Journal of Natural Sciences Research www.iiste.org
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.3, No.3, 2013

Nonparametric Measure of Linear Trend

NWANKWO, CHIKE H.* OYEKA, I. C. A.
Department of Statistics, Nnamdi Azikiwe University, PMB 5025, Awka. Anambra State, Nigeria
*E-mail of the corresponding author:chikeezeoke@yahoo.com

Abstract
This paper proposes and develops a nonparametric statistical procedure for estimating linear trend effect on data using nonparametric regression methods based on ranks, assuming one of the sampled populations is a measurement on a time scale. The populations of interest may be measurements on as low as the ordinal scale. The nonparametric regression-based linear trend estimate is shown to be similar in computation to the Spearman's Rank Correlation Coefficient. Test statistics are developed for testing hypotheses on the linear trend effect. Sample data are used to illustrate the proposed method.
Keywords: Nonparametric, Linear, Test Statistic, Measure, Trend, Ratio

I. Introduction

If two random samples of equal size are independently drawn from some continuous populations X and Y, and if one of the sampled populations X say, is a time sequence, that is, if the observations drawn from X are time ordered, then a simple linear regression of observations drawn from the other population Y on the observations from X may be used to determine whether there exists a linear trend in the values of observations from Y. If furthermore population Y can be reasonably assumed to be normally distributed, then the usual parametric 'Z' test or the 't' test may be used to test for the statistical significance of the linear trend (Neter et-al,1983). However if these assumptions can not be validly made, then the parametric methods may not be properly used. In these situations nonparametric methods such as the runs test, spearman's rank correlation coefficient or the Kendall's tau correlation coefficient (Gibbons,1971; Seigel, 1956) would then be the method of choice. We in this paper, propose to develop an alternative statistical method for estimating linear trend as a nonparametric regression problem based on ranks.

II. The Proposed Method

Suppose x_i is the *ith* observation in a random sample of size n drawn from populations X and y_i is the *ith* observation in a random sample also of size n independently drawn from population Y, for $i = 1,2 \ldots n$.

Populations X and Y may be measurements on as low as the ordinal scale and need not be continuous or assumed to follow any probability distribution form. For the purpose of this paper we assume that all the observation have been converted into ranks and that one of the populations X say, is a time sequence so that the observations x_i are time ordered. We may therefore here assume, without loss of generality, that $x_i = i$ for $i = 1,2, \ldots, n$ where n is indexed in time. Hence the rank of x_i is $r_{ix} = i$. Let r_{iy} be the rank assigned to y_i ranked from the smallest to the largest observation, the i^{th} observation drawn from population Y.

Now to present the estimation of linear trend as a nonparametric regression problem we here fit a simple linear regression model of the ranks r_{iy} of y_i drawn from population Y as the dependent variable regressing on the ranks $r_{ix} = i$ of x_i drawn from the time ordered population X as the independent variable, obtaining the regression model

$$r_{iy} = \propto + \beta i + e_i \quad . \quad . \quad . \quad (1)$$

Where $\propto and\ \beta$ are regression parameters, with β being the slope, and e_i is the error term, with $E(e_i) = 0, for\ i = 1, 2 \ldots\ldots$n

Since population X is a time sequence or time ordered, the slope or regression coefficient β is also interpreted as a measure of linear trend, that is the effect of the time variable X on population Y.

Applying the usual least squares method of estimation to equation 1, we obtain an unbiased estimate of the linear trend β as

$$\frac{\sum_{l=1}^{n} i\, r_{iy} - \sum_{l=1}^{n} i \sum_{l=1}^{n} \frac{r_{iy}}{n}}{\sum_{l=1}^{n} i^2 - \frac{\left(\sum_{t=1}^{n} i\right)^2}{n}} \quad . \quad . \quad . \quad (2)$$

Now since $r_{ix} = l \; and \; r_{iy}$ are each permutations of the first n positive integers, we have that

$$\sum_{l=1}^{n} r_{ix} = \sum_{l=1}^{n} r_{iy} = \frac{n(n+1)}{2} \quad . \quad . \quad . \quad (3)$$

And

$$\sum_{l=1}^{n} r_i^2 = \sum_{l=1}^{n} i^2 = \sum_{l=1}^{n} r_{iy}^2 = \frac{n(n+1)(2n+1)}{6} \quad . \quad . \quad . \quad (4)$$

Hence using Equations 3 and 4 in Equation 2 we have that the estimated simple linear regression coefficient or trend value is

$$b = \frac{\sum_{l=1}^{n} i r_{iy} - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} = \frac{12\left(\sum_{l=1}^{n} i r_{iy} - \frac{n(n+1)^2}{4}\right)}{n(n^2-1)} \quad . \quad . \quad . \quad (5)$$

Equation 5 provides an estimate, b, of the linear trend or time effect of the time ordered sampled population X on the sampled population Y.

Equation 5 however can be further simplified as follows.

Let d_i be the difference between the ranks assigned to x_i and $y_i \; that \; is, let$

$$d_i = i - r_{iy} \quad . \quad . \quad . \quad (6)$$

for $i = 1,2 \ldots n$

Then,

$$\sum_{l=1}^{n} i r_{iy} = \frac{n(n+1)(2n+1)}{6} - \frac{\sum_{l=1}^{n} d_i^2}{2} \quad . \quad . \quad . \quad (7)$$

Now using equation 7 in Equation 5, which yields, after simplification

$$b = 1 - \frac{6\sum_{l=1}^{n} d_i^2}{n(n^2-1)} \quad (8)$$

(Equation 8 is seen to be, infact, the same expression often used for the calculation of Spearman's Rank Correlation Coefficient between two ranked data (Siegel,1956[3]). Viewed as a regression coefficient, the estimated trend value $'b'$ could conceivably assume all possible values on the real line. However, in reality as a correlation coefficient its value would range only from -1 to +1 inclusively. The trend value is 0 if there is no association and assumes the values -1 and +1 respectively if there is perfect indirect and perfect direct association between $Y \; and$ the time axis. The F-test is used to test the null hypothesis that the linear trend model fits, that is, the null hypothesis that no trend exists in the data (Ho: $\beta = 0$). If Ho is rejected, in which case there exist some trend in the data, so that $\beta \neq 0$ then the t- test may be used to determine whether the trend value is consistent with some hypothesised value $\beta_0 \; say.$

The constant term or the so called "intercept" is $a = \hat{a} = \frac{(n+1)(1-b)}{2}$ . . (9)

Hence the fitted or predicted linear trend line is

$$\hat{r}_{iy} = \frac{n+1}{2} + b\left(i - \frac{n+1}{2}\right) = (n+1)(1-b) + bi \quad (10)$$

Note that an interesting feature of the non parametric approach to the estimation of linear trend is that for a given sample size or time point n, once an appropriate trend value 'b' is available, perhaps obtained from a related previous study, then equation 10 may be used to predict the rank and hence estimate the associate value of the time dependent variable y_i at a desired time 'i'. Then if there is no trend in Y the corresponding population value of b, namely β would be expected to be zero. But a more general null hypothesis, Ho that would need to be tested is

$$\text{Ho: } \beta = \beta_0 \; versus \; H_1: \beta \neq \beta_0, \quad . \quad . \quad . \quad (11)$$

Where β_0 may be any real number between $-1 \; and \; +1$ which also includes 0.

Two methods will be presented here for testing Ho, namely the Fisher's F- test using analysis of variance approach and the usual $'t'$ test.

Note that in terms of ranks, the regression sum of squares SSR for a simple linear regression model is

$$SSR = b\left(\sum_{l=1}^{n} ir_{iy} - \sum_{l=1}^{n} i \sum_{l=1}^{n} \frac{r_{iy}}{n}\right) = b\left(\frac{n(n^2-1)}{12} - \frac{\sum_{l=1}^{n} d_i^2}{2}\right) \quad . \quad . \quad . \quad (12)$$

With 1 degree of freedom,

Where d_i is given in equation 6.

Also the total sum of squares is

$$SST = \sum_{l=1}^{n} r_{iy}^2 - \frac{\left(\sum_{L=1}^{N} r_{iy}\right)^2}{n} = \frac{n(n^2-1)}{12} \quad . \quad . \quad . \quad (13)$$

With $n-1 \; degrees \; of \; fredom.$

Finally the sum of squares error SSE is

$$SSE \; = \; SST - \; SSR = \left(\sum_{l=1}^{n} r_{iy}^2 - \frac{\left(\sum_{l=1}^{n} r_{iy}\right)^2}{n}\right) - \mathrm{b}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{i}r_{iy} - \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{i} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{r_{iy}}}{\mathrm{n}}\right)$$

That is

$$SSE \; = \frac{n(n^2-1)(1-b)}{12} + b\sum_{i=1}^{n} \frac{d_i^2}{2} \quad . \quad . \quad . \quad (14)$$

Or equivalently

$$SSE \; = \frac{n(n^2-1)\left(1-b^2\right)}{12} \quad . \quad . \quad . \quad (15)$$

With $(n-1)-1 = n-2$ degrees of freedom, since

$$\sum_{l=1}^{n} ir_{iy} - \sum_{i=1}^{n} \frac{r_{iy}}{n} = b\left(\sum_{l=1}^{n} i^2 - \frac{\left(\sum_{l=1}^{n} i\right)^2}{n}\right) = \frac{bn(n^2-1)}{12} \quad . \quad . \quad . \quad (16)$$

So that an alternative expression for the regression sum of squares (equation 12) is

$$SSR \; = \frac{n(n^2-1)\, b^2}{12} \quad . \quad . \quad . \quad (17)$$

These results are summarized in the Analysis Of Variance (ANOVA) table (table 1) below

Table 1: ANOVA table for use in testing for linear trend

Source of Variation	**Sum of Squares (SS)**	**Degree of Freedom (df)**	**Means Squares (MS)**	**F- Ratio**
Trend (Regression)	$SSR = \frac{n(n^2-1)\,b^2}{12}$	1	$MSR = \frac{n(n^2-1)b^2}{12}$	$F = \frac{b^2}{\frac{(1-b^2)}{n-2}}$
Error	$SSE = \frac{n(n^2-1)(1-b^2)}{12}$	$n-2$	$MSR = \frac{n(n^2-1)(1-b^2)}{12(n-2)}$	
Total	$SST = \frac{n(n^2-1)}{12}$	$n-1$		

Notice that the proportion of the total variation in the ranks of y explained by the fitted trend line is

$$R^2 = \frac{SSR}{SST} = b^2 \quad . \quad . \quad . \quad (18)$$

This is the familiar coefficient of determination in the parametric analysis of variance parlance.

which results of table 1 may be used to test the null hypothesis of no trend. That is the null hypothesis that there is no trend in the values of population Y $(H_o: \beta = 0)$. The null hypothesis H_o is rejected at the $\propto$ level of significance if the calculated F-ratio is greater than the tabulated F value with 1 and $n-2$ degrees of freedom, that is if

$$F \geq F_{(1-\propto,1,n-2)} \quad . \quad . \quad . \quad (19)$$

Otherwise H_o is accepted.

If H_o is rejected then one may proceed to test some null hypothesis that the trend β has some hypothesized value β_0 (equation11). To do this we note that

$$var(b) = \frac{MSE}{\sum_{i=1}^{n}(i-t)^2} = \frac{MSE}{\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_1}{n}\right)^2}$$

Or using Equations 3 and 4

$$var(b) = \frac{MSE}{\frac{n(n^2-1)}{12}} \quad . \quad . \quad . \quad (20)$$

Or using the value of MSE in table 1 we have that

$$var(b) = \frac{1-b^2}{n-2} \quad . \quad . \quad . \quad (21)$$

Now for sufficiently large $n(n \geq 10)$ (Siegel, 1956)

the statistic $t = \frac{b-\beta_0}{\sqrt{var(b)}} = \frac{b-\beta_0}{\sqrt{\frac{1-b^2}{n-2}}} = \frac{\sqrt{n-2}(b-\beta_0)}{\sqrt{1-b^2}}$ . . . (22)

has approximately the student t distribution with $n-2$ degrees of freedom and may be used to test the null hypothesis of equation 11.

Ho is rejected at the $\propto$ level of significance if $t \geq t_{\left(1-\frac{\propto}{2},n-2\right)}$ . . . (23)

Otherwise Ho is accepted

A problem that often arise in nonparametric regression is converting predicted or estimated ranks $\hat{r}_{iy}$ (equation 10) back to their corresponding predicted or estimated values of the original variables Y. This problem may be solved using interpolation.

Thus, suppose y_s and y_z are the observed sample values of the dependent variable Y with assigned ranks r_{sy} and r_{zy} respectively. Furthermore, suppose the rank predicted for a predicted t[th] value $\hat{y}_t$ of the dependent variable Y is $\hat{r}_{ty}$ which is found to be closer in value to r_s than to r_z, then the predicted or estimated value of Y at point or condition namely $\hat{y}_t$ is calculated as

$$\hat{y}_t = \left(\frac{\hat{r}_{ty} - r_{zy}}{r_{sy} - r_{zy}}\right) y_s + \left(1 - \frac{(\hat{r}_{ty} - r_{zy})}{r_{sy} - r_{zy}}\right) y_z \quad . \quad . \quad . \quad (24)$$

$for\ t = 1,2, \ldots..$

This estimate is sufficiently reliable if the value to be predicted is within or close to the possible range of observed values of the dependent variable.

III. Illustrative Example

A research scientist working with students is interested in performing an experiment that requires the result to be strictly confidential. His laboratory can only be used by one student at a time each day. However the researcher has reasons to believe that students who used the laboratory earlier are passing information to those who used the laboratory later. To check his suspicions the researcher selects a random sample of ten students and arranges for them to perform a certain experiment in the laboratory each on a different day. Each student is cautioned not to tell any person about the nature of the laboratory experiment. The scores earned by the students (out of maximum of a 100 points) and the days they perform the experiment are presented in table 2 below

Table 2: Day (X) of Experiments and Score (Y) by a Random Sample of Students in a Laboratory Class

S/No (i)	Day (X) of Experiment (x_i)	Student Score (y_i)	Rank of x_i $(r_{ix} = i)$	Rank of y_i (r_{iy})	Difference between Ranks $(d_i = i - r_{iy})$	Square of difference between Ranks $d_i^2 = (i - r_{iy})^2$
1	1	60	1	1	0	0
2	2	62	2	2	0	0
3	3	64	3	3	0	0
4	4	67	4	6	-2	4
5	5	65	5	4	1	1
6	6	66	6	5	1	1
7	7	68	7	7	0	0
8	8	70	8	9	-1	1
9	9	71	9	10	-1	1
10	10	69	10	8	2	4
					Total	12.0

Note that days, X, of experiment for this example are time ordered. Hence the null hypothesis to be tested here is that students score in the laboratory experiment is not affected by the day in which the experiment is performed, that is that there is no association between day of experiment and students score in the experiment $(H_0: \beta = \beta_0 = 0;\ equation\ 11)$

To test H_o using the proposed method we first rank the days students performed the experiment from day 1, ranked 1 through day 10, ranked 10 and also their corresponding scores ranked from the lowest (60)assigned the rank of 1 through the highest score (71) assigned the rank of 10.The results of these rankings, the differences between the pairs of ranks d_i and the squares of these differences d_i^2 are shown in Table 2.From the last column of this table we calculate that $\sum_{l=1}^{10} d_i^2 = 12.0$.

Now using this value in Equation 8 with $n = 10$ we obtain an estimate of the trend effect of day on students

score as

$$b = 1 - \frac{6(12.0)}{10(100-1)} = 1 - \frac{72}{990} = 1 - 0.073 = 0.927$$

The size and sign of b a positive sign suggest that day a student performs the experiment in the laboratory and the students score are strongly and directly related with an estimated simple regression coefficient or trend effect of 0.927. Interpreted, this would mean that students score on the average increases by about 92.7 percent for every one additional day students perform the experiment.

To determine whether the trend effect is statistically different from zero we would first obtain the sums of squares shown in the analysis of variance table (table 1) as follows. From equation 12 or Table 1 we have that

$SSR = \frac{990}{12}(0.927)^2 = 82.5 \times 0.927^2 = 70.895$ with 1 degree of freedom. Also from equation 11 we have that

$SST = \frac{990}{12} = 82.50$ with 9 degrees of freedom.

And from equation 14 we have that

$SSE = 82.50 - 70.895 = 11.605$ with 8 degrees of freedom.

These results are summarized in Table 3

Table 3: Analysis of Variance Table to Test for Linear Trend in Students score in an Experiment

Sources of Variation	**Sum of Squares (SS)**	**Degree of Freedom (df)**	**Mean Sum of Squares (MS)**	**F-Ratio**	**P-Value**
Trend (Regression)	70.895	1	70.895	48.859	0.0000
Error	11.605	8	1.451		
Total	82.50	9			

We have from Table 3 that $F = 48.859(p - value = 0.0000)$ with 1 and 8 degrees of freedom which is statistically significant. Hence we would reject the null hypothesis $(Ho: \beta = o)$ of no trend effect of day on students Score in the experiment .

Having confirmed his suspicion that students, who perform the experiment earlier may be providing information to students who perform the experiment later, the research scientist may in fact wish to hypothesize the likely effect this would have on students' scores in the experiment.

Suppose the researcher believes that students performance in the laboratory experiment increases on the average by 50 percent for every one additional day students spend performing the experiment.

That is the value of β_0 in Equation 11 is 0.50.

To test this null hypothesis we have from Equation 22 with

$\sqrt{var(b)} = \sqrt{\frac{1-(0.927)^2}{10-2}} = \sqrt{0.018} = 0.134,$ that is

$$t = \frac{0.929 - 0.50}{0.134} = \frac{0.427}{0.134} = 3.187(p - value = 0.0069)$$

which with 8 degrees of freedom is statistically significant at the 5 percent level $(t_{(0.975;8)} = 2.3060)$. Even with $\beta_{0=}0.60$ the null hypothesis is still rejected $(t = 2.440)$ showing that the arrangement by which students serially perform the experiment by days is strongly favorable to students performing the experiment later. In fact it will seem that 62.0 percent is the least values the research scientist could hypothesize for β for the null hypothesis of equation 9 to be accepted. $(t = 2.291)$ Now from equation 9 we have that $a = \hat{\propto} = 5.5(1 - 0.927) = 0.402.$ Hence the fitted trend line

$(equation\ 10)$ is $= \hat{r}_{iy}0.4 + 0.9\ i.$ Using this equation we estimate the score by a student who performed the experiment at the 11^{th} day with $\hat{r}_{iy} = 10.3\ as\ \hat{y}_{11} = (1.3)(71) + (-0.3)(70) = 71.3$ percent.

IV. Conclusion

This paper has presented and discussed a nonparametric statistical method for estimating 'Trend' effects in data converted to ranks, using nonparametric regression procedures. The estimated effect is similar in its calculation to the Spearman's rank correlation coefficient between two rank samples. An analysis of variance F-test is developed and used for testing the consistency of the data with fitted simple linear nonparametric regression model and for testing the existence of linear trend. If a trend is found to exist then that test is used to test the validity of any hypothesized value for this effect.

References

Gibbons Jean Dickinson (1971): Nonparametric Statistical Inference. McGraw-Hill Book Company. New York.

Neter, J; Wasserman, W; Kutner, M.H. (1983): Applied Linear Regression Models. Richard D. Irwin Inc. Illinois.

Siegel, Sydney (1956): Nonparametric Statistics for the behavioral sciences. McGraw-Hill KOGAKUSHA, LTD. International Student Edition.

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage: http://www.iiste.org

CALL FOR PAPERS

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. There's no deadline for submission. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a **fast** manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar