demonstrate the application process of real data for multiple linear regressions.



KEYWORDS : *test statistics , Accurate critical, multiple linear regressions.*

INTRODUCTION :

The exterior is an inconsistent observation. This is an observation that does not fit into the rest of the observation pattern. This is clearly different from the other members not only from the set from which it arises, but also from its fitted value. Such observations usually have a large residue. Outliers meet data analysts during data analysis and in data mining. Domankshi P.D. It was pointed out that there are various reasons for the exit such as human error, incorrect operation of the computer system, sampling error or standardization failure. Outliers usually have a large influence on the estimation of the resulting parameters and their presence adversely affects the results of the statistical findings related to the models. They can reduce the power of statistical tests during analysis. Rajaratinmana and Vinoth advised that if the analyst's outline exists, they need to be identified so that appropriate measures can be taken.

Outsiders need to be identified and corrected or eliminated. The process of identifying and correcting outsiders is not a straightforward one; instead it requires strict adherence to clear competence, competence, caution, and a high degree of scientific objectivity (objectivity). If identified external measures cannot be taken, they must be removed as they contaminate information in the rest of the data set. Testing for external observations in response variables is usually based on the use of test statistics that rely on standardized residues. In the minimally square analysis of the linear regression model, various test statistics have been developed for external testing. However, accurate critical values of some of these test statistics are not available and are not easy to obtain. The approximate ones available are based on first-order Bonferoni upper bound or large-scale simulations. A high limit for critical values of test data to detect the presence of a single extraterrestrial in linear regression has been developed by Prescott P and Lund RE. While drawing these upper boundaries, we show in this paper that these upper boundaries drawn by Prescott P and Lund REare algebraically similar.

The Multiplier Linear Regression is as follows…

$$Y = X\beta + \varepsilon$$

Equation – 1

Where Y is the n X 1 observation vector, X is an n X 1 matrix of constants, β is p X 1 vector of unknown parameter to be estimated and ε is an n X 1 vector of normally distributed errors. Assuming that E(ε) = 0 and Var (ε) = $\sigma^2$1 the least square estimator is β in equation – 1 given as,

$$\hat{\beta} = (X'X)^{-1}X', Y$$

Equation – 2

And the vector of residual is

$$e = Y - X\hat{\beta} = (I - X(X'X)^{-1}X')\varepsilon$$

Equation – 3

The variance of co-variance of matrix e is as follows,

$$var(e) = Y - X\hat{\beta} = (I - X(X'X)^{-1}X')\varepsilon$$

Equation – 4

If σ^2estimated using $\sigma \wedge^2 = e'e(n-p)$, the approximate variation-covariance matrix of **e**is formed

$$\widehat{var}(e) = Y - X\hat{\beta} = (I - X(X'X)^{-1}X')\sigma \wedge^2$$

Residues are important diagnostic tools in regression analysis because no regression analysis is complete without their thorough investigation. They are versatile because most regression diagnoses are written from their point of view. They are used to check the adequacy of the model and the validity of the model assumptions. Therefore a thorough examination of the residue provides valuable information regarding the suitability of the underlying assumptions of the statistical models and helps to determine the appropriate model. A variety of graphic plots (statements) of remains are used for diagnostic purposes. Common residues are not suitable for diagnostic purposes and a standardized version of them is usually preferred. This is because the differences in the residues are not homogeneous and this causes them to deviate. The standard residue represents the form is as follows…

$$R_i = \frac{y_i - \hat{y}_i}{\sigma \wedge^2 \sqrt{1 - h_{ii}}}$$

Equation – 6

Where$\hat{y}_i$ is the approximate value of y_i and $h_{ii}$is the ith component of matrix $X(X'X)^{-1}X'$ of, called the hat matrix. The *ith*converted residualR_i is usually called internal pupil residue. They are tractable and more versatile. They are used as a replacement for common residues in regression diagnostics. Numerous graphical and numerical techniques for examining model assumptions using standardized residues can be found in the regression literature. They are also basic building blocks for known test data studied in the literature for external detection in linear models.

The Statistic Tests:

$$R_n = max \left| \frac{y_i - \hat{y}_i}{\sigma \wedge^2 \sqrt{1 - h_{ii}}} \right| = \max|R_i|$$

Equation – 7

Considered the statistics test by Prescott P. is as follows…

$$R_n^* = max\left|\frac{e_i}{\bar{\sigma}_{e_i}}\right|$$

Equation – 8

Where $\bar{\sigma}_{e_i}$ is the approximate average difference of normal residues. Studies by Andrews and Pregibon have shown that the difference in residues is $(n-p)\,\sigma^2/n$, so that the approximate difference in common residues is $\bar{\sigma}_{e_i} = (n-p)\,\sigma\,\Lambda^2/n$.

Therefore:

$$R_n^* = \frac{n^{1/2}max|e_i|}{\left(\sum e_i^2\right)^{1/2}}$$

Equation – 9

According to Prescott, the relative percentage of $R_0^*$is limited to R_n^* is bounded above by,

$$U = \sqrt{\frac{(n-p)F}{n-p-1+F}}$$

Equation – 10

Where F is the $100(1-\alpha/n)$ percentage point of the F distribution with degrees of freedom 1 and n-p-1, n is the number of observations, and p is the approximate number parameter. Due to the unavailability of the required values of the f-distribution, the Prescott's result was not as wide and wide as a result of the equation.

$$\xi_1 = \frac{R_i}{\sqrt{n-p}}$$

Equation – 11

Ellenberg showed that the combined distribution of $\xi_i' s$has a multivariate inverse-student function, and the probability density function for any ξ_i is a student function with an irreversible inverse-potential density function.

$$f(\xi_i) = C\left(1-\xi_i^2\right)^{(n-p-3)/2}, \qquad \xi_i^2 \le 1$$

Equation – 12

Where,

$$C\frac{\Gamma((n-p)/2)}{\Gamma(1/2)\Gamma((n-p)/2)}$$

Equation – 13

By the reference of Lund's following suggestion of Prescott made us the result of following equation – 14. With the use of Prescott's the first-order Boneferroni inequality is obtained the upper bounds R$_0$ of the critical value of R$_n$.

$$\int_{\xi_0}^{1} 2nf(\xi_i)d\xi_i = \alpha$$

Equation – 14

Where,

$\xi_0 = R_0/\sqrt{n-p}$and then obtained R_0 using the relationship between R_0 and ξ_0 given by the equation

$$R_0 = \xi_0\sqrt{n-p}$$

Equation – 15

Demonstration of similarity of upper boundaries:

In this section, we show that the upper boundaries R ∗ 0 and R0 are algebraically identical. From the equation - 10, we do

$$U = \sqrt{\frac{(n-p)F}{n-p-1+F}}$$

Equation - 16

$$P_r(U < u) = Pr\left(\sqrt{\frac{(n-p)F}{n-p-1+F}} < u\right) = Pr\left(f < \frac{u^2(n-p-1)}{n-p-u^2}\right)$$

Equation - 17

So that,

$$f_U(u) = f_F\left[\left(\frac{u^2(n-p-1)}{n-p-u^2}\right), 1, n-p-1\right]\frac{n(n-p-1)(n-p)u}{(p+u^2-n)^2}$$

Equation – 18

With the distribution of the domain given by $\left(0, \sqrt{(n-p)}\right)$ with explicitly we have,

$$f_U(u) = H\left(1 - \frac{u^2}{n-p}\right)^{(n-p-3)/2} \qquad 0 < u < \sqrt{(n-p)}$$

Equation – 19

Where,

$$H = \frac{2\Gamma((n-p)/2)}{\Gamma(1/2)\sqrt{n-p}\Gamma((n-p-1)/2)}$$

Equation – 20

Then, using the first Bonferoni inequality, one can get the upper limit R_n^* by solving

$$\int_{R_0^*}^{\sqrt{(n-p)}} n f_U(u)du = \alpha$$

Equation – 21

Now from the equation 11 we have,

$$Pr(R_i < r) = Pr\left(\xi_i\sqrt{n-p} < r\right) = Pr\left(\xi_i < \frac{r}{\sqrt{(n-p)}}\right)$$

Equation – 22

So that,

$$f_{R_i}(r) = f_\xi\left[\left(\frac{r}{\sqrt{(n-p)}}\right)\right]\frac{1}{\sqrt{n-p}}$$

Equation - 23

With the distribution domain or with a given range by $\left(-\sqrt{n-p,}\sqrt{n-p,}\right)$ explicitly we have,

$$f_{R_i}(r) = D\left(1-\frac{r^2}{n-p}\right)^{(n-p-3)/2} \qquad -\sqrt{n-p,} < r < \sqrt{n-p}$$

Equation – 24

Where,

$$D\frac{\Gamma((n-p)/2)}{\Gamma(1/2)\sqrt{n-p}\Gamma((n-p-1)/2)}$$

Equation – 25

Let,

$$Y_i = |R_i|$$

Equation – 26

$$Pr(Y_i < y) = Pr(|R_i| < y) = Pr(-y < R_i < y)$$

Due to the symmetry of the distribution of R_i in Equation - 24, we get the distribution of $Y_i = |R_i|$ as follows:

$$Pr(Y_i < y) = 2\,\mathrm{Pr}(R_i < y)$$

Equation – 27

Explicitly we have the following,

$$fY_i = H\left(1-\frac{y^2}{n-p}\right)^{(n-p-3)/2} \qquad 0 < y < \sqrt{n-p}$$

Equation – 28

Then, using the first Bonferoni inequality, a person can get R0 by solving

$$\int_{R_0^*}^{\sqrt{(n-p)}} nf_y(y)dy = \alpha$$

Equation – 29

With the equality of equation – 21 and equation – 29, we have

$$\int_{R_0^*}^{\sqrt{(n-p)}} nf_U(u)du = \alpha \Rightarrow \int_{R_0^*}^{\sqrt{(n-p)}} nf_y(y)dy = \alpha$$

Equation – 30

This means that R_0= R_0^*, which means that Rn and R_n^*have constructs bound by the same distribution.

CONCLUSION:

In this article, we have shown that the upper binding value of the test figure by Equation 7 is R0 and the upper bound of the test figure by Equation 8 is R ∗ 0. Although formal differences exist in the principles used by Prescott to draw R ∗ 0 and are employed by Lund to draw R0, we have shown here that they are algebraically similar. After showing this, we recommend using Equation - 29 to calculate the upper boundaries of Prescott or Lund. This is more tractable than Equation-10 and Equation-14. Some sort of transformation and limitation of the equation - 10 uses tabulated values of F-distribution, accuracy and precision may be lost while using them.

REFERENCES:

1. Andrews D.F. and PregibonD., “Finding the outliers that matter,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 40, no. 1, pp. 85–93, 1978
2. Barnett V. and T. Lewis, Outlier in Statistical Datas, Wiley and Son, Chichester, U.K., 1994.
3. Hawkins D.M., Identification of Outliers, Springer, Dordrecht, The Netherlands, 1980
4. Lund R.E., “Tables for an approximate test for outliers in linear models,” Technometrics, vol. 17, no. 4, pp. 473–476, 1975
5. Prescott P., “An approximate test for outliers in linear models,” Technometrics, vol. 17, no. 1, pp. 129–132, 1975.
6. RangaSuri N.N.R., MurtyM. N., and AthithanG., Outlier Detection: Techniques and Applications: A Data Mining Perspective, Springer, Cham, Germany, 2019.
7. Tietjen G.L., MooreR. H., and BeckmanR. J., “Testing for a single outlier in simple linear regression,” Technometrics, vol. 15, no. 4, pp. 717–721, 1973.
8. Ugah T.E., Ikechukwu E., Eze M.C., Arum K.C. Christy I. and Oranye H.E., “On the Upper Bounds of Test Statistics for a Single Outlier Test in Linear Regression Models”, Journal of Applied Mathematics, Vol-21, pp. 1-5.