IMPROVMENTS ON OUTLIER DETECTION

Antonio Simões Silva
Universidade Federal de Viçosa
Departamento de Engenharia Civil
Setor de Engenharia de Agrimensura
Viçosa MG, BRAZIL
CEP 36570-000
Fax. 031-8992302
email: asimoes@mail.ufv.br

RESUMO

Com o advento da automação na obtenção de dados, a fase de depuração feita pelo observador foi eliminada. Na maioria das vezes os dados passam via eletrônica do instrumento para o computador. Em vista disso, torna-se necessário um processo automático para detectar os erros grosseiros. Um método de detecção de erros grosseiros (*outliers*) em observações de redes horizontais é apresentado neste trabalho. O processo é feito simultaneamente com o ajustamento das observações pelo MMQ. É calculada a razão, T_i , entre o resíduo e o sigma do resíduo para cada observação. O valor crítico da distribuição estatitística t é calculado. Será detectado como possível erro grosseiro a observação cujo T_i for maior que t.

ABSTRACT

The trend for capturing and processing data automatically brought the necessity for automatic statistical methods of detecting outliers. This paper shows a method for detecting outliers using t distribution. The ratio, T_i , of a least squares residual to its standard error is computed. The critical value for t is computed as well. If the ratio, T_i , is greater than t the observation is possible to be an outlier.

1. INTRODUCTION

This paper describes a method for detecting outliers. and how tau test subroutine (TTAU) was introduced in a conventional adjustment program.

Values from standard tau test table was compared against values from subroutine TTAU to validate the implementation. Critical values of tau using two different levels of significance will be calculated in order to compare with the approach used in a program which uses a normalised way of detecting outliers.

2. TEST FOR OUTLIERS

In the former adjustment program to check whether an observation is an outlier the ratio of the residual of the observation to its standard error is computed by:

$$T_i = \frac{V_i}{\sigma_{Vi}} \tag{1}$$

If T_i is greater than 2.0, 2.5 and 3.0 (about 5%,1% and 0.3% probability of making a type I error) the i^{th} observation is flagged as a possible outlier. This approach although simple has the disadvantage of considering v_j and s_{vi} as statistically independent. Pope's tau method that applies tau distribution to assess if an observation is an outlier can be used instead.

There is a drawback on tau method. It is based on the assumption that just one observation is affected by a gross error. It is frequently used a pragmatic approach if more than one observation is flagged as outlier. This approach consists of discarding the greatest T_i. After discarding the greatest T_i the adjustment is repeated with the remained n-1 observations. Then a new test is applied. This process is repeated until all outliers are flagged.

The first assumption for tau test is that all observations are normally distributed with:

$$E(l) = AX \qquad (2)$$

which implies that the residuals of the least squares estimation have

$$E(v) = 0 \qquad (3)$$

So

$$Ho: E(v_i) = 0 \qquad (4)$$

Ha: one residual is an outlier

The probability of type I error of the test is usually chosen at 5%. In fact, instead of one test there are n individual tests and a level of significance a_o of the n one-dimensional test is calculated, ignoring the dependency among residuals, as

$$\alpha_o = 1-(1-\alpha)^{\frac{1}{n}} \qquad (5)$$

where,

a_o = computed level of significance

n = number of observations

a = given level of significance.

3. STUDENT'S T-TEST

To determine tau we need to compute student's t-test first as tau is computed from t. POPE, cited in CASPARY(1987), shows that:

$$\tau_{(f)} = \frac{\sqrt{f}.t_{(f-1)}}{\sqrt{f-1+t^2_{(f-1)}}} \qquad (6)$$

where,

f = degree of freedom

t = Student's t-test

The Student's t-test distribution is represented by:

$$f(t) = \frac{\Gamma(\frac{n+1}{2})\,[\frac{1+\ t^2/_n}{n}]^{-(n+1)/_2}}{\sqrt{n\,\pi}\,\Gamma(n/_2)} \qquad (7)$$

For computer programming, there are more appropriate formulae from ABRAMOWITZ (1970). These are mathematical series based on

$$t = 1-[\frac{(1-2\alpha)\pi}{2}] \qquad (8)$$

for one degree of freedom.

For the degree of freedom, simultaneously odd and greater than one the area under the curve of the function is

$$A=\frac{2}{\pi}\{\theta+\sin\theta[\cos\theta+\frac{2}{3}\cos^2\theta+ \\ ......+2.4(f-3)\cos^{(f-2)}\theta]\} \qquad (9)$$

For the degree of freedom, simultaneously odd and greater than one the area under the curve of the function is

$$A=\frac{2}{\pi}\theta \qquad (10)$$

For an even degree of freedom equal to 1, A is

$$A=\sin\theta\{1+\frac{1}{2}\cos^2\theta+\frac{1.3.\cos^4\theta}{2.4}+..... \\ .+\frac{1.3.5.(f-3)\cos^{(f-2)}\theta}{2.4.6...(f-2)} \qquad (11)$$

From equation 9 to 11,

$$\theta=\arctan\frac{t}{\sqrt{f}}$$

where

f = degree of freedom

t = Student's t-test

A is the probability, i.e., the area under the curve of the function.

A subroutine was written for computing the Student's t-test for a given number of observations, by using these formulae.

4. TAU-TEST

Tau test is computed, for each measurement, by:

$$T_i=\frac{e_i^t Wv}{(e_i^t WC_v We_i)^{0.5}} \qquad (12)$$

where

$e_i=[0\,0\,0\,0\ldots1\ldots0\,0]^T$

W = weight matrix

v = vector of residuals (least squares corrections)

C_V = covariance matrix of residuals

For uncorrelated measurements the weight matrix becomes diagonal and (12) reduces to:

$$T_i=\frac{v_i}{\sigma_{v_i}} \qquad (13)$$

The value of T_i for each residual is tested against a critical value computed from the distribution. The procedure for the tau test is as follows.

a) - T_i is calculated for all observations using equation (13)

b) - $t_{(f-1)}$ is computed in function of (f-1) and a $_o$ which is computed by equation (5)

c) - from $t_{(f-1)}$ tau is calculated using the equation (6)

d) - t and T_i are compared. If any value of T_i exceeds t then the relevant observation can be considered as an outlier.

5. THE PROGRAM

The original program was implemented for adjustment and statistical analysis of a plane network by the method of variation of coordinates. A new statistical analysis was introduced which uses tau distribution to find outliers.

Instead of establishing values (2.0, 2.5 and 3.0) as before, a subroutine

(TTAU) computes critical value for tau according to the number of observations, number of unknowns and probability of type I error

A subroutine calculates the statistic T_i for each observation, then these values are compared with that computed by TTAU. If any T_i is greater than tau the observation is flagged as an outlier. .

Some tests have been made using TTAU. Data from SET 1 of a TEST NETWORK and from other adjusted networks have been used for doing those tests. With the TEST NETWORK two levels of significance were carried out, a = 0.01 and a = 0.05, the critical value of tau were:

for a =0.01

t =3.43

for a =0.05

t =3.10

The greatest T_i was 3.05, so we can accept with 5% probability of making a type I error, the null hypothesis that none of the observations contains an outlier. The same we could say at the 0.01 level of significance.

6. CONCLUSION

A subroutine for checking if an observation is an outlier was implemented and worked successfully.

Although it is possible to do this check by using tables the way of adapting subroutines to the adjustment program proved be more convenient.

The tau test was implemented for horizontal network but it can be extended for three or one dimensional network.

7. REFERENCES

ABROMOWITTZ,M.; STEGUN,I., (Ed), *Handbook of Mathematical Functions,* National Bureau of Standards, AMS Series 55, Washington D. C., 1970.

CASPARY, W.F., *Concepts of network and deformation analysis,* Monograph 11, School of Surveying. The University of New South Wales, Kensington Australia. 1987.

COOPER,M.A.R.; CROSS,P.A., Statistical Concepts and their application in Photogrammetry and Surveying. *Photogrammetric Record* 12(71):637-663., 1988

ETTER,D.M., *Problem solving with structured FORTRAN 77.* The Benjamin/Cumming Publishing Company, Inc. California USA., 1984.

KIRCH, A. M., *Introduction to statistics with FORTRAN.* Holt,Rinehart and Winston. New York., 1973.

SILVA, A. S. *Detection of Outliers in Two-Dimensional Network.* First Semester Report IESSG. University of Nottingham., Nottingham ,UK, 1989.