Regression Analysis with t-distribution for the Error

R P Suresh
Indian Institute of Management Kozhikode
Calicut REC PO., Calicut – 673601, Kerala, India
e-mail : rps@iimk.ren.nic.in

1. Introduction

When we perform least squares regression using n observations, for a p-parameter model $y = X\beta + \varepsilon$, we make certain idealized assumptions about the vector of errors ε, namely, that it is distributed $N(0, I\sigma^2)$. In practice departures from these assumptions occur. If our analysis seems to point to the errors having a non-normal distribution, we might consider a robust regression method, particularly, in cases where the error distribution is heavier tailed than the normal, that is, has more probability in tails than the normal. A least square analysis weights each observation equally in getting parameter estimates. Robust methods enable the observations to be weighted unequally. Essentially observations that produce large residuals are down weighted by a robust estimation method. Any specific robust estimator is usable when it provides (exactly or approximately) maximum likelihood estimation of the parameter under the alternative error assumption believed to be true, when the assumption of normality is not true. However, the use of robust regression method involves certain practical difficulties (see Draper and Smith, 1998 p. 580) viz.

a) What function $\rho(u)$ should we use, if we do not know the distribution of the errors in the model?
b) How should we choose the tuning constants in the $\rho(u)$ we choose?
c) How should we choose robust estimator to the scale factor?

Even with some knowledge about the distribution of the errors, and the corresponding choice of the function $\rho(u)$, the difficulties in (b) and (c) will remain, and different values in (b) and different choices of estimators in (c) would lead to different estimation of the coefficients in the regression.

In this paper, we consider the regression study with security returns from the Lincoln National Insurance Corporation as the dependent variable (Y) and the market returns of the Standard & Poor's 500 Index as the explanatory variable (X) reported in Frees(1996). For this data, the residual plot (of least square regression) suggests that the error is non-normal, but can be modeled by a Student's t-distribution. In this paper, we use the modified maximum likelihood estimators of the parameters derived by Rajarshi and Suresh (2000) to fit a linear regression for the security returns data assuming the error distribution to be a Student's t-distribution. We compare the performance of the proposed estimator with that of Least Squares estimator and also that of a robust estimator with Student's t-distribution for the error using a Monte Carlo simulation study.

2. Analysis of Market Returns data

Consider the regression study with security returns from the Lincoln National Insurance Corporation as the dependent variable (Y) and the market returns of the Standard & Poor's 500 Index as the explanatory variable (X) reported in Frees(1996). Here monthly returns over the 5-year period from January 1986 to December 1990 are considered. The Least Squares method yielded the regression equation $Y = -.00214 + 0.973\ X$, with R^2 of 0.354, and estimated standard error of $s = 0.0696$. It may be noted that the standard error of y (unexplained) is 0.0859. The residual plot corresponding to this regression equation indicate that the assumption of normality for the error may not be satisfied, mainly due to the unusual observations in October and November of 1990. Upon investigation, it was observed that the reason for the price to plummet was an announcement by its competitor that it would take a large write-off in their real estate portfolio due to an unprecedented number of mortgage defaults. Thus, these observations cannot be ignored as such events do take place in the market often.

For the above linear regression fit, we tested the error distribution for Normality using Kolmogorov-Smirinov (K-S) goodness of fit test. The test rejects normality with a p-value of 0.0059. We then tried to fit the Student's t-distribution for the error with 5 degrees of freedom. The p-value is 0.0721.

Rajarshi and Suresh (2000) proposed an iterative procedure to estimate the parameters in the linear regression when the error is distributed as Student's t. This procedure is based on the Modified Maximum

Likelihood (MML) estimators of Tiku and Suresh (1992). The MML estimators are known to be asymptotically fully efficient (Bhattacharya, 1985) and almost fully efficient for small sample sizes (Vaughan, 1992). Here we use the above procedure with m=5 to obtain MML estimators for the parameters. The regression fit obtained thus is Y = -.01065576+ 0.9115140 X, with estimate of the standard error to be 0.044631, thus explaining more variation in Y then the Least Squares Estimate.

3. Comparison of Estimators

In this section, we compare the performance of the proposed estimator with that of Least Squares estimator and also that of a robust estimator with Student's t-distribution for the error (with weight function $w(u)=(m+1)/(m+u^2)$ suggested in Draper and Smith, 1998) using a Monte Carlo simulation study. The simulation study was conducted for the linear regression model $y = \beta x + \varepsilon$, for n=10(5)25, β=0.5,1 and degrees of freedom 3(1)5 and with 1000 simulations. We give below, in Table 2.1, the results of the simulation study corresponding to n=20 and β=1.0. Further results will be presented during the Conference, and can also be obtained from the author.

Table 2.1: Comparison of estimates of β

MEAN/MSE	d.f.=3	d.f.=4	d.f.=5
Mean($\hat{\beta}_{LSE}$)	0.9684	1.005719	.990023
MSE($\hat{\beta}_{LSE}$)	.331241	.266675	.237512
Mean($\hat{\beta}_{MML}$)	.980762	1.000173	.996671
MSE($\hat{\beta}_{MML}$)	.191673	.174758	.200786
Mean($\hat{\beta}_{RS}$)	.9781734	1.003676	.995819
MSE($\hat{\beta}_{RS}$)	.191734	.179261	.201941

The Modified Maximum Likelihood Estimators of the regression parameters discussed here provide an alternative method to the Robust Regression, as the observations that produce large residuals are down weighted as in the case of a robust estimation method. However, this procedure is not affected by the problems indicated in (b)-(c) in Introduction, if the data analyst can possibly fit a Student's t-distribution with a specific degree of freedom for the error. Moreover, this procedure performs better than the robust estimation method.

REFERENCES

Bhattacharya, G.K. (1985). The Asymptotics of Maximum Likelihood and related estimators based on Type II censored data. *J. Amer. Statist. Assoc.* 80, 398-404.

Draper, N.R. and Smith, H. (1998) *Applied Regression Analysis*. 3rd Edition. John Wiley and Sons, Inc.

Frees, E.W. (1996) *Data Analysis using Regression Models: The business Perspective.* Prentice Hall.

Rajarshi, M.B. and Suresh, R.P. (2000) Analysis of Regression and Autoregressive models with t-distribution for the error. (Unpublished)

Tiku, M.L. and Suresh, R.P. (1992) A new method of estimation of location and scale parameters. *J. Stat Plann. And Inf.* 30, 281-292.

Tiku, M.L., Tan, W.Y. and Balakrishnan, N. (1986) *Robust Inference.* Marcel Dekker, New York.

Vaughan, D.C. (1992) On the Tiku-Suresh method of estimation. *Commun. Stat. Theory Meth.* 21, 451-469.

RESUME

Dr. R. Padmanabha Suresh obtained his doctor of philosophy (Ph.D) from the University of Pune, Pune, India. He has published several papers in reputed national / international journals. His areas of research and teaching interests are Statistical Inference, Statistical Reliability and Statistical Quality Control.